UNIVERSIDAD DE GUAYAQUIL

FACULTAD DE CIENCIAS MATEMÁTICAS Y FÍSICAS

CARRERA DE INGENIERÍA EN SISTEMAS COMPUTACIONALES

DISEÑO DE UN MODELO PREDICTIVO-ASISTENCIAL DE PACIENTES INFECTADOS POR COVID-19, MEDIANTE UN MODELO SUPERVISADO DE MACHINE LEARNING BASADO EN CRITERIOS DE DERIVACIÓN HOSPITALARIA O AMBULATORIA

PROYECTO DE TITULACIÓN

Previa a la obtención del Título de:

INGENIERO EN SISTEMAS COMPUTACIONALES

AUTORES:

FUENTES MARMOLEJO MELINA DANIELA

MEDINA PARRA WILMER DAVID

TUTOR:

ING. LORENZO CEVALLOS TORRES, M.Sc.

GUAYAQUIL – ECUADOR

2020

REPOSITORIO NACIONAL EN CIENCIAS Y TECNOLOGÍAS

FICHA DE REGISTRO DE TRABAJO DE TITULACIÓN

TÍTULO: *Diseño de un modelo predictivo-asistencial de pacientes infectados por COVID-19, mediante un modelo supervisado de Machine Learning basado en criterios de derivación hospitalaria o ambulatoria*	
AUTOR(ES): Melina Daniela Fuentes Marmolejo Wilmer David Medina Parra	**REVISOR(A):** Ing. Ángela Yanza Montalván, M.Sc.
INSTITUCIÓN: Universidad de Guayaquil	**FACULTAD: Ciencias Matemáticas y Físicas**
CARRERA: Ingeniería en Sistemas Computacionales	
FECHA DE PUBLICACIÓN:	**N° DE PAGS:** 242
AREA TEMÁTICA: *INVESTIGACIÓN*	
PALABRAS CLAVES: *DERIVACIÓN HOSPITALARIA, COVID-19, APRENDIZAJE AUTOMÁTICO, BOSQUES ALEATORIOS, REDES BAYESIANAS.*	

RESUMEN: Un problema derivado de la pandemia del COVID-19 es la falta de una herramienta digital que pueda predecir la intensidad de la gravedad de un paciente enfermo. El presente proyecto consiste en realizar un modelo predictivo asistencial para pacientes infectados por COVID-19, utilizando herramientas de Machine Learning mediante algoritmos de aprendizaje supervisado como Naive Bayes y Random Forest para obtener un criterio sobre derivación hospitalaria o ambulatoria. Entre los principales objetivos específicos se encuentran la extracción de un conjunto de base de datos con la información vinculada al historial médico de los pacientes diagnosticados con COVID-19, para la depuración y construcción de un dataset con las variables relacionadas, y evaluarlas para mejorar la toma de decisiones a partir de un modelo de algoritmo supervisado. La metodología empleada es "Knowledge Discovery in Databases – KDD", la cual se desarrolla en 6 fases: importación y muestreo de datos, calidad de datos, transformación, modelización, evaluación e implementación; sin embargo, esta última fase no se llevará a cabo, en su lugar se realizará un prototipo desarrollado a nivel de Python. Se utilizó la librería sklearn de la herramienta Python 3.5 para el entrenamiento del algoritmo, la herramienta STAT::FIT para las distribuciones estadísticas, y basándose en la sintomatología del paciente los algoritmos arrojaron un porcentaje de precisión (93,5% Random Forest y 95% Naive Bayes), concluyendo que el mejor predictor es el algoritmo de Naive Bayes, también se demostró que existe relación entre ambos algoritmos con respecto a la derivación hospitalaria o ambulatoria mediante el análisis de correlación de Pearson, haciendo que se cumplan las hipótesis planteadas. Al ser útil este prototipo para la toma de decisiones para la respectiva derivación del paciente, los beneficiarios directos son los doctores, dado que obtienen una herramienta que les agilitará la exhaustiva acción de decidir.

N° DE REGISTRO:	**N° DE CLASIFICACIÓN:**	
DIRECCIÓN URL: www.covid19gye.com		
ADJUNTO PDF	**SI** [x]	**NO** [ ]
CONTACTO CON AUTOR(ES):	**Teléfono:** 0995727633 0968478416	**Email:** **melina.fuentesm@ug.edu.ec** **wilmer.medinap@ug.edu.ec**
CONTACTO DE LA INSTITUCIÓN	**Nombre:** Ab. Juan Chávez Atocha	
	Teléfono: 2307729	
	Email: juan.chaveza@ug.edu.ec	

APROBACIÓN DEL TUTOR

En mi calidad de Tutor del Trabajo de Titulación, "DISEÑO DE UN MODELO PREDICTIVO-ASISTENCIAL DE PACIENTES INFECTADOS POR COVID-19, MEDIANTE UN MODELO SUPERVISADO DE MACHINE LEARNING BASADO EN CRITERIOS DE DERIVACIÓN HOSPITALARIA O AMBULATORIA" elaborado por el(la)Sr.(Srta.) MELINA DANIELA FUENTES MARMOLEJO y WILMER DAVID MEDINA PARRA, **estudiantes no titulados** de la Carrera de Ingeniería en Sistemas Computacionales, Facultad de Ciencias Matemáticas y Físicas de la Universidad de Guayaquil, previo a la obtención del Título de Ingeniero(a) en Sistemas Computacionales, me permito declarar que luego de haber orientado, estudiado y revisado, la **apruebo** en todas sus partes.

Atentamente,

_____

Ing. Lorenzo Cevallos Torres, M.Sc.

TUTOR

DEDICATORIA

A mis padres, por siempre apoyarme y aconsejarme en cada paso que doy, tanto en el ámbito personal como profesional, siendo ellos el pilar fundamental para cumplir mis metas.

Melina Daniela Fuentes Marmolejo

Es mi deseo el dedicar este proyecto de titulación a mis padres, quienes en todo momento me han inculcado buenos valores, también, porque me han ayudado durante todo este proceso con su espíritu alentador aportando su granito de arena para que pueda cumplir con todas mis metas

Wilmer David Medina Parra

AGRADECIMIENTO

En primer lugar, agradezco a Dios, en segundo lugar, a mi familia y amigos que me brindaron la motivación para continuar en este proceso y, en tercer lugar, al Ing. Lorenzo Cevallos, M.Sc. por guiarnos con sus conocimientos.

Melina Daniela Fuentes Marmolejo

Agradezco a Dios por darme las fuerzas de seguir adelante cada día y por acompañarme siempre en las buenas y en las malas. A mis padres, por haberme dado una buena educación que han sido la base para mi futuro, todo lo que he logrado es gracias a ellos. Y muchas gracias al Ing. Lorenzo Cevallos, porque nos ha brindado su ayuda en cada duda que hemos tenido en este proceso de titulación. Gracias a todos por eso y por muchos más.

Wilmer David Medina Parra

TRIBUNAL PROYECTO DE TITULACIÓN

Ing. José González Ruiz, M.Sc.
DECANO DE LA FACULTAD
CIENCIAS MATEMÁTICAS Y FÍSICAS

Ing. Lorenzo Cevallos Torres, M.Sc.
DIRECTOR DE LA CARRERA DE
INGENIERÍA EN SISTEMAS
COMPUTACIONALES

Ing. Lorenzo Cevallos Torres, M.Sc.
PROFESOR TUTOR DEL PROYECTO
DE TITULACIÓN

Ing. Ángela Yanza Montalván, M.Sc.
PROFESORA REVISORA DEL
PROYECTO
DE TITULACIÓN

Ab. Juan Chávez Atocha, Esp.
SECRETARIO

DECLARACIÓN EXPRESA

"La responsabilidad del contenido de este Proyecto de Titulación, me corresponden exclusivamente; y el patrimonio intelectual de la misma a la UNIVERSIDAD DE GUAYAQUIL".

MELINA DANIELA FUENTES MARMOLEJO

WILMER DAVID MEDINA PARRA

CESIÓN DE DERECHOS DE AUTOR

Ingeniero

José González Ruiz, M.Sc.

DECANO DE LA FACULTAD DE CIENCIAS MATEMÁTICAS Y FÍSICAS

Presente.

A través de este medio indico a usted que procedo a realizar la entrega de la cesión de derechos de autor en forma libre y voluntaria del trabajo de titulación "**Diseño de un modelo predictivo-asistencial de pacientes infectados por COVID-19, mediante un modelo supervisado de Machine Learning basado en criterios de derivación hospitalaria o ambulatoria**", realizado como requisito previo para la obtención del Título de Ingeniero(a) en Sistemas Computacionales de la Universidad de Guayaquil.

Guayaquil, abril de 2021.

_____

Melina Daniela Fuentes Marmolejo
C.I. N° 0950786988

_____

Wilmer David Medina Parra
C.I. N° 0940665326

UNIVERSIDAD DE GUAYAQUIL

FACULTAD DE CIENCIAS MATEMÁTICAS Y FÍSICAS

CARRERA DE INGENIERÍA EN SISTEMAS COMPUTACIONALES

DISEÑO DE UN MODELO PREDICTIVO-ASISTENCIAL DE PACIENTES INFECTADOS POR COVID-19, MEDIANTE UN MODELO SUPERVISADO DE MACHINE LEARNING BASADO EN CRITERIOS DE DERIVACIÓN HOSPITALARIA O AMBULATORIA

Proyecto de Titulación que se presenta como requisito para optar por el título de

INGENIERO(A) EN SISTEMAS COMPUTACIONALES

Autor(es): Melina Daniela Fuentes Marmolejo

C.I. N° 0950786988

Wilmer David Medina Parra

C.I. N° 0940665326

Tutor: Ing. Lorenzo Cevallos Torres, M.Sc.

Guayaquil, abril de 2021

CERTIFICADO DE ACEPTACIÓN DEL TUTOR

En mi calidad de Tutor del Proyecto de Titulación, nombrado por el Consejo Directivo de la Facultad de Ciencias Matemáticas y Físicas de la Universidad de Guayaquil.

CERTIFICO:

Que he analizado el Proyecto de Titulación presentado por el/la/los estudiantes(s) **MELINA DANIELA FUENTES MARMOLEJO, WILMER DAVID MEDINA PARRA**, como requisito previo para optar por el Título de Ingeniero(a) en Sistemas Computacionales cuyo proyecto es:

DISEÑO DE UN MODELO PREDICTIVO-ASISTENCIAL DE PACIENTES INFECTADOS POR COVID-19, MEDIANTE UN MODELO SUPERVISADO DE MACHINE LEARNING BASADO EN CRITERIOS DE DERIVACIÓN HOSPITALARIA O AMBULATORIA.

Considero aprobado el trabajo en su totalidad.

Presentado por:

	0950786988
Fuentes Marmolejo Melina Daniela	Cédula de identidad N°
	0940665326
Medina Parra Wilmer David	Cédula de identidad N°

Tutor: _____

Ing. Lorenzo Cevallos Torre, M. Sc.

Guayaquil, abril de 2021

UNIVERSIDAD DE GUAYAQUIL
FACULTAD DE CIENCIAS MATEMÁTICAS Y FÍSICAS
CARRERA DE INGENIERÍA EN SISTEMAS COMPUTACIONALES

AUTORIZACIÓN PARA PUBLICACIÓN DE PROYECTO DE TITULACIÓN EN FORMATO DIGITAL

1. Identificación del Proyecto de Titulación

Nombre del Estudiante: Melina Daniela Fuentes Marmolejo	
Dirección: Floresta 2 Mz. 199 V. 9	
Teléfono: 0995727633	**Email:** melina.fuentesm@ug.edu.ec

Nombre del Estudiante: Wilmer David Medina Parra	
Dirección: Trinipuerto, cooperativa los Ángeles 1 Mz. 486 Sl. 7	
Teléfono: 0968478416	**Email:** wilmer.medinap@ug.edu.ec

Facultad: Ciencias Matemáticas y Físicas
Carrera: Ingeniería en Sistemas Computacionales
Proyecto de Titulación al que opta: Investigación
Profesor Tutor: Ing. Lorenzo Cevallos Torres, M.Sc.

Título del Proyecto de Titulación: Diseño de un modelo predictivo-asistencial de pacientes infectados por COVID-19, mediante un modelo supervisado de Machine Learning basado en criterios de derivación hospitalaria o ambulatoria.

Palabras Claves: DERIVACIÓN HOSPITALARIA, COVID-19, MACHINE LEARNING, RANDOM FOREST, NAIVE BAYES.

2. Autorización de Publicación de Versión Electrónica del Proyecto de Titulación
A través de este medio autorizo a la Biblioteca de la Universidad de Guayaquil y a la Facultad de Ciencias Matemáticas y Físicas a publicar la versión electrónica de este Proyecto de Titulación.

Publicación Electrónica:

Inmediata	**X**	Después de 1 año	

Firma Estudiante:

Fuentes Marmolejo Melina Daniela	0950786988 Cédula de identidad N°
Medina Parra Wilmer David	0940665326 Cédula de identidad N°

3. Forma de envío:
El texto del Proyecto de Titulación debe ser enviado en formato Word, como archivo .docx, .RTF o Puf para PC. Las imágenes que la acompañen pueden ser: .gif, .jpg o .TIFF.

DVDROM ☐ CDROM ☐

ÍNDICE GENERAL

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

ABREVIATURAS

ACE 2	Enzima Convertidora de Angiotensina 2
ALT	Alanina Aminotransferasa
ANOVA	Analysis Of Variance
ANS	Aprendizaje No Supervisado
API	Application Programming Interface
ARN	Ácido Ribonucleico
AS	Aprendizaje Supervisado
CAPTCHA	Completely Automated Public Turing test to tell Computers and Humans Apart
CART	Classification and Regression Trees
CCAES	Centro de Coordinación de Alertas y Emergencias Sanitarias
CDC	Centros para el Control y Prevención de Enfermedades
COE-M	Comité de Operaciones de Emergencia Municipal
COE-N	Comité de Operaciones de Emergencia Nacional
COE-P	Comité de Operaciones de Emergencia Provincial
COVID	Coronavirus
CSV	Comma-Separated Values
EM	Expectation Maximization
FCI	Fondo Competitivo de Investigación
GBM	Gradient Boosting Machine
HTML	HyperText Markup Language
IA	Inteligencia Artificial
IBM	International Business Machines
ICA	Independent Component Analysis
IVR	Interactive Voice Response

JSON	JavaScript Object Notation
KDD	Knowledge Discovery in Databases
KNN	K-Nearest Neighbors
LDA	Linear Discrimiant Analysis
MERS	Síndrome respiratorio de oriente medio
MINSAL	Ministerio de Salud
ML	Machine Learning
MSP	Ministerio de Salud Pública
NDA	Normal Discrimiant Analysis
NP	Nondeterministic Polynomial time
OMS	Organización Mundial de la Salud
OPS	Organización Panamericana de la Salud
PCA	Principal Component Analysis
PME	Public Health and Preventive Medicine
RN	Redes Neuronales
RNA	Redes Neuronales Artificiales
ROC	Receiver Operating Characteristic
SARS-CoV 2	Severe Acute Respiratory Syndrome - Coronavirus
SEDISA	Sociedad Española de Directivos de la Salud
SRAS	Síndrome Respiratorio Agudo Severo
SVM	Support Vector Machines
UCI	Unidad de Cuidados Intesivos
UM	Unidad Médica
Visión CEVECE	Centro Estatal de Vigilancia Epidemiológica y Control de Enfermedades

SIMBOLOGÍA

C°	Grados centígrados
d	Distancia
g.l	Grado de libertad
H_0	Hipótesis nula
H_a	Hipótesis alternativa
n	muestra
O2	Oxígeno
p	Significación asintótica (bilateral)
SpO2	Saturación de Oxígeno
X^2	Chi-cuadrado
α	Correlación

UNIVERSIDAD DE GUAYAQUIL
FACULTAD DE CIENCIAS MATEMÁTICAS Y FÍSICAS
CARRERA DE INGENIERÍA EN SISTEMAS COMPUTACIONALES

DISEÑO DE UN MODELO PREDICTIVO-ASISTENCIAL DE PACIENTES INFECTADOS POR COVID-19, MEDIANTE UN MODELO SUPERVISADO DE MACHINE LEARNING BASADO EN CRITERIOS DE DERIVACIÓN HOSPITALARIA O AMBULATORIA

Autor(es): Melina Daniela Fuentes Marmolejo
C.I. N° 0950786988
Wilmer David Medina Parra
C.I. N° 0940665326

Tutor: Ing. Lorenzo Cevallos Torres, M.Sc.

RESUMEN

Un problema derivado de la pandemia del COVID-19 es la falta de una herramienta digital que pueda predecir la intensidad de la gravedad de un paciente enfermo. El presente proyecto consiste en realizar un modelo predictivo asistencial para pacientes infectados por COVID-19, utilizando herramientas de Machine Learning mediante algoritmos de aprendizaje supervisado como Naive Bayes y Random Forest para obtener un criterio sobre derivación hospitalaria o ambulatoria. Entre los principales objetivos específicos se encuentran la extracción de un conjunto de base de datos con la información vinculada al historial médico de los pacientes diagnosticados con COVID-19, para la depuración y construcción de un dataset con las variables relacionadas, y evaluarlas para mejorar la toma de decisiones a partir de un modelo de algoritmo supervisado. La metodología empleada es "Knowledge Discovery in Databases – KDD", la cual se desarrolla en 6 fases: importación y muestreo de datos, calidad de datos, transformación, modelización, evaluación e implementación; sin embargo, esta última fase no se llevará a cabo, en su lugar se realizará un prototipo desarrollado a nivel de Python. Se utilizó la librería sklearn de la herramienta Python 3.5 para el entrenamiento del algoritmo, la herramienta STAT:FIT para las distribuciones estadísticas, y basándose en la sintomatología del paciente los algoritmos arrojaron un porcentaje de precisión (93,5% Random Forest y 95% Naive Bayes), concluyendo que el mejor predictor es el algoritmo de Naive Bayes, también se demostró que existe relación entre ambos algoritmos con respecto a la derivación hospitalaria o ambulatoria mediante el análisis de correlación de Pearson, haciendo que se cumplan las hipótesis planteadas. Al ser útil este prototipo para la toma de decisiones para la respectiva derivación del paciente, los beneficiarios directos son los doctores, dado que obtienen una herramienta que les agilitará la exhaustiva acción de decidir.

Palabras clave: DERIVACIÓN HOSPITALARIA, COVID-19, APRENDIZAJE AUTOMÁTICO, BOSQUES ALEATORIOS, REDES BAYESIANAS.

UNIVERSIDAD DE GUAYAQUIL
FACULTAD DE CIENCIAS MATEMÁTICAS Y FÍSICAS
CARRERA DE INGENIERÍA EN SISTEMAS COMPUTACIONALES

DESIGN OF A PREDICTIVE-CARE MODEL FOR PATIENTS INFECTED BY COVID-19, THROUGH A SUPERVISED MODEL OF MACHINE LEARNING BASED ON HOSPITAL OR OUTPATIENT REFERRAL CRITERIA

Author(s): Melina Daniela Fuentes Marmolejo
C.I. N° 0950786988
Wilmer David Medina Parra
C.I. N° 0940665326

Tutor: Ing. Lorenzo Cevallos Torres, M.Sc.

ABSTRACT

A problem derived from the COVID-19 pandemic is the lack of a digital tool that can predict the severity of a sick patient. The present project consists of carrying out a predictive model of care for patients infected by COVID-19, using Machine Learning tools using supervised learning algorithms such as Naive Bayes and Random Forest to obtain criteria on hospital or outpatient referral. Among the main specific objectives are the extraction of a set of databases with the information related to the medical history of patients diagnosed with COVID-19, for the purification and construction of a dataset with the related variables and evaluate them to improve the decision making based on a supervised algorithm model. The methodology used is "Knowledge Discovery in Databases - KDD", which is developed in 6 phases: import and data sampling, data quality, transformation, modeling, evaluation and implementation; however, this last phase will not be carried out, instead a prototype developed at the Python level will be made. The sklearn library of the Python 3.5 tool was used for the training of the algorithm, the STAT::FIT tool for the statistical distributions, and based on the patient's symptoms, the algorithms yielded a percentage of precision (93.5% Random Forest and 95% Naive Bayes), concluding that the best predictor is the Naive Bayes algorithm, it was also shown that there is a relationship between both algorithms with respect to hospital or outpatient referral by means of Pearson's correlation analysis, making the hypotheses raised. As this prototype is useful for decision-making for the respective referral of the patient, the direct beneficiaries are the doctors, since they obtain a tool that will expedite the exhaustive decision-making action.

Keywords: HOSPITAL REFERRAL, COVID-19, MACHINE LEARNING, RANDOM FOREST, NAIVE BAYES.

INTRODUCCIÓN

El síndrome respiratorio SARS-CoV-2, también conocido como el nuevo virus llamado COVID-19 es una enfermedad que ha causado gran impacto a nivel mundial, formando parte de las enfermedades más letales de la historia y consigo el caos en la atención hospitalaria por la alta demanda de pacientes con sospecha de padecerlo. El propósito de este proyecto es ofrecer una solución a la gestión de los hospitales, derivando a la atención hospitalaria o ambulatoria a los pacientes bajo las condiciones que se encuentren, examinando los factores más importantes y priorizando los casos que puedan presentar un alto índice de mortalidad en el transcurso de la enfermedad.

Según el autor Liang (2020), indica que "we observed similar results when the severe events were defined both by the above objective events and physician evaluation (nine [50%] of 18 patients vs 245 [16%] of 1572 patients)". (pág. 336). Aproximadamente, el 16% de los pacientes corresponden a casos graves y el 5% a enfermedad crítica, siendo la mortalidad en este último grupo de alrededor del 50%. (Liang, Chen, & Guan, 2020)

Los factores de derivación hospitalaria deberían tener en cuenta una evaluación previa de la comorbilidad, la situación de gravedad, la presencia de deterioro cognitivo grave y la dependencia o la necesidad de soporte ventilatorio en pacientes graves. (Blanco & Sanchez, 2020, p. 38)

Al observar que los elementos principales para la derivación hospitalaria o ambulatoria depende de ciertos factores que conlleva el paciente, la disponibilidad de camas en los centros hospitalarios llegó a ser un inconveniente muy grande cuando la pandemia llegó a su punto más alto, teniendo que acomodar otras áreas en los hospitales para cubrir el mayor número de casos posibles. De cara a la apertura de nuevas unidades en un periodo tan corto de tiempo, los principales problemas afectaron a la disponibilidad de infraestructura, personal y material según lo indican

Bardi, Gómez, Candela, Martínez, De Pablo y Pestaña (2020) en un estudio realizado sobre la derivación hospitalaria y la respuesta al virus en España.

Este proyecto hace uso de una de las ramas de la Inteligencia Artificial (IA), la cual es el Machine Learning o en español Aprendizaje Automático, siendo muy útil en diferentes áreas, tales como: educación, financiera, transporte, salud, entre otros; siendo este último tema central para tratar. Recientemente la IA comenzó a establecerse en la medicina para mejorar la atención al paciente con la aceleración de procedimientos, consiguiendo mayor precisión en el diagnóstico de enfermedades, dando oportunidades de ofrecer mejores condiciones de atención médica (Ávila, Mayer, & Quesada, 2020). Se plantea usar dos algoritmos de aprendizaje de máquina supervisado, los cuales son Naive Bayes y Random Forest cuyo objetivo es obtener un modelo predictivo-asistencial para la derivación hospitalaria o ambulatoria para pacientes con COVID-19 y finalmente realizar una comparación para seleccionar el más preciso. A continuación, se segmenta el tema en cuatros capítulos:

En el *Capítulo I* se muestra de manera detallada la descripción de la situación, ubicación, situación conflicto nudos críticos, delimitación, evaluación, causas y consecuencias y formulación con respecto al problema. Además, los objetivos: general y específicos, alcance del proyecto, justificación e importancia y limitaciones de estudio.

En el *Capítulo II* s*e* exhiben los antecedentes del estudio, fundamentación teórica y legal, dando referencia a investigaciones relevantes, hipótesis y las definiciones conceptuales.

En el *Capítulo III* sobresale la parte estadística del proyecto, estableciendo la metodología de la investigación, tipo de investigación, la población y muestra, metodología del desarrollo del prototipo, beneficiarios, entregables del proyecto, propuesta, criterio dado por expertos y los resultados que arroja la investigación por los algoritmos supervisados de Machine Learning.

En el *Capítulo IV s*e presentan las conclusiones del proyecto en base a los objetivos específicos mencionados y las hipótesis, recomendaciones y trabajos futuros.

CAPÍTULO I

PLANTEAMIENTO DEL PROBLEMA

Descripción de la situación problemática

Este proyecto está basado en un FCI ya aprobado, con nombre "Identificación, segmentación y predicción de personas sintomáticas y asintomáticas de afecciones del coronavirus, mediante técnicas de inteligencia artificial y un nuevo modelo de red neuronal artificial convolucional" perteneciente al grupo de Investigación Inteligencia Artificial y Tecnologías de la Información de la Universidad de Guayaquil, de tal manera que este proyecto de titulación se adapta a las necesidades que busca uno de los objetivos propuestos por el FCI ya nombrado.

El nuevo virus COVID-19 generó un evento sin precedentes de tal magnitud en la historia de los últimos 70 años, generando muertes y crisis económica a nivel mundial. "En Ecuador a fecha de noviembre del 2020 de acuerdo con el Ministerio de Salud Pública existen más de 180.000 contagiados y más de 8.000 fallecidos" (MSP, 2020). En Guayaquil con más de 2 millones de habitantes, el primer caso de COVID-19 a finales de febrero inició la primera ola, en los hospitales el grupo de riesgo se despreocupó, despriorizando unidades de cuidados especiales, saturando la capacidad de atención del hospital y aumentando la probabilidad de muerte del paciente. Actualmente, se mejoró el diagnóstico y clasificación de pacientes; sin embargo, en casos muy concretos surgen errores en la asignación de unidades debido a variables volátiles tales como: desconocimiento del historial médico, síntomas poco comunes, y estatus de la capacidad del hospital.

La atención ambulatoria o también llamada a domicilio tiene como objetivo mejorar la gestión hospitalaria. (Martinez, Regalado de los Cobos, & Ruíz, 2020) describen que:

> Una alternativa asistencial es la hospitalización a domicilio consistente en un modelo organizativo capaz de dispensar a pacientes en su propio domicilio un conjunto de actividades y cuidados sanitarios con complejidad, intensidad y duración comparables a los de una hospitalización convencional cuando todavía precisan de una vigilancia activa y una asistencia compleja. (p. 60)

Bajo este contexto, el Machine Learning siendo una rama científica de la Inteligencia Artificial puede automatizar la derivación hospitalaria hacia un centro médico o domiciliar a través de algoritmos de aprendizaje automático entrenados con datos estadísticos recopilados, generando una solución óptima para la correcta toma de decisiones de derivación hospitalaria, aliviando la congestión por turnos en focos de infección como lo son los hospitales y agilizando la atención a pacientes que requieren de un tratamiento más exhaustivo para recuperarse de la enfermedad.

Ubicación del problema en un contexto

El SARS-CoV-2 inicialmente llamado 2019-nCoV es un virus patógeno perteneciente a la familia de los coronavirus, causante de la enfermedad denominada COVID-19; fue descubierto en la ciudad china de Wuhan en diciembre del 2019 y tiene origen zoonótico (Gorbalenya, 2020). El principal reto de los países ante el virus es tratar de contener la cadena de contagios y suplir la demanda de casos, puesto que, al ser un virus de rápida propagación la capacidad hospitalaria es insuficiente, esto sumado al escepticismo poblacional obligó al personal médico a tomar decisiones drásticas para atender a los infectados de manera que muchas vidas se perdieron en la lucha contra el nuevo virus.

A nivel mundial se han contabilizado a fecha de diciembre del 2020; más de 66 millones de contagiados, con la fatídica cantidad de más de un millón de fallecidos (OMS, 2020). Los

países menos afectados por el virus fueron aquellos quienes tomaron medidas oportunas y eficaces, tales como Nueva Zelanda, Vietnam o sin ir muy lejos Uruguay; sin embargo, una gran cantidad de países no supieron tomar medidas adecuadas para enfrentar a la emergente pandemia siendo víctimas de los peores escenarios posibles. Tras un año de su aparición, el virus ha sido estudiado y relativamente controlado, aumentando la esperanza de vida en los infectados, y hasta la llegada de la vacuna considerada la solución global del problema, los países deben tomar decisiones acertadas para minimizar el impacto de la pandemia.

Una característica que ha hecho que esta pandemia sea una de las más letales a nivel mundial es su capacidad de propagación, y el colapso de las unidades médicas (UM) era algo inminente; sin embargo, no se tuvo un correcto plan para manejar la ola de contagios dentro de los hospitales, junto a ello la desesperación e incertidumbre, las malas decisiones fueron desencadenando caos entre las personas contagiadas acudiendo lo antes posible al hospital más cercano, dando paso al aumento de UM para el tratamiento de la enfermedad y saturándose a pocas semanas de haber empezado la pandemia. El principal factor de la mala administración de las UM fue la falta de un modelo predictivo para la derivación hospitalaria, puesto que, entre los infectados, de acuerdo a diversas variables tendrían mayor posibilidad de sobrevivir a la enfermedad que otros.

Los algoritmos predictivos han demostrado ser el método más viable para manejar la toma de decisiones de cualquier índole, no obstante, la eficacia de su implementación radica en el método de implementación. Computacionalmente, los algoritmos predictivos son viables, la capacidad de procesamiento de una computadora permite resultados precisos y claros que ayudan a escoger la mejor opción para situaciones críticas. A través de aprendizaje supervisado se obtienen datos confiables, y utilizando datos de diferentes variables en torno a los contagiados se obtendría un modelo matemático predictivo que sea capaz de ayudar al personal

médico para la asignación de UM, solventando el problema de sobredemanda en los hospitales y mejorando la atención médica para los internados.

Entre los modelos predictivos que abarcan una varianza bastante amplia como la que se obtienen de datos concernientes a los casos de infectados de SARS-CoV-2, se encuentran dos principales: Naive Bayes y Random Forest. Estos algoritmos arrojan resultados mucho más certeros para casos en donde no se sigue una tendencia lineal, haciendo que sean los más indicados para implementar computacionalmente. Utilizando lenguajes de programación orientados al Machine Learning y Data Science se creará y entrenará el algoritmo para obtener los resultados para la toma de decisiones y la implementación generará resultados óptimos para lo que fue diseñado.

En China, la hermeticidad del gobierno hizo que no se revelara la suficiente información para conocer la verdadera realidad del potencial mortífero del virus; tomando situaciones relajadas hasta enero del 2020 en donde el país entero entró en confinamiento, acatando órdenes restrictivas y mejorando la situación meses después, no obstante, en países occidentales, el relajamiento y tardía respuesta ante el virus condujo a una situación espeluznante en tiempos modernos, tomando medidas drásticas en los meses posteriores a la detección de los primeros casos; por ejemplo, en Estados Unidos las medidas se tomaron a la ligera haciendo que la ciudad de Nueva York sea el epicentro de la pandemia; mientras que en Ecuador, de acuerdo al Ministerio de Salud Pública (MSP), se llegó a la cifra de más de 30 mil contagiados en el mes de abril del 2020 donde en la ciudad de Guayaquil la situación era cada vez más catastrófica con miles de muertos debido a la enfermedad y hospitales colapsados debido a la demanda de UM para atender a los contagiados. A nivel global se han detectado 66 millones de casos y más de un millón de personas fallecidas, siendo una de las pandemias más letales del nuevo milenio. (OMS, 2020)

El Machine Learning (ML) es una tecnología relativamente nueva que está teniendo su auge gracias a la aceleración de la capacidad de procesamiento de las computadoras, automatizando y analizando diversos procesos haciendo que mejoren su eficacia y efectividad. El ML ha hecho que las computadoras sean capaces de recomendar una serie de acuerdo con las preferencias, hasta ganar competencias internacionales de Ajedrez, Go, o incluso el famoso juego Jeopardy, demostrando su capacidad de realizar procesos complejos. Hoy en día, las supercomputadoras que implementan ML pueden llegar a hacer simulaciones de escenarios con una precisión increíble, de manera que estudian fenómenos complejos e incluso enfermedades contagiosas siendo una de las tecnologías que tendrá mayor relevancia en el futuro.

A la fecha, existen muchas alternativas para enfrentar la pandemia, y la implementación de tecnologías se convierten en un factor importante, el uso de estas se vuelve vital e imprescindible en el futuro, haciendo que se invierta bastantes recursos en su perfeccionamiento. Así mismo, en combinación con otras alternativas puede realizar acciones mucho más efectivas para el desarrollo de nuevas soluciones de diversas índoles. El empleo de las tecnologías orientada para predicción de eventos es hacia donde se dirige la mayoría de investigación que utilizan ML para así prevenir y prever catástrofes mundiales que afecten no solo a la población humana si no en general preservar la vida misma.

Situación conflicto nudos críticos

El COVID-19 es una enfermedad letal por su nivel de propagación y dificultad para detectar casos a tiempo, obligando a las personas estar precavidas; sin embargo, el 87,9% de los contagiados logran recuperarse de la enfermedad de los cuales 8 de cada 10 personas no necesitaron atención hospitalaria; mientras que en el 4,68% lamentablemente fallecen (MSP, 2020), de este porcentaje una de las principales causas de sucesos fue la negligencia suscitada en el inicio de la pandemia, en donde los pacientes que llegaban a los hospitales no fueron

debidamente atendidos, entre muchas causas la principal era la mala derivación hospitalaria al no detectar cuales casos necesitaban con urgencia una UM, es así como el problema de diagnóstico correcto se hace más evidente teniendo que ser solventada de alguna manera para tomar decisiones correctas.

Actualmente, ya se tienen definidos los protocolos a tomar para derivar a los pacientes; sin embargo, el margen de error puede llegar a ser bastante grande pues el cansancio de los profesionales de la salud puede hacer que se tome decisiones incorrectas, y por el momento para liberar esa carga se utilizan metodologías remotas de atención médica para derivar correctamente a los posibles contagiados. Con cerca de 879.099 llamadas a los servicios de consulta de COVID-19 se ha logrado minimizar la llegada de más personas a los hospitales, no obstante, la eficacia de esta metodología es todavía cuestionada para llevar un correcto manejo de la derivación. (INEC, 2020)

Delimitación del problema

El principal problema que ahonda la pandemia por el virus COVID-19, es que los centros médicos no proceden a un control estricto para los pacientes que llegan infectados. Cuando un paciente llega a dicho sitio con síntomas iniciales, suelen ser internados, esto provoca que los hospitales se llenen y no haya espacio para ningún paciente con síntomas graves.

Se plantea las limitantes del problema en los diferentes elementos de investigación para facilitar el proceso investigativo detallado en la *Tabla 1*:

Tabla 1

Delimitación del problema

Delimitador	Descripción
Campo	Tecnológico (Machine Learning)
Área	Salud humana
Aspecto	Diseño de un modelo predictivo
Tema	Diseño de un modelo predictivo-asistencial de pacientes infectados por COVID-19, mediante un modelo supervisado de Machine Learning basado en criterios de derivación hospitalaria o ambulatoria

Nota: Se muestra la delimitación del problema del presente proyecto, de acuerdo con la problemática investigada. La elaboración y fuente es propia.

Evaluación del Problema

Los aspectos generales de evaluación son:

- **Delimitado:** El presente proyecto consiste en realizar un modelo predictivo asistencial para pacientes infectados por COVID-19, utilizando herramientas de Machine Learning mediante algoritmos de aprendizaje supervisado como Naive Bayes y Random Forest para obtener un criterio para la derivación hospitalaria o ambulatoria. Este modelo permitirá predecir qué tan grave está el paciente para que sea hospitalizado o deba ser enviado a casa para ser atendido de manera ambulatoria.
- **Evidente:** Se evidencia la falta de una herramienta digital que pueda predecir la intensidad de la gravedad de un paciente que llega al hospital infectado con COVID-19, que sirva como apoyo para el médico donde pueda tener la certeza de si el paciente deberá ser atendido inmediatamente en el hospital o deba ser enviado a su domicilio para que pueda recibir ayuda mediante telemedicina (proceso ambulatorio), logrando de mejor manera que el hospital no se quede sin camas en la Unidad de Cuidados Intensivos y evitar así pérdidas humanas.
- **Concreto:** El problema con el COVID-19 es que esta enfermedad actúa de manera inmediata durando aproximadamente 15 días. Pasado este tiempo, el paciente verá

una mejora o empeorará considerablemente de tal manera que podría fallecer. Cuando el paciente llegue al hospital con síntomas dispuestos por la OMS, el modelo predictivo asistencial determinará si el paciente tendrá una atención hospitalaria o ambulatoria. Se observó que más del 50% de los pacientes desarrollan dificultad respiratoria, lo cual dura desde el inicio de la enfermedad hasta la disnea, 8 días. Luego de esto, los pacientes desarrollan síndrome de dificultad respiratoria aguda seguido de shock séptico. (Morales Navarro, 2020)

- **Contextual:** Se pretende ofrecer una mejor atención a los pacientes infectados por COVID-19, mediante el oportuno pronóstico de la gravedad de su enfermedad, a través de técnicas de Machine Learning y minería de datos. De este modo, se podrá evitar pérdidas humanas, haciendo que en el hospital se atienda a pacientes que estén en estado grave. El 4 de abril del 2020, Guayaquil se transformó en la ciudad con más pacientes infectados. Centenares de pacientes estaban hospitalizados con cuadros graves y para esta fecha, la tasa de mortalidad para las mujeres es menor comparada con la de los hombres. (Labarthe, 2020)
- **Factible:** Este modelo pretende compensar la necesidad tecnológica de los especialistas brindándole herramientas necesarias para el alcance de objetivos establecidos de manera eficiente y así reducir errores de diagnóstico. Las tecnologías informáticas son cruciales en la lucha contra la pandemia. Son fundamentales para vencer el virus, pero son las menos conocidas y las más alejadas de lo habitual. Los gobiernos deben velar por las I.A., supercomputación y la big data para convertirlas en aliadas que puedan combatir esta situación de salubridad; lo cual debe predecir en el menor tiempo posible la propagación o realizar seguimiento del virus en personas: ¿cuánta gente morirá? Lograr la implementación

en tecnología de inteligencia artificial que ayude al diagnóstico prolijo de la enfermedad. (Gil Osuna, Arias Romero, & Gil Ozuna, 2020)

- **Identifica los productos esperados:** Al finalizar este proyecto, se presenta un modelo predictivo asistencial para paciente con COVID-19 en el que implica técnicas de Machine Learning y minería de datos, con algoritmos de aprendizaje supervisado. Este modelo permitirá obtener un criterio óptimo en la derivación hospitalaria o ambulatoria para los pacientes.

Causas y consecuencias del problema

En la *Tabla 2* se describen aquellas causas y consecuencias relacionadas al problema y la importante creación de una herramienta tecnológica para la mejora de la gestión hospitalaria. El problema central es identificado en la causa 5: inexistencia de un prototipo para la derivación hospitalaria o ambulatoria, debido a que presenta como consecuencias la pésima gestión hospitalaria.

Tabla 2

Matriz de causas y consecuencias del problema

Causas	Consecuencias
C1. Desconocimiento de síntomas.	E1. Propagación masiva del virus.
C2. Persona adulto mayor.	E2. Mayor riesgo de mortalidad.
C3. Comorbilidades.	E3. Complicaciones en la salud del paciente.
C4. Desconocimiento de protocolos a seguir.	E4. Negligencia médica.
C5. Inexistencia de un prototipo para la derivación hospitalaria o ambulatoria.	E5. Mala gestión hospitalaria, aglomeraciones en hospitales y nula disponibilidad de camas.

Nota: Esta tabla presenta las causas y consecuencias de acuerdo con la problemática estudiada realizándose por medio de "brainstorming" también conocido como tormenta de ideas y consultándose a un experto en el área médica. La elaboración y fuente es propia.

Formulación del problema

La idea estructurada de manera formal que surge a partir de las investigaciones realizadas conjunto a los conflictos del problema e hipótesis cuestionadas se plasma en la siguiente pregunta, la cual será la guía a la correcta observación del tema planteado.

¿Qué efectos tendrán el uso de un algoritmo de aprendizaje de máquina supervisado como ayuda para obtener un prototipo en base a un modelo predictivo-asistencial de pacientes infectados por COVID-19, para la derivación hospitalaria o ambulatoria?

Objetivos del proyecto

Objetivo general

- Diseñar un modelo predictivo de los factores que se relacionan con el ingreso hospitalario frente al ambulatorio mediante algoritmo supervisado de Machine Learning para la obtención un criterio de derivación entre niveles asistenciales durante la pandemia del COVID-19 dirigido a un hospital público de la ciudad de Guayaquil.

Objetivos específicos

1. Recopilar información bibliográfica de las variables relacionadas a la derivación hospitalaria y ambulatoria por síntomas en pacientes con COVID-19 para la determinación de los posibles valores que ayudarán en la definición los algoritmos de aprendizaje supervisado (Random Forest – Naive Bayes).
2. Extraer un conjunto de una base de datos con la información vinculada al historial médico de los pacientes diagnosticado con COVID-19, para la depuración y construcción de un dataset con las variables relacionadas dirigido a un hospital público de la ciudad de Guayaquil.
3. Evaluar las variables relacionadas a la derivación hospitalaria o ambulatoria para el mejoramiento de la toma de decisiones a partir de un modelo de algoritmo supervisado

de Machine Learning con Random Forest y Naive Bayes dirigido a un hospital público de la ciudad de Guayaquil.

Alcance del proyecto

Dentro del alcance del proyecto de investigación denominado “Diseño de un modelo predictivo-asistencial de pacientes infectados por COVID-19, mediante un modelo supervisado de Machine Learning basado en criterios de derivación hospitalaria o ambulatoria”, se han considerado los siguientes aspectos:

Para la recopilación de datos se elaborará un meta-análisis de artículos científicos tales como Science Direct, Taylor & Francis, Springer, IEEE, Elsevier, entre otros. Además, los datos recolectados en diversas instituciones hospitalarias de dominio público o privado de casos relacionados.

En la extracción del conjunto de una base de datos se hará uso de las técnicas de big data, estadística inferencial, no paramétrica y descriptiva.

Realizando la evaluación de las variables relacionadas a la derivación hospitalaria o ambulatoria se utilizarán herramientas open source como la plataforma de desarrollo Python versión 3.5 y herramientas de análisis de datos estadísticos como STAT::FIT. La base de datos será realizada en un archivo Excel con extensión .csv.

El prototipo que genera este proceso investigativo, el cual sirve para validar los resultados de la investigación estarán visualizados al portal www.covid19gye.com, específicamente en la opción “Derivación Hospitalaria”. Se mostrarán las distribuciones de cada variable, formulario a llenar para la predicción, resultados de ambos algoritmos: Random Forest y Naive Bayes, con su respectiva derivación, porcentaje de predicción, matriz de confusión y curva de ROC.

A través de la elaboración de un artículo científico, como parte evidente de proyectos investigativos, se podrá visualizar la extracción de lo más relevante de la documentación;

dando paso a saber los óptimos resultados arrojados por el prototipo y su proceso de creación.

Justificación e importancia

A través de este prototipo de algoritmo de predicción se quiere minimizar los cuellos de botella, cantidad de recursos económicos, médicos y de atención ocasionados por una mala derivación de personas diagnosticadas con COVID-19 en los hospitales cuya capacidad de unidades de atención es limitada, haciendo que se priorice los pacientes con alto grado de predicción de mortalidad.

La importancia de un prototipo que agilite la toma de decisiones ante un caso positivo de COVID-19 recae en la cantidad de pacientes que se recuperaran del virus tras su correcta derivación y la disminución tanto del esfuerzo del personal médico y gastos económicos que se suman tras la correcta hospitalización de un paciente diagnosticado.

Limitaciones del estudio

Durante el desarrollo del proyecto de investigación, se toma en consideración los posibles inconvenientes que podrían llegar a impedir la ejecución de algunos alcances del proyecto. Estas limitaciones se tomarán en consideración para evitar la afectación de su viabilidad.

1. El tiempo estimado para el acceso a la información podría llevar algunos meses debido a que se recogerá información de plataformas para artículos científicos, también la recolección de datos que se obtendrá mediante fuentes de los hospitales para armar la base de datos que se utilizará en el modelo predictivo. Esta situación podría traer dificultades ya que el modelo predictivo debe tener cuanto antes los datos de entrenamiento.
2. Los datos recolectados podrían tener una gran cantidad de información innecesaria, por lo que la manera de resolver este problema se implementa la minería de datos

para así, solo así, obtener los datos concretos para la base de datos que será el dataset que se utilizará en el modelo predictivo.

3. Python utiliza librerías que son utilizadas para la realización de los algoritmos de aprendizaje automático. Si estas librerías están disponibles para otras versiones de Python no empleadas, habría que adaptarlas para su correcto uso.
4. El algoritmo de Machine Learning que se maneja para la solución al proyecto, utilizará los datos que se creará en el dataset. Si el modelo no obtiene la cantidad esperada de aceptación de la predicción, se deberá añadir nuevas variables al dataset para así obtener el porcentaje de predicción deseada.
5. Al momento de obtener la base de datos de pacientes infectados por COVID-19 se presentan algunos inconvenientes. El principal es que al ser un virus relativamente nuevo y existir cierto desconocimiento, la confidencialidad sobre el historial clínico de pacientes con COVID-19 es sumamente alta.

CAPÍTULO II

MARCO TEÓRICO

En el presente capítulo se muestran los antecedentes de estudio y las herramientas a utilizar para la realización del prototipo de un modelo predictivo-asistencial para la derivación hospitalaria o ambulatoria, además se describe mediante una línea de tiempo el inicio del COVID-19 y cómo ha ido evolucionando hasta la actualidad y su gestión hospitalaria, de igual modo toda la parte conceptual relacionada a la enfermedad y a nivel computacional los modelos a emplear.

Antecedentes del estudio

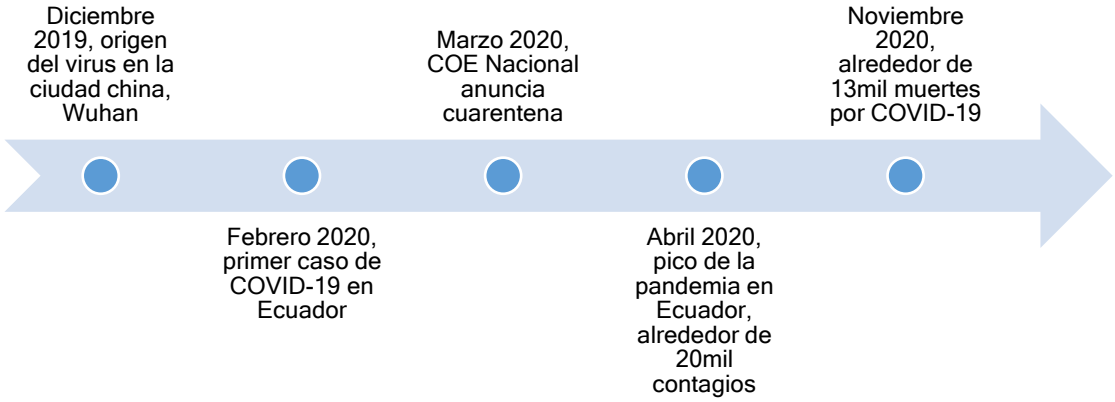
En este subapartado se hace una revisión de los estudios previos, sobre las principales variables de interés, llevados a cabo por otros investigadores en el pasado. Sobre todo, se investigará sobre los resultados producto de estudios empíricos, en la medida de lo posible, bajo contextos más o menos similares al presente estudio. Además, dichos estudios se enmarcarán dentro del contexto global, regional y local. (Castillo, 2018)

Dicho de otra manera, la investigación a realizar debe tener en cuenta el conocimiento previamente construido (Dominguez, 2017), pues esta forma parte de una estructura teórica ya existente. El marco teórico implica analizar teorías, investigaciones, antecedentes que se consideren válidos para el encuadre del estudio, pues la búsqueda y sistematización de aquellas teorías procedentes pueden ayudar en el análisis del problema a investigar. (Dominguez, 2017)

.

Figura 1

Línea de tiempo del COVID-19



Nota: Esta figura presenta el origen del COVID-19 y su llegada y avance en Ecuador. La elaboración es propia y la fuente es proporcionada por las investigaciones realizadas.

Una nueva cepa de coronavirus, el SARS-CoV-2, se detectó por primera vez en diciembre de 2019 en Wuhan, una ciudad de la provincia china de Hubei con una población de 11 millones de habitantes, después de un brote de neumonía sin una causa obvia. El virus se ha extendido ahora a más de 200 países y territorios de todo el mundo, y la Organización Mundial de la Salud (OMS, 2020) lo caracterizó como una pandemia el 11 de marzo de 2020. El SARS-CoV-2 pertenece a una familia de virus de ARN monocatenario conocidos como coronaviridae, un tipo común de virus que afecta a mamíferos, aves y reptiles (Nemecio, 2020). La definición de caso de COVID-19 se basa en los síntomas independientemente del historial de viajes o del contacto con los casos confirmados. El diagnóstico se sospecha en pacientes con tos nueva y continua, fiebre o pérdida o alteración del sentido del gusto o del olfato normal (anosmia). (Huarcaya, 2020)

En un estudio realizado por Wong y Thompson (2020) en el cual indican que, la pandemia de COVID-19 se ha cobrado más de 1,85 millones de vidas en 191 países y regiones desde que se originó en China en diciembre de 2019. De los países asiáticos donde el COVID-19 está aumentando actualmente, Japón es el más preocupante para muchos expertos en enfermedades infecciosas. Las implicaciones clínicas del trastorno hepático pueden variar en

diferentes escenarios clínicos. El Grupo de Trabajo de Asia y el Pacífico para los trastornos hepáticos durante la pandemia COVID-19, en pacientes se puede observar una función hepática anormal o un trastorno hepático, ya sea en forma de hepatitis, colestasis o ambas. En un informe anterior de China, se notificó un aumento de la alanina aminotransferasa (ALT) sérica en 28 (28%) de 99 pacientes con COVID-19 y un aumento de la bilirrubina total en 18 (18%).

El coronavirus no solo tiene a Europa, sino a toda la comunidad mundial bajo su control y es el enemigo común del mundo. Un enemigo que solo podemos derrotar con un enfoque global y una coordinación transfronteriza. (Cotino, 2020)

En un estudio realizado por Mazzucchelli y Dieguez (2020) demuestran que en Europa existe una gran variabilidad en la mortalidad por COVID-19 entre diferentes países. Mientras que algunos países, como Grecia, Bielorrusia o Ucrania, tienen una tasa de mortalidad de menos de 5 casos / 100.000 habitantes, otros países como Bélgica, España o el Reino Unido tienen una tasa de mortalidad de más de 50 casos / 100.000 habitantes. Generalmente se considera que la razón de esta variabilidad es multifactorial (incluidas razones políticas), pero existen pocos estudios que asocien factores relacionados con esta variabilidad. El objetivo del trabajo fue analizar los factores / marcadores de riesgo político que podrían explicar la variabilidad en la mortalidad por COVID-19 entre diferentes países europeos.

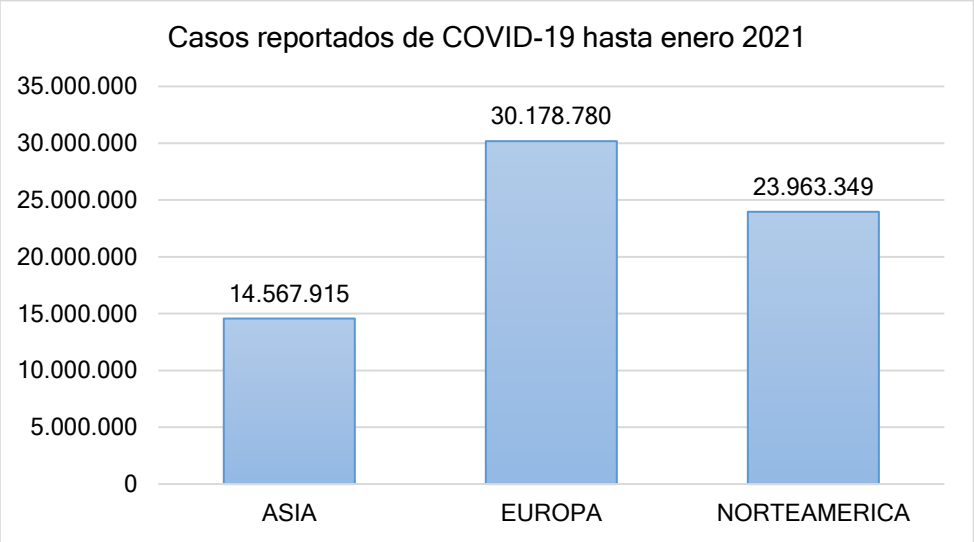
Los primeros casos de la pandemia de coronavirus COVID-19 2019 en América del Norte se notificaron en los Estados Unidos en enero de 2020. El 26 de marzo de 2020, EE.UU. se convirtió en el país con el mayor número de infecciones confirmadas por COVID-19, con más de 82.000 casos. Al 21 de noviembre de 2020, el total de casos de COVID-19 superaba los 13’942.964 con más de 383.084 muertes en total. (Serra, 2020)

En un estudio realizado por Valdés (2020) permitió observar que al cumplirse seis meses del primer contagio por COVID-19 en el mundo, se realizó un estudio de investigación de los resultados de transmisión de la enfermedad por su distribución a nivel mundial como se

muestra en la *Figura 2*, el número de casos reportados en Europa, Asia y Norteamérica, siendo Asia el continente más afectado en los primeros reportes. El objetivo del estudio fue comparar si hay diferencias en los promedios de personas con contagio por COVID-19 entre los diferentes países del mundo, se registra que la mitad de los contagios del mundo se han registrado en Estados Unidos.

Figura 2

Casos reportados de COVID-19 hasta el 15 de enero de 2021



Nota: Esta figura presenta los casos reportados de COVID-19 hasta enero de 2021. La elaboración es propia y la fuente es proporcionada por las investigaciones realizadas.

Tabla 3

Casos reportados de COVID-19.

Descripción	Frecuencia Absoluta	Frecuencia Acumulada	Frecuencia Relativa	Frecuencia Acumulada
Casos reportados de COVID-19 en Asia	14.567.915	14.567.915	21%	0,21
Casos reportados de COVID-19 en Europa	30.178.780	44.746.695	44%	0,65
Casos reportados de COVID-19 en Norteamérica	23.963.349	68.710.044	35%	1,00
TOTAL	68.710.044		100%	

Nota: Esta tabla presenta la frecuencia correspondiente a los casos reportados de COVID-19 hasta enero de 2021. La elaboración es propia y la fuente es proporcionada por las investigaciones realizadas.

La pandemia de COVID-19 en Brasil es parte de la pandemia en curso de la enfermedad por coronavirus 2019 causada por el síndrome respiratorio agudo severo coronavirus 2 (SARS-

CoV-2). Se confirmó que el virus se había propagado a Brasil el 25 de febrero de 2020, cuando un hombre de São Paulo dio positivo por el virus. La enfermedad se había extendido a todos los estados de Brasil el 21 de marzo. (Huarcaya, 2020)

En un artículo realizado por Mejía y Molina (2020) indican que, mega-ciudades, modelos de capitalismo exitoso han sido duramente golpeadas por la pandemia; en Brasil, ciudades emblemáticas como São Paulo y Río de Janeiro donde la realidad ha superado las estadísticas por la cantidad de contagios y muertes imposibles de contar; verdades que permiten vislumbrar el entretejido social y las deformaciones típicas de las sociedades capitalistas en el panorama mundial actual. En Brasil la crisis sanitaria llegó a agravarse considerablemente por las diferencias internas del respectivo gobierno sobre el manejo de la pandemia, la renuncia del ministro de salud dejó al descubierto la real situación inmanejable del gobierno y puso a la población en un estado de completo terror.

Esta nueva enfermedad (COVID-19), ha desnudado de manera cruda y real, la terrible situación sanitaria del Perú: hospitales viejos, falta de materiales, laboratorios especializados, camas, ventiladores, especialistas, y una población geriátrica abandonada, médicos mal remunerados, sin seguro médico, y como nunca, falta de equipos de bioseguridad para combatir a este nuevo flagelo. Esta plaga del coronavirus ha sacado a luz otras verdades como son el papel de la industria y la sociedad en acciones que contaminan ríos lagos y mares, la indiferencia en la sociedad; se ha globalizado el planeta de manera increíble, y cada segundo se sabe cuántos casos nuevos de coronavirus hay y cuántos muertos hay día a día.

Con respecto a la situación del COVID-19, en el Perú, al 24 de marzo del 2020, se tiene 416 casos de coronavirus y 7 muertos, 23 hospitalizados y 9 en Unidad de Cuidados Intensivos (UCI) con ventilación mecánica, indicaba Maguiña (2020) en su artículo.

Peraza (2020) en su trabajo de publicación, indica que los trabajadores del sector de la salud en Ecuador necesitan implementos especiales adecuados para protegerse en su entorno

de trabajo; en la actualidad, estos requisitos adquieren especial relevancia, pues constituyen la garantía para no convertirse ellos mismos en foco de transmisión del COVID-19, poniendo en riesgo su grupo familiar y la comunidad. Ecuador ha sido uno de los tres países de América del Sur con más casos de COVID-19 y 1.564 muertes registradas hasta mediados de mayo de 2020. Cientos de médicos, enfermeras y personal administrativo que trabajan en centros médicos y hospitales se han convertido en casos positivos de la enfermedad, lo que ha dificultado la atención de los pacientes en una emergencia.

En Guayaquil se reportó el primer caso positivo el 29 de febrero. A partir de ese momento, Ecuador inicia el camino en su lucha contra el virus. A partir de esta fecha, se activa el Comité de Operaciones de Emergencia (COE) para atender la pandemia, a nivel nacional (COE-N) y local (provincial, COE-P y cantonal, COE-M). El 12 de marzo se declara la emergencia sanitaria y con esta decisión se suspenden los actos masivos, se ordena el cierre de instituciones educativas y a partir del 16 de marzo se declara el Estado de Excepción en todo el territorio ecuatoriano donde se especifican restricciones de movilidad y limitación de actividades de trabajo. Así, específicamente en Guayaquil (ciudad con una población aproximada de 2,7 millones de habitantes), se reportan casos de contagio comunitario, y la propagación del virus era inevitable, indican (Borja & Cañadas, 2020).

Ferrari, Tonelli y Ghinelli (2020) demuestran que, la pandemia de COVID encontró que todo el sistema de atención médica no estaba adecuadamente preparado e insta a la necesidad de nuevas herramientas para enfrentar esta emergencia clínica y de salud pública sin precedentes. La complejidad clínica del COVID-19 varía desde casos asintomáticos hasta neumonía grave cuya progresión a insuficiencia respiratoria es difícil de predecir. Los métodos de aprendizaje automático (Machine Learning), como los empleados para crear el modelo, han mostrado potencial para producir modelos predictivos que se pueden aplicar para ayudar y

mejorar las decisiones clínicas para una amplia variedad de resultados, y se han utilizado recientemente en respuesta al COVID-19.

Castaneda y Ongkeko (2020) proponen en su trabajo, generar un modelo de diagnóstico más preciso de COVID-19 basado en los síntomas del paciente y los resultados de las pruebas de rutina mediante la aplicación del aprendizaje automático para volver a analizar los datos de COVID-19 de 151 estudios publicados. El objetivo es investigar las correlaciones entre las variables clínicas, agrupar a los pacientes con COVID-19 en subtipos y generar un modelo de clasificación computacional para discriminar entre los pacientes con COVID-19 y los pacientes con influenza basándose únicamente en las variables clínicas. Demostraron que los métodos computacionales entrenados en grandes conjuntos de datos clínicos podrían producir modelos de diagnóstico COVID-19 cada vez más precisos para mitigar el impacto de la falta de pruebas. También se presenta correlaciones de variables clínicas COVID-19 previamente desconocidas y subgrupos clínicos.

Se ha identificado que el aprendizaje automático y la inteligencia artificial son tecnologías prometedoras empleadas por varios proveedores de atención médica, ya que dan como resultado una mejor ampliación, una mayor potencia de procesamiento, confiables e incluso superan a los humanos en tareas específicas de atención médica. (Serra, 2020)

Por lo tanto, las industrias de la salud y los médicos de todo el mundo emplearon diversas tecnologías de ML e IA para afrontar la pandemia de COVID-19 y abordar los desafíos durante el brote. En las industrias médicas, la IA no se aplica para reemplazar las interacciones humanas, sino para brindar apoyo a la toma de decisiones de los médicos. Este documento se centra en la nueva epidemia de COVID-19 y en cómo la tecnología moderna de IA y ML se empleó recientemente para resolver los desafíos durante el estallido. Se mostrarán revisiones exhaustivas de estudios sobre el modelo y la tecnología aplicados para abordar la novedosa pandemia COVID-19.

Por último, es necesario hacer énfasis en la importancia de resolver los problemas de derivación hospitalaria. De acuerdo con el caso de estudio realizado por Capdevilla (2020), publicado en la Revista digital del Programa de Docencia e Investigación en Sistemas de Información Geográfica (PRODISIG) de la Universidad Nacional de Luján, Argentina, se resolvió el problema de derivación hospitalaria. Dada la gran difusión que tuvo el tablero con los casos a nivel mundial desarrollado por la Universidad Johns Hopkins se decidió tomar de base el formato para presentar la información de distribución y evolución del COVID-19 en Argentina. La solución de la derivación y seguimientos de pacientes se presentó a través de la plataforma ArcGIS, con un seguimiento de los casos a domicilio. Al final se concluyó que, a través de dicho tablero informativo digital, es posible que los centros de salud realicen el seguimiento de cada uno de los pacientes derivados para acompañar en su recuperación a los casos que vayan mejorando, solucionando así los diversos problemas de derivación hospitalaria presentados en el país.

Fundamentación teórica

En el presente subapartado se realiza una revisión, en primera instancia, sobre las principales conceptualizaciones en torno a las variables objeto de estudio en la investigación. Así también, se revisa brevemente sus dimensiones y sus principales características. Posteriormente, se hace una revisión de las teorías más relevantes existentes en torno a ellas, se elige la que se relacione de forma más clara con el contexto de la presente investigación y su problemática. De forma ulterior, se desarrolla la perspectiva teórica en base a dicha teoría elegida. Los resultados de este subapartado servirán de guía a lo largo del proceso investigativo, proporcionando luces, sobre todo para la evaluación de los resultados y su respectiva contrastación. (Castillo, 2018)

A continuación, se presenta los principales descriptores de la investigación, los mismos que son: COVID-19, Inteligencia artificial, Machine Learning, Random Forest, Naive Bayes,

Ciencia de datos, Minería de datos, Herramientas de desarrollo de software, Meta-análisis, Hipótesis y Definiciones conceptuales.

COVID-19

Definición

Al inicio el SARS-CoV-2 fue llamado 2019-nCoV, el cual se integra en la familia coronaviridae, siendo este motivo de la famosa enfermedad nombrada como COVID-19. Su origen es zoonótico y radica en Wuhan - China a partir de diciembre del 2019. (Gorbalenya, 2020)

El COVID-19 es un patógeno letal por su nivel de transmisión y complicidad para detectar casos a tiempo, exigiendo a las personas a estar atentas; no obstante, el 87,9% de los contagiados logran recuperarse de la enfermedad de los cuales 8 de cada 10 personas no necesitaron atención hospitalaria; mientras que en el 4,68% lamentablemente fallecen. (MSP, 2020)

Sintomatología

De acuerdo con Carr, Boerner, & Moorman, (2020), los coronavirus son un amplio grupo de virus que tienen la capacidad de causar daños tanto en seres humanos como en animales. Por lo que respecta al ser humano, se conoce que una gran cantidad de coronavirus producen afecciones en las vías respiratorias que van desde el resfriado común hasta patologías de mayor peligrosidad, como es el caso del síndrome respiratorio de Oriente Medio (MERS) y el síndrome respiratorio agudo severo (SRAS), como se citó en (Pérez Abreu , Gómez Tejeda , & Dieguez Guach , 2020)

Cabe mencionar que da lugar a síntomas parecidos a los que son causados por la gripe, incluida fiebre, disnea, tos, fatiga y mialgia. Así mismo, se ha verificado la disminución del sentido del gusto y el olfato (sin que la mucosidad fuese la causa) (Pérez Abreu , Gómez Tejeda

, & Dieguez Guach , 2020). Adicionalmente a lo previamente señalado, existen otros síntomas asociados con la COVID-19, los cuales se exponen en las siguientes líneas:

- **Temperatura:** La fiebre mayor a 37° se ha detectado como uno de los principales síntomas de la COVID-19, así, por ejemplo, en la ciudad de Wuhan, la fiebre era el síntoma más frecuente, en tanto que en otras ciudades se encontró que 43,8% de los internados tenían fiebre en el instante de su entrada al hospital, aunque durante su permanencia en el hospital casi todos la desarrollaron. (CCAES, 2020)
- **Tos:** la presencia de tos (asociada a la neumonía), la cual en la mayor cantidad de casos se inicia con tos seca, es seguida de tos productiva; en algunas ocasiones con disnea, sin datos de hipoxemia, se pueden auscultar crepitantes; otros casos no tienen signos o síntomas clínicos; pese a ello, la tomografía maquinizada evidencia afectaciones a los pulmones que necesitan una atención prioritaria. (Mier, López-Perea, & Calles, 2018)
- **Astenia:** También conocida como fatiga, se conceptualiza como la percepción que tiene un sujeto acerca de la falta de fuerza o cansancio físico o mental, que, a su vez, tiene como efecto la reducción de la capacidad para ejecutar sus tareas rutinarias (Guijarro Sánchez, Royuela García , Guillén González , & Aranburu Aizpiri, 2019)**.** Dicho esto, haciendo referencia de forma específica al COVID-19, se puede expresar lo siguiente: “la astenia se presenta en diferentes grados, desde extrema e invalidante, a moderada o leve”. (SEDISA, 2020)
- **Mialgia:** La mialgia se define como un el dolor muscular que es atendido en consulta médica, un dolor muscular crónico. Estas patologías deben ser bien definidas para conocer cuál es la razón del dolor y por qué la mialgia es persistente y muy fuerte. Numerosos pacientes comenzaron a consultar en Wuhan, provincia de Hubei, China,

a mediados de diciembre de 2019 por una infección respiratoria aguda caracterizada por mialgias y dificultad respiratoria. (Poblete , y otros, 2020)

- **Cefalea:** Es un trastorno del sistema nervioso central caracterizada por la presencia de dolor localizado en la región craneofacial. De acuerdo con Ospina & Volcy (2020) **"**En cuanto a la cefalea, se calcula que la frecuencia en pacientes con COVID-19 puede variar entre 6,5% y 34%". (pág. 27) Adicionalmente, Ospina & Volcy señalan que "Pacientes con antecedentes de cefaleas primarias podrían presentar infección por COVID-19". (pág. 28)
- **Diarrea:** Los síntomas más comunes son fiebre, tos seca y fatiga. Sin embargo, algunos pacientes con COVID-19 desarrollan vómito y diarrea durante el curso de su enfermedad. La diarrea en los pacientes con COVID-19 varía del 2 % al 33 %, y fue el síntoma predominante en el primer paciente diagnosticado en Estados Unidos con COVID-19. (Parra, Flórez, García, & Romero, 2020)
- **Pérdida de olfato:** El mecanismo fisiopatológico por el cual la COVID-19 se asocia a las alteraciones del gusto y del olfato aún no son claras, pero existe evidencia, en el cual se ha informado que el cerebro expresa receptores de la enzima convertidora de angiotensina 2 (ACE 2), receptores dianas del SARS-CoV-2, que se encuentran sobre células gliales y neuronales, lo que los convierte en un objetivo potencial de COVID-19, pudiendo causar daño y muerte neuronal, recorriendo desde las neuronas periféricas a través de la lámina cribosa hasta el bulbo olfatorio. (Huamán & Aparcana , 2020)
- **Pérdida de apetito:** En la misma línea de lo alimentario, se conoce que existe relación entre el gusto y el olfato, y que generalmente estas alteraciones se acompañan; pudiendo verse comprometida la nutrición del paciente al no percibir los sabores de los alimentos, llevando a una conducta restrictiva; es decir,

disminución del apetito al no disfrutar del sabor de la comida, o por el otro lado, a una conducta que no sea favorable para su salud al sobre sazonar los alimentos, esto toma aún mayor peso en los pacientes con diabetes y/o hipertensión arterial. (Sepúlveda , Waissbluth, & González, 2020)

- **Fatiga:** El cansancio o agotamiento, en particular en la o el paciente que fue víctima de COVID-19 y que ya superó la enfermedad, puede llegar a progresar a una circunstancia más compleja que se denomina síndrome de fatiga crónica. Así mismo, según investigaciones recientes, la fatiga es el síntoma más marcado en un/a paciente que sobrevivió al nuevo coronavirus con 53.1%; le sigue la disnea (dificultad para respirar) con un 43.4%, el dolor articular con 27% y el dolor de pecho torácico con 21.7%. (Visión CEVECE, 2020)
- **Pérdida del conocimiento:** En relación con las secuelas neuropsiquiátricas, en casos graves de COVID-19, la respuesta hiper-inflamatoria sistémica podría causar un deterioro cognitivo a largo plazo, como, por ejemplo, deficiencias en la memoria, atención, velocidad de procesamiento y funcionamiento junto con pérdida neuronal difusa, lo que eventualmente, puede originar pérdida del conocimiento. (OPS, 2020)
- **Dolor abdominal:** Los pacientes además pueden presentar manifestaciones gastrointestinales cuya frecuencia varían según la población estudiada y la gravedad del cuadro. En Chile, según un informe del MINSAL al 11 de abril 2020, mostró que el 11% de los casos de SARS-CoV-2 confirmados presentaron algún síntoma gastrointestinal, aproximadamente un 7,3% presentó dolor abdominal. (Díaz & Espino , 2020)
- **Saturación de O2:** Este es un síntoma del COVID-19 descubierto recientemente, la cual indica un decaimiento de los niveles de oxígeno en la sangre. En ese sentido, de acuerdo con la evidencia disponible, los oxímetros de pulso de uso no médico

> tendrían una eficacia comparable a la de los oxímetros de uso médico para descartar la presencia de hipoxemia en pacientes con COVID-19. El valor predictivo negativo para descartar pacientes con hipoxemia (déficit de oxígeno), definida como SpO2 < 94%, es de 99% aunque su precisión disminuye de manera significativa para saturaciones por debajo de 94%. (OPS, 2020, pág. 2)

Personas asintomáticas

Una infección asintomática es aquella en la que una bacteria, virus, hongo o parásito ha invadido el cuerpo, pero aún no ha causado ningún síntoma (como fiebre o tos). El cuerpo puede luchar contra el invasor y es posible que nunca sepan los doctores que el virus estaba dentro del organismo, o puede desarrollar señales de la enfermedad después de una fase asintomática. Dependiendo del patógeno, es posible que el individuo pueda transmitir los gérmenes a otras personas, aunque no tenga síntomas. (Sayampanathan, Heng, & Pin, 2020)

Comorbilidades

Se sabe que el riesgo de COVID-19 es grave si una persona pertenece a la población de edad avanzada y en aquellos con condiciones de salud subyacentes. Conocer la cantidad de individuos con mayor riesgo de COVID-19 grave puede informar el diseño de posibles estrategias de protección, manejo y cuidado de las condiciones crónicas, y orientar la distribución de vacunas para aquellos con mayor riesgo. Además, la herramienta permite a los países estimar el porcentaje de la población sin condiciones, una condición y múltiples condiciones por rangos etarios de 5 años. (CDC, 2020)

Según el Centro de Control de Enfermedades (CDC, 2020) las condiciones de salud subyacentes para COVID-19 grave son condiciones asociadas con un mayor riesgo de COVID-19 grave según las guías publicadas por la OMS, los CDC y Public Health England, incluyendo:

- Enfermedad cardiovascular

- Enfermedad renal crónica
- Tuberculosis (activa)
- Enfermedad respiratoria crónica
- Trastornos neurológicos crónicos
- Enfermedad hepática crónica
- Trastornos de células falciformes
- Diabetes
- Consumo de tabaco fumado
- Cánceres con inmunosupresión directa
- Obesidad severa (Índice la masa corporal ≥40)
- Cánceres sin inmunosupresión directa, pero con posible
- Inmunosupresión causada por el tratamiento
- Hipertensión

Edad

No cabe duda de que la pandemia por COVID-19 además de una emergencia sanitaria sin precedentes ha puesto de manifiesto la capacidad de generación información población de forma masiva y casi en tiempo real. La mayoría de los países han estado proporcionando informes diarios o semanales de las tasas de incidencia de la enfermedad, así como ingresos hospitalarios, en unidades de cuidados intensivos (UCI) y/o mortalidad. Se planteó analizar la información disponible en relación con la edad y mortalidad por COVID-19 con el objetivo aclarar el efecto de cada década de edad. Para esto, se llevó a cabo un meta-análisis con los informes oficiales nacionales de China, Italia, España, Inglaterra y New York que detallaban las tasas de COVID-19 y mortalidad por décadas de edad. (Cordero Fort, 2020)

Los resultados mostraron que la mortalidad por debajo de los 50 años fue muy baja (< 1%) pero se incrementó exponencialmente a partir de esta edad, especialmente a partir de los 60 años. (Cordero Fort, 2020)

Se analizaron un total de 611.1583 sujetos procedentes de 5 registros nacionales, de los que 141.745 (23,2%) tenían ≥ 80 años. El porcentaje de octogenarios fue diferente en los 5 registros, siendo el más bajo en China (3,2%) y el más alto en Reino Unido y el estado de Nueva York. La tasa de mortalidad global fue del 12,10% y varió ampliamente entre países, siendo la más baja en China (3,1%) y la más alta en el Reino Unido (20,8%) y el estado de Nueva York (20,99%). La mortalidad fue < 1,1% en pacientes < 50 años y aumentó exponencialmente a partir de esa edad en los 5 registros nacionales. Como era de esperar, la tasa de mortalidad más alta se observó en pacientes ≥ 80 años, cuyo riesgo fue 6 veces superior al resto. Todos los grupos de edad tuvieron una mortalidad significativamente mayor en comparación con el grupo de edad inmediatamente más joven. El mayor aumento en el riesgo de mortalidad se observó en pacientes de 60 a 69 años en comparación con los de 50 a 59 años (razón de posibilidades 3,13; intervalo de confianza del 95%: 2,61-3,76). (Cordero Fort, 2020)

Inteligencia Artificial

Definición

La inteligencia artificial (IA) es una amplia rama de la informática que se ocupa de la construcción de máquinas inteligentes capaces de realizar tareas que normalmente requieren inteligencia humana. La IA es una ciencia interdisciplinaria con múltiples enfoques, pero los avances en el aprendizaje automático y el aprendizaje profundo están creando un cambio de paradigma en prácticamente todos los sectores de la industria tecnológica. Se refiere a la simulación de la inteligencia humana en máquinas que están programadas para pensar como humanos e imitar sus acciones. El término también se puede aplicar a cualquier máquina que

exhiba rasgos asociados con una mente humana, como el aprendizaje y la resolución de problemas.

La característica ideal de la inteligencia artificial es su capacidad para racionalizar y emprender acciones que tengan las mejores posibilidades de lograr un objetivo específico. Un subconjunto de la inteligencia artificial es el aprendizaje automático, que se refiere al concepto de que los programas informáticos pueden aprender y adaptarse automáticamente a nuevos datos sin la ayuda de humanos. Las técnicas de aprendizaje profundo permiten este aprendizaje automático mediante la absorción de grandes cantidades de datos no estructurados como texto, imágenes o video. (Rouhiainen, 2018)

Cuando la mayoría de la gente escucha el término inteligencia artificial, lo primero en lo que suele pensar es en los robots, eso es porque las películas y novelas de gran presupuesto tejen historias sobre máquinas similares a las humanas que causan estragos en la tierra, pero nada podría estar más lejos de la verdad. La inteligencia artificial se basa en el principio de que la inteligencia humana se puede definir de manera que una máquina pueda imitarla fácilmente y ejecutar tareas, desde las más simples hasta las más complejas. Los objetivos incluyen el aprendizaje, el razonamiento y la percepción.

A medida que avanza la tecnología, los puntos de referencia anteriores que definían la inteligencia artificial se vuelven obsoletos. Por ejemplo, ya no se considera que las máquinas que calculan funciones básicas o reconocen texto a través del reconocimiento óptico de caracteres incorporen inteligencia artificial, ya que esta función ahora se da por sentada como una función informática inherente. La IA evoluciona continuamente para beneficiar a muchas industrias diferentes. Las máquinas están conectadas utilizando un enfoque multidisciplinario basado en matemáticas, informática, lingüística, psicología y más. (Boden, 2017)

Las aplicaciones de la inteligencia artificial son infinitas. La tecnología se puede aplicar a muchos sectores e industrias diferentes. La IA está siendo probada y utilizada en la industria

de la salud para la dosificación de medicamentos y diferentes tratamientos en pacientes y para procedimientos quirúrgicos en el quirófano. Otros ejemplos de máquinas con inteligencia artificial incluyen computadoras que juegan al ajedrez y autos sin conductor. Cada una de estas máquinas debe sopesar las consecuencias de cualquier acción que tomen, ya que cada acción afectará el resultado final.

La inteligencia artificial también tiene aplicaciones en la industria financiera, donde se utiliza para detectar y marcar la actividad en la banca y las finanzas, como el uso inusual de tarjetas de débito y grandes depósitos en cuentas, todo lo cual ayuda al departamento de fraude de un banco. Las aplicaciones para IA también se están utilizando para ayudar a agilizar y facilitar el comercio. Esto se hace facilitando la estimación de la oferta, la demanda y el precio de los valores. (Leyva, 2018)

Técnicas de la Inteligencia Artificial

La inteligencia artificial se puede dividir en diferentes categorías según la capacidad de la máquina para utilizar experiencias pasadas para predecir decisiones, memoria y autoconciencia futuras. IBM ideó Deep Blue, un programa de ajedrez que puede identificar las piezas en el tablero de ajedrez, pero no tiene la memoria para predecir acciones futuras. Este sistema, aunque útil, no se puede adaptar a otra situación. Otro tipo de sistema de IA que usa experiencias pasadas y tiene la ventaja de una memoria limitada para predecir las decisiones. Un ejemplo de este tipo de sistema de IA se puede encontrar en las funciones de toma de decisiones en el caso de los coches autónomos. (Ibargüengoytia, 2018)

Aquí las observaciones ayudan en las acciones que se deben tomar en breve, lo que no se almacena permanentemente ya que las observaciones cambian con frecuencia. Al mismo tiempo con el avance de la tecnología, podría ser posible tener máquinas con un sentido o conciencia donde las máquinas comprendan el estado actual de las cosas, lo que se puede utilizar para inferir lo que se debe hacer, pero tales sistemas no existen.

Automatización y robótica

El propósito de la automatización es lograr que las tareas monótonas y repetitivas sean realizadas por máquinas que también mejoran la productividad y en la obtención de resultados rentables y más eficientes. Muchas organizaciones utilizan el aprendizaje automático, las redes neuronales y los gráficos en la automatización. Dicha automatización puede evitar problemas de fraude durante las transacciones financieras en línea mediante el uso de tecnología CAPTCHA. La automatización robótica de procesos está programada para realizar tareas repetitivas de gran volumen que pueden adaptarse al cambio en diferentes circunstancias. (Salazar, 2018)

Aprendizaje automático (Machine Learning)

> Es una de las aplicaciones de la IA donde las máquinas no están programadas explícitamente para realizar determinadas tareas; más bien, aprenden y mejoran de la experiencia de forma automática. El aprendizaje profundo es un subconjunto del aprendizaje automático basado en redes neuronales artificiales para el análisis predictivo. Hay varios algoritmos de aprendizaje automático, como el aprendizaje no supervisado, el aprendizaje supervisado y el aprendizaje reforzado. (Cerrillo, 2019)

El aprendizaje automático automatizado (ML) es el proceso de automatizar el proceso de aplicar el aprendizaje automático a problemas del mundo real. ML cubre la canalización completa desde el conjunto de datos sin procesar hasta el modelo de aprendizaje automático implementable. ML se propuso como una solución basada en inteligencia artificial para el desafío cada vez mayor de aplicar el aprendizaje automático. El alto grado de automatización en ML permite a los no expertos hacer uso de modelos y técnicas de aprendizaje automático sin necesidad de convertirse primero en un experto en el campo. (Corvalán, 2019, págs. 16 - 20)

La automatización del proceso de aplicación del aprendizaje automático de un extremo a otro ofrece además las ventajas de producir soluciones más simples, una creación más rápida de esas soluciones y modelos que a menudo superan a los modelos diseñados a mano. En una aplicación típica de aprendizaje automático, los profesionales tienen un conjunto de puntos de datos de entrada para entrenar. Es posible que los datos sin procesar no estén en una forma en la que se puedan aplicar todos los algoritmos.

Para que los datos sean aptos para el aprendizaje automático, es posible que un experto tenga que aplicar métodos adecuados de preprocesamiento de datos, ingeniería de características, extracción de características y selección de características. Después de estos pasos, los profesionales deben realizar la selección de algoritmos y la optimización de hiperparámetros para maximizar el rendimiento predictivo de su modelo. Todos estos pasos generan desafíos, que se acumulan hasta convertirse en un obstáculo importante para comenzar con el aprendizaje automático. (Smarandache, 2019)

Ingeniería del conocimiento (Knowledge Engineering)

El objetivo principal de este grupo de investigación es el análisis, diseño, implementación y aplicación de varias técnicas de Inteligencia Artificial, para apoyar el funcionamiento o análisis del comportamiento de sistemas o dominios complejos del mundo real. La investigación se centra en el análisis, diseño, gestión o supervisión de estos dominios, como en el campo de la salud y la medicina, en los procesos y sistemas ambientales, y en el sector industrial y empresarial. Se realizan esfuerzos de investigación específicos en el análisis y desarrollo de agentes inteligentes, comprensión de la dinámica del establecimiento de coaliciones, análisis de la dinámica de la estructura social, construcción de modelos formales de normas y convenciones para el comercio electrónico. (Ramió, 2019)

Lógica difusa (Fuzzy logic)

La lógica difusa es un método de razonamiento que se asemeja al razonamiento humano. El enfoque imita la forma de toma de decisiones en humanos que involucra todas las posibilidades intermedias entre los valores digitales SI y NO (Boden, 2017). El bloque lógico convencional que una computadora puede entender radica en la toma de una entrada precisa y produce una salida definida como VERDADERO o FALSO, que es equivalente al SÍ o NO humano. La lógica difusa trabaja en los niveles de posibilidades de entrada para lograr una salida definida. Ahora, hablando de la implementación de esta lógica:

- Se puede implementar en sistemas con diferentes tamaños y capacidades, como microcontroladores, grandes sistemas en red o basados en estaciones de trabajo.
- Además, se puede implementar en hardware, software o una combinación de ambos.

Redes neuronales artificiales (Artificial Neuronal Networks)

Las redes neuronales artificiales (RNA), generalmente llamadas simplemente redes neuronales (RN), son sistemas informáticos inspirados vagamente en las redes neuronales biológicas que constituyen los cerebros de los animales. Una RNA se basa en una colección de unidades conectadas o nodos llamados neuronas artificiales, que modelan libremente las neuronas en un cerebro biológico. Cada conexión, como las sinapsis en un cerebro biológico, puede transmitir una señal a otras neuronas. (Boden, 2017)

Una neurona artificial que recibe una señal la procesa y puede señalar a las neuronas conectadas a ella. La "señal" en una conexión es un número real, y la salida de cada neurona se calcula mediante alguna función no lineal de la suma de sus entradas. Las conexiones se denominan bordes. Las neuronas y los bordes suelen tener un peso que se ajusta a medida que avanza el aprendizaje. El peso aumenta o disminuye la fuerza de la señal en una conexión.

Las neuronas pueden tener un umbral tal que una señal se envía solo si la señal agregada cruza ese umbral. Normalmente, las neuronas se agregan en capas. Diferentes capas pueden

realizar diferentes transformaciones en sus entradas. Las señales viajan desde la primera capa (la capa de entrada) hasta la última capa (la capa de salida), posiblemente después de atravesar las capas varias veces. (Boden, 2017)

Sistemas reactivos (Reactive System)

Reactive Systems, tal como lo define el Reactive Manifesto, es un conjunto de principios de diseño arquitectónico para construir sistemas modernos que están bien preparados para satisfacer las crecientes demandas a las que se enfrentan las aplicaciones en la actualidad.

Los principios de los sistemas reactivos definitivamente no son nuevos, y se pueden remontar a los años 70 y 80 y el trabajo seminal de Jim Grey y Pat Helland en el Sistema Tandem, y Joe Armstrong y Robert Virding en Erlang. Sin embargo, estas personas se adelantaron a su tiempo y sólo en los últimos 5-10 años la industria de la tecnología se ha visto obligada a repensar las "mejores prácticas" actuales para el desarrollo de sistemas empresariales. Esto significa aprender a aplicar el conocimiento duramente ganado de los principios reactivos en el mundo actual de multinúcleo, computación en la nube e Internet de las cosas. (Cotino, 2020)

La base de un sistema reactivo es el paso de mensajes, que crea un límite temporal entre los componentes que permite desacoplarlos en el tiempo, lo que permite la concurrencia, y el espacio, lo que permite la distribución y la movilidad. Este desacoplamiento es un requisito para el aislamiento total entre componentes, y forma la base tanto para la resiliencia como para la elasticidad.

Sistemas basados en reglas (Rule-Based Systems)

Los sistemas basados en reglas (también conocidos como sistemas de producción o sistemas expertos) son la forma más simple de inteligencia artificial. Un sistema basado en

reglas usa reglas como la representación del conocimiento para el conocimiento codificado en el sistema.

Las definiciones de sistema basado en reglas dependen casi por completo de los sistemas expertos, que son sistemas que imitan el razonamiento del experto humano en la resolución de un problema intensivo en conocimiento. En lugar de representar el conocimiento de una manera declarativa y estática como un conjunto de cosas que son verdaderas, el sistema basado en reglas representa el conocimiento en términos de un conjunto de reglas que dice qué hacer o qué concluir en diferentes situaciones. (Boden, 2017)

Sistemas expertos (Expert Systems)

En inteligencia artificial, un sistema experto es un sistema informático que emula la capacidad de toma de decisiones de un experto humano. Los sistemas expertos están diseñados para resolver problemas complejos mediante el razonamiento a través de cuerpos de conocimiento, representados principalmente como reglas en lugar de a través del código de procedimiento convencional.

Los primeros sistemas expertos se crearon en la década de 1970 y luego proliferaron en la década de 1980. Los sistemas expertos estuvieron entre las primeras formas verdaderamente exitosas de software de inteligencia artificial (IA). Un sistema experto se divide en dos subsistemas: el motor de inferencia y la base de conocimientos. La base de conocimientos representa hechos y reglas. El motor de inferencia aplica las reglas a los hechos conocidos para deducir nuevos hechos. Los motores de inferencia también pueden incluir explicaciones y capacidades de depuración. (Rouhiainen, 2018)

Lingüística computacional

La lingüística computacional es el estudio científico del lenguaje desde una perspectiva computacional. Los lingüistas computacionales están interesados en proporcionar modelos

computacionales de varios tipos de fenómenos lingüísticos. Estos modelos pueden estar "basados en el conocimiento" ("hechos a mano") o "basados en datos" ("estadísticos" o "empíricos").

El trabajo en lingüística computacional está motivado en algunos casos desde una perspectiva científica en el sentido de que uno está tratando de proporcionar una explicación computacional para un fenómeno lingüístico o psicolingüístico particular; y en otros casos, la motivación puede ser más puramente tecnológica en el sentido de que se desea proporcionar un componente funcional de un sistema de habla o lenguaje natural. (Cotino, 2020)

De hecho, el trabajo de los lingüistas computacionales está incorporado en muchos sistemas de trabajo en la actualidad, incluidos los sistemas de reconocimiento de voz, sintetizadores de texto a voz, sistemas de respuesta de voz automatizados, motores de búsqueda web, editores de texto, materiales de instrucción de idiomas, por nombrar solo algunos.

Procesamiento de lenguaje natural (Natural Language Processing)

El procesamiento del lenguaje natural ayuda a las computadoras a comunicarse con los humanos en su propio idioma y escala otras tareas relacionadas con el lenguaje. Por ejemplo, la PNL hace posible que las computadoras lean texto, escuchen el habla, lo interpreten, midan el sentimiento y determinen qué partes son importantes.

Sin embargo, la naturaleza de los lenguajes humanos dificulta el procesamiento del lenguaje natural debido a las reglas que intervienen en el paso de información utilizando el lenguaje natural, y no son fáciles de entender para las computadoras. Entonces, la PNL usa algoritmos para reconocer y abstraer las reglas de los lenguajes naturales, donde los datos no estructurados de los lenguajes humanos se pueden convertir a un formato que la computadora pueda entender. (Manjarrés, 2018)

En PNL, la máquina captura el audio de una conversación humana. Luego se produce la conversación de audio a texto y luego se procesa el texto donde los datos se convierten en

audio. Luego, la máquina usa el audio para responder a los humanos. Las aplicaciones de procesamiento de lenguaje natural se pueden encontrar en aplicaciones IVR (respuesta de voz interactiva) que se utilizan en centros de llamadas, aplicaciones de traducción de idiomas como Google Translate y procesadores de texto como Microsoft Word para verificar la precisión de la gramática en el texto.

Las máquinas de hoy pueden analizar más datos basados en el lenguaje que los humanos, sin fatiga y de forma coherente e imparcial. El lenguaje humano es asombrosamente complejo y diverso. El ser humano se expresa de infinitas formas, tanto verbalmente como por escrito. No solo hay cientos de idiomas y dialectos, sino que dentro de cada idioma hay un conjunto único de reglas, términos y jerga gramaticales y sintácticos. Cuando se escribe, a menudo se hace mal o se abrevian mal las palabras, o se omiten la puntuación. Cuando se habla, se tienen acentos regionales y se murmura, se tartamudea y se toman prestados términos de otros idiomas. (Corvalán, 2019)

Machine Learning

El aprendizaje automático es una rama de la inteligencia artificial (IA) centrada en la creación de aplicaciones que aprenden de los datos y mejoran su precisión con el tiempo sin estar programadas para hacerlo.

En ciencia de datos, un algoritmo es una secuencia de pasos de procesamiento estadístico. En el aprendizaje automático, los algoritmos están 'entrenados' para encontrar patrones y características en cantidades masivas de datos con el fin de tomar decisiones y predicciones basadas en datos nuevos. Cuanto mejor sea el algoritmo, más precisas serán las decisiones y predicciones a medida que procesa más datos. El aprendizaje automático es el concepto de que un programa de computadora puede aprender y adaptarse a nuevos datos sin intervención humana. El aprendizaje automático es un campo de la inteligencia artificial (IA)

que mantiene actualizado los algoritmos integrados de una computadora independientemente de los cambios en la economía mundial. (Carleo, 2019)

Según Carleo (2019) indica que “Machine learning (ML) encompasses a broad range of algorithms and modeling tools used for a vast array of data processing tasks, which has entered most scientific disciplines in recent years”. (pág. 1)

El aprendizaje automático abarca una extensa gama de algoritmos y herramientas de modelado que se usan para infinidad de tareas de procesamiento de datos; ingresando a la mayoría de las áreas científicas en los últimos años. (Carleo, 2019)

El ML se utiliza en diferentes sectores por diversas razones. Los sistemas comerciales se pueden calibrar para identificar nuevas oportunidades de inversión. Las plataformas de marketing y comercio electrónico se pueden ajustar para proporcionar recomendaciones precisas y personalizadas a sus usuarios en función del historial de búsqueda en Internet de los usuarios o de transacciones anteriores. Las instituciones crediticias pueden incorporar el aprendizaje automático para predecir préstamos incobrables y crear un modelo de riesgo crediticio.

Los centros de información pueden utilizar el aprendizaje automático para cubrir grandes cantidades de noticias de todos los rincones del mundo. Los bancos pueden crear herramientas de detección de fraude a partir de técnicas de aprendizaje automático. La incorporación del aprendizaje automático en la era de los conocimientos digitales es interminable a medida que las empresas y los gobiernos se vuelven más conscientes de las oportunidades que presenta el big data.

La calidad de la atención médica es un tema muy importante en el mundo. Decenas de miles de personas mueren cada año, y muchos más sufren de lesiones no mortales debido a errores en el sistema de atención de la salud. Se han dirigido varios enfoques para resolver este problema. Los métodos de aprendizaje automático son bien conocidos por el descubrimiento

de conocimiento. Pueden ayudar a obtener conocimiento (explícito y tácito) a partir de datos y generalizar ese conocimiento a casos nuevos nunca vistos. (Molnar, 2020)

Muchos factores, como las características institucionales, los riesgos del paciente, la distancia de viaje, y las posibilidades de supervivencia y complicaciones deben ser incluidos en la decisión de selección del hospital. Idealmente, cada paciente debe ser tratado individualmente, con el proceso de decisión incluyendo no sólo su condición de ella, sino también sus creencias sobre las compensaciones entre las características deseadas del hospital. Un sistema experto puede ayudar con esta decisión compleja, especialmente cuando se deben considerar numerosos factores.

El algoritmo obtiene conocimiento por sí mismo mediante la construcción de clasificadores de aprendizaje automático a partir de una colección de casos etiquetados. En respuesta a una consulta, el algoritmo da una recomendación personalizada, utilizando un paso de optimización para ayudar al paciente a maximizar la probabilidad de lograr un resultado deseado. En este caso, el hospital recomendado es la solución óptima que maximiza la probabilidad del resultado deseado. Con la formulación adecuada, este sistema experto puede combinar múltiples factores para dar apoyo a la decisión de selección de hospitales a nivel individual. (Raschka, 2019)

Tipos de algoritmos de Machine Learning

Aquí está la lista de algoritmos de aprendizaje automático de uso común. Estos algoritmos se pueden aplicar a casi cualquier problema de datos:

- Regresión lineal
- Regresión logística
- Árbol de decisión
- SVM
- Bayes ingenuos
- K-NN
- K-medias
- Bosque aleatorio
- Algoritmos de reducción de dimensionalidad

- Algoritmos de aumento de gradiente
- GBM

Cada uno de estos algoritmos pertenece a una subdivisión de Machine Learning, se dividen en:

- Aprendizaje supervisado
- Aprendizaje no supervisado

Aprendizaje supervisado

El aprendizaje supervisado es la tarea de aprendizaje automático de aprender una función que asigna una entrada a una salida en función de pares de entrada-salida de ejemplo. Infiere una función a partir de datos de entrenamiento etiquetados que consisten en un conjunto de ejemplos de entrenamiento. En el aprendizaje supervisado, cada ejemplo es un par que consiste en un objeto de entrada (típicamente un vector) y un valor de salida deseado (también llamado la señal de supervisión). Un algoritmo de aprendizaje supervisado analiza los datos de entrenamiento y produce una función inferida, que puede ser utilizada para mapear nuevos ejemplos. Un escenario óptimo permitirá que el algoritmo determine correctamente las etiquetas de clase para instancias invisibles. Esto requiere que el algoritmo de aprendizaje generalice de los datos de entrenamiento a situaciones invisibles de una manera "razonable". (Biamonte, 2018)

Para resolver un problema dado de aprendizaje supervisado, uno tiene que realizar los siguientes pasos:

- **Determinar el tipo de ejemplos de formación.** Antes de hacer cualquier otra cosa, el usuario debe decidir qué tipo de datos se va a utilizar como un conjunto de entrenamiento. En el caso del análisis de escritura a mano, por ejemplo, esto podría ser un solo carácter manuscrito, una palabra escrita a mano completa, o una línea entera de escritura a mano.

- **Reunir un conjunto de entrenamiento.** El conjunto de entrenamiento debe ser representativo del uso real de la función. Por lo tanto, un conjunto de objetos de entrada se reúne y las salidas correspondientes también se reúnen, ya sea de expertos humanos o de mediciones.
- **Determinar la representación de la característica de entrada de la función aprendida.** La precisión de la función aprendida depende fuertemente de cómo se representa el objeto de entrada. Típicamente, el objeto de entrada se transforma en un vector de características, que contiene una serie de características que son descriptivas del objeto. El número de características no debe ser demasiado grande, debido a la maldición de la dimensionalidad; pero debe contener suficiente información para predecir con precisión la salida. (Liakos, 2019)
- **Determinar la estructura de la función aprendida y el algoritmo de aprendizaje correspondiente.** Por ejemplo, el ingeniero puede optar por utilizar máquinas de vectores de soporte o árboles de decisión.
- **Completar el diseño.** Ejecutar el algoritmo de aprendizaje en el conjunto de entrenamiento reunido. Algunos algoritmos de aprendizaje supervisado requieren que el usuario determine ciertos parámetros de control. Estos parámetros pueden ser ajustados optimizando el rendimiento en un subconjunto (llamado un conjunto de validación) del conjunto de entrenamiento, o a través de validación cruzada.
- **Evaluar la precisión de la función aprendida.** Después del ajuste y el aprendizaje de los parámetros, el rendimiento de la función resultante debe medirse en un conjunto de prueba que esté separado del conjunto de entrenamiento.

Aprendizaje no supervisado

El aprendizaje no supervisado es una técnica de aprendizaje automático en la que los usuarios no necesitan supervisar el modelo. En cambio, permite que el modelo funcione por sí

solo para descubrir patrones e información que antes no se había detectado. Se trata principalmente de los datos sin etiquetar. Los algoritmos de aprendizaje no supervisado permiten a los usuarios realizar tareas de procesamiento más complejas en comparación con el aprendizaje supervisado. Sin embargo, el aprendizaje no supervisado puede ser más impredecible en comparación con otros métodos de aprendizaje naturales. Los algoritmos de aprendizaje no supervisados incluyen agrupamiento, detección de anomalías, redes neuronales, etc.

El aprendizaje no supervisado es un tipo de algoritmo que aprende patrones a partir de datos sin etiquetar. La esperanza es que, a través del mimetismo, la máquina se vea obligada a construir una representación interna compacta de su mundo. A diferencia del aprendizaje supervisado (AS) donde los datos son etiquetados por un humano, por ejemplo, como "coche" o "pez", etc., ANS exhibe una autoorganización que captura patrones como preselecciones neuronales o densidades de probabilidad. Los otros niveles en el espectro de supervisión son el aprendizaje por refuerzo, donde a la máquina se le da sólo una puntuación de rendimiento numérica como guía, y el aprendizaje semi-supervisado donde se etiqueta una porción más pequeña de los datos. Dos métodos generales en ANS son las redes neuronales y los métodos probabilísticos. (Dueñas, 2018)

Métodos probabilísticos

Dos de los principales métodos utilizados en el aprendizaje no supervisado son componentes principales y análisis de grupos. El análisis de clústeres se utiliza en el aprendizaje no supervisado para agrupar, o segmentar, conjuntos de datos con atributos compartidos con el fin de extrapolar relaciones algorítmicas. El análisis de clústeres es una rama del aprendizaje automático que agrupa los datos que no han sido etiquetados, clasificados o categorizados. En lugar de responder a la retroalimentación, el análisis de grupos identifica

puntos en común en los datos y reacciona en función de la presencia o ausencia de tales puntos en común en cada nuevo dato.

El único requisito para ser llamado una estrategia de aprendizaje sin supervisión es aprender un nuevo espacio de características que capture las características del espacio original maximizando alguna función objetiva o minimizando alguna función de pérdida. Por lo tanto, generar una matriz de covarianza no es un aprendizaje sin supervisión, pero tomar los vectores propios de la matriz de covarianza se debe a que la operación de descomposición propia del álgebra lineal maximiza la varianza; esto se conoce como análisis de componentes principales. De manera similar, tomar la transformación logarítmica de un conjunto de datos no es un aprendizaje no supervisado, sino pasar datos de entrada a través de múltiples funciones sigmoides, mientras se minimiza alguna función de distancia entre los datos generados y resultantes. (Vázquez, 2018)

Una aplicación central del aprendizaje no supervisado está en el campo de la estimación de densidad en estadística, aunque el aprendizaje no supervisado abarca muchos otros dominios que involucran resumir y explicar características de datos. Podría contrastarse con el aprendizaje supervisado diciendo que mientras que el aprendizaje supervisado intenta inferir una distribución de probabilidad condicional.

Método de momentos

Uno de los enfoques estadísticos para el aprendizaje no supervisado es el método de los momentos. En el método de los momentos, los parámetros desconocidos (de interés) en el modelo están relacionados con los momentos de una o más variables aleatorias, y, por lo tanto, estos parámetros desconocidos pueden ser estimados dados los momentos. Los momentos se estiman usualmente a partir de muestras empíricamente. Los momentos básicos son momentos de primer y segundo orden. Para un vector aleatorio, el momento de primer orden es el vector medio, y el momento de segundo orden es la matriz de covarianza (cuando la media es cero).

Los momentos de orden superior generalmente se representan utilizando tensores que son la generalización de matrices a órdenes superiores como matrices multidimensionales. (Arcila, 2017).

En particular, se muestra que el método de los momentos es efectivo en el aprendizaje de los parámetros de los modelos de variables latentes. Los modelos de variables latentes son modelos estadísticos donde además de las variables observadas, también existe un conjunto de variables latentes que no se observan. Un ejemplo muy práctico de modelos de variables latentes en el aprendizaje automático es el modelado de temas que es un modelo estadístico para generar las palabras (variables observadas) en el documento basado en el tema (variable latente) del documento.

En el modelado de temas, las palabras en el documento se generan de acuerdo a diferentes parámetros estadísticos cuando se cambia el tema del documento. Se muestra que el método de momentos (técnicas de descomposición tensorial) recupera consistentemente los parámetros de una gran clase de modelos de variables latentes bajo algunos supuestos. El algoritmo expectation - maximization (EM) es también uno de los métodos más prácticos para el aprendizaje tardío puede atascarse en los óptimos locales y no se garantiza que el algoritmo converja con los verdaderos parámetros desconocidos del modelo. Por el contrario, para el método de los momentos, la convergencia global está garantizada en algunas condiciones. (Campos, 2018)

Clasificación de aprendizaje supervisado

Clasificación

La clasificación es el proceso de reconocer, comprender y agrupar ideas y objetos en categorías preestablecidas o "subpoblaciones". Usando conjuntos de datos de entrenamiento pre-categorizados, los programas de aprendizaje automático utilizan una variedad de

algoritmos para clasificar conjuntos de datos futuros en categorías. Los algoritmos de clasificación en el aprendizaje automático utilizan datos de entrenamiento de entrada para predecir la probabilidad de que los datos posteriores se incluyan en una de las categorías predeterminadas. Uno de los usos más comunes de la clasificación es filtrar correos electrónicos en "spam" o "no spam".

De acuerdo con Kanavati (2020) indica que “Machine Learning is a supervised learning concept that basically categorizes a set of data into classes. The most common problems are voice recognition, face detection, handwriting recognition, document classification, etc. this can be a binary classification problem”. (pág. 8)

En el aprendizaje automático, la clasificación es un concepto de aprendizaje supervisado que básicamente categoriza un conjunto de datos en clases. Los problemas más comunes son: reconocimiento de voz, detección de rostros, reconocimiento de escritura a mano, clasificación de documentos, etc. Puede ser un problema de clasificación binaria o un problema de varias clases. Hay muchos algoritmos para la clasificación en el aprendizaje automático. (Kanavati, Toyokawa, & Momosaki, 2020)

Regresión

Los algoritmos de aprendizaje automático también se pueden dividir en modelo de aprendizaje paramétrico y modelo de aprendizaje no paramétrico. Los algoritmos que tienen supuestos sólidos en el proceso de aprendizaje y que simplifican la función a la forma conocida se conocen como algoritmos paramétricos de aprendizaje automático. La regresión lineal y la regresión logística son ejemplos de algoritmos paramétricos de aprendizaje automático. Los algoritmos de regresión se ocupan de modelar la relación entre variables que se refinan de forma iterativa utilizando una medida de error en las predicciones realizadas por el modelo. La regresión lineal es un enfoque para modelar la relación entre una variable dependiente escalar

y una o más variables explicativas (o variables independientes) denotadas. Las regresiones lineales y logísticas son los principales algoritmos en el modelado predictivo. (Estrada, 2018)

La regresión lineal es una forma popular de analizar datos descritos en un modelo que es de naturaleza lineal. Es un proceso de encontrar la línea recta de ajuste óptimo a través de los puntos de datos dados. Sin embargo, una representación matemática relaciona la respuesta con las variables predictoras. La regresión lineal es un intento de modelar la relación entre dos variables ajustando una ecuación lineal a los datos observados, donde una variable se considera una variable explicativa y la otra una variable dependiente. Por ejemplo, el estadístico puede querer relacionar los pesos de los individuos con sus alturas usando un modelo de regresión lineal.

Una regresión lineal simple relaciona dos variables (x e y) con una ecuación de línea recta, mientras que una regresión no lineal genera una línea, como si cada valor de y fuera una variable aleatoria. El objetivo de este modelo es hacer que la suma del valor de los cuadrados sea lo más pequeña posible. La regresión lineal es más fácil de usar e interpretar. Sin embargo, si no es posible un buen ajuste con la regresión lineal, entonces se utiliza la regresión no lineal. Las funciones logarítmicas, funciones exponenciales y funciones trigonométricas se encuentran entre los otros métodos de ajuste en una regresión no lineal. (Sandoval, 2018)

Clasificación de aprendizaje no supervisado

Agrupamiento

La agrupación en clústeres es un concepto importante cuando se trata de aprendizaje no supervisado. Se trata principalmente de encontrar una estructura o patrón en una colección de datos sin categorizar. Los algoritmos de agrupación procesarán sus datos y encontrarán clústeres naturales (grupos) si existen en los datos. También puede modificar cuántos clústeres

deben identificar sus algoritmos. Le permite ajustar la granularidad de estos grupos. (Rodríguez N. , 2018)

Agrupamiento K-means

K significa que es un algoritmo de agrupamiento iterativo que le ayuda a encontrar el valor más alto para cada iteración. Inicialmente, se selecciona el número deseado de clústeres. En este método de agrupación, debe agrupar los puntos de datos en K grupos. Una K más grande significa grupos más pequeños con más granularidad de la misma manera. Una K más baja significa grupos más grandes con menos granularidad. La salida del algoritmo es un grupo de "etiquetas". Asigna un punto de datos a uno de los K grupos. En la agrupación de K-means, cada grupo se define creando un centroide para cada grupo. Los centroides son como el corazón del grupo que captura los puntos más cercanos a ellos y los agrega al grupo. (Arcila, 2017)

La agrupación de K-mean define además dos subgrupos:

- **Agrupación aglomerativa:** Este tipo de agrupación de K-medias comienza con un número fijo de agrupaciones. Asigna todos los datos en el número exacto de clústeres. Este método de agrupación no requiere el número de clústeres K como entrada. El proceso de aglomeración comienza formando cada dato como un solo grupo.
- **Dendrograma:** En el método de agrupamiento de dendrogramas, cada nivel representará un posible grupo. La altura del dendrograma muestra el nivel de similitud entre dos clústeres de unión. Cuanto más cerca de la parte inferior del proceso son más similares el grupo que es el hallazgo del grupo del dendrograma que no es natural y en su mayoría subjetivo. (Baviera, 2018)

Reducción de dimensiones

La reducción de dimensionalidad, o reducción de dimensión, es la transformación de datos de un espacio de alta dimensión en un espacio de baja dimensión de modo que la representación de baja dimensión conserva algunas propiedades significativas de los datos originales, idealmente cerca de su dimensión intrínseca. Trabajar en espacios de gran dimensión puede ser indeseable por muchas razones; los datos brutos suelen ser escasos como consecuencia de la dimensionalidad, y el análisis de los datos suele ser difícil de resolver desde el punto de vista computacional. Es común en campos que se ocupan de un gran número de observaciones y / o un gran número de variables, como el procesamiento de señales, el reconocimiento de voz, la neuroinformática y la bioinformática. (Blanc, 2018)

La dimensionalidad es el número de variables, características o características presentes en el conjunto de datos. Estas dimensiones se representan como columnas, y el objetivo es reducir el número de ellas. En la mayoría de los casos, esas columnas están correlacionadas y, por lo tanto, hay alguna información que es redundante que aumenta el ruido del conjunto de datos. Esta información redundante tiene un impacto negativo en el entrenamiento y el rendimiento del modelo de aprendizaje automático y es por eso que el uso de métodos de reducción de dimensionalidad se vuelve de suma importancia. Es una forma muy útil de reducir la complejidad del modelo y evitar el sobreajuste

En la práctica, se construye la matriz de covarianza (y a veces la correlación) de los datos y se calculan los vectores propios de esta matriz. Los eigenvectors (vector propio) que corresponden a los mayores eigenvalues (valor propio) (los componentes principales) ahora pueden ser utilizados para reconstruir una gran fracción de la varianza de los datos originales. Además, los primeros vectores propios a menudo se pueden interpretar en términos del comportamiento físico a gran escala del sistema, porque a menudo contribuyen con la gran mayoría de la energía del sistema, especialmente en sistemas de baja dimensión. Aun así, esto

debe probarse caso por caso, ya que no todos los sistemas muestran este comportamiento. El espacio original (con dimensión del número de puntos) se ha reducido (con pérdida de datos, pero con suerte conservando la varianza más importante) al espacio abarcado por unos pocos autovectores. (Fontalvo, 2018)

Algoritmos de clasificación

Naive Bayes

Se basa en el teorema de Bayes con los supuestos de independencia entre predictores; es decir, asume que la presencia de una característica en una clase no está relacionada con ninguna otra característica. Incluso si estas características dependen unas de otras, o de la existencia de las otras características, todas estas propiedades de forma independiente. (Berrar, 2018)

Es un algoritmo de aprendizaje simple que utiliza la regla de Bayes junto con una fuerte suposición de que los atributos son condicionalmente independientes, dada la clase. Si bien este supuesto de independencia a menudo se infringe en la práctica, ofrece una precisión de clasificación competitiva. Junto con su eficiencia computacional y muchas otras características deseables, esto lleva a que Bayes se aplique ampliamente en la práctica.

Naive Bayes proporciona un mecanismo para usar la información en los datos de la muestra para estimar la probabilidad posterior P (y|x) de cada clase y, dado un objeto x. Una vez obtenidas tales estimaciones, son utilizadas para la clasificación u otras aplicaciones de apoyo a la toma de decisiones. Naive Bayes es un algoritmo de clasificación para problemas de clasificación binarios (de dos clases) y de clases múltiples. La técnica es más fácil de entender cuando se describe utilizando valores de entrada binarios o categóricos. (Guerrero, 2018)

Árboles de decisión

Un árbol de decisiones es un mapa de los posibles resultados de una serie de elecciones relacionadas. Permite a un individuo u organización calcular posibles acciones entre sí en función de sus costos, probabilidades y beneficios. Se pueden utilizar para impulsar una discusión informal o para delinear un algoritmo que prediga la mejor opción matemáticamente.

Según el autor Shanahan (2017) indica que, "Typically, a decision tree starts with a single node, which branches out into possible outcomes. Each of those results leads to additional nodes, which branch out into other possibilities. This gives it a tree-like shape". (p.3)

Comúnmente, un árbol de decisiones comienza con un solo nodo, que se extiende mediante ramificaciones en posibles resultados. Cada resultados conduce a nodos adicionales, que se bifurcan en otras posibilidades, esto le da una forma de árbol. (Shanahan, de Sousa, & Marshall, 2017)

Aunque, un conjunto de datos real tendrá muchas más funciones y esto será solo una rama en un árbol mucho más grande, no se puede ignorar la simplicidad de este algoritmo. La importancia de la característica es clara y las relaciones se pueden ver fácilmente. Esta metodología se conoce más comúnmente como árbol de decisión de aprendizaje a partir de datos y el árbol anterior se denomina árbol de clasificación, ya que un ejemplo sería clasificar al pasajero como sobreviviente o muerto. Los árboles de regresión se representan de la misma manera, solo que predicen valores continuos como el precio de una casa. En general, los algoritmos de árboles de decisión se denominan CART o árboles de clasificación y regresión.

Random Forest

Random Forest es un algoritmo robusto de aprendizaje automático que se puede usar para una variedad de tareas, incluidas regresión y clasificación. Es un método de conjunto, lo que significa que un modelo de bosque aleatorio se compone de un gran número de pequeños árboles de decisión, llamados estimadores, que cada uno produce sus propias predicciones. El

modelo de bosque aleatorio combina las predicciones de los estimadores para producir una predicción más precisa. Los clasificadores de árboles de decisión estándar tienen la desventaja de que tienden a sobreajustarse al conjunto de entrenamiento. (German, Vitale, & Castañeda, 2019)

El diseño de conjunto del bosque aleatorio permite que el bosque aleatorio compense esto y generalice bien a los datos invisibles, incluyendo los datos con valores perdidos. Los bosques aleatorios también son buenos para manejar grandes conjuntos de datos con alta dimensionalidad y tipos de características heterogéneas (por ejemplo, si una columna es categórica y otra numérica). Los bosques aleatorios son muy buenos para los problemas de clasificación, pero son un poco eficientes para los problemas de regresión. En contraste con la regresión lineal, un bosque aleatorio de regresión es incapaz de hacer predicciones fuera del rango de sus datos de entrenamiento. (Camacho, 2018)

Random Forest también son cajas negras: a diferencia de algunos algoritmos de aprendizaje automático más tradicionales, es difícil mirar dentro de un clasificador de bosque aleatorio y entender el razonamiento detrás de sus decisiones. Además, pueden ser lentos de entrenar y ejecutar, y producir archivos de gran tamaño. Debido a que son extremadamente robustos, fáciles de comenzar, buenos para tipos de datos heterogéneos y tienen muy pocos hiperparámetros, los bosques aleatorios son a menudo el primer puerto de escala de un científico de datos al desarrollar un nuevo sistema de aprendizaje automático, ya que permiten a los científicos de datos obtener una descripción general rápida de qué tipo de precisión se puede lograr razonablemente en un problema, incluso si la solución final puede no involucrar un bosque aleatorio. (Athey, 2018)

K-NN

Según el autor Lee (2018) indica que el algoritmo de clasificación no puede predecir valores fuera del rango de los datos mostrados, debido que

> K-NN (categorized as a lazy or instance-based learner) is only able to predict values from within the range of the observed data; that is, it can only predict from experience, and cannot predict wild, or geologically unreason-able values. However, this means K-NN also cannot predict values outside the range of sampled data. (pág. 3)

El algoritmo K-NN es uno de los algoritmos de clasificación más simples y se utiliza para identificar los puntos de datos que se separan en varias clases para predecir la clasificación de un nuevo punto de muestra. K-NN es un algoritmo de aprendizaje perezoso no paramétrico. Clasifica nuevos casos en función de una medida de similitud; es decir, funciones de distancia. (Lee, 2018)

K-Neighbors Neighbors (K-NN) es uno de los algoritmos más simples utilizados en Machine Learning para problemas de regresión y clasificación. La clasificación se realiza por mayoría de votos a sus vecinos. Los datos se asignan a la clase que tiene los vecinos más cercanos. A medida que aumenta el número de vecinos más cercanos, el valor de K, la precisión puede aumentar. En estadística, el algoritmo de los K-vecinos más cercanos (K-NN) es un método de clasificación no paramétrico desarrollado. Se utiliza para la clasificación y la regresión. En ambos casos, la entrada consiste en los K ejemplos de entrenamiento más cercanos en el conjunto de datos. El resultado depende de si se utiliza K-NN para la clasificación o la regresión.

En la clasificación K-NN, la salida es una membresía de clase. Un objeto es clasificado por un voto de pluralidad de sus vecinos, con el objeto siendo asignado a la clase más común entre sus K vecinos más cercanos (K es un entero positivo, típicamente pequeño). Si, entonces el objeto es simplemente asignado a la clase de ese único vecino más cercano. En la regresión K-NN, la salida es el valor de propiedad del objeto. Este valor es el promedio de los valores de K-vecinos más cercanos. Es un tipo de clasificación, donde la función sólo se aproxima localmente y todo el cálculo se difiere hasta la evaluación de la función. (Probst, 2018)

Dado que este algoritmo se basa en la distancia para la clasificación, si las características representan diferentes unidades físicas o vienen en escalas muy diferentes entonces la normalización de los datos de entrenamiento puede mejorar su precisión dramáticamente. Tanto para la clasificación como para la regresión, una técnica útil puede ser asignar ponderaciones a las contribuciones de los vecinos, de modo que los vecinos más cercanos contribuyen más al promedio que los más lejanos. Por ejemplo, un esquema de ponderación común consiste en dar a cada vecino un peso de 1 / d, donde d es la distancia.

Regresión logística

La regresión logística es una especie de regresión lineal, pero se usa cuando la variable dependiente no es un número sino otra cosa (por ejemplo, una respuesta "sí / no"). Se llama regresión, pero realiza una clasificación basada en la regresión y clasifica la variable dependiente en cualquiera de las clases. La regresión logística se utiliza para predecir la salida que es binaria. Por ejemplo, si una compañía de tarjetas de crédito construye un modelo para decidir si emitir o no una tarjeta de crédito a un cliente, modelará si el cliente va a “default” o “no default” en su tarjeta. (Valenzuela, 2019)

La función logística, también llamada función sigmoidea, fue desarrollada por los estadísticos para describir las propiedades del crecimiento de la población en ecología, aumentando rápidamente y llegando al máximo de la capacidad de carga del medio ambiente. Es una curva en forma de S que puede tomar cualquier número de valor real y asignarlo a un valor entre 0 y 1, pero nunca exactamente en esos límites. La regresión logística es un modelo estadístico que en su forma básica utiliza una función logística para modelar una variable dependiente binaria, aunque existen muchas extensiones más complejas. (Alba, 2018)

En el análisis de regresión, la regresión logística (o regresión logit) es la estimación de los parámetros de un modelo logístico (una forma de regresión binaria). Matemáticamente, un modelo logístico binario tiene una variable dependiente con dos valores posibles, como pasa /

no pasa, que está representada por una variable indicadora, donde los dos valores están etiquetados como "0" y "1". En el modelo logístico, el logaritmo de las probabilidades para el valor etiquetado "1" es una combinación lineal de una o más variables independientes ("predictores"); cada una de las variables independientes puede ser una variable binaria (dos clases, codificadas por una variable indicadora) o una variable continua (cualquier valor real).

La probabilidad correspondiente del valor etiquetado "1" puede variar entre 0 (ciertamente el valor "0") y 1 (ciertamente el valor "1"), de ahí el etiquetado; la función que convierte log-odds en probabilidad es la función logística, de ahí el nombre. La unidad de medida para la escala log-odds se llama logit, de unidad logística, de ahí los nombres alternativos. También se pueden utilizar modelos análogos con una función sigmoidea diferente en lugar de la función logística, como el modelo probit; la característica definitoria del modelo logístico es que el aumento de una de las variables independientes escala multiplicativamente las probabilidades del resultado dado a una tasa constante, y cada variable independiente tiene su propio parámetro; para una variable dependiente binaria esto generaliza la relación de probabilidades. (Rodríguez A. , 2018)

Redes neuronales

Las redes neuronales artificiales (RNA), generalmente llamadas simplemente redes neuronales, son sistemas informáticos inspirados vagamente en las redes neuronales biológicas que constituyen los cerebros de los animales. Una RN se basa en una colección de unidades conectadas o nodos llamados neuronas artificiales, que modelan libremente las neuronas en un cerebro biológico. Cada conexión, como las sinapsis en un cerebro biológico, puede transmitir una señal a otras neuronas. Una neurona artificial que recibe una señal la procesa y puede señalar a las neuronas conectadas a ella. La "señal" en una conexión es un número real, y la salida de cada neurona se calcula mediante alguna función no lineal de la suma de sus entradas. (Acosta, 2018)

Las conexiones se llaman bordes. Las neuronas y los bordes suelen tener un peso que se ajusta a medida que avanza el aprendizaje. El peso aumenta o disminuye la fuerza de la señal en una conexión. Las neuronas pueden tener un umbral tal que una señal se envía sólo si la señal agregada cruza ese umbral. Normalmente, las neuronas se agregan en capas. Diferentes capas pueden realizar diferentes transformaciones en sus entradas. Las señales viajan desde la primera capa (la capa de entrada), hasta la última capa (la capa de salida), posiblemente después de atravesar las capas varias veces. (Acosta, 2018)

Las redes neuronales aprenden (o se entrenan) procesando ejemplos, cada uno de los cuales contiene una "entrada" y "resultado" conocidos, formando asociaciones ponderadas por probabilidad entre los dos, que se almacenan dentro de la estructura de datos de la propia red. El entrenamiento de una red neuronal a partir de un ejemplo dado generalmente se realiza determinando la diferencia entre la salida procesada de la red (a menudo una predicción) y una salida objetivo. Este es el error. La red entonces ajusta sus asociaciones ponderadas de acuerdo con una regla de aprendizaje y utilizando este valor de error. Los ajustes sucesivos causarán que la red neuronal produzca una salida que sea cada vez más similar a la salida objetivo. Después de un número suficiente de estos ajustes el entrenamiento puede ser terminado en base a ciertos criterios. Esto se conoce como aprendizaje supervisado.

Dichos sistemas "aprenden" a realizar tareas considerando ejemplos, generalmente sin estar programados con reglas específicas de tareas. Por ejemplo, en el reconocimiento de imágenes, pueden aprender a identificar imágenes que contienen gatos analizando imágenes de ejemplo que se han etiquetado manualmente como "gato" o "sin gato" y utilizando los resultados para identificar gatos en otras imágenes. Hacen esto sin ningún conocimiento previo de los gatos, por ejemplo, que tienen pelaje, colas, bigotes y caras de gatos. En cambio, generan automáticamente características de identificación a partir de los ejemplos que procesan. (Mosquera, 2019)

Algoritmos de regresión

Los algoritmos de aprendizaje automático también se pueden dividir en modelo de aprendizaje paramétrico y modelo de aprendizaje no paramétrico. Los algoritmos que tienen supuestos sólidos en el proceso de aprendizaje y que simplifican la función a la forma conocida se conocen como algoritmos paramétricos de aprendizaje automático. La regresión lineal y la regresión logística son ejemplos de algoritmos paramétricos de aprendizaje automático. Los algoritmos de regresión se ocupan de modelar la relación entre variables que se refinan de forma iterativa utilizando una medida de error en las predicciones realizadas por el modelo. La regresión lineal es un enfoque para modelar la relación entre una variable dependiente escalar y, y una o más variables explicativas (o variables independientes) denotadas x. Las regresiones lineales y logísticas son los principales algoritmos en el modelado predictivo. (Estrada, 2018)

La regresión lineal es una forma popular de analizar datos descritos en un modelo que es de naturaleza lineal. Es un proceso de encontrar la línea recta de ajuste óptimo a través de los puntos de datos dados. Sin embargo, una representación matemática relaciona la respuesta con las variables predictoras. La regresión lineal es un intento de modelar la relación entre dos variables ajustando una ecuación lineal a los datos observados, donde una variable se considera una variable explicativa y la otra una variable dependiente. Por ejemplo, el estadístico puede querer relacionar los pesos de los individuos con sus alturas usando un modelo de regresión lineal.

Una regresión lineal simple relaciona dos variables (x, y) con una ecuación de línea recta, mientras que una regresión no lineal genera una línea, como si cada valor de y fuera una variable aleatoria. El objetivo de este modelo es hacer que la suma del valor de los cuadrados sea lo más pequeña posible. La regresión lineal es más fácil de usar e interpretar. Sin embargo, si no es posible un buen ajuste con la regresión lineal, entonces se utiliza la regresión no lineal.

Las funciones logarítmicas, funciones exponenciales y funciones trigonométricas se encuentran entre los otros métodos de ajuste en una regresión no lineal. (Sandoval, 2018)

Lasso regresión

El “LASSO” significa Operador de Selección y Contracción Mínima Absoluta en español. La regresión de LASSO es una técnica de regularización, se utiliza sobre los métodos de regresión para una predicción más precisa. Este modelo utiliza contracción. La contracción es donde los valores de los datos se reducen hacia un punto central como la media. El procedimiento de lazo fomenta modelos simples y dispersos; es decir, modelos con menos parámetros. Este tipo particular de regresión es adecuado para modelos que muestran altos niveles de multicolinealidad o cuando desea automatizar ciertas partes de la selección del modelo, como la selección de variables / eliminación de parámetros. (Acosta, 2018)

Polinomial regresión

La regresión polinomial es una forma de análisis de regresión en la que la relación entre la variable independiente x y la variable dependiente y se modela como un polinomio de enésimo grado en x. La regresión polinomial se ajusta a una relación no lineal entre el valor de x y la media condicional correspondiente de y, denotada E (y | x). Aunque la regresión polinomial ajusta un modelo no lineal a los datos, como problema de estimación estadística es lineal, en el sentido de que la función de regresión E (y | x) es lineal en los parámetros desconocidos que se estiman a partir de los datos. Por esta razón, la regresión polinomial se considera un caso especial de regresión lineal múltiple. Las variables explicativas (independientes) resultantes de la expansión polinomial de las variables de "línea base" se conocen como términos de grado superior. Estas variables también se utilizan en entornos de clasificación. (Alba, 2018)

Lineal regresión

La regresión lineal es un modelo lineal, p. Ej. un modelo que asume una relación lineal entre las variables de entrada (x) y la variable de salida única (y). Más específicamente, esa y se puede calcular a partir de una combinación lineal de las variables de entrada (x). Cuando hay una única variable de entrada (x), el método se denomina regresión lineal simple. Cuando hay múltiples variables de entrada, la literatura de estadística a menudo se refiere al método como regresión lineal múltiple. Se pueden utilizar diferentes técnicas para preparar o entrenar la ecuación de regresión lineal a partir de datos, la más común de las cuales se llama Mínimos Cuadrados Ordinarios. Por lo tanto, es común referirse a un modelo preparado de esta manera como Regresión lineal de mínimos cuadrados ordinarios o simplemente Regresión de mínimos cuadrados. (Arcila, 2017)

Algoritmos de Agrupamiento

Básicamente es un tipo de método de aprendizaje no supervisado. Un método de aprendizaje no supervisado es un método en el que extraemos referencias de conjuntos de datos que consisten en datos de entrada sin respuestas etiquetadas. Generalmente, se utiliza como un proceso para encontrar una estructura significativa, procesos subyacentes explicativos, características generativas y agrupaciones inherentes a un conjunto de ejemplos. La agrupación es la tarea de dividir la población o los puntos de datos en varios grupos, de modo que los puntos de datos de los mismos grupos sean más similares a otros puntos de datos del mismo grupo y diferentes a los puntos de datos de otros grupos. Es básicamente una colección de objetos sobre la base de similitudes y diferencias entre ellos. (Arcila, 2017)

K-Means

El agrupamiento de K-medias es un método de cuantificación de vectores, originalmente a partir del procesamiento de señales, que tiene como objetivo dividir n

observaciones en K grupos en los que cada observación pertenece al grupo con la media más cercana (centros de grupo o centroide de grupo), sirviendo como un prototipo del racimo. Esto da como resultado una partición del espacio de datos en celdas de Voronoi. La agrupación de K-reduce minimiza las varianzas dentro del grupo (distancias euclidianas cuadradas), pero no distancias euclidianas regulares, que sería el problema de Weber más difícil: la media optimiza los errores cuadrados, mientras que sólo la mediana geométrica minimiza las distancias euclidianas. Por ejemplo, se pueden encontrar mejores soluciones euclidianas usando K-medianas y K-medoides. (Boden, 2017)

El problema es computacionalmente difícil (NP-hard); sin embargo, los algoritmos heurísticos eficientes convergen rápidamente a un óptimo local. Suelen ser similares al algoritmo de maximización de expectativas para mezclas de distribuciones gaussianas a través de un enfoque de refinamiento iterativo empleado por el modelado de mezclas K-medias y gaussianas. Ambos utilizan centros de clústeres para modelar los datos; sin embargo, la agrupación de K-medias tiende a encontrar agrupaciones de extensión espacial comparable, mientras que el mecanismo de maximización de expectativas permite que las agrupaciones tengan formas diferentes. (Campos, 2018)

DBSCAN

La agrupación en clústeres basada en la densidad se refiere a métodos de aprendizaje no supervisados que identifican grupos / clústeres distintivos en los datos, basándose en la idea de que un clúster en el espacio de datos es una región contigua de alta densidad de puntos, separada de otros clústeres de este tipo por regiones contiguas de baja densidad de puntos. El agrupamiento espacial de aplicaciones con ruido basado en densidad (DBSCAN) es un algoritmo básico para el agrupamiento en clústeres basado en densidad. Puede descubrir grupos de diferentes formas y tamaños a partir de una gran cantidad de datos, que contienen ruido y valores atípicos. (Vázquez, 2018)

Agglomerative

Antes de sumergirse en los algoritmos aglomerativos, se debe comprender los diferentes conceptos en las técnicas de agrupación. Entonces, primero, observe el concepto de Clustering en Machine Learning. La agrupación es el amplio conjunto de técnicas para encontrar subgrupos o agrupaciones sobre la base de la caracterización de objetos dentro de un conjunto de datos, de modo que los objetos con grupos son similares pero diferentes del objeto de otros grupos. La directriz principal de la agrupación en clústeres es que los datos dentro de un clúster deben ser muy similares entre sí pero muy diferentes de los que están fuera de los clústeres. Existen diferentes tipos de técnicas de agrupamiento como métodos de particionamiento, métodos jerárquicos. (Valenzuela, 2019)

Algoritmos de reducción de dimensiones

La reducción de la dimensionalidad es común en campos que se ocupan de un gran número de observaciones y / o un gran número de variables, como el procesamiento de señales, el reconocimiento de voz, la neuroinformática y la bioinformática. Los métodos se dividen comúnmente en enfoques lineales y no lineales. Los enfoques también se pueden dividir en la selección de características y la extracción de características. La reducción de dimensionalidad se puede utilizar para la reducción de ruido, visualización de datos, análisis de clústeres, o como un paso intermedio para facilitar otros análisis. (Smarandache, 2019)

Análisis de Componente Principal

El análisis de componentes principales (PCA) es una técnica estadística no paramétrica y no supervisada que se utiliza principalmente para la reducción de dimensionalidad en el aprendizaje automático. Una dimensionalidad alta significa que el conjunto de datos tiene una gran cantidad de características. El problema principal asociado con la alta dimensionalidad en

el campo del aprendizaje automático es el sobreajuste del modelo, que reduce la capacidad de generalizar más allá de los ejemplos del conjunto de entrenamiento. (Sandoval, 2018)

La capacidad de generalizar correctamente se vuelve exponencialmente más difícil a medida que aumenta la dimensionalidad del conjunto de datos de entrenamiento, ya que el conjunto de entrenamiento cubre una fracción cada vez menor del espacio de entrada. Los modelos también se vuelven más eficientes a medida que el conjunto de funciones reducido aumenta las tasas de aprendizaje y disminuye los costos de cálculo al eliminar las funciones redundantes. PCA también se puede utilizar para filtrar conjuntos de datos ruidosos, como la compresión de imágenes. El primer componente principal expresa la mayor cantidad de variación. Cada componente adicional expresa menos varianza y más ruido, por lo que representar los datos con un subconjunto más pequeño de componentes principales conserva la señal y descarta el ruido. (Sandoval, 2018)

Análisis de componentes independientes

En el procesamiento de señales, el análisis de componentes independientes (ICA) es un método computacional para separar una señal multivariante en subcomponentes aditivos. Esto se hace asumiendo que los subcomponentes son señales no gaussianas y que son estadísticamente independientes entre sí. ICA es un caso especial de separación de fuente ciega. El análisis de componentes independientes intenta descomponer una señal multivariante en señales independientes no gaussianas. Como ejemplo, el sonido suele ser una señal que se compone de la adición numérica, en cada tiempo t, de señales de varias fuentes. La pregunta entonces es si es posible separar estas fuentes contribuyentes de la señal total observada. Cuando la suposición de independencia estadística es correcta, la separación ICA ciega de una señal mixta da muy buenos resultados. También se utiliza para señales que no se supone que se generen mediante la mezcla con fines de análisis. (Rouhiainen, 2018)

Análisis discriminante lineal

El análisis discriminante lineal (LDA), el análisis discriminante normal (NDA) o el análisis de función discriminante es una generalización de la discriminante lineal de Fisher, un método utilizado en estadísticas y otros campos, para encontrar una combinación lineal de características que caracterizan o separan dos o más clases de objetos o eventos. La combinación resultante puede ser utilizada como clasificador lineal, o, más comúnmente, para la reducción de dimensionalidad antes de la clasificación posterior. (Rodriguez, 2018)

LDA está estrechamente relacionado con el análisis de varianza (ANOVA) y el análisis de regresión, que también intentan expresar una variable dependiente como una combinación lineal de otras características o medidas. Sin embargo, ANOVA utiliza variables independientes categóricas y una variable dependiente continua, mientras que el análisis discriminante tiene variables independientes continuas y una variable dependiente categórica; es decir, la etiqueta de clase. La regresión logística y la regresión probit son más similares a LDA que ANOVA, ya que también explican una variable categórica por los valores de las variables independientes continuas. Estos otros métodos son preferibles en aplicaciones donde no es razonable asumir que las variables independientes están distribuidas normalmente, lo cual es un supuesto fundamental del método LDA. (Rodríguez N. , 2018)

LDA también está estrechamente relacionado con el análisis de componentes principales (PCA) y el análisis factorial, ya que ambos buscan combinaciones lineales de variables que expliquen mejor los datos. LDA explícitamente intenta modelar la diferencia entre las clases de datos. El PCA, por el contrario, no tiene en cuenta ninguna diferencia de clase, y el análisis factorial construye las combinaciones de características basándose en diferencias en lugar de similitudes. El análisis discriminante también es diferente del análisis factorial en que no es una técnica de interdependencia: se debe hacer una distinción entre las variables independientes y las variables dependientes (también llamadas variables de criterio).

LDA funciona cuando las mediciones realizadas en variables independientes para cada observación son cantidades continuas. Cuando se trata de variables independientes categóricas, la técnica equivalente es el análisis de correspondencia discriminante. El análisis discriminante se utiliza cuando los grupos se conocen a priori (a diferencia del análisis de conglomerados). Cada caso debe tener una puntuación en una o más medidas predictoras cuantitativas, y una puntuación en una medida de grupo. En términos simples, el análisis de función discriminante es la clasificación - el acto de distribuir cosas en grupos, clases o categorías del mismo tipo. (Cerrillo, 2019)

Random Forest

Random Forest es un algoritmo de aprendizaje automático flexible y fácil de usar que produce, incluso sin ajuste de hiperparámetros, un gran resultado la mayor parte del tiempo. También es uno de los algoritmos más utilizados, por su simplicidad y diversidad (se puede utilizar tanto para tareas de clasificación como de regresión). La idea general del método de ensacado es que una combinación de modelos de aprendizaje aumenta el resultado general. El bosque aleatorio es un algoritmo de aprendizaje supervisado. El "bosque" que construye es un conjunto de árboles de decisión, generalmente entrenados con el método de "ensacado". La idea general del método de ensacado es que una combinación de modelos de aprendizaje aumenta el resultado general. (Cotino, 2020)

Random Forest es un algoritmo de clasificación, que permite procesar datos que tengan una varianza significativa; los datos obtenidos por las distintas fuentes de información son bastantes dispersos y en muchos casos inconsistentes y variables. Este algoritmo está especialmente diseñado para procesar este tipo de casos, es así que, su aplicación en la problemática del proyecto resulta fundamental para obtener predicciones eficientes y eficaces, dando así un nivel de confiabilidad alto a los doctores que utilicen la herramienta, obteniendo resultados positivos en la toma de decisiones para la derivación hospitalaria.

Naive Bayes

Es una técnica de clasificación basada en el teorema de Bayes con un supuesto de independencia entre los predictores. En términos simples, un clasificador Naive Bayes asume que la presencia de una característica particular en una clase no está relacionada con la presencia de ninguna otra característica. Naive Bayes es una técnica simple para construir clasificadores: modelos que asignan etiquetas de clase a instancias de problemas, representadas como vectores de valores de entidad, donde las etiquetas de clase se extraen de un conjunto finito. (Cerrillo, 2019)

No hay un solo algoritmo para entrenar tales clasificadores, sino una familia de algoritmos basados en un principio común: todos los clasificadores de Bayes ingenuos asumen que el valor de una característica particular es independiente del valor de cualquier otra característica, dada la variable de clase. Por ejemplo, una fruta puede considerarse una manzana si es roja, redonda y de unos 10 cm de diámetro. Un clasificador de Bayes ingenuo considera que cada una de estas características contribuye de forma independiente a la probabilidad de que esta fruta sea una manzana, independientemente de las posibles correlaciones entre las características de color, redondez y diámetro. (Cerrillo, 2019)

A pesar de su diseño ingenuo y supuestos aparentemente simplificados, los clasificadores de Bayes ingenuos han funcionado bastante bien en muchas situaciones complejas del mundo real. En 2004, un análisis del problema de clasificación bayesiana mostró que existen razones teóricas sólidas para la eficacia aparentemente inverosímil de los clasificadores Bayes ingenuos. Aun así, una comparación integral con otros algoritmos de clasificación en 2006 mostró que la clasificación de Bayes es superada por otros enfoques, como árboles impulsados o bosques aleatorios. (Hutter, 2019)

Algunas de las ventajas que Naive Bayes ofrece en contraste con Random Forest son:

- Simplicidad si la independencia condicional se mantiene convergiendo más rápido que modelos discriminatorios, necesitando menos datos para entrenamiento.
- Reducción de tiempos en creación del modelo.

La principal diferencia entre los algoritmos radica en los ajustes que se les tiene que dar a los modelos, siendo que, el modelo Naive Bayes es bajo y constante, en Random Forest es considerablemente grande y variable. Siendo Naive Bayes el algoritmo óptimo para la implementación del proyecto, debido que, puede adaptarse rápidamente a los cambios y nuevos datos.

Ciencia de datos

La ciencia de datos es un concepto que se utiliza para abordar los macro datos e incluye la limpieza, preparación y análisis de datos. Un científico de datos recopila datos de múltiples fuentes y aplica el aprendizaje automático, el análisis predictivo y el análisis de sentimientos para extraer información crítica de los conjuntos de datos recopilados. Entienden los datos desde un punto de vista empresarial y pueden proporcionar predicciones y conocimientos precisos que se pueden utilizar para impulsar decisiones empresariales críticas. (Dueñas, 2018)

Lenguaje de programación R

Es un programa estadístico y un lenguaje de programación de uso libre, de distribución gratuita y código abierto, desarrollado a partir de un proyecto colaborativo voluntario de investigadores y estadísticos de diversos países y disciplinas. Es un programa basado en comandos, que permite acceder a todos los procedimientos opciones a través de una sintaxis textual. Fue oficialmente presentado en 1997 bajo Licencia General Pública de la Fundación de Software Libre. (Avello & Seisdedo, 2017)

SAS

El sistema SAS se define como un conjunto integrado por programas independientes del Hardware. Su finalidad es el procesamiento y análisis de la información en la industria, gobierno, educación y negocios. Este sistema se ejecuta en diferentes plataformas y sistemas operativos, como Windows, Linux, Unix, macOs, entre otros.

Las tareas desarrolladas por el sistema SAS se encuentran en torno a datos almacenados llamados "datasets" entre las que se encuentran acceso, manejo, análisis y presentación de datos. En el acceso de datos permite adquirir los archivos que se requieran para luego proceder a editarlos, procesarlos, formatearlos o convertirlos. También, en la tarea de análisis de datos, transforma los datos obtenidos en información útil y de gran significancia, desde simples pruebas estadísticas descriptivas e inferenciales hasta programación de modelos lineales y sofisticados. Por último, permite la creación de informes y gráficos de los análisis realizados y obtener soporte físico. (Zambrano, 2016)

Python

Los desarrolladores prefieren Python debido a su capacidad para ejecutarse en múltiples plataformas sin la necesidad de cambiar, a diferencia de otros lenguajes de programación. Python se ejecuta en diferentes plataformas, como Windows, Linux y macOS, por lo que requiere pocos o ningún cambio. Las plataformas son totalmente compatibles con el lenguaje de programación Python, lo que significa que hay poca o ninguna necesidad de que un experto en Python explique el código del programa. La facilidad de ejecución facilita la distribución de software, lo que permite crear y ejecutar software independiente con Python.

El software se puede programar de principio a fin utilizando Python como único idioma. Es una ventaja para los desarrolladores, ya que otros lenguajes de programación requieren la complementación con otros lenguajes antes de que el proyecto se concluya por completo. La independencia de Python en todas las plataformas ahorra tiempo y recursos para los

desarrolladores, que de otro modo incurrirían en muchos recursos para completar un solo proyecto. (Acosta, 2018)

En el presente proyecto se utilizará Python por ser un lenguaje de programación flexible y adaptable. Este lenguaje de programación es de los lenguajes de mejor rendimiento al momento de implementar cualquier proyecto o prototipo que funcione con Machine Learning e Inteligencia Artificial. A continuación, se presentan algunas comparaciones entre Python y otro lenguaje de programación como lo es R:

Tabla 4

Comparación entre características de R y Python.

Tabla de Comparación	
R	Python
• Cuenta con variedad de productos de código abierto. incorpora redes neuronales y otras tecnologías avanzadas. • La instalación cuenta con funcionalidades muy completas • La visualización de datos en R es una de sus principales virtudes.	• Cuenta variedad de opciones y funcionalidades, tiene una sintaxis manejable y sencilla. • Es un lenguaje mucho más fácil de aprender que otros. • Python más común que otros lenguajes de programación, lo que lo convierte en la opción principal a elegir tanto por novatos como por expertos.

Nota: Se presenta las diferencias entre R y Python. La elaboración es propia y la fuente es proporcionada por Acosta (2018).

Minería de datos

La minería de datos se considera el proceso de extraer información útil de una gran cantidad de datos. Se utiliza para descubrir patrones nuevos, precisos y útiles en los datos, buscando significado e información relevante para la organización o la persona que la necesita. Es una herramienta utilizada por los humanos (Fontalvo, 2018). La minería de datos es un proceso de descubrimiento de patrones en grandes conjuntos de datos que involucran métodos en la intersección del aprendizaje automático, las estadísticas y los sistemas de bases de datos. La minería de datos es un subcampo interdisciplinario de la informática y la estadística con el

objetivo general de extraer información (con métodos inteligentes) de un conjunto de datos y transformar la información en una estructura comprensible para su uso posterior.

Es el paso de análisis del proceso de "descubrimiento de conocimiento en bases de datos", o KDD. Aparte del paso de análisis en bruto, también implica aspectos de gestión de datos y bases de datos, preprocesamiento de datos, consideraciones de modelo e inferencia, métricas de interés, consideraciones de complejidad, post-procesamiento de estructuras descubiertas, visualización y actualización en línea. El término "minería de datos" es un nombre inapropiado, porque el objetivo es la extracción de patrones y conocimiento de grandes cantidades de datos, no la extracción (minería) de datos en sí.

Herramientas de desarrollo de software

Python 3.5

Python es un lenguaje de programación interpretado, el cual es orientado a objetos de alto nivel con semántica dinámica. Sus estructuras de datos integradas son combinadas con tipado y enlace dinámico, lo hacen muy atractivo para el desarrollo rápido de aplicaciones, así como para su uso como lenguaje de scripts para conectar componentes existentes.

Según indica Rajkomar (2019) “Python's simple and easy-to-learn syntax emphasizes readability and therefore compresses the cost of maintaining the program. The algorithm accepts modules and packages, which causes the modularity of the program and the reuse of the code”. (pág. 6)

La sintaxis simple y fácil de aprender de Python enfatiza la legibilidad y, por lo consiguiente, comprime el costo de mantenimiento del programa. El algoritmo acepta módulos y paquetes, lo que provoca la modularidad del programa y la reutilización del código. El intérprete estándar está disponible en formato fuente o binario sin cargo para todas las plataformas principales y se pueden distribuir libremente. (Rajkomar, 2019)

Línea de tendencia

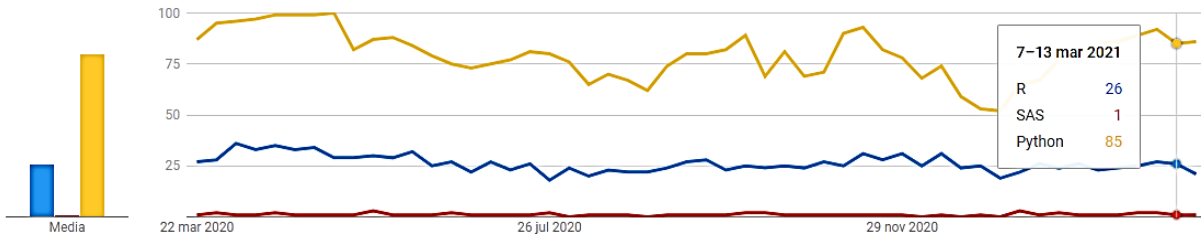
Una vez conociéndose la definición y características que brinda Python, se elabora la comparación con R.

R es utilizado para la elaboración de análisis estadísticos complejos, todo aquello que tenga que ver con procesos matemáticos, debido a su extensa librería dedicada a este propósito y su simplicidad al momento de implementar, mientras que, Python es empleado para códigos más generales por la sencillez del desarrollo y en tiempo de ejecución es más rápido.

En la *Figura 3* se muestra que Python tiene el 85%, mientras que R obtiene el 26% y por último SAS con sólo 1%. ¿Por qué Python? Porque es excelente para la codificación de algoritmos de Machine Learning debido a que contiene una cantidad de librerías bastante amplia para realizar modelos predictivos que ayuden a la toma de decisiones. En este caso, se ha utilizado para realizar los algoritmos de Random Forest y Naive Bayes.

Figura 3

Línea de tendencia entre R, SAS y Python



Nota: Se presenta la línea de tendencia, en la cual se visualiza la herramienta más utilizada entre marzo 2020 y marzo 2021. La elaboración es propia y la fuente es proporcionaba por Google Trends.

Google Colab

Es un servicio gratuito en la nube alojado por Google para fomentar la investigación del aprendizaje automático y la inteligencia artificial, donde a menudo la barrera para el aprendizaje y el éxito es el requisito de un poder computacional. Google Colab permite a los desarrolladores usar y compartir el cuaderno entre sí sin tener que descargar, instalar o ejecutar otra cosa que no sea un navegador. (Bisong, 2019)

Google Colab brinda un servicio en la nube, el cual está basado en las notebooks de Jupiter, gracias a esto, se lo ha utilizado para realizar los algoritmos planteados ya que no se necesita de instalación previa en los dispositivos portátiles o móviles.

STAT::FIT (ProModel 2016)

Es un software dedicado a realizar ajustes de curvas y análisis estadístico de los datos de entrada y salida que serán utilizados para la simulación. Permite alcanzar cinco objetivos que ayudan a comprobar los resultados obtenidos, tales como: ajuste de curvas, determinación del número de réplicas para correr un determinado modelo de terminación, determinación del tamaño de muestra para conocer tiempos de proceso y transportación, graficación de datos de entrada y de distribuciones probabilísticas y difusión de pensamiento estadístico. (Geer Mountain Software Corp, 2016)

Esta herramienta es una de las opciones que brinda el programa de ProModel, el cual es útil para realizar la ulterior simulación de datos. STAT::FIT se lo utilizó con la finalidad de obtener las distribuciones probabilísticas por cada variable del dataset que ayudará más adelante para realizar la Simulación de Montecarlo.

@RISK 8.1

El programa @RISK es una herramienta útil para el uso de Excel, que sirve para la creación de Simulación de Montecarlo. Se lo ha empleado para la obtención de una mayor cantidad de datos, un total de 1400 datos, para posteriormente añadirlos al dataset y empezar a realizar los modelos predictivos.

SPSS 26.0

El programa de SPSS ha sido útil para realizar estadísticas descriptivas, el cual arroja los siguientes resultados: el análisis de correlación Pearson, el chi-cuadrado y las tablas de

contingencia. Este programa se utilizó para las encuestas realizadas a los estudiantes y a los expertos.

Meta-análisis

El meta-análisis es una forma de revisión sistemática cuantitativa, en donde los resultados de una serie de estudios empíricos sobre un mismo tópico de investigación son integrados estadísticamente. Los resultados de los estudios son expresados en una métrica común a través de un índice del tamaño del efecto. La diferencia de medias tipificada es uno de los más utilizados en estudios dónde dos o más grupos son comparados en una variable de resultado continua. (Rubio, Sánchez, Martínez, & López, 2020)

Marco Muestral

El marco muestral está constituido por artículos científicos de diferentes fuentes y autores que fueron previamente seleccionados luego de la investigación realizada. Estos artículos son analizados en el meta-análisis definiendo once variables que se encuentran en la *Tabla 5*.

Tabla 5

Variables analizadas en el meta-análisis

N°	Nombre de la variable
1	Título del artículo científico
2	Autor(es)
3	Palabras claves
4	Resumen
5	Número de veces que se repite la palabra “COVID-19” en los artículos
6	Número de veces que se repite la palabra “Machine Learning” en los artículos
7	Número de veces que se repite la palabra “Random Forest” en los artículos
8	Número de veces que se repite la palabra “Naive Bayes” en los artículos
9	Número de veces que se repite la palabra “Derivación Hospitalaria” en los artículos
10	Bibliografía

Nota: Se presentan las variables que forman parte importante para la construcción del meta-análisis. La elaboración es propia y la fuente es proporcionaba por las investigaciones dadas.

Diseño del meta-análisis

En el diseño del meta-análisis, la población total fue de 50 artículos científicos. Se tomó una muestra de 30 artículos que obedecían al criterio de contener al menos el 50% de las palabras claves establecidas en la investigación. Para la toma de la muestra se utilizó el programa "SurveyMonkey", el cual toma los siguientes datos: tamaño de la población, nivel de confianza y margen de error.

Descripción de las variables utilizadas

1) **Variable 1: Palabras claves.** Palabras que condensan lo más relevante del contenido de un artículo científico. Para la elección de las palabras claves se debe considerar cuáles serían las palabras que utilizaría el usuario para encontrar el artículo en el buscador.

Tabla 6

Palabras claves como variable del meta-análisis

N°	Palabras claves
1	COVID-19
2	Machine Learning
3	Random Forest
4	Naive Bayes
5	Derivación Hospitalaria

Nota: Se presentan las palabras claves que forman parte importante para la creación del meta-análisis. La elaboración es propia y la fuente es proporcionaba por las investigaciones dadas.

2) **Variable 2: Número de veces que se repite la palabra "COVID-19" en los artículos.** Esta variable se define como el número de veces que es utilizada la palabra "COVID-19" en el contenido de los 30 artículos científicos seleccionados.
3) **Variable 3: Número de veces que se repite la palabra "Machine Learning" en los artículos.** Esta variable se define como el número de veces que es utilizada la palabra "Machine Learning" en el contenido de los artículos. Machine Learning es una

disciplina científica dentro del ámbito de la Inteligencia Artificial en la cual se crean sistemas que son capaces de aprender de manera automática.

4) **Variable 4: Número de veces que se repite la palabra "Random Forest" en los artículos. -** Esta variable se define como el número de veces que es utilizada la palabra "Random Forest" en el contenido de los artículos. Random Forest es un método de aprendizaje automático que es capaz de llevar a cabo tareas de regresión y de clasificación realizando una estimación coherente de ambos.
5) **Variable 5: Número de veces que se repite la palabra "Naive Bayes" en los artículos.** Esta variable se define como el número de veces que es utilizada la palabra "Naive Bayes" en el contenido de los artículos. Naive bayes representa una clase especial de algoritmos de clasificación dentro del aprendizaje automático.
6) **Variable 6: Número de veces que se repite la palabra "Derivación Hospitalaria" en los artículos.** Esta variable se define como el número de veces que es utilizada la palabra "Derivación Hospitalaria" en el contenido de los artículos**.** La Derivación Hospitalaria es la acción de remitir a un paciente a través de interconsulta desde un profesional de la salud hacia otro para ofrecerle atención personalizada y complementaria tanto para su diagnóstico como para algún tratamiento que este requiera.
7) **Variable 7: Bibliografía.** Esta variable se define como las fuentes de las cuales fueron consultados y extraídos los 30 artículos científicos establecidos en la muestra. En la *Tabla 7* se muestran las diferentes fuentes revisadas.

Tabla 7

Fuentes bibliográficas consultadas en la investigación.

N°	Bibliografía
1	Google académico
2	PeerJ
3	MDPI
4	Frontiers
5	Elsevier
6	Springer

Nota: Se presentan las palabras claves que forman parte importante para la construcción del meta-análisis. La elaboración es propia y la fuente es proporcionaba por las investigaciones dadas.

Instrumentos utilizados para la recolección de datos

El instrumento utilizado en este meta-análisis fue la exploración y análisis de artículos científicos que contenían información relevante acerca del tema de investigación y con fecha de publicación durante el período 2018-2020. La búsqueda de estos se realizó mediante las palabras claves y el objetivo general del proyecto. Fueron encontrados en diferentes fuentes bibliográficas y leídos de manera rápida para, posteriormente, ser añadidos en una base de datos en Microsoft Excel para analizarlos. En la base de datos de Microsoft Excel se obtuvo información como el título, autores, palabras claves y el número de veces que se repetían las palabras estudiadas en este meta-análisis. Además, se extrajeron las fuentes bibliográficas y se realizaron cálculos estadísticos a cada variable.

Análisis de los resultados

Variable 1. Palabras claves.

Tabla 8

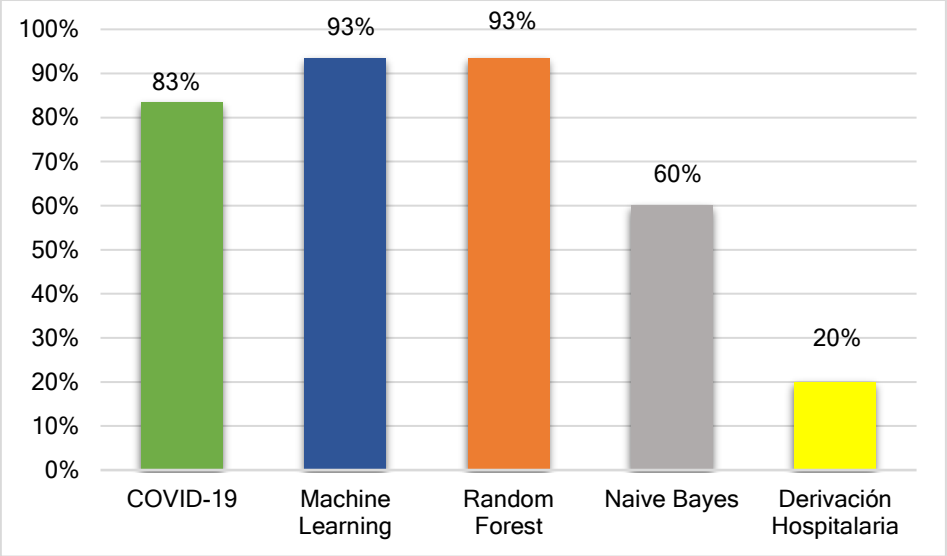
Variable palabras claves

Palabras Claves	Frecuencia absoluta	Frecuencia relativa
COVID-19	25	83%
Machine Learning	28	93%
Random Forest	28	93%
Naive Bayes	18	60%
Derivación Hospitalaria	6	20%

Nota: Se presentan las palabras claves que forman parte importante para la preparación del meta-análisis. La elaboración es propia y la fuente es proporcionaba por las investigaciones dadas.

Figura 4

Variable 1. Gráfico estadístico de las palabras claves



Nota: Se presenta el gráfico estadístico de las palabras claves que forman parte importante para la construcción del meta-análisis. La elaboración es propia y la fuente es proporcionaba por las investigaciones dadas.

Análisis: De acuerdo con los resultados obtenidos, el 83% de los 30 artículos científicos considerados contienen la palabra "COVID-19" a lo largo del texto, presentan una frecuencia de 25 artículos. El 93% de los artículos contienen la palabra clave "Machine Learning" presentando una frecuencia de 28 artículos, por lo tanto, se considera que este ámbito de la

inteligencia artificial aportará en gran medida en la resolución del problema planteado en la investigación. El 93% de los artículos contienen la palabra clave “Random Forest” presentando una frecuencia de 28 artículos, lo que también ayuda a concluir que este método de aprendizaje automático complementará la solución. El 60% de los artículos contienen la palabra clave “Naive Bayes” presentando una frecuencia de 18 artículos, concluyendo que estos algoritmos serán de utilidad para la investigación porque su frecuencia es mayor al 50%. Por último, el 20% de los artículos contienen la palabra clave “Derivación Hospitalaria” presentando una frecuencia de 6 artículos; la frecuencia disminuye puesto que es un concepto que está muy claro y será necesario para complementar la solución debido a que bajo esos criterios se trabajará la misma.

Variable 2. Número de veces que se repite la palabra “COVID-19” en los artículos.

Tabla 9

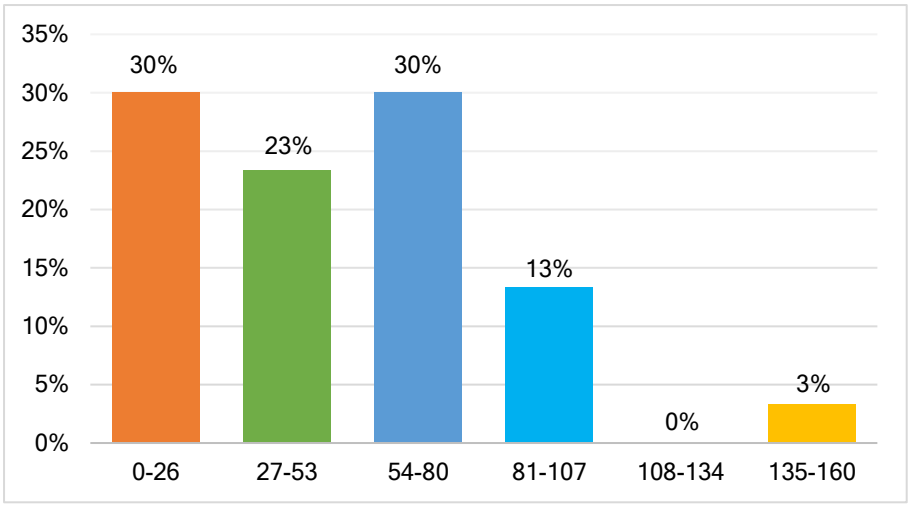
Variable número de veces que se repite la palabra "COVID-19" en los artículos

Número de veces	Frecuencia absoluta	Frecuencia Acumulada	Frecuencia Relativa	Frecuencia Relativa Acumulada
0-26	9	9	30%	30%
27-53	7	16	23%	53%
54-80	9	25	30%	83%
81-107	4	29	13%	97%
108-134	0	29	0%	97%
135-160	1	30	3%	100%
TOTAL	30		100%	

Nota: Se presentan la tabla de frecuencia de la variable COVID-19 que forma parte importante para la creación del meta-análisis. La elaboración es propia y la fuente es proporcionaba por las investigaciones dadas.

Figura 5

Variable 2. Gráfico estadístico de la palabra COVID-19



Nota: Se presenta el gráfico estadístico de la variable COVID-19 que forma parte importante para la realización del meta-análisis. La elaboración es propia y la fuente es proporcionaba por las investigaciones dadas.

Análisis: De acuerdo con los 30 artículos científicos seleccionados, el 30% cuenta con una frecuencia de 0 a 26 veces de la palabra "COVID-19" en su contenido, el 23% cuenta con una frecuencia de 27 a 53 veces, el 30% cuenta con una frecuencia de 54 a 80 veces, el 13% cuenta con una frecuencia de 81 a 107 veces, el 0% cuenta con una frecuencia de 108 a 134 veces y, por último, el 3% cuenta con una frecuencia de 135 a 160 veces.

Variable 3. Número de veces que se repite la palabra "Machine Learning" en los artículos.

Tabla 10

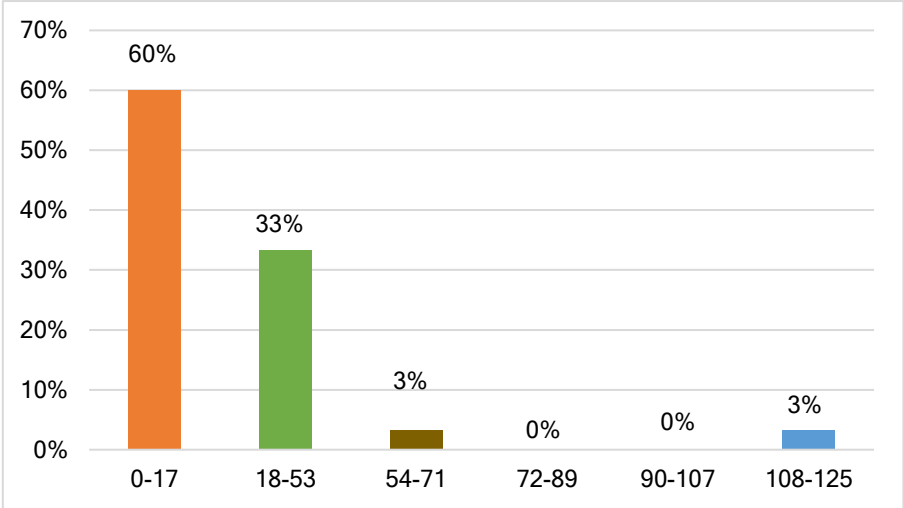
Número de veces que se repite la palabra "Machine Learning" en los artículos

Número de veces	Frecuencia absoluta	Frecuencia Acumulada	Frecuencia Relativa	Frecuencia Relativa Acumulada
0-17	18	18	60%	60%
18-53	10	28	33%	93%
54-71	1	29	3%	97%
72-89	0	29	0%	97%
90-107	0	29	0%	97%
108-125	1	30	3%	100%
TOTAL	30		100%	

Nota: Se presentan la tabla de frecuencia de la variable Machine Learning que forma parte importante para la construcción del meta-análisis. La elaboración es propia y la fuente es proporcionaba por las investigaciones dadas.

Figura 6

Variable 3. Gráfico estadístico de la palabra Machine Learning.



Nota: Se presenta el gráfico estadístico de la variable Machine Learning que forma parte importante para la elaboración del meta-análisis. La elaboración es propia y la fuente es proporcionaba por las investigaciones dadas.

Análisis: De acuerdo con los 30 artículos científicos seleccionados, el 60% cuenta con una frecuencia de 0 a 17 veces de la palabra "Machine Learning" en su contenido, el 33% cuenta con una frecuencia de 18 a 53 veces, el 3% cuenta con una frecuencia de 54 a 71 veces, el 0% cuenta con una frecuencia de 72 a 89 veces, el 0% cuenta con una frecuencia de 90 a 107 veces y, por último, el 3% cuenta con una frecuencia de 108 a 125 veces.

Variable 4. Número de veces que se repite la palabra "Random Forest" en los artículos.

Tabla 11

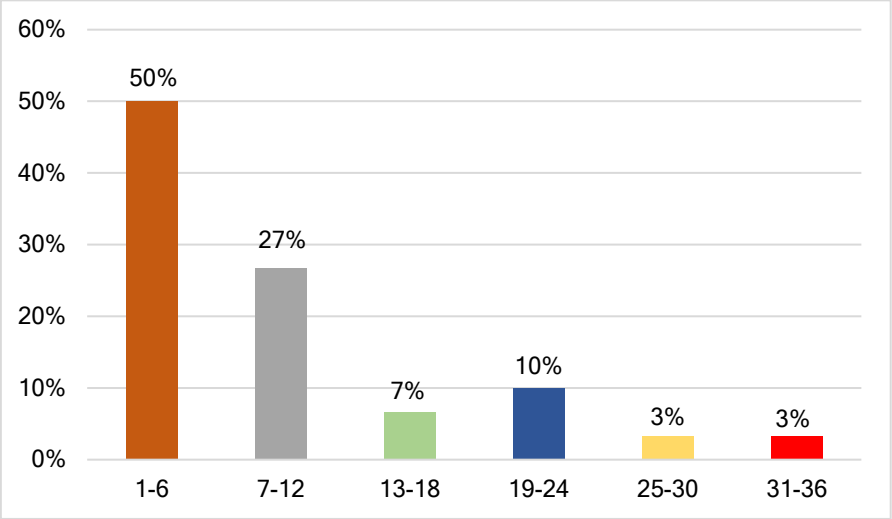
Número de veces que se repite la palabra "Random Forest" en los artículos

Número de veces	Frecuencia absoluta	Frecuencia Acumulada	Frecuencia Relativa	Frecuencia Relativa Acumulada
1-6	15	15	50%	50%
7-12	8	23	27%	77%
13-18	2	25	7%	83%
19-24	3	28	10%	93%
25-30	1	29	3%	97%
31-36	1	30	3%	100%
TOTAL	30		100%	

Nota: Se presentan la tabla de frecuencia de la variable Random Forest que forma parte importante para la realización del meta-análisis. La elaboración es propia y la fuente es proporcionaba por las investigaciones dadas.

Figura 7

Variable 4. Gráfico estadístico de la palabra Random Forest.



Nota: Se presenta el gráfico estadístico de la variable Random Forest que forma parte importante para la preparación del meta-análisis. La elaboración es propia y la fuente es proporcionaba por las investigaciones dadas.

Análisis: De acuerdo con los 30 artículos científicos seleccionados, el 50% cuenta con una frecuencia de 1 a 6 veces de la palabra "Random Forest" en su contenido, el 27% cuenta con una frecuencia de 7 a 12 veces, el 7% cuenta con una frecuencia de 13 a 18 veces, el 10% cuenta con una frecuencia de 19 a 24 veces, el 3% cuenta con una frecuencia de 25 a 30 veces y, por último, el 3% cuenta con una frecuencia de 31 a 36 veces.

Variable 5. Número de veces que se repite la palabra "Naive Bayes" en los artículos.

Tabla 12

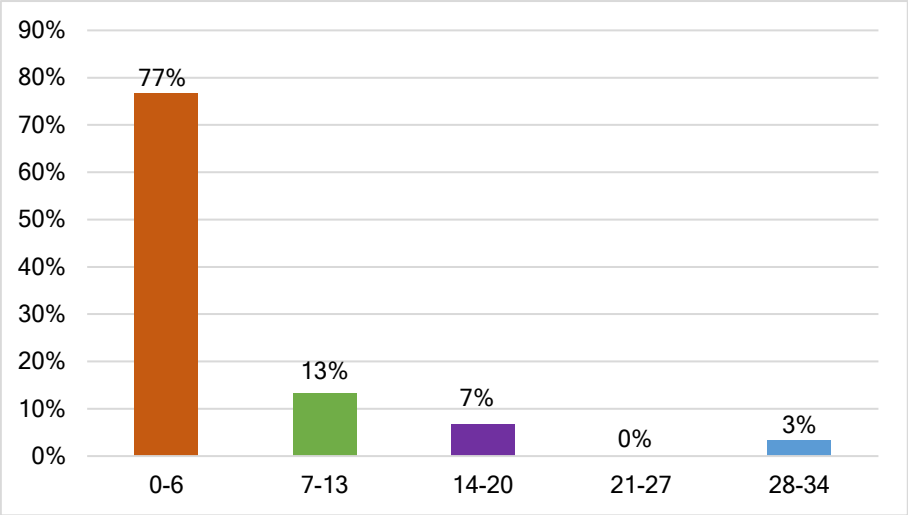
Número de veces que se repite la palabra "Naive Bayes" en los artículos

Número de veces	Frecuencia absoluta	Frecuencia Acumulada	Frecuencia Relativa	Frecuencia Relativa Acumulada
0-6	23	23	77%	77%
7-13	4	27	13%	90%
14-20	2	29	7%	97%
21-27	0	29	0%	97%
28-34	1	30	3%	100%
TOTAL	30		100%	

Nota: Se presentan la tabla de frecuencia de la variable Naive Bayes que forma parte importante para la construcción del meta-análisis. La elaboración es propia y la fuente es proporcionaba por las investigaciones dadas.

Figura 8

Variable 5. Gráfico estadístico de la palabra Naive Bayes.



Nota: Se presenta el gráfico estadístico de la variable Naive Bayes que forma parte importante para la construcción del meta-análisis. La elaboración es propia y la fuente es proporcionaba por las investigaciones dadas.

Análisis: De acuerdo con los 30 artículos científicos seleccionados, el 77% cuenta con una frecuencia de 0 a 6 veces de la palabra "Naive Bayes" en su contenido, el 13% cuenta con una frecuencia de 7 a 13 veces, el 7% cuenta con una frecuencia de 14 a 20 veces, el 0% cuenta con una frecuencia de 21 a 27 veces y, por último, el 3% cuenta con una frecuencia de 28 a 34 veces.

Variable 6. Número de veces que se repite la palabra "Derivación Hospitalaria" en los artículos.

Tabla 13

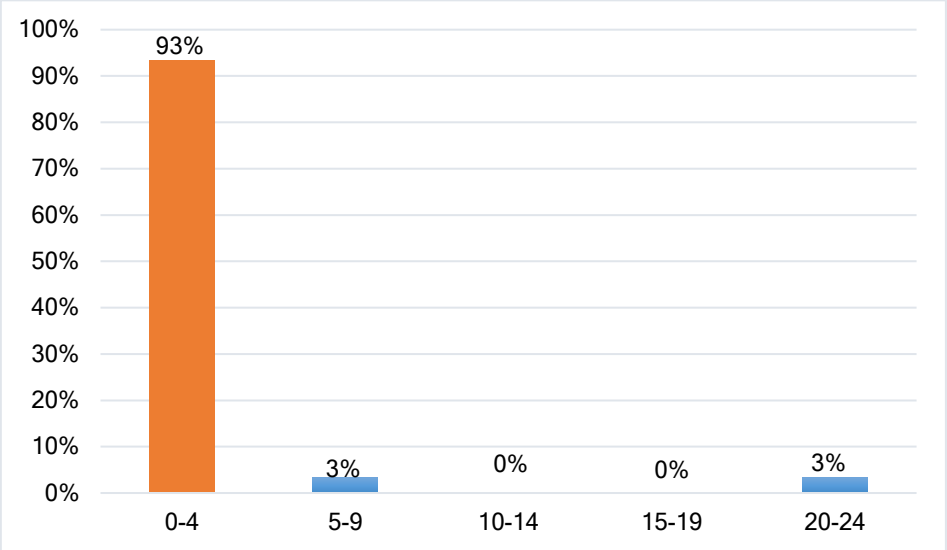
Número de veces que se repite la palabra "Derivación Hospitalaria" en los artículos

Intervalos	Frecuencia absoluta	Frecuencia Acumulada	Frecuencia Relativa	Frecuencia Relativa Acumulada
0-4	28	28	93%	93%
5-9	1	29	3%	97%
10-14	0	29	0%	97%
15-19	0	29	0%	97%
20-24	1	30	3%	100%
TOTAL	30		100%	

Nota: Se presentan la tabla de frecuencia de la variable Derivación Hospitalaria que forma parte importante para la construcción del meta-análisis. La elaboración es propia y la fuente es proporcionaba por las investigaciones dadas.

Figura 9

Variable 6. Gráfico estadístico de la palabra Derivación Hospitalaria.



Nota: Se presenta el gráfico estadístico de la variable Derivación Hospitalaria que forma parte importante para la preparación del meta-análisis. La elaboración es propia y la fuente es proporcionaba por las investigaciones dadas.

Análisis: De acuerdo con los 30 artículos científicos seleccionados, el 93% cuenta con una frecuencia de 0 a 4 veces de la palabra "Derivación Hospitalaria" en su contenido, el 3% cuenta con una frecuencia de 5 a 9 veces, el 0% cuenta con una frecuencia de 10 a 24 veces, el 0% cuenta con una frecuencia de 15 a 19 veces y, por último, el 3% cuenta con una frecuencia de 20 a 24 veces.

Variable 7. Bibliografía.

Tabla 14

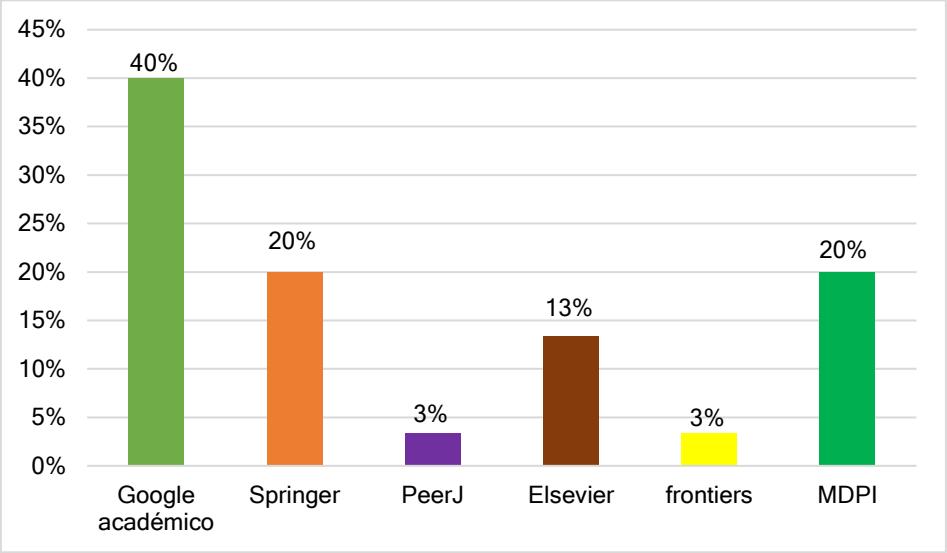
Variable bibliografía

Bibliografía	Frecuencia absoluta	Frecuencia Relativa
Google académico	12	40%
Springer	6	20%
PeerJ	1	3%
Elsevier	4	13%
frontiers	1	3%
MDPI	6	20%
TOTAL	30	100%

Nota: Se presentan la tabla de frecuencia de la variable Bibliografía que forma parte importante para la construcción del meta-análisis. La elaboración es propia y la fuente es proporcionaba por las investigaciones dadas.

Figura 10

Variable 7. Gráfico estadístico de la palabra Bibliografía.



Nota: Se presenta el gráfico estadístico de la variable Bibliografía que forma parte importante para la preparación del meta-análisis. La elaboración es propia y la fuente es proporcionaba por las investigaciones dadas.

Análisis: De acuerdo con los 30 artículos científicos seleccionados, el 40% de los fueron consultados en Google académico, el 20% fueron consultados en la plataforma virtual Springer, el 3% fueron consultados en la plataforma PeerJ, el 13% fueron consultados en la plataforma Elservier, el 3% fueron consultados en la plataforma Frontiers y, por último, el 20% fueron consultados en la plataforma MDPI.

Hipótesis

Si se hace uso de un algoritmo de Machine Learning como es el de Random Forest entonces se podrá tomar una decisión óptima respecto a la derivación hospitalaria o ambulatoria por parte del experto para los pacientes infectados con COVID-19 en un hospital público de la ciudad de Guayaquil.

Si se hace uso de un algoritmo de Machine Learning como es el de Naive Bayes entonces se podrá tomar una decisión óptima respecto a la derivación hospitalaria o ambulatoria por parte del experto para los pacientes infectados con COVID-19 en un hospital público de la ciudad de Guayaquil.

Definiciones conceptuales

Machine Learning

De acuerdo al autor Southgate (2019) se define el Machine Learning como “a subfield of artificial intelligence, machine learning is the science of get machines to learn like humans in an autonomous way”. (pág. 8)

Machine Learning es una rama de la inteligencia artificial que se define como un proceso por medio del cual una computadora analiza datos para automatizar la construcción de diferentes modelos analíticos. Esta rama está basada en la idea de que los diversos sistemas son capaces de aprender datos identificando patrones y participar en la toma de decisiones con mínima intervención humana (Southgate, y otros, 2019). De los múltiples algoritmos que existen en la rama de Machine Learning, se ha enfocado para este presente proyecto de titulación el uso de dos modelos: Random Forest y Naive Bayes.

COVID-19

Enfermedad causada por el virus del síndrome respiratorio agudo severo tipo-2 (SARS-CoV-2). Se detectó como pandemia en marzo del 2020 afectando en mayor medida a los adultos mayores y personas con patologías previas como diabetes, hipertensión, enfermedades cardiovasculares y cáncer. Los síntomas principales son tos, fiebre y dificultad respiratoria. (Díaz & Toro, 2020)

Derivación Hospitalaria

Se establece como derivación hospitalaria al procedimiento utilizado con la finalidad de canalizar a un paciente desde una unidad operativa hacia otra que cuente con mayor capacidad resolutiva tanto para diagnóstico como para tratamiento recibiendo de esta manera atención médica integral. (Díaz, Díaz, Jaraboc, Roig, & Román, 2017)

Random Forest

Random Forest se establece como uno de los algoritmos de clasificación de imágenes más utilizados. Una de sus mayores ventajas es la aportación de estimación interna de exactitud que brinda a través de una validación cruzada. Este modelo hace uso de dos parámetros, siendo estos el número de árboles y el número de predictores que se utilizarán en la partición de cada árbol (Cánovas, Alonso, & Gomariz, 2016). Este algoritmo fue aplicado para el área de salud para la ayuda en la toma de decisiones en relación con la derivación hospitalaria o ambulatoria de pacientes infectados con COVID-19.

Naive Bayes

Es un algoritmo que utiliza el teorema de Bayes para clasificar objetos. Los clasificadores ingenuos de Bayes asumen una independencia fuerte entre los atributos. Los usos populares incluyen filtros de spam, análisis de texto y diagnóstico médico. Los mismos se utilizan ampliamente para el aprendizaje automático porque son simples de implementar (Hutter, 2019). Este algoritmo fue aplicado para el área de salud para la ayuda en la toma de decisiones en relación con la derivación hospitalaria o ambulatoria de pacientes infectados con COVID-19.

Modelo predictivo asistencial

Son aquellos algoritmos de Machine Learning, creados para ayudar en las tomas de decisiones de cualquier ámbito. En mucho de los casos, el suceso que se desea predecir va normalmente para tiempo futuro, queriendo diagnosticar la probabilidad de un conjunto de datos ya establecidos; aun así, el modelo puede emplearse para sucesos inexplorados sin importar el momento exacto. Para el presente proyecto, el modelo será empleado en el ámbito de la salud, dando soporte a los doctores de primera línea en caso de COVID-19, específicamente en la derivación hospitalaria o ambulatoria del paciente infectado por el virus.

CAPÍTULO III

METODOLOGÍA DE LA INVESTIGACIÓN

De acuerdo con Tacillo (2016), la metodología de la investigación se puede definir a través de la siguiente manera: Método, corresponde a una palabra de origen griego (meta = fin; odo = camino= cuyo significado es senda para alcanzar un objetivo. Así, cuando ese camino es planificado, ordenado e imparcial, corresponde al método científico, el cual constituye la metodología de la investigación.

De tal suerte que, la metodología de la investigación hace referencia a seguir un camino o conjunto de pasos ordenados, que se elabora con antelación con la finalidad de alcanzar un cierto propósito. Adicionalmente, se denomina sistemático, debido a ser planificado con base en un problema, metas, proposiciones tentativas, de las cuales se recogen los datos para el análisis, posteriormente se contextualizan e interpretan, y se obtienen conclusiones, los cuales reforzarán, alterarán o contrastarán las explicaciones de estudios precedentes. (Tacillo Yauli, 2016)

Modalidad de la investigación

De acuerdo a Norbis, citado en Iza (2019) indica que “Una modalidad de investigación es una colección de prácticas eclécticas de indagación basada en un conjunto real de suposiciones, e implica preferencias metodológicas, opiniones filosóficas e ideológicas, cuestiones de investigación y resultados de viabilidad” (p. 22). Dicho esto, la presente investigación se muestra con una modalidad de campo bibliográfica. En la investigación de campo se realizaron encuestas para determinar las variables de estudio establecidas. Por otra

parte, del total de la investigación, se realizó aproximadamente un 70% de revisiones bibliográficas de fuentes electrónicas, con el objetivo de extraer información relevante que sustente la realización de la investigación y el proyecto.

Tipo de investigación

Según lo afirmado por Arias (2016) "El tipo o nivel de investigación se refiere al grado de profundidad con que se aborda un fenómeno u objeto de estudio" (p. 10). A partir de esta definición, es posible establecer los diversos tipos o niveles de la presente investigación: investigación exploratoria, comprobación de hipótesis e investigación experimental.

Investigación exploratoria

De acuerdo con lo expresado por el autor Arias (2016), la investigación exploratoria "es aquella que se efectúa sobre un tema u objeto desconocido o poco estudiado, por lo que sus resultados constituyen una visión aproximada de dicho objeto; es decir, un nivel superficial de conocimientos". (p. 23)

Mencionado este concepto, se define que la presente investigación es exploratoria debido a que tanto el tema de investigación, como los datos e información recopilada, son tópicos poco explorados con anterioridad y desconocidos.

Comprobación de hipótesis

Según Espinoza (2018), la comprobación de hipótesis se basa en contrastar dicha hipótesis de una realidad; es decir, el investigador tiene que someter a prueba aquello que ha enunciado en su hipótesis, y para ello ha de establecer, mediante alguna técnica de contrastación si su hipótesis concuerda o no con los datos empíricos. En tal caso, solo se pueden dar dos posibilidades previsibles: o bien la hipótesis puede verse apoyada por datos empíricos y ha sido confirmada, o bien la hipótesis no corresponde con los datos empíricos y se dice entonces que ha sido desconformada o refutada por los datos empíricos.

Dicho esto, la presente investigación se desarrolló con una comprobación de hipótesis, debido a que se buscó y recopiló información que permitió someter las hipótesis de estudio a prueba y determinar si se cumplen, o si las mismas son erróneas, a través de la implementación de técnicas e instrumentos de recolección de datos.

Investigación experimental

De acuerdo con lo establecido por Alonso (2016) citado en Guevara, Verdesoto y Castro (2020), en la investigación de enfoque experimental, la persona encargada de tal cambia una o más variables de estudio, para revisar el aumento o disminución de esas variables y su efecto en las conductas observadas. De igual forma, el experimento se puede definir como un ensayo en el que se manipula deliberadamente una o más variables con el fin de cumplir con los objetivos específicos planteados.

Diseño metodológico de la investigación

Arias (2016) define el diseño metodológico de investigación como "la estrategia general que adopta el investigador para responder al problema planteado. En atención al diseño, la investigación se clasifica en: documental, de campo y experimental" (p. 27). De igual forma, La estrategia de la investigación está definida por la procedencia de los datos primarios en diseños de campo y secundarios en estudios documentales, y por la manipulación o no, de las condiciones en las cuales se realiza el estudio: diseños experimentales y no experimentales o de campo.

La presente investigación se desarrolló con un diseño de campo. En la investigación de campo se desarrollaron técnicas e instrumentos de recolección de datos (Encuesta) para determinar las variables de estudio. De igual forma, se desarrolló revisión bibliográfica de documentos académicos y revistas científicas (investigación documental) para la recolección de información relacionada con la investigación.

Metodología de investigación

El proceso metodológico para emplearse será “Knowledge Discovery in Databases – KDD”, el cual se utiliza para encontrar conocimiento en un conjunto de datos en bruto. Se desarrolla el modelo predictivo en 6 fases:

- **Importación y muestreo de datos:** En esta fase se reconocen y se reúnen los datos con los que se trabajará en un futuro.
- **Calidad de datos:** Identificar si los datos obtenidos están bien clasificados realizando una limpieza a los datos, dejando la base de datos completamente manejable.
- **Transformación:** Realizar la transformación y generación de nuevas variables las cuales son las más importantes que serán añadidas a un dataset que se va a utilizar en el modelo predictivo.
- **Modelización:** Durante esta fase se va a manejar los algoritmos de aprendizaje que serán utilizados para el modelo predictivo, manipulando los datos que se han creado en el dataset de la fase anterior.
- **Evaluación:** Verificar si lo que el modelo predictivo arroja es correcto. También se evalúa qué algoritmo de aprendizaje arroja una mejor certeza y cuál es el más rápido para obtener los resultados.
- **Implementación:** Esta última fase no se llevará a cabo, en su lugar se realizará un prototipo desarrollado a nivel de Python versión 3.5.

Población y muestra

Población

La población hace referencia a un conjunto de todos los componentes o integrantes de la misma naturaleza, que evidencian una cualidad específica o que, en su defecto, se refieren a un mismo concepto y a cuyos componentes se los someterá a análisis referentes a sus características y asociaciones. En este sentido, la población es definida previamente por el investigador y puede estar conformada por personas o por componentes disímiles a personas, tales como: casas, ventanas, tornillos, pacientes de psicología, ordenadores, antecedentes de pacientes, entre otros. (Lerma, Muñoz, Ortiz, & Ramos, 2016)

Población objetivo

Corresponde a la población que se busca específicamente estudiar mediante el proyecto. De manera ideal, la población objetivo debería corresponder a la población desprovista de alguna cualidad. No obstante, ello, por múltiples causas puede suceder que se observe a una población menor que la desprovista. En un caso como ese, la población objetivo sería disímil e inferior a esta última (Pastor, 2019). Dicho esto, para el presente trabajo de titulación se escoge como población objetivo la base de datos adquirida por un hospital público de la ciudad de Guayaquil, donde se recopilaron datos de pacientes con COVID-19 atendidos en el periodo 2020 hasta enero del 2021.

Muestra

De acuerdo con Pastor (2019), la muestra es el conjunto de operaciones que se realizan para estudiar la distribución de determinados caracteres en la totalidad de una población, universo o colectivo, partiendo de la observación de una fracción de la población considerada. Dicho de otra manera, la muestra corresponde a una porción de la población total objeto de

estudio, la cual se considera de ésta última, debido a implicar menores costos en su recolección. Además de ello, muchas veces no es posible acceder a toda la población, por factor tiempo y la cantidad de elementos totales que puede tener. En ese sentido, es preferible tomar muestras y en la medida de lo posible, hacer inferencias de la población. Para el presente proyecto no fue un hecho necesario utilizar la fórmula para el cálculo de la muestro, debido a que es el mismo tamaño de la población.

Muestreo

Según Hernández Sampieri et al., citado en Otzen y Manterola (2017) el muestreo corresponde a la acción de analizar las asociaciones que existen entre la distribución de una variable "y" en una determinada población "z" y la distribución de esta variable en la muestra a estudio. De esta manera, los procedimientos de muestreo de tipo probabilístico posibilitan advertir la posibilidad que cada uno de los elementos del estudio tiene de ser incluido en la muestra con base en una asignación aleatoria. Por contra, en los procedimientos de muestreo no probabilísticos, las observaciones elegidas del total de observaciones en estudio estarán determinados por ciertas cualidades, juicios, entre otros.

Marco Muestral

Respecto al marco muestral se puede definir de la siguiente manera: "consiste en descripciones disponibles con anterioridad del material en forma de mapas, listas, directorios, etc., a partir de los cuales las unidades de la muestra se pueden construir y se puede seleccionar un conjunto de unidades" (Guzmán, 2020). Dicho de una forma simple, el marco muestral corresponde al total de observaciones que integran el universo muestral que se considera válido para el estudio; es decir, se puede definir como aquel conjunto de elementos de los cuales se extraerá la muestra a estudiarse.

Procesamiento y análisis

Para el procesamiento y análisis de datos del presente proyecto se efectuó por medio de encuestas con una muestra de 30 personas procedentes de la ciudad de Guayaquil. Para el posterior análisis de los datos obtenidos se utilizó el programa SPSS, el cual arroja el consolidado de la tabla chi-cuadrado de todas las preguntas para el análisis de correlación de Pearson, tablas de contingencia y gráficos de barra.

Técnicas de recolección de datos

Se entenderá por técnica de investigación, el procedimiento o forma particular de obtener datos o información. Las técnicas son particulares y específicas de una disciplina, por lo que sirven de complemento al método científico, el cual posee una aplicabilidad general (Arias, 2016). Las técnicas de recolección de datos más conocidas son la encuesta, la entrevista, el análisis documental y la observación.

Para el desarrollo de la presente investigación se utilizó la técnica de la encuesta, la cual permitió realizar la recolección de datos directamente de la muestra seleccionada (datos primarios).

Encuesta

La encuesta se define como una investigación realizada sobre una muestra de sujetos representativa de un colectivo más amplio, utilizando procesos definidos de interrogación con el fin de obtener mediciones cuantitativas de una gran variedad de características objetivas y subjetivas de la población, citado en Espinoza y Bahamondes (2020).

Según las investigaciones dadas en variados artículos científicos y guiándose de la hipótesis establecida, se obtuvo una encuesta clara y precisa, basada en el conocimiento del área tecnológica y en la toma de decisiones en el área de salud. El objetivo de esta encuesta fue

la recopilación de las variables más importantes y la correlación entre ellas. La encuesta se la desarrolló vía online mediante la herramienta Google Forms, con un total de 7 preguntas.

Técnicas estadísticas para el procesamiento de la información.

A continuación, se detalla el análisis de correlación de Pearson, el contraste de hipótesis y el análisis de tablas de contingencia:

Análisis de correlación de Pearson

Conocimientos procesos de derivación vs. las demás variables. De las variables examinadas de la encuesta, las cuales son: conocimiento procesos de derivación, conocimiento uso de IA, conocimiento de algoritmos, toma de decisiones, conocimiento de Naive Bayes y conocimiento de Random Forest, es posible identificar las diferencias estadísticas entre cada una de ellas cuando las mismas están asociadas a la variable "conocimientos procesos de derivación" mostradas en la *Tabla 15*; en cada una de las variables se aplicó la prueba o test de Chi-cuadrado partiéndose del establecimiento de la hipótesis nula (H_0), la cual determina que las variables son independientes (es decir, no están correlacionadas). Por otra parte, se contrasta el planteamiento de una hipótesis alternativa (H_a) la cual determina que las variables son dependientes (es decir, están relacionadas). Por lo tanto, es posible determinar que si $\alpha = 0{,}05$ y el valor de p (significación asintótica bilateral) es menor al valor de α, entonces se rechaza la hipótesis nula y se acepta la alterna, es decir, significa que las variables están correlacionadas.

Tabla 15

Pruebas de chi-cuadrado. Conocimientos procesos de derivación vs. las demás variables

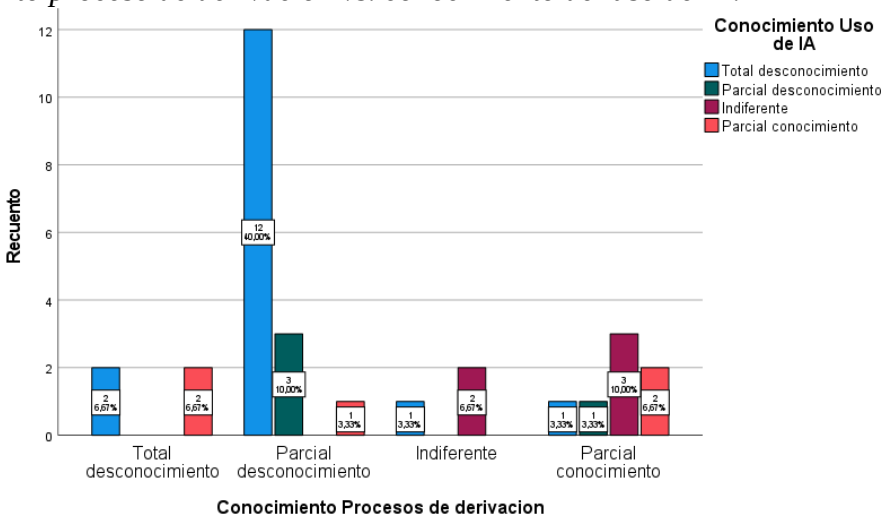
	Conocimientos procesos de derivación			
Características de la muestra	X^2	***g.l.***	***P***	***n***
Conocimientos uso de IA	20,451	9	0,015	30
Manejo de big data	17,484	12	0,132	30
Conocimiento de algoritmos	19,991	12	0,067	30
Toma de decisiones	23,810	9	0,005	30
Conocimiento de Naive Bayes	21,515	12	0,043	30
Conocimiento de Random Forest	21,800	12	0,040	30

Nota. Valor de la prueba de Chi-cuadrado de Pearson para la asociación entre la variable conocimientos procesos de derivación y las demás variables de estudio. La elaboración es propia y la fuente obtenida de las encuestas realizadas. X^2 Chi- cuadrado // g.l. Grados de libertad // p Significación asintótica (bilateral) // n número de encuestados de la muestra. La elaboración es propia y la fuente es obtenida de las encuestas realizadas.

Por lo tanto, para el estudio de las variables, se determina que dos de las seis categorías examinadas (Conocimientos uso de IA, toma de decisiones, conocimiento Naive Bayes y conocimiento de Random Forest) tienen relación o están correlacionados con la variable principal "conocimientos de procesos de derivación", al encontrar que el valor de p (significancia asintótica bilateral) es menor a 0,05.

Figura 11

Conocimiento proceso de derivación vs. conocimiento del uso de IA.

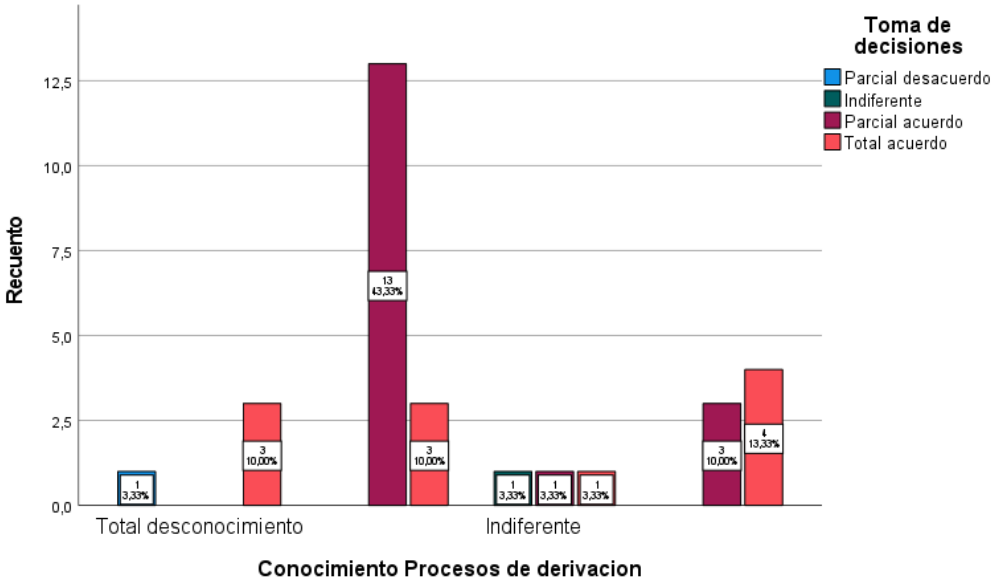


Nota. Gráfico de barras para la asociación entre la variable conocimientos procesos de derivación y conocimiento uso de IA. La elaboración es propia y la fuente obtenida de las encuestas realizadas.

En la *Figura 11* es posible apreciar los resultados obtenidos sobre los conocimientos de uso de inteligencia artificial (IA) en relación con los conocimientos de los procesos de derivación. En donde el total desconocimiento en “conocimiento procesos de derivación” se obtiene un 6,67% en total desconocimiento y un 6,67% en parcial conocimiento en relación con el “conocimiento de uso de IA”. Para parcial desconocimiento en “conocimiento procesos de derivación” se obtiene un 40% en total desconocimiento y un 10% en parcial desconocimiento en relación con el “conocimiento de uso de IA”. Para indiferente en “conocimiento procesos de derivación” se obtiene un 3,33% en total desconocimiento y un 6,67% en indiferente en relación con el “conocimiento de uso de IA”, y para parcial conocimiento en “conocimiento procesos de derivación” se obtiene un 3,33% en total desconocimiento, 3,33% en parcial desconocimiento, 10% en indiferente y un 6,67% en parcial conocimiento en relación con el “conocimiento de uso de IA”. En resumen, 16 encuestados (un 53,33%) respondieron a tener un total desconocimiento. Por el contrario, se determinó que solo un 16,67% (es decir, 5 personas) reconocieron tener conocimientos parciales.

Figura 12

Conocimiento proceso de derivación vs. toma de decisiones

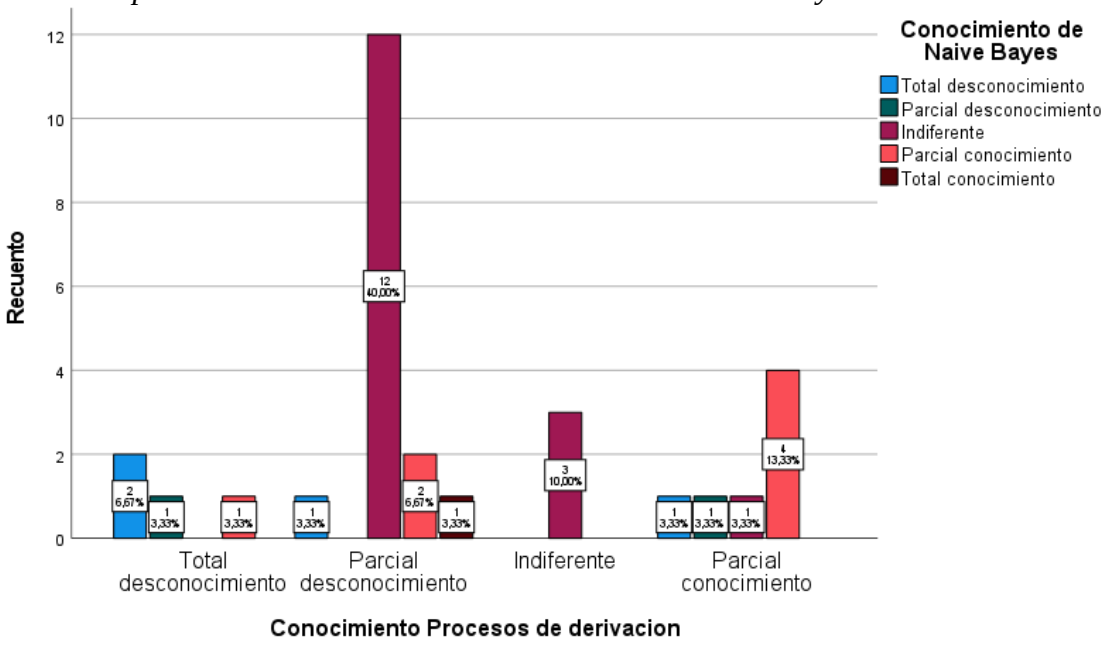


Nota. Gráfico de barras para la asociación entre la variable conocimientos proceso de derivación vs. toma de decisiones. La elaboración es propia y la fuente obtenida de las encuestas realizadas.

De acuerdo con los datos establecidos en la *Figura 12*, es posible apreciar los resultados obtenidos sobre el conocimiento de procesos de derivación y su relación con la toma de decisiones, en donde 17 encuestados (56,66%) respondieron estar parcial de acuerdo con este tema. Por otra parte, 10 encuestados (33,33%) respondieron a estar en total acuerdo con respecto con la pregunta relacionada a la toma de decisiones.

Figura 13

Conocimiento procesos de derivación vs. conocimiento de Naive Bayes

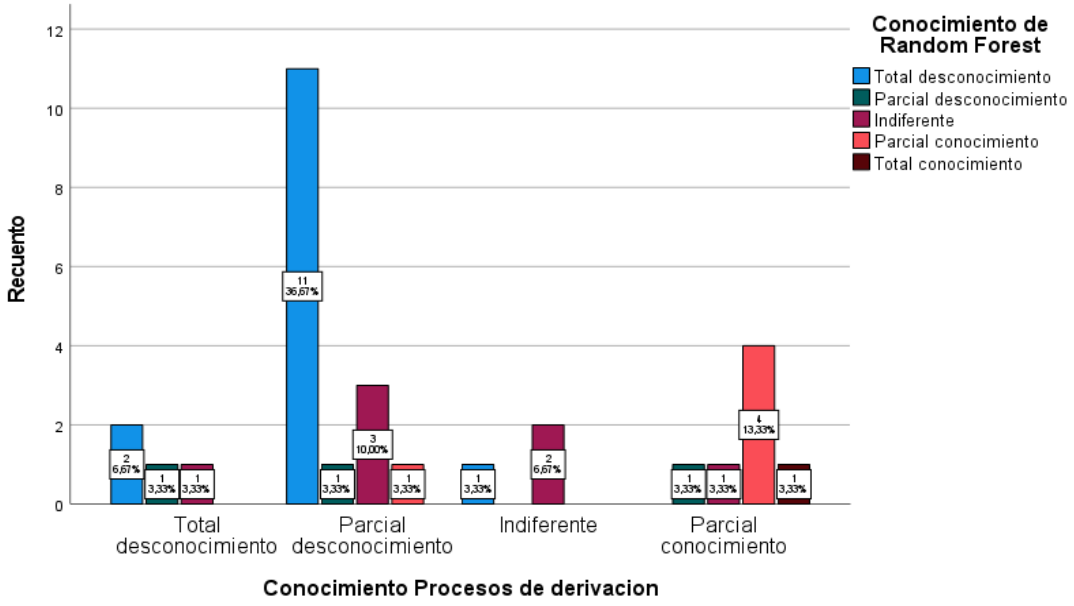


Nota. Gráfico de barras para la asociación entre la variable conocimientos proceso de derivación vs. conocimiento de Naive Bayes. La elaboración es propia y la fuente obtenida de las encuestas realizadas.

A partir de lo establecido en la *Figura 13*, es posible apreciar los resultados obtenidos sobre el conocimiento sobre los procesos de derivación y el conocimiento del algoritmo Naive Bayes, en donde 15 (50%) encuestados son indiferentes a este tema, 4 (13,33%) de los encuestados definieron estar con un total desconocimiento, 2 (6,66%) encuestados en parcial desconocimiento, 7 (23,33%) encuestados en parcial conocimiento y solo 1 (3,33%) encuestado definió un total conocimiento. Esto indica que existe una gran parte de las personas encuestadas que desconocen, tienen muy pocos conocimientos e indiferencia sobre este tema.

Figura 14

Conocimiento procesos de derivación vs. conocimiento de Random Forest



Nota. Gráfico de barras para la asociación entre la variable conocimientos proceso de derivación vs. Conocimiento de Random Forest. La elaboración es propia y la fuente obtenida de las encuestas realizadas.

De acuerdo con los datos establecidos en la *Figura 14*, es posible apreciar los resultados obtenidos sobre los conocimientos del algoritmo Random Forest y los procesos de derivación, 15 (46,67%) encuestados respondieron tener un total desconocimiento, 3 (9,99%) encuestados parcial desconocimiento, 5 (23,33%) encuestados indiferente, 5 (16,66%) encuestados respondieron tener un parcial conocimiento y solo 1 (3,33%) encuestado tiene un total conocimiento. Esto indica que, si bien existe un total desconocimiento por este tema, a un pequeño número de personas que tiene nociones básicas sobre el algoritmo Random Forest.

Conocimientos sobre uso de IA vs. las demás variables. De las variables estudiadas, es posible identificar las diferencias estadísticas entre cada una de ellas cuando las mismas están asociadas a la variable "Conocimiento de uso de IA" visualizadas en la *Tabla 16*, en la cual a cada una de las variables se le aplicó la prueba o test de Chi-cuadrado. Para esta prueba, se parte del establecimiento de la hipótesis nula (H_0), la cual determina que las variables son independientes (es decir, no están correlacionadas). Por otra parte, se contrasta el planteamiento

de una hipótesis alternativa (H_a) la cual determina que las variables son dependientes (es decir, están relacionadas). Por lo tanto, es posible determinar que si $\alpha = 0.05$ y el valor de p (significación asintótica bilateral) es menor al valor de α, entonces se rechaza la hipótesis nula y se acepta la alterna, es decir, significa que las variables están correlacionadas.

Tabla 16

Pruebas de chi-cuadrado. Conocimientos sobre uso de IA vs. las demás variables

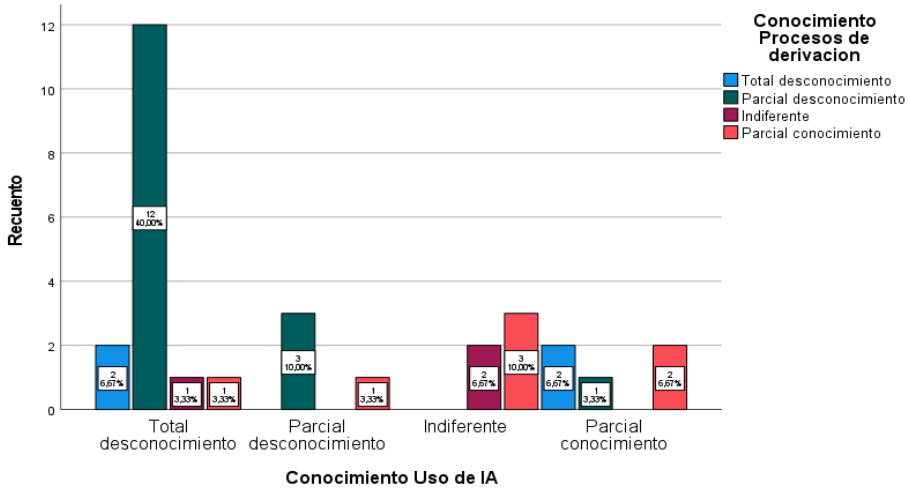
	Conocimientos sobre uso de IA			
Características de la muestra	***X²***	***g.l.***	***p***	***n***
Conocimientos procesos de derivación	20,451	9	0,015	30
Manejo de big data	22,496	12	0,032	30
Conocimiento de algoritmos	20,463	12	0,059	30
Toma de decisiones	16,709	9	0,053	30
Conocimiento de Naive Bayes	17,344	12	0,137	30
Conocimiento de Random Forest	30,243	12	0,003	30

Nota. Valor de la prueba de Chi-cuadrado de Pearson para la asociación entre la variable conocimientos uso de IA y las demás variables de estudio. La elaboración es propia y la fuente obtenida de las encuestas realizadas.
X² Chi- cuadrado // g.l. Grados de libertad // p Significación asintótica (bilateral) // n número de encuestados de la muestra. La elaboración es propia y la fuente es obtenida de las encuestas realizadas.

Por lo tanto, para el estudio de las variables, se determina que tres de las seis categorías examinadas (conocimientos procesos de derivación, manejo de big data y conocimiento de Random Forest) tienen relación o están correlacionados con la variable principal "conocimientos de uso de IA", al encontrar que el valor de p (significancia asintótica bilateral) es menor a 0,05.

Figura 15

Conocimiento sobre uso de IA vs. conocimiento procesos de derivación

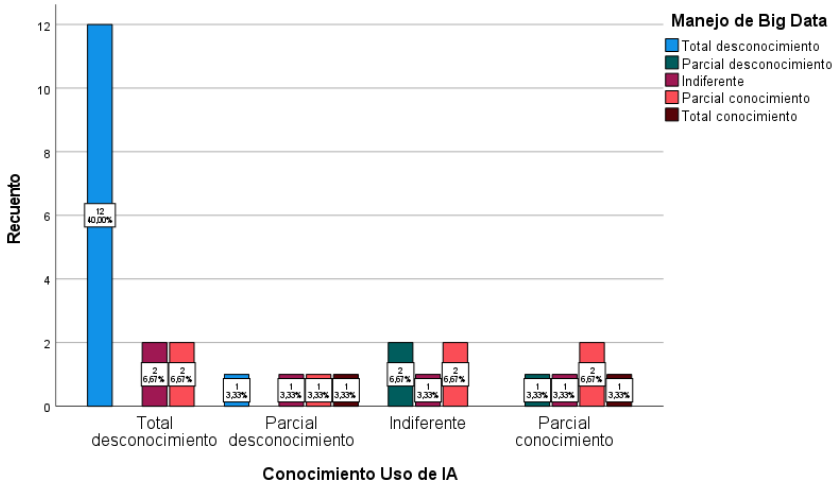


Nota. Gráfico de barras para la asociación entre la variable conocimientos sobre uso de IA vs. conocimiento de procesos de derivación. La elaboración es propia y la fuente obtenida de las encuestas realizadas.

De acuerdo con los datos establecidos en la *Figura 15*, es posible apreciar los resultados obtenidos sobre los conocimientos de uso IA y los conocimientos de procesos de derivación, 4 (13,34%) encuestados con total desconocimiento, 16 (53,3%) encuestados respondieron tener un parcial desconocimiento, 1 (3,33%) encuestado es indiferente y 7 (23,33%) encuestados respondieron tener un parcial conocimiento. Esto indica que, si bien existe desconocimiento, también existe un porcentaje considerable de personas con conocimientos sobre los procesos de derivación y uso de IA.

Figura 16

Conocimiento sobre uso de IA vs. manejo de big data

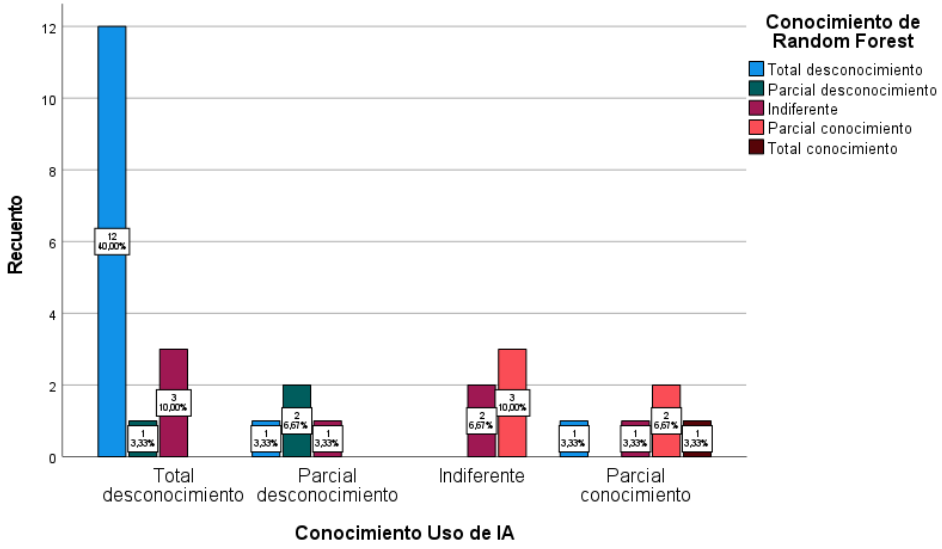


Nota. Gráfico de barras para la asociación entre la variable sobre uso de IA vs. manejo de big data. La elaboración es propia y la fuente obtenida de las encuestas realizadas.

Según los datos establecidos en la *Figura 16*, es posible apreciar los resultados obtenidos sobre los conocimientos sobre uso de IA y manejo de big data arrojando los siguientes resultados: 12 (40%) encuestados respondieron tener un total desconocimiento, 3 (10%) encuestados tienen parcial desconocimiento, 5 (16,66%) encuestados son indiferentes, 7 (23,34%) encuestados un parcial conocimiento y 2 (6,66%) encuestados respondieron tener un total conocimiento.

Figura 17

Conocimiento sobre uso de IA vs. conocimiento Random Forest



Nota. Gráfico de barras para la asociación entre la variable Conocimiento sobre uso de IA vs. conocimiento Random Forest. La elaboración es propia y la fuente obtenida de las encuestas realizadas.

Según la *Figura 17*, 14 (46,66%) encuestados tiene un total desconocimiento sobre la utilización de IA en relación el algoritmo Random Forest, 3 (10%) encuestados tienen parcial desconocimiento, 5 (16,67%) encuestados respondieron a tener conocimientos parciales, 7 (23,33%) encuestados son indiferentes y solo 1 (3,33%) encuestado definió tener un conocimiento total sobre el tema.

Contraste de hipótesis

Conocimiento procesos de derivación vs. las demás variables

Análisis de independencia de las variables. Conocimiento procesos de derivación en relación con el conocimiento del uso de IA

H_0: No hay asociación entre las variables conocimientos procesos de derivación y el conocimiento sobre IA.

H_a: Hay asociación entre las variables conocimientos procesos de derivación y el conocimiento sobre IA.

En este caso, para un nivel de significancia de $\alpha = 0.05$, se rechaza H_0 a favor de Ha si p-valor < 0.05. De acuerdo con los resultados del análisis chi-cuadrados obtenidos y representados en la *Tabla 15*, el valor p (significación asintótica bilateral) es de 0,015, lo que significa que es menor al valor α. Por lo tanto, es posible inferir y determinar que estas dos variables están relacionadas entre sí; es decir, los conocimientos de procesos de derivación están relacionados y dependen de los conocimientos de uso de la IA, con un chi-cuadrado de 20,451 con 9 grados de libertad para la muestra de 30 personas estudiadas.

Análisis de independencia de las variables. Conocimiento procesos de derivación en relación con toma de decisiones.

H_0: Significa que no hay asociación entre las variables conocimientos procesos de derivación y la toma de decisiones.

H_a: Significa que existe asociación entre las variables conocimientos procesos de derivación y la toma de decisiones.

En este caso, para un nivel de significancia de $\alpha = 0.05$, se rechaza H_0 a favor de Ha si p-valor < 0.05. De acuerdo con los resultados del análisis de chi-cuadrado obtenido y representado en la *Tabla 15*, el valor p (significación asintótica bilateral) es de 0,005, lo que significa que es menor al valor α. Por lo tanto, es posible inferir y determinar que las variables

están relacionadas entre sí; es decir, la variable de conocimientos de procesos de derivación está relacionada y depende de la toma de decisiones, con un chi-cuadrado de 23,810 con 9 grados de libertad para la muestra de 30 personas estudiadas.

Análisis de independencia de las variables. Conocimiento procesos de derivación en relación con conocimientos sobre Naive Bayes

H_0: Significa que no hay asociación entre las variables conocimientos procesos de derivación y conocimientos sobre Naive Bayes.

H_a: Significa que existe asociación entre las variables conocimientos procesos de derivación y conocimientos sobre Naive Bayes.

En este caso, para un nivel de significancia de $\alpha = 0.05$, se rechaza H_0 a favor de Ha si p-valor < 0.05. De acuerdo con los resultados del análisis de chi-cuadrado obtenido y representado en la *Tabla 15*, el valor p (significación asintótica bilateral) es de 0,043, lo que significa que es menor al valor α. Por lo tanto, es posible inferir y determinar que las variables están relacionadas entre sí; es decir, la variable de conocimientos de procesos de derivación está relacionada y depende de los conocimientos sobre Naive Bayes, con un chi-cuadrado de 21,515 con 12 grados de libertad para la muestra de 30 personas estudiadas.

Análisis de independencia de las variables. Conocimiento procesos de derivación en relación con conocimientos sobre Random Forest.

H_0: Significa que no hay asociación entre las variables conocimientos procesos de derivación y conocimientos sobre Random Forest.

H_a: Significa que existe asociación entre las variables conocimientos procesos de derivación y conocimientos sobre Random Forest.

En este caso, para un nivel de significancia de $\alpha = 0.05$, se rechaza H_0 a favor de Ha si p-valor < 0.05. De acuerdo con los resultados del análisis de chi-cuadrado obtenido y representado en la *Tabla 15*, el valor p (significación asintótica bilateral) es de 0,040, lo que

significa que es menor al valor α. Por lo tanto, es posible inferir y determinar que las variables están relacionadas entre sí; es decir, la variable de conocimientos de procesos de derivación está relacionada y depende de los conocimientos sobre Random Forest, con un chi-cuadrado de 21,800 con 12 grados de libertad para la muestra de 30 personas estudiadas.

Conocimiento sobre uso de IA vs. las demás variables

Análisis de independencia de las variables. Conocimiento sobre uso de IA en relación con el conocimiento procesos de derivación.

H_0: No hay asociación entre las variables conocimientos procesos de derivación y el conocimiento sobre IA.

H_a: Hay asociación entre las variables conocimientos procesos de derivación y el conocimiento sobre IA.

En este caso, para un nivel de significancia de $\alpha = 0.05$, se rechaza H_0 a favor de Ha si p-valor < 0.05. De acuerdo con los resultados del análisis chi-cuadrados obtenidos y representados en la *Tabla 16*, el valor p (significación asintótica bilateral) es de 0,0032, lo que significa que es menor al valor α. Por lo tanto, es posible inferir y determinar que estas dos variables están relacionadas entre sí; es decir, los conocimientos de uso de IA están relacionados y dependen de los conocimientos de procesos de derivación, con un chi-cuadrado de 20,451 con 9 grados de libertad para la muestra de 30 personas estudiadas.

Análisis de independencia de las variables. Conocimiento sobre uso de IA en relación con el manejo de big data.

H_0: No hay asociación entre las variables conocimientos sobre IA y el manejo de big data.

H_a: Hay asociación entre las variables conocimientos sobre IA y el manejo de big data.

En este caso, para un nivel de significancia de $\alpha = 0.05$, se rechaza H_0 a favor de Ha si p-valor < 0.05. De acuerdo con los resultados del análisis chi-cuadrados obtenidos y representados en la *Tabla 16*, el valor p (significación asintótica bilateral) es 0,032, lo que

significa que es menor al valor α. Por lo tanto, es posible inferir y determinar que estas dos variables están relacionadas entre sí; es decir, los conocimientos de uso de IA están relacionados y dependen de los conocimientos de manejo de big data, con un chi-cuadrado de 22,496 con 12 grados de libertad para la muestra de 30 personas estudiadas.

Análisis de independencia de las variables. Conocimiento sobre uso de IA en relación con el conocimiento de Random Forest.

H_0: No hay asociación entre las variables conocimientos sobre IA y el conocimiento sobre Random Forest.

H_a: Hay asociación entre las variables conocimientos sobre IA y el conocimiento sobre Random Forest.

En este caso, para un nivel de significancia de $\alpha = 0.05$, se rechaza H_0 a favor de Ha si p-valor < 0.05. De acuerdo con los resultados del análisis chi-cuadrados obtenidos y representados en la *Tabla 16*, el valor p (significación asintótica bilateral) es de 0,003, lo que significa que es menor al valor α. Por lo tanto, es posible inferir y determinar que estas dos variables están relacionadas entre sí; es decir, los conocimientos de uso de IA están relacionados y dependen de los conocimientos de Random Forest con un chi-cuadrado de 30,243 con 12 grados de libertad para la muestra de 30 personas estudiadas.

Análisis de las tablas de contingencia

De acuerdo con Herrera (2018), las tablas de contingencia son utilizadas para:

> Establecer si estas relaciones son estadísticamente dependientes o en su defecto independientes, se efectúa el contraste estadístico Chi-cuadrado, prueba que se aplica para comparar si las dos características cualitativas están relacionadas estadísticamente entre sí. Examinando la diferencia que hay entre los valores observados en las celdas y los que se habrían obtenido en el supuesto de no asociación entre las variables (p. 5).

Es decir, el análisis de las tablas de contingencia consiste en la interpretación de la asociación entre los valores porcentuales de las variables empleadas.

Tabla 17

Toma de decisiones vs. conocimiento procesos de derivación

			Conocimiento procesos de derivación				Total
			Total desconocimiento	Parcial desconocimiento	Indiferente	Parcial conocimiento	
Toma de decisiones	Parcial desacuerdo	Recuento	1	0	0	0	1
		% del total	3,3%	0,0%	0,0%	0,0%	3,3%
	Indiferente	Recuento	0	0	1	0	1
		% del total	0,0%	0,0%	3,3%	0,0%	3,3%
	Parcial acuerdo	Recuento	0	13	1	3	17
		% del total	0,0%	43,3%	3,3%	10,0%	56,7%
	Total acuerdo	Recuento	3	3	1	4	11
		% del total	10,0%	10,0%	3,3%	13,3%	36,7%
Total		Recuento	4	16	3	7	30
		% del total	13,3%	53,3%	10,0%	23,3%	100,0 %

Nota. Tabla de contingencia de la variable toma de decisiones y conocimiento de procesos de derivación. La elaboración es propia y la fuente obtenida de las encuestas realizadas.

A partir de la interpretación de estas dos variables (Toma de decisiones vs. conocimiento procesos de derivación) en la *Tabla 17* es posible apreciar que existe un 43,3% (13 personas) que están parcialmente de acuerdo en que existe un parcial desconocimiento entre los conocimientos de derivación y la toma de decisiones. Este porcentaje de personas representa casi la mitad de los encuestados.

Tabla 18

Toma de decisiones vs. conocimiento uso de IA

			Conocimiento uso de IA				Total
			Total desconocimiento	Parcial desconocimiento	Indiferente	Parcial conocimiento	
Toma de decisiones	Parcial desacuerdo	Recuento	0	0	0	1	1
		% del total	0,0%	0,0%	0,0%	3,3%	3,3%
	Indiferente	Recuento	0	0	1	0	1
		% del total	0,0%	0,0%	3,3%	0,0%	3,3%

	Parcial acuerdo	Recuento	11	3	3	0	17
		% del total	36,7%	10,0%	10,0%	0,0%	56,7 %
	Total acuerdo	Recuento	5	1	1	4	11
		% del total	16,7%	3,3%	3,3%	13,3%	36,7 %
Total		Recuento	16	4	5	5	30
		% del total	53,3%	13,3%	16,7%	16,7%	100,0 %

Nota. Tabla de contingencia de la variable toma de decisiones y conocimiento de uso de IA. La elaboración es propia y la fuente obtenida de las encuestas realizadas.

A partir del análisis de la tabla de contingencia "Toma de decisiones vs. conocimiento uso de IA" visualizado en la *Tabla 18*, es posible apreciar que existen 11 personas que están parcialmente de acuerdo en que existe un total desconocimiento entre los conocimientos del uso de inteligencia artificial (IA) y la toma de decisiones. Este porcentaje de personas representa un 36,7%. Por lo tanto, un gran número de encuestados están de acuerdo con esta afirmación.

Tabla 19

Toma de decisiones vs. manejo de big data

			Manejo de big data					**Total**
			Total desconocimiento	**Parcial desconocimiento**	**Indiferente**	**Parcial conocimiento**	**Total conocimiento**	
Toma de decisiones	Parcial desacuerdo	Recuento	0	0	1	0	0	1
		% del total	0,0%	0,0%	3,3%	0,0%	0,0%	3,3%
	Indiferente	Recuento	0	0	1	0	0	1
		% del total	0,0%	0,0%	3,3%	0,0%	0,0%	3,3%
	Parcial acuerdo	Recuento	11	1	1	3	1	17
		% del total	36,7%	3,3%	3,3%	10,0%	3,3%	56,7 %
	Total acuerdo	Recuento	2	2	2	4	1	11
		% del total	6,7%	6,7%	6,7%	13,3%	3,3%	36,7 %
Total		Recuento	13	3	5	7	2	30
		% del total	43,3%	10,0%	16,7 %	23,3%	6,7%	100%

Nota. Tabla de contingencia de la variable toma de decisiones y manejo de big data. La elaboración es propia y la fuente obtenida de las encuestas realizadas.

A partir del análisis de la tabla de contingencia “Toma de decisiones vs. Manejo de big data” visualizado en la *Tabla 19*, es posible apreciar que existen 11 personas que están parcialmente de acuerdo en que existe un total desconocimiento entre los conocimientos del uso de inteligencia artificial (IA) y su relación con la toma de decisiones sobre los nuevos casos de COVID-19. Este porcentaje de personas representa un 36,7%, lo cual es una gran parte de los encuestados. Por otra parte, 4 encuestados (13,3%) estuvieron totalmente de acuerdo en el manejo de big data y su utilización para la toma de decisiones de los nuevos casos de COVID-19.

Tabla 20

Toma de decisiones vs. conocimiento de algoritmos.

			Conocimiento de algoritmos					
			Naive Bayes	**Regresión Logística**	**Random Forest**	**Redes Neuronales**	**No conozco ninguno**	**Total**
Toma de decisiones	Parcial desacuerdo	Recuento	0	0	1	0	0	1
		% del total	0,0%	0,0%	3,3%	0,0%	0,0%	3,3%
	Indiferente	Recuento	1	0	0	0	0	1
		% del total	3,3%	0,0%	0,0%	0,0%	0,0%	3,3%
	Parcial acuerdo	Recuento	8	3	1	4	1	17
		% del total	26,7%	10,0%	3,3%	13,3%	3,3%	56,7%
	Total acuerdo	Recuento	2	2	1	4	2	11
		% del total	6,7%	6,7%	3,3%	13,3%	6,7%	36,7%
Total		Recuento	11	5	3	8	3	30
		% del total	36,7%	16,7%	10,0%	26,7%	10,0%	100,0%

Nota. Tabla de contingencia de la variable toma de decisiones y conocimientos de algoritmos. La elaboración es propia y la fuente obtenida de las encuestas realizadas.

Como se puede apreciar en la *Tabla 20*, un 56,7% de los encuestados estuvieron en parcial acuerdo sobre el conocimiento y la utilización de algoritmos para la toma de decisiones. Un 26,7% (8 personas) respondieron tener conocimientos sobre el algoritmo Naive Bayes, un

10% (3 personas) sobre regresión logística, un 3,3% sobre Random Forest y 4 personas (13,3%) sobre redes neuronales. Lo cual indica que existe un cierto conocimiento sobre de estos algoritmos.

Tabla 21

Toma de decisiones vs. conocimiento Naive Bayes.

			Conocimiento de Naive Bayes					
			Total desconocimiento	**Parcial desconocimiento**	**Indiferente**	**Parcial conocimiento**	**Total conocimiento**	**Total**
Toma de decisiones	Parcial desacuerdo	Recuento	0	1	0	0	0	1
		% del total	0,0%	3,3%	0,0%	0,0%	0,0%	3,3%
	Indiferente	Recuento	0	0	1	0	0	1
		% del total	0,0%	0,0%	3,3%	0,0%	0,0%	3,3%
	Parcial acuerdo	Recuento	1	1	12	2	1	17
		% del total	3,3%	3,3%	40,0%	6,7%	3,3%	56,7 %
	Total acuerdo	Recuento	3	0	3	5	0	11
		% del total	10,0%	0,0%	10,0%	16,7%	0,0%	36,7 %
Total		Recuento	4	2	16	7	1	30
		% del total	13,3%	6,7%	53,3%	23,3%	3,3%	100 %

Nota. Tabla de contingencia de la variable toma de decisiones y conocimiento de Naive Bayes. La elaboración es propia y la fuente obtenida de las encuestas realizadas.

De acuerdo con la *Tabla 21*, un 40% de los encuestados estuvieron en parcial acuerdo sobre su indiferencia sobre los conocimientos del algoritmo Naive Bayes para la toma de decisiones. Por otro lado, 5 personas (16,7%) definieron tener conocimientos parciales sobre Naive Bayes, estando totalmente de acuerdo sobre la toma de decisiones.

Tabla 22

Toma de decisiones vs. conocimiento de Random Forest

			Conocimiento de Random Forest					Total
			Total desconocimiento	**Parcial desconocimiento**	**Indiferente**	**Parcial conocimiento**	**Total conocimiento**	
Toma de decisiones	Parcial desacuerdo	Recuento	0	0	1	0	0	1
		% del total	0,0%	0,0%	3,3%	0,0%	0,0%	3,3%
	Indiferente	Recuento	0	0	1	0	0	1
		% del total	0,0%	0,0%	3,3%	0,0%	0,0%	3,3%
	Parcial acuerdo	Recuento	11	2	2	2	0	17
		% del total	36,7%	6,7%	6,7%	6,7%	0,0%	56,7%
	Total acuerdo	Recuento	3	1	3	3	1	11
		% del total	10,0%	3,3%	10,0%	10,0%	3,3%	36,7%
Total		Recuento	14	3	7	5	1	30
		% del total	46,7%	10,0%	23,3%	16,7%	3,3%	100%

Nota. Tabla de contingencia de la variable toma de decisiones y conocimiento de Random Forest. La elaboración es propia y la fuente obtenida de las encuestas realizadas.

Como se aprecia en la *Tabla 22*, del total de los encuestados, 11 personas están parcialmente de acuerdo en la toma de decisiones en el hospital con un total desconocimiento sobre el algoritmo Random Forest, lo cual representa una gran parte de las decisiones tomadas (36,7%). Por otra parte, un 56,7% (17 personas) fue el total de encuestados que respondieron a estar parcialmente de acuerdo con la toma de decisiones, ya sea teniendo total desconocimiento, parcial desconocimiento, indiferencia o parcial conocimiento sobre este algoritmo.

Fuentes de conocimiento

Fuentes de conocimiento: Este apartado se basa en el conocimiento obtenido por medio de un experto en el área de salud perteneciente a un hospital público de la ciudad de Guayaquil. Además, la información proporcionada por diversos artículos científicos encontrados en páginas confiables y verificadas, los cuales están totalmente relacionados al tema de derivación hospitalaria y ambulatoria de pacientes infectados con COVID-19.

Experto en el área: Un experto en el área de salud forma parte importante para la realización del presente proyecto, al ser un profesional con varios años en el área y por consiguiente un gran conocimiento adquirido, se logra conseguir información más amplia y precisa con respecto a los factores asociados a la derivación hospitalaria y ambulatoria de personas infectadas por COVID-19 y conocer qué criterios utilizan para la respectiva toma de decisiones de cada paciente.

Operatividad de las variables

Tabla 23

Operatividad de las variables.

Conceptualización	Dimensiones	Indicadores	Técnicas y/o instrumentos
Variable Dependiente	Derivación Hospitalaria o Ambulatoria	0 – Ambulatoria 1 - Hospitalaria	Python
Variables Independientes (Factores asociados / Sintomatología)	Dificultad Respiratoria	0 – No 1 – Si	dataset
	Saturación de Oxígeno	≥ 70% & ≤ 96%	dataset
	Dolor Abdominal	0 – No 1 – Si	dataset
	Mialgia	0 – No 1 – Si	dataset
	Tos	0 – No 1 – Si	dataset
	Temperatura	Grados Cº	dataset
	Pérdida de Olfato	0 – No 1 – Si	dataset
	Pérdida de Apetito	0 – No 1 – Si	dataset

Nota. Se presenta la operatividad de cada una de las variables para el algoritmo predictor. La elaboración es propia y la fuente es proporcionada por las investigaciones realizadas.

Estructura del dataset

Para la realización del dataset óptimo para la predicción, se obtuvo información mediante una entrevista vía Zoom con el experto en el área para conocer, en primer lugar, las

variables adecuadas para la respectiva derivación. Posteriormente, se facilitó una base de datos de los historiales clínicos de 20 pacientes infectados por COVID-19, correspondientes al periodo 2020 – 2021, cuya derivación se la realizó enfocándose con mayor importancia en la sintomatología inicial del paciente. Las variables quedan de la siguiente manera:

- **Dificultad Respiratoria:** Falta de aire que puede llegar a ser aguda si se llega a acumular líquido en el área de los pulmones.
- **Saturación de Oxígeno:** Cantidad de oxígeno encontrado en la sangre, no debe ser menor a 94%.
- **Dolor Abdominal:** Síntoma provocado por enfermedades gastrointestinales debido al virus.
- **Mialgia:** Dolor muscular.
- **Tos:** Tos seca es la más común en los síntomas del COVID-19, éste puede llegar a ser tan irritante hasta el punto de producir otros síntomas como lo es el dolor de garganta o pecho.
- **Temperatura:** La temperatura según el COVID-19 es preocupante si llega ser mayor a los 37º.
- **Pérdida de Olfato:** El virus ataca de manera directa las neuronas situadas en el bulbo olfatorio, dando como resultado una disminución del olfato del infectado.
- **Pérdida de Apetito:** Debido a la infección causada por el virus en otras partes del cuerpo.
- **Derivación Hospitalaria o Ambulatoria:** Determinar de acuerdo con la sintomatología del paciente si este necesita de la intervención hospitalaria o ambulatoria de acuerdo con su nivel de grav edad.

Metodología del desarrollo del prototipo

El objetivo del presente proyecto es la creación de un prototipo predictivo, por consiguiente, se ha tomado en cuenta la metodología del desarrollo del prototipo como base fundamental para la explicación detallada del mismo realizándose mediante 6 fases según lo estipula el proceso metodológico "Knowledge Discovery in Databases – KDD".

Fase 1. Importación y muestreo de datos

Los datos obtenidos para la muestra fueron extraídos de un hospital público de la ciudad de Guayaquil; entidad que facilitó la información de sus instalaciones sobre los pacientes diagnosticados con COVID-19 en el periodo correspondiente desde marzo 2020 hasta enero 2021. La muestra se conforma por un total de 20 pacientes cuya derivación fue la siguiente: 10 ambulatorios, es decir su recuperación no requiere atención especializada debido a que no presentan síntomas de gravedad y 10 hospitalizados.

Fase 2. Calidad de datos

La calidad de los datos determina la confiabilidad y el correcto performance de los algoritmos que los implementen, es así como, una vez formado la base de datos, se procedió a hacer el tratamiento correspondiente en 2 fases: identificación y preprocesamiento de los datos; procesos que garantizan resultados precisos y confiables, los cuales se explican de manera detallada en el siguiente apartado:

1. Los registros de la base de datos obtenida se revisaron de manera manual, constatando que había inconsistencias en algunos de ellos, las anomalías encontradas fueron las siguientes:
    a. **Campos vacíos:** En el dataset hay campos importantes tales como: "temperatura y saturación", campos que son variables de acuerdo con la sintomatología del paciente por lo tanto dependiendo si el infectado presentaba

esta sintomatología se indicaba el valor equivalente a ese síntoma, caso contrario se presentaba vacío.

2. Identificadas las inconsistencias en el dataset el siguiente paso fue tratarla de la manera más a adecuada, al tener una muestra que no contenía un gran volumen de datos, se procedió a tratarlos de manera manual de la siguiente manera:
    a. **Campos vacíos:** Los campos que presentaban la inconsistencia son datos cuantitativos, para solucionar este error se procedió a completar los campos con información adecuada acorde a su tipo de derivación y en base a las demás sintomatologías. Este procedimiento fue consultado con el experto del área de salud para su confiable y certera corrección.

Fase 3. Transformación

Esta fase es representada como una de las más importantes debido a que se dedica a la selección de los datos más significativos para el posterior entrenamiento de los algoritmos establecidos. Todo este proceso de transformación tiene un fin, el cual radica en la obtención de las variables que serán utilizadas en la fase de modelización por medio de la minería de datos, de tal forma que se amenoran la cantidad de datos no tan significativos para el entrenamiento.

La base de datos real inicial consistía en los siguientes atributos: “N° paciente, edad, dificultad respiratoria, saturación, presión arterial, pérdida de conocimiento, cefalea (dolor de cabeza), dolor abdominal, mialgia (dolor muscular), odinofagia (dolor de garganta), tos, ronquera, temperatura, diarrea, fatiga, pérdida de olfato, pérdida de apetito, comorbilidad, derivación”. Para seleccionar los atributos/variables más significativas se continuó con los siguientes pasos:

1. Realizando una exhaustiva investigación en fuentes confiables como lo son artículos científicos, libros, revistas científicas (Science Direct, Redalyc, Scielo, Elsevier) y del

criterio del experto, se consiguió descartar aquellas variables/atributos menos significativos en la toma de decisiones para la derivación y dejar únicamente las variables precisas para dicho proceso.

2. Paso siguiente, se creó una nueva base de datos (.csv), la cual consistía solo de aquellas variables con mayor relevancia para el proceso de derivación hospitalaria y ambulatoria. Dejando la modificación de la nueva base de datos con las siguientes variables: “Dificultad Respiratoria, Saturación, Dolor Abdominal, Mialgia, Tos, Temperatura, Pérdida de Olfato, Pérdida de Apetito, Derivación”
3. Por último, ya contando con una base de datos óptima se la exportó a un documento con extensión .xlsx (de Excel) y otra en .csv para su utilización al momento de realizar el algoritmo en Python.

Simulación de datos

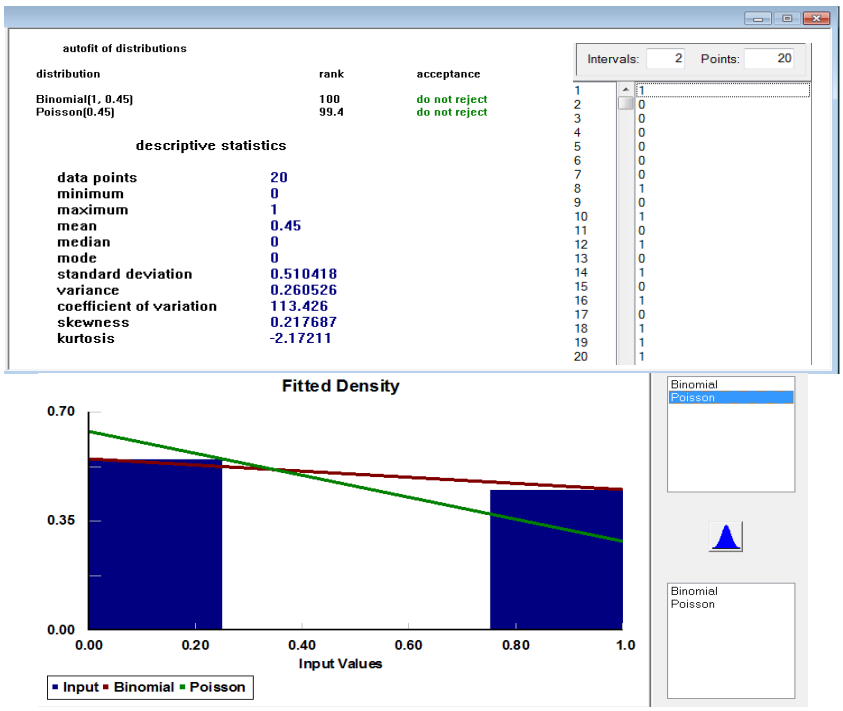
Dada por terminada la fase de la selección de las variables, se concluyó que el volumen de datos, es decir, la cantidad de pacientes derivados de manera hospitalaria y ambulatoria en la base de datos real era insuficiente. Para eso, se realizó una simulación de 1400 datos por medio de los datos reales, esto significa que se duplican los datos y se conseguirá un mejor entrenamiento para los algoritmos elegidos.

Herramientas aplicadas para la simulación de datos. Se utilizaron dos herramientas para realizar la simulación de datos, las cuales son: Microsoft Excel del programa informático Microsoft Office 365 ProPlus, éste a su vez se vincula con la herramienta del campo de desarrollo llamada Visual Basic, cuya finalidad es generar la cantidad de 1400 datos aleatorios. Esta simulación se consigue siguiendo el método de simulación Montecarlo y la herramienta @RISK 8.1 que permitirá realizar la simulación con Excel. Cabe destacar que no arroja el resultado final, derivación.

Para utilizar el método Montecarlo se debe conocer sobre las distribuciones de cada variable de la base de datos, esto se logra utilizando el software ProModel, específicamente haciendo uso de la herramienta tecnológica STAT::FIT. Al tenerse una base de datos real de 20 pacientes, esta herramienta permite la entrada de datos entre la cantidad 10 como mínimo y 50 como máximo. Dado este caso, todos los datos de la base de datos real son ingresados en la herramienta por variable, arrojando la siguiente información: Data, distribuciones, su rango y aceptación, estadística descriptiva, y el gráfico de distribución.

Figura 18

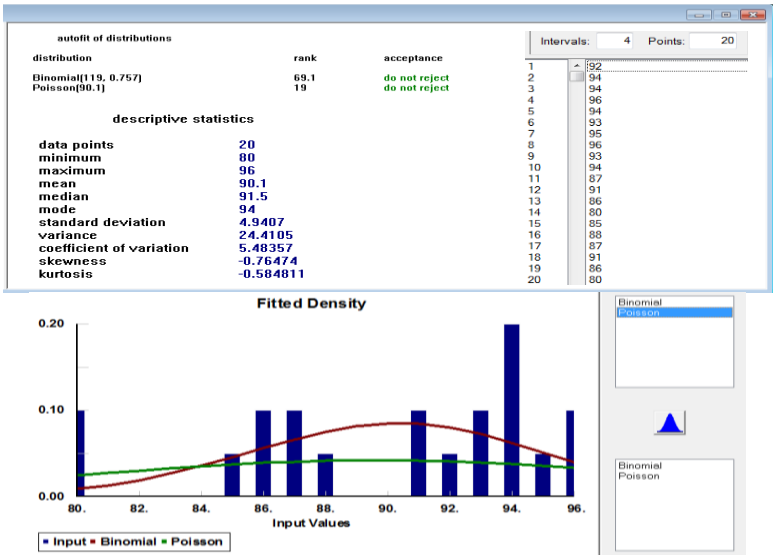
STAT::FIT. Distribución de la variable dificultad respiratoria



Nota: En el presente gráfico se muestra información detalla de la variable dificultad respiratoria del dataset: estadística descriptiva, datos, autofit de las distribuciones y el gráfico de distribución. La elaboración es propia y la fuente obtenida de las investigaciones realizadas.

Figura 19

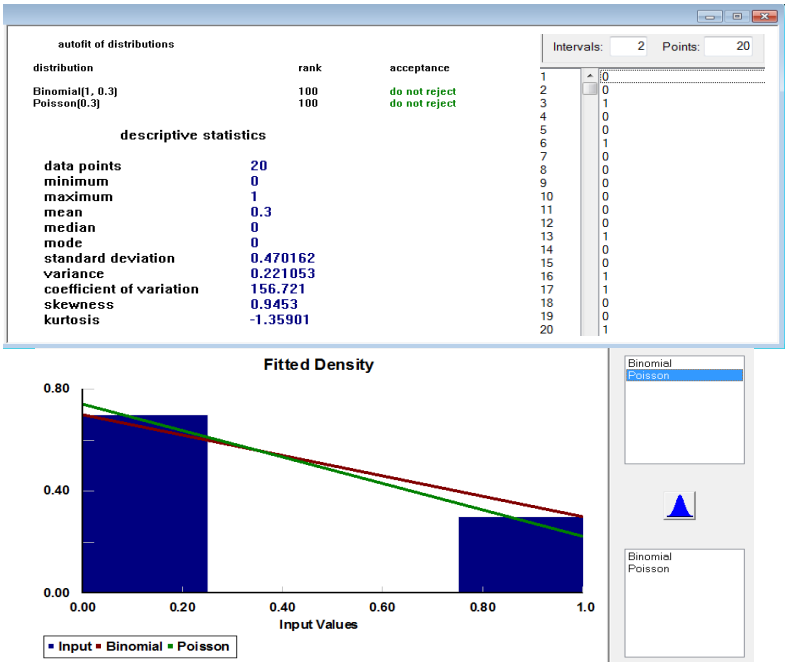
STAT::FIT. Distribución de la variable saturación



Nota: En el presente gráfico se muestra información detalla de acuerdo a las distribuciones de la variable saturación del dataset, entre ellos se encuentra la estadística descriptiva, datos, autofit de las distribuciones y el gráfico de distribución. La elaboración es propia y la fuente obtenida de las investigaciones realizadas.

Figura 20

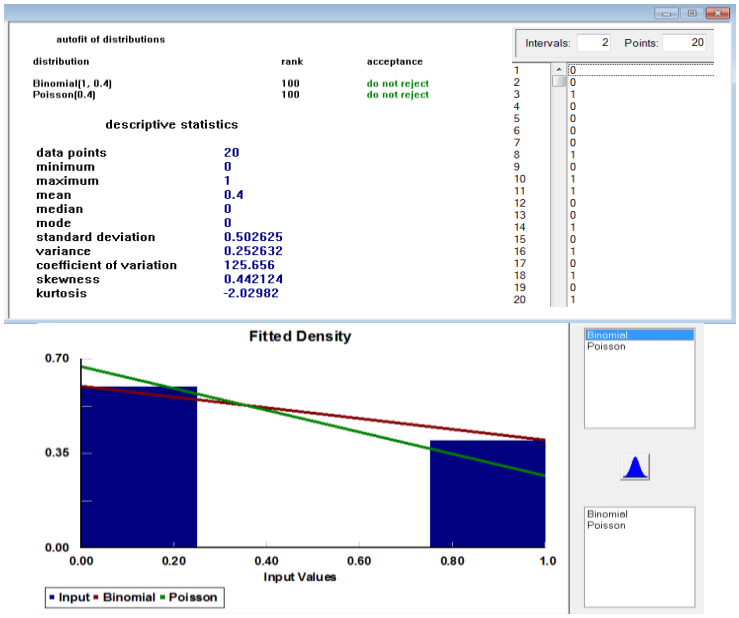
STAT::FIT. Distribución de la variable dolor abdominal



Nota: En el presente gráfico se muestra información detalla de acuerdo a las distribuciones de la variable dolor abdominal del dataset, entre ellos se encuentra la estadística descriptiva, datos, autofit de las distribuciones y el gráfico de distribución. La elaboración es propia y la fuente obtenida de las investigaciones realizadas.

Figura 21

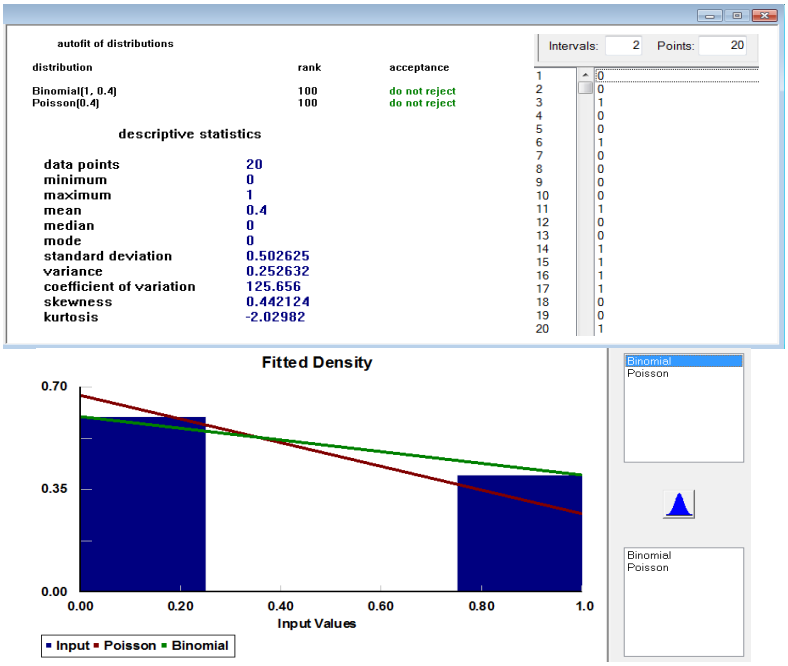
STAT::FIT. Distribución de la variable mialgia



Nota: En el presente gráfico se muestra información detalla de acuerdo a las distribuciones de la variable mialgia del dataset, entre ellos se encuentra la estadística descriptiva, datos, autofit de las distribuciones y el gráfico de distribución. La elaboración es propia y la fuente obtenida de las investigaciones realizadas.

Figura 22

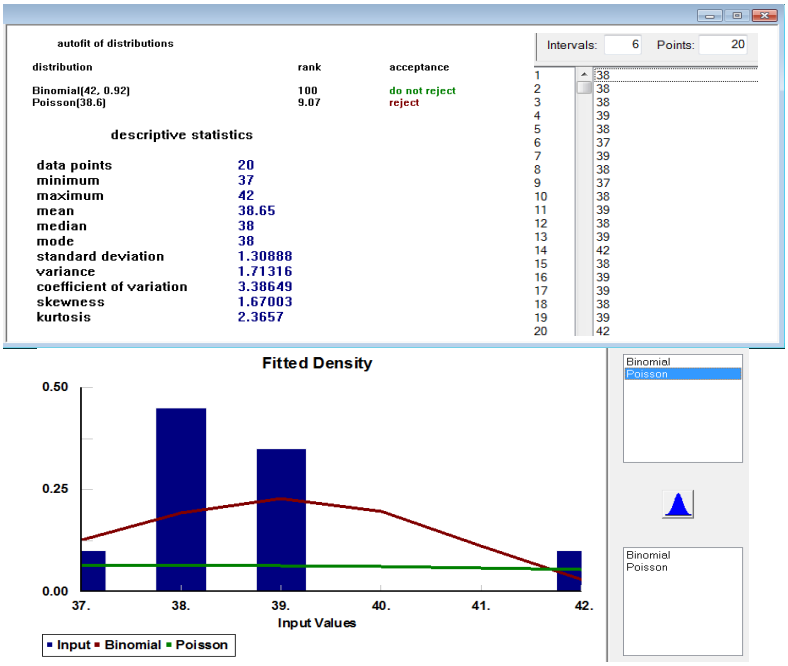
STAT::FIT. Distribución de la variable tos



Nota: En el presente gráfico se muestra información detalla de acuerdo a las distribuciones de la variable tos del dataset, entre ellos se encuentra la estadística descriptiva, datos, autofit de las distribuciones y el gráfico de distribución. La elaboración es propia y la fuente obtenida de las investigaciones realizadas.

Figura 23

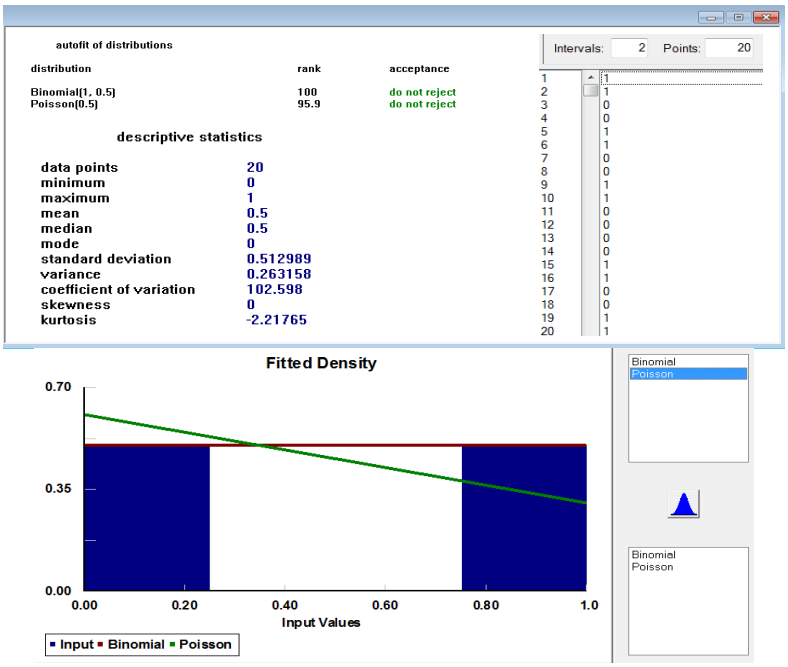
STAT::FIT. Distribución de la variable temperatura



Nota: En el presente gráfico se muestra información detalla de acuerdo a las distribuciones de la variable temperatura del dataset, entre ellos se encuentra la estadística descriptiva, datos, autofit de las distribuciones y el gráfico de distribución. La elaboración es propia y la fuente obtenida de las investigaciones realizadas.

Figura 24

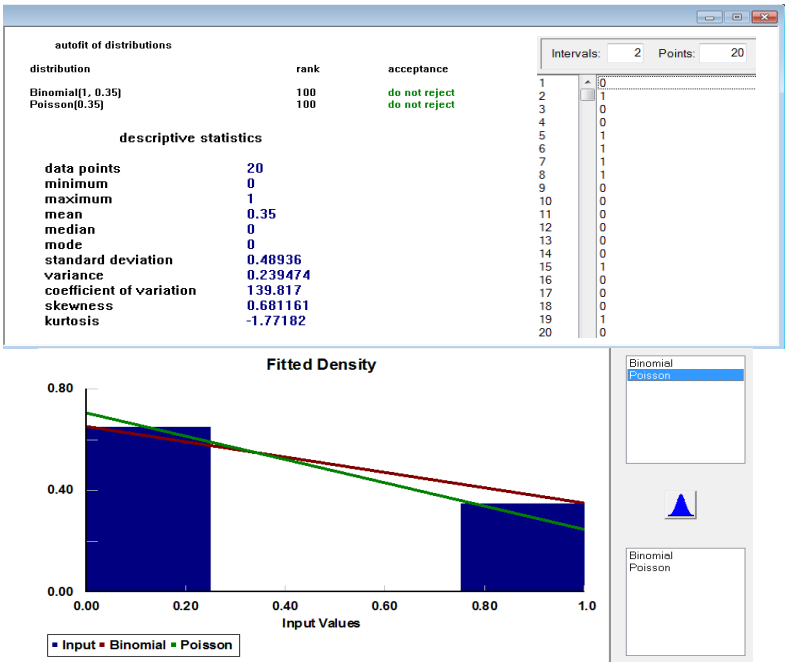
STAT::FIT. Distribución de la variable pérdida de olfato



Nota: En el presente gráfico se muestra información detalla de acuerdo a las distribuciones de la variable pérdida de olfato del dataset, entre ellos se encuentra la estadística descriptiva, datos, autofit de las distribuciones y el gráfico de distribución. La elaboración es propia y la fuente obtenida de las investigaciones realizadas.

Figura 25

STAT::FIT. Distribución de la variable pérdida de apetito



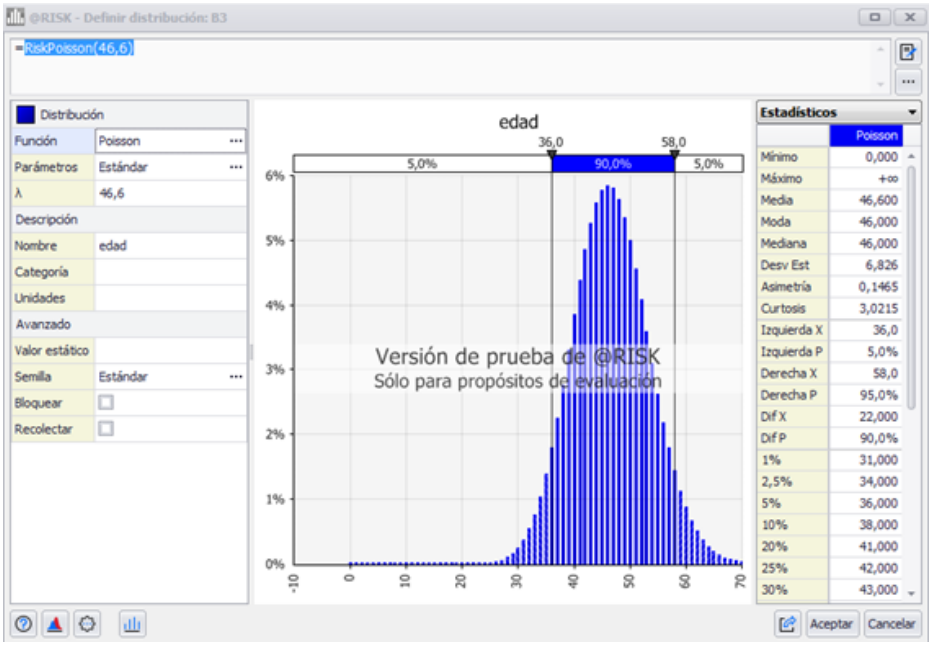
Nota: En el presente gráfico se muestra información detalla de acuerdo a las distribuciones de la variable pérdida de apetito del dataset, entre ellos se encuentra la estadística descriptiva, datos, autofit de las distribuciones y el gráfico de distribución. La elaboración es propia y la fuente obtenida de las investigaciones realizadas.

Simulación Montecarlo. Se define como una herramienta estadística que permite la modelación de resultados de acuerdo con el comportamiento histórico de los datos y de su probabilidad de ocurrencia. (Jiménez & Castro, 2018).

El método de simulación Montecarlo para emplearlo correctamente y que arroje datos precisos acorde a la derivación (dado que este no simula el resultado de la variable dependiente, derivación), se ha dividido los 20 datos en dos categorías: pacientes con derivación hospitalaria y pacientes con derivación ambulatoria. Consiguiendo dos distribuciones de cada variable por cada categoría y así, obtener la simulación de los 1400 datos más certera.

Figura 26

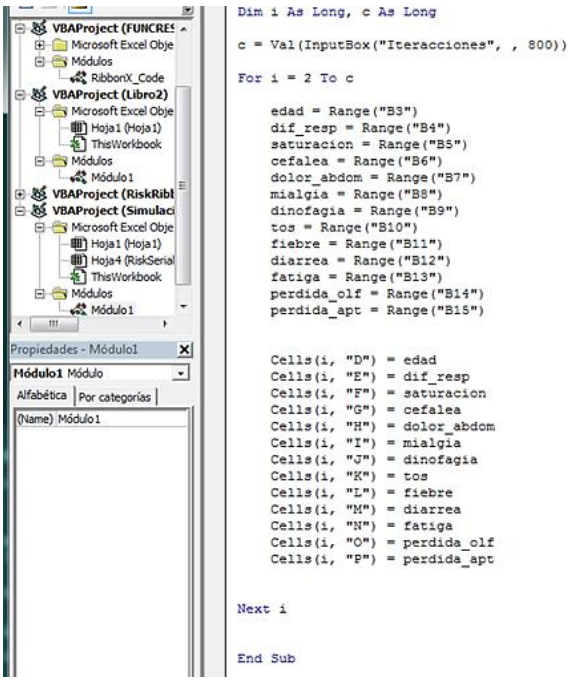
Simulación Montecarlo. Elección de la distribución de STAT::FIT en @RISK8.1



Nota: En el presente gráfico se muestra la elección de las distribuciones de STAT::FIT en @RISK8.1. La elaboración es propia y la fuente obtenida de las investigaciones realizadas.

Figura 27

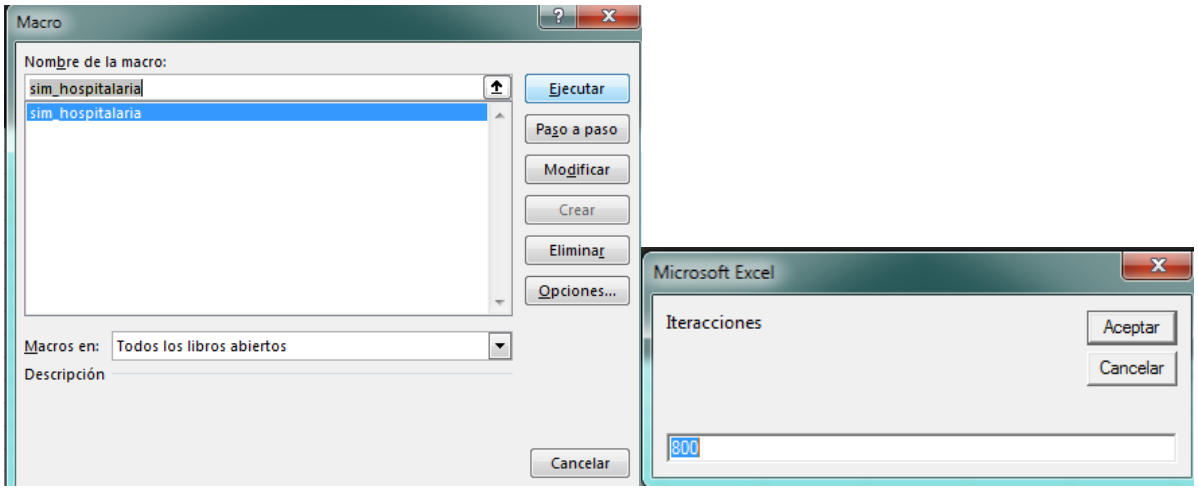
Simulación Montecarlo. Editor de Visual Studio



Nota: En el presente gráfico se muestra el código a ejecutar para la obtención de los 1400 datos. La elaboración es propia y la fuente obtenida de las investigaciones realizadas.

Figura 28

Simulación Montecarlo. Iteraciones por variables



Nota: En el presente gráfico se muestra el número de iteraciones a realizar, en este caso 1400. La elaboración es propia y la fuente obtenida de las investigaciones realizadas.

Figura 29

Simulación Montecarlo. Resultado final de la simulación

			edad	dif_resp	saturacion	cefalea	dolor_abdom	mialgia	dinofagia	tos	fiebre	diarrea	fatiga	perdida_olf	perdida_apt	derivacion
			34	0	89	1	1	0	1	0	39	1	1	0	1	1
edad	49		43	0	83	0	1	1	1	1	37	0	0	0	0	1
dif_resp	0		49	1	87	0	0	0	1	0	38	0	0	0	0	1
saturacion	80		39	1	84	1	1	1	1	1	38	0	1	1	0	1
cefalea	0		52	1	88	1	0	1	1	0	42	1	1	1	0	1
dolor_abdom	0		50	1	90	0	0	1	0	0	40	1	1	0	0	1
mialgia	1		69	0	87	1	0	0	0	0	37	0	0	0	0	1
dinofagia	0		49	0	90	0	1	0	1	1	40	1	1	0	0	1
tos	0		48	0	86	0	1	1	1	0	41	0	1	0	0	1
fiebre	42		62	1	94	1	0	0	1	0	41	0	1	1	0	1
diarrea	0		48	1	81	1	0	0	1	0	40	0	0	1	1	1
fatiga	1		45	1	91	1	0	1	1	0	37	1	1	1	0	1
perdida_olf	0		43	1	84	0	1	1	0	0	36	0	0	0	1	1
perdida_apt	0		35	0	87	1	0	1	1	1	39	0	1	0	0	1
			39	0	79	0	0	1	0	0	41	0	0	0	0	1
			47	0	78	0	1	1	0	0	38	0	0	0	1	1
			44	0	79	1	1	1	1	1	40	0	0	1	0	1
			44	0	89	1	1	0	1	0	38	1	0	1	0	1
			39	1	85	1	0	1	0	0	38	0	1	1	1	1
			47	0	84	1	0	1	1	0	41	0	1	1	0	1
			51	0	90	1	0	0	0	1	36	0	0	1	0	1

Nota: En el presente gráfico se muestran los 1400 datos simulados correspondientes a la derivación ambulatoria. La elaboración es propia y la fuente obtenida de las investigaciones realizadas.

Fase 4. Modelización

Minería de datos

Los algoritmos de Machine Learning (aprendizaje supervisado), específicamente los de aprendizaje supervisado, en este caso Random Forest y Naive Bayes, necesitan previamente datos etiquetados para ser entrenados. Para lograr esto se partió de la distribución de los datos

en dos grupos: datos para el entrenamiento y datos para la prueba a partir de la simulación de datos ya establecida. A continuación, se muestra la distribución en la *Tabla 24*.

Tabla 24

Distribución de los datos para el entrenamiento de los algoritmos

Datos	Absoluto	Relativo
Datos para entrenamiento	1260	90%
Datos para prueba	140	10%
Total	1400	100%

Nota: Se muestra la distribución para el entrenamiento y para la prueba del total de los 1400 datos simulados. La elaboración es propia y la fuente es proporcionada por las investigaciones realizadas.

Posteriormente, se debe hacer la elección entre el algoritmo de Random Forest y Naive Bayes de acuerdo con el porcentaje de precisión más alto para la respectiva derivación hospitalaria o ambulatoria de un paciente con COVID-19, según el dataset previamente proporcionado. Se detallan a continuación porciones de código comentadas para la compresión de las funciones usadas para la predicción.

Elección del algoritmo

Según el meta-análisis empleado en el *capítulo II*, se llegó a la conclusión de que los algoritmos más óptimos para la predicción son Random Forest y Naive Bayes por diversas características. Teniendo claro los algoritmos a usar, se procedió a comparar el resultado del porcentaje de precisión de cada uno, arrojando lo siguiente:

- **Random Forest**

  Se utiliza la función “accuracy_score” para obtener la precisión del algoritmo mencionado. Se puede apreciar en la *Figura 30.*

  Figura 30

  Precisión del algoritmo Random Forest

```
from sklearn.metrics import accuracy_score
accuracy_score(y_test, y_prds)
```

```
0.9357142857142857
```

Nota: Se muestra que el algoritmo de Random Forest arroja una precisión de 0.935 (93,5%). La elaboración es propia y la fuente es proporcionada por las investigaciones realizadas.

- **Naive Bayes**

Se utiliza la función "accuracy_score" para obtener la precisión del algoritmo mencionado. Se puede apreciar en la *Figura 31*.

Figura 31

Precisión del algoritmo Naive Bayes

```
from sklearn.metrics import accuracy_score
accuracy_score(y_test, y_pred)
```

```
0.95
```

Nota: Se muestra que el algoritmo de Naive Bayes arroja una precisión de 0.95 (95%). La elaboración es propia y la fuente es proporcionada por las investigaciones realizadas.

Fase 5. Evaluación

Análisis de los resultados

Ya con los porcentajes de precisión arrojados por cada uno algoritmos de aprendizaje supervisado mostrados en la *Figura 30* y *Figura 31*, se detallan en la siguiente *Tabla 25*.

Tabla 25

Descripción del porcentaje de precisión de cada algoritmo.

Algoritmo	Precisión obtenida	% de la precisión
Random Forest	0,935	93,5%
Naive Bayes	0,95	95%

Nota: Se presentan las precisiones obtenidas por cada algoritmo, de acuerdo con las distribuciones para el entrenamiento. La elaboración es propia y la fuente es proporcionada por las investigaciones realizadas.

Se refleja que ambos algoritmos proporcionan un alto nivel de precisión, siendo éstos factibles para el prototipo, aunque cuentan con una diferencia de milésimas entre los algoritmos analizados. Random Forest cuenta con una precisión de 0,935, es decir, el 93,5% de precisión; mientras que, Naive Bayes cuenta con una precisión de 0,95, es decir, el 95% de precisión.

Interpretación

Se llega a la conclusión que ambos algoritmos arrojan un óptimo resultado de precisión, acercándose bastante a un porcentaje total, claro está que siempre hay un margen de error. Por tal motivo, el algoritmo Naive Bayes, debido a su mayor puntuación, es considerado como el algoritmo más factible para el modelo predictivo asistencial para las personas infectadas con COVID-19 para la derivación hospitalaria o ambulatoria.

Fase 6. Implementación

En los alcances del proyecto, la implementación del algoritmo consiste en su creación y entrenamiento para que pueda ser utilizado en diversas herramientas que lo requieran, por lo cual se explica el proceso de implementación.

Figura 32

Diagrama de flujo para la implementación de los algoritmos.

Nota: Se presenta el diagrama de flujo que se implementa con los algoritmos. La elaboración es propia y la fuente es proporcionada por las investigaciones realizadas.

Arquitectura del diseño

- **Dificultad respiratoria** (0: no, 1: sí)
- **Saturación de oxígeno:** Porcentaje de oxigenación en la sangre.
- **Dolor abdominal:** (0: no, 1: sí)
- **Mialgia:** (0: no, 1: sí)
- **Tos:** (0: no, 1: sí)
- **Temperatura:** Medida en C°.
- **Pérdida de olfato:** (0: no, 1: sí)

- **Pérdida de apetito:** (0: no, 1: sí)

Entrenamiento

El entrenamiento consiste en la clasificación de datos etiquetados de manera supervisada, elaborando un modelo acorde al grupo de datos entrenados y etiquetas de clase, consiguiendo clasificar datos nuevos. En el siguiente apartado se detallan los requerimientos necesarios para crear el entorno de trabajo:

- **Base de datos:** Se utilizó alrededor de 1400 registros para el dataset almacenados en un archivo .csv.
- **Preparación del ambiente de trabajo:** Se lo realiza a través de Google Colab, el cual permite trabajar de manera online sin instalación de IDE, ya que contiene una herramienta de interpretación de Python, que facilita el desarrollo de soluciones en este lenguaje.
- **Creación del ambiente de trabajo:** Para que el ambiente de trabajo se pueda utilizar, es necesario acceder al sitio web: https://colab.research.google.com/, el cual proporciona el paquete de herramientas necesarias para la creación del algoritmo. A continuación, se presentan los fragmentos del código:

```
## Importación de las librerías ##
import pandas as pd
import numpy as np
import seaborn as sns
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import train_test_split
from matplotlib import pyplot as plt
%matplotlib inline
```

```
from sklearn.metrics import confusion_matrix
from sklearn.feature_selection import SelectKBest
## Se asigna a una variable los datos del dataset ##
df = pd.read_csv('dataset_derivacion v2.csv', sep=';', encoding='UTF-8')
## Selección de los predictores más importantes ##
X = df.drop(["derivacion"], axis=1)
y = df["derivacion"]
best = SelectKBest(k=8)
X_new = best.fit_transform(X, y)
X_new.shape
selected = best.get_support(indices=True)
used_features = X.columns[selected]
X.columns[selected]
## División de los datos: 90% para entrenamiento y 10% para prueba ##
X_train, X_test = train_test_split(df, test_size = 0.10, random_state=10)
y_train = X_train["derivacion"]
y_test = X_test["derivacion"]
## Creación del modelo predictivo Random Forest ##
rf = RandomForestClassifier(n_estimators=25)
rf = rf.fit(X_train[used_features], y_train)
y_prds = rf.predict(X_test[used_features])
## Precision del modelo ##
from sklearn.metrics import accuracy_score
accuracy_score(y_test, y_prds)
```

Pruebas

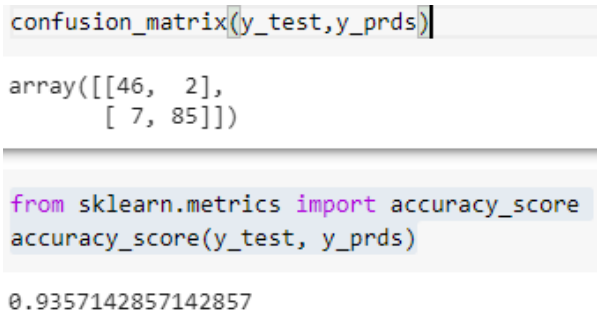
Del conjunto de datos obtenidos se precisó que el 10% de los datos se utilizarían para las pruebas del algoritmo, mientras que el porcentaje restante para el entrenamiento.

Testeo del algoritmo "Random Forest"

En primer lugar, se muestran la matriz de confusión obtenida por las pruebas realizadas por el algoritmo, posteriormente, la matriz de confusión se ve detallada en una tabla. Los datos se trataron a través del conjunto de herramientas proporcionadas por "Google Colab".

Figura 33

Matriz de confusión del algoritmo Random Forest.



Nota: Se presenta la matriz de confusión del algoritmo Random Forest. La elaboración es propia y la fuente es proporcionada por las investigaciones realizadas.

Tabla 26

Matriz de confusión del algoritmo Random Forest.

		Predicción	
		Positivo	**Negativo**
Entrenamiento	Positivo	VP = 46	FN = 2
	Negativo	FP = 7	VN = 85

Nota: Se presenta los falsos positivos y negativos de la matriz de confusión de Random Forest. La elaboración es propia y la fuente es proporcionada por las investigaciones realizadas.

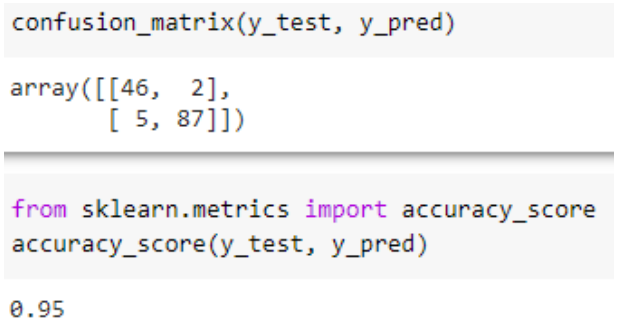
Como se aprecia en la *Tabla 26* y en la *Figura 33*, la precisión obtenida por el algoritmo Random Forest es de 0,9357142857142857, es así como en la matriz de confusión se tiene los siguientes valores: 46 datos verdaderos positivos, 85 datos verdaderos negativos, 7 datos falsos positivos y 2 datos falsos negativos. Para concluir, se puede afirmar que la precisión llega a un rango del 94%.

Testeo del algoritmo "Naive Bayes"

El testeo del algoritmo de Naive Bayes consistió en obtener la precisión y matriz de confusión del modelo al interactuar con los datos entregados, en consecuencia, en el siguiente apartado se muestran la figura y la tabla con los valores obtenidos.

Figura 34

Matriz de confusión del algoritmo Naive Bayes



Nota: Se presenta la matriz de confusión del algoritmo Naive Bayes. La elaboración es propia y la fuente es proporcionada por las investigaciones realizadas.

Tabla 27

Descripción del porcentaje de precisión de cada algoritmo

		Predicción	
		Positivo	**Negativo**
Entrenamiento	**Positivo**	VP = 46	FN = 2
	Negativo	FP = 5	VN = 87

Nota: Se presenta los falsos positivos y negativos de la matriz de confusión de Naive Bayes. La elaboración es propia y la fuente es proporcionada por las investigaciones realizadas.

Como se aprecia en la *Tabla 27* y en la *Figura 34*, la precisión obtenida por el algoritmo Random Forest es de 0,95, es así como en la matriz de confusión se tiene los siguientes valores: 46 datos verdaderos positivos, 87 datos verdaderos negativos, 5 datos falsos positivos y 2 datos falsos negativos. Para concluir, se puede afirmar que la precisión llega a un rango del 95%.

Beneficiarios directos e indirectos del proyecto

En esta sección se da a conocer el personal beneficiario, sea de manera indirecta o directa, del proyecto.

Beneficiarios Directos

Se destacan como beneficiarios directos aquellos profesionales del área de salud, específicamente del área asociada a temas de COVID-19, tales como: epidemiólogos, neumólogos, entre otros. Adicional, las personas que atraviesan por este complicado virus, ayudando que su derivación sea correcta sin posibles perjuicios futuros.

Beneficiarios Indirectos

Se reconocen como beneficiarios indirectos a la comunidad científica, debido a un gran avance en la vinculación de la tecnología con la salud en tiempos de COVID-19, dando paso a trabajos futuros con base a esta exhaustiva investigación para controlar y amenorar el impacto de este virus.

Entregables del proyecto

En este apartado se presenta todo aquel producto entregable del proyecto en cuestión, siendo estos los siguientes:

- Trabajo de titulación de investigación.
- Artículo científico.
- Base de datos del meta-análisis.
- Sitio web del prototipo funcional. (covid19gye.com)
- Esquema de la base de datos (.csv).
- Manual de usuario.
- Manual técnico.

Propuesta

Se propone un modelo predictivo asistencial mediante algoritmos supervisados de Machine Learning, cuya finalidad es apoyar a la toma de decisiones oportuna y correcta en los criterios de derivación hospitalaria o ambulatoria de pacientes infectados por COVID-19.

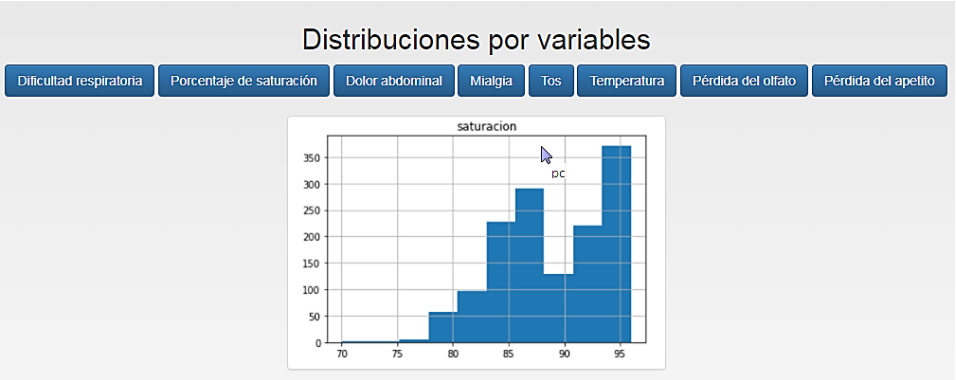
Para realizar los modelos seleccionados se tiene como datos de entradas los síntomas que presentan las personas contagiadas con este virus, las cuales son: dificultad respiratoria, saturación, dolor abdominal, mialgia, tos, temperatura, pérdida de olfato, pérdida de apetito; y como dato de salida la derivación del paciente que puede ser hospitalaria o ambulatoria como se muestra en la *Figura 35 y Figura 36*. Cada algoritmo arrojó un porcentaje de precisión, Random Forest con 93,5% y Naive Bayes con 95%, siendo este último categorizado como el mejor predictor para el prototipado, debido a su simplicidad y reducción en su tiempo de ejecución, tal como se presenta en la *Figura 37*.

La elaboración del prototipo se llevó a cabo mediante el uso de algunas herramientas, tales como: Google Colab, puesto que ofrece un servicio en la nube; es decir, no se necesita de instalaciones previas al uso de este. Está basado en las notebooks de Jupiter, las cuales permiten la creación y ejecución de código en celdas independientes y así probar las porciones de código de manera individual. El lenguaje de programación utilizado fue Python versión 3.5, ya que resulta beneficioso para la codificación de algoritmos de Machine Learning, gracias a su extensa librería que sirve para la realización de modelos predictivos**.** Para empezar con la creación del sitio web donde se muestra los resultados del proyecto, se utiliza como primer recurso, Heroku; plataforma que permite subir los modelos ya entrenados, para emplearlos como un web service, y finalmente entregar una API donde se hará peticiones tipo JSON desde el sitio web www.covid19gye.com/grupo1. La interfaz de usuario se realizó con el lenguaje de programación PHP.

El modelo predictivo asistencial será de gran ayuda para aquellos profesionales del área de salud, precisamente a los que se encuentran en primera línea de respuesta al COVID-19, dado que se les facilitará y agilizará la toma de decisiones en la derivación hospitalaria o ambulatoria de los pacientes diagnosticados con el virus.

Figura 35

Distribuciones por variables



Nota: Se muestran las distribuciones de cada una de las variables del dataset. La elaboración es propia y la fuente es proporcionada por las investigaciones realizadas.

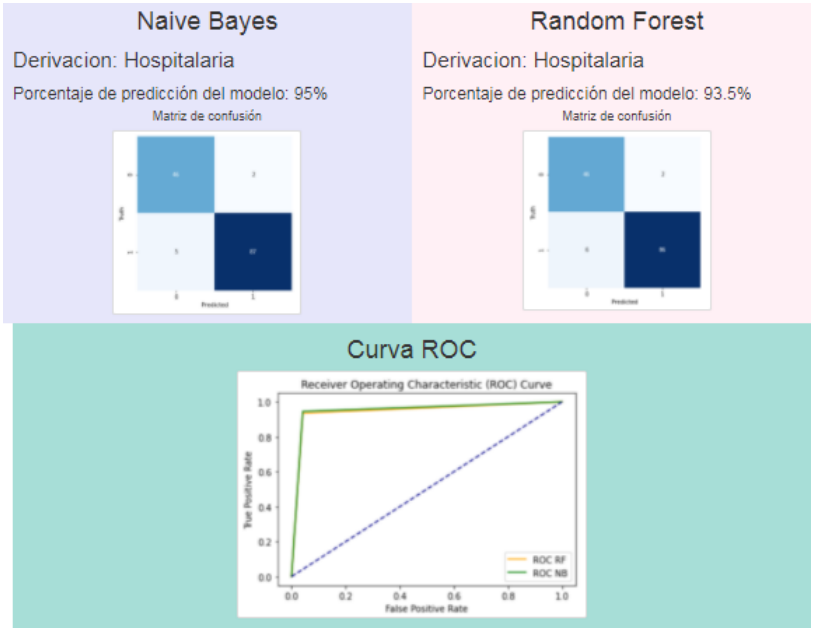
Figura 36

Formulario

Nota: Se presenta el formulario a llenar para posteriormente arrojar el resultado del algoritmo. La elaboración es propia y la fuente es proporcionada por las investigaciones realizadas.

Figura 37

Resultados de la predicción



Nota: Se muestra el resultado de los algoritmos Random Forest y Naive Bayes. La elaboración es propia y la fuente es proporcionada por las investigaciones realizadas.

Criterios de validación de la propuesta

Posteriormente, una vez efectuado el estudio a la problemática definida en el presente trabajo investigativo, se denota una gran necesidad de implementación de una herramienta basada en algoritmos predictivos que ayuden a tomar decisiones críticas como lo es la derivación de un paciente diagnosticado con COVID-19 a su respectivo tratamiento, sea este hospitalario o ambulatorio, gestionando de una manera más eficiente y eficaz la derivación hospitalaria. Además, esta solución contribuye a la agilización de los tratamientos y al desbordamiento de la capacidad de las unidades de cuidados intensivos.

Para realizar un infalible criterio de validación de la propuesta, el juicio de expertos y el contraste de hipótesis estadística y científica son métodos que corroborarán la confiabilidad obtenida en el presente proyecto; el juicio de expertos se realizó en conjunto con dos expertos en el área de la inteligencia artificial, análisis de datos, y uno en ciencias médicas, los cuales

son: Ing. Miguel Botto Tobar, M.Sc., Ing. Darwin Patiño Pérez, PhD., docentes de la Facultad de Ciencias Matemáticas y Físicas de la Universidad de Guayaquil, y el Dr. Eduardo Stay del Hospital de Especialidades Guayaquil "Dr. Abel Gilbert Pontón", véase en el *Anexo4: Anexo 4.1, Anexo 4.2* y *Anexo 4.3.*

Del juicio de expertos se indicó la siguiente sugerencia: utilizar el método "feature_importances_" de la clase "DecisionTreeRegressor" de la librería sklearn para realizar un dataframe que indique las variables más importantes para realizar el algoritmo. En cuanto al contraste de las hipótesis, una vez realizada se denotó una relación significativa entre ambas, corroborando que, la hipótesis científica se cumple de acuerdo con el análisis de Pearson. A continuación, se presenta información de los expertos en la *Tabla 28*.

Tabla 28

Información de los expertos

Nombre y Apellidos	Profesión	Entidad	Años de experiencia
Miguel Botto Tobar	Ingeniero en Sistemas Computacionales	Universidad de Guayaquil	15 años
Darwin Patiño Pérez	Ingeniero en Sistemas Computacionales	Universidad de Guayaquil	24 años
Eduardo Stay Quinde	Nefrólogo	Hospital de Especialidades Guayaquil "Dr. Abel Gilbert Pontón"	55 años

Nota: Se presenta información de cada uno de los expertos, tales como sus nombres, profesión, entidad donde trabaja y años de experiencia. La elaboración es propia y la fuente es proporcionada por las investigaciones realizadas.

Resultados

Los algoritmos de aprendizaje supervisado Random Forest y Naive Bayes fueron aplicados mediante la herramienta tecnológica Python, siendo esta muy utilizada para proyectos que requieran Machine Learning. Se codificó en el entorno de los servicios cloud de Google Colab, con la finalidad a la ayuda de la toma de decisiones para los procesos de derivación hospitalaria o ambulatoria de pacientes infectados por COVID-19.

Siguiendo el proceso requerido para la construcción del modelo predictivo, se obtuvo un exitoso porcentaje de precisión del 93,5% para el algoritmo Random Forest. El segundo algoritmo Naive Bayes arrojó un resultado de precisión del 95%, concluyendo que, en vista que ambos porcentajes son muy buenos, cabe destacar que Naive Bayes es el más eficiente para efectuar un mejor resultado para la derivación de los pacientes; predominando su diferencia en los tiempos de procesamiento, éste requiere menor tiempo.

- **Objetivo 1:** Mediante investigaciones en fuentes científicas se definieron aquellas variables significativas para la respectiva derivación hospitalaria o ambulatoria. Además, la amplia información recolectada dio paso al conocimiento de la estructura del diseño de los algoritmos supervisados de Machine Learning utilizados.
- **Objetivo 2:** Se depuró la base de datos facilitada por un hospital público de Guayaquil, dejándola más limpia y con los campos llenados correctamente. Siguiendo las 6 fases de la metodología KDD se mejoró la calidad de los datos empleados.
- **Objetivo 3:** Ya obteniendo un definido dataset, se evaluaron las variables por medio de los algoritmos supervisados de Machine Learning, definiendo una función en Python que denote aquellas variables más importantes para el resultado de las derivaciones y no presenten discrepancia con los algoritmos seleccionados. Entre ellos se descartaron 7 variables, dejando como relevantes 8 variables.

CAPÍTULO IV

CONCLUSIONES Y RECOMENDACIONES

En esta sección se detallan las conclusiones, recomendaciones y trabajos a futuros encontradas en lo largo del proceso de trabajo de titulación. Siendo estas basadas en los objetivos específicos, alcances e hipótesis del presente proyecto, quedando demostrado su cumplimiento. Adicional, se describen posibles sugerencias asociadas a los alcances que contienen un nivel de complicación más alto de realizar.

Conclusiones

Según cada uno de los objetivos específicos se han determinado las conclusiones y complementando aquellas basadas en las hipótesis demostradas.

1. La recopilación de información bibliográfica de las variables relacionadas a la derivación hospitalaria y ambulatoria por síntomas en pacientes con COVID-19, fue en base a la numerosa información alojada en plataformas confiables que contienen artículos científicos, revistas y libros, tales como: Science Direct, Taylor & Francis, Springer, IEEE, Elsevier, entre otros. Preciso que en dichas plataformas se encontró información válida acorde al tema investigativo, como lo son los factores asociados para la derivación hospitalaria, es decir, los síntomas más relevantes, entre ellos resaltando la saturación de oxígeno en la sangre, temperatura, edad y otros. Cada una fue seleccionada según los protocolos generales del área epidemiológica. Se extrajo de un total de 30 referencias bibliográficas visualizadas en el *Anexo 6.*

2. Se consiguió una base de datos perteneciente a un hospital público de la ciudad de Guayaquil, conformado por el historial clínico de pacientes infectados por COVID-19 y su respectiva derivación sea esta hospitalaria o ambulatoria. Para su respectiva depuración, para la posterior creación del dataset, se hizo uso de la metodología Knowledge Discovery in Databases, KDD, el cual se conforma de 6 fases, arrojando como resultado una base de datos más limpia conteniendo únicamente variables significativas para la derivación del paciente, las cuales son: dificultad respiratoria, saturación, dolor abdominal, mialgia, tos, temperatura, pérdida de olfato, pérdida de apetito y derivación. Además, se logró solucionar todo aquel campo vacío que impidiese más adelante el correcto desempeño del algoritmo. Concluyendo que, esta metodología es excelente para el procesamiento de la información obtenida del objetivo1.
3. Se evaluaron las variables relacionadas a la derivación hospitalaria o ambulatoria para mejorar la toma de decisiones a partir de los algoritmos de aprendizaje supervisados Random Forest y Naive Bayes. Haciendo uso de la librería sklearn del lenguaje de programación Python para el entrenamiento del algoritmo, la herramienta STAT::FIT para las distribuciones estadísticas, y basándose en la sintomatología del paciente los algoritmos arrojaron un gran porcentaje de precisión (93.5% Random Forest y 95% Naive Bayes) para la predicción de la derivación para pacientes con COVID-19. Entre ambos se concluyó que el mejor predictor es el algoritmo de Naive Bayes.
4. De acuerdo al análisis realizado en la hipótesis estadística de las preguntas “Conocimiento procesos de derivación y conocimientos sobre Random Forest” (*Capítulo III*) se pudo dar respuesta a la primera hipótesis científica “ Si se hace uso de un algoritmo de Machine Learning como es el de Random Forest entonces se podrá tomar una decisión óptima respecto a la derivación hospitalaria o ambulatoria por parte

del experto para los pacientes infectados con COVID-19 en un hospital público de la ciudad de Guayaquil" (*Capítulo II*) concluyendo que existe una relación entre ambas, mediante el análisis de correlación de Pearson, se obtiene un valor de *p* =0,040, es decir, menor a 0,05 haciendo que se cumpla la hipótesis.

5. De acuerdo al análisis realizado en la hipótesis estadística de las preguntas "Conocimiento procesos de derivación y conocimientos sobre Naive Bayes" (*Capítulo III*) se pudo dar respuesta a la primera hipótesis científica "Si se hace uso de un algoritmo de Machine Learning como es el de Naive Bayes entonces se podrá tomar una decisión óptima respecto a la derivación hospitalaria o ambulatoria por parte del experto para los pacientes infectados con COVID-19 en un hospital público de la ciudad de Guayaquil" (*Capítulo II*) concluyendo que existe una relación entre ambas, mediante el análisis de correlación de Pearson, se obtiene un valor de p =0,043, es decir, menor a 0,05 haciendo que se cumpla la hipótesis.

Recomendaciones

A continuación, se describen todas aquellas sugerencias a tomar en consideración para un correcto desempeño del proyecto.

- Para la obtención de la base de datos real, y más si es de tratarse de una entidad pública de salud, se debe realizar con antelación una solicitud al Ministerio de Salud Pública para obtener la información confidencial.
- Con el juicio de expertos debe asegurarse que las variables utilizadas para el dataset sean relevantes para el modelo, realizando una codificación con el método "features_importances_" para conocer las variables con un alto valor significativo en el modelo, y perfeccionar la interfaz gráfica del sitio web.
- Comprobar si existe sobreajuste en los algoritmos en concordancia con el total de datos que posee el dataset.
- Verificar si lo datos están correctamente escritos para la posterior conversión de cualitativos a cuantitativos.
- Realizar el apartado "derivación hospitalaria" como un sitio web responsive para ser utilizado en dispositivos móviles, debido que se encuentra desarrollado para escritorio.

Trabajos futuros

Los trabajos que puedan surgir más adelante en el ámbito de desarrollo destacan los siguientes:

- Desarrollar e implementar los algoritmos Random Forest y Naive Bayes, debido a que se trata únicamente de un prototipo, para la ejecución y verificación de su funcionalidad en ambientes reales de trabajo como lo son los hospitales públicos y privados del país.
- Desarrollar un algoritmo no supervisado de Machine Learning para efectuar la comparación de los resultados entre estos tipos de algoritmos y los ya planteados y así conocer el comportamiento y diferencia de precisión de cada uno con respecto a las variables utilizadas.

REFERENCIAS BIBLIOGRÁFICAS

Acosta, N. (2018). Nuevas tecnologías como factor de cambio ante los retos de la inteligencia artificial y la sociedad del conocimiento. *Revista ESPACIOS, 1*, 41.

Alba, M. (2018). Aplicación de técnicas de Machine Learning basado en información sísmica para profundizar la probabilidad de terremotos mediante el uso de regresión logística y redes neuronales. *Dialnet, 12*, 10.

Arcila, C. (2017). Análisis supervisado de sentimientos políticos en español: clasificación en tiempo real de tweets basada en aprendizaje automático. *Profesional de la Información, 3*, 18.

Arias, F. (2016). *El proyecto de investigación: Introducción a la métodología científica 7ma edición.* Caracas: Episteme.

Athey, S. (2018). Generalized random forests. *Annals of Statistics, 2*, 18.

Avello, R., & Seisdedo, A. (2017). El procesamiento estadístico con R en la investigación científica. *Medisur*, 4.

Ávila, J., Mayer, M., & Quesada, V. (2020). La inteligencia artificial y sus aplicaciones en medicina I: introducción antecedentes a la IA y robótica.

Bardi, T., Candela, A., De pablo , R., Martinez, R., & Pestaña, D. (2020). Respuesta rápida a COVID-19, estrategias de escalada y desescalada para ajustar la capacidad suplementaria de camas de UVI a una epidemia de gran magnitud. *Revista Española de Anestesiología y Reanimación, 68(1),*, 21-27.

Baviera, T. (2018). Técnicas para el análisis del sentimiento en Twitter: Aprendizaje Automático Supervisado y SentiStrength. *Dígitos, 12*, 18.

Berrar, D. (2018). Bayes' theorem and naive Bayes classifier. Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics; Elsevier Science Publisher: Amsterdam. *The Netherlands, 1*, 19.

Biamonte, J. (2018). Quantum machine learning. *Nature, 12*, 195.

Bisong, E. (2019). Google colaboratory. In Building Machine Learning and Deep Learning Models on Google Cloud Platform . *Apress, Berkeley, CA.*, 59-64.

Blanc, N. (2018). Modelo computacional de clasificación de aprendizaje de máquina supervisado, para el análisis de datos cardiovas-culares y pronóstico médico. *Ecuadorian Science Journal, 12*, 19.

Blanco, E., & Sanchez, G. (2020). Atención primaria y residencias de ancianos: a propósito de la COVID-19. *Semergen, 46, 26.*

Boden, M. (2017). Inteligencia artificial. *Turner, 12*, 12.

Borja, D., & Cañadas, V. (2020). Sí, la normalidad es el problema: Inequidad, exclusión y fuerza estatal en la crisis de la Covid-19 en guayaquil. *Journal of Latin American Geography, 3*, 19.

Camacho, V. (2018). Aprendizaje auto supervisado para reconocimiento de objetos. *Dialnet, 2*, 10.

Campos, M. (2018). Análisis de datos financieros con técnicas de Machine Learning. *Dialnet, 2*, 18.

Cánovas, F., Alonso, F., & Gomariz, F. (2016). MODIFICACIÓN DEL ALGORITMO RANDOM FOREST PARA SU EMPLEO EN CLASIFICACIÓN DE IMÁGENES DE TELEDETECCIÓN. 10.

Capdevilla, M. (2020). COVID-19: SOLUCIONES DESARROLLADAS MEDIANTE LAS TECNOLOGÍAS DE INFORMACIÓN GEOGRÁFICA. *Revista digital del Programa de Docencia e Investigación en Sistemas de Información Geográfica (PRODISIG). Universidad Nacional de Luján, Argentina.*

Carleo, G. (2019). Machine learning and the physical sciences. *Reviews of Modern Physics, 4*, 19.

Carr, D., Boerner, K., & Moorman, S. (2020). Bereavement in the time of coronavirus: Unprecedented challenges demand novel interventions. . *Journal of Aging & Social Policy, 32(4-5), ,* 425-431. .

Castaneda, G., & Ongkeko, W. (2020). Using machine learning of clinical data to diagnose COVID-19: a systematic review and meta-analysis. *BMC medical informatics and decision making, 1*, 14-20.

Castillo, V. A. (2018). Fundamentación teórica y metodológica del comportamiento del consumidor. *Revista Científica ECOCIENCIA, 5, 1-21.*

CCAES. (2020). *Enfermedad por coronavirus, Covid-19.* Madrid, España: Ministerio de sanidad. Obtenido de https://www.mscbs.gob.es/profesionales/saludPublica/ccayes/alertasActual/nCov/documentos/20200417_ITCoronavirus.pdf

CDC, C. (2020). *CDC*. Obtenido de personas con ciertas afecciones: https://espanol.cdc.gov/coronavirus/2019-ncov/need-extra-precautions/people-with-medical-conditions.html

Cerrillo, K. (2019). El impacto de la inteligencia artificial en el derecho administrativo¿ Nuevos conceptos para nuevas realidades técnicas? *Dialnet*, 12.

Cordero Fort, A. (2020). *Edad y mortalidad por COVID-19. Metaanálisis de 611.583 pacientes*. Obtenido de https://secardiologia.es/blog/11769-edad-y-mortalidad-por-covid-19-metaanalisis-de-611-583-pacientes

Corvalán, J. (2019). Inteligencia artificial: retos, desafíos y oportunidades-Prometea: la primera inteligencia artificial de Latinoamérica al servicio de la Justicia. *Revista de Investigações Constitucionais*, 16-20.

Cotino, L. (2020). Inteligencia artificial, big data y aplicaciones contra la COVID-19: privacidad y protección de datos. *IDP: Revista de Internet, Derecho y Politica*, 31.

Díaz, F., & Toro, A. (2020). SARS-CoV-2/COVID-19: el virus, la enfermedad y la pandemia. 23.

Díaz, J., Díaz, M., Jaraboc, A., Roig, P., & Román, P. (2017). Estudio de derivaciones de Atención Primaria a centros de Salud Mental en pacientes adultos en la Comunidad de Madrid. 7.

Díaz, L., & Espino , A. (2020). Manifestaciones gastrointestinales de pacientes infectados con el nuevo Coronavirus SARS-CoV-2. *Gastroenterol. latinoam, 31*(1), 35-38. Obtenido de https://gastrolat.org/DOI/PDF/10.46613/gastrolat202001-05.pdf

Dominguez, S. (2017). Magnitud del efecto en análisis de regresión. *Interacciones*.

Dueñas, Q. (2018). Aplicación de técnicas de machine learning a la ciberseguridad: Aprendizaje supervisado para la detección de amenazas web mediante clasificación basada en árboles de decisión. *Dialnet, 1*, 19.

Espinoza, A., & Bahamondes, G. (2020). Autoevaluación del profesorado chileno sobre el dominio de textos publicitarios y el tratamiento de la comprensión de los mismos en el aula. *Revista Espacios*.

Espinoza, E. (2018). La hipótesis en la investigación. *MENDIVE Vol. 16 No. 1*, 122-139.

Estrada, N. (2018). Detección de mastitis subclínica en vacas lecheras por modelos de regresión lineal y algoritmos de inteligencia artificial, San Carlos, Costa Rica. *Revista AgroInnovación, 2*, 12.

Ferrari, D., Tonelli, R., & Ghinelli, F. (2020). Machine learning in predicting respiratory failure in patients with COVID-19 pneumonia-challenges, strengths, and opportunities in a global health. *MedRxiv., 8*, 21.

Fontalvo, J. (2018). Aplicación de Minería de Datos para la Clasificación de programas universitarios de Ingeniería Industrial Acreditados en alta calidad en Colombia. *Información tecnológica, 12*, 19.

Geer Mountain Software Corp. (2016). *Stat::Fit®*. Obtenido de https://www.geerms.com/files/114225421.pdf

German, L., Vitale, J., & Castañeda, N. (2019). Estimación de superficie de invernáculos en el Partido de La Plata, mediante dos algoritmos de Inteligencia Artificial en la Plataforma Google Earth Engine. . *In XI Congreso de AgroInformática (CAI)-JAIIO 48 (Salta, 2019).*

Gil Osuna, B., Arias Romero, P., & Gil Ozuna, M. (2020). El coronavirus y la salud como derecho humano al hilo de las TIC: Ecuador y Brasil. *Risti*.

Gorbalenya, B. B. (2020). Severe acute respiratory syndrome-related coronavirus: The species and its viruses–a statement of the Coronavirus Study Group.

Guerrero, M. (2018). Identificación de múltiples especies en paisajes acústicos usando técnicas de aprendizaje no supervisado. *Dialnet, 1*, 10.

Guevara, G., Verdesoto, A., & Castro, N. (2020). Metodologías de investigación educativa (descriptivas, experimentales, participativas, y de investigación-acción). *RECIMUNDO*, 163-173.

Guijarro Sánchez, M. J., Royuela García , L., Guillén González , J., & Aranburu Aizpiri, I. (2019). Cuando no todo es "vejez" en el anciano. Astenia y apatía, ¿por dónde empezar? *Revista de Medicina de Familia y Atención Primaria (fml), 24*(2), 4. Obtenido de http://www.revistafml.es/wp-content/uploads/2019/07/CC-Vejez.pdf

Guzmán, R. (2020). Comparación de dos estrategias metodologicas para la estratificación socioeconomica del marco muestral de viviendas de Costa Rica.

Herrera, R., Palomino, K., Reyes, F., & Valencia, G. (2018). Análisis Estadístico Descriptivo e Inferencial de la Velocidad y Dirección del viento en la Costa Caribe Colombiana. *ESPACIOS*, 11.

Huamán , A., & Aparcana , J. (2020). La anosmia como síntoma temprano en pacientes con covid-19. *Facultad de Medicina Humana URP, 20*(3), 532-533. Obtenido de http://www.scielo.org.pe/pdf/rfmh/v20n3/2308-0531-rfmh-20-03-532.pdf

Huarcaya, J. (2020). Consideraciones sobre la salud mental en la pandemia de COVID-19. *Revista Peruana de Medicina Experimental y Salud Pública, 2*, 27.

Hutter, F. (2019). Automated machine learning: methods, systems, challenges . *Springer Nature*, 12.

Ibargüengoytia, C. (2018). Predicción de potencia eólica utilizando técnicas modernas de Inteligencia Artificial. *Ingeniería, investigación y tecnología, 1*, 10.

INEC. (15 de abril de 2020). Obtenido de https://www.ecuadorencifras.gob.ec/inec-implementa-entrevistas-telefonicas-como-alternativa-a-la-recoleccion-presencial-de-datos-suspendida-por-el-covid-19/

Iza, M. (2019, octubre). *LA GAMIFICACIÓN COMO ESTRATEGIA INNOVADORA PARA LA.*

Jiménez, E. J., & Castro, W. A. (2018). *Aplicación de la simulación Monte Carlo en la proyección del estado de resultados. Un estudio de caso*. Obtenido de http://www.revistaespacios.com/a18v39n51/a18v39n51p11.pdf

Kanavati, F., Toyokawa, G., & Momosaki, S. (2020). Weakly-supervised learning for lung carcinoma classification using deep learning. *Scientific reports, 10*(1), 1-11.

Labarthe, S. (2020). ¿Qué pasa en Ecuador? Covid-19, crisis sanitaria y conflictividad política. *Nueva Sociedad*.

Lee, T. (2018). A machine learning (kNN) approach to predicting global seafloor total organic carbon. *Global Biogeochemical Cycles, 1*, 10.

Lerma, J., Muñoz, M., Ortiz, H., & Ramos, M. (2016). Relación del trastorno limite de la personalidad y estilos de apego en una población mexicana de una institución de salud mental. *Psiquis*.

Leyva, M. (2018). Inteligencia Artificial: retos, perspectivas y papel de la Neutrosofía. *Infinite Study*, 12-19.

Liakos, K. (2019). Machine learning in agriculture: A review. *Sensors, 12*, 19.

Liang, W., Chen, R., & Guan, W. (2020). Cancer patients in SARS-CoV-2 infection: a nationwide analysis in China. *The lancet oncology, 21(3)*, 335-337.

Maguiña, C. (2020). Reflexiones sobre el COVID-19, el Colegio Médico del Perú y la Salud Pública. *Acta Médica Peruana, 7*, 37.

Manjarrés, C. (2018). Asistente virtual académico utilizando tecnologías cognitivas de procesamiento de lenguaje natural. *Revista Politécnica*, 14.

Martinez, M., Regalado de los Cobos, J., & Ruíz, M. (2020). Derivación a hospitalización a domicilio de pacientes con infección por COVID-19. *Hospital a Domicilio, 4(2)*, 59-67.

Mazzucchelli, R., & Dieguez, A. (2020). Democracia y mortalidad por Covid-19 en Europa. *Rev Esp Salud Pública, 94*, 24.

Mier, M. d., López-Perea, N., & Calles, J. M. (2018). *SITUACIÓN DE LA TOS FERINA EN ESPAÑA, 1998-2016 ANÁLISIS PRELIMINAR DEL IMPACTO DE LA VACUNACIÓN DE TOS FERINA EN EMBARAZADAS*. Obtenido de http://revista.isciii.es/index.php/bes/article/view/1060/1303

Molina, N., & Mejia, M. (2020). Impacto social de la COVID-19 en Brasil: donde la realidad supera las estadísticas. *Edumecentro*, 21.

Molnar, C. (2020). Interpretable machine learning. *Dialnet, 1*, 19.

Morales Navarro, D. (2020). Acciones del personal de salud del área estomatológica en relación a la COVID-19. *Revista Cubana Estomatología*.

Mosquera, R. (2019). Predicción de riesgos psicosociales en docentes de colegios públicos colombianos utilizando técnicas de Inteligencia Artificial. *Información tecnológica*, *2*, 29.

MSP. (2020). *Actualizacion de casos de coronavirus en Ecuador*. Obtenido de https://www.salud.gob.ec/actualizacion-de-casos-de-coronavirus-en-ecuador/

Nemecio, J. (2020). Determinaciones socioambientales del COVID-19 y vulnerabilidad económica, espacial y sanitario-institucional. *Revista de ciencias sociales*, *1*, 26.

OMS. (2020). Recomendaciones acerca COVID-19. *Boletín de la Organización Mundial de la Salud*, *8*, 1-10.

OPS. (2020). *Aspectos técnicos y regulatorios sobre el uso de oxímetros de pulso en el monitoreo de pacientes con COVID-19.*

Ospina , C., & Volcy, M. (2020). Enfoque del paciente con cefalea en tiempos de covid-19. *Acta Neurológica Colombiana, 36*(2), 27-38. Obtenido de http://www.scielo.org.co/pdf/anco/v36s1/2422-4022-anco-36-s1-27.pdf

Otzen, T., & Manterola, C. (2017). Técnicas de muestreo sobre una población a estudio. *International Journey of morphology, 35(1)*, 227-232.

Parra, V., Flórez, C., García, F., & Romero, C. (2020). Síntomas gastrointestinales en la enfermedad por covid-19 y sus implicaciones en la enfermedad inflamatoria intestinal. *Rev Colomb Gastroenterol., 35*(1), 45-55. Obtenido de file:///C:/Users/HEWLETT-PACKARD/Downloads/532-Texto%20del%20art%C3%ADculo-3447-1-10-20200506.pdf

Pastor, B. (2019). Población y muestra. *Pueblo continente 30(1)*, 245-247.

Peraza, C. (2020). Salud laboral frente a la pandemia del COVID-19 en Ecuador. *MediSur*, 18.

Pérez Abreu , M., Gómez Tejeda , J., & Dieguez Guach , R. (2020). Características clínico-epidemiológicas de la COVID-19. *Revista Habanera de ciencias médicas, 19*(2), 1-15. Obtenido de http://scielo.sld.cu/pdf/rhcm/v19n2/1729-519X-rhcm-19-02-e3254.pdf

Poblete , R., Peñafiel , F., Sabatini, N., Vite, A., Ceriani, A., Schaffeld, S., . . . Rabagliati , R. (2020). Infección respiratoria aguda por coronavirus Sars-CoV-2 en personal de salud. Implementación de un programa de detección precoz y seguimiento de casos en un hospital universitario. *Rev Med Chile*(148), 724-733. Obtenido de https://scielo.conicyt.cl/pdf/rmc/v148n6/0717-6163-rmc-148-06-0724.pdf

Probst, P. (2018). Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 12*, 12.

Rajkomar, A. (2019). Machine learning in medicine. *New England Journal of Medicine*, 17.

Rajkomar, A. (2019). Machine learning in medicine. *New England Journal of Medicine*, 17.

Ramió, C. (2019). Inteligencia artificial y Administración pública: Robots y humanos compartiendo el servicio público. *Los Libros de la Catarata*, 6.

Raschka, S. (2019). Python Machine Learning: Machine Learning and Deep Learning with Python. *Scikit-Learn, and TensorFlow. Second edition ed, 2*, 19.

Rodríguez, A. (2018). Uso del algoritmo Adaboost y la regresión logística para la predicción de fuga de clientes en una empresa de telefonía móvil. *Natura@ economía, 12*, 9.

Rodriguez, N. (2018). Los métodos de aprendizaje automático supervisado en la clasificación textual según el grado de especialización. *Dialnet, 2*, 29.

Rodríguez, N. (2018). Los métodos de aprendizaje automático supervisado en la clasificación textual según el grado de especialización. *Dialnet, 2*, 29.

Rouhiainen, L. (2018). Inteligencia artificial. *Alienta Editorial, 12*, 12.

Rubio, A., Sánchez, M., Martínez, M., & López, J. (2020). Comunicaciones Orales.- Distribución del tamaño del efecto y del tamaño muestral en los meta-análisis dentro del ámbito psicológico. *Jornadas Doctorales de la Universidad de Murcia.*

Salazar, L. (2018). Los robots y la Inteligencia Artificial: nuevos retos del periodismo. *Dialnet*.

Sandoval, S. (2018). Algoritmos de aprendizaje automático para análisis y predicción de datos. *Revista Tecnológica, 1*, 18.

Sayampanathan, A., Heng, C. S., & Pin, P. H. (2020). Infectivity of asymptomatic versus symptomatic COVID-19. *The Lancet, 397(10269)*, 93-94.

SEDISA. (2020). *Documento para la atención integral del paciente Post-Covid.* Madrid, España: Fundación AstraZeneca. Obtenido de http://www.sepsiq.org/file/InformacionSM/2020-Sedisa-DocumentoAtencionIntegralPost-Covid.pdf

Sepúlveda , V., Waissbluth, S., & González, C. (2020). Anosmia y enfermedad por Coronavirus 2019 (COVID-19): ¿Qué debemos saber? *Rev. Otorrinolaringol. Cir. Cabeza Cuello*(80), 247-258. Obtenido de https://scielo.conicyt.cl/pdf/orl/v80n2/0718-4816-orl-80-02-0247.pdf

Serra, M. (2020). COVID-19. De la patogenia a la elevada mortalidad en el adulto mayor y con comorbilidades. *Revista Habanera de Ciencias Médicas*, 12.

Shanahan, D. R., de Sousa, I. L., & Marshall, D. M. (2017). Simple decision-tree tool to facilitate author identification of reporting guidelines during submission: a before–after study. *Research integrity and peer review*, 1-6.

Smarandache, F. (2019). Fundamentos de la lógica y los conjuntos neutrosóficos y su papel en la inteligencia artificial. Infinite Study. *Dialnet*, 19-22.

Southgate, E., Blackmore, K., Pieschl, S., Grimes, S., McGuire, J., & Smithers, K. (2019). Artificial Intelligence and Emerging Technologies in Schools. 155.

Tacillo Yauli, E. (2016). *Metodología de la investigación cientifica.*

Valdés, S. (2020). Infección respiratoria aguda por COVID-19: una amenaza evidente. *Rev haban cienc méd*, 23.

Valenzuela, M. (2019). Machine Learning y el Control Automático en Chile. *Dialnet, 2*, 19.

Vázquez, A. (2018). Introducción a Machine Learning. *Dialnet, 12*, 19.

Visión CEVECE. (2020). *Secuelas por covid-19.* Estado de México: Secretaría de Salud. Obtenido de https://salud.edomex.gob.mx/cevece/documentos/difusion/tripticos/2020/Semana42.pdf

Wong, G., & Thompson, C. (2020). Management of patients with liver derangement during the COVID-19 pandemic: an Asia-Pacific position statement. *The Lancet Gastroenterology & Hepatology, 6*, 12.

Zambrano, A. A. (2016). *Computación Estadística con SAS.*

ANEXOS

Anexo 1. Planificación de actividades del proyecto

Actividades	Fecha de inicio	Duración en días	Fecha fin
PROYECTO DE TITULACIÓN	26/11/2020	103	09/03/2021
Asignación de tutor	26/11/2020	0	26/11/2020
Inicialización del proyecto	28/11/2020	0	28/11/2020
CAPITULO 1 Planteamiento del proble	29/11/2020	9	08/12/2020
Descripción de la situación problemáti	29/11/2020	1	30/11/2020
Causas y consecuencias del problema	30/11/2020	1	01/12/2020
Formulación del problema	01/12/2020	1	02/12/2020
Objetivos del proyecto	02/12/2020	1	03/12/2020
Alcance del proyecto	03/12/2020	1	04/12/2020
Justificación e importancia	04/12/2020	3	07/12/2020
Limitaciones de estudio	07/12/2020	1	08/12/2020
Reunion para revisión del Capítulo 1	08/12/2020	0	08/12/2020
Corrección del Capítulo 1	09/12/2020	0	09/12/2020
Entrega del cronograma, anexo 0, 1 y 2	10/12/2020	0	10/12/2020
CAPITULO II MARCO TEÓRICO	09/12/2020	5	14/12/2020
Antecedentes del estudio	09/12/2020	1	10/12/2020
Fundamentación teórica	10/12/2020	1	11/12/2020
Revisiones sistemáticas	11/12/2020	0	11/12/2020
Hipótesis	11/12/2020	1	12/12/2020
Variables de la investigación	12/12/2020	1	13/12/2020
Definiciones conceptuales	13/12/2020	1	14/12/2020
Reunion para revisión del Capítulo 2	14/12/2020	0	14/12/2020
Corrección del Capítulo 2	15/12/2020	13	28/12/2020
CAPITULO III METODOLOGÍA DE LA IN	28/12/2020	54	20/02/2021
Tipo de investigación	28/12/2020	6	03/01/2021
Diseño metodológico de la investigaci	03/01/2021	2	05/01/2021
Beneficiarios directos e indirectos del	05/01/2021	3	08/01/2021
Entregables del proyecto	08/01/2021	10	18/01/2021
Propuesta	18/01/2021	10	28/01/2021
Criterios de la validación de la propue	28/01/2021	13	10/02/2021
Resultados	10/02/2021	10	20/02/2021
CAPÍTULO IV CONCLUSIONES Y RECOM	20/02/2021	13	05/03/2021
Conclusiones	20/02/2021	9	01/03/2021
Recomendaciones	01/03/2021	4	05/03/2021
SUSTENTACIÓN PROYECTO DE TITULA	17/03/2021	16	02/04/2021

Elaboración: Fuentes Melina y Medina Wilmer.
Fuente: Propia.

Anexo 2. Fundamentación legal

Las Normas Legales en un Proyecto de Titulación

El presente proyecto de titulación se basa en brindar ayuda a los expertos del área de salud en la toma de decisiones para la derivación hospitalaria o ambulatoria de pacientes infectados con COVID-19 mediante técnicas de algoritmos supervisados de Machine Learning. Este proyecto fundamenta en la constitución, leyes y normas como se detalla a continuación…

ARTÍCULO DE LA LOES	CONTEXTO
¿Qué regula la LOES? ART. 1 ÁMBITO	Esta Ley regula el sistema de educación superior en el país, a los organismos e instituciones que lo integran; determina derechos, deberes y obligaciones de las personas naturales y jurídicas, y establece las respectivas sanciones por el incumplimiento de las disposiciones contenidas en la Constitución y la presente Ley ARTICULO 1
¿Cuál es el Objeto de esta Ley? ART. 2 OBJETO	Esta Ley tiene como objeto definir sus principios, garantizar el derecho a la educación superior de calidad que propenda a la excelencia, al acceso universal, permanencia, movilidad y egreso sin discriminación alguna.
Entre las funciones ART. 4 DERECHO A LA EDUCACION SUPERIOR	a) Garantizar el derecho a la educación superior mediante la docencia, la investigación y su vinculación con la sociedad, y asegurar crecientes niveles de calidad, excelencia académica y pertinencia; n) Garantizar la producción de pensamiento y conocimiento articulado con el pensamiento universal; y, ñ) Brindar niveles óptimos de calidad en la formación.
Principio de Igualdad y Principio de Calidad	El **principio de igualdad** de oportunidades consiste en garantizar a todos los actores del Sistema de Educación Superior las mismas posibilidades en el acceso, permanencia, movilidad y egreso del sistema, sin discriminación de género, credo, orientación sexual, etnia, cultura, preferencia política, condición socioeconómica o discapacidad. El **principio de calidad** consiste en la búsqueda constante y sistemática de la excelencia, la pertinencia, producción óptima, transmisión del conocimiento y desarrollo del pensamiento mediante la autocrítica, la crítica externa y el mejoramiento permanente
ART. 87	Como requisito previo a la obtención del título, los y las estudiantes deberán acreditar servicios a la comunidad mediante prácticas o pasantías preprofesionales. debidamente monitoreadas. en los campos de su especialidad, de conformidad con los lineamientos generales definidos por el Consejo de Educación Superior.
ARTÍCULO 19.- DEL REGLAMENTO. - NÓMINA DE GRADUADOS Y NOTIFICACIÓN A LA SENESCYT	Las instituciones de educación superior notificarán obligatoriamente a la SENESCYT la nómina de los graduados y las especificaciones de los títulos que expida, en un plazo no mayor de treinta días contados a partir de la fecha de graduación. (…) este será el único medio oficial a través del cual se verificará el reconocimiento y validez del título en el Ecuador.
ARTÍCULO 144 PRINCIPIOS	Art. 144.- Tesis Digitalizadas. - Todas las instituciones de educación superior estarán obligadas a entregar las tesis que se elaboren para la obtención de títulos académicos de grado y posgrado en formato digital para ser integradas al Sistema Nacional de Información de la Educación Superior del Ecuador para su difusión pública respetando los derechos de autor.

Elaboración: Fuentes Melina y Medina Wilmer.
Fuente: Ley Orgánica de Educación Superior.

ARTÍCULO DE LA CONSTITUCIÓN	CONTEXTO
ARTÍCULO 22	Establece: las personas tienen derecho a desarrollar su capacidad creativa, al ejercicio digno y sostenido de las actividades culturales y artísticas, y a beneficiarse de la protección de los derechos morales y patrimoniales que les correspondan por las producciones científicas, literarias o artísticas de su autoría.
ARTÍCULO 26	La educación es un derecho de las personas a lo largo de su vida y un deber ineludible e inexcusable del Estado. Constituye un área prioritaria de la política pública y de la inversión estatal, garantía de la igualdad e inclusión social y condición indispensable para el buen vivir.
ARTÍCULO 28	La educación responderá al interés público y no estará al servicio de intereses individuales y corporativos. Se garantizará el acceso universal, permanencia, movilidad y egreso sin discriminación alguna.
ARTÍCULO 322	Se reconoce la propiedad intelectual de acuerdo con las condiciones que señale la ley. Se prohíbe toda forma de apropiación de conocimientos colectivos, en el ámbito de las ciencias, tecnologías y saberes ancestrales. Se prohíbe también la apropiación sobre los recursos genéticos que contienen la diversidad biológica y la agrobiodiversidad.
ARTÍCULO 350	El sistema de educación superior tiene como finalidad la formación académica y profesional con visión científica y humanista; la investigación científica y tecnológica; la innovación, promoción, desarrollo y difusión de los saberes y las culturas; la construcción de soluciones para los problemas del país, en relación con los objetivos del régimen de desarrollo.
ARTÍCULO 351	El sistema de educación superior estará articulado al sistema nacional de educación y al Plan Nacional de Desarrollo; la ley establecerá los mecanismos de coordinación del sistema de educación superior con la Función Ejecutiva. Este sistema se regirá por los principios de autonomía responsable, cogobierno, igualdad de oportunidades, calidad, pertinencia, integralidad, autodeterminación para la producción del pensamiento y conocimiento, en el marco del diálogo de saberes, pensamiento universal y producción científica tecnológica global.
ARTÍCULO 355 primer y segundo inciso	El Estado reconocerá a las universidades y escuelas politécnicas autonomía académica, administrativa, financiera y orgánica, acorde con los objetivos del régimen de desarrollo y los principios establecidos en la Constitución.
ARTÍCULO 385	El sistema nacional de ciencia, tecnología, Innovación y saberes ancestrales, en el marco del respeto al ambiente, la naturaleza, la vida, las culturas y la soberanía, tendrá como finalidad: a) Generar, adaptar y difundir conocimientos científicos y tecnológicos. b) Recuperar, fortalecer y potenciar los saberes ancestrales. c) Desarrollar tecnologías e innovaciones que impulsen la producción nacional, eleven la eficiencia y productividad, mejoren la calidad de vida y contribuyan a la realización del buen vivir.
ARTÍCULO 386	El sistema comprenderá programas, políticas, recursos, acciones, e incorporará a instituciones del Estado, universidades y escuelas politécnicas, institutos de investigación públicos y privados, empresas públicas y privadas, organismos no gubernamentales y personas naturales o jurídicas, en tanto realizan actividades de investigación, desarrollo tecnológico, innovación y aquellas ligadas a los saberes ancestrales. El Estado, a través del organismo competente, coordinará el sistema, establecerá los objetivos y políticas, de conformidad con el Plan Nacional de Desarrollo, con la participación de los actores que lo conforman.
ARTÍCULO 387	Será responsabilidad del Estado: a) Facilitar e impulsar la incorporación a la sociedad del conocimiento para alcanzar los objetivos del régimen de desarrollo. b) Promover la generación y producción de conocimiento, fomentar la investigación científica y tecnológica, y potenciar los saberes ancestrales, para así contribuir a la realización del buen vivir, al sumak kawsay.

	c) Asegurar la difusión y el acceso a los conocimientos científicos y tecnológicos, el usufructo de sus descubrimientos y hallazgos en el marco de lo establecido en la Constitución y la Ley. d) Garantizar la libertad de creación e investigación en el marco del respeto a la ética, la naturaleza, el ambiente, y el rescate de los conocimientos ancestrales. e) Reconocer la condición de investigador de acuerdo con la Ley.
ARTÍCULO 388	El Estado destinara los recursos necesarios para la investigación científica, el desarrollo tecnológico, la innovación, la formación científica, la recuperación y desarrollo de saberes ancestrales y la difusión del conocimiento. Un porcentaje de estos recursos se destinará a financiar proyectos mediante fondos concursables. Las organizaciones que reciban fondos públicos estarán sujetas a la rendición de cuentas y al control estatal respectivo.

Elaboración: Fuentes Melina y Medina Wilmer.
Fuente: Constitución de la Republica del Ecuador.

FACTIBILIDAD LEGAL. - El presente proyecto de titulación es legalmente factible, puesto que no va en contra de ningún lineamiento, normativa o política interna. Se usaron herramientas de código abierto, y aquellas que estén proyectadas en el área del COVID-19 sin infringir ningún parámetro legal.

CODIGO ORGÁNICO DE LA ECONOMÍA SOCIAL DE LOS CONOCIMIENTOS, CREATIVIDAD E INVENCIÓN	
ARTÍCULO DE LA CONSTITUCIÓN	**CONTEXTO**
ARTÍCULO 104 Obras susceptibles de protección	La protección reconocida por el presente Título recae sobre todas las obras literarias, artísticas y científicas, que sean originales y que puedan reproducirse o divulgarse por cualquier forma o medio conocido o por conocerse. Las obras susceptibles de protección comprenden, entre otras, las siguientes: 2.- Colecciones de obras, tales como enciclopedias, antologías o compilaciones y bases de datos de toda clase, que por la selección o disposición de las materias constituyan creaciones intelectuales originales, sin perjuicio de los derechos que subsistan sobre las obras, materiales, información o datos; 12.- Software.
ARTÍCULO 131.- Protección de software	El software se protege como obra literaria. Dicha protección se otorga independientemente de que hayan sido incorporados en un ordenador y cualquiera sea la forma en que estén expresados, ya sea como código fuente; es decir, en forma legible por el ser humano; o como código objeto; es decir, en forma legible por máquina, ya sea sistemas operativos o sistemas aplicativos, incluyendo diagramas de flujo, planos, manuales de uso, y en general, aquellos elementos que conformen la estructura secuencian y organización del programa. Se excluye de esta protección las formas estándar de desarrollo de software. En este sentido, los documentos y textos producidos en las Instituciones de Educación Superior desarrollados con el objeto de obtener sus grados académicos y/o trabajos de facultad, son autores intelectuales con el patrocinio de cada institución, por lo tanto, son acreedores a los derechos de protección intelectual dispuestos en la normativa vigente.

Anexo 3. Formatos de técnicas de recolección de datos

Encuesta

UNIVERSIDAD DE GUAYAQUIL

FACULTAD DE CIENCIAS MATEMÁTICAS Y FÍSICAS

CARRERA DE INGENIERÍA EN SISTEMAS COMPUTACIONALES

Proyecto: Diseño de un modelo predictivo-asistencial de pacientes infectados por COVID-19, mediante un modelo supervisado de Machine Learning basado en criterios de derivación hospitalaria o ambulatoria

Objetivo: Diseñar un modelo predictivo de los factores que se relacionan con el ingreso hospitalario frente al ambulatorio mediante algoritmo supervisado de Machine Learning para obtener un criterio de derivación entre niveles asistenciales durante la pandemia del COVID-19.

Encuesta Nº: ____

Preguntas:

1. ¿Conoce usted sobre los procesos de derivación hospitalaria o ambulatoria que se producen en tiempos de COVID-19?
   - Total conocimiento
   - Parcial conocimiento
   - Indiferente
   - Parcial desconocimiento
   - Total desconocimiento

2. ¿Tiene usted conocimiento sobre la existencia del uso de inteligencia artificial como modelo computacional para resolver de manera automática procesos de derivación hospitalaria o ambulatoria?
   - Total conocimiento
   - Parcial conocimiento
   - Indiferente
   - Parcial desconocimiento
   - Total desconocimiento

3. ¿Qué conocimientos tiene usted, que dentro de la inteligencia artificial existen modelos tales como el de Machine Learning que ayuda al manejo de grandes volúmenes de datos de forma automática?
   - Total conocimiento
   - Parcial conocimiento
   - Indiferente
   - Parcial desconocimiento
   - Total desconocimiento

4. Dentro de Machine Learning ¿qué tipo de algoritmos usted conoce? (Puede marcar más de una respuesta)
   - Naive Bayes (Redes Bayesianas)
   - Regresión Logística
   - Random Forest (Bosques Aleatorios)
   - Redes Neuronales
   - No conozco ninguno
5. ¿Cree usted que este tipo de aplicaciones permita tomar mejores decisiones para saber si la persona entra en un proceso hospitalario o ambulatorio para pacientes con COVID-19?
   - Total acuerdo
   - Parcial acuerdo
   - Indiferente
   - Parcial desacuerdo
   - Total desacuerdo
6. ¿Qué conocimiento tiene usted, dentro del modelo de Machine Learning, sobre el uso del algoritmo de Naive Bayes?
   - Total conocimiento
   - Parcial conocimiento
   - Indiferente
   - Parcial desconocimiento
   - Total desconocimiento
7. ¿Qué conocimiento tiene usted, dentro del modelo de Machine Learning, sobre el uso del algoritmo de Random Forest?
   - Total conocimiento
   - Parcial conocimiento
   - Indiferente
   - Parcial desconocimiento
   - Total desconocimiento

Elaboración: Fuentes Melina, Medina Wilmer.
Fuente: Propia.

Entrevista

UNIVERSIDAD DE GUAYAQUIL

FACULTAD DE CIENCIAS MATEMÁTICAS Y FÍSICAS

CARRERA DE INGENIERÍA EN SISTEMAS COMPUTACIONALES

Proyecto: Diseño de un modelo predictivo-asistencial de pacientes infectados por COVID-19, mediante un modelo supervisado de Machine Learning basado en criterios de derivación hospitalaria o ambulatoria.

La presente entrevista tiene como objetivo realizar un levantamiento de información sobre los procesos de derivación hospitalaria, factores asociados, y conocimiento y uso de Machine Learning vinculados a dichos procesos.

Entrevista Nº: ____

Nombre del entrevistado:_____

Área de trabajo: _____

Cargo: _____-

Preguntas:

1. ¿Usted en qué sintomatología se basa para la toma de decisiones para la derivación (tratamiento) hospitalaria o ambulatoria de personas diagnosticadas con COVID-19?

_____

_____

_____

2. ¿Las personas infectadas con COVID-19, además con comorbilidades deben ser hospitalizadas?

_____

_____

_____

3. ¿Las personas adulto mayor infectadas con COVID-19, debido a su avanzada edad, es necesaria la hospitalización?

_____

_____

_____

4. ¿Hace manejo de alguna herramienta informática que le ayude a la toma de decisiones sobre la derivación hospitalaria o ambulatoria de pacientes infectados con COVID-19?

_____

_____

_____

5. ¿Considera usted que una herramienta informática le ayudaría a la toma de decisiones y apoyaría su criterio médico para la correcta derivación hospitalaria o ambulatoria?

_____

_____

_____

Elaboración: Fuentes Melina, Medina Wilmer.
Fuente: Propia.

Anexo 4. Validación de expertos

Juicios de expertos

Para la validación del proyecto se utilizó el instrumento de juicio de expertos con la finalidad de realizar las pruebas de funcionalidad y porcentaje de validación del software desarrollado, adicional los expertos que realicen la validación correspondiente pueda ofrecer valorización para este proyecto y que las técnicas implementadas sean las adecuadas. (Véase *Anexo 4: Anexo 4.1, Anexo 4.2 y Anexo 4.3*).

ANEXO 4.1. VALIDACIÓN DE EXPERTOS

DATOS GENERALES

APELLIDOS Y NOMBRES DEL EXPERTO	TITULO PROFESIONAL DEL EXPERTO	AUTOR(ES)	
Ing. Miguel Botto Tobar, M.Sc.	Ingeniero en Sistemas Computacionales	Fuentes Marmolejo Melina Daniela	Medina Parra Wilmer David
TÍTULO DEL PROYECTO	DISEÑO DE UN MODELO PREDICTIVO-ASISTENCIAL DE PACIENTES INFECTADOS POR COVID-19, MEDIANTE UN MODELO SUPERVISADO DE MACHINE LEARNING BASADO EN CRITERIOS DE DERIVACIÓN HOSPITALARIA O AMBULATORIA		

INDICADOR	CRITERIO	DEFICIENTE 0-20				REGULAR 21-40				BUENA 41- 60				MUY BUENA 61- 80				EXCELENTE 81 - 100			
		5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
CLARIDAD	Se utiliza el lenguaje de programación apropiado que facilita la comprensión.																				X
OBJETIVIDAD	Está expresado en conductas observables y medibles.																				X
ACTUALIDAD	Esta acorde a los aportes recientes en la disciplina de estudio.																				X
SUFICIENCIA	Son suficientes la cantidad y calidad de ítems presentados en el instrumento.																				X
INTENCIONALIDAD	Es adecuado para valorar la variable seleccionada.																				X
CONSISTENCIA	Está basado en aspectos teóricos y científicos.																				X
METODOLOGÍA	El instrumento se relaciona con el método planteado en el proyecto.																				X
APLICABILIDAD	El instrumento es de fácil aplicación.																				X

CONSTANCIA DE JUICIO DE EXPERTO

Estimado Ingeniero
Lorenzo Cevallos Torres
DOCENTE TUTOR DEL TRABAJO DE TITULACIÓN
Ciudad. -

El presente instrumento certifica que se realizó la revisión del proyecto de titulación "DISEÑO DE UN MODELO PREDICTIVO-ASISTENCIAL DE PACIENTES INFECTADOS POR COVID-19, MEDIANTE UN MODELO SUPERVISADO DE MACHINE LEARNING BASADO EN CRITERIOS DE DERIVACIÓN HOSPITALARIA O AMBULATORIA" cuyos criterios e indicadores empleados permitieron articular el trabajo según se muestra en el *Anexo 4.1*, por tanto, Fuentes Marmolejo Melina Daniela y Medina Parra Wilmer David estudiante(s) no titulados de la Carrera de Ingeniería en Sistemas computacionales de la Universidad de Guayaquil, pueden continuar con el proceso de titulación en vista que no existen observaciones.

Por lo actuado en el *Anexo 4.1*, se procede a validar el trabajo de titulación.

Sin otro particular.

Firmado electrónicamente por:
MIGUEL ANGEL
BOTTO TOBAR

Ing. Miguel Botto Tobar, M.Sc.
C.I. N° 1204824328

Elaboración: Fuentes Melina y Medina Wilmer.
Fuente: Propia.

ANEXO 4.2. VALIDACIÓN DE EXPERTOS

DATOS GENERALES

APELLIDOS Y NOMBRES DEL EXPERTO	TITULO PROFESIONAL DEL EXPERTO	AUTOR(ES)	
Ing. Darwin Patiño Pérez, PhD.	Ingeniero en Sistemas Computacionales	Fuentes Marmolejo Melina Daniela	Medina Parra Wilmer David
TÍTULO DEL PROYECTO	DISEÑO DE UN MODELO PREDICTIVO-ASISTENCIAL DE PACIENTES INFECTADOS POR COVID-19, MEDIANTE UN MODELO SUPERVISADO DE MACHINE LEARNING BASADO EN CRITERIOS DE DERIVACIÓN HOSPITALARIA O AMBULATORIA		

INDICADOR	CRITERIO	DEFICIENTE 0-20				REGULAR 21-40				BUENA 41- 60				MUY BUENA 61- 80				EXCELENTE 81 - 100			
		5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
CLARIDAD	Se utiliza el lenguaje de programación apropiado que facilita la comprensión.																				X
OBJETIVIDAD	Está expresado en conductas observables y medibles.																				X
ACTUALIDAD	Esta acorde a los aportes recientes en la disciplina de estudio.																				X
SUFICIENCIA	Son suficientes la cantidad y calidad de ítems presentados en el instrumento.																				X
INTENCIONALIDAD	Es adecuado para valorar la variable seleccionada.																				X
CONSISTENCIA	Está basado en aspectos teóricos y científicos.																				X
METODOLOGÍA	El instrumento se relaciona con el método planteado en el proyecto.																				X
APLICABILIDAD	El instrumento es de fácil aplicación.																				X

CONSTANCIA DE JUICIO DE EXPERTO

Estimado Ingeniero

Lorenzo Cevallos Torres

DOCENTE TUTOR DEL TRABAJO DE TITULACIÓN

Ciudad. -

El presente instrumento certifica que se realizó la revisión del proyecto de titulación "DISEÑO DE UN MODELO PREDICTIVO-ASISTENCIAL DE PACIENTES INFECTADOS POR COVID-19, MEDIANTE UN MODELO SUPERVISADO DE MACHINE LEARNING BASADO EN CRITERIOS DE DERIVACIÓN HOSPITALARIA O AMBULATORIA" cuyos criterios e indicadores empleados permitieron articular el trabajo según se muestra en el *Anexo 4.2*, por tanto, Fuentes Marmolejo Melina Daniela y Medina Parra Wilmer David estudiante(s) no titulados de la Carrera de Ingeniería en Sistemas computacionales de la Universidad de Guayaquil, pueden continuar con el proceso de titulación en vista que no existen observaciones.

Por lo actuado en el *Anexo 4.2*, se procede a validar el trabajo de titulación.

Sin otro particular.

Firmado electrónicamente por:
DARWIN
GUILLERMO
PATINO PEREZ

Ing. Darwin Patiño Pérez, PhD.
C.I. N° - 0911147999

Elaboración: Fuentes Melina y Medina Wilmer.
Fuente: Propia.

ANEXO 4.3. VALIDACIÓN DE EXPERTOS

DATOS GENERALES

APELLIDOS Y NOMBRES DEL EXPERTO	TITULO PROFESIONAL DEL EXPERTO	AUTOR(ES)	
Dr. Eduardo Stay Quinde	Nefrólogo	Fuentes Marmolejo Melina Daniela	Medina Parra Wilmer David
TÍTULO DEL PROYECTO	DISEÑO DE UN MODELO PREDICTIVO-ASISTENCIAL DE PACIENTES INFECTADOS POR COVID-19, MEDIANTE UN MODELO SUPERVISADO DE MACHINE LEARNING BASADO EN CRITERIOS DE DERIVACIÓN HOSPITALARIA O AMBULATORIA		

INDICADOR	CRITERIO	DEFICIENTE 0-20				REGULAR 21-40				BUENA 41- 60				MUY BUENA 61- 80				EXCELENTE 81 - 100			
		5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
CLARIDAD	El modelo predictivo es fácilmente comprensible.																				X
OBJETIVIDAD	Está expresado en conductas observables y medibles.																				X
ACTUALIDAD	Esta acorde a los aportes recientes en la disciplina de estudio.																				X
SUFICIENCIA	Son suficientes la cantidad y calidad de ítems presentados en el instrumento.																				X
INTENCIONALIDAD	Es adecuado para valorar la variable seleccionada.																				X
CONSISTENCIA	Está basado en aspectos teóricos y científicos.																				X
METODOLOGÍA	El instrumento se relaciona con el método planteado en el proyecto.																				X
APLICABILIDAD	El instrumento es de fácil aplicación.																				X

CONSTANCIA DE JUICIO DE EXPERTO

Estimado Ingeniero
Lorenzo Cevallos Torres
DOCENTE TUTOR DEL TRABAJO DE TITULACIÓN
Ciudad. -

El presente instrumento certifica que se realizó la revisión del proyecto de titulación "DISEÑO DE UN MODELO PREDICTIVO-ASISTENCIAL DE PACIENTES INFECTADOS POR COVID-19, MEDIANTE UN MODELO SUPERVISADO DE MACHINE LEARNING BASADO EN CRITERIOS DE DERIVACIÓN HOSPITALARIA O AMBULATORIA" cuyos criterios e indicadores empleados permitieron articular el trabajo según se muestra en el *Anexo 4.3*, por tanto, Fuentes Marmolejo Melina Daniela y Medina Parra Wilmer David estudiante(s) no titulados de la Carrera de Ingeniería en Sistemas computacionales de la Universidad de Guayaquil, pueden continuar con el proceso de titulación en vista que no existen observaciones.

Por lo actuado en el *Anexo 4.3*, se procede a validar el trabajo de titulación.

Sin otro particular.

Dr. Eduardo Stay Quinde
C.I. N° -0905692612

Elaboración: Fuentes Melina y Medina Wilmer.
Fuente: Propia.

Anexo 5. Criterios éticos a utilizarse en el desarrollo del proyecto

Criterios	Características del criterio	Procedimientos
Credibilidad Valor de la verdad/autenticidad	Aproximación de los resultados de una aproximación frente al fenómeno observado.	– Los resultados son reconocidos "verdaderos" por los participantes. – Observación continua y prolongada del fenómeno.
Conformabilidad o Reflexibilidad Neutralidad/Objetividad	Los resultados de la investigación deben garantizar la veracidad de las descripciones realizadas por los participantes.	– Transcripciones textuales de las entrevistas. – Contrastación de los resultados de la literatura existente. – Revisión de hallazgos por otros investigadores. – Identificación y descripción de limitaciones y alcances del investigador.
Relevancia	Permite evaluar el logro de los objetivos planteados y saber si se obtuvo un mejor conocimiento del fenómeno del estudio.	– Configuración de nuevos planteamientos teóricos o conceptuales. – Comprensión amplia del fenómeno. – Correspondencia entre la justificación y los resultados obtenidos.
Adecuación teórica-epistemológica	Correspondencia adecuada del problema por investigar y la teoría existente.	– Contrastación de la pregunta con los métodos. – Ajustes de diseño.

Elaboración: Fuentes Melina y Medina Wilmer.
Fuente: Propia.

Anexo 6. Tabla del meta-análisis

N°	Título del artículo científico	Autor(es)	Resumen	Año	Revista
1	Aplicación de técnicas de machine learning a la ciberseguridad	José María Dueñas Quesada	El objetivo del TFM es desarrollar un modelo predictivo de machine learning basado en árboles de decisión cuya tarea sea clasificar un conjunto de peticiones HTTP en peticiones normales y anómalas. El TFM incluye un estudio del estado del arte sobre las aplicaciones en ciberseguridad del machine learning, la implementación con Python y Scikit-learn de un modelo clasificador basado en aprendizaje supervisado con árboles de decisión y el análisis de los resultados de aplicar dicho modelo sobre el dataset CSIC-2010. El modelo propuesto en este TFM consigue hasta un 100% de exactitud (accuracy) en la clasificación de las peticiones HTTP.	2020	Google académico

N°	Título del artículo científico	Autor(es)	Resumen	Año	Revista
2	COVID-19 Public Sentiment Insights and Machine Learning for Tweets Classification	Jim Samuel, G. G. Md. Nawaz Ali, Md. Mokhlesur Rahman, Ek Esawi and Yana Samuel	Along with the Coronavirus pandemic, another crisis has manifested itself in the form of mass fear and panic phenomena, fueled by incomplete and often inaccurate information. There is therefore a tremendous need to address and better understand COVID-19's informational crisis and gauge public sentiment, so that appropriate messaging and policy decisions can be implemented. In this research article, we identify public sentiment associated with the pandemic using Coronavirus Specific Tweets and R statistical software, along with its sentiment analysis packages. We demonstrate insights into the progress of fear-sentiment over time as COVID-19 approached peak levels in the United States, using descriptive textual analytics supported by necessary textual data visualizations.	2020	MDPI

N°	Título del artículo científico	Autor(es)	Resumen	Año	Revista
3	Predictive Data Mining Models for Novel Coronavirus (COVID-19) Infected Patients' Recovery	L. J. Muhammad, Md. Milon Islam, Sani Sharif Usman, Safal Islam Ayon	Novel coronavirus (COVID-19 or 2019-nCoV) pandemic has neither clinically proven vaccine nor drugs; however, its patients are recovering with the aid of antibiotic medications, anti-viral drugs, and chloroquine as well as vitamin C supplementation. It is now evident that the world needs a speedy and quicker solution to contain and tackle the further spread of COVID-19 across the world with the aid of non-clinical approaches such as data mining approaches, augmented intelligence and other artificial intelligence techniques so as to mitigate the huge burden on the healthcare system while providing the best possible means for patients' diagnosis and prognosis of the 2019-nCoV pandemic effectively.	2020	Google académico
4	COVID-19 Patient Health Prediction Using Boosted Random Forest Algorithm	Celestine Iwendi, Ali Kashif Bashir, Atharva Peshkar, R. Sujatha, Jyotir Moy Chatterjee, Swetha Pasupuleti, Rishita Mishra, Sofia Pillai and Ohyun Jo	Integration of artificial intelligence (AI) techniques in wireless infrastructure, real-time collection, and processing of end-user devices is now in high demand. It is now superlative to use AI to detect and predict pandemics of a colossal nature. The Coronavirus disease 2019 (COVID-19) pandemic, which originated in Wuhan China, has had disastrous effects on the global community and has overburdened advanced healthcare systems throughout the world. Globally; over 4,063,525 confirmed cases and 282,244 deaths have been recorded as of 11th May 2020, according to the European Centre for Disease Prevention and Control agency.	2020	Frontiers

N°	Título del artículo científico	Autor(es)	Resumen	Año	Revista
5	Machine learning based approaches for detecting COVID-19 using clinical text data	Akib Mohi Ud Din Khanday, Syed Tanzeel Rabani, Qamar Rayees Khan, Nusrat Rouf, Masarat Mohi Ud Din	Technology advancements have a rapid effect on every field of life, be it medical field or any other field. Artificial intelligence has shown the promising results in health care through its decision making by analysing the data. COVID-19 has affected more than 100 countries in a matter of no time. People all over the world are vulnerable to its consequences in future. It is imperative to develop a control system that will detect the coronavirus. One of the solutions to control the current havoc can be the diagnosis of disease with the help of various AI tools.	2020	Springer
6	Detection of COVID-19 Infection from Routine Blood Exams with Machine Learning: A Feasibility Study	Davide Brinati, Andrea Campagner, Davide Ferrari, Massimo Locatelli, Giuseppe Banfi, Federico Cabitza	The COVID-19 pandemia due to the SARS-CoV-2 coronavirus, in its first 4 months since its outbreak, has to date reached more than 200 countries worldwide with more than 2 million confirmed cases (probably a much higher number of infected), and almost 200,000 deaths. Amplification of viral RNA by (real time) reverse transcription polymerase chain reaction (rRT-PCR) is the current gold standard test for confirmation of infection, although it presents known shortcomings: long turnaround times (3-4 hours to generate results), potential shortage of reagents, false-negative rates as large as 15-20%, the need for certified laboratories, expensive equipment and trained personnel.	2020	Springer

N°	Título del artículo científico	Autor(es)	Resumen	Año	Revista
7	A Machine Learning Model to Identify Early Stage Symptoms of SARSCov-2 Infected Patients	JMartuza Ahamad, Sakifa Aktar, Rashed-Al-Mahfuz, Shahadat Uddin, Pietro Lió, Haoming Xu, Matthew A. Summers, Julian M.W. Quinn, Mohammad Ali Moni	The recent outbreak of the respiratory ailment COVID-19 caused by novel coronavirus SARS-Cov2 is a severe and urgent global concern. In the absence of effective treatments, the main containment strategy is to reduce the contagion by the isolation of infected individuals; however, isolation of unaffected individuals is highly undesirable. To help make rapid decisions on treatment and isolation needs, it would be useful to determine which features presented by suspected infection cases are the best predictors of a positive diagnosis.	2020	Elsevier
8	Association between weather data and COVID-19 pandemic predicting mortality rate: Machine learning approaches	Zohair Malki, l-Sayed Atlama, Aboul Ella Hassanienc, Guesh Dagnewd, Mostafa A. Elhosseini, Ibrahim Gadb	Nowadays, a significant number of infectious diseases such as human coronavirus disease (COVID-19) are threatening the world by spreading at an alarming rate. Some of the literatures pointed out that the pandemic is exhibiting seasonal patterns in its spread, incidence and nature of the distribution. In connection to the spread and distribution of the infection, scientific analysis that answers the questions whether the next summer can save people from COVID-19 is required. Many researchers have been exclusively asked whether high temperature during summer can slow down the spread of the COVID-19 as it has with other seasonal flues.	2020	Elsevier

N°	Título del artículo científico	Autor(es)	Resumen	Año	Revista
9	Role of biological Data Mining and Machine Learning Techniques in Detecting and Diagnosing the Novel Coronavirus (COVID-19): A Systematic Review	A. S. Albahri, Rula A. Hamid, Jwan k. Alwan, Z.T. Al-qays,A. A. Zaidan, B. B.Zaidan1, A O. S. Albahrl A. H. AlAmoodi, Jamal Mawlood Khlaf, E. M. Almahdi, Eman Thabet, Suha M. Hadi, K I. Mohammed, M. A. Alsalem, Jameel R. Al-Obaidi, H.T. Madhloom	Coronaviruses (CoVs) are a large family of viruses that are common in many animal species, including camels, cattle, cats and bats. Animal CoVs, such as Middle East respiratory syndrome-CoV, severe acute respiratory syndrome (SARS)-CoV, and the new virus named SARS-CoV-2, rarely infect and spread among humans. On January 30, 2020, the International Health Regulations Emergency Committee of the World Health Organisation declared the outbreak of the resulting disease from this new CoV called ‘COVID-19’, as a ‘public health emergency of international concern’.	2020	springer

N°	Título del artículo científico	Autor(es)	Resumen	Año	Revista
10	COVID-19 Pandemic Prediction for Hungary; A Hybrid Machine Learning Approach	Gergo Pinter, Imre Felde, Amir Mosavi, Pedram Ghamisi and Richard Gloaguen	Several epidemiological models are being used around the world to project the number of infected individuals and the mortality rates of the COVID-19 outbreak. Advancing accurate prediction models is of utmost importance to take proper actions. Due to the lack of essential data and uncertainty, the epidemiological models have been challenged regarding the delivery of higher accuracy for long-term prediction. As an alternative to the susceptible-infected-resistant (SIR)-based models, this study proposes a hybrid machine learning approach to predict the COVID-19, and we exemplify its potential using data from Hungary.	2020	MDPI

N°	Título del artículo científico	Autor(es)	Resumen	Año	Revista
11	How to Improve Compliance with Protective Health Measures during the COVID-19 Outbreak: Testing a Moderated Mediation Model and MachineLearning Algorithms	Paolo Roma, Merylin Monaro, Laura Muzi, Marco Colasanti, Eleonora Ricci, Silvia Biondi, Christian Napoli, Stefano Ferracuti and Cristina Mazza	In the wake of the sudden spread of COVID-19, a large amount of the Italian population practiced incongruous behaviors with the protective health measures. The present study aimed at examining psychological and psychosocial variables that could predict behavioral compliance. An online survey was administered from 18–22 March 2020 to 2766 participants. Paired sample t-tests were run to compare efficacy perception with behavioral compliance. Mediation and moderated mediation models were constructed to explore the association between perceived efficacy and compliance, mediated by self-efficacy and moderated by risk perception and civic attitudes. Machine learning algorithms were trained to predict which individuals would be more likely to comply with protective measures.	2020	MDPI
12	Predicting Perceived Stress Related to the COVID-19 Outbreak through Stable Psychological Traits and Machine Learning Models	Luca Flesia, Merylin Monaro, Cristina Mazza Valentina Fietta, Elena Colicino, Barbara Segatto and Paolo Roma	The global SARS-CoV-2 outbreak and subsequent lockdown had a significant impact on people’s daily lives, with strong implications for stress levels due to the threat of contagion and restrictions to freedom. Given the link between high stress levels and adverse physical and mental consequences, the COVID-19 pandemic is certainly a global public health issue. In the present study, we assessed the effect of the pandemic on stress levels in N = 2053 Italian adults, and characterized more vulnerable individuals on the basis of sociodemographic features and stable psychological traits.	2020	MDPI

N°	Título del artículo científico	Autor(es)	Resumen	Año	Revista
13	Development of a Machine-Learning System to Classify Lung CT Scan Images into Normal/COVID-19 Class	Seifedine Kadry, Venkatesan Rajinikanth, Seungmin Rho, Nadaradjane Sri Madhava Raja Vaddi Seshagiri Rao Krishnan Palani Thanaraj	Recently, the lung infection due to Coronavirus Disease (COVID-19) affected a large human group worldwide and the assessment of the infection rate in the lung is essential for treatment planning. This research aims to propose a Machine-Learning-System (MLS) to detect the COVID-19 infection using the CT scan Slices (CTS). This MLS implements a sequence of methods, such as multi-thresholding, image separation using threshold filter, feature-extraction, feature-selection, feature-fusion and classification. The initial part implements the Chaotic-Bat-Algorithm and Kapur's Entropy (CBA+KE) thresholding to enhance the CTS. The threshold filter separates the image into two segments based on a chosen threshold 'Th'.	2020	Google académico

N°	Título del artículo científico	Autor(es)	Resumen	Año	Revista
14	A review of mathematical modeling, artificial intelligence and datasets used in the study, prediction and management of COVID-19	Youssoufa Mohamadou, Aminou Halidou, Pascalin Tiam Kapen	In the past few months, several works were published in regard to the dynamics and early detection of COVID-19 via mathematical modeling and Artificial intelligence (AI). The aim of this work is to provide the research community with comprehensive overview of the methods used in these studies as well as a compendium of available open source datasets in regard to COVID-19. In all, 61 journal articles, reports, fact sheets, and websites dealing with COVID-19 were studied and reviewed. It was found that most mathematical modeling done were based on the Susceptible-Exposed-Infected-Removed (SEIR) and Susceptible-infected-recovered (SIR) models while most of the AI implementations were Convolutional Neural Network (CNN) on X-ray and CT images.	2020	Springer
15	Machine Learning model to predict the number of cases contaminated by COVID-19	Allae Erraissi and Mouad Banane	This paper presents a dedicated machine learning model to predict the number of cases infected by the Corona Virus; the case of Morocco was chosen to validate this study. Completely realized in Spark ML with the 'Scala' language and tested for a certain number of algorithms generated on datasets coming from dedicated sources to gather Covid19 data in the world. The results show the possibility of achieving better scores prediction after using the proposed method. We tested our model on the case of China and the results were relevant. The proposed Machine	2020	Google académico

N°	Título del artículo científico	Autor(es)	Resumen	Año	Revista
			Learning model can be applied to data from any country in the world. We have applied it in this paper to the case of Morocco and China. We are sending this work to the world to help them fight this 2019 Corona Virus pandemic.		
16	An Interpretable Machine Learning Framework for Accurate Severe vs Non-severe COVID-19 Clinical Type Classification	Yuanfang Chen, Liu Ouyang, Forrest Sheng Bao, Qian Li, Lei Han, Baoli Zhu Yaorong Ge, Patrick Robinson, Ming Xu, Jie Liu and Shi Chen	Effectively and efficiently diagnosing COVID-19 patients with accurate clinical type is essential to achieve optimal outcomes for the patients as well as reducing the risk of overloading the healthcare system. Currently, severe and non-severe COVID-19 types are differentiated by only a few features, which do not comprehensively characterize the complicated pathological, physiological, and immunological responses to SARS-CoV-2 invasion in different types. In this study, we recruited 214 confirmed COVID-19 patients in non-severe and 148 in severe type, from Wuhan, China.	2020	Google académico
17	Predicting COVID-19 community mortality risk using machine learning and development of an online prognostic tool	Ashis Kumar Das, Shiba Mishra and Saji Saraswathy Gopalan	The recent pandemic of COVID-19 has emerged as a threat to global health security. There are very few prognostic models on COVID-19 using machine learning.	2020	PeerJ

N°	Título del artículo científico	Autor(es)	Resumen	Año	Revista
18	A Machine Learning Model Reveals Older Age and Delayed Hospitalization as Predictors of Mortality in Patients with COVID-19	Jit Sarkar1 and Partha Chakrabarti1	The recent pandemic of novel coronavirus disease 2019 (COVID-19) is increasingly causing severe acute respiratory syndrome (SARS) and significant mortality. We aim here to identify the risk factors associated with mortality of coronavirus infected persons using a supervised machine learning approach.	2020	Google académico
19	Using Machine Learning to Predict ICU Transfer in Hospitalized COVID-19 Patients	Fu-Yuan Cheng, Himanshu Joshi, Pranai Tandon, Robert Freeman, David L Reich, Madhu Mazumdar, Roopa Kohli-Seth, Matthew A. Levin, Prem Timsina 1 and Arash Kia	Objectives: Approximately 20–30% of patients with COVID-19 require hospitalization, and 5–12% may require critical care in an intensive care unit (ICU). A rapid surge in cases of severe COVID-19 will lead to a corresponding surge in demand for ICU care. Because of constraints on resources, frontline healthcare workers may be unable to provide the frequent monitoring and assessment required for all patients at high risk of clinical deterioration. We developed a machine learning-based risk prioritization tool that predicts ICU transfer within 24 h, seeking to facilitate efficient use of care providers' efforts and help hospitals plan their flow of operations.	2020	MDPI

N°	Título del artículo científico	Autor(es)	Resumen	Año	Revista
20	COVID-19 and stock market volatility: An industry level analysis	Seungho Baek, Sunil K. Mohanty, Mina Glambosky	COVID-19 has had significant impact on US stock market volatility. This study focuses on understanding the regime change from lower to higher volatility identified with a Markov Switching AR model. Utilizing machine learning feature selection methods, economic indicators are chosen to best explain changes in volatility. Results show that volatility is affected by specific economic indicators and is sensitive to COVID-19 news.	2020	Elsevier
21	Risk of a second wave of COVID-19 infections: using artificial intelligence to investigate stringency of physical distancing policies in North America	Shashank Vaid, Aaron McAdie, Ran Kremer, Vikas Khanduja, Mohit Bhandari	Accurately forecasting the occurrence of future COVID-19-related cases across relaxed (Sweden) and stringent (USA and Canada) policy contexts has a renewed sense of urgency. Moreover, there is a need for a multidimensional county-level approach to monitor the second wave of COVID-19 in the USA.	2020	Springer

N°	Título del artículo científico	Autor(es)	Resumen	Año	Revista
22	Machine Learning for Mortality Analysis in Patients with COVID-19	Manuel Sánchez-Montañés, Pablo Rodríguez-Belenguer, Antonio J. Serrano-López Emilio Soria-Olivas and Yasser Alakhdar-Mohmara	This paper analyzes a sample of patients hospitalized with COVID-19 in the region of Madrid (Spain). Survival analysis, logistic regression, and machine learning techniques (both supervised and unsupervised) are applied to carry out the analysis where the endpoint variable is the reason for hospital discharge (home or deceased). The different methods applied show the importance of variables such as age, O2 saturation at Emergency Rooms (ER), and whether the patient comes from a nursing home. In addition, biclustering is used to globally analyze the patient-drug dataset, extracting segments of patients. We highlight the validity of the classifiers developed to predict the mortality, reaching an appreciable accuracy. Finally, interpretable decision rules for estimating the risk of mortality of patients can be obtained from the decision tree, which can be crucial in the prioritization of medical care and resources.	2020	MDPI

N°	Título del artículo científico	Autor(es)	Resumen	Año	Revista
23	Using Machine Learning Methods to Predict Physician-Hospital Integration	Stuart Craig, Matthew Grennan, Joseph Martinez, and Ashley Swanson	In this document, we propose a new method that combines high-dimensional data with machine learning methods to predict physician-hospital integration. We compare the performance of this method with alternative approaches used in the healthcare economics literature for a large validated sample, finding that it outperforms previous methods by a substantial margin. We also compare the static predictions of this model to a large validated sample recently made available by the Agency for Healthcare Research and Quality (AHRQ) and again document a high degree of accuracy. Finally, we briefly summarize the implications of our method for the growth in physician-hospital integration over the years 2008-2016.	2020	Google académico
24	Predicting the Individual Treatment Effect of Neurosurgery for TBI Patients in the Low Resource Setting: A Machine Learning Approach in Uganda	Syed M. Adil, BS, Cyrus Elahi, MD, Robert Gramer, MD, Charis A. Spears, BA, Anthony T, Fuller, MD, Michael M. Haglund, MD, PhD; Timothy W. Dunn, PhD	Traumatic brain injury (TBI) disproportionately affects low- and middle-income countries (LMICs). In these low-resource settings, effective triage of patients with TBI-including the decision of whether or not to perform neurosurgery-is critical in optimizing patient outcomes and healthcare resource utilization. Machine learning may allow for effective predictions of patient outcomes both with and without surgery. Data from patients with TBI was collected prospectively at Mulago National Referral Hospital in Kampala, Uganda, from 2016 to 2019. One linear and six non-linear machine learning models were designed to	2020	Google académico

N°	Título del artículo científico	Autor(es)	Resumen	Año	Revista
			predict good versus poor outcome near hospital discharge and internally validated using nested five-fold cross-validation.		
25	Building a hospital referral expert system with a Prediction and Optimization-Based Decision Support System algorithm	Chih-Lin Chi, W. Nick Street, Marcia M. Ward	This study presents a new method for constructing an expert system using a hospital referral problem as an example. Many factors, such as institutional characteristics, patient risks, traveling distance, and chances of survival and complications should be included in the hospital-selection decision. Ideally, each patient should be treated individually, with the decision process including not only their condition but also their beliefs about trade-offs among the desired hospital features.	2018	Elsevier

N°	Título del artículo científico	Autor(es)	Resumen	Año	Revista
26	Analysis, Prediction and Evaluation of COVID-19 Datasets using Machine Learning Algorithms	Kolla Bhanu Prakash, S. Sagar Imambi, Mohammed Ismail, T Pavan Kumar, YVR Naga Pawan	COVID-19, Corona Virus Diasease-2019, belongs to genus of Coronaviridae. A virus with no vaccine creating unpredictable havocs in the human lives and financial and economic systems in every country throughout the world. It is precariously halted everything in the society mercilessly. An analysis on COVID-19 datasets to understand which age group is mostly effected due to COVID-19. Different prediction models are built using machine learning algorithms and their performances are computed and evaluated. Random Forest Regressor and Random Forest Classifier outperformed the other machine learning models like SVM, KNN+NCA, Decision Tree Classifier, Gaussian Naïve Bayesian Classifier, Multilinear Regression, Logistic Regression and XGBoost Classifier.	2020	Google académico
27	Machine-learning algorithm that can improve the diagnostic accuracy of septic arthritis of the knee	Eun-Seok Choi, Jae Ang Sim, Young Gon Na, Jong- Keun Seon, Hyun Dae Shin	Prompt diagnosis and treatment of septic arthritis of the knee is crucial. Nevertheless, the quality of evidence for the diagnosis of septic arthritis is low. In this study, the authors developed a machine learning-based diagnostic algorithm for septic arthritis of the native knee using clinical data in an emergency department and validated its diagnostic accuracy.	2021	Springer

N°	Título del artículo científico	Autor(es)	Resumen	Año	Revista
28	Coronavirus Infectious Disease (COVID-19) Modeling: Evidence of Geographical Signals	Ali Keshavarzi	2019 Novel Coronavirus (2019-nCoV) is a virus (more specifically, coronavirus) known as the cause of respiratory disease outbreak that was first detected in Wuhan, China. In the present study, we demonstrated the applicability of four different data mining techniques namely Decision Tree (DT), Random Forests (RF), Logistic Model Trees (LMT) and Naive Bayes (NB) classifiers to model and present the development of Coronavirus disease (COVID-19) based on 482 records of cases. Johns Hopkins University has created an outstanding database using the data from the affected cases (Johns Hopkins Github repository).	2020	Google académico

N°	Título del artículo científico	Autor(es)	Resumen	Año	Revista
29	COVID-19 rapid test by combining a random forest based web system and blood tests	Valter Augusto de Freitas Barbosa, Juliana Carneiro Gomes, Maíra Araújo de Santana, Clarisse Lins de Lima, Raquel Bezerra Calado, Claúdio Roberto Bertoldo Júnior, Jeniffer Emidio de Almeida Albuquerque, Rodrigo Gomes de Souza, Ricardo Juarez Escorel de Araújo, Ricardo Emmanuel de Souza, Wellington Pinheiro dos Santos	The disease caused by the new type of coronavirus, the COVID-19, has posed major public health challenges for many countries. With its rapid spread, since the beginning of the outbreak in December 2019, the disease transmitted by SARS-Cov2 has already caused over 400 thousand deaths to date. The diagnosis of the disease has an important role in combating COVID-19.	2020	Google académico

N°	Título del artículo científico	Autor(es)	Resumen	Año	Revista
30	Accurate Prediction of COVID-19 using Chest X-Ray Images through Deep Feature Learning model with SMOTE and Machine Learning Classifiers	Rahul Kumar, Ridhi Arora, Vipul Bans, Vinodh J Sahayasheela, Himanshu Buckchashaved Imran, Narayanan Narayanan, Ganesh N Pandian and Balasubramanian Raman	According to the World Health Organization (WHO), the coronavirus (COVID-19) pandemic is putting even the best healthcare systems across the world under tremendous pressure. The early detection of this type of virus will help in relieving the pressure of the healthcare systems. Chest X-rays has been playing a crucial role in the diagnosis of diseases like Pneumonia. As COVID-19 is a type of influenza, it is possible to diagnose using this imaging technique. With rapid development in the area of Machine Learning (ML) and Deep learning, there had been intelligent systems to classify between Pneumonia and Normal patients.	2020	Google académico

Elaboración: Fuentes Melina y Medina Wilmer.
Fuente: Propia.

Anexo 7. Acta de entrega y recepción definitiva

En la ciudad de Guayaquil, marzo del 2021

Por el presente documento.

Los estudiantes no titulados de la Carrera de Ingeniería en Sistemas Computacionales Fuentes Marmolejo Melina Daniela con cédula de identidad N° 0950786988 y Medina Parra Wilmer David con cédula de identidad N° 0940665326 hacemos la entrega del código fuente del proyecto de titulación a la Dirección de la Carrera de Ingeniería en Sistemas Computacionales en un medio magnético.

Los códigos del programa/producto que se encargaron por compromiso al estar inserto en el proceso de titulación desde 23 de noviembre del 2020.

Para efectos de dar cumplimiento a la entrega del código fuente, cedo todos los derechos de explotación sobre el programa y, en concreto, los de transformación, comunicación pública, distribución y reproducción, de forma exclusiva, con un ámbito territorial nacional.

Fuentes Marmolejo Melina Daniela	0950786988 Cédula de identidad N°
Medina Parra Wilmer David	0940665326 Cédula de identidad N°

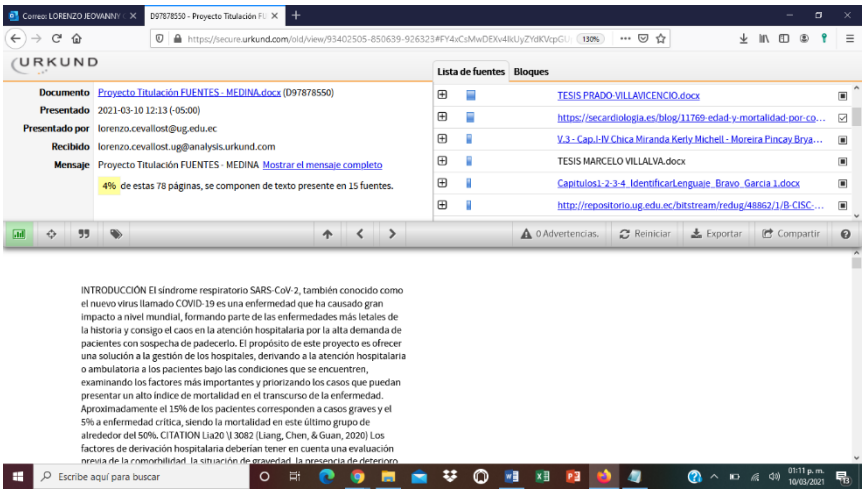
Elaboración: Fuentes Melina y Medina Wilmer.
Fuente: Propia.

Anexo 8. Certificado porcentaje de similitud

CERTIFICADO PORCENTAJE DE SIMILITUD

Habiendo sido nombrado Ing. Lorenzo Cevallos Torres, M.Sc. Tutor del trabajo de titulación certifico que el presente trabajo de titulación ha sido elaborado por Fuentes Marmolejo Melina Daniela con cédula de identidad N° 0950789688 y Medina Parra Wilmer David con cédula de identidad N° 0940665326, con mi respectiva supervisión como requerimiento parcial para la obtención del título de Ingeniería en sistemas computacionales.

Se informa que el trabajo de Titulación: "**Diseño de un modelo predictivo-asistencial de pacientes infectados por COVID-19, mediante un modelo supervisado de Machine Learning basado en criterios de derivación hospitalaria o ambulatoria**", ha sido orientado durante todo el periodo de ejecución en el programa anti plagio URKUND quedando el 4% de coincidencia.



https://secure.urkund.com/old/view/93402505-850639-926323#FY4xCsMwDEXv4lkUyZYdKVcpGUpoi4dmyVh6974MD330n0Df8jnLejdNMVMwqNDAocMQa8x2zQUC8B3XcR3XcZ3e6Z3e6Tu5Mwf7hbywD3KQ48p0yX1yn9xnl2gSLkEYgoaRKmmblHO+j/ma++PYn2XVmzYbWdWt1+rJf78/

Firmado electrónicamente por:
LORENZO JOVANNY CEVALLOS TORRES

Ing. Lorenzo Cevallos Torres, M.Sc.
C.C.: 0914517966
Fecha: 17 de marzo del 2021

Anexo 9. Manual técnico

UNIVERSIDAD DE GUAYAQUIL

FACULTAD DE CIENCIAS MATEMÁTICAS Y FÍSICAS

CARRERA DE INGENIERÍA EN SISTEMAS COMPUTACIONALES

DISEÑO DE UN MODELO PREDICTIVO-ASISTENCIAL DE PACIENTES INFECTADOS POR COVID-19, MEDIANTE UN MODELO SUPERVISADO DE MACHINE LEARNING BASADO EN CRITERIOS DE DERIVACIÓN HOSPITALARIA O AMBULATORIA

PROYECTO DE TITULACIÓN

Previa a la obtención del Título de:

INGENIERO EN SISTEMAS COMPUTACIONALES

AUTORES:

FUENTES MARMOLEJO MELINA DANIELA

MEDINA PARRA WILMER DAVID

TUTOR:

ING. LORENZO CEVALLOS TORRES, M.Sc.

GUAYAQUIL – ECUADOR

2020

ÍNDICE GENERAL

INTRODUCCIÓN

En el manual técnico se detalla toda aquella información y pasos a seguir para el correcto soporte del prototipo alojado en el sitio web www.covid19gye.com, haciendo énfasis en la sección “Derivación Hospitalaria”. Este manual va dirigido para todo personal encargado del área técnica.

Objetivo

Instruir al personal en la estructura del prototipo y del sitio web, mediante un manual técnico para proveer el soporte pertinente en cada una de sus actualizaciones.

Instalación de herramientas

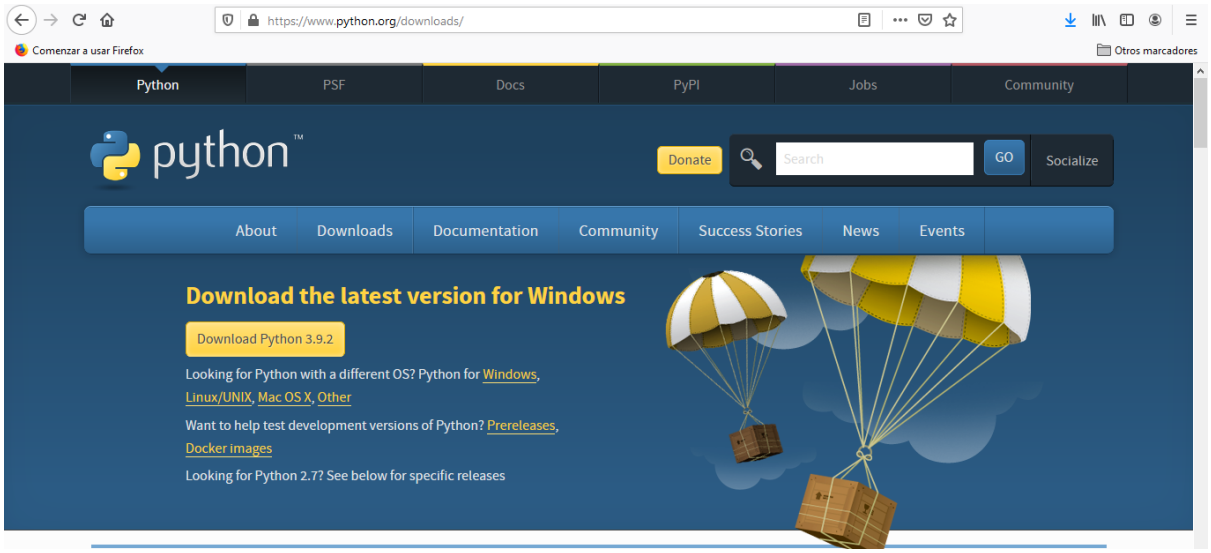
Requisitos de Software: Sistema Operativo Windows 7 de 64 bits.

Requisitos de Hardware: Se debe tomar en consideración las siguientes características mínimas con respecto al hardware para la implementación del prototipo en un ambiente web:

- Procesador Intel i5 de 8va Generación.
- Mínimo 4 GB RAM.
- 500 GB de disco duro en adelante.

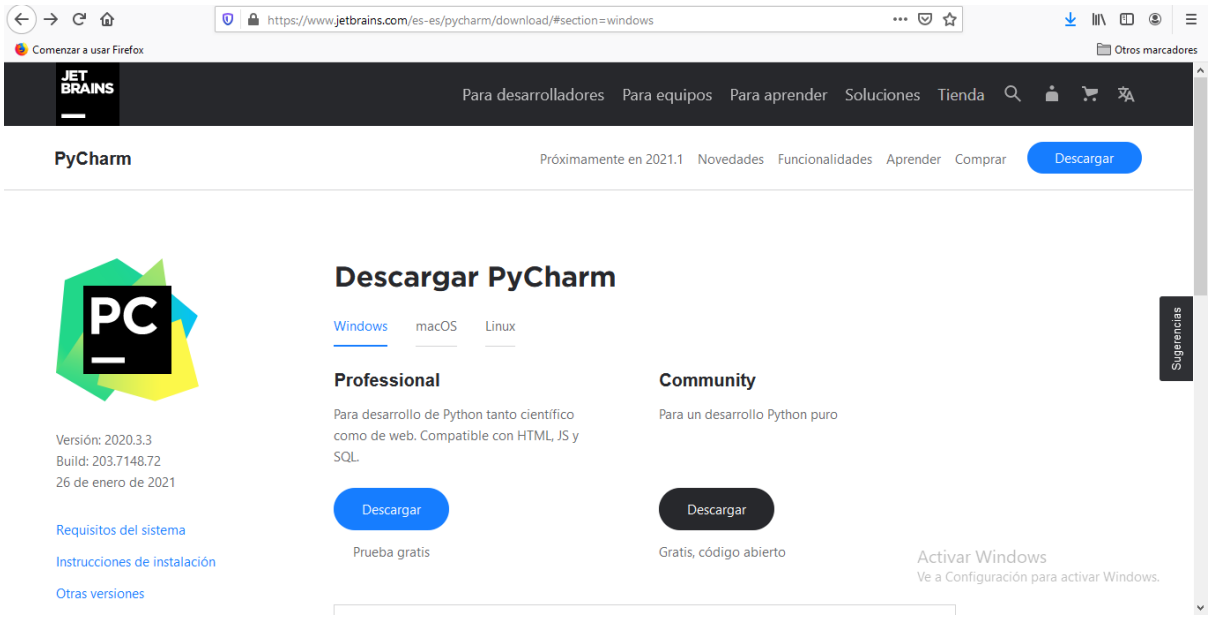
Instalación de Python 3.5: Para la creación de los algoritmos predictivos es esencial la instalación de Python a través de la página oficial: https://www.python.org/downloads/ .

Se escoge la versión de acuerdo a las características del equipo, en este caso, debido a ser Windows 7.



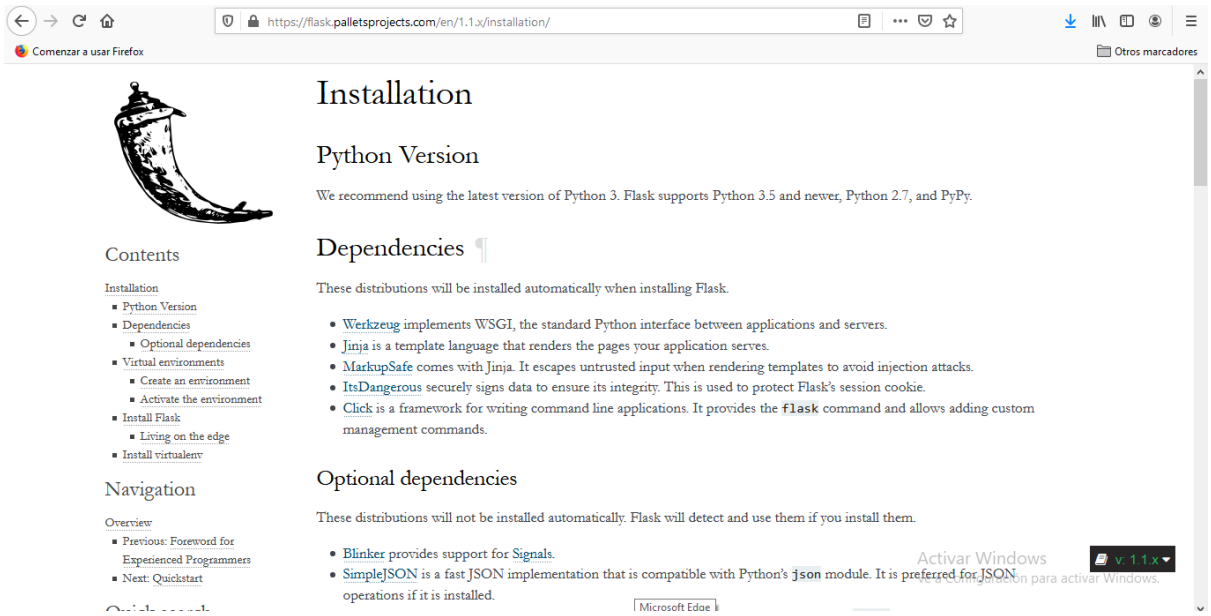
Instalación de Pycharm: Para la edición de código de Python para los algoritmos y creación del sitio web se utilizó el IDE Pycharm, el cual sólo necesita ser descargada mediante el enlace: https://www.jetbrains.com/es-es/pycharm/download/#section=windows

Una vez ingresado a la página de descargas, se selecciona la opción gratuita Community.

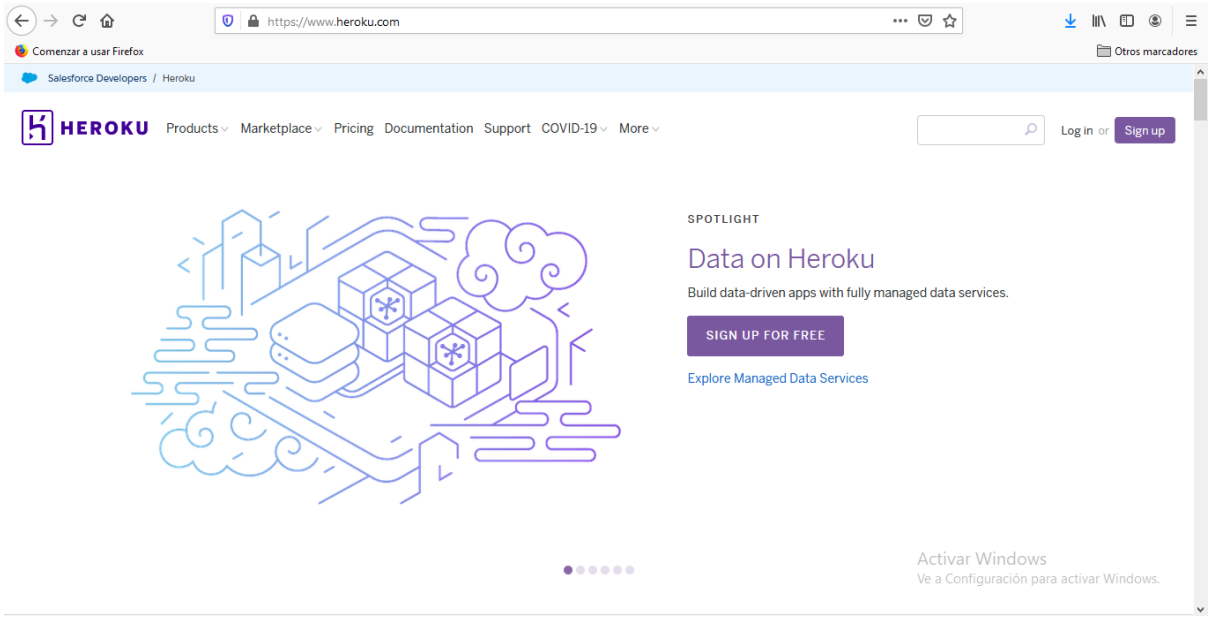


Instalación de Framework Flask: Posterior a la instalación de las anteriores herramientas, también se necesita el framework flask para el entrenamiento del algoritmo y subir al servidor. Para su instalación se escribe el comando “pip install Flask” desde la consola de comando del equipo. Se recomienda tener la última versión de Python 3 aunque soporta

desde la versión 2.7. A continuación, se adjunta el link para acceder a la instalación de la herramienta: https://flask.palletsprojects.com/en/1.1.x/installation/

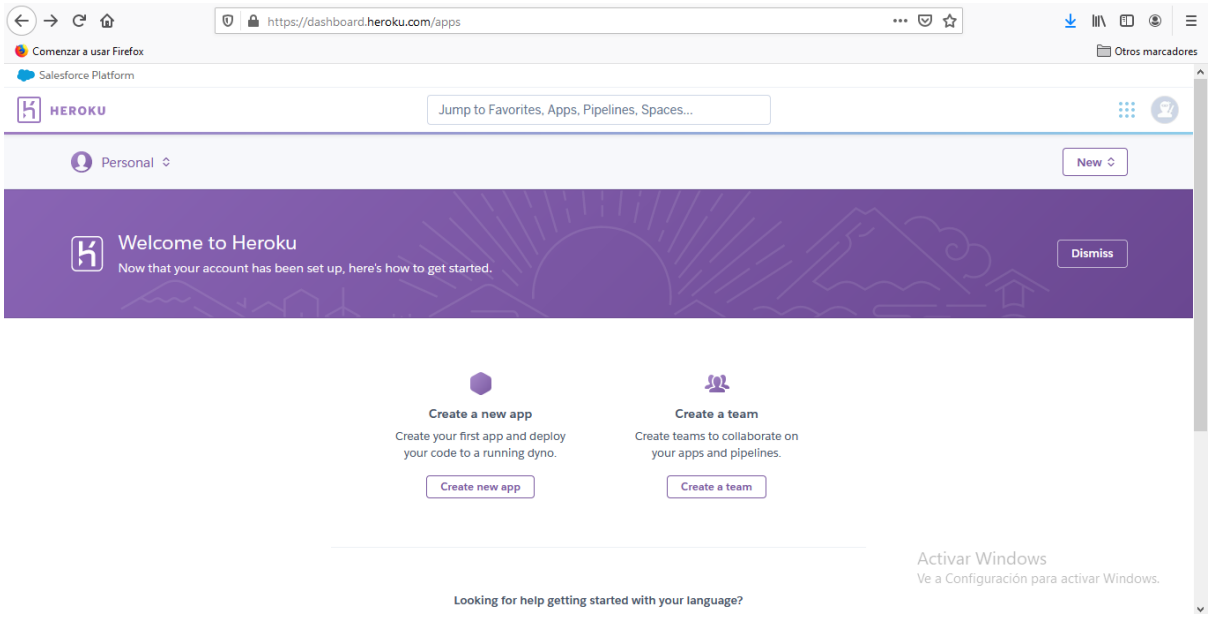


Instalación de Heroku: Para la descarga de esta plataforma en la nube, se necesita crear una cuenta ingresando al siguiente enlace: https://www.heroku.com/ en la opción "sign up" y confirmando el registro vía correo electrónico.

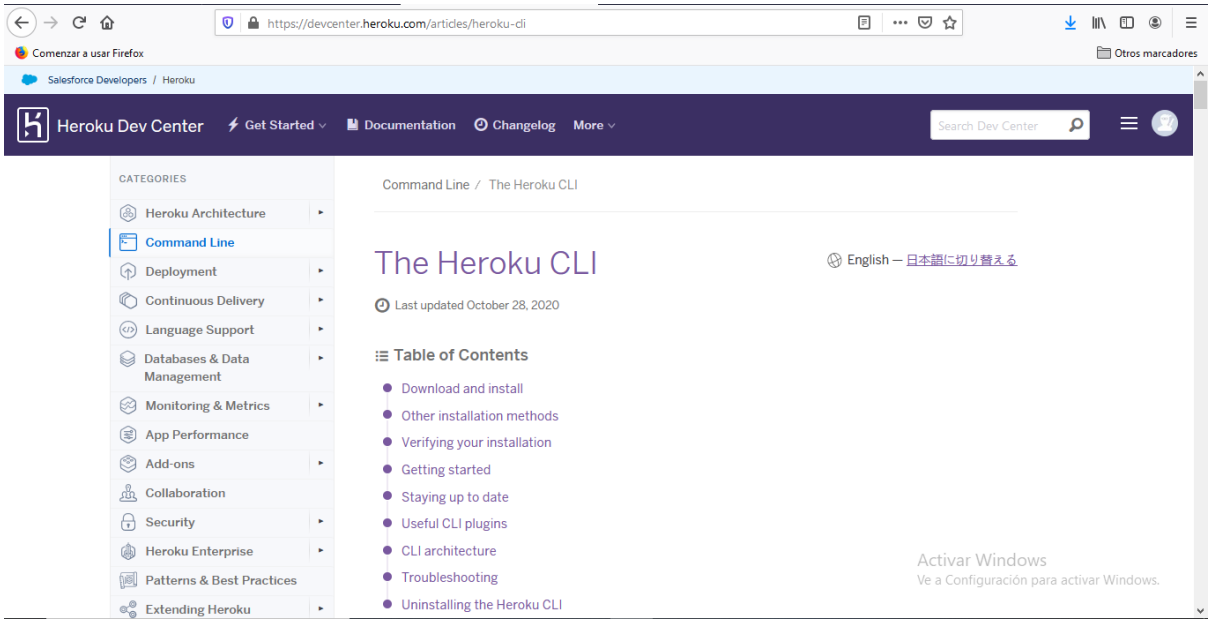


Una vez realizado dicho paso, se procede a elegir "créate a new app" traducido al español como "crear una nueva aplicación", donde se alojará todo el código desarrollado en

Python, en el cual los algoritmos elegidos ya se encuentran debidamente entrenados y preparados para la predicción por medio del uso de un cliente GitHub.

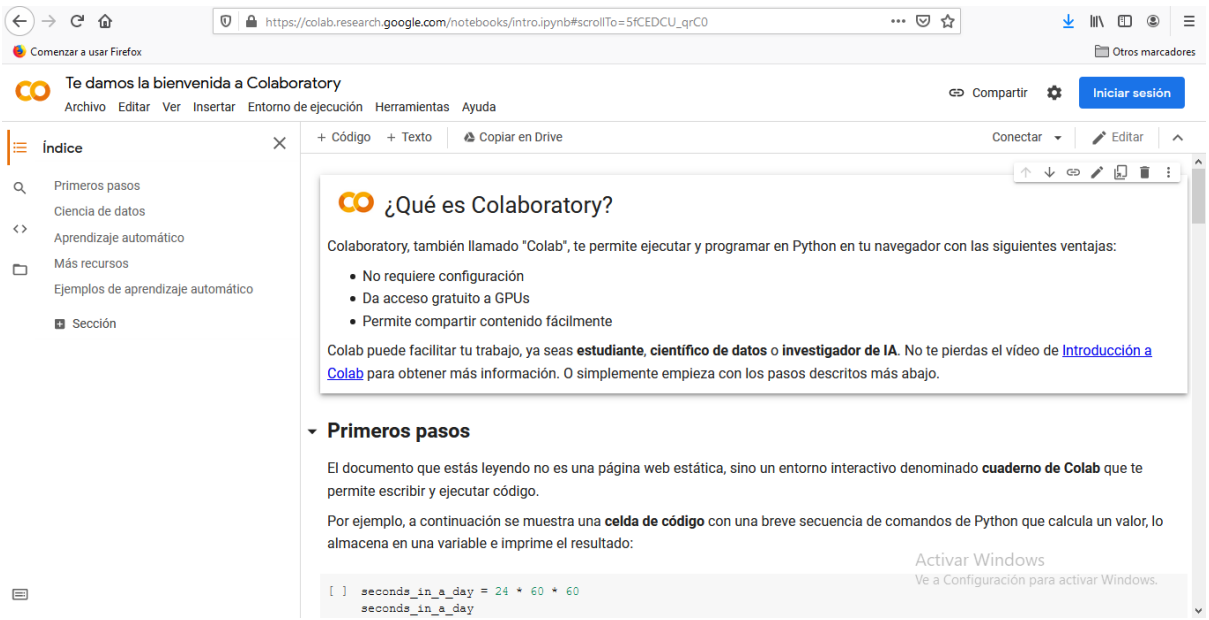


El siguiente enlace https://devcenter.heroku.com/articles/heroku-cli llevará a la instalación del CLI (Coman Line Interface), el cual brindará la opción de deploy de Python a esta plataforma y usar los comandos de Git en la propia consola.



Google Colab: Se debe crear una cuenta Gmail y Google Drive para hacer uso de esta plataforma y ya realizando ese procedimiento se ingresa al siguiente enlace: https://colab.research.google.com/notebooks/intro.ipynb#scrollTo=5fCEDCU_qrC0, este

llevará a la página para la creación de un nuevo proyecto en Python, cabe resaltar que no necesita instalación previa al ser un alojamiento en la nube. Cada que se escriba alguna línea de código y se desee ejecutar sólo deberá hacer clic en el símbolo de reproducir.



PHP: Se utilizó PHP puro para traer los datos del web service para posteriormente mostrarlos en el sitio web www.covid19gye.com, mediante curl para la comunicación con el web service, traer el valor de la predicción, en este caso 0 para ambulatoria y 1 para hospitalaria e imprimirlo en la página. En el siguiente enlace se muestra una guía con respecto a este procedimiento: https://www.php.net/manual/es/book.curl.php.

Anexo 10. Manual de usuario

UNIVERSIDAD DE GUAYAQUIL

FACULTAD DE CIENCIAS MATEMÁTICAS Y FÍSICAS

CARRERA DE INGENIERÍA EN SISTEMAS COMPUTACIONALES

DISEÑO DE UN MODELO PREDICTIVO-ASISTENCIAL DE PACIENTES INFECTADOS POR COVID-19, MEDIANTE UN MODELO SUPERVISADO DE MACHINE LEARNING BASADO EN CRITERIOS DE DERIVACIÓN HOSPITALARIA O AMBULATORIA

PROYECTO DE TITULACIÓN

Previa a la obtención del Título de:

INGENIERO EN SISTEMAS COMPUTACIONALES

AUTORES:

FUENTES MARMOLEJO MELINA DANIELA

MEDINA PARRA WILMER DAVID

TUTOR:

ING. LORENZO CEVALLOS TORRES, M.Sc.

GUAYAQUIL – ECUADOR
2020

ÍNDICE GENERAL

Pantalla inicio

Paso 1.- Ingresar a la página web www.covid19gye.com. Aquí se visualiza información actualizada acerca del COVID-19.

Proyectos de titulación

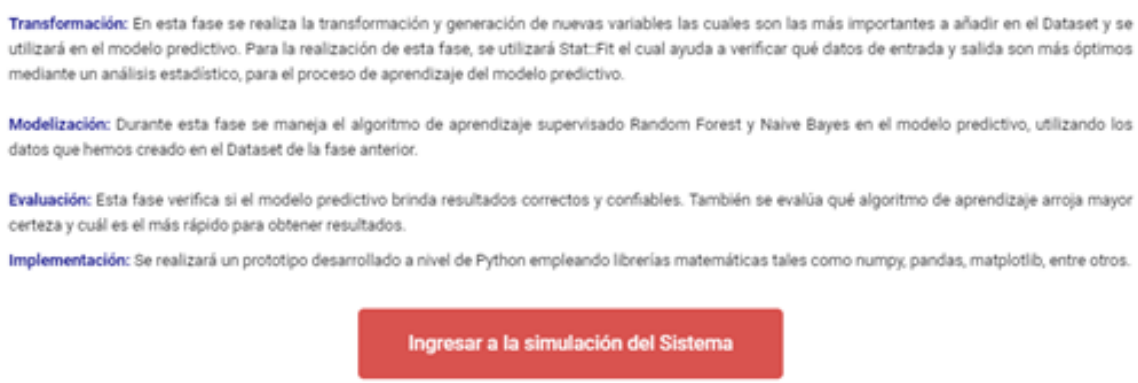
Paso 2.- Bajando en la misma página, se encuentran los proyectos FCI a cargo del Ing. Lorenzo Cevallos. El usuario debe ingresar al proyecto “Derivación Hospitalaria”.

Información sobre Derivación Hospitalaria

Paso 3.- En esta página, el usuario observa información detallada acerca del proyecto de titulación tales como: antecedentes, objetivo general y específicos, y la metodología que se empleó.

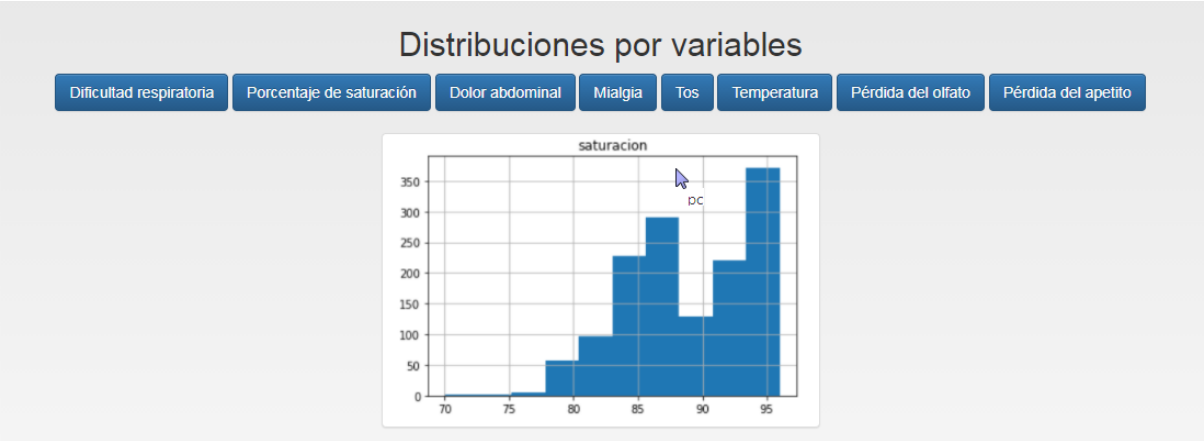
Ingreso a la simulación del sistema

Paso 4.- Al final de la página, el usuario podrá apreciar un botón que lo dirige a la simulación del sistema.



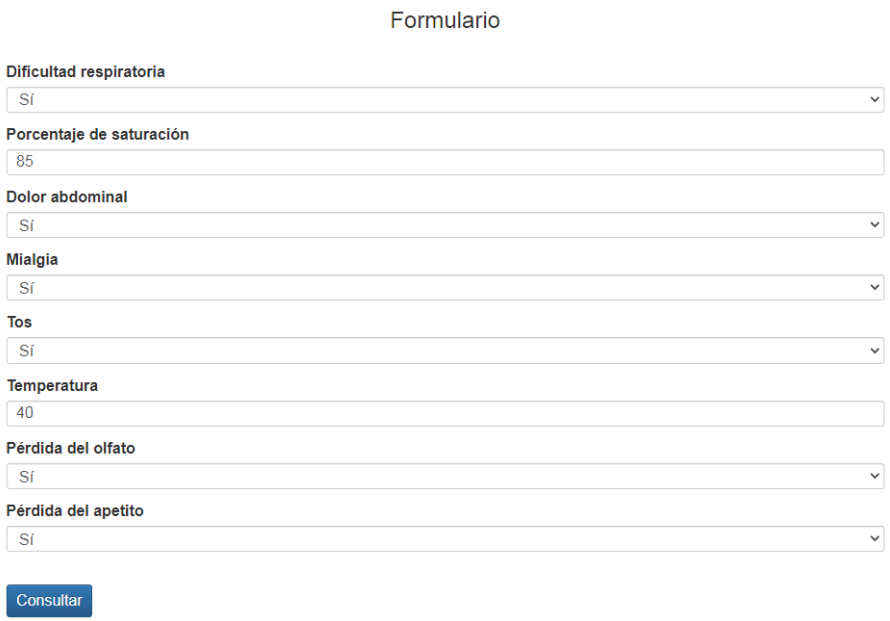
Distribución de las variables

Paso 5.- Dentro de esta página, se visualizan las distribuciones por variables, las cuales han sido utilizadas en el dataset para alimentar a los algoritmos de predicción. Para visualizar cada distribución, basta con dar clic en cada botón.



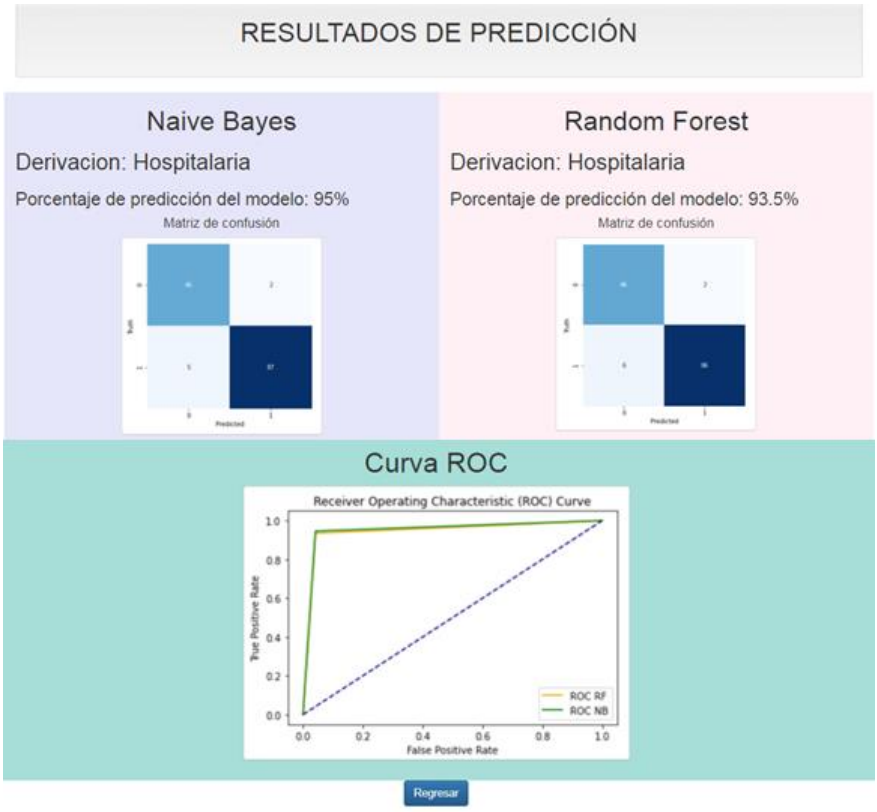
Formulario

Paso 6.- En esta sección, se aprecia el formulario donde el usuario ingresa los síntomas del paciente con COVID-19 para consultar si debe ser hospitalizado o enviado a domicilio para su respectivo tratamiento ambulatorio.



Resultados de los algoritmos

Paso 7.- Una vez realizada la consulta, el usuario visualiza los resultados de predicción por cada algoritmo supervisado, el cual predice en tiempo real si el paciente obtiene una derivación hospitalaria o ambulatoria. También se observa el porcentaje de predicción que se obtiene de cada algoritmo incluyendo su matriz de confusión; estos valores son los resultados que se obtuvieron durante el proceso de entrenamiento de los modelos predictivos. Por último, se aprecia la curva de ROC, donde compara el porcentaje de los dos algoritmos de Machine Learning aplicados en este proyecto.



Anexo 11. Artículo científico

(Primera versión)

ECUADORIAN SCIENCE JOURNAL VOL. xx No. x, xxxx – xxxx (xx–xx)

DOI: https://doi.org/10.46480/esj.x.x.xx

Diseño de un modelo predictivo-asistencial de pacientes infectados por COVID-19, mediante un modelo supervisado de Machine Learning basado en criterios de derivación hospitalaria o ambulatoria

Design of a predictive-care model for patients infected by COVID-19, through a supervised model of Machine Learning based on hospital or outpatient referral criteria

Melina Fuentes[1], Wilmer Medina[2], y Lorenzo Cevallos[3]

RESUMEN

Un problema derivado de la pandemia del COVID-19 es la falta de una herramienta digital que pueda predecir la intensidad de la gravedad de un paciente enfermo. El presente proyecto consiste en realizar un modelo predictivo asistencial para pacientes infectados por COVID-19, utilizando herramientas de Machine Learning mediante algoritmos de aprendizaje supervisado como Naive Bayes y Random Forest para obtener un criterio sobre derivación hospitalaria o ambulatoria. Entre los principales objetivos específicos se encuentran la extracción de un conjunto de base de datos con la información vinculada al historial médico de los pacientes diagnosticados con COVID-19, para la depuración y construcción de un dataset con las variables relacionadas, y evaluarlas para mejorar la toma de decisiones a partir de un modelo de algoritmo supervisado. La metodología empleada es "Knowledge Discovery in Databases – KDD", la cual se desarrolla en 6 fases: importación y muestreo de datos, calidad de datos, transformación, modelización, evaluación e implementación; sin embargo, esta última fase no se llevará a cabo, en su lugar se realizará un prototipo desarrollado a nivel de Python. Se utilizó la librería sklearn de la herramienta Python 3.5 para el entrenamiento del algoritmo, la herramienta Stat:Fit para las distribuciones estadísticas, y basándose en la sintomatología del paciente los algoritmos arrojaron un porcentaje de precisión (93,5% Random Forest y 95% Naive Bayes), concluyendo que el mejor predictor es el algoritmo de Naive Bayes, también se demostró que existe relación entre ambos algoritmos con respecto a la derivación hospitalaria o ambulatoria mediante el análisis de correlación de Pearson, haciendo que se cumplan las hipótesis planteadas. Al ser útil este prototipo para la toma de decisiones para la respectiva derivación del paciente, los beneficiarios directos son los doctores, dado que obtienen una herramienta que les agilitará la exhaustiva acción de decidir.

Palabras clave: DERIVACIÓN HOSPITALARIA, COVID-19, APRENDIZAJE AUTOMÁTICO, BOSQUES ALEATORIOS, REDES BAYESIANAS.

ABSTRACT

A problem derived from the COVID-19 pandemic is the lack of a digital tool that can predict the severity of a sick patient. The present project consists of carrying out a predictive model of care for patients infected by COVID-19, using Machine Learning tools using supervised learning algorithms such as Naive Bayes and Random Forest to obtain criteria on hospital or outpatient referral. Among the main specific objectives are the extraction of a set of databases with the information related to the medical history of patients diagnosed with COVID-19, for the purification and construction of a dataset with the related variables and evaluate them to improve the decision making based on a supervised algorithm model. The methodology used is "Knowledge Discovery in Databases - KDD", which is developed in 6 phases: import and data sampling, data quality, transformation, modeling, evaluation and implementation; however, this last phase will not be carried out, instead a prototype developed at the Python level will be made. The sklearn library of the Python 3.5 tool was used for the training of the algorithm, the Stat: Fit tool for the statistical distributions, and based on the patient's symptoms, the algorithms yielded a percentage of precision (93.5% Random Forest and 95% Naive Bayes), concluding that the best predictor is the Naive Bayes algorithm, it was also shown that there is a relationship between both algorithms with respect to hospital or outpatient referral by means of Pearson's correlation analysis, making the hypotheses raised. As this prototype is useful for decision-making for the respective referral of the patient, the direct beneficiaries are the doctors, since they obtain a tool that will expedite the exhaustive decision-making action.

Keywords: HOSPITAL REFERRAL, COVID-19, MACHINE LEARNING, RANDOM FOREST, NAIVE BAYES.

Fecha de recepción: __, __, 2021.
Fecha de aceptación: __, __, 2021.

Introducción

El síndrome respiratorio SARS-CoV-2, también conocido como el nuevo virus llamado COVID-19 es una enfermedad que ha causado gran impacto a nivel mundial, formando parte de las enfermedades más letales de la historia y consigo el caos en la atención hospitalaria por la alta demanda de pacientes con sospecha de padecerlo. El propósito de este proyecto es ofrecer una solución a la gestión de los hospitales, derivando a la atención hospitalaria o ambulatoria a los pacientes bajo las condiciones que se encuentren, examinando los factores más importantes y priorizando los casos que puedan presentar un alto índice de mortalidad en el transcurso de la enfermedad.

Aproximadamente el 15% de los pacientes corresponden a casos graves y el 5% a enfermedad crítica, siendo la mortalidad en este último grupo de alrededor del 50%, indican Liang [1].

Los factores de derivación hospitalaria deberían tener en cuenta una evaluación previa de la comorbilidad, la situación de gravedad, la presencia de deterioro cognitivo grave y la dependencia o la necesidad de soporte ventilatorio en pacientes graves señalan Blanco [2].

Al observar que los elementos principales para la derivación hospitalaria o ambulatoria depende de ciertos factores que conlleva el paciente, la disponibilidad de camas en los centros hospitalarios llegó a ser un inconveniente muy grande cuando la pandemia llegó a su punto más alto, teniendo que acomodar otras áreas en los hospitales para cubrir el mayor número de casos posibles. De cara a la apertura de nuevas unidades en un periodo tan corto de tiempo, los principales problemas afectaron a la disponibilidad de infraestructura, personal y material según lo indican Bardi [3] en un estudio realizado sobre la derivación hospitalaria y la respuesta al virus en España.

Este proyecto hace uso de una de las ramas de la Inteligencia Artificial (IA), la cual es el Machine Learning o en español Aprendizaje Automático, siendo muy útil en diferentes áreas, tales como: educación, financiera, transporte, salud, entre otros; siendo este último tema central para tratar. Recientemente la IA comenzó a establecerse en la medicina para mejorar la atención al paciente con la aceleración de procedimientos, consiguiendo mayor precisión en el diagnóstico de enfermedades, dando oportunidades de ofrecer mejores condiciones de atención médica, plantean Ávila [4]. Se plantea usar dos algoritmos de aprendizaje de máquina supervisado, los cuales son Naive Bayes y Random Forest cuyo objetivo es obtener un modelo predictivo-asistencial para la derivación hospitalaria o ambulatoria para pacientes con COVID-19 y finalmente realizar una comparación para seleccionar el más preciso.

La estructura del presente trabajo es la siguiente: en el capítulo I: se muestra el Planteamiento del problema, y causas y consecuencias; capítulo II se exhiben el Marco Teórico, dando referencia a investigaciones relevantes e hipótesis; capítulo III sobresale la parte estadística del proyecto, estableciendo la Metodología del desarrollo del prototipo y los resultados que arroja la investigación por los algoritmos supervisados de Machine Learning; y como último capítulo IV se presentan las conclusiones del proyecto y las hipótesis.

[1] Estudiante de Ingeniería en Sistemas Computacionales. Universidad de Guayaquil. Ecuador. E-mail: melina.fuentesm@ug.edu.ec
[2] Estudiante de Ingeniería en Sistemas Computacionales. Universidad de Guayaquil. Ecuador. E-mail: wilmer.medinap@ug.edu.ec
[3] Ing. en Estadística Informática, M.Sc. en Modelado Computacional en Ingeniería. Universidad de Guayaquil, Ecuador. E-mail: lorenzo.cevallost@ug.edu.ec

Como citar: Fuentes Marmolejo, M., Medina Parra, W., Cevallos Torres, L. (2021). Diseño de un modelo predictivo-asistencial de pacientes infectados por COVID-19, mediante un modelo supervisado de Machine Learning basado en criterios de derivación hospitalaria o ambulatoria. Ecuadorian Science Journal. xx(x), xx-xx.
DOI: https://doi.org/10.46480/esj.x.x.xx

Planteamiento del problema

El principal problema que ahonda la pandemia por el virus COVID-19, es que los centros médicos no proceden a un control estricto para los pacientes que llegan infectados. Cuando un paciente llega a dicho sitio con síntomas iniciales, suelen ser internados, esto provoca que los hospitales se llenen y no haya espacio para ningún paciente con síntomas graves.

Causas y consecuencias

En la Tabla 1 se describen aquellas causas y consecuencias relacionadas al problema y la importante creación de una herramienta tecnológica para la mejora de la gestión hospitalaria. El problema central es identificado en la causa 5: inexistencia de un prototipo para la derivación hospitalaria o ambulatoria, debido a que presenta como consecuencias la pésima gestión hospitalaria.

Tabla 1. Matriz de causas y consecuencias del problema

Causas	Consecuencias
C1. Desconocimiento de síntomas.	E1. Propagación masiva del virus.
C2. Persona adulto mayor.	E2. Mayor riesgo de mortalidad.
C3. Comorbilidades.	E3. Complicaciones en la salud del paciente.
C4. Desconocimiento de protocolos a seguir.	E4. Negligencia médica.
C5. Inexistencia de un prototipo para la derivación hospitalaria o ambulatoria.	E5. Mala gestión hospitalaria, aglomeraciones en hospitales y nula disponibilidad de camas.

Fuente: Fuentes, M., Medina, W.

Marco teórico

En el presente capítulo se muestran los factores asociados para determinar la correcta derivación del paciente utilizados en el prototipo del modelo predictivo-asistencial para la derivación hospitalaria o ambulatoria; de igual modo toda la parte conceptual relacionada a la enfermedad y a nivel computacional los modelos a emplear.

Factores asociados a la derivación hospitalaria o ambulatoria

COVID-19

Enfermedad causada por el virus del síndrome respiratorio agudo severo tipo-2 (SARS-CoV-2). Se detectó como pandemia en marzo del 2020 afectando en mayor medida a los adultos mayores y personas con patologías previas como diabetes, hipertensión, enfermedades cardiovasculares y cáncer. Los síntomas principales son tos, fiebre y dificultad respiratoria [5].

Sintomatología

De acuerdo con Carr, Boerner, & Moorman, [6] los coronavirus son un amplio grupo de virus que tienen la capacidad de causar daños tanto en seres humanos como en animales. Por lo que respecta al ser humano, se conoce que una gran cantidad de coronavirus producen afecciones en las vías respiratorias que van desde el resfriado común hasta patologías de mayor peligrosidad, como es el caso del síndrome respiratorio de Oriente Medio (MERS) y el síndrome respiratorio agudo severo (SRAS), como se citó en [7].

Cabe mencionar que da lugar a síntomas parecidos a los que son causados por la gripe, incluida fiebre, disnea, tos, fatiga y mialgia. Así mismo, se ha verificado la disminución del sentido del gusto y el olfato (sin que la mucosidad fuese la causa) [7]. Adicionalmente a lo previamente señalado, existen otros síntomas asociados con la COVID-19, los cuales se exponen en las siguientes líneas:

Temperatura: La fiebre mayor a 37° se ha detectado como uno de los principales síntomas de la COVID-19, así, por ejemplo, en la ciudad de Wuhan, la fiebre era el síntoma más frecuente, en tanto que en otras ciudades se encontró que 43,8% de los internados tenían fiebre en el instante de su entrada al hospital, aunque durante su permanencia en el hospital casi todos la desarrollaron [8].

Tos: la presencia de tos (asociada a la neumonía), la cual en la mayor cantidad de casos se inicia con tos seca, es seguida de tos productiva; en algunas ocasiones con disnea, sin datos de hipoxemia, se pueden auscultar crepitantes; otros casos no tienen signos o síntomas clínicos; pese a ello, la tomografía maquinizada evidencia afectaciones a los pulmones que necesitan una atención prioritaria [9].

Astenia: También conocida como fatiga, se conceptualiza como la percepción que tiene un sujeto acerca de la falta de fuerza o cansancio físico o mental, que, a su vez, tiene como efecto la reducción de la capacidad para ejecutar sus tareas rutinarias [10]. Dicho esto, haciendo referencia de forma específica al COVID-19, se puede expresar lo siguiente: "la astenia se presenta en diferentes grados, desde extrema e invalidante, a moderada o leve" [11].

Mialgia: La mialgia se define como un el dolor muscular que es atendido en consulta médica, un dolor muscular crónico. Estas patologías deben ser bien definidas para conocer cuál es la razón del dolor y por qué la mialgia es persistente y muy fuerte. Numerosos pacientes comenzaron a consultar en Wuhan, provincia de Hubei, China, a mediados de diciembre de 2019 por una infección respiratoria aguda caracterizada por mialgias y dificultad respiratoria [12].

Cefalea: Es un trastorno del sistema nervioso central caracterizada por la presencia de dolor localizado en la región craneofacial. De acuerdo con Ospina & Volcy [13] "En cuanto a la cefalea, se calcula que la frecuencia en pacientes con COVID-19 puede variar entre 6,5% y 34%". [13, p. 27] Adicionalmente, Ospina & Volcy señalan que "Pacientes con antecedentes de cefaleas primarias podrían presentar infección por COVID-19" [13].

Diarrea: Los síntomas más comunes son fiebre, tos seca y fatiga. Sin embargo, algunos pacientes con COVID-19 desarrollan vómito y diarrea durante el curso de su enfermedad. La diarrea en los pacientes con COVID-19 varía del 2 % al 33 %, y fue el síntoma predominante en el primer paciente diagnosticado en Estados Unidos con COVID-19 [14].

Pérdida de olfato: El mecanismo fisiopatológico por el cual la COVID-19 se asocia a las alteraciones del gusto y del olfato aún no son claras, pero existe evidencia, en el cual se ha informado que el cerebro expresa receptores de la enzima convertidora de angiotensina 2 (ACE 2), receptores dianas del SARS-CoV-2, que se encuentran sobre células gliales y neuronales, lo que los convierte en un objetivo potencial de COVID-19, pudiendo causar daño y muerte neuronal, recorriendo desde las neuronas periféricas a través de la lámina cribosa hasta el bulbo olfatorio [15].

Pérdida de apetito: En la misma línea de lo alimentario, sabemos que existe relación entre el gusto y el olfato, y que generalmente estas alteraciones se acompañan; pudiendo verse comprometida la nutrición del paciente al no percibir los sabores de los alimentos, llevando a una conducta restrictiva, es decir, disminución del apetito al no disfrutar del sabor de la comida, o por el otro lado, a una conducta que no sea favorable para su salud al sobre sazonar los alimentos, esto toma aún mayor peso en los pacientes con diabetes y/o hipertensión arterial [16].

Fatiga: El cansancio o agotamiento, en particular en la o el paciente que fue víctima de COVID-19 y que ya superó la enfermedad, puede llegar a progresar a una circunstancia más compleja que se denomina síndrome de fatiga crónica. Así mismo, según investigaciones recientes, la fatiga es el síntoma más marcado en un/a paciente que sobrevivió al nuevo coronavirus con 53.1%; le sigue la disnea (dificultad para respirar) con un 43.4%, el dolor articular con 27% y el dolor de pecho torácico con 21.7% [17].

Pérdida del conocimiento: En relación con las secuelas neuropsiquiátricas, en casos graves de COVID-19, la respuesta hiperinflamatoria sistémica podría causar un deterioro cognitivo a largo plazo, como, por ejemplo, deficiencias en la memoria, atención, velocidad de procesamiento y funcionamiento junto con pérdida neuronal difusa, lo que eventualmente, puede originar pérdida del conocimiento [18].

Dolor abdominal: Los pacientes además pueden presentar manifestaciones gastrointestinales cuya frecuencia varían según la población estudiada y la gravedad del cuadro. En Chile, según un informe del MINSAL al 11 de abril 2020, mostró que el 11% de los casos de SARS-CoV-2 confirmados presentaron algún síntoma gastrointestinal, aproximadamente un 7,3% presentó dolor abdominal [19].

Saturación de O2: Este es un síntoma del COVID-19 descubierto recientemente, la cual indica un decaimiento de los niveles de oxígeno en la sangre. En ese sentido, de acuerdo con la evidencia disponible, los oxímetros de pulso de uso no médico tendrían una eficacia comparable a la de los oxímetros de uso médico para descartar la presencia de hipoxemia en pacientes con COVID-19. El valor predictivo negativo para descartar pacientes con hipoxemia (déficit de oxígeno), definida como $SpO2 < 94\%$, es de 99% aunque su precisión disminuye de manera significativa para saturaciones por debajo de 94% [18].

Inteligencia artificial

La inteligencia artificial (IA) es una amplia rama de la informática que se ocupa de la construcción de máquinas inteligentes capaces de realizar tareas que normalmente requieren inteligencia huma-

na. La IA es una ciencia interdisciplinaria con múltiples enfoques, pero los avances en el aprendizaje automático y el aprendizaje profundo están creando un cambio de paradigma en prácticamente todos los sectores de la industria tecnológica. Se refiere a la simulación de la inteligencia humana en máquinas que están programadas para pensar como humanos e imitar sus acciones. El término también se puede aplicar a cualquier máquina que exhiba rasgos asociados con una mente humana, como el aprendizaje y la resolución de problemas.
La característica ideal de la inteligencia artificial es su capacidad para racionalizar y emprender acciones que tengan las mejores posibilidades de lograr un objetivo específico. Un subconjunto de la inteligencia artificial es el aprendizaje automático, que se refiere al concepto de que los programas informáticos pueden aprender y adaptarse automáticamente a nuevos datos sin la ayuda de humanos. Las técnicas de aprendizaje profundo permiten este aprendizaje automático mediante la absorción de grandes cantidades de datos no estructurados como texto, imágenes o video. [20]

Cuando la mayoría de la gente escucha el término inteligencia artificial, lo primero en lo que suele pensar es en los robots. Eso es porque las películas y novelas de gran presupuesto tejen historias sobre máquinas similares a las humanas que causan estragos en la Tierra. Pero nada podría estar más lejos de la verdad. La inteligencia artificial se basa en el principio de que la inteligencia humana se puede definir de manera que una máquina pueda imitarla fácilmente y ejecutar tareas, desde las más simples hasta las más complejas. Los objetivos incluyen el aprendizaje, el razonamiento y la percepción.

A medida que avanza la tecnología, los puntos de referencia anteriores que definían la inteligencia artificial se vuelven obsoletos. Por ejemplo, ya no se considera que las máquinas que calculan funciones básicas o reconocen texto a través del reconocimiento óptico de caracteres incorporen inteligencia artificial, ya que esta función ahora se da por sentada como una función informática inherente. La IA evoluciona continuamente para beneficiar a muchas industrias diferentes. Las máquinas están conectadas utilizando un enfoque multidisciplinario basado en matemáticas, informática, lingüística, psicología y más [21].

Machine Learning

Es una rama de la inteligencia artificial que se define como un proceso por medio del cual una computadora analiza datos para automatizar la construcción de diferentes modelos analíticos. Esta rama está basada en la idea de que los diversos sistemas son capaces de aprender datos identificando patrones y participar en la toma de decisiones con mínima intervención humana [22].

Se ha identificado que el aprendizaje automático y la inteligencia artificial son tecnologías prometedoras empleadas por varios proveedores de atención médica, ya que dan como resultado una mejor ampliación, una mayor potencia de procesamiento, confiables e incluso superan a los humanos en tareas específicas de atención médica [23].

Random Forest

Se establece como uno de los algoritmos de clasificación de imágenes más utilizados. Una de sus mayores ventajas es la aportación de estimación interna de exactitud que brinda a través de una validación cruzada. Este algoritmo hace uso de dos parámetros, siendo estos el número de árboles y el número de predictores que se utilizarán en la partición de cada árbol [24].

Random Forest es un algoritmo de clasificación, que permite procesar datos que tengan una varianza significativa; los datos obtenidos por las distintas fuentes de información son bastantes dispersos y en muchos casos inconsistentes y variables. Este algoritmo está especialmente diseñado para procesar este tipo de casos, es así que, su aplicación en la problemática del proyecto resulta fundamental para obtener predicciones eficientes y eficaces, dando así un nivel de confiabilidad alto a los doctores que utilicen la herramienta, obteniendo resultados positivos en la toma de decisiones para la derivación hospitalaria.

Naive Bayes

Es un algoritmo que utiliza el teorema de Bayes para clasificar objetos. Los clasificadores ingenuos de Bayes asumen una independencia fuerte o ingenua entre los atributos de los puntos de datos. Los usos populares incluyen filtros de spam, análisis de texto y diagnóstico médico. Los mismos se utilizan ampliamente para el aprendizaje automático porque son simples de implementar [25].

A pesar de su diseño ingenuo y supuestos aparentemente simplificados, los clasificadores de Bayes ingenuos han funcionado bastante bien en muchas situaciones complejas del mundo real. En 2004, un análisis del problema de clasificación bayesiana mostró que existen razones teóricas sólidas para la eficacia aparentemente inverosímil de los clasificadores Bayes ingenuos. Aun así, una comparación integral con otros algoritmos de clasificación en 2006 mostró que la clasificación de Bayes es superada por otros enfoques, como árboles impulsados o bosques aleatorios [25].

Algunas de las ventajas que Naive Bayes ofrece en contraste con Random Forest son:

- Simplicidad si la independencia condicional se mantiene convergiendo más rápido que modelos discriminatorios, necesitando menos datos para entrenamiento.
- Reducción de tiempos en creación del modelo.

La principal diferencia entre los algoritmos radica en los ajustes que se les tiene que dar a los modelos, siendo que, el modelo Naive Bayes es bajo y constante, en Random Forest es considerablemente grande y variable. Siendo Naive Bayes el algoritmo óptimo para la implementación del proyecto, debido que, puede adaptarse rápidamente a los cambios y nuevos datos.

Herramientas tecnológicas

Python

El software se puede programar de principio a fin utilizando Python como único idioma. Es una ventaja para los desarrolladores, ya que otros lenguajes de programación requieren la complementación con otros lenguajes antes de que el proyecto se concluya por completo. La independencia de Python en todas las plataformas ahorra tiempo y recursos para los desarrolladores, que de otro modo incurrirían en muchos recursos para completar un solo proyecto [26].

En el presente proyecto se utilizará Python por ser un lenguaje de programación flexible y adaptable. Este lenguaje de programación es de los lenguajes de mejor rendimiento al momento de implementar cualquier proyecto o prototipo que funcione con Machine Learning e Inteligencia Artificial.

Google Colab

Es un servicio gratuito en la nube alojado por Google para fomentar la investigación del aprendizaje automático y la inteligencia artificial, donde a menudo la barrera para el aprendizaje y el éxito es el requisito de un poder computacional. Google Colab permite a los desarrolladores usar y compartir el cuaderno entre sí sin tener que descargar, instalar o ejecutar otra cosa que no sea un navegador [27].

Google Colab brinda un servicio en la nube, el cual está basado en las Notebooks de Jupiter, gracias a esto, se lo ha utilizado para realizar los algoritmos planteados ya que no se necesita de instalación previa en los dispositivos portátiles o móviles.

STAT::FIT

Es un software dedicado a realizar ajustes de curvas y análisis estadístico de los datos de entrada y salida que serán utilizados para la simulación. Permite alcanzar cinco objetivos que ayudan a comprobar los resultados obtenidos, tales como: ajuste de curvas, determinación del número de réplicas para correr un determinado modelo de terminación, determinación del tamaño de muestra para conocer tiempos de proceso y transportación, graficación de datos de entrada y de distribuciones probabilísticas y difusión de pensamiento estadístico [28].

Esta herramienta es una de las opciones que brinda el programa de ProModel, el cual es útil para realizar la ulterior simulación de datos. STAT::FIT se lo utilizó con la finalidad de obtener las distribuciones probabilísticas por cada variable del Dataset que ayudará más adelante para realizar la Simulación de Montecarlo.

@RISK

El programa @RISK es una herramienta útil para el uso de Excel, que sirve para la creación de Simulación de Montecarlo. Se lo ha empleado para la obtención de una mayor cantidad de datos, un total de 1400 datos, para posteriormente añadirlos al dataset y empezar a realizar los modelos predictivos.

SPSS

El programa de SPSS ha sido útil para realizar estadísticas descriptivas, el cual arroja los siguientes resultados: el análisis de correlación Pearson, el chi-cuadrado y las tablas de contingencia. Este programa se utilizó para las encuestas realizadas a los estudiantes y a los expertos.

Hipótesis

Si se hace uso de un algoritmo de Machine Learning como es el de Random Forest entonces se podrá tomar una decisión óptima respecto a la derivación hospitalaria o ambulatoria por parte del experto para los pacientes infectados con COVID-19 en un hospital público de la ciudad de Guayaquil.

Si se hace uso de un algoritmo de Machine Learning como es el de Naive Bayes entonces se podrá tomar una decisión óptima respecto a la derivación hospitalaria o ambulatoria por parte del experto para los pacientes infectados con COVID-19 en un hospital público de la ciudad de Guayaquil.

Metodología

Metodología de la investigación

El presente trabajo se lo realizó por medio del tipo de investigación comprobación de hipótesis.

Según Espinoza [29], la comprobación de hipótesis se basa en contrastar dicha hipótesis de una realidad. Es decir, el investigador tiene que someter a prueba aquello que ha enunciado en su hipótesis, y para ello ha de establecer, mediante alguna técnica de contrastación si su hipótesis concuerda o no con los datos empíricos. En tal caso, solo se pueden dar dos posibilidades previsibles: o bien la hipótesis puede verse apoyada por datos empíricos y ha sido confirmada, o bien la hipótesis no corresponde con los datos empíricos y decimos entonces que ha sido desconformada o refutada por los datos empíricos.

Dicho, esto, la investigación se desarrolló con una comprobación de hipótesis, debido a que se buscó y recopiló información que permitió someter las hipótesis de estudio a prueba y determinar si se cumplen, o si las mismas son erróneas, a través de la implementación de técnicas e instrumentos de recolección de datos, tales como encuestas.

El objetivo de esta encuesta fue la recopilación de las variables más importantes y la correlación entre ellas. Se desarrolló vía online mediante la herramienta Google Forms, con un total de 7 preguntas a 30 encuestados. Los resultados obtenidos de acuerdo con la correlación de Pearson se aprecian en la Tabla 2.

Tabla 2. Pruebas de chi-cuadrado. Conocimientos procesos de derivación Vs. las demás variables

	Conocimientos procesos de derivación			
Características de la muestra	X^2	*g.l.*	*p*	*n*
Conocimientos uso de IA	20,451	9	0,015	30
Manejo de Big Data	17,484	12	0,132	30
Conocimiento de algoritmos	19,991	12	0,067	30
Toma de decisiones	23,810	9	0,005	30
Conocimiento de Naive Bayes	21,515	12	0,043	30
Conocimiento de Random Forest	21,800	12	0,040	30

Fuente: Fuentes, M., Medina, W.

Por lo tanto, para el estudio de las variables, se determina que dos de las seis categorías examinadas (Conocimientos uso de IA, Toma de decisiones, Conocimiento Naive Bayes y Conocimiento de Random Forest) tienen relación o están correlacionados con la variable principal "conocimientos de procesos de derivación", al encontrar que el valor de p (significancia asintótica bilateral) es menor a 0,05.

Análisis de independencia de las variables

Conocimiento procesos de derivación en relación con conocimientos sobre Random Forest.

H0 (Hipótesis nula): Significa que no hay asociación entre las variables conocimientos procesos de derivación y conocimientos sobre Random Forest.

Ha (Hipótesis alternativa): Significa que existe asociación entre las variables conocimientos procesos de derivación y conocimientos sobre Random Forest.

En este caso, para un nivel de significancia de $\alpha = 0.05$, se rechaza H0 a favor de Ha si p-valor < 0.05. De acuerdo con los resultados del análisis de chi-cuadrado obtenido y representado en la Tabla 2, el valor p (Significación asintótica bilateral) es de 0,040, lo que significa que es menor al valor α. Por lo tanto, es posible inferir y determinar que las variables están relacionadas entre sí. Es decir, la variable de conocimientos de procesos de derivación está relacionada y depende de los conocimientos sobre Random Forest, con un chi-cuadrado de 21,800 con 12 grados de libertad para la muestra de 30 personas estudiadas.

Conocimiento procesos de derivación en relación con conocimientos sobre Naive Bayes

H0 (Hipótesis nula): Significa que no hay asociación entre las variables conocimientos procesos de derivación y conocimientos sobre Naive Bayes.

Ha (Hipótesis alternativa): Significa que existe asociación entre las variables conocimientos procesos de derivación y conocimientos sobre Naive Bayes.

En este caso, para un nivel de significancia de $\alpha = 0.05$, se rechaza H0 a favor de Ha si p-valor < 0.05. De acuerdo con los resultados del análisis de chi-cuadrado obtenido y representado en la Tabla 2, el valor p (Significación asintótica bilateral) es de 0,043, lo que significa que es menor al valor α. Por lo tanto, es posible inferir y determinar que las variables están relacionadas entre sí. Es decir, la variable de conocimientos de procesos de derivación está relacionada y depende de los conocimientos sobre Naive Bayes, con un chi-cuadrado de 21,515 con 12 grados de libertad para la muestra de 30 personas estudiadas.

Metodología del desarrollo del prototipo

El proceso metodológico para emplearse será "Knowledge Discovery in Databases – KDD", el cual se utiliza para encontrar conocimiento en un conjunto de datos en bruto. Se desarrolla el modelo predictivo en 6 fases:

Fase 1. Importación y muestreo de datos: Los datos obtenidos para la muestra fueron extraídos de un hospital público de la ciudad de Guayaquil; entidad que facilitó la información de sus instalaciones sobre los pacientes diagnosticados con COVID-19 en el periodo correspondiente desde marzo 2020 hasta enero 2021. La muestra se conforma por un total de 20 pacientes cuya derivación fue la siguiente: 10 ambulatorios, es decir su recuperación no requiere atención especializada debido a que no presentan síntomas de gravedad y 10 hospitalizados.

Fase 2. Calidad de datos: La calidad de los datos determina la confiabilidad y el correcto performance de los algoritmos que los implementen, es así como, una vez formado la base de datos, se procedió a hacer el tratamiento correspondiente en 2 fases: identificación y preprocesamiento de los datos; procesos que garantizan resultados precisos y confiables, los cuales se explican de manera detallada en el siguiente apartado:

1. Los registros de la base de datos obtenida se revisaron de manera manual, constatando que había inconsistencias en algunos de ellos, las anomalías encontradas fueron las siguientes:

a. *Campos vacíos:* En el dataset hay campos importantes tales como: "Temperatura y Saturación de oxígeno", campos que son variables de acuerdo con la sintomatología del paciente por lo tanto dependiendo si el infectado presentaba esta sintomatología se indicaba el valor equivalente a ese síntoma, caso contrario se presentaba vacío.

2. Identificadas las inconsistencias en el dataset el siguiente paso fue tratarla de la manera más a adecuada, al tener una muestra que no contenía un gran volumen de datos, se procedió a tratarlos de manera manual de la siguiente manera:

a. *Campos vacíos:* Los campos que presentaban la inconsistencia son datos cuantitativos, para solucionar este error se procedió a completar los campos con información adecuada acorde a su tipo de derivación y en base a las demás sintomatologías. Este procedimiento fue consultado con el experto del área de salud para su confiable y certera corrección.

Fase 3. Transformación: Esta fase es representada como una de las más importantes debido a que se dedica a la selección de los datos más significativos para el posterior entrenamiento de los algoritmos establecidos. Todo este proceso de transformación tiene un fin, el cual radica en la obtención de las variables que serán utilizadas en la fase de modelización por medio de la minería de datos, de tal forma que se amenoran la cantidad de datos no tan significativos para el entrenamiento.

La base de datos real inicial consistía en los siguientes atributos: "N° paciente, edad, dificultad respiratoria, saturación < 94, presión arterial, pérdida de conocimiento, cefalea (dolor de cabeza), dolor abdominal, mialgia (dolor muscular), odinofagia (dolor de garganta), tos, ronquera, temperatura, diarrea, fatiga, pérdida de olfato, pérdida de apetito, comorbilidad, derivación". Para seleccionar los atributos/variables más significativas se continuó con los siguientes pasos:

1. Realizando una exhaustiva investigación en fuentes confiables como lo son artículos científicos, libros, revistas científicas (ScienceDirect, Redalyc, Scielo, Elsevier) y del criterio del experto, se consiguió descartar aquellas variables/atributos menos significativos en la toma de decisiones para la derivación y dejar únicamente las variables precisas para dicho proceso.

2. Paso siguiente, se creó una nueva base de datos, la cual consistía solo de aquellas variables con mayor relevancia para el proceso de derivación hospitalaria y ambulatoria. Dejando la modificación de la nueva base de datos con las siguientes variables: "Dificultad Respiratoria, Saturación < 94, Dolor Abdominal, Mialgia, Tos, Temperatura, Pérdida de Olfato, Pérdida de Apetito, Derivación"

3. Por último, ya contando con una base de datos óptima se la exportó a un documento con extensión .xlsx (de Excel) y otra en .csv para su utilización al momento de realizar el algoritmo en Python.

Simulación de datos

Dada por terminada la fase de la selección de las variables, se concluyó que el volumen de datos, es decir, la cantidad de pacientes derivados de manera hospitalaria y ambulatoria en la base de datos real era insuficiente. Para eso, se realizó una simulación de 1400 datos por medio de los datos reales, esto significa que

se duplican los datos y se conseguirá un mejor entrenamiento para los algoritmos elegidos.

Herramientas aplicadas para la simulación de datos: Se utilizaron dos herramientas para realizar la simulación de datos, las cuales son: Microsoft Excel del programa informático Microsoft Office 365 ProPlus, éste a su vez se vincula con la herramienta del campo de desarrollo llamada Visual Basic, cuya finalidad es generar la cantidad de 1400 datos aleatorios. Esta simulación se consigue siguiendo el método de simulación Montecarlo y la herramienta @RISK 8.1 que permitirá realizar la simulación con Excel. Cabe destacar que no arroja el resultado final, derivación.

Para utilizar el método Montecarlo se debe conocer sobre las distribuciones de cada variable de la base de datos, esto se logra utilizando el software ProModel, específicamente haciendo uso de la herramienta tecnológica STAT::FIT. Al tenerse una base de datos real de 20 pacientes, esta herramienta permite la entrada de datos entre la cantidad 10 como mínimo y 50 como máximo. Dado este caso, todos los datos de la base de datos real son ingresados en la herramienta por variable, arrojando la siguiente información: Data, distribuciones, su rango y aceptación, estadística descriptiva, y el gráfico de distribución.

Simulación Montecarlo: La simulación Montecarlo se define como una herramienta estadística que permite la modelación de resultados de acuerdo con el comportamiento histórico de los datos y de su probabilidad de ocurrencia [30].

El método de simulación Montecarlo para emplearlo correctamente y que arroje datos precisos acorde a la derivación (dado que este no simula el resultado de la variable dependiente, derivación), se ha dividido los 20 datos en dos categorías: pacientes con derivación hospitalaria y pacientes con derivación ambulatoria. Consiguiendo dos distribuciones de cada variable por cada categoría y así, obtener la simulación de los 1400 datos más certera.

Fase 4. Modelización: Los algoritmos de Machine Learning (Aprendizaje Supervisado), específicamente los de aprendizaje supervisado, en este caso Random Forest y Naive Bayes, necesitan previamente datos etiquetados para ser entrenados. Para lograr esto se partió de la distribución de los datos en dos grupos: datos para el entrenamiento y datos para la prueba a partir de la simulación de datos ya establecida. A continuación, se muestra la distribución en la Tabla 3.

Tabla 3 Distribución de los datos para el entrenamiento de los algoritmos

Datos	Absoluto	Relativo
Datos para entrenamiento	1260	90%
Datos para prueba	140	10%
Total	1400	100%

Fuente: Fuentes, M., Medina, W.

Posteriormente, se debe hacer la elección entre el algoritmo de Random Forest y Naive Bayes de acuerdo con el porcentaje de precisión más alto para la respectiva derivación hospitalaria o ambulatoria de un paciente con COVID-19, según el dataset previamente proporcionado. Se detallan a continuación porciones de código comentadas para la compresión de las funciones usadas para la predicción.

Elección del algoritmo: Según el meta-análisis empleado en el capítulo II, se llegó a la conclusión de que los algoritmos más óptimos para la predicción son Random Forest y Naive Bayes por diversas características. Teniendo claro los algoritmos a usar, se procedió a comparar el resultado del porcentaje de precisión de cada uno, arrojando lo siguiente (ver Algoritmo I y II):

Algoritmo Random Forest

```
#Uso de accuracy_score para obtener la precisión del algoritmo Random Forest
accuracy_score (y_test, y_rf_prds)

0,9357142857142857
```

Algoritmo Naive Bayes

```
#Uso de accuracy_score para obtener la precisión del algoritmo Random Forest
accuracy_score (y_test, y_rf_prds)

0,95
```

Fase 5. Evaluación

Análisis de los resultados

Ya con los porcentajes de precisión arrojados por cada uno algoritmos de aprendizaje supervisado se obtiene la Tabla 4

Tabla 4 Descripción del porcentaje de precisión de cada algoritmo

Algoritmo	Precisión obtenida	% de la precisión
Random Forest	0,935	93,5%
Naive Bayes	0,95	95%

Fuente: Fuentes, M., Medina, W.

Se refleja que ambos algoritmos proporcionan un alto nivel de precisión, siendo éstos factibles para el prototipo, aunque cuentan con una diferencia de milésimas entre los algoritmos analizados. Random Forest cuenta con una precisión de 0,935, es decir, el 93,5% de precisión, mientras que Naive Bayes cuenta con una precisión de 0,95, es decir, el 95% de precisión.

Interpretación: Se llega a la conclusión que ambos algoritmos arrojan un óptimo resultado de precisión, acercándose bastante a un porcentaje total, claro está que siempre hay un margen de error. Por tal motivo, el algoritmo Naive Bayes, debido a su mayor puntuación, es considerado como el algoritmo más factible para el modelo predictivo asistencial para las personas infectadas con COVID-19 para la derivación hospitalaria o ambulatoria.

Fase 6. Implementación: En los alcances del proyecto, la implementación del algoritmo consistirá en su creación y entrenamiento para que pueda ser utilizado en diversas herramientas que lo requieran, por lo cual se explicará el proceso de implementación.

Figura 1 Diagrama de flujo para la implementación de los algoritmos

Fuente: Fuentes, M., Medina, W.

Arquitectura del diseño

- Dificultad respiratoria (0: no, 1: sí)
- Saturación de oxígeno: Porcentaje de oxigenación en la sangre.
- Dolor abdominal: (0: no, 1: sí)
- Mialgia: (0: no, 1: sí)
- Tos: (0: no, 1: sí)
- Temperatura: Medida en C°.
- Pérdida de olfato: (0: no, 1: sí)
- Pérdida de apetito: (0: no, 1: sí)

Entrenamiento

El entrenamiento consiste en la clasificación de datos etiquetados de manera supervisada, elaborando un modelo acorde al grupo de datos entrenados y etiquetas de clase, consiguiendo clasificar datos nuevos. En el siguiente apartado se detallan los requerimientos necesarios para crear el entorno de trabajo:

Base de datos: Se utilizará alrededor de 1400 registros para el dataset almacenados en un archivo .csv.

Preparación del ambiente de trabajo: Se lo realiza a través de Google Colab, el cual permite trabajar de manera online sin instalación de IDE, ya que contiene una herramienta de interpretación de Python, que facilita el desarrollo de soluciones en este lenguaje.

Creación del ambiente de trabajo: Para que el ambiente de trabajo se pueda utilizar, es necesario acceder al sitio web: https://colab.research.google.com/, el cual proporciona el paquete de herramientas necesarias para la creación del algoritmo.

A continuación, se presentan los fragmentos del código en Python:

```
## Importación de las librerías ##
import pandas as pd
import numpy as np
import seaborn as sns
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import train_test_split
from matplotlib import pyplot as plt
%matplotlib inline
from sklearn.metrics import confusion_matrix
from sklearn.feature_selection import SelectKBest
## SE ASIGNA UNA VARIABLE LOS DATOS DEL DATASET ##
df = pd.read_csv('dataset_derivacion v2.csv', sep=';', encoding='UTF-8')
## Se realiza la selección de los predictores más importantes ##
X = df.drop(["derivacion"], axis=1)
y = df["derivacion"]
best = SelectKBest(k=8)
X_new = best.fit_transform(X, y)
X_new.shape
selected = best.get_support(indices=True)
used_features = X.columns[selected]
X.columns[selected]
## División de los datos. 90% para entrenamiento y 10% para prueba ##
X_train, X_test = train_test_split(df, test_size = 0.10, random_state=10)
y_train = X_train["derivacion"]
y_test = X_test["derivacion"]
## Creación del modelo predictivo Random Forest ##
rf = RandomForestClassifier(n_estimators=25)
rf = rf.fit(X_train[used_features], y_train)
y_prds = rf.predict(X_test[used_features])
## Precisión del modelo ##
from sklearn.metrics import accuracy_score
accuracy_score(y_test, y_prds)
```

Pruebas

Del conjunto de datos obtenidos se precisó que el 10% de los datos se utilizarían para las pruebas del algoritmo, mientras que el porcentaje restante para el entrenamiento.

Testeo del algoritmo "Random Forest": En primer lugar, se muestran la matriz de confusión obtenida por las pruebas realizadas por el algoritmo, posteriormente, la matriz de confusión se ve detallada en una tabla. Los datos se trataron a través del conjunto de herramientas proporcionadas por "Google Colab".

Tabla 5 Matriz de confusión del algoritmo Random Forest.

		Positivo	**Negativo**
Entrenamiento	**Positivo**	VP = 46	FN = 2
	Negativo	FP = 7	VN = 85

Fuente: Fuentes, M., Medina, W.

Como se aprecia en la Tabla 5, la precisión obtenida por el algoritmo Random Forest es de 0,9357142857142857, es así como en la matriz de confusión se tiene los siguientes valores: 46 datos verdaderos positivos, 85 datos verdaderos negativos, 7 datos falsos positivos y 2 datos falsos negativos. Para concluir, se puede afirmar que la precisión llega a un rango del 94%.

Testeo del algoritmo "Naive Bayes": El testeo del algoritmo de Naive Bayes consistirá en obtener la precisión y matriz de confusión del modelo al interactuar con los datos entregados, en consecuencia, en el siguiente apartado se mostrarán la figura y la tabla con los valores obtenidos.

Tabla 6 Matriz de confusión del algoritmo Naive Bayes

		Positivo	Negativo
Entrenamiento	**Positivo**	VP = 46	FN = 2
	Negativo	FP = 5	VN = 87

Fuente: Fuentes, M., Medina, W.

Como se aprecia en la Tabla 6, la precisión obtenida por el algoritmo Random Forest es de 0,95, es así como en la matriz de confusión se tiene los siguientes valores: 46 datos verdaderos positivos, 87 datos verdaderos negativos, 5 datos falsos positivos y 2 datos falsos negativos. Para concluir, se puede afirmar que la precisión llega a un rango del 95%.

Resultados

Los algoritmos de aprendizaje supervisado Random Forest y Naive Bayes fueron aplicados mediante la herramienta tecnológica Python, siendo esta muy utilizada para proyectos que requieran Machine Learning. Se codificó en el entorno de los servicios cloud de Google Colab, con la finalidad a la ayuda de la toma de decisiones para los procesos de derivación hospitalaria o ambulatoria de pacientes infectados por COVID-19.

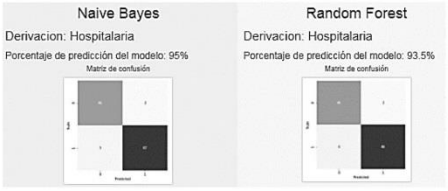
Siguiendo el proceso requerido para la construcción del modelo predictivo, se obtuvo un exitoso porcentaje de precisión del 93,5% para el algoritmo Random Forest. El segundo algoritmo Naive Bayes arrojó un resultado de precisión del 95%, como se muestra en la Figura 2; concluyendo que, en vista que ambos porcentajes son muy buenos, cabe destacar que Naive Bayes es el más eficiente para efectuar un mejor resultado para la derivación de los pacientes; predominando su diferencia en los tiempos de procesamiento, éste requiere menor tiempo.

1. Mediante investigaciones en fuentes científicas se definieron aquellas variables significativas para la respectiva derivación hospitalaria o ambulatoria. Además, la amplia información recolectada dio paso al conocimiento de la estructura del diseño de los algoritmos supervisados de Machine Learning utilizados.

2. Se depuró la base de datos facilitada por un hospital público de Guayaquil, dejándola más limpia y con los campos llenados correctamente. Siguiendo las 6 fases de la metodología KDD se mejoró la calidad de los datos empleados.

3. Ya obteniendo un definido dataset, se evaluaron las variables por medio de los algoritmos supervisados de Machine Learning, definiendo una función en Python que denote aquellas variables más importantes para el resultado de las derivaciones y no presenten discrepancia con los algoritmos seleccionados. Entre ellos se descartaron 7 variables, dejando como relevantes 8 variables.

Figura 2. Resultado de los algoritmos Random Forest y Naive Bayes



Fuente: Fuentes, M., Medina, W.

Conclusiones

1. La recopilación de información bibliográfica de las variables relacionadas a la derivación hospitalaria y ambulatoria por síntomas en pacientes con COVID-19, fue en base a la numerosa información alojada en plataformas confiables que contienen artículos científicos, revistas y libros, tales como: Science Direct, Taylor & Francis, Springer, IEEE, Elsevier, entre otros. Preciso que en dichas plataformas se encontró información válida acorde al tema investigativo, como lo son los factores asociados para la derivación hospitalaria, es decir, los síntomas más relevantes, entre ellos resaltando la saturación de oxígeno en la sangre, temperatura, edad y otros. Cada una fue seleccionada según los protocolos generales del área epidemiológica. Se extrajo de un total de 30 referencias bibliográficas.

2. Se consiguió una base de datos perteneciente a un hospital público de la ciudad de Guayaquil, conformado por el historial clínico de pacientes infectados por COVID-19 y su respectiva derivación sea esta hospitalaria o ambulatoria. Para su respectiva depuración, para la posterior creación del dataset, se hizo uso de la metodología Knowledge Discovery in Databases, KDD, el cual se conforma de 6 fases, arrojando como resultado una base de datos más limpia conteniendo únicamente variables significativas para la derivación del paciente, las cuales son: dificultad respiratoria, saturación, dolor abdominal, mialgia, tos, temperatura, pérdida de olfato, pérdida de apetito y derivación. Además, se logró solucionar todo aquel campo vacío que impidiese más adelante el correcto desempeño del algoritmo. Concluyendo que, esta metodología es excelente para el procesamiento de la información.

3. Se evaluaron las variables relacionadas a la derivación hospitalaria o ambulatoria para mejorar la toma de decisiones a partir de los algoritmos de aprendizaje supervisados Random Forest y Naive Bayes. Haciendo uso de la librería sklearn del lenguaje de programación Python para el entrenamiento del algoritmo, la herramienta STAT::FIT para las distribuciones estadísticas, y basándose en la sintomatología del paciente los algoritmos arrojaron un gran porcentaje de precisión (93.5% Random Forest y 95% Naive Bayes) para la predicción de la derivación para pacientes con COVID-19. Entre ambos se concluyó que el mejor predictor es el algoritmo de Naive Bayes.

4. De acuerdo al análisis realizado en la hipótesis estadística de las preguntas "Conocimiento procesos de derivación y conocimientos sobre Random Forest" (Capítulo III) se pudo dar respuesta a la primera hipótesis científica " Si se hace uso de un algoritmo de Machine Learning como es el de Random Forest entonces se podrá tomar una decisión óptima respecto a la derivación hospitalaria o ambulatoria por parte del experto para los pacientes infectados con COVID-19 en un hospital público de la ciudad de Guayaquil" (Capítulo II) concluyendo que existe una relación entre ambas, mediante el análisis de correlación de Pearson, se obtiene un valor de p =0,040, es decir, menor a 0,05 haciendo que se cumpla la hipótesis.

5. De acuerdo al análisis realizado en la hipótesis estadística de las preguntas "Conocimiento procesos de derivación y conocimientos sobre Naive Bayes" (Capítulo III) se pudo dar respuesta a la primera hipótesis científica "Si se hace uso de un algoritmo de Machine Learning como es el de Naive Bayes entonces se podrá tomar una decisión óptima respecto a la derivación hospitalaria o ambulatoria por parte del experto para los pacientes infectados con COVID-19 en un hospital público de la ciudad de Guayaquil" (Capítulo II) concluyendo que existe una relación

entre ambas, mediante el análisis de correlación de Pearson, se obtiene un valor de p =0,043, es decir, menor a 0,05 haciendo que se cumpla la hipótesis.

Recomendaciones

A continuación, se describen todas aquellas sugerencias a tomar en consideración para un correcto desempeño del proyecto:

- Para la obtención de la base de datos real, y más si es de tratarse de una entidad pública de salud, se debe realizar con antelación una solicitud al Ministerio de Salud Pública para obtener la información confidencial.
- Con el juicio de expertos debe asegurarse que las variables utilizadas para el dataset sean relevantes para el modelo, realizando una codificación con el método "features_importances_" para conocer las variables con un alto valor significativo en el modelo, y perfeccionar la interfaz gráfica del sitio web.
- Comprobar si existe sobreajuste en los algoritmos en concordancia con el total de datos que posee el dataset.
- Verificar si lo datos están correctamente escritos para la posterior conversión de cualitativos a cuantitativos.
- Realizar el apartado "derivación hospitalaria" como un sitio web responsive para ser utilizado en dispositivos móviles, debido que se encuentra desarrollado para escritorio.

Agradecimientos

Agradezco a Dios por darme las fuerzas de seguir adelante cada día y por acompañarme siempre en las buenas y en las malas. A mis padres, por haberme dado una buena educación que han sido la base para mi futuro, todo lo que he logrado es gracias a ellos. Y muchas gracias al Ing. Lorenzo Cevallos, porque nos ha brindado su ayuda en cada duda que hemos tenido en este proceso de titulación. Gracias a todos por eso y por muchos más. *Fuentes, M., Medina, W.*

Referencias Bibliográficas

[1] W. Liang, R. Chen y W. Guan, «Cancer patients in SARS-CoV-2 infection: a nationwide analysis in China.,» The lancet oncology, 21(3), pp. 335-337, 2020.

[2] E. Blanco and G. Sanchez, "Atención primaria y residencias de ancianos: a propósito de la COVID-19.," Semergen, 46, 26., 2020.

[3] T. Bardi, A. Candela, R. De pablo , R. Martinez y D. Pestaña, «Respuesta rápida a COVID-19, estrategias de escalada y desescalada para ajustar la capacidad suplementaria de camas de UVI a una epidemia de gran magnitud.,» Revista Española de Anestesiología y Reanimación, 68(1),, pp. 21-27, 2020.

[4] J. Ávila, M. Mayer y V. Quesada, «La inteligencia artificial y sus aplicaciones en medicina I: introducción antecedentes a la IA y robótica,» 2020.

[5] F. J. Díaz-Castrillón y A. I. Toro-Montoya, «SARS-CoV-2/COVID-19: el virus, la enfermedad y la pandemia,» p. 23, 2020.

[6] D. Carr, K. Boerner y S. Moorman, «Bereavement in the time of coronavirus: Unprecedented challenges demand novel interventions.,» Journal of Aging & Social Policy, 32(4-5), , pp. 425-431. , 2020.

[7] M. R. Pérez Abreu , J. J. Gómez Tejeda y R. A. Dieguez Guach , «Características clínico-epidemiológicas de la COVID-19,» Revista Habanera de ciencias médicas, vol. 19, n° 2, pp. 1-15, 2020.

[8] CCAES, «Enfermedad por coronavirus, Covid-19,» Ministerio de sanidad, Madrid, España, 2020.

[9] M. d. V. T. d. Mier, N. López-Perea y J. M. Calles, «SITUACIÓN DE LA TOS FERINA EN ESPAÑA, 1998-2016 ANÁLISIS PRELIMINAR DEL IMPACTO DE LA VACUNACIÓN DE TOS FERINA EN EMBARAZADAS,» 2018. [En línea]. Available: http://revista.isciii.es/index.php/bes/article/view/1060/1303.

[10] M. J. Guijarro Sánchez, L. Royuela García , J. Guillén González y I. Aranburu Aizpiri, «Cuando no todo es "vejez" en el anciano. Astenia y apatía, ¿por dónde empezar?,» Revista de Medicina de Familia y Atención Primaria (fml), vol. 24, n° 2, p. 4, 2019.

[11] SEDISA, «Documento para la atención integral del paciente Post-Covid,» Fundación AstraZeneca, Madrid, España, 2020.

[12] R. Poblete , F. Peñafiel , N. Sabatini, A. Vite, A. Ceriani, S. Schaffeld, L. Letelier, J. Gran scheuch y R. Rabagliati , «Infección respiratoria aguda por coronavirus Sars-CoV-2 en personal de salud. Implementación de un programa de detección precoz y seguimiento de casos en un hospital universitario,» Rev Med Chile, n° 148, pp. 724-733, 2020.

[13] C. Ospina y M. Volcy, «Enfoque del paciente con cefalea en tiempos de covid-19,» Acta Neurológica Colombiana, vol. 36, n° 2, pp. 27-38, 2020.

[14] V. Parra, C. Flórez, F. García y C. Romero, «Síntomas gastrointestinales en la enfermedad por covid-19 y sus implicaciones en la enfermedad inflamatoria intestinal,» Rev Colomb Gastroenterol., vol. 35, n° 1, pp. 45-55, 2020.

[15] A. Huamán y J. Aparcana , «La anosmia como síntoma temprano en pacientes con covid-19,» Facultad de Medicina Humana URP, vol. 20, n° 3, pp. 532-533, 2020.

[16] V. Sepúlveda , S. Waissbluth y C. González, «Anosmia y enfermedad por Coronavirus 2019 (COVID-19): ¿Qué debemos saber?,» Rev. Otorrinolaringol. Cir. Cabeza Cuello, n° 80, pp. 247-258, 2020.

[17] Visión CEVECE, «Secuelas por covid-19,» Secretaría de Salud, Estado de México, 2020.

[18] OPS, «Aspectos técnicos y regulatorios sobre el uso de oxímetros de pulso en el monitoreo de pacientes con COVID-19,» 2020.

[19] L. Díaz y A. Espino , «Manifestaciones gastrointestinales de pacientes infectados con el nuevo Coronavirus SARS-CoV-2,» Gastroenterol. latinoam, vol. 31, n° 1, pp. 35-38, 2020.

[20] L. Rouhiainen, «Inteligencia artificial,» Alienta Editorial, vol. 12, p. 12, 2018.

[21] M. Boden, «Inteligencia artificial,» Turner, vol. 12, p. 12, 2017.

[22] E. Southgate, K. Blackmore, S. Pieschl, S. Grimes, J. McGuire y K. Smithers, «Artificial Intelligence and Emerging Technologies in Schools,» p. 155, 2019.

[23] M. Serra, «COVID-19. De la patogenia a la elevada mortalidad en el adulto mayor y con comorbilidades,» Revista

Habanera de Ciencias Médicas, p. 12, 2020.

[24] F. Cánovas-García, F. Alonso-Sarría y F. Gomariz-Castillo, «MODIFICACIÓN DEL ALGORITMO RANDOM FOREST PARA SU EMPLEO EN CLASIFICACIÓN DE IMÁGENES DE TELEDETECCIÓN,» p. 10, 2016.

[25] F. Hutter, «Automated machine learning: methods, systems, challenges,» Springer Nature, p. 12, 2019.

[26] N. Acosta, «Nuevas tecnologías como factor de cambio ante los retos de la inteligencia artificial y la sociedad del conocimiento,» Revista ESPACIOS, vol. 1, p. 41, 2018.

[27] E. Bisong, «Google colaboratory. In Building Machine Learning and Deep Learning Models on Google Cloud Platform,» Apress, Berkeley, CA., pp. 59-64, 2019.

[28] Geer Mountain Software Corp, «Stat::Fit®,» 2016. [En línea]. Available: https://www.geerms.com/files/114225421.pdf.

[29] E. Espinoza, «La hipótesis en la investigación,» MENDIVE Vol. 16 No. 1, pp. 122-139, 2018.

[30] E. J. S. Jiménez y W. A. A. Castro, «Aplicación de la simulación Monte Carlo en la proyección del estado de resultados. Un estudio de caso,» 2018. [En línea]. Available: http://www.revistaespacios.com/a18v39n51/a18v39n51p11.pdf.

[31] J. Díaz, M. Díaz-de-Neira, A. Jaraboc, P. Roig y P. Román, «Estudio de derivaciones de Atención Primaria a centros de Salud Mental en pacientes adultos en la Comunidad de Madrid,» p. 7, 2017.

Elaboración: Fuentes Melina y Medina Wilmer.
Fuente: Propia.