UNIVERSIDAD TECNOLÓGICA DE LA MIXTECA

Identificación de nombres confusos de medicamentos

por simetría fonética y ortográfica mediante algoritmos de Aprendizaje Computacional

Documento de tesis que presenta:

Miguel Ángel Uribe Matus

Para obtener el título de:

Ingeniero en Computación

Director de tesis:

Dr. Christian Eduardo Millán Hernández

Codirector de tesis:

Dr. René Arnulfo García Hernández

H. Cd. Huajuapan de León, Oaxaca. Abril de 2024

Contenido

Dedicatoria

El esfuerzo de 5 años de estudios se ven reflejados en este trabajo de investigación, es un precedente de mi vida universitaria, mi dedicación y mi carrera profesional, así mismo este trabajo jamás se hubiese podido realizar sin todo el apoyo recibido por mis amigos, mis allegados y familiares.

Agradezco a mi director por confiar en mí para llevar a cabo la tesis, lo que alguna vez se platicó como una propuesta interesante siguió un desarrollo hasta converger en el trabajo presente. Agradezco a mis amigos por haberme dado la oportunidad de crear lazos significativos con ellos, por impulsarme siempre que lo necesité y por estar conmigo todos estos años. Agradezco a mi familia por darme su apoyo incondicional fueron la razón por la que decidí iniciar una carrera, es infinito el aprecio que les tengo.

Esta tesis está dedicada a mí mismo. Gracias a todos por contribuir en mi formación y ayudarme a ser quien soy ahora.

Agradecimientos

Primero, quiero agradecer a mi asesor, Christian Eduardo Millán Hernández, por su guía experta, paciencia y dedicación a lo largo de este proyecto. Sus conocimientos y orientaciones fueron fundamentales para dar forma a esta investigación y superar los obstáculos encontrados en el camino.

También quiero agradecer a la Universidad Tecnológica de la Mixteca por brindarme los recursos necesarios para llevar a cabo este estudio y por su constante apoyo académico a lo largo de todos mis años de estudio.

Agradezco sinceramente a todos los participantes que colaboraron en este estudio. Muchas gracias a Luis René Morales Velasco, Austin López Ituarte, Melissa Alexandra Bello Cruz y Octavio Martínez Hernández por su contribución, fue esencial para la recopilación de datos y la realización de análisis significativos.

Mi más profundo agradecimiento a mi familia y amigos por su apoyo incondicional, comprensión y ánimo durante todo este proceso. Sus palabras de aliento fueron un motor constante para avanzar en este camino.

Resumen

Dada la amplia variedad de fármacos disponibles en el mercado, surge la posibilidad de situaciones de confusión entre medicamentos, ya sea por sus similitudes visuales o por las similitudes fonéticas en sus nombres. Estos pares de nombres de medicamentos, que son propensos a ser confundidos debido a dichas similitudes, se denominan pares LASA. En trabajos previos, se han empleado algoritmos de aprendizaje computacional para identificar pares LASA, utilizando dos enfoques distintos: uno *a priori*, que busca determinar un valor de similitud entre dos nombres, y otro *a posteriori*, que clasifica si un par de nombres es confuso o no.

Esta tesis propone el uso de algoritmos de aprendizaje computacional supervisados para mejorar la identificación de pares LASA, abordando ambos enfoques, tanto el *a priori* como el *a posteriori*, en tres bases de datos diferentes que contienen medicamentos de Estados Unidos, España y Brasil. Para lograr este objetivo, se entrenaron modelos utilizando ocho algoritmos de aprendizaje computacional supervisado y se aplicaron ocho técnicas de balanceo de datos para cada una de las bases de datos mencionadas.

Capítulo 1 Introducción

Una de las principales preocupaciones de la Organización Mundial de la Salud es mejorar la seguridad del paciente, incluido la prevención de errores de medicación. A nivel internacional, cada año se registran decenas de denominaciones de fármacos (Rocco & Garrido, 2017). Por consecuencia, si se considera la cantidad de medicamentos existentes en el mercado, es posible que ocurran similitudes visuales o en la pronunciación de los nombres, lo que puede causar potenciales errores de confusión de nombres durante la administración y uso de medicamentos.

Los medicamentos poseen dos nombres, el nombre de marca y el nombre genérico. El nombre de marca está asociado a la farmacéutica que desarrolla el medicamento o poseen la patente; en cambio el nombre genérico está asociado a los ingredientes activos que componen al medicamento (Australia, 2022). El nombre genérico es establecido por alguna institución reguladora con la finalidad de indicar el uso terapéutico. En contraste, un nombre de marca es elegido por la farmacéutica con fines de mercadotecnia y está sujeto a la aprobación de las leyes regulatorias de cada país para evitar la confusión de nombres (FDA, 2023a).

Dado el problema que representan los errores de medicación por confusión de nombres, existen Institutos de salud pública que analizan las incidencias de casos de errores de medicación con la finalidad de prevenir futuros casos. En Estados Unidos, el Instituto para el Uso Seguro de los Medicamentos (ISMP por sus siglas en inglés, *Institute for Safe Medication Practices*) en colaboración con la Administración de Alimentos y Medicamentos (FDA, por sus siglas en inglés, *Food And Drug Administration*) ha recopilado casos de errores de medicación por confusión de nombres. El ISMP tiene instituciones filiales a nivel internacional, entre ellos se encuentras la comisión española y la comisión brasileña. En cada país, el ISMP a través de un sitio web reporta una lista de pares de nombres de medicamentos confusos (ISMP, 2023a). Los nombres de medicamentos potenciales a confundirse reciben la clasificación de LASA (LASA, por sus siglas en inglés *Look-alike Sound-alike*) (Abdellatif et al., 2007)

Para prevenir la aparición de nuevos pares LASA, un medicamento nuevo debe pasar por varios procesos antes de llegar al mercado ante una autoridad regulatoria. En uno de estos procesos se evalúa el nombre propuesto del medicamento (FDA, 2023a), donde se determina si el nombre propuesto cumple con las normas del instituto regulador, y no pueda ser potencialmente confundido por algún otro nombre de medicamento ya existente en el mercado (Pfizer, 2023).

En Estados Unidos, la FDA evalúa los nombres de medicamentos con el apoyo de la herramienta de Análisis Computacional Ortográfico y Fonético (POCA, por sus siglas en inglés de *Phonetic and Orthographic Computer Analysis*). El software POCA implementa algoritmos de correspondencia de cadenas para comparar la propuesta del nombre de medicamento contra una base de datos de nombres registrados, dichas fuentes de datos reciben actualizaciones y consideran incluso otros nombres que también están en proceso de revisión paralelamente evaluado (FDA, 2023b).

El uso de la computación ha mostrado una evolución en el poder de cálculo, capacidad de almacenamiento y procesamiento de cantidades inmensas de datos. Como resultado, la cantidad

de tareas que pueden automatizarse se ha expandido, abriendo nuevas posibilidades gracias a la inteligencia artificial. En la industria se utiliza el aprendizaje computacional para la automatización de tareas que incluyen: la clasificación de documentos, análisis de textos, toma de decisiones y predicciones; los algoritmos y modelos resultantes dependen del área y la tarea asignada (M. I. Jordan & Mitchell, 2015).

En el sector salud se destaca el uso del aprendizaje computacional como una herramienta de apoyo para los profesionales de la salud y expertos del área. Sus aplicaciones comprenden tanto en la detección y diagnóstico de enfermedades, como en la asistencia de toma de decisiones para tratamientos y mejoras de los procesos de salud. El impacto del uso de modelos de aprendizaje computacional se ve reflejado en un ahorro en recursos humanos, económicos y de tiempo (Sidey-Gibbons & Sidey-Gibbons, 2019). En particular para esta tesis, el área de interés corresponde a la aplicación de algoritmos de aprendizaje computacional para la identificación de nombres de medicamentos que entren en la categoría LASA.

Dentro de la literatura, resaltan dos enfoques para la identificación de pares confusos, utilizando aprendizaje computacional y algoritmos de correspondencia de cadenas, los cuales se utilizan en dos momentos o situaciones distintas. El primer enfoque es utilizado a priori a la aprobación del medicamento, en donde se requiere determinar un valor de parecido entre el nombre propuesto y una lista de nombres registrados, para que en una serie de pruebas adicionales se determine si el nombre es aceptado o rechazado (Kondrak & Dorr, 2006a; Lambert et al., 1999a; Millán-Hernández, García-Hernández, & Ledeneva, 2019a). El segundo enfoque es implementado a posteriori de la aprobación, cuando el medicamento está a disposición de los usuarios. En este caso se requiere que los sistemas de administración y dispensación de medicamentos envíe advertencias a los usuarios cuando existe el potencial error de confusión de un medicamento recetado con otros nombres de medicamentos (Chen et al., 2011a; Lambert et al., 2004a).

En (Lambert et al., 1999a) se implementó un modelo de regresión logística para predecir si dos nombres de medicamentos pueden o no formar un par LASA. La primera fase del experimento consistió en implementar veintidós algoritmos de correspondencia de cadenas, calculando el valor de parecido de una base de datos con 2254 nombres de medicamentos (incluyendo casos ya registrados de confusión y nombres únicos con la posibilidad de encontrar nuevos casos). El resultado de esta investigación es un modelo de regresión logística para la clasificación de pares de nombres confusos que depende de tres medidas basadas en similitud ortográfica (Trigram-2b), distancia ortográfica (Distancia de Edición Normalizada) y distancia fonética (Editex). Reportó una sensibilidad del 93.7%, una especificidad de 95.9% y una precisión de 94.8%.

En 2006, (Kondrak & Dorr, 2006a) realizó una investigación en la cual propone dos nuevas medidas de correspondencia de cadenas. A su vez realizó una evaluación de la sensibilidad de un conjunto de medidas (incluyendo las medidas propuestas y una medida que calcula el promedio del resultado de: Prefix, NED, BI-SIM y Aline). La investigación se llevó a cabo utilizando una base de datos la cual consta de 363 pares confusos y 582 nombres únicos; a su vez se calculan los valores de similitud para crear una lista ordenada, donde las primeras

posiciones son ocupadas por los pares de nombres de medicamentos con mayor potencial a pertenecer a la categoría LASA. La forma en que se evalúan los resultados consiste en comparar la sensibilidad, con un corte de umbral para las primeras diez posiciones de la lista generada. Los resultados muestran un modelo lineal que combina cuatro medidas: Prefix, Aline, Bisim y NED, con una mejora para la recuperación de pares de nombres confusos; obteniendo una exhaustividad/recuperación del 85% para las primeras 10 posiciones de la lista. El modelo lineal al utilizarse como valor de similitud para la recuperación de pares de nombres confusos apunta a un mejor desempeño que el uso de las medidas de forma individuales extraídos en (Lambert et al., 1999b).

En (Millán-Hernández, García-Hernández, & Ledeneva, 2019a) se utilizó un método basado en una regresión logística optimizada mediante un algoritmo genético (OLRM por sus siglas en inglés, *Optimized Logistic Regression Method*), donde se utiliza la salida de la función sigmoide como valor de similitud entre pares de nombres de medicamento. El experimento se realizó con una base de datos la cual consta de 630 nombres únicos de medicamentos, 858 pares de medicamentos confusos y un total de 396,000 pares de medicamentos (incluyendo pares confusos y no confusos). El mejor modelo obtenido de OLRM utiliza 21 medidas de correspondencia de cadenas. Los resultados de la evaluación muestran que OLRM21 y la regresión logística de 21 medidas supera en rendimiento a la regresión de 3 medidas propuestas por Lambert con una significancia estadística del 95% en el proceso de aprendizaje sobre el proceso de entrenamiento.

En (Lambert et al., 2004a) se realizó un prototipo de un sistema para la comparación multi-atributo entre medicamentos. La interacción con el sistema se realiza mediante una consulta realizada por el usuario, ingresando datos como el nombre del medicamento, dosis de administración, concentración. El sistema regresa una lista ordenada de los 50 medicamentos que más posibilidades de confusión tienen en comparación de la entrada del usuario. Para generar la lista de casos de similitud el sistema utiliza una regresión lineal con múltiples medidas de similitud para realizar la predicción y clasificación entre medicamentos. La evaluación del prototipo demostró un rendimiento, en términos de la sensibilidad, una recuperación variada que oscila entre un 40% a 60% de los casos relevantes dentro de la lista ordenada. Esto quiere decir que el usuario debe buscar más allá de las 50 posiciones para encontrar todos los casos relevantes de confusión.

En (Chen et al., 2011a) se propuso un nuevo sistema de dispensación con alerta de pares LASA utilizando una base datos con 915 nombres únicos de medicamento y 7862 pares de medicamentos de los cuales 3931 son pares LASA (equivalente al 50% del total de los datos). La arquitectura del sistema utiliza dos conjuntos de datos, el primero el cual contiene los casos de error de nombres de medicamentos y el segundo la base de datos de cada medicamento (color, tamaño, ubicación dentro de la farmacia, entre otros). Los nombres son utilizados para el entrenamiento de una regresión logística y un árbol de decisiones mientras la base de datos de medicamentos es utilizada para generar una matriz de disimilitud. La finalidad de la investigación es generar un sistema para evitar los errores de dispensación dando advertencias al usuario acerca de similitudes de nombre o similitudes físicas o de espacio.

En conclusión, existe la viabilidad del uso de aprendizaje computacional en la identificación de nombres de medicamentos confusos. En el enfoque a priori se requiere predecir un valor numérico que indica el potencial de error antes de salir al mercado, por lo que un valor de similitud genera una lista ordenada y se puede determinar si entre un par de medicamentos con valores altos sea necesario realizar pruebas adicionales con el nombre propuesto del medicamento. Mientras que, en el enfoque a posteriori es necesario predecir si el medicamento es potencialmente confuso con otro. La literatura también señala que utilizar medidas de correspondencia de cadenas como atributo fundamental para el entrenamiento de modelos de aprendizaje computacional otorga la capacidad necesaria para discernir similitudes entre los nombres de medicamento.

1.1 Planteamiento del problema

Los métodos de aprendizaje computacional pueden ser vistos como la exploración de un conjunto de posibles soluciones destinadas a descubrir una función adecuada. Estos métodos se clasifican en cuatro tipos: algoritmos supervisados, no supervisados, semi-supervisado y aprendizaje de refuerzo. En el caso de los algoritmos de aprendizaje supervisado, se puede identificar dos categorías principales: de clasificación (o etiquetado), y de regresión los cuales determinan un valor de similitud entre objetos (M. I. Jordan & Mitchell, 2015).

En este contexto, la implementación de aprendizaje supervisado ha permitido atacar el problema de la detección de nombres de medicamento confusos. De acuerdo con los trabajos previos, se utilizan algoritmos de clasificación para indicar si un par es confuso o no (Lambert et al., 1999a). En (Chen et al., 2011a) se utilizan algoritmos de regresión para determinar el valor de similitud entre dos nombres de medicamentos. En el caso de (Kondrak & Dorr, 2006a) no utiliza aprendizaje computacional, pero el método utiliza una combinación lineal de medidas que obtiene un número que representa la similitud. Por último, en (Millán-Hernández, García-Hernández, & Ledeneva, 2019a) si se utiliza aprendizaje computacional en el método de OLRM, pero al final se genera un valor numérico y no una etiqueta.

La diferencia de enfoques está estrechamente relacionado a los procesos por los que un medicamento debe pasar para durante el proceso de aprobación mencionado en la sección anterior. *Grosso* modo se puede referir a los trabajos de Chen y de Lambert como clasificadores binarios mientras, los trabajos de Millán y Kondrak se pueden percibir como una regresión. No obstante, en los trabajos revisados en la literatura no se menciona implementaciones de otros algoritmos de aprendizaje computacional diferentes a la regresión logística como modelo para la identificación, excepto por el trabajo de Chen quién utilizó en conjunto un modelo de árboles de decisión. Por lo tanto, se plantea la siguiente pregunta investigación:

¿Es posible mejorar la identificación de nombres confusos de medicamento por simetría fonética y ortográfica mediante la implementación de algoritmos de Aprendizaje Computacional diferentes de los propuestos en los trabajos previos?

1.2 Justificación

Dada la gran cantidad de fármacos existentes en el mercado, se puede dar lugar a situaciones de confusión entre nombres de medicamentos, tanto por sus similitudes visuales como por sus similitudes fonéticas. Esta problemática puede desencadenar potenciales errores durante la administración de medicamentos, lo que representa un riesgo significativo para la seguridad y bienestar de los pacientes. El problema de los medicamentos LASA (*Look-alike Sound-alike*) es especialmente delicado, ya que las confusiones en su prescripción, dispensación o administración pueden tener graves repercusiones en la salud de las personas.(Abdellatif et al., 2007).

En Estados Unidos, uno de cada cuatro errores de medicación es atribuidos a similitudes fonéticas u ortográficas entre los nombres de los medicamentos (Emmerton & Rizk, 2012). Por lo que es de alto interés para los organismos de salud prevenir estos errores presentes y a futuro. Por lo anterior, el ISMP en afiliación con FDA han realizado una lista de nombres de medicamentos que entran en la categoría LASA, no obstante, la lista se ha generado a través de reportes de casos donde ya hubo errores de medicación por confusión de los nombres (ISMP, 2023b).

Las listas de medicación no son exclusivas de Estados Unidos, El ISMP también tiene comisiones en otros países. Por lo tanto, existen diferentes listas de pares de medicamentos confusos de diferentes países alrededor del mundo; entre los cuáles se incluye ISMP España e ISMP Brasil (ISMP Brasil, 2014; ISMP España, 2020). En México los errores de medicación por nombres en la categoría LASA son un problema por tratar considerados en el plan Nacional de Desarrollo 2013-2018 que en su meta “México Incluyente” objetivo 2.3.4 establece “Garantizar el acceso efectivo a servicios de salud y de calidad” en el cual se contemplan acciones para tratar posibles errores de medicación con nombres de medicamento que entren en la categoría LASA (Rocío et al., 2018).

En el planteamiento del problema se menciona en el trabajo de (Lambert et al., 1999b) se plantea desarrollar una herramienta para la identificación de pares de nombres confusos durante el proceso de dispensación. A su vez en los trabajos de (Millán-Hernández et al., 2020b; Millán-Hernández, García-Hernández, & Ledeneva, 2019b) se plantea el uso del aprendizaje computacional para generar un valor de similitud entre nombres de medicamentos. Una herramienta con dichas características es principalmente útil para las farmacéuticas que buscan liberar nuevos medicamentos en el mercado sin convertirse en nombres potencialmente confusos con otros ya existentes.

Con lo anterior establecido, la posibilidad de apoyar a la seguridad del paciente utilizando el aprendizaje computacional, generando ahorro de horas hombre que hubiesen sido invertidas en revisar individualmente nombres de medicamento, lo que también equivale a un ahorro económico. Aunado a lo anterior, las ventajas del aprendizaje computacional en la identificación de pares de medicamento LASA asisten en la prevención de errores de medicación lo que cual se traduce como una mejor atención médica.

Es importante mencionar que al utilizar el aprendizaje computacional no se busca reemplazar al personal encargado de las pruebas, en su lugar se busca definir las bases para la mejora de una herramienta que pueda auxiliar en las actividades del proceso.

Por lo anterior, en esta tesis se busca comparar el rendimiento de diversos algoritmos de aprendizaje computacional para la asistencia de identificación de pares de nombres de medicamentos que entren en la categoría LASA. Se buscará la implementación de los algoritmos basados en regresión y de algoritmos basados en clasificación realizando un entrenamiento en diferentes bases de datos de distintos idiomas aplicando técnicas de balanceo de datos para mejorar la recuperación de pares LASA.

1.3 Hipótesis

H_i: La implementación de un de balanceo de datos para el entrenamiento de los modelos de aprendizaje computacional mejorar la clasificación y ***predicción de valor de similitud*** *de pares de nombres confusos de medicamento por simetría fonética y ortográfica.*

1.4 Objetivos

Mejorar la clasificación y predicción de valor de similitud de pares de nombres confusos de medicamento por simetría fonética y ortográfica mediante la implementación algoritmos de Aprendizaje Computacional diferentes de los propuestos en los trabajos previos.

1.4.1 Objetivos específicos

Para lograr el objetivo general es necesario llevar a cabo las siguientes actividades:

1. Investigar acerca de algoritmos de aprendizaje computacional utilizados para clasificación de nombres de medicamentos.
2. Investigar acerca de las medidas utilizadas para determinar similitud entre nombres de medicamentos.
3. Establecer los algoritmos de aprendizaje computacional a implementar para la experimentación.
4. Establecer las técnicas de balanceo a utilizar para el entrenamiento de modelos.
5. Realizar el entrenamiento de los modelos y evaluar su desempeño.
6. Comparar los resultados obtenidos de los algoritmos de aprendizaje computacional.

1.5 Metas

1. Elaboración de un reporte sobre algoritmos de aprendizaje computacional para la clasificación de nombres de medicamentos confusos.
2. Elaboración de un reporte de las medidas basadas en correspondencia de cadenas utilizadas para determinar similitud entre nombres de medicamentos.
3. Elaboración de un reporte de las técnicas de balanceo seleccionados para la modificación del conjunto de datos.
4. Elaboración de un reporte sobre los algoritmos de aprendizaje computacional seleccionados.
5. Análisis de resultados obtenidos de los algoritmos de clasificación y regresión. Se seleccionará a los de mejor resultado de cada grupo y al peor de cada grupo.
6. Comparación de resultados de los modelos obtenidos de las distintas técnicas de balanceo de datos en distintas bases de datos de pares LASA.
7. Elaboración del documento de tesis.

1.6 Limitaciones

1. Los conjuntos de datos únicamente abarcan tres idiomas: español, inglés y portugués.
2. La evaluación de los resultados del entrenamiento de los algoritmos de aprendizaje computacional depende de una lista de pares de nombres confusos de medicamentos que son obtenidos de reportes en errores de medicación.
3. El tiempo disponible resulta insuficiente para llevar a cabo pruebas adicionales con una mayor cantidad de conjuntos de datos y algoritmos adicionales.
4. El enfoque de los modelos no reemplaza el proceso de toma de decisiones de las entidades regulatorias, sino que opera como una herramienta de asistencia.
5. El trabajo se limita a la aplicación de una biblioteca de algoritmos de aprendizaje computacional para la obtención de modelos y no incluye el desarrollo de un software.

1.7 Estructura de la tesis

El presente trabajo de investigación aborda el uso de estrategias de balanceo de datos y la exploración de algoritmos de aprendizaje computacional supervisado para atacar el problema de los nombres de medicamentos confusos.

El proceso de la investigación descrito en este documento es presentado bajo la siguiente estructura:

El capítulo 2 Marco teórico recopila las herramientas necesarias para la investigación. Dando un contexto preciso de las estrategias de balanceo de datos y los algoritmos de aprendizaje supervisado. Así mismo se informa de los algoritmos de correspondencia de cadenas utilizadas para abordar el desafío de los nombres de medicamentos confusos.

El Capítulo 3 Método, se describe en detalle los pasos para llevar a cabo los experimentos, incluyendo el proceso de recopilación de datos, la preparación de los conjuntos de entrenamiento y prueba, y los parámetros utilizados para la evaluación de los modelos.

El Capítulo 4 resultados, se presentan los hallazgos obtenidos a partir del proceso descrito en el método. Los resultados se analizan críticamente para identificar tendencias y patrones significativos que ayuden a comprender mejor el problema de los nombres de medicamentos confusos y a sugerir posibles soluciones.

Finalmente, Capítulo 5 Conclusiones. se resumen los principales hallazgos de la investigación y se discuten sus implicaciones prácticas y teóricas.

Capítulo 2 Marco Teórico

En este capítulo, se proporciona un contexto detallado sobre los fundamentos teóricos, las herramientas y técnicas utilizadas en esta investigación. Se inicia de la definición de inteligencia artificial y del aprendizaje computacional supervisado. A continuación, se presenta una revisión de los modelos de aprendizaje computacional destacados en la literatura, propuestos para la experimentación en este proyecto. También se describen las medidas de correspondencia de cadenas utilizadas como características durante el entrenamiento de los modelos de aprendizaje computacional. Al final, el capítulo concluye abordando las métricas de evaluación, las técnicas de balanceo y los trabajos relacionados con esta investigación.

2.1. Inteligencia artificial

Previo a la definición de los modelos de aprendizaje supervisado, es necesario establecer el concepto de una máquina con la habilidad de aprender. La inteligencia artificial (IA) se refiere a la simulación de procesos de inteligencia humana mediante la programación de sistemas computacionales. Implica la creación de algoritmos y modelos que permiten a las máquinas realizar tareas que normalmente requerirían inteligencia humana, como el aprendizaje, el razonamiento, la resolución de problemas, el reconocimiento de patrones y la toma de decisiones. Si se hace énfasis en el aprendizaje; la inteligencia artificial ha permitido la automatización de tareas que originalmente son complejas de generalizar o automatizar y para conseguirlo se hace uso de un área de la Inteligencia artificial conocida como aprendizaje computacional (Microsoft Azure, 2023).

2.2 Aprendizaje computacional o aprendizaje automático

El aprendizaje computacional o aprendizaje automático consiste en la extracción de conocimiento a través de datos, un campo de estudio del cual interactúan técnicas de otras áreas como las ciencias computacionales y la estadística. Visto de otro modo, se puede decir que el aprendizaje computacional se asemeja a dar la instrucción a la computadora de que generalice o construya sus propios criterios de decisión a partir de la experiencia que este obtiene de los datos y reconocimiento de patrones (Müller & Guido, 2017).

Las aplicaciones del aprendizaje automático han tomado mayor relevancia y renombre en los últimos años, por ejemplo, la sugerencia de películas en plataformas digitales, productos de compra o la capacidad de reconocer rostros de amigos o conocidos en las fotos. Las aplicaciones de mayor popularidad son los que automatizan trabajos para la toma de decisiones. Para conseguir una automatización o generalización, existen algoritmos realizan el entrenamiento de un modelo a partir de ejemplos o casos conocidos, es decir cuándo ya se tiene la respuesta esperada. Los modelos que realizan este tipo de entrenamiento se dice que manejan un aprendizaje supervisado (Müller & Guido, 2017).

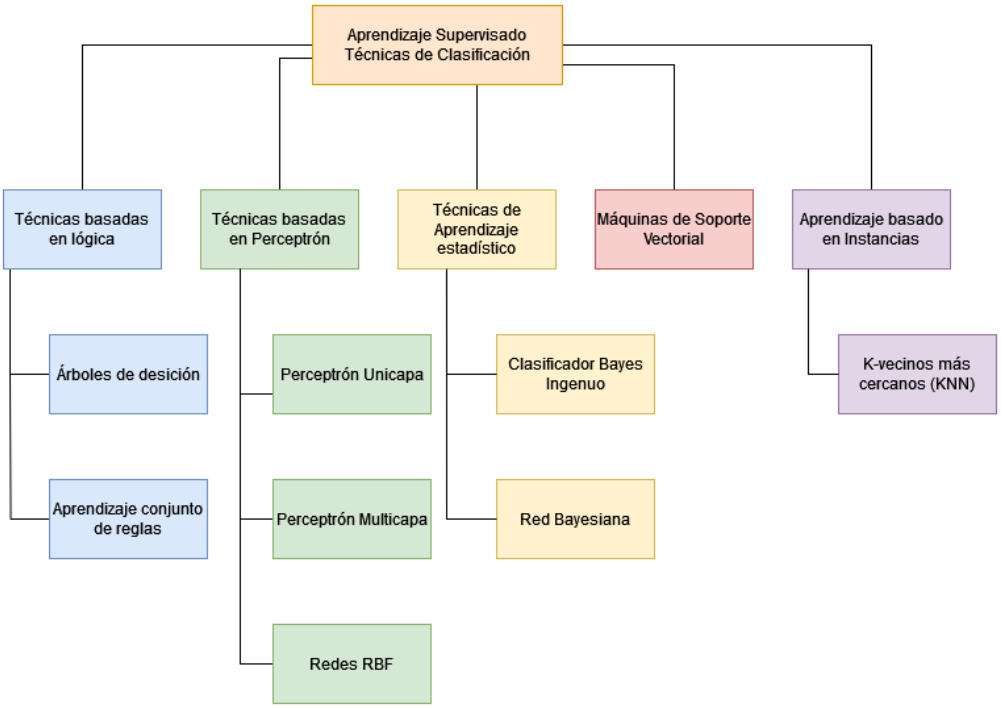
2.3 Aprendizaje computacional supervisado

En el aprendizaje supervisado se realiza el entrenamiento de un modelo donde el usuario proporciona al modelo ejemplos etiquetados, es decir pares de entradas y las salidas deseadas. A través de procesos internos el algoritmo encuentra una forma de producir la salida deseada dada una entrada para lograr una generalización. En concreto, el objetivo del algoritmo es que sea capaz de crear una salida para una entrada que nunca ha visto antes sin ayuda humana.(Müller & Guido, 2017), (Russell et al., 2004).

En la figura 2-1 se muestra clasificación simplificada de los tipos de modelos de aprendizaje supervisado.

Figura 2-1

Clasificación de Técnicas de aprendizaje Supervisado (Soofi & Awan, 2017)



2.2.1 Diferencia entre clasificación y regresión

Es importante establecer que los problemas de aprendizaje supervisado se clasifican en dos tipos, la clasificación y la regresión (Müller & Guido, 2017).

Cuando se habla de clasificación el objetivo del algoritmo es obtener una etiqueta, la cual, es un resultado discreto o categórico que se determina dentro de una lista de opciones iniciales, por ejemplo, si estuviésemos trabajando con la clasificación de frutas entre maduras y no maduras, el objetivo final de un algoritmo de aprendizaje supervisado regresaría la etiqueta “madura” o “no madura” a partir del conjunto de características de cada fruta. En contraste la regresión tiene la tarea de determinar un valor numérico real continuo, siguiendo con el ejemplo

de las frutas, se puede utilizar el valor de la regresión como un estimado de qué tan madura está una fruta a partir de sus características en una escala continua de valores entre 0 y 10 (Müller & Guido, 2017).

En las siguientes secciones se profundiza sobre como los algoritmos pueden ser utilizados tanto en problemas de clasificación como en tareas de regresión.

2.4 Modelos utilizados para identificación de pares confusos de medicamentos

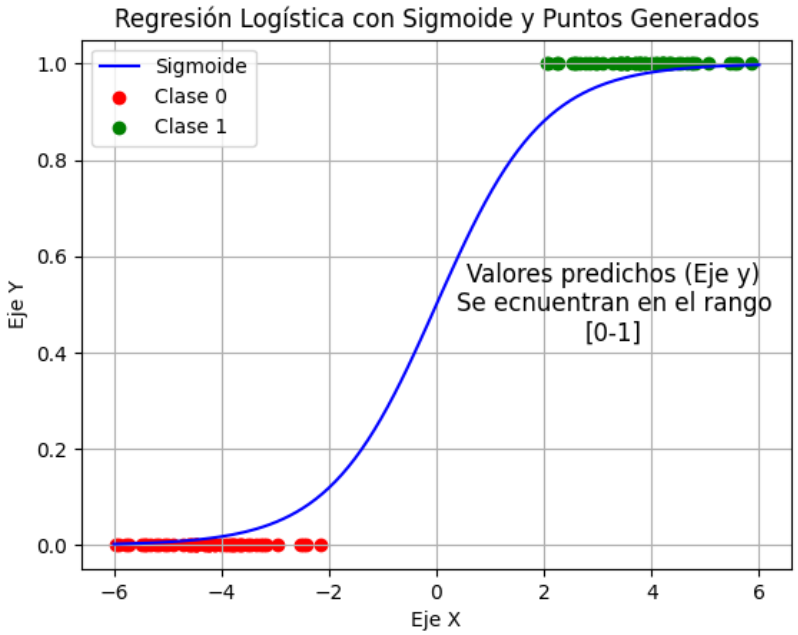
En el estado del arte se han implementado distintos modelos para realizar la clasificación o la regresión en el problema de nombres confusos de medicamentos. A continuación, se hará un resumen de dichos modelos.

2.4.1 Regresión Logística

La regresión logística es un método estadístico utilizado en el campo del aprendizaje automático y la estadística. Aunque el nombre incluye la palabra “regresión”, en realidad se utiliza también para problemas de clasificación (ver figura 2-2). Específicamente, la regresión logística se implementa para predecir la probabilidad de que una instancia pertenezca a una de dos o más categorías, siendo comúnmente utilizada en problemas de clasificación binaria (Hosmer Jr et al., 2013).

Figura 2-2

Función de activación de Regresión logística (Ayush Pant, 2019)



La Regresión Logística es similar a la Regresión Lineal convencional, pero en lugar de regresar la suma del producto de las características con sus respectivos pesos, se aplica un umbral de

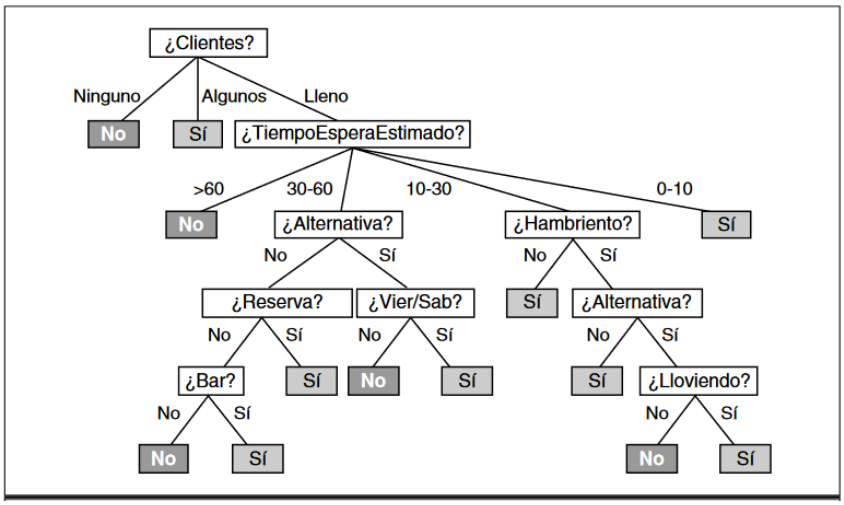
decisión para generar una etiqueta; el umbral de decisión en este caso es: si el valor de la suma es mayor que cero se considera como clase positiva o 1, si el valor de la suma es menor que 0 entonces el valor de la etiqueta es -1, por lo que pertenece a la clase negativa (Müller & Guido, 2017).

2.4.2 Árboles de decisión

Un árbol de decisión es un modelo de aprendizaje automático que toman un objeto o situación como entrada, descrito mediante un conjunto de atributos, y genera una "decisión" como salida, que representa el valor previsto considerando la entrada (ver figura 2-3). Los atributos de entrada pueden ser discretos o continuos, y en este contexto se asume que son discretos. La salida resultante puede ser discreta o continua, y se enfoca en clasificaciones booleanas, donde los ejemplos se clasifican como verdadero (positivo) o falso (negativo). (Russell et al., 2004).

Figura 2-3

Representación de un árbol de decisión (Russell et al., 2004).



Los árboles de decisión esencialmente aprenden al construir jerarquías de preguntas y condicionales de tipo, "*if-else*", las cuales concluyen en una decisión. Para construir un árbol de decisión el algoritmo internamente prueba combinaciones de características para tomar una decisión; a través de ensayo y error generan los nodos. Cada nodo interno del árbol corresponde a una prueba sobre el valor de una propiedad particular, y las ramas que se desprenden del nodo están etiquetadas con los posibles valores de esa propiedad. Una vez formado el árbol, para la predicción de nuevos datos, el algoritmo aplica las condiciones creadas en el árbol para determinar a qué región pertenece el nuevo dato, es decir determinar a qué clase pertenece (Müller & Guido, 2017).

2.5 Modelos de aprendizaje propuestos para el experimento

En el presente trabajo, se procederá a realizar una síntesis sobre los algoritmos de aprendizaje computacional diseñados con el propósito de entrenar modelos de clasificación y regresión destinados a abordar el desafío de discernir entre pares de nombres de medicamentos que

puedan generar confusión por simetría fonética u simetría ortográfica (Pares de nombre de medicamento LASA).

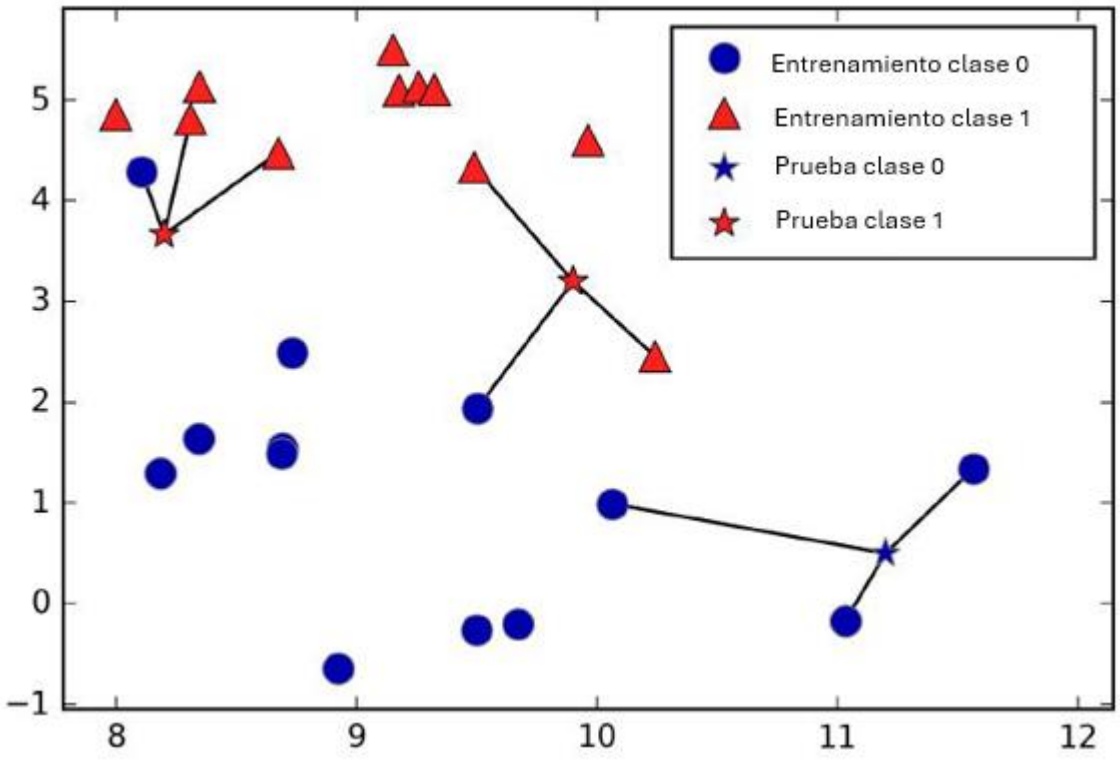
2.5.1 K-vecinos más cercanos (KNN)

K-vecinos más cercanos (KNN por sus siglas en inglés *"K-nearest neighbors*") es considerado uno de los algoritmos de aprendizaje computacional más sencillos. Su aprendizaje implica únicamente almacenar el conjunto de datos en el modelo sin la necesidad de una fase de entrenamiento. La predicción con nuevos datos se efectúa considerando los datos más cercanos al nuevo dato y se genera la etiqueta a partir de las etiquetas de los vecinos más cercanos (Müller & Guido, 2017).

En su versión más simple, utiliza la etiqueta del vecino más cercano asignando el valor de K como 1, incrementar el valor de K incrementa la cantidad de vecinos cercanos para hacer la clasificación, por dar un ejemplo si se consideran tres vecinos o $k = 3$ (ver figura 2-4) el algoritmo tomará la decisión con base a la etiqueta dominante, es decir la etiqueta que más se repita entre los vecinos cercanos (Müller & Guido, 2017).

Figura 2-4

Representación de K vecinos más cercanos (Müller & Guido, 2017).



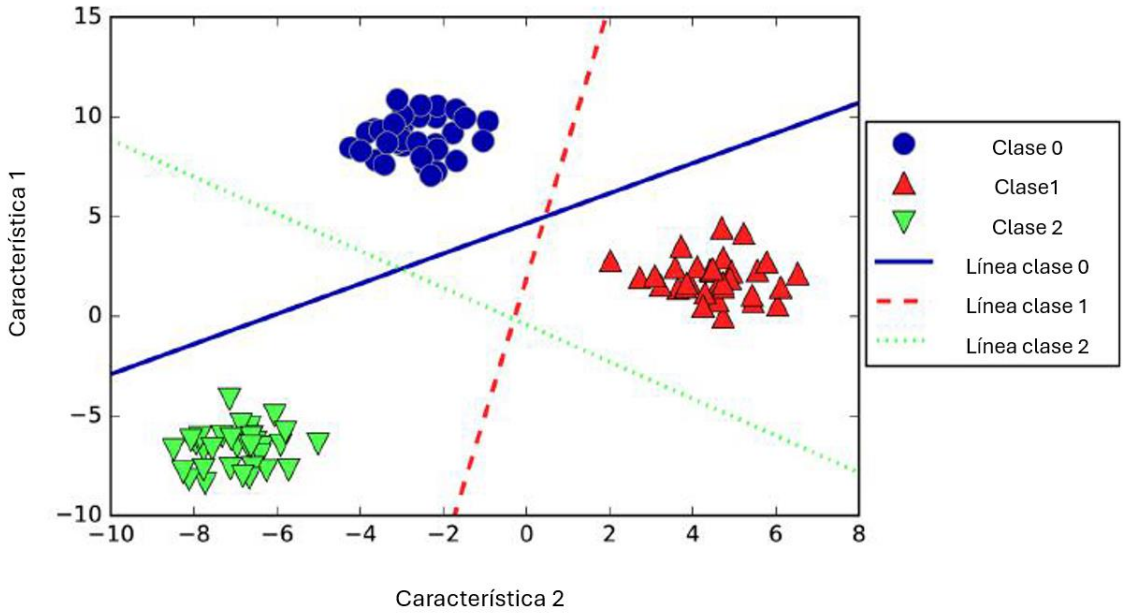
2.5.2 Máquinas de soporte vectorial lineales

Las máquinas de soporte vectorial (SVM por sus siglas en inglés "*Support Vector Machine*") se definen como algoritmos especializados para en la identificación de hiperplanos para encontrar la separación entre clases. Existen dos tipos de máquina de soporte vectorial las

lineales y la no lineales. Las SVM lineales (ver figura 2-5) se enfocan en la separación lineal de la información. En otras palabras, el modelo busca crear una línea dentro del plano para separar las clases, esto con la ayuda del hiperplano.(Müller & Guido, 2017)

Figura 2-5

Separación Lineal en un hiperplano (Müller & Guido, 2017).



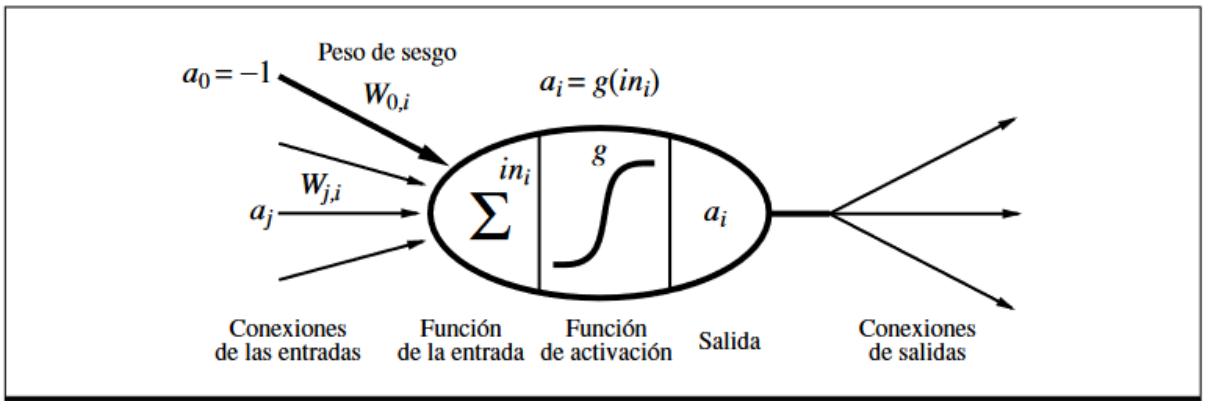
2.5.3 Redes Neuronales

Una neurona (ver figura 2-6) es la unidad básica de procesamiento. Está inspirada en las neuronas del cerebro humano, pero simplificada para el propósito de modelar y resolver problemas mediante computadoras. En términos simples, una neurona toma varias entradas con sus propias características, asigna un peso a cada característica, las suma y luego aplica una función de activación (por ejemplo, una función sigmoide) para producir una salida. Este proceso se repite en capas sucesivas de la red neuronal, permitiendo que la red aprenda patrones y realice tareas específicas, como reconocimiento de patrones, clasificación, o predicción. (Russell et al., 2004)

En una arquitectura neuronal en la que todas las entradas están directamente conectadas a las salidas se conoce como red neuronal de una sola capa o perceptrón. Dado que cada salida es independiente de las demás (cada peso solo influye en una salida), podemos restringir nuestro análisis a perceptrones que cuenten con una única unidad de salida.(Russell et al., 2004).

Figura 2-6

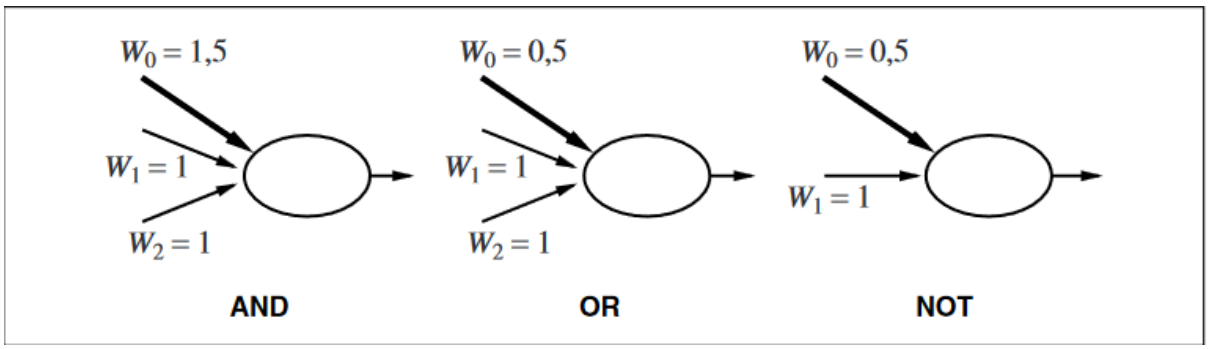
Estructura del Perceptrón (Russell et al., 2004)



Mediante el uso de una función de activación umbral, el perceptrón es capaz de modelar funciones booleanas (ver figura 2-7). Además de las funciones booleanas fundamentales como AND, OR y NOT. El perceptrón tiene la capacidad de representar funciones booleanas un tanto más "sofisticadas" de manera concisa (Russell et al., 2004).

Figura 2-7

Tipos que puede representar un Perceptrón (Russell et al., 2004)



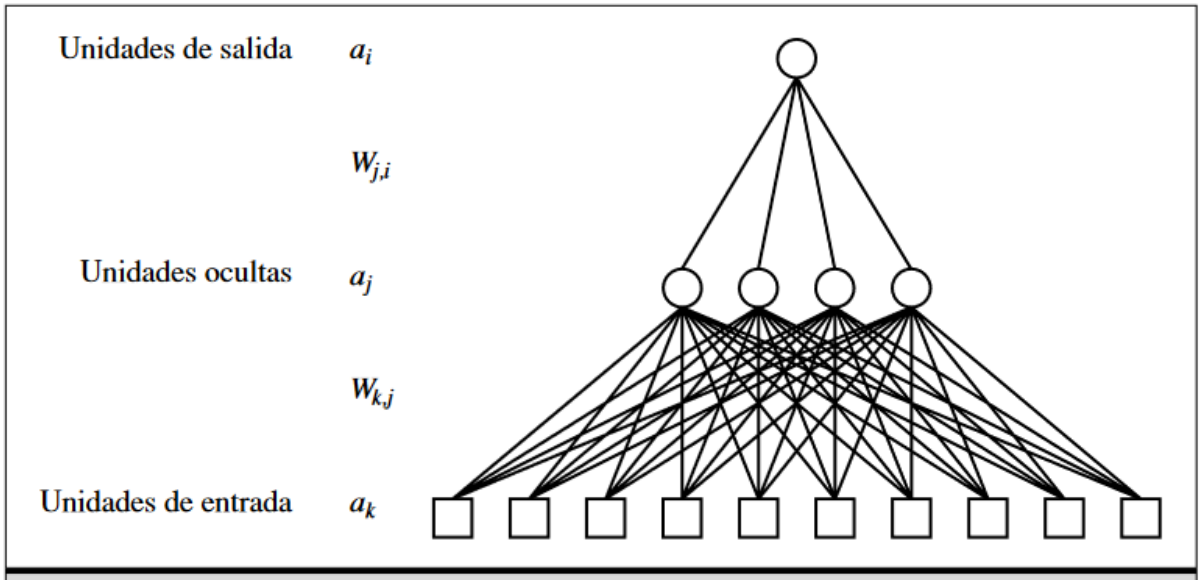
El **Perceptrón Multicapa** (o MLP por sus siglas en inglés *Multi-Layer Perceptron*) es una red a la cual se le han añadido capas ocultas (comúnmente se añade una capa oculta). La ventaja de añadir capas ocultas a la red unicapa (Perceptrón) es la ampliación del espacio de hipótesis que puede representar la red. Una red multicapa simple está compuesta por tres capas, una capa de entrada, la capa oculta y la capa de salida (Russell et al., 2004).

Cada unidad en una capa está conectada a todas las unidades de la capa siguiente, y las conexiones entre las unidades tienen pesos asociados. Cada capa oculta puede contener múltiples unidades que funcionan como nodos o neuronas artificiales (ver figura 2-8). Estas unidades suelen aplicar una función de activación no lineal a una combinación lineal de las salidas de las unidades en la capa anterior (Russell et al., 2004).

Las redes multicapa son poderosas porque pueden aprender representaciones complejas y no lineales de los datos(Russell et al., 2004).

Figura 2-8

Estructura de una red multicapa (Russell et al., 2004)



2.5.4 Ensambles de árboles

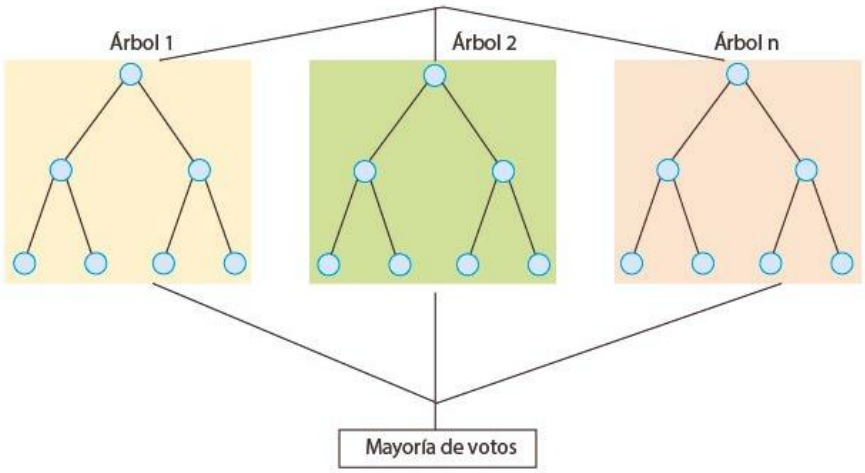
Los ensambles son métodos los cuales combinan múltiples modelos para crear un modelo más robusto. Debido a que los árboles de decisión tienden a caer en un sobreajuste o volverse demasiado complejos debido a su profundidad, el ensamble de árboles ha demostrado ser una alternativa que evita los inconvenientes mencionados. Las técnicas más comunes para el ensamble de árboles es la agrupación o *bagging* presente en los Bosques aleatorios, y el impulso o *boosting* presente en Árboles de Regresión con Impulso del Gradiente (Müller & Guido, 2017).

2.5.4.1 Bosque Aleatorio

Un Bosque Aleatorio es un algoritmo de aprendizaje automático que se basa en la combinación de múltiples árboles de decisión individuales para crear un modelo más robusto y preciso. Cada árbol en el bosque se entrena con una parte aleatoria del conjunto de datos y luego sus resultados se combinan para tomar una decisión final (ver figura 2-9). La idea detrás de los Bosques Aleatorios es aprovechar la diversidad de los árboles para evitar el sobreajuste y mejorar la capacidad de generalización del modelo. Cada árbol en el bosque se entrena en una submuestra de los datos de entrenamiento y, además, en cada división del árbol solo se considera un subconjunto aleatorio de las características. Esto introduce aleatoriedad en el proceso y evita que los árboles se especialicen demasiado en el conjunto de datos de entrenamiento (Müller & Guido, 2017) .

Figura 2-9

Ensamble de Árboles de decisión para un Bosque aleatorio (IBM, 2023)

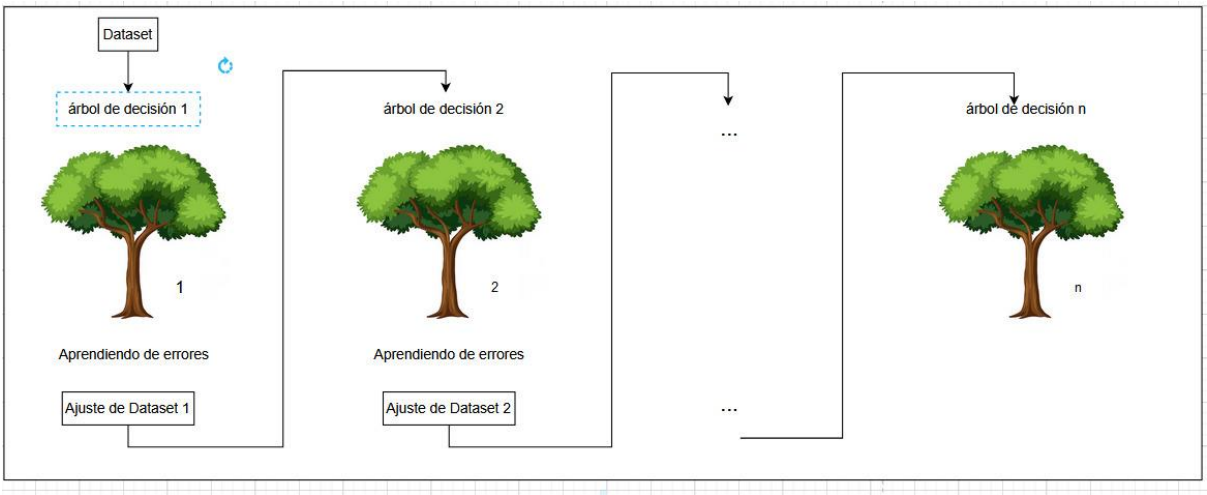


2.5.2 Árboles de Regresión con Impulso del Gradiente

Los Árboles de regresión con impulso del Gradiente (o GBRT por sus siglas en inglés *Gradiant Boosting Regression Tree*) a diferencia del bosque aleatorio, se construyen de manera secuencial, donde cada árbol construido intenta corregir los errores generados por el árbol anterior (ver figura 2-10), por lo tanto, al momento de construirse cada árbol no se maneja la mezcla de los datos ya que así permite hacer correcciones sobre su predecesor (Müller & Guido, 2017).

Figura 2-10

Representación del entrenamiento de un modelo con algoritmo Árboles de Decisión con Impuso del Gradiente



El modelo se genera al combinar modelos simples conocidos como aprendices débiles, donde un árbol es un aprendiz débil que posee pocas ramificaciones. Cada árbol resultante provee

buenas predicciones en partes concretas de los datos, de este modo en cada iteración del algoritmo se añaden más árboles para mejorar el rendimiento del bosque como modelo.(Müller & Guido, 2017).

2.6 Algoritmos de correspondencia de cadenas

Un algoritmo de correspondencia de cadenas es un conjunto de pasos lógicos diseñados para determinar la similitud o equivalencia entre dos cadenas de caracteres. Estos algoritmos son fundamentales en el procesamiento de texto y la búsqueda de patrones.

En la tabla 2-1 y la tabla 2-2 se dará a conocer los algoritmos utilizados para determinar la similitud entre los nombres de medicamentos en este proyecto.

Tabla 2-1

Medidas de similitud fonéticas

Nombre	**Resumen**
Soundex	Soundex es un algoritmo fonético que indexa y busca palabras según su pronunciación en inglés. Asigna códigos alfabéticos y numéricos a palabras basándose en su fonética, agrupando palabras similares para facilitar la búsqueda, incluso si se escriben de forma diferente(Gupta et al., 2014)
Editext	Editext es una medida de distancia fonética que fusiona la distancia de edición con técnicas de agrupación de letras. Define la distancia mediante la relación de recurrencia con funciones r (a, b) y d (a, b). La función r (a, b) devuelve 0 para idénticos, 1 para el mismo grupo, y 2 para otros casos. La función d (a, b) es similar, con excepciones para las letras h y w (a menudo silenciosas). Cuando a = b, d (a, b) es 1 (Gupta et al., 2014).
Phonix	Phonix es un algoritmo fonético que asigna un valor numérico a nombres con la misma pronunciación. Realiza sustituciones fonéticas y aplica reglas como conservar el primer carácter, sustituir por 'v' ciertas letras, ajustar el final del nombre, eliminar vocales y ciertas letras, y reemplazar consonantes con valores numéricos. El código de recuperación resulta de este proceso (Gadd, 1990).

Nota. Las medidas fonéticas utilizan una codificación fija de las palabras y no contempla variaciones en la pronunciación de regiones específicas

Tabla 2-2

Medidas de Similitud Ortográfica

Nombre	Descripción
NED	El algoritmo de distancia de edición, conocido como algoritmo de Levenshtein, mide la similitud entre cadenas de texto mediante operaciones mínimas (Insertar, Cambiar o Eliminar carácter). La distancia normalizada se calcula dividiendo el coste total de edición por la longitud de la cadena más larga. (Gupta et al., 2014), (Kondrak & Dorr, 2006a).
TED	La "Distancia de Edición Cónica" (TED) ajusta penalizaciones según la ubicación de errores, priorizando diferencias iniciales y variando la penalización según la posición. Se destaca que dos errores reciben más penalización que uno solo, independientemente de la posición, pero las cadenas con un solo error se clasifican según la ubicación del error (Gupta et al., 2014).
N-Grama	La técnica de N-gramas analiza texto en fragmentos de N elementos consecutivos. Pueden ser por caracteres o palabras, determinando la longitud de los fragmentos. Por ejemplo, si se considera la cadena “hola” y N igual a 2, los N-gramas resultantes son “ho”, “ol” y “la”. La similitud entre dos cadenas usando N-gramas se establece calculando la similitud entre los N-gramas de ambas cadenas (Grigori Sidorov, 2013).
RLCS	La Relación de la Subsecuencia Común Más Larga (LCSR) se obtiene dividiendo la longitud de la subsecuencia común más larga entre la longitud de la cadena más extensa. Las subsecuencias no requieren caracteres contiguos. Se relaciona estrechamente con la distancia de edición normalizada. Si el costo de sustitución es al menos el doble del de inserción/eliminación, la ecuación LCSR (X, Y) = 1 - NED (X, Y) se cumple para cadenas X e Y de igual longitud (Kondrak & Dorr, 2006a).
N-SIM	N-sim utiliza el algoritmo de N-gramas considerando parámetros como el valor de N, la restricción No-crossing-Links, la longitud del factor de normalización, y el número de símbolos al inicio y final de la cadena, junto con la escala de similitud. Esta escala mide la similitud entre dos N-gramas, evaluando la identidad de letras en posiciones correspondientes y distinguiendo entre capas de similitud, asignando 1 a N-gramas idénticos y 0 a los completamente distinto (Kondrak & Dorr, 2006a).
Soft-Bisim	Soft-Bisim es una medida basada en Bisim (Proveniente de N-sim) la cuál añade una penalización al valor obtenido de Bisim dividiéndolo entre la longitud máxima de la cadena más larga. Fue realizada específicamente como medida de similitud para pares de nombres de medicamentos.(Millán-Hernández, García-Hernández, Ledeneva, et al., 2019).
Prefix	La medida de similitud Prefix devuelve la longitud del prefijo común dividido por la longitud de la cadena más larga. Por ejemplo, el prefijo compartido de "Tobradex" y "Torecan" tiene una longitud de 2 ("To-"), lo que, dividido por la longitud de 8, resulta en 0,25. (Kondrak & Dorr, 2006a)
Omission-Key	Omission-Key se basa en la omisión frecuente de consonantes al escribir, siguiendo un orden específico. La clave de omisión de una palabra sigue este orden inverso para consonantes y luego vocales. La similitud se mide mediante la distancia en una lista ordenada alfabéticamente, considerando la inversa de la similitud (Ulrich Pfeifer et al., 1996).
Skeleton-Key	Skeleton-Key se fundamenta en que las consonantes aportan más información que las vocales. La clave de una palabra incluye su primera letra, las consonantes restantes y, finalmente, las vocales, sin duplicados. La medida de similitud se establece a través de la posición en una lista ordenada alfabéticamente (Ulrich Pfeifer et al., 1996).

2.7 Técnicas y Métricas de evaluación

En esta sección se hará mención y descripción de las métricas que se usarán para evaluar durante la experimentación del proyecto.

2.7.1 Precisión

En el contexto de modelos de aprendizaje computacional, la precisión se refiere a la medida de exactitud en un modelo en sus predicciones. Es una métrica que evalúa la proporción de predicciones correctas en relación con el total de predicciones realizadas por el modelo. La precisión es especialmente útil en problemas de clasificación, donde el objetivo es asignar instancias a diferentes categorías (scikit-learn, 2023).

La fórmula para calcular la precisión es (scikit-learn, 2023):

$$Precisión = \frac{Verdaderos\ Positivos}{Verdaderos\ Positivos + Falsos\ Positivos}$$

Donde:

- Verdaderos Positivos (True Positives) son los casos en los que el modelo predijo correctamente la clase positiva.
- Falsos Positivos (False Positives) son los casos en los que el modelo predijo incorrectamente la clase positiva cuando en realidad debería haber sido negativa.

2.7.2 Exhaustividad

La exhaustividad (también conocida como *recall* o sensibilidad) es una métrica que mide la capacidad de un modelo para identificar correctamente todas las instancias positivas en un conjunto de datos. En otras palabras, la exhaustividad evalúa qué tan bien el modelo evita pasar por alto ejemplos positivos. Es especialmente relevante en problemas donde es crucial identificar todos los casos positivos, incluso a costa de cometer más falsos positivos (scikit-learn, 2023)..

La fórmula para calcular la exhaustividad es (scikit-learn, 2023).:

$$Exhaustividad = \frac{Verdaderos\ Positivos}{Verdaderos\ Positivos + Falsos\ Negativos}$$

Donde:

- Verdaderos Positivos (True Positives) son los casos en los que el modelo predijo correctamente la clase positiva.
- Falsos Negativos (False Negatives) son los casos en los que el modelo predijo incorrectamente la clase negativa cuando en realidad debería haber sido positiva.

Para lograr una alta exhaustividad, el modelo debe minimizar los falsos negativos, asegurándose de que la mayoría de las instancias positivas sean detectadas.

2.7.3 Medida F

La Medida-F, también conocida como F-score es una métrica que combina tanto la precisión como la exhaustividad para proporcionar una medida única del rendimiento del modelo. El F-score se utiliza comúnmente cuando hay un desequilibrio entre las clases en el conjunto de datos y se busca un equilibrio entre la precisión y la exhaustividad (scikit-learn, 2023).

El F-score se calcula utilizando la siguiente fórmula (scikit-learn, 2023):

$$Fscore = 2 * \frac{Precisión * Exhaustividad}{Precisión + Exhaustividad}$$

Donde:

- Precisión es la proporción de verdaderos positivos en relación con el total de instancias positivas predichas.
- Exhaustividad es la proporción de verdaderos positivos en relación con el total de instancias positivas reales.

El F-score proporciona un equilibrio entre la precisión y la exhaustividad, y puede ser particularmente útil cuando se busca una métrica que considere ambos aspectos del rendimiento del modelo (scikit-learn, 2023).

2.7.4 Evaluación por posiciones

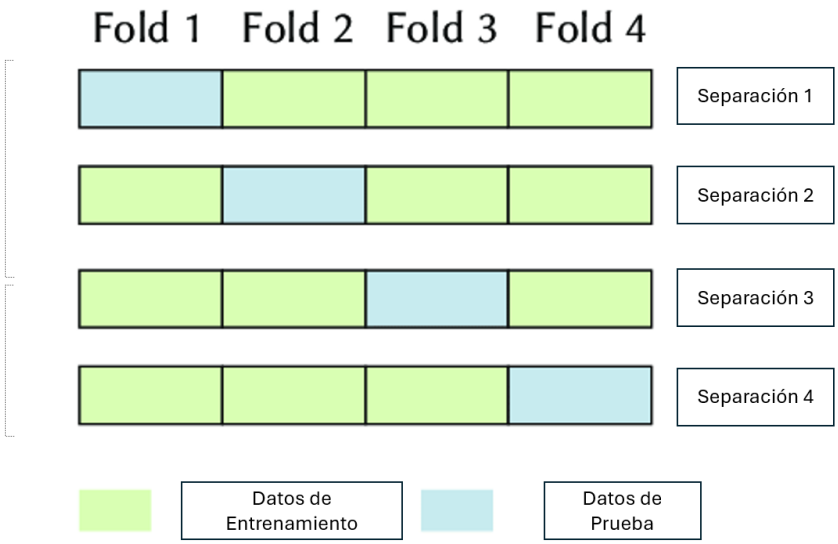
La métrica de Evaluación por posiciones (*Ranking Evaluation* en inglés) se refiere a la evaluación del rendimiento de un modelo en la tarea de ordenar o clasificar elementos en función de su relevancia o valor esperado. Esta métrica es comúnmente utilizada en problemas de recomendación, búsqueda y recuperación de información, donde la calidad de la clasificación de los elementos es fundamental. La evaluación por posiciones busca medir qué tan bien un modelo clasifica los elementos en el orden correcto en comparación con un orden de referencia, como puede ser el ranking manual creado por humanos o el ranking de relevancia. Las métricas utilizadas para evaluar la clasificación de posiciones a menudo se basan en conceptos como la precisión en el top k (cuántos elementos relevantes están presentes en las primeras k posiciones), el MAP (*Mean Average Precision*, que promedia la precisión en todas las posiciones relevantes) y el NDCG (*Normalized Discounted Cumulative Gain*, que considera la relevancia y la posición del elemento en el ranking). La métrica de Evaluación por Posiciones es esencial para medir la eficacia de los modelos en proporcionar resultados relevantes en un orden significativo, lo que es crucial en aplicaciones como la recomendación de productos, la búsqueda de información y el filtrado de contenido (M. Jordan et al., 2006).

2.7.5 Validación Cruzada

La validación cruzada es un método estadístico utilizado para la evaluación generalizada del rendimiento que a su vez resulta ser más estable y minucioso que utilizar una simple división entre un conjunto de prueba y otro de entrenamiento. En la validación cruzada, los datos son divididos repetidamente y con ello varios modelos son entrenados. La versión más comúnmente utilizada de validación cruzada es la *K-Fold*, donde k representa un valor dado por un usuario (por lo general se utiliza el número 5 o el 10). Al ejecutar la validación cruzada con k igual a 5, los datos primero son divididos en 5 partes de igual tamaño llamados *folds*. Posteriormente una secuencia de modelos es entrenada donde el primer modelo utiliza el primero *fold* como conjunto de pruebas y los otros 4 (los *folds* 2 al 5) son utilizados como conjunto de entrenamiento. Para el siguiente modelo el segundo *fold* es utilizado como conjunto de pruebas y el resto como conjunto de entrenamiento. Este proceso se repite por hasta que cada *fold* haya sido utilizado individualmente como conjunto de pruebas (Müller & Guido, 2017).

Figura 2-11

Representación de división por Folds (Müller & Guido, 2017)



2.7.6 Error Cuadrático Medio

El error cuadrático medio (o MSE por sus siglas en inglés de *Mean squared error*) es una medida utilizada en modelos de regresión para evaluar la precisión. El cálculo se realiza mediante la media de los cuadrados de la diferencia entre los valores predichos por el modelo y los valores reales aplicando la siguiente fórmula (Müller & Guido, 2016):

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

Donde:

- MSE representa el error cuadrático medio.
- n es el número de observaciones.
- y_i son los valores observados.
- $\widehat{y_i}$ son los valores predichos.

El MSE penaliza las diferencias altas sobre las pequeñas, es decir si la diferencia entre los valores predichos y los valores reales es alta el MSE acumulará un valor más alto, en contraste las diferencias pequeñas tendrás aportes poco significativos (Müller & Guido, 2016).

2.8 Balanceo de datos

En esta Sección se abordará acerca del desequilibrio en la proporción de datos en un conjunto. A su vez se tratará acerca de las técnicas para balancear la proporción de estos.

2.8.1 Conjunto de datos desbalanceados

Cuando hay una diferencia significativa en la proporción de datos entre clases, se dice que los datos están desbalanceados. En otras palabras, se dice que un conjunto está desbalanceado cuando existe una predominancia de datos de una clase sobre otra. A la clase predominante se le denomina clase mayoritaria; por lo tanto, la clase con menos datos recibe el nombre de clase minoritaria. Al realizar el entrenamiento con un conjunto de datos desbalanceado pueden surgir distintos tipos de problemas como generar un sesgo en el modelo al momento de hacer clasificación de objetos de la clase minoritaria, otro problema de tener un conjunto de datos desbalanceado se ve reflejado en un rendimiento engañoso obteniendo una precisión alta de forma general, pero resultados pobres en su sensibilidad o precisión de la clase minoritaria. Para evitar problemas por un desequilibrio de datos se hace uso de técnicas de balanceo de datos o “resampling” (García Abad Joaquín, 2021).

2.8.2 Técnicas de balanceo o resampling

El balanceo de datos o *resampling* consiste en la modificación de la distribución de los datos en una clase o en ambas. Las soluciones para realizar *resampling* se categorizan en dos grupos, el primer grupo se denomina *undersampling* el cual presenta como solución al desbalanceo la eliminación de instancias u objetos provenientes de la clase mayoritaria. En contraste el segundo grupo de soluciones se llama *oversampling* y consiste en replicar o generar nuevas instancias de la clase minoritaria. A su vez las técnicas de balanceo se pueden categorizar en aquellas que siguen una heurística o un proceso controlado y las técnicas que manejan procesos no controlados o aleatorios, es decir, no heurísticos (García Abad Joaquín, 2021).

2.8.3 Selección aleatoria

La selección aleatoria puede ser utilizada tanto para *undersampling* y *oversampling*. Para realizar *undersampling* se selecciona de manera aleatoria datos pertenecientes a la clase

mayoritaria para igualar a los datos de la clase minoritaria. En contraste cuando se usa la selección aleatoria para *oversampling* se hace una repetición de los datos pertenecientes a la clase minoritaria para que igualar a los de la clase mayoritaria, no obstante, la repetición de datos no fomenta la generalización si no la memorización de los datos (Menardi & Torelli, 2014) .

2.8.4 *Near miss*

El *Near miss* es una técnica de *undersampling* diseñada para estabilizar la distribución de datos al eliminar muestras de la clase mayoritaria. Se enfoca en casos que estuvieron cerca de ser clasificados incorrectamente, pero que finalmente no lo fueron. Estos casos se consideran complicados y se seleccionan con el objetivo de aumentar la robustez del modelo. (referencia de enlace). En este enfoque se selecciona los casos positivos (clase mayoritaria) cuya distancia promedio a las N muestras más cercanas de la clase minoritaria sea la más corta (Inderjeet Mani. & I Zhang, 2003).

2.8.5 *Edited Nearest Neighbour*

Edited Nearest Neighbour o ENN es una técnica de balanceo de datos similar al *Near miss.* Para poder hacer la selección aplica el algoritmo de vecinos más cercanos y edita el conjunto de datos removiendo los vecinos que no encajan dentro apropiadamente dentro del vecindario. Por cada muestra del conjunto de datos a ser reducida, se calculan sus vecinos más cercanos y si el criterio de selección no es cumplido, la muestra es eliminada del conjunto (Wilson, 1972).

2.8.6 *Tomek's Links*

Los enlaces de Tomek o *Tomek's Links*, es un algoritmo para realizar undersampling, el cual se enfoca en identificar enlaces de Tomek, los cuales son pares de instancias cercanas de clases diferentes, es decir si los dos datos son el vecino más cercano uno del otro. Una vez identificados los enlaces de Tomek, se elimina una de las instancias (Por lo general solo la perteneciente a la clase mayoritaria) con la finalidad de mejorar la frontera de decisión entre las clases (Ivan Tomek., 1976).

2.8.7 SMOTE

Synthetic Minority Oversampling Technique o SMOTE es un método de *oversampling* el cual se centra en la creación de datos sintéticos a través de la interpolación de características de datos pertenecientes a la clase minoritaria. Así la proporción de datos entre clases queda equilibrada. El proceso de generación de datos sintéticos consiste en que por cada uno de los elementos pertenecientes a la clase minoritaria se selecciona de forma aleatoria uno de sus vecinos más cercanos. Posteriormente se crea un dato sintético en el espacio de muestra entre el elemento y su vecino seleccionado. El dato sintético contiene una combinación convexa entre ambos elementos (Nitesh V. Chawla et al., 2002).

2.8.8 ADASYN

Adaptive Synthetic o ADASYN es método de *oversampling* que comparte características con SMOTE en la manera en cómo genera un dato sintético, sin embargo, ADASYN se centra en la creación de datos sintéticos considerando la densidad en el espacio de muestra. En otras palabras, crea casos sintéticos en las zonas de menor densidad dentro del espacio de muestra de los datos pertenecientes a la clase minoritaria. Para realizar la tarea el algoritmo primero calcula la densidad de la muestra y determina las zonas de menor densidad. La cantidad de datos sintéticos generados por región es inversamente proporcional a la densidad de cada zona, es decir a menor densidad haya en una región más datos sintéticos serán generados (Haibo He et al., 2008).

2.8.9 *Combination Sampling*

Las técnicas de balanceo *oversampling* como SMOTE al crear datos sintéticos llegan a generar muestras que pueden ser catalogadas como ruido, esto se debe a que durante la interpolación no se hace una distinción entre datos típicos y datos atípicos (o *outliers*). Esto puede ser fácilmente enmendado al utilizar técnicas de limpieza de datos como *Tomek's Links* y *Edited Nearest Neighbour*. Por lo tanto, *Combination Sampling* es el nombre dado al proceso de crear datos sintéticos y eliminar el ruido generado. *SMOTE-Tomek's Link* (Batista et al., 2004) es el nombre que recibe utilizar SMOTE en conjunto a *Tomek's Links*. *SMOTE-Edited Nearest Neighbour* (Gustavo EAPA Batista et al., 2003) es el nombre que recibe utilizar SMOTE en conjunto a *Edited Nearest Neighbour*.

2.9 Estado de Arte

A continuación, se hace un resumen general de las investigaciones previas a esta. El propósito es dar contexto general de qué se ha hecho previamente, señalar la distinción de enfoques y tener un punto de comparación para el alcance del proyecto de investigación.

2.9.1 Enfoque de modelos para generar un valor de similitud

En (Lambert et al., 1999a) se implementó un modelo de regresión logística para predecir si dos nombres de medicamentos pueden o no formar un par LASA. La primera fase del experimento consistió en implementar veintidós algoritmos de correspondencia de cadenas, calculando el valor de parecido de una base de datos con 2254 nombres de medicamentos (incluyendo casos ya registrados de confusión y nombres únicos con la posibilidad de encontrar nuevos casos). En los resultados, el modelo utiliza tres medidas (Distancia de Edición Normalizada, Editex y Trigrama-2b), reporta una sensibilidad del 93.7%, una especificidad de 95.9% y una precisión de 94.8%. El resultado de esta investigación es un modelo para la clasificación de pares de nombres confusos que depende de tres medidas basadas en similitud ortográfica (Trigram-2b), distancia ortográfica (Distancia de Edición Normalizada) y distancia fonética (Editex).

En 2006, (Kondrak & Dorr, 2006a) realizó una investigación en la cual propone dos nuevas medidas de correspondencia de cadenas. A su vez realizó una evaluación de la sensibilidad de un conjunto de medidas (incluyendo las medidas propuestas y una medida que calcula el promedio del resultado de: Prefix, NED, BI-SIM y Aline). En la investigación se utiliza los valores de similitud para crear una lista ordenada, donde las primeras posiciones son ocupadas por los pares de nombres de medicamentos con mayor potencial a pertenecer a la categoría LASA. La forma en que se evalúan los resultados consiste en comparar la exhaustividad, con un corte de umbral para las primeras diez posiciones de la lista generada. Los resultados muestran una medida combinada (el promedio de cuatro medidas: Prefix, Aline, Bisim y NED) con una mejora para la recuperación de pares de nombres confusos; obteniendo una exhaustividad/recuperación del 85% para las primeras 10 posiciones de la lista. La medida combinada al utilizarse como valor de similitud para la recuperación de pares de nombres confusos apunta a un mejor desempeño que el uso de las medidas de forma individuales.

En (Millán-Hernández, García-Hernández, & Ledeneva, 2019a) se utiliza un método basado en una regresión logística optimizada mediante un algoritmo genético (OLRM por sus siglas en inglés, *Optimized Logistic Regression Method*), donde se utiliza la salida en crudo de la predicción como valor de similitud entre pares de nombres de medicamento. El mejor modelo obtenido de OLRM utiliza 21 medidas de correspondencia de cadenas. La evaluación a los que se llegan muestra que la regresión de 21 medidas supera en rendimiento a la regresión de 3 medidas propuestas por Lambert con una significancia estadística del 95% en el proceso de aprendizaje sobre el proceso de entrenamiento.

Posteriormente en (Millán-Hernández et al., 2020a) se retomaría el modelo entrenado en (Millán-Hernández, García-Hernández, & Ledeneva, 2019a) y se evaluaría su desempeño con una base de datos de pares confusos en el idioma español. La finalidad de la investigación era determinar si el modelo ya entrenado podía reutilizarse para otros idiomas o si se debía reentrenar el modelo para una base de datos distintas. Para su experimentación realizó la comparativa entre el rendimiento del modelo OLRM con 21 medidas entrenado con la base de datos en inglés y un modelo OLRM con 25 medidas entrenado con la base de datos en español. Sus resultados muestran que el mejor rendimiento se obtiene al reentrenar el modelo con la base de datos en español (OLRM 25 medidas) obteniendo una medida *F-macro-averaging* del 44.14% en las primeras 5 posiciones.

Por último, en (Vázquez et al., 2020) se desarrolló una regresión simbólica para predecir un valor de similitud mediante algoritmos evolutivos sobre la misma base de datos utilizada en (Millán-Hernández, García-Hernández, & Ledeneva, 2019b). Para la creación de la regresión simbólica fueron necesarias utilizar 12 medidas de correspondencia de cadenas. Sus resultaron mostraron una mejora respecto a (Millán-Hernández, García-Hernández, & Ledeneva, 2019b) alcanzado una medida *F-macro-averaging* de 45.35% en las primeras 4 posiciones.

2.9.2 Enfoque de modelos para la clasificación de pares confusos

En (Lambert et al., 2004a) se realizó un prototipo de un sistema para la comparación multi-atributo entre medicamentos. La interacción con el sistema se realiza mediante una consulta

realizada por el usuario, ingresando datos como el nombre del medicamento, dosis de administración, concentración. El sistema regresa una lista ordenada de los 50 medicamentos que más posibilidades de confusión tienen en comparación del *input* del usuario. Para generar la lista de casos de similitud el sistema utiliza una regresión lineal con múltiples medidas de similitud para realizar la predicción y clasificación entre medicamentos. La evaluación del prototipo demostró un rendimiento, en términos de la recuperación/exhaustividad, una recuperación variada que oscila entre un 40% a 60% de los casos relevantes dentro de la lista ordenada. Esto quiere decir que el usuario debe buscar más allá de las 50 posiciones para encontrar todos los casos relevantes de confusión.

En (Chen et al., 2011a) se propone un nuevo sistema de dispensación utilizando acercamiento híbrido. La arquitectura del sistema utiliza dos conjuntos de datos, el primero el cual contiene los casos de error de nombres de medicamentos y el segundo la base de datos de cada medicamento (Color, tamaño, ubicación dentro de la farmacia, entre otros). Los nombres son utilizados para el entrenamiento de una regresión logística y un árbol de decisiones mientras la base de datos de medicamentos es utilizada para generar una matriz de disimilitud. La finalidad de la investigación es generar un sistema para evitar los errores de dispensación dando advertencias al usuario acerca de similitudes de nombre o similitudes físicas o de espacio.

La tabla 2-3 muestra de forma general las aportaciones de los trabajos relacionados, clasificados según el enfoque del trabajo donde a su vez se muestra el modelo implementado y las métricas obtenidas en dichas investigaciones.

Tabla 2-3

Comparativa de resultados de trabajos relacionados

Enfoque	**Autor**	**Año**	**Medidas de similitud**	**Modelo**	**Métrica**	**Objetivo**
A priori	Lambert	1999	Ned, Editex Trigram2B	Regresión Logística	Precisión del 94.8%	Clasificación
	Kondrak	2006	Ned, Aline, BiSim, Prefix	Utiliza el promedio	Recall del 85% top 10	Valor de similitud
	Millán	2019	21 medidas	Regresión Logística	F. Macro averaging 175.57 (43.89%)	Valor de similitud
	Millán	2020	25 medidas	Regresión Logística	F. Macro averaging 220.71(44.14%)	Valor de similitud
	Vázquez	2020	12 medidas	Regresión Simbólica	F. Macro averaging 181.43 (45.35%)	Valor de similitud
A posteriori	Lambert	2004	Ned, Editex, Trigram2B	Regresión Lineal	Recall del 40% top 10	Valor de similitud
	Chen	2011	Ned	Árboles de decisión Regresión Logística	Precisión entre el 83% y 86%	Clasificación y valor de similitud

Resumen del capítulo

La inteligencia artificial se refiere a la simulación de procesos de inteligencia humana mediante la programación de sistemas computacionales. Una de las ramas de la inteligencia artificial es el aprendizaje computacional o aprendizaje automático, esta rama está enfocada en la automatización de tareas. Los algoritmos de aprendizaje automático se dividen en cuatro tipos: aprendizaje supervisado, no supervisado, semi-supervisado y aprendizaje por reforzamiento. De los cuatro tipos de algoritmos, el aprendizaje supervisado se caracteriza por realizar entrenamientos con etiquetas, es decir, se tiene las entradas y las salidas esperadas y son mayormente utilizadas para hacer una generalización basada en la experiencia y predicciones con nuevos datos.

Los algoritmos de aprendizaje supervisado dependen de dos parámetros para poder realizar el entrenamiento de un modelo: las características de los datos (que son las entradas) y las etiquetas de estos (que son las salidas esperadas). Existen diferentes tipos de algoritmos para el entrenamiento de modelos de aprendizaje supervisado. Cada uno con sus respectivas ventajas como desventajas. No obstante, es importante resaltar que las tareas que puede realizar un modelo de aprendizaje supervisado no solo se limitan a la obtención de una etiqueta el cual es un valor discreto o categórico, También son aptos para determinar un valor continúo conocido como regresión. El modelo será entrenado en función de la tarea asignada ya sea clasificación o regresión.

Para el entrenamiento de modelos dedicados al problema de nombres confusos de medicamentos, en trabajos previos se ha utilizado como características (para los pares de nombres de medicamento) los valores de similitud obtenidos de diferentes algoritmos de correspondencia de cadenas. Un algoritmo de correspondencia de cadenas es un conjunto de pasos lógicos diseñados para determinar la similitud o equivalencia entre dos cadenas de caracteres los cuáles cubren similitudes fonéticas u ortográficas.

Existen varias formas de evaluar el rendimiento de un modelo de clasificación. Entre ellos se destaca la medida F. La medida F es un balance entre las medidas convencionalmente más utilizadas que son la precisión y la exhaustividad; un valor alto indica en la medida F indica que el modelo obtiene buenos resultados en su precisión y su exhaustividad. La precisión evalúa la cantidad de aciertos que obtuvo un modelo para clasificar elementos correctamente. La exhaustividad evalúa la capacidad de un modelo para identificar correctamente todas las instancias positivas en un conjunto de datos. Para un modelo de regresión se opta por una evaluación por posiciones. La evaluación por posiciones evalúa el rendimiento de un modelo en la tarea de ordenar o clasificar elementos en función de su relevancia o valor esperado. Esta métrica es comúnmente utilizada en problemas de recomendación, búsqueda y recuperación de información, como lo es el problema de pares de nombres confusos de medicamentos, donde la calidad de la clasificación de los elementos es fundamental.

Por último, este trabajo considera la proporción de los datos como un problema a tratar y por ende se hace la apuesta por la implementación de técnicas de balanceo de datos. El balanceo de datos o *resampling* como su nombre lo indica, se utiliza para generar una proporción

equitativa entre los datos, es decir que no haya una etiqueta o clase predominante. El balanceo de datos se cataloga en dos tipos, "*undersampling*" y "*oversampling*". El "*undersampling*" consiste en seleccionar elementos de la clase predominante para igualar en proporción a los de la clase minoritaria. En contraste el "*oversampling*" consiste en generar más datos de la clase minoritaria para igualar a los de la clase predominante, por lo general este se lleva a cabo mediante la creación de datos sintéticos. El objetivo de aplicar técnicas de balanceo de datos es mejorar el rendimiento de los modelos para la clasificación y regresión de pares de medicamentos confusos.

Capítulo 3 Método Propuesto

A nivel internacional, cada año se registran decenas de denominaciones de fármacos, por lo que es posible que ocurran similitudes visuales o fonéticas en los nombres, lo que causa potenciales errores de confusión de nombres durante la administración de medicamentos. En los trabajos previos para la identificación de pares LASA, se muestran métodos basados en la implementan un modelo obtenido por una regresión logística, exceptuando en el trabajo de Chen, donde se combina la regresión logística y los árboles de decisión. Por lo que, en esta investigación surge la siguiente interrogante de investigación: *¿Es posible mejorar la identificación de nombres confusos de medicamento por simetría fonética y ortográfica mediante la implementación de algoritmos de Aprendizaje Computacional diferentes de los propuestos en los trabajos previos?*

En este capítulo se detalla el método propuesto para la identificación de pares LASA ya sea desde un enfoque *a priori* o *a posteriori* mediante la implementación de serie de pasos que incluyen el balanceo del conjunto de datos para mejorar el desempeño de modelos obtenidos por diferentes algoritmos de Aprendizaje Computacional.

La Figura 3-1 muestra un diagrama donde se detalla el método a seguir en esta investigación la cual consta de 3 etapas, siendo la primera etapa correspondiente a la creación del conjunto de datos. La segunda etapa la experimentación con técnicas de balanceo y la tercera etapa correspondiente a la evaluación de los modelos para la identificación de pares LASA en ambos enfoques (*a priori* y *a posteriori*).

Etapa 1: Creación del conjunto de datos

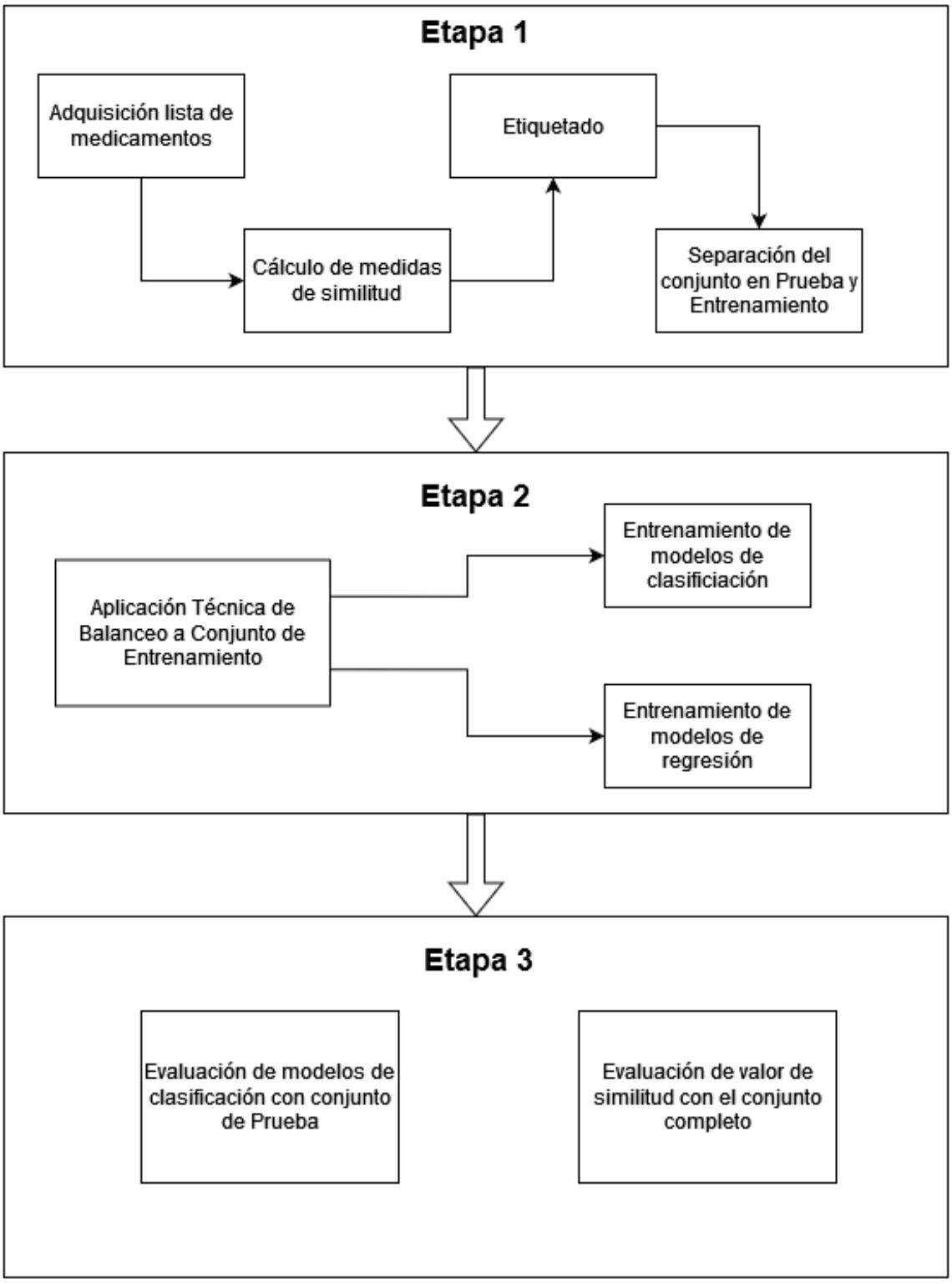
En este apartado se describe el proceso necesario para la creación del conjunto de datos a utilizar en las siguientes etapas, ver Figura 3-2. Para la creación del conjunto de datos se parte de (Millán-Hernández, García-Hernández, & Ledeneva, 2019b) al extraer los nombres únicos y generar un producto cartesiano entre estos. Esta propuesta permite obtener una lista de pares LASA lo suficientemente grande para representar de forma sólida la relación entre nombres de medicamento.

3.1.1 Adquisición de los datos

En este paso, la adquisición de datos consta de obtener una lista de medicamentos LASA (L), es decir, pares de nombres de medicamentos que han sido reportados como casos de errores de medicación por confusión de nombres de medicamentos. Una vez obtenida la lista, es necesario obtener el conjunto de los nombres únicos ($U = \{a_1, a_2, a_3, \dots a_n\}$ donde cada a_i es un nombre único) para construir una lista de pares de nombres de medicamentos D, mediante un producto cartesiano de estos nombres únicos ($U \times U = \{(a_i, a_j) |\ a_i\ y\ a_j\ \in U\ \wedge i\ \neq j\ \}$).

Figura 3-1

Diagrama del Método para la investigación.



3.1.2 Cálculo medidas de similitud

A partir de la lista obtenida D se procede al cálculo del valor de similitud y su normalización, para cada par de nombres de medicamento (a_i, a_j) empleando los algoritmos de correspondencia de cadena de la Tabla 3-1. Posterior al cálculo, se consolidan todas las medidas para todos los nombres de medicamentos en un único conjunto de datos.

Tabla 3-1

Lista de algoritmos de correspondencia de cadenas para el cálculo de medidas de similitud

Tipo de medida	Nombre del algoritmo
Fonética	Soundex
	Editex
	Phonix
Ortográfica	NED
	TED
	Bigrama 0b-0a
	Bigrama 0b-1a
	Bigrama 1b-0a
	Bigrama 1b-1a
	Trigrama 0b-0a
	Trigrama 0b-1a
	Trigrama 0b-2a
	Trigrama 1b-0a
	Trigrama 1b-1a
	Trigrama 1b-2a
	Trigrama 2b-0a
	Trigrama 2b-1a
	Trigrama 2b-2a
	Bisim
	Trisim
	Soft-Bisim
	Omission-key
	Skeleton-key
	Prefix
	TED

3.1.3 Etiquetado del conjunto de datos

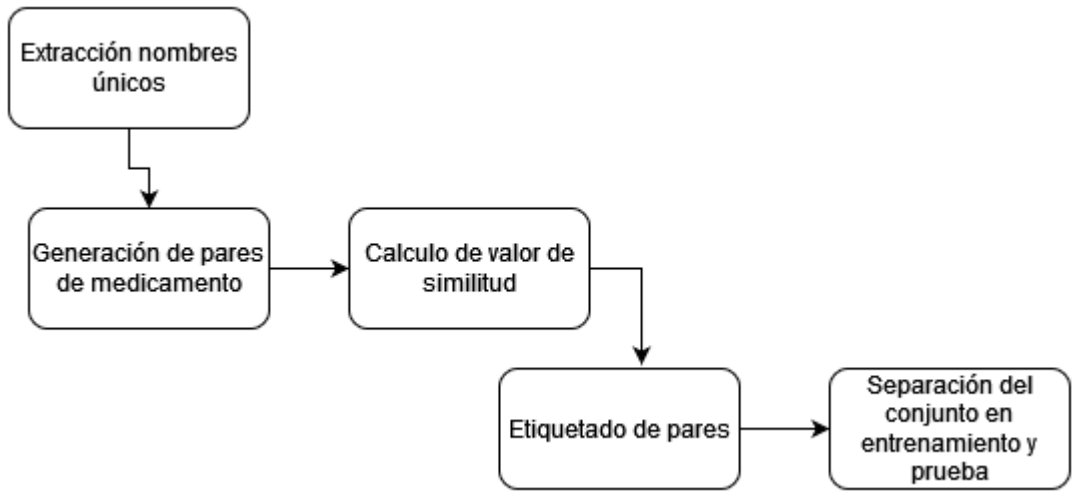
El etiquetado de los pares de la lista D se genera a partir de la lista L, en dónde es asigna un uno a todos los pares que estén incluidos en las listas publicadas, representando su pertenencia como un par LASA. El resto de los pares son etiquetados con cero (representando un par No-LASA). Es decir, $\forall\ (a_i, a_j)\ \in L$ etiquetar con uno, caso contrario etiquetar con cero.

3.1.4 Separación de los datos.

En el último paso de la etapa uno, se realiza una separación estratificada de la lista D en una proporción de 80% para un conjunto de entrenamiento y del 20% para el conjunto de prueba. El conjunto de entrenamiento $D_{Entrenamiento}$ es lo que se utilizará para construir el modelo y el conjunto de prueba D_{Prueba} son utilizados para evaluar el modelo obtenido.

Figura 3-2

Representación de la Etapa 1 del método



Nota: La figura resume el proceso llevado para la etapa uno del método, el método fue aplicado con las tres listas de forma separada.

Etapa 2: Entrenamiento de los algoritmos de aprendizaje computacional

En este apartado se describe el proceso para realizar el entrenamiento de un modelo a partir de los algoritmos de aprendizaje supervisado de la Tabla 3-2, para cada una de las técnicas de balanceo (Tabla 3-3) sobre el conjunto de entrenamiento $D_{Entrenamiento}$. Para validar los resultados se utiliza la técnica de validación cruzada *k-fold.*

Tabla 3-2

Algoritmos de aprendizaje supervisado para entrenamiento

Algoritmos de Clasificación	**Algoritmos de Regresión**
Regresión Logística	Regresión Logística-Regresor
Árboles de Decisión	Árboles de Decisión-Regresor
K-vecinos más cercanos	K-vecinos más cercanos-Regresor
Máquina de Soporte Vectorial	Máquina de Soporte Vectorial-Regresor
Redes Neuronales MLP	Redes Neuronales MLP-Regresor
Bosque Aleatorio	Bosque Aleatorio-Regresor
Árboles de Regresión con Impulso del Gradiente	Árboles de Regresión con Impulso del Gradiente-Regresor

Es importante mencionar que, en el entrenamiento, se realiza una separación del conjunto $D_{Entrenamiento}$ en 10 *folds* de forma estratificada para mantener las mismas proporciones de ambas clases (LASA y No-LASA) en cada *fold*. Además, de que en el entrenamiento de cada iteración los *fold* seleccionados se aplica la técnica de balanceo seleccionada y de que el *fold* restante se mantiene desbalanceado para representar la distribución real del problema de confusión por pares LASA durante la prueba en cada iteración.

Tabla 3-3

Técnicas de Balanceo de datos

Tipo de Balanceo	**Nombre de la Técnica**
Undersampling	Selección aleatoria
	Near miss
	Tomek´s Link
	Edited Nearest Neighbour
Oversampling	SMOTE
	ADASYN
Combination Sampling	SMOTE-*Tomek's Link*
	SMOTE-*Edited Nearest Neighbour*

La separación entre *folds* y el uso de técnicas de balanceo se ve reflejado en la figura 3-3. Dónde para la validación cruzada solo se utiliza el conjunto $D_{Entrenamiento}$ para la creación y evaluación de modelos, es importante mencionar que, en este paso, el *fold* de prueba no es modificado por la técnica de balanceo de datos.

3.2.1 Entrenamiento de modelos de clasificación

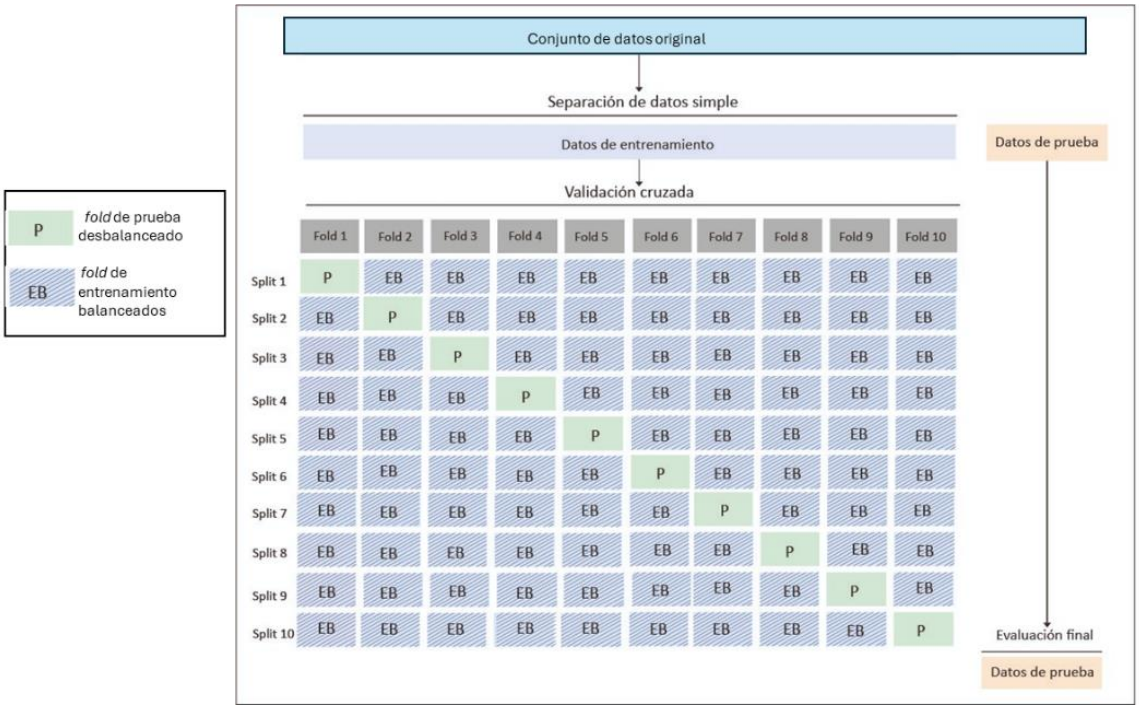
Para los modelos de clasificación de pares LASA, se realizará un entrenamiento con validación cruzada con K-*folds* sobre el conjunto $D_{Entrenamiento}$, dónde los *folds de* entrenamiento se aplicará técnicas de balanceo y el *fold* de prueba estará desbalanceado. Se determinará el F1 promedio de la clase de interés y se calculará la desviación estándar del promedio para determinar si existen sesgos durante el entrenamiento.

3.2.2 Entrenamiento de modelos de regresión

Para los modelos de regresión, se realizará un entrenamiento con validación cruzada con K-*folds* sobre el conjunto $D_{Entrenamiento}$, dónde los *folds de* entrenamiento se aplicará técnicas de balanceo y el *fold* de prueba estará desbalanceado. Se determinará el Error Cuadrático Medio (MSE) promedio de la regresión y se calculará la desviación estándar del promedio para determinar si existen sesgos durante el entrenamiento.

Figura 3-3

Diagrama de separación de datos para entrenamiento de modelos con K-fold y Validación cruzada



Etapa 3: Evaluación de los modelos en el conjunto de datos de prueba

En esta etapa de realiza la evaluación de los modelos obtenidos en la etapa dos, dicha evaluación se realizar por separado para los modelos de clasificación y de regresión.

3.3.1 Evaluación de los modelos de clasificación

En esta etapa, a partir del conjunto de entrenamiento $D_{Entrenamiento}$, es obtenido un modelo para cada uno los algoritmos de clasificación de la Tabla 3-2 con cada una de las técnicas de balanceo (Tabla 3-3) con el fin de entrenar el modelo. En seguida, se evalúa el modelo utilizando los datos de prueba D_{Prueba}. Mediante la medida F1 de la clase de relevancia (caso LASA).

3.3.2 Evaluación de los modelos de regresión

Para evaluar el desempeño de los modelos de regresión, al igual que en el paso 3.3.1 se entrenan los modelos de regresión de la Tabla 3-2 para las técnicas de la Tabla 3-3 mediante el conjunto de datos $D_{Entrenamiento}$. La evaluación del desempeño de los modelos obtenidos se realiza con el conjunto de datos D con la medida *F-macro-averaging* de las primeras n posiciones propuesto en el trabajo de (Millán-Hernández, García-Hernández, & Ledeneva, 2019b). El método de evaluación requiere calcular el valor de similitud para un conjunto de pares de nombres de medicamentos (predicción del modelo). Una vez obtenidos los valores se similitud son ordenados de forma descendente. A partir de esta lista ordenada, se evalúa el F1 en cada

posición, por lo que se considera que, a mayor número de pares LASA al principio de la lista, mayor es el valor de F1. Para resumir la eficacia en la recuperación de todos los pares LASA en las primeras n posiciones se utiliza la medida *F-macro-averaging*. Por lo que, n es determinado por la cantidad máxima de pares LASA en lo que en una la lista de pares de nombres confusos de medicamentos.

Resumen del capítulo

El método propuesto en este capítulo se pude resumir en el siguiente esquema donde:

- La Etapa 1 corresponde a la creación del conjunto de datos
- La Etapa 2 corresponde al Entrenamiento de modelos con técnicas de balanceo
- La Etapa 3 corresponde a la Evaluación para clasificar y predecir un valor de similitud

La Etapa 1 conlleva generar el producto cartesiano entre todos los nombres de medicamentos de la lista de pares LASA y calcular las medidas de similitud mediante algoritmos de correspondencia de cadenas. Hacer el etiquetado entre los pares LASA y no LASA y finalmente para crear el conjunto de datos *D* y separar los datos en conjuntos de D_{Prueba}. y $D_{Entrenamiento}$.

En la Etapa 2 se realiza el entrenamiento de modelos para regresión y modelos para clasificación de pares LASA mediante validación cruzada. El entrenamiento de los modelos es realizado con los algoritmos de aprendizaje supervisados listados en la tabla 3-2 y aplicando las técnicas de balanceo mencionadas en la tabla 3-3 sobre el conjunto $D_{Entrenamiento}$. Dónde el conjunto $D_{Entrenamiento}$ es dividido en K-*folds* con $K = 10$ para la validación cruzada. Los *folds* de entrenamiento les es aplicado una técnica de balanceo y el *fold* de prueba se mantiene desbalanceado.

La Etapa 3 consiste en el entrenamiento de modelos utilizando el conjunto de $D_{Entrenamiento}$ aplicando las técnicas de balanceo. La evaluación de los modelos de clasificación se realiza con el conjunto D_{Prueba} desbalanceado, considerando el F1 de la clase de interés (pares LASA) como métrica de evaluación. En el caso de la regresión se evaluará considerando el *F-macro-averaging* mediante una evaluación por posiciones para predecir un valor de similitud sobre el conjunto *D*.

Capítulo 4 Resultados

En los trabajos previos, revisados en el Capítulo 2, se proponen métodos basados en la Regresión Logística para identificar pares LASA. En algunos casos, se muestran experimentaciones con conjuntos de datos donde el número de casos LASA y no LASA están balanceados (Chen et al., 2011b; Lambert et al., 1999b, 2004b). En últimas investigaciones (Kondrak & Dorr, 2006b; Millán-Hernández et al., 2020b; Millán-Hernández, García-Hernández, & Ledeneva, 2019b) se utilizan conjuntos de datos desbalanceado, pero se optan por un enfoque evolutivo (Millán-Hernández, García-Hernández, & Ledeneva, 2019b; Vázquez et al., 2020).

En la hipótesis propuesta al inicio de esta tesis se dice que: *"La implementación de un de balanceo de datos para el entrenamiento de los modelos de aprendizaje computacional puede mejorar la clasificación y predicción de valor de similitud de pares de nombres confusos de medicamento por simetría fonética y ortográfica"*. Por lo tanto, el método del capítulo anterior se propone la implementación de distintas técnicas de balanceo con el objetivo de mejorar tanto los resultados de los modelos para clasificación como para la regresión. Además, de evaluar distintos modelos obtenidos. Para ello se utilizan tres bases de datos pertenecientes a listas de medicamentos confusos de diferentes países. Los conjuntos de datos serán referenciados en este capítulo como:

- EU-USP76 creada a partir de la lista de pares de nombres de medicamentos confusos por cómo se ven o como suenan (LASA) publicados por la USP en Estados Unidos.
- ESP-ISMP2018 creada a partir de la lista de pares de nombres medicamentos confusos por cómo se ven o como suenan (LASA) publicados por la ISMP de España.
- BRA-ISMP2014 creada a partir de la lista de pares de nombres de medicamentos por cómo se ven o como suenan (LASA) confusos publicados por la ISMP de Brasil.

La experimentación se realizó con *Python* 3.12.0, *Scikit-Learn* 1.3.1, *Imbalance-Learn* 0.11.0. En un equipo de cómputo con procesador Ryzen 5600 y 16 GB de memoria RAM; en un sistema operativo *Windows 11 Home Edition*.

4.1 Identificación de pares LASA en la base de datos EU-USP76

En la etapa uno del método propuesto en el capítulo anterior, a partir de la lista de medicamentos confusos de la USP publicada en (USP, 2001) se creó el conjunto de datos EU-USP76. La lista de la USP consta de 596 nombres únicos de medicamentos y un nombre de medicamento participan hasta ocho pares LASA. El conjunto de datos EU-USP76 se compone de 354,620 pares, obtenido a partir del producto cartesiano de los nombres únicos y tiene 25 características que fueron obtenidas mediante el cálculo de 25 medidas de similitud (Millán-Hernández et al., 2020b) con un valor en rango numérico entre 0 y 1. Después de etiquetar los datos la distribución de los datos en las dos clases es: 814 pares de nombres confusos (clase 1) y 353,806 pares no confusos (clase 0), es decir, existe una relación de 1:435 entre los pares confusos y no confusos.

A partir del conjunto de datos EU-USP76 se realizó una separación de los datos en dos conjuntos de datos, uno de entrenamiento y uno de prueba. En la Tabla 4-1 se muestra la distribución de los ejemplos tanto de los pares LASA y No-LASA.

Tabla 4-1

Distribución de los datos de EU-USP76

Categoría	EU-USP76-Entrenamiento	EU-USP76-Prueba
LASA	651	163
No-LASA	283,045	70,761

4.1.1 Primer Experimento: EU-USP76-Entrenamiento sin balanceo de clases

En la primera experimentación se utilizan el conjunto de datos EU-USP76-Entrenamiento obtenido en la etapa uno.

4.1.1.1. Clasificación LASA y No-LASA en EU-USP76-Entrenamiento sin balanceo de clases

Este experimento se enfoca en el problema *a posteriori*, es decir, donde el objetivo es predecir si dos nombres de medicamentos forman un par LASA o No-LASA. Los modelos: Bosque Aleatorio (RF), Árboles de Decisión (DT), K-Vecinos más cercanos (KNN), Regresión Logística (LR), Perceptrón Multicapa (MLP), Árboles de Decisión con Impulso de Gradiente (GBRT) y Máquina de Soporte Vectorial Lineal (SVM-L) fueron entrenados con el propósito de conocer el desempeño sobre el conjunto de datos EU-USP-Entrenamiento desbalanceado y ver su desempeño en contraste de usar técnicas de balanceo. En los resultados de la evaluación se utilizó la medida F1, dado que representa un equilibrio entre la precisión y la exhaustividad ideal para datos donde existe un desbalance entre las clases. En los resultados de la Tabla 4-2 solo se muestra el F1 obtenido en la clase de interés (LASA con etiqueta 1) utilizando la técnica de validación cruzada *K-fold* con un $k = 10$. Los resultados del promedio de los diez pliegues o *folds* muestran un entrenamiento sin sobreajuste. Además, la desviación estándar confirma que los resultados obtenidos son independientes de la forma en que se separan los datos para cada uno de los diez *folds*.

Tabla 4-2

Resultados de clasificación con k-fold en EU-USP76-Entrenamiento desbalanceado

Modelo	**Entrenamiento**		**Prueba**	
	F1 LASA	**Desv. Est.**	**F1 LASA**	**Desv. Est.**
RF	**0.98**	**0.00099**	**0.83**	**0.03491**
DT	0.98	0.00101	0.78	0.04605
KNN	0.54	0.00918	0.44	0.05617
LR	**0.35**	**0.00816**	**0.35**	**0.04975**
MLP	0.29	0.07137	0.30	0.10227
GBRT	0.30	0.15701	0.27	0.15319
SVM-L	0.15	0.01782	0.16	0.03667

Nota. Resultados de evaluación mediante validación cruzada k-folds de los siete modelos utilizados (Bosque Aleatorio (RF), Árboles de Decisión (DT), K-Vecinos más cercanos (KNN), Regresión Logística (LR), Perceptrón Multicapa (MLP), Árboles de Decisión con Impulso de Gradiente (GBRT) y Máquina de Soporte Vectorial Lineal (SVM-L)) con el promedio de la métrica F1 para los pares LASA (clase 1) sobre el conjunto EU-USP76 desbalanceado.

4.1.1.1. Regresión en EU-USP76-Entrenamiento sin balanceo de clases

En esta sección se muestra la experimentación con siete modelos de regresión: (Regresor de Bosque Aleatorio (RF-R), Regresor de Árboles de Decisión (DT-R), Regresor de K-Vecinos más cercanos *(KNN-R)*, Regresión Logística (LR-R), Regresor de Perceptrón Multicapa (MLP-R), Regresor de Árboles de Decisión con Impulso de Gradiente (GBRT-R) y Regresor de Máquina de Soporte Vectorial Lineal (SVM-L-R)), ver Tabla 4-3. La finalidad de estos modelos de predicción es calcular un valor de similitud para el problema *a priori* (para evaluar si la propuesta para un nombre de medicamento puede resultar en una potencial confusión). De acuerdo con la evaluación de los resultados obtenidos mediante validación cruzada con un $fold = 10$, los promedios de error MSE obtenidos muestran una desviación estándar en un rango entre 0.000156 y 0.000425. Por lo que, la forma de separar los datos no influye con respecto a los resultados obtenidos.

Tabla 4-3

Resultados de regresión con k-fold en EU-USP76-Entrenamiento desbalanceado

Modelo	**Entrenamiento**		**Prueba**	
	MSE	**Desv. Est.**	**MSE**	**Desv. Est.**
RF-R	0.00013	0.000003	0.00074	0.000108
DT-R	0.00003	0.000002	0.00101	0.000235
KNN-R	0.00099	0.000013	0.00153	0.000156
LR-R	0.00167	0.000028	0.00168	0.000265
MLP-R	0.0017	0.000044	0.0017	0.000425
GBRT-R	0.00126	0.000025	0.00150	0.000241
SVM-L-R	0.00229	0.000037	0.00229	0.000341

Nota. Resultados de evaluación en validación cruzada k-folds de los modelos utilizados (Regresor de Bosque Aleatorio (RF-R), Regresor de Árboles de Decisión (DT-R), Regresor de K-Vecinos más cercanos (KNN-R), Regresión Logística (LR-R), Regresor de Perceptrón Multicapa (MLP-R), Regresor de Árboles de Decisión con Impulso de Gradiente (GBRT-R) y Regresor de Máquina de Soporte Vectorial Lineal (SVM-L-R)) con el promedio de la métrica MSE sobre el conjunto EU-USP76 desbalanceado.

4.1.2 Segundo Experimento: EU-USP76-Entrenamiento con *undersampling*

En el segundo experimento se realizó el entrenamiento de los modelos mediante la implementación de balanceo al conjunto de datos EU-USP76 mediante las técnicas de *undersampling*: selección aleatoria (RS), *Near miss* (NM), *Tomek's Link* (TL) y *Edited Nearest Neighbour* (ENN). En la Tabla 4-4 se muestra la distribución de los datos de EU-USP76-Entrenamiento al aplicar cada técnica de *undersampling*.

Tabla 4-4

Número de ejemplos de entrenamiento EU-USP76-Entrenamiento con undersampling.

Clase	Técnica de undersampling			
	RS	NM	TL	ENN
LASA	651	651	651	651
No-LASA	651	651	283,031	228,494

Nota. Número de ejemplos de entrenamiento y prueba para EU-USP76-Entrenamiento mediante las técnicas de undersampling: selección aleatoria (RS), Near miss (NM), Tomek's Link (TL) y Edited Nearest Neighbour (ENN).

4.1.2.1 Clasificación LASA y No-LASA en EU-USP76 con undersampling

Para esta primera parte, el conjunto de datos EU-USP76-Entrenamiento balanceado mediante las técnicas de *undersampling* (ver Tabla 4-7) es utilizado. En la Tabla 4-5 se muestran los resultados del desempeño de los modelos obtenidos utilizando una validación cruzada *k-folds*, con un $k = 10$. Es importante destacar que el *fold* de prueba, correspondiente en cada iteración, no fue balanceado. Los resultados obtenidos muestran que el modelo de RF con TL obtiene el mejor resultado con un F1sobre la muestra de prueba para la clase de relevancia (clase LASA) con un valor de 0.82. En general, las técnicas TL y ENN se encuentra en las primeras posiciones en comparación con RS y NM.

Tabla 4-5

Resultados de clasificación con k-fold en EU-USP76-Entrenamiento con undersampling

Modelo	**Undersampling**	**Entrenamiento**		**Prueba**	
		F1 LASA	**Desv. Est.**	**F1 LASA**	**Desv. Est.**
RF	**TL**	**0.98**	**0.00123**	**0.82**	**0.03742**
DT	TL	0.98	0.00101	0.77	0.03555
RF	ENN	0.99	0.00072	0.76	0.06037
DT	ENN	0.99	0.0007	0.69	0.05348
KNN	ENN	0.65	0.00599	0.48	0.05329
KNN	TL	0.54	0.00925	0.44	0.05659
MLP	ENN	0.46	0.05943	0.42	0.08124
LR	**ENN**	**0.44**	**0.02309**	**0.40**	**0.04573**
GBRT	ENN	0.48	0.11657	0.39	0.07999
LR	TL	0.35	0.01644	0.35	0.06843
MLP	TL	0.34	0.07282	0.34	0.09648
SVM-L	ENN	0.32	0.01233	0.29	0.06298
GBRT	TL	0.23	0.17603	0.20	0.14979
SVM-L	TL	0.16	0.01922	0.17	0.03703
RF	RS	0.99	0.00026	0.08	0.01076
KNN	RS	0.95	0.00398	0.08	0.00921
SVM-L	RS	0.95	0.00497	0.08	0.01110
MLP	RS	0.95	0.00774	0.08	0.01186
LR	RS	0.94	0.00614	0.08	0.0142
GBRT	RS	0.99	0.00173	0.07	0.01323
DT	RS	0.99	0.00027	0.06	0.00993
LR	NM	0.88	0.00905	0.04	0.00561
SVM-L	NM	0.89	0.01293	0.01	0.00399
RF	NM	0.99	0.00079	0.00	0.00067
DT	NM	0.99	0.00078	0.00	0.00062
GBRT	NM	0.96	0.00553	0.00	0.00075
MLP	NM	0.91	0.01394	0.00	0.0023
KNN	NM	0.89	0.00767	0.00	0.00069

Nota. Resultados de evaluación mediante validación cruzada k-folds de los modelos utilizados (Bosque Aleatorio (RF), Árboles de Decisión (DT), K-Vecinos más cercanos (KNN), Regresión Logística (LR), Perceptrón Multicapa (MLP), Árboles de Decisión con Impulso de Gradiente (GBRT) y Máquina de Soporte Vectorial Lineal (SVM-L)) con el promedio de la métrica F1para los pares LASA (clase 1) sobre el conjunto EU-USP76-Entrenamiento balanceado mediante undersampling: selección aleatoria (RS), *Near miss* (NM), *Tomek's Link* (TL) y *Edited Nearest Neighbour* (ENN).

4.1.2.2 Regresión LASA y No-LASA en EU-USP76-Entrenamiento con undersampling

Al igual que en la sección 4.1.1.1 donde se realizó el experimento para calcular un valor de similitud para el problema *a priori* utilizando el conjunto de datos balanceado con *undersampling*. En la tabla 4-6 se muestra los promedios de error MSE obtenidos en validación cruzada *k-fold* con un $fold = 10$, con una desviación estándar en un rango entre 0.000156 y 0.000341. Por lo que, se puede considerar que no existe dependencia de la forma de separar los

datos con respecto a los resultados obtenidos. Las técnicas TL y ENN nuevamente mostraron mejores resultados, pero para el caso de los modelos de regresión.

Tabla 4-6

Resultados de regresión con k-fold en EU-USP76-Entrenamiento con undersampling

Modelo	**Undersampling**	**Entrenamiento**		**Prueba**	
		MSE	**Desv. Est.**	**MSE**	**Desv. Est.**
RF-R	**TL**	**0.00013**	**0.000004**	**0.00076**	**0.000108**
DT-R	TL	0.00003	0.00002	0.00106	0.000009
RF-R	ENN	0.00008	0.000003	0.00105	0.000156
GBRT-R	TL	0.00126	0.000027	0.00150	0.000244
KNN-R	TL	0.00098	0.000012	0.00153	0.000157
GBRT-R	ENN	0.00109	0.000025	0.00162	0.000230
LR-R	**TL**	**0.00166**	**0.000029**	**0.00168**	**0.000266**
DT-R	ENN	0.0	0.000001	0.00169	0.000234
KNN-R	ENN	0.00064	0.000010	0.00170	0.000177
MLP-R	TL	0.00169	0.000042	0.00172	0.000239
LR-R	ENN	0.00148	0.000027	0.00173	0.000271
MLP-R	ENN	0.00153	0.000064	0.00174	0.000251
SVM-L-R	ENN	0.00229	0.000038	0.00229	0.000341
SVM-L-R	TL	0.00229	0.000037	0.00229	0.000341
LR-R	RS	0.03773	0.002788	0.03659	0.002005
SVM-L-R	RS	0.07372	0.002877	0.03860	0.000782
KNN-R	RS	0.02831	0.002031	0.04015	0.003064
RF-R	RS	0.00459	0.000402	0.04066	0.002961
GBRT-R	RS	0.00820	0.001684	0.04086	0.003014
MLP-R	RS	0.04039	0.005220	0.04337	0.003582
DT-R	RS	0.0	0	0.06451	0.007151
LR-R	NM	0.09298	0.006201	0.06782	0.006917
SVM-L-R	NM	0.11528	0.006654	0.11123	0.006299
MLP-R	NM	0.07526	0.008024	0.41902	0.101852
KNN-R	NM	0.06483	0.004876	0.73522	0.060977
RF-R	NM	0.00914	0.000653	0.82582	0.048581
GBRT-R	NM	0.03255	0.003642	0.81122	0.078037
DT-R	NM	0.00518	0.000384	0.94549	0.020211

Nota. Resultados de evaluación en validación cruzada k-fold de los modelos utilizados (Regresor de Bosque Aleatorio (RF-R), Regresor de Árboles de Decisión (DT-R), Regresor de K-Vecinos más cercanos (KNN-R), Regresión Logística (LR-R), Regresor de Perceptrón Multicapa (MLP-R), Regresor de Árboles de Decisión con Impulso de Gradiente (GBRT-R) y Regresor de Máquina de Soporte Vectorial Lineal (SVM-L-R)) con el promedio de la métrica MSE sobre el conjunto EU-USP76 utilizando técnicas de undersampling: selección aleatoria (RS), Near miss (NM), Tomek's Link (TL) y Edited Nearest Neighbour (ENN).

4.1.3 Tercer Experimento: EU-USP76-Entrenamiento con *oversampling*

En este experimento se realizó el entrenamiento mediante la implementación de las técnicas de balanceo de *oversampling*: SMOTE (SM) y ADASYN (AS) al conjunto de datos EU-USP76-Entrenamiento. En la Tabla 4-7, se muestra las distribuciones de los datos para cada técnica de *oversampling*.

Tabla 4-7

Número de ejemplos de entrenamiento y prueba para EU-USP76 con oversampling.

Clase	Técnica de oversampling	
	SM	AS
LASA	283045	282933
No-LASA	283045	283045

Nota. Número de ejemplos de entrenamiento y prueba para EU-USP76-Entrenamiento mediante las técnicas de oversampling: SMOTE (SM) y ADASYN (AS).

4.1.3.1 Clasificación LASA y No-LASA en EU-USP76-Entrenamiento con oversampling

En la Tabla 4-8 se muestran el desempeño de los modelos, obtenidos al utiliza el conjunto de datos EU-USP76-Entrenamiento balanceado con las técnicas de *oversampling*, mediante validación cruzada *k-folds* con un $k = 10$. El *fold* correspondiente a la prueba, dentro de cada iteración, no fue balanceado. Los resultados obtenidos muestran que el modelo de RF tanto para la técnica SM y AS obtuvieron los mejores resultados con un F1 para la clase de relevancia (clase LASA) con un valor de 0.81. Además, se puede aprecia que el desempeño de LR se ve afectado por estas técnicas de balanceo de datos.

Tabla 4-8

Resultados de clasificación con k-fold en EU-USP76-Entrenamiento con oversampling

Modelo	***Oversampling***	**Entrenamiento**		**Prueba**	
		F1 LASA	**Desv. Est.**	**F1 LASA**	**Desv. Est.**
RF	**SM**	**0.99**	**0.0**	**0.81**	**0.03807**
RF	AS	0.99	0.0	0.81	0.03426
DT	SM	0.99	0.0	0.72	0.03903
DT	AS	0.99	0.0	0.71	0.06191
KNN	SM	0.99	0.00017	0.46	0.05156
KNN	AS	0.99	0.00015	0.46	0.05255
MLP	AS	0.99	0.00066	0.37	0.03468
MLP	SM	0.99	0.00051	0.35	0.05352
GBRT	SM	0.98	0.00054	0.15	0.02062
GBRT	AS	0.98	0.00068	0.13	0.01932
LR	**SM**	**0.95**	**0.00159**	**0.08**	**0.01206**
SVM-L	SM	0.95	0.00148	0.08	0.01200
SVM-L	AS	0.95	0.00176	0.07	0.01041
LR	AS	0.94	0.00198	0.07	0.01072

Nota. Resultados de evaluación mediante validación cruzada k-folds de los modelos: Bosque Aleatorio (RF), Árboles de Decisión (DT), K-Vecinos más cercanos (KNN), Regresión Logística (LR), Perceptrón Multicapa (MLP), Árboles de Decisión con Impulso de Gradiente (GBRT) y Máquina de Soporte Vectorial Lineal (SVM-L) con el promedio de la métrica F1 para los pares LASA (clase 1) sobre el conjunto EU-USP76 balanceado mediante oversampling: SMOTE (SM) y ADASYN (AS).

4.1.3.2 Regresión LASA y No-LASA en EU-USP76 con oversampling

A continuación, los modelos para el problema *a posterior*, sobre el conjunto de datos balanceado con técnicas de *oversampling*, fueron obtenidos y evaluados. Los resultados de la evaluación en validación cruzada *k-fold* con un $fold = 10$ se muestran en la Tabla 4-9. Los

promedios de error MSE calculados presentan una desviación estándar en un rango entre 0.000148 y 0.0010581. Por lo que, la forma de separar los no existe dependencia en los resultados obtenidos, ya que se mantiene una dispersión baja en cada iteración del *k-fold*.

Tabla 4-9

Resultados de regresión con k-fold en EU-USP76-Entrenamiento con oversampling

Modelo	***Oversampling***	**Entrenamiento**		**Prueba**	
		MSE	**Desv. Est.**	**MSE**	**Desv. Est.**
RF-R	**SM**	**0.00008**	**0.000003**	**0.00103**	**0.000148**
RF-R	AS	0.00008	0.000003	0.00104	0.000141
DT-R	AS	0.00002	0.000002	0.00133	0.000191
DT-R	SM	0.00002	0.000002	0.00134	0.000327
KNN-R	SM	0.00098	0.000041	0.00320	0.000381
KNN-R	AS	0.00098	0.000044	0.00321	0.000278
MLP-R	SM	0.01327	0.000832	0.01984	0.001424
MLP-R	AS	0.01333	0.000822	0.01989	0.000781
GBRT-R	SM	0.01503	0.000356	0.02046	0.000428
GBRT-R	AS	0.01700	0.000413	0.02390	0.000556
LR-R	**SM**	**0.03341**	**0.000971**	**0.03510**	**0.000952**
LR-R	AS	0.03826	0.001027	0.04137	0.001058
SVM-*L-R*	SM	0.08480	0.001608	0.05349	0.002520
SVM-*L-R*	AS	0.08932	0.001145	0.05967	0.002346

Nota. Resultados de evaluación en validación cruzada k-folds de los modelos utilizados (Regresor de Bosque Aleatorio (RF-R), Regresor de Árboles de Decisión (DT-R), Regresor de K-Vecinos más cercanos (KNN-R), Regresión Logística (LR-R), Regresor de Perceptrón Multicapa (MLP-R), Regresor de Árboles de Decisión con Impulso de Gradiente (GBRT-R) y Regresor de Máquina de Soporte Vectorial Lineal (SVM-L-R)) con el promedio de la métrica MSE sobre el conjunto EU-USP76 utilizando técnicas de oversampling: SMOTE (SM) y ADASYN (AS).

4.1.4 Cuarto Experimento: EU-USP76-Entrenamiento con *combination sampling*

En este experimento se utilizaron las técnicas de balanceo de datos, conocidas como *combination sampling* al conjunto de datos EU-USP76-Entrenamiento. Las combinaciones implementadas como técnicas de balanceo fueron: *SMOTE-Tomek's Link* (SM-TL) y como segunda combinación *SMOTE-Edited Nearest Neighbour* (SM-ENN). En la Tabla 4-10, se muestran las distribuciones de los datos para cada combinación.

Tabla 4-10.

Número de ejemplos de entrenamiento y prueba para EU-USP76-Entrenamiento al aplicar combination sampling

Clase	Técnica de *combination sampling*	
	SM-TL	SM-ENN
LASA	283045	282295
No-LASA	283045	281663

4.1.4.1 Clasificación LASA y No-LASA en EU-USP76-Entrenamiento con *combination-sampling*

Al utiliza el conjunto de datos EU-USP76-Entrenamiento balanceado mediante las técnicas de *combination-sampling*. La Tabla 4-11. muestran los resultados del desempeño de los modelos

obtenidos mediante una validación cruzada *k-folds* con $k = 10$. Es importante destacar que el *fold* correspondiente a la prueba, en cada iteración, no fue balanceado. Los resultados obtenidos muestran que el modelo de RF con SM-TL y DT-TL obtiene el mejor resultado con un F1 para la clase de relevancia (clase LASA) con un valor de 0.79 y 0.73, respectivamente. El desempeño de LR se ve afectado por estas técnicas de balanceo de datos obteniendo un F1 de 0.08.

Tabla 4-11

Resultados de clasificación con k-fold en EU-USP76-Entrenamiento con combination sampling

Modelo	**Com. Samp.**	**Entrenamiento**		**Prueba**	
		F1 LASA	**Desv. Est.**	**F1 LASA**	**Desv. Est.**
RF	**SM-TL**	**0.99**	**0**	**0.79**	**0.05418**
DT	SM-TL	0.99	0	0.73	0.07731
RF	SM-ENN	1.0	0	0.65	0.03954
DT	SM-ENN	1.0	0	0.56	0.04455
KNN	SM-TL	0.99	0.00019	0.46	0.02899
KNN	SM-ENN	0.99	0.00002	0.39	0.04688
MLP	SM-TL	0.99	0.00084	0.36	0.05482
MLP	SM-ENN	0.99	0.00050	0.34	0.06198
GBRT	SM-TL	0.98	0.00068	0.15	0.01362
GBRT	SM-ENN	0.98	0.00042	0.14	0.02188
SVM-L	SM-TL	0.95	0.00188	0.08	0.00818
LR	**SM-TL**	**0.95**	**0.00271**	**0.08**	**0.00836**
SVM-L	SM-ENN	0.95	0.00154	0.08	0.01192
LR	SM-ENN	0.95	0.00144	0.08	0.01188

Nota. Resultados de evaluación mediante validación cruzada k-folds de los modelos: Bosque Aleatorio (RF), Árboles de Decisión (DT), K-Vecinos más cercanos (KNN), Regresión Logística (LR), Perceptrón Multicapa (MLP), Árboles de Decisión con Impulso de Gradiente (GBRT) y Máquina de Soporte Vectorial Lineal (SVM-L) con el promedio de la métrica F1 para los pares LASA (clase 1) sobre el conjunto EU-USP76 balanceado mediante combination-sampling: SMOTE-Tomek's Link (SM-TL) y SMOTE-Edited Nearest Neighbour (SM-ENN).

4.1.4.2 Regresión LASA y No-LASA en EU-USP76-Entrenamiento con *combination-sampling*

A continuación, se muestran el desempeño de los modelos para el problema *a posteriori*, con las técnicas de balanceo de datos de *combination-sampling*. Los promedios de error MSE de los resultados obtenidos mediante validación cruzada con un $fold = 10$ son mostrados en la Tabla 4-12. Los errores presentaron una desviación estándar en un rango entre 0.000148 y 0.0010581, por lo que, se puede considerar que no existe dependencia de la forma de separar los datos con respecto a los resultados obtenidos.

Tabla 4-12

Resultados de regresión con k-fold en EU-USP76-Entrenamiento con combination sampling

Modelo	**Com. Samp.**	**Entrenamiento**		**Prueba**	
		MSE	**Desv. Est.**	**MSE**	**Desv. Est.**
RF-R	**SM-TL**	**0.00008**	**0.000003**	**0.00102**	**0.000146**
DT-R	SM-TL	0.00002	0.000002	0.00137	0.000330
RF-R	SM-ENN	0.00004	0.000002	0.00199	0.000196
DT-R	SM-ENN	0.0	0.0	0.00273	0.000453
KNN-R	SM-TL	0.00098	0.000044	0.00320	0.000283
KNN-R	SM-ENN	0.00013	0.000009	0.00536	0.000338
MLP-R	SM-ENN	0.01155	0.000729	0.01943	0.001779
MLP-R	SM-TL	0.01288	0.001619	0.01976	0.003037
GBRT-R	SM-TL	0.01517	0.000334	0.02061	0.000579
GBRT-R	SM-ENN	0.01361	0.000320	0.02092	0.000515
LR-R	**SM-TL**	**0.03343**	**0.001053**	**0.03511**	**0.001061**
LR-R	SM-ENN	0.03182	0.000897	0.03594	0.001011
SVM-L-*R*	SM-TL	0.08475	0.001494	0.05345	0.002530
SVM-L-*R*	SM-ENN	0.08356	0.001490	0.05353	0.002510

Nota. Resultados de evaluación en validación cruzada k-folds de los modelos utilizados (Regresor de Bosque Aleatorio (RF-R), Regresor de Árboles de Decisión (DT-R), Regresor de K-Vecinos más cercanos (KNN-R), Regresión Logística (LR-R), Regresor de Perceptrón Multicapa (MLP-R), Regresor de Árboles de Decisión con Impulso de Gradiente (GBRT-R) y Regresor de Máquina de Soporte Vectorial Lineal (SVM-L-R)) con el promedio de la métrica MSE sobre el conjunto EU-USP76 utilizando técnicas de combination-sampling: SMOTE-Tomek's Link (SM-TL) y SMOTE-Edited Nearest Neighbour (SM-ENN).

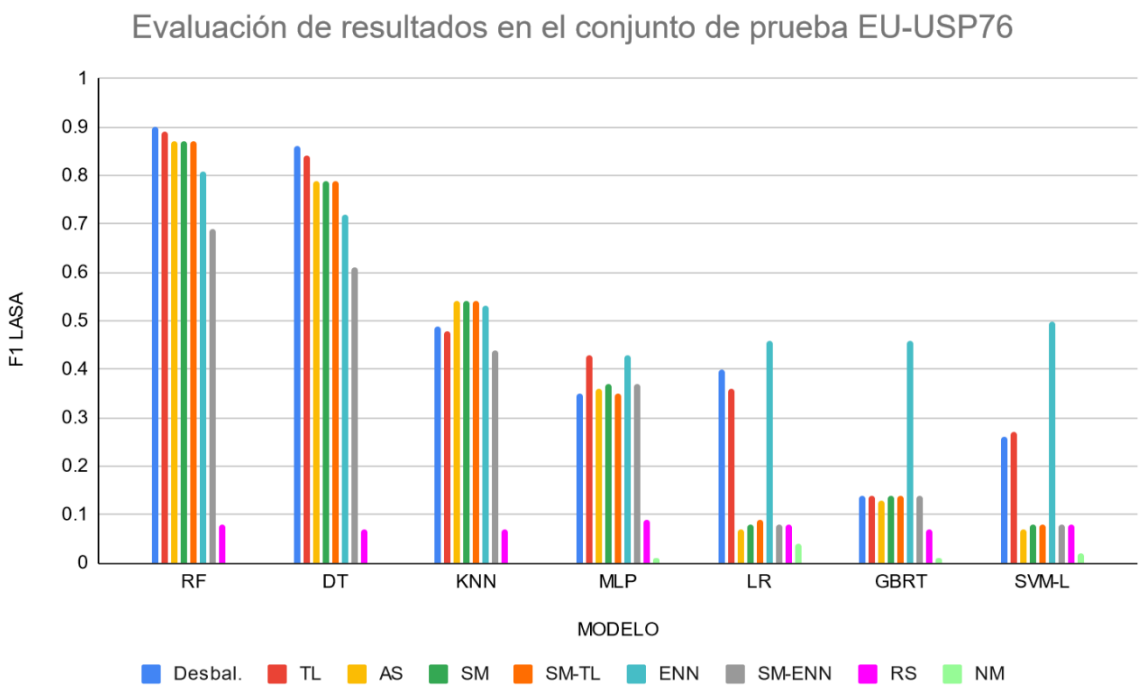
4.1.5 Evaluación de resultados en conjunto de datos EU-USP76-Prueba

A continuación, se presentan los resultados de la evaluación de los modelos obtenidos en los experimentos 4.1.1 al 4.1.4. Primero, los resultados para el caso de detección de nombres confusos *a posteriori* (clasificación), sobre el conjunto de datos de prueba EU-USP76-Prueba, se muestra en la Figura 4-1.

El mejor resultado en la evaluación de los modelos mediante el F1 de la clase de relevancia usando técnicas de balanceo fue obtenido por RF y DT con TL (ver apéndice A-1, para ver la matriz de confusión). Lo que representa una mejora en la tarea de clasificación de pares LASA y No-LASA. En cuanto al modelo LR, muestra valores de F1 sobre la clase LASA bajos con la mayoría de las técnicas de balanceo con excepción de aplicar *undersampling* con ENN y TL, no obstante, los resultados siguen siendo bajos comparados con RF con técnicas de balanceo con excepción de RS y NM.

Figura 4-1

Evaluación de los resultados de clasificación en el conjunto de prueba EU-USP76-Prueba



Nota. Evaluación de los resultados de clasificación de los modelos: Bosque Aleatorio (RF), Árboles de Decisión (DT), K-Vecinos más cercanos (KNN), Regresión Logística (LR), Perceptrón Multicapa (MLP), Árboles de Decisión con Impulso de Gradiente (GBRT) y Máquina de Soporte Vectorial Lineal (SVM-L) con el promedio de la métrica F1 para los pares LASA (clase 1) sobre el conjunto EU-USP76 en las versiones: Desbalanceada, selección aleatoria (RS), Near miss (NM), Tomek's Link (TL) y Edited Nearest Neighbour (ENN), SMOTE (SM), ADASYN (AS), SMOTE-Tomek's Link (SM-TL) y SMOTE-Edited Nearest Neighbour (SM-ENN).

En los resultados es posible observar que las técnicas de balanceo RS y NM muestran el desempeño más bajo para entrenar modelos en comparación de usar otra técnica de balanceo, para el F1 de la clase LASA. En el caso de la técnica de *undersampling* ENN, se puede observar en la Figura 4-1 que modelos como MLP, LR, GBRT y SVM-L obtuvieron mejoras significativas en comparación de otras técnicas de *oversampling* y *combination sampling*.

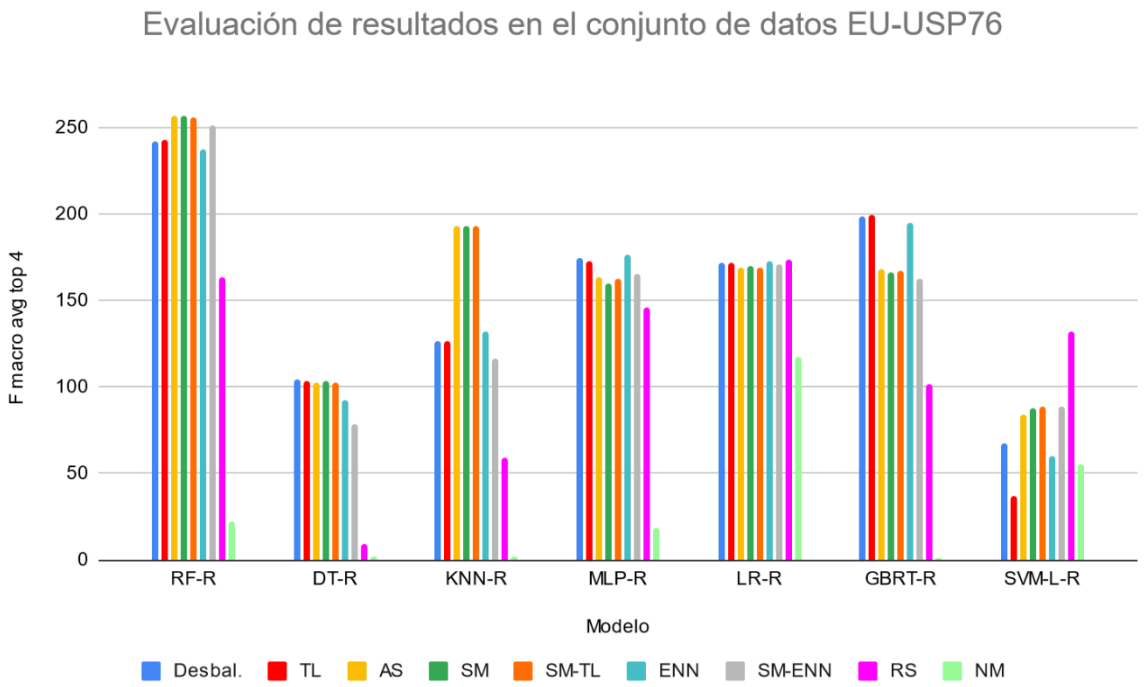
Posteriormente se realizó un análisis estadístico de los resultados mediante una prueba ANOVA (Devore, 2012) con la finalidad de conocer si el desempeño es mejor de al menos una de las técnicas con respecto a las demás. La prueba ANOVA obtuvo *p-value* de 0.00118 (para confiabilidad 99.9%) se puede decir que existe una significancia estadística entre las técnicas de balanceo para el F1 de pares LASA, para más información consultar apéndice A-1. Por lo que, se concluye que al aplicar técnicas de balanceo diferentes a RS si existe una mejora.

4.1.6 Evaluación de resultados en conjunto de datos EU-USP76

En esta última sección se muestran los resultados de la evaluación de los modelos obtenidos en los experimentos 4.1.1 al 4.1.4. para el caso *a prior* (regresión) sobre el conjunto de datos EU-USP76. La medida *F-macro-averaging* mediante una evaluación por posiciones para las primeras 4 posiciones, propuesta por (Millán-Hernández, García-Hernández, & Ledeneva, 2019b; Vázquez et al., 2020), sobre el conjunto de datos de prueba EU-USP76 es calculada (ver Figura 4-2). Los resultados de RF-R con *oversampling* destacan como los modelos que mejores valores obtienen, superando el valor obtenido en trabajos previos (Millán-Hernández, García-Hernández, & Ledeneva, 2019a) para predecir valores de similitud. Los resultados de RF-R muestran un desempeño alto tanto en los datos desbalanceados como en las diferentes técnicas de balanceo estudiadas en este trabajo, exceptuando nuevamente por NM.

Figura 4-2

Evaluación de los resultados de regresión en el conjunto de datos EU-USP76



Nota. Evaluación de los resultados de regresión de los modelos: Regresor de Bosque Aleatorio (RF-R), Regresor de Árboles de Decisión (DT-R), Regresor de K-Vecinos más cercanos (KNN-R), Regresión Logística (LR-R), Regresor de Perceptrón Multicapa (MLP-R), Regresor de Árboles de Decisión con Impulso de Gradiente (GBRT-R) y Regresor de Máquina de Soporte Vectorial Lineal (SVM-L-R) con el promedio de la métrica F1 sobre el conjunto EU-USP76 en las versiones: desbalanceada, selección aleatoria (RS), Near miss (NM), Tomek's Link (TL) y Edited Nearest Neighbour (ENN), SMOTE (SM), ADASYN (AS), SMOTE-Tomek's Link (SM-TL) y SMOTE-Edited Nearest Neighbour (SM-ENN).

Al realizar el análisis estadístico de los datos mediante una prueba ANOVA (Devore, 2012) se determinó que al menos una de las técnicas es mejor que RS. Con una *p-value* de 5.835354e-11 se determinó una significancia, para más información consultar apéndice A-1.

Los resultados obtenidos a su vez muestran una mejora para predecir un valor de similitud mediante la regresión con respecto a los trabajos con un enfoque *a priori* que utilizan *F-macro-averaging* como métrica de evaluación.

Tabla 4-13

F-macro-averaging top 4 comparativo con el conjunto EU-USP76

Modelo	***F-macro-averaging acumulado top 4***
RF-R SM	256.50
Regresión Simbólica-12 (Vázquez et al., 2020).	181.43
OLRM-21 (Millán-Hernández, García-Hernández, & Ledeneva, 2019b).	175.57
LRM-3 (Millán-Hernández, García-Hernández, & Ledeneva, 2019b).	167.52

Nota: Comparación de los resultados de regresión de los modelos utilizados en trabajos previos: Regresor de Bosque Aleatorio con SMOTE (RF-R SM), Regresión Simbólica-12, Optimized Logistic Regression Measures with 21 measures (OLRM-21) y Logistic Regression Measures with 3 Measures (LRM-3).

4.2 Identificación de pares LASA en la base de datos ISMP-ESP 2018

El conjunto ESP-ISMP2018 fue construido a partir de la lista de medicamentos confusos de España publicada por la ISMP (ISMP España, 2020), de la misma forma que se realizó el conjunto de pares LASA de Estados Unidos en la sección 4.1. La lista de la ISMP España cuenta con 586 nombres únicos de medicamento, donde cada nombre único participa hasta en 5 pares LASA. El conjunto de datos ESP-ISMP2018 consta de 342,810 pares. La distribución de los datos después del etiquetado es de: 784 pares de nombres confusos (pares LASA) y 342,026 pares de nombres no confusos (pares No LASA). Los datos presentan un desbalance, e cual tienen una relación de 1:436 entre pares confusos y no confusos.

A continuación, se realizó la separación de los datos en entrenamiento y prueba. En la tabla *4-13* se muestra la distribución de los pares LASA y No-LASA para cada conjunto.

Tabla 4-14

Distribución de los datos de ESP-ISMP2018

Categoría	**ESP-ISMP2018-Entrenamiento**	**ESP-ISMP2018-Prueba**
LASA	627	157
No-LASA	273,621	68,405

4.2.1 Experimentos: ESP-ISMP2018-Entrenamiento: sin balanceo de clases, *undersampling*, *oversampling* y *combination sampling*.

En esta sección se realizaron cinco experimentos con el conjunto de datos ESP-ISMP2018-Entrenamiento, utilizando la base de datos desbalanceada y posteriormente aplicando las técnicas de balance de datos de *undersampling*, *oversampling* y *combination sampling* (al igual que en los experimentos de la base de datos EU-USP76). De la misma forma, se obtuvieron siete modelos para el problema *a priori* y *a posteriori*, es decir, modelos de clasificación y de regresión.

En los modelos de clasificación, para el caso *a priori*, se obtuvieron resultados de evaluación mediante una validación cruzada k-fold con un *k=10.* Los resultados de dichos experimentos se pueden consultar en el Apéndice B-1. En general, los resultados muestran que las técnicas de *undersampling* ENN obtuvo el mejor resultado del F1 sobre la clase LASA, en comparación con las otras técnicas de balanceo de datos, sin embargo, el modelo RF obtuvo un 0.99 en su entrenamiento promedio y un 0.87 en su evaluación en el F1 de la clase LASA con los datos desbalanceados.

En los modelos de regresión, para el caso *a posteriori*, los resultados de la evaluación de los modelos pueden ser consultados en el Apéndice B-1. En el caso de la regresión, RF-R desbalanceado obtiene los mejores resultados con un MSE promedio prueba de 0.00065.

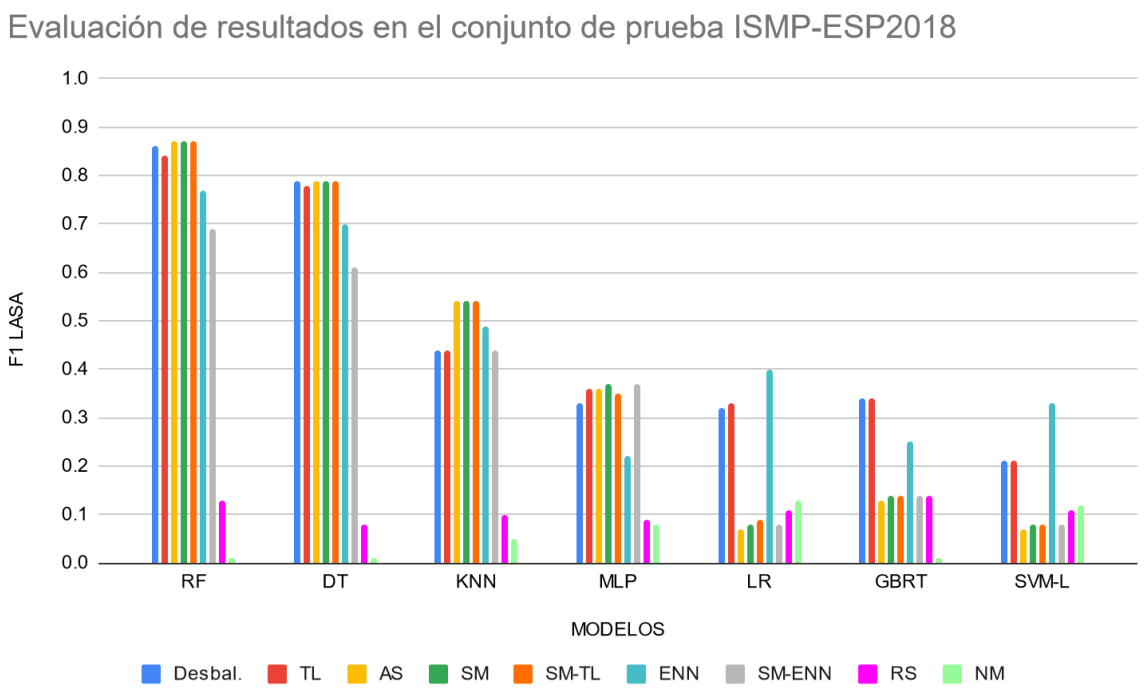
4.2.2 Evaluación de resultados en conjunto de datos ESP-ISMP2018-Prueba

A continuación, se presentan los resultados de la evaluación de los modelos sobre el conjunto ESP-ISMP-2018-Prueba, al utilizar los modelos entrenados en la sección 4.2.1. Los resultados para la detección de pares de nombres confusos *a posteriori* (clasificación), sobre el conjunto de datos de prueba se muestra en la Figura 4-3.

El mejor resultado de evaluación de modelos mediante el F1 de la clase de relevancia fue obtenido por RF y DT con el conjunto de entrenamiento desbalanceado (ver apéndice B-1, para ver la matriz de confusión), en todas las técnicas de *oversampling*, *combination sampling* y en las técnicas de *undersampling* ENN y TL. En el caso del modelo LR, los resultados de la métrica F1 sobre la clase LASA resultaron inferiores a los obtenidos por los modelos RF Y DT. Lo que muestra que existen modelos a partir de otros algoritmos de aprendizaje computacional que pueden mejorar la tarea de identificación de nombres confusos de medicamentos.

Figura 4-3

Evaluación de los resultados de clasificación en el conjunto de prueba ESP-ISMP2018-Prueba



Nota. Evaluación de los resultados de clasificación de los modelos: Bosque Aleatorio (RF), Árboles de Decisión (DT), K-vecinos más cercanos (KNN), Regresión Logística (LR), Perceptrón Multicapa (MLP), Árboles de Decisión con Impulso de Gradiente (GBRT) y Máquina de Soporte Vectorial Lineal (SVM-L) con el promedio de la métrica F1 para los pares LASA (clase 1) sobre el conjunto ESP-ISMP2018 en las versiones: Desbalanceada, Selección Aleatoria (SM), Near miss (NM), Tomek's Link (TL) Edited Nearest Neighbour (ENN), SMOTE (SM), ADASYN (AS), SMOTE-Tomek's Link (SM-TL) y SMOTE-Edited Nearest Neighbour (SM-ENN).

En los resultados de la Figura 4-3, se observa que las técnicas de balanceo de datos RS y NM muestran el desempeño más bajo en comparación de otras técnicas de balanceo, para la recuperación de la clase LASA. En el caso de la técnica de *undersampling* ENN, se puede observar en la Figura 4-3 que modelos como LR, GBRT y SVM-L obtuvieron mejoras significativas en comparación de otras técnicas de *oversampling* y *combination sampling*. Para el modelo MLP, las técnicas de *oversampling* y *combination sampling* presentan un mejor valor F1 que el uso de técnicas de *undersampling* y entrenamiento desbalanceado.

Posteriormente se realizó el análisis estadístico de los datos mediante una prueba ANOVA (Devore, 2012) en el cual se obtuvo *p-value* de 0.000001. Por lo que se puede nuevamente establecer una significancia estadística con respecto al desempeño de las técnicas de balanceo, para conocer más detalles, consultar apéndice B-1.

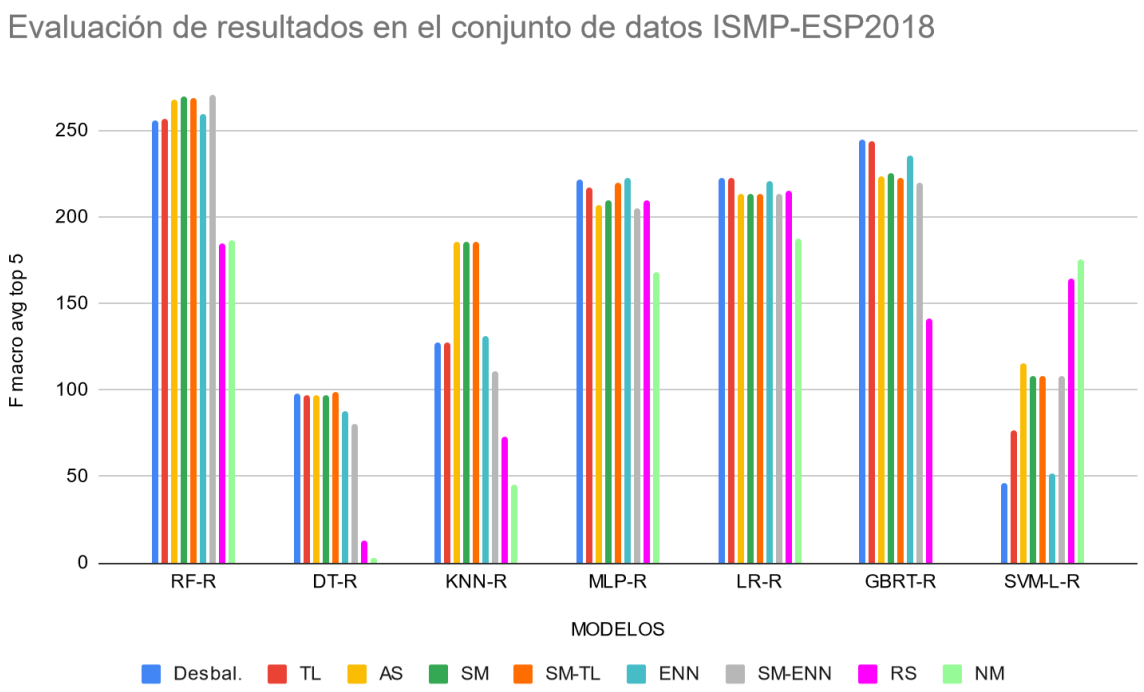
4.2.3 Evaluación de resultados en conjunto de datos ESP-ISMP2018

Para concluir la experimentación con el conjunto de datos ESP-ISMP2018, se muestran los resultados de evaluación de los modelos obtenidos en los experimentos para el caso *a priori* (regresión), utilizando la medida *F-macro-averaging* mediante una evaluación por posiciones para las primeras 5 posiciones propuesta por (Millán-Hernández et al., 2020a)(ver Figura 4-4).

Los resultados muestran que RF-R destaca como uno de los mejores modelos, teniendo un mejor desempeño que LR utilizado en trabajos previos, para predecir valores de similitud. RF-R presenta un desempeño alto en todas las técnicas de balanceo aplicadas con la excepción de NM y RS.

Figura 4-4

Evaluación de los resultados de regresión en el conjunto de datos ESP-ISMP2018



Nota. Evaluación de los resultados de regresión de los modelos: Regresor de Bosque Aleatorio (RF-R), Regresor de Árboles de Decisión (DT-R), Regresor de K-Vecinos más cercanos (KNN-R), Regresión Logística (LR-R), Regresor de Perceptrón Multicapa (MLP-R), Regresor de Árboles de Decisión con Impulso de Gradiente (GBRT-R) y Regresor de Máquina de Soporte Vectorial Lineal (SVM-L-R) con el promedio de la métrica F1 sobre el conjunto EU-USP76 en las versiones: desbalanceada, selección aleatoria (RS), Near miss (NM), Tomek's Link (TL) y Edited Nearest Neighbour (ENN), SMOTE (SM), ADASYN (AS), SMOTE-Tomek's Link (SM-TL) y SMOTE-Edited Nearest Neighbour (SM-ENN).

Posteriormente se realizó un análisis estadístico de los resultados de la Figura 4-4 mediante una prueba ANOVA (Devore, 2012) la cual obtuvo *p-value* de 4.699723e-08. Por lo que se puede establecer una significancia estadística entre las técnicas de balanceo para la recuperación de pares LASA, para más información consultar apéndice B-1.

Los resultados obtenidos a su vez muestran una mejora para predecir un valor de similitud mediante la regresión con respecto a los trabajos con un enfoque *a priori* que utilizan *F-macro-averaging* como métrica de evaluación.

Tabla 4-15

F-macro-averaging top 5 comparativo con el conjunto ESP-ISMP2018

Modelo	***F-macro-averaging acumulado top 5***
RF-R SM-ENN	270.84
OLRM-25 (Millán-Hernández et al., 2020b).	220.71

Nota: Comparación de los resultados de regresión de los modelos utilizados en trabajos previos: Regresor de Bosque Aleatorio con SMOTE-Edited Nearest Neighbour (RF-R SM-ENN) y Optimized Logistic Regression Measures with 25 measures (OLRM-25).

4.3 Identificación de pares LASA en la base de datos BRA-ISMP2014

El conjunto BRA-ISMP2014 fue construido a partir de la lista de medicamentos confusos de Brasil publicada por la ISMP (ISMP Brasil, 2014) , de la misma forma que se realizó el conjunto de pares LASA de Estados Unidos y España en la sección 4.1 y 4.2 respectivamente. La lista de la ISMP Brasil cuenta con 122 nombres únicos de medicamentos donde cada nombre único participa hasta en 4 pares LASA. El conjunto de datos BRA-ISMP2014 consta de 14,762 pares. La distribución de los datos después del etiquetado en ambas clases se compone de 170 pares de nombres confusos (LASA) y 14,592 pares no confusos (No LASA). Existiendo una relación de 1:86 entre pares confusos y no confusos.

Posteriormente, se realizó una separación de los datos en subconjuntos de entrenamiento y de prueba. En la tabla 4-14 se muestra la distribución de los pares LASA y No-LASA.

Tabla 4-16

Distribución de los datos de BRA-ISMP2014

Categoría	BRA-ISMP2014-Entrenamiento	BRA-ISMP2014-Prueba
LASA	136	34
No-LASA	11673	2919

4.3.1 Experimentos: BRA-ISMP2014-Entrenamiento: sin balanceo de clases, *undersampling*, *oversampling* y *combination sampling*.

En esta sección se realizaron cinco experimentos con el conjunto de datos BRA-ISMP2014-Entrenamiento, utilizando la base de datos desbalanceada y posteriormente aplicando las técnicas de balance de datos *undersampling*, *oversampling* y *combination sampling* (al igual que en los experimentos de la base de datos EU-USP76 y ESP-ISMP2018). De la misma forma, se obtuvieron siete modelos para el problema *a priori* y *a posteriori*, es decir, modelos de clasificación y de regresión.

En los modelos de clasificación, para el caso *a priori*, se obtuvieron resultados de evaluación mediante una validación cruzada k-fold con $k = 10$. Los resultados de dichos experimentos se pueden consultar en el Apéndice C-1. En general, los resultados muestran que las técnicas de *undersampling* TL obtuvo el mejor resultado del f1 sobre la clase LASA, en comparación a

otras técnicas de balanceo de datos, sin embargo, RF obtuvo un 1.0 en su entrenamiento promedio y un 0.89 en su evaluación en el F1 de la clase LASA con los datos desbalanceados.

En los modelos de regresión, para el caso *a posteriori*, los resultados de la evaluación de los modelos pueden ser consultados en el Apéndice C-1. En el caso de la regresión, RF-R con TL obtiene los mejores resultados con un MSE promedio prueba de 0.00283

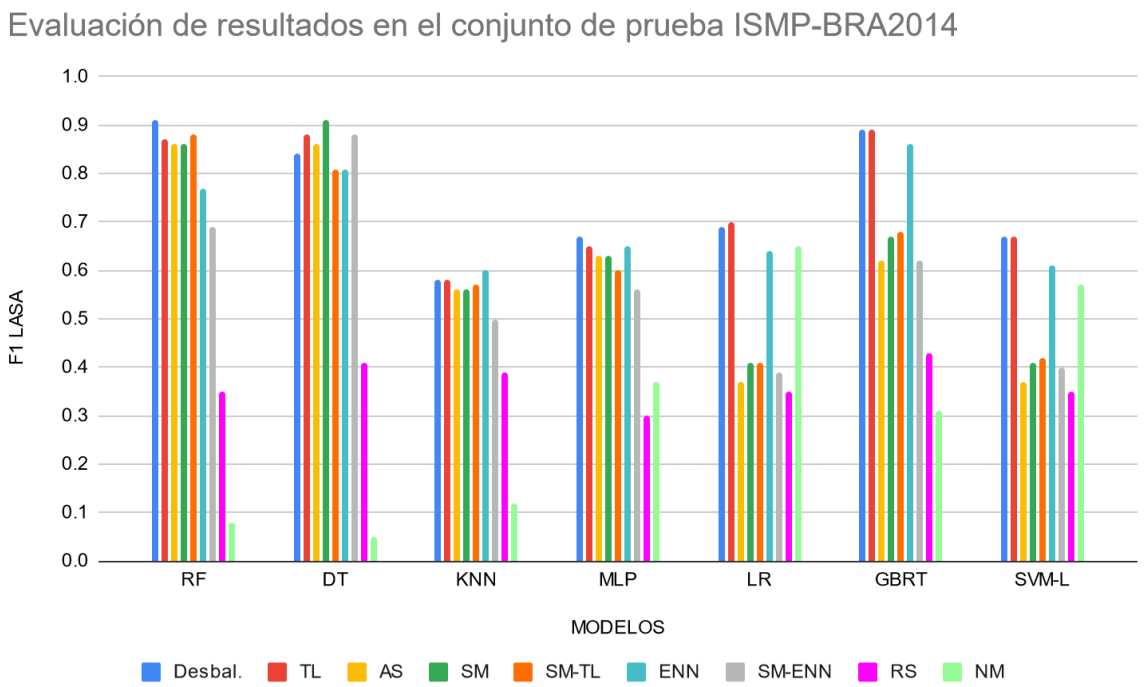
4.3.2 Evaluación de resultados en conjunto de datos BRA-ISMP2014-Prueba

A continuación, se presenta la evaluación de los modelos utilizando el conjunto de datos BRA-ISMP2014-Prueba. Los resultados de la detección de pares de nombres confusos a posteriori (clasificación) en el conjunto de datos de prueba se exhiben en la Figura 4-5.

El modelo que mejor valor obtuvo en la métrica F1 para la clase de relevancia fue RF con el conjunto desbalanceado (ver apéndice C-1, para ver la matriz de confusión), DT con SM como técnica de balanceo aplicada y GBRT con el conjunto desbalanceado y GBRGT con TL como técnica de balanceo. Dando resultados por encima de LR utilizado por (Lambert et al., 1999a).

Figura 4-5

Evaluación de los resultados de clasificación en el conjunto de prueba BRA-ISMP2014-Prueba



Nota. Evaluación de los resultados de clasificación de los modelos: Bosque Aleatorio (RF), Árboles de Decisión (DT), K-Vecinos más cercanos (KNN), Regresión Logística (LR), Perceptrón Multicapa (MLP), Árboles de Decisión con Impulso de Gradiente (GBRT) y Máquina de Soporte Vectorial Lineal (SVM-L) con el promedio de la métrica F1 para los pares LASA (clase 1) sobre el conjunto EU-USP76 en las versiones: Desbalanceada, selección aleatoria (RS), Near miss (NM), Tomek's Link (TL) y Edited Nearest Neighbour (ENN), SMOTE (SM), ADASYN (AS), SMOTE-Tomek's Link (SM-TL) y SMOTE-Edited Nearest Neighbour (SM-ENN).

En la Figura 4-5 Se aprecia que aplicar las técnicas de balanceo NM y RS obtienen resultados bajos en comparación de no utilizar una técnica de balanceo. GBRT con el conjunto BRA-ISMP2014 obtiene resultados similares a RF, siendo que sus valores en F1 sin balanceo son cercanos entre sí, a diferencia del entrenamiento con conjunto de datos en español e inglés donde los resultados exhiben una disparidad considerable.

Posteriormente se realizó la prueba ANOVA [(Devore, 2012) donde se obtuvo *p-value* de 0.000055. Es decir, que nuevamente existe significancia estadística entre las técnicas de balanceo para el F1 de pares LASA, para más información consultar apéndice C-1.

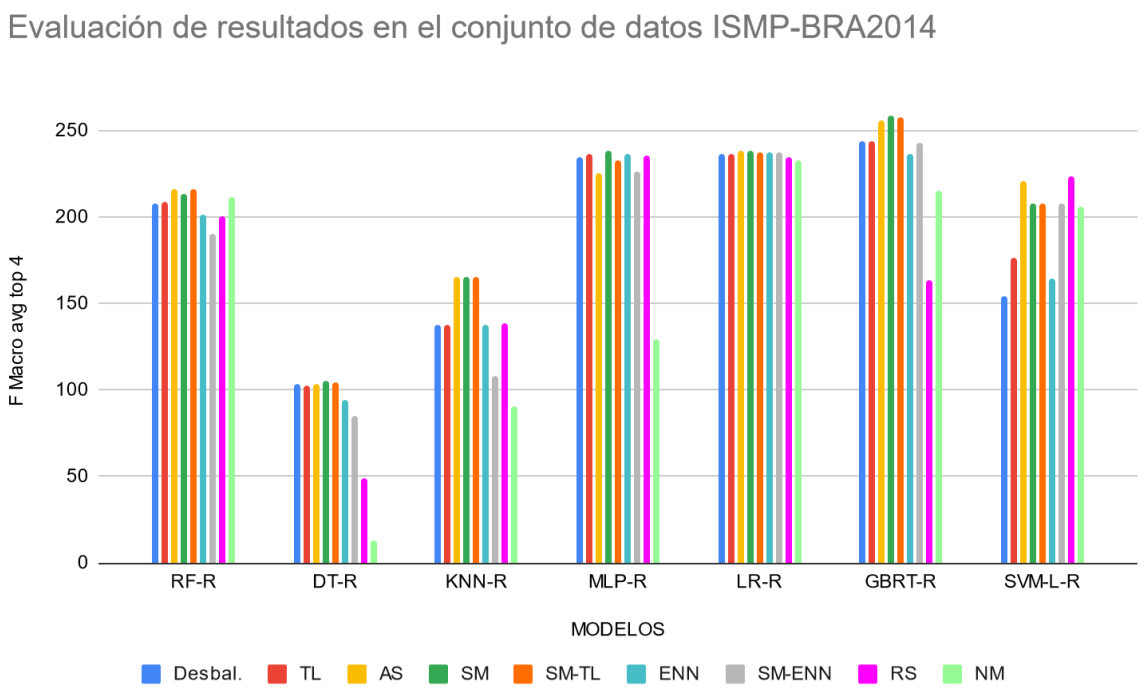
4.3.3 Evaluación de resultados en conjunto de datos BRA-ISMP2014

Para finalizar las evaluaciones sobre BRA-ISMP2014. A continuación, muestra la evaluación de modelos entrenados para el caso *a priori* (regresión) y generar un valor de similitud. En la evaluación de los resultados se calculó la medida *F-macro-averaging* mediante una evaluación por posiciones para las primeras 4 posiciones (Millán-Hernández et al., 2020a; Millán-Hernández, García-Hernández, & Ledeneva, 2019a) (ver Figura 4-6).

Los resultados apuntan a que GBRT obtiene los valores más altos al utilizar la mayoría de las técnicas de balanceo sobre todo con técnicas de *oversampling*, con la excepción de RS y NM. A su vez LR se mantiene equilibrado en resultados sin importar la técnica de balanceo aplicada para su entrenamiento siendo qué es LR el segundo modelo con los valores más altos.

Figura 4-6

Evaluación de los resultados de regresión con el conjunto de datos BRA-ISMP2014



Nota. Evaluación de los resultados de regresión de los modelos: Regresor de Bosque Aleatorio (RF-R), Regresor de Árboles de Decisión (DT-R), Regresor de K-Vecinos más cercanos (KNN-R), Regresión Logística (LR-R), Regresor de Perceptrón Multicapa (MLP-R), Regresor de Árboles de Decisión con Impulso de Gradiente (GBRT-R) y Regresor de Máquina de Soporte Vectorial Lineal (SVM-L-R) con el promedio de la métrica F1 sobre el conjunto EU-USP76 en las versiones: desbalanceada, selección aleatoria (RS), Near miss (NM), Tomek's Link (TL) y Edited Nearest Neighbour (ENN), SMOTE (SM), ADASYN (AS), SMOTE-Tomek's Link (SM-TL) y SMOTE-Edited Nearest Neighbour (SM-ENN).

Posteriormente se realizó la prueba ANOVA (Devore, 2012) con *p-value* de 4.699723e-08. Por lo que se puede establecer una significancia estadística entre las técnicas de balanceo en el desempeño *F-macro-averaging* para las primeras 4 posiciones para mejora de la recuperación de pares LASA (para más información consultar apéndice C-1).

Resumen del Capítulo

A lo largo de la experimentación se observó que los modelos entrenados con RF y RF-R para el conjunto de datos EU-USP76 Y ESP-ISMP2018 obtienen resultados positivos al utilizar alguna técnica de balanceo diferente a RS Y NM. A su vez se destaca DT como una alternativa para el enfoque *a posteriori* (clasificación) y GBRT-R para el enfoque *a priori* (regresión) para generar un valor de similitud. Cabe aclarar que el mejor modelo tanto para clasificación como para regresión se obtuvo al entrenar RF desbalanceado.

Con el conjunto de datos BRA-ISMP2014 se observa que el modelo de DT con SM obtiene los valores más altos en F1 de la clase de relevancia entre las técnicas de balanceo en la clasificación. No obstante, para la generar un valor de similitud se observa que el GBRT-R utilizando técnicas de *oversampling* obtiene un valor más alto en su *F-macro-averaging* que utilizar DT-R y RF-R con cualquier técnica de balanceo lo que supone un contraste con los otros dos conjuntos de datos. Así mismo se observa que MLP-R con casi todas las técnicas de balanceo aplicadas a excepción de NM, obtiene puntuaciones más altas. Quién repite este patrón es LR que aún con NM es consistente en sus resultados suponiendo un contraste amplio con LR entrenador con las bases de datos EU-USP76 e ESP-ISMP2018.

Se observa también que SVM-L, KNN, SVM-L-R y KNN-R no obtuvieron rendimientos satisfactorios para la clasificación o para generar un valor de similitud con los conjuntos EU-USP76 y ESP-ISMP2018. Con el conjunto BRA-ISMP2014 se obtuvieron resultados positivos, pero sin alcanzar el rendimiento de que se obtiene al entrenar con los algoritmos de RF, GBRT, RF-R y GBRT-R.

Adicionalmente se señala un rendimiento alto en algunos de los modelos al entrenar con el conjunto desbalanceado como lo es en el caso de RF y RF-R en los conjuntos de datos provenientes de Estados Unidos y España. En el caso del conjunto proveniente de Brasil es RF y GBRT-R.

Finalmente, al aplicar pruebas ANOVA en cada evaluación en los conjuntos de datos de prueba, para EU-USP76, ESP-ISMP2018 y BRA-ISMP2014 se mostró que al menos una de las técnicas de balanceo es estadísticamente significativa mejor que el resto. Esto permite, entender que la mejora de los modelos obtenidos es resultado de la aplicación de una técnica en particular.

Capítulo 5 Conclusiones

Los errores de medicación por confusión de nombres confusos actualmente siguen siendo un problema de interés para las entidades regulatorias. Debido a la existencia de pares LASA y por lo tanto se buscan estrategias como métodos computacionales que ayuden a identificar potenciales formación de pares LASA a partir de nombres de medicamentos propuestos (caso a priori) y alertar sobre casos de confusión durante la administración de medicamentos (enfoque a posteriori). Por lo tanto, en este trabajo se abordó el problema del uso de algoritmos de aprendizaje computacional para la identificación de nombres confusos de medicamentos por su simetría fonética y ortográfica.

En los trabajos previos revisados en esta tesis, la Regresión Logística es utilizado para la identificación de pares LASA. Por lo que, en esta tesis se planteó mejorar la identificación de pares LASA mediante la implementación de algoritmos de Aprendizaje Computacional como: Bosque Aleatorio, Árboles de Decisión, K-vecinos más cercanos, Árboles de Regresión con Impulso del Gradiente, Redes Neuronales MLP y Máquina de Soporte Vectorial Lineal. Los resultados de esta investigación muestran que sí es posible mejorar la identificación de nombres de medicamento confusos tanto para el caso *a priori* como el caso *a posteriori.* Entonces, el objetivo de esta investigación fue alcanzado al mostrar que existen algoritmos como Bosque Aleatorio que mejoran la clasificación de pares LASA. En el caso de predecir un valor de similitud, Bosque Aleatorio Regresor y Árboles de Regresión con Impulso del Gradiente Regresor muestran los mejores resultados.

Existen alternativas al uso de la regresión logística para la mejora de la identificación de nombres confusos de medicamentos por simetría fonética y ortográfica.

Dado que en lo experimentos realizados en las tres bases de datos obtuvieron *p-values* por debajo de 0.01, se puede descartar una hipótesis nula, es decir, *no existe una diferencia significativa entre técnicas de balanceo para identificación de nombres confusos de medicamento*). Por lo tanto, no existe evidencias en la experimentación para rechazar la hipótesis de este trabajo. Sin embargo, es importante aclarar que las pruebas ANOVA fueron realizadas solo entre las técnicas de balanceo. Si se tomará en cuenta en la prueba de hipótesis, la evaluación los modelos obtenidos con el conjunto desbalanceado, los resultados muestran que la proporción de los datos es lo suficientemente robusta para que algunos algoritmos de aprendizaje supervisado generen modelos de Bosque Aleatorio con resultados similares a los modelos obtenidos por técnicas de balanceo de datos para el enfoque *a posteriori*.

Es relevante mencionar que, por limitantes de tiempo durante la experimentación, la cantidad de algoritmos de aprendizaje supervisado probados es reducida, lo mismo se puede decir de las técnicas de balanceo aplicadas en este proyecto. No obstante, los resultados obtenidos apuntan a una viabilidad al uso de algoritmos distintos a la Regresión Logística y el uso de técnicas de balanceo por mejorar la identificación de nombres confusos de medicamentos.

Una limitante, de esta investigación, es la implementación de los modelos evaluados, los cuáles no están diseñados para ser integrados a un software especializado o de uso comercial. El proyecto está planteado como una guía para futuras investigaciones que busquen implementar

modelos de aprendizaje supervisado para tratar errores de medicación por confusión de nombres.

Finalmente, Los hallazgos de esta investigación refuerzan el impacto del uso de técnicas de balanceo de datos en el entrenamiento de modelos de aprendizaje computacional, evidenciando su mejora en el rendimiento de clasificación y predicción de valores de similitud de nombres de medicamento LASA.

5.1 Aportaciones

- Los resultados de esta tesis muestran el impacto de técnicas de balanceo de datos para mejorar el rendimiento de los modelos de aprendizaje para la clasificación y regresión en el problema la identificación de pares LASA.
- La obtención de modelos que mejoran los resultados del estado del arte para la clasificación y regresión en el problema la identificación de pares LASA.

5.2 Trabajo Futuro

A partir de los resultados obtenidos se abren nuevas interrogantes que pueden ser exploradas por el lector o por un tercero en futuras investigaciones o proyectos. El primero, que, a raíz de obtener los mejores valores en las métricas de evaluación, es un análisis de los modelos entrenados con el algoritmo de bosque aleatorio. Al realizar su entrenamiento con la creación de varios árboles de decisión de baja profundidad, se propone la recuperación de cada uno de los árboles creados para hacer una generalización de qué es lo que se considera para la creación de un árbol que pueda identificar pares LASA.

La segunda propuesta es aluza a mantener la indagación y experimentación con algoritmos de aprendizaje supervisado diferentes a los utilizados en este proyecto a su vez de la aplicación de diferentes técnicas de balanceo, con una premisa de existir algún algoritmo de aprendizaje supervisado que en conjunto de una técnica de balanceo se obtenga mejores resultados a los presentados.

Por último, se considera la posibilidad de usar técnicas de optimización mediante metaheurísticas como se hizo en (Millán-Hernández et al., 2020b; Millán-Hernández, García-Hernández, & Ledeneva, 2019b) para el ajuste de pesos. El distintivo radica en enfocar la optimización al balanceo del conjunto en lugar del ajuste de pesos. Con la finalidad de generar un conjunto más pequeño y que a su vez sea lo suficientemente representativo para mejorar la identificación de pares LASA.

Referencias.

Abdellatif, A., Bagian, J. P., Barajas, E. R., Cohen, M., Cousins, D., Denham, C. R., Essinger, K., Gegelashvili, G., Glenister, H., Hoffman, C., & others. (2007). Look-alike, sound-alike medication names: Patient safety solutions, volume 1, solution 1, May 2007. *Joint Commission Journal on Quality and Patient Safety*, *33*(7), 430–433.

Australia, H. (2022). *Generic vs. brand-name medicines*. https://www.healthdirect.gov.au/generic-medicines-vs-brand-name-medicines

Ayush Pant. (2019). *Introduction to Logistic Regression*. Towards Data Science. https://towardsdatascience.com/introduction-to-logistic-regression-66248243c148

Batista, G. E. A. P. A., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*, *6*(1), 20–29. https://doi.org/10.1145/1007730.1007735

Chen, L.-C., Chen, C.-H., Chen, H.-M., & Tseng, V. S. (2011a). Hybrid data mining approaches for prevention of drug dispensing errors. *Journal of Intelligent Information Systems*, *36*(3), 305–327. https://doi.org/10.1007/s10844-009-0107-6

Chen, L.-C., Chen, C.-H., Chen, H.-M., & Tseng, V. S. (2011b). Hybrid data mining approaches for prevention of drug dispensing errors. *Journal of Intelligent Information Systems*, *36*(3), 305–327. https://doi.org/10.1007/s10844-009-0107-6

Devore, J. L. (2012). *Probabilidad y Estadística para ingenierías y ciencias* (6th ed.).

Emmerton, L. M., & Rizk, M. F. S. (2012). Look-alike and sound-alike medicines: Risks and "solutions." *International Journal of Clinical Pharmacy*, *34*(1), 4–8. https://doi.org/10.1007/S11096-011-9595-X/TABLES/2

FDA. (2023a). *Medicamentos Genéricos: Preguntas y Respuestas | FDA*. https://www.fda.gov/drugs/generic-drugs/medicamentos-genericos-preguntas-y-respuestas

FDA. (2023b). *Phonetic and Orthographic Computer Analysis (POCA) Program | FDA*. https://www.fda.gov/drugs/information-industry-drugs/phonetic-and-orthographic-computer-analysis-poca-program

Gadd, T. N. (1990). PHONIX: The algorithm. In *Program* (Vol. 24, Issue 4, pp. 363–366). https://doi.org/10.1108/eb047069

García Abad Joaquín. (2021). *Comparativa de balanceo de datos. Aplicación a un caso real para la predicción de fuga de clientes*. Universidad de Oviedo.

Grigori Sidorov. (2013). N-GRAMAS SINTÁCTICOS Y SU USO EN LA LINGÜÍSTICA COMPUTACIONAL. *Revista Vectores de Investigación*, *6*.

Gupta, S., Srivastava, A. P., & Awasthi, S. (2014). Fast and Effective Searches of Personal Names in an International Environment. In *International Journal of Innovative Research in Engineering & Management (IJIREM)* (Vol. 1, Issue 1).

Gustavo EAPA Batista, Ana LC Bazzan, & Maria Carolina Monard. (2003). Balancing training data for automated annotation of keywords: a case study. *WOB*, 10–18.

Haibo He, Yang Bai, Garcia, E. A., & Shutao Li. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, 1322–1328. https://doi.org/10.1109/IJCNN.2008.4633969

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (Vol. 398). John Wiley & Sons.

IBM. (2023). *What is Random Forest?* IBM. https://www.ibm.com/topics/random-forest

Inderjeet Mani., & I Zhang. (2003). *KNN Approach to Unbalanced Data distributions: A Case Study involving Information Extraction* (Vol. 126).

ISMP. (2023a). *Institute for Safe Medication Practices (ISMP). ISMP List of Confused Drug Names...* https://www.ismp.org/recommendations/confused-drug-names-list

ISMP. (2023b). *List of Confused Drug Names | Institute for Safe Medication Practices*. https://www.ismp.org/recommendations/confused-drug-names-list

ISMP Brasil. (2014). *Veja a lista completa a partir da página 5 ou acesse: www.boletimismpbrasil.org NOMES DE MEDICAMENTOS COM GRAFIA OU SOM SEMELHANTES: COMO EVITAR OS ERROS?* www.ismp-brasil.org

ISMP España. (2020). *LISTA DE NOMBRES SIMILARES DE MEDICAMENTOS QUE SE PRESTAN A CONFUSIÓN*.

Ivan Tomek. (1976). Two modifications of cnn. In *IEEE Trans. Systems, Man and Cybernetics* (Vol. 6, pp. 769–772).

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255–260. https://doi.org/10.1126/science.aaa8415

Jordan, M., Kleinberg, J., Schölkopf, B., & Bishop, C. (2006). *Pattern Recognition and Machine Learning*.

Kondrak, G., & Dorr, B. (2006a). Automatic identification of confusable drug names. *Artificial Intelligence in Medicine*, *36*(1), 29–42. https://doi.org/10.1016/J.ARTMED.2005.07.005

Kondrak, G., & Dorr, B. (2006b). Automatic identification of confusable drug names. *Artificial Intelligence in Medicine*, *36*(1), 29–42. https://doi.org/10.1016/J.ARTMED.2005.07.005

Lambert, B. L., Lin, S.-J., Chang, K.-Y., & Gandhi, S. K. (1999a). Similarity As a Risk Factor in Drug-Name Confusion Errors. *Medical Care*, *37*(12), 1214–1225. https://doi.org/10.1097/00005650-199912000-00005

Lambert, B. L., Lin, S.-J., Chang, K.-Y., & Gandhi, S. K. (1999b). Similarity As a Risk Factor in Drug-Name Confusion Errors. *Medical Care*, *37*(12), 1214–1225. https://doi.org/10.1097/00005650-199912000-00005

Lambert, B. L., Yu, C., & Thirumalai, M. (2004a). A System for Multiattribute Drug Product Comparison. In *Journal of Medical Systems* (Vol. 15, Issue 1).

Lambert, B. L., Yu, C., & Thirumalai, M. (2004b). A System for Multiattribute Drug Product Comparison. In *Journal of Medical Systems* (Vol. 15, Issue 1).

Menardi, G., & Torelli, N. (2014). Training and assessing classification rules with imbalanced data. *Data Mining and Knowledge Discovery*, *28*(1), 92–122. https://doi.org/10.1007/s10618-012-0295-5

Microsoft Azure. (2023). *¿Qué es la inteligencia artificial? | Microsoft Azure*. https://azure.microsoft.com/es-es/resources/cloud-computing-dictionary/what-is-artificial-intelligence/#how

Millán-Hernández, C. E., García-Hernández, R. A., & Ledeneva, Y. (2019a). An evolutionary logistic regression method to identify confused drug names. *Journal of Intelligent & Fuzzy Systems*, *36*(5), 4609–4619. https://doi.org/10.3233/JIFS-179012

Millán-Hernández, C. E., García-Hernández, R. A., & Ledeneva, Y. (2019b). An evolutionary logistic regression method to identify confused drug names. *Journal of Intelligent & Fuzzy Systems*, *36*(5), 4609–4619. https://doi.org/10.3233/JIFS-179012

Millán-Hernández, C. E., García-Hernández, R. A., & Ledeneva, Y. (2020a). Improving the identification of confused drug names in Spanish. *Journal of Intelligent and Fuzzy Systems*, *39*(2), 2027–2036. https://doi.org/10.3233/JIFS-179869

Millán-Hernández, C. E., García-Hernández, R. A., & Ledeneva, Y. (2020b). Improving the identification of confused drug names in Spanish. *Journal of Intelligent and Fuzzy Systems*, *39*(2), 2027–2036. https://doi.org/10.3233/JIFS-179869

Millán-Hernández, C. E., García-Hernández, R. A., Ledeneva, Y., & Hernández-Castañeda, Á. (2019). Soft Bigram Similarity to Identify Confusable Drug Names. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *11524 LNCS*, 433–442. https://doi.org/10.1007/978-3-030-21077-9_40

Müller, A. C., & Guido, S. (2016). *Introduction to Machine Learning with Python A GUIDE FOR DATA SCIENTISTS Introduction to Machine Learning with Python*.

Müller, A. C., & Guido, S. (2017). *Introduction to Machine Learning with Python A GUIDE FOR DATA SCIENTISTS Introduction to Machine Learning with Python*.

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, & W. Philip Kegelmeyer. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*, 321–357.

Pfizer. (2023). *Ever Wonder How Drugs Are Named? Read On | Pfizer*. https://www.pfizer.com/news/articles/ever_wonder_how_drugs_are_named_read_on

Rocco, C., & Garrido, A. (2017). SEGURIDAD DEL PACIENTE Y CULTURA DE SEGURIDAD. *Revista Médica Clínica Las Condes*, *28*(5), 785–795. https://doi.org/10.1016/J.RMCLC.2017.08.006

Rocío, Q. F. B., Vázquez, M., Tomás, M. F., Cruz, D., Héctor, Q. F. B., Schoelly, S., Antonio, J., & Forzán, K. (2018). Errores de Medicación con Medicamentos L.A.S.A. *Boletín CIM 2018-1*.

Russell, S. J. (Stuart J., Norvig, Peter., Corchado Rodríguez, J. Manuel., & Joyanes Aguilar, Luis. (2004). *Inteligencia artificial: un enfoque moderno*. Pearson Prentice Hall.

scikit-learn. (2023). *Metrics and scoring: quantifying the quality of predictions*. Scikit-Learn 1.3.0 Documentation. https://scikit-learn.org/stable/modules/model_evaluation.html

Sidey-Gibbons, J. A. M., & Sidey-Gibbons, C. J. (2019). *Machine learning in medicine: a practical introduction*. https://doi.org/10.1186/s12874-019-0681-4

Soofi, A. A., & Awan, A. (2017). Classification Techniques in Machine Learning: Applications and Issues. *Journal of Basic & Applied Sciences*, *13*, 459–465.

Ulrich Pfeifer, Thomas Poersch, & Norbert Furh. (1996). Retrieval Effectivenes of Proper Name Search Method. *Information Processing & Management*, *32*, 667–679.

USP. (2001). *USP quality review (76). US Pharmacopeia*. https://www.pbm.va.gov/vacenterformedicationsafety/othervasafetyprojects/appendixiiusplasa.pdf

Vázquez, E. V., Ledeneva, Y., & García-Hernández, R. A. (2020). Combination of similarity measures based on symbolic regression for confusing drug names identification. *Journal of Intelligent & Fuzzy Systems, 39*(2), 2093–2103. https://doi.org/10.3233/JIFS-179875

Wilson, D. L. (1972). Asymptotic Properties of Nearest Neighbor Rules Using Edited Data. In *EEE Transactions on Systems, Man, and Cybernetics* (Issue 2, pp. 408–421). https://sci2s.ugr.es/keel/dataset/includes/catImbFiles/1972-Wilson-IEEETSMC.pdf

Apéndices

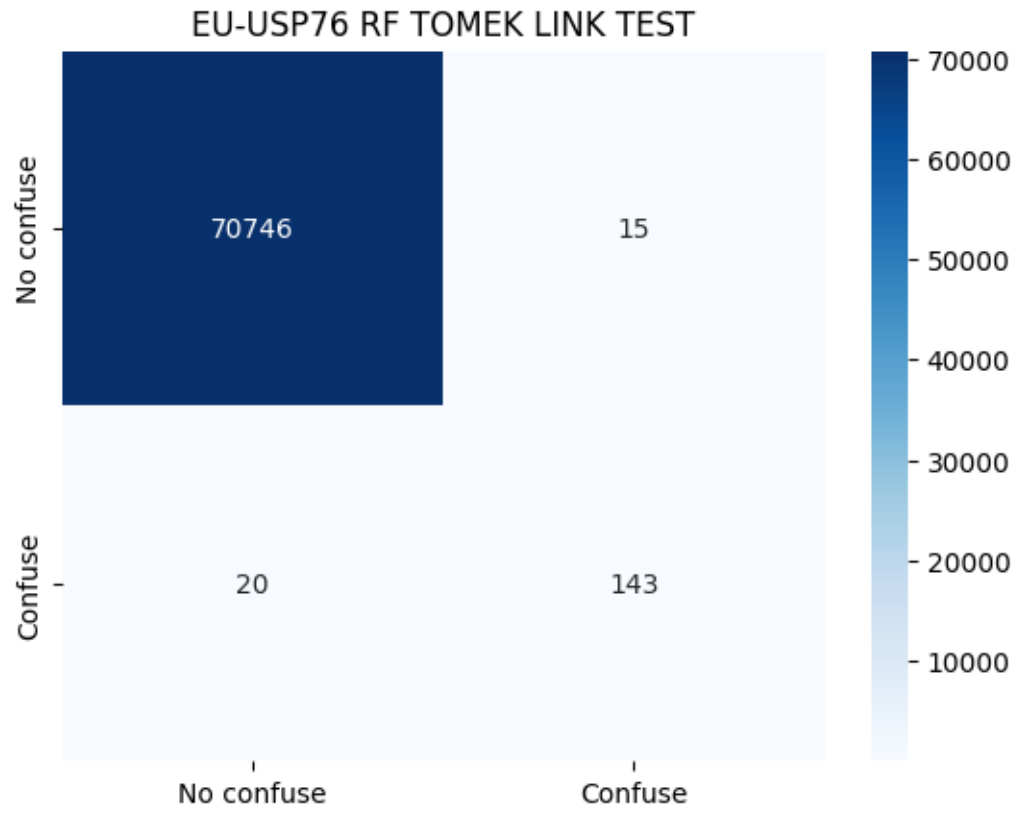
Apéndice A-1

El apéndice A-1 corresponde a los resultados obtenidos en los experimentos con el conjunto de datos EU-USP76.

Matriz de confusión conjunto de Prueba

Figura 0-1

Matriz de confusión Bosque Aleatorio (RF) con Tomek's Link conjunto de Prueba



Pruebas ANOVA EU-USP76 Clasificación

La prueba ANOVA fue utilizada para determinar la existe una diferencia significativa entre las técnicas de balanceo de datos para el entrenamiento de modelos de clasificación con el conjunto de datos EU-USP76. Para realizar una prueba ANOVA es necesario determinar si existe una distribución normal entre los resultados de, la cual se realizó mediante una prueba de normalidad Shapiro-Wilk.

Tabla 0-1

Prueba de normalidad Shapiro-Wilk para F1 obtenidos de modelos a los que se aplicó técnicas de balanceo sobre conjunto EU-USP76

Muestra	W	p-value	Normalidad
TL	0.910726	0.400925	True
AS	0.876037	0.209431	True
SM	0.877130	0.214010	True
SM-TL	0.876130	0.209818	True
ENN	0.817962	0.061376	True
SM-ENN	0.891333	0.281690	True
RS	0.833385	0.086137	True
NM	0.816275	0.059115	True

Debido a que los datos pasan la prueba de normalidad se puede proceder con la prueba ANOVA obteniendo un P-*value*.

Tabla 0-2

Prueba de ANOVA aplicada sobre los F1 obtenidos por las técnicas de balanceo sobre EU-USP-76

Suma de cuadrados	Grados de libertad	Media cuadrada	F-values	p-unc	np2
1.821313	7	0.260188	4.167048	0.00118	0.377991

Pruebas ANOVA EU-USP76 Regresión

La prueba ANOVA fue utilizada para determinar la existe una diferencia significativa entre las técnicas de balanceo de datos para el entrenamiento de modelos de regresión con el conjunto de datos EU-USP76. Para realizar una prueba ANOVA es necesario determinar si existe una distribución normal entre los resultados de, la cual se realizó mediante una prueba de normalidad Shapiro-Wilk.

Tabla 0-3

Prueba de normalidad Shapiro-Wilk para F-macro-averaging obtenidos de modelos a los que se aplicó técnicas de balanceo sobre EU-USP76

Muestra	W	p-value	Normalidad
TL	0.894509	0.299016	True
AS	0.820513	0.064952	True
SM	0.852636	0.129897	True
SM-TL	0.881908	0.235054	True
ENN	0.819233	0.063134	True
SM-ENN	0.905916	0.368318	True
RS	0.763193	0.017391	False
NM	0.813270	0.055278	True

Debido a que RS no pasa la prueba de normalidad, se determina la homocedasticidad para determinar una igualdad de varianzas con el método Levene y poder proceder con la prueba ANOVA.

Tabla 0-4

Prueba Levene para determinar homocedasticidad entre los datos sobre EU-USP76

Levene Test	P-value	Igualdad de varianzas
1.940053	0.083597	True

Al demostrar homocedasticidad en los datos, se procede con la prueba ANOVA, obteniendo un P-value.

Tabla 0-5

Prueba de ANOVA aplicada sobre e1 F-macro-averaging obtenidos por las técnicas de balanceo sobre EU-USP76

Suma de cuadrados	Grados de libertad	Media cuadrada	F-values	p-unc	np2
1.821313	7	0.260188	4.167048	0.00118	0.377991

Apéndice B-1

El apéndice B-1 corresponde a los resultados obtenidos en los experimentos con el conjunto de datos ESP-ISMP2018

Experimentos: ESP-ISMP2018-Entrenamiento: sin balanceo de clases, *undersampling*, *oversampling* y *combination sampling*.

Tabla 0-6

Resultados de clasificación con k-fold en ESP-ISMP2018 -Entrenamiento desbalanceado

Modelo	**Entrenamiento**		**Prueba**	
	F1 LASA	**Desv. Est.**	**F1 LASA**	**Desv. Est.**
RF	**0.99**	**0.00049**	**0.87**	**0.03896**
DT	0.99	0.00037	0.80	0.03051
KNN	0.54	0.01010	0.43	0.09090
LR	**0.41**	**0.00742**	**0.41**	**0.07332**
MLP	0.40	0.06050	0.39	0.08806
GBRT	0.30	0.12554	0.26	0.13385
SVM-L	0.28	0.01132	0.29	0.07160

Nota. Resultados de evaluación mediante validación cruzada k-folds de los siete modelos utilizados (Bosque Aleatorio (RF), Árboles de Decisión (DT), K-Vecinos más cercanos (KNN), Regresión Logística (LR), Perceptrón Multicapa (MLP), Árboles de Decisión con Impulso de Gradiente (GBRT) y Máquina de Soporte Vectorial Lineal (SVM-L)) con el promedio de la métrica F1 para los pares LASA (clase 1) sobre el conjunto ESP-ISMP2018-Entrenamiento desbalanceado

Tabla 0-7

Resultados de regresión con k-fold en ESP-ISMP2018-Entrenamiento desbalanceado

Modelo	**Entrenamiento**		**Prueba**	
	MSE	**Desv. Est.**	**MSE**	**Desv. Est.**
RF-R	**0.00009**	**0.000003**	**0.00065**	**0.000106**
DT-R	0.0	0.000001	0.00093	0.000184
KNN-R	0.00096	0.000014	0.00152	0.000254
LR-R	**0.00153**	**0.000030**	**0.00155**	**0.000280**
MLP-R	0.00155	0.000050	0.00157	0.000263
GBRT-R	0.00113	0.000025	0.00137	0.000241
SVM-L-R	0.00228	0.000034	0.00228	0.000310

Nota. Resultados de evaluación en validación cruzada k-folds de los modelos utilizados (Regresor de Bosque Aleatorio (RF-R), Regresor de Árboles de Decisión (DT-R), Regresor de K-Vecinos más cercanos (KNN-R), Regresión Logística (LR-R), Regresor de Perceptrón Multicapa (MLP-R), Regresor de Árboles de Decisión con Impulso de Gradiente (GBRT-R) y Regresor de Máquina de Soporte Vectorial Lineal (SVM-L-R)) con el promedio de la métrica MSE sobre el conjunto ESP-ISMP2018 desbalanceado

Tabla 0-8

Resultados de clasificación con k-fold en ESP-ISMP2018-Entrenamiento con undersampling

Modelo	**Undersampling**	**Entrenamiento**		**Prueba**	
		F1 LASA	**Desv. Est.**	**F1 LASA**	**Desv. Est.**
RF	**TL**	**0.99**	**0.00050**	**0.86**	**0.03035**
DT	TL	0.99	0.00037	0.78	0.02551
RF	ENN	0.99	0.00037	0.75	0.02513
DT	ENN	0.99	0.00028	0.71	0.02030
KNN	ENN	0.68	0.01223	0.47	0.07059
KNN	TL	0.55	0.00883	0.44	0.09385
MLP	ENN	0.55	0.04986	0.47	0.06446
LR	**ENN**	**0.53**	**0.01091**	**0.47**	**0.05769**
GBRT	ENN	0.38	0.13781	0.32	0.13439
LR	TL	0.41	0.00850	0.41	0.07083
MLP	TL	0.42	0.08131	0.43	0.08709
SVM-L	ENN	0.47	0.00999	0.43	0.07895
GBRT	TL	0.27	0.09201	0.24	0.12567
SVM-L	TL	0.29	0.01071	0.29	0.06826
RF	RS	1.0	0.0	0.11	0.01681
KNN	RS	0.97	0.00413	0.10	0.02121
SVM-L	RS	0.96	0.00316	0.10	0.01340
MLP	RS	0.96	0.00428	0.10	0.01332
LR	RS	0.96	0.00264	0.10	0.01658
GBRT	RS	0.99	0.00083	0.11	0.01987
DT	RS	1.0	0.0	0.08	0.00682
LR	NM	0.87	0.01317	0.15	0.02143
SVM-L	NM	0.87	0.01015	0.13	0.01784
RF	NM	0.99	0.00028	0.0	0.00960
DT	NM	0.99	0.00028	0.0	0.00224
GBRT	NM	0.95	0.00506	0.0	0.00530
MLP	NM	0.90	0.01652	0.04	0.02356
KNN	NM	0.88	0.01038	0.06	0.01627

Nota. Resultados de evaluación mediante validación cruzada k-folds de los modelos utilizados (Bosque Aleatorio (RF), Árboles de Decisión (DT), K-Vecinos más cercanos (KNN), Regresión Logística (LR), Perceptrón Multicapa (MLP), Árboles de Decisión con Impulso de Gradiente (GBRT) y Máquina de Soporte Vectorial Lineal (SVM-L)) con el promedio de la métrica F1para los pares LASA (clase 1) sobre el conjunto ESP-ISMP2018-Entrenamiento balanceado mediante undersampling: selección aleatoria (RS), Near miss (NM), Tomek's Link (TL) y Edited Nearest Neighbours (ENN).

Tabla 0-9

Resultados de regresión con k-fold en ESP-ISMP2018-Entrenamiento con undersampling

Modelo	**Undersampling**	**Entrenamiento**		**Prueba**	
		MSE	**Desv. Est.**	**MSE**	**Desv. Est.**
RF-R	**TL**	**0.00009**	**0.000003**	**0.00066**	**0.000100**
DT-R	TL	0.0	0.000001	0.00098	0.000144
RF-R	ENN	0.00007	0.000002	0.00098	0.000128
GBRT-R	TL	0.00112	0.000022	0.00138	0.000248
KNN-R	TL	0.00095	0.000014	0.00152	0.000251
GBRT-R	ENN	0.00091	0.000023	0.00149	0.00264
LR-R	TL	0.00153	0.000031	0.00155	0.000280
DT-R	ENN	0.0	0.0	0.00157	0.000291
KNN-R	ENN	0.00059	0.000009	0.00166	0.000224
MLP-R	TL	0.00154	0.000048	0.00157	0.000286
LR-R	**ENN**	**0.00129**	**0.000030**	**0.00162**	**0.000265**
MLP-R	ENN	0.00133	0.000043	0.00159	0.000297
SVM-L-R	ENN	0.00229	0.000034	0.00228	0.000310
SVM-L-R	TL	0.00228	0.000034	0.00228	0.000310
LR-R	RS	0.02662	0.002080	0.02968	0.002291
SVM-L-R	RS	0.05804	0.003710	0.03872	0.001228
KNN-R	RS	0.02000	0.003714	0.03263	0.002471
RF-R	RS	0.00300	0.000406	0.02878	0.002172
GBRT-R	RS	0.00264	0.000618	0.03092	0.002738
MLP-R	RS	0.03277	0.002822	0.03375	0.003546
DT-R	RS	0.0	0.0	0.04826	0.005422
LR-R	NM	0.09069	0.007461	0.01953	0.002905
SVM-L-R	NM	0.11260	0.008008	0.07943	0.010491
MLP-R	NM	0.08239	0.008504	0.07248	0.007217
KNN-R	NM	0.06903	0.005863	0.06878	0.014751
RF-R	NM	0.00606	0.000572	0.40727	0.104922
GBRT-R	NM	0.03749	0.003662	0.77135	0.337864
DT-R	NM	0.00050	0.000186	0.69700	0.215661

Nota. Resultados de evaluación en validación cruzada k-fold de los modelos utilizados (Regresor de Bosque Aleatorio (RF-R), Regresor de Árboles de Decisión (DT-R), Regresor de K-Vecinos más cercanos (KNN-R), Regresión Logística (LR-R), Regresor de Perceptrón Multicapa (MLP-R), Regresor de Árboles de Decisión con Impulso de Gradiente (GBRT-R) y Regresor de Máquina de Soporte Vectorial Lineal (SVM-L-R)) con el promedio de la métrica MSE sobre el conjunto ESP-ISMP2018 utilizando técnicas de undersampling: selección aleatoria (RS), Near miss (NM), Tomek's Link (TL) y Edited Nearest Neighbour (ENN).

Tabla 0-10

Resultados de clasificación con k-fold en ESP-ISMP2018-Entrenamiento con oversampling

Modelo	**Oversampling**	**Entrenamiento**		**Prueba**	
		F1 LASA	**Desv. Est.**	**F1 LASA**	**Desv. Est.**
RF	**SMOTE**	**0.99**	**0.0**	**0.84**	**0.02500**
RF	ADASYN	0.99	0.0	0.84	0.02581
DT	SMOTE	0.99	0.0	0.76	0.02219
DT	ADASYN	0.99	0.0	0.76	0.02777
KNN	SMOTE	0.99	0.00002	0.51	0.03337
KNN	ADASYN	0.99	0.00004	0.51	0.03335
MLP	ADASYN	0.99	0.00057	0.40	0.05192
MLP	SMOTE	0.99	0.00046	0.40	0.03171
GBRT	SMOTE	0.99	0.00040	0.21	0.02015
GBRT	ADASYN	0.99	0.00051	0.20	0.01736
LR	**SMOTE**	**0.97**	**0.00217**	**0.12**	**0.01687**
SVM-L	SMOTE	0.97	0.00211	0.12	0.01548
SVM-L	ADASYN	0.96	0.00274	0.10	0.01399
LR	ADASYN	0.96	0.00266	0.10	0.01410

Nota. Resultados de evaluación mediante validación cruzada k-folds de los modelos: Bosque Aleatorio (RF), Árboles de Decisión (DT), K-Vecinos más cercanos (KNN), Regresión Logística (LR), Perceptrón Multicapa (MLP), Árboles de Decisión con Impulso de Gradiente (GBRT) y Máquina de Soporte Vectorial Lineal (SVM-L) con el promedio de la métrica F1 para los pares LASA (clase 1) sobre el conjunto ESP-ISMP2018-Entrenamiento balanceado mediante oversampling: SMOTE (SM) y ADASYN (AS).

Tabla 0-11

Resultados de regresión con k-fold en ESP-ISMP2018-Entrenamiento con oversampling

Modelo	**Oversampling**	**Entrenamiento**		**Prueba**	
		MSE	**Desv. Est.**	**MSE**	**Desv. Est.**
RF-R	**SMOTE**	**0.00005**	**0.000003**	**0.00087**	**0.000101**
RF-R	ADASYN	0.00005	0.000002	0.00085	0.000121
DT-R	ADASYN	0.0	0.0	0.00116	0.000207
DT-R	SMOTE	0.0	0.0	0.00111	0.000132
KNN-R	SMOTE	0.00070	0.000012	0.00276	0.000226
KNN-R	ADASYN	0.00069	0.000012	0.00276	0.000235
MLP-R	SMOTE	0.00974	0.000504	0.01511	0.000992
MLP-R	ADASYN	0.00982	0.000706	0.01546	0.001561
GBRT-R	SMOTE	0.00990	0.000335	0.01435	0.000726
GBRT-R	ADASYN	0.01077	0.000484	0.01605	0.000888
LR-R	**SMOTE**	**0.02206**	**0.001298**	**0.02412**	**0.001156**
LR-R	ADASYN	0.02529	0.001489	0.02807	0.001251
SVM-L-R	SMOTE	0.06920	0.002442	0.05030	0.001140
SVM-L-R	ADASYN	0.07428	0.003460	0.05115	0.001002

Nota. Resultados de evaluación en validación cruzada k-folds de los modelos utilizados (Regresor de Bosque Aleatorio (RF-R), Regresor de Árboles de Decisión (DT-R), Regresor de K-Vecinos más cercanos (KNN-R), Regresión Logística (LR-R), Regresor de Perceptrón Multicapa (MLP-R), Regresor de Árboles de Decisión con Impulso de Gradiente (GBRT-R) y Regresor de Máquina de Soporte Vectorial Lineal (SVM-L-R)) con el promedio de la métrica MSE sobre el conjunto ESP-ISMP2018 utilizando técnicas de oversampling: SMOTE (SM) y ADASYN (AS).

Tabla 0-12

Resultados de clasificación con k-fold en ESP-ISMP2018-Entrenamiento con combination sampling

Modelo	Com. Samp	Entrenamiento		Prueba	
		F1 LASA	Desv. Est.	F1 LASA	Desv. Est.
RF	**SMTL**	**0.99**	**0.0**	**0.82**	**0.01355**
DT	SMTL	0.99	0.0	0.75	0.02720
RF	SMENN	1.0	0.0	0.67	0.02596
DT	SMENN	1.0	0.0	0.62	0.02181
KNN	SMTL	0.99	0.00003	0.50	0.04863
KNN	SMENN	0.99	0.00001	0.44	0.02900
MLP	SMTL	0.99	0.00043	0.39	0.05438
MLP	SMENN	0.99	0.00110	0.35	0.05687
GBRT	SMTL	0.99	0.00029	0.21	0.02777
GBRT	SMENN	0.99	0.00037	0.21	0.02163
SVM-L	SMTL	0.97	0.00093	0.12	0.01994
LR	**SMTL**	**0.97**	**0.00106**	**0.12**	**0.01905**
SVM-L	SMENN	0.97	0.00220	0.12	0.01568
LR	SMENN	0.97	0.00225	0.12	0.01670

Nota. Resultados de evaluación mediante validación cruzada k-folds de los modelos: Bosque Aleatorio (RF), Árboles de Decisión (DT), K-Vecinos más cercanos (KNN), Regresión Logística (LR), Perceptrón Multicapa (MLP), Árboles de Decisión con Impulso de Gradiente (GBRT) y Máquina de Soporte Vectorial Lineal (SVM-L) con el promedio de la métrica F1 para los pares LASA (clase 1) sobre el conjunto EU-USP76 balanceado mediante combination-sampling: SMOTE-Tomek's Link (SM-TL) y SMOTE-Edited Nearest Neighbours (SM-ENN).

Tabla 0-13

Resultados de regresión con k-fold en ESP-ISMP2018-Entrenamiento con combination sampling

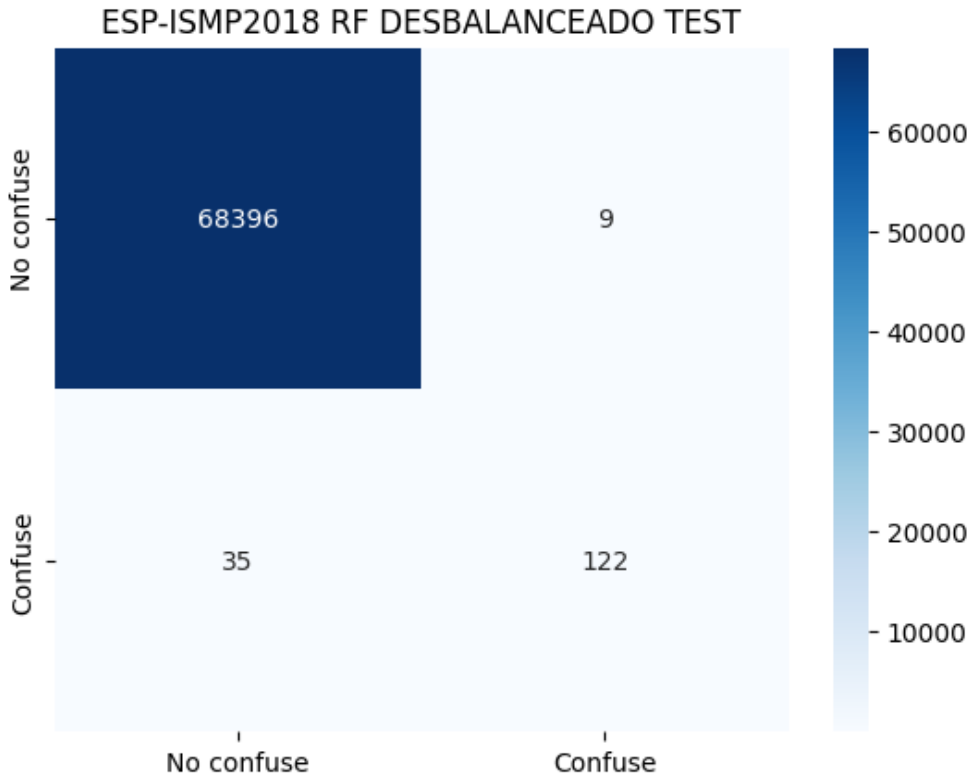
Modelo	Com. Samp	Entrenamiento		Prueba	
		MSE	Desv. Est.	MSE	Desv. Est.
RF-R	**SMTL**	**0.00005**	**0.000003**	**0.00086**	**0.000109**
DT-R	SMTL	0.0	0.0	0.00115	0.000209
RF-R	SMENN	0.00003	0.000002	0.00183	0.000201
DT-R	SMENN	0.0	0.0	0.00243	0.000221
KNN-R	SMTL	0.00070	0.000012	0.00266	0.000227
KNN-R	SMENN	0.00009	0.000005	0.00438	0.000375
MLP-R	SMENN	0.00848	0.000872	0.01470	0.000934
MLP-R	SMTL	0.00959	0.000738	0.01468	0.001847
GBRT-R	SMTL	0.00963	0.000499	0.01438	0.000765
GBRT-R	SMENN	0.00860	0.000305	0.01466	0.000706
LR-R	**SMTL**	**0.02200**	**0.001257**	**0.02407**	**0.001128**
LR-R	SMENN	0.02040	0.001315	0.02468	0.001189
SVM-L-R	SMTL	0.06935	0.002401	0.05030	0.001146
SVM-L-R	SMENN	0.06802	0.002488	0.05034	0.001131

Nota. Resultados de evaluación en validación cruzada k-folds de los modelos utilizados (Regresor de Bosque Aleatorio (RF-R), Regresor de Árboles de Decisión (DT-R), Regresor de K-Vecinos más cercanos (KNN-R), Regresión Logística (LR-R), Regresor de Perceptrón Multicapa (MLP-R), Regresor de Árboles de Decisión con Impulso de Gradiente (GBRT-R) y Regresor de Máquina de Soporte Vectorial Lineal (SVM-L-R)) con el promedio de la métrica MSE sobre el conjunto ESP-ISMP2018 utilizando técnicas de combination-sampling: SMOTE-Tomek's Link (SM-TL) y SMOTE-Edited Nearest Neighbours (SM-ENN).

Matriz de confusión conjunto de Prueba

Figura 0-2

Matriz de confusión Bosque Aleatorio (RF) desbalanceado conjunto de Prueba



Pruebas ANOVA ESP-ISMP2018 Clasificación

La prueba ANOVA fue utilizada para determinar la existe una diferencia significativa entre las técnicas de balanceo de datos para el entrenamiento de modelos de clasificación con el conjunto de datos ESP-ISMP2018. Para realizar una prueba ANOVA es necesario determinar si existe una distribución normal entre los resultados de, la cual se realizó mediante una prueba de normalidad Shapiro-Wilk.

Tabla 0-14

Prueba de normalidad Shapiro-Wilk para F1 obtenidos de modelos a los que se aplicó técnicas de balanceo sobre conjunto ESP-ISMP2018

Muestra	W	p-value	Normalidad
TL	0.842310	0.104398	True
AS	0.876037	0.209431	True
SM	0.877130	0.214010	True
SM-TL	0.876130	0.209818	True
ENN	0.910837	0.401697	True
SM-ENN	0.891333	0.281690	True
RS	0.965365	0.863217	True
NM	0.844381	0.109113	True

Debido a que los datos pasan la prueba de normalidad se puede proceder con la prueba ANOVA y determinar el P-*value*.

Tabla 0-15

Prueba de ANOVA aplicada sobre los F1 obtenidos por las técnicas de balanceo sobre ESP-ISMP2018

Suma de cuadrados	Grados de libertad	Media cuadrada	F-values	p-unc	np2
1.229971	7	0.175710	2.77659	0.016563	0.288215

Pruebas ANOVA ESP-ISMP2018 Regresión

La prueba ANOVA fue utilizada para determinar la existe una diferencia significativa entre las técnicas de balanceo de datos para el entrenamiento de modelos de regresión con el conjunto de datos ESP-ISMP2018. Para realizar una prueba ANOVA es necesario determinar si existe una distribución normal entre los resultados de, la cual se realizó mediante una prueba de normalidad Shapiro-Wilk.

Tabla 0-16

Prueba de normalidad Shapiro-Wilk para F-macro-averaging obtenidos de modelos a los que se aplicó técnicas de balanceo sobre ESP-ISMP2018

Muestra	W	p-value	Normalidad
TL	0.912010	0.409966	True
AS	0.888610	0.267491	True
SM	0.912519	0.413590	True
SM-TL	0.900109	0.331641	True
ENN	0.868144	0.178800	True
SM-ENN	0.885622	0.252596	True
RS	0.866846	0.174158	True
NM	0.882082	0.235851	True

Debido a que los datos pasan la prueba de normalidad se puede proceder con la prueba ANOVA y determinar el P-value.

Tabla 0-17

Prueba de ANOVA aplicada sobre e1 F-macro-averaging obtenidos por las técnicas de balanceo sobre ESP-ISMP2018

Suma de cuadrados	Grados de libertad	Media cuadrada	F-values	p-unc	np2
115.094141	7	16.442020	8.323132	0.000001	0.548286

Apéndice C-1

El apéndice C-1 corresponde a los resultados obtenidos en los experimentos con el conjunto de datos BRA-ISMP2014

Experimentos: BRA-ISMP2014-Entrenamiento: sin balanceo de clases, *undersampling*, *oversampling* y *combination sampling*.

Tabla 0-18

Resultados de clasificación con k-fold en BRA-ISMP-Entrenamiento desbalanceado

Modelo	**Entrenamiento**		**Prueba**	
	F1 LASA	**Desv. Est.**	**F1 LASA**	**Desv. Est.**
RF	**1.0**	**0.0**	**0.89**	**0.02556**
DT	1.0	0.0	0.83	0.05657
GBRT	0.99	0.00498	0.83	0.05677
KNN	0.75	0.01909	0.68	0.16229
MLP	0.67	0.01074	0.62	0.15116
SVM-L	0.67	0.01851	0.63	0.15248
LR	**0.65**	**0.01649**	**0.61**	**0.15205**

Nota. Resultados de evaluación mediante validación cruzada k-folds de los siete modelos utilizados (Bosque Aleatorio (RF), Árboles de Decisión (DT), K-Vecinos más cercanos (KNN), Regresión Logística (LR), Perceptrón Multicapa (MLP), Árboles de Decisión con Impulso de Gradiente (GBRT) y Máquina de Soporte Vectorial Lineal (SVM-L)) con el promedio de la métrica F1 para los pares LASA (clase 1) sobre el conjunto BRA-ISMP2014 desbalanceado.

Tabla 0-19

Resultados de regresión con k-fold en BRA-ISMP2014-Entrenamiento desbalanceado

Modelo	**Entrenamiento**		**Prueba**	
	MSE	**Desv. Est.**	**MSE**	**Desv. Est.**
RF-R	**0.00032**	**0.000020**	**0.00283**	**0.001430**
GBRT-R	0.00076	0.000049	0.00315	0.001837
DT-R	0.0	0.0	0.00423	0.002555
LR-R	**0.00516**	**0.000206**	**0.00527**	**0.001967**
KNN-R	0.00335	0.000193	0.00546	0.002586
MLP-R	0.00568	0.000290	0.00589	0.002002
SVM-L-R	0.01151	0.000456	0.01151	0.004113

Nota. Resultados de evaluación en validación cruzada k-folds de los modelos utilizados (Regresor de Bosque Aleatorio (RF-R), Regresor de Árboles de Decisión (DT-R), Regresor de K-Vecinos más cercanos (KNN-R), Regresión Logística (LR-R), Regresor de Perceptrón Multicapa (MLP-R), Regresor de Árboles de Decisión con Impulso de Gradiente (GBRT-R) y Regresor de Máquina de Soporte Vectorial Lineal (SVM-L-R)) con el promedio de la métrica MSE sobre el conjunto BRA-ISMP2014 desbalanceado.

Tabla 0-20

Resultados de clasificación con k-fold en BRA-ISMP2014-Entrenamiento con undersampling

Modelo	**Undersampling**	**Entrenamiento**		**Prueba**	
		F1 LASA	**Desv. Est.**	**F1 LASA**	**Desv. Est.**
RF	**TL**	**1.0**	**0.0**	**0.87**	**0.06961**
DT	TL	1.0	0.0	0.83	0.06491
GBRT	TL	0.99	0.00506	0.83	0.04533
RF	ENN	1.0	0.0	0.81	0.06698
GBRT	ENN	1.0	0.0	0.79	0.07725
DT	ENN	1.0	0.0	0.76	0.07608
KNN	TL	0.75	0.02003	0.69	0.14320
KNN	ENN	0.88	0.01280	0.66	0.13335
SVM-L	ENN	0.77	0.01557	0.65	0.10634
LR	NM	0.72	0.01732	0.65	0.10922
MLP	TL	0.67	0.01496	0.65	0.14231
MLP	ENN	0.67	0.02349	0.64	0.10132
LR	ENN	0.73	0.01889	0.62	0.14681
SVM-L	TL	0.67	0.01879	0.62	0.17077
LR	TL	0.65	0.01606	0.61	0.15205
SVM-L	NM	0.75	0.01253	0.60	0.12069
LR	RS	0.97	0.00676	0.40	0.08512
RF	RS	1.0	0.0	0.41	0.09503
KNN	RS	0.96	0.00839	0.40	0.08512
SVM-L	RS	0.97	0.01011	0.38	0.08014
GBRT	RS	1.0	0.0	0.36	0.07983
DT	RS	1.0	0.0	0.36	0.10913
MLP	RS	0.96	0.00927	0.35	0.09010
MLP	NM	0.74	0.02201	0.34	0.18819
GBRT	NM	1.0	0.0	0.25	0.18893
DT	NM	1.0	0.0	0.18	0.17828
KNN	NM	0.77	0.01802	0.16	0.11894
RF	NM	1.0	0.0	0.11	0.16459

Nota. Resultados de evaluación mediante validación cruzada k-folds de los modelos utilizados (Bosque Aleatorio (RF), Árboles de Decisión (DT), K-Vecinos más cercanos (KNN), Regresión Logística (LR), Perceptrón Multicapa (MLP), Árboles de Decisión con Impulso de Gradiente (GBRT) y Máquina de Soporte Vectorial Lineal (SVM-L)) con el promedio de la métrica F1para los pares LASA (clase 1) sobre el conjunto BRA-ISMP2014-Entrenamiento balanceado mediante undersampling: selección aleatoria (RS), Near miss (NM), Tomek's Link (TL) y Edited Nearest Neighbours (ENN).

Tabla 0-21

Resultados de regresión con k-fold en BRA-ISMP2014-Entrenamiento con undersampling

Modelo	**Undersampling**	**Entrenamiento**		**Prueba**	
		MSE	**Desv. Est.**	**MSE**	**Desv. Est.**
RF-R	**TL**	**0.00029**	**0.000025**	**0.00283**	**0.001370**
GBRT-R	TL	0.00076	0.000047	0.00346	0.002047
DT-R	TL	0.0	0.0	0.00381	0.002403
GBRT-R	ENN	0.00029	0.000045	0.00465	0.002798
RF-R	ENN	0.00018	0.000025	0.00466	0.002647
LR-R	**TL**	**0.00514**	**0.000194**	**0.00527**	**0.001963**
KNN-R	TL	0.00331	0.000191	0.00546	0.002602
MLP-R	ENN	0.00449	0.000282	0.00576	0.002210
MLP-R	TL	0.00563	0.000290	0.00579	0.002046
LR-R	ENN	0.00376	0.000285	0.00585	0.002488
KNN-R	ENN	0.00152	0.000154	0.00634	0.002972
DT-R	ENN	0.0	0.0	0.00643	0.003705
SVM-L-R	ENN	0.01118	0.000608	0.01121	0.004016
SVM-L-R	TL	0.01151	0.000456	0.01151	0.000411
LR-R	NM	0.17237	0.005050	0.01170	0.001642
KNN-R	RS	0.01772	0.005337	0.02705	0.005584
RF-R	RS	0.00281	0.000448	0.02809	0.003992
LR-R	RS	0.02444	0.005730	0.02894	0.003624
MLP-R	RS	0.02976	0.005014	0.03380	0.006205
MLP-R	NM	0.17156	0.011255	0.03382	0.014994
SVM-L-R	RS	0.05050	0.005795	0.03538	0.004121
GBRT-R	RS	0.0	0.000018	0.03748	0.011360
DT-R	RS	0.0	0.0	0.04098	0.012042
KNN-R	NM	0.13334	0.005443	0.08063	0.051674
SVM-L-R	NM	0.17358	0.005756	0.09611	0.056433
RF-R	NM	0.01027	0.001017	0.20297	0.097254
GBRT-R	NM	0.01039	0.001375	0.24140	0.231780
DT-R	NM	0.0	0.0	0.35569	0.308097

Nota. Resultados de evaluación en validación cruzada k-fold de los modelos utilizados (Regresor de Bosque Aleatorio (RF-R), Regresor de Árboles de Decisión (DT-R), Regresor de K-Vecinos más cercanos (KNN-R), Regresión Logística (LR-R), Regresor de Perceptrón Multicapa (MLP-R), Regresor de Árboles de Decisión con Impulso de Gradiente (GBRT-R) y Regresor de Máquina de Soporte Vectorial Lineal (SVM-L-R)) con el promedio de la métrica MSE sobre el conjunto BRA-ISMP2014 utilizando técnicas de undersampling: selección aleatoria (RS), Near miss (NM), Tomek's Link (TL) y Edited Nearest Neighbours (ENN).

Tabla 0-22

Resultados de clasificación con k-fold en BRA-ISMP2014-Entrenamiento con oversampling

Modelo	**Oversampling**	**Entrenamiento**		**Prueba**	
		F1 LASA	**Desv. Est.**	**F1 LASA**	**Desv. Est.**
DT	**SM**	**1.0**	**0.0**	**0.88**	**0.08536**
RF	AS	1.0	0.0	0.85	0.06118
RF	SM	1.0	0.0	0.83	0.08563
DT	AS	1.0	0.0	0.81	0.08219
GBRT	SM	0.99	0.00046	0.66	0.08667
MLP	AS	0.99	0.00098	0.62	0.09438
GBRT	AS	0.99	0.00024	0.62	0.07366
KNN	SM	0.99	0.00044	0.61	0.11892
KNN	AS	0.99	0.00046	0.61	0.12060
MLP	SM	0.99	0.00075	0.61	0.12009
SVM-L	SM	0.98	0.00126	0.45	0.09810
LR	**SM**	**0.97**	**0.00248**	**0.45**	**0.09982**
SVM-L	AS	0.97	0.00116	0.41	0.09822
LR	AS	0.97	0.00122	0.40	0.09454

Nota. Resultados de evaluación mediante validación cruzada k-folds de los modelos: Bosque Aleatorio (RF), Árboles de Decisión (DT), K-Vecinos más cercanos (KNN), Regresión Logística (LR), Perceptrón Multicapa (MLP), Árboles de Decisión con Impulso de Gradiente (GBRT) y Máquina de Soporte Vectorial Lineal (SVM-L) con el promedio de la métrica F1 para los pares LASA (clase 1) sobre el conjunto BRA-ISMP2014-Entrenamiento balanceado mediante oversampling: *SMOTE (SM) y ADASYN (AS).*

Tabla 0-23

Resultados de regresión con k-fold en BRA-ISMP2014-Entrenamiento con oversampling

Modelo	**Oversampling**	**Entrenamiento**		**Prueba**	
		MSE	**Desv. Est.**	**MSE**	**Desv. Est.**
RF-R	**AS**	**0.00017**	**0.000016**	**0.00367**	**0.001962**
RF-R	SM	0.00017	0.000021	0.00381	0.002030
DT-R	SM	0.0	0.0	0.00457	0.002466
DT-R	AS	0.0	0.0	0.00474	0.002911
KNN-R	SM	0.00251	0.000199	0.00939	0.003114
GBRT-R	SM	0.00358	0.000252	0.00944	0.002665
KNN-R	AS	0.00253	0.000195	0.00954	0.003085
GBRT-R	AS	0.00413	0.000324	0.01109	0.002445
MLP-R	SM	0.01019	0.001464	0.01663	0.002974
MLP-R	AS	0.01021	0.002717	0.01737	0.003417
LR-R	**SM**	**0.01624**	**0.001335**	**0.02092**	**0.003404**
LR-R	AS	0.01873	0.001138	0.02551	0.003822
SVM-L-R	AS	0.06076	0.003198	0.04960	0.004068
SVM-L-R	SM	0.05295	0.001835	0.3344	0.008542

Nota. Resultados de evaluación en validación cruzada k-folds de los modelos utilizados (Regresor de Bosque Aleatorio (RF-R), Regresor de Árboles de Decisión (DT-R), Regresor de K-Vecinos más cercanos (KNN-R), Regresión Logística (LR-R), Regresor de Perceptrón Multicapa (MLP-R), Regresor de Árboles de Decisión con Impulso de Gradiente (GBRT-R) y Regresor de Máquina de Soporte Vectorial Lineal (SVM-L-R)) con el promedio de la métrica MSE sobre el conjunto BRA-ISMP2014 utilizando técnicas de oversampling: SMOTE (SM) y ADASYN (AS).

Tabla 0-24

Resultados de clasificación con k-fold en BRA-ISMP2014-Entrenamiento con combination sampling

Modelo	**Com. Samp**	**Entrenamiento**		**Prueba**	
		F1 LASA	**Desv. Est.**	**F1 LASA**	**Desv. Est.**
RF	SMTL	1.0	0.0	0.88	0.07174
DT	SMTL	1.0	0.0	0.86	0.08844
RF	SMENN	1.0	0.0	0.69	0.08506
DT	SMENN	1.0	0.0	0.67	0.10932
GBRT	SMTL	0.99	0.00040	0.66	0.14634
GBRT	SMENN	0.99	0.00015	0.62	0.11544
MLP	SMTL	0.99	0.00050	0.61	0.10847
MLP	SMENN	0.99	0.00049	0.60	0.11057
KNN	SMTL	0.99	0.00020	0.58	0.12858
KNN	SMENN	0.99	0.00015	0.56	0.09985
SVM-L	SMTL	0.98	0.00261	0.45	0.11194
SVM-L	SMENN	0.98	0.00148	0.44	0.11091
LR	SMTL	0.97	0.00268	0.44	0.11452
LR	SMENN	0.98	0.00243	0.42	0.09322

Nota. Resultados de evaluación mediante validación cruzada k-folds de los modelos: Bosque Aleatorio (RF), Árboles de Decisión (DT), K-Vecinos más cercanos (KNN), Regresión Logística (LR), Perceptrón Multicapa (MLP), Árboles de Decisión con Impulso de Gradiente (GBRT) y Máquina de Soporte Vectorial Lineal (SVM-L) con el promedio de la métrica F1 para los pares LASA (clase 1) sobre el conjunto BRA-ISMP2014-Entrenamiento balanceado mediante combination-sampling: SMOTE-Tomek's Link (SM-TL) y SMOTE-Edited Nearest Neighbours (SM-ENN).

Tabla 0-25

Resultados de regresión con k-fold en BRA-ISMP2014-Entrenamiento con combination sampling

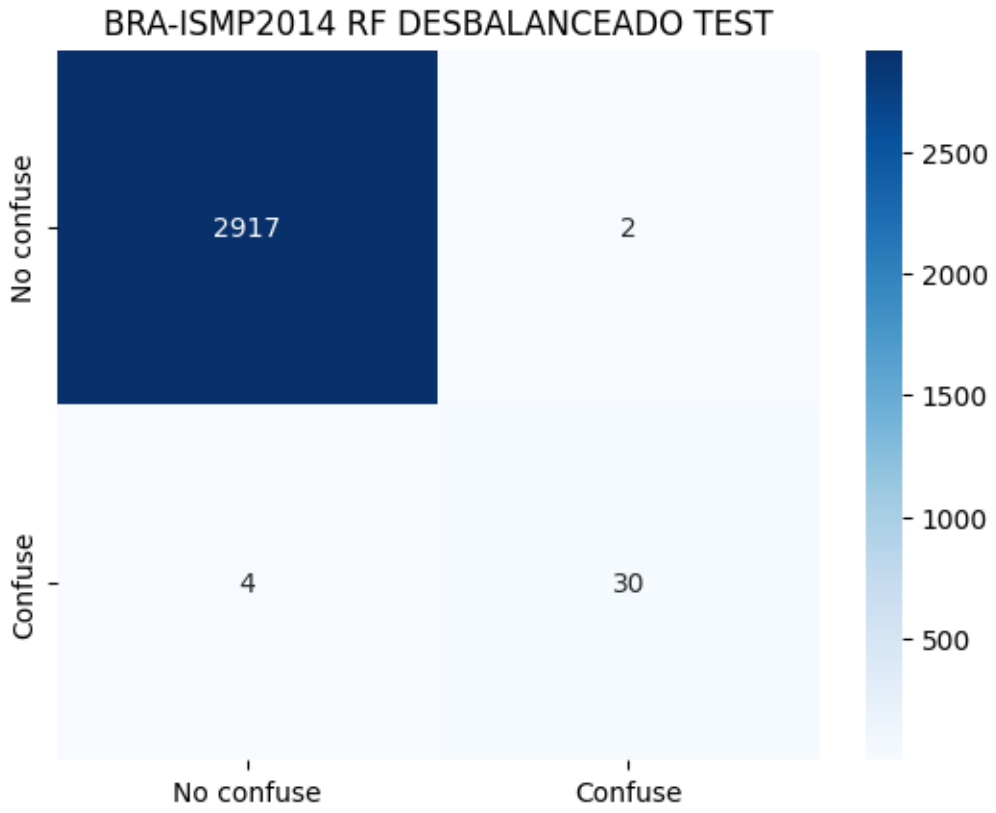
Modelo	**Com. Samp**	**Entrenamiento**		**Prueba**	
		MSE	**Desv. Est.**	**MSE**	**Desv. Est.**
RF-R	**SMTL**	0.00017	0.000021	0.00391	0.002044
DT-R	**SMTL**	0.0	0.0	0.00499	0.002863
RF-R	SMENN	0.00008	0.000021	0.00865	0.003128
KNN-R	**SMTL**	0.00249	0.000189	0.00946	0.003099
GBRT-R	**SMTL**	0.00355	0.000188	0.00951	0.002851
DT-R	SMENN	0.0	0.0	0.01050	0.003507
GBRT-R	SMENN	0.00134	0.000095	0.01154	0.003118
KNN-R	SMENN	0.00037	0.000082	0.01493	0.003548
MLP-R	**SMTL**	0.01119	0.001545	0.01651	0.003882
MLP-R	SMENN	0.00608	0.000597	0.01923	0.003762
LR-R	**SMTL**	0.01626	0.001086	0.02093	0.003587
LR-R	SMENN	0.01090	0.001310	0.02304	0.003473
SVM-L-R	**SMTL**	0.05298	0.001920	0.03330	0.008580
SVM-L-R	SMENN	0.05020	0.002202	0.03376	0.08556

Nota. Evaluación en validación cruzada k-folds de los modelos utilizando los Regresores de: Bosque Aleatorio (RF-R). Árboles de Decisión (DT-R), K-Vecinos más cercanos (KNN-R), Regresión Logística (LR-R), Perceptrón Multicapa (MLP-R), Árboles de Decisión con Impulso de Gradiente (GBRT-R). Máquina de Soporte Vectorial Lineal (SVM-L-R)) con el promedio de la métrica MSE sobre el conjunto BRA-ISMP2014 utilizando técnicas de combination-sampling: SMOTE-Tomek's Link (SM-TL) y SMOTE-Edited Nearest Neighbour (SM-ENN).

Matriz de confusión conjunto de Prueba

Figura 0-3

Matriz de confusión Bosque Aleatorio (RF) Desbalanceado conjunto de Prueba



Pruebas ANOVA BRA-ISMP2014 Clasificación

La prueba ANOVA fue utilizada para determinar la existe una diferencia significativa entre las técnicas de balanceo de datos para el entrenamiento de modelos de clasificación con el conjunto de datos BRA-ISMP2014. Para realizar una prueba ANOVA es necesario determinar si existe una distribución normal entre los resultados de, la cual se realizó mediante una prueba de normalidad Shapiro-Wilk.

Tabla 0-26

Prueba de normalidad Shapiro-Wilk para F1 obtenidos de modelos a los que se aplicó técnicas de balanceo sobre conjunto BRA-ISMP2014

Muestra	W	p-value	Normalidad
TL	0.858579	0.147010	True
AS	0.884049	0.245034	True
SM	0.916382	0.441815	True
SM-TL	0.937086	0.612633	True
ENN	0.872160	0.193860	True
SM-ENN	0.940018	0.638879	True
RS	0.940410	0.642414	True
NM	0.905645	0.366545	True

Debido a que los datos pasan la prueba de normalidad se puede proceder con la prueba ANOVA obteniendo un P-*value*.

Tabla 0-27

Prueba de ANOVA aplicada sobre los F1 obtenidos por las técnicas de balanceo sobre BRA-ISMP2014

Suma de cuadrados	Grados de libertad	Media cuadrada	F-values	p-unc	np2
1.181971	7	0.168853	5.911828	0.000055	0.462984

Pruebas ANOVA BRA-ISMP2014 Regresión

La prueba ANOVA fue utilizada para determinar la existe una diferencia significativa entre las técnicas de balanceo de datos para el entrenamiento de modelos de regresión con el conjunto de datos BRA-ISMP2014. Para realizar una prueba ANOVA es necesario determinar si existe una distribución normal entre los resultados de, la cual se realizó mediante una prueba de normalidad Shapiro-Wilk.

Tabla 0-28

Prueba de normalidad Shapiro-Wilk para F-macro-averaging obtenidos de modelos a los que se aplicó técnicas de balanceo sobre BRA-ISMP2014

Muestra	W	p-value	Normalidad
TL	0.943573	0.671081	True
AS	0.915784	0.437363	True
SM	0.874061	0.201363	True
SM-TL	0.871440	0.191082	True
ENN	0.923571	0.497645	True
SM-ENN	0.932893	0.575765	True
RS	0.925308	0.511742	True
NM	0.923360	0.495949	True

Debido a que los datos pasan la prueba de normalidad se puede proceder con la prueba ANOVA y determinar el P-*value*.

Tabla 0-29

Prueba de ANOVA aplicada sobre e1 F-macro-averaging obtenidos por las técnicas de balanceo sobre BRA-ISMP2014

Suma de cuadrados	Grados de libertad	Media cuadrada	F-values	p-unc	np2
1.821313	7	0.260188	4.167048	0.00118	0.377991