UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Comunicaciones en Estadística
Julio 2025, Vol. 18, No. 1, pp. 51–85

Validación Cruzada: una herramienta crucial para mejorar la eficiencia de modelos de clasificación con datos biomédicos

Cross-Validation: A Crucial Tool to Enhance the Efficiency of Classification Models in the Medical Field

Daniel Felipe Zuleta Fuerte[a]
danielzuleta@usantotomas.edu.co

Osnamir Elias Bru Cordero[b]
oebruc@unal.edu.co

Karina Susana Pastor Sierra[c]
karinapastor@unisinu.edu.co

Resumen

El presente trabajo se centra en la implementación de técnicas de validación cruzada para comparar diversos modelos de clasificación en distintos escenarios relacionados con datos biomédicos. Estos métodos robustos de clasificación son esenciales para garantizar diagnósticos precisos y tratamientos efectivos. Sin embargo, la variabilidad inherente en los datos reales del ámbito biomédico y la complejidad de los conjuntos de datos requieren enfoques sólidos de validación.

Como propuesta investigativa, el estudio aborda la aplicación de técnicas de validación cruzada, incluyendo *k-fold* (validación cruzada con partición en k subconjuntos), *Leave-One-Out* (validación excluyendo una observación por iteración), *StratifiedKFold* (validación estratificada en k subconjuntos) y *Shuffle Split* (validación cruzada basada en particiones aleatorias). Estas técnicas, comúnmente utilizadas, buscan mejorar la precisión y generalización de los modelos de clasificación, así como identificar y mitigar posibles sesgos y problemas de sobreajuste.

Además, se presentan diversos algoritmos de clasificación, como el método de los k vecinos más cercanos (*K-Nearest Neighbors, KNN*), la regresión logística (*Logistic Regression*), bosques aleatorios (*Random Forest*) y los árboles de decisión (*Decision Tree*), para enfrentar los retos asociados a la naturaleza heterogénea de los datos biomédicos recolectados en cuatro municipios de Colombia bajo exposición a metales pesados.

Finalmente, se analiza cómo la validación cruzada puede contribuir a incrementar la robustez de los modelos, favoreciendo una aplicación más efectiva en entornos clínicos. Este artículo busca proporcionar una visión clara y significativa de los métodos de validación cruzada aplicados a algoritmos de clasificación en este tipo de datos, con el objetivo de adoptar modelos que se ajusten a las particularidades del contexto.

***Palabras clave*:** Validación cruzada, *Cross-Validation*, Métodos de clasificación, *Machine learning*, aprendizaje supervisado, aprendizaje no supervizado.

[a]Estudiante Maestría en Estadística Aplicada
[b]Director, Profesor asistente Universidad Nacional de Colombia , sede de La Paz Cesar
[c]Docente investigadora, facultad de ciencias de la salud, laboratorio de investigaciones biomédicas y Biología Molecular, Universidad del Sinú, Montería, Colombia.

Abstract

This study focuses on the implementation of cross-validation techniques to compare various classification models in different scenarios related to biomedical data. These robust classification methods are essential to ensure accurate diagnoses and effective treatments. However, the inherent variability of real-world data in the biomedical field and the complexity of datasets require robust validation approaches.

As an investigative proposal, the study addresses the application of cross-validation techniques, including *k-fold*, *Leave-One-Out*, *StratifiedKFold*, and *Shuffle Split*. These techniques, widely used, aim to enhance the accuracy and generalization of classification models, as well as to identify and mitigate potential biases and overfitting issues.

Additionally, various classification algorithms are presented, such as the *K-Nearest Neighbors, KNN*, *Logistic Regression*, *Random Forest*, and *Decision Tree*, to address the challenges posed by the heterogeneous nature of biomedical data collected from four municipalities in Colombia under exposure to heavy metals.

Finally, this study analyzes how cross-validation can help increase the robustness of models, enabling a more effective application in clinical environments. This article aims to provide a clear and meaningful overview of cross-validation methods applied to classification algorithms for this type of data, with the goal of adopting models tailored to the specific context.

***Keywords*:** Cross-validation, *Cross-Validation*, Classification Methods, *Machine learning*, supervised learning, unsupervised learning..

1. Introducción

1.1. Contexto y justificación

En el campo del aprendizaje automático, los modelos de clasificación desempeñan un papel crucial al permitir la toma de decisiones basadas en datos categorizados. Un modelo de clasificación tiene como objetivo asignar una etiqueta o clase a cada observación dentro de un conjunto de datos, en función de ciertas características o atributos. Este tipo de modelos son ampliamente utilizados en diversas aplicaciones, desde la detección de fraudes en transacciones financieras y el diagnóstico médico, hasta el reconocimiento de imágenes y la categorización de correos electrónicos como spam. La capacidad de estos modelos para aprender patrones a partir de datos históricos y generalizar sobre nuevas entradas los convierte en herramientas poderosas para la toma de decisiones automatizada [1].

Para evaluar un modelo de aprendizaje automático, se emplean métricas como la precisión, exactitud, *recall*, *F1-score* y *AUC-ROC*. Con la validación cruzada, estas métricas se promedian en varias divisiones del conjunto de datos, proporcionando una evaluación más robusta y generalizable. Sin este enfoque, las métricas se calculan en una única partición, lo que puede sesgar los resultados. La validación cruzada permite realizar comparaciones más confiables entre modelos y configuraciones, mostrando resultados en tablas o gráficos comparativos [2].

En este trabajo se aborda de manera detallada la implementación de 4 de estas estas técnicas, *k-fold* (validación cruzada con partición en k subconjuntos), *Leave-One-Out* (validación excluyendo una observación por iteración), *StratifiedKFold* (validación estratificada en k subconjuntos) y *Shuffle Split* (validación cruzada basada en particiones aleatorias), como un método esencial para refinar las estimaciones de diversos modelos de clasificación en bases de datos biomédicas. La investigación se centrará en la aplicación de estas técnicas, reconocidas por su capacidad para mejorar la precisión y robustez de los modelos, así como para identificar y mitigar posibles sesgos y problemas de sobreajuste que puedan surgir en entornos médicos complejos.

Se explorará una variedad de algoritmos de clasificación, como el método de los k vecinos más cercanos (*K-Nearest Neighbors, KNN*), la regresión logística (*Logistic Regression*), bosques aleatorios (*Random*

Forest) y los árboles de decisión (*Decision Tree*), evaluando cómo la validación cruzada contribuye significativamente a la optimización y mejora su aplicación en contextos clínicos. El objetivo principal es ofrecer una comprensión más profunda de los métodos de validación cruzada aplicados a algoritmos de clasificación en datos biomédicos, y cómo estos pueden facilitar la adopción de modelos más eficaces y adaptables [3].

Por otro lado, se hace necesario explicar lo que se entiende por inestabilidad cromosómica (NIC), definida como el desequilibrio genómico que ocurre cuando una célula presenta un número anormal de cromosomas, ha sido un tema de interés en estudios de salud ocupacional, especialmente en contextos de exposición a contaminantes ambientales. Esta condición puede ser causada por eventos como el entrecruzamiento inesperado de cromosomas o la presencia de fragmentos de ADN extra cromosómico [**?**].

Para evaluar esta inestabilidad, la técnica de micronúcleos con bloqueo de citocinesis (*MNBN*) ha sido ampliamente utilizada, especialmente en poblaciones expuestas a metales debido a actividades mineras [4]. El presente estudio tiene como objetivo aplicar métodos avanzados de validación cruzada y técnicas de *machine learning* para la clasificación de daño celular, utilizando datos obtenidos del estudio "Micronuclei frequency and exposure to chemical mixtures in three Colombian mining populations". Los datos de este estudio permiten investigar la relación entre la exposición a contaminantes mineros y la inestabilidad cromosómica en individuos de distintas regiones mineras de Colombia.

Es importante resaltar que, aunque este estudio utiliza datos provenientes del artículo "Micronuclei frequency and exposure to chemical mixtures in three Colombian mining populations", el objetivo principal no es profundizar en los aspectos investigativos relacionados con la frecuencia de micronúcleos o la exposición a mezclas químicas en estas poblaciones. Más bien, el propósito central es llevar a cabo pruebas de concepto que demuestren la relevancia y efectividad de las técnicas de validación cruzada en la evaluación y optimización de modelos de clasificación.

1.2. Antecedentes

La minería artesanal, caracterizada por la exposición a agentes químicos sin protección adecuada, dada su naturaleza simple y de manual, representa un riesgo significativo para la salud de los trabajadores [**?**].

La exposición al mercurio líquido utilizado para separar el oro de la arenilla genera efectos genotóxicos documentados, como la formación de micronúcleos en células epiteliales de la mucosa bucal. Estos micronúcleos, indicadores de daño cromosómico numérico y estructural, permiten evaluar la aneugenicidad y clastogenicidad inducidas por agentes tóxicos. [**?**]

Estudios previos, como el realizado por Rosales-Rimache, evidencian que un 15 % de los trabajadores expuestos al mercurio presentaron micronúcleos, además de otros eventos genotóxicos como puentes nucleoplásmicos y binucleaciones. Dichos hallazgos subrayan la capacidad del mercurio para inducir inestabilidad genética, un mecanismo que, según la literatura, podría aumentar el riesgo de padecer cáncer en poblaciones expuestas. Este tipo de evaluaciones se posiciona como una herramienta clave en programas de vigilancia molecular y epidemiológica para mitigar riesgos en el sector minero artesanal [**?**].

El presente trabajo llena un vacío en la literatura al centrarse en la aplicación de técnicas de validación cruzada para modelos de clasificación basados en la detección de daño celular asociado a la NIC, particularmente en poblaciones expuestas a contaminantes mineros. A diferencia de investigaciones previas que se enfocaron principalmente en el diagnóstico y la medición de la NIC, este estudio profundiza en la optimización de modelos de clasificación utilizando el conteo generado por ténicas como la MNBN para la identificación de poblaciones con posibles NIC.

La validación cruzada (*cross-validation, CV*) es un enfoque ampliamente utilizado en aprendizaje automático para evaluar el rendimiento de los modelos y prevenir problemas como el sobreajuste. En el contexto de datos biomédicos, donde la precisión y la generalización de los modelos son esenciales para la toma de decisiones clínicas, el uso adecuado de técnicas de validación cruzada es aún más crítico. Los modelos que predicen eventos médicos, como la probabilidad de un accidente cerebrovascular o

cualquier condición que afecte la salud humana, deben ser altamente precisos y generalizables para evitar diagnósticos erróneos o subóptimos que puedan tener consecuencias negativas para la vida de los pacientes [2].

En los resultados y discusiones presentados en la investigación realizada por Isaac Kofi Nti, Owusu Nyarko-Boateng y Justice Aning en "Performance of Machine Learning Algorithms with Different K Values in K-fold CrossValidation" [5], se justifica la evaluación de diferentes valores de k (3, 5, 7, 10, 15 y 20) en el desempeño de varios modelos de aprendizaje automático, incluidos máquina de Refuerzo Gradiente (GBM por Gradient Boosting Machine), Regresión Logística (LR), Árbol de Decisión (DT) y K Vecinos más cercano (KNN). Con el artículo se puede concluir que la elección del valor óptimo de k es crucial para equilibrar la precisión del modelo y la complejidad computacional. El estudio demostró que no existe un valor de k generalizable para todos los algoritmos, ya que la respuesta varía dependiendo del modelo y la tarea de clasificación específica [5].

Se observó que el valor de k influye de manera distinta en el rendimiento de los algoritmos evaluados. Por ejemplo, mientras que el Regresión Logística mantuvo una precisión constante sin importar el valor de k (0,959), la métrica de AUC disminuyó levemente a medida que k aumentaba de 3 a 15, sugiriendo que $k = 3$ es el más adecuado para este algoritmo. En cambio, máquina de Refuerzo Gradiente (GBM) y K Vecino más cercano (KNN) mejoraron su rendimiento a medida que k aumentaba hasta 7, siendo este el valor óptimo para ambos. Para el Árbol de Decisión, el mejor rendimiento se obtuvo con $k = 15$. Los valores de k más pequeños, como $k = 3$ o $k = 5$, tienden a tener menor complejidad computacional y ofrecen resultados comparables, mientras que valores más altos, como $k = 15$ o $k = 20$, aunque aumentan la complejidad computacional, no necesariamente mejoran el rendimiento. Por ejemplo, el LOOCV mostró una ligera ventaja en precisión sobre los valores de k analizados (7 y 10), pero su costo computacional es significativamente mayor. En conclusión, el estudio muestra que seleccionar un valor de k adecuado depende del equilibrio entre la precisión del modelo y la complejidad computacional. Aunque $k = 10$ es comúnmente recomendado, los resultados sugieren que $k = 7$ es óptimo para obtener un rendimiento robusto con menor costo computacional [5].

En el estudio realizado por Ilias Tougui , Abdelilah Jilbab , Jamal El Mhamdi, [6], el estudio tiene como objetivo comparar dos estrategias de validación cruzada en el contexto de la inteligencia artificial y el aprendizaje automático aplicados a la salud humana utilizando la tecnología disponible, específicamente en la clasificación de la enfermedad de Parkinson (EP) a partir de grabaciones de audio. Se busca detectar la EP utilizando grabaciones de voz de sujetos diagnosticados con y sin la enfermedad, que fueron recopiladas a través de un estudio clínico móvil. El conjunto de datos utilizado consiste en grabaciones de audio de teléfonos inteligentes de sujetos diagnosticados con EP y controles sanos. Las grabaciones se recopilaron mediante un protocolo que incluía tareas específicas de voz y encuestas demográficas, y se filtraron para asegurar que los sujetos fueran diagnosticados profesionalmente y que sus grabaciones tuvieran un ruido ambiental mínimo.

El estudio dividió el conjunto de datos en dos enfoques: por sujeto y por registro. En el enfoque por sujeto, las grabaciones de cada individuo se incluyeron en un solo conjunto (entrenamiento o reserva), simulando un estudio clínico real y evitando la contaminación de datos. En el enfoque por registro, el conjunto de datos se dividió aleatoriamente sin considerar la pertenencia de los registros a los mismos sujetos, lo que puede sobreestimar el rendimiento del modelo y no refleja adecuadamente su capacidad de generalización en un contexto clínico.

Los resultados mostraron que las técnicas de CV por registro superaron a las de CV por sujeto en términos de precisión; por ejemplo, la precisión utilizando CV por registro fue de aproximadamente 73,5 %, mientras que la precisión utilizando técnicas por sujeto fue significativamente más baja, alrededor del 63 %. Sin embargo, cuando se evaluó el rendimiento en datos no vistos, se observó un mayor error de clasificación utilizando técnicas por registro, indicando un riesgo de sobreajuste. La división por sujeto es más adecuada para reflejar un estudio clínico real, ya que evita la contaminación de los datos, lo que asegura que el conjunto de entrenamiento y el de prueba sean independientes entre sí [6].

En conclusión, la validación cruzada (CV) se presenta como una herramienta esencial en el ámbito del

aprendizaje automático, especialmente al trabajar con datos clínicos. La correcta implementación de técnicas de CV no solo permite evaluar el rendimiento de los modelos en conjuntos de datos limitados, sino que también ayuda a seleccionar bases de datos adecuadas para la predicción en escenarios clínicos reales y no observados.

Por otro lado, al simular la forma en que se recogerían los datos en un estudio clínico, la validación cruzada ayuda a garantizar que los modelos aprendan a clasificar condiciones médicas basándose en características verdaderas de la enfermedad, en lugar de en peculiaridades de los participantes. Además, la revisión del sobreajuste se vuelve crítica, ya que los modelos pueden mostrar un rendimiento optimista en conjuntos de datos que comparten registros del mismo sujeto. Sin una adecuada división de los datos y un manejo cuidadoso del sobreajuste, se corre el riesgo de obtener resultados que no se traducen en un rendimiento efectivo en la práctica clínica. Por lo tanto, es vital que los investigadores presten atención a la selección de la metodología de validación adecuada, asegurando que los modelos sean verdaderamente generalizables y aplicables a situaciones del mundo real. Esto no solo mejora la fiabilidad de los modelos predictivos, sino que también contribuye a la calidad y la seguridad de la atención médica, al permitir decisiones más informadas basadas en evidencia robusta [6].

Concluyendo entonces la validación cruzada ha demostrado ser fundamental para garantizar la generalización de los modelos de clasificación en estudios biomédicos, reduciendo riesgos de sobreajuste y mejorando la precisión de las predicciones en contextos clínicos.

Los modelos de clasificación han sido ampliamente empleados en la predicción y diagnóstico médico, proporcionando herramientas esenciales para la toma de decisiones clínicas. Entre estos, los algoritmos como bosques aleatorios y regresión logística han mostrado un desempeño notable en problemas de clasificación binaria y multicategoría

Akinsola en su estudio sobre algoritmos supervisados de aprendizaje automático, artículo titulado "Supervised Machine Learning Algorithms: Classification and Comparison", utilizó el conjunto de datos de diabetes de PIMA India, con 768 instancias y 9 atributos, explora distintas técnicas de aprendizaje supervisado, centrándose en la comparación de varios algoritmos de clasificación para identificar el más eficiente según las características del conjunto de datos, el número de instancias y variables. Se evaluaron siete algoritmos: Decision Table, Random Forest, Naïve Bayes, Support Vector Machine (SVM), Redes Neuronales (Perceptron), JRip y Decision Tree (J48), utilizando el entorno WEKA para el análisis. los autores analizaron precisión, error absoluto medio (MAE), estadístico kappa y tiempo de ejecución en validación cruzada de 10 pliegues. Los resultados mostraron que el algoritmo SVM presentó el mayor nivel de precisión y exactitud, seguido por los algoritmos de Naïve Bayes y Random Forest. La investigación también destacó que factores como el tiempo requerido para construir el modelo y la precisión influyen en el rendimiento de los algoritmos.

En el análisis con un conjunto de datos más pequeño (384 instancias y seis atributos), los resultados indicaron que los algoritmos SVM y Naïve Bayes ofrecieron los mejores niveles de precisión tanto para el diagnóstico positivo como negativo de diabetes. La comparación entre los algoritmos mostró una diferencia notable en el tiempo de construcción del modelo, siendo el SVM el más rápido y preciso en ambos escenarios. En general, el estudio concluye que la selección adecuada del algoritmo de clasificación y la configuración de sus parámetros son esenciales para lograr modelos supervisados efectivos, especialmente en problemas de clasificación donde se requiere alta precisión y mínima tasa de error. [**?**].

Por otro lado, y en esenarios parecidos dada los finales investigativos y los datos utilizados, Jobeda Jamal Khanam en su articlo "A comparison of machine learning algorithms for diabetes prediction", explora el uso de técnicas de aprendizaje automático (ML) para la predicción de diabetes utilizando el conjunto de datos PIMA India. Se evaluaron siete algoritmos de ML: Decision Tree (DT), k-Nearest Neighbors (KNN), Random Forest (RF), Naive Bayes (NB), AdaBoost (AB), Logistic Regression (LR) y Support Vector Machines (SVM). Los resultados mostraron que los modelos basados en redes neuronales (NN) con dos capas ocultas y 400 épocas alcanzaron una precisión del 88.6 %, lo que representa el mejor rendimiento entre los métodos evaluados. Además, tanto los modelos de LR como los de SVM mostraron precisiones superiores al 78 %, indicando un buen desempeño en comparación con estudios previos. Estos

hallazgos sugieren que los modelos de NN con dos capas ocultas son los más eficaces para el análisis de datos PIMA y la predicción temprana de diabetes, superando los resultados de estudios previos en términos de precisión y generalización. [**?**]

De este modo, podemos concluir que los resultados obtenidos en diferentes estudios varían en función del enfoque metodológico y los ajustes realizados en los experimentos. Cada experimento con el mismo conjunto de datos y algoritmos puede generar resultados diferentes, lo que depende del diseño y la configuración empleada. No podemos asegurar que un modelo sea el mejor para clasificar problemas de este tipo o de cualquier otro tipo en especifico, ya que la selección óptima de un algoritmo depende de diversos factores, como las capacidades computacionales, el diseño del experimento, y los parámetros ajustados. En definitiva, el rendimiento de los modelos varía según el contexto y las decisiones tomadas durante el proceso investigativo [**?**] y [**?**].

2. Objetivos

2.1. Objetivo general

Evaluar la eficacia de técnicas como *k-fold* (validación cruzada con partición en k subconjuntos), *Leave-One-Out* (validación excluyendo una observación por iteración), *StratifiedKFold* (validación estratificada en k subconjuntos) y *Shuffle Split* (validación cruzada basada en particiones aleatorias) en la mejora de la generalización de modelos de clasificación en datos biomédicos.

2.2. Objetivos específicos

1) Evaluar diferentes algoritmos de clasificación: Comparar el desempeño de algoritmos como árboles de decisión (*Árbol de Decisión*), bosques aleatorios (*Random Forest*), vecinos más cercanos (*k-Nearest Neighbors*), regresión logística (*Regresión Logística*), en la clasificación de datos médicos mediante validación cruzada.

2) Identificar y mitigar sesgos y sobreajustes: Utilizar técnicas de validación cruzada para identificar patrones de sobreajuste y sesgos en modelos de clasificación, proponiendo ajustes necesarios para mejorar la generalización del modelo.

3) Aplicar técnicas avanzadas de validación cruzada: Implementar técnicas como la validación cruzada y llevar a cabo pruebas de concepto en conjuntos de datos médicos específicos para evaluar su efectividad.

4) Generar recomendaciones prácticas: Proponer recomendaciones basadas en los resultados obtenidos sobre el uso adecuado de técnicas de validación cruzada para investigadores y profesionales del campo médico interesados en mejorar la precisión de sus modelos de clasificación.

3. Marco teórico

3.1. Inteligencia Artificial

Según Stuart J. Russell y Peter Norvig, la inteligencia artificial se define como el estudio y la construcción de programas de agentes que se desempeñan bien en una clase dada de entornos [7].

La Inteligencia Artificial (IA) según John McCarthy uno de los pioneros en este campo, se define como la ciencia e ingeniería de hacer máquinas inteligentes, especialmente programas informáticos inteligentes [8].

En esencia, la IA busca crear mecanismos que operen eficientemente en entornos específicos, gestionando

la incertidumbre mediante el aprendizaje a partir de datos o cualquier otra fuente de información que les permita generar predicciones [7]. Se puede afirmar que la IA resulta indispensable cuando la capacidad humana para establecer patrones o procesar grandes volúmenes de información es limitada o insuficiente.

La IA no se restringe a imitar procesos biológicamente observables, lo que abre la posibilidad de emplear enfoques que difieran de los métodos humanos. Se espera que las máquinas piensen de manera racional o similar a los humanos, utilizando grandes y robustos conjuntos de datos para resolver problemas [8].

Si bien la inteligencia artificial abarca un amplio espectro de disciplinas destinadas a crear sistemas inteligentes, el aprendizaje automático (Machine Learning) se posiciona como una de sus áreas fundamentales, proporcionando las herramientas y algoritmos necesarios para que las máquinas aprendan de los datos y mejoren su desempeño con el tiempo. A continuación, se exploran los conceptos clave.

3.2. Aprendizaje Automático

El Aprendizaje Automático, conocido como *Machine Learning* (ML), puede entenderse como un proceso que imita la forma en que los humanos adquieren habilidades para adaptarse a diversas situaciones a lo largo de sus vidas. En palabras de Japkowicz, "Machine learning tiene un objetivo similar al proceso humano de aprendizaje, aplicado a las computadoras, enfocado en la perfección de la inferencia inductiva, donde se observa un fenómeno y se generaliza para hacer predicciones sobre fenómenos futuros" [9]. En términos prácticos, se emplea un conjunto de datos con múltiples variables para predecir el comportamiento del objeto de estudio.

Para comprender la relación entre la inteligencia artificial (IA) y el aprendizaje automático (ML), es fundamental explorar conceptos clave de ML. Según McCarthy, "El aprendizaje automático implica el desarrollo de algoritmos que permiten a las computadoras aprender y mejorar su rendimiento basándose en datos. Estos algoritmos pueden emplear técnicas de estadística, teoría de la probabilidad y neurociencia para alcanzar sus objetivos" [8].

En conclusión, la IA y el ML están estrechamente relacionados. Mientras que la IA busca desarrollar mecanismos o entornos inteligentes, el ML proporciona las herramientas y técnicas necesarias para que las máquinas aprendan y mejoren su desempeño de manera continua.

En el mundo del ML existen tres grandes técnicas que son:

- Aprendizaje supervisado: se puede describir como la técnica que utiliza conjuntos de entrenamiento, en los que se tiene control sobre las etiquetas de los datos, para encontrar la relación existente entre ellos.
- Aprendizaje no-supervisado: encontraste del aprendizaje supervisado, en esta técnica no tenemos control de las etiquetas de nuestros datos, por ende, lo que se busca es que se encuentre propiedades únicas en los datos, que permita generar conclusiones a partir del objeto a estudiar, ya sea agrupamientos o clasificaciones, entre otros.
- Aprendizaje por refuerzo: El aprendizaje por refuerzo es una combinación entre el supervisado y no-supervisado, donde el algoritmo o agente aprende interactuando con su entorno, y recibiendo recompensas y castigos según el comportamiento tomado por el agente [10].

3.3. Modelos de clasificación

El presente trabajo, se hará uso del aprendizaje supervisado, el cual podemos ejecutar a través de modelos de clasificación, por lo cual se hace necesario dar una pequeña explicación acerca de los modelos de clasificación y que algoritmos existen.

Los modelos de clasificación se pueden definir, como técnicas o algoritmos dentro del *Machine Learning*, que tienen como objetivo predecir o prever la etiqueta en un conjunto de datos, la clasificación implica

asignar dicha etiqueta a una, observación desconocida, basándose en patrones y relaciones identificadas en las variables predictoras durante el entrenamiento del modelo.

"En clasificación, el objetivo es predecir una etiqueta de clase, que es una elección de una lista predefinida de posibilidades" [3].

Es importante saber que para el presente trabajo se utilizarán diferentes modelos de ML por ende, se hace necesario explicar que es un árbol de decisión, "Los árboles de decisión o de clasificación son un modelo surgido en el ámbito del aprendizaje automático (ML) y de la IA al que, partiendo de una base de datos, crea diagramas de construcciones lógicas que nos ayudan a resolver problemas. A esta técnica también se la denomina segmentación jerárquica" [11]. Podemos también definir los árboles como "método de aprendizaje supervisado no paramétrico que se utiliza para clasificación y regresión. El objetivo es crear un modelo que prediga el valor de una variable objetivo aprendiendo reglas de decisión simples inferidas de las características de los datos. Un árbol puede verse como una aproximación constante por partes." [12]

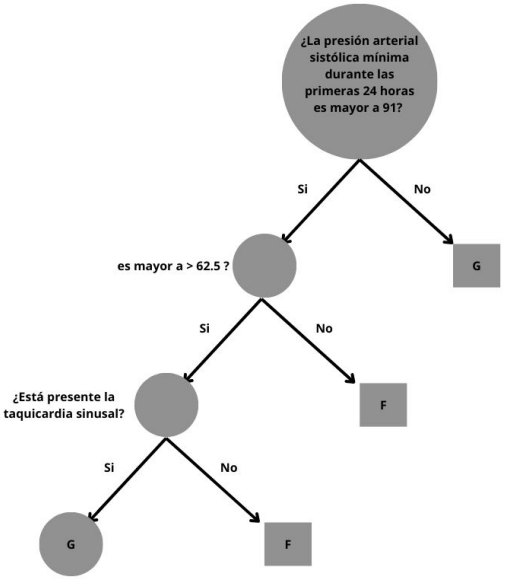

Figura 1: Árbol de decisión Recuperado de [Figura 1.1]. En Breiman, L., *Classification and regression tree*. California: Chapman & Hall. [**?**]

En la Figura 1 el Autor Breiman L, [**?**] expone un ejemplo de cómo trabaja la lógica detrás de un árbol, a partir de diferentes caminos acerca de la condición médica de un paciente, podríamos llegar a determinar si está dentro de un grupo de alto riesgo "G", o si por el contrario, no pertenece a este grupo de alto riesgo "F", dado el conocimiento médico acerca de la condición médica "ataque cardíaco", tenemos en un nodo inicial, ¿Es la presión arterial sistólica mínima durante las primeras 24 horas mayor a 91?, dando como respuesta no, vemos que será clasificado en el grupo de alto riesgo, y si la respuesta es sí, volvemos a realizar segmentación binaria, con otra variable, que en este caso está representado como una pregunta, si el paciente es mayor de 62,5 años, y así sucesivamente.

Resumiendo, acerca de los árboles de decisiones, los podríamos entender como algoritmos de aprendizaje supervisado, que no hacen parte de *"un modelo estadístico basado en la estimación de los parámetros de la ecuación propuesta"* [11], en pocas palabras un método no paramétrico, basado en decisiones jerárquicas, que tienen como ventaja su carácter descriptivo otorgando interpretabilidad de las decisiones tomadas para llevar a cabo la clasificación. Una de las principales ventajas de este método frente a otros es su fácil representación visual, lo que permite identificar claramente los principales segmentadores del modelo y obtener una comprensión más detallada de los datos. Además, no se ve tan afectado por la presencia de valores atípicos que puedan existir en los datos. [13]

En este contexto, resulta fundamental comprender la estructura y los componentes clave de los árboles de decisión, dado que su diseño interno define la forma en que se toman las decisiones jerárquicas para clasificar los datos. Esto permite no solo interpretar fácilmente los resultados obtenidos, sino también identificar los elementos principales que sustentan el modelo.

La estructura de dichos árboles consta de un nodo raíz, ramas, nodos internos y nodos hoja.

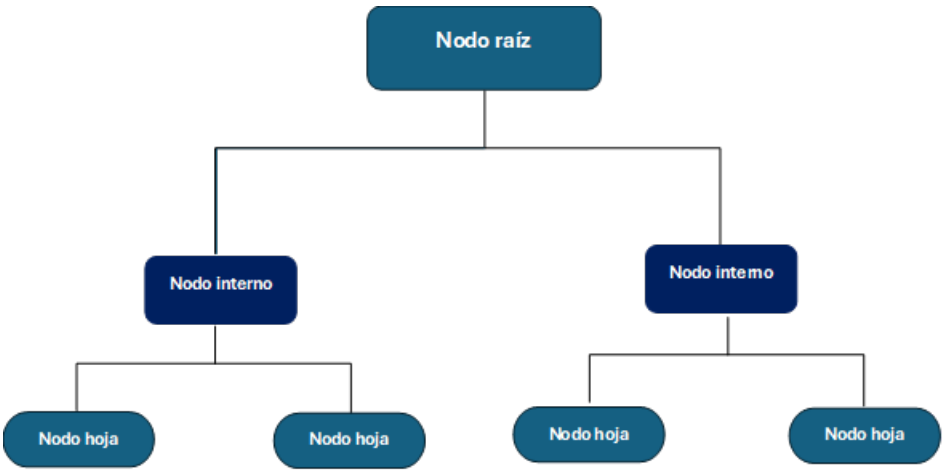

Figura 2: Estructura de árbol de desiciones. Recuperado de https://www.ibm.com/es-es/topics/decision-trees [**?**]

La figura 2, [**?**] nos ayuda a identificar las partes de un árbol de decisiones, siendo el nodo raíz, el nodo principal, con la variable más descriptiva o característica para la segmentación (root node), debajo de este están los nodos internos (internal node), conocidos también como nodos de decisiones, estos están en función a las características disponibles proporcionadas, estos dos nodos realizan evaluaciones que permiten agrupar los individuos dependiendo de su características, las ramas o los conectores de los nodos, son las encargadas de establecer la decisión posible para el nodo interno, estas alimentan a esos nodos, y los nodos hojas o nodos terminales, representan el posible resultado del camino tomado para la decisión.

En este trabajo se emplean algoritmos como el clasificador de bosques aleatorios (Random Forest Classifier), el método de los vecinos más cercanos (Nearest Neighbors) y la regresión logística (Logistic Regression). A continuación, se detallan las características principales de cada uno de estos enfoques.

Random Forest Classifier(RFC): Es un clasificador que consiste en la colección de clasificadores basado en árboles de decisión, cada árbol de entrada recibe una submuestra aleatoria del conjunto de datos, estos árboles en el bosque emiten votos por la clase más popular en la entrada x, y la predicción final se obtiene mediante el promedio de las predicciones de todos los árboles en el bosque [3] y [12].

En otras palabras, podríamos tomar como ejemplo, que debemos resolver un problema complejo, y nuestro algoritmo de RFC, trabaja bajo la premisa de varios árboles donde, cada árbol es diferente a los otros, dada la muestra obtenida para cada árbol, dando como resultado, que cada árbol puede resultar ser mejor para predecir cierta clase de nuestra variable estudio, *"cada árbol de decisión es*

como un experto, proporcionando su opinión sobre cómo clasificar los datos. Las predicciones se realizan calculando la predicción para cada árbol de decisión, luego tomando el resultado más popular" [14]. Una de las principales ventajas del árbol de decisión es que, gracias a su capacidad para realizar estimaciones basadas en múltiples árboles, reduce el riesgo de sobreajuste y la varianza, ya que no depende de una única estimación realizada por un solo árbol de decisión, el cual podría ajustarse demasiado a los datos de entrenamiento [13].

Figura 3: Estructura de Bosques aleatorios Recuperado de https://www.datacamp.com/tutorial/random-forests-classifier-python [14]

En la Figura 3 [14], tenemos un bosque aleatorio con n árboles de decisión, que es una represenacion grafica de como el calsificador de bosques aleatorios, se aproxima a sus estimaciones. Se han mostrado los primeros 5 árboles, junto con sus predicciones (ya sea Perro o Gato). Cada árbol está expuesto a un número diferente de características y una muestra diferente del conjunto de datos original, y como tal, cada árbol generar una predicción diferente, o la misma, basado en las características obtenidas en el conjunto de datos con el que cada árbol se entrenó.

Así como bosques aleatorios utilizan múltiples árboles de decisión entrenados en subconjuntos del conjunto de datos para generar predicciones, otros algoritmos, como K-Nearest Neighbors (KNN) o K vecino más cercano, emplean estrategias distintas para clasificar datos en función de su proximidad a puntos de referencia en el espacio de características. A continuación, se detalla el funcionamiento del algoritmo KNN.

Nearest Neighbors (KNN): El algoritmo K-Nearest Neighbors (KNN) o k vecino más cercano es un clasificador basado en distancias que asigna una categoría a un punto de datos según la proximidad de sus vecinos más cercanos. Utiliza la distancia euclidiana para medir similitudes y predice que las instancias más cercanas tienen más probabilidades de pertenecer a la misma clase. KNN es fácil de implementar, con dos parámetros principales: el número de vecinos k y la métrica de distancia. Sin embargo, puede sufrir sobreajuste cuando k es pequeño y suavizar las predicciones cuando k es grande. "Describimos el algoritmo de KNN, que realiza una predicción asignando la etiqueta de clase o el valor objetivo continuo del ejemplo de entrenamiento más similar al punto de consulta (donde la similitud se mide típicamente utilizando la métrica de distancia euclidiana para características continuas). En lugar de basar la predicción en el único ejemplo de entrenamiento más similar, KNN considera los k vecinos más cercanos al predecir una etiqueta de clase (en clasificación) o un valor objetivo continuo (en regresión)" [15].

Este método presenta varias ventajas frente a otros, no asume una distribución específica de los datos, lo que le permite adaptarse a diferentes problemas sin restricciones previas. Además, agiliza el tiempo de entrenamiento al simplemente almacenar el conjunto de datos y aprender de él al realizar predicciones, lo que proporciona flexibilidad y rapidez al actualizarse con nuevos datos sin necesidad de un reentrenamiento completo [13].

Tras analizar las características y ventajas del algoritmo K-Nearest Neighbors (KNN), es importante abordar otro enfoque ampliamente utilizado en problemas de clasificación: la regresión logística. A diferencia de KNN, que se basa en la proximidad a puntos vecinos, la regresión logística utiliza un modelo probabilístico para predecir la pertenencia a categorías, lo que la hace especialmente adecuada para problemas de clasificación binaria y multinomial [**?**] y [**?**].

Regresión logística: en machine learning es un algoritmo de clasificación utilizado para predecir la probabilidad de que una instancia pertenezca a una de dos categorías posibles (en el caso binario) o más de

dos categorías (en el caso multinomial). A pesar de su nombre, la regresión logística es principalmente un modelo de clasificación y no de regresión en el sentido clásico de predecir valores continuos.

Es importante saber que este algoritmo está especialmente optimizado para la clasificación binaria, y es por esta razón que la regresión logística está basada en una función sigmoidea y no en una regresión lineal clásica. Dado que intentar clasificar una variable binaria en un modelo lineal puede resultar en clasificaciones alejadas de la realidad debido a la influencia de valores atípicos [**?**] y [**?**].

En términos generales, podemos decir que la regresión logística es la transformación logística aplicando la función sigmoide

$$P(x) = \frac{1}{1 + e^{-x}}$$

tomando

$$X = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

que son las combinaciones lineales de las variables independientes, dando como resultado a la regresión logística

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k)}}$$

Donde

- P: La probabilidad estimada de que la instancia pertenezca a la clase positiva.
- e^{-z}: La exponencial del valor negativo de la combinación lineal Z.
- $\frac{1}{1+e^{-z}}$: La función sigmoide que transforma el valor lineal Z en una probabilidad entre 0 y 1.

Dentro de la regresion logística, podemos encontrar diferentes tipos de análisis:

Regresión logística binaria: En este enfoque, la respuesta o variable dependiente es de naturaleza dicotómica, es decir, sólo tiene dos resultados posibles (por ejemplo, 0 o 1)

Regresión Logística Multinomial: Se extiende para manejar más de dos clases utilizando el enfoque `one-vs-rest`(uno contra todos) o "softmax", que calcula las probabilidades de pertenencia a cada clase.

Regresión Logística Ordinal: Se utiliza cuando las categorías de la variable dependiente tienen un orden inherente. Ajusta múltiples umbrales en la función logística para modelar la probabilidad acumulativa de pertenencia a una categoría o a una inferior [**?**], [**?**] y [**?**].

En base a lo ya definido para las metodologias a usar, se hace necesario explicar porque se seleccionaron los algoritmos Decision Tree (*Árbol de Decisión*), bosques aleatorios (*Random Forest*), vecinos más cercanos (*k-Nearest Neighbors*), regresión logística (*Regresión Logística*) para este estudio. Esta elección se fundamenta en que estos modelos representan diferentes metodologías de clasificación: Random Forest y Decision Tree destacan por su enfoque basado en árboles y su capacidad para manejar relaciones complejas y datos heterogéneos; Logistic Regression, como modelo paramétrico, es eficaz para problemas con relaciones lineales; mientras que KNN, de naturaleza no paramétrica, se basa en la proximidad entre instancias para realizar clasificaciones. Esta variedad garantiza una evaluación integral de técnicas de clasificación en el contexto médico

3.4. Métodos de Evaluación y Mejora de Modelos

Dentro de los modelos de clasificación, una de las formas más comunes de evaluar su desempeño es dividir el conjunto de datos disponible en dos partes: una para entrenar el modelo y otra para probar su capacidad de predicción. Este enfoque busca medir qué tan bien el modelo generaliza al enfrentarse a datos nuevos que no se utilizaron durante el entrenamiento. "No estamos interesados en qué tan bien se

ajusta nuestro modelo al conjunto de entrenamiento, sino en qué tan bien puede hacer predicciones para datos que no fueron observados durante el entrenamiento." [3]

Aunque esta técnica es útil, puede no ser suficiente para garantizar una evaluación adecuada en todos los casos. Es aquí donde la validación cruzada, un método estadístico más robusto, cobra importancia. Este enfoque no solo evalúa el rendimiento de generalización del modelo, sino que también contribuye a mejorar sus estimaciones, especialmente en situaciones donde los patrones de los datos no son claros o consistentes.

¿Qué es validación cruzada?

La validación cruzada es una técnica fundamental en la evaluación y mejora de modelos, que supera las limitaciones de la simple división de datos en entrenamiento y prueba. Este método permite evaluar el rendimiento del modelo de manera más precisa y obtener generalizaciones más confiables. Utilizando diferentes particiones de los datos, logra robustecer el modelo y maximizar la información obtenida en el proceso.

La versión más comúnmente utilizada de la validación cruzada es la validación cruzada ***k-fold***: donde k es un número especificado por el usuario, pero también representa el número de pliegues que tendrá la base de datos. Cada iteración de la validación cruzada *k-fold* implica usar un pliegue diferente como conjunto de prueba y los demás como conjunto de entrenamiento, de manera que cada pliegue se utiliza exactamente una vez como conjunto de prueba. Esto asegura una evaluación exhaustiva del rendimiento del modelo. Por ejemplo, si se establece $k = 5$, habrá exactamente 5 iteraciones, dado que cada pliegue debe ser utilizado una vez para probar el modelo generado con los otros pliegues de entrenamiento [3].

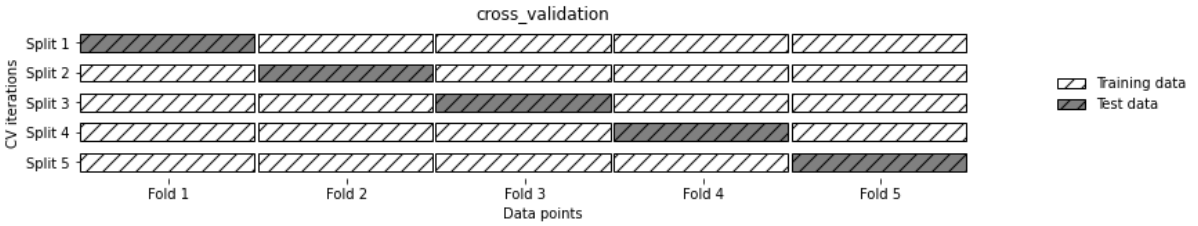

Figura 4: Validación Cruzada K-fold. Recuperado de Guido, A. C. (2016). Introduction to [Figura 5.1]. En Guido, A. C., Introduction to. O'Reilly Media, Inc [3] y [**?**].

La representación visual de lo anteriormente explicado se encuentra en la Gráfica 4, donde cada pliegue, destacado en un color más oscuro, se utiliza una única vez como conjunto de datos de prueba y el resto de entrenamiento.

Otro método de validación cruzada es ***la validación cruzada estratificada k-fold***: esta técnica surge como una alternativa en bases de datos donde cada *pliegue* no podría representar información suficiente para una categoría específica, reduciendo su capacidad de generalización. Por ejemplo, en una base de datos con una variable a estudiar (denotada como "y") que es binaria (0 o 1), y donde una de estas dos clases (1) solo representa el 10 % de las observaciones, si se utiliza la técnica *k-fold*, es muy probable que algún *pliegue* no contenga observaciones de la clase 1. Por esta razón, nace *la validación cruzada estratificada k-fold*, porque, a diferencia de *k-fold*, esta asegura que cada *pliegue* tenga la misma proporción de clases en la variable "y", lo que permite que cada *pliegue* proporcione información al modelo y fortalezca su capacidad de generalización.

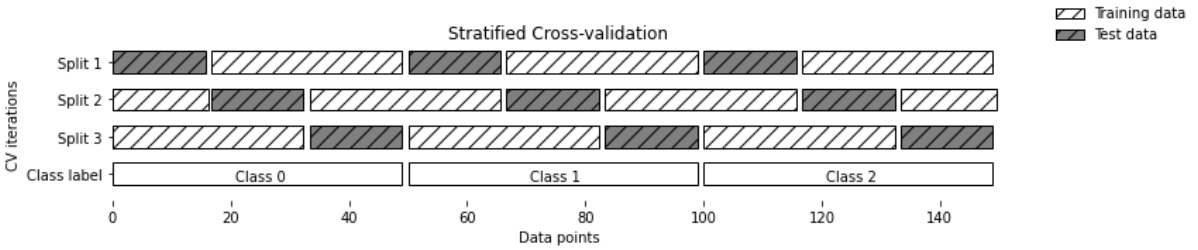

Figura 5: Validación Cruzada Estratificada tomado de Guido, A. C. (2016). Introduction to [Figura 5.2]. En Guido, A. C., Introduction to. O'Reilly Media, Inc [3] y [**?**].

La Figura 5 ofrece una representación visual de lo explicado anteriormente. En ella se observa cómo se extrae una porción de cada clase para el entrenamiento del modelo, garantizando, incluso en escenarios altamente desbalanceados, la inclusión de información de todas las clases.

Por otro lado, la **Validación Cruzada *Leave-one-out* (LOOCV)** es recomendable usarla en conjunto de datos pequeños, dada su naturaleza y costo computacional, este tipo de validación cruzada utiliza cada observación del conjunto de datos como testeo del modelo, y el resto del conjunto de datos como entrenamiento, es esta la razón por la cual este método es tan costoso computacionalmente, dado que entrena N modelos como observaciones existen en el conjunto de datos.

Para finalizar la **Validación cruzada con división y mezcla** A diferencia de LOOCV que está pensada para conjuntos de datos pequeños, está pensada para conjuntos muy grandes de datos, ya que dada su naturaleza nos permite experimentar en estos conjuntos de datos para encontrar mejores rendimientos a costos no tan altos de computación, ***"La validación cruzada con división y mezcla permite controlar el número de iteraciones de forma independiente de los tamaños de entrenamiento y prueba"*** [3]

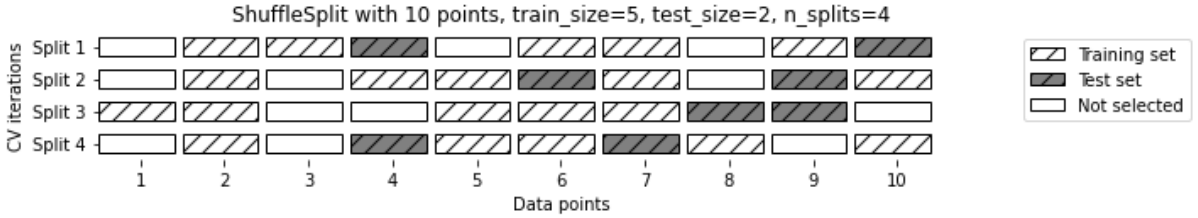

Figura 6: Validación cruzada con división y mezcla Recuperado de Guido, A. C. (2016). Introduction to [Figura 5.3]. En Guido, A. C., Introduction to. O'Reilly Media, Inc [3] y [**?**].

En la Figura 6, vemos como el conjunto esta divido en 10 *fold*, pero solo 5 de estos se utilizan para el entrenamiento del modelo y se prueban en solo 2, podemos seleccionar la cantidad de *fold* de prueba y entrenamiento y jugar con esta característica para experimentar en el conjunto de datos.

Para la validación cruzada con división y mezcla, también existe la versión estratificada, que tiene mejores rendimientos en tareas de clasificación, dado que para cada *fold* guarda las proporciones de la variable objeto de estudio, tal cual como la versión estratificada de *k-fold*.

4. Datos: recolección y descripción

Es importante aclarar que la Inestabilidad cromosómica o NIC por sus siglas en inglés, está definida como el desequilibrio genómico que se presenta cuando una célula tiene un número anormal de cromosomas. La causa puede ser un entrecruzamiento de cromosomas inesperado o la presencia de pequeños fragmentos de

ADN extracromosómico [**?**] y la técnica de micronúcleos con bloqueo de citocinesis (MNBN) mencionadas en el artículo han sido utilizadas anteriormente para evaluar la exposición a metales en poblaciones con problemas ocupacionales [4], tiene como objetivo evaluar la capacidad de ciertas sustancias químicas para causar daño en el material genético de las células, ya sea de forma numérica (cromosomas adicionales o faltantes) o estructural (daños en los cromosomas) [**?**].

El presente trabajo de grado tiene como objetivo aplicar técnicas de validación cruzada y diversos algoritmos de machine learning para la clasificación e identificación de daño celular o inestabilidad cromosómica (NIC) a través de las técnicas mencionadas, los datos analizados se tomaron del estudio "Micronuclei frequency and exposure to chemical mixtures in three Colombian mining populations" [4]. Se cuenta con plena autorización para el uso de estos datos y la aplicación de las técnicas mencionadas anteriormente. En ningún momento se pretende continuar con la investigación original de los autores, sino utilizar los datos recolectados con fines predictivos en la inestabilidad cromosómica (NIC) en pacientes de la región expuestos a actividades mineras.

La investigación hace énfasis en los profundos de los efectos de las actividades de extracción minera en diferentes municipios de Colombia, en el estudio, se tomaron muestras de los municipios de Montelíbano (Córdoba), Nechí (Antioquia) y Aranzazu (Caldas) para analizar los efectos de la minería en la salud.

Montelíbano, con minería a gran escala de ferroníquel activa actualmente, minería de carbón a mediana escala y minería artesanal y de pequeña escala (ASGM) para oro, presentó minerales como Selenio (Se), Mercurio (Hg), Manganeso (Mn), Plomo (Pb) y Magnesio (Mg). Nechí, con minería artesanal y de pequeña escala (ASGM) para oro, mostró la presencia de Cromo (Cr), Níquel (Ni), Mercurio (Hg), Selenio (Se) y Magnesio (Mg). Aranzazu, con una mina de mercurio cerrada que operó de 1948 a 1974, identificó Mercurio (Hg) y Níquel (Ni) en el ambiente residual. Montería, establecida como área de referencia y sin actividad minera, permite una comparación adecuada con las zonas mineras.

Para identificar la inestabilidad cromosómica (NIC), se empleó la técnica de micronúcleos con bloqueo de citocinesis (MNBN), técnica que es considerada uno de los mejores biomarcadores validados de inestabilidad cromosómica [**?**], siguiendo los criterios de Fenech [**?**]. Se tomaron muestras de sangre de 407 individuos en total: 99 de Montelíbano, 103 de Aranzazu, 104 de Nechí y 101 de Montería, siguiendo un procedimiento estandarizado. Las muestras se analizaron para detectar micronúcleos, evaluando 2000 células binucleadas por muestra, con el fin de detectar la presencia de inestabilidad cromosómica (NIC), asociada con la exposición a contaminantes mineros.

Los investigadores en su análisis de datos utilizaron el programa Microsoft Excel para Windows, para realizar un exhaustivo de control de calidad de los datos para la gestión de errores, datos faltantes y valores atípicos. Las variables categóricas se expresaron como porcentajes, y las variables continuas se resumieron con medidas de tendencia central y dispersión. Debido a la distribución sesgada de las concentraciones de metales, se emplearon cuartiles para resumir los datos. Se utilizó la prueba de Mann-Whitney [16] para comparar las concentraciones medianas de metales entre los tipos de exposición (residencial y ocupacional) en Montelíbano y Nechí, y frente al área de referencia. Esta prueba no paramétrica permite comparar dos grupos independientes sin asumir normalidad en los datos, evaluando si hay diferencias significativas entre las distribuciones.

Los investigadores realizaron un análisis de regresión de Poisson que se aplicó para investigar la asociación entre la exposición a metales, las variables sociodemográficas y el recuento de micronúcleos en los grupos de exposición y el grupo de control. Se realizaron análisis univariados y multivariados, utilizando el Criterio de Información de Akaike (AIC) para identificar las variables que mejor explicaran la variabilidad del recuento de micronúcleos. Los resultados se interpretaron examinando la razón de prevalencia y sus intervalos de confianza del 95 %. Además, se generaron gráficos de correlación y análisis de componentes principales (PCA) para identificar patrones en las concentraciones de metales. En Montelíbano, donde se observó una fuerte correlación entre los metales, se incorporó el primer componente principal (PC-1) en el análisis de regresión de Poisson multivariado, junto con las variables sociodemográficas. Todas las pruebas estadísticas se realizaron con un nivel de significación de 0,05 utilizando la versión 3.3.0 de R [**?**].

Para garantizar que las muestras no estuvieran sesgadas por variables sociodemográficas y de com-

portamiento, los investigadores llevaron a cabo controles exhaustivos en relación con la edad, hábitos alimenticios, tabaquismo y consumo de alcohol. La distribución demográfica de las poblaciones estudiadas mostró que, en general, la mayoría de los participantes eran hombres con una mediana de edad de 49 años, mientras que las mujeres tenían una mediana de 45 años. No se observaron diferencias significativas en la edad promedio, nivel socioeconómico o hábitos alimentarios entre las poblaciones expuestas y de referencia. El análisis incluyó un ajuste por variables como el consumo de alcohol y tabaco, asegurando que los resultados no estuvieran sesgados por estos factores. El 54,1 % de los participantes se identificaron como no bebedores y el 75,7 % como no fumadores, mientras que la mayoría de los bebedores reportaron un consumo bajo. La frecuencia de consumo de pescado, que puede influir en la exposición a metales pesados, se evaluó detalladamente para asegurar que las diferencias observadas no se debieran a estos hábitos. Las referencias utilizadas para el análisis incluyeron directrices sobre la evaluación de factores de confusión en estudios epidemiológicos, como las establecidas por la ATSDR y el Instituto Nacional de Abuso de Alcohol y Alcoholismo (NIAAA), para asegurar que la exposición medida reflejara adecuadamente los efectos de los contaminantes sin estar influenciada por factores externos como el consumo de alcohol o tabaco.

4.1. Metodología para el tratamiento de la base de datos

La base de datos fue obtenida mediante una encuesta estructurada, en la cual varias preguntas dependían de respuestas previas. Este diseño condicional estableció que, si un encuestado respondía negativamente a una pregunta inicial (por ejemplo, ¿Fuma actualmente?), las respuestas a preguntas de seguimiento relacionadas (como ¿Cuántos cigarrillos fuma al día?) fueran asignadas con un valor de 0, indicando la ausencia de consumo.

4.1.1. Carga y selección de variables

Se cargó la base de datos desde el archivo `Base_Proyecto_905_Unisinu.xlsx` y, tras revisar la totalidad de las columnas, se seleccionaron aquellas de mayor relevancia para el análisis, con un enfoque en variables sociodemográficas, patrones de exposición a sustancias, y condiciones laborales y ambientales. La lista de variables seleccionadas incluyó:

- **Variables demográficas:** Edad (`N4Edad`), Sexo (`N5Sexo`), y lugar de residencia (`MPIO_AJ`).
- **Variables de exposición y ocupación:** Relativas al trabajo en minería (`N20Trabajaactivminera`), trabajo agrícola (`N28_Trabaj_agricultura`), uso de metales y plaguicidas, consumo de sustancias, entre otras.
- **Variables de consumo y frecuencia:** Relacionadas con la dieta, incluyendo el consumo de pescado, carnes, frutas, verduras y alimentos enlatados.

4.1.2. Imputación de valores condicionales

En el proceso de imputación de valores condicionales para preguntas con dependencias, se aplicaron métodos específicos que respetan la lógica subyacente del cuestionario. Estas imputaciones aseguran que, cuando un encuestado responde negativamente a una pregunta principal, los valores asociados a preguntas dependientes se asignen de manera coherente. Los tratamientos realizados incluyen:

- **Ocupación y exposición en minería:** Para encuestados que respondieron "No"en la pregunta ¿Trabaja en actividades mineras? (`N20Trabajaactivminera`), se asignó el valor de 0 en las columnas dependientes:
  - `N22tiempotrabajaconmetales` (tiempo de trabajo con metales),

- `N24frecuenciausametales` (frecuencia de uso de metales),
- `N25horas_aldia` (horas de exposición diaria).

- **Trabajo agrícola:** Si el encuestado indicó que no trabaja en actividades agrícolas (`N28_Trabaj_agricultura` = 0), las columnas:

  - `N29_tiempo_trabaja_plaguicida` (tiempo trabajando con plaguicidas),
  - `N31_frecuencia_usa_plaguicidas` (frecuencia de uso de plaguicidas),
  - `N32_horas_dia` (horas de exposición diaria a plaguicidas),

  fueron imputadas con 0.

- **Consumo de tabaco:** Los encuestados que indicaron que nunca han fumado (`N50H_afumado_alguna_vez` = 0) recibieron el valor de 0 en:

  - `N51_Fuma_actualmente` (fuma actualmente),
  - `N52_Hace_cuanto_tiempo_fuma` (hace cuánto fuma),
  - `N53_Cuantos_cigarrillos_fuma_al_dia` (cigarrillos al día).

  Además, para quienes respondieron que no fuman actualmente (`N51_Fuma_actualmente` = 0), se asignó 0 en `N53_Cuantos_cigarrillos_fuma_al_dia`.

- **Consumo de licor:** Si el encuestado indicó que no consume licor (`N57_Usted_consume_licor` = 0), las variables relacionadas:

  - `N58_frecuencia_consume_licor` (frecuencia de consumo),
  - `N59_Usted_consume_licor` (confirmación de consumo),

  fueron asignadas a 0.

4.1.3. Transformación y limpieza de columnas en ppm

Las columnas de concentración de elementos en partes por millón (ppm), como Li 7 (ppm), Be 9 (ppm), B 11 (ppm), Na 23 (ppm), entre otras, contenían valores en formatos variados, como 0,010 (valor numérico estándar con coma como separador decimal) y <0,002 (indicando valores por debajo del umbral de detección). Para asegurar la consistencia y prepararlos para el análisis, se aplicó un proceso que incluyó: la limpieza de símbolos, donde los valores con el símbolo "<"fueron ajustados multiplicándolos por 0,8 para representarlos como ligeramente menores al límite detectable, la conversión de separadores, cambiando la coma por un punto para hacer los valores consistentes con el formato reconocido por el software, y la imputación de valores faltantes, donde se calculó la media de cada columna para completar los valores nulos sin generar sesgos. Estas transformaciones resultaron en un conjunto de datos homogéneo y listo para el análisis, eliminando inconsistencias en los formatos y completando los valores ausentes de manera efectiva.

4.1.4. Transformación y escalado de datos

- **Escalado de variables numéricas:** Las columnas numéricas seleccionadas, como `N4Edad`, `N19Tiempooficio` (tiempo en el oficio), entre otras, se estandarizaron mediante el método de escalado estándar.

- **Codificación de variables categóricas:** Las variables categóricas seleccionadas fueron codificadas mediante *one-hot encoding*.

4.1.5. Variable objetivo y clasificación por cuartiles

La variable de interés `Bi-Nucleada` se categorizó en cuartiles para efectos del objetivo del trabajo, generando cuatro categorías: "Bajo", "Medio-Bajo", "Medio-Alto", y "Alto". Esta transformación permitió definir una variable de respuesta adecuada para los modelos de clasificación empleados en la investigación.

Valor	Frecuencia
Medio-Bajo	104
Bajo	102
Alto	102
Medio-Alto	99

Tabla 1: Distribución de categorías y frecuencias

4.1.6. Variable objetivo numérica descriptivo

La variable **Bi-Nucleada** presenta un total de 407 observaciones con una media de 178,08 y una desviación estándar de 95,69, lo cual indica una dispersión considerable alrededor de la media.

Los valores de esta variable oscilan entre un mínimo de 0 y un máximo de 468, lo que refleja una amplitud significativa. Los percentiles muestran que el 25 % de los valores están por debajo de 118, el 50 % (mediana) se encuentra en 190, y el 75 % está por debajo de 252,5.

En cuanto a la forma de la distribución, la asimetría de -0,29 indica una ligera asimetría negativa, sugiriendo que la mayoría de los valores tienden a agruparse levemente hacia la derecha del promedio. La curtosis de -0,62 indica una distribución con colas menos gruesas que una normal, lo que sugiere una menor presencia de valores extremos o atípicos en comparación con una distribución normal.

Con base en los descriptivos realizados, podemos afirmar que la división de la variable objetivo `Bi-Nucleada` en cuartiles simplifica una variable numérica continua en categorías discretas, facilitando tanto la interpretación como el análisis mediante modelos de clasificación. Además, la categorización basada en cuartiles garantiza una distribución balanceada entre las clases, como lo demuestra la Tabla 1, al presentar frecuencias similares para cada categoría. Por otra parte, la considerable dispersión y amplitud de la variable original respaldan esta transformación, ya que los cuartiles dividen el rango amplio de valores en intervalos representativos, reflejando posibles patrones relevantes que los modelos de clasificación pueden aprender.

5. Modelo, resultados y conclusiones

En este estudio, se decidió emplear los hiperparámetros básicos de cada modelo de clasificación, dado que el objetivo principal era evaluar el impacto de las diferentes técnicas de validación cruzada en el desempeño de los modelos. Ajustar hiperparámetros específicos, especialmente en modelos con un alto grado de personalización como Random Forest o k-Nearest Neighbors, podría introducir sesgos que afectarían la comparación objetiva entre las técnicas de validación. Este enfoque permite centrarse en la evaluación de la robustez y generalización de los modelos bajo las condiciones establecidas, asegurando una comparación más equilibrada.

La metodología aplicada en este estudio, que incluyó el uso de los métodos de los k vecinos más cercanos (*K-Nearest Neighbors, KNN*), la regresión logística (*Logistic Regression*), bosques aleatorios (*Random Forest*) y los árboles de decisión (*Decision Tree*), así como diversas técnicas de validación cruzada, incluyendo *k-fold* (validación cruzada con partición en k subconjuntos), *Leave-One-Out* (validación excluyendo una observación por iteración), *StratifiedKFold* (validación estratificada en k subconjuntos) y

Shuffle Split (validación cruzada basada en particiones aleatorias), se diseñó con el propósito de evaluar el desempeño de estos modelos en un entorno clínico y con datos sensibles. La selección de estas técnicas permitió abordar posibles sesgos, mejorar la capacidad de generalización de los modelos y garantizar una estimación robusta. La categorización de la variable objetivo en cuartiles, propuesta para mitigar el sesgo inherente al manejo de datos clínicos, se sometió a pruebas de granularidad con 3, 4 y 6 categorías, con el fin de evaluar su impacto en el desempeño de los modelos. La validación cruzada se implementó como una herramienta crucial no solo para mejorar las estimaciones, sino también como método comparativo entre los modelos, permitiendo identificar diferencias en su rendimiento y proporcionando una base sólida para futuras investigaciones. En conjunto, la metodología aplicada contribuyó a optimizar el análisis, permitiendo una evaluación más precisa y confiable de los modelos en datos clínicos sensibles.

5.1. Configuración del experimento

En el presente análisis, se utilizaron diversas librerías para implementar modelos de machine learning y métodos de validación cruzada, lo que permitió una evaluación robusta de los algoritmos seleccionados. Las principales librerías empleadas incluyen `sklearn`, que ofrece herramientas esenciales como `train_test_split`, `LeaveOneOut`, `LeavePOut`, y `ShuffleSplit` para la validación cruzada. Además, se utilizó `pandas` para la manipulación de datos y `StandardScaler` para la normalización de características. Los modelos aplicados abarcan `DecisionTreeClassifier` y `RandomForestClassifier` y así como el `KNeighborsClassifier` y `LogisticRegression`. Las métricas de evaluación, tales como `accuracy_score`, `recall_score`, y `roc_auc_score`, se calcularon mediante funciones de `sklearn`. En cuanto a la infraestructura, el análisis se llevó a cabo en un sistema equipado con un procesador `AMD Ryzen 7 3700X` de 8 núcleos a 3.59 GHz, con 32.0 GB de RAM y una tarjeta gráfica `Nvidia RTX 4060 TI` de 16 GB, lo que permitió un procesamiento eficiente y rápido de los datos y modelos.

Model	**hiperparametros**
Árbol de Decisión	criterion='gini', max_depth=None, min_samples_split=2, random_state=226
Bosque Aleatorio	n_estimators=100, criterion='gini', max_depth=None, bootstrap=True, random_state=226
Vecinos Más Cercanos	n_neighbors=5, weights='uniform', algorithm='auto'
Regresión Logística	penalty='l2', solver='lbfgs', max_iter=100, C=1.0, random_state=42

Tabla 2: Hiperparámetros base de los modelos

En la tabla 2, podemos resumir los hiperparametros usados:

- **Decision Tree:** Usa el criterio `gini`, sin límite en la profundidad del árbol (`max_depth=None`), con un mínimo de 2 muestras para dividir nodos (`min_samples_split=2`) y una semilla aleatoria fija (`random_state=226`).
- **Random Forest:** Configurado con 100 árboles (`n_estimators=100`), criterio `gini`, sin límite de profundidad, habilitando el muestreo con reemplazo (`bootstrap=True`) y con la misma semilla aleatoria.
- **K Neighbors:** Considera 5 vecinos (`n_neighbors=5`), con pesos uniformes (`weights='uniform'`) y algoritmo automático para la búsqueda de vecinos (`algorithm='auto'`).
- **Logistic Regression:** Aplica regularización L2 (`penalty='l2'`), el solucionador `lbfgs`, un máximo de 100 iteraciones (`max_iter=100`), una regularización inversa estándar (`C=1.0`) y una semilla aleatoria distinta (`random_state=42`).

5.2. Árbol de decisión

5.2.1. Árbol de decisión con Validación Cruzada ShuffleSplit

Tabla 3: Resultados del modelo de árbol de decisión con ShuffleSplit

Valores K	Accuracy	F1 Score	Recall	AUC
2	0,293	0,274	0,280	0,519
3	0,268	0,250	0,266	0,510
4	0,280	0,266	0,279	0,518
5	0,268	0,255	0,268	0,511
6	0,285	0,273	0,284	0,522
7	0,275	0,261	0,276	0,517
8	0,271	0,258	0,269	0,513
9	0,257	0,244	0,255	0,504
15	0,246	0,232	0,242	0,496
20	0,245	0,232	0,243	0,496
Modelo Base	**0,333**	**0,331**	**0,333**	**0,556**

Interpretación: El modelo de árbol de decisión muestra un rendimiento moderado en términos de precisión, F1, y recall, con valores que tienden a estar por debajo del modelo base. La métrica de AUC oscila entre 0,496 y 0,522, lo que indica una capacidad limitada para distinguir entre las clases. En general, el rendimiento disminuye ligeramente conforme el valor de k aumenta.

5.2.2. Árbol de decisión con KFold

Tabla 4: Resultados del modelo de árbol de decisión con KFold

Valores K	Accuracy	F1 Score	Recall	AUC
2	0,258	0,258	0,259	0,506
3	0,273	0,271	0,276	0,517
4	0,292	0,288	0,296	0,530
5	0,297	0,293	0,296	0,531
6	0,319	0,311	0,325	0,549
7	0,268	0,262	0,271	0,513
8	0,275	0,267	0,276	0,517
9	0,288	0,284	0,290	0,526
15	0,290	0,272	0,282	0,523
20	0,292	0,268	0,284	0,525
Modelo Base	**0,333**	**0,331**	**0,333**	**0,556**

Interpretación: Los resultados del modelo de árbol de decisión con KFold son ligeramente mejores que los obtenidos con ShuffleSplit, especialmente en valores de k más altos (como 6 y 9), donde el modelo se acerca al rendimiento del modelo base en términos de precisión y recall. La AUC muestra una mejora en comparación con ShuffleSplit, alcanzando un máximo de 0,549 para $k = 6$.

5.2.3. Árbol de decisión con StratifiedKFold

Tabla 5: Resultados del modelo de árbol de decisión con StratifiedKFold

Valores K	Accuracy	F1 Score	Recall	AUC
2	0,265	0,264	0,265	0,510
3	0,270	0,269	0,271	0,514
4	0,273	0,271	0,273	0,515
5	0,295	0,291	0,295	0,530
6	0,287	0,286	0,288	0,525
7	0,282	0,280	0,283	0,522
8	0,278	0,270	0,277	0,518
9	0,283	0,283	0,283	0,522
15	0,312	0,303	0,314	0,543
20	0,273	0,263	0,273	0,516
Modelo Base	**0,333**	**0,331**	**0,333**	**0,556**

Interpretación: En StratifiedKFold, el modelo de árbol de decisión muestra resultados que oscilan cerca de los obtenidos con KFold, con valores de precisión, F1 y recall que aumentan ligeramente con k. Para $k = 15$, se observa un rendimiento notable con un AUC de 0,543, cercano al modelo base, aunque sigue siendo inferior en precisión.

5.2.4. Árbol de decisión con LeavePOut (LOOCV)

Tabla 6: Resultados del modelo de árbol de decisión con LeavePOut (LOOCV)

Valores K	Accuracy	F1 Score	Recall	AUC
LOOCV	0,275	0,275	0,275	-
Modelo Base	**0,333**	**0,331**	**0,333**	**0,556**

Interpretación: El resultado de Validación cruzada Leave-One-Out(LOOCV) muestra que el modelo de árbol de decisión tiene una precisión, F1 y recall de aproximadamente 0,275, lo cual es menor al rendimiento del modelo base. La ausencia de valores AUC limita la evaluación de la capacidad de discriminación del modelo, aunque los valores de precisión indican que el modelo no supera al modelo base.

5.2.5. Análisis del Rendimiento del Modelo de Árbol de Decisión

En general, el modelo de Árbol de Decisión mostró un rendimiento inferior al modelo base en todas las métricas principales (Accuracy, F1, Recall, AUC), independientemente del método de validación cruzada.

- **Mejores resultados:** KFold y StratifiedKFold se destacan ligeramente, alcanzando un AUC máximo de 0,549 (para KFold, k=6) y 0,543 (para StratifiedKFold, k=15), aunque siguen estando por debajo del modelo base.
- **Comportamiento con k:** A medida que k aumenta, los resultados tienden a estabilizarse sin una mejora notable, lo que sugiere que el modelo no se beneficia considerablemente de valores de k más altos.
- **Interpretación:** El Árbol de Decisión parece ser menos robusto y más sensible a los cambios en las divisiones de datos, mostrando resultados limitados en su capacidad de generalización, incluso con validación cruzada.

5.3. Boques Aleatorios

5.3.1. Boques Aleatorios con Validación Cruzada ShuffleSplit

Tabla 7: Resultados del modelo Boques Aleatorios con Validación Cruzada ShuffleSplit

Valores K	Accuracy	F1 Score	Recall	AUC
2	0,293	0,289	0,315	0,566
3	0,285	0,281	0,295	0,536
4	0,293	0,289	0,300	0,548
5	0,298	0,294	0,304	0,541
6	0,309	0,303	0,313	0,539
7	0,321	0,314	0,328	0,553
8	0,326	0,314	0,325	0,555
9	0,331	0,313	0,325	0,564
15	0,319	0,303	0,327	0,557
20	0,324	0,310	0,331	0,560
Modelo Base	**0,293**	**0,281**	**0,290**	**0,561**

Interpretación: El modelo Boques Aleatorios con Validación Cruzada ShuffleSplit muestra una mejora significativa en comparación con el modelo base. La precisión, el F1 y el recall aumentan gradualmente con el valor de k. A partir de $k = 7$, las métricas se estabilizan, alcanzando un AUC máximo de 0,564 para $k = 9$. La tendencia general sugiere que el modelo mejora con el incremento de k, lo que indica una mayor capacidad para generalizar y distinguir entre las clases en comparación con el modelo base.

5.3.2. Boques Aleatorios con Validación Cruzada KFold

Tabla 8: Resultados del modelo Boques Aleatorios con Validación Cruzada KFold

Valores K	Accuracy	F1 Score	Recall	AUC
2	0,307	0,301	0,310	0,547
3	0,310	0,305	0,314	0,551
4	0,324	0,320	0,334	0,566
5	0,322	0,317	0,327	0,570
6	0,319	0,311	0,325	0,570
7	0,292	0,285	0,304	0,559
8	0,278	0,268	0,291	0,552
9	0,305	0,297	0,312	0,577
15	0,315	0,302	0,323	0,560
20	0,302	0,278	0,293	0,560
Modelo Base	**0,293**	**0,281**	**0,290**	**0,561**

Interpretación: Boques Aleatorios con Validación Cruzada KFold también mejora respecto al modelo base, alcanzando una precisión máxima de 0,324 para $k = 4$, y un AUC máximo de 0,577 para $k = 9$. Las métricas como el F1 y el recall muestran una mejora más constante en comparación con el modelo base. Sin embargo, el rendimiento tiende a estabilizarse a medida que aumentan los valores de k, con los valores de AUC siendo bastante altos, especialmente en $k = 9$.

5.3.3. Boques Aleatorios con Validación Cruzada StratifiedKFold

Tabla 9: Resultados del modelo Boques Aleatorios con Validación Cruzada StratifiedKFold

Valores K	Accuracy	F1 Score	Recall	AUC
2	0,280	0,279	0,280	0,547
3	0,310	0,307	0,310	0,555
4	0,307	0,306	0,306	0,561
5	0,293	0,282	0,291	0,581
6	0,298	0,287	0,296	0,558
7	0,339	0,333	0,339	0,570
8	0,307	0,306	0,308	0,563
9	0,300	0,295	0,299	0,564
15	0,300	0,288	0,302	0,543
20	0,355	0,336	0,353	0,570
Modelo Base	**0,293**	**0,281**	**0,290**	**0,561**

Interpretación: El modelo de Boques Aleatorios con Validación Cruzada StratifiedKFold muestra una mejora notable en comparación con los otros métodos de validación cruzada, especialmente con $k = 20$, donde se observa una precisión de 0,355 y un AUC de 0,570. El F1 y el recall también muestran mejoras consistentes a lo largo de los valores de k, lo que sugiere una mayor capacidad para manejar el desequilibrio de clases y mejorar la capacidad de clasificación.

5.3.4. Boques Aleatorios con Validación Cruzada LeavePOut (LOOCV)

Tabla 10: Resultados del Boques Aleatorios con Validación Cruzada LeavePOut (LOOCV)

Valores K	Accuracy	F1 Score	Recall	AUC
LOOCV	0,346	0,346	0,346	-
Modelo Base	**0,293**	**0,281**	**0,290**	**0,561**

Interpretación: El método LeavePOut (LOOCV) muestra un rendimiento superior al modelo base, con una precisión, F1 y recall de 0,346. Aunque no se proporciona el valor de AUC, las métricas de clasificación indican una mejora significativa en el desempeño del modelo, lo que sugiere que LeavePOut ha permitido una mayor generalización y precisión en la clasificación de las muestras.

5.3.5. Resumen Comparativo de los Resultados

El modelo de Boques Aleatorios mejoró sustancialmente en comparación con el modelo base, destacándose en particular con los métodos StratifiedKFold y LeavePOut (LOOCV).

- **Mejores resultados:** StratifiedKFold alcanzó la mayor precisión (Accuracy de 0,355 y AUC de 0,570 para k=20), mientras que LeavePOut obtuvo valores elevados de precisión, F1 y Recall de 0,346.
- **Comportamiento con K:** Boques Aleatorios muestra una mejora constante y estable con un aumento de k, alcanzando un rendimiento óptimo en valores de k intermedios (como k=9 en KFold) y altos (k=20 en StratifiedKFold), por otro lado, ShuffleSplit y KFold presentan un comportamiento más estable pero menos eficiente en términos de aumento de rendimiento con aumentos de k
- **Interpretación:** StratifiedKFold y LOOCV permiten que Random Forest maneje mejor las clases desbalanceadas, indicando su robustez y mayor capacidad de generalización. Esto sugiere que, en

general, los métodos que equilibran el tamaño de las particiones y preservan la distribución de clases, como StratifiedKFold, parecen ser más adecuados para este modelo de Random Forest.

5.4. Vecino más Cercano

5.4.1. Vecino más Cercano con Validación Cruzada ShuffleSplit

Tabla 11: Resultados del modelo Vecino más Cercano con Validación Cruzada ShuffleSplit

Valores K	Accuracy	F1 Score	Recall	AUC
2	0,341	0,328	0,353	0,570
3	0,341	0,333	0,364	0,566
4	0,323	0,310	0,337	0,573
5	0,346	0,335	0,362	0,578
6	0,350	0,341	0,363	0,568
7	0,331	0,323	0,344	0,559
8	0,335	0,318	0,340	0,564
9	0,350	0,329	0,352	0,567
15	0,353	0,334	0,364	0,569
20	0,346	0,326	0,355	0,574
Modelo Base	**0,244**	**0,232**	**0,241**	**0,490**

Interpretación: Los resultados del modelo Vecino más Cercano con Validación Cruzada ShuffleSplit muestran una mejora significativa en todas las métricas comparado con el modelo base. La precisión oscila entre 0,323 y 0,353, con el valor máximo alcanzado en $k = 15$. La F1 y el recall siguen una tendencia similar, alcanzando valores más altos a medida que k aumenta. La AUC también muestra una mejora notable, alcanzando un valor máximo de 0,574 en $k = 20$. Esto indica que el modelo K Neighbors mejora considerablemente con ShuffleSplit en términos de precisión y recall.

5.4.2. Vecino más Cercano con Validación Cruzada KFold

FloatBarrier

Tabla 12: Resultados del modelo Vecino más Cercano con Validación Cruzada KFold

Valores K	Accuracy	F1 Score	Recall	AUC
2	0,283	0,273	0,286	0,526
3	0,253	0,239	0,255	0,533
4	0,302	0,292	0,308	0,547
5	0,314	0,302	0,320	0,545
6	0,327	0,315	0,335	0,561
7	0,310	0,299	0,328	0,551
8	0,310	0,298	0,324	0,551
9	0,315	0,301	0,331	0,559
15	0,324	0,302	0,339	0,557
20	0,327	0,302	0,346	0,554
Modelo Base	**0,244**	**0,232**	**0,241**	**0,490**

Interpretación: En el caso Vecino más Cercano con Validación Cruzada KFold, se observa una mejora progresiva en la precisión, F1 y recall conforme aumenta el valor de k. La precisión oscila entre 0,253 y 0,327, destacándose un valor de 0.327 en $k = 6$. La AUC también muestra una tendencia creciente,

alcanzando su valor máximo de 0,561 en $k = 6$, resaltando de igual manera k=20 con resultados muy parecidos a k=6. A pesar de estas mejoras, los resultados no superan significativamente los obtenidos con ShuffleSplit, y la mejora es más modesta.

5.4.3. Vecino más Cercano con Validación Cruzada StratifiedKFold

Tabla 13: Resultados del modelo Vecino más Cercano con Validación Cruzada StratifiedKFold

Valores K	Accuracy	F1 Score	Recall	AUC
2	0,312	0,297	0,312	0,541
3	0,295	0,276	0,295	0,539
4	0,307	0,292	0,307	0,536
5	0,332	0,321	0,331	0,562
6	0,305	0,295	0,305	0,546
7	0,332	0,317	0,330	0,558
8	0,324	0,309	0,324	0,550
9	0,341	0,325	0,343	0,568
15	0,331	0,310	0,332	0,549
20	0,325	0,303	0,325	0,555
Modelo Base	**0,244**	**0,232**	**0,241**	**0,490**

Interpretación: El modelo Vecino más Cercano con Validación Cruzada StratifiedKFold muestra mejoras en comparación con el modelo base, con una precisión que varía entre 0,295 y 0,341. Las métricas de F1 y recall siguen una tendencia similar, destacando el valor de 0,341 en $k = 9$, lo que indica un mejor rendimiento en cuanto a la capacidad del modelo para manejar las clases desbalanceadas. La AUC también presenta una mejora, alcanzando 0.568 en $k = 9$, el valor más alto.

5.4.4. Vecino más Cercano con Validación Cruzada LeavePOut (LOOCV)

Tabla 14: Resultados del modelo Vecino más Cercano con Validación Cruzada LeavePOut (LOOCV)

Valores K	Accuracy	F1 Score	Recall	AUC
LOOCV	0,324	0,324	0,324	-
Modelo Base	**0,244**	**0,232**	**0,241**	**0,490**

Interpretación: El modelo con LeavePOut (LOOCV) muestra una mejora significativa en comparación con el modelo base, con una precisión, F1 y recall de 0,324. Aunque no se dispone de un valor de AUC, los resultados en cuanto a precisión y recall sugieren que el modelo tiene un rendimiento robusto con este método de validación cruzada.

5.4.5. Resumen Comparativo de los Resultados

K Neighbors con validación cruzada superó consistentemente al modelo base, con mejoras notables en todas las métricas, especialmente en ShuffleSplit y StratifiedKFold.

- **Mejores resultados:** ShuffleSplit alcanzó valores de precisión de hasta 0,353 (para $k = 15$) y AUC de 0,574, mientras que StratifiedKFold presentó un AUC máximo de 0,568 ($k = 9$).
- **Comportamiento con K:** En todos los métodos, el rendimiento del modelo mejora con el aumento de K, lo que indica que K Neighbors se adapta bien a particiones más amplias y es relativamente robusto frente a la variabilidad de la partición.

- **Interpretación:** K Neighbors demuestra un buen manejo de la estratificación y se beneficia de métodos de validación cruzada que conservan la distribución de clases. Esto sugiere que es un modelo que gana en estabilidad y desempeño con validación cruzada, especialmente en métodos que manejan bien las clases desbalanceadas.

5.5. Regresión logística

5.5.1. Regresión logística con ShuffleSplit

Tabla 15: Resultados del modelo Regresión logística con ShuffleSplit

Valores K	Accuracy	F1 Score	Recall	AUC
2	0,256	0,241	0,254	0,474
3	**0,244**	**0,234**	**0,246**	**0,458**
4	0,268	0,261	0,272	0,499
5	0,259	0,252	0,262	0,512
6	0,260	0,252	0,264	0,510
7	0,265	0,255	0,267	0,511
8	0,262	0,250	0,261	0,503
9	0,274	0,260	0,269	0,508
15	0,276	0,265	0,282	0,526
20	0,287	0,276	0,290	0,537
Modelo Base	**0,325**	**0,317**	**0,338**	**0,552**

Interpretación: Los resultados de Regresión logística con ShuffleSplit muestran un rendimiento que se incrementa ligeramente a medida que aumenta k. La precisión, F1 y recall varían de manera limitada, con el valor máximo de precisión en 0,287 para $k = 20$. La AUC también presenta una tendencia de mejora, alcanzando su máximo de 0,537 en k= 20. Sin embargo, todos los resultados se mantienen por debajo del modelo base, indicando que ShuffleSplit no logra mejorar significativamente el rendimiento de este modelo en comparación con la configuración sin validación.

5.5.2. Regresión logística con KFold

Tabla 16: Resultados del modelo Regresión logística con KFold

Valores K	Accuracy	F1 Score	Recall	AUC
2	**0,263**	**0,262**	**0,264**	**0,531**
3	0,314	0,307	0,318	0,556
4	0,268	0,263	0,266	0,538
5	0,305	0,297	0,306	0,529
6	0,285	0,277	0,282	0,537
7	0,285	0,277	0,291	0,535
8	0,278	0,270	0,285	0,537
9	0,290	0,279	0,290	0,539
15	0,290	0,266	0,277	0,529
20	0,290	0,272	0,289	0,544
Modelo Base	**0,325**	**0,317**	**0,338**	**0,552**

Interpretación: Con KFold, el modelo de Regresión logística logra su mejor rendimiento en $k = 3$, alcanzando un valor de precisión de 0,314 y un AUC de 0,556. Aunque estos resultados son mejores en

comparación con ShuffleSplit, el modelo sigue sin alcanzar el rendimiento del modelo base, especialmente en recall y precisión. Los valores de k más altos no mejoran significativamente el desempeño.

5.5.3. Regresión logística con StratifiedKFold

Tabla 17: Resultados del modelo Regresión logística con StratifiedKFold

Valores K	Accuracy	F1 Score	Recall	AUC
2	**0,263**	**0,260**	**0,264**	**0,520**
3	0,263	0,262	0,263	0,494
4	0,280	0,277	0,280	0,509
5	0,288	0,285	0,288	0,532
6	0,287	0,281	0,289	0,521
7	0,295	0,291	0,295	0,527
8	0,292	0,288	0,292	0,536
9	0,290	0,282	0,290	0,541
15	0,297	0,286	0,296	0,530
20	0,308	0,292	0,306	0,531
Modelo Base	**0,325**	**0,317**	**0,338**	**0,552**

Interpretación: El método StratifiedKFold ofrece el mejor desempeño en $k = 20$, donde se alcanza una precisión de 0,308 y una AUC de 0,531. Aunque los resultados mejoran en comparación con ShuffleSplit y se aproximan a los de KFold, el rendimiento todavía se encuentra por debajo del modelo base en todas las métricas. La consistencia en los valores de F1 y recall indica que StratifiedKFold proporciona un balance adecuado entre precisión y capacidad de detección, pero no logra superar al modelo base.

5.5.4. Logistic Regression con LeavePOut (LOOCV)

Tabla 18: Resultados del modelo Regresión logística con LeavePOut (LOOCV)

Valores K	Accuracy	F1 Score	Recall	AUC
LOOCV	0,287	0,287	0,287	-
Modelo Base	**0,325**	**0,317**	**0,338**	**0,552**

Interpretación: El modelo de Regresión logística con LeavePOut (LOOCV) muestra un rendimiento limitado, con valores de precisión, F1 y recall de 0,287. No se proporciona un valor de AUC para este método, lo que dificulta la comparación en términos de equilibrio entre sensibilidad y especificidad. Este método, aunque es exhaustivo, no ofrece una mejora notable sobre los otros métodos de validación cruzada y se encuentra por debajo del rendimiento del modelo base.

5.5.5. Resumen Comparativo

En general, el modelo de Regresión Logística mostró un rendimiento que no superó al modelo base en las principales métricas de evaluación, aunque presentó algunas variaciones dependiendo del método de validación cruzada.

- **Mejores resultados:** El modelo alcanzó su mejor rendimiento en términos de precisión y AUC con el método StratifiedKFold para $k = 20$, obteniendo una precisión de 0,308 y AUC de 0,531. Esto fue un ligero avance en comparación con ShuffleSplit y KFold, aunque aún por debajo de los resultados del modelo base.

- **Comportamiento con K:** En general, el rendimiento mejora levemente a medida que k aumenta, especialmente en ShuffleSplit y StratifiedKFold, pero sin una mejora significativa que indique que un mayor k proporcione grandes beneficios. La precisión y AUC se estabilizan sin alcanzar el nivel del modelo base.

- **Interpretación:** La Regresión Logística mostró un desempeño moderado con los diferentes métodos de validación cruzada, destacándose ligeramente en StratifiedKFold, pero sin superar los resultados del modelo base. La precisión y recall se mantuvieron en niveles bajos, sugiriendo que el modelo no captó de manera efectiva las relaciones en los datos. A pesar de ser un modelo estadísticamente simple, no se observó una mejora notable al usar métodos de validación cruzada más complejos como LeavePOut (LOOCV), lo que indica que la capacidad de generalización no mejoró significativamente con la validación cruzada en este caso.

6. Consideraciones generales

Los resultados obtenidos en el análisis permiten concluir que el uso de técnicas avanzadas de validación cruzada ha demostrado ser efectivo en la mejora de la generalización de los modelos de clasificación en datos biomédicos, cumpliendo así con los objetivos propuestos en el estudio.

k vecinos más cercanos (*K-Nearest Neighbors, KNN*), la regresión logística (*Logistic Regression*), bosques aleatorios (*Random Forest*) y los árboles de decisión (*Decision Tree*)

6.0.1. Resumen Comparación de algoritmos de clasificación:

El modelo bosques aleatorios (*Random Forest*) destacó con un rendimiento superior, especialmente al usar StratifiedKFold y Leave-One-Out (LOOCV). k vecinos más cercanos (*K-Nearest Neighbors, KNN*) también mostró mejoras, principalmente en la métrica de AUC, cuando se aplicaron estas técnicas. Los algoritmos árbol de decisión (*Decision Tree*) y regresión logística (*Logistic Regression*) tuvieron rendimientos menos sobresalientes, aunque KFold y StratifiedKFold lograron acercarlos al modelo base en ciertas métricas.

Así mismo podemos destacar que bosques aleatorios (*Random Forest (RF)*) haya obtenido los mejores resultados implica que su enfoque basado en su capacidad para combinar predicciones de múltiples árboles permite abordar problemas de sobreajuste y variabilidad, características críticas en escenarios donde los datos pueden ser ruidosos o incompletos. Además, su naturaleza no paramétrica le permite adaptarse a diversos tipos de datos sin requerir suposiciones estrictas, lo que aumenta su aplicabilidad en diferentes escenarios médicos, sin dejar de lado su capacidad para generar métricas de importancia de variables, lo que permite a los investigadores extraer conclusiones sobre los patrones subyacentes y cómo se llevaron a cabo las segmentaciones

En cuanto al buen desempeño también del algoritmo KNN sugiere que las instancias similares en términos de proximidad son indicadores fiables para la clasificación, lo que puede ser útil en diagnósticos basados en mediciones o características comparables o similares. KNN no asume una distribución específica de los datos, adaptándose bien a problemas clínicos con estructuras desconocidas. siendo fácil de implementar y entender, lo que permite su uso directo en diversas situaciones sin entrenar un modelo.

Si hablamos de la Regresión Logística podríamos decir que las relaciones entre las variables y la etiqueta objetivo no son exactamente lineales o que no se aproximan a una combinación lineal transformada, Así mismo podríamos concluir acerca del Árbol de Decisión, que las relaciones entre las variables no son relativamente simples y jerárquicas. Esto sugiere que las decisiones clínicas no pueden estructurarse de manera lógica y segmentada.

6.0.2. Identificación y mitigación de sesgos y sobreajustes:

Las técnicas de validación estratificada y ShuffleSplit principalmente, resultaron más efectivas en especial para Random Forest, permitiéndole manejar mejor los datos de naturaleza real o heterogénea de la investigación. Este comportamiento sugiere que estos métodos ayudan a mejorar la estabilidad y capacidad de generalización de los modelos en datos biomédicos.

6.1. Gráficos comparativos promedio

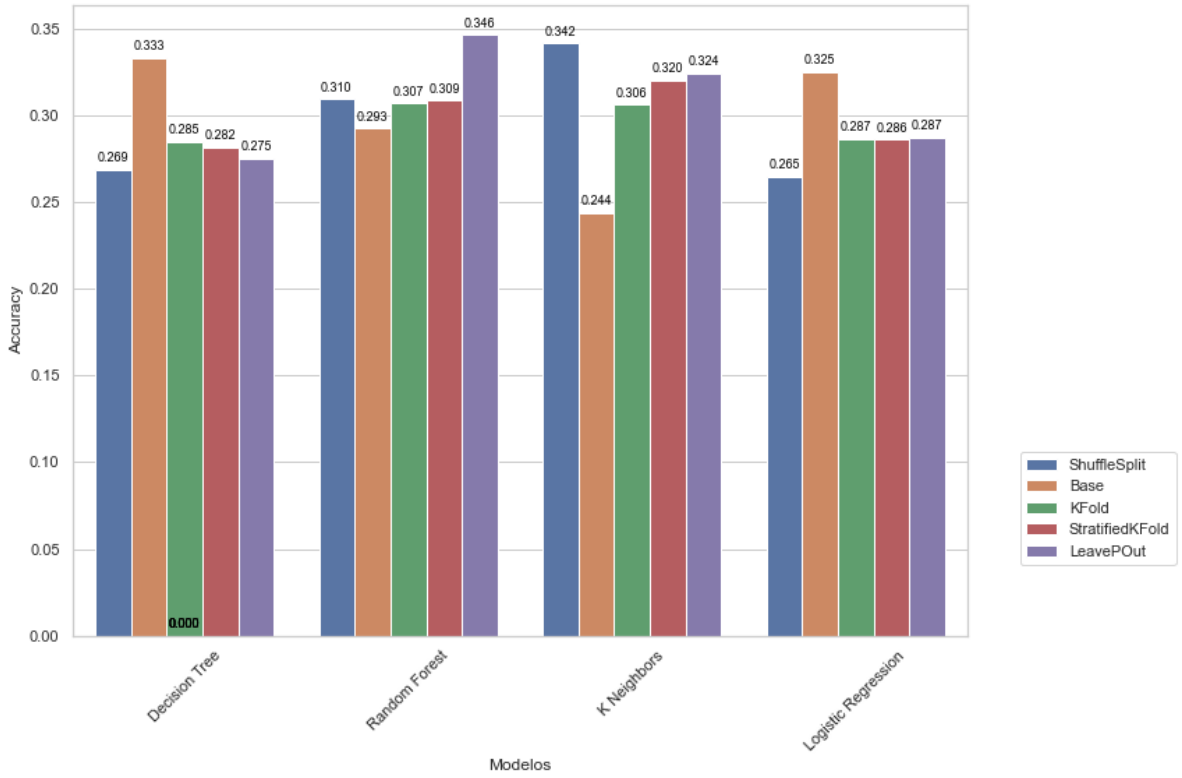

Figura 7: Comparación de Accuracy por Modelo y Método de Validación Cruzada

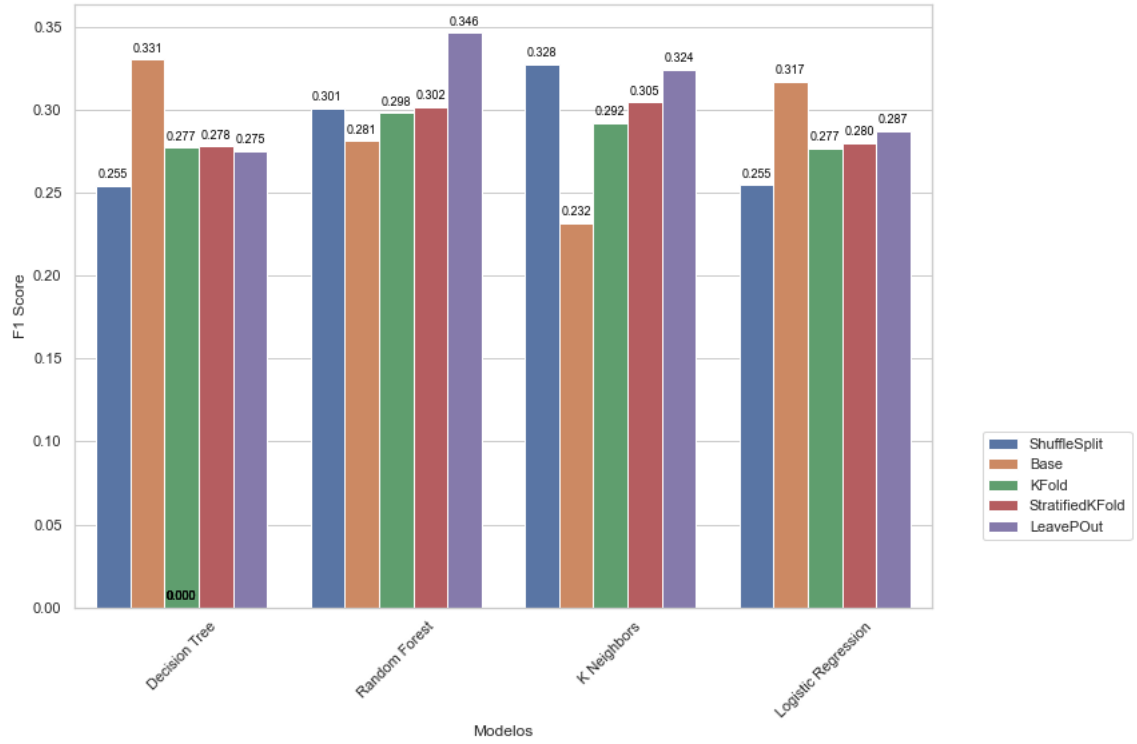

Figura 8: Comparación de F1 Score por Modelo y Método de Validación Cruzada

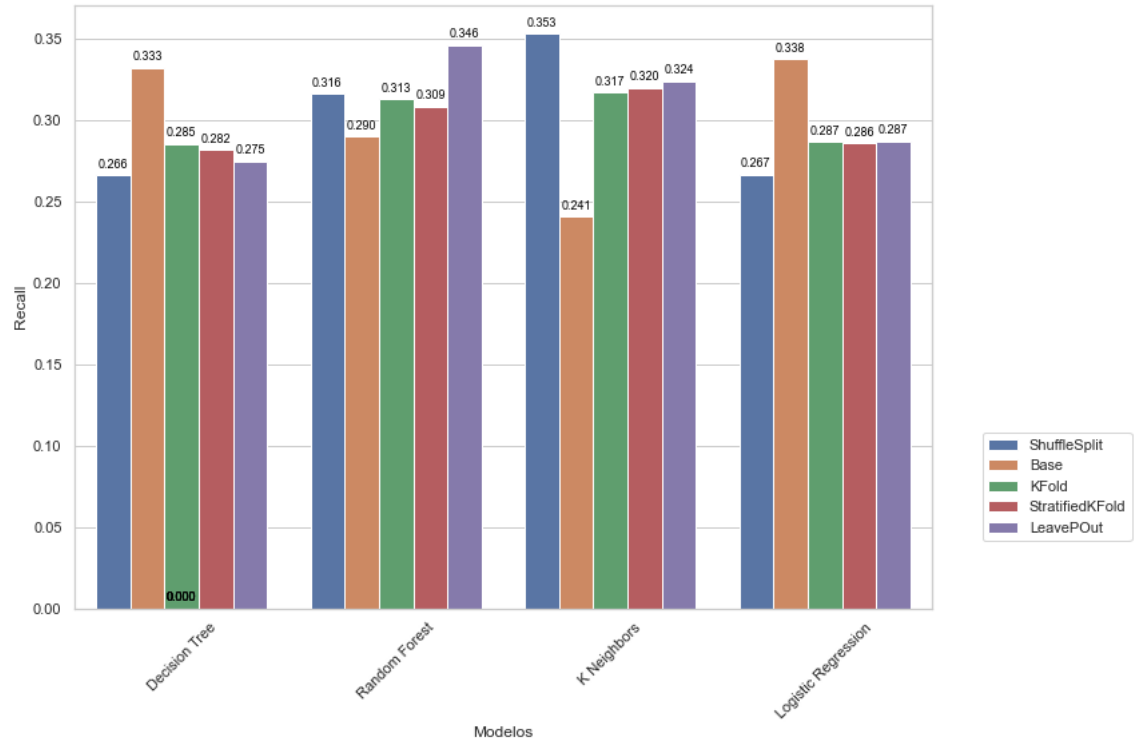

Figura 9: Comparación de Recall por Modelo y Método de Validación Cruzada

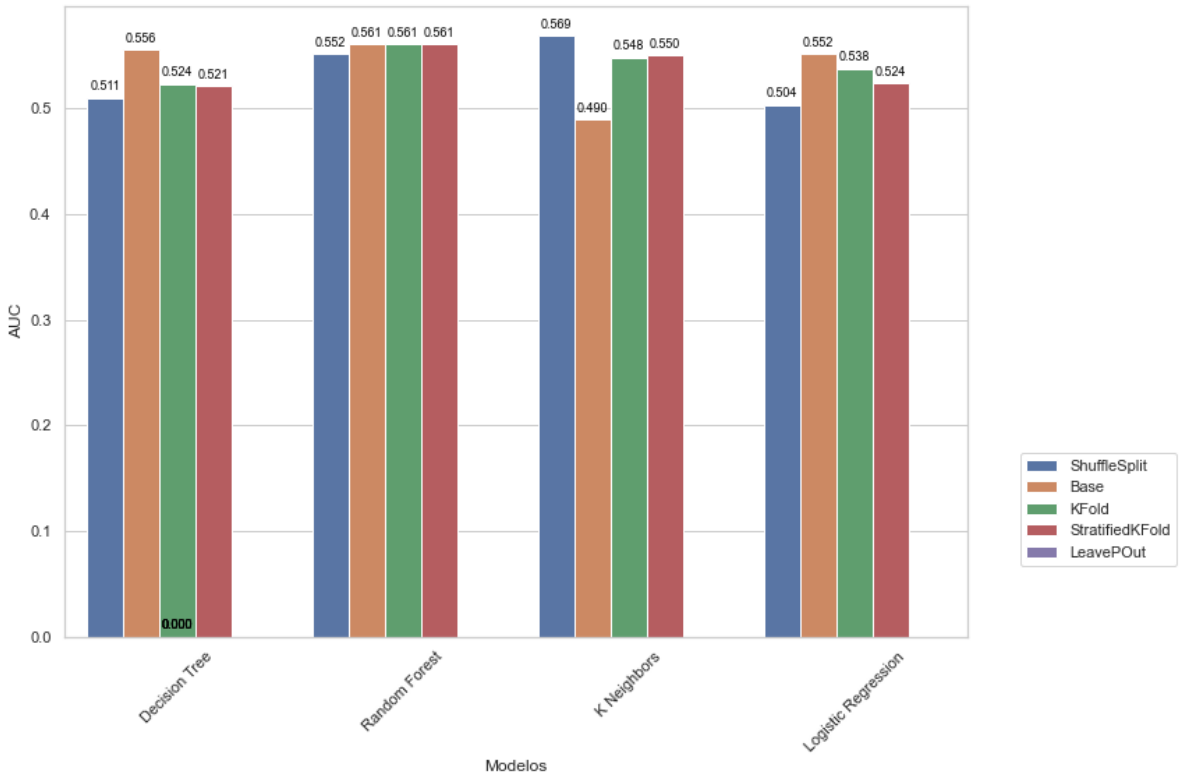

Figura 10: Comparación de AUC por Modelo y Método de Validación Cruzada

6.2. Interpretación gráficos comparativos promedio

Podemos reafirmar, a través de los gráficos 7, 8, 9 y 10, algunas de nuestras conclusiones sobre la efectividad de los métodos de validación cruzada y los modelos más eficientes en términos de métricas. En particular, destaca la estabilidad y calidad del modelo Random Forest, que mantiene un rendimiento consistente y alto con todos los métodos de validación cruzada. Este comportamiento resalta la robustez del Random Forest frente a las variaciones en los conjuntos de entrenamiento y validación.

Asimismo, es notable el rendimiento del modelo K Neighbors (KNN), que muestra una mejora significativa al utilizar los métodos de validación cruzada, en comparación con el modelo base. Aunque las métricas en los gráficos corresponden a promedios de todos los valores de k utilizados (2, 3, 4, 5, 6, 7, 8, 9, 15 y 20), se evidencia una tendencia positiva al aplicar estos métodos de validación. Esto sugiere que la validación cruzada ayuda a mitigar el sobreajuste y mejora la capacidad de generalización del modelo.

7. Granularidad Variable Objetivo

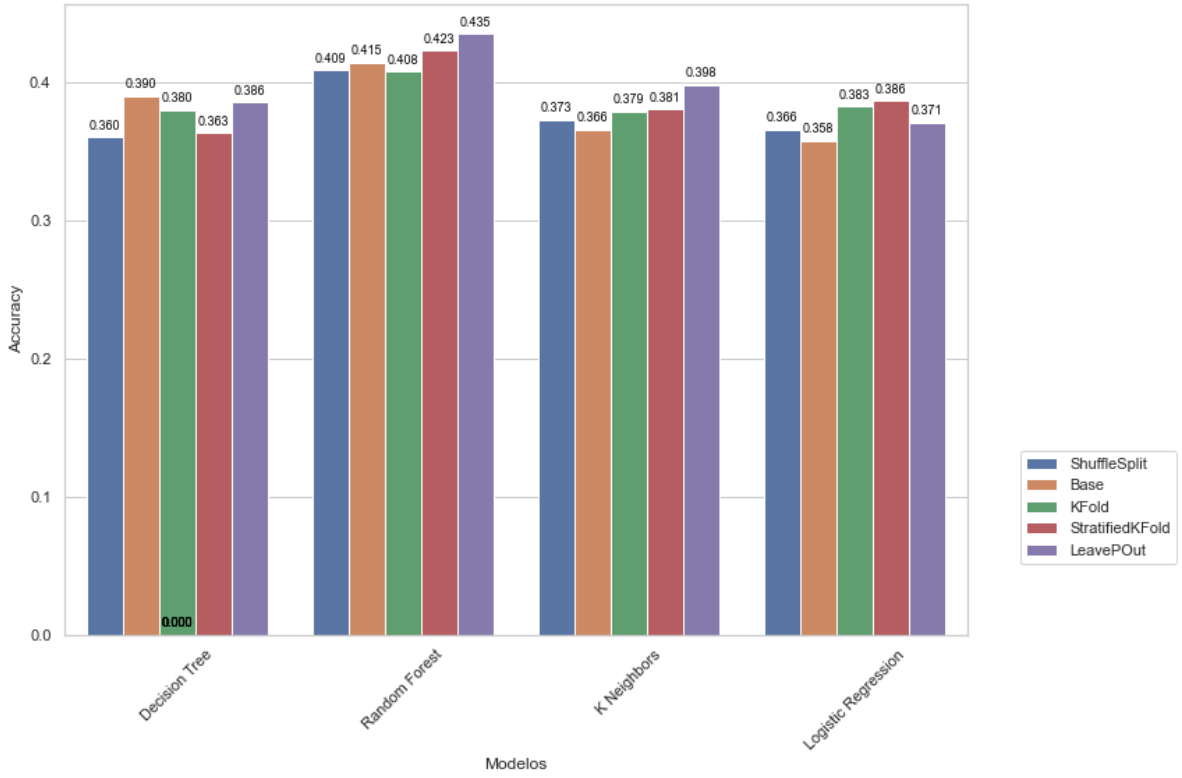

Figura 11: Comparación de Accuracy por Modelo y Método de Validación Cruzada con 3 categorias

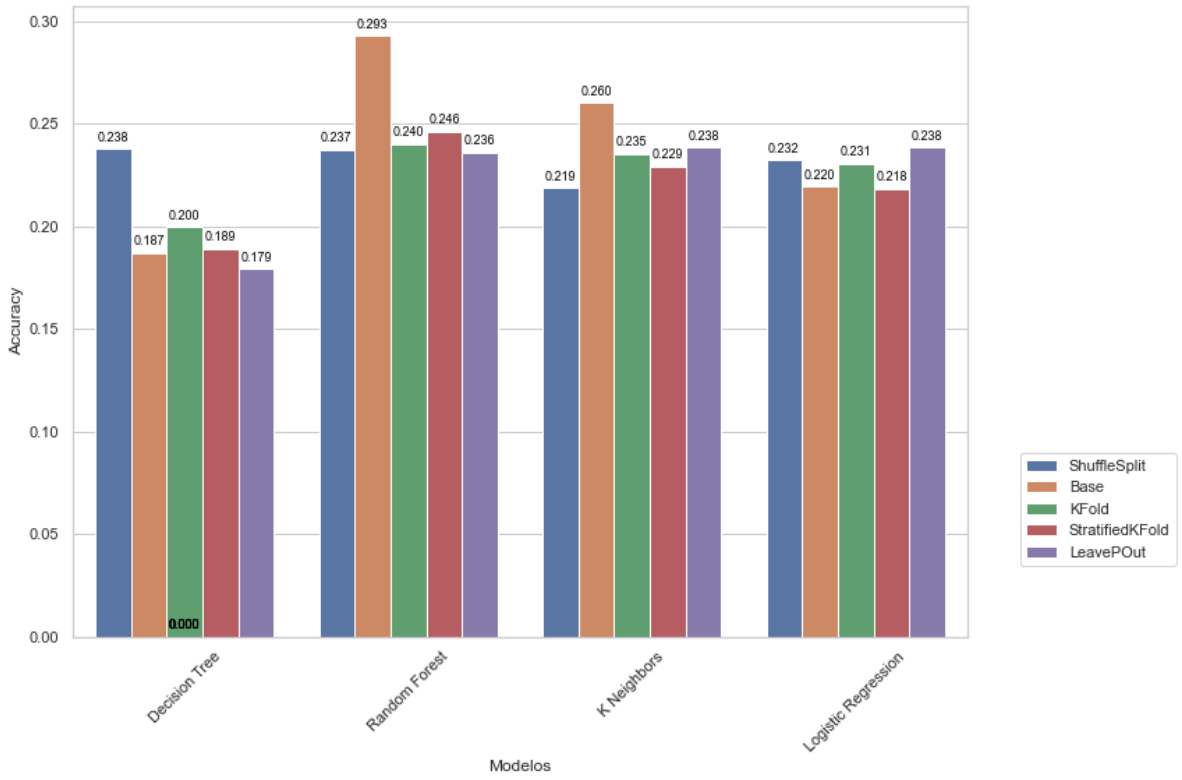

Figura 12: Comparación de Accuracy por Modelo y Método de Validación Cruzada con 6 categorias

Las Gráficas 7, 11 y 12 proporcionan una visión del desempeño de nuestros experimentos en términos de precisión (accuracy) para diferentes cantidades de categorías en la variable objetivo. La Figura 7 muestra los resultados del experimento con 4 categorías, la Figura 11 corresponde a los resultados con 3 categorías, y la Figura 12 presenta los resultados con 6 categorías.

El desempeño decreciente de las métricas de evaluación (accuracy, F1-score, recall y AUC) al aumentar el número de categorías de la variable objetivo puede atribuirse a la mayor complejidad que enfrenta el modelo al distinguir entre más clases, lo que reduce su capacidad predictiva. Por el contrario, la ganancia de desempeño observada al disminuir el número de categorías podría atribuirse a una mayor simplicidad del problema de clasificación. Al haber menos clases, los modelos tienen menos fronteras de decisión que aprender, lo que reduce el riesgo de sobreajuste y facilita una mejor generalización.

Sin embargo, esto no implica una pérdida de validez en el análisis, ya que el propósito principal es evaluar el impacto de las técnicas de validación cruzada en la mejora del desempeño del modelo. Es importante destacar que, en ejercicios supervisados similares, no siempre se tiene la potestad de modificar la granularidad de la variable objetivo, ya que los grupos o etiquetas suelen estar predefinidos por el diseño experimental o las características del dominio del problema. Este estudio ofrece la oportunidad de explorar cómo diferentes niveles de granularidad influyen en el rendimiento, lo cual enriquece el análisis.

A pesar de las diferencias en granularidad (3, 4 o 6 categorías), los resultados muestran consistentemente mejoras significativas al aplicar técnicas avanzadas de validación cruzada, lo que refuerza su utilidad para mitigar sesgos y sobreajustes. Por lo tanto, el número de categorías puede considerarse una configuración específica para explorar la sensibilidad del enfoque, sin afectar las conclusiones generales. En este contexto, el experimento se presenta como una prueba de concepto que demuestra que la validación cruzada mejora la generalización de los modelos, incluso en escenarios con diferentes niveles de granularidad en la variable objetivo.

Es importante señalar que esta investigación queda abierta a la exploración de diversos factores que podrían contribuir a mejorar las estimaciones. Entre ellos, se incluyen la implementación de más modelos, el ajuste y control de los hiperparámetros de cada uno, la selección adecuada del número de categorías para la variable objetivo, y la investigación en metodologías que promuevan la mejora continua de las estimaciones.

8. Recomendaciones prácticas:

Los hallazgos sugieren el uso de StratifiedKFold y ShuffleSplit en investigaciones biomédicas, debio a los beneficios de conservar posibles proporciones en bases desbalanceadas con el uso de StratifiedKFold y la reudcción de sesgos en ShuffleSplit que podemos obtener al mezclar las bases, esto junto con Random Forest o KNN como modelos preferentes debido a su robustez y capacidad de generalización bajo estas técnicas. También se recomienda ajustar el valor de k de acuerdo con el modelo y la técnica de validación seleccionada para optimizar el rendimiento.

Este estudio deja abierta la posibilidad de aplicar otros métodos adicionales, a pesar de que se han utilizado modelos como Random Forest, Decision Tree, K-Nearest Neighbors (KNN) y Logistic Regression, así como diferentes métodos de validación cruzada, incluyendo K-Fold, validación cruzada estratificada, ShuffleSplit y Leave-One-Out (LOOCV). Aunque los métodos empleados demostraron mejorar el rendimiento en términos de precisión y capacidad de generalización, se reconocen oportunidades para explorar técnicas alternativas que podrían complementar o enriquecer aún más los hallazgos.

Por ejemplo, el uso de métodos avanzados como redes neuronales podría ofrecer nuevas perspectivas, especialmente en el manejo de datos clínicos complejos y heterogéneos. Asimismo, no se profundizó en el impacto de la granularidad de la variable objetivo (3, 4 y 6 categorías) en los resultados, lo que sugiere un espacio para futuras investigaciones que puedan explorar cómo diferentes niveles de granularidad pueden influir en el rendimiento de los modelos, especialmente al combinarse con distintas técnicas de validación cruzada.

Además, la optimización de hiperparámetros de los modelos utilizados, como Random Forest, Decision Tree, KNN y Logistic Regression, también podría contribuir a mejorar aún más las estimaciones. La elección adecuada de hiperparámetros puede ayudar a controlar el sesgo y la variabilidad, lo que se traduce en una mejora del desempeño y la generalización de los modelos. Este aspecto también merece atención en futuras investigaciones, ya que puede potenciar el rendimiento al optimizar los parámetros según las características específicas de los datos clínicos.

Este estudio establece un punto de partida sólido, pero deja abiertas las puertas para futuras investigaciones que puedan explorar métodos adicionales, incluyendo el ajuste de hiperparámetros, con el objetivo de mejorar la precisión y generalización de los modelos en contextos clínicos.

9. Limitaciones

A pesar de los beneficios de las técnicas de validación cruzada en la mejora de la generalización de los modelos de clasificación, su aplicación tiene ciertas limitaciones y potenciales fuentes de sesgo que deben ser reconocidas:

1) Sensibilidad de Validación Cruzada Leave-One-Out(LOOCV) a conjuntos de datos grandes: Si bien LOOCV puede ser valioso en conjuntos de datos pequeños, su aplicación en conjuntos de datos de gran tamaño resulta computacionalmente costosa y puede no ser práctica. Este método implica entrenar el modelo tantas veces como observaciones haya en el conjunto de datos, lo que demanda significativos recursos computacionales y tiempo. Además, LOOCV tiende a producir una varianza elevada en las estimaciones del error, lo que puede llevar a resultados inconsistentes y a una posible sobreestimación de la capacidad del modelo para generalizar.

2) Sesgo potencial en k-fold cross-validation: Al dividir los datos en k subconjuntos para la validación cruzada, es posible que algunos subconjuntos no representen adecuadamente la distribución completa de los datos, especialmente si el valor de k es bajo o si los datos presentan un alto grado de heterogeneidad. Esto podría inducir sesgo en la estimación de error, afectando la fiabilidad de los resultados. Además, si los datos no están bien distribuidos entre los folds, el modelo puede mostrar un desempeño inconsistente debido a la variabilidad en la composición de los subconjuntos de entrenamiento y prueba.

10. Implicaciones éticas

En el contexto de este estudio, es fundamental abordar las implicaciones éticas asociadas con el manejo de datos sensibles. En todo momento, se garantizaron medidas estrictas para proteger la privacidad y la confidencialidad de los datos, asegurando que no se revelarán identidades de los individuos involucrados. Los datos recopilados fueron tratados de manera anónima y únicamente con fines de investigación, respetando las normativas éticas y legales aplicables, como las directrices de protección de datos personales. Además, se puede estar seguro al respecto del riesgo de sesgos demográficos, dado que los datos utilizados provienen de otro estudio, el cual ya había implementado criterios rigurosos para la selección de participantes. Esto incluyó el emparejamiento por edad, sexo, nivel socioeconómico y etnia, además de la clasificación de los hábitos de consumo de tabaco y alcohol, siguiendo criterios estandarizados reconocidos a nivel internacional. Estas medidas aseguran que los posibles sesgos demográficos fueron controlados adecuadamente, garantizando la validez y la precisión de los resultados, al tiempo que se resalta el compromiso con la ética y la transparencia en el tratamiento de datos sensibles.

Recibido: Marzo de 2025
Aceptado: Junio de 2025

Referencias

[1] M. Kuhn and K. Johnson. *Applied Predictive Modeling*. Springer, 2013.

[2] A. Geron. *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media, 2 edition, 2019.

[3] A. C. Guido. *Introduction to Python*. O'Reilly Media, Inc., 2016.

[4] ... Micronuclei frequency and exposure to chemical mixtures in three colombian mining populations. *Science of the Total Environment*, 889:165789, 2023.

[5] I. K. Nti, O. Nyarko-Boateng, and J. Aning. Performance of machine learning algorithms with different k values in k-fold cross-validation. *International Journal of Information Technology and Computer Science*, 6:61–71, 2021.

[6] I. Tougui, A. Jilbab, and J. El Mhamdi. Impact of the choice of cross-validation techniques on the results of machine learning-based diagnostic applications. *Healthcare Informatics Research*, 27(3):189–199, 2021.

[7] S. J. Norvig. *Instructor's Solution Manual Artificial Intelligence*. Pearson Education, Hoboken, 2022.

[8] J. McCarthy. What is artificial intelligence?, 2007. Accessed: 2024-07-21.

[9] N. Japkowicz. *Evaluating Learning Algorithms: A Classification Perspective*. Cambridge University Press, Cambridge, 2011.

[10] O. Simeone. arxiv preprint, 2018.

[11] F. Parra. *Estadística y Machine Learning con R*. ICANE, Santander, 2019.

[12] scikit-learn. Decision trees (dts), 2024. Accessed: 2024-03-12.

[13] A. Huertas Mora. Algoritmos de aprendizaje supervisado utilizando datos de monitoreo de condiciones: un estudio para el pronóstico de fallas en máquinas, 2020. Tesis de maestría, Universidad Santo Tomás Colombia.

[14] Datacamp. Random forests classifier in python, 2024. Accessed: 2024-03-12.

[15] S. Raschka. Stat 479: Machine learning, 2024.

[16] T. Emura and J.-H. Hsu. Estimation of the mann–whitney effect in the two-sample problem under dependent censoring. *Computational Statistics & Data Analysis*, 152:106990, 2020.