

# INTELIGENCIA ARTIFICIAL

## APLICADA EN EL PROCESAMIENTO DE LENGUAJE NATURAL CON PYTHON Y MACHINE LEARNING



# IA

Sangacha-Tapia, Lady Mariuxi  
Celi, Ricardo Javier  
Acosta-Guzmán, Ivan Leonel  
Varela-Tapia, Eleanor Alexandra



# **Inteligencia Artificial Aplicada a Procesamiento de Lenguaje Natural (NLP) con Python y Machine Learning.**

## **Autor/es:**

**Sangacha-Tapia, Lady Mariuxi**

*Instituto Superior Tecnológico del Azuay con condición de Superior  
Universitario*

**Celi, Ricardo Javier**

*Universidad Técnica Luis Vargas Torres de Esmeraldas*

**Acosta-Guzmán, Ivan Leonel**

*Universidad de Guayaquil*

**Varela-Tapia, Eleanor Alexandra**

*Universidad de Guayaquil*



Datos de Catalogación Bibliográfica

Sangacha-Tapia, L. M.  
Celi, R. J.  
Acosta-Guzmán, I. L.  
Varela-Tapia, E. A.

**Inteligencia Artificial Aplicada a Procesamiento de Lenguaje Natural  
(NLP) con Python y Machine Learning.**

Editorial Grupo AEA, Ecuador, 2024  
ISBN: 978-9942-651-43-3  
Formato: 210 cm X 270 cm

129 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 Jacqueline Herrera Sánchez

**Supervisor de Producción:** Prof. José Luis Vera

**Diseño:** Tnlg. Oscar J. Ramírez P.

**Consejo Editorial** Editorial Grupo AEA

Primera Edición, 2024

D.R. © 2024 por Autores y Editorial Grupo AEA Ecuador.

Cámara Ecuatoriana del Libro con registro editorial No 708

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

*Los contenidos de este libro pueden ser descargados, reproducidos difundidos e impresos con fines de estudio, investigación y docencia o para su utilización en productos o servicios no comerciales, siempre que se reconozca adecuadamente a los autores como fuente y titulares de los derechos de propiedad intelectual, sin que ello implique en modo alguno que aprueban las opiniones, productos o servicios resultantes. En el caso de contenidos que indiquen expresamente que proceden de terceros, deberán dirigirse a la fuente original indicada para gestionar los permisos.*

**Título del libro:**

**Inteligencia Artificial Aplicada a Procesamiento de Lenguaje Natural (NLP) con Python y Machine Learning.**

© Sangacha-Tapia, Lady Mariuxi; Celi, Ricardo Javier; Acosta-Guzmán, Ivan Leonel; Varela-Tapia, Eleanor Alexandra.

© Agosto, 2024

Libro Digital, Primera Edición, 2024

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

**ISBN: 978-9942-651-43-3**



<https://doi.org/10.55813/egaea.l.88>

**Como citar (APA 7ma Edición):**

Sangacha-Tapia, L. M., Celi, R. J., Acosta-Guzmán, I. L., Varela-Tapia, E. A. (2024). *Inteligencia Artificial Aplicada a Procesamiento de Lenguaje Natural (NLP) con Python y Machine Learning*. Editorial Grupo AEA.  
<https://doi.org/10.55813/egaea.l.88>

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; Instituto Tecnológico  
Superior Los Andes; Idrix  
Technology S.A – 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.

## Derechos de autor ©

Este documento se publica bajo los términos y condiciones de la licencia Creative Commons Reconocimiento-NoComercial-CompartirlGual 4.0 Internacional (CC BY-NC-SA 4.0).



El “copyright” y todos los derechos de propiedad intelectual y/o industrial sobre el contenido de esta edición son propiedad de la Editorial Grupo AEA y sus Autores. Se prohíbe rigurosamente, bajo las sanciones en las leyes, la producción o almacenamiento total y/o parcial de esta obra, ni su tratamiento informático de la presente publicación, incluyendo el diseño de la portada, así como la transmisión de la misma de ninguna forma o por cualquier medio, tanto si es electrónico, como químico, mecánico, óptico, de grabación o bien de fotocopia, sin la autorización de los titulares del copyright, salvo cuando se realice confines académicos o científicos y estrictamente no comerciales y gratuitos, debiendo citar en todo caso a la editorial. Las opiniones expresadas en los capítulos son responsabilidad de los autores.



# AUTORES

## RESEÑA DE AUTORES



**Sangacha Tapia Lady Mariuxi**



Instituto Superior Tecnológico del Azuay  
con condición de Superior Universitario



[lady.sangacha@tecazuay.edu.ec](mailto:lady.sangacha@tecazuay.edu.ec)  
[lady\\_tapia@hotmail.com](mailto:lady_tapia@hotmail.com)



<https://orcid.org/0000-0002-5169-8918>



Docente investigadora acreditada por SENESCYT-ECUADOR, líneas de investigación: En áreas de investigación de Ciencias Físicas y Matemáticas, Ciencias de la Computación e Inteligencia Artificial. Candidata doctorando de Inteligencia Artificial y Robótica en España (Universidad de Jaén.) Magister en Seguridad de Informática Aplicada (Espol-Ecuador). Máster Universitario en Ingeniería de software y Sistemas Informáticos (Universidad Internacional de la Rioja-España). Ingeniera en sistemas computacionales (Universidad de Guayaquil). Directora de Proyectos Sociales (Instituto Superior Universitario Tecnológico del Azuay). Jefa de Proyectos I+D+i (Instituto Superior Universitario Tecnológico del Azuay). Directora de Proyecto de investigación I+D (Universidad de Guayaquil). Directora de Proyectos I+D+i (Instituto Superior Universitario Tecnológico del Azuay). Coordinadora de Proyectos Sociales (Hogar de Cristo). Líder ganadora Primer Lugar "Scratch Day Ecuador" en Corporación Hogar de Cristo. Co-líder ganador Primer Lugar "DEModay de la 2da Edición Latam Online" en IA Saturday 2022. Editora en Jefa de ATENAS Revista Científica Técnica y Tecnológica. Fundadora "MEGCI".



**Celi, Ricardo Javier**



Universidad Técnica Luis Vargas  
Torres de Esmeraldas



[ricardo.celi@utelvt.edu.ec](mailto:ricardo.celi@utelvt.edu.ec)



<https://orcid.org/0000-0002-8525-5744>



Ingeniero en sistemas informáticos (UTLVTE) Ecuador, Master en ingeniería del software y sistemas informáticos (UNIR) España, Master en matemática mención modelación y docencia (UTLVTE) Ecuador. Docente de la Universidad Técnica Luis Vargas Torres de Esmeraldas. Investigador auxiliar 1 acreditado por la SENESCYT-ECUADOR. Líneas de investigación: programación, inteligencia artificial, matemática.

# AUTORES

## RESEÑA DE AUTORES



**Acosta Guzmán Ivan Leonel**



Universidad de Guayaquil



[ivan.acostag@ug.edu.ec](mailto: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).



**Varela Tapia Eleanor Alexandra**



Universidad de Guayaquil



[eleanor.varelat@ug.edu.ec](mailto: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). Magíster en Docencia y Gerencia en Educación Superior (UG, Ecuador). Magíster en Sistemas de Información Gerencial (Escuela Superior Politécnica del Litoral ESPOL, Ecuador). Ingeniera en Computación (ESPOL, Ecuador).

## Índice

|   |      |
|---|------|
| Reseña de Autores .....   | ix   |
| Índice .....  | xi   |
| Índice de Tablas.....   | xiii |
| Índice de Figuras .....   | xiv  |
| Agradecimiento .....  | xvii |
| Introducción .....  | xix  |
| Capítulo I: Conociendo la Inteligencia Artificial.....  | 1    |
| 1.1.    ¿A qué llamamos inteligencia artificial? .....  | 3    |
| 1.1.1.    Centrados en el hombre .....  | 3    |
| 1.1.2.    Centrados en torno a la racionalidad .....  | 4    |
| 1.2.    ¿De dónde Nace el procesamiento de Lenguaje Natural? .....                                  | 4    |
| 1.3.    Los primeros pasos en NLP .....   | 5    |
| 1.4.    Concepto de procesamiento de lenguaje Natural.....  | 6    |
| 1.5.    Relación de la IA y NLP .....   | 6    |
| 1.5.1.    Categorías de NLP .....   | 7    |
| 1.6.    Machine Learning .....  | 9    |
| 1.7.    Tipos de aprendizaje.....   | 9    |
| 1.7.1.    Regresión y Clasificación.....  | 10   |
| 1.8.    ¿Cómo modelar y entrenar algoritmos con texto?.....   | 16   |
| Capítulo II: Conociendo técnicas para un modelo .....   | 19   |
| 2.1.    A que llamoas técnicas .....  | 21   |
| 2.1.1.    Técnicas de preprocesamiento de datos.....  | 21   |
| 2.1.2.    Preprocesamiento en el texto .....  | 26   |
| Capítulo III: Machine Learning: Algoritmos de aprendizaje supervisado y métricas de evaluación..... | 35   |
| 3.1.    Pasos para construir un modelo de machine Learning.....                                     | 37   |

|  |    |
|--|----|
| 3.2. Algoritmos de Aprendizaje Supervisado.....                    | 39 |
| 3.2.1. Regresión lineal simple .....                               | 39 |
| 3.2.2. Regresión lineal múltiple .....                             | 40 |
| 3.2.3. Regresión Logística .....                                   | 42 |
| 3.2.4. Naive Bayes.....  | 44 |
| 3.2.5. KNN .....   | 45 |
| 3.2.6. Árboles de decisión.....                                    | 47 |
| 3.2.7. Super Vector Machine .....                                  | 50 |
| 3.3. Evaluación de modelos .....                                   | 52 |
| 3.3.1. Métricas de evaluación en algoritmos de clasificación ..... | 52 |
| 3.3.2. Métricas de evaluación en algoritmos de regresión .....     | 55 |
| Capítulo IV: IA aplicada con NLP y machine learning .....          | 57 |
| 4.1. Modalidad de la investigación .....                           | 59 |
| 4.2. Tipo de investigación .....                                   | 59 |
| 4.2.1. Investigación Exploratoria .....                            | 59 |
| 4.2.2. Investigación diagnóstica .....                             | 60 |
| 4.2.3. Investigación cuasi-experimental .....                      | 60 |
| 4.2.4. Investigación evaluativa .....                              | 60 |
| 4.3. Diseño metodológico de la investigación .....                 | 61 |
| 4.4. Metodología de investigación.....                             | 61 |
| 4.4.1. Definición del problema .....                               | 61 |
| 4.4.2. Objetivo principal .....                                    | 62 |
| 4.4.3. Técnicas de Procesamiento de Lenguaje Natural .....         | 62 |
| 4.4.4. Entrenamiento y aprendizaje del modelo .....                | 62 |
| 4.4.5. Medición y validación de resultados.....                    | 62 |
| 4.4.6. Población y muestra .....                                   | 62 |
| 4.4.6.1. Población.....  | 62 |

|   |     |
|---|-----|
| 4.4.6.2. Muestra.....   | 63  |
| 4.4.7. Técnica de recolección de datos.....                     | 63  |
| 4.4.7.1. Encuesta.....  | 64  |
| 4.4.8. Instrumento de medición.....                             | 64  |
| 4.4.8. Cuestionario .....                                       | 64  |
| 4.5. Técnica para el procesamiento y análisis de los datos..... | 64  |
| 4.5.8. Encuesta para personas contagiadas de Covid-19.....      | 64  |
| 4.5.9. Análisis de las preguntas cerradas .....                 | 65  |
| 4.5. Desarrollo de la investigación .....                       | 69  |
| 4.5.1. Fase 1: Recolección de data.....                         | 69  |
| 4.5.2. Fase 2: Limpieza de datos y depuración.....              | 70  |
| 4.5.3. Fase 3: Formato y visualización de datos .....           | 71  |
| 4.5.4. Fase 4: Preprocesamiento .....                           | 74  |
| 4.5.5. Fase 5: Etiquetación .....                               | 76  |
| 4.5.6. Fase 6: Entrenamiento.....                               | 79  |
| 4.5.7. Fase 7: Evaluación .....                                 | 86  |
| 4.5.8. Fase 8: Predicción .....                                 | 101 |
| 4.6. Anexo 1 .....  | 102 |
| Referencias Bibliográficas.....                                 | 105 |

## Índice de Tablas

|   |    |
|---|----|
| <b>Tabla 1:</b> Edad.....   | 65 |
| <b>Tabla 2:</b> Género .....  | 65 |
| <b>Tabla 3:</b> Variante del virus lo contagio.....   | 66 |
| <b>Tabla 4:</b> Nivel de intensidad que tuvo los síntomas .....   | 66 |
| <b>Tabla 5:</b> Lugar o evento considera que se contagió .....  | 67 |
| <b>Tabla 6:</b> En caso de haber estado vacunado al momento de contagiarse,<br>¿Cuántas dosis tenía aplicadas al contagiarse? ..... | 68 |

|  |    |
|--|----|
| <b>Tabla 7:</b> En caso de haber estado vacunado al momento de contagiarse ¿Qué vacuna recibió?..... | 68 |
| <b>Tabla 8:</b> Descripción de los campos del dataset.....   | 71 |
| <b>Tabla 9:</b> Parámetros de versiones de cada modelo.....  | 85 |

## Índice de Figuras

|   |    |
|---|----|
| <b>Figura 1:</b> Ramas de la Inteligencia Artificial.....   | 3  |
| <b>Figura 2:</b> Relación de la Inteligencia Artificial y Procesamiento de Lenguaje Natural .....                             | 7  |
| <b>Figura 3:</b> Nos va mostrando como se va empoderando el Procesamiento de Lenguaje Natural con la lingüística y la AI..... | 8  |
| <b>Figura 4:</b> Tipo de Aprendizaje .....  | 10 |
| <b>Figura 5:</b> Tipo de aprendizaje.....   | 15 |
| <b>Figura 6:</b> Presenta el resultado de datos.....  | 15 |
| <b>Figura 7:</b> Diferencia de aprendizaje.....   | 16 |
| <b>Figura 8:</b> Plataforma web para aplicar los diferentes modelos .....   | 16 |
| <b>Figura 9:</b> Modelo clásico de Procesamiento de Lenguaje Natural .....  | 17 |
| <b>Figura 10:</b> Tipos de datos.....   | 22 |
| <b>Figura 11:</b> Transformaciones de datos en el proceso de machine learning ....  | 25 |
| <b>Figura 12:</b> Versión abreviada en inglés .....   | 27 |
| <b>Figura 13:</b> Abreviación en español .....  | 27 |
| <b>Figura 14:</b> Caracteres no alfanuméricos.....  | 27 |
| <b>Figura 15:</b> Indica cuando eliminar el stop Word.....  | 28 |
| <b>Figura 16:</b> Diferencia entre Stemming y Lemmatization .....   | 29 |
| <b>Figura 17:</b> Se conoce los 2 errores principales.....  | 29 |
| <b>Figura 18:</b> Ambiente de google colab.....   | 31 |
| <b>Figura 19:</b> Resultados de la carga de datos .....   | 31 |
| <b>Figura 20:</b> Resultados a .....  | 32 |
| <b>Figura 21:</b> Resultados b .....  | 33 |
| <b>Figura 22:</b> Resultados c .....  | 33 |
| <b>Figura 23:</b> Resultados d .....  | 33 |

|  |    |
|--|----|
| <b>Figura 24:</b> Resultados e .....   | 34 |
| <b>Figura 25:</b> Flujo de Tareas de NLP .....   | 34 |
| <b>Figura 26:</b> Pasos de construcción de modelo de machine learning .....                                    | 37 |
| <b>Figura 27:</b> Modelo de Regresión Lineal.....  | 39 |
| <b>Figura 28:</b> Modelo de Regresión logística .....  | 42 |
| <b>Figura 29:</b> Modelo de Naive Bayes .....  | 44 |
| <b>Figura 30:</b> Modelo KNN.....  | 46 |
| <b>Figura 31:</b> Modelo de clasificación con árboles de decision .....  | 48 |
| <b>Figura 32:</b> Modelo de regresión con árboles de decisión .....  | 48 |
| <b>Figura 33:</b> Modelo de SVM.....   | 50 |
| <b>Figura 34:</b> Matriz de confusión .....  | 52 |
| <b>Figura 35:</b> Curva ROC .....  | 54 |
| <b>Figura 36:</b> Área AUC.....  | 54 |
| <b>Figura 37:</b> Diseño metodológico.....   | 61 |
| <b>Figura 38:</b> Encuesta en Google Form .....  | 70 |
| <b>Figura 39:</b> Dataset recopilado .....   | 71 |
| <b>Figura 40:</b> Número de registros del dataset de síntomas en el preprocesamiento de datos .....            | 74 |
| <b>Figura 41:</b> Número de registros del dataset de recomendaciones en el preprocesamiento de datos.....      | 75 |
| <b>Figura 42:</b> Función que aplica el pre-procesamiento.....   | 75 |
| <b>Figura 43:</b> Pregunta 10 depurada del dataset .....   | 76 |
| <b>Figura 44:</b> Pregunta 15 depurada del dataset .....   | 76 |
| <b>Figura 45:</b> Conteo de la cantidad de síntomas por registro .....   | 77 |
| <b>Figura 46:</b> Conteo de la cantidad de recomendaciones por registro.....                                   | 78 |
| <b>Figura 47:</b> Cantidad de registros del dataset de síntomas pre-aplicación de técnicas de NLP.....         | 78 |
| <b>Figura 48:</b> Cantidad de registros del dataset de recomendaciones pre-aplicación de técnicas de NLP ..... | 78 |
| <b>Figura 49:</b> Tokenizado de la pregunta 10 del dataset.....  | 79 |
| <b>Figura 50:</b> Tokenizado de la pregunta 15 del dataset.....  | 80 |
| <b>Figura 51:</b> Visualización de las stopwords del corpus en español .....                                   | 80 |
| <b>Figura 52:</b> Eliminación de stopwords en la pregunta 10 del dataset de síntomas .....                     | 81 |

|  |     |
|--|-----|
| <b>Figura 53:</b> Eliminación de stopwords para la pregunta 15 del dataset de recomendaciones..... | 81  |
| <b>Figura 54:</b> Aplicación de la técnica lematización en la pregunta 10 de síntomas .....        | 82  |
| <b>Figura 55:</b> Aplicación de la técnica lematización en la pregunta 15 de recomendaciones.....  | 82  |
| <b>Figura 56:</b> POS Tagging aplicado a la pregunta 10 de síntomas.....                           | 83  |
| <b>Figura 57:</b> POS Tagging aplicado a la pregunta 15 de recomendaciones .....                   | 83  |
| <b>Figura 58:</b> Columna POS sin palabras auxiliares en dataset de síntomas.....                  | 84  |
| <b>Figura 59:</b> Columna POS sin palabras auxiliares en dataset de recomendaciones.....           | 84  |
| <b>Figura 60:</b> Matriz de validación para el dataset de síntomas .....                           | 86  |
| <b>Figura 61:</b> Matriz de validación para el dataset de recomendaciones .....                    | 88  |
| <b>Figura 62:</b> Matriz de confusión de las clases de síntomas .....                              | 91  |
| <b>Figura 63:</b> Matriz de confusión de las clases de recomendaciones .....                       | 96  |
| <b>Figura 64:</b> Curva de ROC de las clases de síntomas .....                                     | 100 |
| <b>Figura 65:</b> Curva de ROC de las clases de recomendaciones .....                              | 100 |
| <b>Figura 66:</b> Prueba de predicción de test para síntomas .....                                 | 101 |
| <b>Figura 67:</b> Prueba de predicción de datos de test para recomendaciones ....                  | 101 |
| <b>Figura 68:</b> Prueba de predicción de datos ingresados para síntomas. ....                     | 102 |
| <b>Figura 69:</b> Prueba de predicción de datos ingresados para recomendaciones.                   |     |
|  | 102 |

## Agradecimiento

El plasmar el conocimiento en este libro adquirido de la investigación, se agradece a todo el equipo que participó, docentes investigadores de la Universidad de Guayaquil, Instituto Superior Tecnológico del Azuay con condición de Superior Universitario, ESPOL, Universidad Técnica Luis Vargas de Esmeraldas y Sociedad Estadística Ecuatoriana, que hicieron posible la obtención de los resultados. También, al Sr. Ángel Jesús Cuesta Chipre y el Sr. Luis André Holmes Montero, estudiantes de titulación de la Carrera de Ingeniería en Sistemas Computacionales de la Facultad de Ciencias Matemáticas y Físicas de Universidad de Guayaquil. A todos ustedes, muchas gracias por su compromiso en el desarrollo de la investigación, donde en cada integrante del equipo, se evidenció en el desarrollo del proyecto, la capacidad investigativa de cada uno al proyecto de investigación 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”.



## Introducción

Un trabajo investigativo plasmado, por un grupo de docentes investigadores de varias instituciones de educación superior de Ecuador junto con estudiantes que levantaron con información tecnológica de Inteligencia Artificial en conversaciones de texto de personas contagiadas de COVID-19. Antes de explicar el presente trabajo es importante conocer definiciones básicas y aclarar otras definiciones indispensables de las diferentes ramas de la inteligencia artificial enfocado al Procesamiento de Lenguaje Natural con su inicial en español PLN aplicando en lenguaje PYTHON en machine learning, las siglas en inglés es NLP que significa Natural Language Processing. En este libro se refleja el trabajo realizado de los fondos competitivos de investigación con el afán de que sea útil al lector para que conozca como puede predecir al momento de entrenar un algoritmo clasificado de texto en PLN en machine learning.

Se podrá encontrar 4 capítulos los cuales están conformados con la utilidad para quienes están iniciando el mundo de la IA de la rama de procesamiento de lenguaje natural con Python en machine learning, para entender es necesario tener nociones básicas de Python, pero en conocer lo demás no es necesario tener algún conocimiento previo, a continuación, se presenta lo que contiene cada capítulo de este libro:

**Capítulo 1** – Redacción de la inteligencia artificial, menciona lo que hoy en día ha ido evolucionando las varias ramas o que conocimiento abarca en este campo de conocimiento, además se menciona que, para alcanzar el entendimiento de procesamiento de lenguaje natural, aquí podrá conocer el concepto de machine learning, los tipos de aprendizaje para resolver problemas como el supervisado, no supervisado y refuerzo. Explicar la diferencia entre machine learning en español aprendizaje de máquina y Deep learning en español aprendizaje profundo finalmente el proceso para entrenar un algoritmo en procesamiento de lenguaje natural.

**Capítulo 2** – Se profundiza el procesamiento de lenguaje natural para ello se conoce los contenidos básicos de clasificación como: Las técnicas y diseño de LSTM, tokeización, stopword, lematización, bag of Word (part of speech tagging). La funcionalidad de cada una como se ha mencionado.

**Capítulo 3** – Se denota la estructuración de este capítulo el concocer los modelos de tipo supervisado más aplicados que son útiles en procesamiento de lenguaje natural orientado a la clasificación de texto.

**Capítulo 4** – Un caso práctico de predicciones o presentación del grado de asertividad del modelamiento de un algoritmo que se tomó, la intención es demostrar la utilización de un modelo y varias técnicas aplicando procesamiento de lenguaje natural aplicando machine learning.



CAPITULO

01

# CONOCIENDO LA INTELIGENCIA ARTIFICIAL



## Conociendo la Inteligencia Artificial

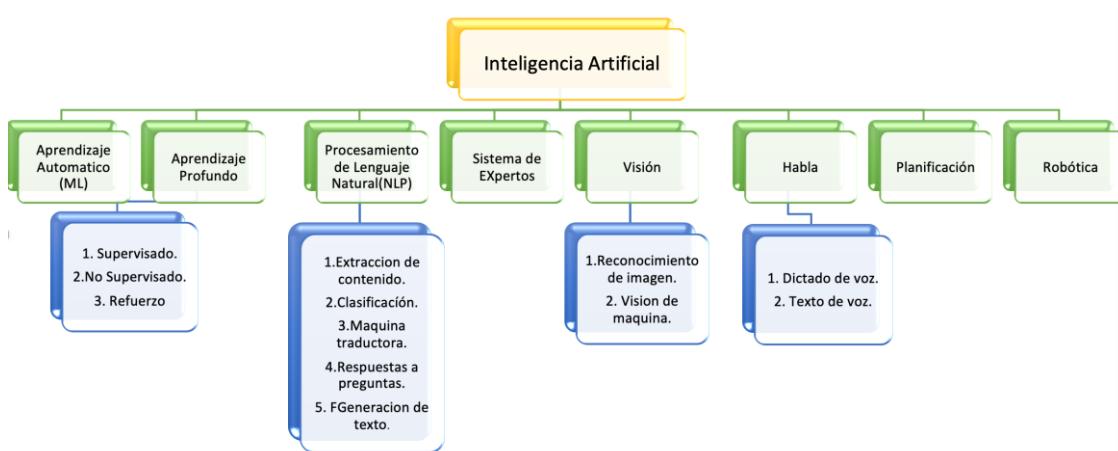
### 1.1. ¿A qué llamamos inteligencia artificial?

Sus siglas de IA, un campo que se pretende estudiar la construcción de máquinas inteligentes, se revisó de varios libros donde indica que el trabajo inicio luego del suceso de la segunda guerra mundial donde en los años 50 se introduce el nombre partiendo su estudio de tratar de entender como unos datos puede percibir, entender, predecir y manipular lo que el entorno del mundo ofrece (Gallastegui, 2008).

En pleno siglo XXI se sigue estudiando, descubriendo más ramas y subrama del campo de inteligencia artificial. Como podemos apreciar:

**Figura 1:**

*Ramas de la Inteligencia Artificial*



*Nota:* Se muestra todas las ramas de la IA, en base al autor Cerdas Méndez en el año 2017, actualizado por Alejandro Haro. 2023.

En la figura superior indica la finalidad de que todos los sistemas sean capaces de interpretar, memorizar y razonar cualquier clase de información. Sin embargo, a lo largo de la historia aún se considera los 4 enfoques que son centrados en el hombre y centrados en torno a la racionalidad. A continuación, podemos conocer un poco más a profundidad de cada enfoque. (Martínez-Comesaña, 2023)

#### 1.1.1.Centrados en el hombre

Se trata en el estudio del comportamiento de humano, entender el lenguaje, la

emoción y la conducta, con la intención de ofrecer experiencia eficaz entre humanos y robots. Existe beneficios como la ayuda a tomar decisiones, fiabilidad y escalabilidad, desarrollo de productos de software más exitoso, algunos autores mencionan ejemplos de ello (Bucalo et al., 2018).

### 1.1.2. Centrados en torno a la racionalidad

Hace referencia a la combinación de las matemáticas e ingenierías, esto permitirá que las máquinas actúen según lo correcto

A continuación, se menciona los 4 aspectos que a raíz de la historia aun prevalece:

- A. Sistemas que piensan como humanos
- B. Sistemas que piensan racionalmente
- C. Sistemas que actúan como humanos
- D. Sistemas que actúan racionalmente

Algunos autores mencionan que hace referencia a las leyes el pensamiento, por lo que uno de los primeros de poder detectar la forma correcta de pensar fue Aristóteles un griego filosofo, donde indicaba que es el proceso del razonamiento, mencionaba argumentaciones correctas, otro personaje fue Sócrates que también realizaba filosofías adecuadas. (Atencia López, J. M. 2000)

## 1.2. ¿De dónde Nace el procesamiento de Lenguaje Natural?

En el año del 1950 se realizó la primera prueba de Turing por Alan Turing, capaz de calcular cualquier tipo de cálculo con la intención de diseñarse de forma autómata, pero de manera de estados finitos, una máquina que no piensa, ni sienten ni mucho menos la intangibilidad de interacción con el ser humano (Pineda, 2022).

A través del tiempo el PLN o NLP se identifica como una disciplina, también conocida como la lingüística computacional, puesto que hace referencia al estudio del texto donde hoy en día por medio de las conversaciones se pueden

descubrir patrones de comportamientos que aportan a nuevos campos de conocimiento, el lenguaje hablado, la interacción de conversaciones. (Nadkarni,2011)

En los años 57 el entendimiento del NLP era más complejo, pero hoy en día se ha encontrado nuevos mecanismos para entenderlo, aunque todavía se encuentran en estudios y pruebas al menos ya existe alternativas para realizar nuevos estudios de predicciones o porcentajes de asertividad.

En el año 99 se presentó para la línea informática un programa que permitía resolver un crucigrama, desde ese entonces se han desarrollado otros programas que con la detección de patrones han permitido encontrar simulaciones o descubrimiento por medio de las palabras.

En la actualidad el gran reto o problema no es conocer el texto sino la comprensión del lenguaje la misma que se puede mencionar los siguientes desafíos:

1. Se menciona ambigüedad del lenguaje
2. La extensa dimensionalidad del lenguaje

Con lo mencionado cabe recalcar que el procesamiento de lenguaje natural no es solo de texto sino en la voz, gestos, expresiones que las máquinas aún no están listos. Por lo que en toda la explicación de este libro nos vamos a enfocar a la parte textual. Lo interesante que podemos aplicar cualquier tipo de aprendizaje. Mas adelante detallaremos referente a lo mencionado que ayudara a los lectores e investigadores que nos siguen.

En el siguiente punto da a conocer los primeros pasos de tener en cuenta para procesar NLP, esto ayuda principalmente cuando estas incursionando en este mundo de procesamiento de datos en texto.

### **1.3. Los primeros pasos en NLP**

Si estas interesado emergente en el mundo del procesamiento del lenguaje natural se te recomienda considerar lo siguiente:

1. Conocer las definiciones de NLP

2. Identificar términos de dimensionalidad y ambigüedades del lenguaje.
3. Identificar el objetivo para resolver el problema
4. Pretender reducir la dimensionalidad del texto analizar.

## 1.4. Concepto de procesamiento de lenguaje Natural

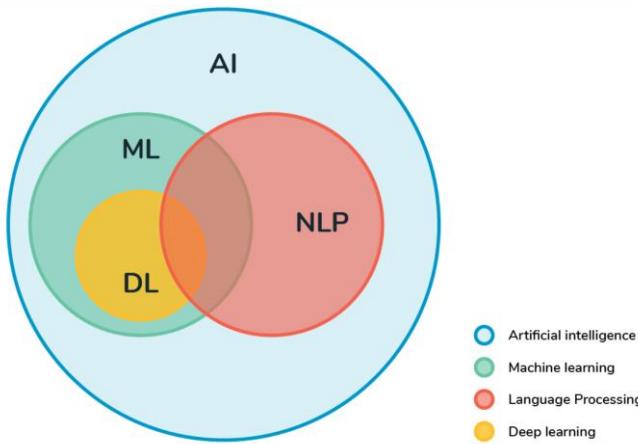
Es toda aquella acción que te conlleva a la comprensión del lenguaje natural donde es procesado para que la maquina interprete y puede existir la interacción entre el ser humano y la máquina. Hay que recordar que la maquina no entiende de letras o números, sino que de ceros y, de encendidos que representa “1” y los apagados que representa “0”.

## 1.5. Relación de la IA y NLP

Según Sarkar (2019), menciona que *“El procesamiento del lenguaje natural siempre ha captado mi atención debido a que el cerebro humano y nuestras habilidades cognitivas son realmente fascinantes. La habilidad para comunicar información, pensamientos complejos y emociones con tan poco esfuerzo es asombrosa una vez que piensas en cómo replicarla en las máquinas. Es un hecho que estamos avanzando por saltos y con límites en lo que se refiere a la computación cognitiva y la inteligencia artificial, pero aún no llegamos a la meta. Tal vez el pasar la prueba de Turing no es suficiente ¿Puede una máquina en realidad replicar al humano en todos sus aspectos?”*. Conocer los actores que intervienen en el proceso natural nos proporciona una visión de sus funciones.

**Figura 2:**

*Relación de la Inteligencia Artificial y Procesamiento de Lenguaje Natural*



*Nota:* Se puede apreciar que la AI abarca machine learning, Deep Learning y Procesamiento de Lenguaje Natural. Sarkar 2019.

### 1.5.1. Categorías de NLP

Se dividen en 2 categorías que son:

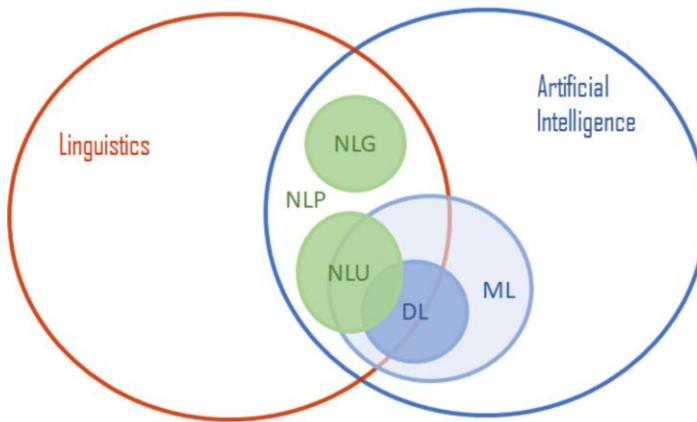
- A. La comprensión del lenguaje Natural (NLU o en inglés Natural Language Understanding)
- B. Generación del Lenguaje Natural (NLG o en inglés Natural Language Generation)

Estas categorías permiten alcanzar tres áreas principales como la informática, lenguaje humano e inteligencia artificial. El procesamiento de lenguaje natural permite fusionar los focos lingüísticos, técnicas de ML y DL.

A continuación, podemos apreciar la función de la Ai con las categorías y la lingüística.

**Figura 3:**

Nos va mostrando como se va empoderando el *Procesamiento de Lenguaje Natural con la lingüística y la AI*



*Nota:* Permite identificar la interacción de la comunicación lingüística para la inteligencia artificial con la inclusión de las categorías de NLP. Tomado de Manel Mezghanni 2020.

Se conoce que uno de los grandes problemas que se presenta en el lenguaje con la máquina es la ambigüedad y la dimensionalidad o sub dimensionado del texto.

Con respecto a la ambigüedad hace referencia que una acción podría referirse a una misma situación de algunas formas, por ejemplo:

Hola la palabra se puede interpretar en voz de texto como un saludo o de ola como la acción del mar.

En cambio, la dimensionalidad o sub dimensionado es referente a la gran cantidad de palabras que se presenta, sobre todo si son de diferentes lenguajes, la complejidad permite tener varias interpretaciones. Por tal razón la construcción de una maquina en el proceso del lenguaje es complejo. Imaginémonos una conversación con una maquina donde el humano testeá palabras de jergas urbanas, como la maquina entenderá cuando se le es sarcástico o la aplicación de palabras que tienen un mismo significado y a eso agregarle conversaciones extensas, por eso sería un reto gigantesco aplicar algoritmos que permita resolver este inconveniente y que la maquina entienda, pudiendo entablar una conversación fluida. En la actualidad las maquinas solo han sido capaz de crear e intercambiar diálogos.

## 1.6. Machine Learning

“Trabajan con **modelos y técnicas** que les permite comprender los datos de entrada a través de **algoritmos para automatizar los procesos de autoaprendizaje** en sus agentes, estos a su vez tienen la capacidad de tomar sus propias decisiones en base a lo que ya han aprendido de sus antiguas conversaciones” (Zuñiga & Humberto, 2018).

El aprendizaje automático o aprendizaje automatizado o aprendizaje de máquinas (del inglés, **machine learning**) es el **subcampo de las ciencias de la computación** y una **rama de la inteligencia artificial**, cuyo objetivo es desarrollar técnicas que permitan que las computadoras aprendan (Saturday AI, 2021).

## 1.7. Tipos de aprendizaje

Se presenta algunos tipos de aprendizaje en algunos textos mencionan 4 pero se realizó un cuadro comparativo de algunos autores y se consideró tomar 3 que se mencionan a continuación:

- Aprendizaje supervisado
- Aprendizaje no supervisado
- Aprendizaje de refuerzo

### Aprendizaje supervisado

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

algoritmo se mide mediante el llamado error de generalización.

Al considerar una nueva aplicación, el ingeniero puede comparar múltiples algoritmos de aprendizaje y determinar experimentalmente cual funciona mejor en el problema en cuestión (ver validación cruzada). Ajustar el rendimiento de un algoritmo de aprendizaje puede llevar mucho tiempo. Dados los recursos fijos, a menudo es mejor dedicar más tiempo a recopilar datos de entrenamiento adicionales y características más informativas que dedicar más tiempo a ajustar los algoritmos de aprendizaje.

En este tipo de aprendizaje se puede trabajar dos diferentes modelos que son:

- Regresiones
- Clasificación

Existe una extensa gama de lista de modelos en este tipo de aprendizaje sin embargo podremos nombrar las más importante en el siguiente punto.

**Figura 4:**

*Tipo de Aprendizaje*



*Nota:* A la máquina se debe aplicar el tipo de aprendizaje en base al objetivo del proyecto. 2024

### 1.7.1. Regresión y Clasificación

Los algoritmos más utilizados en los problemas de Machine Learning son los siguientes:

1. Regresión Lineal simple y múltiple
2. Regresión Logística
3. Árboles de Decisión
4. Random Forest
5. SVM o Máquinas de vectores de soporte.
6. KNN o K vecinos más cercanos.

## 7. K-means

El análisis de regresión es un subcampo del aprendizaje automático supervisado cuya meta es proporcionar un método para encontrar la relación entre una cierta cantidad numérica de características y una variable objetivo-continua.

La aplicación de la regresión es predecir propiedad de valor continuo que son añadido en base a un objeto. En cambio, la aplicación de clasificación es la identificación a que categorización (estado o evento) está perteneciendo el objeto.

En la actualidad existe algoritmos muy interesantes que son deseados y más utilizados, pero no está demás de dar a conocer que existen varios algoritmos considerados algoritmos básicos que se puede encontrar en <https://scikit-learn.org/> como:

### 1.1. Linear Models

- 1.1.1. Ordinary Least Squares
- 1.1.2. Ridge regression and classification
- 1.1.3. Lasso
- 1.1.4. Multi-task Lasso
- 1.1.5. Elastic-Net
- 1.1.6. Multi-task Elastic-Net
- 1.1.7. Least Angle Regression
- 1.1.8. LARS Lasso
- 1.1.9. Orthogonal Matching Pursuit (OMP)
- 1.1.10. Bayesian Regression
- 1.1.11. Logistic regression
- 1.1.12. Generalized Linear Regression
- 1.1.13. Stochastic Gradient Descent - SGD
- 1.1.14. Perceptron
- 1.1.15. Passive Aggressive Algorithms
- 1.1.16. Robustness regression: outliers and modeling errors
- 1.1.17. Quantile Regression
- 1.1.18. Polynomial regression: extending linear models with basis functions

## 1.2. Linear and Quadratic Discriminant Analysis

- 1.2.1. Dimensionality reduction using Linear Discriminant Analysis
- 1.2.2. Mathematical formulation of the LDA and QDA classifiers
- 1.2.3. Mathematical formulation of LDA dimensionality reduction
- 1.2.4. Shrinkage and Covariance Estimator
- 1.2.5. Estimation algorithms

## 1.3. Kernel ridge regression

## 1.4. Support Vector Machines

- 1.4.1. Classification
- 1.4.2. Regression
- 1.4.3. Density estimation, novelty detection
- 1.4.4. Complexity
- 1.4.5. Tips on Practical Use
- 1.4.6. Kernel functions
- 1.4.7. Mathematical formulation
- 1.4.8. Implementation details

## 1.5. Stochastic Gradient Descent

- 1.5.1. Classification
- 1.5.2. Regression
- 1.5.3. Online One-Class SVM
- 1.5.4. Stochastic Gradient Descent for sparse data
- 1.5.5. Complexity
- 1.5.6. Stopping criterion
- 1.5.7. Tips on Practical Use
- 1.5.8. Mathematical formulation
- 1.5.9. Implementation details

## 1.6. Nearest Neighbors

- 1.6.1. Unsupervised Nearest Neighbors
- 1.6.2. Nearest Neighbors Classification
- 1.6.3. Nearest Neighbors Regression
- 1.6.4. Nearest Neighbor Algorithms
- 1.6.5. Nearest Centroid Classifier
- 1.6.6. Nearest Neighbors Transformer
- 1.6.7. Neighborhood Components Analysis

## 1.7. Gaussian Processes

- 1.7.1. Gaussian Process Regression (GPR)
- 1.7.2. GPR examples
- 1.7.3. Gaussian Process Classification (GPC)
- 1.7.4. GPC examples
- 1.7.5. Kernels for Gaussian Processes

## 1.8. Cross decomposition

- 1.8.1. PLSCanonical
- 1.8.2. PLSSVD
- 1.8.3. PLSRegression
- 1.8.4. Canonical Correlation Analysis

## 1.9. Naive Bayes

- 1.9.1. Gaussian Naive Bayes
- 1.9.2. Multinomial Naive Bayes
- 1.9.3. Complement Naive Bayes
- 1.9.4. Bernoulli Naive Bayes
- 1.9.5. Categorical Naive Bayes
- 1.9.6. Out-of-core naive Bayes model fitting

## 1.10. Decision Trees

- 1.10.1. Classification
- 1.10.2. Regression
- 1.10.3. Multi-output problems
- 1.10.4. Complexity
- 1.10.5. Tips on practical use
- 1.10.6. Tree algorithms: ID3, C4.5, C5.0 and CART
- 1.10.7. Mathematical formulation
- 1.10.8. Minimal Cost-Complexity Pruning

## 1.11. Ensemble methods

- 1.11.1. Bagging meta-estimator
- 1.11.2. Forests of randomized trees
- 1.11.3. AdaBoost
- 1.11.4. Gradient Tree Boosting
- 1.11.5. Histogram-Based Gradient Boosting
- 1.11.6. Voting Classifier

1.11.7. Voting Regressor

1.11.8. Stacked generalization

## 1.12. Multiclass and multioutput algorithms

1.12.1. Multiclass classification

1.12.2. Multilabel classification

1.12.3. Multiclass-multioutput classification

1.12.4. Multioutput regression

## 1.13. Feature selection

1.13.1. Removing features with low variance

1.13.2. Univariate feature selection

1.13.3. Recursive feature elimination

1.13.4. Feature selection using SelectFromModel

1.13.5. Sequential Feature Selection

1.13.6. Feature selection as part of a pipeline

## 1.14. Semi-supervised learning

1.14.1. Self-Training

1.14.2. Label Propagation

## 1.15. Isotonic regression

## 1.16. Probability calibration

1.16.1. Calibration curves

1.16.2. Calibrating a classifier

1.16.3. Usage

## 1.17. Neural network models (supervised)

1.17.1. Multi-layer Perceptron

1.17.2. Classification

1.17.3. Regression

1.17.4. Regularization

1.17.5. Algorithms

1.17.6. Complexity

1.17.7. Mathematical formulation

1.17.8. Tips on Practical Use

1.17.9. More control with warm\_start

Nota: Los algoritmos básicos lo puedes encontrar en la plataforma que proporciona los algoritmos [https://scikit-learn.org/stable/supervised\\_learning.html#supervised-learning](https://scikit-learn.org/stable/supervised_learning.html#supervised-learning).

## Aprendizaje no supervisado

Anteriormente se listo sobre los algoritmos de aprendizaje supervisado. Ahora, pasamos a algoritmos de aprendizaje no supervisados:

Los modelos de aprendizaje no supervisados se utilizan para tres tareas principales:

- Clustering o agrupamiento,
- Asociación
- Reducción de dimensionalidad.

**Figura 5:**

*Tipo de aprendizaje*

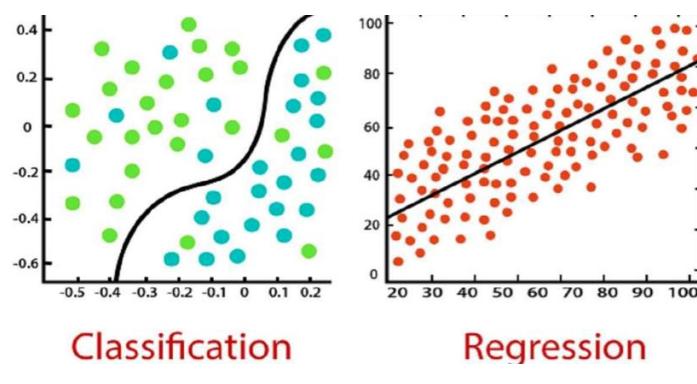


*Nota:* A la máquina se debe aplicar el tipo de aprendizaje en base al objetivo del proyecto. 2024.

En el siguiente punto podemos observar diferencias de gráfica de aprendizaje supervisado

**Figura 6:**

*Presenta el resultado de datos*

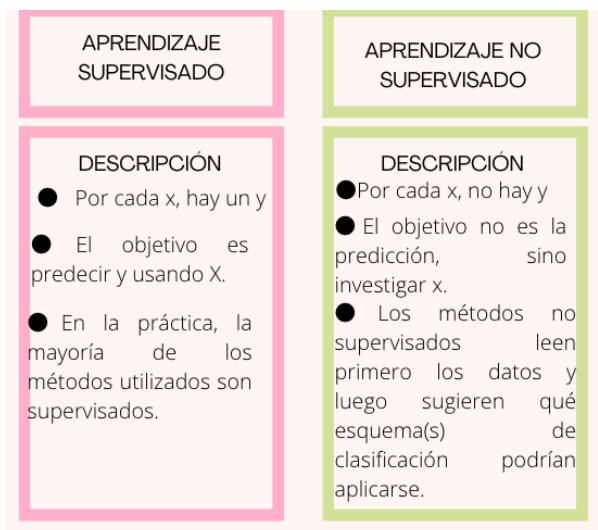


*Nota:* Se puede notar el resultado de datos donde el modelo de clasificación incluye desde modelos lineales como de la regresión logística, SVM, como otros no lineales como K-NN, Kernel SVM y bosques aleatorios. Academic 2022

¿qué son, cuándo son útiles y qué papel juegan en el futuro del aprendizaje automático?

**Figura 7:**

*Diferencia de aprendizaje*



*Nota:* Diferencias para identificar cada tipo de aprendizaje. Saturday 2021.

Esto permite identificar cómo funciona cada aprendizaje al momento de decidir que aprendizaje escoger, recuerda que dependerá del objetivo del proyecto y la data que tienes.

**Figura 8:**

*Plataforma web para aplicar los diferentes modelos*

The screenshot shows the official website for scikit-learn, a Python library for machine learning. The top navigation bar includes links for 'Install', 'User Guide', 'API', 'Examples', 'Community', and 'More'. Below the header, there's a large orange banner with the text 'scikit-learn' and 'Machine Learning in Python'. On the right side of the banner, there's a bulleted list of features: 'Simple and efficient tools for predictive data analysis', 'Accessible to everybody, and reusable in various contexts', 'Built on NumPy, SciPy, and matplotlib', and 'Open source, commercially usable - BSD license'. The main content area is divided into three sections: 'Classification' (Identifying which category an object belongs to), 'Regression' (Predicting a continuous-valued attribute associated with an object), and 'Clustering' (Automatic grouping of similar objects into sets). Each section includes a brief description, applications, and algorithms.

*Nota:* Modelos en la plataforma web para revisar la aplicación con los tipos de aprendizaje.  
<https://scikit-learn.org/stable/>

## 1.8. ¿Cómo modelar y entrenar algoritmos con texto?

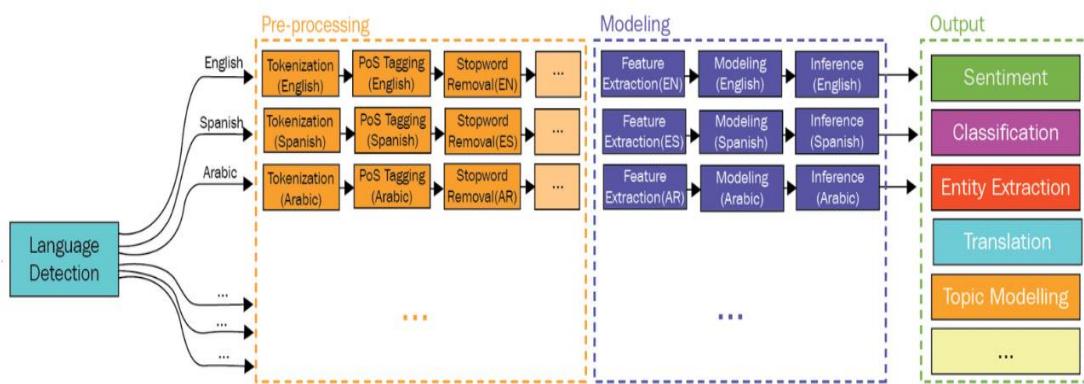
Se conoce que existe una guía de los pasos para entrenar un modelo con la intención de generar algoritmos que permita predecir o dar porcentaje de assertividad, la respuesta es que si, son los mis pasos que se aplica para

procesamiento de lenguaje natural pero en base a texto con la diferencia que se profundiza la utilización del tipo de aprendizaje supervisado la misma que se aplica la técnica de clasificación y finalmente de etiquetación, más adelante podrás ver este proceso indispensable antes de ejecutar el algoritmo. A continuación, podrás apreciar un bosquejo donde te proporciona un diseño para conocer el modelo clásico de NLP.

**Figura 9:**

*Modelo clásico de Procesamiento de Lenguaje Natural*

Classical NLP



*Nota:* Un modelo guía para detectar el lenguaje y pasar al preprocesamiento para entrenar el modelo que finalmente ofrece una salida. Se cumple con las etapas de modelar Saturday AI. 2022.

Sin embargo, para la aplicación de este modelo se debe seguir los pasos para modelar basado en machine learning que te presento a continuación:

**Recolectar los datos.** En este paso es la recolección de los datos desde muchas fuentes, podemos, por ejemplo: extraer de un sitio web público o privado o captar los datos utilizando una aplicación API, también puede ser desde una base de datos o dispositivos que realicen la extracción de manera automatizada. Lo que se puede aseverar que este paso es inmenso, tedioso y consume mucho tiempo, dependiendo desde proyectos grandes que demora años o pequeños que existe demora de 3 a 6 meses.

**Preprocesar o procesar datos.** En este paso ya se tienen la data set que puede ser convertido en. json, .CSV, .XML, HTML etc. Debes asegurarte de que los datos obtenidos deben tener el formato adecuado. Por tal razón la aplicación de

técnicas de preprocesamiento antes de utilizarse esos datos para el algoritmo en aplicar es sencillo.

**La Exploración de los datos (EDA).** En este paso el preanálisis es lo idóneo, incluso permite verificar que tiene la data set, asegurarse si no hay valores faltantes o si se debiera aplicar alguna técnica necesaria como el de imputación de datos, eso se trata de llenar con la aplicación de procesos estadísticas. Además, se puede detectar valores atípicos.

**Entrenar el algoritmo.** Poner en acción el proceso de técnicas de machine learning, le das el aprendizaje a la máquina con la data que se ha ido procesando en los pasos anteriores para realizar las respectivas predicciones o porcentajes de asertividad.

**Evaluación del algoritmo.** Una vez obtenido el algoritmo de entrenamiento, es necesaria evaluar los resultados para verificar si es el idóneo. Lo interesante es que si presenta inconsistencia se puede ir al paso anterior para volver a entrenar hasta alcanzar con el rendimiento aceptable.

**Utilización del algoritmo.** Se puede conocer el resultado del problema.

**¿Ahora la duda es que se debe modelar o entrenar con datos clasificados, o como se realiza?**

Esta explicación se realiza antes de realizar la aplicación del entrenamiento del modelo, tener claro el objetivo del proyecto. Para esta lectura se pretende conocer el porcentaje de asertividad o de predicción de los síntomas de una persona que tienen COVID-19. Se conoce que se realizó un proceso de métodos de investigación para obtener la data set, con datos de conversaciones de personas que nos indicaba los síntomas de COVID-19, también con datos personales. Listo conocido el contenido, la idea era por medio de la ayuda del experto es identificar las palabras que hace referencia al síntoma, ejemplo: Una mujer de 55 años escribió un texto que se ha sentido agotada y tenía mucha fiebre, había otros casos donde un hombre de 30 años escribió que tenía mucho dolor de cabeza y fiebre. Si te fijas un patrón de detección es la fiebre. Mas adelante del capítulo podrás conocer el proceso que se llevó porque el capítulo 1, 2 y 3 son bases fundamentales para entender el modelo de texto.

CAPITULO

02

# CONOCIENDO TÉCNICAS PARA UN MODELO



## Conociendo técnicas para un modelo

### 2.1. A que llamamos técnicas

Se puede definir como un conjunto de procedimientos a cumplir para alcanzar lo planteado.

#### 2.1.1. Técnicas de preprocessamiento de datos

En el capítulo anterior se vio los pasos para modelar un algoritmo, explicando que se realizó una recopilación de datos, para ello se aplicó metodologías de investigación cuantitativa y cualitativa. Se recopilo conversaciones de texto en redes sociales como Facebook, tweeter. Además, con la ayuda de los formularios que se preguntaba datos como edad, género, vacunas, dosis, que le recomendó el médico, que transcriba cuales fueron los síntomas y como se recuperó entre otros. Si nos animamos a volver a revisar el capítulo se podrá observar que el paso 1 del capítulo 1 con referente a la de colectar datos es válido con la ayuda de la tecnología.

Se preguntarán que tiene que ver con las técnicas, bueno realmente mucho porque una vez que se obtuvo la data set con las conversaciones y otros datos pasa al siguiente paso que es el preprocessamiento y procesamiento. Se trata de tomar esos datos y empezar a realizar un preanálisis, incluso cualquier data set que tengas se debe realizar este proceso, pero podemos explicar de forma general lo que se debería considerar.

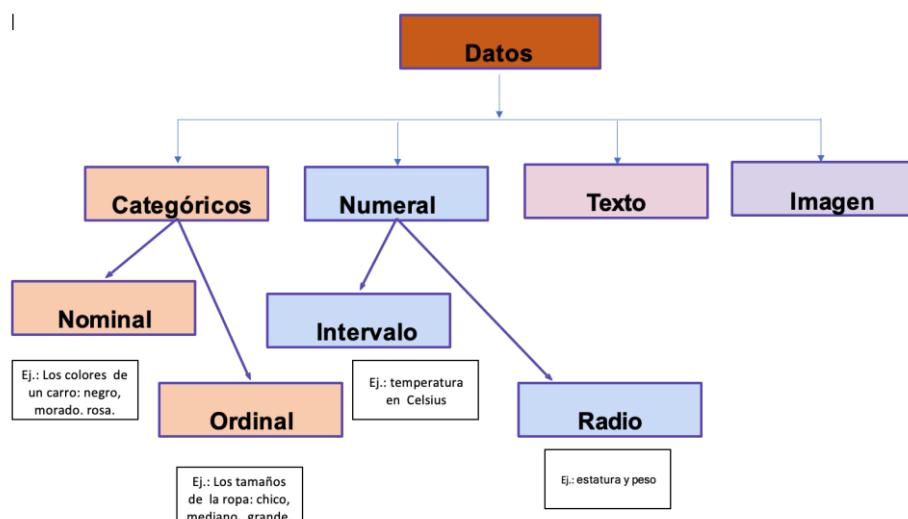
1. Verificar que tipos de datos existente en la data set.
  2. Pasos para aplicar preprocessamiento
  3. Elegir una herramienta de ciencia estadístico de datos como Google colab, R-studio, anaconda etc.
  4. Seleccionar un lenguaje de programación sin embargo en este texto solo se mencionará Python.
- 1. Verificar que tipos de datos existente en la data set.**

La importancia de conocer la data es para evaluar qué es lo que tienes, pero siempre considerando el objetivo del proyecto, asumiendo con la consecución

de este libro con referente en conocer los síntomas para detectar que un paciente tiene covid-19, se indicó que se tenía conversaciones de textos de personas que mencionaba como se sintieron, indicando los síntomas. Por otra parte, hay que recordar que la maquina no entiende de letras ni de números, pero sí de solo binarios, los ceros “ceros” y los unos “1”, por lo cual se procedió aplicar un proceso de clasificación con el texto y luego etiquetarla. Para conocer un poco más de este proceso en el capítulo 4 te lo explicara a más detalle. Siguiendo con la explicación esos datos de la data se debe evaluarlas donde pueden ser de tipo categórico, numeral, texto o imagen.

**Figura 10:**

*Tipos de datos*



*Nota:* En cualquier data set se puedes encontrar con estos tipos de datos, sin embargo, hay que recordar que esos datos pueden estar en diferentes tipos de archivos, ejemplo en extenciones .csv, .json, .sql, .html etc.

## 2. Pasos para aplicar preprocesamiento

La etapa de preprocesamiento de datos resulta esencial tanto para la minería de datos como para el aprendizaje automático, puesto que los datos crudos suelen contener ruido y ambigüedad, factores que pueden afectar negativamente la calidad de los modelos generados.

Para quienes están iniciando, este aporte es clave para que identifique ciertas diferencias en aplicar el preprocesamiento a manera general, por ahora se menciona para la aplicación de cualquier data, pero más adelante con ejercicio práctico se planteará los pasos de preprocesamiento de datos, pero con texto e

incluso con imágenes.

1. La evaluación de la calidad de los datos
2. La agregación de funciones
3. Muestreo de características
4. Reducción de dimensionalidad
5. Codificación de características

La **evaluación de la calidad de los datos**, la información obtenida debe ser confiables a pesar de que en base al problema que debes plantear no puedes esperar que los datos sean perfectos puesto que el error humano es inevitable o las fallas de las aplicaciones. En todo caso podrás encontrar lo siguiente:

**a) Los valores faltantes:**

Es muy común encontrar un data con valores faltantes pero la pregunta es, ¿qué hacer con ello?, pues puedes aplicar dos cosas, la primera eliminar esa fila de datos faltantes o realizar una estimación de valores que faltan, esto quiere decir que se aplicaría el método de interpolación como poner la fórmula de la mediana, modo, valor medio, promedio, etc.

**b) Los valores inconsistentes:**

Puedes encontrar una data set donde observes que existe valores que no tienen ninguna relación a la data que tratas o con el objetivo a tratar, por ejemplo: Supongamos que en una data tengas una columna de edad y en vez de encontrar número encuentras string, ¿que puedes hacer con ello?, sin son letras numéricas no habría problema, solo lo transformas, pero si no es nada que ver entonces podridas colocar valor cero o eliminar, seria dependiendo lo que deseas alcanzar.

**c) Los valores duplicados o repetidos**

Esto puede ocurrir cuando en el proceso de colección de datos una persona ingresa dos veces, ya sea en un formulario o en una encuesta. En la mayoría de los casos los valores duplicados son eliminados para no reflejar objetos de datos, un sesgo cuando se ejecutan algoritmos supervisados.

La **agregación de funciones** es la selección de características, cuando en una

data eliges atributos que sean de gran relevancia que serían útiles en una aplicación. Esta agrupación de funciones en evaluación se puede realizar por filtros y de envolvente, usualmente se utiliza en las transacciones de ventas.

El **Muestreo de características** para el cambio de datos en base a una porción de datos porque trabajar con datos completos puede ser demasiado costoso ya sea de uso de memoria y tiempo, ayudando reducir el tamaño del conjunto de datos. Una es la aplicación de muestreo aleatorio simple teniendo en cuenta en dos variaciones:

- a) Muestreo sin reemplazo
- b) Muestro con reemplazo

La **Reducción de dimensionalidad** esto sucede cuando en tu data tienes una gran cantidad de características, podría darse en el caso de utilización de imágenes. En la aplicación de reducción de dimensiones se asigna el conjunto de datos a un lugar de dimensión menor.

La **Codificación de características** aquí se realiza transformaciones en los valores de la data que permita identificar con facilidad las entradas para el proceso del algoritmo del aprendizaje, para ello se debe de considerar ciertas reglas para las variables continuas en las nominales, ordinal. En cambio, en las variables numéricas se debe considerar los intervalos, proporción.

Estos pasos son aplicados basados en machine learning. Ahora en el preprocesamiento con manejo de datos con texto.

Anteriormente se indicó que las maquinas no entienden de texto libre, ni las imágenes o datos de video como están porque solo entienden de 1 o 0. En las datas del mundo real a menudo están incompletos, inconsistentes, presentan errores donde la calidad de datos afecta en la aplicación del modelo para aprender. Por tal razón contar con datos que sean completos, precisos y válidos, al menos lo más cercano a eso. Se recomienda que se aplique *transformaciones de datos* en el proceso de machine learning para llevarlos a un estado que las maquinas puedan analizarlos fácilmente como puedes apreciar en la figura siguiente.

**Figura 11:**

*Transformaciones de datos en el proceso de machine learning*



Nota: Aplicar preprocessamiento en el proceso de machine learning es posible. 2024.

A continuación, se describen algunas técnicas comunes de preprocessamiento de datos, junto con algunos de sus autores destacados.

- a) Normalización y estandarización: estas técnicas se utilizan para escalar los datos para que estén en una escala común. La normalización es la técnica que escala los datos en un rango de 0 a 1, mientras que la estandarización escala los datos para que tengan una media de cero y una desviación estándar de uno. Estas técnicas fueron propuestas por varios autores, como Leo Breiman y Richard A. Olshen en su libro "Classification and Regression Trees" y Ilya Sutskever en su artículo "Training recurrent neural networks".
- b) Eliminación de ruido: esta técnica se utiliza para eliminar los valores atípicos y los datos faltantes. La eliminación de valores atípicos se realiza mediante el uso de técnicas como la eliminación de umbrales o la eliminación basada en el percentil, mientras que la eliminación de datos faltantes se realiza mediante la imputación de valores faltantes utilizando técnicas como la media, la mediana, el modo, entre otros. Esta técnica fue propuesta por varios autores, como Peter J. Huber en su libro "Robust Statistics" y Geoffrey Hinton en su artículo "Reducing the dimensionality of data with neural networks".

- c) Selección de características: esta técnica se utiliza para seleccionar las características más relevantes de un conjunto de datos. La selección de características puede realizarse mediante técnicas como la correlación, la prueba t y la eliminación recursiva de características. Esta técnica fue propuesta por varios autores, como John R. Quinlan en su artículo "Induction of Decision Trees" y Yoshua Bengio en su artículo "Deep Sparse Rectifier Neural Networks".
- d) Reducción de dimensionalidad: esta técnica se utiliza para reducir la cantidad de características de un conjunto de datos. La reducción de dimensionalidad se puede realizar mediante técnicas como el análisis de componentes principales (PCA) y el análisis de factores. Esta técnica fue propuesta por varios autores, como Harold Hotelling en su artículo "Analysis of a complex of statistical variables into principal components" y Geoffrey E. Hinton en su artículo "A practical guide to training restricted Boltzmann machines".

### 2.1.2. Preprocesamiento en el texto

En la aplicación del preprocesamiento con texto se trabaja con un corpus para la estructuración textual. El preprocesamiento de texto se trata de una serie de técnicas empleadas para acondicionar el texto original, de manera que pueda ser procesado, analizado o modelado en un sistema de procesamiento de lenguaje natural.

- a) Texto de preprocesamiento o texto de normalización, este proceso convierte los datos para que la computadora pueda entender. Tratando de *normalizar* los elementos del corpus del texto:
  - Convertir todas a mayúscula o minúsculas
  - Expandir contracciones
  - Eliminar caracteres especiales
  - Stemming y Lemmatization
  - Entre otras

Para **Convertir las mayúsculas o minúsculas** en este proceso es muy entendible y no hay problema de ello.

El **Expandir contracciones**, en esta explicación se indicará con idioma en inglés y en español, pero básicamente son versiones abreviadas de palabras o sílabas, se crean eliminando letras en específico. En el siguiente ejemplo hace referencia de los 2 idiomas.

### En idioma inglés

**Figura 12:**

*Versión abreviada en inglés*

**don't** -----> **do not**  
**I would** -----> **I'd.**

Nota: Podemos encontrar que la abreviación de inglés es el apóstrofo. 2024

### En idioma en español

**Figura 13:**

*Abreviación en español*

|   |
|---|
| Preposición + Artículo<br>a + el= al<br>Yo voy <b>al</b> teatro |
|---|

Nota: Podemos encontrar que la abreviación de español, se utiliza el concepto de preposición más artículo. 2024

La **Eliminación de caracteres especiales** los caracteres no alfanuméricos o incluso ocasionalmente caracteres numéricos, suman al ruido adicional en el texto no estructurado. La utilización de expresiones regulares simples para eliminarlas.

**Figura 14:**

*Caracteres no alfanuméricicos*

Nota: Existe muchos caracteres no alfanuméricos. 2024

El **Stopwords** se realiza la filtración de datos inútiles, se refiere de las palabras que no es considerado por ser extremadamente comunes que aportan poco valor.

### Figura 15:

*Indica cuando eliminar el stop Word*

| Remove Stopwords  | Avoid Stopword Removal  |
|---|---|
| <ul style="list-style-type: none"> <li>▪ Text Classification           <ul style="list-style-type: none"> <li>▫ Spam Filtering</li> <li>▫ Language Classification</li> <li>▫ Genre Classification</li> </ul> </li> <li>▪ Auto-Tag Generation</li> </ul> | <ul style="list-style-type: none"> <li>▪ Machine Translation</li> <li>▪ Language Modeling</li> <li>▪ Text Summarization</li> <li>▪ Question-Answering problems</li> </ul> |

*Nota:* Presenta las formas de remover o eliminación del stop Word. 2024

Existe estrategias para manejar el tema de Stop Word, en este caso se dará a conocer lo que se maneja en español. Conoce uno de ellos.

- El orden de términos por frecuencia de recopilación, indica el número total de veces e cada termino se conoce en el corpus.
- Es ideal considerar los términos más frecuentes, esto se recomienda que sea a mano por su contenido semántico en relación con el dominio de los documentos.
- El tener guardado en un STOP LIST.
- Conoce el manejo de las palabras de afijos
- La identificación de la flexión (lingüística), el Cambio que sufren las palabras a través de los morfemas.

El **Stemming y Lemmatization** la aplicación de esta técnica es para reducir las palabras a sus formas base o raíces. El stemming consiste en eliminar sufijos, por ejemplo, las palabras “correr”, “corre”, “corriendo” se reducirán a su raíz “corr”. En la lematización es un proceso más complejo que se refiere a la reducción de las palabras a su forma o base, se debe considerar su contexto gramatical y sintáctico, por otra parte, se puede obtener resultados más precisos al conservar el significado original de las palabras, por ejemplo: las palabras “correr”, “corre”, “corriendo” se reducirán a su lema “correr”.

### Figura 16:

#### Diferencia entre Stemming y Lemmatization

##### Stemming

- Proceso 'crudo'
- Corta los extremos de las palabras (incluye la eliminación de afijos derivados)

##### Lemmatization

- Análisis morfológico de las palabras
- Eliminar solo las terminaciones flexivas y regresar a la forma básica o de diccionario de una palabra lema.

Nota: Pequeñas diferencias de la técnica. 2024

Existen dos errores en stemming que son el Over-stemming y el under-stemming.

### Figura 17:

#### Se conoce los 2 errores principales

##### Over-stemming

- Algoritmo
  - 2 palabras -> 1 raíz
- Realidad
  - 2 palabras -> 2 raíces
- Puede considerarse como **falsos positivos**.

##### Under-stemming

- Algoritmo
  - 2 palabras -> 2 raíz
- Realidad
  - 2 palabras -> 1 raíz
- Puede interpretarse como **falsos negativos**.

Nota: Diferenciación de over-stemming y under-stemming. La aplicación basada en redes neuronales da mejores resultados y simples.

### 3. Elegir una herramienta de ciencia estadístico de datos como google colab, R-studio, anaconda etc

La elección de la herramienta va a depender si se usara para la visualización o para el modelamiento, en nuestro caso se considerará para correr el modelo. Existe muchas herramientas que se puede utilizar, a continuación, se presenta algunas:

- a) Anaconda, una herramienta académica para uso científico. Se puede conocer el sitio oficial para conocerlo <https://www.anaconda.com/>.

- b) Jupyter notebook, su uso es libre para la ciencia de dato aplicada soporta lenguaje de R y Python. Lo puede revisar en la página oficial <https://jupyter.org/>
- c) Google Colab, permite programar y ejecutar Python en su navegador, su mayor beneficio es que no requiere configuración.
- d) Entre otros.

La elección de la herramienta será dependiendo lo que se requiere u el objetivo a seguir, debido a que este texto es para la ciencia científicas y porque además es una herramienta que se puede utilizar de forma colaborativa y a continuación podemos aplicar un ejemplo de la aplicación de este. Primero es conocer la herramienta, en este caso será google colab, ayudará a trabajar a crear una arquitectura para aplicar los pasos de modelar un modelo. En el siguiente ejemplo se da a conocer un ejemplo práctico en la aplicación de preprocesamiento con la aplicación de imputación de datos, como parte de la evaluación de calidad de datos.

#### **4. Seleccionar un lenguaje de programación sin embargo en este texto solo se mencionará Python.**

**Python** es un lenguaje de programación de alto nivel, tiene un concepto de que el diseño enfatiza la legibilidad del código. Se conoce que su simplicidad ayuda a los grandes científicos de datos por su fácil lenguaje de aprender, esto se debe que es parecido a la sintaxis de conversación en inglés.

Una vez que se explicó cada paso se procede en mostrar algunas de la utilización de preprocesamiento de datos.

Primero que nada, es ir al sitio oficial de google colab, para ello recuerda que debes tener una cuenta de Gmail registrado para su uso, revisa <https://colab.research.google.com/>.

**Figura 18:**

Ambiente de google colab



Nota: Se utilizará la herramienta colaborativa. 2024

Como siguiente paso es cargar la data que sería el paso 1 para modelar un modelo, sin embargo, esa colección de datos es colocarla en un archivo, para esta explicación será de extensión .CSV.

```
#cargar la librería
import pandas as pd
#Cargar base de datos
data = pd.read_csv(filepath_or_buffer= "titanic_custom.csv", sep=",")
```

Se debe utilizar la librería panda y se usa el método de read\_csv

Detección de valores faltantes

```
#Nulos en toda la base = True
pd.isnull(data).head(3)
```

**Figura 19:**

Resultados de la carga de datos

|   | Unnamed: 0 | pclass | survived | name  | sex   | age   | sibsp | parch | ticket | fare  | cabin | embarked | boat | body  | home.dest |
|---|------------|--------|----------|-------|-------|-------|-------|-------|--------|-------|-------|----------|------|-------|-----------|
| 0 | False      | False  | False    | False | False | False | False | False | False  | False | False | False    | True | False |           |
| 1 | False      | False  | False    | False | False | False | False | False | False  | False | False | False    | True | False |           |
| 2 | False      | False  | False    | False | False | False | False | False | False  | False | False | False    | True | False |           |

Nota: Se visibiliza los resultados después de colocar el comando

```
#Número de nulos en cada columna
pd.isnull(data).sum()
```

## Figura 20:

### Resultados a

```

    Unnamed: 0      0
    pclass          0
    survived        0
    name            0
    sex             0
    age            263
    sibsp           0
    parch           0
    ticket          0
    fare            1
    cabin          1014
    embarked        2
    boat            823
    body           1188
    home.dest       564
    dtype: int64

```

Nota: Autores (2024)

```
#Número de nulos en la columna body
pd.isnull(data['body']).sum()
```

### Borrar valores faltantes

Los valores faltantes se denotan como NA (not available). En python podemos usar el comando *dropna* para eliminarlos de la base de datos.

### Paraméetros:

```

axis      =      0      borrar      fila,      1      borrar      columna
how      ='all'  borra  cuando  toda  la  fila  (o  columna)  sea  NA
how      ='any' Si al menos uno de los valores de la fila (o columna) es NA.
#Borrar de la base "data", todas las filas
#que se NA en todas sus columnas:
data.dropna(axis=0, how ='all')
```

**Figura 21:**

**Resultados b**

| Unnamed: 0 | pclass | survived | name  | sex    | age     | sibsp | parch | ticket | fare     | cabin   | embarked | boat | body  | home.dest                       |
|------------|--------|----------|---|--------|---------|-------|-------|--------|----------|---------|----------|------|-------|---------------------------------|
| 0          | 0      | 1        | Allen, Miss. Elisabeth Walton                   | female | 29.0000 | 0     | 0     | 24160  | 211.3375 | B5      | S        | 2    | NaN   | St Louis, MO                    |
| 1          | 1      | 1        | Allison, Master. Hudson Trevor                  | male   | 0.9167  | 1     | 2     | 113781 | 151.5500 | C22 C26 | S        | 11   | NaN   | Montreal, PQ / Chesterville, ON |
| 2          | 2      | 1        | Allison, Miss. Helen Loraine                    | female | 2.0000  | 1     | 2     | 113781 | 151.5500 | C22 C26 | S        | NaN  | NaN   | Montreal, PQ / Chesterville, ON |
| 3          | 3      | 1        | Allison, Mr. Hudson Joshua Creighton            | male   | 30.0000 | 1     | 2     | 113781 | 151.5500 | C22 C26 | S        | NaN  | 135.0 | Montreal, PQ / Chesterville, ON |
| 4          | 4      | 1        | Allison, Mrs. Hudson J C (Bessie Waldo Daniels) | female | 25.0000 | 1     | 2     | 113781 | 151.5500 | C22 C26 | S        | NaN  | NaN   | Montreal, PQ / Chesterville, ON |
| ...        | ...    | ...      | ...   | ...    | ...     | ...   | ...   | ...    | ...      | ...     | ...      | ...  | ...   | ...                             |
| 1304       | 1304   | 3        | Zabour, Miss. Hileni                            | female | 14.5000 | 1     | 0     | 2665   | 14.4542  | NaN     | C        | NaN  | 328.0 | NaN                             |
| 1305       | 1305   | 3        | Zabour, Miss. Thamine                           | female | NaN     | 1     | 0     | 2665   | 14.4542  | NaN     | C        | NaN  | NaN   | NaN                             |
| 1306       | 1306   | 3        | Zakarian, Mr. Mapriededer                       | male   | 26.5000 | 0     | 0     | 2656   | 7.2250   | NaN     | C        | NaN  | 304.0 | NaN                             |
| 1307       | 1307   | 3        | Zakarian, Mr. Ortin                             | male   | 27.0000 | 0     | 0     | 2670   | 7.2250   | NaN     | C        | NaN  | NaN   | NaN                             |
| 1308       | 1308   | 3        | Zimmerman, Mr. Leo                              | male   | 29.0000 | 0     | 0     | 315082 | 7.8750   | NaN     | S        | NaN  | NaN   | NaN                             |

Nota: Autores (2024)

**Reemplazar valores faltantes**

Muchas veces los datos son difíciles o costosos de conseguir, por dicha razón nos puede interesar sustituir un NA en lugar de borrar toda la fila. Existen diversas técnicas para esto.

#Rellernar cada NA de la base con 0

```
data.fillna(0).head(3)
```

**Figura 22:**

**Resultados c**

| pclass | survived | name                             | sex    | age     | sibsp | parch | ticket | fare     | cabin   | embarked | boat | body | home.dest                       |
|--------|----------|----------------------------------|--------|---------|-------|-------|--------|----------|---------|----------|------|------|---------------------------------|
| 0      | 1        | 1 Allen, Miss. Elisabeth Walton  | female | 29.0000 | 0     | 0     | 24160  | 211.3375 | B5      | S        | 2    | 0.0  | St Louis, MO                    |
| 1      | 1        | 1 Allison, Master. Hudson Trevor | male   | 0.9167  | 1     | 2     | 113781 | 151.5500 | C22 C26 | S        | 11   | 0.0  | Montreal, PQ / Chesterville, ON |
| 2      | 1        | 0 Allison, Miss. Helen Loraine   | female | 2.0000  | 1     | 2     | 113781 | 151.5500 | C22 C26 | S        | 0    | 0.0  | Montreal, PQ / Chesterville, ON |

Nota: #Rellenar NA con otra palabra, en este caso "Desconocido"

```
data.fillna('Desconocido').head(3)
```

**Figura 23:**

**Resultados d**

| pclass | survived | name                             | sex    | age    | sibsp | parch | ticket | fare    | cabin   | embarked | boat        | body        | home.dest                       |
|--------|----------|----------------------------------|--------|--------|-------|-------|--------|---------|---------|----------|-------------|-------------|---------------------------------|
| 0      | 1        | 1 Allen, Miss. Elisabeth Walton  | female | 29     | 0     | 0     | 24160  | 211.338 | B5      | S        | 2           | Desconocido | St Louis, MO                    |
| 1      | 1        | 1 Allison, Master. Hudson Trevor | male   | 0.9167 | 1     | 2     | 113781 | 151.55  | C22 C26 | S        | 11          | Desconocido | Montreal, PQ / Chesterville, ON |
| 2      | 1        | 0 Allison, Miss. Helen Loraine   | female | 2      | 1     | 2     | 113781 | 151.55  | C22 C26 | S        | Desconocido | Desconocido | Montreal, PQ / Chesterville, ON |

Nota: #Rellenar una columna numérica con el promedio de toda columna

```
promedio_age = data['age'].mean()
```

```
data['age'].fillna(promedio_age)
```

**Figura 24:**

*Resultados e*

```

0      29.000000
1      0.916700
2      2.000000
3      30.000000
4      25.000000
...
1304   14.500000
1305   29.881135
1306   26.500000
1307   27.000000
1308   29.000000
Name: age, Length: 1309, dtype: float64

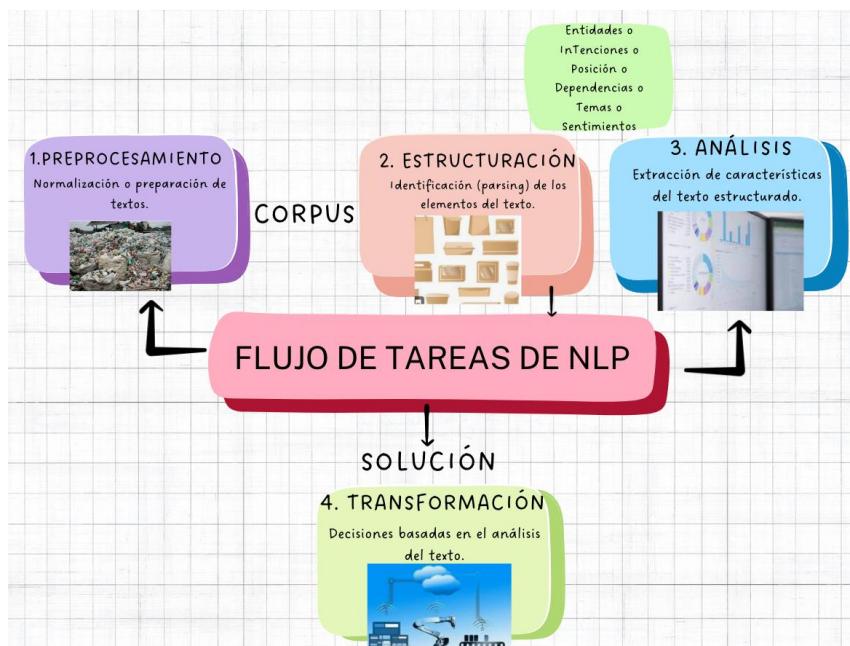
```

*Nota:* Autores (2024)

A continuación, se presenta un esquema de cómo se maneja el flujo de tareas de NLP

**Figura 25:**

*Flujo de Tareas de NLP*



*Nota:* Se indica la recepción del corpus para ser preprocesado, luego se procesa con la estructuración que identifica para el análisis obteniendo la solución de la transformación.

# CAPITULO

# 03

# **MACHINE LEARNING: ALGORITMOS DE APRENDIZAJE SUPERVISADO Y MÉTRICAS DE EVALUACIÓN**



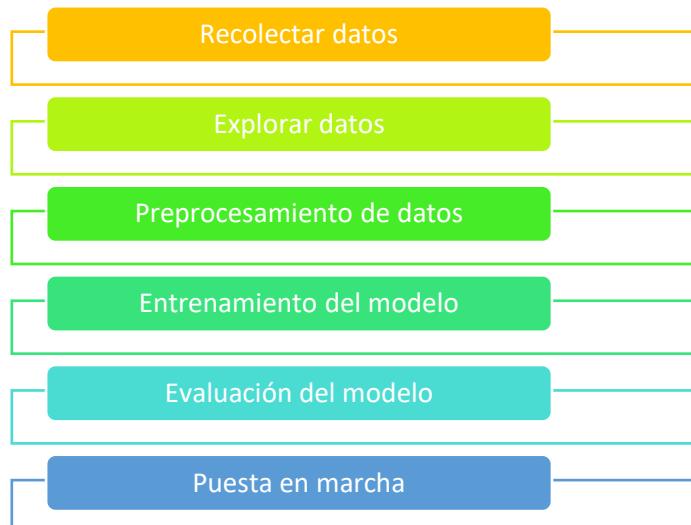
## Machine Learning: Algoritmos de aprendizaje supervisado y métricas de evaluación

### 3.1. Pasos para construir un modelo de machine Learning

Desarrollar un modelo de aprendizaje automático es más que simplemente usar un algoritmo de aprendizaje automático o una librería de aprendizaje automático; es un proceso completo que generalmente requiere los siguientes pasos:

**Figura 26:**

*Pasos de construcción de modelo de machine learning*



Nota: Autores (2024)

1. Recolectar datos. Podemos obtener los datos de una variedad de fuentes, como extraerlos de un sitio web o usando una API o una base de datos. Además, podemos utilizar otros dispositivos que recolectan datos por nosotros mismos; o podemos utilizar datos de dominio público. Las opciones de recolección de datos pueden venir de diversas fuentes como ficheros .csv, .xls, bases de datos relacionales, bases de datos no relacionales, entre otros. Aunque parece obvio, este paso es uno de los más complicados y exige más tiempo.
2. Explorar los datos. Podemos realizar un análisis previo para identificar los casos de valores faltantes o intentar encontrar algún patrón en los datos para facilitar la construcción del modelo una vez que tenemos los datos y están en el

formato correcto. En esta etapa, las medidas estadísticas y los gráficos en 2 o 3 dimensiones suelen ser muy útiles para tener una idea visual de cómo se comportan los datos. En este momento, podemos encontrar valores extraños o identificar las características que tienen el mayor impacto en la realización de una predicción.

3. Preprocesamiento de datos. Para ejecutar el algoritmo de aprendizaje, debemos asegurarnos de que los datos estén en el formato correcto. Antes de poder usar los datos, es prácticamente inevitable tener que completar una serie de tareas de preprocesamiento. Aquí implica, limpieza de datos como eliminación o imputación de nulos, tratamiento de valores atípicos, transformación de datos, codificación de variables categóricas e ingeniería de características.
4. Entrenar el modelo. En este paso, se utiliza realmente las técnicas de aprendizaje automático. En el entrenamiento, se alimenta los algoritmos de aprendizaje con los datos que procesamos en las etapas previas. La idea es que los algoritmos puedan hacer predicciones extrayendo información útil de los datos que les pasamos. Para el entrenamiento los datos deben ser particionados en datos de train y datos de pruebas, generalmente entre 70% y 30% o 80% y 20% respectivamente.
5. Evaluar el modelo. En esta etapa, ponemos a prueba la información o conocimiento que el algoritmo obtuvo del entrenamiento del paso anterior. Podemos volver a la etapa anterior y continuar entrenando el algoritmo cambiando algunos parámetros hasta lograr un rendimiento aceptable si no estamos muy conformes con su rendimiento. Durante la evaluación se aplica métricas de clasificación y métricas de regresión.
6. Poner el modelo en producción. En esta última etapa, ya preparamos nuestro modelo para abordar el problema real. Aquí también podemos evaluar su desempeño, lo que podría llevarnos a revisar todos los pasos anteriores.

## 3.2. Algoritmos de Aprendizaje Supervisado

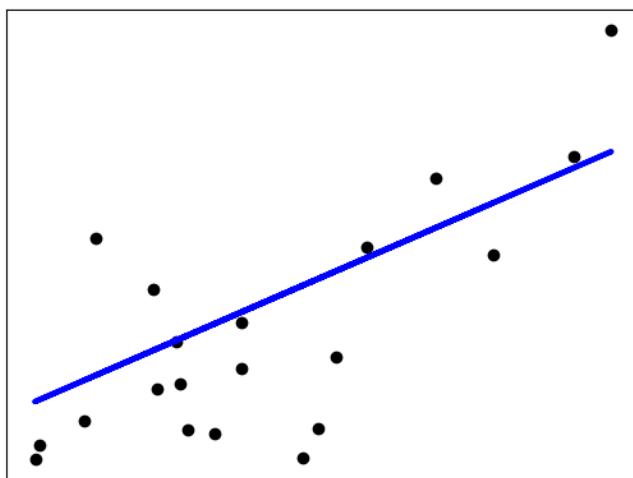
### 3.2.1. Regresión lineal simple

#### Definición

La regresión lineal simple es una técnica estadística que se utiliza para modelar la relación lineal entre una variable dependiente continua (que puede tomar cualquier valor dentro de un rango determinado) y una variable independiente. El objetivo de la regresión lineal simple es encontrar la ecuación de la recta que mejor se ajuste a los puntos de datos en el diagrama de dispersión (Draper et al., 1981) (Kutner et al., 2005) (Montgomery et al., 2012).

**Figura 27:**

*Modelo de Regresión Lineal*



Nota: [https://scikit-learn.org/stable/modules/linear\\_model.html#](https://scikit-learn.org/stable/modules/linear_model.html#)

#### Formulas asociadas al modelo

1. **Ecuación de la recta.** La ecuación de la recta de regresión lineal simple se expresa como:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

**Donde:**

- $y$  es el valor de la variable dependiente para un valor específico de la variable independiente  $x$ .
- $\beta_0$  es la intersección de la recta con el eje  $y$ , que representa el valor de  $y$  cuando  $x = 0$ .

- $\beta_1$  es la pendiente de la recta, que representa la tasa de cambio de  $y$  con respecto a  $x$ .
  - $x$  es el valor de la variable independiente.
  - $\epsilon$  es el término de error, que representa la diferencia entre el valor observado de  $y$  y el valor predicho por la ecuación de regresión.
2. **Estimación de parámetros.** Los parámetros de regresión lineal simple ( $\beta_0$  y  $\beta_1$ ) se estimaron utilizando mínimos cuadrados ordinarios (OLS). Este método tiene como objetivo encontrar los valores  $\beta_0$  y  $\beta_1$  tales que la suma de los cuadrados de los residuos sea mínima (la diferencia entre el valor de  $y$  observado y el valor predicho por la ecuación de regresión).
3. **Coeficiente de determinación ( $R^2$ ).** El coeficiente de determinación ( $R^2$ ) es una medida de la bondad de ajuste del modelo. Indica la porción de la varianza total de  $y$  que se explica por la variable independiente  $x$ . Un valor de  $R^2$  cercano a 1 indica que el modelo explica la mayor parte de la variación en  $y$ , mientras que un valor cercano a 0 significa que el modelo explica una pequeña porción de la variación en  $y$ .

### Aplicaciones de la regresión lineal simple

- Análisis de Tendencias: Examinamos cómo la variable dependiente ( $y$ ) cambia con el tiempo o en función de otra variable independiente ( $x$ ).
- Predicción: Predice el valor de la variable independiente ( $y$ ) para un nuevo valor de la variable independiente ( $x$ ).
- Causalidad: Investiga la posible relación causal entre dos variables ( $x$  e  $y$ ).

### 3.2.2. Regresión lineal múltiple

#### Definición

La regresión lineal múltiple es una técnica estadística que amplía la regresión lineal simple al considerar dos o más variables independientes para explicar una variable dependiente continua. El objetivo de la regresión lineal múltiple es encontrar la ecuación de hiperplano que mejor se ajuste a los puntos de datos en el espacio multidimensional (Afifi et al., 2020) (Bhattacherjee & Mukhopadhyay, 2020) (Shmueli & Crosley, 2019).

## Formulas asociadas al modelo

1. **Ecuación de la regresión.** La ecuación de la regresión lineal múltiple se expresa como:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \epsilon$$

Donde:

- $y$  es el valor de la variable dependiente para un conjunto específico de valores de las variables independientes ( $x_1, x_2, \dots, x_k$ ).
  - $\beta_0$  es la intersección del hiperplano con el eje  $y$ , que representa el valor de  $y$  cuando todas las variables independientes son iguales a 0.
  - $\beta_1, \beta_2, \dots, \beta_k$  son los coeficientes de regresión, que representan la tasa de cambio de  $y$  con respecto a cada variable independiente, manteniendo las demás constantes.
  - $x_1, x_2, \dots, x_k$  son los valores de las variables independientes.
  - $\epsilon$  es el término de error, que representa la diferencia entre el valor observado de  $y$  y el valor predicho por la ecuación de regresión.
2. **Estimación de parámetros.** Los parámetros de regresión lineal múltiple ( $\beta_0, \beta_1, \beta_2, \dots, \beta_k$ ) se estiman utilizando el método de mínimos cuadrados ordinarios (OLS). El objetivo de este método es encontrar los valores de  $\beta_0, \beta_1, \beta_2, \dots, \beta_k$  que minimicen la suma de los cuadrados de los residuos.
  3. **Coeficiente de determinación múltiple ( $R^2$ ).** El coeficiente de determinación múltiple ( $R^2$ ) es una medida del ajuste del modelo. Expresa la parte de la variación total de  $y$  explicada por las variables independientes ( $x_1, x_2, \dots, x_k$ ). Un valor de  $R^2$  cercano a 1 indica que el modelo explica una gran parte de la variación en  $y$ , mientras que un valor cercano a 0 indica que el modelo explica una pequeña parte de la variación en  $y$ .

## Aplicaciones de la regresión lineal múltiple

- Análisis factorial: Determinar los factores que más afectan la variación de la variable dependiente.

- **Predicción:** Predice el valor de la variable independiente ( $y$ ) para un nuevo conjunto de valores de las variables independientes ( $x_1, x_2, \dots, x_k$ ).
- **Modelo de negocio:** Desarrollar modelos para predecir la demanda de productos o servicios, el comportamiento del consumidor o el riesgo financiero.

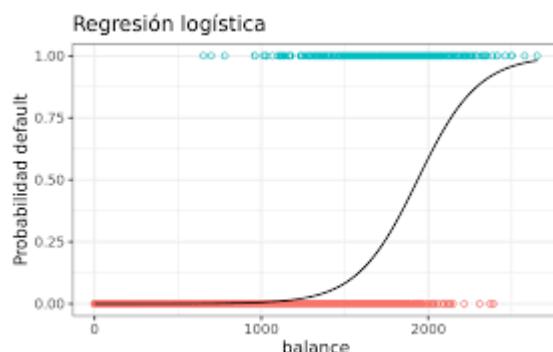
### 3.2.3. Regresión Logística

#### Definición.

La regresión logística es un método estadístico para representar la relación entre una variable dependiente binaria (que solo puede tener dos valores, como "sí" o "no", "éxito" o "fracaso", etc.) y una o más variables independientes. La regresión logística predice la probabilidad de que ocurra un evento específico, utilizado en problemas de clasificación; en contraste con la regresión lineal, que predice valores continuos (Harrell, 2015) (MacQueen et al., 2013) (Ng, 2016).

**Figura 28:**

*Modelo de Regresión logística*



Nota: [https://cienciadedatos.net/documentos/27\\_regresion\\_logistica\\_simple\\_y\\_multiple](https://cienciadedatos.net/documentos/27_regresion_logistica_simple_y_multiple)

#### Fórmulas asociadas al modelo.

1. **Función de hipótesis.** En la regresión logística, la función de hipótesis indica la probabilidad de que ocurra el evento de interés (por ejemplo, "sí") para una combinación específica de valores de las variables independientes:

$$h(x) = P(y = 1 | x) = 1 / (1 + \exp(-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n)))$$

Donde:

- $h(x)$  es la probabilidad predicha de que  $y = 1$  (evento de interés) para un conjunto de valores de entrada  $x = (x_1, x_2, \dots, x_n)$ .
  - $\beta_0, \beta_1, \beta_2, \dots, \beta_n$  son los parámetros del modelo que se estiman utilizando datos de entrenamiento.
  - $x_1, x_2, \dots, x_n$  son los valores de las variables independientes para el punto de datos en cuestión.
2. **Función de costo.** La función de costo evalúa la discrepancia entre los resultados reales y las predicciones del modelo. La función de costo de la entropía cruzada binaria se utiliza en la regresión logística:
- $$L(\beta) = -\sum_{i=1}^N [y_i * \log(h(x_i)) + (1 - y_i) * \log(1 - h(x_i))]$$
- Donde:
- $L(\beta)$  es el valor de la función de costo para un conjunto de datos dado.
  - $N$  es el número de puntos de datos en el conjunto de entrenamiento.
  - $y_i$  es el valor real de la variable dependiente para el punto de datos  $i$  (1 si ocurre el evento de interés, 0 en caso contrario).
  - $h(x_i)$  es la probabilidad predicha de que  $y = 1$  para el punto de datos  $i$ .
3. **Optimización del modelo.** El objetivo del entrenamiento de un modelo de regresión logística es encontrar los valores de los parámetros ( $0, 1, 2, \dots, n$ ) que minimizan la función de costo. Esto se puede lograr mediante el uso de algoritmos de optimización como el descenso del gradiente.

### Aplicaciones de la regresión logística.

- Previsión de riesgo crediticio: Estimar la probabilidad de que el cliente incumpla con el préstamo.
- Detección de fraude: identifique transacciones fraudulentas con tarjetas de crédito.
- Diagnóstico médico: predecir la probabilidad de enfermedad en función de los síntomas y factores de riesgo.
- Marketing Analytics: Determinar la efectividad de las campañas publicitarias

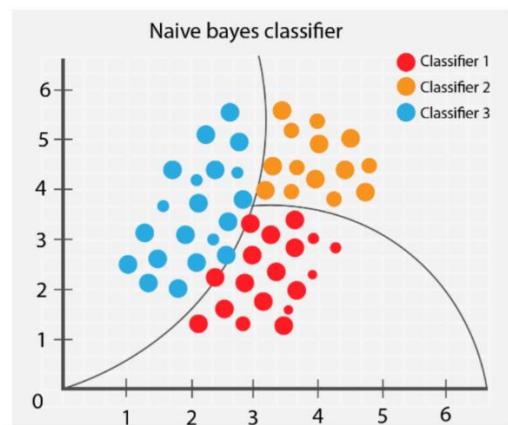
### 3.2.4. Naive Bayes

#### Definición.

El clasificador Naive Bayes, también conocido como Bayesiano ingenuo, es un algoritmo probabilístico de aprendizaje basado en el teorema de Bayes y el supuesto de independencia condicional de características. Este algoritmo se utiliza para tareas de clasificación que tienen como objetivo asignar un conjunto de datos a una de varias clases predefinidas (Mishra & Sharma, 2020) (Singh & Kaur, 2019) (Zhang & Sun, 2020).

**Figura 29:**

Modelo de Naive Bayes



*Nota:* [https://www.researchgate.net/figure/Naive-Bayes-Classifier-The-Naive-Bayes-algorithm-is-a-statistical-method-in-machine\\_fig1\\_372837562](https://www.researchgate.net/figure/Naive-Bayes-Classifier-The-Naive-Bayes-algorithm-is-a-statistical-method-in-machine_fig1_372837562)

#### Formulas asociadas al modelo.

1. **Teorema de Bayes.** El teorema de Bayes proporciona la base para las probabilidades condicionales en Naive Bayes. Se expresa como:

$$P(H | E) = (P(E | H) * P(H)) / P(E)$$

Donde:

- $P(H | E)$  es la probabilidad de la hipótesis  $H$  dado el evento  $E$ .
- $P(E | H)$  es la probabilidad del evento  $E$  dado la hipótesis  $H$ .
- $P(H)$  es la probabilidad a priori de la hipótesis  $H$ .
- $P(E)$  es la probabilidad a priori del evento  $E$ .

2. **Probabilidad a priori.** La probabilidad a priori de una hipótesis o evento indica su probabilidad antes de que se observe la evidencia. En Naive

Bayes, las probabilidades previas de cada clase (C) se calculan utilizando la frecuencia relativa de cada clase en el conjunto de datos de entrenamiento

3. **Probabilidad condicional.** La probabilidad condicional de un evento dada una hipótesis representa la probabilidad de que el evento sepa que la hipótesis es verdadera. En Naive Bayes, las probabilidades condicionales de cada característica (X) para cada clase (C) se calculan utilizando la frecuencia relativa de cada característica en cada clase.
4. **Clasificación.** Para clasificar los nuevos datos (x), se calcula la probabilidad de cada clase de datos (C) y se determina la clase con mayor probabilidad. Se utiliza la siguiente fórmula:

$$P(C | x) = (P(x | C) * P(C)) / P(x)$$

Donde:

- $P(C | x)$  es la probabilidad de la clase C dado el dato x.
- $P(x | C)$  es la probabilidad del dato x dado la clase C.
- $P(C)$  es la probabilidad a priori de la clase C.
- $P(x)$  es la probabilidad a priori del dato x.

### Aplicaciones de Naive Bayes.

- Clasificación de correo electrónico: se utiliza ampliamente para clasificar correos electrónicos como spam o no spam.
- Detección de intrusiones: se puede utilizar para detectar intrusiones en redes de datos.
- Análisis de sentimiento: se puede utilizar para clasificar el sentimiento de un texto como positivo, negativo o neutral.
- Sistemas de recomendación: se puede utilizar para recomendar productos o servicios a los usuarios en función de sus compras o preferencias anteriores

### 3.2.5. KNN

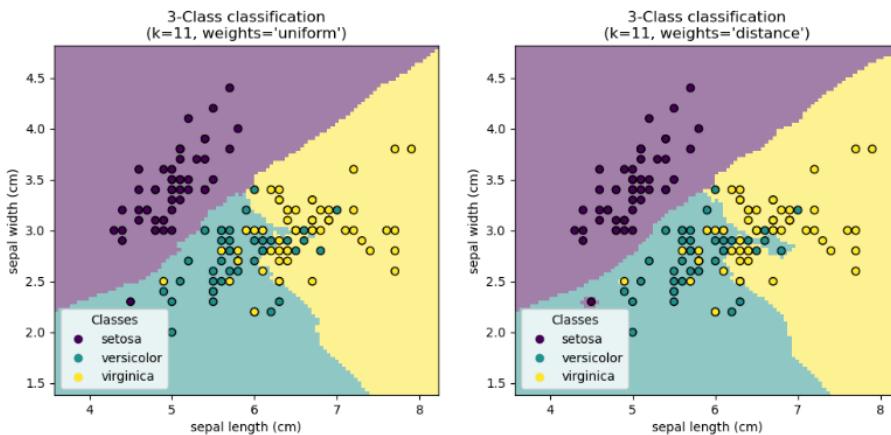
#### Definición.

El algoritmo K-Nearest Neighbors (KNN) es un algoritmo de aprendizaje automático supervisado ampliamente utilizado para tareas de clasificación y regresión. Durante la clasificación, KNN clasifica un nuevo punto de datos según las clases de la mayoría de sus vecinos más cercanos en el espacio de

características (Bashir & Augenthaler, 2020) (Carmona & Torres, 2019) (Dong et al., 2020).

**Figura 30:**

*Modelo KNN*



Nota: <https://scikit-learn.org/stable/modules/neighbors.html>

### Fórmulas asociadas al modelo.

1. **Distancia entre puntos.** Para determinar los vecinos más cercanos, KNN utiliza un medidor de distancia para calcular la distancia entre el nuevo punto de datos ( $x$ ) y cada punto de datos en el conjunto de entrenamiento. Algunas medidas de distancia comunes son:

#### Distancia euclíadiana

$$d(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

Donde:

- $d(x, y)$  es la distancia euclíadiana entre los puntos  $x$  e  $y$ .
- $n$  es el número de dimensiones en el espacio de características.
- $x_i$  y  $y_i$  son las coordenadas de los puntos  $x$  e  $y$  en la dimensión  $i$ .

#### Distancia de Manhattan

$$d(x, y) = \sum_{i=1}^n |x_i - y_i|$$

Donde:

- $d(x, y)$  es la distancia de Manhattan entre los puntos  $x$  e  $y$ .
- $n$  es el número de dimensiones en el espacio de características.
- $x_i$  y  $y_i$  son las coordenadas de los puntos  $x$  e  $y$  en la dimensión  $i$ .

2. **Clasificación.** Una vez que se identifican los K vecinos más cercanos del nuevo punto de datos, KNN asigna la clase más común entre esos vecinos al nuevo punto. En caso de empate se pueden utilizar diferentes estrategias de suavizado, como elegir la clase de mayoría absoluta o la clase con la distancia media más cercana al nuevo punto.
3. **Selección de K.** El valor de K, que representa el número de vecinos más cercanos observados, es un parámetro importante de KNN. Un valor de K demasiado pequeño puede dar lugar a una clasificación sobreajustada, mientras que un valor de K demasiado grande puede dar lugar a una clasificación inadecuada. La elección óptima de K generalmente se realiza mediante técnicas de validación cruzada o búsqueda de hiperparámetros

### Aplicaciones de KNN.

- Clasificación de imágenes: se puede utilizar para clasificar imágenes en diferentes categorías, como gatos, perros o paisajes.
- Sistemas de recomendación: se puede utilizar para recomendar productos o servicios a los usuarios según sus preferencias o compras anteriores.

Detección de valores atípicos: se puede utilizar para detectar puntos de datos anormales o atípicos en un conjunto de datos

### 3.2.6. Árboles de decisión

#### Definición.

Los árboles de decisión son un algoritmo de aprendizaje automático supervisado ampliamente utilizado para tareas de clasificación y regresión. En relación con la clasificación, los árboles de decisión se forman dividiendo recursivamente el espacio de características en áreas más pequeñas y homogéneas. Cada partición se realiza en función de una característica específica y un valor umbral, clasificando los datos según la ruta en el árbol (De Luca & Piemontese, 2019) (Rokach, 2019) (Sahoo & Panda, 2019).

**Figura 31:**

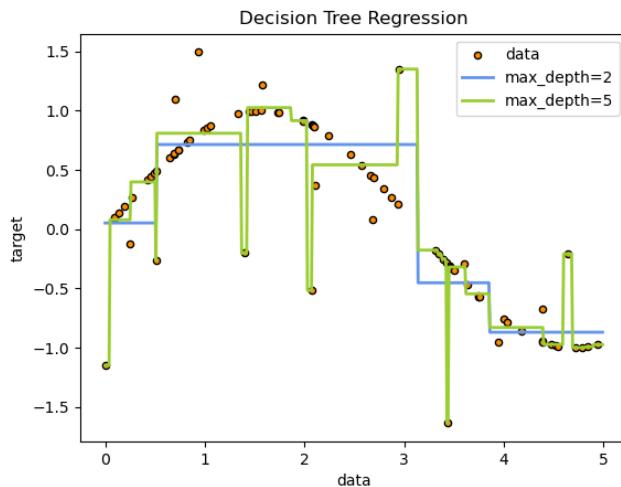
*Modelo de clasificación con árboles de decisión*



Nota: <https://scikit-learn.org/stable/modules/tree.html>

**Figura 32:**

*Modelo de regresión con árboles de decisión*



Nota: <https://scikit-learn.org/stable/modules/tree.html>

**Fórmulas asociadas al modelo.**

1. **Ganancia de información.** La validación de datos se utiliza para medir la calidad de la distribución del árbol de decisión. Se calcula como la disminución de la entropía después de la división. La entropía se define

como una medida de incertidumbre en un conjunto de datos. La fórmula para acceder a la información es:

$$\text{Gain}(S, A) = \text{Entropy}(S) - \sum_{v \in \text{Values}(A)} |S_v|/|S| * \text{Entropy}(S_v)$$

Donde:

- Gain(S, A) es la ganancia de información para dividir el conjunto de datos S en función del atributo A.
- Entropy(S) es la entropía del conjunto de datos S.
- Values(A) es el conjunto de valores posibles para el atributo A.
- S<sub>v</sub> es el subconjunto de datos S que contiene los puntos de datos con el valor v para el atributo A.
- |S| es el tamaño del conjunto de datos S.
- |S<sub>v</sub>| es el tamaño del subconjunto de datos S<sub>v</sub>.

2. **Entropía.** La entropía se calcula utilizando la siguiente fórmula:

$$\text{Entropy}(S) = -\sum_{i=1}^c p_i * \log_2(p_i)$$

Donde:

- Entropy(S) es la entropía del conjunto de datos S.
- c es el número de clases en el conjunto de datos S.
- p<sub>i</sub> es la proporción de puntos de datos en la clase i.

3. **Selección de la división.** Para cada nodo del árbol de decisión, se selecciona la distribución que maximiza la ganancia de información. Esto asegura que el árbol se construya de manera que minimice la incertidumbre en cada paso.

4. **Criterio de parada.** El árbol de decisión se expande hasta que se cumple el criterio de parada. Los criterios de parada comunes son:

- Alcanzar la profundidad máxima del árbol.
- Llegue al nodo con el número mínimo de puntos de datos.
- Los datos aumentan por debajo de un cierto umbral.

### Aplicaciones de los árboles de decisión.

- Clasificación de clientes: los árboles de decisión se pueden utilizar para clasificar a los clientes en diferentes segmentos de riesgo crediticio o para predecir su comportamiento de compra.

- Diagnóstico médico: los árboles de decisión se pueden utilizar para ayudar a los médicos a diagnosticar enfermedades basándose en los síntomas de los pacientes.
  - Análisis de datos financieros: los árboles de decisión se pueden utilizar para analizar datos financieros y tomar decisiones de inversión.
- Detección de fraude: los árboles de decisión se pueden utilizar para detectar transacciones fraudulentas con tarjetas de crédito.

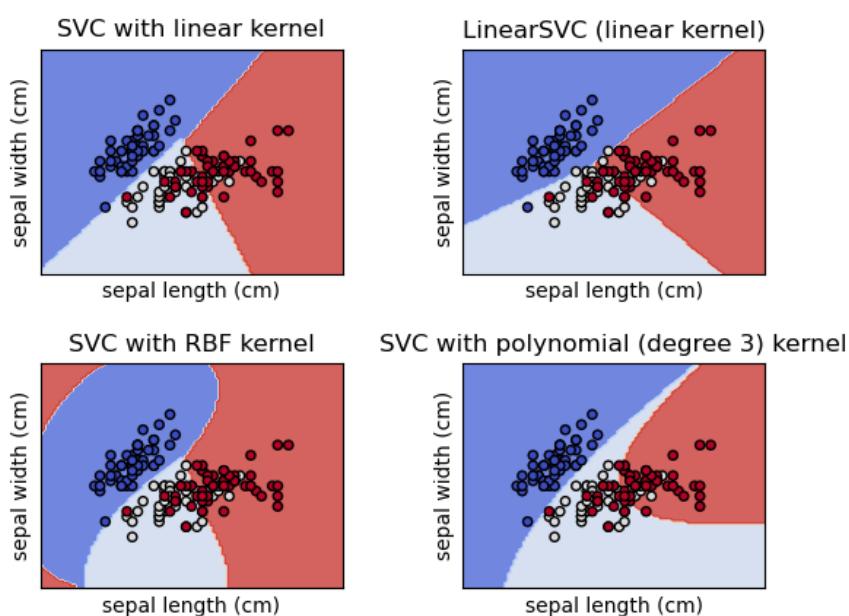
### 3.2.7. Super Vector Machine

#### Definición.

Support Vector Machines (SVM) es un algoritmo de aprendizaje automático supervisado ampliamente utilizado para tareas de clasificación y regresión. En clasificación, SVM intenta encontrar un hiperplano en el espacio de características que maximice el margen entre puntos de datos de diferentes clases. Este hiperplano representa la mejor separación de clases, lo que permite clasificar nuevos puntos de datos en función de su ubicación en relación con el hiperplano (Cristianini & Schölkopf, 2020) (Géron, 2019) (Li & Sun, 2020).

**Figura 33:**

Modelo de SVM



Nota: <https://scikit-learn.org/stable/modules/svm.html>

## Fórmulas asociadas al modelo.

1. **Función objetivo.** La función objetivo SVM intenta maximizar el margen, que se define como la distancia mínima entre el hiperplano y los puntos de datos más cercanos (vectores de soporte) de cada clase. La fórmula de la función objetivo es:

$$\text{maximize margin} = 1/\|w\|$$

Donde:

- $w$  es el vector normal al hiperplano.
- $\|w\|$  es la norma del vector  $w$ .

2. **Restricciones.** Las restricciones de SVM garantizan que los puntos de datos de clase principales estén en un lado del hiperplano de al menos 1, y los puntos de datos de clase menores estén en el lado opuesto o hiperplano. Los límites se expresan de la siguiente manera:

$$y_i * (w^T x_i + b) \geq 1 \text{ para } i = 1, \dots, n$$

Donde:

- $y_i$  es la etiqueta de clase del punto de datos  $i$  (-1 para la clase menor, 1 para la clase mayor).
- $x_i$  es el vector de características del punto de datos  $i$ .
- $b$  es el sesgo del hiperplano.

3. **Solución.** El problema de optimización SVM se resuelve utilizando el método de Lagrange o algoritmos de programación cuadrática. El resultado es un hiperplano óptimo que maximiza el margen y separa clases de manera efectiva.

## Aplicaciones de SVM.

- Clasificación de imágenes: SVM se usa ampliamente para clasificar imágenes en diferentes categorías, como gatos, perros o paisajes.
- Detección de spam: SVM se puede utilizar para clasificar correos electrónicos como spam o no spam.
- Análisis de ADN: SVM se puede utilizar para clasificar secuencias de ADN en diferentes clases, como promotores, intrones o exones.
- Reconocimiento de voz: SVM se puede utilizar para identificar diferentes hablantes o palabras en grabaciones de audio

### 3.3. Evaluación de modelos

#### 3.3.1. Métricas de evaluación en algoritmos de clasificación

##### Matriz de confusión

La matriz de confusión es una herramienta básica de aprendizaje automático que se puede utilizar para evaluar el rendimiento de un modelo de clasificación. Esta matriz tabular resume los resultados del modelo de clasificación y muestra el número de casos clasificados correctamente y mal clasificados en cada clase.

**Figura 34:**

*Matriz de confusión*

|                  |     | Actual Values  |                |
|------------------|-----|----------------|----------------|
|                  |     | Yes            | No             |
| Predicted Values | Yes | True Positive  | False Positive |
|                  | No  | False Negative | True Negative  |

Nota: <https://www.datasource.ai/es/data-science-articles/comprendiendo-la-matriz-de-confusion-y-como-implementarla-en-python>

##### Métricas.

A partir de la matriz de confusión, se pueden calcular importantes medidas de desempeño para evaluar la precisión del modelo de clasificación. Algunas de las medidas más comunes incluyen:

1. **Precisión:** Representa la proporción de instancias correctamente clasificadas en una clase específica. Se calcula como:  

$$\text{Precisión} = (\text{Verdaderos Positivos}) / (\text{Verdaderos Positivos} + \text{Falsos Positivos})$$
2. **Exactitud:** Representa la proporción de instancias correctamente clasificadas en todas las clases. Se calcula como:  

$$\text{Exactitud} = (\text{Verdaderos Positivos} + \text{Verdaderos Negativos}) / (\text{Total})$$

3. **Recall:** Representa la proporción de instancias positivas correctamente identificadas por el modelo. Se calcula como:  

$$\text{Recall} = (\text{Verdaderos Positivos}) / (\text{Verdaderos Positivos} + \text{Falsos Negativos})$$
4. **F1-Score:** Es una medida que combina la precisión y el recall, proporcionando una evaluación general del rendimiento del modelo. Se calcula como:  

$$\text{F1-Score} = 2 * (\text{Precisión} * \text{Recall}) / (\text{Precisión} + \text{Recall})$$
5. **Especificidad:** Representa la proporción de instancias negativas correctamente identificadas por el modelo. Se calcula como:  

$$\text{Especificidad} = (\text{Verdaderos Negativos}) / (\text{Falsos Positivos} + \text{Verdaderos Negativos})$$

### Curva ROC.

Una curva ROC (curva operativa del receptor) es un gráfico que muestra el rendimiento de un modelo de clasificación en todos los umbrales de clasificación. Esta curva representa dos parámetros (Fawcett, 2020) (Hand & Henley, 2019):

- Tasa de verdaderos positivos (TPR) es sinónimo de exhaustividad y se define de la siguiente manera:

$$\text{TPR} = \text{TP} / (\text{TP} + \text{FN})$$

Donde:

- TP: total positivos
- FN: falsos negativos

- Tasa de falsos positivos (FPR) y se define de la siguiente manera:

$$\text{FPR} = \text{FP} / (\text{FP} + \text{TN})$$

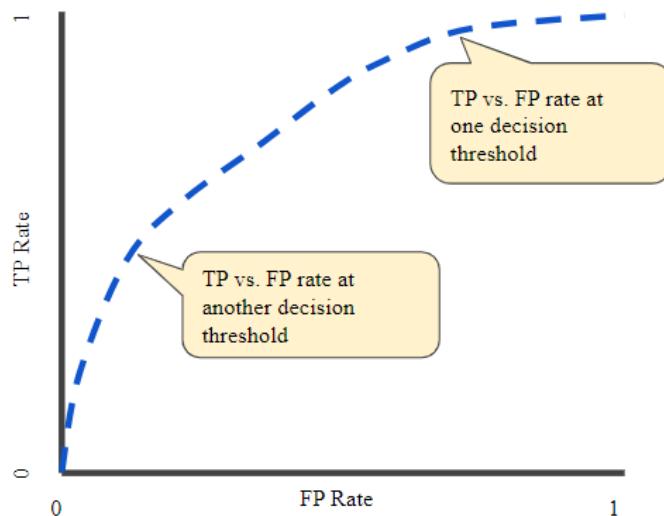
Donde:

- FP: falsos positivos
- TN: total negativos

La curva ROC representa la TPR y la FPR en diferentes umbrales de clasificación. Al reducir el umbral de clasificación, se clasifican más objetos como positivos, lo que aumenta tanto los valores falsos positivos como los verdaderos positivos. En la siguiente figura se muestra una curva ROC típica.

**Figura 35:**

Curva ROC



Nota: <https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc?hl=es-419>

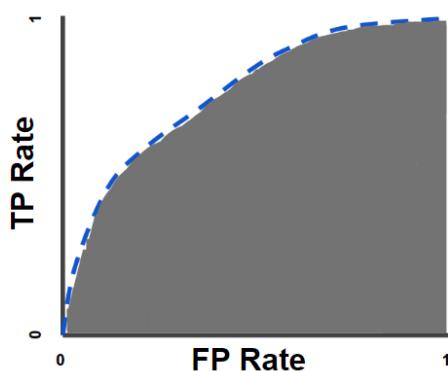
Para calcular las puntuaciones de la curva ROC, podríamos estimar el modelo de regresión logística varias veces con diferentes umbrales de clasificación, pero esto sería ineficiente. Afortunadamente, existe un potente algoritmo basado en órdenes que puede proporcionarnos esta información, llamado AUC.

### Área AUC

AUC significa "área bajo la curva ROC". En otras palabras, el AUC mide toda el área bidimensional bajo toda la curva ROC como un cálculo integral entre (0,0) y (1,1) (Fawcett, 2020) (Hand & Henley, 2019).

**Figura 36:**

Área AUC



Nota: <https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc?hl=es-419>

AUC proporciona una medida del rendimiento general en todos los umbrales de clasificación posibles. Una forma de interpretar el AUC es como la probabilidad de que el modelo clasifique mejor un ejemplo aleatorio positivo que un ejemplo aleatorio negativo.

El valor AUC es 0-1. Un modelo con predicciones 100% incorrectas tiene un AUC de 0,0; cuyas predicciones son 100% correctas tienen un AUC de 1,0.

### **3.3.2. Métricas de evaluación en algoritmos de regresión**

Las métricas de regresión son métricas que se utilizan para evaluar el rendimiento de los modelos de aprendizaje automático en tareas de regresión. Estas métricas muestran con qué precisión el modelo predice los valores continuos en comparación con los valores reales (Chai & Draxler, 2019) (Fernández-Escudero & Zazo, 2020).

#### **Métricas.**

##### **1. Error Cuadrático Medio (MSE):**

$$MSE = (1/n) * \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Donde:

- MSE: Error Cuadrático Medio.
- n: Número de muestras en el conjunto de datos.
- $y_i$ : Valor real de la i-ésima muestra.
- $\hat{y}_i$ : Valor predicho por el modelo para la i-ésima muestra.

##### **2. Raíz Cuadrada Media del Error (RMSE):**

$$RMSE = \sqrt{MSE}$$

La RMSE es la raíz cuadrada del MSE y se expresa en las mismas unidades que los valores reales.

##### **3. Error Absoluto Medio (MAE):**

$$MAE = (1/n) * \sum_{i=1}^n |y_i - \hat{y}_i|$$

Donde:

- MAE: Error Absoluto Medio.
- n: Número de muestras en el conjunto de datos.
- $y_i$ : Valor real de la i-ésima muestra.
- $\hat{y}_i$ : Valor predicho por el modelo para la i-ésima muestra.

#### 4. Coeficiente de Determinación ( $R^2$ ):

$$R^2 = 1 - (\sum_{i=1}^n (y_i - \hat{y}_i)^2) / (\sum_{i=1}^n (y_i - \bar{y})^2)$$

Donde:

- $R^2$ : Coeficiente de Determinación.
- $n$ : Número de muestras en el conjunto de datos.
- $y_i$ : Valor real de la  $i$ -ésima muestra.
- $\hat{y}_i$ : Valor predicho por el modelo para la  $i$ -ésima muestra.
- $\bar{y}$ : Media de los valores reales.

El coeficiente de determinación ( $R^2$ ) es una medida estadística ampliamente utilizada para evaluar el ajuste de un modelo de regresión lineal. Se expresa como un valor entre 0 y 1, donde:

- $R^2 = 0$ : indica que el modelo no explica ninguna parte de la variación en los datos de respuesta alrededor de la media. Es decir, el modelo no es mejor que simplemente promediar los datos para predecir valores futuros.
- $R^2 = 1$ : indica que el modelo explica toda la variación en los datos de respuesta alrededor de la media. Es decir, el modelo predice perfectamente valores futuros.

#### Interpretación de $R^2$ .

Un valor de  $R^2$  cercano a 1 indica que el modelo de regresión lineal se ajusta bien a los datos y tiene un buen poder predictivo. Sin embargo, es importante señalar que un valor alto de  $R^2$  no garantiza necesariamente que el modelo sea causal o que no haya otras variables significativas en el modelo.

CAPITULO

03

# IA APLICADA CON NLP Y MACHINE LEARNING



## IA aplicada con NLP y machine learning

### 4.1. Modalidad de la investigación

Vera et al. (2018) expresan que la metodología de la investigación se define como una disciplina que mediante la sistematización y evaluación de diversos métodos técnicos que facilitan la búsqueda de información y construcción del conocimiento científico. La metodología consta de un conjunto de procedimientos que tienen como objetivo, diferentes procesos como recolección, clasificación y validación de datos.

En este capítulo se detallarán las fases y secuencias realizadas en el proyecto de carácter investigativo, se identificará la modalidad y tipos de investigación empleadas, se señalará las técnicas empleadas en cada prueba realizada.

La modalidad de investigación es de tipo bibliográfico que constó aproximadamente de un 40%, basada en la búsqueda de documentos científicos que permitan obtener la información necesaria de los diferentes puntos que componen al trabajo y que aborde el tema general y un 60% de modalidad de campo correspondiente a la encuesta realizada a las personas contagiadas de Covid-19.

### 4.2. Tipo de investigación

Existen diferentes tipos y niveles de investigación, en este proyecto de investigación, donde se proponen el uso de los siguientes tipos de investigación:

#### 4.2.1. Investigación Exploratoria

La investigación exploratoria como lo indican Castro et al. (2018), se basa en el análisis e investigación de aspectos relevantes los cuales no han sido analizados con un mayor grado profundidad, en pocas palabras se trata de indagar o explorar lo que permite que futuras investigaciones tengan acceso a información de temas ya tratados.

Para el proyecto de investigación se utilizó la investigación exploratoria para la

extracción, análisis e identificación de información teórica acerca de las técnicas de procesamiento de lenguaje natural, definición, que funciones realizan, que permita poseer un enfoque general de todos ellos para la realización de la propuesta.

#### **4.2.2. Investigación diagnóstica**

González (2020) da a conocer que la investigación de diagnóstico es un procedimiento que determina lo que sucede en un escenario o situación establecida, en pocas palabras se analiza la cadena de sucesos con el propósito de identificar aquellos factores que intervinieron e iniciaron en la aparición de un fenómeno.

El presente trabajo investigativo aplicó la investigación diagnóstica para la estimación de la población contagiada por Covid-19 de la zona 8 del Guayas (Guayaquil, Durán y Samborondón) mediante uso de encuestas para determinar características tales como síntomas, recomendaciones, edad, variante de contagio, dosis de vacunas aplicadas

#### **4.2.3. Investigación cuasi-experimental**

Ramos (2021) menciona que para el diseño cuasi-experimental, las variables de investigación son en base a una asignación no aleatoria del grupo de estudio, es decir no se asigna al azar, en cambio se trabaja con grupos ya establecidos o formados.

En el proyecto, se utiliza una investigación cuasi experimental, donde la variable dependiente corresponde a la categorización del texto de conversaciones textuales de personas con Covid-19 y la variable independiente corresponde a técnicas NLP con algoritmos de clasificación de Machine Learning ya establecidas siendo el sujeto de estudio.

#### **4.2.4. Investigación evaluativa**

Garduño et al. (2019) expresan que la investigación evaluativa es un recurso valioso para el proceso de toma de decisiones en el mejoramiento de procesos de diversos indoles ya que se respalda en metodologías diversas.

Para el proyecto, se aplica la investigación evaluativa para evaluar y estimar los

resultados obtenidos en la ejecución del modelo experimentado individualmente por cada conjunto de técnicas de NLP, la evaluación se realiza mediante métricas y matriz de confusión, con la intención de determinar la efectividad y precisión por cada técnica.

### 4.3. Diseño metodológico de la investigación

A través de la representación gráfica de la figura 37, se da a conocer las fases del diseño metodológico de esta investigación.

**Figura 37:**

*Diseño metodológico*



*Nota:* Diseño metodológico. Autores (2024)

### 4.4. Metodología de investigación

#### 4.4.1. Definición del problema

Existen diferentes técnicas de NLP, el uso adecuado de las mismas en la identificación y clasificación de texto, en donde el mejor resultado es lo que se espera de la aplicación de un algoritmo o técnica, para ello se experimenta y se realiza una evaluación o validación para visualizar que tan competente es la técnica.

#### 4.4.2. Objetivo principal

Construir un análisis evolutivo de las diferentes técnicas de Procesamiento de Lenguaje Natural (NLP) a través de una herramienta tecnológica de IA para identificar el algoritmo eficaz en el Procesamiento del Lenguaje Natural de conversaciones textuales etiquetadas de personas contagiadas de Covid-19

#### 4.4.3. Técnicas de Procesamiento de Lenguaje Natural

Las técnicas aplicadas en el modelo son:

- Técnica de Tokenización
- Técnica de Stopwords
- Técnica de Part of Speech
- Técnica de Lematización

#### 4.4.4. Entrenamiento y aprendizaje del modelo

Se utiliza el dataset preprocesado y aplicado las técnicas tales como: técnica de tokenización, técnica de de stopwords, técnica de lematización y la técnica de Part of Speech. Esta data servirá para el entramiento correspondiente en distintos modelos como: Modelo Naive Bayes, Modelo k-nearest neighbors y Modelo de Regresión Logística.

#### 4.4.5. Medición y validación de resultados

A través de las diferentes métricas, se analiza que modelo tiene mejores resultados en base a la data set procesada.

#### 4.4.6. Población y muestra

##### 4.4.6.1. Población

Castro (2019) indica que la población también considerada como universo de estudio, es el conjunto total de individuos, objetos o medidas que poseen características comunes para ser considerado para un estudio. También indica que trabajar con la población conlleva a algunos problemas como: Dificultad para ejecutar toda la población, consumo elevado de recursos como tiempo y dinero.

La población, correspondiente a las personas que fueron contagiadas de Covid-

19 en la zona 8 del Guayas de las ciudades de Guayaquil, Durán y Samborondón. Las encuestas se realizaron a este sector de población sin definir el tamaño específico inicial de la misma, considerando que el tamaño de los encuestados corresponde a 5570 personas encuestadas, además de 190 encuestas correspondientes a la extracción de las redes sociales de Twitter y Facebook, el data set unificado final para los registros de síntomas y recomendaciones corresponde a 4140 registros. Además, se considera la opinión de expertos en Inteligencia Artificial y Sistemas como población correspondiente para el análisis de factibilidad del proyecto.

#### 4.4.6.2. Muestra

La técnica de muestreo usada es del tipo aleatorio simple, donde la población de estudio es escogida con la misma probabilidad de forma al azar. Para el cálculo de la población se usó la fórmula de población infinita que se muestra a continuación:

$$n = \frac{Z_{\infty}^2 \times p \times q}{e^2}$$

Donde:

Z = Parámetro estadístico que depende del nivel de confianza

n = Tamaño de la muestra buscado

e = Error de estimación

p = Probabilidad de Éxito

q = Probabilidad de Fracaso

Para la población del caso de estudio, tenemos los siguientes datos.

$$n = \frac{(2.57^2) \times 0.5 \times 0.5}{(0.02)^2} = 4128.06$$

Donde:

Z = 6.6

e = 0.02

p = 0.5

q = 0.5

#### 4.4.7. Técnica de recolección de datos.

Considerado como una serie de acciones que se realizan para poder reunir información que permita conseguir los objetivos propuestos en una investigación. (Arispe et al., 2020)

#### 4.4.7.1. Encuesta

Esta técnica usada para investigaciones permite realizar observaciones de forma indirecta ante un hecho, mediante la intervención de los entrevistados o encuestados, usado para recopilar información en grandes cantidades o masiva. (Ramos et al., 2017)

Para el proyecto de titulación presente se utilizó a la encuesta como técnica de recolección de datos e información, aplicada a personas contagiadas de Covid-19 de la zona 8 del Guayas que corresponde a los cantones de Guayaquil, Durán y Samborondón.

#### 4.4.8. Instrumento de medición

Arispe et al. (2020) indican que gracias a los instrumentos hacen posible la aplicación de las técnicas teniendo en cuenta los diversos indicativos. Es imprescindible contar con la validez del contenido específico y confiabilidad de los datos.

#### 4.4.8. Cuestionario

Las encuestas usan como instrumentos a los cuestionarios, se tiene que un cuestionario es utilizado de forma común en investigaciones de carácter científico, compuesto de preguntas dirigidas hacia un encuestado. (Arias, 2019)

### 4.5. Técnica para el procesamiento y análisis de los datos

#### 4.5.8. Encuesta para personas contagiadas de Covid-19

En este apartado se presentarán estadísticamente las preguntas de tipo abiertas que estaban presente en la encuesta de donde está basado el data set. Existe la presencia de preguntas de tipo abiertas y cerradas en el data set. En el Anexo 1 se adjunta las preguntas abiertas.

#### 4.5.9. Análisis de las preguntas cerradas

A continuación, se describe el análisis de las preguntas cerradas a las personas contagiadas de Covid-19.

##### Pregunta 2: Seleccione la edad

**Tabla 1:**

*Edad*

| Opciones de respuesta | Frecuencia Absoluta | Frecuencia Relativa |
|-----------------------|---------------------|---------------------|
| Entre 18 a 25 años    | 2430                | 58.70%              |
| Entre 26 a 40 años    | 894                 | 21.60%              |
| Entre 41 a 64 años    | 684                 | 16.52%              |
| De mayor a igual a 65 | 70                  | 1.69%               |
| Vacías                | 62                  | 1.49%               |
| <b>TOTAL</b>          | <b>4140</b>         | <b>100.00%</b>      |

*Nota:* En esta tabla se muestran los valores absolutos y relativos que corresponden al proceso de tabulación de la pregunta 2 de la encuesta dirigida a personas contagiadas de covid-19 de la Zona 8 de la provincia Guayas (Guayaquil, Durán y Samborondón). Autores (2024).

Análisis: De un total de 4140 encuestados se puede observar que el 58.70% corresponde la cantidad de personas con un rango de edad de 18 a 25 años, el 21.60% corresponde a la cantidad de personas con un rango de edad de 26 a 40 años, el 16.52% a la cantidad de personas con un rango de edad de 26 a 64 años y el 1.69% corresponde a personas con rango que sea mayor o igual a los 65 años.

##### Pregunta 3: Género

**Tabla 2:**

*Género*

| Opciones de respuesta | Frecuencia Absoluta | Frecuencia Relativa |
|-----------------------|---------------------|---------------------|
| Femenino              | 2336                | 56.42%              |
| Masculino             | 1804                | 43.58%              |
| <b>TOTAL</b>          | <b>4140</b>         | <b>100.00%</b>      |

*Nota:* En esta tabla se muestran los valores absolutos y relativos que corresponden al proceso de tabulación de la pregunta 3 de la encuesta dirigida a personas contagiadas de covid-19 de la Zona 8 de la provincia Guayas (Guayaquil, Durán y Samborondón). Autores (2024).

Análisis: Los resultados que presenta este grafico indica que 56.42% de personas corresponden al sexo femenino, el cual es superior al porcentaje de las personas que del sexo masculino que se le atribuye a un porcentaje del 43.58%.

#### Pregunta 4: ¿Qué variante del virus lo contagió?

**Tabla 3:**

*Variante del virus lo contagio*

| Opciones de respuesta | Frecuencia Absoluta | Frecuencia Relativa |
|-----------------------|---------------------|---------------------|
| Alfa                  | 2790                | 67.39%              |
| No sabe               | 1009                | 24.37%              |
| Delta                 | 224                 | 5.41%               |
| Gamma                 | 86                  | 2.08%               |
| Omicron               | 28                  | 0.68%               |
| Vacías                | 3                   | 0.07%               |
| <b>TOTAL</b>          | <b>4140</b>         | <b>100,00%</b>      |

*Nota:* En esta tabla se muestran los valores absolutos y relativos que corresponden al proceso de tabulación de la pregunta 4 de la encuesta dirigida a personas contagiadas de covid-19 de la Zona 8 de la provincia Guayas (Guayaquil, Durán y Samborondón). Autores (2024).

**Análisis:** Los resultados presentes en este grafico muestran un gran porcentaje que corresponde al 67.39% de personas que padecían la variante de Alfa del virus covid-19, el 24.37% corresponde a personas que desconocían que variante tenían en ese momento, el 5.41% pertenecían a las personas que padecían la variante delta del virus, el 2.08% a personas que padecían la variante Gamma del virus, el 0.68% a personas que padecían la variante Ómicron del virus y por ultimo con un total mínimo del 0.07% corresponde a celdas que no contenían información.

#### Pregunta 6: ¿Nivel de intensidad que tuvo los síntomas?

**Tabla 4:**

*Nivel de intensidad que tuvo los síntomas*

| Opciones de respuesta | Frecuencia Absoluta | Frecuencia Relativa |
|-----------------------|---------------------|---------------------|
| Leve                  | 1420                | 34.30%              |
| Medio                 | 1668                | 40.29%              |
| Fuerte                | 614                 | 14.83%              |
| Sin síntomas          | 433                 | 10.46%              |
| Vacías                | 5                   | 0.12%               |
| <b>TOTAL</b>          | <b>4140</b>         | <b>100,00%</b>      |

*Nota:* En esta tabla se muestran los valores absolutos y relativos que corresponden al proceso de tabulación de la pregunta 6 de la encuesta dirigida a personas contagiadas de covid-19 de la Zona 8 de la provincia Guayas (Guayaquil, Durán y Samborondón). Autores (2024).

**Análisis:** Los resultados presentes en este grafico muestran un gran porcentaje que corresponde al 40.29% de personas que estipularon que la intensidad de los

síntomas que tenían era Medio, el 34.30% corresponde a las personas que estipularon que la intensidad de los síntomas que tenían era Leve, el 14.83% corresponde a las personas que estipularon que la intensidad de los síntomas que tenían era Fuerte, el 10.46% corresponde a las personas que estipularon que no tenían síntomas y el 0.12% corresponde al total de celdas vacías.

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

**Tabla 5:**

*Lugar o evento considera que se contagió*

| Opciones de respuesta     | Frecuencia Absoluta | Frecuencia Relativa |
|---------------------------|---------------------|---------------------|
| Casa                      | 1032                | 24.93%              |
| Centro de Atención Medica | 157                 | 3.79%               |
| Fiesta                    | 299                 | 7.22%               |
| Medio de transporte       | 890                 | 21.50%              |
| Otro                      | 650                 | 15.7%               |
| Trabajo                   | 930                 | 22.46%              |
| Vacíos                    | 182                 | 4.40%               |
| <b>TOTAL</b>              | <b>4140</b>         | <b>100,00%</b>      |

Nota: En esta tabla se muestran los valores absolutos y relativos que corresponden al proceso de tabulación de la pregunta 7 de la encuesta dirigida a personas contagiadas de covid-19 de la Zona 8 de la provincia Guayas (Guayaquil, Durán y Samborondón). Autores (2024).

Análisis: Los resultados presentes en este grafico muestran un gran porcentaje que corresponde al 24.93% de personas que se contagieron en casa, el 22.46% corresponde a la cantidad de personas que se contagieron en el trabajo, el 21.50% corresponde a la cantidad de personas que se contagieron en el medio de transporte, el 15.70% corresponde a la cantidad de personas que respondieron "Otro", el 7.22% corresponde a la cantidad de personas que se contagieron en una fiesta, el 4.40% corresponde a la cantidad de personas que se contagieron a celdas vacías y el 3.79% corresponde a la cantidad de personas que se contagieron en un Centro de atención Medica.

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

**Tabla 6:**

*En caso de haber estado vacunado al momento de contagiarse, ¿Cuántas dosis tenía aplicadas al contagiarse?*

| Opciones de respuesta | Frecuencia Absoluta | Frecuencia Relativa |
|-----------------------|---------------------|---------------------|
| 0 dosis               | 3245                | 78.38%              |
| 1 dosis               | 326                 | 7.87%               |
| 2 dosis               | 359                 | 8.67%               |
| 3 dosis               | 16                  | 0.39%               |
| 4 dosis               | 15                  | 0.36%               |
| Vacías                | 179                 | 4.32                |
| <b>TOTAL</b>          | <b>4140</b>         | <b>100,00%</b>      |

Nota: En esta tabla se muestran los valores absolutos y relativos que corresponden al proceso de tabulación de la pregunta 8 de la encuesta dirigida a personas contagiadas de covid-19 de la Zona 8 de la provincia Guayas (Guayaquil, Durán y Samborondón). Autores (2024).

Análisis: Los resultados presentes en este grafico muestran un gran porcentaje que corresponde al 78.38% de personas que no se aplicaron ninguna dosis, el 8.67% corresponde a las personas que se han aplicado 2 dosis, el 7.87% corresponde a personas que se han aplicado 1 dosis, el 4.32% corresponde a celdas vacías, el 0.39% corresponde a personas que se han aplicado 3 dosis y el 0.36% corresponde a personas que se han aplicado 4 dosis.

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

**Tabla 7:**

*En caso de haber estado vacunado al momento de contagiarse ¿Qué vacuna recibió?*

| Opciones de respuesta          | Frecuencia Absoluta | Frecuencia Relativa |
|--------------------------------|---------------------|---------------------|
| Astrazeneca                    | 152                 | 3.67%               |
| No tenía ninguna vacuna puesta | 2723                | 65.77%              |
| Otra                           | 51                  | 1.23%               |
| Pfizer                         | 427                 | 10.31%              |
| Sinovac                        | 602                 | 14.54%              |
| Vacías                         | 185                 | 4.47%               |

| <b>TOTAL</b> | <b>4140</b> | <b>100,00%</b> |
|--------------|-------------|----------------|
|--------------|-------------|----------------|

Nota: En esta tabla se muestran los valores absolutos y relativos que corresponden al proceso de tabulación de la pregunta 9 de la encuesta dirigida a personas contagiadas de covid-19 de la Zona 8 de la provincia Guayas (Guayaquil, Durán y Samborondón). Autores (2024).

Análisis: Los resultados presentes en este grafico muestran un gran porcentaje que corresponde al 65.77% que corresponde a personas que no se aplicaron ningún tipo de vacuna, el 14.54% corresponde a personas que se han aplicado la vacuna Sinovac, el 10.31 corresponde a personas que se han aplicado la vacuna Pfizer, el 4.47% corresponde a celdas vacías, el 3.67% corresponde a personas que se han aplicado la vacuna Astrazeneca y el 1.23% corresponde a personas que se han aplicado otra vacuna.

## 4.5. Desarrollo de la investigación

### 4.5.1. Fase 1: Recolección de data

#### Determinación del público encuestado

La población encuestada corresponde a personas que tuvieron Covid-19 y que además habitan en la zona 8 del Guayas, correspondiente a los cantones de Guayaquil, Durán y Samborondón.

#### Recolección a través de encuestas y redes sociales

En el presente trabajo de titulación se procede a la realización de una encuesta que permita recopilar información referente a los síntomas y recomendaciones de las personas que fueron contagiadas por el virus Covid-19, utilizando la herramienta Google Form de uso gratuito para la realización de la encuesta, estableciendo al dataset inicial con un total de 5570 encuestas, además se extrajo información de redes sociales tales como Twitter y Facebook, resultando en 190 registros válidos.

**Figura 38:**

*Encuesta en Google Form*

Sección 1 de 3

Encuesta para personas contagiadas de COVID-19

La universidad de Guayaquil cuenta con un grupo de investigadores y expertos para el desarrollo de una creación de un algoritmo de procesamiento de lenguaje natural en conversaciones textuales de personas contagiadas covid -19, para el desarrollo de planes futuros que ayude a la comunidad local a combatir el virus.

0. Nombre y Apellido del encuestador (Persona que te ha pedido que llene la encuesta) \*

Texto de respuesta corta

1. Ha tenido coronavirus? \*

Si

No

Después de la sección 1 Ir a la siguiente sección

Sección 2 de 3

No te preocupes!!, no estas solo!!, te tomará

Nota: Formato de encuesta. Autores (2024).

### **Recolección a través de encuestas y web scrapping**

Se obtuvo 190 registros provenientes de las redes sociales Twitter y Facebook, realizando una búsqueda manual de conversaciones en grupos y publicaciones referentes al tema de Covid-19, filtrando aquellas personas que habiten en la zona 8 de la provincia del Guayas.

#### **4.5.2.Fase 2: Limpieza de datos y depuración**

##### **Análisis y proceso de limpieza de la data**

Se realiza un análisis manual de los registros contenidos en el dataset, para establecer solamente los registros de aquellas personas que tuvieron Covid-19, siendo el público objetivo de la encuesta, cuyo tamaño final del dataset es de 4140 registros.

**Figura 39:**

*Dataset recopilado*

|    | E                         | F                     | G         | H                                       | I                             | J                        | K                           | L                   |
|----|---------------------------|-----------------------|-----------|---|-------------------------------|--------------------------|-----------------------------|---------------------|
| 1  | 1. Ha tenido coronavirus? | 2. Seleccione la Edad | 3. Género | 4. ¿Qué variante del Virus lo contagio? | 5. ¿En qué fecha se contagió? | 6. ¿Nivel de intensidad? | 7. ¿En qué lugar o espacio? | 8. ¿En caso de que? |
| 2  | Si                        | Entre 26 a 40 años    | Femenino  | Alfa (Original, covid-19)               | 10/03/2020                    | Leve                     | Trabajo                     |                     |
| 3  | Si                        | Entre 26 a 40 años    | Masculino | Alfa (Original, covid-19)               | 14/10/2020                    | Leve                     | Medio de Transporte         |                     |
| 4  | Si                        | Entre 18 a 25 años    | Masculino | Alfa (Original, covid-19)               | 13/08/2020                    | Leve                     | Trabajo                     |                     |
| 5  | Si                        | Entre 18 a 25 años    | Femenino  | Alfa (Original, covid-19)               | 15/12/2020                    | Leve                     | Casa                        |                     |
| 6  | Si                        | Entre 26 a 40 años    | Femenino  | Alfa (Original, covid-19)               | 08/12/2021                    | Leve                     | Medio de Transporte         |                     |
| 7  | Si                        | Entre 18 a 25 años    | Femenino  | No Sabe                                 | 09/06/2021                    | Leve                     | OTRO                        |                     |
| 8  | Si                        | Entre 18 a 25 años    | Masculino | Alfa (Original, covid-19)               | 15/03/2021                    | Leve                     | Trabajo                     |                     |
| 9  | Si                        | Entre 26 a 64 años    | Femenino  | Alfa (Original, covid-19)               | 01/05/2020                    | Leve                     | Trabajo                     |                     |
| 10 | Si                        | Entre 18 a 25 años    | Femenino  | Alfa (Original, covid-19)               | 25/02/2021                    | Medio                    | Casa                        |                     |
| 11 | Si                        | Entre 26 a 64 años    | Masculino | Alfa (Original, covid-19)               | 05/06/2020                    | Leve                     | OTRO                        |                     |
| 12 | Si                        | Entre 18 a 25 años    | Masculino | Alfa (Original, covid-19)               | 08/07/2020                    | Medio                    | Trabajo                     |                     |
| 13 | Si                        | Entre 18 a 25 años    | Femenino  | Alfa (Original, covid-19)               | 08/12/2021                    | Medio                    | Medio de Transporte         |                     |
| 14 | Si                        | Entre 26 a 64 años    | Femenino  | Alfa (Original, covid-19)               | 12/06/2020                    | Medio                    | Casa                        |                     |
| 15 | Si                        | Entre 18 a 25 años    | Femenino  | Alfa (Original, covid-19)               | 07/04/2021                    | Fuerte                   | Casa                        |                     |
| 16 | Si                        | Entre 18 a 25 años    | Masculino | Alfa (Original, covid-19)               | 05/03/2020                    | Sin Sintomas             | OTRO                        |                     |
| 17 | Si                        | Entre 26 a 40 años    | Masculino | Alfa (Original, covid-19)               | 11/03/2021                    | Leve                     | OTRO                        |                     |

Nota: Dataset de la encuesta para personas contagiadas de Covid-19. Autores (2024).

#### 4.5.3.Fase 3: Formato y visualización de datos

##### Descripción de tablas y atributos del dataset

Para obtener los datos con los cuales se va a trabajar para la posterior aplicación de técnicas y modelo de entrenamiento, se realiza el análisis de los datos más relevantes que sirvan para cumplir con el propósito del trabajo de titulación, se etiqueta los datos de forma que se generan columnas correspondientes a las diversas recomendaciones y síntomas obtenidos de la encuesta, se utilizaron las preguntas 10: “Describa lo más detallado ¿Qué síntomas ha tenido?” y la pregunta 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”.

**Tabla 8:**

*Descripción de los campos del dataset*

| Campo                                   | Descripción  | Tipo  |
|---|--|-------|
| 1. ¿Ha tenido coronavirus?              | Atributos de selección: Identificación si tuvo el virus o no | Texto |
| 2. Seleccione la Edad                   | Identificación de rango de edad                              | Texto |
| 3. Género                               | Identificación de género de la persona                       | Texto |
| 4. ¿Qué variante del Virus lo contagio? | Corresponde a la variante del virus covid-19                 | Texto |
| 5. ¿En qué fecha se contagió?           | Fecha que contrajo el virus                                  | Fecha |
| 6. ¿Nivel de intensidad                 | Corresponde que grado de                                     | Texto |

|  |   |       |
|--|---|-------|
| que tuvo los síntomas?   | intensidad de los síntomas que tuvo una persona al momento de contraer el virus                                 |       |
| 7. ¿En qué lugar o evento considera que se contagió?   | Lugar o establecimiento que una persona contagio el virus   | Texto |
| 8. ¿En caso de haber estado vacunado al momento de contagiarse cuantas dosis tenía aplicadas al contagiarse?                             | Cantidad de dosis que se aplicó la persona  | Texto |
| 9. ¿En caso de haber estado vacunado al momento de contagiarse Qué vacuna recibió?   | Tipo de vacuna que recibió la persona   | Texto |
| 10. Describa lo más detallado ¿Qué síntomas ha tenido?   | Descripción de síntomas que ha tenido la persona con el virus   | Texto |
| 11. Describa ¿Qué medicamentos considera que le ayudo en su recuperación?  | Descripción de medicamentos que ayudaron en la recuperación y control del virus                                 | Texto |
| 12. Describa ¿Qué cuidados aplico durante el proceso de recuperación del Covid, y cuánto tiempo en días tomo su recuperación?            | Descripción de los cuidados que opto la persona contagiada y el tiempo de recuperación                          | Texto |
| 13. Describa a más detalle ¿Qué alimentos y/o vitaminas considera que le ayudaron a fortalecerse y superar el Covid?                     | Descripción de los alimentos y/o vitaminas que ayudan en el fortalecimiento y control contra el virus           | 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 sobre el estado emocional de la persona   | Texto |
| 15. Describa ¿Qué Recomendaciones saludables le dio a conocer su médico?   | Descripción de las recomendaciones saludables que ayudaron en la rehabilitación de persona que padecía el virus | Texto |

|  |   |       |
|--|---|-------|
| 16. Finalmente,<br>describa ¿Qué<br>información le gustaría<br>que esté disponible<br>fácilmente para ser<br>consultada por usted<br>(con respecto al<br>COVID19)? | Opinión al encuestado referente a<br>disponibilidad de información<br>sobre el virus Covid-19 | Texto |
| Lugar que reside de la<br>zona 8   | Lugar donde proviene la persona<br>encuestada   | Texto |

*Nota: Descripción de los campos del dataset. Autores (2024)*

### **Etiquetas de síntomas**

Con la etiqueta de la data del dataset de síntomas, se obtuvieron 76 etiquetas relacionadas a síntomas, tales como: Afectación psicológica, Alergia, Alucinación, Amigdalitis, Ansiedad, Apatía, Arritmia, Asintomatismo, Bronquitis, Cambio coloración de piel, Cansancio, Colitis, Congestión nasal, Conjuntivitis, Convulsión, Debilidad, Desmayo, Diarrea, Dificultad al hablar, Dificultad para moverse, Dificultad para Respirar, Disfonía, Disgeusia, Dolor articular, Dolor de cabeza, Dolor de encías, Dolor de espalda, Dolor de estómago, Dolor de garganta, Dolor de huesos, Dolor en extremidades, Dolor muscular, Dolor ocular, Dolor renal, Dolor torácico Erucción de piel, Escalofrio, Espasmo muscular, Estreñimiento, Falta de oxígeno, Fátiga, Fiebre, Gripe, Hematemesis, Hipotensión, Infección urinaria, Insomnio, Lagrimeo, Lupus, Malestar general, Manchas rojas, Mareo, Náusea, Neumonía, Otalgia, Parálisis corporal, Parálisis facial, Pérdida de apetito, Pérdida de cabello, Pérdida de Gusto, Pérdida de memoria, Pérdida de Olfato, Picazón de garganta, Presión elevada, Presión en el pecho, Problema circulatorio, Prurito, Resfriado, Retención de líquidos, Sarpullido, Saturación baja al respirar, Sudoración excesiva, Taquicardia, Tos, Visión borrosa, Vómito.

### **Etiquetas de recomendaciones**

Con el etiquetado de la data del dataset de recomendaciones, se obtuvieron 26 etiquetas relacionadas a recomendaciones, tales como: Ejercicios, Reposo, Vitaminas, Terapia respiratoria, Hidratado, Alimentación, Mascarilla, Ejercicios respiratorios, No esfuerzo físico, Desestresarme, Solearme, Gárgaras, Aislarme, Aseo, Infusión, Nebulizaciones, Paracetamol, Eucalipto, Antibiótico,

Vaporizaciones, Legumbres, Distanciamiento, Control médico, No visite al médico, No automedicarse, Ninguna.

#### 4.5.4. Fase 4: Preprocesamiento

##### Preprocesamiento de los datos

Se utiliza el ambiente de programación otorgado por Google conocido como Google Colaboratory, se determinaron las columnas correspondientes a la pregunta 10: “Describa lo más detallado ¿Qué síntomas ha tenido?” y pregunta 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” como objetivos para el preprocesamiento.

Los registros contenidos en el dataset contienen campos no relevantes o que directamente no contienen información, por lo que se realiza la limpieza de valores perdidos y nulos en los campos referidos a la pregunta 10: “Describa lo más detallado ¿Qué síntomas ha tenido?” finalizando con 4104 registros válidos para la actual pregunta y para la pregunta 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” finalizando con 3955 registros válidos para dicha pregunta.

**Figura 40:**

Número de registros del dataset de síntomas en el preprocesamiento de datos

|      |      |         |            |   |    |                       |   |                                |            |      |     |   |   |   |   |
|------|------|---------|------------|---|----|-----------------------|---|--------------------------------|------------|------|-----|---|---|---|---|
| 4135 | 4136 | tesis 5 | 23/12/2020 | Isaias Elieco<br>Villalob<br>Tipantasig | si | Entre 25 a<br>35 años | F | Gama((Variante<br>covid - 19)) | 25/07/2021 | éve  | ... | 0 | 0 | 0 | 0 |
| 4136 | 4137 | tesis 5 | 23/12/2020 | Isaias Elieco<br>Villalob<br>Tipantasig | si | Entre 25 a<br>35 años | F | Gama((Variante<br>covid - 19)) | 23/07/2021 | LEVE | ... | 0 | 0 | 0 | 0 |
| 4137 | 4138 | tesis 5 | 23/12/2020 | Isaias Elieco<br>Villalob<br>Tipantasig | si | Entre 20 a<br>30 años | F | Gama((Variante<br>covid - 19)) | 24/07/2021 | LEVE | ... | 0 | 0 | 0 | 0 |
| 4138 | 4139 | tesis 5 | 23/12/2020 | Isaias Elieco<br>Villalob<br>Tipantasig | si | Entre 25 a<br>35 años | F | Gama((Variante<br>covid - 19)) | 20/07/2021 | leve | ... | 0 | 0 | 0 | 0 |
| 4139 | 4140 | tesis 5 | 23/12/2020 | Isaias Elieco<br>Villalob<br>Tipantasig | si | Entre 45 a<br>55 años | F | Gama((Variante<br>covid - 19)) | 20/03/2021 | leve | ... | 0 | 0 | 0 | 0 |

4104 rows × 91 columns

Nota: Representación de los registros del dataset de Síntomas para su posterior preprocesamiento.Código establecido en Google Colab. Autores (2024)

**Figura 41:**

*Número de registros del dataset de recomendaciones en el preprocesamiento de datos*

|      |      |         |                  |  |    |                       |           |                                 |            |        |     |   |   |   |   |   |
|------|------|---------|------------------|--|----|-----------------------|-----------|---------------------------------|------------|--------|-----|---|---|---|---|---|
|      |      |         |                  |  |    |                       |           |                                 |            |        |     |   |   |   |   |   |
| 3993 | 3994 | tesis 5 | 12/20/21<br>1:07 | Priogote<br>Pérez<br>Edzzar<br>Alberto | Si | Entre 26 a<br>64 años | Masculino | alfa<br>(original,<br>covid-19) | NaN        | medio  | ... | 1 | 0 | 1 | 0 | 0 |
| 4036 | 4037 | tesis 5 | 12/18/21<br>7:06 | Rivera<br>Carrón<br>Rafael<br>Ignacio  | si | NaN                   | femenino  | alfa<br>(original,<br>covid-19) | 11/15/2020 | fuerte | ... | 1 | 0 | 0 | 0 | 0 |
| 4037 | 4038 | tesis 5 | 12/18/21<br>7:12 | Rivera<br>Carrón<br>Rafael<br>Ignacia  | si | NaN                   | femenino  | alfa<br>(original,<br>covid-19) | 11/7/2020  | fuerte | ... | 0 | 0 | 0 | 0 | 0 |
| 4039 | 4040 | tesis 5 | 12/18/21<br>7:50 | Rivera<br>Carrón<br>Rafael<br>Ignacio  | si | NaN                   | femenino  | alfa<br>(original,<br>covid-19) | 6/18/2021  | medio  | ... | 0 | 0 | 0 | 0 | 0 |
| 4042 | 4043 | tesis 5 | 12/18/21<br>8:08 | Rivera<br>Carrón<br>Rafael<br>Ignacio  | si | NaN                   | masculino | alfa<br>(original,<br>covid-19) | 7/18/2020  | medio  | ... | 0 | 0 | 0 | 0 | 0 |

Nota: Representación de los registros del dataset de Recomendaciones para su posterior preprocesamiento. Código establecido en Google Colab. Autores (2024)

## Preprocesamiento del texto

Se utilizó mediante código de programación la aplicación de mecanismos que permiten depurar el texto, consiste en la eliminación de caracteres especiales, signos de puntuación, exclamación e interrogación, establecer en minúscula al texto, cuyo fin es la de tener el texto de la data lo más limpia posible de información no relevante para el futuro entrenamiento.

**Figura 42:**

*Función que aplica el pre-procesamiento*

*Nota:* Función escrita en lenguaje Python que ayuda en el preprocesamiento de los datos en el dataset. Autores (2024). Fuente: código establecido en Google Colab.

Al aplicar las técnicas de preprocesamiento, el tamaño de los dataset a utilizar para la posterior aplicación de las técnicas de Procesamiento de lenguaje natural consiste en 4104 registros para la columna correspondiente a la pregunta #10 y 3955 registros para la columna correspondiente a la pregunta #15 como se visualizan en la figura 5 y figura 6.

**Figura 43:**

Pregunta 10 depurada del dataset

| 10. Describa lo más detallado ¿Qué síntomas ha tenido?-DEPURADO |   |
|---|---|
| 0   | perdida de gusto y olfato                         |
| 1   | dolor leve de espalda y fiebre                    |
| 2   | dolor al momento de respirar por algunos días     |
| 3   | dolor de cabeza tos seca y mucha fiebre no ten... |
| 4   | sin sabor   |
| ...   | ...   |
| 4135  | como una alergia con estornudos constantemente... |
| 4136  | dolor de cabeza tos dolor de espalda en fin de... |
| 4137  | con dolor de garganta se me quito                 |
| 4138  | siente muy apachurrado su pecho como flemas mu... |
| 4139  | neumonia bilateral cocida ya soy asmatica sigo... |
| 4104 rows × 1 columns   |   |

Nota: Visualización de la columna “10. Describa lo más detallado ¿Qué síntomas ha tenido?” en donde se aprecia la versión depurada y sin depurar. Autores (2024). Fuente: Código establecido en Google Colab.

**Figura 44:**

Pregunta 15 depurada del dataset

| 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 - DEPURADA |   |
|--|---|
| 0  | dormir boca abajo                                 |
| 1  | me dijo que haga terapias de respiración cuand... |
| 2  | ninguna   |
| 3  | me dijo que aguntara la respiración y soltara ... |
| 4  | si  |
| ...  | ...   |
| 3993   | muchos ejercicios respiratorios y medicinas       |
| 4036   | recomendaciones para el dolor de cabeza           |
| 4037   | alguien de verdad me puede decir que esto term... |
| 4039   | como se han recuperado                            |
| 4042   | tener paciencia comer bien hacer ejercicio baj... |
| 3955 rows × 1 columns  |   |

Nota: Visualización de la columna “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” en donde se aprecia la versión depurada y sin depurar. Autores (2024). Fuente: Código establecido en Google Colab.

#### 4.5.5. Fase 5: Etiquetación

##### Definición de las entradas y salidas

De las 76 etiquetas iniciales correspondientes al dataset de síntomas, se consideraron aquellas etiquetas que contengan más de 200 registros válidos,

resultando en 13 etiquetas finales a utilizar para el entrenamiento del modelo referente al dataset de síntomas, para las 26 etiquetas correspondientes al dataset de recomendaciones, se consideraron aquellas etiquetas que posean alrededor de 200 o más registros válidos, como resultante, dando 10 etiquetas finales válidas para el entrenamiento del modelo del dataset de recomendaciones. Posteriormente, se mantuvieron aquellos registros de los datasets que posean al menos un síntoma y una recomendación.

Las etiquetas establecidas para síntomas son las siguientes: Asintomatismo, Cansancio, Dificultad para Respirar, Dolor de cabeza, Dolor de garganta, Dolor muscular, Escalofrío, Fiebre, Gripe, Malestar general, Pérdida de Gusto, Pérdida de Olfato, Tos.

Las etiquetas establecidas para recomendaciones son las siguientes: Ejercicios, Reposo, Vitaminas, Hidratado, Alimentación, Ejercicios respiratorios, Aislarme, Aseo, No visite al médico, Ninguna.

**Figura 45:**

*Conteo de la cantidad de síntomas por registro*

| Asintomatismo | Cansancio | Dificultad para Respirar | Dolor de cabeza | Dolor de garganta | Dolor muscular | Escalofrío | Fiebre | Gripe | Malestar general | Pérdida de Gusto | Pérdida de Olfato | Tos | total |
|---------------|-----------|--------------------------|-----------------|-------------------|----------------|------------|--------|-------|------------------|------------------|-------------------|-----|-------|
| 0             | 0         | 0                        | 0               | 0                 | 0              | 0          | 0      | 0     | 0                | 1                | 1                 | 0   | 2     |
| 1             | 0         | 0                        | 0               | 0                 | 0              | 0          | 0      | 1     | 0                | 0                | 0                 | 0   | 1     |
| 2             | 0         | 0                        | 0               | 0                 | 0              | 0          | 0      | 0     | 0                | 0                | 0                 | 0   | 0     |
| 3             | 0         | 0                        | 0               | 0                 | 0              | 0          | 0      | 1     | 0                | 0                | 1                 | 1   | 1     |
| 4             | 0         | 0                        | 0               | 0                 | 0              | 0          | 0      | 0     | 0                | 0                | 1                 | 0   | 0     |
| ...           | ...       | ...                      | ...             | ...               | ...            | ...        | ...    | ...   | ...              | ...              | ...               | ... | ...   |
| 4135          | 0         | 0                        | 0               | 0                 | 0              | 0          | 0      | 1     | 0                | 0                | 1                 | 1   | 0     |
| 4136          | 0         | 0                        | 0               | 1                 | 0              | 0          | 0      | 0     | 0                | 0                | 0                 | 0   | 2     |
| 4137          | 0         | 0                        | 0               | 0                 | 1              | 0          | 0      | 0     | 0                | 0                | 0                 | 0   | 1     |
| 4138          | 0         | 1                        | 1               | 0                 | 0              | 0          | 0      | 1     | 0                | 0                | 0                 | 0   | 3     |
| 4139          | 0         | 1                        | 1               | 0                 | 0              | 0          | 0      | 0     | 0                | 0                | 0                 | 0   | 2     |

4104 rows × 14 columns

Nota: Visualización de la cantidad de Síntomas están presente por registro. Autores (2024). Código establecido en Google Colab.

### Figura 46:

*Conteo de la cantidad de recomendaciones por registro*

|      | ejercicios | Reposo | Vitaminas | Hidratado | alimentacion | ejercicios respiratorios | aislarame | aseo | No visite al médico | Ninguna | total |
|------|------------|--------|-----------|-----------|--------------|--------------------------|-----------|------|---------------------|---------|-------|
| 0    | 0          | 1.0    | 0         | 0         | 0            |                          | 0         | 0    | 0                   | 0       | 0     |
| 1    | 0          | 0.0    | 0         | 0         | 0            |                          | 0         | 0    | 0                   | 0       | 0     |
| 2    | 0          | 0.0    | 0         | 0         | 0            |                          | 0         | 0    | 0                   | 0       | 0     |
| 3    | 0          | 0.0    | 0         | 0         | 0            |                          | 1         | 0    | 0                   | 0       | 1.0   |
| 4    | 0          | 0.0    | 0         | 0         | 0            |                          | 0         | 0    | 0                   | 0       | 0     |
| ...  | ...        | ...    | ...       | ...       | ...          |                          | ...       | ...  | ...                 | ...     | ...   |
| 3993 | 0          | 0.0    | 0         | 0         | 0            |                          | 1         | 0    | 0                   | 0       | 1.0   |
| 4036 | 0          | 0.0    | 0         | 0         | 0            |                          | 0         | 0    | 0                   | 0       | 0     |
| 4037 | 0          | 0.0    | 0         | 0         | 0            |                          | 0         | 0    | 0                   | 0       | 0     |
| 4039 | 0          | 0.0    | 0         | 0         | 0            |                          | 0         | 0    | 0                   | 0       | 0     |
| 4042 | 1          | 1.0    | 0         | 0         | 1            |                          | 0         | 0    | 0                   | 0       | 3.0   |

3955 rows x 11 columns

*Nota:* Visualización de la cantidad de Recomendaciones están presente por registro. Autores (2024). Código establecido en Google Colab.

El dataset correspondiente a síntomas, pre-aplicación de técnicas de procesamiento de lenguaje natural, resultando en 3925 registros para el dataset de síntomas y 3509 registros para el dataset de recomendaciones.

### Figura 47:

*Cantidad de registros del dataset de síntomas pre-aplicación de técnicas de NLP*

```
#DIMENSIONES DE LA PREGUNTA 10
y.shape
(3925, 13)
```

*Nota:* Se presenta la dimensión del dataset de Síntomas. Autores (2024). Código establecido en Google Colab.

### Figura 48:

*Cantidad de registros del dataset de recomendaciones pre-aplicación de técnicas de NLP*

```
[104] #DIMENSIONES DE LA PREGUNTA 15
X.shape
(3509, 1)

[105] #DIMENSIONES DE RECOMENDACIONES
y.shape
(3509, 10)
```

*Nota:* Se presenta la dimensión del dataset de Recomendaciones. Autores (2024). Código establecido en Google Colab.

#### 4.5.6. Fase 6: Entrenamiento

##### Aplicación de técnicas de NLP

Para el desarrollo de los algoritmos que hacen uso de las técnicas de Procesamiento del Lenguaje Natural, se hizo uso de las librerías NLTK y Spacy, las técnicas empleadas son tokenización, eliminación de stopwords, lematización y PoS tagging.

##### Tokenización

En la tokenización correspondientes a las columnas de la pregunta 10 del dataset de síntomas y de la pregunta 15 del dataset de recomendaciones, se utilizó el siguiente algoritmo:

##### Algoritmo para síntomas

```
x_temp['Tokenizado']=x_temp.apply(lambda row: nltk.word_tokenize(row['10. Describa lo más detallado ¿Qué síntomas ha tenido?-DEPURADO'], "spanish"), axis=1)
```

##### Algoritmo para recomendaciones

```
x_temp['Tokenizado']=x_temp.apply(lambda row: nltk.word_tokenize(row['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 - DEPURADA'], "spanish"), axis=1)
```

##### Figura 49:

##### Tokenizado de la pregunta 10 del dataset

| 10. Describa lo más detallado ¿Qué síntomas ha tenido?-DEPURADO |   | Tokenizado  |
|---|---|---|
| 0   | perdida de gusto y olfato                         | [perdida, de, gusto, y, olfato]                     |
| 1   | dolor leve de espalda y fiebre                    | [dolor, leve, de, espalda, y, fiebre]               |
| 3   | dolor de cabeza los seca y mucha fiebre no ten... | [dolor, de, cabeza, los, seca, y, mucha, fiebr...]  |
| 4   | sin sabor   | [sin, sabor]  |
| 5   | solamente no tenía sabor                          | [solamente, no, tenía, sabor]                       |
| ...   | ...   | ...   |
| 4135  | como una alergia con estornudos constantemente... | [como, una, alergia, con, estornudos, constant...]  |
| 4136  | dolor de cabeza los dolor de espalda en fin de... | [dolor, de, cabeza, los, dolor, de, espalda, e...]  |
| 4137  | con dolor de garganta se me quito                 | [con, dolor, de, garganta, se, me, quito]           |
| 4138  | siente muy apachurrido su pecho como flemas mu... | [siente, muy, apachurrido, su, pecho, como, f...]   |
| 4139  | neumonia bilateral cocida ya soy asmatica sigo... | [neumonia, bilateral, cocida, ya, soy, asmatica...] |

3925 rows × 2 columns

Nota: Aplicación de la técnica de tokenización en la pregunta 10 del dataset. Autores (2024). Código establecido en Google Colab.

## Figura 50:

*Tokenizado de la pregunta 15 del dataset*

| x_temp                |  | Tokenizado   |
|-----------------------|--|--|
| 0                     | 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 - DEPURADA | [dormir, boca, abajo]                                  |
| 3                     | me dijo que aguntara la respiración y soltara ...  | [me, dijo, que, aguntara, la, respiracion, y, ...]     |
| 6                     | me recomiendo que me cuidara que me podria dar ...   | [me, recomiendo, que, me, cuidara, que, me, pod, ...]  |
| 7                     | evitar lugares con aglomeraciones distanciamie...  | [evitar, lugares, con, aglomeraciones, distanc, ...]   |
| 8                     | no me atendi con un medico   | [no, me, atendi, con, un, medico]                      |
| ...                   |  | ...  |
| 3948                  | que tenga mucho cuidado  | [que, tenga, mucho, cuidado]                           |
| 3949                  | ejercicio fisico trotar  | [ejercicio, fisico, trotar]                            |
| 3952                  | me recomendaron comer toronja asada  | [me, recomendaron, comer, toronja, asada]              |
| 3993                  | muchos ejercicios respiratorios y medicinas  | [muchos, ejercicios, respiratorios, y, medicinas]      |
| 4042                  | tener paciencia comer bien hacer ejercicio baj...  | [tener, paciencia, comer, bien, hacer, ejercicio, ...] |
| 3509 rows × 2 columns |  |  |

Nota: Aplicación de la técnica de tokenización en la pregunta 15 del dataset. Autores (2024) Código establecido en Google Colab.

## Stopwords

Para la aplicación de la técnica Stopwords, se importa un corpus en español que contiene todas las palabras de paradas a utilizar para la limpieza de los datos.

```
from nltk.corpus import stopwords
```

```
stop_words = set(stopwords.words('spanish'))
```

## Figura 51:

*Visualización de las stopwords del corpus en español*

```
#CREAR LISTA DE STOPWORDS EN ESPAÑOL
from nltk.corpus import stopwords
stop_words = set(stopwords.words('spanish'))
```

```
[ ] stop_words
```

```
{'a',
 'al',
 'algo',
 'algunas',
 'algunos',
 'ante',
 'antes',
 'como',
 'con',
 'contra',
 'cuál',
 'cuando',
 'de',
 'del',
 'desde',
 'donde',
 'durante',
 'e',
 'el',
 'ella',
 'ellas',
 'ellos',
 'en',
 'entre',
 'era',
 'erais',
 'eran',
 'eras'}
```

Nota: Visualización del corpus adquirido mediante la librería Nltk. Autores (2024). Código establecido en Google Colab.

Para limpiar el contenido de stopwords se aplica el siguiente algoritmo a la columna previamente aplicada con tokenización.

```
x_temp['SinStopwords']=x_temp['Tokenizado'].apply(lambda z: ' '.join([word for word in z if word not in (stop_words)]))
```

**Figura 52:**

*Eliminación de stopwords en la pregunta 10 del dataset de síntomas*

| [ ] #LIMPIANDO EL CONTENIDO DE x_temp DE stopwords<br>x_temp['SinStopwords']=x_temp['Tokenizado'].apply(lambda z: ' '.join([word for word in z if word not in (stop_words)]))<br>x_temp |   |  |   |
|---|---|--|---|
|   | 10. Describa lo más detallado ¿Qué síntomas ha tenido?-DEPURADO | Tokenizado   | SinStopwords                                      |
| 0   | perdida de gusto y olfato                                       | [perdida, de, gusto, y, olfato]                    | perdida gusto olfato                              |
| 1   | dolor leve de espalda y fiebre                                  | [dolor, leve, de, espalda, y, fiebre]              | dolor leve espalda fiebre                         |
| 3   | dolor de cabeza los seca y mucha fiebre no ten...               | [dolor, de, cabeza, los, seca, y, mucha, fiebr...] | dolor cabeza los seca mucha fiebre tenia gusto... |
| 4   | sin sabor   | [sin, sabor]                                       | sabor   |
| 5   | solo no tenia sabor   | [solo, no, tenia, sabor]                           | solo tenia sabor                                  |
| ...   | ...   | ...  | ...   |
| 4135  | como una alergia con estornudos constantemente...               | [como, una, alergia, con, estornudos, constant...] | alergia estornudos constantemente luego perdi ... |
| 4136  | dolor de cabeza los dolor de espalda en fin de...               | [dolor, de, cabeza, los, dolor, de, espalda, e...] | dolor cabeza los dolor espalda fin cintura hac... |
| 4137  | con dolor de garganta se me quito                               | [con, dolor, de, garganta, se, me, quito]          | dolor garganta quito                              |
| 4138  | siente muy apachurrado su pecho como flemas mu...               | [siente, muy, apachurrado, su, pecho, como, fl...] | apachurrado pecho flemas cansada agitada          |
| 4139  | neumonia bilateral cocida ya soy asmatica sigo...               | [neumonia, bilateral, cocida, ya, soy, asmatic...] | neumonia bilateral cocida asmatica sigo simpto... |
| 3925 rows x 3 columns   |   |  |   |

Nota: Aplicación de la técnica de Eliminación de stopwords en la pregunta 10 del dataset. Autores (2024). Código establecido en Google Colab.

**Figura 53:**

*Eliminación de stopwords para la pregunta 15 del dataset de recomendaciones*

| [ ] x_temp            |  |  |  |
|-----------------------|--|--|--|
|                       | 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 - DEPURADA | Tokenizado   | SinStopwords   |
| 0                     | me dijeron que dormir boca abajo   | [dormir, boca, abajo]                                  | dormir boca abajo                                    |
| 3                     | me dijeron que aguantara la respiración y soltar...  | [me, dijeron, que, aguantara, la, respiracion, y, ...] | dijo aguantara respiracion soltar lentamente         |
| 6                     | me recomendaron que me cuidara que me podria dar...  | [me, recomendaron, que, me, cuidara, que, me, pod...]  | recomendaron cuidara podria dar nuevamente tomara... |
| 7                     | evitar lugares con aglomeraciones distanciamie...  | [evitar, lugares, con, aglomeraciones, distanci...]    | evitar lugares aglomeraciones distanciamiento ...    |
| 8                     | no me atendí con un medico   | [no, me, atendi, con, un, medico]                      | atendi medico  |
| ...                   | ...  | ...  | ...  |
| 3948                  | que tenga mucho cuidado  | [que, tenga, mucho, cuidado]                           | cuidado  |
| 3949                  | ejercicio fisico trotar  | [ejercicio, fisico, trotar]                            | ejercicio fisico trotar                              |
| 3952                  | me recomendaron comer toronja asada  | [me, recomendaron, comer, toronja, asada]              | recomendaron comer toronja asada                     |
| 3993                  | muchos ejercicios respiratorios y medicinas  | [muchos, ejercicios, respiratorios, y, medicinas]      | ejercicios respiratorios medicinas                   |
| 4042                  | tener paciencia comer bien hacer ejercicio baj...  | [tener, paciencia, comer, bien, hacer, ejercicio...]   | tener paciencia comer bien hacer ejercicio baj...    |
| 3599 rows x 3 columns |  |  |  |

Nota: Aplicación de la técnica de Eliminación de stopwords en la pregunta 15 del dataset. Autores (2024). Código establecido en Google Colab.

## Lematización

Usando la librería Spacy, se importó un corpus que contiene el modelo de palabras en español para aplicar la lematización, usando el siguiente algoritmo:

```
!python -m spacy download es_core_news_md
x_temp['Lemmatizacion']=x_temp['SinStopwords'].apply(lambda x: ' '.join([(word.lemma_) for word in nlp(str(x))]))
```

**Figura 54:**

*Aplicación de la técnica lematización en la pregunta 10 de síntomas*

| 10. Describa lo más detallado ¿Qué síntomas ha tenido? - DEPURADO |  | Tokenizado   | SinStopwords                                       | Lematización   |
|---|--|--|--|--|
| 0   | perdida de gusto y olfato                          | [perdida, de, gusto, y, olfato]                      | perdida gusto olfato                               | perdido gusto olfacto                                |
| 1   | dolor leve de espalda y fiebre                     | [dolor, leve, de, espalda, y, fiebre]                | dolor leve espalda fiebre                          | dolor leve espalda fiebre                            |
| 3   | dolor de cabeza los seca y mucha fiebre no ten...  | [dolor, de, cabeza, los, seca, y, mucha, fiebre...]  | dolor cabeza los seca mucha fiebre tenia gusto...  | dolor cabeza los seco mucho fiebre tenia gusto.      |
| 4   | sin sabor  | [sin, sabor]   | sabor  | sabor  |
| 5   | soló no tenía sabor                                | [soló, no, tenia, sabor]                             | soló tenia sabor                                   | soló tenía sabor                                     |
| ...   | ...  | ...  | ...  | ...  |
| 4135  | como una alergia con estornudos constantemente...  | [como, una, alergia, con, estornudos, constante...]  | alergia estornudos constantemente luego perdi...   | alergiar estornudo constantemente luego perder...    |
| 4136  | dolor de cabeza los dolor de espalda en fin de...  | [dolor, de, cabeza, los, dolor, de, espalda, e...]   | dolor cabeza los dolor espalda fin cintura hac...  | dolor cabeza a dolor espalda fin cintura hacer...    |
| 4137  | con dolor de garganta se me quito                  | [con, dolor, de, garganta, se, me, quito]            | dolor garganta quito                               | dolor garganta quitar                                |
| 4138  | siente muy apachurrado su pecho como flemas mu...  | [siente, muy, apachurrado, su, pecho, como, r...]    | apachurrado pecho flemas cansada agitada           | apachurrando pecho flemas cansado agitado            |
| 4139  | neumonia bilateral coicida ya soy asmatica sigo... | [neumonia, bilateral, coicida, ya, soy, asmatico...] | neumonia bilateral coicida asmatica sigo simpto... | neumonia bilateral coicido asmatico seguir simple... |

Nota: Aplicación de la técnica de Lematización en la pregunta 10 del dataset. Autores (2024). Código establecido en Google Colab.

**Figura 55:**

*Aplicación de la técnica lematización en la pregunta 15 de recomendaciones*

| 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 - DEPURADA |  | Tokenizado   | SinStopwords                                      | Lematización                                       |
|--|--|--|---|--|
| 0  | dormir boca abajo                                    | [dormir, boca, abajo]                                | dormir boca abajo                                 | dormir boca abajo                                  |
| 3  | me dijo que aguantara la respiración y soltarla ...  | [me, dijo, que, aguantara, la, respiracion, y, ...]  | dijo aguantara respiracion soltarla lentamente    | dejar aguantar respiracion soltar lentamente       |
| 6  | me recomiendo que me cuidara que me podria dar ...   | [me, recomiendo, que, me, cuidara, que, me, pod...]  | recomiendo cuidara podria dar nuevamente tomar... | recomiendo cuidar podria dar nuevamente tomar...   |
| 7  | evitar lugares con aglomeraciones distanciamiento... | [evitar, lugares, con, aglomeraciones, distanci...]  | evitar lugares aglomeraciones distanciamiento...  | evitar lugar aglomeraciones distanciamiento...     |
| 8  | no me atendí con un medico                           | [no, me, atendí, con, un, medico]                    | atendi medico                                     | atender medico                                     |
| ...  | ...  | ...  | ...   | ...  |
| 3948   | que tenga mucho cuidado                              | [que, tenga, mucho, cuidado]                         | cuidado   | cuidado  |
| 3949   | ejercicio fisico trotar                              | [ejercicio, fisico, trotar]                          | ejercicio fisico trotar                           | ejercicio fisico trotar                            |
| 3952   | me recomendaron comer toronja asada                  | [me, recomendaron, comer, toronja, asada]            | recomendaron comer toronja asada                  | recomendar comer toronja asada                     |
| 3993   | muchos ejercicios respiratorios y medicinas          | [muchos, ejercicios, respiratorios, y, medicinas]    | ejercicios respiratorios medicinas                | ejercicio respiratorio medicina                    |
| 4042   | tener paciencia comer bien hacer ejercicio baj...    | [tener, paciencia, comer, bien, hacer, ejercicio...] | tener paciencia comer bien hacer ejercicio baj... | tener paciencia comer bien hacer ejercicio bajo... |

Nota: Aplicación de la técnica de Lematización en la pregunta 15 del dataset. Autores (2024). Código establecido en Google Colab.

### Part-of-Speech Tagging (Pos Tagging)

Para la aplicación de la técnica de Part of speech tagging, se aplicó el siguiente algoritmo en la columna previamente creada de lematización.

```
x_temp['POS'] = x_temp['Lemmatizacion'].apply(lambda x: [(token.text, token.pos_) for token in nlp(str(x))])
```

**Figura 56:**
*POS Tagging aplicado a la pregunta 10 de síntomas*

| 10. Describa lo más detallado ¿Qué síntomas ha tenido? -DEPURADO |   |   |  |  |  |
|--|---|---|--|--|--|
|  |   | Tokenizado  | SinStopwords                             | Lemmatizacion                            | POS  |
| 0  | perdida de gusto y olfato                         | [perdida, de, gusto, y, olfato]                       | perdida gusto olfato                     | perdido gusto olfato                     | [[perdido, ADJ], (gusto, NOUN), (olfato, ADJ)]     |
| 1  | dolor leve de espalda y fiebre                    | [dolor, leve, de, espalda, y, fiebre]                 | dolor leve espalda fiebre                | dolor leve espalda fiebre                | [(dolor, NOUN), (leve, ADJ), (espalda, NOUN), ...] |
| 3  | dolor de cabeza los seca y mucha fiebre no ten... | [dolor, de, cabeza, los, seca, y, mucha, fiebre, ...] | dolor cabeza los seca mucha fiebre       | dolor cabeza los seco mucho fiebre       | [(dolor, NOUN), (cabeza, NOUN), (tos, NOUN), ...]  |
| 4  | sin sabor   | [sin, sabor]  | sabor                                    | sabor                                    | [(sabor, NOUN)]                                    |
| 5  | soló no tenía sabor                               | [soló, no, tenía, sabor]                              | soló tenía sabor                         | soló tenía sabor                         | [[soló, ADJ], (tenía, NOUN), (sabor, NOUN)]        |
| ...  | ...   | ...   | ...                                      | ...                                      | ...  |
| 4135   | como una alergia con estornudos constantemente... | [como, una, alergia, con, estornudos, constan...      | alergia estornudos constantemente        | alergiar estornudo constantemente        | [[alergiar, VERB], (estornudo, ADJ), (constan...   |
| 4136   | dolor de cabeza los dolor de espalda en fin de... | [dolor, de, cabeza, los, dolor, de, espalda, e...     | dolor cabeza los dolor espalda fin       | dolor cabeza to dolor espalda fin        | [(dolor, NOUN), (cabeza, NOUN), (to, ADJ), (do...  |
| 4137   | con dolor de garganta se me quito                 | [con, dolor, de, garganta, se, me, quito]             | dolor garganta quito                     | dolor garganta quito                     | [(dolor, NOUN), (garganta, PROPN), (quito, PRO...  |
| 4138   | siente muy apachurrido su pecho como flemas mu... | [siente, muy, apachurrido, su, pecho, como, fl...     | apachurrido pecho flemas cansado agitado | apachurrado pecho flemas cansado agitado | [[apachurrido, ADJ], (pecho, NOUN), (flemas, N...  |

*Nota: Aplicación de la técnica de POS Tagging en la pregunta 15 del dataset. Autores (2024). Código establecido en Google Colab.*

**Figura 57:**
*POS Tagging aplicado a la pregunta 15 de recomendaciones*

| 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 - DEPURADA |  |  |   |   |   |
|--|--|--|---|---|---|
|  |  | Tokenizado   | SinStopwords                                      | Lemmatizacion                                     | POS   |
| 0  | dormir boca abajo                                  | [dormir, boca, abajo]                              | dormir boca abajo                                 | dormir boca abajo                                 | [(dormir, VERB), (boca, NOUN), (abajo, ADV)]        |
| 3  | me dijo que aguntara la respiración y soltara ...  | [me, dijo, que, aguntara, la, respiración, y, ...] | dijo aguntara respiracion soltar...               | decir aguntar respiracion soltar...               | [(decir, VERB), (aguntar, VERB), (respiracion, ...) |
| 6  | me recomiendo que me cuidara que me podria dar ... | [me, recomiendo, que, me, cuidara, que, me, pod... | recomiendo cuidara podria dar                     | recomiendo cuidar podria dar                      | [(recomiendo, VERB), (cuidar, VERB), (podrio, A...  |
| 7  | evitar lugares con aglomeraciones distanciamie...  | [evitar, lugares, con, aglomeraciones, distanc...  | evitar lugares aglomeraciones                     | evitar lugar aglomeraciones                       | [(evitar, VERB), (lugar, NOUN), (aglomeracione...   |
| 8  | no me atendi con un medico                         | [no, me, atendi, con, un, medico]                  | atendi medico                                     | atender medico                                    | [(atender, VERB), (medico, ADJ)]                    |
| ...  | ...  | ...  | ...   | ...   | ...   |
| 3948   | que tenga mucho cuidado                            | [que, tenga, mucho, cuidado]                       | cuidado   | cuidado   | [(cuidado, NOUN)]                                   |
| 3949   | ejercicio fisico trotar                            | [ejercicio, fisico, trotar]                        | ejercicio fisico trotar                           | ejercicio fisico trotar                           | [(ejercicio, NOUN), (fisico, ADJ), (trotar, VE...]  |
| 3952   | me recomendaron comer toronja asada                | [me, recomendaron, comer, toronja, asada]          | recomendaron comer toronja asada                  | recomendar comer toronja asado                    | [(recomendar, VERB), (comer, VERB), (toronja, ...)  |
| 3993   | muchos ejercicios respiratorios y medicinas        | [muchos, ejercicios, respiratorios, y, medicinas]  | ejercicios respiratorios medicinas                | ejercicio respiratorio medicina                   | [(ejercicio, NOUN), (respiratorio, ADJ), (medi...   |
| 4042   | tener paciencia comer bien hacer ejercicio baj...  | [tener, paciencia, comer, bien, hacer, ejercic...  | tener paciencia comer bien hacer ejercicio baj... | tener paciencia comer bien hacer ejercicio baj... | [(tener, VERB), (pacienza, NOUN), (comer, VER...]   |

*Nota: Aplicación de la técnica de POS Tagging en la pregunta 15 del dataset. Autores (2024). Código establecido en Google Colab.*

Se identificó que de las etiquetas dadas con la técnica de POS Tagging, las palabras con etiquetas AUX de auxiliares, son las que proporcionan menor información relevante al dataset y se procedió con su eliminación mediante el siguiente algoritmo:

for i in x\_temp.index:

```
x_temp['POS'][i] = ''.join([x for (x,y) in x_temp['POS'][i] if y not in ('AUX')])
```

**Figura 58:**
*Columna POS sin palabras auxiliares en dataset de síntomas*

|      | 10. Describa lo más detallado ¿Qué síntomas ha tenido? -DEPURADO | Tokenizado   | SinStopwords                                      | Lemmatización                                     | POS   |
|------|--|--|---|---|---|
| 0    | perdida de gusto y olfato  | [perdida, de, gusto, y, olfato]                    | perdida gusto olfato                              | perdido gusto olfato                              | perdido gusto olfato                              |
| 1    | dolor leve de espalda y fiebre                                   | [dolor, leve, de, espalda, y, fiebre]              | dolor leve espalda fiebre                         | dolor leve espalda fiebre                         | dolor leve espalda fiebre                         |
| 3    | dolor de cabeza los seca y mucha fiebre no ten...                | [dolor, de, cabeza, los, seca, y, mucha, fiebr...] | dolor cabeza los seco mucha fiebre tenia gusto... | dolor cabeza los seco mucho fiebre tenia gusto... | dolor cabeza los seco mucho fiebre tenia gusto... |
| 4    | sin sabor  | [sin, sabor]                                       | sabor   | sabor   | sabor   |
| 5    | solamente tenia sabor  | [solamente, tenia, sabor]                          | solamente tenia sabor                             | solamente tenia sabor                             | solamente tenia sabor                             |
| ...  | ...  | ...  | ...   | ...   | ...   |
| 4135 | como una alergia con estornudos constantemente...                | [como, una, alergia, con, estornudos, constant...] | alergia estornudos constantemente luego perdi...  | alergiar estornudo constantemente luego perder... | alergiar estornudo constantemente luego perder... |
| 4136 | dolor de cabeza los dolor de espalda en fin de...                | [dolor, de, cabeza, los, dolor, de, espalda, e...] | dolor cabeza los dolor espalda fin cintura hac... | dolor cabeza lo dolor espalda fin cintura hac...  | dolor cabeza lo dolor espalda fin cintura hac...  |
| 4137 | con dolor de garganta se me quito                                | [con, dolor, de, garganta, se, me, quito]          | dolor garganta quito                              | dolor garganta quito                              | dolor garganta quito                              |
| 4138 | siente muy apachurrado su pecho como flemas mu...                | [siente, muy, apachurrado, su, pecho, como, fl...] | apachurrado pecho flemas cansado agitado          | apachurrado pecho flemas cansado agitado          | apachurrado pecho flemas cansado agitado          |
| 4139 | neumonia bilateral cocida ya soy asmatica sigo...                | [neumonia, bilateral, cocida, ya, soy, asmatic...] | neumonia bilateral cocida asmatica sigo simp...   | neumonia bilateral cocido asmatico seguir simp... | neumonia bilateral cocido asmatico seguir simp... |

Nota: Visualización después de la eliminación de los AUX presentes en la columna POS correspondiente del dataset de Síntomas. Autores (2024). Código establecido en Google Colab.

**Figura 59:**
*Columna POS sin palabras auxiliares en dataset de recomendaciones*

|      | 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 - DEPURADA | Tokenizado  | SinStopwords                                       | Lemmatizacion                                     | POS   |
|------|--|---|--|---|---|
| 0    | dormir boca abajo  | [dormir, boca, abajo]                               | dormir boca abajo                                  | dormir boca abajo                                 | dormir boca abajo                                 |
| 3    | me dijo que aguantara la respiración y soltara ...   | [me, dijo, que, aguantara, la, respiracion, y, ...] | dijo aguantara respiracion soltara lentamente      | decir aguntar respiracion soltar lentamente       | decir aguntar respiracion soltar lentamente       |
| 6    | me recomiendo que me cuidara que me podria dar ...   | [me, recomiendo, que, me, cuidara, que, me, pod...] | recomendo cuidara podria dar nuevamente tomar v... | recomendo cuidar podrio dar nuevamente tomar v... | recomendo cuidar podrio dar nuevamente tomar v... |
| 7    | evitar lugares con aglomeraciones distanciamie...  | [evitar, lugares, con, aglomeraciones, distanc...]  | evitar lugares aglomeraciones distancia...         | evitar lugar aglomeraciones distancia...          | evitar lugar aglomeraciones distancia...          |
| 8    | no me atendi con un medico   | [no, me, atendi, con, un, medico]                   | atendi medico                                      | atender medico                                    | atender medico                                    |
| ...  | ...  | ...   | ...  | ...   | ...   |
| 3948 | que tenga mucho cuidado  | [que, tenga, mucho, cuidado]                        | cuidado  | cuidado   | cuidado   |
| 3949 | ejercicio fisico trotar  | [ejercicio, fisico, trotar]                         | ejercicio fisico trotar                            | ejercicio fisico trotar                           | ejercicio fisico trotar                           |
| 3952 | me recomendaron comer toronja asada  | [me, recomendaron, comer, toronja, asada]           | recomendaron comer toronja asada                   | recomendar comer toronja asado                    | recomendar comer toronja asado                    |
| 3993 | muchos ejercicios respiratorios y medicinas  | [muchos, ejercicios, respiratorios, y, medicinas]   | ejercicios respiratorios medicinas                 | ejercicio respiratorio medicina                   | ejercicio respiratorio medicina                   |
| 4042 | tener paciencia comer bien hacer ejercicio baj...  | [tener, paciencia, comer, bien, hacer, ejercic...]  | tener paciencia comer bien hacer ejercicio baj...  | tener paciencia comer bien hacer ejercicio baj... | tener paciencia comer bien hacer ejercicio baj... |

Nota: Visualización después de la eliminación de los AUX presentes en la columna POS correspondiente del dataset de Recomendaciones. Autores (2024). Código establecido en Google Colab.

### Aplicación de algoritmos de machine Learning

En el entrenamiento de los respectivos modelos, se utilizaron diversos clasificadores, entre ellos Naive Bayes, K-Neighbors y Regresión logística, la salida del modelo es del tipo Multi-Output Text Classification en base a las distintas etiquetas proporcionadas en los datasets y utilizadas para el entrenamiento de los modelos.

Para ambos datasets correspondientes a síntomas y recomendaciones se aplicó los distintos clasificadores para determinar el que mejores resultados en predicción resulte en base a las técnicas aplicadas.

En cada versión de los modelos aplicados, se realizó tres modelos o versiones

de dataframes, dónde el primer dataframe se le aplicó Tokenización, eliminación de stopwords, lematización y Pos tagging, el segundo dataframe posee las técnicas de Tokenización, eliminación de stopwords y lematización y, por último, el tercer dataframe contiene las técnicas de Tokenización y eliminación de stopwords.

Para cada modelo y variante de estos, se utilizó las etiquetas establecidas para cada dataset de síntomas y recomendaciones.

**Tabla 9:**

*Parámetros de versiones de cada modelo*

| Versión | Clasificador | Técnicas       | Test_Size |
|---------|--------------|----------------|-----------|
| 1       | NBM          | Tok/St/Lem/Pos | 0.2       |
| 2       | NBM          | Tok/St/Lem     | 0.2       |
| 3       | NBM          | Tok/St         | 0.2       |
| 1       | KNN          | Tok/St/Lem/Pos | 0.2       |
| 2       | KNN          | Tok/St/Lem     | 0.2       |
| 3       | KNN          | Tok/St         | 0.2       |
| 1       | LR           | Tok/St/Lem/Pos | 0.2       |
| 2       | LR           | Tok/St/Lem     | 0.2       |
| 3       | LR           | Tok/St         | 0.2       |

*Nota:* Representación de las variantes de cada algoritmo utilizado. Autores (2024). Código establecido en Google Colab.

### Clasificador Naive Bayes

Algoritmo empleado que utiliza Naive Bayes como clasificador

```
pipe_nb=Pipeline(steps=[('cv', CountVectorizer()),
('nb_multi', MultiOutputClassifier(MultinomialNB()))])
```

### Clasificador K-nearest neighbors

```
pipe_knn=Pipeline(steps=[('cv', CountVectorizer()),
('knn', KNeighborsClassifier(n_neighbors=5))])
```

### Clasificador Regresión Logística

```
pipe_lr = Pipeline(steps=[('cv', CountVectorizer()),
('lr', MultiOutputClassifier(LogisticRegression()))])
```

Para el conjunto de datos de entrenamiento y prueba correspondiente a X\_train,

X\_test, y\_train, y\_test en todos los modelos y sus variantes, se dividieron los datos para entrenamiento y pruebas, con una relación del 80% como datos de entrenamiento y 20% para datos de prueba

Train Test Split de los modelos de clasificación

```
from sklearn.model_selection import train_test_split
X_train, X_test , y_train, y_test = train_test_split(Xfeature, ylabels, test_size=0.2,
random_state=42)
```

#### 4.5.7. Fase 7: Evaluación

##### Matriz de validación

Para determinar que variante de aplicación de técnicas junto al modelo de clasificación, se realiza la matriz de validación cruzada de cada versión de los modelos, con sus respectivas métricas tales como: Accuracy train, Accuracy test, Precision train, Precision test, Recall train, Recall test, F1 train y F1 test.

El conjunto de datos de los datasets se divide en dos, tales como:

- 80% de los datos para train o entrenamiento.
- 20% de los datos para test o prueba.

De forma que los algoritmos puedan clasificar textos de síntomas y recomendaciones respectivamente.

##### Modelo con la etiqueta de síntoma

**Figura 60:**

Matriz de validación para el dataset de síntomas

| df_metricas |              |                |               |                 |                |              |             |          |          |          |
|-------------|--------------|----------------|---------------|-----------------|----------------|--------------|-------------|----------|----------|----------|
| Version     | Clasificador | Accuracy_train | Accuracy_test | Precision_train | Precision_test | Recall_train | Recall_test | F1_train | F1_test  |          |
| 0           | 1            | NBM            | 0.579618      | 0.496815        | 0.859900       | 0.852169     | 0.888455    | 0.844037 | 0.866966 | 0.830122 |
| 1           | 2            | NBM            | 0.579936      | 0.496815        | 0.859852       | 0.853045     | 0.888455    | 0.844037 | 0.866944 | 0.830333 |
| 2           | 3            | NBM            | 0.571019      | 0.475159        | 0.860042       | 0.846796     | 0.883611    | 0.839180 | 0.863697 | 0.822928 |
| 3           | 1            | KNN            | 0.678025      | 0.568153        | 0.946669       | 0.902287     | 0.832616    | 0.764166 | 0.879432 | 0.811908 |
| 4           | 2            | KNN            | 0.675478      | 0.565605        | 0.944879       | 0.898924     | 0.829656    | 0.767404 | 0.876768 | 0.813126 |
| 5           | 3            | KNN            | 0.685350      | 0.592357        | 0.965524       | 0.942069     | 0.830867    | 0.780356 | 0.882969 | 0.832725 |
| 6           | 1            | LR             | 0.845860      | 0.810191        | 0.975319       | 0.967260     | 0.934876    | 0.919050 | 0.953337 | 0.940663 |
| 7           | 2            | LR             | 0.845860      | 0.810191        | 0.974945       | 0.966871     | 0.935145    | 0.919050 | 0.953364 | 0.940449 |
| 8           | 3            | LR             | 0.846815      | 0.794904        | 0.976022       | 0.964528     | 0.934069    | 0.914733 | 0.953293 | 0.937104 |

Nota: Representación de los diferentes modelos mediante una Matriz de Validación Cruzada correspondiente al dataset de Síntomas. Autores (2024). Código establecido en Google Colab.

### Modelo de síntomas con Naive Bayes

La primera versión del modelo Naive Bayes se ejecutaron diversas técnicas NLP el cual consta de: Tokenizacion, Stopwords, Lematizacion y Part of Speech (POS). Teniendo como resultados proporcionados por la matriz de validación corresponde a las siguientes métricas: Con accuracy de 0.496815, una precisión de 0.852169, una sensibilidad de 0.844037 y F1 de 0.830122.

La segunda versión del modelo Naive Bayes se ejecutaron diversas técnicas NLP el cual consta de: Tokenizacion, Stopwords y Lematizacion. Teniendo como resultados proporcionados por la matriz de validación corresponde a las siguientes métricas: Con accuracy de 0.496815, una precisión de 0.853045, una sensibilidad de 0.844037 y F1 de 0.830333.

La tercera versión del modelo Naive Bayes se ejecutaron diversas técnicas NLP el cual consta de: Tokenizacion y Stopwords. Teniendo como resultados proporcionados por la matriz de validación corresponde a las siguientes métricas: Con accuracy de 0.475159, una precisión de 0.846796, una sensibilidad de 0.839180 y F1 de 0.822928.

### Modelo de síntomas con k-nearest neighbors

La primera versión del modelo k-nearest neighbors se ejecutaron diversas técnicas NLP el cual consta de: Tokenización, Stopwords, Lematizacion y Part of Speech (POS). Teniendo como resultados proporcionados por la matriz de validación corresponde a las siguientes métricas: Con accuracy de 0.568153, una precisión de 0.902287, una sensibilidad de 0.764166 y F1 de 0.811908.

La segunda versión del modelo k-nearest neighbors se ejecutaron diversas técnicas NLP el cual consta de: Tokenización, Stopwords y Lematizacion. Teniendo como resultados proporcionados por la matriz de validación corresponde a las siguientes métricas: Con accuracy de 0.565605, una precisión de 0.898924, una sensibilidad de 0.767404 y F1 de 0.813126.

La tercera versión del modelo k-nearest neighbors se ejecutaron diversas técnicas NLP el cual consta de: Tokenización y Stopwords. Teniendo como resultados proporcionados por la matriz de validación corresponde a las siguientes métricas: Con accuracy de 0.592357, una precisión de 0.942069, una

sensibilidad de 0.780356 y F1 de 0.832725.

### Modelo de síntomas con Regresión Logística

La primera versión del modelo Regresión Logística se ejecutaron diversas técnicas NLP el cual consta de: Tokenizacion, Stopwords, Lematizacion y Part of Speech (POS). Teniendo como resultados proporcionados por la matriz de validación corresponde a las siguientes métricas: Con accuracy de 0.810191, una precisión de 0.967260, una sensibilidad de 0.919050 y F1 de 0.940663.

La segunda versión del modelo Regresión Logística se ejecutaron diversas técnicas NLP el cual consta de: Tokenizacion, Stopwords y Lematizacion. Teniendo como resultados proporcionados por la matriz de validación corresponde a las siguientes métricas: Con accuracy de 0.810191, una precisión de 0.966871, una sensibilidad de 0.919050 y F1 de 0.940449.

La tercera versión del modelo Regresión Logística se ejecutaron diversas técnicas NLP el cual consta de: Tokenizacion y Stopwords. Teniendo como resultados proporcionados por la matriz de validación corresponde a las siguientes métricas: Con accuracy de 0.794904, una precisión de 0.964528, una sensibilidad de 0.914733y F1 de 0.937104.

### Modelo con la etiqueta de recomendaciones

**Figura 61:**

*Matriz de validación para el dataset de recomendaciones*

| df_metricas |   |         |              |                |               |                 |                |              |             |          |          |
|-------------|---|---------|--------------|----------------|---------------|-----------------|----------------|--------------|-------------|----------|----------|
|             |   | Version | Clasificador | Accuracy_train | Accuracy_test | Precision_train | Precision_test | Recall_train | Recall_test | F1_train | F1_test  |
| 0           | 1 | NBM     | NBM          | 0.752048       | 0.672365      | 0.897914        | 0.877242       | 0.852360     | 0.773377    | 0.872559 | 0.812958 |
| 1           | 2 | NBM     | NBM          | 0.750980       | 0.673789      | 0.898418        | 0.879003       | 0.851537     | 0.771177    | 0.872319 | 0.812083 |
| 2           | 3 | NBM     | NBM          | 0.773424       | 0.668091      | 0.907532        | 0.889887       | 0.863063     | 0.770077    | 0.883202 | 0.814404 |
| 3           | 1 | KNN     | KNN          | 0.652298       | 0.616809      | 0.908022        | 0.874656       | 0.695664     | 0.654565    | 0.770297 | 0.720717 |
| 4           | 2 | KNN     | KNN          | 0.655860       | 0.623932      | 0.910300        | 0.895863       | 0.695115     | 0.655666    | 0.769925 | 0.725880 |
| 5           | 3 | KNN     | KNN          | 0.626648       | 0.591168      | 0.910434        | 0.895570       | 0.665752     | 0.613861    | 0.744569 | 0.688951 |
| 6           | 1 | LR      | LR           | 0.827218       | 0.739316      | 0.967399        | 0.935436       | 0.873216     | 0.797580    | 0.916325 | 0.856020 |
| 7           | 2 | LR      | LR           | 0.828643       | 0.739316      | 0.967436        | 0.935313       | 0.874314     | 0.795380    | 0.916999 | 0.854647 |
| 8           | 3 | LR      | LR           | 0.828643       | 0.725071      | 0.970664        | 0.937053       | 0.872667     | 0.782178    | 0.917463 | 0.848340 |

*Nota:* Representación de los diferentes modelos mediante una Matriz de Validación Cruzada correspondiente al dataset de Recomendaciones. Autores (2024). Código establecido en Google Colab.

### Modelo de recomendaciones con Naive Bayes

La primera versión del modelo Naive Bayes se ejecutaron diversas técnicas NLP el cual consta de: Tokenizacion, Stopwords, Lematizacion y Part of Speech (POS). Teniendo como resultados proporcionados por la matriz de validación corresponde a las siguientes métricas: Con accuracy de 0.672365, una precisión de 0.877242, una sensibilidad de 0.773377 y F1 de 0.812958.

La segunda versión del modelo Naive Bayes se ejecutaron diversas técnicas NLP el cual consta de: Tokenizacion, Stopwords y Lematizacion. Teniendo como resultados proporcionados por la matriz de validación corresponde a las siguientes métricas: Con accuracy de 0.673789, una precisión de 0.879003, una sensibilidad de 0.771177 y F1 de 0.812083.

La tercera versión del modelo Naive Bayes se ejecutaron diversas técnicas NLP el cual consta de: Tokenizacion y Stopwords. Teniendo como resultados proporcionados por la matriz de validación corresponde a las siguientes métricas: Con accuracy de 0.668091, una precisión de 0.889887, una sensibilidad de 0.770077 y F1 de 0.814404.

### Modelo de recomendaciones con k-nearest neighbors

La primera versión del modelo k-nearest neighbors se ejecutaron diversas técnicas NLP el cual consta de: Tokenizacion, Stopwords, Lematizacion y Part of Speech (POS). Teniendo como resultados proporcionados por la matriz de validación corresponde a las siguientes métricas: Con accuracy de 0.616809, una precisión de 0.874656, una sensibilidad de 0.654565 y F1 de 0.720717.

La segunda versión del modelo k-nearest neighbors se ejecutaron diversas técnicas NLP el cual consta de: Tokenizacion, Stopwords y Lematizacion. Teniendo como resultados proporcionados por la matriz de validación corresponde a las siguientes métricas: Con accuracy de 0.623932, una precisión de 0.895863, una sensibilidad de 0.655666 y F1 de 0.725880.

La tercera versión del modelo k-nearest neighbors se ejecutaron diversas técnicas NLP el cual consta de: Tokenizacion y Stopwords. Teniendo como resultados proporcionados por la matriz de validación corresponde a las siguientes métricas: Con accuracy de 0.591168, una precisión de 0.895570, una

sensibilidad de 0.613861 y F1 de 0.688951.

### Modelo de recomendaciones con Regresión Logística

La primera versión del modelo Regresión Logística se ejecutaron diversas técnicas NLP el cual consta de: Tokenización, Stopwords, Lematización y Part of Speech (POS). Teniendo como resultados proporcionados por la matriz de validación corresponde a las siguientes métricas: Con accuracy de 0.739316, una precisión de 0.935436, una sensibilidad de 0.797580 y F1 de 0.856020.

La segunda del modelo Regresión Logística se ejecutaron diversas técnicas NLP el cual consta de: Tokenización, Stopwords y Lematización. Teniendo como resultados proporcionados por la matriz de validación corresponde a las siguientes métricas: Con accuracy de 0.739316, una precisión de 0.935313, una sensibilidad de 0.795380 y F1 de 0.854647.

La tercera versión del modelo Regresión Logística se ejecutaron diversas técnicas NLP el cual consta de: Tokenización y Stopwords. Teniendo como resultados proporcionados por la matriz de validación corresponde a las siguientes métricas: Con accuracy de 0.725071, una precisión de 0.937053, una sensibilidad de 0.782178 y F1 de 0.848340.

### Elección de modelo de clasificación y técnicas NLP

Con los modelos propuestos para el entrenamiento junto con las técnicas de procesamiento de lenguaje natural, mediante las métricas definidas, se consideraron a aquellas referentes a los datos de test, para la clasificación de valores nuevos.

Entre los clasificadores utilizados referentes a Naive Bayes, KNN y Regresión logística, se aprecia que, de los 3 modelos de clasificación, el modelo con Regresión Logística con técnicas de NLP tales como Stopwords, Tokenización, Lemmatización y PoS Tagging, dio resultados mayores a los demás clasificadores.

### Matriz de confusión

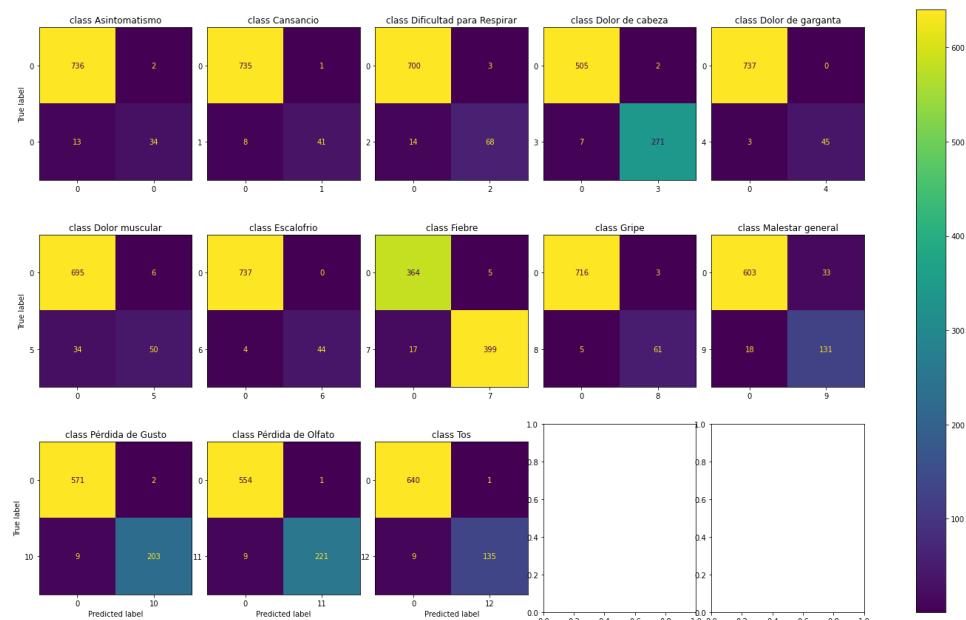
Por cada una de las etiquetas de síntomas, se realiza la matriz de confusión indicando los TN y TP correspondiente a los valores del modelo

### Matriz de confusión de las clases de síntomas

Por cada una de las etiquetas de síntomas, se realiza la matriz de confusión indicando los TN y TP correspondiente a los valores del modelo.

**Figura 62:**

*Matriz de confusión de las clases de síntomas*



*Nota:* Matriz de confusión de cada etiqueta de Síntomas que corresponde al modelo de Regresión Logística. Autores (2024). Código establecido en Google Colab.

### Clase Asintomatismo

En la etiqueta o clase Asintomatismo, se presentan valores correspondientes a TN = 736 y TP = 34 que muestra los valores que estima el modelo como forma correcta. En FN = 13 y FP = 2 para aquellos valores que el modelo no estimó de forma correcta.

$$precision = \frac{34}{34 + 2} = 0.9444444444$$

$$recall = \frac{34}{34 + 13} = 0.7234042553$$

$$F1 = 2 \cdot \frac{0.9444444444 \cdot 0.7234042553}{0.9444444444 + 0.7234042553} = 0.8192771084$$

### Clase Cansancio

En la etiqueta o clase Cansancio, se presentan valores correspondientes a TN = 735 y TP = 41 que muestra los valores que estima el modelo como forma

correcta. En FN = 8 y FP 1 = para aquellos valores que el modelo no estimó de forma correcta.

$$\text{precisión} = \frac{41}{41 + 1} = 0.9761904762$$

$$\text{recall} = \frac{41}{41 + 8} = 0.8367346939$$

$$F1 = 2 \cdot \frac{0.9761904762 \cdot 0.8367346939}{0.9761904762 + 0.8367346939} = 0.9010989011$$

### Clase Dificultad para respirar

En la etiqueta o clase Dificultad para respirar, se presentan valores correspondientes a TN = 700 y TP = 68 que muestra los valores que estima el modelo como forma correcta. En FN = 14 y FP = 3 para aquellos valores que el modelo no estimó de forma correcta.

$$\text{precisión} = \frac{68}{68 + 3} = 0.9577464789$$

$$\text{recall} = \frac{68}{68 + 14} = 0.8292682927$$

$$F1 = 2 \cdot \frac{0.9577464789 \cdot 0.8292682927}{0.9577464789 + 0.8292682927} = 0.8888888889$$

### Clase Dolor de cabeza

En la etiqueta o clase Dificultad de cabeza, se presentan valores correspondientes a TN = 505y TP = 271 que muestra los valores que estima el modelo como forma correcta. En FN = 7 y FP = 2 para aquellos valores que el modelo no estimó de forma correcta.

$$\text{precisión} = \frac{271}{271 + 2} = 0.9926739927$$

$$\text{recall} = \frac{271}{271 + 7} = 0.9748201439$$

$$F1 = 2 \cdot \frac{0.9926739927 \cdot 0.9748201439}{0.9926739927 + 0.9748201439} = 0.9836660617$$

### Clase Dolor de garganta

En la etiqueta o clase Dolor de garganta, se presentan valores correspondientes a TN = 737 y TP = 45 que muestra los valores que estima el modelo como forma correcta. En FN = 3 y FP = 0 para aquellos valores que el modelo no estimó de forma correcta.

$$\text{precisión} = \frac{45}{45 + 0} = 1$$

$$\text{recall} = \frac{45}{45 + 3} = 0.9375$$

$$F1 = 2 \cdot \frac{1 \cdot 0.9375}{1 + 0.9375} = 0.9677419355$$

### Clase Dolor muscular

En la etiqueta o clase Dolor muscular, se presentan valores correspondientes a TN = 695 y TP = 50 que muestra los valores que estima el modelo como forma correcta. En FN = 34 y FP = 6 para aquellos valores que el modelo no estimó de forma correcta.

$$\text{precisión} = \frac{50}{50 + 6} = 0.8928571429$$

$$\text{recall} = \frac{50}{50 + 34} = 0.5952380952$$

$$F1 = 2 \cdot \frac{0.8928571429 \cdot 0.5952380952}{0.8928571429 + 0.5952380952} = 0.7142857143$$

### Clase Escalofrío

En la etiqueta o clase Escalofrío, se presentan valores correspondientes a TN = 737 y TP = 44 que muestra los valores que estima el modelo como forma correcta. En FN = 4 y FP = 0 para aquellos valores que el modelo no estimó de forma correcta.

$$\text{precisión} = \frac{44}{44 + 0} = 1$$

$$\text{recall} = \frac{44}{44 + 4} = 0.9166666667$$

$$F1 = 2 \cdot \frac{1 \cdot 0.9166666667}{1 + 0.9166666667} = 0.9565217391$$

### Clase Fiebre

En la etiqueta o clase Fiebre, se presentan valores correspondientes a TN = 364 y TP = 399 que muestra los valores que estima el modelo como forma correcta. En FN = 17 y FP = 5 para aquellos valores que el modelo no estimó de forma correcta.

$$\text{precisión} = \frac{399}{399 + 5} = 0.9876237624$$

$$\text{recall} = \frac{399}{399 + 17} = 0.9591346154$$

$$F1 = 2 \cdot \frac{0.9876237624 \cdot 0.9591346154}{0.9876237624 + 0.9591346154} = 0.9731707317$$

### Clase Gripe

En la etiqueta o clase Gripe, se presentan valores correspondientes a TN = 716 y TP = 61 que muestra los valores que estima el modelo como forma correcta. En FN = 5 y FP = 3 para aquellos valores que el modelo no estimó de forma correcta.

$$\text{precisión} = \frac{61}{61 + 3} = 0.953125$$

$$\text{recall} = \frac{61}{61 + 5} = 0.9242424242$$

$$F1 = 2 \cdot \frac{0.953125 \cdot 0.9242424242}{0.953125 + 0.9242424242} = 0.9384615384$$

### Clase Malestar general

En la etiqueta o clase Malestar general, se presentan valores correspondientes a TN = 603 y TP = 131 que muestra los valores que estima el modelo como forma correcta. En FN = 18 y FP = 33 para aquellos valores que el modelo no estimó de forma correcta.

$$\text{precisión} = \frac{131}{131 + 33} = 0.7987804878$$

$$\text{recall} = \frac{131}{131 + 18} = 0.8791946309$$

$$F1 = 2 \cdot \frac{0.7987804878 \cdot 0.8791946309}{0.7987804878 + 0.8791946309} = 0.8370607029$$

### Clase Pérdida de gusto

En la etiqueta o clase Pérdida de gusto, se presentan valores correspondientes a TN = 571 y TP = 203 que muestra los valores que estima el modelo como forma correcta. En FN = 9 y FP = 2 para aquellos valores que el modelo no estimó de forma correcta.

$$\text{precisión} = \frac{203}{203 + 2} = 0.9902439024$$

$$\text{recall} = \frac{203}{203 + 9} = 0.9575471698$$

$$F1 = 2 \cdot \frac{0.9902439024 \cdot 0.9575471698}{0.9902439024 + 0.9575471698} = 0.9736211031$$

### Clase Pérdida de olfato

En la etiqueta o clase Pérdida de olfato, se presentan valores correspondientes a TN = 554 y TP = 221 que muestra los valores que estima el modelo como forma correcta. En FN = 9 y FP = 1 para aquellos valores que el modelo no estimó de forma correcta.

$$\text{precisión} = \frac{221}{221 + 1} = 0.9954954955$$

$$\text{recall} = \frac{221}{221 + 9} = 0.9608695652$$

$$F1 = 2 \cdot \frac{0.9954954955 \cdot 0.9608695652}{0.9954954955 + 0.9608695652} = 0.9778761062$$

### Clase Tos

En la etiqueta o clase Tos, se presentan valores correspondientes a TN = 640 y TP = 135 que muestra los valores que estima el modelo como forma correcta. En FN = 9 y FP = 1 para aquellos valores que el modelo no estimó de forma correcta.

$$\text{precisión} = \frac{135}{135 + 1} = 0.9926470588$$

$$\text{recall} = \frac{135}{135 + 9} = 0.9375$$

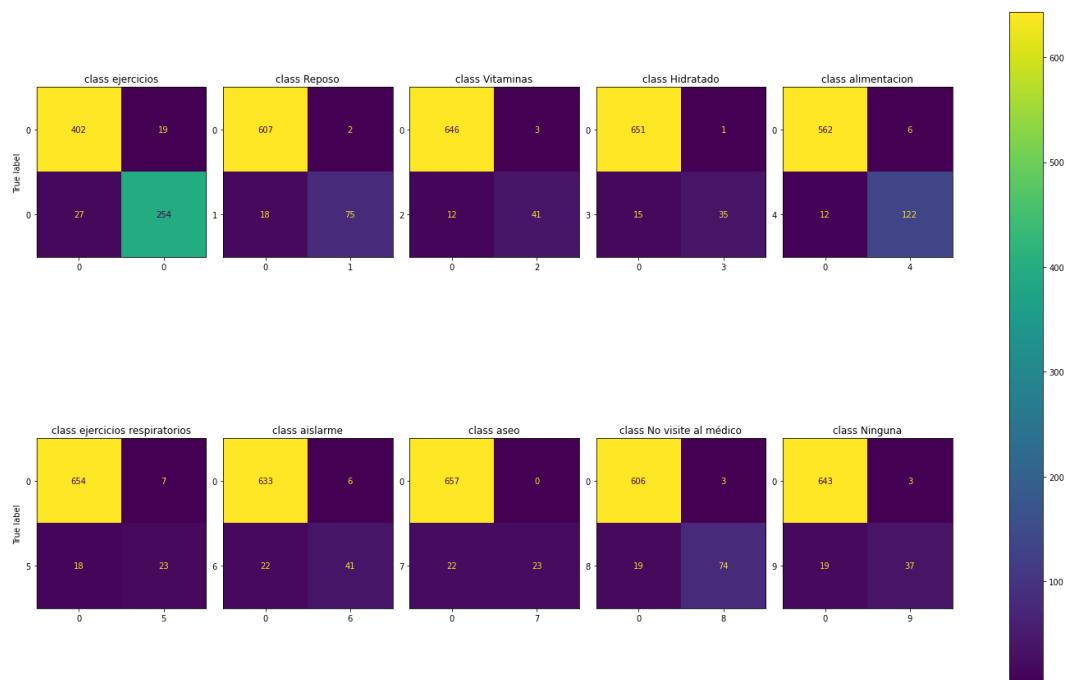
$$F1 = 2 \cdot \frac{0.9926470588 \cdot 0.9375}{0.9926470588 + 0.9375} = 0.9642857143$$

## Matriz de confusión de las clases de recomendaciones

Por cada una de las etiquetas de recomendaciones, se realiza la matriz de confusión indicando los TN y TP correspondiente a los valores del modelo.

**Figura 63:**

*Matriz de confusión de las clases de recomendaciones*



*Nota:* Matriz de confusión de cada etiqueta de Recomendaciones que corresponde al modelo de Regresión Logística. Autores (2024). Código establecido en Google Colab.

### Clase Ejercicios

En la etiqueta o clase Ejercicios, se presentan valores correspondientes a TN = 402 y TP = 254 que muestra los valores que estima el modelo como forma correcta. En FN = 27 y FP = 19 para aquellos valores que el modelo no estimó de forma correcta.

$$precision = \frac{254}{254 + 19} = 0.9304029304$$

$$recall = \frac{254}{254 + 27} = 0.9039145907$$

$$F1 = 2 \cdot \frac{0.9304029304 \cdot 0.9039145907}{0.9304029304 + 0.9039145907} = 0.916967509$$

### Clase Reposo

En la etiqueta o clase Reposo, se presentan valores correspondientes a TN =

607 y TP = 75 que muestra los valores que estima el modelo como forma correcta. En FN = 18 y FP = 2 para aquellos valores que el modelo no estimó de forma correcta.

$$\text{precisión} = \frac{75}{75 + 2} = 0.974025974$$

$$\text{recall} = \frac{75}{75 + 18} = 0.8064516129$$

$$F1 = 2 \cdot \frac{0.974025974 \cdot 0.8064516129}{0.974025974 + 0.8064516129} = 0.8823529412$$

### Clase Vitaminas

En la etiqueta o clase Vitaminas, se presentan valores correspondientes a TN = 646 y TP = 41 que muestra los valores que estima el modelo como forma correcta. En FN = 12 y FP = 3 para aquellos valores que el modelo no estimó de forma correcta.

$$\text{precisión} = \frac{41}{41 + 3} = 0.9318181818$$

$$\text{recall} = \frac{41}{41 + 12} = 0.7735849057$$

$$F1 = 2 \cdot \frac{0.9318181818 \cdot 0.7735849057}{0.9318181818 + 0.7735849057} = 0.8453608248$$

### Clase Hidratado

En la etiqueta o clase Hidratado, se presentan valores correspondientes a TN = 651 y TP = 35 que muestra los valores que estima el modelo como forma correcta. En FN = 15 y FP = 1 para aquellos valores que el modelo no estimó de forma correcta.

$$\text{precisión} = \frac{35}{35 + 1} = 0.9722222222$$

$$\text{recall} = \frac{35}{35 + 15} = 0.7$$

$$F1 = 2 \cdot \frac{0.9722222222 \cdot 0.7}{0.9722222222 + 0.7} = 0.8139534884$$

### Clase Alimentación

En la etiqueta o clase Alimentación, se presentan valores correspondientes a TN

= 562 y TP = 122 que muestra los valores que estima el modelo como forma correcta. En FN = 12 y FP = 6 para aquellos valores que el modelo no estimó de forma correcta.

$$\text{precisión} = \frac{122}{122 + 6} = 0.953125$$

$$\text{recall} = \frac{122}{122 + 12} = 0.9104477612$$

$$F1 = 2 \cdot \frac{0.953125 \cdot 0.9104477612}{0.953125 + 0.9104477612} = 0.9312977845$$

### Clase Ejercicios respiratorios

En la etiqueta o clase Ejercicios respiratorios, se presentan valores correspondientes a TN = 654 y TP = 23 que muestra los valores que estima el modelo como forma correcta. En FN = 18 y FP = 7 para aquellos valores que el modelo no estimó de forma correcta.

$$\text{precisión} = \frac{23}{23 + 7} = 0.7666666667$$

$$\text{recall} = \frac{23}{23 + 18} = 0.5609756098$$

$$F1 = 2 \cdot \frac{0.7666666667 \cdot 0.5609756098}{0.7666666667 + 0.5609756098} = 0.647887324$$

### Clase Aislarme

En la etiqueta o clase Aislarme, se presentan valores correspondientes a TN = 633 y TP = 41 que muestra los valores que estima el modelo como forma correcta. En FN = 22 y FP = 6 para aquellos valores que el modelo no estimó de forma correcta.

$$\text{precision} = \frac{41}{41 + 6} = 0.8723404255$$

$$\text{recall} = \frac{41}{41 + 22} = 0.6507936508$$

$$F1 = 2 \cdot \frac{0.8723404255 \cdot 0.6507936508}{0.8723404255 + 0.6507936508} = 0.7454545454$$

### Clase Aseo

En la etiqueta o clase Aseo, se presentan valores correspondientes a TN = 657

y TP = 23 que muestra los valores que estima el modelo como forma correcta. En FN = 22 y FP = 0 para aquellos valores que el modelo no estimó de forma correcta.

$$\text{precisión} = \frac{23}{23 + 0} = 1$$

$$\text{recall} = \frac{23}{23 + 22} = 0.5111111111$$

$$F1 = 2 \cdot \frac{1 \cdot 0.5111111111}{1 + 0.5111111111} = 0.6764705882$$

### Clase No visité al médico

En la etiqueta o clase No visité al médico, se presentan valores correspondientes a TN = 606 y TP = 74 que muestra los valores que estima el modelo como forma correcta. En FN = 19 y FP = 3 para aquellos valores que el modelo no estimó de forma correcta.

$$\text{precisión} = \frac{74}{74 + 3} = 0.961038961$$

$$\text{recall} = \frac{74}{74 + 19} = 0.7956989247$$

$$F1 = 2 \cdot \frac{0.961038961 \cdot 0.7956989247}{0.961038961 + 0.7956989247} = 0.8705882353$$

### Clase Ninguna

En la etiqueta o clase Ninguna, se presentan valores correspondientes a TN = 643 y TP = 37 que muestra los valores que estima el modelo como forma correcta. En FN = 19 y FP = 3 para aquellos valores que el modelo no estimó de forma correcta.

$$\text{precisión} = \frac{37}{37 + 3} = 0.925$$

$$\text{recall} = \frac{37}{37 + 19} = 0.6607142857$$

$$F1 = 2 \cdot \frac{0.925 \cdot 0.6607142857}{0.925 + 0.6607142857} = 0.7708333333$$

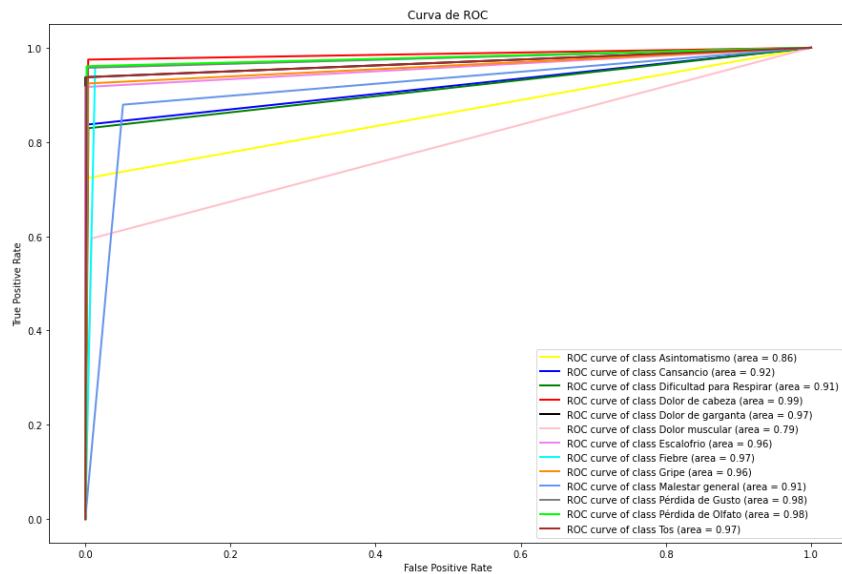
### Curva de ROC

Para determinar el desempeño de un problema de clasificación, se hace uso de

la Curva de AUC-ROC, la cual es una curva probabilística que permite representar si es modelo es capaz de distinguir entre clases, donde entre mayor sea el valor de AUC, mejor es el modelo en predecir clases 0 y clases 1. (Narkhede, 2018)

**Figura 64:**

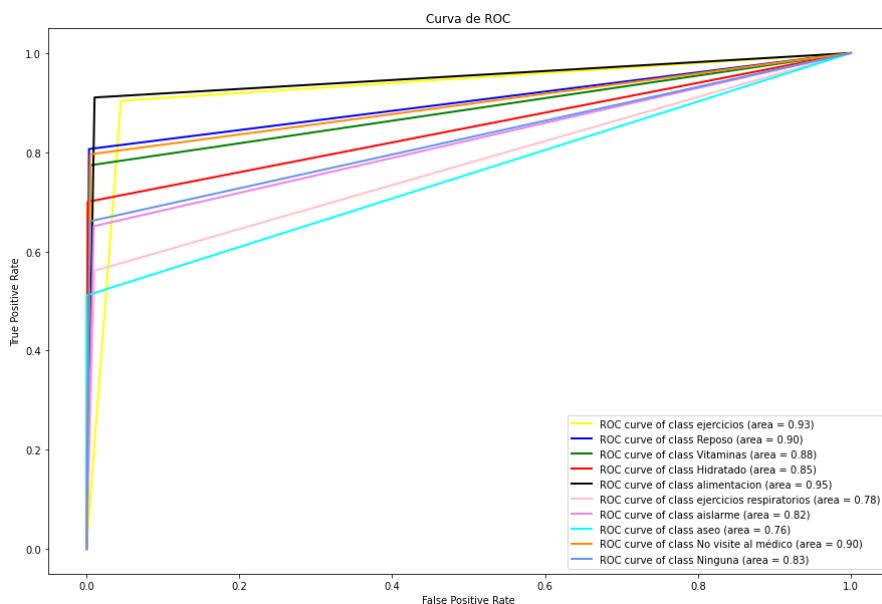
*Curva de ROC de las clases de síntomas*



Nota: Curva de ROC de cada etiqueta de Síntomas que corresponde al modelo de Regresión Logística. Autores (2024). Código establecido en Google Colab.

**Figura 65:**

*Curva de ROC de las clases de recomendaciones*



Nota: Curva de ROC de cada etiqueta de Síntomas que corresponde al modelo de Regresión Logística. Autores (2024). Código establecido en Google Colab.

#### 4.5.8. Fase 8: Predicción

##### Prueba de predicción con datos de test

En base a la división del 80% de los datos para el entrenamiento, el 20% se usa para pruebas, con ello se permite apreciar el desenvolvimiento del modelo de clasificación para datos no conocidos, es decir con datos con los que no fue entrenado.

**Figura 66:**

*Prueba de predicción de test para síntomas*

```
[114] ex_v1=X_test_v1.iloc[150]
lista_consulta=pipe_lr_v1.predict([ex_v1])
print(ex_v1)
print(pipe_lr_v1.predict([ex_v1]))

tos dolor cuerpo dolor cabeza neumonia atípico
[[0 0 0 1 0 0 0 0 1 0 0 1]]
```

[115] df = pd.DataFrame(lista\_consulta,columns=["Asintomatismo", "Cansancio", "Dificultad para Respirar", "Dolor de cabeza", "Dolor de garganta", "Dolor muscular", "Escalofrío", "Fiebre", "Gripe", "Malestar general", "Pérdida de Gusto", "Pérdida de Olfato", "Tos"])

|   | Asintomatismo | Cansancio | Dificultad para Respirar | Dolor de cabeza | Dolor de garganta | Dolor muscular | Escalofrío | Fiebre | Gripe | Malestar general | Pérdida de Gusto | Pérdida de Olfato | Tos |
|---|---------------|-----------|--------------------------|-----------------|-------------------|----------------|------------|--------|-------|------------------|------------------|-------------------|-----|
| 0 | 0             | 0         | 0                        | 1               | 0                 | 0              | 0          | 0      | 0     | 1                | 0                | 0                 | 1   |

Nota: Visualización de resultados de predicción de test de síntomas. Autores (2024). Código establecido en Google Colab.

**Figura 67:**

*Prueba de predicción de datos de test para recomendaciones*

```
[115] exv1=X_testv1.iloc[40]
lista_consulta=pipe_lr_v1.predict([exv1])
print(exv1)
print(pipe_lr_v1.predict([exv1]))

coma saludable hacer ejercicio jugos verde
[[1 0 0 1 0 0 0 0]]
```

[116] df = pd.DataFrame(lista\_consulta,columns=["ejercicios", "Reposo", "Vitaminas", "Hidratado", "alimentacion", "ejercicios respiratorios", "aislarime", "aseo", "No visite al médico", "Ninguna"])

|   | ejercicios | Reposo | Vitaminas | Hidratado | alimentacion | ejercicios respiratorios | aislarime | aseo | No visite al médico | Ninguna |
|---|------------|--------|-----------|-----------|--------------|--------------------------|-----------|------|---------------------|---------|
| 0 | 1          | 0      | 0         | 0         | 1            | 0                        | 0         | 0    | 0                   | 0       |

Nota: Visualización de resultados de predicción de test de recomendaciones. Autores (2024). Código establecido en Google Colab.

##### Prueba de laboratorio con datos nuevos

Para comprobar el funcionamiento efectivo del modelo de clasificación de texto para una futura puesta en predicción, se ingresa texto haciendo referencia a síntomas y recomendaciones o hábitos saludables para visualizar los resultados de predicción del modelo.

**Figura 68:**

Prueba de predicción de datos ingresados para síntomas.

| [122] PruebaSintomas="Tenia dolor muscular y algo de gripe"   |           |                          |                 |                   |                |            |        |       |                  |                  |                   |     |
|---|-----------|--------------------------|-----------------|-------------------|----------------|------------|--------|-------|------------------|------------------|-------------------|-----|
| lista_pred = pipe_lr_v1.predict([PruebaSintomas])   |           |                          |                 |                   |                |            |        |       |                  |                  |                   |     |
| df_pred = pd.DataFrame(lista_pred,columns=[ "Asintomatismo", "Cansancio", "Dificultad para Respirar", "Dolor de cabeza", "Dolor de garganta", "Dolor muscular", "Escalofrio", "Fiebre", "Gripe", "Malestar general", "Pérdida de Gusto", "Pérdida de Olfato", "Tos"]) |           |                          |                 |                   |                |            |        |       |                  |                  |                   |     |
| Asintomatismo   | Cansancio | Dificultad para Respirar | Dolor de cabeza | Dolor de garganta | Dolor muscular | Escalofrio | Fiebre | Gripe | Malestar general | Pérdida de Gusto | Pérdida de Olfato | Tos |
| 0   | 0         | 0                        | 0               | 0                 | 0              | 1          | 0      | 0     | 1                | 0                | 0                 | 0   |

Nota: Visualización de resultados de predicción de test de síntomas. Autores (2024). Código establecido en Google Colab.

**Figura 69:**

Prueba de predicción de datos ingresados para recomendaciones.

| [ ] PruebaRecomendacion="Hay que tomar agua todos los días"   |        |           |           |              |                          |           |      |                     |         |   |   |
|---|--------|-----------|-----------|--------------|--------------------------|-----------|------|---------------------|---------|---|---|
| lista_pred=pipe_lr_v1.predict([PruebaRecomendacion])  |        |           |           |              |                          |           |      |                     |         |   |   |
| df_pred=pd.DataFrame(lista_pred,columns=[ "ejercicios", "Reposo", "Vitaminas", "Hidratado", "alimentacion", "ejercicios respiratorios", "aislarime", "aseo", "No visite al médico", "Ninguna"]) |        |           |           |              |                          |           |      |                     |         |   |   |
| ejercicios  | Reposo | Vitaminas | Hidratado | alimentacion | ejercicios respiratorios | aislarime | aseo | No visite al médico | Ninguna |   |   |
| 0   | 0      | 0         | 0         | 1            | 0                        | 0         | 0    | 0                   | 0       | 0 | 0 |

Nota: Visualización de resultados de predicción mediante datos ingresados de recomendaciones. Autores (2024). Código establecido en Google Colab.

## 4.6. Anexo 1

Cuestionario a personas contagiadas de Covid-19 de la zona 8 (preguntas no tabuladas)

- Pregunta 1. ¿Ha tenido coronavirus?
- Pregunta 5. ¿En qué fecha se contagió?
- Pregunta 10. Describa lo más detallado ¿Qué síntomas ha tenido?
- Pregunta 11. Describa ¿Qué medicamentos considera que le ayudó en su recuperación?
- Pregunta 12. Describa ¿Qué cuidados aplicó durante el proceso de recuperación del Covid-19, y cuánto tiempo en días tomó su recuperación?
- Pregunta 13. Describa a más detalle ¿Qué alimentos y/o vitaminas considera que le ayudaron a fortalecerse y superar el Covid-19?
- 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?
- Pregunta 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

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



# **REFERENCIAS BIBLIOGRÁFICAS**



## Referencias Bibliográficas

- Arias, J. (2019). *Técnicas e instrumentos de investigación científica* (1st ed.). Enfoques Consulting EIRL. <http://hdl.handle.net/20.500.12390/2238>
- Arispe, C., Yangali, J., Guerrero, M., Lozada, O., Acuña, L., & Arellano, C. (2020). *La investigación científica*. Universidad Internacional del Ecuador. <https://repositorio.uide.edu.ec/handle/37000/4310>
- Bucalo, M. L., Barbieri, C., Roca, S., Ion Titapiccolo, J., Ros Romero, M. S., Ramos, R., Albaladejo, M., Manzano, D., Mari, F., & Molina, M. (2018). *El modelo de control de anemia: ¿ayuda al nefrólogo en la decisión terapéutica para el manejo de la anemia?* *Nefrología*, 38(5), 491–502. <https://doi.org/10.1016/j.nefro.2018.03.004>
- Castro, A., Meléndez, L., López, G., Soto, I., & Muñoz, R. (2018). La investigación exploratoria aplicada como estrategia didáctica en el laboratorio. *Revista Electrónica Sobre Cuerpos Académicos y Grupos de Investigación*, 5(10). <https://www.cagi.org.mx/index.php/CAGI/article/view/184>
- Castro, M. (2019). Bioestadística aplicada en investigación clínica: conceptos básicos. *Revista Médica Clínica Las Condes*, 30(1), 50–65. <https://doi.org/10.1016/j.rmclc.2018.12.002>
- Erazo-Luzuriaga, A. F. (2024). Integración de las TICs en el aula: Un análisis de su impacto en el rendimiento académico. *Revista Científica Zambos*, 3(1), 56-72. <https://doi.org/10.69484/rpz/v3/n1/12>
- Erazo-Luzuriaga, A. F., Ramos-Secaira, F. M., Galarza-Sánchez, P. C., & Boné-Andrade, M. F. (2023). La inteligencia artificial aplicada a la optimización de programas informáticos. *Journal of Economic and Social Science Research*, 3(1), 48–63. <https://doi.org/10.55813/gaea/jessr/v3/n1/61>
- Gallastegui, L. M. G. (2008). *Inteligencia Artificial: In Miradas sobre el emprendimiento ante la crisis del coronavirus*. <https://doi.org/10.2307/j.ctv2gz3w9c.97>
- Garduño, E., Albarrán, D., & Damián, 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>
- González, G. (2020, 2 marzo). Investigación diagnóstica: características, técnicas, tipos, ejemplos. Lifeder. Recuperado 6 de marzo de 2022, de <https://www.lifeder.com/investigacion-diagnostica/>
- Montalván-Vélez, C. L., Mogrovejo-Zambrano, J. N., Romero-Vitte, I. J., & Pinargote-Carrera, M. L. D. C. (2024). Introducción a la Inteligencia Artificial: Conceptos Básicos y Aplicaciones Cotidianas . *Journal of Economic and Social Science Research*, 4(1), 173–183. <https://doi.org/10.55813/gaea/jessr/v4/n1/93>
- Nadkarni PM, Ohno-Machado L, Chapman WW. *Natural language processing: an introduction*. J Am Med Inform Assoc. 2011 Sep-Oct;18(5):544-51. doi: [10.1136/amiajnl-2011-000464](https://doi.org/10.1136/amiajnl-2011-000464). PMID: 21846786; PMCID: PMC3168328.
- Narkhede, S. (2018, 26 junio). *Understanding AUC - ROC Curve - Towards Data Science*. Medium. Recuperado 20 de julio de 2022, de <https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5>
- Pineda, J. M. (2022). Modelos predictivos en salud basados en aprendizaje de maquina ( machine learning ). *Revista Clínica Las Condes*, 33(6), 583–590. <https://doi.org/10.1016/j.rmclc.2022.11.002>

Referencias:

- Quinatoa-Chasi, W. D., Cepeda-Valente, W. M., Chasi-Chela, A. V., Chasi-Chela, N. F., Casanova-Villalba, C. I., Salgado-Ortiz, P. J., Guerrero-Freire, E. I., Guerrero-Freire, A. E., Herrera-Sánchez, M. J., Mina-Bone, S. G., Santana-Torres, A. A., Ríos-Gaibor, C. G., Calero-Cherres, R. V., López-Salinas, C. M., Mora-Estrada, I. A., & Chuchuca-Peñaiza, P. M. (2024). *Fronteras del Futuro: Innovación y Desarrollo en Ciencia y Tecnología*. Editorial Grupo AEA. <https://doi.org/10.55813/eqaea.1.69>
- Ramos, C. (2021). Diseños de Investigación Experimental. *CienciAmérica Revista de Divulgación Científica de La Universidad Tecnológica Indoamérica*, 10(1), 1–7. <https://doi.org/10.33210/ca.v10i1.356>
- Ramos, J. R., Del Águila, V., & Bazalar, A. (2017). ESTADÍSTICA BÁSICA PARA LOS NEGOCIOS (1st ed.). Universidad de Lima. Fondo Editorial. <https://hdl.handle.net/20.500.12724/10771>
- Robalino-Latorre, M. C., Ramírez-Klinger, W. N., Guadalupe-Copa, R. C., & Cuello-García, S. A. (2023). Aplicación del Método Montecarlo en flujo de potencias a través del Software Octave. *Journal of Economic and Social Science Research*, 3(1), 31–47. <https://doi.org/10.55813/gaea/jessr/v3/n1/60>
- Solano-Gutiérrez, G. A. (2024). La Tecnología en la Educación a Distancia: Revisión de Progresos y Obstáculos a Superar. *Revista Científica Zambos*, 3(2), 48-73. <https://doi.org/10.69484/rcz/v3/n2/17>
- Solano-Gutiérrez, G. A., Núñez-Freire, L. A., Mendoza-Loor, J. J., Choez-Calderón, C. J., & Montaño-Cabezas, L. J. (2023). *Evolución del Computador: desde el ABC de su Arquitectura hasta la Construcción de una PC Gamer*. Editorial Grupo AEA. <https://doi.org/10.55813/eqaea.1.2022.24>
- Vera, J., Castaño, R., & Torres, Y. (2018). *Fundamentos de metodología de la investigación científica* (Grupo Compás). <http://142.93.18.15:8080/jspui/bitstream/123456789/274/3/libro.pdf>

# RESUMEN

Este libro refleja el trabajo realizado bajo investigación entre docentes investigadores con el afán de que sea útil al lector, el uso de predicciones al momento de entrenar un algoritmo clasificado de texto en procesamiento de lenguaje natural (PLN) basado en machine learning. Conformado de 4 capítulos con la utilidad para el inicio al mundo de la IA de la rama de procesamiento de lenguaje natural con Python en machine learning. El Capítulo 1 menciona conceptos y la evolución de las diferentes ramas de conocimiento que abarca la inteligencia artificial (AI), el entendimiento del NLP, machine learning, tipos de aprendizaje para resolver problemas como el supervisado, no supervisado y refuerzo. Capítulo 2 se profundiza el NLP conociendo los contenidos básicos de clasificación como: Las técnicas y diseño de LSTM, tokeización, stopword, lematización, bag of Word (part of speech tagging). Capítulo 3 es la estructuración de este capítulo el conocer las definiciones de los modelos de aprendizaje supervisado que son útiles en NLP orientado a la clasificación de texto. Capítulo 4 un caso de predicción o grado de asertividad del modelamiento de un algoritmo, la intención es demostrar la utilización de un modelo y varias técnicas aplicando NLP basado en machine learning.

**Palabras Clave:** Procesamiento de lenguaje natural, modelo, predicción, aprendizaje supervisado, inteligencia artificial.

## Abstract

This book reflects the work done under research between teachers and researchers with the aim of being useful to the reader, the use of predictions when training a classified text algorithm in natural language processing (NLP) based on machine learning. Conformed of 4 chapters with the usefulness for the beginning to the world of the AI of the branch of natural language processing with Python in machine learning. Chapter 1 mentions concepts and the evolution of the different branches of knowledge that encompasses artificial intelligence (AI), the understanding of NLP, machine learning, types of learning to solve problems such as supervised, unsupervised and reinforcement. Chapter 2 deepens the NLP knowing the basic classification contents such as: LSTM techniques and design, tokeization, stopword, lemmatization, bag of Word (part of speech tagging). Chapter 3 is the structuring of this chapter to know the definitions of supervised learning models that are useful in NLP oriented to text classification. Chapter 4 is a case of prediction or degree of assertiveness of an algorithm modeling, the intention is to demonstrate the use of a model and several techniques applying NLP based on machine learning.

**Keywords:** Natural language processing, model, prediction, supervised learning, artificial intelligence.



<http://www.editorialgrupo-aea.com>



[Editorial Grupo AeA](#)



[editorialgrupoaea](#)



[Editorial Grupo AEA](#)

ISBN: 978-9942-651-43-3



9 789942 651433