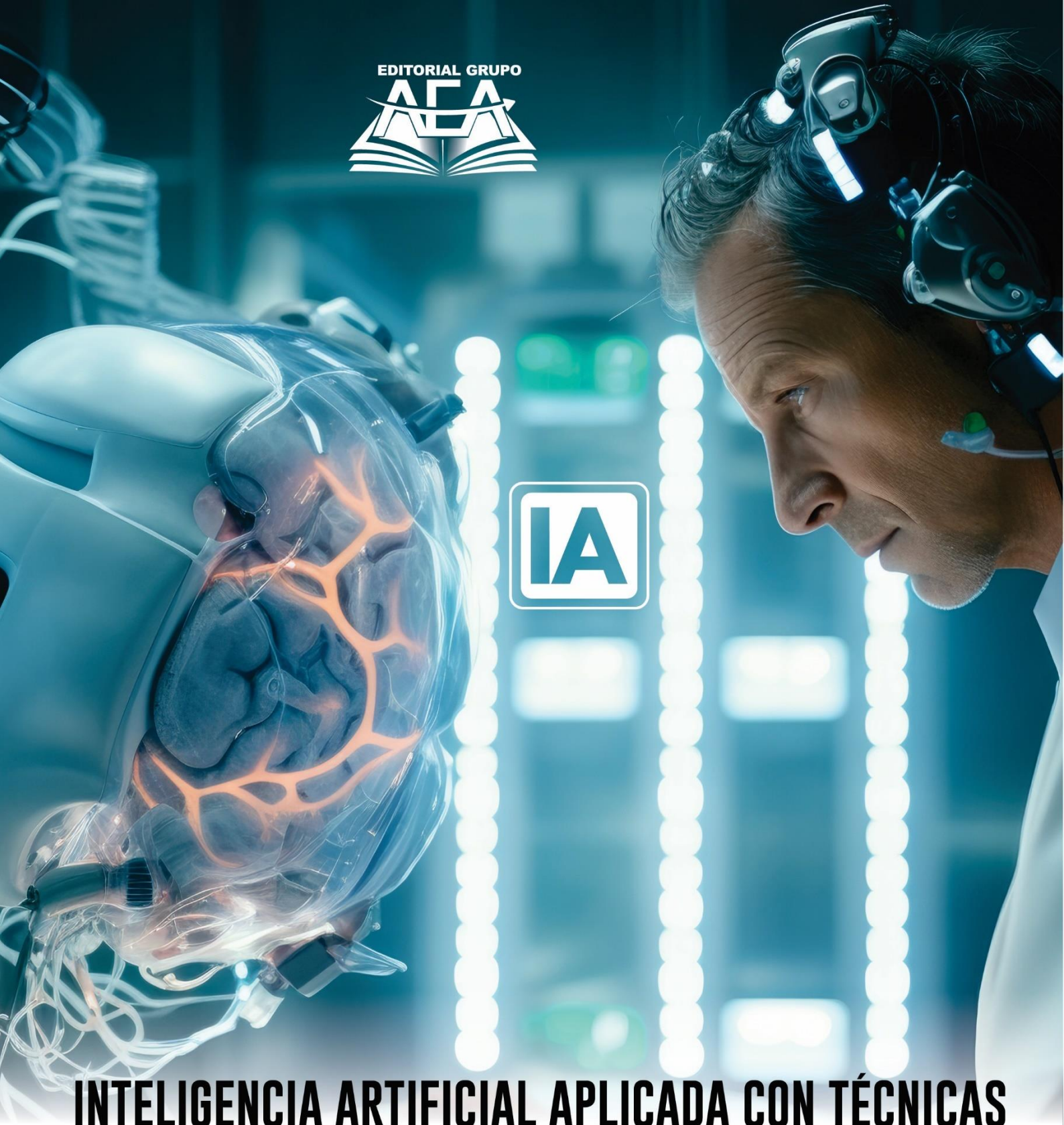


INTELIGENCIA ARTIFICIAL APLICADA CON TÉCNICAS DE PROCESAMIENTO DE LENGUAJE NATURAL Y MACHINE LEARNING EN EL CAMPO DE LA SALUD.

Varela-Tapia, Eleanor Alexandra
Acosta-Guzmán, Ivan Leonel
Fajardo-Romero, Inés Janellys
Oviedo-Peñafiel, Jorge Alberto

www.editorialgrupo-aea.com

Inteligencia Artificial Aplicada con técnicas de Procesamiento de Lenguaje Natural y Machine Learning en el campo de la salud

Autor/es:

Varela-Tapia, Eleanor Alexandra
Universidad de Guayaquil

Acosta-Guzmán, Ivan Leonel
Universidad de Guayaquil

Fajardo-Romero, Inés Janellys
Universidad de Guayaquil

Oviedo-Peñafiel, Jorge Alberto
Universidad de Guayaquil

Datos de Catalogación Bibliográfica

Varela-Tapia, E. A.
Acosta-Guzmán, I. L.
Fajardo-Romero, I. J.
Oviedo-Peñafiel, J. A.

Inteligencia Artificial Aplicada con técnicas de Procesamiento de Lenguaje Natural y Machine Learning en el campo de la salud.

Editorial Grupo AEA, Ecuador, 2024
ISBN: 978-9942-651-38-9
Formato: 210 cm X 270 cm 185 págs.

Publicado por Editorial Grupo AEA
Ecuador, Santo Domingo, Vía Quinindé, Urb. Portón del Río.
Contacto: +593 983652447; +593 985244607
Email: info@editorialgrupo-aea.com
https://www.editorialgrupo-aea.com/

Director General:	*Prof. César Casanova Villalba.*
Editor en Jefe:	*Prof. Giovanni Herrera Enríquez*
Editora Académica:	*Prof. Maybelline Jaqueline Herrera Sánchez*
Supervisor de Producción:	*Prof. José Luis Vera*
Diseño:	*Tnlgo. Oscar J. Ramírez P.*
Consejo Editorial	*Editorial Grupo AEA*

Primera Edición, 2024

Cámara Ecuatoriana del Libro con registro editorial No 708

Disponible para su descarga gratuita en https://www.editorialgrupo-aea.com/

Título del libro:
Inteligencia Artificial Aplicada con técnicas de Procesamiento de Lenguaje Natural y Machine Learning en el campo de la salud.

Libro Digital, Primera Edición, 2024
Editado, Diseñado, Diagramado y Publicado por Comité Editorial del Grupo AEA, Santo Domingo de los Tsáchilas, Ecuador, 2024

ISBN: 978-9942-651-38-9

https://doi.org/10.55813/egaea.l.83

Como citar (APA 7ma Edición):

Varela-Tapia, E. A., Acosta-Guzmán, I. L., Fajardo-Romero, I. J., & Oviedo-Peñafiel, J. A. (2024). *Inteligencia Artificial Aplicada con técnicas de Procesamiento de Lenguaje Natural y Machine Learning en el campo de la salud*. Editorial Grupo AEA. https://doi.org/10.55813/egaea.l.83

Cada uno de los textos de Editorial Grupo AEA han sido sometido a un proceso de evaluación por pares doble ciego externos (double-blindpaperreview) con base en la normativa del editorial.

Revisores:

Ing. García Peña Víctor René, PhD.	Universidad Laica Eloy Alfaro de Manabí – Ecuador
Ing. Ramos Secaira Francisco Marcelo, Mgs.	Pontificia Universidad Católica del Ecuador – Ecuador

Los libros publicados por **"Editorial Grupo AEA"** cuentan con varias indexaciones y repositorios internacionales lo que respalda la calidad de las obras. Lo puede revisar en los siguientes apartados:

Editorial Grupo AEA

http://www.editorialgrupo-aea.com

Editorial Grupo AeA

editorialgrupoaea

Editorial Grupo AEA

Aviso Legal:

La información presentada, así como el contenido, fotografías, gráficos, cuadros, tablas y referencias de este manuscrito es de exclusiva responsabilidad del/los autor/es y no necesariamente reflejan el pensamiento de la Editorial Grupo AEA.

RESEÑA DE AUTORES

AUTORES

Varela Tapia Eleanor Alexandra

Universidad de Guayaquil

eleanor.varelat@ug.edu.ec

https://orcid.org/0000-0002-5357-4046

Docente Titular de la Universidad de Guayaquil (UG). Investigadora Agregado 2 acreditada por SENESCYT-ECUADOR, líneas de investigación: Inteligencia Artificial, Natural Language Processing (NLP), Ciencia de Datos, Big Data e Ingeniería de Software. Directora e investigadora principal en proyectos de investigación (2018-2024) en la Carrera de Software y la Carrera de Teleinformática de la UG. Directora de proyectos de tesis de pregrado y posgrado (2010-2024) en la Carrera de Software, Carrera de Sistemas Computacionales y en la Maestría de Ingeniería de Software de la UG. Desempeño en cargos como Gestora de Investigación y Resultados Científicos (2019-2022), Gestora de Bienestar Estudiantil (2017-2018) y Gestora de Prácticas Pre-profesionales (2016) en la UG. Maestrante en Big Data y Ciencia de Datos (Universidad Internacional de Valencia, VIU-España). Magister en Docencia y Gerencia en Educación Superior (UG, Ecuador). Magister en Sistemas de Información Gerencial (Escuela Superior Politécnica del Litoral ESPOL, Ecuador). Ingeniera en Computación (ESPOL, Ecuador).

Acosta Guzmán Ivan Leonel

Universidad de Guayaquil

ivan.acostag@ug.edu.ec

https://orcid.org/0000-0002-1589-1825

Docente Titular de la Universidad de Guayaquil (UG). Investigador Auxiliar 1 acreditado por SENESCYT-ECUADOR, líneas Inteligencia Artificial, Machine Learning, Natural Language Processing (NLP) e Inteligencia de Negocios. Desempeño en cargos como Gestor de Posgrado, Gestor de Proyectos de Vinculación y Gestor de Integración Curricular en la UG. Docente y Gestor de Acreditación de la Carrera de Ingeniería en Sistemas (Universidad Politécnica Salesiana sede Guayaquil, UPS). Jefe de Departamento de Aseguramiento de Ingresos, a cargo de proyectos de Auditoria Informática, Inteligencia de Negocios, en procesos de Aprovisionamiento y Facturación (Conecel S.A.). Maestrante en Big Data y Ciencia de Datos (Universidad Internacional de Valencia, VIU-España). Magíster en Administración de Empresas (UEES-Ecuador). Magíster en Sistemas de Información Gerencial (ESPOL-Ecuador). Ingeniero en Computación (ESPOL-Ecuador).

AUTORES

RESEÑA DE AUTORES

Fajardo Romero Inés Janellys

Universidad de Guayaquil

ines.fajardoro@ug.edu.ec

https://orcid.org/0009-0005-3185-5807

Ingeniera en Sistemas Computacionales de la Universidad de Guayaquil (UG). Trabajo de titulación en el FCI 010-2021 “Inteligencia Artificial Conversacional al servicio del Bien Social en un sector vulnerable de la Coordinación Zonal 8 frente a personas contagiados de Covid-19” de la UG.

Oviedo Peñafiel Jorge Alberto

Universidad de Guayaquil

jorge.oviedop@ug.edu.ec

https://orcid.org/0009-0001-0798-7026

Ingeniero en Sistemas Computacionales de la Universidad de Guayaquil (UG). Trabajo de titulación en el FCI 010-2021 “Inteligencia Artificial Conversacional al servicio del Bien Social en un sector vulnerable de la Coordinación Zonal 8 frente a personas contagiados de Covid-19” de la UG.

Índice

Índice de Tablas

Índice de Figuras

Introducción

El Covid-19 es una enfermedad que afecta drásticamente el sistema inmunológico respiratorio de quien la padece, esta enfermedad es considerada pandemia mundial la cual afectó a toda la población alrededor del mundo. Mientras esta enfermedad avanzaba muchos países optaron por incluir soluciones tecnológicas que ayudaran a mejorar la vida cotidiana, puesto que el Covid-19 al ser declarada pandemia altamente peligrosa y de fácil contagio todas las personas debieron aislarse por un tiempo. A pesar del aislamiento se obtuvo como resultado el colapso total de clínicas y hospitales siendo evidente la falta de personal médico y espacios adecuados para solventar dicha emergencia además de la desinformación acerca de síntomas y las recomendaciones que se debían seguir.

Por lo cual se plantea mediante el presente proyecto brindar una herramienta que ayude a la clasificación de síntomas y recomendaciones, y de esta manera poder ayudar a la ciudadanía a conocer que pasos se deben seguir cuando padezcamos de esta enfermedad.

A continuación, se detalla lo que se presentará en cada capítulo de este trabajo de titulación:

- **Capítulo I – El problema –** En este capítulo se presenta una descripción detallada del contenido relacionado al problema en cuestión, sus causas y consecuencia, además de los objetivos específicos que se plantearon para alcanzar la finalización del proyecto y se muestra la relevancia que tiene el presente proyecto y su desarrollo.
- **Capítulo II – Marco Teórico –** Aquí se presentan los antecedentes de la investigación, la fundamentación teórica que debe ser conocida por cada uno de los autores para el desarrollo correcto del proyecto y con pautas base a este.
- **Capítulo III – Metodología de la investigación –** en este capítulo se determina la factibilidad del proyecto, los procesos realizados en la investigación, el meta análisis y los resultados obtenidos por el modelo de Machine Learning para modelos de clasificación textual NLP.
- **Capítulo IV – Resultados, Conclusiones y Recomendaciones –** Finalmente en este último capítulo se detallan las conclusiones, recomendaciones y trabajos futuros del presente proyecto

01 CAPITULO

PLANTEAMIENTO DEL PROBLEMA

Planteamiento del problema

1.1. Descripción de la situación problemática

1.1.1. Ubicación del problema en un contexto

A finales del año 2019, en Wuhan, provincia de Huebi República Popular de China se ve afectada por la aparición de un nuevo tipo de coronavirus llamado SARS-COV-2 (Covid-19), según los estudios realizados uno de los primeros casos se dio en un mercado de animales vivos en Wuhan, la primera persona contagiada presentó síntomas comunes como la fiebre, tos y dificultad para respirar. La nueva enfermedad infecciosa se fue propagando rápidamente dentro República Popular de China, llegando así a convertirse de esta manera epidemia, misma que fue confirmada 30 de enero del 2020 por la Organización Mundial de la Salud (OMS) quienes la dieron a conocer como una Emergencia de Salud Pública de Interés Internacional (Trilla, 2020).

El 11 de marzo del 2020 la Organización Mundial de la Salud (OMS) declara al Covid-19 como pandemia mundial, puesto que este virus es altamente contagioso logró expandirse rápidamente alrededor del mundo, dentro de los indicadores de contagios para el nuevo SARS-COV-2 (Covid-19) se tiene que, el radio de contagio es de 2 metros de distancia mediante gotas respiratorias que se producen al momento de estornudar o de toser. Como consecuencia de esto muy rápidamente se produjo el desbordamiento de la atención médica en todos los centros de salud a nivel global, además del fuerte impacto económico que afecto a todos los países (Cuero César, 2020).

En Ecuador, el Ministerio de Salud Pública (MSP) anuncia el primer caso sospechoso proveniente de un ciudadano chino el 26 de enero del 2020, el cual no se confirmó como un caso positivo para Covid-19, el 29 de febrero del mismo año se confirma el primer caso de Covid-19 en el país, ocasionando la conmoción y desesperación dentro de la población (Guerrero, 2020). Poco a poco los casos de Covid-19 se fueron incrementando dentro del territorio nacional, siendo Guayas una de las provincias más afectadas, en especial el cantón Guayaquil que fue considerado como el epicentro de la pandemia a nivel nacional, teniendo el colapso total de clínicas y hospitales lo cual trae como resultado la evidente falta de personal médico, espacios adecuados para

solventar emergencias y el déficit en la atención para las personas contagiadas o con sospecha de contagio de Covid-19 (Fernández-Garza & Marfil, 2020).

Durante estos últimos años la tecnología y la medicina siguen un camino paralelo, los avances tecnológicos van modificando el concepto de salud y las necesidades sanitarias están influyendo en el desarrollo de la tecnología (Ávila-Tomás et al., 2020). Actualmente los sistemas informáticos permiten mediante la Inteligencia Artificial el análisis predictivo de procesos complejos e imperfectos como el control inteligente de enfermedades y la prevención para el cuidado de la salud (Graf César, 2020).

La Inteligencia Artificial (IA) es una disciplina científica que se enfoca en comprender y crear algoritmos informáticos capaces de realizar tareas similares a los que pueden hacer los humanos, facilitando una mayor accesibilidad, relevancia y capacidad de acción de la información. Esta ciencia, ha recorrido un largo camino, obteniendo avances significativos que incluyen el procesamiento de lenguaje natural, agentes virtuales, aprendizaje automático, etc., brindando así confianza en los algoritmos informáticos específicamente en el aprendizaje profundo, tanto por la información concreta y objetiva, y por la posibilidad de predecir eventos futuros (Joison A et al., 2021).

La aplicación de inteligencia artificial también presenta varias dificultades y desventajas a tomar en consideración, por ejemplo se puede mencionar los altos costes, elevado esfuerzo y tiempo de desarrollo, baja tolerancia a fallas y actualización de funcionalidades muy complejas entre otros problemas existentes que dificulta en gran medida su implementación además del hecho de que en la actualidad no existe una forma de desarrollar un sistema que tenga la capacidad de resolver problemas de forma general en situación ambiguas donde aún la capacidad humana y el sentido común siguen jugando un papel indispensable (Torres, 2019).

Dentro del amplio campo de la tecnología donde se encuentra incluida la inteligencia artificial, atiende el Procesamiento del Lenguaje Natural (NLP) la cual tiene cada vez mayor aceptación en diferentes disciplinas que trabajen con altos volúmenes de datos, entre ellas se encuentra el sector de la salud (Sancho Escrivá et al., 2020).

El Procesamiento del Lenguaje Natural (NLP) es una rama de la Inteligencia Artificial que ha ganado protagonismo en un mundo cada vez más digital, es

gracias a las técnicas de NPL que actualmente se pueden realizar traducciones automáticas, revisión de ortografía o conteo de palabras (Escobar Ariana, 2019). A pesar de tener esta gran aceptación el Procesamiento de Lenguaje Natural, hoy en día se sigue investigando distintas ramas de la Inteligencia Artificial entre ellas el NLP, ya que en esta rama existen dos grandes problemáticas las cuales son: la ambigüedad y la dimensionalidad de los textos; por lo que, estos aspectos hacen que el proceso de NLP se transforme en un problema complejo como en el entendimiento del lenguaje, y su correspondiente procesamiento de textos de forma conversacional (Lady M. Sangacha Tapia et al., 2021) .

El conjunto de técnicas que comprende el NLP consiste en analizar y representar textos naturales mediante software y algoritmos en uno o diferentes niveles de análisis lingüístico con la finalidad de obtener una apariencia humana en el procesamiento de lenguaje para tareas concretas (Sancho Escrivá et al., 2020). En la informática el NLP es un problema difícil, porque el lenguaje humano rara vez es preciso, por lo cual entender al ser humano no solo es comprender las palabras, sino también los conceptos y cómo están vinculados para crear su significado. Por lo tanto, la ambigüedad del lenguaje es lo que hace que el procesamiento del lenguaje natural sea un problema difícil de dominar para las computadoras (Lloret, 2019).

El Procesamiento de Lenguaje Natural (NLP) se desarrolla mediante el aprendizaje automático integrado o técnicas de Embedded Machine Learning tras la recolección de los datos. El Machine Learning es una rama dedicada al estudio de agentes o software que evoluciona en función de su experiencia, para desempeñarse cada vez mejor hacia alguna tarea determinada. El objetivo principal es desarrollar técnicas que permitan a las computadoras aprender. Hay 4 modos de aprendizaje diferentes: mediante el aprendizaje no supervisado, el aprendizaje supervisado o el aprendizaje por refuerzo (Cedeno-Moreno & Vargas, 2020).

El proyecto FCI “Inteligencia Artificial Conversacional al Servicio del Bien Social en un Sector Vulnerable de la Coordinación Zonal 8 Frente a Personas Contagiadas de Covid-19”, fue aprobado durante el ciclo I 2020-2021, con el código FCI 010-2021 de la Facultad de Ciencias Matemáticas y Físicas; estableciendo como objetivo general modelar un Procesamiento del Lenguaje Natural (NLP), para que pueda utilizarse de manera más efectiva al momento de

analizar el lenguaje textual, interpretando y dándole significado especialmente a las terminologías de Covid-19.

Al no contar con un modelo conversacional adecuado a las necesidades previamente mencionadas es muy difícil el desarrollo de una solución que permita brindar un apoyo real a esta problemática. El uso de herramientas como: Machine Learning para el aprendizaje automático de algoritmos aplicando IA, modelos predictivos que permitan precisar en la toma de decisiones y técnicas de análisis; brindara mayores oportunidades para la creación de modelos de conversación que permitan una mejor interacción con personas que manifiesten síntomas relacionados al Covid-19.

Sin embargo, necesitan un gran volumen de datos de entrenamiento. Puesto que, la carencia de una correcta validación del modelo durante el entrenamiento puede llevar a un sobreajuste, el cual puede añadir más ruido aleatorio que datos reales, obteniendo un modelo no generalizable con datos que no son relevantes o que son incorrectos, además, la generación de datos de entrenamiento requiere mucho tiempo, costes y recursos (Martín Noguerol et al., 2019).

Para efectos de estudio, se considera importante tener como muestra poblacional a los ciudadanos ecuatorianos habitantes de la provincia del Guayas, siendo esta una de las provincias más afectadas por el Covid-19, en especial la Zona 8, la cual está conformada por los cantones de Guayaquil, Durán y Samborondón.

1.1.2. Situación conflicto nudos críticos

El Covid-19 es una enfermedad infecciosa, originada por el coronavirus. El cuadro clínico de este se manifiesta entre los días dos y catorce posterior a la exposición del virus, dentro de los posibles síntomas están: tos, fiebre y escalofríos, dolor de garganta, cansancio, malestar general, cefalea y dolor de pecho, sin embargo la sintomatología del Covid-19 puede ir variando y se manifiesta de manera muy diversa entre los portadores del virus llegando a incluir síntomas como las náuseas, diarrea y erupciones en la piel, dependiendo incluso del tipo de variante de la enfermedad, dentro de lo que se conoce hasta la fecha (Narro-Cornelio & Vásquez-Tirado, 2021).

La presencia de síntomas y su perdurabilidad es una cuestión que sigue planteando muchas incógnitas alrededor del Covid-19, llegando a existir incluso casos de pacientes asintomáticos. El desconocimiento generalizado de la

población contribuye a la confusión de la enfermedad con otras, esto debido a la similitud que existe entre los síntomas y debido a que algunos de estos son inespecíficos de la enfermedad del coronavirus (Sosa García, 2020). La consecuencia evidente de esto es el tratamiento inadecuado de las anomalías en la salud del cuerpo humano causado en gran parte por la ausencia de asistencia médica provisional que sea de ayuda a las personas con indicadores de un posible contagio.

El uso de modelos conversacionales alimentados mediante la recolección y análisis de datos masivos para conversaciones textuales es una solución que toma cada vez más fuerza para la interacción aplicación-usuario mediante el uso de los agentes inteligentes además de otras tecnologías disruptivas (Ávila-Tomás et al., 2020). La inexistencia de un modelo altamente eficaz viene dada por las limitantes y dificultades presentes en el desarrollo de los mismo, temas tales como el limitado entendimiento del lenguaje, terminología ambigua, el habla con sobreentendidos y la ausencia de procesos de razonamiento para el examen de expresiones en la medición de las palabras hacen que puedan presentarse errores perjudiciales en el trabajo ejecutado por los modelos (Moreira, Cruz, Gonzalez, & Quirumbay, 2020).

1.1.3. Delimitación del problema

En esta área se plantea las delimitantes del problema, de manera específica en los

diferentes elementos de la investigación. El estudio del presente proyecto forma parte del área del Procesamiento de Lenguaje Natural (NPL) en el campo de la Inteligencia Artificial. Esta investigación propone el diseño de un modelo de Procesamiento del Lenguaje Natural (NLP) para la aplicación de una conversación textual clasificada de personas contagiadas de Covid-19, tomando como población principal a las personas de la Zona 8 que comprenden los cantones de Guayaquil, Durán y Samborondón.

A continuación, en la Tabla 1, se detalla la delimitación del problema:

Tabla 1:

Delimitación del problema

Delimitador	Descripción
Campo	Inteligencia Artificial
Área	Procesamiento de Lenguaje Natural (NLP)
Aspecto	Conversaciones textuales de personas contagiadas de Covid-19
Tema	Diseño de un modelo de Procesamiento del Lenguaje Natural (NLP) para la aplicación de una conversación textual clasificada de Covid-19 del FCI 010-2021.

Nota: En esta tabla se plantean los términos de análisis aplicados para la delimitación del problema conforme al contexto en donde se desarrolla la problemática. La elaboración es propia y la fuente corresponde a los datos de la investigación.

1.1.3.1. Evaluación del Problema

La problemática existente en un contexto tiene como principal finalidad el propósito de ser resuelto evaluando diferentes aspectos que podrían incidir en la interpretación de la hipótesis presentada, a continuación, se realiza la evaluación del problema tomando en cuenta los siguientes factores:

- **Delimitado:** Delimitar un tema de estudio significa, enfocar en términos concretos el área de interés, especificar sus alcances, determinar sus límites (Espinoza Freire, 2018).
- Los modelos para el procesamiento de lenguaje natural pueden ser diversos, pero igualmente complejos como los que usan redes neuronales convolucionales y redes neuronales recurrentes hasta los últimos modelos desarrollados como es el caso de los Transformers que demuestran ser más efectivos, pero a su vez son mucho más complejos que los modelos convolucionales y recurrentes.
- **Claro:** La claridad del problema que se quiere resolver permite la elaboración de una hipótesis clara y concreta a fin de que cualquier investigador que quiera replicar la investigación pueda hacerlo (Espinoza Freire, 2018).

- Elaborar un modelo de procesamiento de lenguaje natural con Python para asistencia de personas con síntomas del Covid-19 usando redes neuronales recurrentes a fin de facilitar asistencia virtual mediante conversaciones.
- **Concreto:** Seleccionar un tema o una idea no lo coloca inmediatamente en la posición de considerar qué información habrá de recolectar, con cuales métodos y cómo analizará los datos que obtenga. Antes necesita formular el problema específico en términos concretos y explícitos (Diaz, 2021).
- Entrenamiento y evaluación de un modelo de procesamiento de lenguaje natural en Python con corpus de datos de conversaciones con personas contagiados de Covid-19 obtenidos mediante encuestas.
- **Relevante:** A la hora de acometer lo que es la valoración de un problema como estos, es imprescindible basarse en aspectos tales como si es real, si se puede calificar como relevante (Saing, n.d.).
- Dar un tratamiento provisional a personas con posible contagio por Covid-19 sin reemplazar la atención médica profesional.
- **Factible:** Se necesita establecer la problemática de la investigación, debe ser usado para definir el alcance del estudio, debe formularse claramente, deberá expresar una relación entre dos o más variables, que sea factible (Niño & Mendoza, 2021).
- **Variables:** la identificación del problema es el paso más importante del método científico y se presenta como la etapa más complicada en la formulación de un estudio de investigación, esto es debido a la cantidad de variables correlacionadas que intervienen en el dominio de este (Niño & Mendoza, 2021).
- Las variables de entrada de modelo son los textos depurados de las conversaciones de personas contagiadas de Covid-19, juntos con variables de peso y categóricas.

1.1.4. Causas y consecuencias del problema

Con el fin de una mejor comprensión por parte de los lectores, a continuación, se da a conocer las causas y consecuencias del problema principal y sus

derivaciones presentes a lo largo de esta investigación, se utilizará la tabla de división para presentar la información.

Tabla 2:

Matriz de causas y consecuencias del problema

Causas	Consecuencias
C1. Carencia de conocimiento en la estructura del modelo con NLP en conversaciones textuales de Covid-19 en la Zona 8 de la provincia de Guayas.	**E1.** Desconocimiento de textos en las conversaciones textuales de Covid-19 en la Zona 8 de la provincia de Guayas para los médicos y futuras personas contagiadas de Covid-19
C2. Desinformación clasificada acerca de los síntomas de Covid-19	**E2.** Puede ocasionar que se imparta información errónea y realizar recomendaciones poco adecuadas.
C3. Desinformación clasificada acerca de los hábitos saludables que se deben llevar en caso de sospechas de tener el Covid-19	**E3**. Poco o ningún control sobre los hábitos saludables que se deben seguir en caso de tener sospechas de Covid-19
C4. Ambigüedad al momento de procesar el texto dentro del modelo NLP conversacional textual en español	**E4**. Puede generar errores al momento de proporcionar información acerca de síntomas o hábitos saludables por los múltiples significados que puede tener una palabra.
C5. Desactualización de modelos existentes NLP de conversación textual en español	**E5**. Limitar el desarrollo de asistentes virtuales con conversación textual en español.
C6. Alta dimensionalidad al procesar el texto dentro del modelo NLP conversacional textual en español	**E6**. Puede ocurrir que el proceso de aprendizaje sea más complicado.
C7. Deficiencia del uso de las técnicas para la depuración de los datos por la falta de información científicas en español.	**E7**. Preparación incorrecta de los datos que pueden causar confusión al modelo

Nota: La tabla indica 7 causas y efectos que producen el poco conocimiento de técnicas de preparación de los datos, también se menciona el mal procesamiento del lenguaje natural aplicado a la conversación textual en español. Elaboración: Jorge Alberto Oviedo Peñafiel e Inés Janellys Fajardo Romero. Fuente: Datos de la Investigación

1.1.5. Formulación del problema

Después de un profundo estudio del tema de investigación, pueden formularse preguntas científicas, el proceso que más peso tiene en este aspecto es la formulación del problema que es el primer paso para hallar una solución, la siguiente pregunta es el fundamento de la presente investigación:

¿Cuál es el impacto de un modelo con Procesamiento del Lenguaje Natural mediante la Inteligencia Artificial utilizando una Dataset para la aplicación de conversación textual categorizada de personas contagiadas de Covid-19 en un sector vulnerable de la Zona 8 del Guayas (Guayaquil, Durán y Samborondón) para el FCI 010-2021?

1.1.6. Objetivos del proyecto

1.1.6.1. Objetivo general

Diseñar un modelo de Procesamiento del Lenguaje Natural (NLP) para la evaluación de la eficacia del análisis del tratamiento de datos en una conversación textual clasificada de personas contagiadas con Covid-19 por medio del lenguaje de programación Python.

1.1.6.2. Objetivos específicos

- Recopilar información de los contenidos del Procesamiento del Lenguaje Natural para el diseño de la aplicación del algoritmo en el modelo de NLP por medio de las técnicas investigativas.
- Preparar el Dataset mediante carga, limpieza y depuración de datos recolectados de encuestas a personas que fueron contagiadas de Covid-19 en la Zona 8 del Guayas (Guayaquil, Durán y Samborondón) para el entrenamiento del modelo de NLP.
- Identificar las entradas y posibles salidas en conversaciones textuales en español clasificadas de personas contagiadas de Covid-19 para el diseño del modelo de NLP.
- Diseñar un modelo con NLP basados en los algoritmos de Procesamiento del Lenguaje Natural para la evaluación eficiente de las entradas identificadas de texto clasificadas.
- Evaluar el modelo por medio de Python con la librería Sklearn para entrenar el modelo de NLP.

1.1.7. Alcance del proyecto

Dentro del alcance del proyecto de titulación se ha considerado lo siguiente:

Levantamiento de información llevado a cabo mediante encuestas virtuales destinada a personas que estén contagiadas o hayan padecido el Covid-19 en la zona 8 de la provincia del Guayas correspondientes a los cantones de Guayaquil, Duran y Samborondón, esta información tiene la finalidad de servir como entrada de datos para la alimentación del modelo de conversación textual.

Limpieza y depuración de la data obtenida mediante la eliminación de cualquier incidencia que pueda afectar el correcto funcionamiento del modelo conversacional como por ejemplos faltas ortográficas, uso inadecuado de signos de puntuación y acentos, incongruencia semántica y posibles vacíos en los registros de la información.

Realizar un análisis, investigación y estudio de información como artículos científicos, libros, material audiovisual de portales académicos mediante el criterio de clasificación de temas relacionados referentes a la investigación, con el cual se podrá hacer uso de toda información relevante que permita la creación de un óptimo algoritmo de modelo para el procesamiento de lenguaje natural de conversación textual.

Entrenamiento del o los modelos creados con toda la información obtenida, previamente depurada y etiquetada, probando diferentes valores de entrada haciendo uso de herramientas en la nube como Google Colab con entornos colaborativos en lenguaje Python como es el caso de Júpiter aprovechando las capacidades y recursos que nos proporcione el entorno de Google.

Desarrollo de interfaz web para la visualización del desempeño del modelo entrenado y con mejor resultados durante su evaluación, con la finalidad de mostrar una página intuitiva que no requiera conocimientos muy detallados del modelo desarrollado.

Validación del modelo de conversación textual por expertos en la materia cuya valoración determinara si es adecuado para posteriores implementaciones en proyectos futuros.

1.1.8. Justificación e importancia

Muchas de las complicaciones relacionadas al Covid-19 tienen que ver con las prácticas poco adecuadas que realizan las personas al momento de presentar síntomas de un posible contagio, a esto se suma un escaso

conocimiento de las prácticas y hábitos saludables que pueden ayudar a sobrellevar dicho padecimiento. La adecuada implementación de un modelo de conversación textual permitirá dar información acertada a personas que presenten un posible contagio de Covid-19 que permitan en lo sumo posible preservar su salud con recomendaciones oportunas para sobrellevar determinados síntomas manifestados. El presente proyecto de investigación servirá como insumo para implementaciones prácticas de asistentes virtuales.

La justificación de la investigación está en función de las siguientes cuestiones

- **La conveniencia** ¿Para qué sirve la investigación?
  Para poder suministrar información adecuada, prácticas y recomendar hábitos saludables a personas que presenten diversos síntomas de un posible contagio.
- **Relevancia Social** ¿Cuál es la transcendencia para la sociedad?
  Mediante la creación de un modelo de procesamiento de lenguaje natural en conversaciones textuales se explorarán nuevas formas de implementar ciencias informáticas e inteligencia artificial a la medicina.
- **Implicaciones prácticas** ¿Ayudará a resolver algún problema practico?
  Ayudará a las personas que no sepan qué hacer con respecto a algún síntoma que presente y que esta acción este a su alcance hasta que acudan a la atención médica profesional, esto es de beneficio a la población en general logrando que muchas personas tengan una asistencia accesible para atender sus dudas.
- **Valor teórico** ¿En qué campo de la teoría sentara alguna pauta?
  La creación del modelo óptimo para las conversaciones textuales podrá ser usado e incluso optimizado para los diversos propósitos que puedan darles investigaciones o proyectos futuros.
- **Utilidad** ¿Qué utilidad tendrá la solución de la investigación?
  Asistencia y ayuda antes la aparición de sintomatología relacionada al Covid-19.

1.1.9. Limitaciones del estudio

A continuación, se dará a conocer las limitantes del presente proyecto de investigación.

- La presente investigación está basada en artículos científicos, revistas, tesis, libros y paginas académicas.
- Las encuestas se realizarán de manera virtual, debido a la pandemia para evitar posibles contagios de Covid-19. Lo que implica que la Dataset puede contener registros no válidos, ya que no se tiene un control sobre los encuestados, los cuales pueden no prestar la atención debida y no respondan de manera correcta.
- Las encuestas se realizarán dentro de la provincia del Guayas, específicamente en los cantones que comprenden la Zona 8, los cuales son: Guayaquil, Durán y Samborondón.
- Se hará uso de herramientas de software libre, por los bajos costos que estos tienen.
- Se utilizará Google Colab como entorno de desarrollo gratuito en la nube para la elaboración del modelo de Procesamiento de Lenguaje Natural en código Python.
- Recursos y equipo computacionales con requerimientos mínimos.
- Por falta de presupuesto se hará uso de espacio en la nube de manera libre para el almacenamiento de la data y alojamiento de la página web.

02 CAPITULO

Marco teórico

Marco teórico

2.1. Antecedentes del estudio

Se realiza una extensa revisión literaria sustentada en investigaciones de temas específicos como: el aprendizaje automático para análisis cognitivo relacionado con la ciberseguridad, el uso de Modelos de Procesamiento de Lenguaje Natural explicable y profundo para combatir la información errónea y realizar predicciones sobre diagnósticos finales en la salud, la desambiguación léxica automática cuyos estudios demuestran resultados significativos en el área de Procesamiento de Lenguaje Natural (NLP), siendo estos apoyo principal para la realización del presente trabajo.

En los últimos años, ha aumentado el interés por diseñar e implementar soluciones que utilicen modelos de Procesamiento de Lenguaje Natural basados en aprendizaje automático (Machine Learning). A raíz de estos avances con NPL (Georgescu, 2020) desarrolló e implementó un modelo NPL basado en Machine Learning especializado en el área de ciberseguridad, este modelo se utiliza en un sistema de indexación semántica diseñado para monitorear y extraer automáticamente la información más relevante para la ciberseguridad, este sistema recopila los datos de texto, y los analiza utilizando el algoritmos de NPL, para almacenar solo documentos relevantes que están indexados semánticamente y mismos que están a disposición de una plataforma donde los usuarios pueden realizar búsquedas semánticas. Este modelo fue desarrollado a través de ontología de dominio utilizando un enfoque de dos etapas: la etapa de simetría y la etapa del ajuste de la máquina. En esta ontología, el modelo se entrenó en un conjunto de unas 300.000 palabras. Para modelos similares informados en la literatura, se logró una puntuación de 0,81 para el reconocimiento de entidades con nombre y de 0,58 para la extracción de relaciones.

Uno de los inconvenientes que existen dentro de Procesamiento del Lenguaje Natural NPL es la ambigüedad del lenguaje, pues esto ocasiona que no se pueda comprender al ser humano no solo en las palabras, sino también en los conceptos y como están asociados para generar un significado. Por ello Haga clic o pulse aquí para escribir texto.(Núñez-Torres, 2021) diseño y desarrollo un

modelo desambiguación léxica automática aplicado al Procesamiento del Lenguaje Natural (NPL). Para la creación de dicho modelo se realizó revisión del fenómeno lingüístico de la ambigüedad léxica, junto con los métodos para la desambiguación léxica automática más representativos en similitud y relación semántica, y basados en conocimiento contextual, también expuso una panorámica cronológica de la utilización del corpus en el análisis lingüístico, junto con recursos lingüísticos informatizados. Posteriormente se ejecutó un experimento de desambiguación léxica automática basado en el corpus SENSEVAL-3 (Evaluating Word Sense Disambiguation Systems), utilizando un método de aprendizaje automático supervisado. Dando paso al montaje de un corpus basado en una submuestra de CODICACH (Corpus Dinámico del Castellano de Chile). Este proceso ha permitido el desarrollo de un modelo de desambiguación léxica automática basado en una medida híbrida, como lingüística basada en la interacción de los dos enfoques taxonómicos de búsqueda de distancia entre caminos y contenido de información.

La desinformación en durante la pandemia es un problema latente a nivel mundial, que impulsaron en la propagación de noticias falsas que dificultaron el control del Covid-19. Ante esta problemática un grupo de expertos de la Universidad de Michigan propuso un modelo de procesamiento de lenguaje natural basado en DistilBERT y SHAP (modelos de arquitectura transformer) para combatir la información errónea acerca del virus, cuya efectividad y eficiencia demostraron ser mejores que modelos antecesores a esta solución basados en la misma arquitectura (Ayoub et al., 2021). La prueba del modelo funciono correctamente en una muestra de datos de 984 afirmaciones iniciales de Covid-19 verificadas, luego replicados con traducción inversa para la detección de noticias falsas, con una precisión del 93.8% muy cercana a su modelo base el DistilBERT que obtuvo una precisión del 97.2%, los resultados manifiestan un mejor trabajo que los modelos tradicionales de Machine Learning. El uso de procesamiento de lenguaje natural en la medicina es una ciencia cada vez más popular, especialmente en lo que respecta a la asistencia médica. Factores como el error humano, el tiempo, la sobrefacturación y los costos de reembolso impiden el desarrollo flexible y rápido de diagnósticos y procedimientos basados en el historial médico. Un grupo de expertos en la salud de la universidad de Indiana presentaron un modelo de NLP para la identificación

de diagnósticos y procedimientos con redes neuronales (Nuthakki et al., 2019). Basados en previos estudios que demuestran la eficacia del aprendizaje profundo, el modelo desarrollado puede asignar notas y códigos médicos para la predicción de un diagnóstico final a partir de entradas no estructuradas de la historia de la enfermedad actual y lo síntomas al momento de ingreso. Usando datos seleccionados para entrenamiento y evaluación junto con más de un millón de notas clínicas, el modelo es puede predecir los principales diagnósticos y procedimientos con una precisión del 80.3% y el 80.5% superando notablemente a soluciones existentes cuya precisión está en el 70.7% y el 63.9%, el objetivo a posterior del equipo es que el modelo supere a la capacidad del profesional en diagnósticas y procedimientos médicos.

A nivel nacional el impacto de la pandemia fue sumamente alto, llegando al punto de colapsar los servicios médicos para atención de los pacientes con casos probables de Covid-19, además de ser notoria y preocupante la ausencia de herramientas digitales que ayuden a sostener la situación fue notable y preocupante. La dificultad expuesta llevó al desarrollo de varios modelos entre ellos aquellos destinados a la predicción y derivación ambulatoria que faciliten la labor del personal profesional en medicina (Fuentes Marmolejo & Medina Parra, 2020). Al analizar diversos algoritmos con diferentes predicciones, se concluyó que el mejor modelo recomendado en este estudio es el modelo de Naive Bayes con una precisión del 95% de precisión en el diagnóstico de Covid -19. Este modelo es de suma utilidad para la acción de decisión de los doctores en carácter de derivación.

2.2. Fundamentación teórica

2.2.1. SARS-CoV-2 (Covid-19)

El SARS-CoV-2 es parte de la familia de coronavirus que están estrechamente relacionado con el SARS-CoV, este recibe diferentes nombres como: 2019-nCoV, Virus Wuhan y Nuevo coronavirus de Wuhan (WN-CoV), COVID-19, etc. En diciembre del 2019 aparece por primera vez en la ciudad de Wuhan, provincia de Hubei, República Popular de China el primer caso de Covid-19, el cual se propaga por todo el país asiático, ocasionando una enfermedad respiratoria aguda. A raíz de su alto índice de contagio dentro de dicho país el

30 de enero de 2020, la Organización Mundial de la Salud (OMS) declara al Covid-19 como una emergencia de salud pública de importancia internacional. Poco a poco el Covid-19 fue expandiéndose a los distintos países, dejando como consecuencia el colapso en los sistemas sanitarios, afectaciones económicas y millones de personas contagiadas y fallecidas, siendo así que el 11 de marzo del 2020 la (OMS) declara al Covid-19 pandemia mundial. (Cuero César, 2020).

Como lo indica (Maguiña Vargas et al., 2020) en su estudio, el virus SARS-CoV-2 es altamente contagioso y se propaga rápidamente de persona en persona por medio de la tos o secreciones respiratorias; este virus es capaz de viajar hasta dos metros de distancia mediante las gotas respiratorias de más de cinco micras, además las manos o los fómites son foco de contagio, puesto que podrían tener secreciones contaminantes, las cuales al tener contacto con la mucosa de la boca, nariz u ojos se podría contagiar de este virus.

El tiempo de incubación de este virus es de 4 a 7 días, pero la mayoría de los casos son de 12 días, en este periodo de tiempo los portadores del virus pueden contagiar a más personas sin saber que lo posee. Por otra parte, se han presentado miles de casos en los que el portador de dicho virus es asintomático, es decir que no presentan ningún tipo de síntoma, a pesar de ello estos pueden contagiar a los demás; los pacientes sintomáticos suelen presentar diferentes niveles de sintomatología los cuales son: cuadros leves con malestar general y tos ligera; cuadros moderados como fiebre, tos seca persistente, fatiga y cuadro severo que se caracteriza por fiebre constante, tos, disnea severa, neumonía, este cuadro puede llegar a ocasionar daño cardiovascular, falla multiorgánica, y pueden llegar a provocar el fallecimiento de la persona (Maguiña Vargas et al., 2020).

Uno de los órganos más afectados por la infección por Covid-19 son los pulmones ya que ocasiona dificultades respiratorias; no obstante, también puede llegar a ocasionar daños cardiovasculares, gastrointestinales, renales, hepáticos del sistema nervioso central y oculares que deben ser monitoreados de cerca. Estas afecciones suelen ser consideradas secuelas a raíz del contagio por del virus. Si estas no son tratadas de manera correcta pueden empeorar rápidamente y morir por insuficiencia orgánica múltiple (Ciotti et al., 2020) .

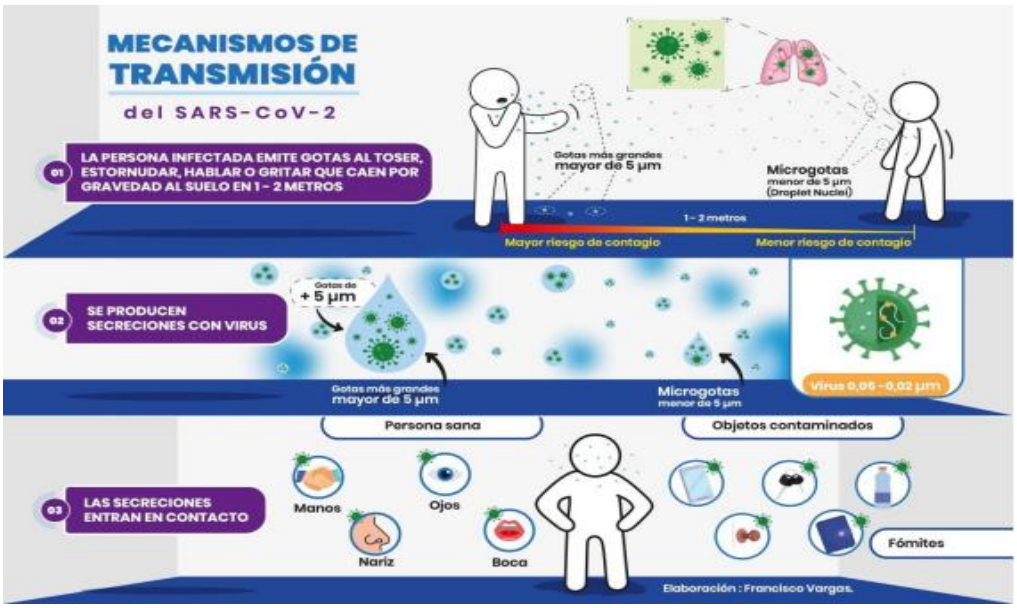
En el estudio realizado por (Palacios Cruz et al., 2021) explica que el Covid-19 es capaz de infectar a personas de diferentes edades sin distinción, aunque en

su mayoría las personas mayores de 70 años son más vulnerables a ser contagiadas y de presentar sintomatología más severa. Sin embargo, según informes el mayor índice de fallecidos tenía 56 años en promedio, y gran parte de estos padecían otras afecciones médicas como: asma, diabetes, enfermedades cardiacas, accidente cerebrovascular, etc., ocasionando que estos sean mucho más vulnerables al virus.

Según indica (Khanna et al., 2020) que al ser el Covid-19 un virus muy contagioso y de fácil propagación se recomendó tomar medidas preventivas para reducir la trasmisión de la infección. La OMS sugirió el lavado de manos frecuente con desinfectante para manos a base de alcohol o con agua y jabón, evitar tocarse los ojos, la nariz, y la boca, y practicar la higiene respiratoria. Además, se recomienda el uso de equipo de protección personal (EPP) con máscaras de tres capas o máscaras N95, a esto se añade el distanciamiento social (un metro como mínimo) tanto a nivel individual como comunitario. En caso de estar contagiado y evitar la propagación se sugiere el aislamiento y el rastreo de personas que estuvieron en contacto con el caso positivo, seguido de la cuarenta de dichas personas.

Figura 1:

Mecanismos de transmisión del SARS-CoV-2



Nota: El siguiente gráfico representa la manera existente de transmitir el SARS-CoV-2 (Covid-19). Adaptado de “Transmisión del SARS-CoV-2 por gotas respiratorias, objetos contaminados y aerosoles (vía aérea)” (p. 8), por Vargas Marcos, 2020, Sociedad Española de Sanidad Ambiental.

2.2.2. Inteligencia Artificial

La Inteligencia Artificial (IA) tiene la capacidad de simular la inteligencia humana, es una ciencia multidisciplinaria con diversos enfoques como: el aprendizaje automático y el aprendizaje profundo, estos han creado nuevos paradigmas en diversos sectores de la industria tecnológica. La IA cuenta con diversos métodos, técnicas y herramientas para modelar y solucionar problemas simulando el proceder de los sujetos cognoscentes, lo que le permite hacerse cargo del diseño de sistemas inteligentes relacionados con el comportamiento humano. Además, tiene la capacidad de instruirse a través de datos usando algoritmos que le permiten hacer uso de la información y tomar las mejores decisiones, así como lo haría el ser humano (Ocaña-Fernández et al., 2019).

La Inteligencia Artificial a pesar de ser un término acuñado hace poco tiempo, tuvo su primera aparición en 1950 por Alan Turing, quien describe a la IA como un simulador del comportamiento humano. Alan Turing en su libro Computers and Intelligence determina que las máquinas son capaces de alcanzar la inteligencia humana. Aproximadamente seis años después, John McCarthy describió a la IA como la ciencia y la ingeniería de crear máquinas inteligentes. Desde entonces la IA con simples reglas como: “si, entonces” ha ido evolucionando a lo largo de varias décadas hasta añadir algoritmos complejos, los cuales han llegado a funcionar de forma similar al cerebro humano. Hoy en día la IA no solo a transformado el mundo empresarial, sino también el ámbito social con sus diversos usos (Kaul et al., 2020).

A menudo es común que cuando se menciona el término Inteligencia Artificial automáticamente es asociada de manera directa con robots, ya que en la actualidad es lo más vendido por la industria del cine y televisión, creando una idea sobres las máquinas en donde estas se parecen a los humanos y causan daño, haciendo que esto se aleje de la realidad y genere incertidumbre en las personas, quienes comienzan a considerar a la IA como algo negativo. Sin embargo, el objetivo de la IA no es causar daño; puesto que existen tres leyes de la robótica creadas por Isaac Asimov que resumidamente indican que los robots no causaran daño, ni permitirá que el ser humano sufra, seguirá todas las órdenes de un ser humano y se protegerá a si mismo respetando las leyes anteriores. La Inteligencia Artificial se basa en el principio de la inteligencia humana ya que busca que una máquina pueda imitar y ejecutar tareas humanas

ya sean desde las más simples hasta las más complejas, teniendo como objetivo el aprendizaje, el razonamiento, y la percepción (Fuentes Marmolejo & Medina Parra, 2020).

Las técnicas y tecnología basada en inteligencia artificial son usadas en toda una gama de aspectos cuyo objetivo en dar el mayor beneficio a la labor humana que contribuyan a potenciar la eficiencia en producción de la industria humana así también dar solución a cuestiones de ámbito social que son consideradas difíciles y tediosas e inclusive peligrosas, cuya consecuencia será en la realización de tareas se pensaban imposibles (Rouhiainen, 2018).

Algunas de las aplicaciones más comunes son:

2.2.2.1. Empresa

La complejidad del mundo actual delimitada por marcos como la globalización, comercio masivo, riesgo y toma de decisiones en los mercados presentan nuevos paradigmas que desestiman la perspectiva de un mundo viejo en el que existían las certezas. Aplicaciones de la inteligencia artificial como el manejo automático de la información, predicción de mercados, interpretación de publicidad y devoción a la atención personalizada son algunos de los ejemplos que demuestran con claridad el impacto positivo que ha marcado la IA en su incursión por la empresa e industria (Valverde Bourdié & Sandra, 2019).

2.2.2.2. Medicina

La incorporación de la inteligencia artificial en la medicina es un aspecto de estudio contemporáneo limitado a las ultima décadas en las que se demostrado la utilidad que tiene aplicarla por ejemplo para la mejora de la precisión diagnostica, aceleración de procesos, procesamiento de imagen radiológicas, estudio de patologías y uso de machine Learning para el registro de atenciones de médicas que ayuden al proceso de atención y procedimientos en pacientes, así mismo están en desarrollo actualmente números proyectos dedicados a explorar la aplicación de IA en todos los aspectos de la medicina (Ávila-Tomás et al., 2021).

2.2.2.3. Marketing

El análisis y estudio de las masas con fines de ventas y publicidad es una ciencia con larga trascendencia en la economía de las empresas y naciones enteras, sin embargo gracias a la inteligencia artificial se pueden actuar sobre objetivos mucho más ambiciosos como la predicción de comportamiento y atención de

necesidades de masa que son aprovechados en gran parte por la industria y diversas entidades de índole comercial, aplicaciones como la maximización de las estrategias publicitarias, programática digital, análisis de datos y la optimización de campañas son tareas en la que el ser humano empleo un tiempo mucho mayor que la IA y que revelan lo útil que es para los objetivos de esta ciencia (Costalgo, 2019).

2.2.2.4. Economía

La naturaleza de la inteligencia artificial le permite aplicada a campos en los que prima la producción, como en el estudio y análisis de los mercados y precios por lo que su aplicación no se limita a una industria en concreto debido a que en relación de propósitos cumple con los requisitos GPT (Generalización, Dinamismo, Innovación) estas características le permiten ser aplicada a todas las áreas de comportamiento en base a capital y valor, la progresiva implementación potenciara la economía global (Gallardo, 2018).

2.2.3. Ventajas y desventajas de la inteligencia artificial

Tabla 3:

Matriz de ventajas y desventajas de la IA

Ventajas	Desventajas
V1. Procesamiento de información masiva es mucho más efectivo que la labor humana	D1. El desarrollo de las IA requiere un conocimiento amplio en varias disciplinas por lo que hay necesidad de personal capacitado
V2. capacidad para trabajar con información incompleta y predecir comportamientos	D2. Alto consumo de recursos computacionales para el procesamiento masivo de datos.
V3. Capacidad de aprendizaje profundo mediante algoritmo según la necesidad y efectividad	D3. Muchas veces las IA resultan en comportamientos inentendibles debido a la complejidad de las redes de su desarrollo lo que limita la intervención técnica de la mano humana
	D4. El aprendizaje de una inteligencia artificial requiere de extensos corpus

V4. Mayor precisión en las predicciones y una degradación muy lenta en relación con su tiempo de funcionamiento	de datos que sean usados de entrada para su desarrollo y entrenamiento
V5. Capacidad de retroalimentación con la información trabajada y las entradas proporcionadas al sistema experto	D5. Eventualmente reemplaza determinadas necesidades en que ya no será necesaria la intervención humanan
V6. Aplicabilidad a diferentes ciencias para un mejor trabajo que ha sido demostrado su eficacia cada vez más elevada	D6. Las estructuras de redes para su elaboración pueden resultar muy complejas y alargan considerablemente el tiempo de entrenamiento

Nota: Matriz de ventajas y desventajas de la IA. Elaboración: Jorge Alberto Oviedo Peñafiel e Inés Janellys Fajardo Romero. Fuente: información tomada de (Gallardo, 2018), (Costalgo, 2019), (Ávila-Tomás et al., 2021), (Rouhiainen, 2018) Elaborada por los autores.

2.2.4. Inteligencia artificial conversacional

En anteriores párrafos se ahondo en la definición y conceptos relacionados de la inteligencia artificial, sin embargo, uno de los aspectos en los que más utilidad se ha demostrado son en las entidades conversacionales, los cuales son programas diseñados para tener conversaciones ya sean verbales o textuales con las personas y que esta sea lo más natural posible asemejándose a una conversación normal entre humanos, cuya interfaz de comunicación normalmente suele ser alguna plataforma o programa (Jimenez Flores et al., 2020).

Según (David & Cortés, 2020) una IA conversacional debe tener las siguientes características:

Procesamiento y comprensión del lenguaje natural: la IA conversacional debe ser capaz de entender y procesar una salida a las respectivas entradas en una conversación natural con su interlocutor humano.

Integración de aplicaciones externas: debe poseer una interfaz de comunicación e integración con proveedores externos de información.

Inserción sencilla de datos: la inserción de datos tiene que ser sencilla sin la necesidad de reprogramar la IA conversacional

Contexto conversacional: debe tener la capacidad de situarse en contexto de la conversación que el usuario está teniendo con esta.

Base de conocimientos de preguntas y respuestas: la IA conversacional debe tener una base de parejas de preguntas y respuestas que será consumido por la IA para dar respuestas acertadas a los usuarios.

Memoria conversacional: debe poder soportar cambio de situaciones sin olvidar información de contextos anteriores que sean de utilidad al usuario.

El fundamento principal para la inteligencia artificial conversacional es la imitación de la conversación natural entre personas, de este proceso dinámico se destacan aspectos como los temas y categorías de las conversaciones que condicionan sustancialmente las respuestas emitidas por cualquiera de los participantes de las conversaciones. Este estudio es la base de la construcción de los sistemas conversacionales para asistencia virtual o comúnmente conocido como chatbots cuya denominación es la implementación práctica de una IA a la interacción por el humano mediante la conversación (Colomo Magaña et al., 2020).

2.3. Machine Learning

El Machine Learning o aprendizaje automático es un campo de la Inteligencia Artificial que, a través de algoritmos permite que una máquina pueda aprender por si sola y sea capaz de realizar predicciones o sugerencias de manera autónoma. Es decir, que estos cada vez van mejorando y presentados resultados óptimos tras el procesamiento de una gran cantidad de datos sin tener ningún tipo de instrucción externa (Núñez Raíz et al., 2019) .

El aprendizaje automático (ML) es una técnica que incluyen algoritmos que permite que las máquinas adquieran conocimiento y estos vayan aprendiendo mediante el análisis de grandes volúmenes de datos y en base a sus anteriores experiencias. Estos algoritmos ayuden a predecir y tomar mejores decisiones basado en los patrones de comportamiento.

En el estudio realizado por (Choi et al., 2020) explica que en Machine Learning, existen tres diferentes métodos de aprendizaje que son frecuentemente usados,

cada uno de ellos es útil para resolver un problema distinto, los cuales están: el aprendizaje supervisado, el aprendizaje no supervisado y el aprendizaje por refuerzo.

En la actualidad los algoritmos de Machine Learning son un recurso tecnológico que ha sido incluido en herramientas que se utilizan de manera cotidiana, por ejemplo: las redes sociales son capaces de analizar los diversos comportamientos en la sociedad mediante el análisis de los datos, los asistentes virtuales hoy en día están siendo muy utilizado en muchas empresas ya que han optado por la atención al cliente mediante estos, los cuales son capaces de responder como un ser humano lo haría, otros de los ejemplos claros en la actualidad son aquellas aplicaciones de entretenimiento que mediante el uso de algoritmos de Machine Learning son capaces de conocer las preferencias de sus usuarios ya sean musicales, películas o series.

2.3.1. Ciclo de vida del Machine Learning

El Machine Learning es un tipo de inferencia estadística capaz de aprender por medio de datos existentes y datos desconocidos. Es por ello, que mientras mayor cantidad de datos se proporcione, su conocimiento será aún mayor y podrá realizar predicciones mucho más precisas.

El aprendizaje automático tiene un ciclo de vida el cual es un proceso cíclico usados para construir proyectos de aprendizaje automático eficiente. El ciclo de vida del Machine Learning comprende de 5 pasos los cuales son:

2.3.2. Data Collection

La recolección de dato es el primer paso del ciclo de vida del aprendizaje automático, estos son un conjunto de datos que se utilizan para identificar una población y medir sus características y etiquetas. De los datos recopilados no es recomendable incluir toda la población objetivo, sin embargo, entre las características y etiquetas se muestrean de un subconjunto (Suresh & Guttag, 2021).

2.3.3. Data Preparation

Una vez obtenidos los datos, es frecuente realizar una serie de pasos de preprocesamiento de datos los cuales son: el tratamiento de datos duplicados, datos perdidos, la normalización, el aumento y control de calidad. Al finalizar esta etapa de preparación de datos se obtendrá un conjunto de datos de entrenamiento que serán usados en el modelado (Spjuth et al., 2021).

2.3.4. Model Development

En esta etapa se lleva a cabo la construcción de un modelo optimo y eficiente el cual pueda minimizar los errores de predicción y etiquetas reales. Estos modelos se construyen a partir de datos de entrenamiento obtenidos en la preparación de los datos, excluyendo los datos de validación retenidos. Existen distintos tipos de modelos los cuales se van a comparar según la capacidad de comprender diversos patrones, reglas y características, además de su rendimiento y predicciones más exactas, eligiéndose de esta manera el mejor modelo.

2.3.5. Model Evaluation

Una vez finalizada la construcción del modelo, se evalúa el rendimiento de este realizando pruebas con un conjunto de datos no vistos para garantizar que el rendimiento del modelo sea una representación real de cómo se comporta. En esta etapa se prueba la seguridad y fiabilidad del modelo evaluándolo con datos recolectados en distintas condiciones y datos generados para simular eventos improbables. Así, de esta manera se puede asegurar que el modelo cumple los requisitos necesarios y es económicamente viable a comparación de los otros modelos (Gärtler et al., 2021).

2.3.6. Model Deployment

En esta etapa el modelo se implementa en un sistema de mundo real, este modelo que ha sido preparado con anterioridad genera un resultados precisos, rápidos y aceptables permitiendo de esta manera se despliegue finalmente en un sistema. Sin embargo, puede ser necesario modificar el modelo en aplicabilidad o coherencia aparente de los resultados ya que se debe verificar el mejoramiento del rendimiento. Esta etapa es el informe final del proyecto (Zaharia et al., 2018) .

2.4. Tipos de Machine Learning

2.4.1. Aprendizaje Supervisado

El aprendizaje supervisado o Supervised Learning es un tipo de Machine Learning que se encarga de aprender, posee una función cuya entrada y salida se basa en pares de ejemplos. En los algoritmos de aprendizaje supervisado, sus variables de entrada se dividen en dos grupos de datos los cuales son: entrenamiento y prueba. Los datos de entrenamiento constan de una salida que

debe predecirse y clasificarse y estos, a través de patrones aprendidos durante el entrenamiento, estos son aplicados al conjunto de datos de prueba (Mahesh, 2018).

El aprendizaje supervisado usa un conjunto de datos para su entrenamiento mediante etiquetas permitiéndole realizar predicciones y corregir los datos si están equivocados. Teniendo en cuenta que, si dotamos de suficientes datos al algoritmo este podrá agregar nuevos datos sin tener que etiquetarlo puesto que, el proceso de entrenamiento de este algoritmo continúa aprendiendo hasta que el modelo alcanza el nivel deseado de precisión.

2.4.2. Aprendizaje no supervisado

El aprendizaje no supervisado o comúnmente denominado aprendizaje de clustering consiste en técnicas de agrupación de objetos similares que en lugar de dar predicciones directamente son más bien usadas como herramientas complementarias de los algoritmos de aprendizaje supervisado, un ejemplo de esto son las investigaciones a los modelos SVM agrupados usando algoritmos de aprendizaje no supervisado como los K-means jerárquicos divisorios (DHK) (Bao et al., 2019).

Los algoritmos de aprendizaje no supervisado se basan en la premisa de clasificación por atributos específicos que permitan crear grupos con características similares, o en términos más sencillos, la agrupación de objetos en base a sus atributos por lo que a esta clase de algoritmos se los conoce como algoritmos de clustering. Esta agrupación tiene beneficios que permiten por ejemplo a la detección de anomalías o a la segregación de corpus de datos simplificando así su procesamiento.

2.4.3. Reforzamiento

El aprendizaje por reforzamiento es un algoritmo de optimización que realiza la interacción con un modelo y su entorno, para aprender acciones optimas que se maximicen mediante las recompensas. Las recompensas son usadas para obtener el comportamiento deseado, mantener un ajuste de la parametrización y la optimización del mismo proceso empleado. El algoritmo de aprendizaje reforzado intenta maximizar la recompensa obtenida que recibe (Machalek et al., 2021).

El aprendizaje por refuerzo es un algoritmo que basa su efectiva en recompensas y castigos, a diferencia de los algoritmos supervisados y no supervisados, el

aprendizaje por refuerzo consiste en trazar una ruta de decisiones que llevaron al objetivo deseado, obteniendo una recompensa por cada decisión acertada y un castigo por aquellas que alejan el resultado de la meta planteada inicialmente. Estos algoritmos basan también su efectividad a ser robusto en un entorno cuyas variables tienen un constante cambio según avanzan los procesos del entorno.

2.4.4. Regresión Logística

La regresión logística es un método estadístico utilizado para resolver problemas de clasificación binaria, que utiliza una función logística para modelar una variable dependiente, que puede ser binomial (o binaria), ordinal o polinomial. Es un método de clasificación supervisada la encuentra la probabilidad de una nueva instancia pertenezca a una determina clase. La Regresión Logística es una probabilidad la cual da como resultados valores entre 0 y 1 (Giraldo Ossa & Jaramillo Marín, 2021).

La regresión logística ayuda a predecir si un evento ocurrirá o no, ya que solo existen dos posibilidades, “que ocurra” o “que no ocurra”, ya que posee una clasificación binaria. Usar el algoritmo de regresión logística en un modelo usa los datos de entrenamiento para encontrar valores que reduzcan errores entre resultados previstos y resultado reales.

2.4.5. Naives Bayes

El algoritmo de Bayes es un clasificador de probabilidad basado en el teorema de Bayes, suponiendo que cada función contribuye por igual a la clase destino de forma independiente y no interactúa entre sí. Por lo cual contribuye de manera independiente y equitativa a que una muestra probablemente pertenezca a una clase en particular. Es favorable para usar en aplicaciones y tiempo real ya que es sencillo de aplicar y computacionalmente rápido, además tiene resultados positivos en grandes conjuntos de datos ya que tiene una gran dimensionalidad y no es afectado por el ruido. Estos algoritmos toman un conjunto de datos los cuales calcular la probabilidad de la clase y probabilidad de condiciones que define la frecuencia del valor de cada característica para un valor de clase dividido por la frecuencia de la instancia con ese valor de clase (Misra & Li, 2020).

2.4.6. Reglas de Asociación

Las reglas de asociación son algoritmo que se crean mediante la búsqueda de datos patrones frecuentes como “si-entonces” y utilizan los criterios de soporte y

confianza para para determinar las relaciones más importantes. El soporte mide la frecuencia con la que aparece un elemento en los datos. La confiabilidad mide el número de veces que una afirmación como "si, entonces" se considera verdadera. Se puede usar una tercera métrica, llamada ajuste, para comparar la confianza con la confianza esperada, o el número de veces que una declaración "si-entonces" puede ser verdadera (IBM, 2021).

2.4.7. Support Vector Machines

Support Vector Machines o Máquina de Vectores de Apoyo (SVM) es una técnica de aprendizaje de modelos para regresión o clasificación de datos desconocidos. Estos tienen dos fases: la de entrenamiento en donde se procesa la información creando un modelo basado en datos entrenados o conocidos obtenidos desde la experimentación y de prueba la cual usa la clasificación de datos desconocidos en base al modelo de entrenamiento (Muthukrishnan et al., 2020).

2.4.8. Decision Trees

Decision Trees o árboles de decisión son algoritmos que pueden predecir variables objetivas a partir de otras variables. Al hacer uso de una estructura de árbol para tomar decisiones es transparente y comprensible de manera especial cuando el árbol no es demasiado grande. Este modelo se entrena mediante datos de entrenamiento y se validan en datos de prueba para ajustar y mejorar el modelo. Al automatizar un modelo es una característica clave dentro del aprendizaje automático, mientras que la creación de un modelo en estadística generalmente significa que el ser humano es esencial para configurar un modelo estadístico. Una vez realizado el algoritmo valida el proceso y se proporcionar datos válidos y útiles y se evalúa la utilidad del modelo entregado por el algoritmo de aprendizaje automático (Biehler & Fleischer, 2021).

2.4.9. Ensemble methods

Ensemble Methods o métodos de conjunto es una técnica que permite unir predicciones de varios estimadores de base construidos con un algoritmo de aprendizaje dado para mejorar la generalización y robustes para un solo estimador. Existen dos métodos utilizados en los métodos de conjunto que son: los métodos de promediación y los métodos de boosting. Los métodos de promediación construye estimadores de manera independiente y promedia sus predicciones ya que reduce la varianza. Los métodos boosting se crean de

manera secuencial y trata de reducir el sesgo del estimador combinado (Scikit-learn, 2021).

2.5. Algoritmos de aprendizaje no supervisado

2.5.1. K-Means Clustering Algorithm

Este es un algoritmo de aprendizaje no supervisado el cual es comúnmente usado para la resolución de problemas de agrupamiento en aprendizaje automático. Este algoritmo es capaz de unir un conjunto de datos sin necesidad de etiquetarlos en distintos grupos, ya que K es el número de clústeres que deben crearse en el proceso, de esta manera cada conjunto de datos solo pertenece a un solo grupo con propiedades similares (Sinaga & Yang, n.d.) .

Tabla 4:

Matriz de beneficios de los Algoritmos de Aprendizaje Supervisado

Algoritmo	Beneficios	Ejemplos de aplicación
Regresión Lineal	- Sencillo de entender y de explicar - Es muy sencillo en su desarrollo - Son usados para datos con variadas características - Sencillos en su aplicación	-Usado en el análisis predictivo -Estudio de mercados
Regresión Logística	- Pueden ser actualizados con facilidad - No hace suposiciones sobre las variables independientes - Buenos resultados probabilísticos	- Análisis y edición de textos - Procesamiento de imágenes
Reglas de asociación	- Son útiles para descubrir patrones en los datos - Parten de una conclusión para generar muchas situaciones	- predicción de análisis de datos médicos - aplicaciones en la minería de datos

Máquinas de soporte vectorial	- son usados para la minería de datos - Son modelos muy robustos - pueden trabajar sobre múltiples espacios - Poseen un menor riesgo de Overfitting -Son excelentes para tareas de clasificación	- categorización de texto e hipertexto - reconocimiento de escritura
Árboles de decisiones	- La clasificación es mucho más fácil de interpretar -la depuración de los datos es más fácil - puede trabajar con diferentes tipos de datos como los numéricos y los nominales - Crea robustos clasificadores evaluados mediante métricas	- Minería de datos - son muy aplicados a problemas de clasificación
Métodos de ensamblaje	- Permite combinar varios modelos para mejorar la predicción - reducen los errores en entrenamiento y prueba	- Teledetección en cambios de superficies y cartografía - detección de malware
Naive Bayes	- Sencillos de usar en grandes corpus de datos - puede ser usado para clasificaciones - Necesita mucho menos datos de entrenamiento - Puede trabajar de manera efectiva con datos discretos	- Clasificación de texto - Sistemas de recomendación

Nota: Matriz de beneficios de los Algoritmos de Aprendizaje Supervisado. Elaboración: Jorge Alberto Oviedo Peñafiel e Inés Janellys Fajardo Romero. Fuente: (Jet & O, 2017)

Tabla 5:

Matriz de beneficios de los Algoritmos de Aprendizaje No Supervisado

Algoritmo	Beneficios	Ejemplos de aplicación
Clustering	- Fácil implementación - Fundamentos matemáticos sencillos - Son muy usados para tareas de clasificación - Rápida convergencia	- Análisis y clasificador de texto - Construcción de bases documentales

Nota: Matriz de beneficios de los Algoritmos de Aprendizaje No Supervisado. Elaboración: Jorge Alberto Oviedo Peñafiel e Inés Janellys Fajardo Romero. Fuente: (Jet & O, 2017)

2.6. Evaluación de métricas y modelos

2.6.1. Métricas de clasificación

Las métricas de clasificación son una de las principales temáticas de investigación en esta rama de las métricas, cada conocimiento procesado requiere sistemas de calificaciones sistemática para estructuras los datos y el contenido, algunos ejemplos de las métricas más comunes de este tipo son la matriz de confusión, Accuracy y precisión los cuales están constantemente bajo escrutinio científico (Botchkarev, 2019).

Las métricas de clasificación son una parte de las multidisciplinarias métricas de rendimiento de los modelos, son una parte vital en la evaluación de los desarrollos de modelos de aprendizaje automático que permiten evaluar lo acertado que son las predicciones o decisiones en la tarea de la segregación de conjunto de datos, que permiten muchas cosas como la detección de patrones de comportamientos.

2.6.2. Matriz de Confusión

Tabla 6:

Matriz de Confusión

		Valores Predichos	
		Clase 1	**Clase 2**
Valores reales	**Clase 1**	Verdadero Positivo	False Negativo
	Clase 2	Falso Positivo	Verdadero Negativo

Nota: Matriz de Confusión. Elaboración: Jorge Alberto Oviedo Peñafiel e Inés Janellys Fajardo Romero. Fuente: (Botchkarev, 2019)

Los posibles valores son:

- Verdaderos Positivos (TP)
- Falsos Positivos (FN)
- Falsos Negativos (FP)
- Verdaderos Negativos (TN)

Accuracy:

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

Precisión:

$$precision = \frac{TP}{TP+FP}$$

2.6.3. Métricas de Regresión

Si bien pueden ser aplicadas ampliamente las métricas de clasificación para la regresión, existen muchos huecos dejados por estas que son lidiados con técnicas como el error absoluto (|A-P|) que trabaja con valores no negativos y el error cuadrático, sin embargo, en este ámbito hay métricas mucho más efectivas y minuciosamente revisadas tales como la métrica R, R2, MAE, MAPE y RMSE que son más efectivas para la adecuación en modelos de regresión (Naser & Alavi, 2021).

Estas métricas son complementadas con métodos de verificación multicriterio que pueden hacer uso de métricas más tradicionales así también métricas mucho más modernas. A estar en verificación constantes estas métricas permiten una evaluación mucho más acertada de los modelos de regresión y a su vez combatir las limitaciones que pudieran presentarse al momento de evaluar la cercanía que tienen los resultados del modelo a los resultados objetivos.

Error Medio Absoluto

$$EMA = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} = \frac{\sum_{i=1}^{n} |e_i|}{n}$$

Error Cuadrático Medio

$$ECM = \frac{1}{n}\sum_{i-1}^{n}(Y_i \acute{} - Y_i)^2$$

Sesgo / Varianza

Gran parte de las métricas de sesgo involuntario trabajan mediante la división de los datos para las pruebas en subgrupos que son comparados con otros datos también segregados en subgrupos o comúnmente llamados “datos de fondo”, algunas de las métricas más conocidas de este tipo son: métrica de ligero cambio de puntuación, métrica de alto cambio de puntuación, desplazamiento de puntuación y desviación de tamaño, cambio de puntuación a la izquierda, baja separabilidad de subgrupos, amplio rango de puntuación de subgrupos sin solapamiento y amplio rango de puntuación de subgrupos con solapamiento (Borkan et al., 2019).

El sesgo podría definirse en que tan diferente son los valores predichos a los valores esperados y la varianza en que tanto cambiarían los valores de nuestra función objetivo si se utilizan diferentes datos de entrenamiento. Todos estos cambios producidos por el dinamismo del entorno del modelo pueden ser evaluados mediante la extracción de grupos de datos que permiten la comparación de resultados versus resultados objetivos.

El sesgo es la diferencia entre el valor esperado del que fue estimado y la varianza es la estimación de cuanto varia la predicción del modelo en dependencia de los datos usados para el entrenamiento, así un error total se determina de la siguiente manera:

$$Error\ total = \ Bias^2 + Varianza + Error\ irreducible$$

2.6.4. Overfitting/Underfitting

El exceso y escases de entrenamiento es una de las problemáticas al más comunes en esta etapa, ocurren dos fenómenos conocidos como Overfitting y Underfitting. Overfitting se denomina cuando el modelo empieza a aprender características no relevantes que lo hacen más efectivo en su predicción, pero no útil para diferentes circunstancias, algo como una memorización en lugar de aprendizaje, Underfitting se puede describir como escases de datos de

entrenamiento que no permite al modelo extraer información relevante (Robissout et al., 2021).

Estos indicadores de evaluación son usados para medir el rendimiento interno de los modelos de aprendizaje profundo, estas métricas de rendimiento como también se las conoce sirven para encontrar un punto medio en que el modelo aprenda como es debido, muchas veces esto suele resulta una tarea compleja debido a los números valores de parametrización a probar para la optimización del modelo.

2.6.5. Regularización

La regularización es una técnica de evaluación que permite la optimización de la predicción de los modelos en forma general, estos bien pueden ser implementados controlando el número de parámetros en los modelos o establecer una estructura de los parámetros del modelo de manera forzosa. Algunos de los ejemplos de regularización son: regularización de la norma de Frobenius, norma del L-21, regularización de la norma nuclear y la regularización de cierre consecutivo (Zhu et al., 2018).

El ajuste forzoso de la estructura de la red del modelo es una técnica de evaluación que permite la medición de precisión para diferentes circunstancias a la que el modelo se vea sometido y deba realizar sus predicciones, existen actualmente muchos modelos basados en regularización de evaluación implementados por ejemplo en detección de calidad.

Regularización de la norma de Frobenius:

$$\varphi(W) = \lambda||W||_F^2$$

Donde λ es un parámetro de regularización

Regularización de la norma L-21:

$$||W||_{2,1} = \sum_{j=1}^{d} ||W_{j,*}||_2$$

2.6.6. Validación cruzada

La validación cruzada es un método probabilístico de evaluación que consiste en la separación de los datos en datos de entrenamiento y datos de evaluación del modelo, de manera tradicional estos datos son rotados secuencialmente para que cada dato tenga la oportunidad de ser validado en las predicciones del modelo, esta forma tradicional de validación cruzada es denominada validación por N capas. Existen formas más complejas y con mejor

rendimiento como k-fold cross-validation o la validación por rondas repetidas. Esta métrica de evaluación es empleada en extensos corpus de datos para obtener resultados más confiables (Sahli, 2020).

Si bien este es un método de evaluación ampliamente conocido y usado varios estudios han demostrado que tiene numerosas limitaciones como la cuantificación del rendimiento del modelo, identificación de muestras atípicas y un cuestionado rendimiento (Xiong et al., 2020). Algunas de las alternativas que se basan en la validación cruzada y demuestran tener mejores resultados son: validación cruzada hacia adelante (FCV), validación cruzada hacia adelante con exclusión de uno de los casos (LOOFCV) y validación cruzada paso a paso (k-fold-m-step FCV).

Error de validación cruzada K-fold (interacciones):

$$E = \frac{1}{K}\sum_{i=1}^{k} E_i$$

2.6.7. Hiperparametros de ajustes

Todos los algoritmos de aprendizaje automático tienen valores a ser modificados con la finalidad de mejorar el rendimiento del modelo para las predicciones de los valores esperados, estos valores son conocidos como hiperparametros. En independencia del modelo estos hiperparametros pueden ser usados en diferentes estrategias de búsqueda para encontrar a su vez mas hiperparametros, cuya modificación afectara de manera sustancial los resultados del modelo (Palacio-Niño & Berzal, 2019).

Dentro del aprendizaje automático el ajuste de hiperparametros, es la modificación de estos para controlar la etapa de aprendizaje. La cuestión más importante en esta evaluación en la elección del hiperparametro adecuado para los resultados deseados en el modelo. Incógnitas para evaluar como cuantas neuronas o cuantas capas es adecuado poner en una red de aprendizaje automático son los temas tratados en esta métrica de evaluación.

2.7. Técnicas de Machine Learning

2.7.1. Clasificación

La clasificación es una técnica aplicada a una amplia gama de actividades, ya sea de texto, imágenes, datos, direcciones y cualquier información susceptible

de ser segregada en clases. En el mundo de Machine Learning existen una amplia variedad de clasificadores con distinto propósito que sin embargo todo usan el mismo fundamento. Estos clasificadores no son más que elaborados algoritmos o modelos los cuales son maleables mediante parámetros de ajuste, algunos de los ejemplos de algoritmos de clasificación son SVM, DT, Boosted y ANNs (Maxwell et al., 2018).

En la actualidad la elección de un modelo adecuado de clasificación es una tarea que debe ser pensada en miras del propósito deseado. Uno de los inconvenientes más comunes de este tipo de técnicas es excesiva generalización de los datos debido al Overfitting ocurrido durante la etapa de entrenamiento, por lo que los clasificadores son muy susceptibles a ser "demasiado" precisos, lo que los limita al trabajar con información más variada a la acostumbrada.

2.7.2. Regresión

La regresión es una técnica muy usada en el Machine Learning, existiendo modelos de regresión lineal, multivariable y polinomial. Una regresión es una técnica usada especialmente en la previsión y la predicción de la información así también sirve para estudiar la relación entre variables dependientes e independientes. Existen aplicaciones de esta técnica en los datos que incluso permiten el análisis con una colección entera de variables.

La regresión es un concepto con fundamento en las matemáticas que permite la creación de modelos predictivos y relacionales para la elaboración de hipótesis en el análisis de los datos. Tan poderosa herramienta es aplicada muy ampliamente en el aprendizaje automático, siendo de mucha utilidad en su implementación en los corpus de datos, cabe señalar que es una técnica muy sencilla de aplicar y explicar.

2.7.3. Agrupación

Una técnica muy similar a la clasificación es la agrupación que básicamente consiste en la búsqueda de objetos en un especto de unidades que guarden cierta similitud con un objeto inicial u objeto base. Cada grupo debe ser identificado por atributos únicos como en el caso de los textos podrían ser palabra y frases que engloben una idea central del grupo. Es una técnica muy usada en la agrupación de documentos para análisis en estudios, un ejemplo de la implementación de esta técnica es la agrupación difusa de K-means.

Esta técnica es muy común no solamente en la clasificación de documentos, sino también por ejemplo en la clasificación de imágenes, siguiendo el mismo fundamento de la agrupación por atributos claves y únicos. Esta es una implementación mucho más pragmática del clustering, siendo este último más bien una etapa dentro de un proceso más complejo.

2.7.4. Clusterizacion

La clusterización es la agrupación de objetos en un espacio multidimensional. A cada agrupación de se le denomina “Clúster” que es un grupo cuyos objetos integrantes poseen características similares que los diferencian a otros objetos ubicados en otros Clusters. Cabe recalcar que esta agrupación es parte de un algoritmo de aprendizaje por refuerzo debido a que la etiqueta de estos objetos no es conocida inicialmente. La similitud de un objeto con otro es determinada por una función con capacidad multidimensional. La aplicación de clusterización puede llegar a revelar información relevante como patrones en los objetos.

La clusterización es empleada en múltiples propósitos en los cuales ha demostrado tener una alta eficiencia, por ejemplo, algunas de sus implementaciones abarcan desde la recomendación de servicios por internet hasta la agrupación de documentación en categorías relevantes a la necesidad del usuario. Es muy conocida su aplicación es la atención personalizada a los clientes que son segregados en base a sus preferencias y tendencias.

2.7.5. Procesamiento del Lenguaje Natural (NLP)

El procesamiento de lenguaje natural es una ciencia cuya finalidad es hacer nuestro lenguaje humano accesible y entendible a las computadoras. Hoy en día encontramos su aplicación en muchos aspectos de la web, que van desde traductores hasta los clasificadores de correos de nuestra bandeja de entrada que hacen uso del procesamiento del texto. Todos estos logros y muchos más son posibles gracias al uso de sofisticados algoritmos y aplicaciones con fundamentación matemática, lingüística y multidisciplinaria que permiten la elaboración de herramientas que facilitan la navegación de los usuarios en el internet (Eisenstein, 2018).

Los inicios del NLP se remontan a la mitad del siglo XX cuando fueron mencionados como unos de los aspectos en la prueba de Turing. Hasta entonces han existido a lo largo de los años considerables avances en el

perfeccionamiento de esta ciencia junto con la inteligencia artificial. En las últimas décadas si bien no se existen conceptos como tales innovadores, es cierto que el empleo del NLP ha sido muy elevado, llegando a estar presente en asistentes virtuales, analizadores de texto e incluso la industria en general con propósitos de atención personalizada a usuarios (Bouabdallaoui et al., 2020).

Actualmente el Procesamiento de Lenguaje Natural como ciencia tiene dos campos principales en los cuales se centran la mayoría de los profesionales y expertos en tema, que son: áreas centrales y sus aplicaciones. Las áreas centrales son todas aquellas dedicadas a la investigación y desarrollo de soluciones y modelos de NLP que aporten un valor agregado a la sociedad, como el estudio de algoritmos y modelos matemáticos en combinación con lingüística. La aplicación de NLP viene de la mano con el aprendizaje automático para la codificación de los respectivos modelos. Existen muchas técnicas y arquitecturas que con el paso de los años se van perfeccionado gradualmente según las necesidades y limitantes de la época en cuestión (Otter et al., 2021).

El estudio de NLP es de suma importancia en tiempos que la tecnología avanza a la imitación y perfeccionamiento de habilidades intrínsecas al ser humano. Toda esta gama de aplicaciones al NLP en miras al beneficio humano revelan una necesidad urgente e inherente a la era de información: la importación de trabajar los datos. Un conocimiento detallado de NLP es un paso necesario para hacer frente a los desafíos actuales.

2.8. Conceptos Básicos

2.8.1. Word Embedding

Uno de los problemas en el procesamiento del lenguaje natural, es la representación de los textos para los computadores, los cuales solo son capaces de procesar números. En base a esta problemática dentro del NLP existe la técnica del Word Embedding, el cual consiste en hacer una representación vectorial multidimensional con valores numéricos que mediante operaciones de algebra lineal se pueden establecer similitud y cercanía entre palabras. Todo este entramado de operaciones da paso a la creación de vocabularios para el modelo que es la base de procesamiento de los textos (Bouabdallaoui et al., 2020).

2.8.2. Modelos de NLP

Los modelos de NLP son algoritmos cuya principal funcionalidad es la interpretación del lenguaje humano (lenguaje natural) con la capacidad de realizar predicciones, construidos a partir de arquitecturas en aprendizaje automático y aprendizaje profundo con posibilidad a ser evaluados mediante métricas de rendimiento que permiten determinan su efectividad y limitaciones (DeYoung et al., 2020).

2.8.3. Preprocesamiento

Una parte muy importante del NLP es el procesamiento de los datos. Este consiste en la depuración, separación y limpieza de la información, con el objetivo de eliminar todas aquellas fallas y falencias en los corpus de datos que puedan alterar la etapa de entramiento de los modelos. El preprocesamiento de los textos no es más que la corrección de las faltas en las oraciones que a su vez están separadas en palabras, todo esto con técnicas especializadas como por ejemplo NLTK (Natural Language Tool kit) que es muy usada en el lenguaje de programación Python(Ramachandran & Parvathi, 2019).

2.8.4. Lingüística Computacional

Las relaciones entre las ciencias lingüísticas y las ciencias computacionales contienen variados matices aplicativos a ambas disciplinas, ambos términos de la denominación “lingüística computacional” ponen de relieve los propósitos científicos de dicha relación. En concepto es la aplicación de técnicas y modelos postulados mediante formalizaciones del lenguaje humano con el propósito de ser procesable para los programas informáticos, todo esto sin la premisa de la imitación de la mente humana en los procesos lingüísticos (Rodrigo & Bonino, 2019).

La lingüística computacional, también conocida como lingüística de la informática, es un campo que se enfoca en aspecto teóricos para resolver cuestiones centrales de la lingüística en general y sus diferentes ramas y aplicaciones. Esta conceptualización contrasta con el enfoque más práctico y pragmático del procesamiento del lenguaje natural, cuya finalidad es la creación de soluciones a problemas inherentes del lenguaje como la ambigüedad y el contexto.

Si bien es cierto que existen números diferencias sustanciales entre ambas ciencias, también comparte muchos aspectos que en muchos casos se

complementan el uno con otro (procesos de entrenamiento, aprendizaje automático, modelos y lingüística). Es probable que eventualmente en un futuro ambas ciencias sean empleadas con objetivos académicos que impliquen todos los aspectos a resolver de ambas cuestiones.

2.9. Principales retos del NLP

A través de los años el Procesamiento del Lenguaje Natural (NLP) se ha convertido en una herramienta muy utilizada con avances sumamente significativos, ya que es capaz de permitir una comunicación efectiva entre el computador y los seres humanos. Sin embargo, a pesar del tiempo aún sigue existiendo diferentes inconvenientes al momento de usar el NLP ya que la diversidad y la dimensionalidad de los datos pueden hacer que algunos casos la implementación de estos sea un verdadero desafío.

Según (Chowdhary, 2020) explica que diseñar un programa que comprenda el lenguaje natural del ser humano es una tarea difícil. Muchos de los lenguajes naturales son extensos y contienen oraciones largas. Además, el lenguaje natural tiene mucha ambigüedad, palabras y frases con diversos significados según su contexto. Por ello lograr que un programa entienda al ser humano ya que las sintaxis de un idioma a determinar se combinan para crear significados muy amplios.

2.9.1. Ambigüedad

La ambigüedad es uno de los inconvenientes de la interacción hombre-máquina. El lenguaje humano suele ser muy complejo y variado en su estructura, precisión y elaboración de reglas de formación, ocasionando que existan diferentes significados dentro de una misma oración, provocando que la interpretación de la máquina este abierta a fallos (Jusoh, 2018).

2.9.2. Sinonimia

En el NLP otro de los fenómenos que debe enfrentar es la sinonimia la cual es una relación de igualdad entre el significado de dos palabras, ya que estas pueden estar dentro del mismo campo semántico y la terminología puede cambiar entre si según el contexto. Esto ocasiona que la sinonimia sea una tarea compleja al momento de implementar el NLP puesto que si existe una relación

equivalente entre dos o más palabras se deberán evaluar los conceptos para encontrar una semejanza entre ellas (Tovar et al., 2018).

La sinonimia es un reto del NLP puesto que el humano es capaz de expresar diversas situaciones o palabras de diferentes maneras pero que al final mantienen la misma idea, ya que estos poseen una variedad de términos en los cuales depende el contexto específico en el que se esté hablando.

2.9.3. Sintaxis

La sintaxis es otra peculiaridad dentro del Procesamiento de Leguaje Natural, este procedimiento define una estructura a una frase u oración. Anteriormente en los algoritmos se usaban gramáticas libres de contexto como conjunto de instrucciones para el análisis sintáctico, lo que hace que sea difícil especificar un lenguaje completo mediante un grupo de reglas (Dordevic & Stojkovic, 2020).

En el Procesamiento del Lenguaje Natural la sintaxis es uno de sus desafíos al momento de implementarse ya que la sintaxis se encarga de determinar las estructura que tienen diversas reglas las cuales pueden ocasionar confusión e irregularidades en distintos casos. Además, es posible que la sintaxis de una oración o frase pueda restructurarse de una manera distinta.

2.9.4. Correferencia

Es otro de los desafíos que debe atravesar el NLP, la correferencia posee un alcance mucho más extenso y tiene una terminología extralingüística. El término correferencia se refiere a palabras o frases que se relacionan a una única o varias entidades en un entorno operativo. Estas pueden llegar a tener una estructura gramatical y una función completamente distintas en el género y la parte de la oración. Este punto al igual que los demás suele ser crucial en el NLP ya que comprende en la compresión del texto completo (Sukthanker et al., 2020)

2.9.5. Normalización vs Información

En el Procesamiento de Lenguaje Natural se debe normalizar la información, es decir pasar cualquier texto informal a un formato legible por la máquina. Los textos informales que normalmente los seres humanos utilizan son: cantidades de dinero, números, abreviaturas, etc. Estos son fáciles de ser identificados por ellos ya que su cerebro esta entrenado, sin embargo, la máquina no comprende estos y es por ello, por lo que la información debe ser normalizada es decir pasarla a un formato de cadena o texto plano para que esta tenga una predicción más exacta. Cuando se normaliza se suele perder una pequeña parte de la

información a cambio de generalizar de mejor manera. Lo cual es común en el Procesamiento del Lenguaje Natural (Rahate & Chandak, 2019).

2.10. Técnicas Clásicas de NLP

2.10.1. Text2Dec

Una de las técnicas mayor apogeo en el procesamiento del lenguaje natural y con mayor facilidad implementada es la conversión de texto a numérico (Text2Dec) la cual consiste en una solución teórica y con aplicaciones prácticas para la extracción de decisiones de dependencias entre las palabras. La implementación más común de esta técnica es en el lenguaje de programación Python en conjunto con otras técnicas de NLP de forma secuencial (Etikala, 2021).

La técnica de Text2Dec está dividida en tres etapas. La selección del texto, que no es más que la elección o creación del corpus de datos que servirán como datos de entrada para los modelos de NLP. Posteriormente a eso viene la etapa del procesamiento de los datos y extracción de la información. finalmente, la creación del modelo conversacional que a su vez se divide en dos subetapas: construcción de tablas y constructor de DRD.

2.10.2. Tokenización

La tokenización se describe como la etapa en la cual el texto es dividido en la unidad más simple de información. En la mayoría de los casos esta unidad mínima es la palabra. Esta división se la realiza mediante el establecimiento de los delimitadores que comúnmente suelen ser los espacios en blanco, una vez hecho esto se hace un proceso de indexación de las palabras con una representación numérica (Reese, 2018).

Comúnmente la tokenización es solo un proceso dentro de un conjunto de fases a un propósito mucho mayor. La información resultante de esta etapa puede ser usada para múltiples propósitos como tareas simples, procesos de búsquedas sencillas, detección de oraciones, clasificación de textos e identificación y sustratos de discursos. Uno de los usos de la tokenización corresponde al Análisis sintáctico superficial cuya finalidad es agrupar los tokens en bloques que representan estructuras gramaticales (Celi-Parraga et al., 2021).

2.10.3. Expresiones regulares

Las expresiones regulares o también conocidas con su abreviatura Regex se definen como un mecanismo para buscar y comparar patrones. Las expresiones regulares son muy utilizadas en el campo de la informática, lenguajes de programación, el NLP y bases de datos ya que cuentan con una gran precisión y eficacia (Y. Li et al., 2021).

Las expresiones regulares son la representación de una cadena de texto de manera abreviada que se ajusta a un patrón específico. Estas expresiones sirven para analizar una secuencia de caracteres e identificar que cumplan con cierta regla gramatical.

2.10.4. NER (Named Entity Recognition)

El reconocimiento de entidades con nombre o NER por sus siglas en ingles es una técnica de NLP que consiste en identificar todo aquello que pueda ser catalogado como persona, organización y ubicación para denotarlo como una entidad. Algunos de los ejemplos de entidades pueden ser fechas, ubicaciones geográficas e incluso nombres de monedas. Esta clasificación e identificación es sujeta a personalización según las necesidades del caso (Shelar et al., 2020).

La técnica NER es muy usada en modelos para identificar y clasificar entidades que pueden ser usados en muchos propósitos. Desde los auspiciantes de modas en la web hasta los proveedores de noticias así también las recomendaciones en base a comentarios y la atención al cliente. Cada propósito requiero de una clasificación diferente, sin embargo, en términos generales esta herramienta ha demostrado ser muy útil.

2.10.5. Part-of-speech (POS)

El etiquetador POS es un sistema computacional que toma una oración como entrada de su sistema y etiqueta cada palabra como salida, produciendo así una etiquetación detallada de cada elemento del texto analizado. Una falencia o desafío con el que se enfrenta esta técnica es la ambigüedad de las palabras que hace que el significado varie según el contexto además del hecho de que una palabra no representa por si sola el sentido gramatical de un texto entero (Warjri et al., 2018).

Actualmente esta técnica es perfeccionada ya tomando en cuenta el contexto gramatical de las palabras dentro del texto también tomando en cuenta su definición como tal. Se puede asemejar al proceso de aprendizaje de los infantes

que identificas los elementos de la oración para posteriormente ordenarlos semánticamente y darles un propósito textual significativo.

2.10.6. Stopwords

Para el análisis de texto en el Procesamiento de Lenguaje Natural es necesario la eliminación de aquellas palabras que no son informativas o stopwords como normalmente se las conoce. Este procedimiento ayuda a garantizar la precisión y la eficacia de las tareas NLP ya sea en la indexación, modelado de temas y recuperación de la información (Sarica & Luo, 2020).

2.10.7. Lematización

En el Procesamiento de Lenguaje Natural se realiza la lematización como un proceso para unir las partes flexionadas de una palabra, esto se realiza para que esta pueda ser identificada como un solo elemento llamado lema de la palabra o en su forma léxica. Este proceso agrega significado a una expresión en particular, uniendo el texto en una palabra simple (Khyani et al., 2020).

2.10.8. Representación en bolsa de palabras

Una de las formas más comunes de representación del texto para técnicas de NLP son las bolsas de palabras (BOW por sus siglas en ingles), que consiste en el tratamiento de las palabras como una colección desordenada de las mismas mediante el uso de un vector multidimensional de longitud fija que contiene un conteo de todas las concurrencias de las palabras por separado (Walkowiak et al., 2019).

Actualmente algunas variaciones de esta técnica están disponibles que potencian su capacidad inicial en la representación de las palabras, un ejemplo de esto es la capacidad de reducir el total de las palabras contabilizadas hacia una única representación genérica de las mismas, así también la asignación de pesos para relaciones gramaticales.

2.11. Modelos probabilísticos

2.11.1. Modelos Discretos

Dentro de los modelos probabilístico se tiene a los discretos los cuales se emplean dentro de sistemas donde la alteración de un estado se da de forma instantánea en puntos aleatorios o especificados en el tiempo, esto obedece a la

aparición de eventos discretos (Pulido-Rojano et al., 2021). A continuación, algunos de los modelos discretos más comunes.

2.11.2. Ensayos de Bernoulli

Es un modelo probabilístico que se basa en la eventualidad de un ensayo de x experimento con solamente dos posibles resultados: ausencia o presencia del evento y estos resultados son mutuamente excluyentes. La distribución de este modelo con resultados aleatorios es determinada por un parámetro p que es la probabilidad de que el evento ocurra o no. Un ejemplo muy común de este modelo discreto es el lanzamiento de una moneda cuyo resultado poder únicamente un lado de la moneda, que automáticamente excluye el otro resultado (Palmer & Jiménez, 2022).

La probabilidad de éxito es:

$$P(E) = p$$

Y la probabilidad de fracaso es:

$$P(F) = q$$

Donde

$$q = 1 - p$$

2.11.3. Distribución Binomial

Si bien el modelado de Bernoulli aplica únicamente a un ensayo experimental, la distribución binomial tiene un enfoque más extendido aplicándose a un conjunto de ensayos de Bernoulli y con un resultado igualmente discreto pudiendo tener únicamente dos posibilidades mutuamente excluyentes: todos los experimentos han sido un fracaso o todos los experimentos han sido un éxito (Palmer & Jiménez, 2022).

La probabilidad de tener r éxitos en n intentos se determina en el siguiente modelo

$$P(X = x) = \binom{n}{k} p^k (1-p)^{n-k} = \frac{n!}{k!\,(n-k)!} p^k (1-p)^{n-k}$$

2.12. Modelos Continuos

2.12.1. Distribución uniforme

Esta distribución o modelo es catalogada como una de los más sencillos en su clasificación. Se define como la representación de un escenario en el que todos los resultados de un rango comprendido entre valores mínimos y máximos

poseen la misma probabilidad. Este modelo tiene muchas aplicaciones como el análisis de riesgos y generación de muestras aleatorias (González-Hernández et al., 2020).

La función de distribución uniforme continua viene dada por la siguiente representación:

$$Fx(x) = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & a \leq x < b \\ 1 & x \geq b \end{cases}$$

2.12.2. Distribución Normal

La distribución normal es conocida como campana de Gauss cuya principal característica es la simetría respecto a su media y la pronunciada caída al valor cero en ambos extremos. Se puede deducir de esta particularidad que los valores atípicos son muy poco probables de aparecer en esta distribución por lo que esencialmente este modelo probabilístico normaliza los valores sin tener extremos que alteren la simetría inicialmente mencionada (Fontanelli Espinoza et al., 2020).

La distribución normal se representa de la siguiente forma:

$$f(x) = \frac{1}{\sqrt{2\pi^2}} e^{\frac{-(z-\mu)^2}{2\sigma^2}}$$

2.12.3. Teorema central del limite

Uno de los pilares fundamentales de la teoría de la probabilidad es el teorema central del límite, cuya aplicación se dedica a la investigación del comportamiento en de extensas variables aleatorias disponiendo así la convergencia de estas a una distribución normalizada. Algunas de las aplicaciones más comunes de este teorema se dan en la inferencia estadística, suma de variables aleatorias y aproximación de distribuciones (Oa & Dieser, 2020). Según (Cepeda, 2018) Sean X1, X2, ..., Xn variables aleatorias independientes distribuidas idénticamente con E(Xi) = μ y V ar(Xi) = σ2, σ2 < ∞. Entonces,

$$Z_n = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n\sigma}}$$

2.12.4. Modelos N-gramas

Usado principal en la predicción de palabras, este modelo de NLP se basa principalmente en la N palabra anteriores para realizar la predicción de la siguiente palabra, donde N en si determina el número de probabilidades de ser

estimada. Este tipo de modelo son muy usados en la corrección ortográfica y reconocimiento del habla (Hamarashid et al., 2021).

El funcionamiento de este tipo de modelo se basa en probabilidades que sea asignada a cada palabra de una oración o texto para que sea predicha en el modelo. Otra forma de funcionamiento en otro escenario es la probabilidad de predicción de una palabra basándose en la concurrencia de las palabras anteriores en la oración, lo que con un producto de la probabilidad de cada palabra con la de su concurrencia se puede determinar cuál será la siguiente en aparecer.

2.12.5. Modelo Unigrama

El modelo más simple de los n-gramas es el unigrama (debido a que su cardinalidad n es igual a 1), posee así mismo una funcionalidad de predecir la siguiente palabra en base a la secuencia de n palabras. El modelo unigrama tiene la capacidad de sin tener en cuenta palabras anteriores, calcular la probabilidad de una siguiente palabra y así mismo asignar una probabilidad a una secuencia entera. Estas probabilidades pueden ser estimadas desde el corpus de datos de entrenamiento (Hariyanti et al., 2019).

Este modelo de predicción es muy útil debido a su relativa sencilles en comparación a otros modelos de la rama de los n-gramas. Cuando el modelo se topa con una palabra que no existe en el corpus de dato lo trata como una etiqueta “None”. En algunas aplicaciones específicas el modelo puede llegar a ser muy deficiente con cálculos muy alejados de los razonables.

2.12.6. Modelo Bigrama

Si bien el unigrama hacia predicciones independientemente de las secuencias anteriores, el modelo bigrama si hace uso de estas mediante las etiquetas, las cuales asigna a la palabra predicha en cuestión. Este uso de las etiquetas permite al modelo maneja una especie de contexto basado en probabilidades que le permite tener una mayor eficiencia que el unigrama. La probabilidad de las etiquetas de ver afectada por etiqueta predecesoras (Kanakaraddi & Nandyal, 2018).

En términos generales para casos más amplios el modelo bigrama realiza un mejor trabajo que el modelo unigrama debido al uso de un “contexto” que permite asignar correctamente las probabilidades, esto se puede ver en los casos en que

el unigrama no diferencia la importancia del orden semántico de las palabras en una oración e interpreta cualquier orden como válido.

2.12.7. Modelo Trigrama

El modelo trigrama sigue la misma filosofía que los n-grama anteriores, teniendo un conteo consecutivo de concurrencias anteriores representados en N, por lo que es de esperar que en este caso la cantidad de palabras predecesoras a ser tomadas en cuenta para el cálculo probabilístico sea 3, así mismo son estas tres palabras las que pueden ser consideradas como el contexto que permite realizar la predicción de la palabra requerida en cuestión (Lady M. Sangacha Tapia et al., 2021).

2.13. Modelos Lógicos

2.13.1. Modelo de Márkov Oculto (MMO)

Este es un modelo también catalogado como probabilístico por algunas fuentes que es comúnmente usado para la captura de secuencias con estados ocultos. Tiene una especial aplicación en la representación de un conjunto de variables aleatorias. Es un modelo aplicado de forma muy diversa en diferentes campos como la informática, ingeniería y el procesamiento de lenguaje natural, en este último ha sido de gran utilidad en la implementación de técnica tales como reconocimiento del habla, POS (part-of-speech) y la extracción de la información (Leopold et al., 2019).

2.14. Modelos de NLP

2.14.1. BERT

La unidireccionalidad de algunos modelos presenta limitantes en su aprovechamiento de los recursos como oraciones y contexto de palabras, como es el caso de modelos en los cuales es necesario todo el contenido de una oración (tanto izquierdo y derecho en relación con la palabra actual) para un óptimo entendimiento de esta. Una solución a esto son los modelos bidireccionales basado en la arquitectura transformer, el ejemplo más representativo de estos es el modelo BERT (Gillioz et al., 2020).

Este tipo de modelos tiene la capacidad de hacer fusión de contexto de las oraciones lo que se traduce en una representación bidireccional que permite una

mejor extracción del contexto para tareas de razonamiento. Este modelo es entrenado en sus debidas fases con algoritmos nos supervisados de aprendizaje automático a lo que sumado la bidireccionalidad permite que cada palabra se analice a sí misma en si para realizar predicciones de las siguientes representaciones de palabras.

2.14.2. GPT

Este modelo es uno de los más llamativos debido a su alta capacidad de procesamiento demostrada en los últimos años. Se puede ver como una generación de modelos creado de forma sucesiva que suplen y mejoran las necesidades de cada antecesor, de allí que existan a la fecha el GPT, GPT-2 y GPT-3 siendo este ultimo la versión más popular de las tres. GPT por sus siglas en inglés (Generative pre-trained transformer) es un modelo basado en transformadores autorregresivo que han sido aplicados en el procesamiento del lenguaje natural y se han vuelto muy relevantes en esta disciplina (Chiu et al., 2021).

El modelo GPT hace uso de una estructura con decodificadores, al no tener su contraparte como codificadores, lo componente que está en relación con el codificador son removido simplificando así la estructura del modelo basado en transformer. El modelo GPT cuenta en su estructura con 12 bloques cada uno con sus respectivos parámetros que permite la salida de probabilidades que sirve alimentación a los siguientes bloques (Zheng et al., 2021).

2.14.3. ELMO

Este modelo llamado así por sus siglas en inglés (Embeddings from language models) fue uno de los primeros en mostrar significativos avances en los problemas de la comprensión del lenguaje natural. Su fundamento teórico se basa en que cada palabra debe incorporar características de esta a nivel de la semántica contextual. Cada capa del modelo se reduce a un único vector manejando por pesos en las tareas de capas (Ulčar & Robnik-ˇ, 2021).

ELMO es un claro ejemplo de los modelos de última generación pre-entrenados de situación contextual. Implementado con algoritmos de redes neuronales. El manejo de las palabras se puede realizar en independencia del contexto en cada capa por lo que cada palabra recibe su propio Embeddings. Este tipo de modelos son construidos para estado bidireccionales de entrenamiento.

2.14.4. Comparación de los modelos de procesamiento de lenguaje natural

Tabla 7:

Matriz de Comparación de Modelos NLP

	GPT	BERT	ELMO
Procesamiento de texto	Autorregresivo - Unidireccional	Bidireccional	Bidireccional – capas múltiples
Segmentación	Decoder	Encoder	Word Vectors
Algoritmo	Supervisado	No supervisado	No supervisado
Uso Comunes	Aplicaciones embebidas, sitios web, artículos y documentos	Documentos, documentos en la web	Q&A, Análisis de sentimientos, resolución de conferencias

Nota: Matriz de Comparación de Modelos NLP. Elaboración: Jorge Alberto Oviedo Peñafiel e Inés Janellys Fajardo Romero. Fuente: Información tomada de (Gillioz et al., 2020), (Chiu et al., 2021), (Ulčar & Robnik-ˇ, 2021) . Elaborada por los autores

2.14.5. NLP – Workflow

Los Workflow o flujos de trabajos son procesos de automatización en el cual se realizan un conjunto de instrucciones que van interactuando una atrás de la otra de acuerdo con un conjunto de reglas.

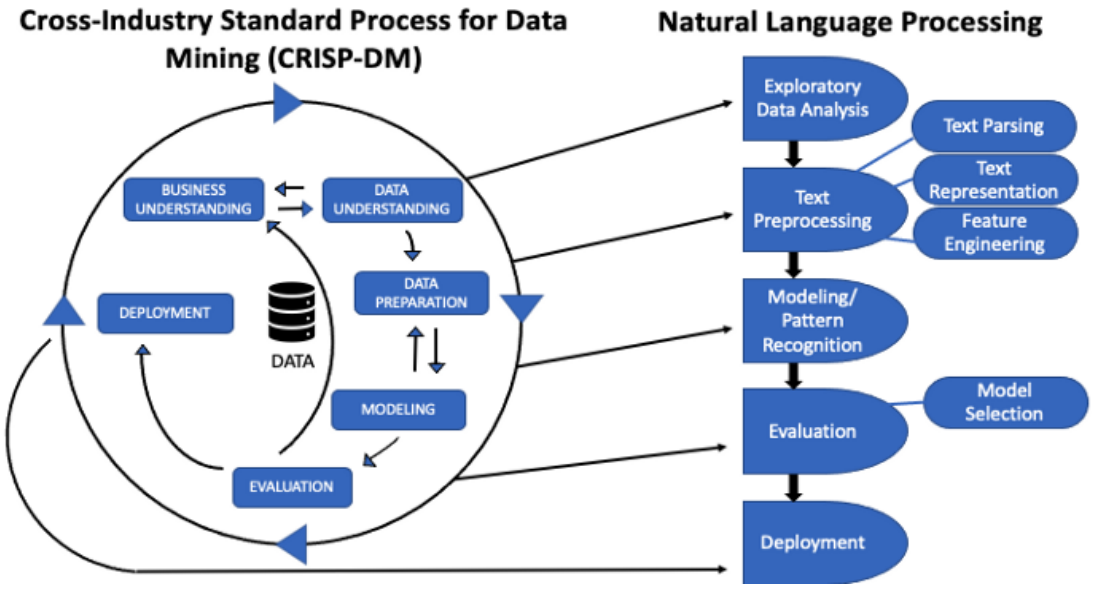
En NLP se requieren de múltiples pasos o tareas de procesamiento, lo cuales se incluyen en un flujo de trabajo o proceso de NLP. Este se emplea ya que el procesamiento de la información se realiza en grandes volúmenes y se realiza una serie de tareas como: la identificación de la lengua, la detección de frases, la segmentación de frases y el análisis sintáctico y dependencias, lo cual está sujeto a fallos que deberían reajustarse o recuperarse. Es por ello, por lo que los desarrolladores y usuarios finales debe especificar el flujo de trabajo (Gudivada & Arbabifard, 2018).

Según los estudios de (Landolt et al., 2021) comenta que el modelo de flujo de trabajo estándar utilizado en proyectos de NLP suele basarse en el modelo CRISP-DM, este modelo permite incrementar la tasa de respuesta y reducir los tiempos de costos. Este marco de trabajo tiene directrices uniformes para los

mineros de datos. Por lo tanto, para resolver cualquier inconveniente basado en NLP se utiliza un flujo metódico el cual cuenta con una secuencia de pasos. Los principales pasos son lo siguiente:

Figura 2:

Workflow Procesamiento de Lenguaje Natural



Nota: El siguiente gráfico se visualiza el modelo CRISP-DM, este flujo de trabajo es en el que se basa el NLP para llevar el procesamiento de la información (p. 8), por **Referencia**

2.14.6. Datos de Texto Exploratorios NLP - Text EDA & Clustering

El análisis exploratorio de datos más conocido como EDA es un proceso utilizado por los científicos de datos para analizar un conjunto de datos y comprender su naturaleza y propiedades para así de esta manera extraer las ideas principales de estos. La EDA se aplica de manera interactiva varias operaciones de análisis como: el filtrado la agregación y visualización; los resultados de cada operación son validados por el usuario. El análisis exploratorio es un proceso que requiere mucho tiempo ya que es complejo en especial para usuarios con poca experiencia, teniendo en cuenta aquello se han creado múltiples líneas de trabajo para que el proceso sea más fácil en múltiples dimensiones (Milo & Somech, 2020).

El objetivo principal de EDA es sugerir una hipótesis acerca de la causa de los fenómenos observados y evaluarlos para la recopilación de datos. El resultado obtenido de EDA se visualiza a través de gráficos como: diagramas, histogramas, escalamiento multidimensional, etc. Es usado normalmente en el

NLP en el preprocesamiento para obtener faltantes y obtener las correlaciones de datos o variables, ya que esta realiza una limpieza de los datos mejorando la calidad y reduciendo la cantidad haciendo el proceso mucho más eficiente (Maylawati et al., 2020).

Como lo explica (Hidayat et al., 2020) en su estudio la EDA utiliza técnicas de minería de datos las cuales pueden ser bien de aprendizaje supervisado (clasificación), el aprendizaje no supervisado (Clustering) o también el aprendizaje semi-supervisado (clasificación y Clustering). El Clustering es un procedimiento que puede producir más conocimiento de información. Los algoritmos de Clustering que se pueden utilizar son: K-Means, KMeans++, MiniBatch K-Means, y MiniBatch K-Means++. Al hacer uso de estos algoritmos se puede evaluar los datos y obtener visión del conocimiento de la información Haga clic o pulse aquí para escribir texto.(Hidayat et al., 2020) .

2.14.7. Preprocesamiento de Texto

Este paso es esencial en cualquier sistema de NLP ya que tanto como caracteres, palabras y oraciones que se identifiquen son fundamentales para todas las etapas que vienen a continuación. En esta etapa mediante un conjunto de tareas se procesa los corpus de texto ya que suelen tener caracteres especiales, formatos de números, fechas y palabras que no sean útiles en la minería de texto, por ende, pueden ser eliminados (Hickman et al., 2022).

2.14.8. Modelado/Reconocimiento de Patrones - Modeling/Pattern Recognition

En esta etapa se escoge y se utilizan distintas técnicas de modelado o algoritmos, que se ajustan a los parámetros para obtener los valores óptimos. Existen diferentes técnicas para un mismo problema de minería de datos, sin embargo, hay que tener en cuenta que cada uno de estos modelos pueden tener diferentes requisitos al momento de presentarse los datos de entrada, por lo cual es probable que se deban realizar pasos adicionales de preparación de los datos (Schröer et al., 2021).

2.14.9. Evaluación/Evaluation

Como lo indica (Espinosa-Zúñiga, 2020) en su estudio, después de haber construido el modelo que represente la más alta calidad en términos de análisis de datos, se evalúa y a analiza cada uno de los pasos posteriores antes de continuar con la implementación y ejecución.

En esta fase se comienza a evaluar todos los procedimientos realizados con anterioridad, y de esta manera poder garantizar que el modelo que se ha construido cumpla con todos los aspectos y objetivos establecidos al inicio del proyecto, en esta etapa se decide si continua el proceso o debe ir nuevamente a una etapa anterior.

2.14.10. Despliegue/Deployment

Esta es la fase final en la cual el modelo ha sido creado y evaluado de manera positiva, permitiendo así el paso al despliegue y que los usuarios puedan hacer uso de este. Una vez realizada esta fase se tiene constante supervisión y mantenimiento de este (Schröer et al., 2021).

2.14.11. Algoritmos de STEMMING

STEMMING es un conjunto de palabras que se reducen a su stem o raíz común. Para implementar stemming se puede realizar mediante algoritmos con reglas gramaticales de origen morfológico para un idioma en cuestión, o mediante diccionarios informatizados que asocian a cada forma con una palabra representativa (González, 2018).

2.14.12. Lovin's Stemmer

Según (Haroon, 2018) explica que este algoritmo fue propuesto por Lovins, es usado para el análisis de la lengua inglesa. Este algoritmo ejecuta dos pasos que son:

- La subcadena más larga posible es recortada del lado derecho y por coincidencia de la lista de terminaciones se almacena en el sistema.
- Las fallas ortográficas se las trata como excepciones. En este paso suele haber ciertas complicaciones, que se resuelven aplicando algunas técnicas post-stemming como la concordancia parcial o la grabación.

Este algoritmo tiene como ventaja la velocidad y la posibilidad de eliminar letras dobles y los plurales irregulares. Sin embargo, también tiene como desventaja que dentro del listado de terminaciones no se encuentren todos los sufijos ocasionando que sea inestable y falle al formar palabras a raíz de los stem.

2.14.13. Porter'S Stemmer

Como lo explica (Gupta & Jivani, 2018) el algoritmo Porter corrobora la existencia de los sufijos dentro de una palabra y la compara con los sufijos que ya están almacenados. Este algoritmo tiene cinco pasos a seguir los cuales son:

1. Reconoce los sufijos y los recodifica generando un token raíz

2. Cambia las “y” por la letra “i” dado el caso que exista otra vocal en la raíz.
3. Elimina los sufijos dobles y recodifica en base a la lista de sufijos dados.
4. Elimina las terminaciones “-ance”, “-er”, “-able”, "-ant", "-iti" de la palabra principal.
5. Elimina la terminación “-e” de la palabra principal.

El algoritmo Porter es muy cuidadoso al eliminar palabras y evita la eliminación si la longitud de la palabra es demasiado corta.

2.14.14. Paice & Husk’s Stemmer

Este generador de raíz itera a través de los dos pasos del algoritmo. Según (Haroon, 2018) dice que el primer paso elimina o reemplaza las terminaciones de las cadenas y el segundo paso vuelve a escribir las reglas. Hay un archivo separado que enumera las finalizaciones en las que puede buscar y tomar medidas adicionales. Todo algoritmo tiene que detenerse en algún punto. Dado que es un algoritmo iterativo, también tiene algunos estados de terminación de la siguiente manera:

- Palabras que comienzan con una vocal y tienen solo dos consonantes.
- Una palabra que comienza con una consonante y tiene solo tres caracteres.

Este algoritmo tiene como desventaja que es relativamente lento porque las iteraciones a veces conducen a raíces excesivas.

2.14.15. Dawson

El algoritmo Stemmer de Dawson es una versión más extensa del algoritmo Stemmer de Lovins, con la diferencia de que este tiene una lista mucho más completa de sufijos. Este Stemmer es un algoritmo de una sola pasada por lo cual es rápido. Los sufijos se guardan y organizan en orden inverso, indexados por la longitud y la letra final. Dawson no usa encriptación en su algoritmo para tratar con la raíz, sino que usa una extensión del proceso de coincidencia parcial también definido en la raíz de Lovins. El fundamento del algoritmo de Dawson es que dos tallos tienen la misma forma y si estas coinciden con un cierto número de caracteres, y los caracteres restantes de cada miembro pertenecen a la misma clase final. La ventaja es que contiene más sufijos que Lovins y se ejecuta más rápido. La desventaja es que es muy complejo y carece de una implementación estándar reutilizable (Sumamo, 2018).

2.14.16. N-Gram Stemmer

Como explica (Memon et al., 2020) en su estudio es que este algoritmo tiene un enfoque muy interesante y es independiente del lenguaje. En este caso, el método de similitud de cadenas se usa para reemplazar la extensión de la palabra con su raíz. N-gram es una secuencia de n caracteres extraídos de un fragmento de texto continuo. Específicamente, n-gram es un conjunto de n caracteres consecutivos extraídos de una palabra. La idea principal detrás de este método es que palabras similares tendrán una gran proporción de n-gramas. Si n es 2 o 3, las palabras extraídas se denominan diagramas o trigramas, respectivamente. Este algoritmo tiene como ventaja que es independiente del idioma y por ello puede ser utilizado en muchas aplicaciones. Sin embargo, tiene como desventaja el alto consumo de memoria y almacenamiento al momento de almacenar los n-gramas por lo cual hace que este método no sea muy práctico.

2.15. Datos y Códigos

2.15.1. Kaggle Dataset

Kaggle es una plataforma de Google que fomenta la competencia de expertos en todos los niveles de Machine Learning, además de contener información y retos en la ciencia de los datos. Kaggle se caracteriza por almacenar un gran número de corpus de datos de diferente naturaleza en la nube que está a disposición de todos aquellos que lo necesiten, sobre estos mismos datos disponibles también poseen un amplia y activa comunidad que se encargan de actualizarlos constantemente (Quaranta et al., 2021).

El conjunto de datos ofrecidos por la plataforma Kaggle es uno de los tantos servicios disponibles en esta. Consta de un ambiente de colaboración para la creación de modelos que sean usados en la web, esta creación de soluciones se ve impulsada por los retos de competencia creados por la plataforma. Kaggle también cuenta con material informativo de como emplear los datasets, además de proporcional kernel de funcionamiento que puedan ser aplicado a los modelos construidos en las competencias.

2.15.2. Dataset Search

Data Search es un proyecto de Google el cual consiste en un motor de búsqueda especializado en datasets que facilita la labor de encontrar estos corpus en la

web, con una funcionalidad parecida a la de Google scholar. La búsqueda de estos datasets se realiza mediante la inspección de los metadatos embebidos de los proveedores de datos, que de manera estructurada es la mejor herramienta que el motor de búsqueda tiene para encontrar estos datos (Canino, 2019).

Este motor de búsqueda se convierte en una herramienta muy útil para todos aquellos expertos en la ciencia de datos e inteligencia artificial que se encuentre en la búsqueda de datos para los modelos. Google data Search provee resultados muchos más específicos y especializados para la búsqueda de recursos de datos. A diferencia de otros motores de esta naturaleza Google data Search es capaz de presentar muchos más formatos en base a las necesidades requeridas.

2.15.3. Papers with code

Papers with code es una plataforma de iniciativa académica la cual se caracteriza por ser un repositorio de artículos científicos que están enlazados con su código correspondiente en la plataforma de GitHub. La mayoría de los sitios de esta naturaleza poseen sus propios medios de almacenamiento de códigos (SOTA). esta plataforma nace de la necesidad de que experto en su mayoría del aprendizaje profundo compartan las fuentes técnicas de sus trabajos y desarrollos científicos (Wattanakriengkrai et al., 2022).

Papers with code como la mayoría de las plataformas de este tipo posee múltiples enlaces de acceso público que ayuda de mucho en los resultados a las búsquedas científicas. Estas iniciativas de la comunidad se ven apoyada por otras plataformas de artículos académicos como arXiv que se dedica también a compartir todos los repositorios de código de los Papers presentes en su almacenamiento.

2.15.4. Big Bad NLP Database

Esto es un catálogo de recursos disponibles para corpus de datos cuya finalidad está enfocada en el uso de estos en tareas del procesamiento del lenguaje natural. Contando con más de 800 enlaces esta plataforma pone a disposición las publicaciones científicas relacionas con los datasets. La mayoría de la información está en el idioma inglés, sin embargo, también hay información disponible en otros idiomas como el ruso, chino, alemán e indonesio (Сикулер, 2021).

Este en un sitio de parada obligatoria para todos aquellos que quieran mejorar sus habilidades en tareas de desarrollo de modelos de NLP, existen números corpus de datos de variada naturaleza como clasificación de documentos, captura de diálogos, modelado de lenguaje, traducción de textos, preguntas y respuesta, subtitulación de imágenes y corpus en clustering.

2.15.5. Awesome NLP

Una de las fuentes de aprendizaje empírico y recursos para desarrollo de modelos son todos aquellos repositorios alojados en el sitio GitHub, un alojamiento de información y códigos fuentes con libertad para todos los usuarios de disponer material al público. Awesome-NLP es uno de los más útiles y populares de estos repositorios, ya que provee referencias a números recursos enfocados principalmente a la ciencia del procesamiento del lenguaje natural (S. Wang et al., 2018).

Dentro de este repositorio de acceso público hay una variedad muy amplia de librerías en diferentes lenguajes de programación, además de sumarse a esto que el material disponible está en múltiples idiomas que van desde el principal que es el inglés hasta otros menos conocidos como el coreano, chino, alemán y danés.

2.15.6. SQuAD

Uno de los datasets más conocido es SQuAD, el corpus de datos de preguntas y respuesta de la Universidad de Standford. Este es uno de los más grandes datasets conformado por un gran conjunto de artículos de la Wikipedia los cuales en total contiene más 100000 preguntas hechas por personas asiduas a la página, las respuestas a estas preguntas están remarcadas en el documento mismo del Dataset, convirtiendo así a este corpus como uno de los más útiles la ciencia del NLP (C. H. Li et al., 2018).

El Dataset ha sido modificado constantemente por los mismos autores en cuantos a temas de partición de datos, manteniendo hasta la fecha un orden de 80% para entrenamiento, un 10% para desarrollo y un 10% para la evaluación de modelos. Si bien la mayoría de los datos están accesibles al público en general, los autores no se han dispuesto a liberar los datos de evaluación hasta el momento.

2.15.7. Amazon Product Reviews

Muchas de las empresas de e-commerce actualmente poseen mecanismos para garantizar la productividad de sus actividades, el caso de la mundialmente conocida Amazon se basa en los reviews (o valoraciones de los usuarios) que consiste en una opinión acompañada de una puntuación y otros datos. Estos datos accesibles a través de un api se cuentan por más de 142 billones de reviews siendo uno de los corpus más grandes hasta el momento para tareas de NLP (Shrestha & Nasoz, 2019).

Cada uno de los registros de los reviews de los datos de producto de Amazon contienen campos como id del producto, id del usuario, hora, fecha, comentarios, etc. Estos campos son de utilidad para varios propósitos, sin embargo, en la mayoría de los casos para el desarrollo de modelos de NLP esta data es preprocesada mediante diferentes técnicas como la conversión a vectores de longitud fija y la tokenización.

2.15.8. Movie Review Dataset

Muchas de las opiniones emitidas por las personas se basan en la experiencia de cierto producto o servicio, las películas son un medio por el cual los usuarios son capaces de expresar opiniones de diferente índole debido al efecto social que pueden llegar a tener. Las opiniones en muchos casos pueden ser clasificadas en diferentes categorías como buenas, malas, discriminadoras, ofensivas, etc. El uso de plataformas de valoración como IMDB, Rotten Tomatoes o Netflix permiten acceder a grandes corpus de datos sobre las valoraciones de las películas y así aplicarlas a modelos de NLP (Bandana, 2018). Uno de los ejemplos más conocidos de este tipo de Dataset es Large Movie Review Dataset es cual es un Dataset de clasificación binaria de sentimientos con más de 25000 reviews de opiniones polarizadas para entrenamiento de modelos de NLP, así mismo cuenta con la misma cantidad para los datos de evaluación. Este corpus tiene a su disposición datos no etiquetados para variados propósitos de quien lo use además que ya poseer técnicas de preprocesamiento del texto aplicado a los datos como es BOW (Bag of words).

2.15.9. Yelp Dataset Challenge

Este es un reto para los estudiantes que consiste en hacer uso del Dataset de Yelp para buscar formas nuevas de sacar provecho de esta información mediante la investigación científica. El Dataset se cataloga por rounds los cuales

tienen un conjunto de datos necesarios según el año, por ejemplo, el round 13 tiene 6.68 millones de reviews de negocios de todo tipo presentes en la plataforma (Rafay et al., 2020).

El Dataset usado para el reto está conformado por entidades como negocios, usuarios, miembros, valoraciones, comentarios que por lo general se remiten a las opiniones de los usuarios en los servicios brindados por las empresas. Todos estos datos se encuentran en formato JSON y cada archivo posee diferentes atributos dependiendo de la estrategia de los negocios relacionados.

2.16. Web Scraping: Extracción de Texto

El Web Scraping es el proceso de usar bots para la recopilación de datos contenidos en páginas web mediante técnicas automatizadas. Lo que distingue al Web Scraping es que, a simple vista, los datos parecen no estructurados. Por lo tanto, el científico de datos debe determinar el patrón que siguen los datos, para que luego se construya y se ejecute un algoritmo para descomprimir y procesar los mismos (López, 2018) .

Con el proceso de Web Scraping se puede leer textos de una página web para extraer la información y guardarla, se puede decir que es algo muy similar al proceso automatizado de copiar y pegar.

2.16.1. Ética en Web Scraping

La ética en Web Scraping es un tema que se ha tomado en cuenta de manera muy limitada a pesar de que en tribunales y documentos legales se habla acerca del este. El Web Scraping puede causar daños y perjuicio a entidades sensibles que están asociadas a un sitio web en concreto, tanto como propietarios y clientes de estos (Krotov, 2018).

Según (Krotov et al., 2020) a continuación nos indica algunas de las posibles consecuencias perjudiciales del Web Scraping.

2.16.2. Privacidad individual y derechos del sujeto de la investigación

Los datos que son recolectado en los sitios web en ocasiones pueden violar de manera inadvertida la privacidad de las personas que usan los diferentes servicios que brindan los sitios web, además no solo puede violar la privacidad

sino también pueden hacer uso de los datos por terceras personas sin el consentimiento de los clientes y por lo tanto violentar su derecho a la privacidad.

2.16.3. Privacidad organizativa y secretos comerciales

Las búsquedas en la web automatizada pueden revelar inadvertidamente información confidencial sobre las actividades de una organización. Es decir, pueden rastrear y contar automáticamente los trabajos en un sitio y puede obtener una idea general de la audiencia objetivo, la participación de mercado y los ingresos del sitio.

2.16.4. Valor decreciente para la organización

Si una persona accede al sitio web a través de una interfaz web no estándar, evitará los anuncios que el sitio web utiliza para monetizar el contenido. Estos pueden dar lugar a la creación de productos de datos que el cliente no puede permitirse comprar al propietario original de los datos sin violar la ley de derechos de autor.

2.16.5. Discriminación y prejuicios

La información de las búsquedas en la web puede conducir a prácticas discriminatorias, conclusiones basadas en prejuicios y etiquetado dañino estos pueden ser tanto social, como financiero. Los investigadores deben anticipar y evitar estos usos potenciales de los datos de la red.

2.16.6. Calidad de los datos e impacto en la toma de decisiones

Las organizaciones a menudo toman decisiones estratégicas en función de los datos que recopilan en la web, lo que puede conducir a decisiones equivocadas debido a la precisión, integridad y relevancia de los datos en la web.

2.16.7. Restricciones de rastreo de la web proporcionadas

En el Web Scraping una de las cosas que se debe tomar en cuenta es puede existir una serie de razones legítimas (como preocupaciones de privacidad) que llevan a un desarrollador de sitios web a prohibir los robots de indexación automática de su página web. El incumplimiento de estas prohibiciones de seguimiento web puede provocar daños no deseados a los usuarios y propietarios de sitios web.

Una de las herramientas más populares para la implementación de Web Scraping es el lenguaje de programación Python. Este se usa en proyectos relacionados con la ciberseguridad, las pruebas de penetración y el análisis

forense digital. Usando los conceptos básicos de la programación de Python, el raspado web se puede hacer sin usar otras herramientas de terceros.

El lenguaje Python es muy adecuado para el desarrollo de proyectos de navegación web por las siguientes razones:

- Simplicidad sintáctica
- Módulos incorporados
- Lenguaje de programación de código abierto
- Amplia gama de aplicaciones

2.16.8. Historia del Lenguaje de Programación de Python

Python es un lenguaje de programación creado por Guido Van Rossum fue lanzado en febrero de 1991. Python fue originalmente creado a partir de un lenguaje interpretado conocido como ABC. El objetivo de crear Python era corregir errores que existían en el lenguaje ABC y a su vez conservar las características positivas del mismo. Este lenguaje de programación fue elaborado en el sistema operativo Amoeba Distributed, en donde se quería que la sintaxis sea parecida a la de ABC con lenguaje de scripts. Este proyecto fue realizado por una sola persona la cual en su momento no tuvo apoyo y su creación no tenía acogida, lo cual no fue un impedimento y en por ello que en 1989 nace lo que hoy conocemos como Python (Cabutto et al., 2018).

2.16.9. Versiones de Python

Tabla 8:

Matriz de Versiones de Python

Características	Python 3	Python 2
Impresión de contenido	Se maneja como una función integrada de Python	Se maneja como una declaración del lenguaje
Operación de división	para la división se utiliza el operador de barra "/" para una división con resultado de coma flotante y una doble barra para un resultado de tipo entero	se utiliza el operador barra indistintamente y el resultado depende de los operandos en la división

Entrada de datos	la entrada de datos es manejada por la función input ()	La entrada de datos es manejada por la función raw_input ()
Rangos	el reemplazo de range () en Python 2 es xrange ()	la función range () devuelve una lista como resultado
Cadenas de texto	Se hace uso del alfabeto Unicode	Se utiliza el alfabeto ASCII
Lanzamiento de excepciones	Se debe encerrar el argumento de la excepción entre paréntesis	los argumentos de la excepción pueden o no ir entre paréntesis
Captura de excepciones	Se requiere el uso de la palabra clave as	El uso de palabras clave no es necesario
Archivos	el uso de ficheros es manejado por la función open ()	la función file () es la encargada del manejo de archivos

Nota: Matriz de Versiones de Python. Elaboración: Jorge Alberto Oviedo Peñafiel e Inés Janellys Fajardo Romero. Fuente: (Nanjekye, 2017)

2.16.10. Python para Ciencia de Datos

La ciencia de datos es un campo que incluye diversas disciplinas relacionadas con la estadística, informática, comunicaciones, gestión y sociología. La ciencia de datos se enfoca en la compresión de datos complejos, ya que su objetivo principal es transformar esos datos en conocimiento e inteligencia para la toma de decisiones (Longbing Cao, 2017).

En la búsqueda de un buen lenguaje de programación que pueda crear muchas aplicaciones de ciencia de datos, Python se ha convertido en una solución de programación completa, ya que, debido a su baja curva de aprendizaje y flexibilidad, Python se ha convertido en uno de los lenguajes de más rápido crecimiento. La creciente biblioteca de Python lo hace ideal para el análisis de datos (Nagpal & Gabrani, 2019).

En el estudio de la ciencia de datos se exige un lenguaje muy versátil y flexible que sea sencillo y que pueda manejar un procesamiento matemático muy complejo. Es por ello por lo que Python es considerado como el lenguaje más adecuado para esta actividad, ya que a pesar de que Python no es un lenguaje

que específicamente es para el análisis de datos o la computación científica. En las últimas dos décadas, Python se ha convertido en la herramienta más avanzada para tareas computacionales, incluido el análisis y la visualización de datos.

2.17. Librerías Python usadas en la Ciencia de Datos

Python contiene una gran cantidad de bibliotecas de extensiones usadas para la ciencia de datos se describen según su utilidad:

2.17.1. Exploración y análisis de datos

2.17.1.1. Numpy

Es una librería de Python usada en el análisis de los datos esta librería proporciona una matriz multidimensional y una serie de funcionalidades para operaciones de matriz rápidas que incluyen matemáticas, lógica, manipulación de formas, clasificación, selección, E/S, también incluyen operaciones de estadística básica y más (Harris et al., 2020).

2.17.1.2. SciPy

Esta librería es un conjunto de algoritmos matemáticos y funciones creadas sobre la extensión NumPy en Python. Esta librería es usada para manipular y mostrar datos. SciPy se convierte en un entorno para el procesamiento de datos (Virtanen et al., 2020).

2.17.1.3. Pandas

Pandas es una librería de Python usada en la ciencia de datos, esta es usada cuando se desea trabajar con datos tabulares, como los que se almacena en hojas de cálculos o bases de datos. Pandas es una herramienta que ayuda a explorar limpiar y procesar los datos, por medio de DataFrame (Hagedorn et al., 2021).

2.18. Visualización de datos

2.18.1. Matplotlib

Es una extensión que se utiliza para crear gráficos estadísticos de dos dimensiones en Python, esta puede crear visualizacisacioens estáticas, animadas e interactiva permitiendo que las cosas sean fáciles y posibles (Lemenkova et al., 2019).

2.19. Machine Learning Clásico

2.19.1. Scikit – Learn

Scikit-Learn es una librería que se utiliza para el ajuste, selección, evaluación de modelos, además del preprocesamiento de datos, etc. Esta biblioteca es de código abierto dirigido hacia el aprendizaje automático ya sea supervisado o no supervisado (Bisong, 2019).

2.19.2. Keras

Es un API creada en Python para realizar pruebas rápidas y ser capaz de transformar ideas en resultados lo más rápido posible. Esta API se ejecuta en la plataforma TensorFlow (Grattarola & Alippi, 2021).

2.19.3. TensorFlow

Es una biblioteca de código abierto que se utiliza para desarrollar y entrenar modelos de aprendizaje automático. La biblioteca realiza cálculos utilizando gráficos de flujo de datos simbólicos. Esta es un marco para manipular arreglos de N-dimesiones similar a Numpy (Lagouvardos et al., 2020).

2.20. Probabilidades y extremos

2.20.1. Spacy

Esta es una extensión de código abierto para el Procesamiento del Lenguaje Natural en Python, esta librería es usada para construir aplicaciones capaces de procesar y entender un gran conjunto de texto (Delgado López, 2021).

2.20.2. NLTK

Natural Language Toolkit es una extensión de Python que permite trabajar con datos de lenguaje humano. Esta herramienta tiene un gran número de corpus incorporados y recursos léxicos que, en conjunto con otras bibliotecas de procesamiento de texto, tokenización, lematización, etiquetado análisis y razonamiento semántico potencia al NLP (M. Wang & Hu, 2021).

2.21. Otras librerías

2.21.1. Wordcloud

Este permite visualizar palabras comunes que suelen ser usadas para representar metadatos de palabras claves o etiquetas en los sitios web o para poder visualizar texto de forma libre (Murthy & Scholar, 2020).

2.21.2. NeuralCoref

Es una biblioteca la cual entrena redes neuronales paralelas para evaluar referencias individuales y pares de referencia en función de una variedad de características gramaticales, de espaciado y de incrustación de palabras (Cho et al., 2018).

2.21.3. Gensim

Gensim es una librería popular de código abierto usada para el aprendizaje automático no supervisado. Su funcionalidad principal es la comparación entre diferentes documentos. Ya que permite realizar el análisis que determine las similitudes entre dos documentos (Kalyoncu et al., 2018).

2.21.4. ¿Por qué Python?

Si bien es cierto que temas como aprendizaje automático, modelos de predicción e Inteligencia artificial no son nada nuevo y han sido objeto de estudio por varias décadas, en los años más actuales el avance de la matemática computacional ha potenciado mucho estas áreas y han permitido la implementación de nuevas formas de implementación. Una de esas soluciones es el lenguaje de programación Python el cual es simple, robusto y con una amplia comunidad de desarrolladores que dan cabida a abundante documentación (Sodhi et al., 2019). Python es un lenguaje de programación sumamente sencillo en su sintaxis y con abundante cantidad de librerías proporcionadas de forma libre para su acceso y uso, a diferencia de otros lenguaje como java o C, Python posee la característica de permitir un fácil manejo de diferentes tipos de datos como arreglos, listas y cadenas de texto, lo cual resulta muy útil al momento de manejar datos e información, lo cual no ha convertido en el lenguaje por excelencia para el desarrollo de modelos de aprendizaje automático y profundo. Es un lenguaje tan sencillo que incluso aquellos que no son expertos en la programación pueden manejarlo con gran habilidad.

2.21.5. Argumentos en contra

Según (Mihajlović et al., 2020) existen diversos factores que como el resto de los lenguajes de programación juegan en contra de la implementación de Python, estos son:

- El dinamismo del lenguaje conduce a eventual lentitud por el trabajo implícito del lenguaje

- Utiliza grandes cantidades de memoria, lo que causa inconvenientes en tareas de optimización
- Se utiliza del lado del servidor y generalmente no tiene una presencia notoria en aplicaciones móviles
- Posee poca capacidad de procesamiento en comparación de otros lenguajes
- No es robusto en comunicaciones con base de datos
- Al ser de tipado dinámico las variables pueden cambiar de tipo según la secuencia del programa lo que puede generar confusión al desarrollador y requiere pruebas exhaustivas

2.21.6. Herramientas para el manejo de dependencias

2.21.6.1. Anaconda

Esta herramienta proporciona una versión libre del lenguaje de programación Python además de proveer todos los paquetes necesarios para las tareas de ciencias de datos en independencia del sistema operativos en el que se trabaje. Posee múltiples librerías científicas con una alta facilidad de instalación y manejo de estas. Es una herramienta muy usada por desarrolladores de software y científicos de todas las disciplinas que requieran el uso de Python para sus objetivos (Aiken et al., 2018).

Anaconda puede definirse en simples palabras con un proveedor de paquetería para los lenguajes de programación Python y R, los cuales con muy usados para el estudio de la ciencia de datos con aplicaciones muy comunes en la inteligencia artificial. Es un programa de distribución libre ampliamente usado en todo el mundo con paquetes y librerías disponibles en varios sistemas operativos como Windows, Linux Y MacOS.

2.21.6.2. Google Colab

Google Colab cuya última palabra es una abreviación de Colaboratory es un servicio brindado por la empresa Google el cual solo es necesario poseer una cuenta Gmail activa. Este servicio consiste en proporcionar un procesador y memoria en la nube para el desarrollo de programa en lenguaje de programación Python, para que se usa el entorno de desarrollo Jupyter, que facilita la ejecución por bloques del código, ahorrando así tiempo y memoria en procesos repetitivos que son inevitables en un ambiente común y corriente (Kanani & Padole, 2019).

Los ambientes de desarrollo en la nube han demostrado aportar muchas ventajas como la abstracción de instalación de dependencia en los equipos, incluso asignación de memorias en programas especializados. Google Colab da la facilidad de almacenar contenido relacionado a los desarrollos en la cuenta Google Drive anexada al correo Gmail, lo cual permite aprovechar muchas el almacenamiento sin importar el equipo desde el que se trabaja.

2.21.6.3. Jupyter

El proyecto Jupyter es un programa interactivo que permite el desarrollo en diferentes lenguajes de programación, aunque en la mayoría de los casos el lenguaje más usado es Python. Jupyter permite la ejecución de código por bloques lo que facilita el desarrollo en la ciencia de datos y el desarrollo de modelos de inteligencia artificial, una característica como esta es muy útil en etapas de entrenamiento. Jupyter es accesible mediante cualquier navegador web y también en aplicaciones de escritorio (Bilheux et al., 2019).

Jupyter permite además de la ejecución por bloques, que estos tengan diferentes lenguajes de programación en la misma hoja y aun así poder trabajar sobre los mismos datos. La interfaz de usuario es sumamente sencilla basándose en una hoja en la cual se pueden introducir bloque códigos y comentarios, así mismo separar estos en orden según sea la necesidad haciendo el programa muchas portable y entendible para retomar los desarrollos en casos de proyectos extensos.

2.22. Framework web para Python

2.22.1. Django

Es un framework web Python de alto nivel que impulsa al desarrollo rápido y el diseño limpio y pragmático. Es gratuito y de código abierto.

2.22.2. Flask

Flask es un framework web, que le permite desarrollar aplicaciones web fácilmente con código Python. Tiene un pequeño núcleo y es fácil de extender: es un micromarco que no hace uso de ORM (Object Relational Manager) o tales características (Visual Studio Code, 2021).

Figura 3:

Comparación entre Django vs Flask

Nota: El siguiente gráfico se visualiza cada de la comparación de interés a lo largo del tiempo con el framework Django y Flask, obtenido desde Google Trends.

2.22.3. Diferencia entre Flask vs Django

Flask y Django son dos herramientas de Python populares que tienen diferencias importantes entre sí.

Tabla 9:

Comparación entre Django vs Flask

Flask	Django
Permite el diseño web Python para desarrollo Rápido	Permite el desarrollo web Python para proyectos fáciles y simples
Framework WSGI	Framework web Full Stack
Proporciona soporte para API	No cuenta con soporte API
Permite depuración visual	No es compatible con Visual Debug
Permite el uso de múltiples bases de datos	No permite múltiples bases de datos
No ofrece páginas HTML dinámicas	Ofrece paginas HTML dinámicas
El despachador de URL del marco web flask es una solicitud RESTful.	El despachador de URL de este marco de Django se basa en controller-regex.

Las mejores características de Flask es que es ligero, de código abierto y ofrece una codificación mínima para desarrollar una aplicación.	Las mejores características de Django son Desarrollo rápido, Código abierto, Gran comunidad, Fácil de aprender.

Nota: La tabla posee las diferencias que existen entre el framework Django vs Flask

2.23. Revisiones sistemáticas

La revisión sistemática del presente proyecto de titulación se realizó mediante el buscador Google Scholar. La selección de los trabajos de referencia se basó en artículos científicos, revistas, libros educativos, tesis de grado, con un intervalo de fechas que van desde el 2017 al 2022, usando filtros de búsqueda como: autores, títulos entre comillas y operadores lógicos como: AND, OR, &&.

La información esencial se obtuvo realizó mediante lectura rápida, teniendo en cuenta puntos específicos en el documento como lo es el resumen, proceso, resultado y conclusiones. Además, se consideró el uso de las palabras claves como: Covid-19, Inteligencia Artificial, Machine Learning, Procesamiento de Lenguaje Natural, modelos de Machine Learning y NLP, lenguaje de programación Python, entre otros.

2.23.1. Meta-análisis

Para el presente trabajo de titulación se utilizó 128 artículos de investigación entre los cuales hay artículos científicos y tesis que forman parte de las referencias bibliográficas, estos se detallan en el anexo 4. la información utilizada en el presente trabajo consta de 97 artículos en inglés y 31 en español. el detalle anexado de la información útil a la investigación permite visualizar las técnicas y metodología de trabajo.

2.24. Preguntas científicas por contestarse

¿Es posible que un modelo NLP con algoritmos de machine Learning pueda categorizar el texto de una conversación textual de personas contagiadas con Covid-19?

2.25. Variables de la investigación

2.25.1. Variable independiente

Características de la información sujeta a clasificación, etiquetas o categorías en las cuales se pueden clasificar las entradas textuales.

2.25.2. Variable dependiente

La clasificación que el modelo realizará en base a las características definidas con anterioridad ósea las etiquetas de la información.

2.26. Definiciones conceptuales

2.26.1. Inteligencia Artificial

La inteligencia artificial es una rama de la ciencia que se encarga de simular el comportamiento humano en las máquinas o sistemas informáticos (Sánchez-caro & Sánchez, 2021).

2.26.2. Machine Learning

Machine Learning es una rama de la inteligencia artificial la cual pretende que las máquinas aprendan sin necesidad de ser programadas (Bi et al., 2019).

2.26.3. Procesamiento de Lenguaje Natural

El NLP es una rama de la inteligencia artificial la cual se basa en la habilidad que tiene una máquina para procesar la información que se obtuvo mediante el uso del lenguaje natural, ya sea escrito o por voz facilitando el entendimiento humano-máquina (Moreira, Cruz, Gonzalez, Quirumbay, et al., 2020).

2.26.4. Dataset

Es un conjunto de datos que contienen mediciones las cuales se asocian con métricas y entidades. Además, estos conjuntos de datos no se limitan solo a números y textos, pueden añadir imágenes o videos.

2.26.5. Algoritmos

Los algoritmos son aquellos que permitan evaluar y entrenar un conjunto de datos capaz de predecir un nuevo conjunto (Cevallos et al., 2020).

2.26.6. Métricas de Evaluación

Son aquella que permiten analizar y verificar el desempeño de los algoritmos utilizados, son útiles al momento de evaluar los patrones obtenidos en términos de compresibilidad (Mendoza Olguín et al., 2019).

2.26.7. Probabilidad

La probabilidad es una rama de las matemáticas que se encarga de estudiar fenómenos aleatorios, cuyos resultados son impredecibles (Andrés Pérez, 2019).

2.26.8. Bosques Aleatorios

Algoritmo matemático que consiste en un conjunto de árboles de clasificación que sirven para dar una predicción (Beltrán & Barbona, 2022).

2.26.9. Máquina de Soporte Vectorial

Modelo de aprendizaje supervisado que genera separaciones en el hiperplano que permite distinguir una clase de la otra (Galindo et al., 2020).

2.26.10. Matriz de Confusión

Tabla que permite visualizar la distribución de errores presentes en el trabajo de un modelo de clasificador (del Castillo Collazo, 2020).

03 CAPITULO

METODOLOGÍA DE LA INVESTIGACIÓN

Metodología de la investigación

En el presente capítulo se realiza un estudio de las fases que fueron necesarias para el trabajo investigativo. Estas fases constan en primer lugar del análisis y preprocesamiento de los datos, desarrollo del modelo conversacional, etapa de entrenamiento con datos tomados del corpus, evaluación del desempeño del modelo mediante métricas que permitan juzgar la eficacia de las predicciones. Se presentarán todos los modelos conversacionales por cada prueba y experimentación. El modelo es sujeto de modificaciones determinadas por la necesidad que se presenten en el proceso, estas modificaciones engloban a toda la variación de valor en los parámetros del modelo. Finalizados los procesos anteriores se evaluará cada solución con tal de hacer elección del modelo óptimo.

3.1. Tipo de investigación

3.1.1. Investigación Exploratoria

La investigación exploratoria es aquella que tiene como objetivo el acercamiento a fenómenos novedosos sobre algún tema u objeto desconocido o poco estudiado que genere interés. La investigación exploratoria por lo general determina tendencias, identifican áreas, ambientes, contextos y situaciones de estudio, relaciones potenciales entre variables; o establecen el "tono" de investigaciones posteriores más elaboradas y rigurosas (Nieto, 2018).El presente proyecto utiliza la investigación exploratoria para establecer bases sólidas, la cuales permitirán el entendimiento adecuado y en profundidad sobre los temas a tratar como: la Inteligencia Artificial, el Procesamiento del Lenguaje Natural, Machine Learning, entre otros. Además, de la indagación con los antecedentes adecuados que nos permitan orientar esta investigación de manera correcta.

3.1.2. Investigación de diagnóstico

La investigación de diagnóstico es un método de estudio el cual se encarga de conocer lo que ocurre en una situación específica. Su objetivo es analizar una serie de sucesos para identificar los distintos factores que promovieron la aparición de un fenómeno y determinar la frecuencia con la que algo ocurre en asociación con otra cosa (González, 2020).

En el presente estudio se utiliza la investigación de diagnóstico con la finalidad de analizar e identificar las características que tengan los modelos de Procesamiento de Lenguaje Natural, puesto que de esta manera se puede establecer si las distintas características de un modelo son diferentes o semejantes. Esto permite tomar la decisión adecuada sobre cuál es el modelo más eficiente.

3.1.3. Investigación Descriptiva

La investigación descriptiva es aquella utilizada cuando el propósito principal de la investigación en es la descripción de todos los componentes que describen una realidad. Tiene como finalidad la descripción de características del objeto de estudio o fenómeno en cuestión (Guevara et al., 2020).

El presente proyecto hace uso de la investigación descriptiva debido a que en el desarrollo del modelo surge la necesidad de explicar cada componente que lo conforma, desde las técnicas de procesamiento de palabras utilizadas, hasta el modelo y su arquitectura en si tomando en cuenta aspecto como los parámetros y capas.

3.1.4. Investigación Evaluativa

La investigación evaluativa es una metodología que beneficia a toma de decisiones con el fin de optimizar los procesos multidisciplinarios debido a que se apoya en los resultados y hace uso de estos para evaluar el cambios o permanencia de una solución en cuestión a una situación o problemática (Garduño Teliz et al., 2019).

El uso de este tipo de investigación se ve matizado en la necesidad de la evaluación de los resultados del modelo del presente proyecto. Existen múltiples técnicas de evaluación abordadas con anterioridad que permiten determinar qué tan eficaz puede llegar a ser una solución para los objetivos propuestos. El desarrollo de este estudio deberá evaluarse con las correspondientes técnicas y así determinar cuál es más viable a los propósitos.

3.1.5. Investigación Cuasi Experimental

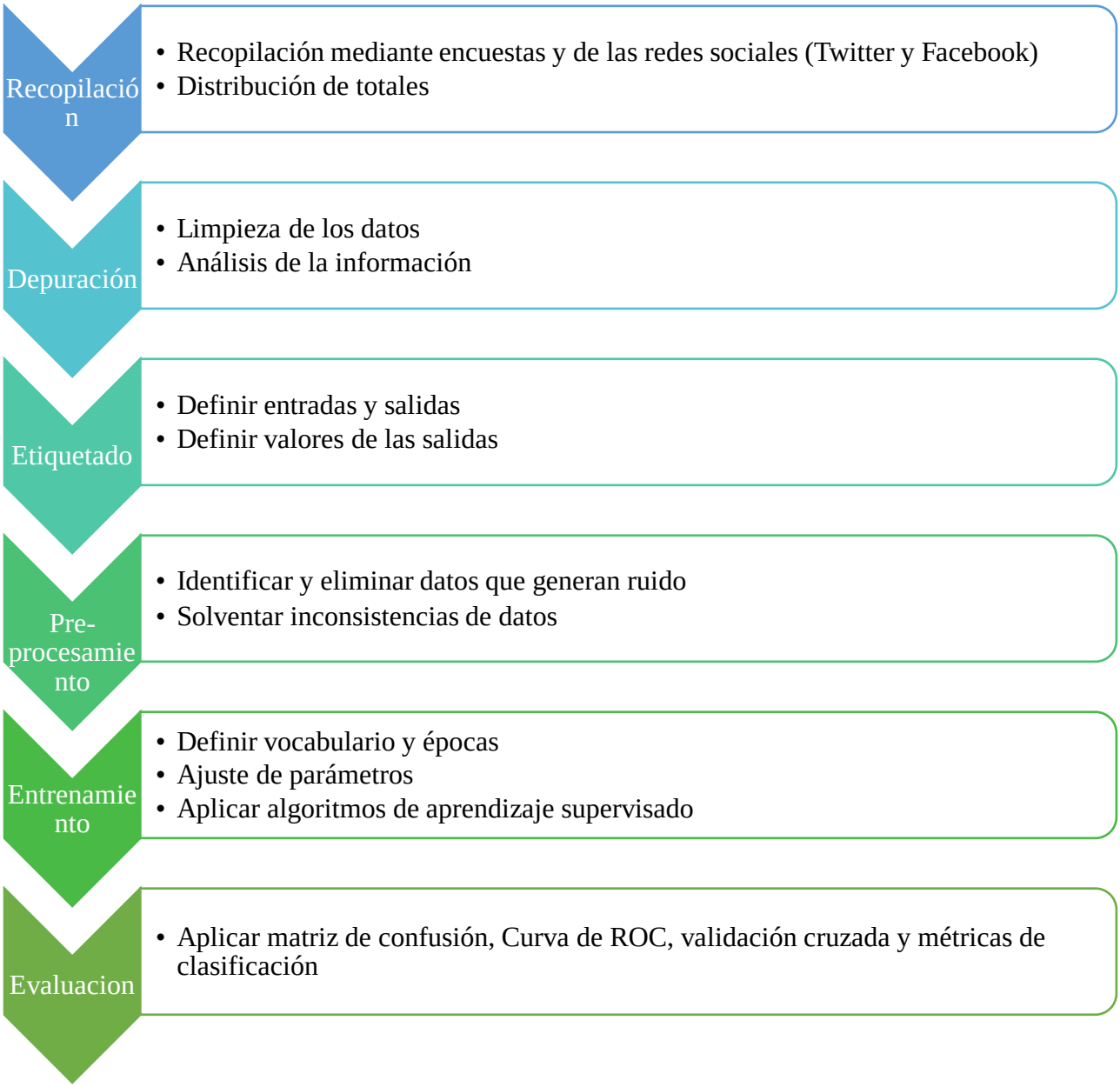
La investigación cuasi-experimental se enfoca en la recolección de información donde se comparan las mediciones del comportamiento que tiene un grupo de control con aquellas medidas de un grupo experimental (Agüero et al., 2021). Es denominado cuasi-experimental debido a la manipulación por parte del investigador de las variables objetivo de verificar si esta modificación surte algún efecto en el comportamiento de nuestro objeto de estudio.

El uso de la investigación cuasi-experimental es parte fundamental de esta investigación debido a que de esta forma se puede llegar a dictaminar la efectividad de un modelo con la modificación de sus variables independientes (parámetros) sobre sus variables dependientes (resultados de las métricas).

3.2. Diseño metodológico de la investigación

Figura 4:

Etapas de desarrollo del Modelo PNL



Nota: El siguiente gráfico se visualiza cada de las etapas necesarias para el desarrollo del modelo de Procesamiento de Lenguaje Natural basado en Machine Learning para la clasificación textual.

3.3. Metodología de la investigación

3.3.1. Definición del problema

El desarrollo de un modelo de NLP para la clasificación de texto representa un reto en el cual se deben tener el objetivo claramente planteado. Una forma de identificar las necesidades que el modelo suplirá es con la modularizarían del objetivo en sí, lo cual se logra mediante cuestionamientos pertinentes a la etapa de desarrollo, algunas de los puntos a mencionar para la obtención de la meta planteada son:

3.3.1.1. Objetivo Principal

Se debe elaborar un modelo de procesamiento de lenguaje natural para la clasificación de textos obtenidos mediante conversaciones con personas contagiadas de covid-19

3.3.1.2. El objeto de la clasificación

Se debe clasificar los textos de entrada del modelo que permita identificar palabras clave que deriven en un mayor entendimiento de la problemática para el modelo

3.3.1.3. Aprendizaje del modelo

El modelo debe ser entrenado para que aprenda identificar y clasificar sintomatología del covid-19 y recomendaciones de conversacionales textuales

3.3.1.4. Alcance del objetivo con datos existentes

Los datos existentes tomados de las encuestas y de las redes sociales serán debidamente etiquetados y distribuidos en datos de entrenamiento y datos de testeo.

3.3.1.5. Medición de los resultados

Los resultados del modelo serán evaluados mediante la matriz de confusión, curva de ROC, validación cruzada y métricas de evaluación de clasificación.

3.4. Fase 1

3.4.1. Recopilación de la Data

Para el presente estudio de investigación se procede a la creación de una encuesta para lograr obtener un conjunto de datos de información necesaria sobre los síntomas y recomendaciones de personas que estuvieron contagiadas de Covid-19, dicha encuesta se realizó mediante el uso de la herramienta Google Forms, el uso de esta aplicación es gratuito y de fácil administración. Se logró obtener un conjunto de datos de 4140 personas que padecieron dicha enfermedad.

3.4.2. Recopilación mediante encuestas

Mediante un conjunto de preguntas se realizó una encuesta, de la cual se obtuvo información relevante sobre los síntomas y recomendaciones de personas que fueron contagiadas de Covid-19. Esta información es almacenada mediante un enlace entre Google Form y la Hoja de cálculo de Google, lo cual permitía el fácil acceso a las respuestas recopiladas en dicha encuesta.

3.4.3. Distribución de totales

La distribución de totales se realizó mediante las diferentes redes sociales, específicamente a personas que pertenecía a la Zona 8 de la provincia del Guayas que comprende Guayaquil, Durán y Samborondón. De esta manera se logró la recopilación total de 4140 encuestas validas.

3.5. Fase 2

3.5.1. Depuración de los datos

Una vez obtenidos los datos necesarios para el trabajo de la presente investigación, estos tienden a encontrarse desordenados, con información sin relevancia e incluso sin sentido semántico, por lo que es necesario intervención para la limpieza y depuración de la información. Este proceso es sumamente necesario para el desarrollo eficiente de modelos.

3.5.2. Limpieza de los datos

En la recopilación de la información se especificó la herramienta usada que fue la encuesta. El uso de este método tiene como resultados que los datos pueden venir con diversas falencias, como faltas ortográficas, datos en blanco, sin sentido semántico etc. la existencia de situaciones como estas contribuyen a que posteriores etapas como el entrenamiento del modelo no sea de forma óptima y la tarea de clasificación de texto tenga resultados muy cuestionables.

Figura 5:

Dataset con errores semánticos

Paracetamol	Una. Semana con vaporozaciones	Miel de abeja, vitamina	Ningún malestar
Analgan,parasetamol, vitar	Cuarentena durante 1 mes	Vitamina c , frutos secc	Me siento muy bien no e tenido ningún probl
Paracetamol	Jugos de tomate y de limon, vapor	Tomate, limon	Me siento bien recuperado
Inyecciones para subir las	Estar aislada, 2 meses en recuper	Bebida de gelatina	Mucho mejor que nunca
Sinceramente no recuerdo	Aislamiento (no todas las persona	Me recetaron Vitamina	Bien
Paracetamol y bajos en la	Tomo 15 días y con cuidados de a	Complejo b . Paracetan	Muy bien
Parasetamol y enterogerm	Aislamiento total 15 días	Vitamina c	Me siento bien sin ninguna molestia
Paracetamol y nebulizacio	Aislamiento total, dos semanas.	Vitamina C, miel, menta	Anímicamente bien, mi sentido del gusto qu
Antibióticos recetados	Dos semanas, me aislé	Vitamina C, Birm	Agradecida con mi familia y a los doctores
Paracetamol enterogermin	Aislamiento y descanso	Vitamina c	Sin ningún sintoma
medicamentes de gripe	tome tes y muchas cosas caliente	vitamina c	mucho mejor
Azitromicina	Infusiones y analgesicos	Frutas y jugos citricos	Normal
Paracetamol	Estar en reposo 3 semanas	Sopita de Poio	Estoy muy bien,full amor propio
Vitamina C Redoxon	4 días	Redoxon	Un poco debil
Nose	No salir de casa, lavarme las man	Frutas	Bien
simgripal la trate como un	baños calientes y te calientes	pura vitamina c y limon	muchisimo mejor
Vitamina C	Vaporizaciones con eucalipto	Birm	No quedo ningún síntoma luego del covid
Paracetamol	En un cuarto aislado solo duró una	Vitamina C	Me siento bien no dejo ninuna secuela que a

Nota: El siguiente gráfico se visualiza el Dataset con información errónea, sin sentido semántico, y errores ortográficos.

3.5.3. Análisis de la información

En la revisión de todos los datos necesarios para la investigación, se conoce que la información esta almacenada en una carpeta de Google drive la cual contiene el total de las encuestas en general de forma depurada y también clasificada por los conjuntos que se asignó a cada grupo de tesis para tareas de limpieza.

3.5.4. Formato y visualización de los datos

La información presente en la Dataset tiene sus respectivos atributos que deben ser analizados en características como nombre, descripción y tipo de dato, a continuación, se visualizara un cuadro en el cual se describen cada uno de los atributos:

Tabla 10:

Atributos del Dataset

Nombre	Descripción	Tipo de Dato
Cantidad	Identificador numérico del registro, aumento de manera secuencial	Numérico
Marca temporal	Fecha en la que la encuesta fue realizada	Fecha
Nombre y apellido del Encuestador	Nombres de la persona encargada de hacer la encuesta	Texto

1. ¿Ha tenido Coronavirus?	Atributo de selección entre opciones	Texto
2. Seleccione la edad	Opciones de rango de edades para el encuestado	Texto
3. genero	Genero con el que se identifica el encuestado	Texto
4. ¿Qué variante del virus lo contagió?	Especificación de la variante en caso de sí haber tenido Covid	Texto
5. ¿En qué fecha se contagió?	Fecha tentativa en la cual el encuestado afirma haber infectado	Fecha
6. ¿Nivel de intensidad que tuvo los síntomas?	Atributo de opción para seleccionar la intensidad de los síntomas del encuestado	Texto
7. ¿En qué lugar o evento considera que se contagió?	Atributo textual en el cual el encuestado específica el lugar que considera que ocurrió su contagio	Texto
8. ¿En caso de haber estado vacunado al momento de contagiarse cuantas dosis tenía aplicadas al contagiarse?	Atributo para especificar el número de vacunas recibidas	Numérico
9. ¿En caso de haber estado vacunado al momento de contagiarse Qué vacuna recibió?	Campo textual para especificar la vacuna que el encuestado recibió	Texto
10. Describa lo más detallado ¿Qué síntomas ha tenido?	Aquí el encuestado tiene vía libre para describir sus síntomas	Texto
11. Describa ¿Qué medicamentos considera que le ayudo en su recuperación?	El encuestado puede especificar el nombre de los medicamentos que le fueron de ayuda en su recuperación	Texto
12. Describa ¿Qué cuidados aplico durante el proceso de	El encuestado puede especificar los cuidados	Texto

recuperación del Covid, y cuánto tiempo en días tomo su recuperación?	que le fueron de ayuda en su recuperación	
13. Describa a más detalle ¿Qué alimentos y/o vitaminas considera que le ayudaron a fortalecerse y superar el Covid?	El encuestado puede especificar los alimentos que le fueron de ayuda en su recuperación	Texto
14. De haber superado el Covid, describa ¿Cómo se siente en su estado de ánimo, autoestima o algún otro malestar que haya usted sentido?	Descripción de estado de ánimo del encuestado una vez superado el Covid	Texto
15. Describa ¿Qué Recomendaciones saludables le dio a conocer su médico? ejemplo: Mi doctor me recomendó que caminará 4 veces por semana por 40 minutos	El encuestado puede especificar los hábitos saludables que le fueron de ayuda en su recuperación	Texto
16. Finalmente, describa ¿Qué información le gustaría que esté disponible fácilmente para ser consultada por usted (con respecto al COVID19)?	El encuestado puede especificar los conocimientos que le gustaría tener a disposición del Covid	Texto
Lugar que reside de la zona 8	Lugar de residencia de la persona encuestada	Texto

Nota: La tabla posee el nombre, descripción y tipo de dato de cada uno de los atributos del Dataset

Figura 6:

Visualización de registros del Dataset

CANTIDAD	TESISTA	0. Nombre y Apellido del encuestador (Pe	1. Ha tenido coronavirus?	2. Seleccione la Edad	3. Género	4. ¿Qué variante del Virus l	5. ¿En qué fecha se contag
1	PA-62-18-6	KerlY Navarro Santos	Si	Entre 26 a 40 años	Femenino	Alfa (Original, covid-19)	10/
2	PA-62-18-6	Kerly Navarro Santos	Si	Entre 26 a 40 años	Masculino	Alfa (Original, covid-19)	14/
3	no	Luis Eduardo Rivera Armijo	Si	Entre 18 a 25 años	Masculino	Alfa (Original, covid-19)	13/
4	no	Luisa Armijos	No				
5	BD-31-5-1	Karellyss Ariadne Terán Otero	Si	Entre 18 a 25 años	Femenino	Alfa (Original, covid-19)	15/
6	no	Jane Coque	Si	Entre 26 a 40 años	Femenino	Alfa (Original, covid-19)	8/
7	no	Estefani triana	No				
8	no	STEFANIA TRIANA	Si	Entre 18 a 25 años	Femenino	No Sabe	9/
9	no	Luis rivera	No lo se	Entre 18 a 25 años	Masculino	No Sabe	17/
10	no	Jose Sanchez	No				

Nota: Los 10 primeros registros del Dataset

3.6. Fase 3

3.6.1. Etiquetado

Los datos ya una vez depurados son insumo para la fase de etiquetación la cual consiste en la separación de etiquetas en función de un atributo específico. Se aplico la creación de etiquetas que determinan la ausencia o presencia de esta en el atributo objeto de análisis.

3.6.2. Definir entradas y salidas

La creación de etiquetas tiene implícita la tarea de repetir esta creación por cada uno de los atributos que se consideren necesarios. En la presente investigación y con su respectivo Dataset, se ha tomado a consideración la elección de dos atributos, los cuales son el atributo **"10. Describa lo más detallado ¿Qué síntomas ha tenido?"** y **"15. Describa ¿Qué Recomendaciones saludables le dio a conocer su médico? ejemplo: Mi doctor me recomendó que caminará 4 veces por semana por 40 minutos"**. Los atributos anteriormente señalados cumplen su función de entrada siendo la salida las etiquetas respectivas que son creadas en función del contenido de cada uno de los registros de las entradas.

3.6.3. Definir los valores de las salidas

Una vez establecidas las entradas y con eso la creación de etiquetas es importante definir lo valores diferenciadores de cada posible situación. Los dos escenarios que se pueden dar es que una etiqueta pueda o no estar presenta en la entrada por lo que se determinó que el valor 1 será para presencia y el valor 0 para ausencia de la etiqueta.

3.7. Fase 4

3.7.1. Preprocesamiento

Una etapa necesaria es el preprocesamiento de la información el cual consiste en la eliminación de todos aquellos elementos y anomalías que puedan interferir en el flujo normal del algoritmo, aspectos como los tipos de datos, registros vacíos, caracteres raros y más son asunto que resolver en esta etapa que es indispensable para el trabajo posterior a realizar.

3.7.2. Identificar y eliminar datos que generan ruido

El ruido se puede conceptualizar como todos aquellos elementos que interfieren con el funcionamiento adecuado de un proceso, en este caso el ruido consiste en

anomalías del Dataset que puedan ocasionar que la información de entrada al modelo este contaminada y este se perjudique de esta información en sí. Una de estas anomalías es la inconsistencia en los tipos de datos que puede variar según como este lleno el Dataset, esto es algo que se debe resolver de forma obligatoria.

Otra forma de mitigar ruido en los datos es mediante la eliminación de caracteres especiales que no den una significación relevante a las palabras que entenderá el modelo, mediante la aplicación de funciones sobre todas las entradas se puede hacer este proceso de manera simple

Figura 7:

Función en Python para eliminación de caracteres anormales

```
[ ]  def clean_text(text):
         text = str(text)
         text = text.lower()
         text = re.sub('á', 'a', text)
         text = re.sub('é', 'e', text)
         text = re.sub('í', 'i', text)
         text = re.sub('ó', 'o', text)
         text = re.sub('ú', 'u', text)
         text = re.sub('\[.*?¿\]\%', ' ', text)
         text = re.sub('[%s]' % re.escape(string.punctuation), ' ', text)
         text = re.sub('\w*\d\w*', '', text)
         text = re.sub('[‘’“”…«»]', '', text)
         text = re.sub('\n', ' ', text)
         text = re.sub(r'\s+', ' ', text)
         text = re.sub('[^a-zA-Z]', ' ', text)
         return text
```

```
[ ]  #APLICANDO LA FUNCIÓN clean_text AL DATAFRAME TEMPORAL DE SINTOMAS
     df_rec_usar['Sintomas']=df_rec_usar.apply(lambda x: clean_text( x["Sintomas"]), axis=1)
```

Nota: La implementación de una función de eliminar caracteres especiales con el uso de expresiones regulares

3.7.3. Solventar inconsistencia de datos

Una de las problemáticas a resolver al momento de revisar las entradas establecidas en pasos anteriores consiste en resolver inconsistencias. Existe un conjunto de situaciones que podrían darse en los registros como por ejemplo entradas vacías, demasiado cortas, con puros espacios en blanco, todo esto representa un obstáculo a solventar para el desarrollo normal de pasos posteriores.

Esta anomalía no solo se debe revisar en las entradas, las salidas, es decir las etiquetas podrían tener así mismos datos inconsistentes en su contenido, si tomamos

en cuenta que solo se admiten valores 1 y 0 no debería existir ningún otro valor aparte de estos.

Figura 8:

Agrupación de datos para analizar entrada y salidas

```
[ ]  #AGRUPACIÓN DE LOS DATOS NUMERICOS A UTILIZAR
     df_rec_usar.iloc[:, 1:77]
```

```
[ ]  df_rec_usar.iloc[:, 1:77] = df_rec_usar.iloc[:, 1:77].fillna(0)
```

Nota: Agrupación de los Dataframe en Python que separan a la preguntan de las etiquetas

Dataset – Sintomatología

3.7.4. Análisis de los atributos del Dataset

La existencia de dos datasets tiene como objetivo la etiquetación de ambos en base a una entrada, que en cuyo caso el presente corresponde a la entrada de los síntomas, específicamente el atributo **"10. Describa lo más detallado ¿Qué síntomas ha tenido?".** Los atributos definidos como salida vienen dato por un compendio de los síntomas identificados en la entrada y en cada uno de los registros que conforman esa columna.

3.7.5. Aplicación de técnicas NLP

Una vez obtenida la Dataset etiquetada para los síntomas es necesario aplicar técnicas de procesamiento de lenguaje natural a la variable independientes, es decir a la entrada ósea los síntomas. La eliminación de palabras de articulación o nexos dejando únicamente las palabras clave es parte de lo que conoce limpieza stopwords el cual consiste en dejar únicamente las palabras que representen una significación alta para el modelo.

Figura 9:

Aplicación de stopwords a la entrada establecida

```
#TECNICAS DE NLP
X = df_rec['Sintomas'].apply(nfx.remove_stopwords)
y = df_rec.loc[ : ,columnas]
```

Nota: se remueven los stopwords de la entrada

3.8. Fase 5

3.8.1. Entrenamiento

El uso de etapa o fases anteriores fue por completo necesario para poder entrar a la creación del modelo en sí. El uso de librerías y sus respectivas clases facilita mucho la implementación del algoritmo implícito en un modelo además de aportar encapsulación de elementos que se simplifican considerablemente lo que se ve reflejado en que no es necesario escribir paso por paso el modelo sino más bien se reduce a unas pocas líneas de código. Los modelos escogidos para su implementación son máquina de soporte vectorial (SVM) y bosques aleatorios (RF).

3.8.2. Definición de vocabulario y vectorización

Las entradas preprocesadas son sin más un conjunto de palabras, que son el resultado de etapas anteriores. Los modelos de inteligencia artificial no entiendes palabras, entienden números y es necesario hacer esta transformación que forma parte del insumo necesario para el entrenamiento del modelo. Se realiza la etapa de vectorización que realiza la transformación necesaria, una vez realizado este paso se deben dividir los datos en un conjunto de entrenamiento y otro para el testeo.

Figura 10:

Vectorización y división de los datos

```
#Extraemos las caracteristicas y propiedades de la variable X en este caso los sintomas
vectorizer = CountVectorizer(max_features=10000, min_df=5, max_df=0.7, stop_words=stopwords.words('spanish'))
Xfeatures = vectorizer.fit_transform(X).toarray()

#División del conjunto de los datos para el entrenamiento
X_train,X_test,y_train,y_test = train_test_split(Xfeatures,y,test_size= param['TestSize'] ,random_state=42)
```

Nota: aplicación de técnicas de NLP como vectorización y creación de vocabularios

3.8.3. Ajuste de Parámetros - SVM

La creación e instanciación del modelo viene dado por clases propias de la librería que reducen en sobremanera el código, sin embargo, es necesario aun especificar el valor de los parámetros de entrada para la creación del modelo. Parámetros como el tipo de núcleo (kernel), tamaño de los datos de entramientos y otros más son especificados antes de la instanciación. Después de probar 12 entrenamientos para SVM y 18 para RF se llegó a la conclusión de que los parámetros más adecuados para obtener una exactitud (Accuracy) más elevada fueron los siguientes:

Kernel = linear

Ocurrencia de etiquetas = 200

Tamaño de datos de entrenamiento = 80%

Figura 11:

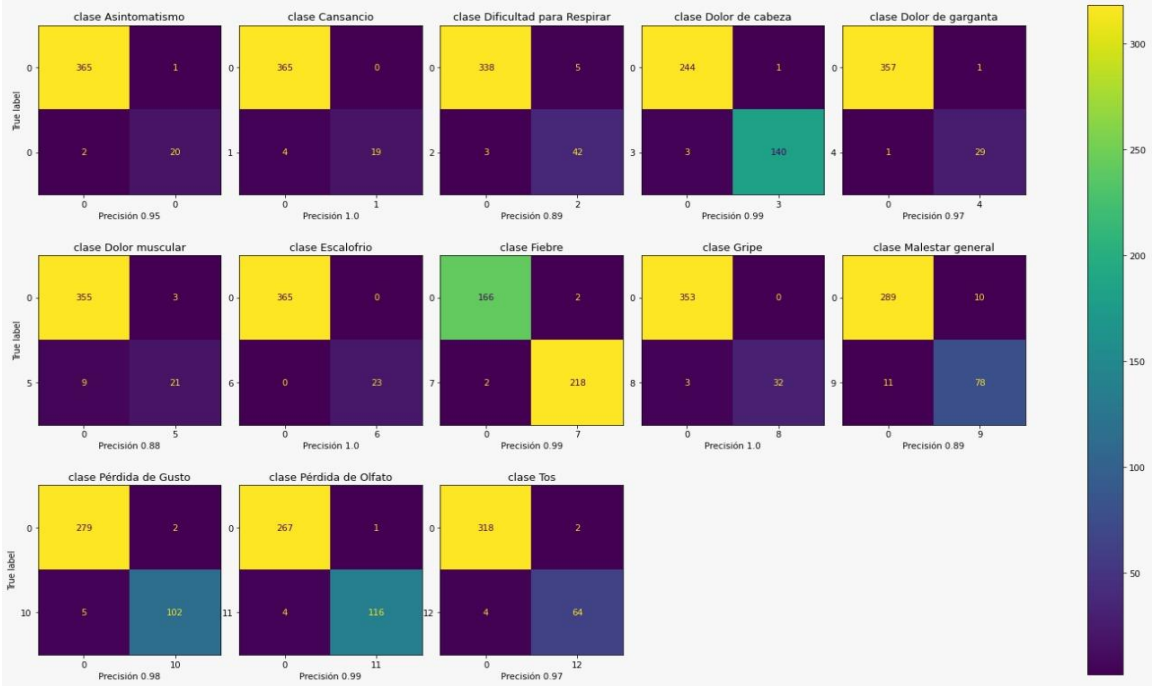
Construcción del modelo en lenguaje Python

```
#CONTRUCCION DEL MODELO
binary_rel_clf = SVC(kernel=param['kernel'],probability=True,random_state=0)
obj_multi_target_forest = MultiOutputClassifier(binary_rel_clf, n_jobs=-1)
label = y_train.apply(lambda x: x.argmax(), axis=1).values
```

Nota: creación del modelo con clases de la librería Scikit-learn

Figura 12:

Matriz de confusión del algoritmo de Soporte de Máquina Vectorial – Síntomas



Nota: la figura posee los resultados obtenidos en la matriz de confusión de cada una de las etiquetas con el algoritmo de Soporte de Máquina Vectorial - Síntomas de la cual se obtuvieron los mejores resultados.

3.8.4. Ajuste de Parámetros – RF

Criterio = entropy

Max Feacture = sqrt

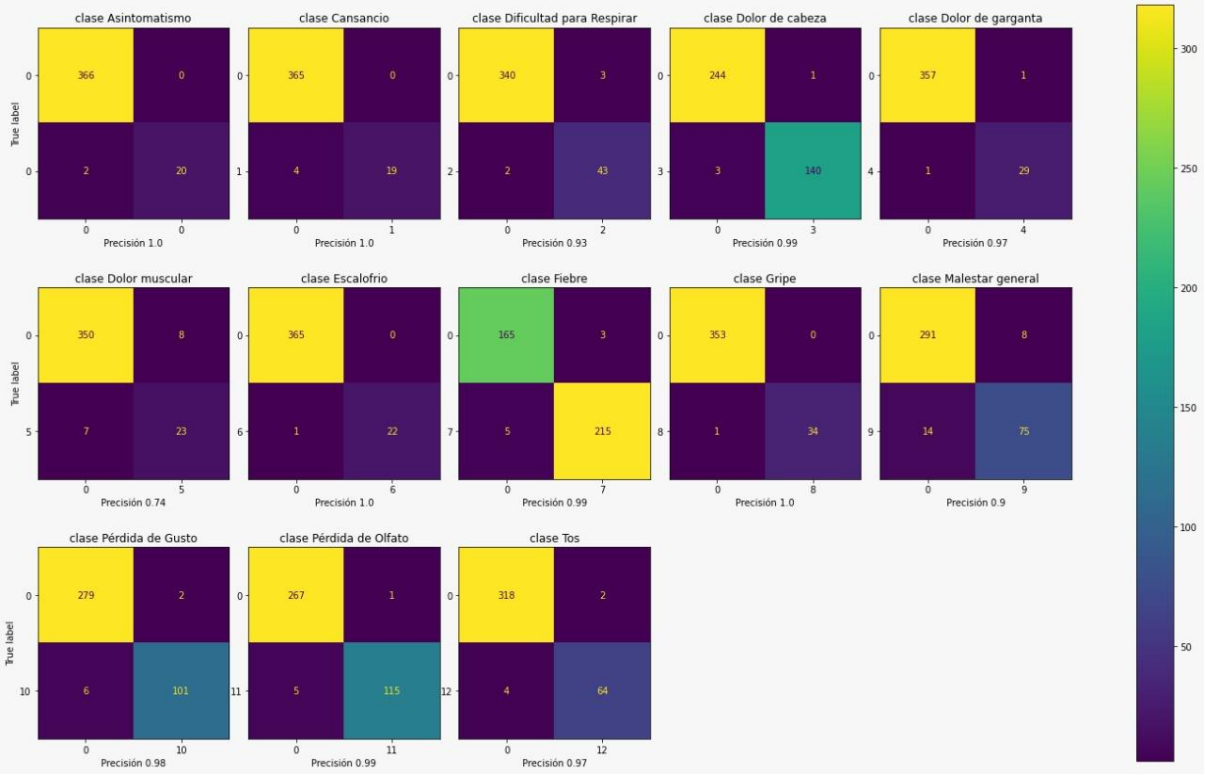
Estimadores = 260

Ocurrencia de etiquetas = 200

Tamaño de datos de entrenamiento = 90%

Figura 13:

Matriz de confusión del algoritmo de Random Forest – Síntomas



Nota: la figura posee los resultados obtenidos en la matriz de confusión de cada una de las etiquetas con el algoritmo de Random Forest-Síntomas de la cual se obtuvieron los mejores resultados.

3.8.5. Aplicación de algoritmo de aprendizaje

Una vez establecida la parametrización se procede a la parte de entrenamiento que consiste en un código que permite al modelo recibir los datos de entrada y salida de entramiento y así tener un entrenamiento interno con los parámetros establecido con anterioridad

3.9. Fase 6

3.9.1. Evaluación

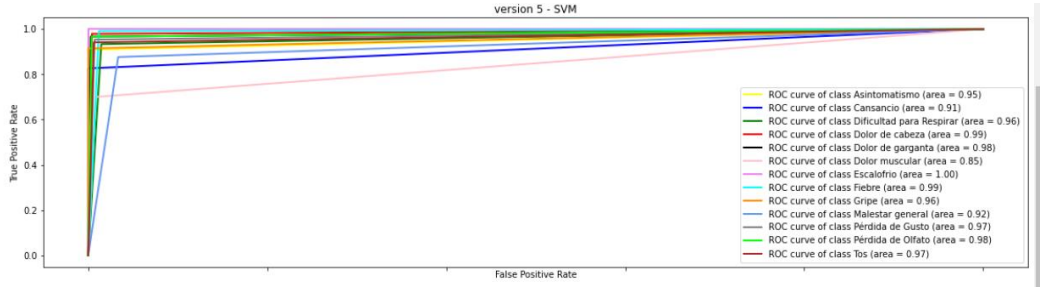
La culminación del proceso de creación del modelo viene dada por la evaluación de los resultados, cuyo análisis es realizado por las métricas de evaluación esto a su vez sirve como comparador entre los resultados de los dos modelos para elección del que posea los resultados óptimos.

3.9.1.1. Curva ROC

Una métrica de detección efectiva en los resultados arrojados por cada uno de los modelos es la curva ROC, la cual evalúa la sensibilidad de un sistema de clasificación, los resultados son óptimos cuando se acercan lo sumo posible al verdadero positivo (TP) lo que daría como resultado un conjunto de curva de apariencia casi perpendicular para cada una de las etiquetas de la clasificación.

Figura 14:

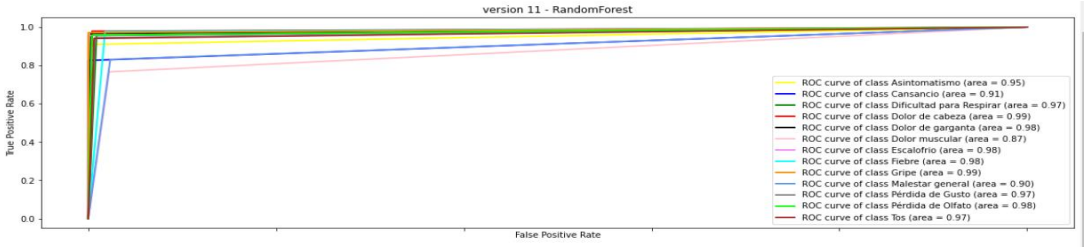
Curva ROC – SVM



Nota: resultados de la curva ROC para SVM en síntomas y cada una de las etiquetas

Figura 15:

Curva ROC - RF



Nota: resultados de la curva ROC para RF en síntomas y cada una de las etiquetas

3.9.2. Métricas de clasificación

Los resultados en métricas de evaluación para el modelo de máquina de soporte vectorial (SVM):

Tabla 11:

Métricas de Evaluación del algoritmo Soporte de Máquina Vectorial - Síntomas

Métricas de Evaluación Soporte de Máquina Vectorial			
Acurracy	**Precisión**	**Recall**	**F1**
0.8685587	0.969784	0.946597	0.957516

Nota: La tabla posee las métricas de evaluación que se obtuvieron en el algoritmo de Random Forest

Los resultados en métricas de evaluación para el modelo de bosques aleatorios (RF):

Tabla 12:

Métricas de Evaluación del algoritmo Random Forest – Síntomas

Métricas de Evaluación Random Forest			
Acurracy	**Precisión**	**Recall**	**F1**
0.865979	0.970071	0.941361	0.955154

Nota: La tabla posee las métricas de evaluación que se obtuvieron en el algoritmo de Random Forest

Figura 16:

Matriz de resultados obtenidos en el entrenamiento con Soporte de Máquina Vectorial y Random Forest – Síntomas.

	Version	Clasificador	Criterio	kernel	Max Features	Estimadores	Ocurrencia de Etiquetas	# columnas	TrainSize	TestSize	Accuracy	Precision	Recall	f1
0	1.0	SVM		linear		0.0	100.0	22.0	0.8	0.2	0.802993	0.950102	0.928158	0.937021
1	2.0	SVM		linear		0.0	200.0	13.0	0.8	0.2	0.856774	0.962384	0.942966	0.951400
2	3.0	SVM		linear		0.0	300.0	9.0	0.8	0.2	0.846591	0.961653	0.940512	0.949393
3	4.0	SVM		linear		0.0	100.0	22.0	0.9	0.1	0.783042	0.945694	0.932510	0.936415
4	5.0	SVM		linear		0.0	200.0	13.0	0.9	0.1	0.868557	0.969784	0.946597	0.957516
5	6.0	SVM		linear		0.0	300.0	9.0	0.9	0.1	0.840909	0.960088	0.938799	0.948059
6	7.0	SVM		rbf		0.0	100.0	22.0	0.8	0.2	0.764339	0.952026	0.881871	0.911154
7	8.0	SVM		rbf		0.0	200.0	13.0	0.8	0.2	0.820645	0.963393	0.914720	0.936849
8	9.0	SVM		rbf		0.0	300.0	9.0	0.8	0.2	0.826705	0.964513	0.925640	0.942127
9	10.0	SVM		rbf		0.0	100.0	22.0	0.9	0.1	0.738155	0.941147	0.890684	0.910236
10	11.0	SVM		rbf		0.0	200.0	13.0	0.9	0.1	0.829897	0.970194	0.922513	0.945029
11	12.0	SVM		rbf		0.0	300.0	9.0	0.9	0.1	0.832386	0.965133	0.928406	0.944691
12	1.0	RandomForest	entropy		sqrt	100.0	100.0	22.0	0.8	0.2	0.784289	0.944370	0.917068	0.929890
13	2.0	RandomForest	entropy		sqrt	100.0	200.0	13.0	0.8	0.2	0.838710	0.955739	0.935361	0.945289
14	3.0	RandomForest	entropy		sqrt	100.0	300.0	9.0	0.8	0.2	0.842330	0.948852	0.941106	0.944615
15	4.0	RandomForest	entropy		sqrt	200.0	100.0	22.0	0.8	0.2	0.784289	0.943056	0.916586	0.929005
16	5.0	RandomForest	entropy		sqrt	200.0	200.0	13.0	0.8	0.2	0.837419	0.955535	0.933732	0.944363
17	6.0	RandomForest	entropy		sqrt	200.0	300.0	9.0	0.8	0.2	0.842330	0.948748	0.941701	0.944747
18	7.0	RandomForest	entropy		sqrt	100.0	100.0	22.0	0.9	0.1	0.780549	0.939781	0.931559	0.934584
19	8.0	RandomForest	entropy		sqrt	100.0	200.0	13.0	0.9	0.1	0.865979	0.969287	0.941361	0.954711
20	9.0	RandomForest	entropy		sqrt	100.0	300.0	9.0	0.9	0.1	0.843750	0.950804	0.939954	0.945131
21	10.0	RandomForest	entropy		sqrt	200.0	100.0	22.0	0.9	0.1	0.780549	0.940171	0.930608	0.934082
22	11.0	RandomForest	entropy		sqrt	200.0	200.0	13.0	0.9	0.1	0.865979	0.970071	0.941361	0.955154
23	12.0	RandomForest	entropy		sqrt	200.0	300.0	9.0	0.9	0.1	0.849432	0.950850	0.942263	0.946331
24	13.0	RandomForest	entropy		sqrt	260.0	300.0	9.0	0.8	0.2	0.846591	0.951056	0.942296	0.946124
25	14.0	RandomForest	entropy		sqrt	260.0	200.0	13.0	0.8	0.2	0.837419	0.955263	0.934275	0.944500
26	15.0	RandomForest	entropy		sqrt	260.0	100.0	22.0	0.8	0.2	0.784289	0.942955	0.916586	0.928920
27	16.0	RandomForest	entropy		sqrt	260.0	100.0	22.0	0.9	0.1	0.780549	0.941033	0.931559	0.935216
28	17.0	RandomForest	entropy		sqrt	260.0	200.0	13.0	0.9	0.1	0.868557	0.969142	0.942408	0.955294
29	18.0	RandomForest	entropy		sqrt	260.0	300.0	9.0	0.9	0.1	0.849432	0.950941	0.942263	0.946413

Nota: la figura posee los resultados obtenidos en evaluación de los algoritmos de Random Forest y Soporte de Máquina Vectorial - Síntomas en el proceso de entrenamiento con diferentes parámetros en cada uno de los modelos.

Dataset – Recomendaciones

3.9.3. Análisis de los atributos del Dataset

La existencia de dos datasets tiene como objetivo la etiquetación de ambos en base a una entrada, que en cuyo cayo el presente corresponde a la entrada de las recomendaciones, específicamente el atributo **“15. Describa ¿Qué Recomendaciones saludables le dio a conocer su médico? ejemplo: Mi doctor me recomendó que caminará 4 veces por semana por 40 minutos”.** Los atributos

definidos como salida vienen dato por un compendio de las recomendaciones identificadas en la entrada y en cada uno de los registros que conforman esa columna.

3.9.4. Aplicación de técnicas NLP

Una vez obtenida la Dataset etiquetada para las recomendaciones es necesario aplicar técnicas de procesamiento de lenguaje natural a la variable independientes, es decir a la entrada ósea las recomendaciones. La eliminación de palabras de articulación o nexos dejando únicamente las palabras clave es parte de lo que conoce limpieza stopwords el cual consiste en dejar únicamente las palabras que representen una significación alta para el modelo.

3.10. Fase 5

3.10.1. Entrenamiento

El uso de etapa o fases anteriores fue por completo necesario para poder entrar a la creación del modelo en sí. El uso de librerías y sus respectivas clases facilita mucho la implementación del algoritmo implícito en un modelo además de aportar encapsulación de elementos que se simplifican considerablemente lo que se ve reflejado en que no es necesario escribir paso por paso el modelo sino más bien se reduce a unas pocas líneas de código. Los modelos escogidos para su implementación son máquina de soporte vectorial (SVM) y bosques aleatorios (RF).

3.10.2. Definición de vocabulario y vectorización

Las entradas preprocesadas son sin más un conjunto de palabras, que son el resultado de etapas anteriores. Los modelos de inteligencia artificial no entienden palabras, entienden números y es necesario hacer esta transformación que forma parte del insumo necesario para el entrenamiento del modelo. Se realiza la etapa de vectorización que realiza la transformación necesaria, una vez realizado este paso se deben dividir los datos en un conjunto de entrenamiento y otro para el testeo.

3.10.3. Ajuste de Parámetros - SVM

La creación e instanciación del modelo viene dado por clases propias de la librería que reducen en sobremanera el código, sin embargo, es necesario aun especificar el valor de los parámetros de entrada para la creación del modelo. Parámetros como el tipo de núcleo (kernel), tamaño de los datos de entramientos y otros más son especificados antes de la instanciación.

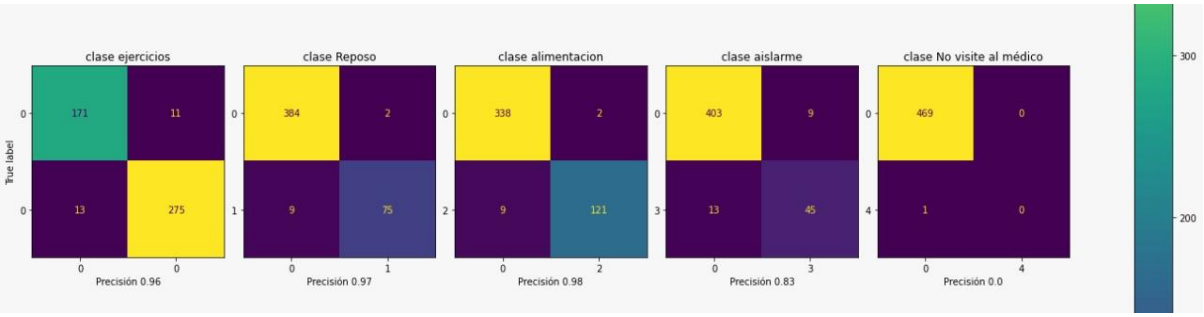
Kernel = linear

Ocurrencia de etiquetas = 300

Tamaño de datos de entrenamiento = 80%

Figura 17:

Matriz de confusión del algoritmo de soporte de máquina vectorial – recomendaciones



Nota: la figura posee los resultados obtenidos en la matriz de confusión de cada una de las etiquetas con el algoritmo de Soporte de Máquina Vectorial de la cual se obtuvieron los mejores resultados.

Ajuste de Parámetros – RF

Criterio = entropy

Max Feacture = sqrt

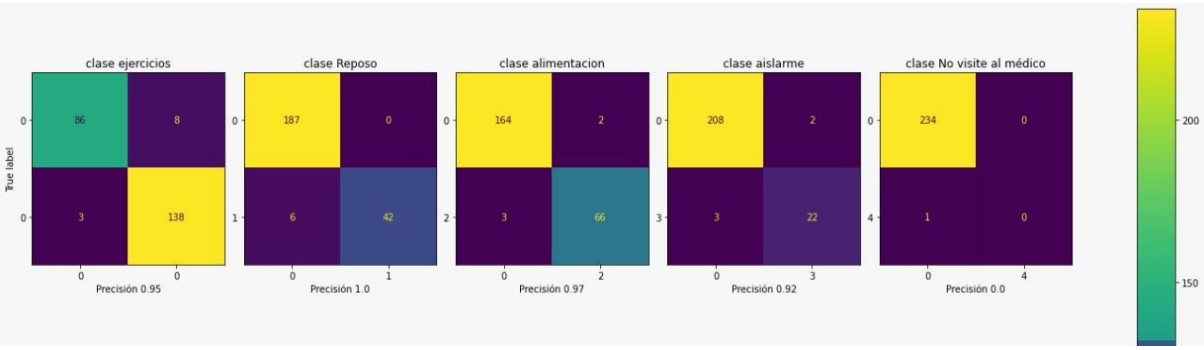
Estimadores = 200

Ocurrencia de etiquetas = 300

Tamaño de datos de entrenamiento = 90%

Figura 18:

Matriz de confusión del algoritmo de Random Forest – recomendaciones



Nota: la figura posee los resultados obtenidos en la matriz de confusión de cada una de las etiquetas con el algoritmo de Random Forest de la cual se obtuvieron los mejores resultados.

3.10.4. Aplicación de algoritmo de aprendizaje

Una vez establecida la parametrización se procede a la parte de entrenamiento que consiste en un código que permite al modelo recibir los datos de entrada y salida de entramiento y así tener un entrenamiento interno con los parámetros establecido con anterioridad.

3.11. Fase 6

3.11.1. Evaluación

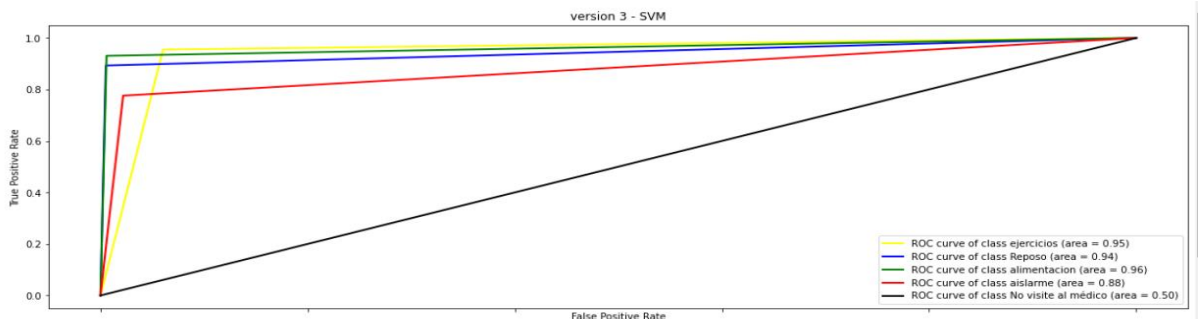
La culminación del proceso de creación del modelo viene dada por la evaluación de los resultados, cuyo análisis es realizado por las métricas de evaluación esto a su vez sirve como comparador entre los resultados de los dos modelos para elección del que posea los resultados óptimos.

3.11.2. Curva ROC

Una métrica de detección efectiva en los resultados arrojados por cada uno de los modelos es la curva ROC, la cual evalúa la sensibilidad de un sistema de clasificación, los resultados son óptimos cuando se acercan lo sumo posible al verdadero positivo (TP) lo que daría como resultado un conjunto de curva de apariencia casi perpendicular para cada una de las etiquetas de la clasificación.

Figura 19:

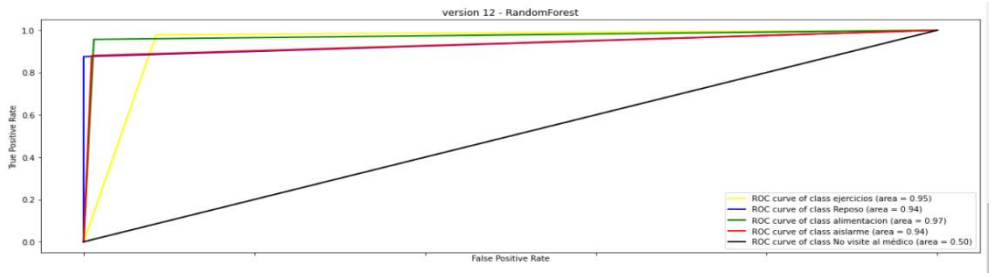
Curva ROC – SVM



Nota: resultados de la curva ROC para SVM en Recomendaciones y cada una de las etiquetas

Figura 20:

Curva ROC - RF



Nota: resultados de la curva ROC para RF en Recomendaciones y cada una de las etiquetas

3.11.3. Métricas de clasificación

Los resultados en métricas de evaluación para el modelo de máquina de soporte vectorial (SVM):

Tabla 13:

Métricas de Evaluación del algoritmo Soporte de Máquina Vectorial - Recomendaciones

Métricas de Evaluación Soporte de Máquina Vectorial			
Acurracy	**Precision**	**Recall**	**F1**
0.874468	0.953584	0.919786	0.936139

Nota: La tabla posee las métricas de evaluación que se obtuvieron en el algoritmo de Soporte de Máquina Vectorial

Los resultados en métricas de evaluación para el modelo de bosques aleatorios (RF):

Tabla 14:

Métricas de Evaluación del algoritmo Random Forest - Recomendaciones

Acurracy	**Precision**	**Recall**	**F1**
0.897872	0.954793	0.943662	0.948333

Nota: La tabla posee las métricas de evaluación que se obtuvieron en el algoritmo de Random Forest

Figura 21:

Matriz de resultados obtenidos en el entrenamiento con Soporte de Máquina Vectorial y Random Forest – Recomendaciones.

	Version	Clasificador	Criterio	kernel	MaxFeatures	Estimadores	Ocurrencia	# columnas	TestSize	Accuracy	Precision	Recall	f1
0	1.0	SVM		linear		0.0	100.0	17.0	0.2	0.776353	0.912367	0.853085	0.879980
1	2.0	SVM		linear		0.0	200.0	10.0	0.2	0.784615	0.914214	0.861872	0.885490
2	3.0	SVM		linear		0.0	300.0	5.0	0.2	0.874468	0.953584	0.919786	0.936139
3	4.0	SVM		linear		0.0	100.0	17.0	0.1	0.763533	0.919564	0.845560	0.878292
4	5.0	SVM		linear		0.0	200.0	10.0	0.1	0.781538	0.911432	0.879908	0.892536
5	6.0	SVM		linear		0.0	300.0	5.0	0.1	0.868085	0.943173	0.926056	0.934429
6	7.0	SVM		rbf		0.0	100.0	17.0	0.2	0.700855	0.941467	0.728697	0.809756
7	8.0	SVM		rbf		0.0	200.0	10.0	0.2	0.720000	0.933596	0.759132	0.827928
8	9.0	SVM		rbf		0.0	300.0	5.0	0.2	0.851064	0.966382	0.889483	0.922678
9	10.0	SVM		rbf		0.0	100.0	17.0	0.1	0.672365	0.950361	0.696911	0.787246
10	11.0	SVM		rbf		0.0	200.0	10.0	0.1	0.710769	0.917473	0.748268	0.811774
11	12.0	SVM		rbf		0.0	300.0	5.0	0.1	0.851064	0.960570	0.908451	0.930424
12	1.0	RandomForest	entropy		sqrt	100.0	100.0	17.0	0.2	0.762108	0.891005	0.846229	0.865107
13	2.0	RandomForest	entropy		sqrt	100.0	200.0	10.0	0.2	0.780000	0.881678	0.873288	0.874910
14	3.0	RandomForest	entropy		sqrt	100.0	300.0	5.0	0.2	0.882979	0.960118	0.918004	0.937410
15	4.0	RandomForest	entropy		sqrt	200.0	100.0	17.0	0.2	0.759259	0.896908	0.846229	0.868690
16	5.0	RandomForest	entropy		sqrt	200.0	200.0	10.0	0.2	0.790769	0.882843	0.876712	0.876514
17	6.0	RandomForest	entropy		sqrt	200.0	300.0	5.0	0.2	0.885106	0.960481	0.921569	0.939692
18	7.0	RandomForest	entropy		sqrt	100.0	100.0	17.0	0.1	0.769231	0.914954	0.847490	0.875277
19	8.0	RandomForest	entropy		sqrt	100.0	200.0	10.0	0.1	0.775385	0.864842	0.891455	0.873079
20	9.0	RandomForest	entropy		sqrt	100.0	300.0	5.0	0.1	0.889362	0.951378	0.940141	0.944950
21	10.0	RandomForest	entropy		sqrt	200.0	100.0	17.0	0.1	0.766382	0.914375	0.841699	0.869397
22	11.0	RandomForest	entropy		sqrt	200.0	200.0	10.0	0.1	0.769231	0.865452	0.882217	0.869084
23	12.0	RandomForest	entropy		sqrt	200.0	300.0	5.0	0.1	0.897872	0.954793	0.943662	0.948333
24	13.0	RandomForest	entropy		sqrt	260.0	300.0	5.0	0.2	0.887234	0.960734	0.923351	0.940837
25	14.0	RandomForest	entropy		sqrt	260.0	200.0	10.0	0.2	0.784615	0.882857	0.873288	0.875045
26	15.0	RandomForest	entropy		sqrt	260.0	100.0	17.0	0.2	0.762108	0.896122	0.848188	0.869141
27	16.0	RandomForest	entropy		sqrt	260.0	100.0	17.0	0.1	0.772080	0.912662	0.845560	0.871580
28	17.0	RandomForest	entropy		sqrt	260.0	200.0	10.0	0.1	0.772308	0.864077	0.884527	0.869336
29	18.0	RandomForest	entropy		sqrt	260.0	300.0	5.0	0.1	0.893617	0.954605	0.940141	0.946530

Nota: la figura posee los resultados obtenidos en evaluación de los algoritmos de Random Forest y Soporte de Máquina Vectorial en el proceso de entrenamiento con diferentes parámetros en cada uno de los modelos.

3.11.4. Herramientas de investigación y recolección de datos

3.11.4.1. Encuesta

La encuesta es un instrumento de la investigación científica, la cual se considera una técnica de recolección de datos mediante de la interrogante de los sujetos cuyo fin es obtener de manera sistemática conceptos claros sobre la problemática de una investigación previamente construida (Falcón et al., 2019).

Para el presente proyecto de investigación se aplicó una encuesta a personas de la zona 8 que comprende a Guayaquil, Duran y Samborondón, sin distinción de sexo, ocupación o estatus, lo cual denomina a la encuesta como de público general siendo únicamente necesario distinguir si ha padecido el covid-19 o no, clasificación que se hace dentro de la misma encuesta en sí, se alcanzó un total de 5568 encuestados y comentarios de redes sociales. De este total se tomaron en cuenta solo aquellos que afirmaron haber dado positivo para Covid lo que dejo un total 4140 registros en el Dataset.

3.11.4.2. Análisis de los resultados

Una vez aplicada la recolección de la información, se procedió a realizar el tratamiento y análisis de la información de una manera ordenada, a través de la tabulación de las encuestas aplicadas a las personas que estuvieron contagiadas de Covid-19, estos

resultados ayudaran de mejor manera a apreciar los resultados obtenidos. Cabe mencionar que esta encuesta fue realizada con la finalidad de obtener información necesaria acerca de las personas que fueron contagiadas por el Covid-19, cuáles fueron sus síntomas y recomendaciones que recibieron para sobrellevar dicha enfermedad, esta información es utilizada para entrenar el modelo de clasificación del presente proyecto de investigación.

3.11.4.3. Interpretación de los resultados

Una vez procesada la información obtenida en las encuestas aplicadas a las personas Contagiadas de Covid-19 los resultados son los siguientes:

Pregunta 1. ¿Ha tenido coronavirus?

Tabla 15:

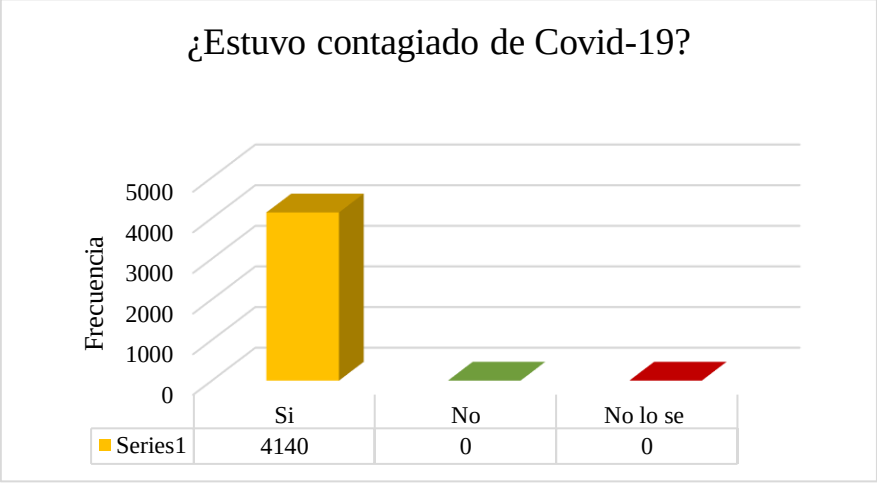
Tabla de frecuencia de la variable Contagiado_Covid-19

	Frecuencia	Porcentaje	Porcentaje Válido	Porcentaje Acumulado
Si	4140	100.00	100.00	100.00
No	0	0.00	0.00	
No lo se	0	0.00	0.00	
Total	**4140**	**100**	**100**	

Nota: Jorge Oviedo Peñafiel e Inés Fajardo Romero

Figura 22:

Frecuencia de la variable Contagiado_Covid-19



Nota: Jorge Oviedo Peñafiel e Inés Fajardo Romero

Análisis: como objetivo principal de la investigación se debe tener en cuenta a todos aquellos que confirmaron ser positivos al Covid-19. El enfoque principal de este

análisis estriba en las personas que dieron positivo de cuales se cuentan a 4140 que representan un 100% total de los encuestados.

Pregunta 2. ¿Cuál es su género?

Tabla 16:

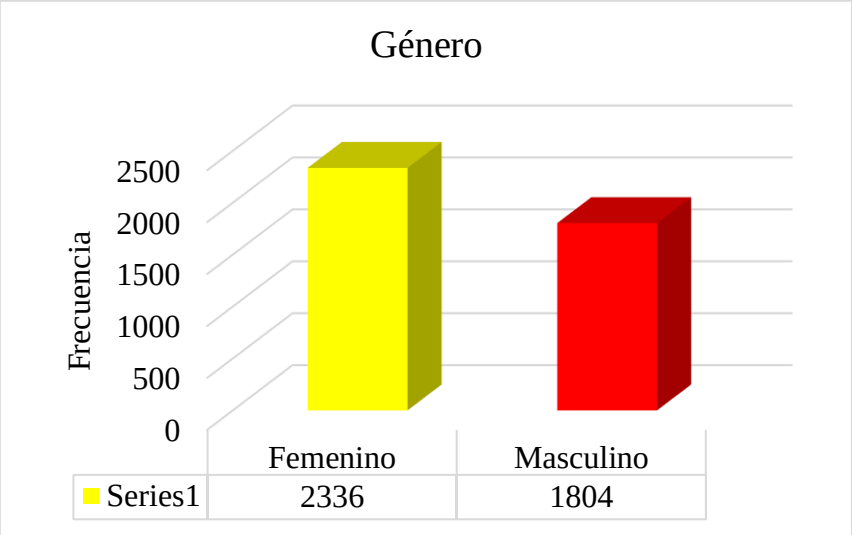
Tabla de frecuencia de la variable Género

	Frecuencia	Porcentaje	Porcentaje Válido	Porcentaje Acumulado
Femenino	2336	56.43	56.43	56.43
Masculino	1804	43.57	43.57	100.00
Total	**4140**	**100**	**100**	

Nota: Jorge Oviedo Peñafiel e Inés Fajardo Romero

Figura 23:

Frecuencia de la variable Género



Nota: Jorge Oviedo Peñafiel e Inés Fajardo Romero

Análisis: Se determinó la variable genero para analizar cuantas personas del género masculino y femenino contrajeron la enfermedad, de la cual se obtiene como resultado que del 100% de los encuestados el 56.43% de personas del género femenino se contagiaron de Covid-19 y un 43.57% de personas del género masculino también dijeron que se habían contagiado de dicha enfermedad.

Pregunta 3. ¿Rango de edad a la que pertenece?

Tabla 17:

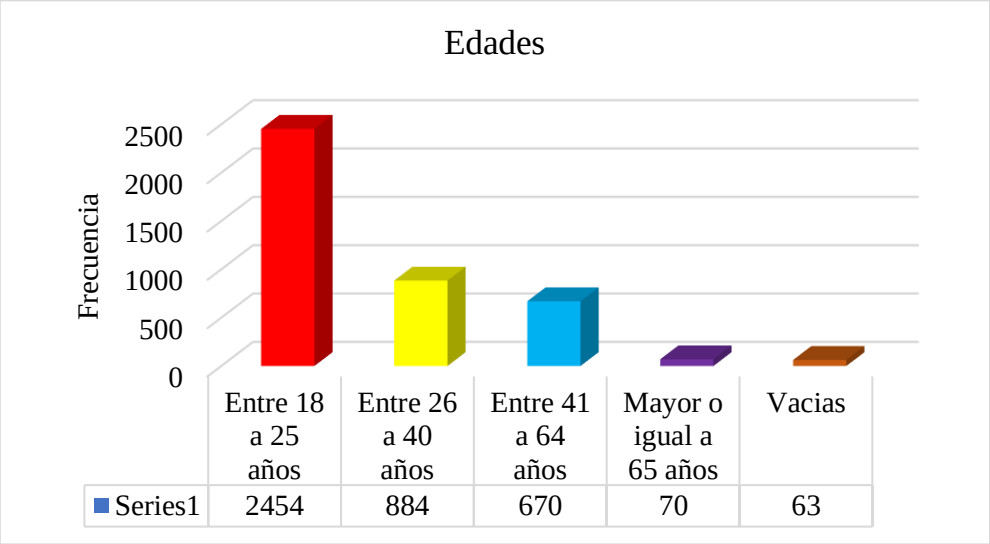
Tabla de frecuencia de la variable Edades

	Frecuencia	Porcentaje	Porcentaje Válido	Porcentaje Acumulado
Entre 18 a 25 años	2454	59.26	59.26	59.26
Entre 26 a 40 años	884	21.35	21.35	80.61
Entre 41 a 64 años	670	16.18	16.18	96.79
Mayor o igual a 65 años	70	1.69	1.69	98.48
Vacías	63	1.52	1.52	100.00
Total	**4141**	**100**	**100**	

Nota: Jorge Oviedo Peñafiel e Inés Fajardo Romero

Figura 24:

Frecuencia de la variable Edades



Nota: Jorge Oviedo Peñafiel e Inés Fajardo Romero

Análisis: Otra de las variables que se tomó en cuenta es la edad de los encuestados ya que según reportes médicos se indicaba edad era un factor importante con respecto al contagio de Covid-19. De esta variable se obtuvieron como resultado que 2454 personas entre 18 a 25 años se contagiaron de Covid-19 representando el 59.26% de las personas encuestadas, sin embargo, también un 21.35% de las personas entre 26 a 40 años confirmaron haber contraído dicha enfermedad, además,

se confirmó que el 1.69% de personas mayor o igual a 65 años también fueron contagiadas.

Pregunta 5. ¿Qué variante del virus lo contagio?

Tabla 18:

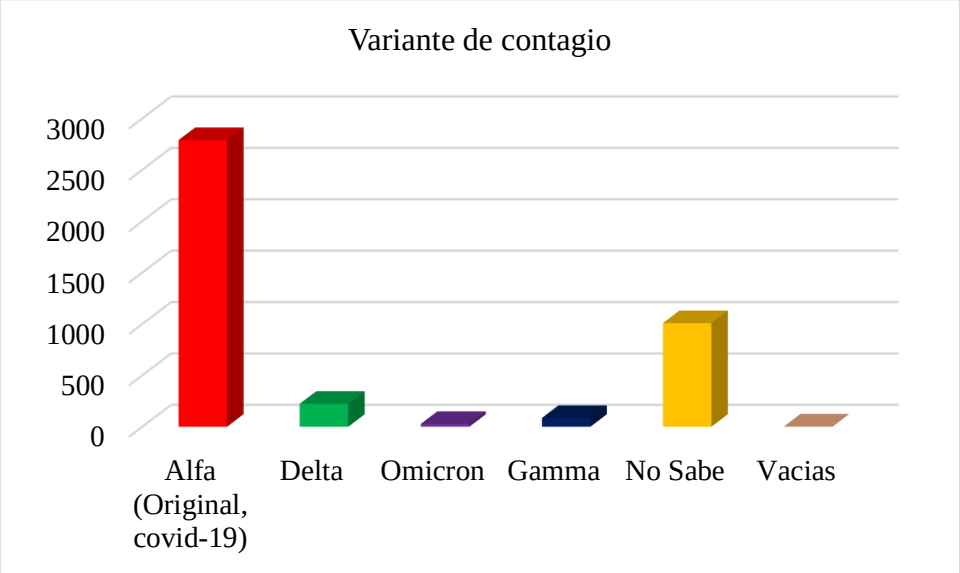
Tabla de frecuencia de la variable variante_contagio

	Frecuencia	Porcentaje	Porcentaje Válido	Porcentaje Acumulado
Alfa (Original, covid-19)	2790	67.39	67.39	67.39
Delta	224	5.41	5.41	72.80
Ómicron	28	0.68	0.68	73.48
Gamma	86	2.08	2.08	75.56
No Sabe	1009	24.37	24.37	99.93
Vacías	3	0.07	0.07	100.00
Total	**4140**	**100**	**100**	

Nota: Jorge Oviedo Peñafiel e Inés Fajardo Romero

Figura 25:

Frecuencia de la variable variante_contagio



Nota: Jorge Oviedo Peñafiel e Inés Fajardo Romero

Análisis: El virus ataca y lo hace en diferentes variantes, hasta el momento de la presente investigación se han detectado 4 variantes a nivel nacional y más específicamente en la zona 8. Una gran mayoría de los encuestados se contagió de la variante Alfa la primera conocida, en menor proporción se tiene que menos del 8% fueron contagiado por alguna de las variantes Ómicron, Delta y Gamma. Se observa

también que casi un 24.37% no tiene conocimiento de que variante fue la que lo contagió.

Pregunta 6. ¿Cuál es el nivel de intensidad que tuvo los síntomas?

Tabla 19:

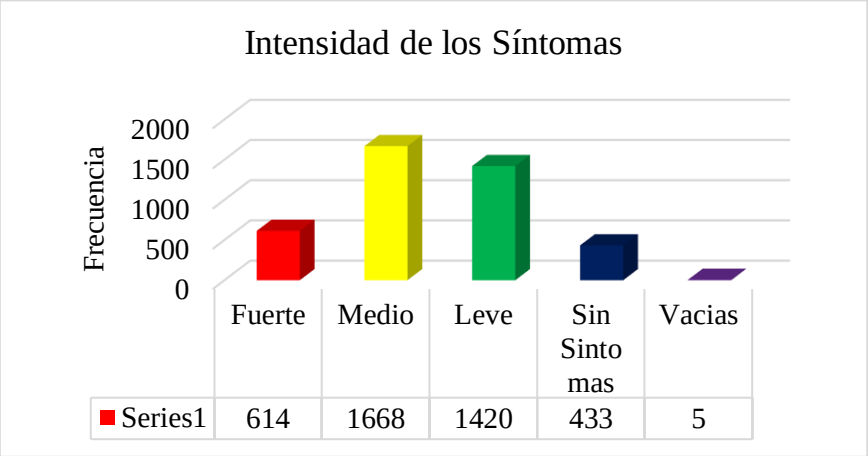
Tabla de frecuencia de la variable intensidad_sintomas

	Frecuencia	Porcentaje	Porcentaje Válido	Porcentaje Acumulado
Fuerte	614	14.83	14.83	14.83
Medio	1668	40.29	40.29	55.12
Leve	1420	34.30	34.30	89.42
Sin Síntomas	433	10.46	10.46	99.88
Vacías	5	0.12	0.12	100.00
Total	**4140**	**100**	**100**	

Nota: Jorge Oviedo Peñafiel e Inés Fajardo Romero

Figura 26:

Frecuencia de la variable intensidad_sintomas



Nota: Jorge Oviedo Peñafiel e Inés Fajardo Romero

Análisis: como parte de la investigación se debe tener en cuenta cual fue la intensidad

con la que se presentaron los síntomas en personas contagiadas por Covid-19. Obteniendo como resultado que a 1668 personas tuvieron síntomas con intensidad media durante el proceso de la enfermedad lo cual representa un 40.27% del total de los encuestados. Aunque no se debe descartar del todo a aquellas personas cuya respuesta fue “Leve” que representan un poco más del 34.30%, del estudio, así como el 14.83% de personas que dijeron tener síntomas fuertes y el 10.46% confirmaron no haber presentado síntoma alguno.

Pregunta 7. ¿En qué lugar o evento considera que se contagió?

Tabla 20:

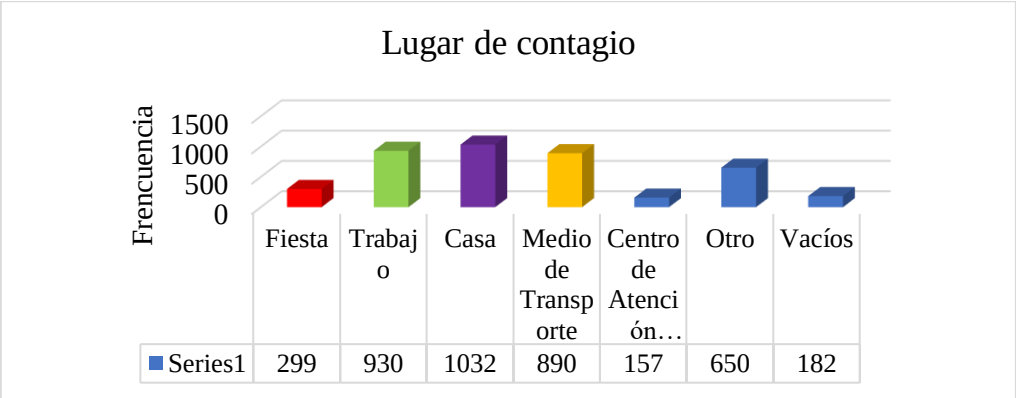
Tabla de frecuencia de la variable lugar_contagio

	Frecuencia	Porcentaje	Porcentaje Válido	Porcentaje Acumulado
Fiesta	299	7.22	7.22	7.22
Trabajo	930	22.46	22.46	29.69
Casa	1032	24.93	24.93	54.61
Medio de Transporte	890	21.50	21.50	76.11
Centro de Atención Médica	157	3.79	3.79	79.90
Otro	650	15.70	15.70	95.60
Vacíos	182	4.40	4.40	100.00
Total	**4140**	**100**	**100**	

Nota: Jorge Oviedo Peñafiel e Inés Fajardo Romero

Figura 27:

Frecuencia de la variable lugar_contagio



Nota: Jorge Oviedo Peñafiel e Inés Fajardo Romero

Análisis: teniendo en cuenta que el virus es altamente contagioso y que se puede contraer dicha enfermedad en diversos lugares se procedió a identificar en que sitios es más probable contraer el Covid-19, a lo cual se obtuvo como resultado que más de 1032 personas fueron contagiadas en sus propios hogares por parientes cercanos lo cual representa a un 24.93% de las personas que respondieron dicha encuesta. A pesar de esto existe un 22.46% de personas que indicaron que contrajeron la enfermedad en el lugar de trabajo, otras indicaron que fueron contagiadas en el medio de transporte siendo estos un 21.50% de las personas encuestadas. Sin embargo, el 15.70% indico haberse contagiado en otro lugar, así como un 7.22% de la población aseguro que se había infectado en una fiesta, una pequeña parte de la población

indico que fueron contagiados en los centros de salud mismo lo cual representa a un 3.79% de las personas que respondieron la presente encuesta.

Pregunta 8. ¿En caso de haber estado vacunado al momento de contagiarse cuantas dosis tenía aplicadas al contagiarse?

Tabla 21:

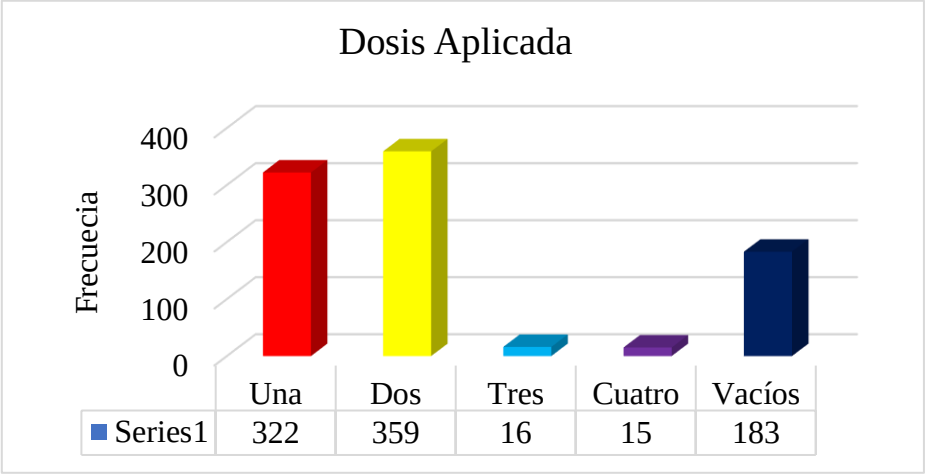
Tabla de frecuencia de la variable dosis_aplicada

	Frecuencia	Porcentaje	Porcentaje Válido	Porcentaje Acumulado
Cero	3245	78.38	78.38	78.38
Una	322	7.78	7.78	86.16
Dos	359	8.67	8.67	94.83
Tres	16	0.39	0.39	95.22
Cuatro	15	0.36	0.36	95.58
Vacíos	183	4.42	4.42	100.00
Total	**4140**	**100**	**100**	

Nota: Jorge Oviedo Peñafiel e Inés Fajardo Romero

Figura 28:

Frecuencia de la variable dosis_aplicada



Nota: Jorge Oviedo Peñafiel e Inés Fajardo Romero

Análisis: en la actualidad muchas personas han sido vacunas y tienen diversos números de dosis aplicadas, es por ello que se tomó en consideración el número de dosis de vacunas aplicadas tiene el encuestado y conocer si la vacuna ayudo a mejorar la inmunidad en las personas, lo cual se logró obtener como resultado que 3245 personas aun no contaban con ninguna dosis de la vacuna aplicada lo cual representa un 78.38% el alto porcentaje en esta respuesta corresponde a que muchos

fueron contagiados al inicio de la pandemia. Sin embargo, el 8.67% confirman haberse contagiado teniendo dos dosis de la vacuna aplicada, hubo personas que aseguraron haberse aplicado una dosis de la vacuna y haberse contagiado siendo estas un 7.78% de personas encuestadas, otro grupo que comprende un 0.39% de la población encuestada dijeron que cuando se contagiaron tenían tres dosis de la vacuna aplicada y un 0.36% de personas confirmaron que fueron contagiadas con 4 dosis de vacuna aplicada.

Pregunta 9. ¿En caso de haber estado vacunado al momento de contagiarse Qué vacuna recibió?

Tabla 22:

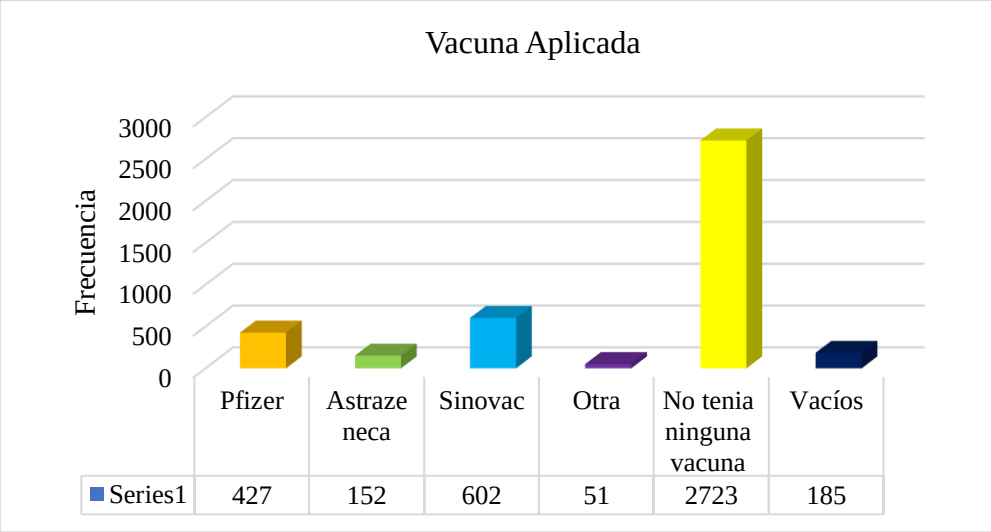
Tabla de frecuencia de la variable tipo_vacuna_aplicada

	Frecuencia	Porcentaje	Porcentaje Válido	Porcentaje Acumulado
Pfizer	427	10.31	10.31	10.31
Astrazeneca	152	3.67	3.67	13.99
Sinovac	602	14.54	14.54	28.53
Otra	51	1.23	1.23	29.76
No tenía ninguna vacuna	2723	65.77	65.77	95.53
Vacíos	185	4.47	4.47	100.00
Total	**4140**	**100**	**100**	

Nota: Jorge Oviedo Peñafiel e Inés Fajardo Romero

Figura 29:

Frecuencia de la variable tipo_vacuna_aplicada



Nota: Jorge Oviedo Peñafiel e Inés Fajardo Romero

Análisis: como parte de esta investigación también se toma como una de las variables el tipo de vacuna aplicada, y de esta manera identificar si las personas

fueron más inmunes antes o después de la vacuna y según que la vacuna que se aplicó. Es por ello por lo que de esta manera se obtuvo como resultado que 2723 personas aun no contaban con ninguna dosis de la vacuna aplicada lo cual representa un 65.77% estos resultados son reflejados con un índice elevado ya que muchos fueron contagiados al inicio de la pandemia y aún no había una vacuna contra el Covid-19. Sin embargo, el 14.54% indicaron tener puesta la vacuna Sinovac, hubo personas que aseguraron haberse aplicado la vacuna Pfizer siendo estas un 10.31% de personas encuestadas, otro grupo que comprende un 3.67% de la población encuestada dijeron que se aplicaron Astrazeneca, un 1.23% de personas confirmaron se aplicaron otro tipo de vacuna y porcentaje mínimo del 4.47% de datos vacíos.

Pregunta 10: Describa lo más detallado ¿Qué síntomas ha tenido?

uno de los factores más relevantes de esta investigación como se ha mencionado es la investigación, preguntas de este tipo de redacción abierta por parte del encuestado permiten mediante su análisis la categorización de las respuestas en si, como por ejemplo de la actual pregunta que dio paso a identificar los principales síntomas presentes en todos aquellos que estuvieron contagiados de covid-19. El haber hecho el proceso mencionado con anterioridad aporta con el objetivo de etiquetación que servirá para entrenar al modelo de clasificación.

Pregunta 11: Describa ¿Qué medicamentos considera que le ayudo en su recuperación?

El uso de medicamentos por parte de las personas que estuvieron contagiadas de covid-19 es un factor para analizar en los resultados de las encuestas. Diversas respuestas indicando el nombre específico de medicamentos y otras señalando medidas alternativas son indicios que permiten hacerse una idea de nombres comunes en todas las respuestas que abran la posibilidad de creación de etiquetas para provecho de entrenamiento del modelo.

Pregunta 12: Describa ¿Qué cuidados aplico durante el proceso de recuperación del Covid-19, y cuánto tiempo en días tomo su recuperación?

El uso de cuestionamientos como estos en el proceso de recolección de información permite obtener una idea general de los métodos y técnicas más comunes empleadas por las personas para hacer frente al virus protagonista de la pandemia actual. Una visión más detallada de las respuestas permite ver que existe gran responsabilidad de la población en general al hacer uso de la atención médica como fuente principal de información para los métodos empleados.

Pregunta 13: Describa a más detalle ¿Qué alimentos y/o vitaminas considera que le ayudaron a fortalecerse y superar el Covid-19?

Las vitaminas y todos aquellos alimentos de reforzamiento inmunológico fueron de vital importancia en muchos casos para las personas que buscaban medidas para hacer frente al virus, nombres comunes en las repuestas tiene un eso y provecho muy parecido a los medicamentos mencionados en la pregunta 11. La identificación de nombres permite la creación de etiquetas para entrenamiento del modelo propuesto. Tener una herramienta que proporcione ayuda en el uso adecuado vitaminas es una solución muy práctica que podría derivar del presente trabajo para un futuro proyecto.

Pregunta 14: De haber superado el covid-19, describa ¿Cómo se siente en su estado de ánimo, autoestima o algún otro malestar que haya usted sentido?

Uno de los factores más relevantes manifestados en la pandemia y que demostró mucho la carencia de metodologías aplicativas para su tratamiento fue la salud mental de las personas que se vio gravemente afectada por temas como el aislamiento, cuarentena, muertes, crisis, etc. la encuesta demostró que la salud mental de las personas en la zona 8 se vio notablemente afectada, teniendo en su mayoría casos de desanimo y depresión, males para los cuales existen muy pocas vías de tratamiento en medio de una crisis mundial.

Pregunta 15: Describa ¿Qué Recomendaciones saludables le dio a conocer su médico?

Siendo este uno de los factores más importantes en la encuesta, las recomendaciones se convierten así en unos de los aspectos principales a ser tomados en cuenta para el procesamiento de la información. Identificar recomendaciones generales dentro de las respuestas también permiten un etiquetado de estas, puntos como el reposo o la actividad física permiten armar un conjunto de recomendaciones que son de provecho en los propósitos de combatir la pandemia.

Pregunta 16: Finalmente, describa ¿Qué información le gustaría que esté disponible fácilmente para ser consultada por usted (con respecto al COVID19)?

El desconocimiento del virus y la enfermedad que esta causa es un mal en sí atribuido en muchos casos a la aun escasa información del covid-19, esta problemática se ha abordado en la encuesta pidiendo a los encuestados mencionar que información quisieran tener acceso. Aspectos como síntomas, tratamientos, medicamentos y noticias se encuentran entre las respuestas más comunes que manifiestan en cierta

medida situación preocupante de ignorancia que pueden derivar en males peores como la renuencia médica y la automedicación.

3.11.5. Entrevista

La entrevista, son una herramienta de investigación científica la cual es utilizada para la recolección de datos, es una de las herramientas más utilizadas para la investigación cualitativa, ya que permite la obtención de datos o información del sujeto de estudio a través de la interacción oral con el investigador (Troncoso-Pantoja & Amaya-Placencia, 2017).

Mediante la intervención de especialista abordados mediante entrevista se pudo obtener respuestas validadas por sus conocimientos, las preguntas pertinentes a esta herramienta tocan temas como modelos, tiempos, visión a futuro y disciplinas de la inteligencia artificial y el procesamiento del lenguaje natural. Este conocimiento adquirido por este medio sirve para un mayor apoyo teórico y empírico en el desarrollo de modelos de clasificación.

Los profesionales entrevistados fueron el ingeniero en mecatrónica Carlos Johnny Gonzales Chancay quien cuenta con más de un año de experiencia en el área, El Magister en Inteligencia Artificial Ricardo Manuel Prieto Galarza quien cuenta con más de 3 años de experiencia y la ingeniera en sistemas computacionales Liseth Estefanía Jiménez Valencia quien tiene una experiencia de más de 4 años.

3.11.6. Análisis de los Resultados

Una vez aplicada la entrevista a expertos de la materia, se procedió a realizar el tratamiento y análisis de la información de una manera ordenada, por medio de la tabulación de las respuestas a las preguntas. Este análisis permite tener una apreciación más detallada de los resultados obtenidos.

3.11.7. Interpretación de los resultados

Una vez procesada la información obtenida de las entrevistas, los resultados son los siguientes:

Pregunta 1: Seleccione el mayor grado de educación que tiene

La pregunta pertinente al nivel de Educación corresponde a la necesidad de establecer aval de estudios de las ciencias correspondientes a la actual investigación. El menor nivel es la de tercer nivel culminada y en la cúspide esta quien posee un doctorado, a nivel medio del total esta quien posee un nivel cuarto de Educación.

Pregunta 1.1: Seleccione el título de educación

Todos los títulos poseídos por los entrevistados están en el rango de especialidades que se inmiscuyen directamente en el estudio de la inteligencia artificial. Desde Ingeniería en sistemas hasta estudios de cuarto nivel. Está concentrada muestra de talento capacitado es de especial provecho para el estudio de la naturaleza de las respuestas posteriores.

Pregunta 1.1.1: Si selecciono la opción "Otro" especifique caso contrario colocar NA

La gama de estudios de los entrevistados se expando a varios en múltiples niveles, dos de los encuestados tienen especialización en la electrónica y mecatrónica, incluse con estudios directos de un cuarto nivel en inteligencia artificial.

Pregunta 1.2: Describir brevemente su experiencia laboral

La variedad de especialidades no es una excepción de los profesionales entrevistados ya que empieza desde electrónica aplicada a IA hasta el desarrollo de software y programación. Todos presentan altos rangos de estudios de la ingeniería con aplicaciones a las ciencias informáticas, poseen varios años de experiencia en la aplicación de la profesión en empresas públicas y trasnacionales.

3.11.8. Pregunta 1.3: Seleccione la Edad

El rango de edades de los entrevistados va desde el más joven con 22 años hasta el más longevo con 33 años cumplidos. Esta generación de jóvenes profesionales son parte de la tendencia del mayor apogeo que tienen las ciencias informáticas en la formación profesional de las personas en la actualidad.

Tabla 23:

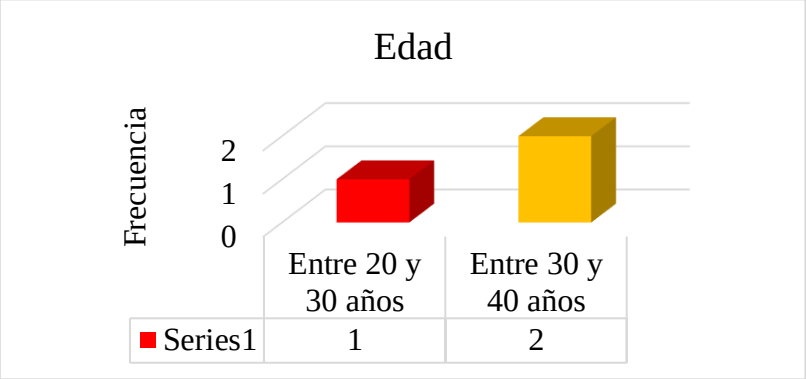
Tabla de frecuencia de la variable edad

	Frecuencia	Porcentaje	Porcentaje Válido	Porcentaje Acumulado
Entre 20 y 30 años	1	33.33	33.33	33.33
Entre 30 y 40 años	2	66.67	66.67	100.00
Total	**3**	**100**	**100**	

Nota: Jorge Oviedo Peñafiel e Inés Fajardo Romero

Figura 30:

Frecuencia de la variable edad



Nota: Jorge Oviedo Peñafiel e Inés Fajardo Romero

3.11.9. Pregunta 1.4: Genero

El género de los entrevistado tiene una representación 2-1: dos masculinos y un femenino.

Tabla 24:

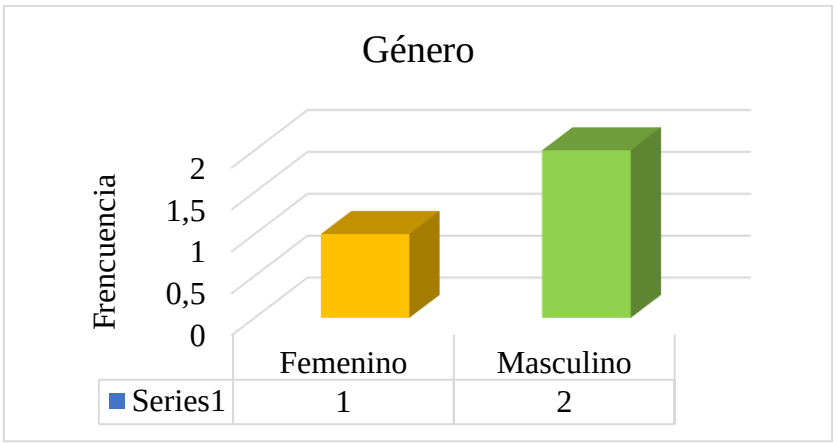
Tabla de frecuencia de la variable género

	Frecuencia	Porcentaje	Porcentaje Válido	Porcentaje Acumulado
Femenino	1	33.33	33.33	33.33
Masculino	2	66.67	66.67	100.00
Total	**3**	**100**	**100**	

Nota: Jorge Oviedo Peñafiel e Inés Fajardo Romero

Figura 31:

Frecuencia de la variable género



Nota: Jorge Oviedo Peñafiel e Inés Fajardo Romero

3.11.10. Pregunta 1.5: Lugar que reside de la zona 8

Del total del entrevistados solo un reside en Guayaquil a la actualidad, en la zona 8, zona de interés del presente estudio. El restante de los profesionales reside en otra ciudad del Ecuador lo que se ve acompañado por su experiencia en otras parte e instituciones del País.

Tabla 25:

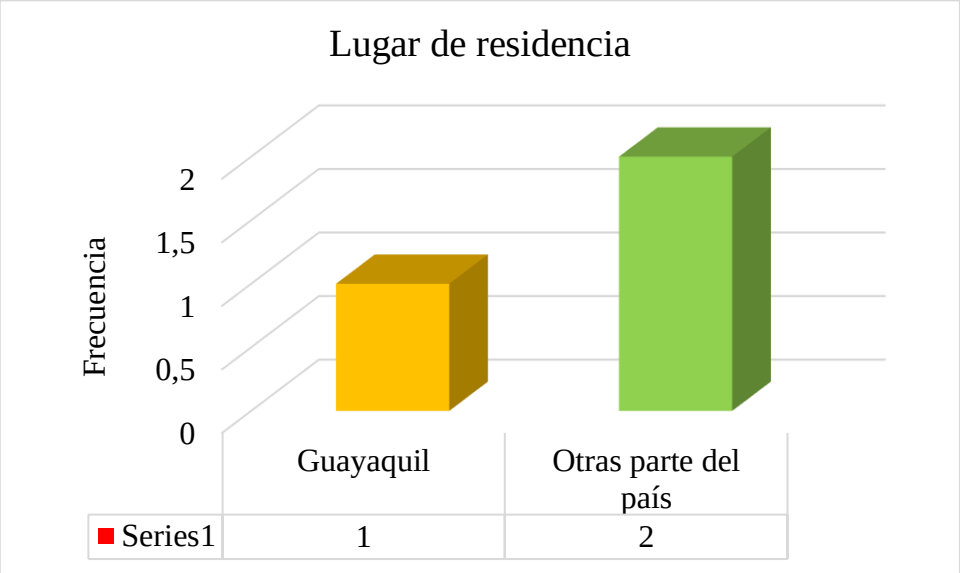
Tabla de frecuencia de la variable lugar_residencia

	Frecuencia	Porcentaje	Porcentaje Válido	Porcentaje Acumulado
Guayaquil	1	33.33	33.33	33.33
Otras partes del país	2	66.67	66.67	100.00
Total	**3**	**100**	**100**	

Nota: Jorge Oviedo Peñafiel e Inés Fajardo Romero

Figura 32:

Frecuencia de la variable lugar_residencia



Nota: Jorge Oviedo Peñafiel e Inés Fajardo Romero

3.11.11. Pregunta 1.6: En caso de haber indicado otra zona de residencia, indique país, provincia, y ciudad (o cantón donde reside)

De todos los casos de los entrevistados, dos de ellos residen en otra provincia, específicamente en Manabí y Azuay que dan razón que estos no habiten en la zona 8.

Pregunta 2: ¿Tiene conocimientos de Inteligencia Artificial?

Todos los entrevistados dieron una respuesta afirmativa a esta cuestión que se presenta como condición necesaria para continuar la entrevista con normalidad. El

hecho de que todos conozcan de la inteligencia artificial abre paso a que los demás cuestionamientos sean respondidos debidamente por una persona competente.

Tabla 26:

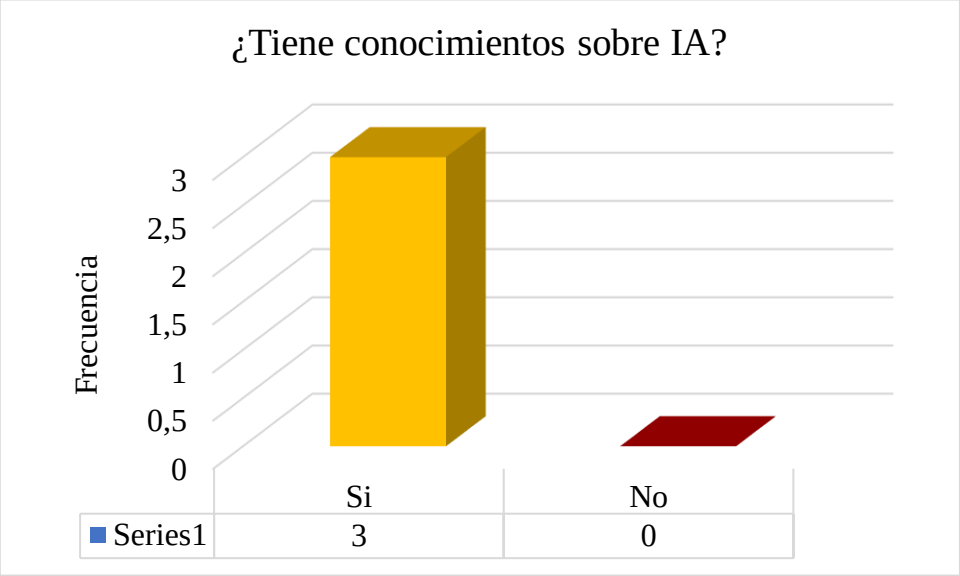
Tabla de frecuencia de la variable conocimiento_IA

	Frecuencia	Porcentaje	Porcentaje Válido	Porcentaje Acumulado
Si	3	100.00	100.00	100.00
No	0	0.00	0.00	100.00
Total	**3**	**100**	**100**	

Nota: Jorge Oviedo Peñafiel e Inés Fajardo Romero

Figura 33:

Frecuencia de la variable conocimiento_IA



Nota: Jorge Oviedo Peñafiel e Inés Fajardo Romero

3.11.12. Pregunta 2.2: ¿Cuánto tiempo en años de experiencia posee trabajando en temas o proyectos de Inteligencia Artificial?

En contraposición a lo que las edades pudieran reflejar no existe una diferencia sustancial en el número de años que tienen de experiencia los entrevistados. Siendo el más bajo un año y el más alto 4 años de experiencia. Lo que no muestra una brecha muy pronunciada entre los profesionales si se toma en especial consideración este aspecto de la entrevista.

Tabla 27:

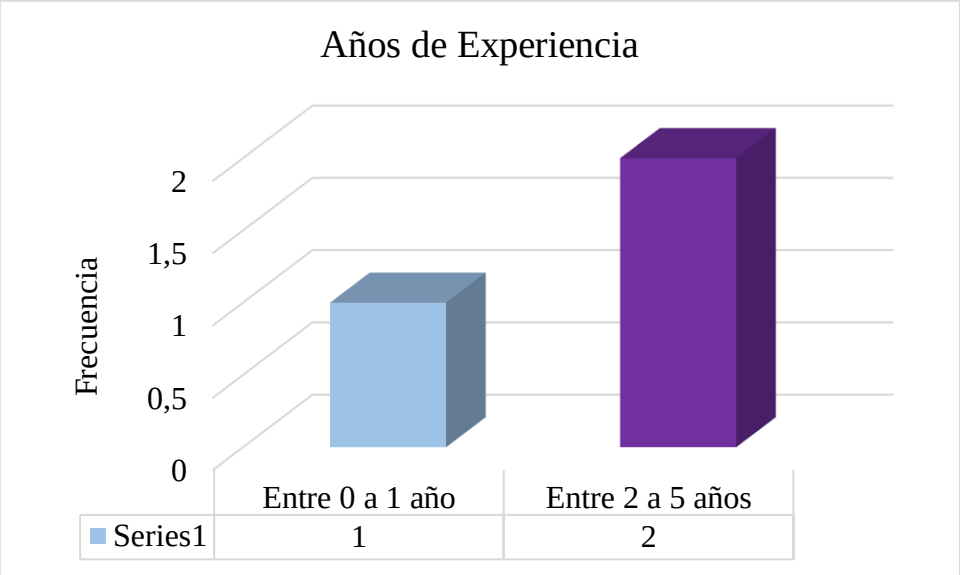
Tabla de frecuencia de la variable años_experiencia_IA

	Frecuencia	Porcentaje	Porcentaje Válido	Porcentaje Acumulado
Entre 0 a 1 año	1	33.33	33.33	33.33
Entre 2 a 5 años	2	66.67	66.67	100.00
Total	**3**	**100**	**100**	

Nota: Jorge Oviedo Peñafiel e Inés Fajardo Romero

Figura 34:

Frecuencia de la variable años_experiencia_IA



Nota: Jorge Oviedo Peñafiel e Inés Fajardo Romero

3.11.13. Pregunta 2.3: ¿Tiene conocimientos de la rama de Inteligencia Artificial llamada Machine Learning (Aprendizaje automático)?

Dos de los tres profesionales entrevistados afirman directamente tener conocimientos de IA en Aprendizaje automático lo cual se ve reflejado en la respuesta de la pregunta 1.2. Uno de los entrevistados manifestó tener un conocimiento a medias de lo abordado en esta pregunta.

Tabla 28:

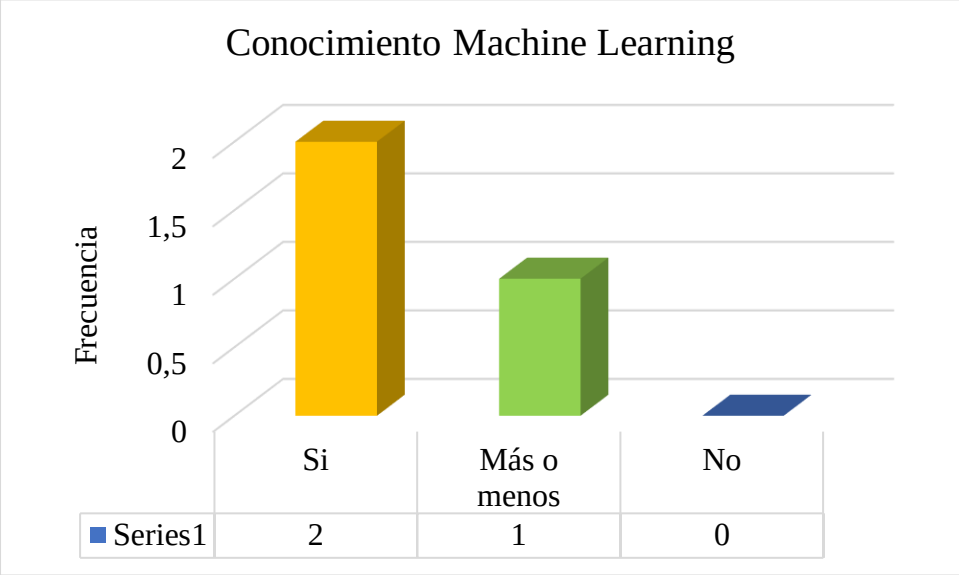
Tabla de frecuencia de la variable conocimiento_machineLearning

	Frecuencia	Porcentaje	Porcentaje Válido	Porcentaje Acumulado
Si	1	50.00	50.00	50.00
Más o menos	1	50.00	50.00	100.00
No	0	0.00	0.00	100.00
Total	**2**	**100.00**	**100**	

Nota: Jorge Oviedo Peñafiel e Inés Fajardo Romero

Figura 35:

Frecuencia de la variable conocimiento_machineLearning



Nota: Jorge Oviedo Peñafiel e Inés Fajardo Romero

3.11.14. Pregunta 2.4: ¿Posee conocimientos de la rama de Inteligencia Artificial denominada Procesamiento de Lenguaje Natural (NLP)?

Todos los entrevistados respondieron afirmativamente a la pregunta de si posee conocimientos del procesamiento de lenguaje natural. La familiaridad de los entrevistados con estas ciencias en cuestión indispensable ya que representa un eje central del presente proyecto.

Tabla 29:

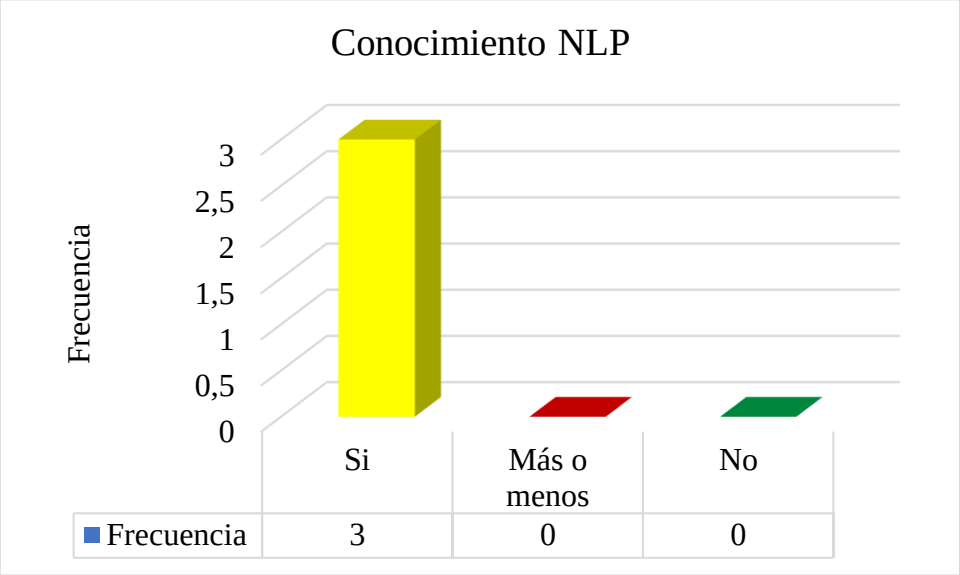
Tabla de frecuencia de la variable conocimiento_NLP

	Frecuencia	Porcentaje	Porcentaje Válido	Porcentaje Acumulado
Si	3	100.00	100.00	100.00
Más o menos	0	0.00	0.00	100.00
No	0	0.00	0.00	100.00
Total	**3**	**100**	**100**	

Nota: Jorge Oviedo Peñafiel e Inés Fajardo Romero

Figura 36:

Frecuencia de la variable conocimiento_NLP



Nota: Jorge Oviedo Peñafiel e Inés Fajardo Romero

3.11.15. Pregunta 3: ¿Qué tan importante considera usted el uso de tecnologías como la Inteligencia Artificial y soluciones de NLP para la superación de la actual de la pandemia?

Todos los entrevistados has manifestado una opinión que deja entrever su percepción de importancia con respecto a la pregunta: Definitivamente es importante. Que esta sea una respuesta general de todos permite seguir manejando el enfoque del proyecto, ya que es muy oportuno que consideren la solución tan importante como se lo tenido en cuenta para esta investigación

Tabla 30:

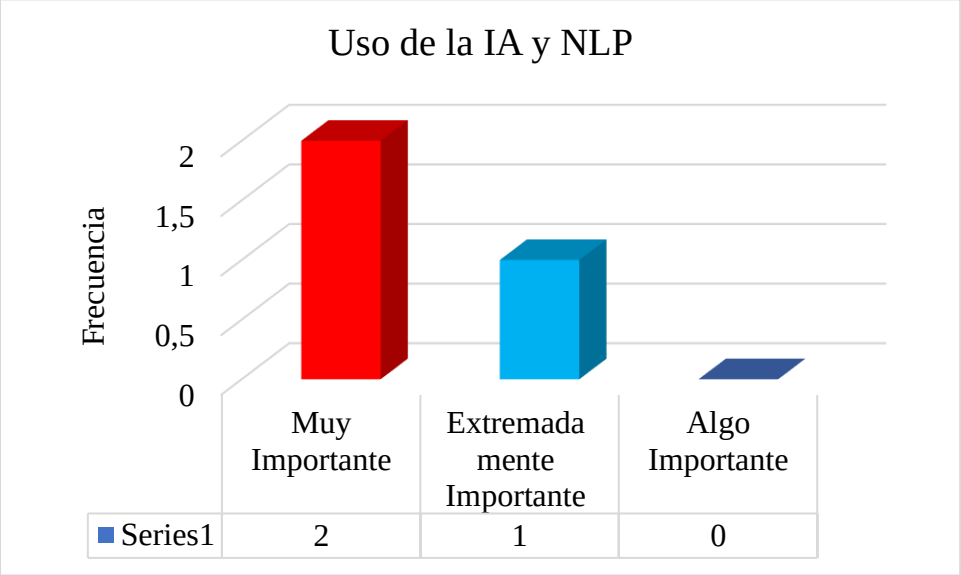
Tabla de frecuencia de la variable uso_IA_NLP

	Frecuencia	Porcentaje	Porcentaje Válido	Porcentaje Acumulado
Muy Importante	2	66.67	66.67	66.67
Extremadamente Importante	1	33.33	33.33	100.00
Algo Importante	0	0.00	0.00	100.00
Total	**3**	**100**	**100**	

Nota: Jorge Oviedo Peñafiel e Inés Fajardo Romero

Figura 37:

Frecuencia de la variable uso_IA_NLP



Nota: Jorge Oviedo Peñafiel e Inés Fajardo Romero

3.11.16. Pregunta 3.2: ¿Describa su opinión de la importancia escogida en la pregunta previa?

La justificación de la pregunta previa a esta es una forma más detallada de entender el contexto personal del entrevistado que lo llevo a determinar su nivel de importancia subjetiva. Aspectos tales como la experiencia, proyectos existentes, opiniones populares o simplemente importancia en el contexto actual son en resumen los argumentos de los entrevistados.

3.11.17. Pregunta 3.3: ¿Qué Algoritmo considera usted más adecuado para usarlo en una arquitectura NLP a ser creada para Clasificación de conversaciones de textos de personas contagiadas de covid-19 (marque los 3 más relevantes)?

Cada persona tuvo una visión sutilmente variada acerca de los algoritmos más importantes para las conversaciones textuales, no obstante, el denominador común de todos ellos radica en el acuerdo de manifestar a las redes neuronales como importante de entre todas sus elecciones. Si bien en la presente pregunta se indica los algoritmos más importantes, solo en la siguiente se pide una justificación.

3.11.18. Pregunta 3.3.1: ¿De los algoritmos escogidos, indique sus motivos o razones por las cuales los escogió?

Basado en diferentes aspectos como la cantidad de datos, complejidad de los algoritmos, experiencia en trabajos anteriores e inclusive tiempos de respuesta, los

entrevistados expresaron que toman en cuenta todo lo mencionado para decantarse por un algoritmo en específico. Caso como árboles de decisiones, Naive bayes y las redes neuronales tiene implementación directa en las experiencias de los profesionales.

3.11.19. Pregunta 3.4: ¿Cuál es su opinión referente a: ¿De la lista de modelos NLP cuál considera usted que seria los más adecuados usarlos con información textual clasificada relacionada con el COVID, para el descubrimiento de tendencias en la población que está enfrentando el Covid-19? (marque los 3 más relevantes)?

Dentro de las opciones del formulario, los más relevantes (o adecuados) según la opinión del entrevistado, son Transformer, Bert y GPT; Resaltando el hecho de uno de los profesionales se abstuvo de opinar justificando esta acción en la respuesta de la pregunta subsecuente a la actual. Los modelos seleccionados por el resto basan su arquitectura en la tecnología Transformer.

3.11.20. Pregunta 3.4.1: Describa el por qué escogió las respuestas de la pregunta anterior

lo común de un modelo dentro de la comunidad de personas dedicadas a la inteligencia artificial puede jugar un papel importante en el número de implementaciones de los proyectos de la comunidad en sí. La entrevista deja entrever esta hipótesis por el hecho de que la justificación en la mayoría de los casos está ligada a que tan conocido es el algoritmo en sí que tantos recursos de ayuda haya para el mismo a la hora de trabajar con él.

3.11.21. Pregunta 3.5: ¿Considera usted que aún se necesita más investigación y nuevas propuestas de construcción de NLP para crear modelos más efectivos de conversaciones de textos con respecto a los existentes relacionados para combatir el Covid-19?

A manera de un acuerdo común entre las partes, la respuesta de los entrevistados es semejante a un acuerdo o entente sobre un asunto en específico, expresando así un acuerdo total acerca de la importancia del desarrollo e investigación de nuevo modelos con optimizaciones en miras a la efectividad. Es de suma importancia también observar el hecho de que cada tiene uno justificación expuesta en la siguiente pregunta.

Tabla 31:

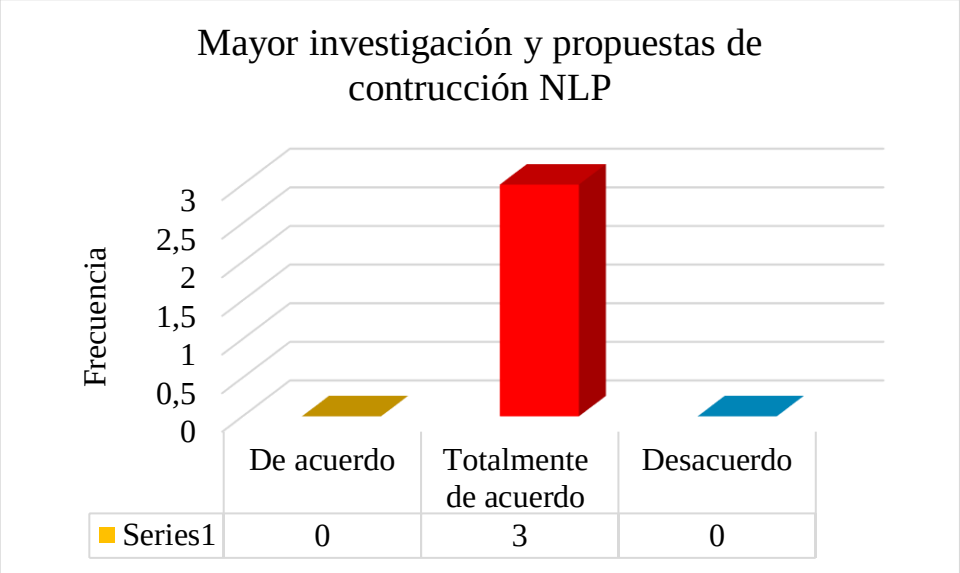
Tabla de frecuencia de la variable mayor_investigacion

	Frecuencia	Porcentaje	Porcentaje Válido	Porcentaje Acumulado
De acuerdo	0	0.00	0.00	0.00
Totalmente de acuerdo	3	100.00	100.00	100.00
Desacuerdo	0	0.00	0.00	100.00
Total	**3**	**100**	**100**	

Nota: Jorge Oviedo Peñafiel e Inés Fajardo Romero

Figura 38:

Frecuencia de la variable mayor_investigacion



Nota: Jorge Oviedo Peñafiel e Inés Fajardo Romero

3.11.22. Pregunta 3.5.1: Describe el porqué de la respuesta anterior

Las respuestas varían desde un enfoque en problemáticas sociales como la pandemia, gobernabilidad y corrupción, hasta conflictos mucho más intrínsecos como lo es la complejidad del lenguaje en sí. Interesante resulta el hecho de que se hace la observación acerca de la ausencia de investigación nacional y escases de trabajo e información para implementación de soluciones en otros idiomas.

3.11.23. Pregunta 4: ¿Sabía usted que uno de los beneficios de aplicar Técnicas de NLP es la simplificación de interacción entre la máquina y el ser humano?

Acerca de esta cuestión relación e interacción humano-máquina, el avance del procesamiento del lenguaje natural representa efectivamente un puente del cual los expertos y profesionales dedicados deben estar completamente informados. En el

caso de las entrevistas existe una adecuada información respecto al tema más aun en aquellos que se han visto inmiscuidos en trabajos pertinentes a la esta ciencia.

Tabla 32:

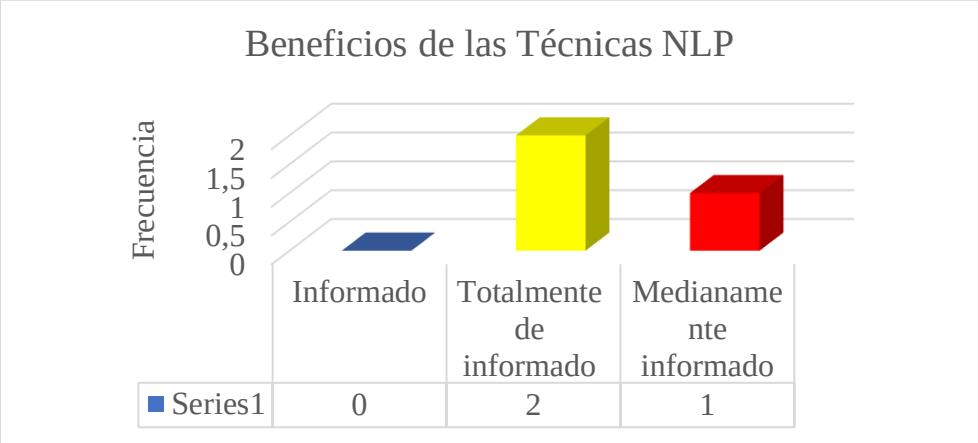
Tabla de frecuencia de la variable técnicas_NLP

	Frecuencia	Porcentaje	Porcentaje Válido	Porcentaje Acumulado
Informado	0	0.00	0.00	0.00
Totalmente de informado	2	66.67	66.67	66.67
Medianamente informado	1	33.33	33.33	100.00
Total	**3**	**100**	**100**	

Nota: Jorge Oviedo Peñafiel e Inés Fajardo Romero

Figura 39:

Frecuencia de la variable técnicas_NLP



Nota: Jorge Oviedo Peñafiel e Inés Fajardo Romero

3.11.24. Pregunta 4.1: De la lista de Técnicas de procesamiento de lenguaje natural NLP ¿Cuál cree usted que funciona mejor para el manejo de información textual clasificada relacionada con el Covid-19? (marque los 3 más relevantes)

En la lista se mencionó las principales técnicas, permitiendo a los profesionales elegir más de una (aunque en un caso fueron incluso más de tres) resultando que en común que todos concordaban en que la segmentación de palabras es de importancia general. Así mismo caso como la tokenización y la técnica MER estuvieron entre las más elegidas.

3.11.25. Pregunta 4.1.1: Describa el porqué de la respuesta anterior

La temática de esta justificación radica mayoritariamente en el hecho de que la segmentación y tokenización deber estar implementada transversalmente a la necesidad de abarcar la mayor información posible. Hay quien expresa que todas las

técnicas expuestas en la lista son importantes por lo cual no señala tres en específico sino más bien antepone la necesidad para remarcar la necesidad de cada una en una situación en específico.

3.11.26. Pregunta 4.2: ¿Considera usted que en futuros proyectos o trabajos investigativos será de utilidad el análisis de técnicas de procesamiento de lenguaje natural NLP para clasificación de texto de conversaciones textuales de personas contagiadas con Covid-19?

Totalmente de acuerdo es la respuesta común de todos los entrevistados quienes arguyen de forma variada en la siguiente pregunta. Sin embargo, esta situación es predecible en cierta manera ya que la importancia de la investigación es un hecho que a cualquier profesional le resulta innegable en exponer su importancia.

Tabla 33:

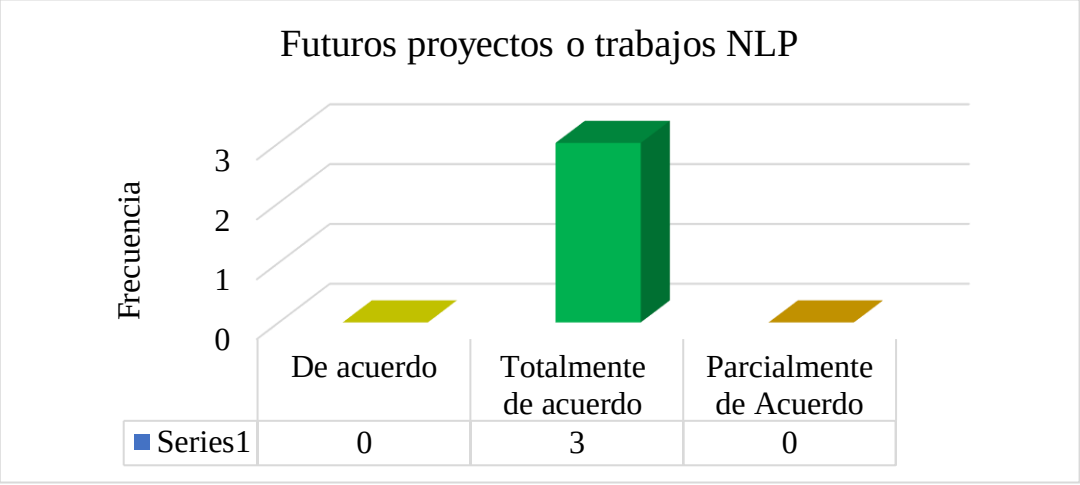
Tabla de frecuencia de la variable futuras_investigaciones

	Frecuencia	Porcentaje	Porcentaje Válido	Porcentaje Acumulado
De acuerdo	0	0.00	0.00	0.00
Totalmente de acuerdo	3	100.00	100.00	100.00
Parcialmente de Acuerdo	0	0.00	0.00	100.00
Total	**3**	**100**	**100**	

Nota: Jorge Oviedo Peñafiel e Inés Fajardo Romero

Figura 40:

Frecuencia de la variable futuras_investigaciones



Nota: Jorge Oviedo Peñafiel e Inés Fajardo Romero

3.11.27. Pregunta 4.2.1: Describa el porqué de la respuesta anterior

Aunque todas las respuestas fueron de naturaleza afirmativa acerca de la importancia de la investigación, uno de los entrevistados recalcó un punto muy interesante: A nivel

nacional hacer este tipo de investigaciones es extremadamente difícil debido al hecho que las organizaciones no manejan la cultura de los datos que permita avanzar en proyectos para la aplicación de IA. La forma primitiva de manejo de la información es aún un obstáculo para el desarrollo de tecnologías a nivel nacional.

3.12. Metodología de desarrollo del proyecto

Para el desarrollo del presente proyecto se realizó en base a la metodología del modelo de prototipado, se define esta metodología por su proceso iterativo el cual se enfoca en diseñar, medir y ajustar el desarrollo de un proyecto.

En esta metodología se lleva a cabo mediante 5 etapas la cuales se mencionan a continuación:

3.12.1. Recolección y refinamiento de requisitos

En el desarrollo del presente proyecto se realizó el levantamiento de los requerimientos finales:

- Se requiere que el sistema de clasificación de conversacional por texto sea implementado en un sitio web.
- Se requiere cuatro módulos, los cuales consisten en el inicio de la página web con una pequeñas definiciones y descripciones del proyecto en mención, además un apartado de clasificación de síntomas y recomendaciones, también un módulo de datos estadísticos y encuéntranos.
- Se requiere que se clasifique la información de síntomas o recomendaciones ingresada por el paciente o el médico en un cuadro de texto.
- Los datos deben ser depurados con minería de datos.
- Se requiere que se presente un apartado de ver más una vez realizada la clasificación en el cual se muestre la curva de ROC y la matriz de confusión tanto de síntomas como recomendaciones.
- Se debe implementar en con el lenguaje Python en un entorno virtual de Flask.

3.12.2. Diseño del prototipo

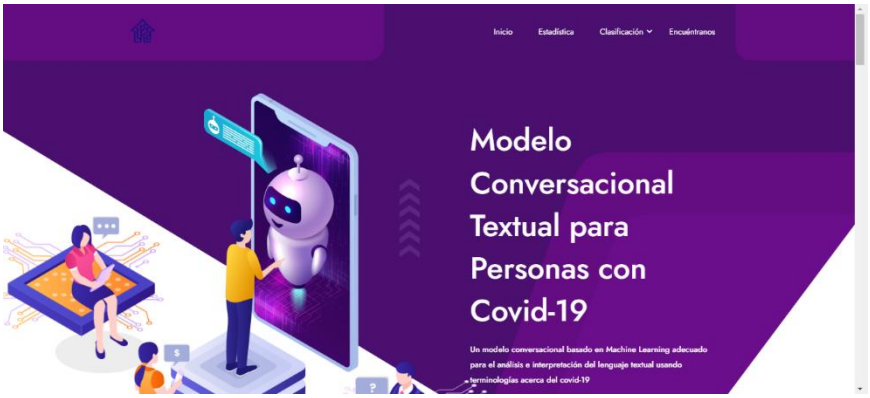
El diseño de la página web se divide en varias secciones como lo son:

3.12.2.1. Inicio

En esta sección se podrá visualizar información relevante acercar del proyecto, conceptos claves, beneficios, testimonios, etc. Así como se puede visualizar en la siguiente figura

Figura 41:

Pestaña de Inicio de prototipo web



Nota: Fajardo Romero Inés, Oviedo Peñafiel Jorge

Figura 42:

Sección del inicio – acerca del proyecto

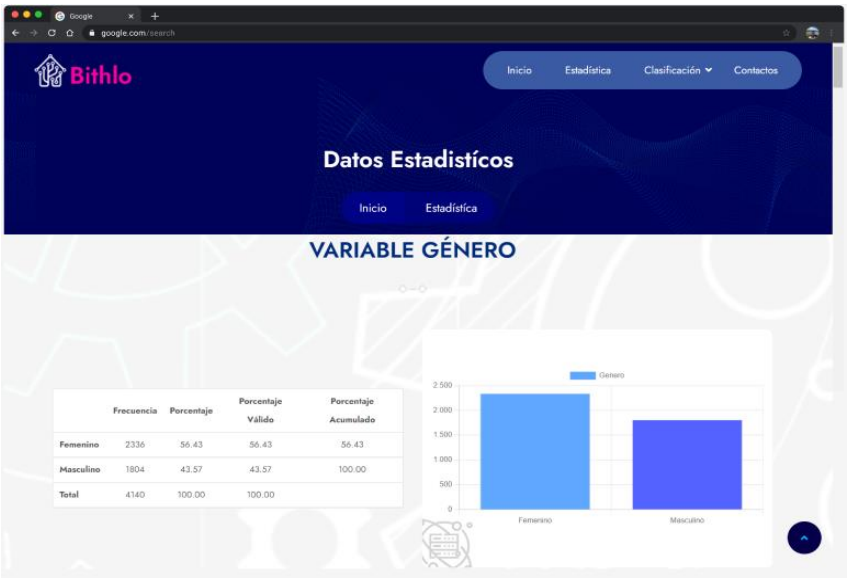
Nota: Fajardo Romero Inés, Oviedo Peñafiel Jorge

3.12.2.2. Estadísticas

En esta pestaña se encuentra las gráficas estadísticas y las tablas con la información general como sexo, numero de vacunas, lugares de contagio etc., datos recolectados mediante encuestas para el entrenamiento del modelo.

Figura 43:

Pestaña de estadísticas



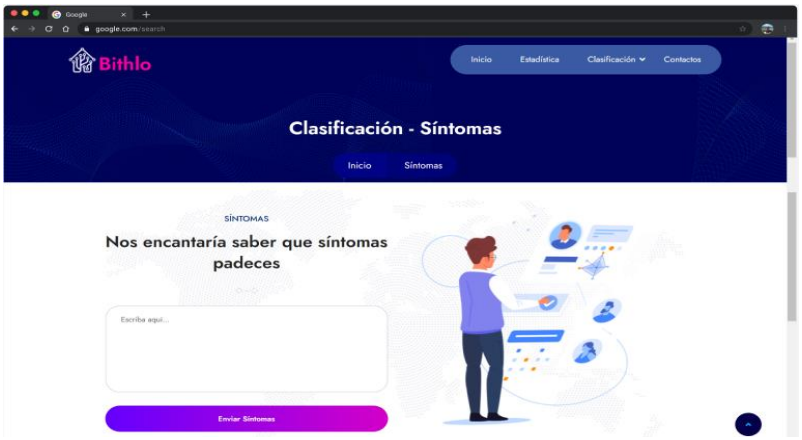
Nota: Fajardo Romero Inés, Oviedo Peñafiel Jorge

3.12.2.3. Clasificación

En esta sección tenemos dos opciones las cuales son los síntomas y las recomendaciones. En este apartado se visualiza un pequeño formulario en el cual sirve para el ingreso de síntomas y recomendaciones para que sean clasificados de manera adecuada.

Figura 44:

Pestaña de Clasificación – Síntomas



Nota: Fajardo Romero Inés, Oviedo Peñafiel Jorge

Figura 45:

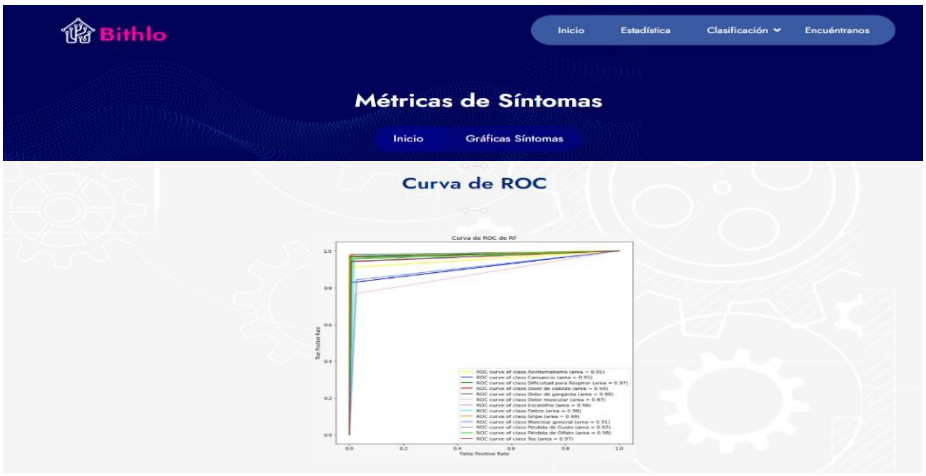
Pestaña de Clasificación – Recomendaciones

Nota: Fajardo Romero Inés, Oviedo Peñafiel Jorge

Dentro de este apartado también encontramos la opción para visualizar las métricas de evaluación tanto para síntomas y recomendaciones.

Figura 46:

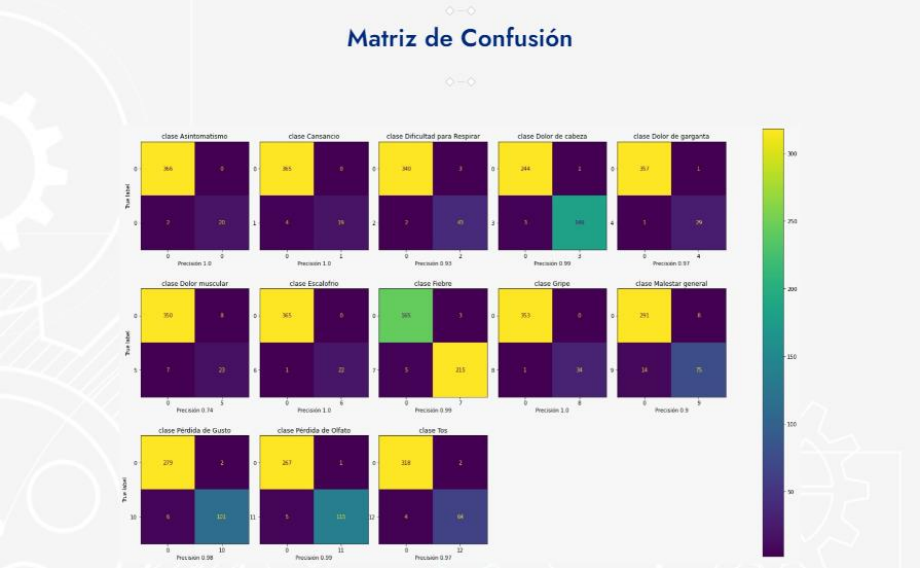
Pestaña de Métricas Síntomas – Curva Roc



Nota: Fajardo Romero Inés, Oviedo Peñafiel Jorge

Figura 47:

Pestaña de Métricas Síntomas – Matriz de Confusión



Nota: Fajardo Romero Inés, Oviedo Peñafiel Jorge

3.12.2.4. Construcción del prototipo

A continuación, en este punto se presentará el código por fragmento detallando cada una de las partes claves para la implementación del modelo en la página web.

```
#Importación del modelo síntomas
clf_rf_sintomas = load('sintomas\\rf\RandomForest.joblib')
rf_sintomas_y_test = pd.read_excel('sintomas\\rf\RandomForest-y_test.xlsx')
rf_sintomas_vectorizador = pickle.load(open("sintomas\\rf\RandomForest-vectorizador.pickel", "rb"))

#Importación del modelo recomendaciones
clf_rf_recomendaciones = load('recomendaciones\\rf\RandomForest-Recomendaciones.joblib')
rf_recomendaciones_y_test = pd.read_excel('recomendaciones\\rf\RandomForest-Recomendaciones-y_test.xlsx')
rf_recomendaciones_vectorizador = pickle.load(open("recomendaciones\\rf\RandomForest-Recomendaciones.pickle", "rb"))
```

Los datos necesarios para la elaboración del caso de prueba puntual están presentes en formato de archivo Excel por lo que es útil las funciones de la librería para los Dataframe, importarlos desde su lugar de almacenamiento en el servidor para

obtener información de las categorías, consecuentemente a esto la importación del modelo ya entrenado se realiza con la librería pickle, es importante también importar el vectorizador que es necesario para alimentar con la entrada al modelo.

```
#Proceso de clasificación
def clasificador():
    try:
        body = request.get_json()
        validarRequestBody(body)
        variablesInicio.definirVariablesModelo(body['modelo'], body['dataset'])

        entrada = procesarEntrada(body['entrada'])
        x_entry = vectorizarEntrada(entrada)
        y_predict = clasificar(x_entry)
        columnas, salidas = obtenerResultados(y_predict)

        res = {"columnas": columnas, "entrada": body['entrada'],
               "salidaModelo": y_predict[0].tolist(), "prediccion": salidas}

        return {"codigo": 0, "mensaje": "Operación Realizada con éxito", "Respuesta": res}
    except ValueError as e:
        return {"codigo": random.randint(1, 100), "mensaje": str(e), "body": None}
```

La comunicación con el modelo entrenado se realiza mediante la exposición de un punto de acceso por protocolo rest, por lo que es necesaria la creación de una función en el framework Flask para exponer la funcionalidad con determinados parámetros de entrada. Se estableció el método POST y un objeto Json de entrada que contenga el algoritmo y el Dataset que se desea que procese la entrada, además de poseer la entrada en sí mismo, posterior a esto ocurre una etapa de validación para el correcto formato de los datos para seguir a un procesamiento y normalización de la entrada. Es de suma importancia la vectorización para pasar a la clasificación.

```
#Procesamiento de la entrada
def procesarEntrada(entrada):
    entrada_limpia = clean_text(entrada)
    entrada_separada_espacios = entrada_limpia.split(" ")
```

```
    important_words = filter(lambda x: x not in stopwords.words(
        'spanish'), entrada_separada_espacios)
    entry = " ".join(list(important_words))
    return entry
def vectorizarEntrada(entrada):
    entrada_vectorizada = variablesInicio.vectorizador.transform(
        [entrada]).toarray()
    return entrada_vectorizada
def clasificar(entrada_vectorizada):
    y_predict = variablesInicio.clasificador.predict(entrada_vectorizada)
    return y_predict
def clean_text(text):
    text = str(text)
    text = text.lower()
    text = re.sub('á', 'a', text)
    text = re.sub('é', 'e', text)
    text = re.sub('í', 'i', text)
    text = re.sub('ó', 'o', text)
    text = re.sub('ú', 'u', text)
    text = re.sub('\[.*?¿\]\%', ' ', text)
    text = re.sub('[%s]' % re.escape(string.punctuation), ' ', text)
    text = re.sub('\w*\d\w*', '', text)
    text = re.sub('[‘“”…«»]', '', text)
    text = re.sub('\n', ' ', text)
    text = re.sub(r'\s+', ' ', text)
    text = re.sub('[^a-zA-Z]', ' ', text)
    return text
```

El algoritmo que se acaba de mostrar permite procesar la información de entrada, realizando un preprocesamiento antes de ser enviada a la clasificación respectiva.

La creación de múltiples funciones para modularizar las tareas de manera que tengas naturaleza atómica es parte de la estructura del trabajo, se crearon funciones para el procesamiento y normalización de la entrada, para la validación del objeto en la solicitud de procesamiento, la vectorización de la entrada y la clasificación en sí.

3.12.3. Evaluación del prototipo

En esta etapa se evalúo el prototipo su interactividad y funcionamiento el cual corresponda a los lineamientos planteados inicialmente de tal forma que los resultados obtenidos en tiempo real sean los correctos. Además, se realizó observaciones las cuales se ajustarán al prototipo del diseño web.

3.12.4. Refinamiento del prototipo

En esta etapa se aplicaron todas aquellas observaciones realizadas durante la evaluación del prototipo web, para mejorar tanto su la funcionalidad como el contenido de esta.

3.13. Entregables del proyecto

A continuación, se indica los entregables del presente proyecto de titulación, los cuales son los siguientes:

- **Código fuente del modelo de clasificación:** el código fuente del modelo desarrollado ser entregado de manera digital.
- **Artículo científico:** el articulo científico que detalla de manera técnica la composición y funcionamiento del modelo.
- **Base de datos:** se entregará de manera digital la base de datos de síntomas y recomendaciones con su respectivo etiquetado.
- **Trabajo de Titulación:** el presente trabajo de titulación en el cual se detalla de manera más específica todo lo relacionado con el diseño de la solución.

3.14. Beneficiarios directos e indirectos del proyecto

3.14.1. Directos

Los beneficiarios directos del presente proyecto son todas aquellas personas que deseen conocer acerca de los síntomas y recomendaciones del Covid-19.

Además, para todas aquellas personas que en un futuro quieran aprovechar los beneficios de este modelo, para el desarrollo de soluciones que aporten contra la lucha del Covid-19.

3.14.2. Indirectos

Se identifica como beneficiarios indirectos a los médicos que pertenecen a la Zona 8 que comprende Guayaquil, Durán y Samborondón y médicos en general.

3.15. Propuesta

En este presente proyecto se presenta un conjunto de datos que está compuesto de 4140 síntomas y recomendaciones pertenecientes a todas aquellas personas de la Zona 8 del Guayas que comprende Guayaquil, Durán y Samborondón, estos datos fueron recolectado por medio de encuestas y las redes sociales como lo son Facebook y Twitter.

El proceso de clasificación textual con terminologías del Covid-19 se desarrollará mediante la creación de un modelo de Procesamiento de Lenguaje Natural basado en Machine Learning, el cual permitirá la clasificación de texto escrito acerca de síntomas y recomendaciones que presentaron todas aquellas personas con Covid-19. Además se realizará un aplicativo web que permitirán a pacientes y médicos interactuar con el modelo de una manera fácil y sencilla y de esta manera redactar sus síntomas y recomendaciones saludables que siguieron durante el proceso de la enfermedad, de esta manera mediante la ejecución del modelo entrenado y los datos de entrada se realizara la clasificación de la información ingresada y su respectiva evaluación de los resultados obtenidos a través de las métricas de evaluación como lo son la Curva ROC y matriz de confusión.

3.15.1. Tratamiento del Dataset de síntomas y recomendaciones

Para la recolección del Dataset de síntomas y recomendaciones se realizó encuestas con personas que padecieron de Covid-19 de la Zona 8 del Guayas, además de ello se recolecto información de las redes sociales como lo son Twitter y Facebook. Una vez recolectada la información se llevó a cabo el tratamiento de esta con la corrección de los datos y el etiquetado de 4140 registros recopilados.

Una vez concluido el proceso de corrección y etiquetado se realizó el procesamiento de los datos en el cual se definían las columnas que se iban a utilizar para el entrenamiento de los datos, se definió el tipo de variables, se eliminaron mediante código tildes y caracteres especiales, etc. Finalmente, luego de este arduo proceso se logró obtener un Dataset listo para el proceso de entrenamiento.

3.15.2. Técnicas de NLP

Las técnicas NLP que se utilizaron en el presente proyecto son las de stopwords y vectorización. Los stopwords se encargaron de filtrar las oraciones añadidas por los encuestados acerca de los síntomas y recomendaciones ya que dentro de estas existen palabras que no tienen un significado relevante para tomar en

consideración al momento del entrenamiento. La vectorización esta técnica fue utilizada para la transformación de la información escrita en texto a vectores para facilitar el consumo y el entendimiento por parte del algoritmo de aprendizaje automático.

3.15.3. Algoritmos de Machine Learning

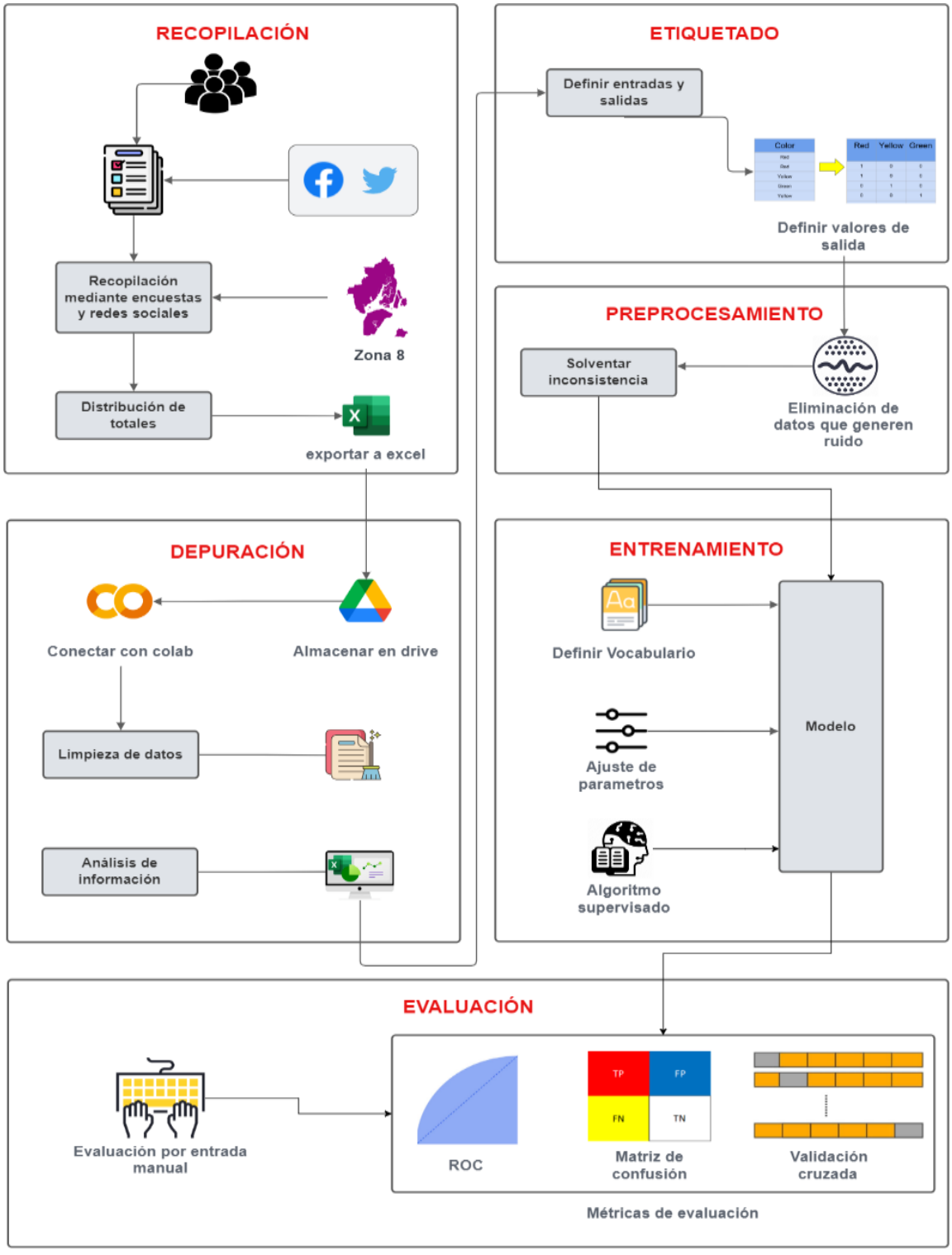
Los algoritmos de Machine Learning utilizados en el presente trabajo de titulación son los algoritmos de Soporte de Máquina Vectorial los cuales se enfocó en aprender a discriminar entre miembros positivos y negativos de las n-etiquetas dadas para su respectivo entrenamiento, sin embargo, el algoritmo de Random Forest se enfocó en el crecimiento y la combinación de múltiples árboles y así generar un bosque de datos en el cual se crea un camino hacia las mejores decisiones o resultados.

3.15.4. Arquitectura del modelo NLP

En la siguiente figura se visualiza la arquitectura planteada para el presente proyecto de investigación donde se puede observar a breves rasgos todos los elementos que conforman parte de la creación del modelo de Procesamiento de Lenguaje Natural basado en Machine Learning.

Figura 48:

Diseño arquitectónico del modelo NLP



Nota: Se muestra la arquitectura implementada para el desarrollo del modelo

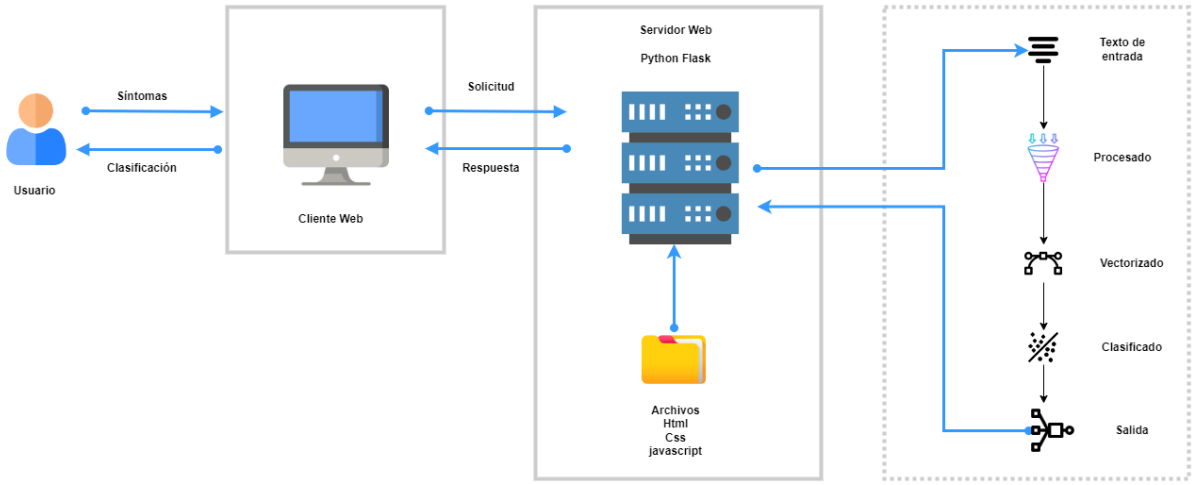
Fuente: Fajardo Romero Inés, Oviedo Peñafiel Jorge

3.15.5. Arquitectura del aplicativo web

En la ***figura 28*** se puede observar la arquitectura planteada para el desarrollo del aplicativo web del proyecto de investigación donde se puede visualizar los elementos que forman parte de este.

Figura 49:

Diseño arquitectónico del aplicativo web



Nota: Se muestra la arquitectura implementada para el desarrollo del aplicativo web

Elaboración: Fajardo Romero Inés, Oviedo Peñafiel Jorge

3.16. Criterios de validación de la propuesta

En el presente proyecto, se hace uso de la técnica de juicio de expertos para la validar la propuesta del proyecto. Para la realización de este juicio se ha recurrido a la presentación del modelo entrenado, el aplicativo web y posteriormente la calificación respectiva por los diferentes profesionales con experiencia en el área de Ingeniería en Sistemas, Ciencias de Datos e Inteligencia Artificial.

Para la validación de la propuesta contamos con la presencia de cuatro expertos, con los cuales para la presentación del proyecto se realizaron sesiones de manera virtual por medio de la plataforma Teams de Microsoft. A continuación, en la ***tabla 34*** detallamos la información de cada uno de ellos.

Tabla 34:

Acerca de los expertos

N°	Apellidos y Nombres	Título Profesional	Nacionalidad
1	Sangacha Tapia Lady Mariuxi	Máster Universitario En Ingeniería De Software Y Sistemas Informáticos	Ecuatoriana
2	De Anda De la Torre Nayelhi Yajaira	Maestra en Big Data y Data Science	Mexicana
3	Ordonez Chávez Frank Emerson	Ingeniero en Sistemas Computacionales	Ecuatoriana
4	Navarro Zurita Henry Alberto	Ingeniero en Sistemas Computacionales	Ecuatoriana

Nota: Jorge Oviedo Peñafiel e Inés Fajardo Romero

A continuación, en la **Tabla 35**, se presentan los indicadores y criterios que fueron tomados en cuenta al momento de realizar la validación de los expertos.

Tabla 35:

Indicadores y Criterios

N°	Indicador	Criterio
1	Usabilidad	La interfaz de interacción con el modelo es de fácil uso y sencilla
2	Capacidad de Respuesta	Los tiempos de respuestas no son muy elevados
3	Robustez	El modelo fue desarrollado en un lenguaje de programación de amplio uso
4	Open Source	Para el desarrollo del modelo se hizo uso de herramienta de libre acceso
5	Portabilidad	El de fácil instalación en cualquier entorno
6	Fiabilidad	Los datos empleados para el entrenamiento son de fuentes verídicas y confiables
7	Integridad	Los datos no fueron alterados en el proceso de desarrollo

8	Métrica de Evaluación	El modelo fue evaluado con las métricas correspondientes a los algoritmos empleados
9	Funcionalidad	El modelo es capaz de hacer una clasificación efectiva en función de la entrada que reciba
10	Tamaño	El Dataset para entrenamiento contiene un conjunto extenso de entradas y salidas para el modelo

Nota: Jorge Oviedo Peñafiel e Inés Fajardo Romero

En la **tabla 36** se puede visualizar los rango y escala de puntuación definidos para la evaluación, rangos que van desde 0-20 siendo este la escala más deficiente a 81-100 siendo esta la escala mejor escala.

Tabla 36:

Rango y escalas de puntuación

Rango de Puntuación	Escala de Puntuación
0-20	Deficiente
21-40	Regular
41-60	Buena
61-80	Muy Buena
81-100	Excelente

Nota: Jorge Oviedo Peñafiel e Inés Fajardo Romero

A continuación, en la tabla 37, se muestra el resumen de la evaluación de los expertos por cada indicador.

Tabla 37:

Evaluación de los expertos

N°	Indicador	Experto 1	Experto 2	Experto 3	Experto 4	Promedio
1	Usabilidad	85	85	100	90	90.00
2	Capacidad de Respuesta	90	90	90	95	91.25
3	Robustez	100	100	90	95	96.25
4	Open Source	100	100	100	100	100.00
5	Portabilidad	100	90	100	90	95.00
6	Fiabilidad	100	100	95	90	96.25
7	Integridad	100	100	90	100	97.50
8	Métrica de Evaluación	100	100	90	100	97.50
9	Funcionalidad	100	100	95	100	98.75

10	Tamaño	100	100	90	90	95.00
Promedio Total						95.75

Nota: Jorge Oviedo Peñafiel e Inés Fajardo Romero

Según los resultados obtenido, se determina que el modelo de Procesamiento de Lenguaje Natural de conversacional textual basado en Machine Learning cumple con los requerimientos necesarios por lo cual fue realizado. Además de acuerdo con los criterios y validaciones de los expertos el presente proyecto ha obtenido un puntaje promedio total de 95.75 siendo un valor que corresponde a la escala de Excelente, por lo cual se puede concluir que el proyecto es totalmente aceptables, viable y factible para ellos.

3.17. Resultados

Se implementaron dos modelos de Procesamiento de Lenguaje Natural de clasificación textual mediante el uso de Machine Learning, en el cual se tuvo como datos de entradas los síntomas y recomendaciones de personas que padecieron de Covid-19. Para la implementación de estos modelos se utilizó la librería Sckit-Learn del lenguaje de programación Python, la cual tiene las funciones necesarias para la creación y entrenamiento de los algoritmos estudiados como lo son el algoritmo de Máquina de Soporte Vectorial y el algoritmo de Bosques Aleatorios.

Mediante esta investigación se pudo obtener como resultado un 86% en el Acurracy y una precisión del 96% en ambos modelos, lo que indica que para la clasificación de texto ambos algoritmos son fiables. Los mejores resultados de los modelos entrenados son los siguientes:

Figura 50:

Matriz de validación cruzada

Version	Algoritmo	Accuracy_Test	Hamming_Loss_Test	log_loss_Test	Precision_Weighted_Test	Recall_Weighted_Test	f1_Weighted_Test	promedio
5.0	svm	0.868557	0.015662	6.459337	0.969784	0.946597	0.957516	0.935613
17.0	rf	0.868557	0.016653	7.127941	0.969142	0.942408	0.955294	0.933850
11.0	rf	0.865979	0.016653	7.131417	0.970071	0.941361	0.955154	0.933141
8.0	rf	0.865979	0.016852	7.292822	0.969287	0.941361	0.954711	0.932834
2.0	svm	0.856774	0.017072	6.428427	0.962384	0.942966	0.951400	0.928381
3.0	svm	0.846591	0.025410	6.678893	0.961653	0.940512	0.949393	0.924537

Nota: Elaboración de Jorge Oviedo e Inés Fajardo Romero

Durante el estudio se realizó la recopilación de información de los contenidos del Procesamiento del Lenguaje Natural para el diseño de la aplicación de los algoritmos

usados en el modelo de NLP mediante la revisión sistemáticas apoyadas en artículos científicos, paginas científicas, tesis de grado, postgrados, etc.

Una vez revisada la información acerca del Procesamiento de Lenguaje Natural se realizó la preparación del Dataset por medio de la carga, limpieza y depuración de datos recolectados en encuesta a personas que fueron contagiadas de Covid-19 en la Zona 8 del Guayas.

Se identificaron las entradas y posible salida que tendría el modelo conversacional textual en español que fueron utilizadas para el diseño del modelo de NLP.

Se llevo a cabo el diseño del modelo con NLP basados en los algoritmos de Procesamiento del Lenguaje Natural, el cual permitió la evaluación eficiente de las entradas identificadas de los textos clasificados. Para el entrenamiento de los modelos se tomaron en cuenta el número de ocurrencias que existía en cada etiqueta obteniendo como resultados que en cada etiqueta debió tener 200 ocurrencias, dejando de esta manera un total de 13 etiquetas en la Dataset de síntomas y 7 etiquetas en el Dataset de recomendaciones.

Se realizó la respectiva evaluación del modelo NLP por medio de las métricas de evaluación como lo son la validación cruzada de los cuales podemos visualizar en la figura 29, además de ello se evaluó por medio de la curva de ROC y la matriz de confusión.

04 CAPITULO

Conclusiones y Recomendaciones

Conclusiones y Recomendaciones

En el presente capítulo se detallarán las conclusiones para los diferentes objetivos específicos planteados en el Capítulo I del presente trabajo de investigación. Además, se presentan las diferentes recomendaciones para impulsar la realización de trabajos futuros y de esta manera seguir mejorando la presente investigación.

4.1. Conclusiones

Una vez concluido el estudio del modelo de Procesamiento de Lenguaje Natural basado en Machine Learning para la clasificación de textos con terminología acerca del Covid-19 se concluye que:

- Se realizó una investigación sobre los algoritmos que se pueden aplicar para la clasificación textual que permitió identificar posibles opciones a tomar en consideración entre las cuales se encuentran las máquinas de soporte vectorial y los bosques aleatorios que fueron los electos al final.
- Se obtuvieron 4140 datos que representan a todas aquellas personas que estuvieron contagiadas de Covid-19 en la Zona 8 que comprende Guayaquil, Durán y Samborondón.
- Se estableció como entrada los síntomas padecidos por las personas encuestadas que los expusieron de manera textual en la encuesta de obtención de información.
- Se realizó la aplicación de dos diferentes algoritmos lo cuales son Soporte de Máquina Vectorial (SVM) y Random Forest (RF) para el diseño del modelo con la finalidad de realizar la comparación entre estos y así determinar qué modelo tiene mejor capacidad de clasificación de la información que ingresa.
- Para la evaluación del modelo se realizó mediante las diferentes métricas de medida como lo son la curva de ROC, tabla cruzada y la matriz de confusión, estas permitieron que podamos validar la eficacia de cada uno de los modelos. Los resultados obtenidos entre la comparación de los modelos de Soporte de Máquina Vectorial y Random Forest pudimos detectar que la precisión del modelo de Soporte de Máquina Vectorial y Random Forest es de un 0.96 (96%) lo cual nos indica que ambos modelos realizan una

clasificación del texto de manera precisa al momento de realizar las diferentes pruebas.

4.2. Recomendaciones

Una vez terminado el presente trabajo se puede enlistar las diferentes recomendaciones que permitan realizar mejoras a largo plazo de este, es por ello por lo que:

- A pesar de tener 4140 registros que permitieron el entrenamiento del modelo. Se recomienda que se sigan recopilando información y que esta recolección se expanda a otras ciudades del Ecuador.
- En este proyecto se trabajaron con los algoritmos de Máquina de Soporte Vectorial y Random Forest por lo cual se recomienda que se realicen más comparaciones con otros tipos de algoritmos supervisados como Regresión Logísticas, Naive Baye, Knn, entre otros; y así verificar la efectividad que tienen otros algoritmos sobre los analizados en este trabajo.
- Se recomienda el uso del modelo desarrollado para la clasificación de síntomas y recomendaciones de otro tipo de enfermedades.
- En el estudio realizado se recomienda tener el menor margen de error al momento de realizar un análisis manual de la información y esta sea procesada de manera correcta.

Referencias Bibliográficas

Referencias Bibliográficas

Agüero, V., Martin, J., Huarcaya, S., & Gloria, J. (2021). *Machine Learning en la mejora del proceso de operaciones comerciales en la empresa Redondos, Lima - 2020*. Universidad César Vallejo.

Aguilar, D., & Camargo, J. (2021). *Sistema inteligente basado en redes neuronales, máquina de soporte vectorial y random forest para la predicción de deserción de clientes en microcréditos de bancos* [Universidad Nacional Mayor de San Marcos]. http://cybertesis.unmsm.edu.pe/bitstream/handle/20.500.12672/16390/Aguilar_vd.pdf?sequence=1&isAllowed=y

Aiken, J. M., Aiken, C., & Cotton, F. (2018). A python library for teaching computation to seismology students. *Seismological Research Letters*, *89*(3). https://doi.org/10.1785/0220170246

Andrés Pérez, F. (2019). *El enfoque probabilístico en Inteligencia Artificial*. www.seguridadinternacional.es

Arango, C., & Osorio, C. (2021). *Aislamiento social obligatorio: un análisis de sentimientos mediante machine learning*. https://doi.org/10.14349/sumneg/2021.V12.N26.A1

Ariza-López, F. J., Rodríguez-Avi, J., & Alba-Fernández, V. (2018). CONTROL ESTRICTO DE MATRICES DE CONFUSIÓN POR MEDIO DE DISTRIBUCIONES MULTINOMIALES. *GeoFocus Revista Internacional de Ciencia y Tecnología de La Información Geográfica*. https://doi.org/10.21138/gf.591

Avila-Tomás, J. F., Mayer-Pujadas, M. A., & Quesada-Varela, V. J. (2020). Artificial intelligence and its applications in medicine I: introductory background to AI and robotics. *Atencion Primaria*, *52*(10), 778–784. https://doi.org/10.1016/j.aprim.2020.04.013

Ayoub, J., Yang, X. J., & Zhou, F. (2021). Combat COVID-19 infodemic using explainable natural language processing models. *Information Processing and Management*, *58*(4). https://doi.org/10.1016/j.ipm.2021.102569

Bandana, R. (2018). Sentiment analysis of movie reviews using heterogeneous features. *2018 2nd International Conference on Electronics, Materials Engineering and Nano-Technology, IEMENTech 2018*. https://doi.org/10.1109/IEMENTECH.2018.8465346

Bao, W., Lianju, N., & Yue, K. (2019). Integration of unsupervised and supervised machine learning algorithms for credit risk assessment. *Expert Systems with Applications*, *128*. https://doi.org/10.1016/j.eswa.2019.02.033

Barrionuevo, C., Ierache, J. S., & Sattolo, I. I. (2020). Reconocimiento de emociones a través de expresiones faciales con el empleo de aprendizaje supervisado aplicando regresión logística. *XXVI Congreso Argentino de*

Ciencias de La Computación (CACIC), 491–500. http://sedici.unlp.edu.ar/handle/10915/114089

Bi, Q., Goodman, K. E., Kaminsky, J., & Lessler, J. (2019). Practice of Epidemiology What is Machine Learning? A Primer for the Epidemiologist. *Oxford University Press on Behalf of the Johns Hopkins Bloomberg School of Public Health*. https://doi.org/10.1093/aje/kwz189

Biehler, R., & Fleischer, Y. (2021). *Introducing students to machine learning with decision trees using CODAP and Jupyter Notebooks*. https://doi.org/10.1111/test.12279

Bilheux, J.-C., Bilheux, H., Lin, J., Lumsden, I., & ZhangOak, Y. (2019). *Neutron imaging analysis using jupyter Python notebook*. https://doi.org/10.1088/2399-6528/ab3bea

Bisong, E. (2019). Introduction to Scikit-learn. *Building Machine Learning and Deep Learning Models on Google Cloud Platform*, 215–229. https://doi.org/10.1007/978-1-4842-4470-8_18

Borkan, D., Dixon, L., Sorensen, J., Thain, N., & Vasserman, L. (2019). Nuanced metrics for measuring unintended bias with real data for text classification. *The Web Conference 2019 - Companion of the World Wide Web Conference, WWW 2019*. https://doi.org/10.1145/3308560.3317593

Botchkarev, A. (2019). A new typology design of performance metrics to measure errors in machine learning regression algorithms. *Interdisciplinary Journal of Information, Knowledge, and Management*, *14*. https://doi.org/10.28945/4184

Bouabdallaoui, Y., Lafhaj, Z., Yim, P., Ducoulombier, L., & Bennadji, B. (2020). Natural language processing model for managing maintenance requests in buildings. *Buildings*, *10*(9). https://doi.org/10.3390/BUILDINGS10090160

Cabutto, T. A., Heeney, S. P., Ault, S. v., Mao, G., & Wang, J. (2018). An overview of the Julia programming language. *ACM International Conference Proceeding Series*, 87–91. https://doi.org/10.1145/3277104.3277119

Camacho Valladares, A. (2020). *Aprendizaje auto supervisado para reconocimiento de objetos* [Universidad Autónoma de Madrid]. https://repositorio.uam.es/bitstream/handle/10486/692807/camacho_valladares_alejandro_tfg.pdf?sequence=1

Canino, A. (2019). Deconstructing Google Dataset Search. *Public Services Quarterly*, *15*(3). https://doi.org/10.1080/15228959.2019.1621793

Cedeno-Moreno, D., & Vargas, M. (2020). Aprendizaje automático aplicado al análisis de sentimientos. *I+D Tecnológico*, *16*(2). https://doi.org/10.33412/idt.v16.2.2833

Celi-Parraga, R. J., Varela-Tapia, E. A., Acosta-Guzmán, I. L., & Montaño-Pulzara, N. R. (2021). Técnicas de procesamiento de lenguaje natural en la inteligencia artificial conversacional textual. *AlfaPublicaciones*, *3*(4.1). https://doi.org/10.33262/ap.v3i4.1.123

Cepeda, E. (2018). *Estadística Matemática*.

Cervantes, J., Garcia-Lamont, F., Rodríguez-Mazahua, L., & Lopez, A. (2020). A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing*. https://doi.org/10.1016/j.neucom.2019.10.118

Cevallos, H. V., Romero, H. C., Unda, S. B., Cevallos, V., Romero, C., & Barrezueta, &. (2020). APPLICATION OF AUTOMATIC LEARNING ALGORITHMS TO CLASSIFY THE FERTILITY OF A BANANA SOIL. *Revista Conrado*, 15–19.

Chiu, K.-L., Alexander, R., Farrow, H., Chen, J., Giorgi, M., Vargas Sepúlveda, M., Alexander, N., & Kolt, T. (2021). *Detecting Hate Speech with GPT-3 **. https://github.com/kelichiu/GPT3-hate-speech-detection.

Cho, Y. S., Sharpe, S., & Smith, H. (2018). *Stylometry in the Modern Era: Coreference and Voice for Authorship Attribution*. https://github.com/ssharpe42/AuthorStyle

Choi, R. Y., Coyner, A. S., Kalpathy-Cramer, J., Chiang, M. F., & Peter Campbell, J. (2020). Introduction to machine learning, neural networks, and deep learning. *Translational Vision Science and Technology*, *9*(2). https://doi.org/10.1167/tvst.9.2.14

Chowdhary, K. R. (2020). Fundamentals of artificial intelligence. In *Fundamentals of Artificial Intelligence*. Springer India. https://doi.org/10.1007/978-81-322-3972-7

Ciotti, M., Ciccozzi, M., Terrinoni, A., Jiang, W. C., Wang, C. bin, & Bernardini, S. (2020). The COVID-19 pandemic. *Critical Reviews in Clinical Laboratory Sciences*, *57*, 365–388. https://doi.org/10.1080/10408363.2020.1783198

Colomo Magaña, E., Sánchez Rivas, E., Ruiz Palmero, J., & Sánchez Rodríguez, J. (2020). La tecnología como eje del cambio metodológico. In *Umaeditorial*.

Cuero César. (2020). COVID-NIVEL-MUNDIAL. *Publicación de La Acádemia Panameña de Medicina y Cirugia* , 1–2. https://doi.org/http://dx.doi.org/10.37980/im.journal.rmdp.2020872

David, J., & Cortés, N. (2020). *IMPLEMENTACIÓN DE UNA APLICACIÓN WEB CON SERVICIO DE CHATBOT CON INTELIGENCIA ARTIFICIAL QUE PERMITA LA AUTOGESTIÓN DE CUENTAS POR PAGAR DE LOS PROVEEDORES DE LA UNIVERSIDAD AUTÓNOMA DE BUCARAMANGA*.

Delgado López, S. (2021). *Procesamiento de Lenguaje Natural sobre textos antiguos*. Universidad de La Laguna.

DeYoung, J., Jain, S., Rajani, N. F., Lehman, E., Xiong, C., Socher, R., & Wallace, B. C. (2020). *ERASER: A Benchmark to Evaluate Rationalized NLP Models*. https://doi.org/10.18653/v1/2020.acl-main.408

Diaz Morales, A., Fonseca Gómez, L., & González Veloza, F. (2020). *Caracterización por árboles de decisión para datos categóricos:*

diagnóstico de virus respiratorios en aves Characterization by decision trees for categorical data: diagnosis of respiratory viruses in birds.

Diaz, M. (2021). *Escuela de posgrado*. 0–1.

Dordevic, T., & Stojkovic, S. (2020). Syntax Analysis of Serbian Language using Context-free Grammars. *2020 55th International Scientific Conference on Information, Communication and Energy Systems and Technologies, ICEST 2020 - Proceedings*, 50–53. https://doi.org/10.1109/ICEST49890.2020.9232872

Eisenstein, J. (2018). *Introduction to Natural Language Processing*. https://cseweb.ucsd.edu//~nnakashole/teaching/eisenstein-nov18.pdf

Escobar Ariana. (2019). *ANÁLISIS DEL USO DEL PROCESAMIENTO DEL LENGUAJE NATURAL*.

Espinosa-Zúñiga, J. J. (2020). Aplicación de metodología CRISP-DM para segmentación geográfica de una base de datos pública. *Ingeniería, Investigación y Tecnología*, *21*(1), 1–13. https://doi.org/10.22201/FI.25940732E.2020.21N1.008

Espinoza Freire, E. E. (2018). La hipótesis en la investigación . *MENDIVE*, *16*(1).

Etikala, V. (2021). Extracting decision model components from natural language text for automated business decision modelling. *CEUR Workshop Proceedings*, *2956*.

Falcón, V., Pertile, V., & Ponce, B. (2019). La encuesta como instrumento de recolección de datos sociales: Resultados diagnostico para la intervención en el Barrio Paloma de la Paz (La Olla) - ciudad de Corrientes (2017-2018). *XXI Jornadas de Geografía de La UNLP*, 3. http://www.memoria.fahce.unlp.edu.ar/trab_eventos/ev.13544/ev.13544.pdfInformaciónadicionalenwww.memoria.fahce.unlp.edu.ar

Fernández-Garza, L. E., & Marfil, A. (2020). Neurological aspects that should not be forgotten during the COVID-19 pandemic. *InterAmerican Journal of Medicine and Health*, *3*. https://doi.org/10.31005/iajmh.v3i0.89

Fontanelli Espinoza, O., Mansilla Corona, R. L., & Miramontes Vidal, P. E. (2020). Distribuciones de probabilidad en las ciencias de la complejidad: una perspectiva contemporánea. *INTERdisciplina*, *8*(22). https://doi.org/10.22201/ceiich.24485705e.2020.22.76416

Fuentes Marmolejo, M. D., & Medina Parra, W. D. (2020). 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. *Universidad de Guayaquil*.

Gallardo, A. O. (2018). Relación entre economía y algunos paradigmas de inteligencia artificial. *TRASCENDER, CONTABILIDAD Y GESTIÓN*, *7*, 26–33. https://doi.org/10.36791/TCG.V0I7.10

Garduño Teliz, E., Albarrán Millán, D. F., & Damián Julián, F. (2019). Investigación evaluativa para la inclusión educativa. *REVISTA CIENCIAS*

PEDAGÓGICAS E INNOVACIÓN, *7*(2), 56–68. https://doi.org/10.26423/rcpi.v7i2.312

Gärtler, M., Khaydarov, V., Klöpper, B., & Urbas, L. (2021). The Machine Learning Life Cycle in Chemical Operations – Status and Open Challenges. In *Chemie-Ingenieur-Technik* (Vol. 93, Issue 12). https://doi.org/10.1002/cite.202100134

Georgescu, T. M. (2020). Natural language processing model for automatic analysis of cybersecurity-related documents. *Symmetry*, *12*(3). https://doi.org/10.3390/sym12030354

Gillioz, A., Casas, J., Mugellini, E., & Khaled, O. A. (2020). Overview of the Transformer-based Models for NLP Tasks. *Proceedings of the 2020 Federated Conference on Computer Science and Information Systems, FedCSIS 2020*. https://doi.org/10.15439/2020F20

Giraldo Ossa, W., & Jaramillo Marín, V. (2021). *Machine Learning para la estimación del riesgo de crédito en una cartera de consumo*. 1–57. https://repository.eafit.edu.co/handle/10784/29589

González, G. (2020, March 2). *Investigación diagnóstica: características, técnicas, tipos, ejemplos*. https://www.lifeder.com/investigacion-diagnostica/

González, N. M. (2018). *Infraestructura para la evaluación intrínseca de algoritmos de stemming (Noel Molina González)*.

González-Hernández, I. J., Simon-Marmolejo, I., Granillo-Macias, R., Santana-Robles, F., Rondero-Guerrero, C., & Soto-Campos, C. A. (2020). Simulación de la distribución uniforme generalizada. *Ingenio y Conciencia Boletín Científico de La Escuela Superior Ciudad Sahagún*, *7*(13), 23–28. https://doi.org/10.29057/ESCS.V7I13.4931

Graf César. (2020). Tecnologías de información y comunicación (TICs).Primer paso para la implementación de TeleSalud y Telemedicina. *Revista Paraguaya de Reumatología*, 2–2.

Grattarola, D., & Alippi, C. (2021). Graph Neural Networks in TensorFlow and Keras with Spektral. *IEEE COMPUTATIONAL INTELLIGENCE MAGAZINE*, 99. https://doi.org/10.1109/MCI.2020.3039072

Gudivada, V. N., & Arbabifard, K. (2018). Open-Source Libraries, Application Frameworks, and Workflow Systems for NLP. In *Handbook of Statistics* (Vol. 38, pp. 31–50). Elsevier B.V. https://doi.org/10.1016/bs.host.2018.07.007

Guerrero, S. (2020). Coronavirus in Ecuador: An opinion from the academia. *Granja*, *32*(2), 124–130. https://doi.org/10.17163/lgr.n32.2020.10

Guevara, G., Verdesoto, A., & Castro, N. (2020). Metodologías de investigación educativa (descriptivas, experimentales, participativas, y de investigación-acción). *RECIMUNDO*, *4*(3).

Gupta, R., & Jivani, A. G. (2018). Analyzing the stemming paradigm. *Smart Innovation, Systems and Technologies*, *84*, 333–342. https://doi.org/10.1007/978-3-319-63645-0_37

Hagedorn, S., Kläbe, S., & Sattler, K.-U. (2021). *Putting Pandas in a Box*. 10–13. https://www.python.org/dev/peps/pep-0249/

Hamarashid, H. K., Saeed, S. A., & Rashid, T. A. (2021). Next word prediction based on the N-gram model for Kurdish Sorani and Kurmanji. *Neural Computing and Applications*, *33*(9). https://doi.org/10.1007/s00521-020-05245-3

Hariyanti, T., Aida, S., & Kameda, H. (2019). Samawa Language Part of Speech Tagging with Probabilistic Approach: Comparison of Unigram, HMM and TnT Models. *Journal of Physics: Conference Series*, *1235*(1). https://doi.org/10.1088/1742-6596/1235/1/012013

Haro Rivera, S., Zúñiga Lema, L., Meneses Freire, A., Vera Rojas, L., & Escudero Villa, A. (2018). MÉTODOS DE CLASIFICACIÓN EN MINERÍA DE DATOS METEOROLÓGICOS. *Perfiles*, *2*(20). https://doi.org/10.47187/perf.v2i20.40

Haroon, M. M. (2018). Comparative Analysis of Stemming Algorithms for Web Text Mining. *Modern Education and Computer Science*, *9*, 20–25. https://doi.org/10.5815/ijmecs.2018.09.03

Harris, C. R., Jarrod Millman, K., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., Fernández del Río, J., Wiebe, M., Peterson, P., … Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*, 357. https://doi.org/10.1038/s41586-020-2649-2

Hickman, L., Thapa, S., Tay, L., Cao, M., & Srinivasan, P. (2022). Text Preprocessing for Text Mining in Organizational Research: Review and Recommendations. *Organizational Research Methods*, *25*(1), 114–146. https://doi.org/10.1177/1094428120971683

Hidayat, A., Jamaluddin, D., Sa, D., & Adillah Maylawati, ". (2020). Data Analytics for Effectiveness Evaluation of Islamic Higher Education using K-Means Algorithm. *International Journal of Advanced Science and Technology*, *29*(3), 4149–4161.

IBM. (2021). *Reglas de asociación - Documentación de IBM*. https://www.ibm.com/docs/es/spss-modeler/SaaS?topic=nodes-association-rules

Jet, A., & O, H. J. (2017). Supervised Machine Learning Algorithms: Classification and Comparison. *International Journal of Computer Trends and Technology*, *48*. https://doi.org/10.14445/22312803/IJCTT-V48P126

Jimenez Flores, V. J., Jimenez Flores, O. J., Jimenez Flores, J. C., & Jimenez Castilla, J. U. (2020). ENTIDAD CONVERSACIONAL DE INTELIGENCIA ARTIFICIAL Y CALIDAD PERCIBIDA DEL SERVICIO DE ATENCIÓN A LOS ESTUDIANTES DE LA UNIVERSIDAD JOSÉ CARLOS

MARIÁTEGUI FILIAL TACNA, 2018-II. *REVISTA CIENCIA Y TECNOLOGÍA-Para El Desarrollo-UJCM*. https://doi.org/http://dx.doi.org/10.37260/rctd.v1i2.30.g27

Joison A, N., Barcudi R, J., Majul E, A., Ruffino S, A., de Mateo Rey J, J., Joison A M., & Baiardi G. (2021). La inteligencia artificial en la educación médica y la predicción en salud. *Methodo. Investigación Aplicada a Las Ciencias Biológicas*, *6*(1). https://doi.org/10.22529/me.2021.6(1)07

Jusoh, S. (2018). A STUDY ON NLP APPLICATIONS AND AMBIGUITY PROBLEMS. *Journal of Theoretical and Applied Information Technology*, *31*(6). www.jatit.org

Kalyoncu, F., Zeydan, E., Yigit, I. O., & Yildirim, A. (2018). A customer complaint analysis tool for mobile network operators. *Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2018*, *January 2016*, 609–612. https://doi.org/10.1109/ASONAM.2018.8508289

Kanakaraddi, S. G., & Nandyal, S. S. (2018). Survey on Parts of Speech Tagger Techniques. *Proceedings of the 2018 International Conference on Current Trends towards Converging Technologies, ICCTCT 2018*. https://doi.org/10.1109/ICCTCT.2018.8550884

Kanani, P., & Padole, M. (2019). Deep learning to detect skin cancer using google colab. *International Journal of Engineering and Advanced Technology*, *8*(6). https://doi.org/10.35940/ijeat.F8587.088619

Kannan, R., & Vasanthi, V. (2019). Machine learning algorithms with ROC curve for predicting and diagnosing the heart disease. In *SpringerBriefs in Applied Sciences and Technology*. https://doi.org/10.1007/978-981-13-0059-2_8

Kaul, V., Enslin, S., & Gross, S. A. (2020). History of artificial intelligence in medicine. *Gastrointestinal Endoscopy*, *92*(4), 807–812. https://doi.org/10.1016/J.GIE.2020.06.040

Khanna, R. C., Cicinelli, M. V., Gilbert, S. S., Honavar, S. G., & Murthy, G. V. S. (2020). COVID-19 pandemic: Lessons learned and future directions. *Indian Journal of Ophthalmology*, *68*(5), 703–710. https://doi.org/10.4103/ijo.IJO_843_20

Khyani, D., Siddhartha, B., Niveditha, N. M., & Divya, B. M. (2020). *An Interpretation of Lemmatization and Stemming in Natural Language Processing*. 355–356. https://www.researchgate.net/publication/348306833

Krotov, V. (2018). *Legality and Ethics of Web Scraping*. https://www.researchgate.net/publication/324907302

Krotov, V., Johnson, L., & Silva, L. (2020). Tutorial: Legality and Ethics of Web Scraping. *Communications of the Association for Information Systems*, *47*, pp-pp. https://doi.org/10.17705/1CAIS.04724

Lady M. Sangacha Tapia, Ricardo Javier Celi Párraga, Eleanor Varela Tapia, & Iván L. Acosta Guzmán. (2021). *Modelos probabilísticos IA del procesamiento de lenguaje natural en conversaciones de personas contagiadas con Covid-19 Probabilistic AI models of natural language processing in conversations of people infected with Covid-19 Modelos probabilísticos de IA de processamento de linguagem natural em conversas de pessoas infectadas com Covid-19 Ciencias de la Salud Artículos de investigación*. https://doi.org/10.23857/pc.v6i9.3099

Lagouvardos, S., Dolby, J., Grech, N., Antoniadis, A., & Smaragdakis, Y. (2020). *Static Analysis of Shape in TensorFlow Programs*. *15*, 1–15. https://doi.org/10.4230/LIPIcs.ECOOP.2020.15

Landolt, S., Wambsganß, T., & Söllner, M. (2021). *A Taxonomy for Deep Learning in Natural Language Processing*. https://fasttext.cc

Lemenkova, P., Dursun, Z., Şeker, Ş., Kaya, A., Tanık, A., & Demir, V. (2019). Generic Mapping Tools and Matplotlib Package of Python for Geospatial Data Analysis in Marine Geology Polina LEMENKOVA Generic Mapping Tools and Matplotlib Package of Python for Geospatial Data Analysis in Marine Geology. *International Journal of Environment and Geoinformatics (IJEGEO*, 225–237. https://doi.org/10.30897/ijegeo.567343

Leopold, H., van der Aa, H., Offenberg, J., & Reijers, H. A. (2019). Using Hidden Markov Models for the accurate linguistic analysis of process model activity labels. *Information Systems*, *83*. https://doi.org/10.1016/j.is.2019.02.005

Li, C. H., Wu, S. L., Liu, C. L., & Lee, H. Y. (2018). Spoken SQuAD: A study of mitigating the impact of speech recognition errors on listening comprehension. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, *2018-September*. https://doi.org/10.21437/Interspeech.2018-1714

Li, Y., Is, S. L., Xu, Z., Cao, J., Chen' L, Z., Hu, Y., Ghent, H., & Cheung, S.-C. (2021). Tr a n s Re g e x : Multi-modal Regular Expression Synthesis by Generate-and-Repair. *2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE)*. https://doi.org/10.1109/IC

Lizcano-Jaramillo, P. A., Camacho-Cogollo, J. E., Lizcano-Jaramillo, P. A., & Camacho-Cogollo, J. E. (2019). Evaluación de Tecnologías en Salud: Un Enfoque Hospitalario para la Incorporación de Dispositivos Médicos. *Revista Mexicana de Ingeniería Biomédica*, *40*(3), 1–8. https://doi.org/10.17488/RMIB.40.3.10

Lloret, E. J. A. (2019). *Procesamiento del Lenguaje Natural (PLN) | Natural Language Processing (NLP)*.

Longbing Cao, U. F. (2017). Data Science : Challenges and Directions. *Communications of the ACM*, *Vol. 60*(No. 8), 59–68. http://delivery.acm.org/10.1145/3020000/3015456/p59-cao.pdf?ip=108.240.47.215&id=3015456&acc=OPEN&key=4D4702B0C3E38B35.4D4702B0C3E38B35.6978DEEF473775AF.6D218144511F34

37&CFID=791759728&CFTOKEN=87852454&__acm__=1501772575_4b9f4ec3dac2f9c27d6decb5123e617d

Lopez, J. (2018). Web scraping. *Academia Accelerating the World's Research*, 1–6. www.programacion.net

Machalek, D., Quah, T., & Powell, K. M. (2021). A novel implicit hybrid machine learning model and its application for reinforcement learning. *Computers and Chemical Engineering*, *155*. https://doi.org/10.1016/j.compchemeng.2021.107496

Maguiña Vargas, C., Gastelo Acosta, R., & Tequen Bernilla, A. (2020, July 31). El nuevo Coronavirus y la pandemia del Covid-19. *Revista Medica Herediana*, *31*(2), 125–131. https://doi.org/10.20453/rmh.v31i2.3776

Mahesh, B. (2018). Machine Learning Algorithms-A Review Machine Learning Algorithms-A Review View project Self Flowing Generator View project Batta Mahesh Independent Researcher Machine Learning Algorithms-A Review. *International Journal of Science and Research*. https://doi.org/10.21275/ART20203995

Martín Noguerol, T., Paulano-Godino, F., Martín-Valdivia, M. T., Menias, C. O., & Luna, A. (2019). Strengths, Weaknesses, Opportunities, and Threats Analysis of Artificial Intelligence and Machine Learning Applications in Radiology. *Journal of the American College of Radiology*, *16*(9), 1239–1247. https://doi.org/10.1016/j.jacr.2019.05.047

Maxwell, A. E., Warner, T. A., & Fang, F. (2018). Implementation of machine-learning classification in remote sensing: An applied review. In *International Journal of Remote Sensing* (Vol. 39, Issue 9). https://doi.org/10.1080/01431161.2018.1433343

Maylawati, D. S., Priatna, T., Sugilar, H., & Ramdhani, M. A. (2020). Data science for digital culture improvement in higher education using K-means clustering and text analytics. *International Journal of Electrical and Computer Engineering*, *10*(5), 4569–4580. https://doi.org/10.11591/IJECE.V10I5.PP4569-4580

Medina-Merino, R. F., & Ñique-Chacón, C. I. (2017). Bosques aleatorios como extensión de los árboles de clasificación con los programas R y Python. *Interfases*, *0*(010), 165. https://doi.org/10.26439/interfases2017.n10.1775

Memon, S., Benazirabad Sindh, S., Ghulam Ali Mallah, P., KNMemon, P., Shaikh, A., KAasoori, S., & Ul Hussain Dehraj, F. (2020). Comparative Study of Truncating and Statistical Stemming Algorithms. *IJACSA) International Journal of Advanced Computer Science and Applications*, *11*(2). www.ijacsa.thesai.org

Mendoza Olguín, G., Laureano De Jesús, Y., & Pérez de Celis Herrero, M. C. (2019). Métricas de similaridad y evaluación para sistemas de recomendación de filtrado colaborativo. *Revista de Investigación En Tecnologías de La Información*, *7*(14), 224–240. https://doi.org/10.36825/riti.07.14.019

Migliorelli, L., Moccia, S., Pietrini, R., Carnielli, V. P., & Frontoni, E. (2020). The babyPose dataset. *Data in Brief*, *33*. https://doi.org/10.1016/j.dib.2020.106329

Mihajlović, S., Kupusinac, A., Ivetić, D., & Berković, I. (2020). *The Use of Python in the field of Artifical Intelligence*.

Milo, T., & Somech, A. (2020). Automating Exploratory Data Analysis via Machine Learning: An Overview. *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 2617–2622. https://doi.org/10.1145/3318464.3383126

Misra, S., & Li, H. (2020). Noninvasive fracture characterization based on the classification of sonic wave travel times. *Machine Learning for Subsurface Characterization*, 243–287. https://doi.org/10.1016/B978-0-12-817736-5.00009-0

Montes, R. H., Alejandra, C., Pinto, M., & Navarro Jiménez, S. G. (2019). Redes neuronales y árboles de decisión para la clasificación de objetos astronómicos Neural Networks and Decision Trees for the Classification of Astronomical Objects. *Research in Computing Science*, *148*(7), 477–489.

Moreira, D., Cruz, I., Gonzalez, K., & Quirumbay, A. (2020). Análisis del Estado Actual de Procesamiento de Lenguaje Natural. *Iberian Journal of Information Systems and Technologies*.

Moreira, D., Cruz, I., Gonzalez, K., Quirumbay, A., Magallan, C., Guarda, T., Andrade, A., & Castillo, C. (2020). *Análisis del Estado Actual de Procesamiento de Lenguaje Natural Analysis of the Current State of Natural Language Processing*. 126–136.

Moreno-Pallares, M. G., Moreno-Pallares, R. R., & Mejia-Peñafiel, E. F. (2022). *Entrenamiento de Redes Neuronales Artificiales con Aprendizaje No Training of Artificial Neural Networks with Unsupervised Learning in the quality control of the analysis Treinamento de Redes Neurais Artificiais com Aprendizado Não Supervisionado no contr*. *7*(5), 1584–1593. https://doi.org/10.23857/pc.v7i5.4048

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). *Interpretable machine learning: definitions, methods, and applications*.

Murillo, C. & Ortiz, A. (2018). Máquinas de soporte vectorial para clasificación supervisada de imágenes en bases de datos espaciales. *Revista Ibérica de Sistemas e Tecnologias de Informação Iberian Journal of Information Systems and Technologies Recebido/Submission:*, *June*.

Murthy, K. N., & Scholar, P. (2020). WORD CLOUD IN PYTHON. *Complexity International Journal (CIJ)*, *24*. http://cij.org.in/Currentvolumeissue2401.aspx

Muthukrishnan, S., Krishnaswamy, H., Thanikodi, S., Sundaresan, D., & Venkatraman, V. (2020). Support vector machine for modelling and simulation of heat exchangers. *Thermal Science*, *24*(1PartB), 499–503. https://doi.org/10.2298/TSCI190419398M

Nagpal, A., & Gabrani, G. (2019). Python for Data Analytics , Scientific and Technical Applications. *2019 Amity International Conference on Artificial Intelligence (AICAI)*, 140–145.

Nanjekye, J. (2017). Python 2 and 3 Compatibility. In *Python 2 and 3 Compatibility*. https://doi.org/10.1007/978-1-4842-2955-2

Narro-Cornelio, K. M., & Vásquez-Tirado, G. A. (2021). Características clínico-epidemiológicas en pacientes con diagnóstico covid-19. Red de salud Virú, marzo - mayo 2020. *Revista Del Cuerpo Médico Del HNAAA*, *13*(4). https://doi.org/10.35434/rcmhnaaa.2020.134.772

Naser, M. Z., & Alavi, A. H. (2021). Error Metrics and Performance Fitness Indicators for Artificial Intelligence and Machine Learning in Engineering and Sciences. *Architecture, Structures and Construction*. https://doi.org/10.1007/s44150-021-00015-8

Nieto, N. T. E. (2018). TIPOS DE INVESTIGACIÓN. *Universidad Santo Domingo de Guzmán*, 1–4. https://scholar.google.com/citations?view_op=view_citation&hl=en&user=gskIDR8AAAAJ&pagesize=100&citation_for_view=gskIDR8AAAAJ:738O_yMBCRsC

Niño, J. S., & Mendoza, M. L. (2021). *LA INVESTIGACIÓN CIENTÍFICA EN EL CONTEXTO ACADÉMICO*.

Núñez Reiz, A., Armengol de la Hoz, M. A., & Sánchez García, M. (2019). Big Data Analysis and Machine Learning in Intensive Care Units. *Medicina Intensiva*, *43*(7), 416–426. https://doi.org/10.1016/j.medin.2018.10.007

Núñez-Torres, F. (2021). *DISEÑO Y DESARROLLO DE UN MODELO DE DESAMBIGUACIÓN LÉXICA AUTOMÁTICA PARA EL PROCESAMIENTO DEL LENGUAJE NATURAL*.

Nuthakki, S., Neela, S., Gichoya, J. W., & Purkayastha, S. (2019). *Natural language processing of MIMIC-III clinical notes for identifying diagnosis and procedures with neural networks*.

Oa, U., & Dieser, M. P. (2020). *TEOREMA CENTRAL DEL LÍMITE*.

Ocaña-Fernández, Y., Valenzuela-Fernández, L. A., & Garro-Aburto, L. L. (2019). Inteligencia artificial y sus implicaciones en la educación superior. *Propósitos y Representaciones*, *7*(2). https://doi.org/10.20511/pyr2019.v7n2.274

Ofer, D., Brandes, N., & Linial, M. (2021). The language of proteins: NLP, machine learning & protein sequences. *Computational and Structural Biotechnology Journal*, *19*, 1750–1758. https://doi.org/10.1016/J.CSBJ.2021.03.022

Otter, D. W., Medina, J. R., & Kalita, J. K. (2021). A Survey of the Usages of Deep Learning for Natural Language Processing. *IEEE Transactions on Neural Networks and Learning Systems*, *32*(2). https://doi.org/10.1109/TNNLS.2020.2979670

Palacio-Niño, J.-O., & Berzal, F. (2019). *Evaluation Metrics for Unsupervised Learning Algorithms*.

Palacios Cruz, M., Santos, E., Velázquez Cervantes, M. A., & León Juárez, M. (2021, January 1). COVID-19, a worldwide public health emergency. *Revista Clinica Espanola*, *221*(1), 55–61. https://doi.org/10.1016/j.rce.2020.03.001

Palmer, A., & Jiménez, R. (2022). *Cálculo de probabilidades en las distribuciones comunes en el análisis de datos*.

Pulido-Rojano, A. D., Ballestas-Acosta, C., Del-Castillo-Herrera, K., Navarro-Rosales, M., Fuentes-Ávila, T., Pizarro-Rada, A., & Rodriguez-Ospino, Y. (2021). Simulación de un Sistema de Inventarios para la Determinación de Niveles de Reposición y de Servicios: Un caso de estudio. *CICIC 2022 - Decima Segunda Conferencia Iberoamericana de Complejidad, Informatica y Cibernetica En El Contexto de the 13th International Multi-Conference on Complexity, Informatics, and Cybernetics, IMCIC 2022 - Memorias*. https://doi.org/10.54808/CICIC2022.01.191

Quaranta, L., Calefato, F., & Lanubile, F. (2021). KGTorrent: A dataset of python jupyter notebooks from kaggle. *Proceedings - 2021 IEEE/ACM 18th International Conference on Mining Software Repositories, MSR 2021*. https://doi.org/10.1109/MSR52588.2021.00072

Rafay, A., Suleman, M., & Alim, A. (2020). Robust Review Rating Prediction Model based on Machine and Deep Learning: Yelp Dataset. *2020 International Conference on Emerging Trends in Smart Technologies, ICETST 2020*. https://doi.org/10.1109/ICETST49965.2020.9080713

Rahate, P. M., & Chandak, M. B. (2019). Text normalization and its role in speech synthesis. *International Journal of Engineering and Advanced Technology*, *8*(5 Special Issue 3), 115–122. https://doi.org/10.35940/ijeat.E1029.0785S319

Ramachandran, D., & Parvathi, R. (2019). Analysis of Twitter Specific Preprocessing Technique for Tweets. *Procedia Computer Science*, *165*. https://doi.org/10.1016/j.procs.2020.01.083

Ray, S. (2019). A Quick Review of Machine Learning Algorithms. *Proceedings of the International Conference on Machine Learning, Big Data, Cloud and Parallel Computing: Trends, Prespectives and Prospects, COMITCon 2019*. https://doi.org/10.1109/COMITCon.2019.8862451

Reese, R. M. (2018). *Natural language processing with Java : explore various approaches to organize and extract useful text from unstructured data using Java*. https://books.google.com/books/about/Natural_Language_Processing_with_Java.html?hl=es&id=q7y4BwAAQBAJ

Robissout, D., Zaid, G., Colombier, B., Bossuet, L., & Habrard, A. (2021). Online Performance Evaluation of Deep Learning Networks for Profiled Side-Channel Analysis. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in*

Bioinformatics), *12244 LNCS*. https://doi.org/10.1007/978-3-030-68773-1_10

Rodrigo, A., & Bonino, R. (2019). *Aprendo con NooJ: de la lingüística computacional a la enseñanza de la lengua*. http://rephip.unr.edu.ar/xmlui/handle/2133/22285

Rodriguez, V. (2020). *CLASIFICACIÓN DEL ESTADO DE RUPTURA DE ANEURISMAS CEREBRALES BASADA EN LA CARACTERIZACIÓN MORFOLÓGICA Y HEMODINÁMICA MEDIANTE MACHINE LEARNING* [UNIVERSIDAD DE CHILE]. https://repositorio.uchile.cl/bitstream/handle/2250/175427/Clasificaci%c3%b3n-del-estado-de-ruptura-de-aneurismas-cerebrales-basada-en-la-caracterizaci%c3%b3n-morfologica-y.pdf?sequence=1&isAllowed=y

Rodríguez-Orejuela, A., Montes-Mora, C. L., & Osorio-Andrade, C. F. (2022). Sentimientos hacia la vacunación contra la covid-19: panorama colombiano en Twitter. *Palabra Clave*, *25*(1). https://doi.org/10.5294/pacla.2022.25.1.4

Sahli, H. (2020). An Introduction to Machine Learning. In *TORUS 1 - Toward an Open Resource Using Services: Cloud Computing for Environmental Data*. https://doi.org/10.1002/9781119720492.ch7

Saing, R. H. (n.d.). *Métodos y Técnicas Aplicadas a la Investigación en Atención Primaria de Salud . Autores : Héctor Bayarre Vea .*

Sánchez-caro, J., & Sánchez, F. A. (2021). Vida e Inteligencia Artificial en el campo de la Salud. In *RevistaeSalud. com*.

Sancho Escrivá, J. V., Fanjul Peyró, C., de la Iglesia Vayá, M., Montell, J. A., & Escartí Fabra, M. J. (2020). Aplicación de la Inteligencia Artificial con Procesamiento del Lenguaje Natural para textos de investigación cualitativa en la relación médico-paciente con enfermedad mental mediante el uso de tecnologías móviles. *Revista de Comunicación y Salud*, *10*(1), 19–41. https://doi.org/10.35669/rcys.2020.10(1).19-41

Sarica, S., & Luo, J. (2020). *Stopwords in Technical Language Processing*. https://doi.org/10.1371/journal.pone.0254937

Schonlau, M., & Zou, R. Y. (2020). The random forest algorithm for statistical learning. *The Stata Journal*, *20*(1), 3–29. https://doi.org/10.1177/1536867X20909688

Schröer, C., Kruse, F., & Gómez, J. M. (2021). A systematic literature review on applying CRISP-DM process model. *Procedia Computer Science*, *181*, 526–534. https://doi.org/10.1016/j.procs.2021.01.199

Scikit-learn. (2021). *Ensemble methods — scikit-learn 1.0.2 documentation*. https://scikit-learn.org/stable/modules/ensemble.html

Shelar, H., Kaur, G., Heda, N., & Agrawal, P. (2020). Named Entity Recognition Approaches and Their Comparison for Custom NER Model. *Science and Technology Libraries*, *39*(3). https://doi.org/10.1080/0194262X.2020.1759479

Sheykhmousa, M., Mahdianpari, M., Ghanbari, H., Mohammadimanesh, F., Ghamisi, P., & Homayouni, S. (2020). Support Vector Machine Versus Random Forest for Remote Sensing Image Classification: A Meta-Analysis and Systematic Review. In *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* (Vol. 13). https://doi.org/10.1109/JSTARS.2020.3026724

Shrestha, N., & Nasoz, F. (2019). Deep Learning Sentiment Analysis of Amazon.Com Reviews and Ratings. *International Journal on Soft Computing, Artificial Intelligence and Applications*, *8*(1). https://doi.org/10.5121/ijscai.2019.8101

Sinaga, K. P., & Yang, M.-S. (n.d.). *Unsupervised K-Means Clustering Algorithm*. https://doi.org/10.1109/ACCESS.2020.2988796

Sodhi, P., Awasthi, N., & Sharma, V. (2019). Introduction to Machine Learning and Its Basic Application in Python. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3323796

Sosa García, J. O. (2020). Atención de pacientes con COVID-19 en el consultorio médico. *Revista CONAMED*, *25*(S1). https://doi.org/10.35366/97343

Spjuth, O., Frid, J., & Hellander, A. (2021). The machine learning life cycle and the cloud: implications for drug discovery. *Expert Opinion on Drug Discovery*, *16*(9), 1071–1079. https://doi.org/10.1080/17460441.2021.1932812

Sukthanker, R., Poria, S., Cambria, E., & Thirunavukarasu, R. (2020). Anaphora and coreference resolution: A review. *Information Fusion*, *59*(May 2019), 139–162. https://doi.org/10.1016/j.inffus.2020.01.010

Sumamo, J. S. (2018). *DESIGNING A STEMMING ALGORITHM FOR KAMBAATA TEXT: A RULE BASED APPROACH*.

Suresh, H., & Guttag, J. (2021). *A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle; A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle*. https://doi.org/10.1145/3465416.3483305

Torres, M. E. (2019). *Derechos y desafíos de la Inteligencia Artificial*. http://www.cyta.com.ar/biblioteca/bddoc/bdlibros/derechos_ia/derechos_ia_torres.htm

Tovar, M., Flores, G., Reyes-Ortiz, J. A., & Contreras, M. (2018). Validation of Semantic Relation of Synonymy in Domain Ontologies Using Lexico-Syntactic Patterns and Acronyms. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *10880 LNCS*, 199–208. https://doi.org/10.1007/978-3-319-92198-3_20

Trilla, A. (2020). One world, one health: The novel coronavirus COVID-19 epidemic. *Medicina Clinica*, *154*(5), 175–177. https://doi.org/10.1016/j.medcli.2020.02.002

Troncoso-Pantoja, C., & Amaya-Placencia, A. (2017). Interview: A practical guide for qualitative data collection in health research. *Revista Facultad de Medicina*, *65*(2), 329–332. https://doi.org/10.15446/revfacmed.v65n2.60235

Ulčar, M., & Robnik-ˇ, M. (2021). *Cross-lingual alignments of ELMo contextual embeddings*. https://github.com/MatejUlcar/elmogan

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Jarrod Millman, K., Mayorov, N., J Nelson, A. R., Jones, E., Kern, R., Larson, E., … van Mulbregt, P. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*. https://doi.org/10.1038/s41592-019-0686-2

Visual Studio Code. (2021, October 27). *Python and Flask Tutorial in Visual Studio Code*. Flask Tutorial in Visual Studio Code. https://code.visualstudio.com/docs/python/tutorial-flask

Walkowiak, T., Datko, S., & Maciejewski, H. (2019). Bag-of-words, bag-of-topics and word-to-vec based subject classification of text documents in Polish - A comparative study. *Advances in Intelligent Systems and Computing*, *761*. https://doi.org/10.1007/978-3-319-91446-6_49

Wang, M., & Hu, F. (2021). The application of nltk library for python natural language processing in corpus research. *Theory and Practice in Language Studies*, *11*(9), 1041–1049. https://doi.org/10.17507/tpls.1109.09

Wang, S., Mao, X., & Yu, Y. (2018). An initial step towards organ transplantation based on github repository. *IEEE Access*, *6*. https://doi.org/10.1109/ACCESS.2018.2872669

Warjri, S., Pakray, P., Lyngdoh, S., & Kumar Maji, A. (2018). *Khasi Language as dominant Part-Of-Speech(POS) ascendant in NLP*. 2018. https://ssrn.com/abstract=3354480

Wattanakriengkrai, S., Chinthanet, B., Hata, H., Kula, R. G., Treude, C., Guo, J., & Matsumoto, K. (2022). GitHub repositories with links to academic papers: Public access, traceability, and evolution. *Journal of Systems and Software*, *183*. https://doi.org/10.1016/j.jss.2021.111117

Xiong, Z., Cui, Y., Liu, Z., Zhao, Y., Hu, M., & Hu, J. (2020). Evaluating explorative prediction power of machine learning algorithms for materials discovery using k-fold forward cross-validation. *Computational Materials Science*, *171*. https://doi.org/10.1016/j.commatsci.2019.109203

Zaharia, M., Chen, A., Davidson, A., Ghodsi, A., Hong, S. A., Konwinski, A., Murching, S., Nykodym, T., Ogilvie, P., Parkhe, M., Xie, F., & Zumar, C. (2018). Accelerating the Machine Learning Lifecycle with MLflow. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, 39–45.

Zheng, X., Zhang, C., & Woodland, P. C. (2021). *ADAPTING GPT, GPT-2 AND BERT LANGUAGE MODELS FOR SPEECH RECOGNITION*.

Zhu, D., Cai, C., Yang, T., & Zhou, X. (2018). A machine learning approach for air quality prediction: Model regularization and optimization. *Big Data and Cognitive Computing*, *2*(1). https://doi.org/10.3390/bdcc2010005

Сикулер, Д. В. (2021). RESOURCES PROVIDING DATA FOR MACHINE LEARNING AND TESTING ARTIFICIAL INTELLIGENCE TECHNOLOGIES. *Информационные и Математические Технологии в Науке и Управлении*, *2(22)*, 39–52. https://doi.org/10.38028/ESI.2021.22.2.004

RESUMEN

La Inteligencia Artificial (IA), el Procesamiento del Lenguaje Natural (NLP) y el Aprendizaje Automático (ML) han jugado un papel crucial en la lucha contra la pandemia de Covid-19, proporcionando herramientas tecnológicas valiosas para el diagnóstico, seguimiento y control de la enfermedad, implementándose soluciones con IA para mitigar sus efectos. Se propone el diseño de un modelo de ML aplicando técnicas NLP en el preprocesamiento de texto para poder evaluar la eficacia del análisis de datos en conversaciones de personas contagiadas del coronavirus SARS-CoV-2. Se recopiló información de redes sociales como Twitter y Facebook, y encuestas a contagiados de Covid-19 en la Zona 8 de la provincia del Guayas. Con estos datos, se entrenó un sistema de clasificación textual utilizando los algoritmos de Soporte de Máquina Vectorial y Random Forest. El estudio resultó en una precisión del 96% en ambos modelos, demostrando su viabilidad para la creación e implementación de clasificadores de texto. Se logró mejorar el rendimiento del modelo, reduciendo las categorías con más de 200 ocurrencias, lo que resultó en una precisión más elevada sin diferencias significativas entre ambos modelos. Por último, se desarrolló un sitio web capaz de clasificar correctamente los síntomas y recomendaciones comentadas por los pacientes.

Palabras Clave: Aprendizaje Supervisado, Conversación Textual, Inteligencia Artificial, Machine Learning, Modelos de Clasificación, Procesamiento de Lenguaje Natural.

Abstract

Artificial Intelligence (AI), Natural Language Processing (NLP) and Machine Learning (ML) have played a crucial role in the fight against the Covid-19 pandemic, providing valuable technological tools for the diagnosis, monitoring and control of the disease, implementing AI solutions to mitigate its effects. We propose the design of an ML model applying NLP techniques in text preprocessing in order to evaluate the effectiveness of data analysis in conversations of people infected with the SARS-CoV-2 coronavirus. Information was collected from social networks such as Twitter and Facebook, and surveys of people infected with Covid-19 in Zone 8 of the province of Guayas. With these data, a textual classification system was trained using the Support Vector Machine and Random Forest algorithms. The study resulted in an accuracy of 96% in both models, demonstrating their viability for the creation and implementation of text classifiers. Model performance was improved by reducing categories with more than 200 occurrences, resulting in higher accuracy with no significant differences between the two models. Finally, a website capable of correctly classifying the symptoms and recommendations commented by patients was developed.

Keywords: Supervised Learning, Textual Conversation, Artificial Intelligence, Machine Learning, Classification Models, Natural Language Processing, Machine Learning.

http://www.editorialgrupo-aea.com

Editorial Grupo AeA

editorialgrupoaea

Editorial Grupo AEA

ISBN: 978-9942-651-38-9