Pontificia Universidad
JAVERIANA
Cali

Clasificación basada en Machine Learning para la identificación de marcadores genéticos utilizando patrones estructurales asociados con cáncer de mama

Lina Yojana Gonzalez Martinez
Código 8993439
Carlos Eduardo Hurtado Siabato
Código 8992506
Camilo Andrés Pérez Ruiz
Código 8992309

Proyecto Aplicado para optar al título de
Magister en Ciencia de Datos

Director
Fabian Tobar Tosse

Facultad de Ingeniería y Ciencias
Maestría en Ciencia de Datos
Santiago de Cali, Junio 9 de 2025

TABLA DE CONTENIDO

LISTA DE FIGURAS

LISTA DE TABLAS

INTRODUCCIÓN

En el campo de la genómica, los Estudios de Asociación de Genoma Completo (GWAS, por sus siglas en inglés) y, más específicamente, los estudios de Secuenciación Exómica Completa (WES, por sus siglas en inglés), han revolucionado la capacidad para identificar variantes genéticas que pueden predisponer a diversas enfermedades. Estas variantes, que incluyen mutaciones y alteraciones estructurales en el Ácido Desoxirribonucleico (ADN), juegan un papel crucial en la comprensión de la predisposición, progresión y respuesta a las enfermedades.

Durante las últimas dos décadas, ha habido un crecimiento significativo en la recopilación de datos genéticos a gran escala, presentando desafíos en su gestión y análisis. Para abordar esta complejidad, se ha vuelto crucial emplear técnicas avanzadas de Machine Learning. En estudios WES, el Machine Learning se utiliza para descubrir relaciones complejas entre variantes genéticas y factores ambientales o hereditarios, aplicándose a enfermedades de alta morbilidad como el cáncer de mama. La aparición de este o cualquier otro tipo de cáncer implica variaciones estructurales en la composición que conducen a la pérdida o la sobreexpresión de los genes que son los encargados de regular el ciclo celular. Es por esto, que identificar con precisión marcadores genéticos -asociados a variantes genética (MG-VG)- es fundamental para avanzar en la medicina personalizada y mejorar estrategias en diagnóstico, tratamiento y prevención de este tipo de cáncer.

Teniendo en cuenta la riqueza que existe de bases de datos, la necesidad de identificar estos MG-VG y con el objetivo de maximizar el aprovechamiento de la información proveniente de la secuenciación, este proyecto empleó Machine Learning en el desarrollo de un método que permitió la identificación de MG-VG asociados al cáncer de mama familiar en una muestra representativa de 5 países latinoamericanos (Colombia, Perú, Guatemala, México y Argentina), utilizando modelos de clasificación basados en patrones estructurales o de composición aprendidos a partir de datos genómicos. Estos modelos evaluaron la probabilidad de que determinadas variantes genéticas actuaran como verdaderos factores patogénicos, considerando su contexto específico dado por las característica genómicas de su loci (posición en el genoma), y que a su ves definen los marcadores MG-VG [1].

Los modelos de clasificación facilitaron la identificación de características de variantes asociadas con la patogenicidad y el diagnóstico de cáncer en pacientes a partir de información estructural y funcional de las mismas, optimizando su desempeño mediante métricas críticas como precisión, exactitud, f1-score y matrices confusión. Además, este estudio puede sentar las bases para una arquitectura adaptable que podría extenderse a otros tipos de cáncer, ampliando así la exploración de marca-

dores genéticos y promoviendo avances significativos en medicina personalizada y estrategias de prevención de enfermedades.

1. DEFINICIÓN DEL PROBLEMA

1.1 PLANTEAMIENTO DEL PROBLEMA

En genómica, los estudios de asociación de genoma completo han demostrado ser herramientas invaluables para detectar e identificar variantes genéticas asociadas con diversas enfermedades. Estas variantes pueden incluir alteraciones o mutaciones en la información codificada en el genoma, desempeñando un papel crucial en la comprensión de la susceptibilidad y la progresión de las enfermedades [2].

Durante las últimas dos décadas, numerosas organizaciones e instituciones a nivel global han desarrollado extensas bases de datos sobre el genoma humano. Entre las más destacadas se encuentran el National Center for Biotechnology Information (NCBI), el European Bioinformatics Institute (EBI), The human genome browser at UCSC y Functional Annotation of Variants (FAVOR). Estas bases de datos almacenan un volumen masivo de información que supera los Petabytes, lo que presenta desafíos significativos en términos de gestión y análisis de datos [3].

Dada la inmensa cantidad de información, se ha vuelto imperativo utilizar técnicas avanzadas de Machine Learning para analizar y extraer valor de estas bases de datos genómicas. En el contexto específico de los estudios de asociación de genoma completo, el Machine Learning se emplea para identificar relaciones complejas entre variantes genéticas y factores tanto ambientales como hereditarios, aplicándose comúnmente a enfermedades de alta morbilidad.

Entre las enfermedades con mayor número de diagnósticos se destaca el cáncer de mama, que, según la Organización Mundial de la Salud, registró más de 2 millones de casos solo en 2022 [4]. Dado que la aparición de cualquier tipo de cáncer está asociada con variaciones estructurales que conducen a la inestabilidad en la expresión de los genes encargados de regular el crecimiento y la supervivencia de las células cancerosas [5], es esencial identificar con precisión marcadores genéticos asociados a esta variación, o patrones sobre las mismas. Esto a futuro permite avanzar en la medicina personalizada y mejorar las estrategias de prevención, diagnóstico y tratamiento de enfermedades.

Considerando los problemas mencionados anteriormente, el Machine Learning ofrece una solución potente para manejar y analizar estos datos de manera eficiente, pero requiere el desarrollo de modelos especializados para maximizar su eficacia en el contexto de los estudios WES.

Por lo tanto, este proyecto desarrolló modelos de clasificación basados en Machine Learning que permitieron identificar marcadores genéticos asociados a variantes genéticas (MG-VG) relacionados con la patogénesis de cáncer de mama familiar.

1.2 FORMULACIÓN DEL PROBLEMA

¿Cómo identificar marcadores genéticos a partir de patrones estructurales asociados con cáncer de mama, teniendo en cuenta su contexto genómico (en el loci), mediante modelos de Machine Learning?

2. OBJETIVOS DEL PROYECTO

2.1 OBJETIVO GENERAL

Implementar un modelo de clasificación basado en Machine Learning que permita identificar marcadores genéticos a partir de patrones estructurales asociados con cáncer de mama, teniendo en cuenta su contexto (loci), con el fin de evaluar la probabilidad de que se genere un factor patogénico.

2.2 OBJETIVOS ESPECÍFICOS

- Aplicar algoritmos de Machine Learning adecuados para el análisis y modelado de datos genómicos, incluyendo técnicas de clasificación supervisada.
- Identificar asociaciones en datos genómicos mediante modelos de clasificación y clustering, con el fin de enriquecer el conjunto de características relacionadas con los genes y sus alteraciones en el cáncer de mama
- Evaluar el rendimiento de los modelos en la identificación de marcadores genéticos asociados con enfermedades utilizando métricas de evaluación apropiadas, como precisión, sensibilidad y especificidad.

3. MARCO TEÓRICO Y ANTECEDENTES

3.1 MARCO TEÓRICO

3.1.1 Cáncer de mama

El cáncer de mama es una enfermedad en la que las células con alteraciones presentes en el seno se multiplican sin control y forman tumores que, si no son tratados, pueden propagarse al resto del cuerpo y causar la muerte. Según la Organización Mundial de la Salud, en 2022 se diagnosticaron 2.3 millones de casos de cáncer de mama, y aproximadamente 670,000 pacientes fallecieron a causa de esta enfermedad [4]. El cáncer de mama puede afectar a cualquier persona sin importar su edad y sexo, aunque el 99 % de los casos se da en mujeres en población adulta. Los riesgos de padecer esta enfermedad están vinculados a varios factores, entre ellos:

- **Factores genéticos:** son representados por alteraciones cromosómicas y de secuencia del ADN; algunos pueden segregarse de forma familiar. Aproximadamente del 5 % al 10 % de los cánceres de mama son atribuibles a factores genéticos. Los genes de predisposición al cáncer de mama más frecuentes son BRCA1, BRCA2, PTEN y TP53 [6].

  La expresión normal de estos genes puede inhibir el crecimiento de células tumorales, pero sus mutaciones eliminan esta capacidad inhibidora. Estas mutaciones genéticas en diversos genes explican hasta el 85 % de los casos de cáncer de mama en mujeres con antecedentes familiares [7].

- **Factores epigenéticos:** estos factores no alteran la secuencia del ADN, pero modifican la expresión del gen. Las modificaciones epigenéticas son referidas como cambios en la expresión génica sin afectar la secuencia nucleotídica. Los procesos epigenéticos alterados se centran en metilación del ADN, modificación de histonas, accesibilidad a la cromatina y la modificación postranscripcional del ARN, además de la traducción [8].

- **Factores ambientales modificables:** son aquellos que contribuyen al desarrollo de cáncer como el tipo de dieta, la ingesta de alcohol, el fumar, la obesidad y la inactividad física; estos factores también influyen en la recurrencia y tiempo de supervivencia [8].

3.1.2 Marcadores Genéticos y Variantes

Los marcadores genéticos representan secuencias definidas de ADN asociadas a características fenotípicas específicas. Estos elementos comprenden, entre otros, genes, repeticiones de ADN, elementos reguladores, patrones de metilación y regiones de variación que, en determinados contextos, han sido vinculadas a patologías

particulares. En el estudio del cáncer de mama, las variantes genéticas emergen como uno de los biomarcadores más frecuentes, englobando desde mutaciones puntuales hasta polimorfismos de un solo nucleótido (SNPs) de carácter patogénico, así como variaciones en el número de copias de segmentos del ADN. La complejidad inherente a la identificación de estos marcadores reside en la necesidad de integrar múltiples factores hereditarios y ambientales, lo que ha conducido al desarrollo del marco teórico de los Marcadores Genéticos asociados a Variantes Genéticas (MGVG). Por consiguiente, es indispensable evaluar tanto el impacto individual de cada marcador como las implicaciones sinérgicas derivadas de su interacción, en aras de dilucidar de manera precisa su contribución a la etiología y progresión de la enfermedad.

Además, es crucial reconocer que estos marcadores son factores de riesgo y no deben ser interpretados como diagnósticos definitivos. En el caso de enfermedades multifactoriales, los biomarcadores genéticos no pueden usarse de manera aislada como información genotípica diagnóstica de enfermedad. También es esencial evaluar su validez, es decir, la probabilidad de que quienes portan el biomarcador positivo desarrollen la enfermedad frente a la probabilidad de que aquellos que desarrollan la enfermedad porten el biomarcador [9].

Por otra parte, un ejemplo de marcador genético clave son los oncogenes y los genes supresores de tumores, los cuales son genes directamete relacionados con los diferentes tipos de cáncer. Específicamente, los oncogenes, cuando están mutados o sobre expresados, pueden causar el crecimiento descontrolado de las células; mientras que los genes supresores de tumores normalmente previenen el crecimiento celular descontrolado y promueven la reparación del ADN [10], por tanto su mutación genera un caos de descontrol. Distinguir entre ambos tipo de marcador es fundamental para comprender los mecanismos de patogenicidad en el cáncer de mama. Los modelos de clasificación pueden incorporar esta característica para mejorar la predicción de si una variante es patogénica o no patogénica.

3.1.4 Modelos de Machine Learning

La genómica involucra el estudio de los genes de un individuo, así como las interacciones entre diferentes genes y de ellos con su ambiente. Este campo de conocimiento produce conjuntos de datos diversos y valiosos para su aprovechamiento [11]. En lo referente a genómica y las relaciones evolutivas que surgen en la evaluación de genes, los algoritmos y métodos de machine learning toman mayor importancia. Respecto a los principales modelos usados en clasificación para una tarea supervisada, hay un gran abanico de modelos que permitirían derivar una predicción con

datos de entrenamiento lo suficientemente robusto. Entre estos se encuentran:

- **Support Vector Machine (SVM)**

  El algoritmo de Support Vector Machine (SVM) es una técnica de aprendizaje supervisado ampliamente utilizada para tareas de clasificación, debido a su capacidad para generar modelos con buen desempeño incluso sin un ajuste exhaustivo de parámetros. El funcionamiento de SVM se basa en la identificación de un hiperplano separador que divide el espacio de características en dos regiones, cada una asociada a una clase diferente. En los casos ideales donde los datos son linealmente separables, el algoritmo busca el hiperplano que maximiza la margen, es decir, la distancia entre dicho hiperplano y los puntos más cercanos de cada clase. Esta estrategia, conocida como clasificador de margen máximo, tiene como objetivo mejorar la capacidad de generalización del modelo [12].

  Cuando los datos no son linealmente separables, SVM permite la inclusión de cierto grado de error mediante la introducción de variables de holgura, lo que da lugar al clasificador de soporte vectorial. Esta extensión admite observaciones mal clasificadas para mejorar la flexibilidad del modelo. Además, el algoritmo puede ampliarse mediante el uso de funciones núcleo (kernels), que permiten transformar el espacio original de características a uno de mayor dimensión, donde la separación entre clases sea más factible mediante fronteras no lineales [12].

  En el ámbito de la medicina aplicada, el algoritmo SVM ha demostrado ser una herramienta útil para abordar problemas de clasificación complejos, incluyendo el análisis de imágenes médicas, la predicción de diagnósticos, la clasificación de tumores y la estimación de mortalidad hospitalaria. Su capacidad para manejar relaciones no lineales y trabajar eficazmente con conjuntos de datos de alta dimensionalidad lo hace especialmente adecuado para contextos clínicos donde la precisión es crítica.

  Un estudio comparó el desempeño de SVM con la regresión logística multivariada (MLR) en la predicción de mortalidad hospitalaria en pacientes críticamente enfermos con neoplasias hematológicas [13]. Los resultados mostraron que SVM superó ligeramente a MLR en términos del área bajo la curva ROC (AUC): 0.802 frente a 0.768 en el primer modelo, y 0.808 frente a 0.781 en el segundo modelo más complejo, aunque sin diferencias estadísticamente significativas.

No obstante, un aspecto clave fue la eficiencia del modelo SVM, que logró realizar sus predicciones utilizando solo cuatro variables, mientras que MLR requirió siete y ocho variables, respectivamente. Esto evidencia no solo la competitividad del modelo en cuanto a desempeño, sino también su capacidad de generar predicciones precisas con un menor número de variables, lo que resulta altamente ventajoso en entornos clínicos con datos limitados o costosos de obtener.

- **Random Forest (RF)**

Para comprender el modelo de Random Forest, es necesario partir de la definición de un árbol de decisión, ya que este constituye su unidad básica. Un árbol de decisión es un clasificador que divide recursivamente el espacio de atributos, estructurándose en forma jerárquica. En esta estructura, el nodo raíz representa el punto de partida sin conexiones entrantes, mientras que los nodos internos realizan particiones basadas en los valores de una única variable o atributo. Por su parte, los nodos terminales, también llamados hojas, representan las predicciones finales, ya sea como clases o como distribuciones de probabilidad sobre la variable objetivo [14].

En cada nodo de decisión se selecciona una variable y un punto de corte que mejor separen las observaciones según un criterio de impureza, como el índice de Gini o la entropía. El objetivo es construir divisiones que generen subgrupos homogéneos respecto a la variable de salida [12].

El modelo de Random Forest se basa en la construcción de un conjunto de árboles de decisión, combinados para mejorar la precisión y reducir la varianza del modelo individual. Esta técnica, propuesta inicialmente por Breiman, introduce dos fuentes de aleatoriedad: (i) cada árbol es entrenado sobre una muestra aleatoria con reemplazo del conjunto de entrenamiento (bootstrap), y (ii) en cada división dentro del árbol, se considera un subconjunto aleatorio de variables en lugar de todas [12]. Esta estrategia genera árboles diversificados, disminuyendo la correlación entre ellos y, en consecuencia, aumentando la estabilidad del modelo final.

En cuanto al proceso de predicción, Random Forest combina los resultados individuales de cada árbol para generar una predicción final. En tareas de clasificación, cada árbol emite un voto por una clase, y la clase con más votos se

selecciona como predicción final del modelo (votación por mayoría). En tareas de regresión, se calcula el promedio de las predicciones individuales de todos los árboles. Este enfoque de combinación ayuda a reducir la varianza del modelo final, ofreciendo predicciones más estables y precisas [12].

Desde el punto de vista de aplicación, Random Forest ha demostrado una amplia utilidad en el ámbito de las ciencias biológicas, especialmente en medicina y genómica, en tareas como la clasificación de expresiones génicas, la identificación de genes asociados a enfermedades y la predicción de riesgos clínicos.

Un ejemplo destacado su aplicación para predecir el riesgo de ocho categorías de enfermedades utilizando datos altamente desbalanceados. En este trabajo, el modelo superó a enfoques como SVM, bagging y boosting en términos del área bajo la curva ROC (AUC), demostrando su eficacia en entornos complejos [15]. Además, una de las principales ventajas de Random Forest es su capacidad para estimar la importancia relativa de cada variable, lo que permite identificar los factores más influyentes en el proceso de clasificación e interpretar mejor la relevancia clínica o genética de los datos.

- **Gradient Boosted Decision Tree (GBDT) y Extreme Gradient Boosting (XGBoost)**

  Gradient Boosted Decision Tree (GBDT) y Extreme Gradient Boosting (XGBoost) Los modelos de Extreme Gradient Boosting son una implementación eficiente de los árboles de decisión impulsados por gradiente (GBDT). La implementación, disponible a través de Python como la librería XGBoost ofrece procesamiento distribuído y es fácilmente escalable, convirtiéndolo en una opción atractiva al trabajar con conjuntos de datos de gran tamaño y con tareas de alto rendimiento [16].

  Los modelos XGBoost optimizan una función objetivo que incluye una función de pérdida y un término de regularización. La función de pérdida asegura que la diferencia entre los valores predichos por el modelo y los valores reales se minimicen de modo que se guíe al modelo hacia una mejor precisión en sus predicciones. Por otro lado, el término de regularización controla la complejidad del modelo al prevenir el sobreajuste mediante la penalización de modelos de muy alta complejidad [17].

  Estos modelos han sido implementados con éxito en estudios enfocados en

la predicción de enfermedades, diagnóstico asistido por computadora, y pronósticos clínicos como la evaluación de riesgos de mortalidad. Un ejemplo representativo de su utilidad es el estudio titulado Construction of the XGBoost model for early lung cancer prediction based on metabolic indices [18], donde se propuso un enfoque interdisciplinario que combina datos de metabolómica con el modelo XGBoost para la predicción temprana del cáncer de pulmón.

En dicho estudio, los metabolitos ornitina y palmitoilcarnitina emergieron como biomarcadores clave, permitiendo al modelo alcanzar una precisión del 75.29 %, una sensibilidad del 74 %, y un valor AUC de 0.81. Estos resultados reflejan el potencial del modelo XGBoost no solo para identificar indicadores tempranos de la enfermedad, sino también como una alternativa diagnóstica más rápida, precisa y menos invasiva que los métodos tradicionales.

Xgbfir (XGBoost Feature Interactions and Importance Ranking)

Xgbfir es una herramienta diseñada para analizar la interpretabilidad de modelos construidos con XGBoost, mediante la extracción y el análisis detallado de la importancia de las características y sus interacciones [19]. Esta herramienta funciona como un parseador del volcado de un modelo de XGBoost y evalúa tanto las características individuales como las combinaciones de características (interacciones) utilizando diversas métricas.

Entre las métricas calculadas se incluyen:

- *Gain:* ganancia total atribuida a una característica o interacción.
- *FScore:* número de divisiones realizadas.
- *wFScore:* FScore ponderado por la probabilidad de cada división.
- *Average Gain* y *Average wFScore:* Promedios de Gain y wFScore respectivamente.

Además, xgbfir proporciona estadísticas sobre la posición media de las divisiones dentro de los árboles (*Average Tree Index* y *Average Tree Depth*), así como histogramas de los valores de corte (*Split Value Histograms*) y estadísticas de hojas. Esta información resulta fundamental para comprender la lógica interna de los modelos de gradiente boosting, identificar características relevantes o redundantes, y mejorar la transparencia de los modelos en aplicaciones críticas [19].

Adicionalmente, muchas etapas iniciales usan técnicas de aprendizaje no supervisado para obtener patrones no visibles en la representación real de los datos, algunas técnicas que podrían ser exploradas en el desarrollo del presente proyecto son:

- **HDBSCAN* (Hierarchical Density-Based Spatial Clustering of Applications with Noise)**

  HDBSCAN* es una técnica de agrupamiento no supervisado que extiende el algoritmo DBSCAN mediante un enfoque jerárquico basado en la densidad. Su objetivo principal es identificar clústeres de diferentes densidades en los datos, mientras maneja de manera explícita la detección de ruido [20]. A diferencia de DBSCAN, que requiere un único parámetro de densidad, HDBSCAN* construye una estructura jerárquica de agrupamientos basada en la variación de la densidad a través de múltiples escalas.

  El funcionamiento del algoritmo se desarrolla a través de los siguientes pasos:

  1. Cálculo del grafo de vecinos más cercanos ponderado por la distancia mutua, ajustada a la densidad local de cada punto (distancia de alcance mutuo).
  2. Construcción de un árbol de dendrograma mínimo (Minimum Spanning Tree) a partir de las distancias de alcance mutuo entre los puntos.
  3. Condensación de la jerarquía: se recorre el árbol para identificar y preservar agrupamientos persistentes a través de distintas escalas de densidad, eliminando aquellas ramas que representan ruido o clústeres inestables.
  4. Extracción de clústeres planos mediante la optimización de una medida de estabilidad, que cuantifica la persistencia de los clústeres a través de las distintas escalas de densidad.

  HDBSCAN* presenta varias ventajas significativas dado no requiere especificar el número de clústeres a priori, puede identificar agrupamientos de formas arbitrarias y es robusto frente a la presencia de ruido y variaciones de densidad en los datos. El algoritmo maximiza la estabilidad de los clústeres, buscando particiones que reflejen agrupamientos verdaderamente persistentes en la estructura subyacente de los datos [20].

  Debido a estas características, HDBSCAN se ha consolidado como una herramienta versátil para el análisis exploratorio de datos complejos, especialmente en escenarios donde los patrones subyacentes no son evidentes y los datos contienen ruido o presentan agrupamientos con densidades variables. Esta

capacidad robusta trasciende el ámbito computacional para aplicarse en problemas reales del campo de la salud, donde la heterogeneidad y el ruido en los datos son comunes.

Un ejemplo destacado de su aplicación médica se encuentra en el análisis de datos de secuenciación de ARN a nivel de células individuales obtenidas del líquido broncoalveolar en pacientes con COVID-19 [21]. En dicho estudio, HDBSCAN permitió identificar tres grupos celulares principales, además de varios grupos minoritarios, detectando diferencias significativas en la abundancia celular según la severidad clínica de los pacientes. Teniendo en cuenta que los macrófagos, células T y células epiteliales constituyen los tipos celulares predominantes en estas muestras, se concluyó que HDBSCAN resulta especialmente útil para diferenciar y clasificar tipos celulares heterogéneos en conjuntos de datos biomédicos sin requerir etiquetas previas.

Además, la combinación de HDBSCAN con métodos neurodifusos permitió caracterizar tanto las diferencias entre estados moderados y graves de COVID-19 como la variabilidad interpaciente y el efecto de mutaciones virales en la respuesta inmune. Esta capacidad para capturar la estructura jerárquica y heterogénea de los datos clínicos facilitó a los investigadores clasificar con mayor precisión a los pacientes según su gravedad, brindando un valioso soporte para la toma de decisiones médicas en entornos clínicos complejos

- **Unified Manifold Approximation and Projection (UMAP)**

UMAP es una técnica de reducción de dimensionalidad que se basa en la teoría de la topología y la geometría. Su objetivo es preservar la estructura local y global de los datos mientras los proyecta en un espacio de menor dimensión. Las representaciones en menor dimensión, se construyen a partir de una representación de los datos en un espacio de alta dimensión, donde se modela la distancia entre puntos utilizando un gráfico de vecinos [22]. UMAP se utiliza a menudo como un paso previo a técnicas de clustering. Al reducir la dimensionalidad de los datos, UMAP facilita la identificación de grupos o patrones en los datos que pueden no ser evidentes en su forma original. Esto es particularmente útil en datos complejos, como imágenes o datos genómicos [23].

UMAP no solo se limita a la reducción de dimensionalidad para datos tabulares, sino que también ha demostrado ser eficaz en el análisis de imágenes biomédicas con estructuras espaciales complejas. En [24] se aplicó una variante denominada spatial UMAP para explorar datos de inmunofluorescencia

múltiple (mIF) en muestras tumorales de pacientes con melanoma metastásico. Esta técnica permitió conservar la información espacial a nivel celular, facilitando la identificación de vecindarios inmunes y patrones topográficos de expresión proteica relacionados con el pronóstico clínico.

Por ejemplo, se observó que la mayor expresión de la proteína PD-L1 se concentraba en células CD163+ situadas en zonas con alta densidad de células CD8+, un hallazgo que podría no haber sido evidente sin el uso de UMAP. Esto demuestra cómo UMAP puede ser una herramienta valiosa en entornos donde la estructura espacial de los datos es crítica para la interpretación y la toma de decisiones.

Métricas de distancia en UMAP

UMAP permite el uso de diversas métricas de distancia para modelar la similitud entre puntos en el espacio de alta dimensión, lo que influye directamente en la construcción del gráfico de vecinos y en la estructura del espacio embebido. Entre las métricas más comunes se encuentran: *euclidiana*, *manhattan*, *minkowski*, *coseno*, *correlación* y *chebyshev*.

La distancia euclidiana mide la separación en línea recta entre dos puntos, mientras que la distancia manhattan suma las diferencias absolutas en cada dimensión, lo que puede hacerla más robusta frente a valores atípicos [25]. La distancia minkowski generaliza ambas anteriores mediante un parámetro p, que permite ajustar la sensibilidad al tipo de diferencias entre coordenadas. Las métricas basadas en ángulos, como la distancia coseno y la distancia de correlación, son útiles cuando la magnitud de los datos es menos importante que su dirección o patrón [25].

Por su parte, la distancia chebyshev mide la mayor diferencia en cualquier dimensión, formando vecindarios de forma hipercúbica. La elección de la métrica impacta en la forma de los agrupamientos, la preservación de la estructura local, y la sensibilidad al ruido en el proceso de reducción de dimensionalidad [22].

3.1.5 Desbalance en bases de datos en Machine Learning

El aprendizaje automático con conjuntos de datos desbalanceados constituye un reto frecuente en ámbitos tan diversos como la minería de textos, el análisis biomédico

y la detección de fraudes financieros. Este problema surge cuando la distribución de las clases es desigual: la clase minoritaria, que agrupa los ejemplos más escasos pero a menudo más determinantes, aparece ampliamente subrepresentada frente a la clase mayoritaria. En entornos clínicos, por ejemplo, los registros de pacientes enfermos suelen ser más abundantes que los de pacientes sanos, ya que las personas con síntomas reciben atención médica y pruebas diagnósticas con mayor frecuencia, mientras que quienes gozan de buena salud no se someten de forma rutinaria a estos exámenes. Esta desproporción puede inducir un sesgo en los modelos de clasificación, disminuyendo su capacidad para detectar correctamente instancias pertenecientes a la clase minoritaria. Por esta razón, el desarrollo y la aplicación de técnicas de remuestreo, tanto a nivel de datos como de algoritmos, se han convertido en una línea de investigación clave para garantizar un aprendizaje más justo y representativo, particularmente en contextos donde los errores de predicción pueden tener consecuencias significativas.

- **Oversampling (sobre-muestreo)**

  Es una técnica utilizada para abordar el desbalance en conjuntos de datos, especialmente en problemas de clasificación binaria en los que una de las clases (la clase mayoritaria) está sobrerrepresentada en comparación con la clase minoritaria. El objetivo principal del submuestreo es reducir el número de instancias pertenecientes a la clase mayoritaria con el fin de equilibrar la distribución de clases y mejorar el desempeño de los algoritmos de aprendizaje automático, que tienden a sesgarse hacia la clase más frecuente [26].

  Existen diversos métodos de submuestreo. Uno de los más conocidos es el de Tomek Links [27]., que elimina pares de muestras cercanas entre clases opuestas para limpiar el límite de decisión y reducir el solapamiento entre clases. Otro enfoque es el uso de centroides de clústeres [28]., donde las instancias de la clase mayoritaria se agrupan mediante algoritmos de clustering (como K-means) y se seleccionan los centroides como representación condensada del grupo, disminuyendo así el tamaño del conjunto sin perder diversidad estructural.

- **Undersampling (sub-muestreo)**

  Es una técnica empleada para abordar el desbalance de clases en conjuntos de datos, particularmente en tareas de clasificación donde la clase minoritaria se encuentra significativamente subrepresentada. Consiste en aumentar artificialmente la proporción de instancias de dicha clase con el objetivo de

equilibrar la distribución de clases y permitir que los modelos de aprendizaje automático aprendan patrones relevantes sin verse sesgados hacia la clase mayoritaria[26].

Este aumento puede lograrse de dos formas principales: repitiendo instancias existentes de la clase minoritaria (sobremuestreo aleatorio), o generando nuevas instancias sintéticas a partir de las ya existentes. La segunda estrategia, más sofisticada, busca enriquecer la representación de la clase minoritaria sin introducir duplicaciones que puedan conducir al sobreajuste [29].

Uno de los métodos más conocidos en esta categoría es Borderline-SMOTE (Synthetic Minority Over-sampling Technique)[29], el cual se centra en generar nuevas instancias sintéticas específicamente en las regiones limítrofes entre clases, donde la clasificación suele ser más ambigua. Esto se hace con el objetivo de reforzar el aprendizaje en los márgenes de decisión y mejorar la capacidad del modelo para distinguir entre clases en escenarios complejos.

- **SMOTE (Synthetic Minority Oversampling Technique) y SMOTENC (SMOTE for Nominal and Continuous data)**

  SMOTE es una técnica ampliamente empleada en el aprendizaje supervisado para contrarrestar el desequilibrio de clases. Cuando una clase está poco representada en los datos de entrenamiento, los algoritmos tradicionales tienden a favorecer a la clase mayoritaria. SMOTE soluciona este problema generando ejemplos sintéticos para la clase minoritaria en lugar de simplemente replicar los existentes, lo que podría conducir a un sobreajuste.

  Para cada muestra de la clase minoritaria, SMOTE comienza calculando sus k vecinos más cercanos utilizando una métrica de distancia en el espacio de atributos. Una vez identificados estos vecinos, el algoritmo genera nuevos ejemplos sintéticos de la siguiente manera:

  1. Determina la diferencia entre el vector de atributos de la muestra original y el de uno de sus vecinos.
  2. Multiplica dicho vector de diferencias por un número aleatorio entre 0 y 1.
  3. Suma el resultado al vector de atributos original.

  De este modo se crea un nuevo punto sintético ubicado a lo largo del segmento de la interpolación lineal que une la muestra original con su vecino. El proceso

se repite hasta alcanzar el nivel deseado de sobremuestreo, rellenando el espacio de la clase minoritaria con ejemplos plausibles y distintos. Este método no solo equilibra la distribución de las clases, sino que también facilita la generalización del límite de decisión, reduciendo la probabilidad de sobreajuste que implicaría la simple replicación de ejemplos [30].

Por otro lado, SMOTENC es una variante de SMOTE diseñada para trabajar con conjuntos de datos que integran tanto atributos numéricos como categóricos. Mientras que SMOTE opera únicamente sobre variables continuas, generando nuevos ejemplos sintéticos mediante la interpolación lineal entre instancias de la clase minoritaria, SMOTENC adapta este proceso para datos mixtos. Específicamente, para los atributos continuos se aplica el mismo procedimiento de interpolación y para los atributos categóricos se emplean técnicas de codificación o funciones de distancia especializadas. De este modo, al generar nuevos valores para las variables categóricas se selecciona aleatoriamente una categoría entre los k vecinos más cercanos garantizando que el nuevo valor sea coherente con el patrón observado evitando combinaciones no plausibles [31].

- **ADASYN (Adaptive Synthetic Sampling)**

ADASYN (Adaptive Synthetic Sampling) es una técnica de sobremuestreo que, al igual que SMOTE, busca mitigar el desequilibrio de clases en problemas de clasificación supervisada. Su principal particularidad radica en que adapta la generación de ejemplos sintéticos en función de la dificultad que presenta cada observación minoritaria para ser aprendida correctamente por el clasificador. Es decir, prioriza la generación de nuevos datos sintéticos en aquellas regiones del espacio de atributos donde la clase minoritaria se encuentra menos representada o rodeada de muestras mayoritarias, lo cual contribuye a mejorar el desempeño del modelo en zonas conflictivas [32].

El procedimiento de ADASYN comienza evaluando la distribución de los vecinos más cercanos para cada instancia de la clase minoritaria. En particular, para cada una de estas instancias, se identifican sus k vecinos más próximos en el espacio de características. A continuación, se calcula la proporción de vecinos que pertenecen a la clase mayoritaria. Este valor se interpreta como una medida de dificultad: a mayor proporción, mayor complejidad para clasificar correctamente dicha muestra.

Con base en esta información, ADASYN determina la cantidad de nuevos ejem-

plos que se deben generar para cada instancia minoritaria. Aquellas con mayor proporción de vecinos mayoritarios recibirán un mayor número de ejemplos sintéticos. Posteriormente, el algoritmo genera estos ejemplos mediante interpolación lineal, de manera similar a SMOTE:

1. Se selecciona aleatoriamente uno de los k vecinos más cercanos de la muestra minoritaria.
2. Se calcula el vector diferencia entre la muestra original y su vecino.
3. Se multiplica este vector por un número aleatorio entre 0 y 1.
4. Se suma el resultado al vector de atributos original.

Este enfoque adaptativo permite centrar la generación sintética en las regiones más complejas del espacio, lo que contribuye a un límite de decisión más flexible y efectivo. Al enfocarse en las zonas de mayor confusión entre clases, ADASYN mejora la capacidad de generalización del modelo sin incurrir en un sobreajuste innecesario en áreas donde la clase minoritaria ya está bien representada [32].

3.1.6 Optimización de hiperparámetros en Modelos de Machine Learning

La optimización de hiperparámetros es un paso fundamental en el desarrollo de modelos de machine learning, ya que permite ajustar configuraciones que afectan directamente su rendimiento. A diferencia de los parámetros que se aprenden durante el entrenamiento, los hiperparámetros deben definirse de antemano.

Entre los métodos más sofisticados destaca la optimización bayesiana [33], que modela la función de rendimiento del modelo para seleccionar de manera inteligente los nuevos puntos a evaluar. En este marco, el Tree-structured Parzen Estimator (TPE) [34] se ha consolidado como una técnica eficaz, ya que utiliza estimadores de densidad no paramétricos para modelar de forma independiente la probabilidad de obtener conjuntos de hiperparámetros buenos o malos, lo que permite explorar de manera dirigida y eficiente espacios complejos.

Asimismo, otros enfoques avanzados incluyen la optimización bayesiana basada en procesos gaussianos [33], que recurre a procesos estocásticos para predecir el rendimiento mediante funciones de adquisición, y los algoritmos evolutivos [35], que imitan procesos de selección natural para evolucionar poblaciones de configuraciones. Por otro lado, métodos híbridos como Hyperband [36] combinan estrategias de asignación de recursos y eliminación progresiva, focalizando el cómputo en las

configuraciones más prometedoras y reduciendo significativamente el costo computacional. En conjunto, cada uno de estos métodos aprovecha la información sobre la distribución del rendimiento para guiar la búsqueda de configuraciones óptimas de forma más efectiva que los enfoques tradicionales.

3.1.7 Métricas de evaluación de desempeño

- **Aprendizaje supervisado**

  - **Exactitud:** La exactitud (Accuracy) es la proporción de predicciones correctas realizadas sobre el total de casos evaluados. Se calcula como el cociente entre la suma de verdaderos positivos (TP) y verdaderos negativos (TN), y el total de instancias del conjunto.

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

  - **Precisión:** La precisión mide la capacidad del modelo para identificar correctamente las instancias positivas de entre todas aquellas clasificadas como positivas. Es decir, se enfoca en minimizar los falsos positivos. Su fórmula es:

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

  - **Sensibilidad/Recall:** El recall (sensibilidad o recuperación) evalúa la capacidad del modelo para detectar correctamente todas las instancias reales positivas, reduciendo así los falsos negativos. Se expresa como:

$$Recall = \frac{TP}{TP + FN}$$

  - **F1-Score:** El F1-Score es la media armónica entre la precisión y el recall, ofreciendo un balance entre ambas métricas, especialmente útil en contextos donde existe un desbalance entre las clases. Se calcula mediante:

$$F1 - Score = 2 \cdot \frac{Precision \times Recall}{Precision + Recall}$$

- **Matriz de confusión:** La matriz de confusión es una tabla que permite visualizar el desempeño de un modelo de clasificación comparando las etiquetas reales con las predichas. Cada fila representa las instancias de una clase real y cada columna las correspondientes predicciones, facilitando el cálculo de métricas como accuracy, precisión y recall.

■ **Aprendizaje no supervisado**

- **Coeficiente de Silueta:** El coeficiente de silueta evalúa la calidad del agrupamiento midiendo qué tan similares son los objetos dentro de un mismo cluster en comparación con los objetos de otros clusters. Su valor varía entre -1 y 1, donde valores cercanos a 1 indican una mejor definición del agrupamiento [37].

  Para cada punto i, se definen:

  - $a(i)$: Promedio de la distancia entre el punto i y los demás puntos dentro de su mismo cluster C_i.

$$a(i) = \frac{1}{|C_i| - 1} \sum_{\substack{j \in C_i \\ j \neq i}} d(i, j)$$

  - $b(i)$: Mínimo promedio de distancia entre el punto i y los puntos de otro cluster $C \neq C_i$.

$$b(i) = \min_{C \neq C_i} \left(\frac{1}{|C|} \sum_{j \in C} d(i, j) \right)$$

  Entonces, el coeficiente de silueta para el punto i se calcula como:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

  El coeficiente de silueta global es el promedio de $s(i)$ para todos los puntos:

$$S = \frac{1}{N} \sum_{i=1}^{N} s(i)$$

- **Índice de Calinski-Harabasz:** Este índice compara la dispersión intra-cluster con la dispersión inter-cluster. Un valor mayor indica agrupamientos más compactos y mejor separados [38].

  Sean:

- K: Número de clusters.
- N: Número total de muestras.
- C_i: Conjunto de puntos del cluster i.
- μ_i: Centroide del cluster i.
- μ: Centroide global del conjunto de datos.

La dispersión entre clusters es:

$$B_K = \sum_{i=1}^{K} |C_i| \cdot \|\mu_i - \mu\|^2$$

La dispersión intra-cluster es:

$$W_K = \sum_{i=1}^{K} \sum_{x \in C_i} \|x - \mu_i\|^2$$

Así, el índice se define como:

$$CH = \frac{B_K/(K-1)}{W_K/(N-K)}$$

- **Índice de Davies-Bouldin:** Este índice mide la similitud entre clusters, considerando la dispersión interna de cada uno y la distancia entre clusters. Valores más bajos indican mejor separación y compacidad [39].

  Para cada cluster i, se define la dispersión interna:

  $$s_i = \frac{1}{|C_i|} \sum_{x \in C_i} \|x - \mu_i\|$$

  Para cada par de clusters i y j, se calcula:

  $$R_{ij} = \frac{s_i + s_j}{\|\mu_i - \mu_j\|}$$

  Luego, para cada cluster i, se toma:

  $$R_i = \max_{j \neq i} R_{ij}$$

  Finalmente, el índice es:

  $$DB = \frac{1}{K} \sum_{i=1}^{K} R_i$$

3.2 ANTECEDENTES

Se llevó a cabo una búsqueda bibliográfica centrada en la implementación de modelos de clasificación basados en Machine Learning para identificar marcadores genéticos asociados al cáncer. Los estudios seleccionados, que se detallan a continuación, proporcionan una base sólida para comprender las investigaciones previas en este campo, ofreciendo perspectivas valiosas para el desarrollo de proyectos enfocados en la detección y diagnóstico del cáncer.

Un estudio pionero desarrolló un método de Machine Learning para diferenciar datos genómicos de ARN provenientes de tejido mamario canceroso y no canceroso. Utilizando el algoritmo Naive Bayesian de Sci-Kit Learn, se analizaron 1270 muestras obtenidas del Cancer Genome Atlas (TCGA) y del proyecto Genotype-Tissue Expression (GTEx). El análisis incluyó un examen exhaustivo del teorema de Bayes, abordando desde la selección de características hasta la optimización y evaluación del modelo. Los resultados fueron sobresalientes, con una precisión, sensibilidad y especificidad del 100 % en la clasificación de células cancerosas y no cancerosas. Este enfoque demostró un gran potencial para la detección temprana del cáncer de mama, lo que podría reducir la mortalidad mediante tratamientos oportunos. Además, se destacó la relevancia del Machine Learning en los ámbitos médico y farmacéutico, sugiriendo que este modelo podría adaptarse para identificar otras enfermedades mediante la secuenciación de ARN [40].

En una línea complementaria, otro trabajo exploró la predicción del cáncer utilizando datos genómicos completos y técnicas de aprendizaje profundo [41]. Los datos, también provenientes del TCGA, fueron preprocesados mediante la clasificación de información individual, la selección de variables relevantes y su transformación en un formato adecuado para el aprendizaje profundo. El modelo se basó en una red neuronal convolucional (CNN) con conexiones residuales, diseñada para evitar problemas de gradientes vanishing y exploding. Esta red empleó capas de convolución para extraer características genómicas, capas de pooling para reducir la dimensionalidad y prevenir el sobreajuste, y una capa completamente conectada con funciones de activación ReLU y softmax para predecir la probabilidad de cáncer. La evaluación, basada en métricas como precisión, sensibilidad, especificidad y F1-score, mostró resultados de 74.1 %, 75.7 %, 73.8 %, 97.6 % y 74.1 %, respectivamente. Comparado con otros modelos de redes neuronales, este enfoque resultó superior, destacando la importancia de analizar patrones genómicos completos en lugar de genes específicos. Los autores sugirieron que este método podría aplicarse a otras enfermedades, consolidando su relevancia para la medicina predictiva.

Por otro lado, un tercer estudio desarrolló un clasificador avanzado de tipos de tumores basado en aprendizaje profundo, utilizando datos genómicos de 39,787 tumores secuenciados [42]. Este modelo, fundamentado en un ensamble de diez redes neuronales entrenadas con características de un panel clínico de genes asociados al cáncer, alcanzó una precisión del 93 % en predicciones de alta confianza para 38 tipos de cáncer, compitiendo con métodos de secuenciación completa del exoma (WES). Además, demostró ser útil para diagnosticar cánceres raros o de origen primario desconocido, integrando información clínica como el sitio de la biopsia y el historial del paciente para optimizar las predicciones. A pesar de emplear un panel de genes específico, el modelo logró una clasificación precisa gracias al gran volumen de datos de entrenamiento, con un 71.9 % de predicciones de alta confianza. La incorporación de un ensamble de Multi-Layer Perceptron (MLP) permitió identificar errores de etiquetado y mejorar la generalización. Diseñado para entornos clínicos, este modelo destaca por su capacidad de integrar datos adicionales mediante un método de priorización adaptable, lo que incrementa su precisión en escenarios reales y sugiere un impacto significativo en la práctica clínica.

La revisión de estos estudios establece una base robusta para proyectos de clasificación basados en Machine Learning orientados a identificar marcadores genéticos del cáncer de mama. Los enfoques analizados, que incluyen algoritmos como Naive Bayesian, redes convolucionales y ensambles de aprendizaje profundo, ofrecen alternativas prometedoras para el análisis de datos genómicos. Sin embargo, un desafío recurrente es la naturaleza de “caja negra” de estos modelos, que dificulta la interpretación de las predicciones, un aspecto crucial para su aplicación clínica. La evaluación de los modelos mediante métricas como precisión, sensibilidad y especificidad resulta fundamental para garantizar su validez y eficacia en el diagnóstico de pacientes con cáncer. Estas métricas no solo validan la robustez de los algoritmos, sino que también aseguran su relevancia en contextos clínicos. En conjunto, estas investigaciones proporcionan perspectivas enriquecedoras que pueden guiar el diseño y los objetivos de futuros proyectos, impulsando avances en la detección y tratamiento del cáncer.

4. MACHINE LEARNING EN EL ANÁLISIS DE VARIANTES GENÉTICAS ASOCIADAS AL CÁNCER DE MAMA

4.1 ENTENDIMIENTO DE DATOS

Los marcadores genómicos incluyendo variantes, fueron obtenidos de dos tipos de fuentes principales. En relación con los datos de pacientes, estos se obtuvieron del estudio exómico de pacientes con cáncer de mama familiar procedentes de Colombia, Perú, Guatemala, México y Argentina, que equivalen aproximadamente a 1000 pacientes con información de variantes relacionadas con cáncer. Los datos restantes se obtuvieron a partir de repositorios públicos de genómica como UCSC Genome Browser, para el reconocimiento de categorías que describen los loci de cada variante encontrada en los pacientes así como variantes ya documentadas en dichos repositorios. A continuación se muestra cómo se integran las distintas fuentes en los datasets de Patogenicidad y de Diagnóstico de Cáncer.

4.1.1 Dataset de Patogenicidad

La patogenicidad representa principalmente si una variante o cambio en el ADN puede conducir a una enfermedad, siendo el sitio donde se encuentra la variación un marcador genético de diagnóstico. Los datos sobre variantes genéticas y sus anotaciones de patogenicidad se obtuvieron utilizando la herramienta Table Browser del UCSC Genome Browser. Para ello, se seleccionaron las siguientes configuraciones:

- *Clado:* Mammal (Mamíferos)
- *Ensamblaje:* GRCh38/hg38
- *Grupo:* Phenotypes, Variants and Literature (Fenotipos, Variantes y Literatura)
- *Pista:* ClinVar Variants
- *Tabla:* ClinVar SNVs

Estas configuraciones permiten acceder a variantes genéticas específicas del ser humano, anotadas en la base de datos ClinVar, que incluyen información sobre su relevancia clínica y patogenicidad. ClinVar es un archivo público que recopila interpretaciones de la importancia clínica de variantes en el genoma humano, facilitando la investigación y comprensión de su relación con diversas enfermedades.

Además, este conjunto de datos fue enriquecido con clasificaciones específicas por gen para las variantes disponibles, identificando si cada variante está asociada a un

gen con características de oncogén, supresor de tumor, doble función o desconocidas [43]. Posteriormente, el dataset se contrastó con el recurso Variation Information dbSNP [44], lo que permitió añadir información detallada sobre los identificadores de variantes encontradas en los pacientes, para la asociación de enfermedades relacionadas, especificidades de cada variante, su clasificación clínica, tipo y consecuencias moleculares.

A continuación, se describen las variables presentes en el conjunto de datos inicial que se procesó para obtener el Dataset de Patogenicidad, que se utilizó en los modelos posteriores:

- **Cancermama clinvarmain**

  Esta base de datos se extrajo de UCSC Genome Browser con los pasos ya descritos anteriormente. UCSC Genome Browser proporciona acceso a datos genómicos y anotaciones relevantes como ClinVar que contiene información sobre la relación entre variaciones genéticas y su relevancia clínica. Esta base incluye registros de variantes asociadas con el cáncer de mama, destacando su clasificación en términos de patogenicidad y condiciones relacionadas [45]. El dataset tiene un tamaño de 338362 registros y 44 variables. Su estructura está compuesta por las siguientes variables:

  - #chrom: Cromosoma donde se encuentra la variante genética. Tipo categórico.
  - chromStart: Posición genómica de inicio de la variante en el cromosoma. Tipo numérico.
  - chromEnd: Posición genómica de finalización de la variante en el cromosoma. Tipo numérico.
  - name: Proporciona detalles sobre cambio de nucleótidos. Tipo categórico.
  - score: Valor numérico relacionado con la puntuación de la variante. Tipo numérico.
  - strand: Cadena de ADN en la que se encuentra la variante genética. Tipo categórico.
  - thickStart: Inicio de la región que se considera significativa para la anotación. Tipo numérico.
  - thickEnd: Fin de la región que se considera significativa para la anotación. Tipo numérico.
  - reserved: Sin información precedente.

- blockCount: Número de bloques en la anotación de la variante. Tipo numérico.
- blockSizes: Tamaño de cada bloque en la anotación de la variante. Tipo numérico.
- chromStarts: Comienzo de cada bloque en relación al inicio del cromosoma. Tipo numérico.
- origName: Proporciona detalles específicos sobre la variante genética, describiendo el gen, donde hay un cambio de nucleótido y la posición. Tipo categórico.
- clinSign: Clasificación clínica de la variante genética. Tipo categórico.
- reviewStatus: Estado de revisión de la variante, indicando cuántos revisores han evaluado la anotación. Tipo categórico.
- type: Tipo específico de la variante genética. Tipo categórico.
- geneId: Identificador del gen afectado. Tipo categórico.
- molConseq: Consecuencia molecular de la variante. Tipo categórico.
- snpId: Identificador de referencia para esta variante genética en la base de datos dbSNP. Tipo categórico.
- nsvId: Sin información precedente.
- rcvAcc: Número de acceso de la variante en la base de datos de condiciones reportadas. Tipo categórico.
- testedInGtr: Indicación de si la variante ha sido probada en una prueba genética. Tipo categórico.
- phenotypeList: Lista de fenotipos asociados con la variante, incluyendo enlaces a bases de datos relevantes. Tipo categórico.
- phenotype: Fenotipo principal asociado con la variante. Tipo categórico.
- origin: Origen de la variante, como genética o somática. Tipo categórico.
- assembly: Versión del ensamblado del genoma humano utilizado para la anotación. Tipo categórico.
- cytogenetic: Localización citogenética de la variante. Tipo categórico.
- jsonHgvsTable: Tabla en formato JSON con las anotaciones HGVS. Tipo categórico.
- _hgvsProt: Sin información precedente.
- numSubmit: Número de presentaciones de la variante. Tipo numérico.
- lastEval: Fecha de la última evaluación de la variante. Tipo Fecha.

- guidelines: Pautas utilizadas para la evaluación de la variante. Tipo categórico.
- otherIds: Otros identificadores relacionados con la variante. Tipo categórico.
- mouseOver: Información mostrada al pasar el cursor sobre la anotación. Tipo categórico.
- vcfDesc: Sin información precedente.
- somImpactDesc: Descripción del impacto somático. Tipo categórico.
- oncogenDesc: Descripción del oncogén. Tipo categórico.
- clinSignCode: Código de clasificación de la variante genética en términos de su relevancia clínica o biológica. Tipo categórico.
- originCode: Código del origen de la variante. Tipo categórico.
- allTypeCode: Código del tipo de variante. Tipo numérico.
- varLen: Longitud de la variante. Tipo numérico.
- starCount: Número de estrellas en la revisión. Tipo numérico.
- variantId: Identificador único de la variante. Tipo numérico.
- dbVarSsvId: Identificador en la base de datos dbVar. Tipo numérico.

- **Variation Information dbSNP**

  La base de datos fue obtenida de Variation Information dbSNP, una herramienta de anotación funcional de variantes de acceso abierto que proporciona información detallada sobre datos de secuenciación de WES. Esta base compila información sobre diversas variaciones genéticas, como polimorfismos de nucleótido único (SNPs), inserciones y deleciones, que pueden influir en la función genética y en la salud del individuo [44]. La estructura de la base de datos contiene 338532 registros y 11 variables descritas a continuación:

  - Disease: Enfermedad o condición médica relacionada. Tipo categórico.
  - Gene: Nombre del gen asociado con la variante genética. Tipo categórico.
  - Chr: Cromosoma donde se encuentra la variante genética. Tipo categórico.
  - Start: Posición genómica de inicio de la variante. Tipo numérico.
  - End: Posición genómica de finalización de la variante. Tipo numérico.
  - ClinInfo: Proporciona detalles específicos sobre la variante genética, describiendo el gen, el cambio de nucleótido y la posición. Tipo categórico.

- ClinClass: Clasificación clínica de la variante genética. Tipo categórico.
- Type: Tipo específico de la variante genética. Tipo categórico.
- SubType: Consecuencia molecular de la variante. Tipo categórico.
- rsID: Identificador de referencia para esta variante genética en la base de datos dbSNP. Tipo categórico.
- Classification: Clasificación más detallada de la variante genética en términos de su relevancia clínica o biológica. Tipo categórico.

- **Clasificación de genes**

  El dataset fue elaborado mediante la integración y curación de datos provenientes de múltiples fuentes especializadas [43]. La información se recopiló de diversas bases de datos, además de estudios científicos revisados por pares y reportes experimentales publicados. Entre las fuentes consultadas se incluyen repositorios de genes y portales de datos genómicos, que aportaron tanto información sobre la expresión y función de los genes como evidencias clínicas y experimentales. Cada fuente fue evaluada según criterios de relevancia, veracidad y actualidad, lo que permitió contrastar y validar los datos, asegurando así una clasificación robusta de los genes como oncogen, supresor tumoral o de doble función. La estructura de la base de datos tiene un tamaño de 4393 registros y dos variables, descritas a continuación:

  - gen: Nombre del gen asociado con una variante genética específica. Tipo categórica.
  - classification: Clasificación de la variante genética. Tipo categórica.

La **Tabla 1** presenta un resumen de los tres conjuntos de datos base utilizados para la creación del Dataset de Patogenicidad. Para cada dataframe se indica su tamaño (número de filas y columnas), la cantidad de variables categóricas (VC) y numéricas (VN), así como el porcentaje total de valores faltantes (% VF). Además, se especifica la variable con mayor número de valores faltantes (Top VF Variable) y el número de variables que exceden el 50 % de valores faltantes (NVF50). Estos indicadores permiten evaluar la calidad y completitud de los datos antes de su procesamiento y análisis.

DF	Tamaño	VC	VN	% VF	Top VF Variable	NVF50
Cancermama clinvarmain	338362 × 44	31	13	14 %	_hgvsProt	7
Variation Information dbSNP	338532 × 11	9	2	0.01 %	subType	0
Clasificación de genes	4393 × 2	2	0	0 %	No aplica	0

DF: Dataframe. **Tamaño:** Filas x Columnas. **VC:** Variables categóricas. **VN:** Variables numéricas. **% VF:** Porcentaje total de valores faltantes. **Top VF Variable:** Variable con mayor número de valores faltantes. **NVF50:** Número de variables con porcentaje de valores faltantes mayor al 50 %.

Tabla 1: Descripción de conjuntos de datos base para la creación de dataset de patogenicidad

4.1.2 Dataset de Variantes a partir de archivos de llamado de variantes (VCF) para pacientes con y sin cáncer de mama familiar

A partir del conjunto de datos previamente enriquecido con anotaciones funcionales y clasificaciones génicas, se integró la base de datos que contenía de pacientes descritos en el Anexo 1. Esta incluía un identificador único por individuo, el estado clínico respecto a la presencia o ausencia de cáncer, y las variantes genéticas asociadas a cada caso. Utilizando los identificadores de variantes compartidos entre ambos conjuntos, se generó un nuevo recurso mediante la plataforma FAVOR (Functional Annotation of Variants Online Resource), que permitió ampliar la información disponible con anotaciones moleculares y clínicas adicionales.

A continuación, se describen las variables presentes en el conjunto de datos que se procesó para obtener el Dataset de Diagnóstico de Cáncer, que se utilizó en los modelos posteriores:

- **FAVOR**

  La base de datos al igual que Variation Information fue obtenida de Functional Annotation of Variants - Online Resource (FAVOR) [46]. No obstante, se diferencia de esta última porque incluye variables que no están exclusivamente asociadas al cáncer de mama. Está conformada por 165 variables y 428683 registros de variantes genéticas. Ofrece información detallada sobre funcionalidad y efectos genéticos, predicciones de conservación, regiones reguladoras, frecuencia y distribución de variantes, asociaciones clínicas, características geográficas y poblacionales, predicción de patogenicidad y funcionalidad, densidad de nucleótidos y recombinación. A continuación, se presenta una descripción general agrupada de las variables:

  - Identificación de variantes: Incluye información básica sobre la identificación de la variante genética, como su posición en el genoma, el cromosoma al que pertenece y su identificador único.

- Información sobre la variante: Proporciona datos adicionales sobre las características de la variante, como las frecuencias alélicas y su estado en bases de datos de calidad o filtros.
- Anotación de Genecode: Describe la localización y categoría de la variante dentro del contexto genético según la anotación de Genecode, incluyendo detalles sobre regiones exónicas y no codificantes.
- Anotación UCSC: Contiene información relacionada con la localización de la variante según la base de datos UCSC Genome Browser, con un enfoque en regiones exónicas.
- Función proteica: Incluye predicciones computacionales que evalúan el impacto funcional de la variante en proteínas y estructuras genéticas.
- Referencias de anotación de genes: Proporciona información sobre las variantes en el contexto de anotaciones genealógicas, especialmente según la base de datos RefSeq.
- Elementos regulatorios: Identifica elementos genómicos, como potenciadores y promotores, que pueden regular la expresión de genes en las proximidades de la variante.
- Clasificación clínica: Relaciona la variante con su relevancia clínica, describiendo posibles condiciones de salud asociadas y el nivel de validación de estos datos.
- Orígenes y bases de datos adicionales: Incluye detalles sobre las bases de datos de enfermedades y la información genética del gen en cuestión.
- Puntuación integrada: Proporciona información sobre la conservación evolutiva de la región donde se encuentra la variante, así como su densidad genética y proximidad a elementos genómicos clave.
- Predicciones de SIFT: Ofrece predicciones específicas de SIFT sobre el efecto potencial de la variante en proteínas, categorizando el posible daño.
- Conservación: Evalúa la conservación de la variante en diferentes niveles evolutivos, como primates, mamíferos y vertebrados.
- Estadísticas y análisis de ChIP-seq: Incluye datos relacionados con la actividad de regiones genómicas basadas en estudios de epigenética y marcas regulatorias detectadas por ChIP-seq.
- Epigenética: Contiene datos sobre la regulación genética y epigenética, especialmente relacionados con marcas de histonas y la accesibilidad cromatínica.
- Frecuencia de variantes: Describe la frecuencia con la que la variante aparece en diferentes escalas genómicas, desde 100 hasta 10,000 pares de bases.

- Factores de transcripción: Relaciona la variante con factores de transcripción y clústeres de actividad reguladora en el genoma.
- Análisis de puntuaciones de CADD: Proporciona puntuaciones predictivas que miden la probabilidad de que una variante tenga un impacto funcional significativo.
- Análisis de conservación y diversidad nucleotídica de APC: Incluye datos sobre la conservación y funcionalidad del gen Apc, junto con la diversidad genética de la región afectada.
- Distribución de frecuencias por poblaciones: Describe la frecuencia de la variante en diferentes poblaciones globales y en subgrupos específicos, como hombres y mujeres.
- Mapeabilidad: Proporciona mapas genómicos de variabilidad y accesibilidad en diferentes regiones del genoma, generados por diferentes enfoques de análisis.
- Diversidad local: Describe la tasa de recombinación genética y la diversidad nucleotídica en la región donde se encuentra la variante.
- Otros indicadores adicionales: Incluye análisis funcionales avanzados que evalúan el impacto de las variantes en la funcionalidad genética y la regulación.

- **Variants**
  La base de datos de variantes genéticas asociadas al cáncer de mama se construyó a partir de la cohorte de 1098 mujeres de América Latina que cumplían los criterios de la guía NCCN v2.2018 para el síndrome HBOC como se describe en el Anexo 1. Una vez parseada la base de datos esta adquiere un tamaño de 359804 variantes y tres variables. A continuación, se presenta una descripción general de las variables agrupadas por categoría:

  - Columna 1: Representa variantes genéticas en formato HGVS, indicando el cromosoma, la posición genómica y el cambio de nucleótido . Tipo categórico.
  - Columna 2: Identificador de referencia para esta variante genética en la base de datos dbSNP. Tipo categórico.
  - Columna 3: Códigos de pacientes asociados a la presencia de la variante en el genoma humano. Tipo categórico.

La **Tabla 2** muestra las características de los dos conjuntos de datos base empleados para la generación del Dataset de Variantes a partir de archivos VCF de pacientes con y sin cáncer de mama familiar. Para cada dataframe se indica su tamaño (número de filas y columnas), la cantidad de variables categóricas (VC) y numéricas (VN), así como el porcentaje total de valores faltantes (% VF). Asimismo, se detalla la variable que presenta el mayor número de valores faltantes (Top VF Variable) y el número de variables cuyo porcentaje de datos ausentes supera el 50 % (NVF50).

DF	Tamaño	VC	VN	% VF	Top VF Variable	NVF50
FAVOR	428683 × 165	14	151	20.7 %	Clndisdbincl	31
Variants Parseado	359804 × 3	3	0	0 %	No aplica	0

DF: Dataframe. **Tamaño:** Filas x Columnas. **VC:** Variables categóricas. **VN:** Variables numéricas. **% VF:** Porcentaje total de valores faltantes. **Top VF Variable:** Variable con mayor número de valores faltantes. **NVF50:** Número de variables con porcentaje de valores faltantes mayor al 50 %.

Tabla 2: Descripción de conjuntos de datos base para la creación de dataset de variantes a partir de archivos de llamado de variantes (VCF) para pacientes con y sin cáncer de mama familiar

4.2 PREPARACIÓN DE DATOS

4.2.1 Dataset de Patogenicidad

Se cargaron los conjuntos de datos Cancermama clinvarmain y Variation Information dbSNP. A continuación, se aplicó la función cleaning() para depurarlos, mediante una serie de transformaciones orientadas a garantizar la calidad y consistencia de la información. Inicialmente, los datos se transformaron del formato pandas a Polars para optimizar el rendimiento computacional. Se eliminaron las columnas con más del 50 % de valores faltantes, aquellas que presentaban un único valor constante.

Para la columna origName, se extrajo una parte de su contenido mediante expresiones regulares y se convirtió a minúsculas, almacenándose en una nueva variable llamada ClinInfo. Asimismo, la columna reviewStatus se limpio respecto a sus fragmentos en formato HTML y se normalizó su contenido mediante un diccionario de mapeo. La columna _jsonHgvsTable fue simplificada usando la función simplify_json, eliminando caracteres innecesarios.

Para garantizar la homogeneidad en las variables categóricas, se reagruparon categorías poco frecuentes bajo etiquetas genéricas. Por ejemplo, las categorías de la variable simplified_hgvs con menos de dos registros fueron etiquetadas como other, y un procedimiento similar se aplicó a la variable origin para aquellas categorías con

tres o menos ocurrencias. Todas las variables de texto se convirtieron a minúsculas y se recodificaron como variables categóricas. Una vez completadas estas transformaciones en Polars, los datos se convirtieron nuevamente a pandas.

A continuación, se construyó una clave compuesta denominada real_id para permitir la fusión de los dos conjuntos, eliminando previamente columnas duplicadas para evitar redundancias. Tras esta integración, se generó una nueva variable binaria llamada bin_class, que clasificó las variantes genéticas como patogénicas (valor 1) si pertenecían a la categoría PG, y como no patogénicas (valor 0) en caso contrario.

Como parte de la depuración, se eliminaron columnas irrelevantes o redundantes, como identificadores repetidos, enlaces web sin utilidad analítica, y campos sin valor informativo. Luego, se imputaron valores faltantes en distintas columnas utilizando valores específicos: clinSign, ClinClass y molConseq fueron completadas con el valor unknown; phenotypeList con not provided; y rcvAcc con el identificador genérico rcv000000000, garantizando la integridad del conjunto.

Posteriormente, se integró información proveniente de la base Clasificación de genes, lo que permitió clasificar los genes según su rol funcional en cáncer: como oncogenes, supresores de tumor, de función desconocida o de doble función. Para asegurar la coherencia, se normalizaron los nombres de los genes a minúsculas y se eliminaron variables innecesarias tras la fusión. En total, se clasificaron 273,593 genes como supresores de tumor, 27,616 como oncogenes, 22,058 con función desconocida y 3,745 con doble función.

Para la selección de las variables mas relevantes, se implementaron varias técnicas con el propósito de reducir la dimensionalidad, minimizar redundancias y preservar la calidad del análisis. A continuación, se describen los tres enfoques utilizados:

- Correlación de Theil - Spearman: Se utilizó una muestra aleatoria equivalente al 10 % del conjunto de datos para calcular las asociaciones entre variables. Para las variables categóricas se aplicó la correlación de Theil, mientras que para las numéricas se empleó el coeficiente de Spearman. En la matriz de correlacion de la **Figura 1**, se visualizan las variables en los ejes X,Y, aquellas variables que tienen una alta correlación muestran colores desde naranja a rojo (0.75 a 1.0 de correlación) y las variables que tienen poca o nada de correlación entre si presentan colores mas verdes a azules (0.3 a 0.0) respectivamente, esto permitió identificar pares redundantes de variables, donde se seleccionó vcfDesc como principal candidata para ser eliminada.

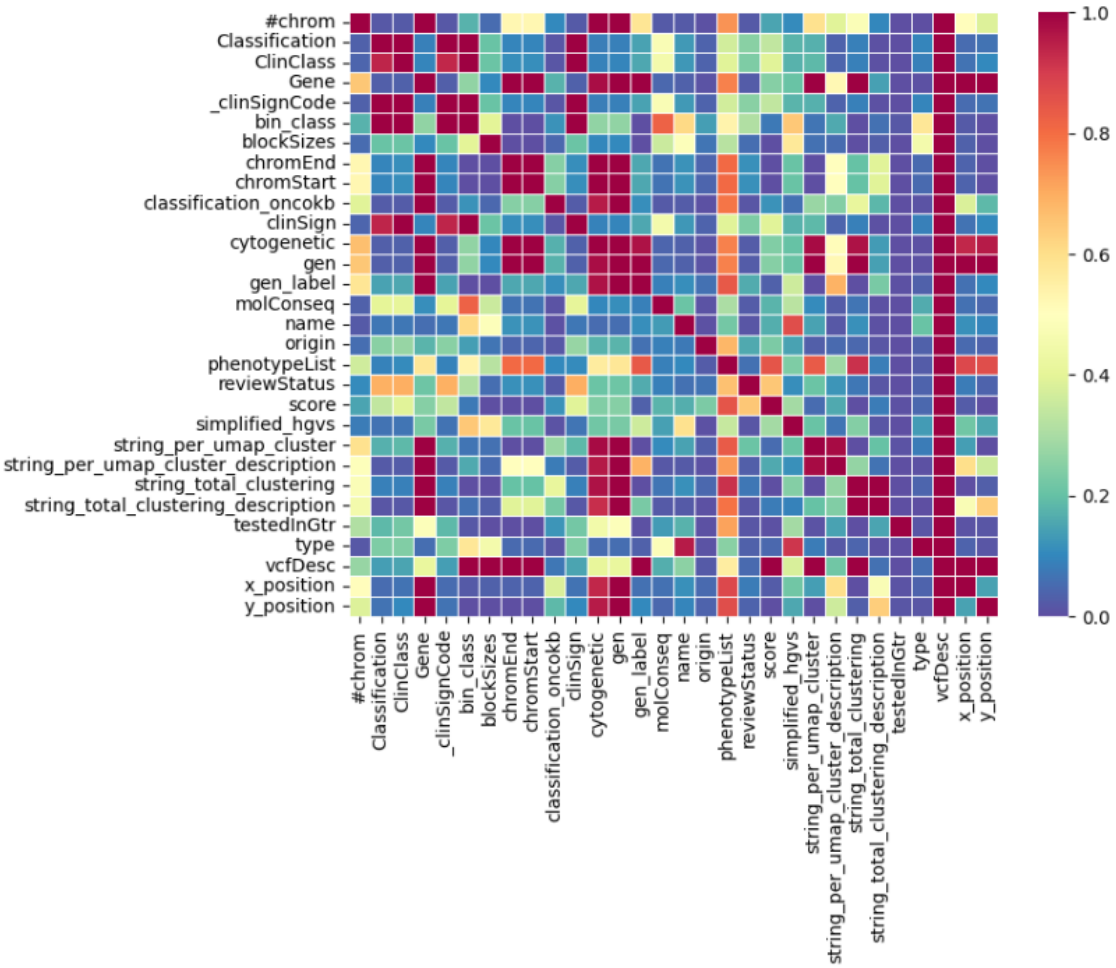


Figura 1: Matriz de correlación Theil - Spearman para dataset de patogenicidad

- Correlación de Cramer - Spearman: Se aplicó un enfoque similar utilizando el coeficiente de Cramer para las variables categóricas y el coeficiente de Spearman para las variables numéricas. A través de este análisis se generó la matriz de correlación presentada en la **Figura 2**, en donde se identificó que la variable gen era redundante, lo que llevó a recomendar su eliminación.

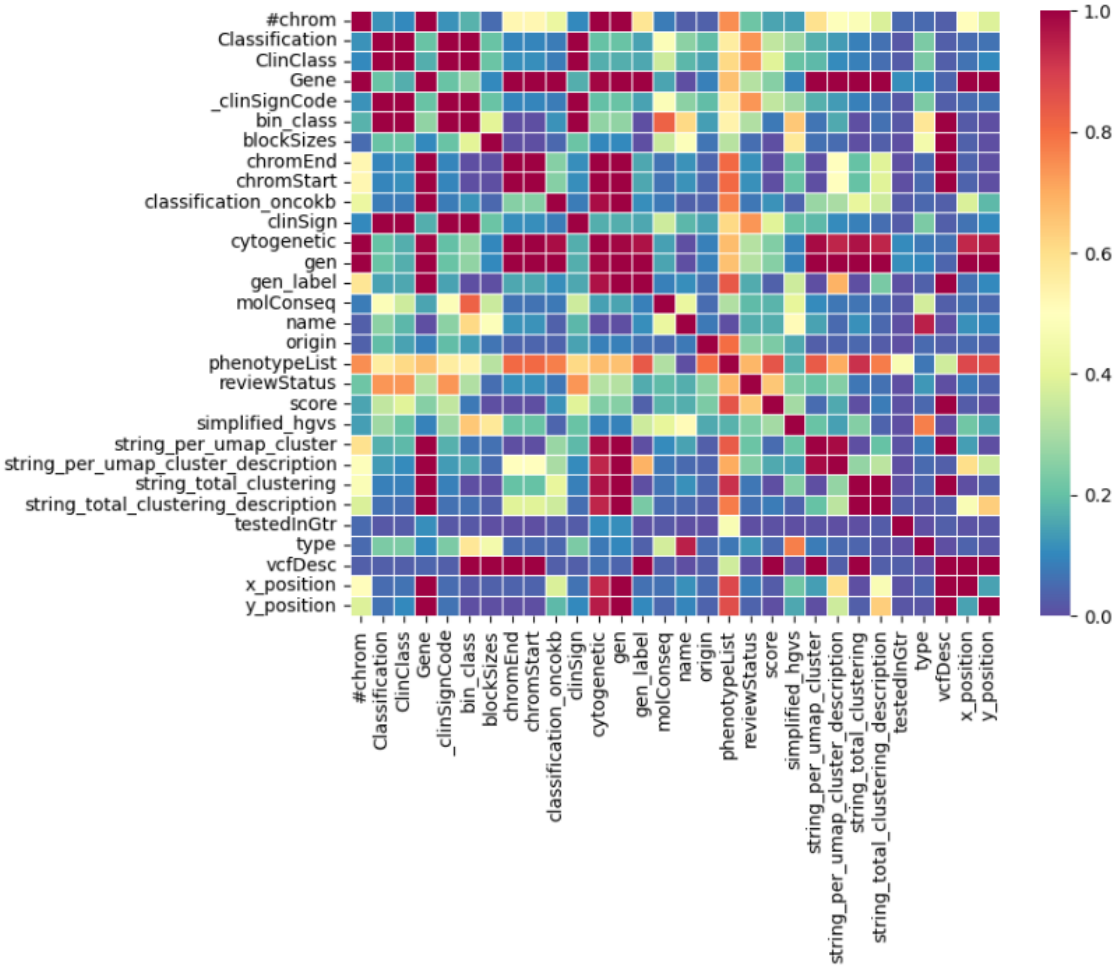


Figura 2: Matriz de correlación Cramer - Spearman para dataset de patogenicidad

- Información Mutua: Para capturar dependencias no lineales entre variables, se utilizó la matriz de información mutua. Las variables categóricas se convirtieron a tipo categórico y se codificaron numéricamente, permitiendo el uso de la función mutual_info_classif de scikit-learn. La matriz resultante se reordenó para facilitar su interpretación a través de un heatmap plasmado en la **Figura 3**. A partir del análisis, se seleccionaron chromStart y chromEnd como las variables redundantes para su eliminación.

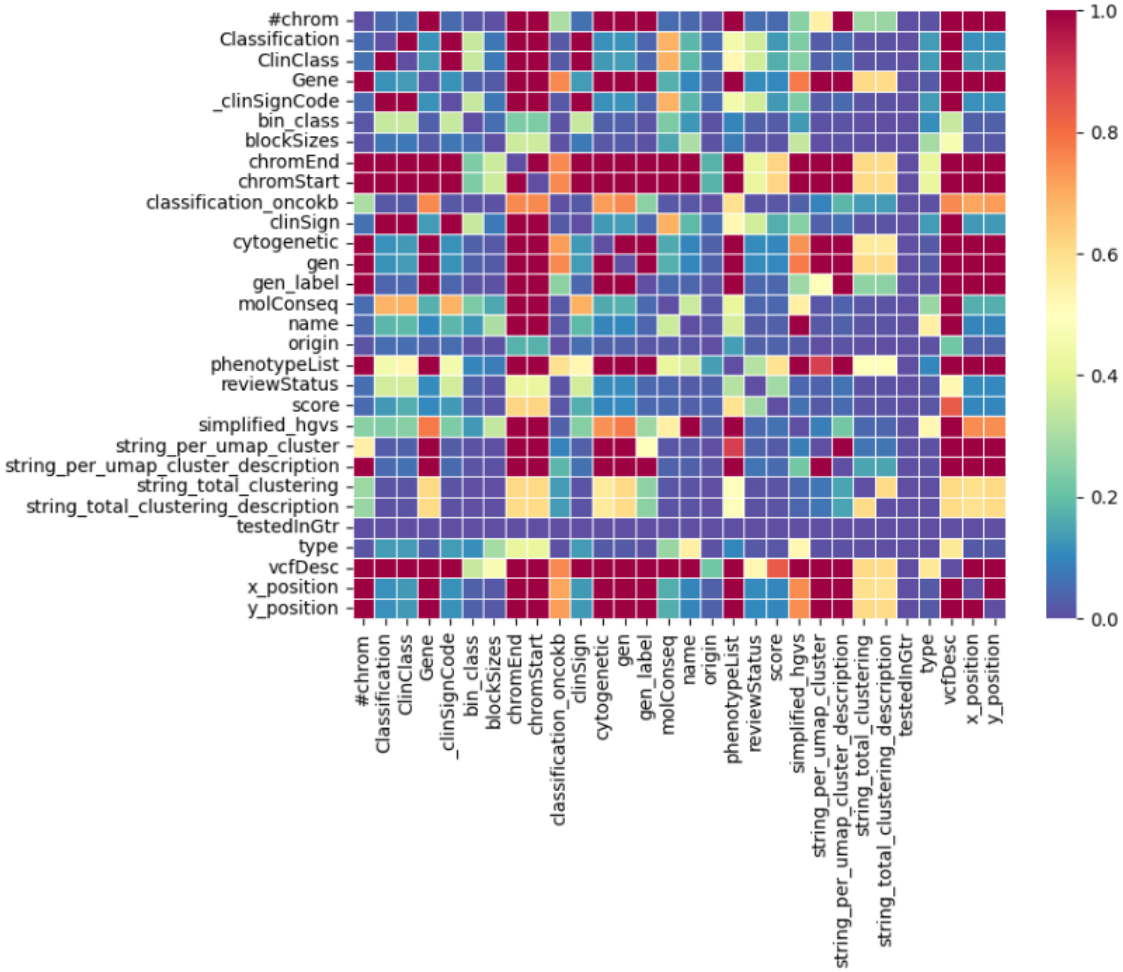


Figura 3: Matriz de información mutua para dataset de patogenicidad

De acuerdo con los enfoques mencionados, se eliminaron las variables recomendadas para optimizar el análisis. Adicionalmente, se consideraron otras variables con asociaciones fuertes, cuya información podía ser representada por atributos más relevantes, evitando redundancias innecesarias. Las columnas eliminadas fueron: Classification, clinSignCode, clinSign, ClinClass, testedInGtr, string per umap cluster y string total clustering, contribuyendo así a mejorar la calidad del modelo al reducir la complejidad del conjunto de datos. El dataset final contiene 328116 registros y 19 variables, de las cuales 13 son categóricas y 6 son numéricas.

4.2.2 Dataset de Variantes a partir de archivos de llamado de variantes (VCF) para pacientes con y sin cáncer de mama familiar

El conjunto de datos de diagnóstico de cáncer fue construido mediante la integración de información clínica de pacientes con anotaciones genéticas derivadas del dataset de patogenicidad previamente descrito. Para cada gen presente en el conjunto objetivo (obtenido a partir de FAVOR Genohub), se realizó un mapeo de su nombre contra el primer conjunto de datos para incorporar la fracción de variantes patogénicas asociadas y su clasificación funcional (oncogén, supresor tumoral, doble función o desconocido).

Posteriormente, se procesaron los datos clínicos para establecer una relación uno-a-muchos entre pacientes y variantes mediante expansión longitudinal, representando cada combinación variante-paciente en una fila independiente. El estado clínico de cada individuo, que luego será la variable objetivo en este dataset, fue codificado como variable binaria (1: cáncer, 0: no cáncer), mientras que las variantes fueron identificadas por su correspondiente rsID, consistente con los otros conjuntos.

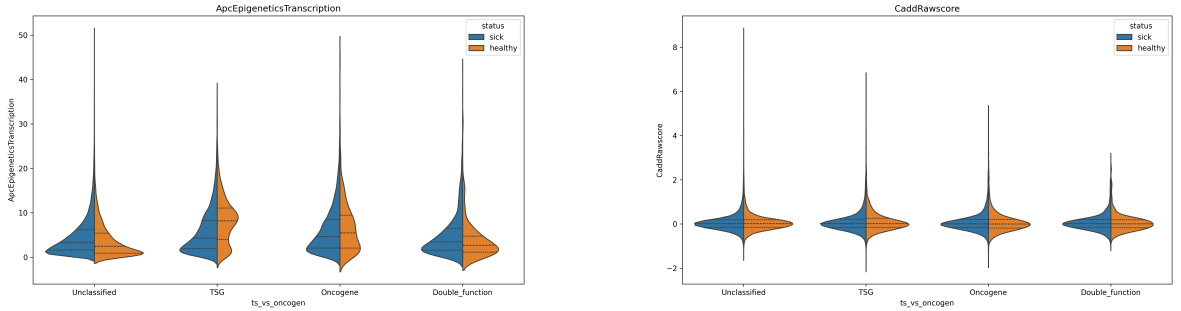
Se calcularon métricas de agregación como el número de variantes por paciente, número de pacientes por variante, y la fracción de pacientes con diagnóstico positivo o negativo por gen. A partir de la anotación estructural de cada variante (por ejemplo, 1-21251006-CA-C), se extrajeron k-mers representativos tanto para la base original como para la mutada, capturando así patrones potencialmente relevantes de alteración genómica. En cuanto al tratamiento de variables, las variables numéricas fueron escaladas usando Scikit-Learn MinMaxScaler, mientras que las categóricas se codificaron empleando label encoding para variables binarias y binary encoding para aquellas con más de dos categorías cuando XGBoost no podía procesarlas nativamente.

El proceso de selección de variables se realizó en dos etapas. Primero, con apoyo de expertos del dominio, se eliminaron variables no informativas para la tarea de clasificación, particularmente aquellas relacionadas con la distribución de frecuencias

poblacionales descritas en la sección anterior. Posteriormente, se realizó un análisis comparativo de distribuciones evaluando cada variable respecto a los cuatro tipos de genes (sin clasificación “Unclassified”, supresor tumoral “TSG”, oncogen “Oncogen” y función dual “Double_function”) y al estado clínico del paciente (sano vs. enfermo).

Como se ilustra en la **Figura 4**, los gráficos de violín revelan patrones discriminatorios distintivos entre variables. Por ejemplo, ApcEpigeneticsTranscription mostró valores consistentemente mayores en muestras sanas, especialmente en genes supresores tumorales y oncogenes, donde las medianas e intervalos intercuartílicos de muestras sanas se situaron por encima de los casos enfermos. En contraste, en la categoría sin clasificación, ambos grupos se agruparon cerca de cero con solo un ligero incremento en casos enfermos. CaddRawScore, por su parte, se centró alrededor de cero para todas las categorías, pero las muestras enfermas exhibieron colas moderadamente más gruesas (valores extremos positivos y negativos) en la mayoría de categorías genéticas.

Estos resultados sugieren que ApcEpigeneticsTranscription proporciona separación clara y específica por categorías entre controles sanos y casos enfermos, mientras que los puntajes CADD muestran diferencias sutiles principalmente en varianza y valores atípicos, indicando que la transcripción epigenética constituye un discriminador más robusto y consistente, especialmente en genes supresores tumorales y oncogenes.



(a) ApcEpigeneticsTranscription

(b) CaddRawScore

Figura 4: Análisis comparativo de distribuciones respecto a tipo de gen y estado clínico del paciente para dos variables de ejemplo.

De este modo, se mantuvieron variables como BravoAc, Linsight, FathmmXf, y múltiples marcadores epigenéticos (Encodeh3k4me1Sum, Encodeh3k9ac3Sum, entre otros) que exhibieron patrones discriminatorios consistentes. Por el contrario, se eliminaron variables como patient_count, Mamphcons, Mamphylop y varios marcadores ENCODE que no mostraron diferencias apreciables entre grupos. Este método

de selección permitió una evaluación integral y simultánea de la capacidad discriminatoria de cada variable considerando tanto la variabilidad biológica (tipo de gen) como el estado clínico lo que permitió obtener un dataset final que contiene 349490 registros y 195 variables, de las cuales 4 son categóricas y 191 numéricas.

4.3 ANÁLISIS EXPLORATORIO

4.3.1 Dataset de Patogenicidad

La **Tabla 3** presenta la distribución de las variantes genéticas presentes en el dataset, clasificadas según su significancia clínica. Se evidencia que el 43 % corresponde a variantes de significancia desconocida (VUS), seguidas por variantes probablemente benignas (31 %), patogénicas (11 %), con conflicto de evidencia (8 %), probablemente patogénicas (4 %) y benignas (3 %). Este hallazgo resalta la gran incertidumbre que aún persiste en la interpretación clínica de las variantes genéticas, ya que una proporción considerable no puede clasificarse con certeza como patogénica o benigna, lo cual representa un desafío para la precisión diagnóstica.

Categoría	Frecuencia	Porcentaje (%)
Significancia desconocida (VUS)	146,312	43.0
Probablemente benigno (LB)	103,710	31.0
Patogénico (PG)	36,155	11.0
Conflicto de evidencia (CF)	25,602	8.0
Probablemente patogénico (LP)	12,911	4.0
Benigno (BN)	9,486	3.0
Otro (OT)	4,594	1.0
Reclasificado (RF)	29	0.0

Tabla 3: Distribución de variantes genéticas según su categoría de clasificación clínica

Desde la perspectiva genómica o de distribución de las variantes en el genoma humano, bajo los principios biológicos de organización (cromosoma, y loci específicos), como se observa en la **Figura 5** las variantes genéticas se distribuyeron a lo largo de todos los cromosomas sin una concentración exclusiva en una sola región, aunque se observaron diferencias notables en la carga mutacional entre ellos. El cromosoma 17 se destacó con la mayor cantidad de variantes clasificadas como de significancia desconocida (VUS), alcanzando aproximadamente 17.500 casos, además de albergar 8.500 variantes patogénicas (PG) y 1.250 benignas (BN), lo que indica una alta complejidad clínica y biológica en esta región. En contraste, el cromosoma 6 presentó la menor cantidad de variantes, reflejando una carga mutacional más baja.

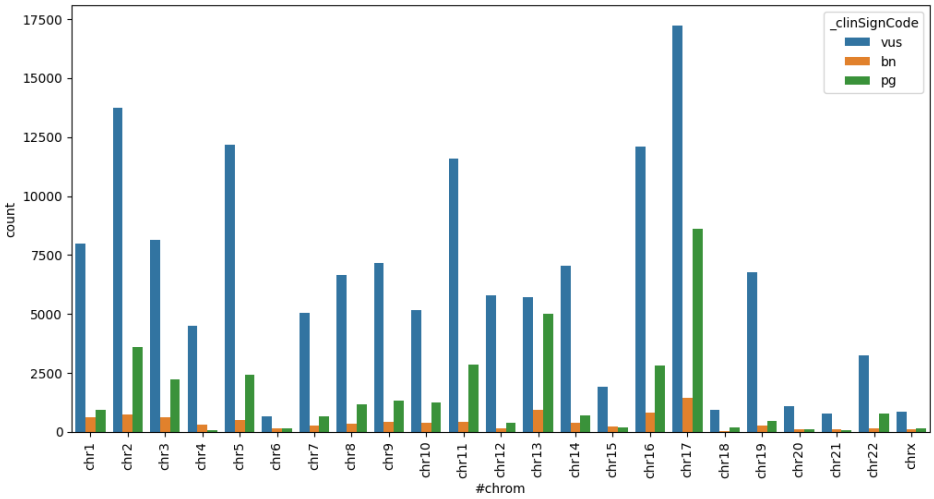


Figura 5: Distribución de variantes genéticas por cromosoma y clasificación clínica

Por otra parte, las variantes se clasifican con base en el efecto en la función de las proteínas que se expresan a partir de los genes, es decir como un cambio afectano solo el gen sino la estructura de las proteínas; en donde las variantes más severas son aquellas que cambian completamente la proteína (nonsense) y las menos severas, aquellas que cambian sutilmente la estructura (synonyms). Como se observa en la **Figura 6**, la gran mayoría de las variantes de significancia desconocida (VUS) corresponden a variantes missense, las cuales modifican un solo aminoácido en la secuencia de proteínas. Este tipo de variantes suele tener una consecuencia funcional difícil de predecir, lo que dificulta su clasificación como variantes benignas (BN) o patogénicas (PG). La distribución presentada en la figura resalta la prevalencia de estas variantes en comparación con otros tipos, lo que refuerza la dificultad de su interpretación clínica.

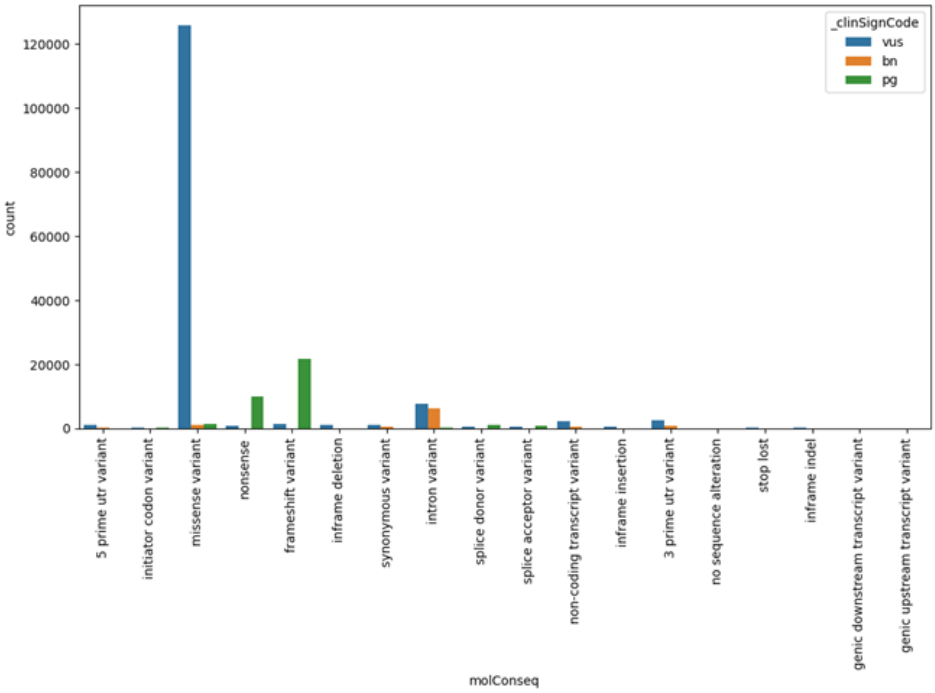


Figura 6: Distribución de variantes genéticas según su clasificación clínica y consecuencia molecular

Por otro lado, las variantes patogénicas (PG), que están comúnmente asociadas con enfermedades genéticas, se concentraron principalmente en categorías de alto impacto, como las variantes nonsense y frameshift. Estas variantes están asociadas con pérdida de función, generando proteínas truncadas o no funcionales, lo que explica su clasificación más clara como patogénicas. La gráfica de la **Figura 6** muestra una clara diferenciación entre las variantes de alto impacto y las de menor severidad, destacando la relevancia de las primeras en la determinación de la patogenicidad.

En términos de tipo de variante, en la **Figura 7** se presenta la frecuencia de distintos tipos de variantes genéticas categorizadas por su significado clínico: variantes de significado incierto (VUS), benignas (BN) y patogénicas (PG). El gráfico revela que las variaciones de un solo nucleótido (SNVs) son, por amplio margen, el tipo de variante más frecuente, destacándose especialmente en la categoría de VUS con más de 140,000 registros. En contraste, las deleciones explicaron gran parte de las PG detectadas. La categoría intron variant mostró la mayor proporción conjunta de BN y VUS, sugiriendo que algunas VUS intrónicas podrían en realidad corresponder a cambios benignos y merecen estudio funcional adicional.

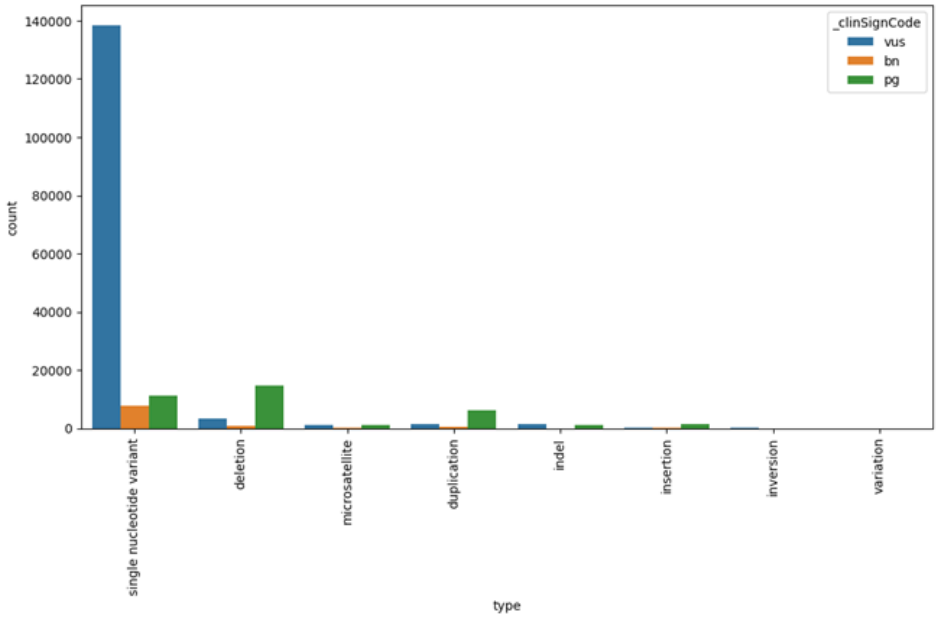


Figura 7: Distribución de variantes genéticas según su clasificación clínica y tipo de mutación

A nivel de genes y tipos de genes como marcadores genómicos, la **Figura 8** evidencia dos perfiles contrastantes. Por un lado, genes supresores de tumores como APC y ATM concentraron una alta proporción de variantes VUS, lo cual pone de manifiesto la complejidad inherente a la interpretación de variantes en genes altamente conservados y funcionalmente esenciales. Por otro lado, genes con un rol ampliamente documentado en la prevención de la oncogénesis, como BRCA1, BRCA2 y NF1, presentaron una mayor carga de variantes clasificadas como patogénicas, lo que resulta coherente con su implicación directa en procesos de reparación del ADN y control del ciclo celular. De manera interesante, tanto BRCA1 como BRCA2 también mostraron variantes benignas, lo cual resalta la importancia de considerar el contexto clínico y familiar específico al momento de interpretar su impacto, especialmente en entornos de asesoramiento genético.

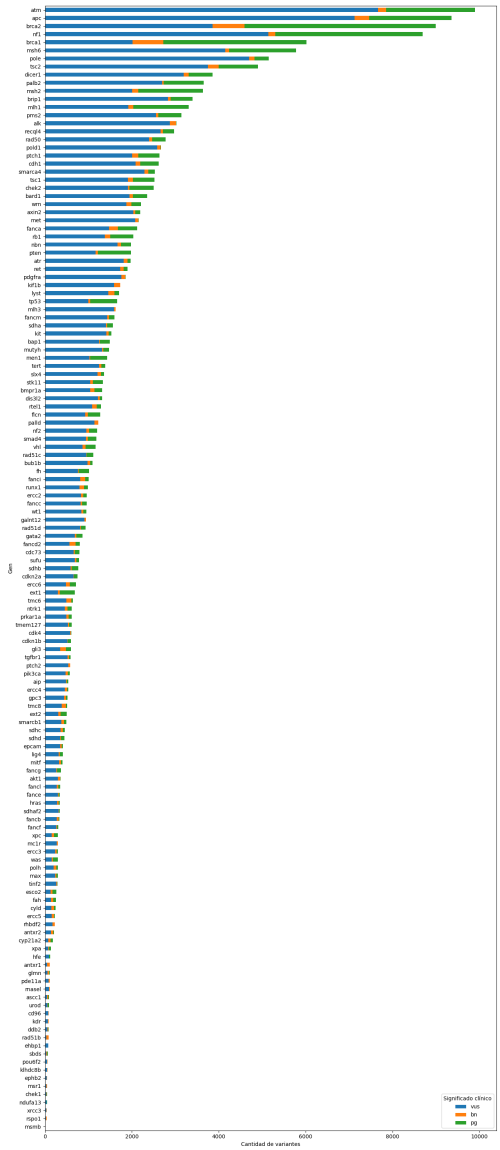


Figura 8: Distribución de genes según su clasificación clínica

4.3.2 Dataset de Variantes a partir de archivos de llamado de variantes (VCF) para pacientes con y sin cáncer de mama familiar

El conjunto de datos de diagnóstico de cáncer está compuesto por un total de 230 variables, de las cuales 12 son categóricas y 218 son numéricas. La estructura del dataset incluye información sobre combinaciones de bases genéticas, fracción de genes afectados por grupo (pacientes sanos y enfermos), así como variables asociadas a características clínicas de los pacientes. Este conjunto contiene información sobre variantes genéticas correspondientes a aproximadamente 1.000 pacientes, lo que representa un total cercano a 600.000 variantes. En la **Figura 9** se presenta el top 100 de pacientes con mayor número de variantes identificadas, donde se observa un rango que va desde aproximadamente 5.000 hasta 800 variantes por paciente. A partir de ese punto, la cantidad de variantes por individuo disminuye gradualmente, oscilando entre 800 y 1.

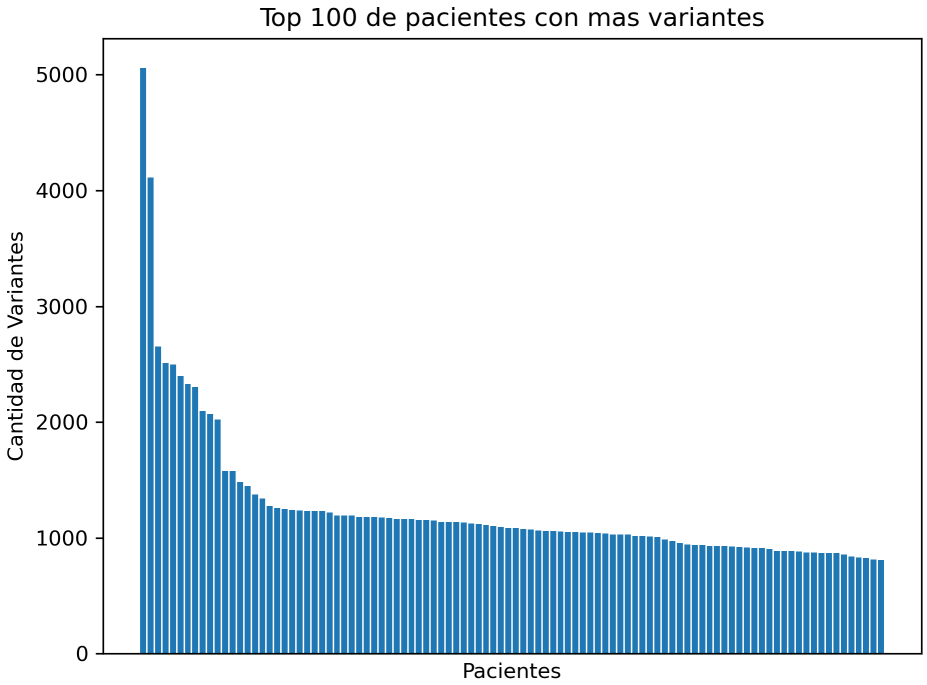


Figura 9: Top 100 de pacientes con más variantes genéticas.

Por otro lado, se estableció una relación entre la variable gen y la cantidad de variantes observadas, con el objetivo de identificar aquellos genes con mayor recurrencia en el conjunto de datos. En la **Figura 10** se presenta el top 20 de los genes asociados al mayor número de variantes. Se destaca el gen Rf00019, el cual se encuentra vinculado a más de 3.500 variantes. Le siguen los genes RBFOX1 y CSMD1, con

aproximadamente 1.000 variantes asociadas cada uno. A partir de este punto, la cantidad de variantes por gen disminuye progresivamente, oscilando entre 500 y 1. Un hallazgo relevante es la ausencia de los genes BRCA1 y BRCA2 en este ranking, a pesar de su reconocida implicación en el desarrollo del cáncer de mama. Esta ausencia podría deberse a particularidades del conjunto de datos, en especial teniendo en cuenta que este dataset se encuentra centrado en variantes para pacientes en Latinoamérica.

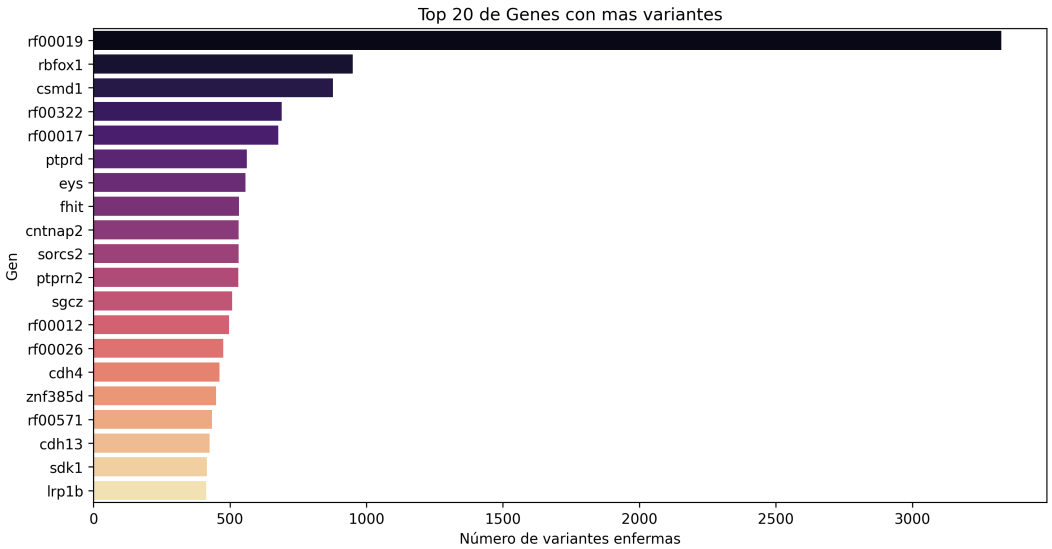


Figura 10: Top 20 de genes vinculados con más variantes genéticas.

4.3.2.1 Distribución de categorías de variable objetivo status

La variable status indica la condición de salud de cada paciente, clasificándolos como sanos o enfermos. En la **Figura 11** correspondiente se presenta el análisis de frecuencia de esta variable, el cual revela un desbalance de clases significativo: aproximadamente el 90 % de los registros corresponde a pacientes enfermos, mientras que menos del 10 % restante corresponde a pacientes sanos. Esta distribución desigual debe ser considerada en las etapas posteriores del análisis, ya que puede afectar el desempeño tanto de modelos de clasificación supervisada, al sesgar las métricas de evaluación, como de modelos no supervisados, al influir en la conformación de grupos dominados por la clase mayoritaria. Por lo tanto, será fundamental aplicar estrategias que permitan mitigar este efecto, tales como técnicas de sobremuestreo (oversampling) o submuestreo (undersampling).

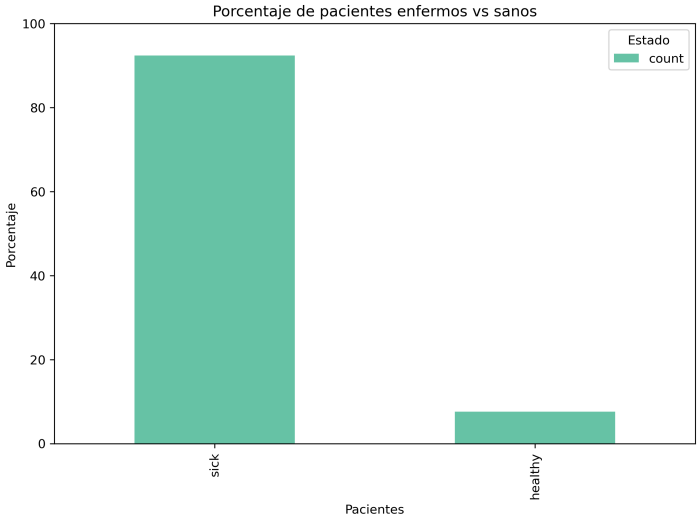


Figura 11: Desbalance de clases en el dataset de variantes a partir de archivos de llamado de variantes (VCF) para pacientes con y sin cáncer de mama familiar

4.3.2.2 Análisis comparativo de las frecuencias de nucleótidos por status

El análisis comparativo de las frecuencias de nucleótidos entre pacientes con cáncer de mama y pacientes sanos revela diferencias significativas en las distribuciones de k-mer, particularmente en k=3. La **Figura 12** muestra diferentes formas de secuencias caracterizadas por patrones de enriquecimiento diferencial de varios trinucleótidos clave, especialmente ACG, GCG y TCG, que presentan alteraciones de frecuencia pronunciadas en las muestras patológicas. Estos trinucleótidos se manifiestan como rasgos prominentes demarcados por líneas de intersección en ambos paneles, aunque sus relaciones contextuales con secuencias adyacentes exhiben diferencias notables entre condiciones.

Específicamente, los gradientes de intensidad adyacentes a dichos motivos exhiben una mayor concentración (coloración roja más intensa) en las muestras de cáncer, particularmente a lo largo de bloques cromosómicos que contienen secuencias GCG y TCG. La agrupación evidente en ambas matrices revela que estas alteraciones de trinucleótidos no son fenómenos aislados, sino que se manifiestan dentro de vecindarios de secuencias más amplias que exhiben cambios de frecuencia coordinados.

Además, los patrones de distribución asimétrica observables en la matriz triangular inferior sugieren posibles procesos mutacionales específicos de la cadena o sesgos transcripcionales que pueden contribuir a la carcinogénesis. Estos hallazgos concuerdan con la literatura genómica sobre el cáncer, que demuestra que los contextos

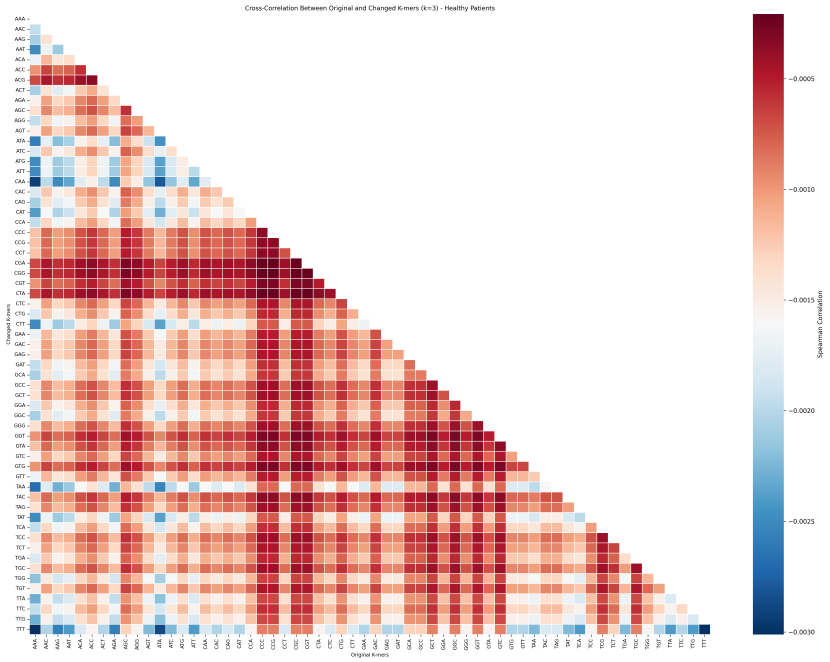
de trinucleótidos que rodean los puntos calientes mutacionales a menudo albergan valor diagnóstico y pronóstico, reflejando potencialmente el daño y la reparación subyacentes del ADN.

Buscando profundizar aún más en las características de los cambios mutacionales del presente dataset se generó un mapa de calor que permita correlacionar los k-mers y los genes de modo que se pueda mapear cambios más comunes por gen. El mapa de calor obtenido para los k-mers generados presentado en la **Figura 13** revela complejos patrones de agrupación de variantes, destacando distintas distribuciones asociadas al estado de "sano" y "enfermo" de los pacientes. La agrupación jerárquica (dendrogramas en la parte superior e izquierda) delinea varios módulos principales de expresión génica, con notables patrones de bandas transcripcionales verticales que sugieren una regulación coordinada de programas genéticos específicos.

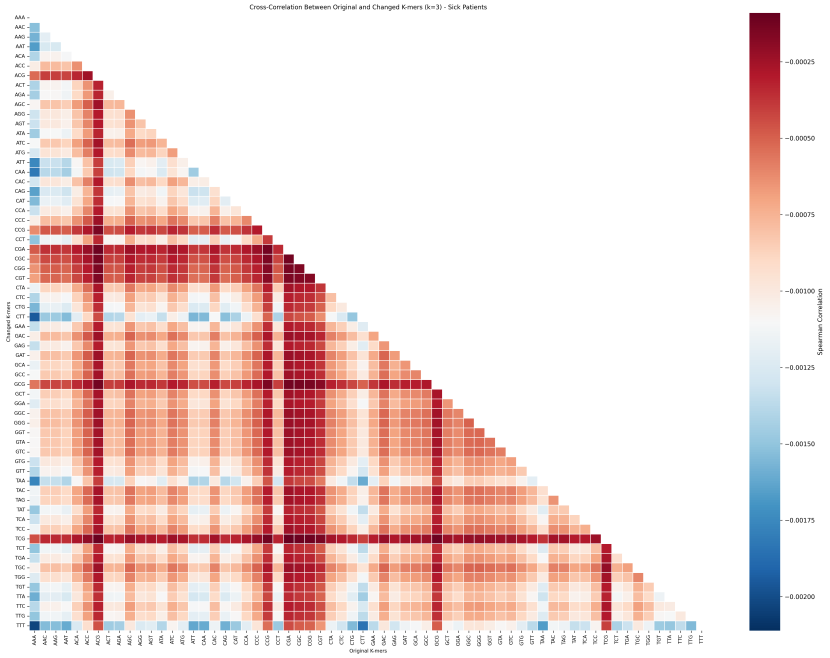
El gradiente de intensidad (de blanco a azul oscuro) indica niveles de expresión diferenciales, con destacados grupos de regulación al alza observados especialmente en genes como KAZN, ERBB4 y CEMIP1, que muestran sólidos patrones de corelación en múltiples tipos de muestras. Surgen varios vecindarios de expresión discretos, concentrados notablemente en torno a los reguladores del ciclo celular (MCM2, MCM6) y los moduladores de la matriz extracelular (TNFAIP6, CEMIP1). La clara organización jerárquica de los genes observada en el dendrograma de la izquierda se divide en aproximadamente 5-7 grupos principales, con genes que muestran una notable corelación.

Este paisaje de asociaciones revela además posibles interacciones novedosas entre genes no asociados previamente, en particular dentro de los módulos teñidos de oscuro en el cuadrante inferior derecho que abarcan tanto genes caracterizados asociados al cáncer como reguladores putativos novedosos. Finalmente, se comparan los mapas de calor para los k-mers de mutaciones en pacientes enfermos y sanos. Al examinar estos mapas de calor (**Figuras 14a** y **14b**), que representan variantes sanas (verde) y variantes asociadas al cáncer de mama (naranja) respectivamente, surgen varias perspectivas adicionales que complementan el análisis anterior:

- La estructura jerárquica de agrupación sigue siendo coherente entre ambos conjuntos de datos, lo que sugiere que la arquitectura genética subyacente se conserva a pesar del estado de la enfermedad. Sin embargo, los patrones de distribución de la intensidad de las mutaciones revelan diferencias críticas con posible importancia clínica. En el mapa de variantes de pacientes sanos (**Figuras 14a**), los clusters muestran un patrón más disperso con distintos puntos calientes localizados, particularmente evidentes en las columnas más a la izquierda correspondientes a contextos trinucleotídicos específicos. Por el con-



(a) K-mers para variantes asociadas a pacientes sanos.



(b) K-mers para variantes asociadas a pacientes enfermos.

Figura 12: K-mers generados a partir de variantes genéticas en formato HGVS

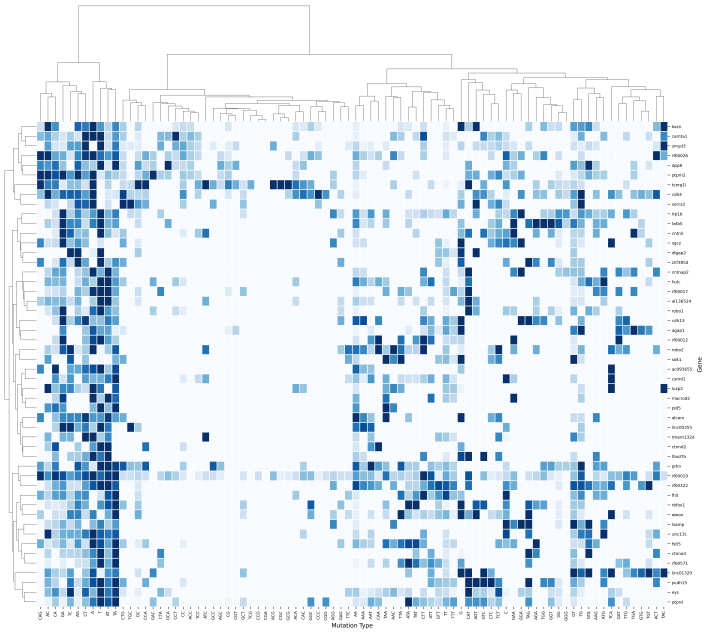
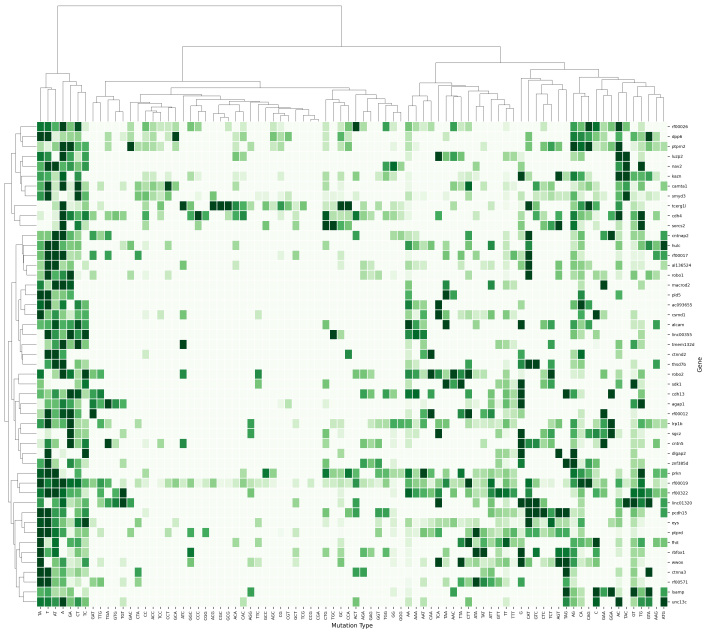


Figura 13: Mapa de calor para k-mers obtenidos asociados con Top 50 de genes más comunes.

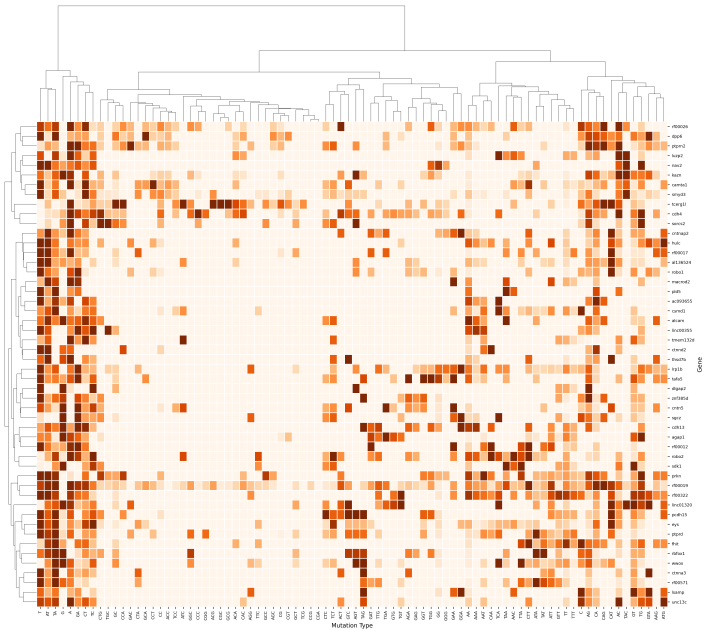
trario, el mapa de variantes de pacientes con cáncer de mama (**Figura 14b**) muestra patrones más generalizados e intensificados en regiones genómicas similares, con señales notablemente intensificadas en genes previamente implicados en vías de malignización (KAZN, ERBB4, PTPRN2).

- El análisis comparativo revela además distribuciones diferenciales de la carga mutacional entre condiciones. Las variantes asociadas al cáncer muestran señales enriquecidas en todos los tipos de mutación, con una concentración particular en las regiones asociadas a las transiciones C>T y C>G (visibles en las partes centrales de la **Figura 14b**), en consonancia con las firmas mutacionales mediadas por APOBEC que se observan con frecuencia en los carcinomas de mama. Además, la matriz de variantes del cáncer muestra patrones de co-ocurrencia más fuertes entre grupos de genes específicos, lo que sugiere vías particulares en lugar de eventos mutacionales estocásticos.

- Resulta especialmente interesante la aparente relación recíproca entre determinados perfiles de mutación en estados sanos y enfermos. Las regiones que muestran una actividad mínima suelen corresponderse con una mayor carga mutacional en las muestras de cáncer, lo que podría identificar nodos de vulnerabilidad críticos. Estos patrones comparativos proporcionan objetivos valiosos para el desarrollo de biomarcadores de diagnóstico, en particular para las vías



(a) K-mers en pacientes sanos.



(b) K-mers en pacientes enfermos.

Figura 14: Comparación de mapas de calor para k-mers según su diagnóstico.

que implican conductores oncogénicos conocidos como ERBB4 y candidatos emergentes como TNC0T329 y MCM2, que muestran una correlación pronunciada entre las condiciones.

4.4 MODELAMIENTO

4.4.1 Dataset de Patogenicidad

Se exploraron múltiples enfoques de clasificación supervisada con el objetivo de identificar patrones discriminativos dados por marcadores genéticos entre variantes patogénicas y no patogénicas como se muestra en la **Figura 15**. En particular, se evaluaron tres algoritmos: Random Forest (RF), Support Vector Machine (SVM) y XGBoost Classifier (XGB) aplicados tanto sobre la base de datos original (no balanceada, con 328.116 instancias) como sobre una versión balanceada mediante undersampling (con 71.480 instancias). Con el fin de optimizar el desempeño predictivo de cada modelo, se implementó un proceso sistemático de ajuste de hiperparámetros basado en búsquedas bayesianas.

Dicho proceso comprendió: (1) la partición estratificada del conjunto de datos en subconjuntos de entrenamiento y validación (70 % y 30 %, respectivamente); (2) la definición de un espacio de exploración de hiperparámetros específico para cada modelo; (3) la formulación de una función objetivo enfocada en maximizar el F1-score promedio a través de validación cruzada estratificada con 10 particiones (StratifiedKFold); y (4) la realización de 50 iteraciones de búsqueda utilizando el algoritmo Tree-structured Parzen Estimator (TPE) mediante la librería Hyperopt. Finalmente, los modelos fueron ajustados con los mejores hiperparámetros identificados (ver Anexo 2) y se generaron métricas de desempeño en clasificación para su posterior análisis comparativo.

4.4.2 Dataset de Variantes a partir de archivos de llamado de variantes (VCF) para pacientes con y sin cáncer de mama familiar

Para el conjunto de datos de Variantes a partir de archivos de llamado de variantes (VCF) para pacientes con y sin cáncer de mama familiar, se empleó el algoritmo XGBoost Classifier como único modelo base, dada su capacidad inherente para manejar variables categóricas sin necesidad de codificación explícita, lo que evita el incremento de dimensionalidad común en técnicas como binary encoding. Aunque esta elección unitaria podría representar una limitación comparativa, se priorizó la eficiencia computacional y la robustez del modelo frente a datos heterogéneos. El proceso de entrenamiento y validación consistió en una partición estratificada del conjunto de

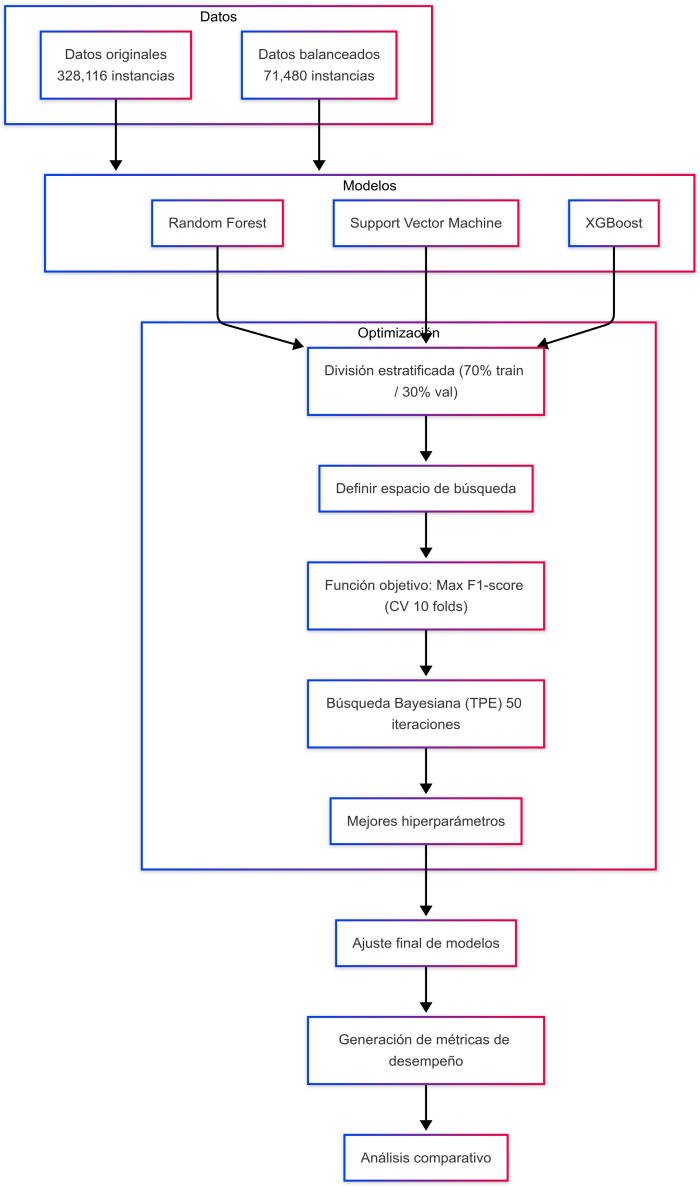


Figura 15: Diagrama de flujo del modelado del dataset de patogenicidad

datos en subconjuntos de entrenamiento/validación (70 %) y prueba (30 %), aplicando una validación cruzada con RepeatedStratifiedKFold (5 particiones, 3 repeticiones), sumando un total de 100 evaluaciones de modelos en una búsqueda bayesiana de hiperparámetros utilizando Hyperopt y el algoritmo Tree-structured Parzen Estimator (TPE). Los hiperparámetros obtenidos para los mejores modelos se pueden encontrar en el Anexo 2. Finalmente, se realizó una validación final con variantes de un conjunto de datos externo al pipeline con el fin de caracterizar la capacidad de generalización del modelo.

La métrica objetivo en todas las iteraciones fue el F1-score promedio. Considerando el desbalance inherente en la variable objetivo (diagnóstico de cáncer), se evaluaron múltiples estrategias de balanceo de clases aplicadas a la partición de entrenamiento incluyendo undersampling, oversampling, SMOTE, SMOTENC y ADASYN, además del conjunto original no balanceado. Cada uno de estos conjuntos fue sometido al mismo proceso de partición, ajuste y evaluación, con el fin de identificar la estrategia más adecuada para mejorar el desempeño del modelo bajo condiciones realistas de desbalance.

Adicionalmente, se implementó una estrategia de reducción de dimensionalidad y agrupamiento no supervisado con el objetivo de identificar patrones latentes no capturados mediante clasificación supervisada. Para ello, se aplicó el algoritmo Uniform Manifold Approximation and Projection (UMAP) en tres dimensiones, optimizando los hiperparámetros clave (n_neighbors y metric) de forma acoplada con el modelo de agrupamiento HDBSCAN a partir de 50 iteraciones. La selección conjunta de estos parámetros se basó en la maximización del silhouette score, con el fin de preservar tanto la estructura local como la separabilidad global en el espacio reducido. Posteriormente y con los hiperparámetros ya obtenidos para HDBSCAN, se determinaron los clústeres iniciales y a través de clustering jerárquico se refinaron clústeres no incluidos originalmente por el modelo. Esta estrategia permitió explorar la presencia de subgrupos fenotípicos o genómicos dentro del conjunto de variantes. Todo el flujo expermiental se resume en la **Figura 16.**

4.5 EVALUACIÓN

Para la evaluación del desempeño del modelo de clasificación aplicado a los datasets anteriormente descritos, se utilizaron métricas estándar en contextos de aprendizaje supervisado, incluyendo accuracy, precisión, sensibilidad (recall), especificidad y F1-score, esta última empleada como criterio principal de optimización. Estas métricas permitieron evaluar de forma integral la capacidad de los modelos para distinguir entre pacientes con y sin diagnóstico de cáncer o variantes patogénicas y no pato-

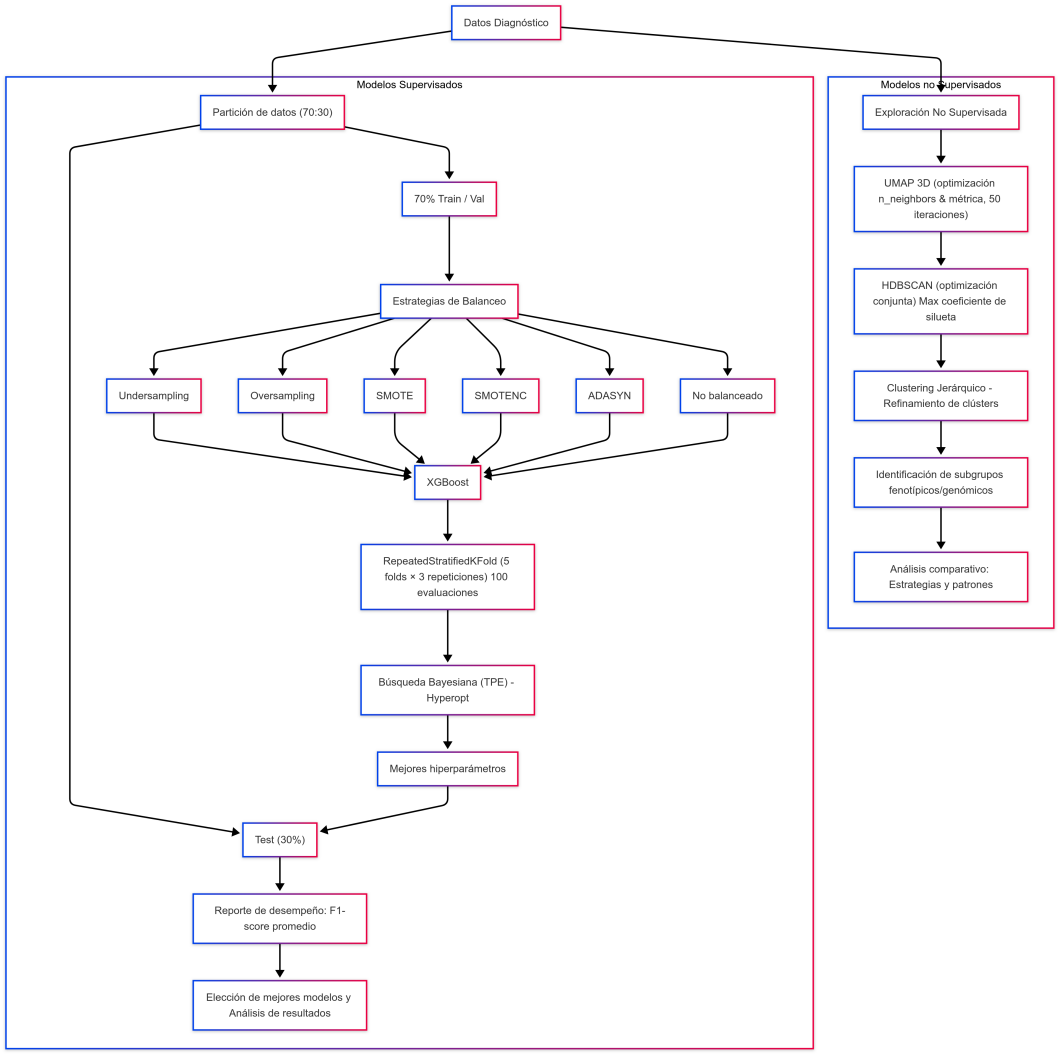


Figura 16: Diagrama de flujo del modelado del dataset de variantes a partir de archivos de llamado de variantes (VCF) para pacientes con y sin cáncer de mama familiar

génicas, priorizando el equilibrio entre verdaderos positivos y falsos negativos.

En especial, para el dataset de variantes a partir de archivos de llamado de variantes (VCF) para pacientes con y sin cáncer de mama familiar, el análisis de desempeño se llevó a cabo comparando múltiples escenarios de clasificación, en los cuales el conjunto de datos de entrada en su partición de entrenamiento fue modificado mediante distintas estrategias de balanceo, tales como undersampling, oversampling, SMOTE, SMOTENC y ADASYN, además del conjunto original no balanceado. Cada una de estas técnicas fue evaluada de forma sistemática mediante validación cruzada estratificada, aplicando las métricas mencionadas con el objetivo de identificar el enfoque de balanceo que maximizara la capacidad predictiva del modelo sin comprometer su generalización. Este enfoque permitió establecer un marco comparativo robusto para seleccionar la configuración más adecuada en términos de desempeño y estabilidad del clasificador.

Para los modelos de aprendizaje no supervisado aplicados al dataset de diagnóstico de cáncer se emplearon métricas especializadas de clustering, tales como el coeficiente de Silueta, el índice de Calinski-Harabasz y el índice de Davies-Bouldin, consideradas estándares cuando no se dispone de etiquetas de referencia. Estas métricas permiten evaluar de forma integral la cohesión interna y la separación entre los grupos generados, facilitando el análisis de la capacidad del modelo para identificar patrones intrínsecos en los datos.

4.5.1 Dataset de Patogenicidad

Modelo	Bal.*	TN	FP	FN	TP	Acc	Precision	Recall	F1	t**
XGB	-	86285	1428	1203	9519	0.9733	0.8696	0.8878	0.8786	145.5
RF	-	86086	1627	1282	9440	0.9704	0.8530	0.8804	0.8665	113.8
XGB	**US**	**10228**	**494**	**317**	**10405**	**0.9622**	**0.9547**	**0.9704**	**0.9625**	**29.1**
RF	US	10263	459	326	10396	0.9637	0.9577	0.9696	0.9636	187.6
SVM	US	10340	506	465	10382	0.9552	0.9535	0.9571	0.9553	$>$720

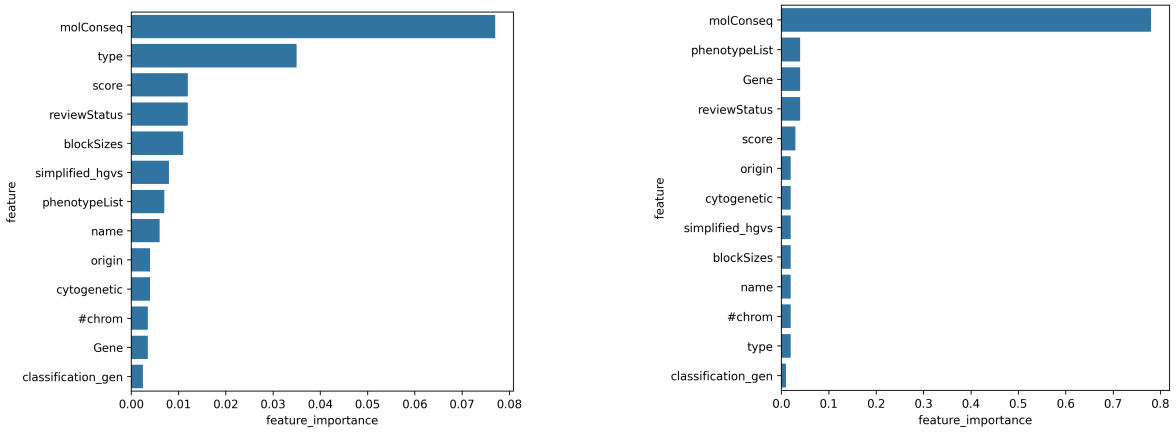
* **Bal.** Describe la técnica de balanceado usada: - (None), US (Undersampling).
** **t** Describe la cantidad de tiempo en minutos usado para la optimización con el espacio de hiperparámetros definido. Los tiempos fueron obtenidos con AMD Ryzen 7 3700U with Radeon Vega Mobile Gfx 2.30 GHz y 20.0 GB de RAM.

Tabla 4: Resultados obtenidos para clasificación en el dataset de patogenicidad

Con el objetivo de predecir la patogenicidad de variantes genéticas, se evaluaron tres algoritmos de clasificación: XGBoost (XGB), Random Forest (RF) y Support Vector Machine (SVM), considerando dos configuraciones de datos: la base original no balanceada y una versión balanceada mediante undersampling (US). El modelo XGB entrenado sobre la base no balanceada obtuvo una precisión de 0.8696, un recall

de 0.8878 y un F1-score de 0.8786, posicionándose como el modelo con mejor desempeño en esta configuración. No obstante, al aplicar la técnica de balanceo, se evidenció una mejora considerable en el recall del modelo para todos los algoritmos, destacando el rendimiento de RF (recall = 0.9696; F1 = 0.9612) y SVM (recall = 0.9571; F1 = 0.9553). Esta mejora estuvo acompañada por una reducción sustancial en los falsos negativos, aspecto crítico en escenarios clínicos donde la omisión de variantes patogénicas puede comprometer la toma de decisiones médicas.

El balanceo por undersampling también impactó el tiempo de cómputo requerido para la optimización de hiperparámetros. En particular, el modelo SVM presentó un tiempo de optimización superior a 720 minutos, lo cual limitó su entrenamiento sobre la base original no balanceada. Dado que esta versión contiene un mayor número de observaciones y atributos, se estimó que el proceso sería computacionalmente inviable bajo las condiciones experimentales disponibles. Por esta razón, se excluyó la evaluación de SVM en dicha configuración. Este hallazgo subraya una limitación práctica del uso de clasificadores basados en márgenes en contextos con alta dimensionalidad y desbalance de clases, especialmente cuando se requiere un ajuste fino de hiperparámetros mediante búsquedas exhaustivas.



(a) RF (b) XGB

Figura 17: Importancia de características para modelos RF y XGB

Adicionalmente, se examinó la contribución individual de las variables predictoras mediante análisis de importancia de características. Las **Figuras 17a** y **17b** ilustran los resultados obtenidos para los modelos XGB y RF entrenados sobre la base balanceada. En ambos casos, la variable molConseq emergió como la característica más influyente. En XGB, le siguieron phenotypeList y Gene, mientras que en RF destacó type. Las diferencias observadas en los perfiles de importancia reflejan las distintas

estrategias de particionamiento y asignación de peso que cada algoritmo aplica internamente, y aportan elementos interpretables sobre los determinantes genómicos más relevantes en la predicción de patogenicidad.

4.5.2 Dataset de Variantes a partir de archivos de llamado de variantes (VCF) para pacientes con y sin cáncer de mama familiar

En esta sección se abordó la tarea de clasificar variantes genéticas asociadas con cáncer, utilizando el modelo XGBoost (XGB) como clasificador base. Se exploraron cinco estrategias de balanceo de clases: undersampling (US), duplicación de la clase minoritaria (OS), SMOTE, SMOTENC y ADASYN. Los resultados se presentan en términos de matrices de confusión, desempeño en el conjunto de prueba y tiempos de optimización.

Bal.*	TN	FP	FN	TP	Acc	Precision	Recall	F1	t**
-	56	1479	35	103277	0.9856	0.9805	0.9856	0.9792	23.5
US	1413	122	426	1109	0.8215	0.8346	0.8215	0.8197	1.5
OS	**102313**	**0**	**633**	**102679**	**0.9969**	**0.9970**	**0.9969**	**0.9969**	**36.7**
SMOTE	101853	1460	127	103185	0.9923	0.9924	0.9923	0.9923	39.2
SMOTENC	101891	1422	123	103189	0.9917	0.9918	0.9917	0.9917	44.1
ADASYN	**102150**	**119**	**119**	**103194**	**0.9923**	**0.9923**	**0.9923**	**0.9923**	**41.5**

* **Bal.** Describe la técnica de balanceado usada: - (None), US (Undersampling), OS (Oversampling usando Repetición de categoría minoritaria).
** **t** Describe la cantidad de tiempo en minutos usado para la optimización con el espacio de hiperparámetros definido. Los tiempos fueron obtenidos con AMD Ryzen 7 3700U with Radeon Vega Mobile Gfx 2.30 GHz y 20.0 GB de RAM.

Tabla 5: Resultados obtenidos para el dataset de variantes a partir de archivos de llamado de variantes (VCF) para pacientes con y sin cáncer de mama familiar

El modelo entrenado sin técnicas de balanceo mostró un alto F1-score en el conjunto de prueba (0.9792), pero con una cantidad de falsos negativos relativamente alta para la cantidad total de variantes que corresponden con esta categoría (FN = 35), lo que representa un riesgo clínico al omitir variantes relevantes para la clasificación. La técnica de undersampling, aunque efectiva para reducir falsos positivos, sacrificó parte del rendimiento general (F1 = 0.8197; Accuracy = 0.8215), probablemente a causa de la pérdida de información al eliminar ejemplos de la clase mayoritaria.

En contraste, las técnicas de sobremuestreo demostraron mejores resultados tanto en términos de recall como de precisión. De forma destacada, la estrategia de oversampling por duplicación directa de la clase minoritaria (OS) obtuvo cero falsos negativos (FN = 0) y presentó el mejor rendimiento en precisión (0.9969), recall (1.0000) y F1-score (0.9984), con un tiempo de optimización relativamente bajo (36.7 minutos). Esta observación sugiere que la repetición controlada de ejemplos reales

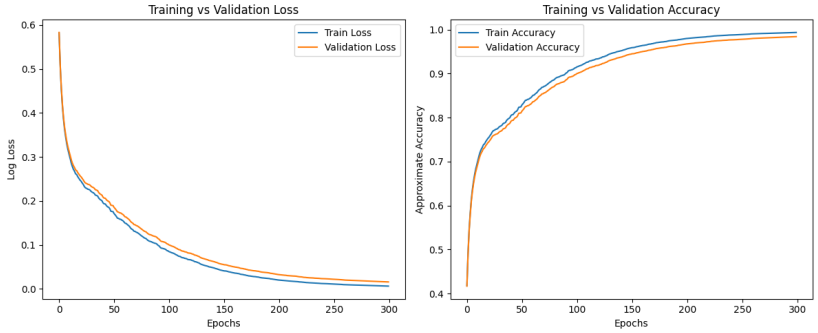
permite al modelo capturar de forma robusta los patrones de la clase minoritaria, minimizando el riesgo de omisiones. Por otro lado, los métodos basados en generación sintética como SMOTE, SMOTENC y ADASYN también mostraron un excelente desempeño, con F1-scores en el rango de 0.9917 a 0.9923 y muy pocos falsos negativos (entre 119 y 127). De estos, ADASYN destacó al lograr un equilibrio entre la preservación de la información funcional y la incorporación de nuevos ejemplos en áreas subrepresentadas, lo que permitió al modelo alcanzar una alta sensibilidad sin comprometer la precisión.

4.5.2.1 Dinámicas de aprendizaje para OS, SMOTE y ADASYN

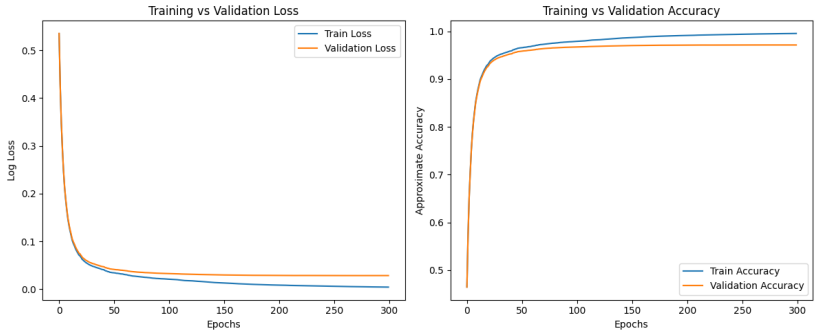
Las curvas de aprendizaje presentadas para los métodos de balanceo OS (duplicación de la clase minoritaria), ADASYN y SMOTE revelan patrones distintivos en el comportamiento del modelo XGBoost durante el entrenamiento y la validación. En el caso de balanceo OS, la **Figura 18a** muestra una convergencia más gradual tanto en pérdida como en precisión en comparación con ADASYN y SMOTE.

La curva de pérdida presenta una disminución constante pero más lenta, requiriendo aproximadamente 200 épocas para estabilizarse completamente. Esta característica sugiere un proceso de aprendizaje más lento, donde el modelo va ajustándose progresivamente a los patrones de la clase minoritaria duplicada. La brecha entre el rendimiento de entrenamiento y validación es ligeramente mayor que en los otros métodos, particularmente en las épocas intermedias (50-150), lo que podría indicar que esta ralentización en el aprendizaje podría usarse para evitar el sobreajuste mientras se capturan patrones complejos.

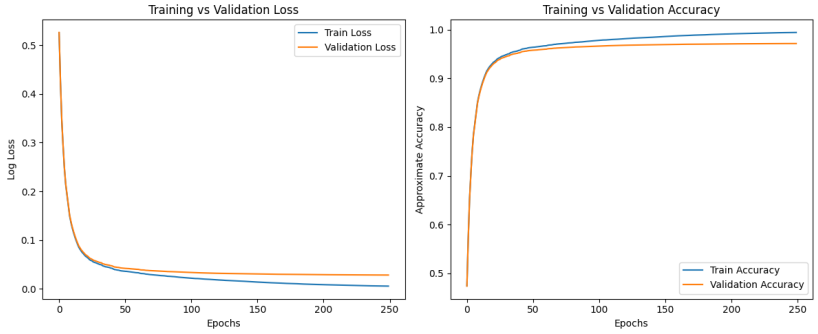
En el caso de SMOTE y ADASYN (**Figura 18b** y **18c**), las curvas para estos dos métodos exhiben comportamientos muy similares, caracterizados por una convergencia extremadamente rápida. Ambos muestran una caída pronunciada en la pérdida y un aumento acelerado en la precisión durante las primeras 25 épocas, seguidos de una estabilización temprana. Esta rápida convergencia podría atribuirse a que los ejemplos sintéticos generados crean un espacio de características más uniforme y fácilmente separable para el clasificador. Es notable que tanto ADASYN como SMOTE mantienen una diferencia muy pequeña entre las curvas de entrenamiento y validación, lo que sugiere una buena generalización sin signos evidentes de sobreajuste.



(a) OS



(b) SMOTE

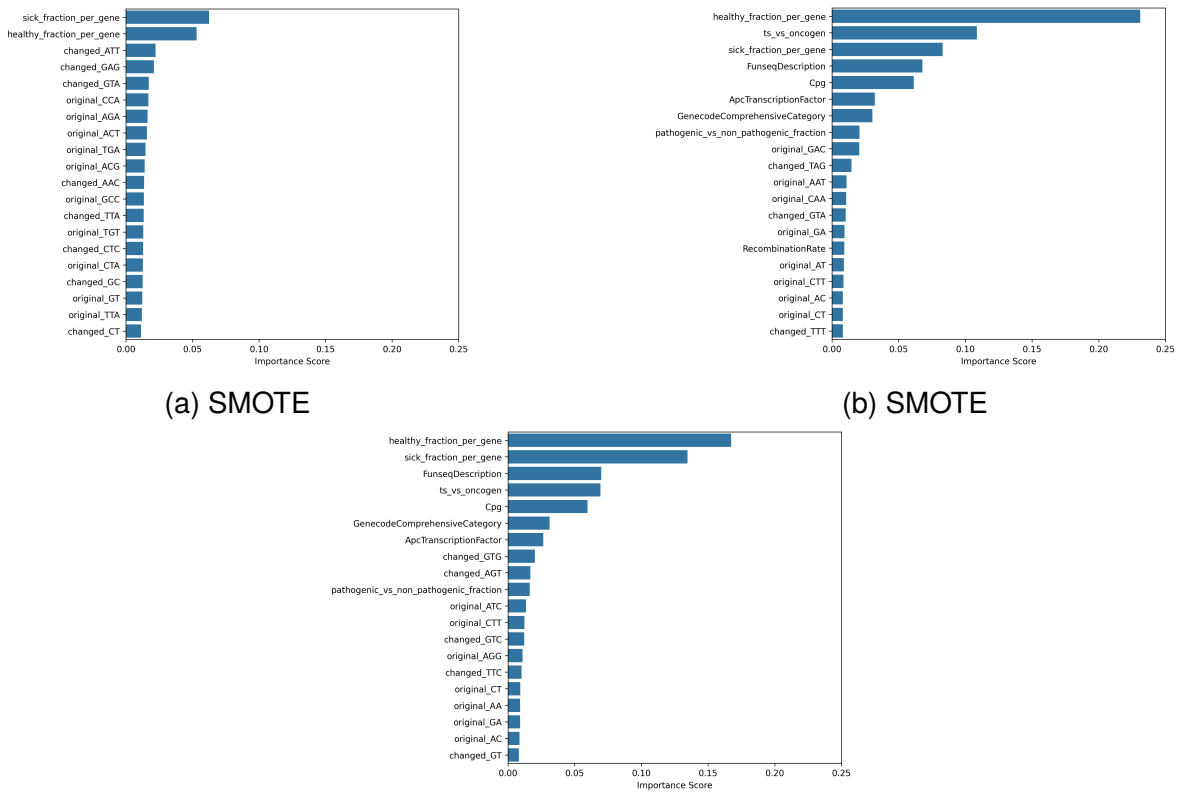


(c) ADASYN

Figura 18: Dinámicas de aprendizaje para XGB y técnicas de sobremuestreo

4.5.2.2 Impacto en la importancia de atributos

Como se puede apreciar en la **Figura 19**, la técnica de balanceo aplicada no solo afectó las métricas globales del modelo, sino que también modificó la importancia relativa de las variables predictoras. En el caso del conjunto sin balancear, el modelo priorizó fuertemente las fracciones por gen (por ejemplo, sick_fraction_per_gene y healthy_fraction_per_gene), lo que puede interpretarse como un sesgo hacia patrones macro que reflejan el desbalance inherente en los datos. En cambio, con la estrategia de oversampling mediante duplicación (OS), se observó un notable incremento en la relevancia de atributos derivados de los k-mers (por ejemplo, variables relacionadas con secuencias mutacionales como changed_ATT o original_CTA). Este hallazgo indica que la repetición directa de ejemplos de la clase minoritaria permite al modelo enfocarse en patrones secuenciales específicos, minimizando el "ruido" generado por la clase mayoritaria.



(a) SMOTE

(b) SMOTE

(c) ADASYN

Figura 19: Importancia de atributos para OS, SMOTE y ADASYN

Por su parte, la técnica de sobremuestreo sintético SMOTE incrementó la visibilidad de variables funcionales y categóricas (tales como FunseqDescription, CpG o ts_vs_oncogen), junto con la persistencia de la señal proveniente de las fracciones por gen. Esto sugiere que la generación de nuevos ejemplos en el espacio de características favorece la identificación de atributos más abstractos y generalizables. Finalmente, con ADASYN se observó un patrón híbrido, en el cual tanto las variables funcionales como algunos atributos basados en k-mers adquieren relevancia, lo que evidencia la capacidad de esta técnica para explorar de forma equilibrada las regiones del espacio de datos más difíciles de clasificar. En conjunto, estos análisis indican que la elección de la estrategia de balanceo no solo tiene un impacto directo sobre las métricas de desempeño, sino que también modifica la interpretación biológica del modelo, resaltando diferentes aspectos de la información contenida en las variantes genéticas asociadas con cáncer.

4.5.2.3 Clustering

nn.	ms.	mcs.	metric	sil. score	db score	ch score
1082	787	496	chebyshev	**0.6064**	0.6063	63276.03
1177	928	646	chebyshev	0.6063	0.6064	**63278.98**
183	626	875	chebyshev	0.6023	**0.6060**	61878.42
1179	691	1022	chebyshev	0.5906	0.6433	60129.44
1303	863	658	chebyshev	0.5864	0.6451	58888.75

* Todas las combinaciones en el top 5 muestran la métrica Chebyshev como la mejor métrica de distancia.
** La abreviaturas incluyen: nn. (n_neighbors), ms. (min_samples), mcs. (min_cluster_size), sil. score (Silhouette score), db score (Davies-Bouldin) y ch score (Calinski-Harabasz).

Tabla 6: Resultados de optimización para parámetros de clustering con optimizador acoplado UMAP/HDBSCAN

Para explorar la estructura latente de los pacientes y sus variantes, se implementó un proceso acoplado de reducción de dimensionalidad y agrupamiento no supervisado. Inicialmente, se optimizó conjuntamente los hiperparámetros de UMAP (n_neighbors y metric) y de HDBSCAN (min_samples y min_cluster_size) mediante una búsqueda bayesiana de 50 iteraciones, utilizando el silhouette score como función objetivo. Con los valores óptimos identificados, se generaron embeddings tridimensionales sobre los cuales HDBSCAN detectó ocho clústeres (**Tabla 6**).

La **Figura 20** resume los resultados de las 50 iteraciones de optimización, revelando que la métrica de distancia Chebyshev produjo consistentemente los puntajes de silueta más elevados, seguida por las distancias Euclidiana y Minkowski. Esta superioridad de Chebyshev sugiere que, en el espacio reducido, las separaciones entre clústeres se caracterizan principalmente por diferencias máximas en dimensiones

individuales más que por diferencias acumulativas en todas las dimensiones.

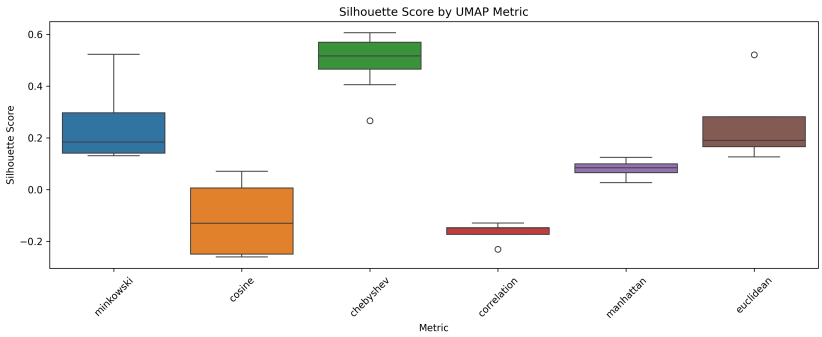


Figura 20: Silhouette scores por métrica de distancia

Con el fin de comparar el performance del modelo en el resto de métricas de distancia y respecto a los índices de calidad de clustering se generó la **Figura 21** donde se confirma que las configuraciones con Chebyshev no solo maximizan el silhouette score, sino que también minimizan el Davies-Bouldin index y maximizan el Calinski-Harabasz score, validando así la robustez de esta métrica según múltiples criterios.

Complementariamente, se realizó un análisis de clustering jerárquico para identificar subestructuras más finas dentro del espacio reducido por UMAP y no capturadas por HDBSCAN con la configuración inicial (**Figura 22**). Al incluir todos los datos (incluyendo outliers clasificados como -1 por HDBSCAN), el clustering jerárquico mostró un silhouette score medio de 0.4262, un puntaje de Davies-Bouldin de 5.2071 y un Calinski-Harabasz score de 56923.28. El refinamiento sobre los outliers mejoró significativamente la cohesión y separación de los clústeres, elevando el silhouette score a 0.6477, reduciendo el Davies-Bouldin index a 0.4580 y aumentando el Calinski-Harabasz score a 233823.88. Estos resultados demuestran que el clustering jerárquico complementa efectivamente a HDBSCAN en la caracterización de subestructuras relevantes, particularmente cuando se controla la influencia de los outliers.

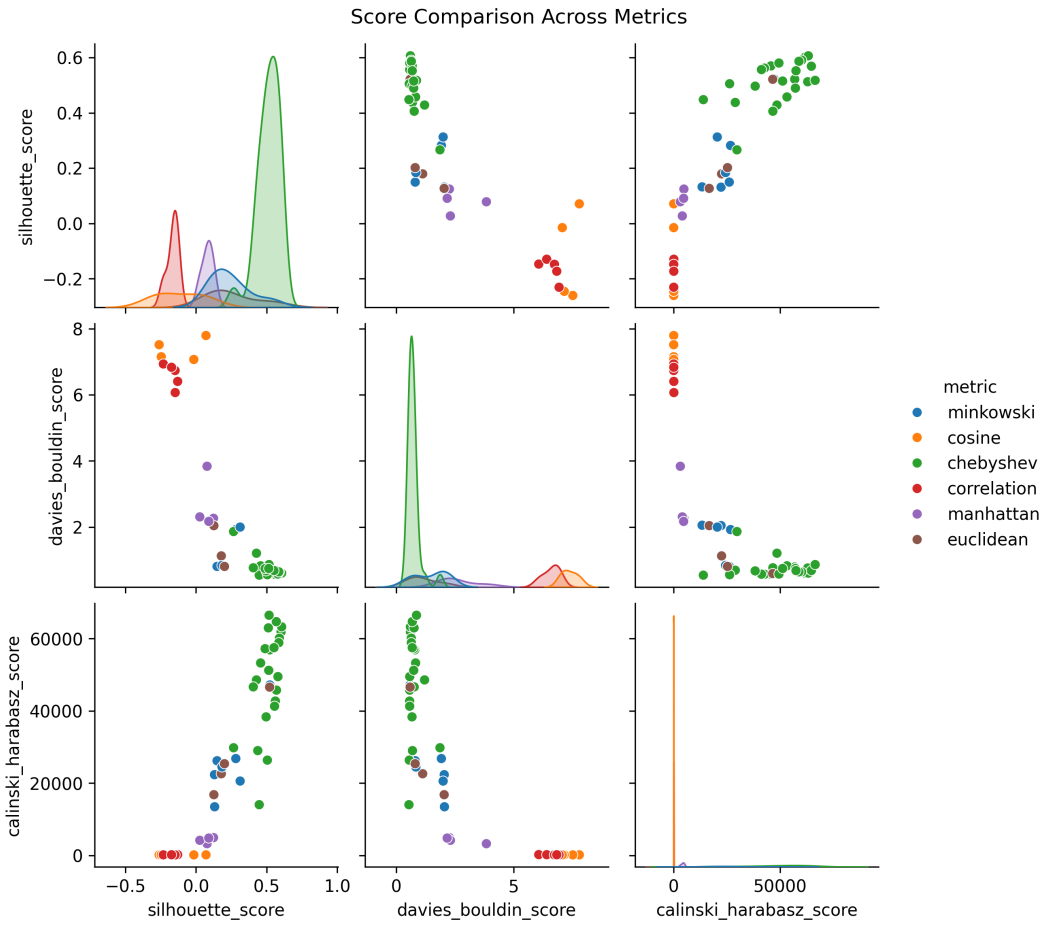


Figura 21: Diagnóstico para optimización de hiperparámetros

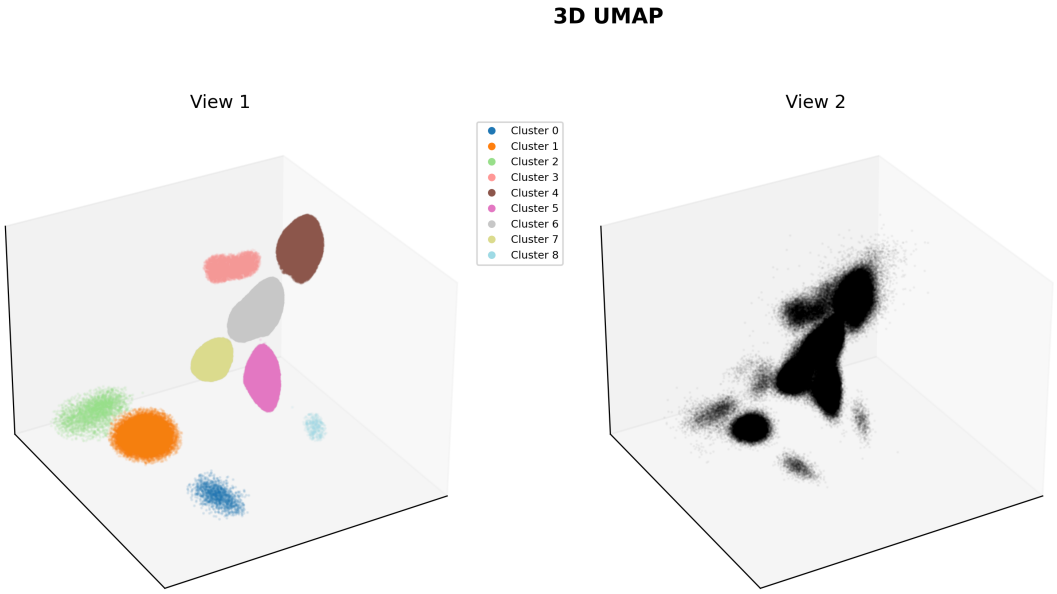


Figura 22: Clusters obtenidos para el dataset de variantes a partir de archivos de llamado de variantes (VCF) para pacientes con y sin cáncer de mama familiar

4.6 DISCUSIÓN

4.6.1 Dataset de Patogenicidad

Los resultados obtenidos en la clasificación de patogenicidad permiten extraer varias conclusiones tanto metodológicas como biológicas. En primer lugar, es necesario enfatizar que la identificación de una variante como patogénica (PG) no equivale de forma directa al diagnóstico o estado clínico del paciente, pues la penetrancia y expresividad varían ampliamente según contexto molecular y ambiental. De hecho, en el conjunto de datos el 43 % de todas las variantes observadas en pacientes con cáncer correspondía a la categoría de significancia desconocida (VUS) lo que subraya la gran incertidumbre clínica aún presente en la interpretación de variantes genéticas.

Desde la perspectiva genómica, las mutaciones se distribuyen por todos los cromosomas sin una concentración exclusiva, destacando que el cromosoma 17 presentó el mayor número de variantes (17500 VUS, 1250 BN y 8500 PG) y el cromosoma 6 albergó la menor carga mutacional. La gran mayoría de variantes de significancia desconocida (VUS) correspondieron a variantes missense, que modifican un solo aminoácido y cuya consecuencia funcional suele ser difícil de predecir, razón por la cual no se clasifican claramente como BN o PG.

Por el contrario, las variantes PG se concentraron especialmente en categorías de alto impacto, como nonsense y frameshift, a menudo asociadas a pérdida de función. En todos los casos, las variaciones de un solo nucleótido (SNVs) fueron predominantes, con casi 140000 registros de VUS, mientras que las deleciones explicaron gran parte de las PG detectadas. La categoría intron variant mostró la mayor proporción conjunta de BN y VUS, sugiriendo que algunas VUS intrónicas podrían en realidad corresponder a cambios benignos y merecen estudio funcional adicional.

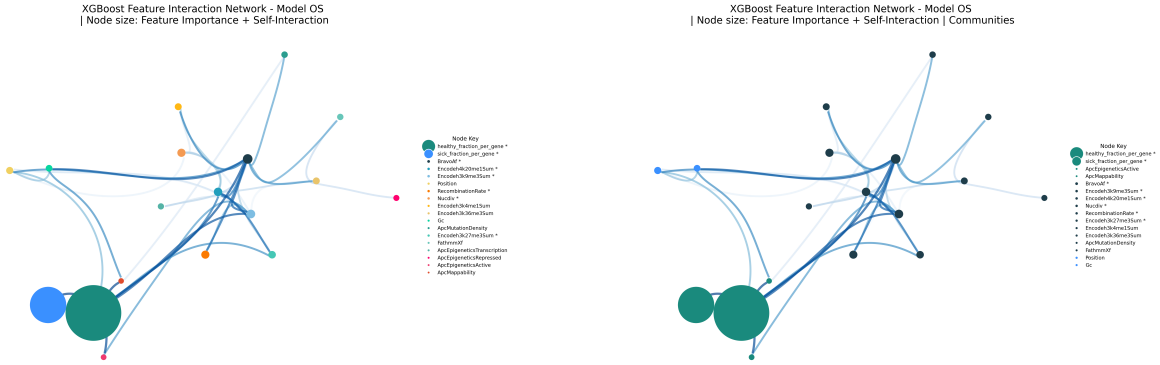
A nivel de genes, destacaron dos perfiles opuestos: los supresores de tumores como APC y ATM acumularon gran parte de las VUS, lo cual refleja la complejidad de interpretar variantes en genes de alta conservación funcional; en cambio, genes como BRCA1, BRCA2 y NF1 se asociaron mayoritariamente a mutaciones PG, coherente con su rol crítico en la prevención de oncogénesis. Curiosamente, BRCA1 y BRCA2 también registraron variantes BN, lo que apunta a la necesidad de evaluar cada variante en su contexto clínico y familiar.

En cuanto a la interpretación del modelo, la variable molConseq o consecuencia molecular emergió como el predictor más influyente en ambos clasificadores, indicando que la categoría de variante (*missense, nonsense, frameshift, etc.*) es el factor más determinante para discriminar una variante patogénica. En XGBoost, le siguieron en importancia phenotypeList (asociaciones fenotípicas conocidas) y Gene, mientras que en Random Forest destacó la variable type (tipo de variante). Estas diferencias reflejan la forma en que cada algoritmo pondera la información: XGBoost integra de manera más fuerte la historia fenotípica y génica, mientras que Random Forest se apoya en la clasificación básica de la variante.

4.6.2 Dataset de Variantes a partir de archivos de llamado de variantes (VCF) para pacientes con y sin cáncer de mama familiar

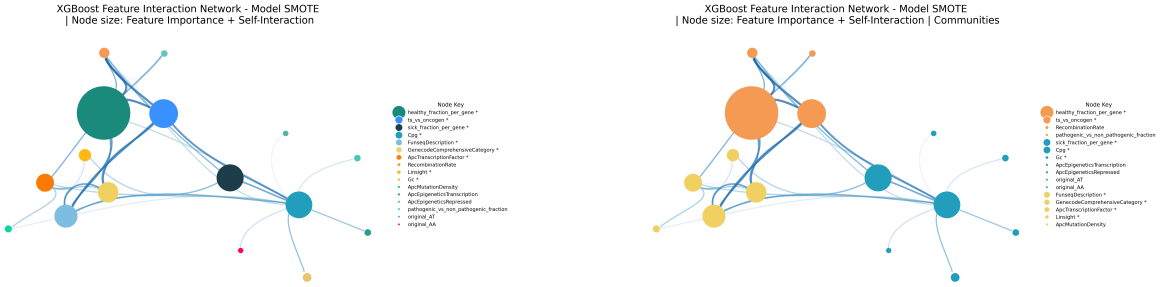
4.6.2.1 Análisis de redes de interacción de características

El análisis de las redes de interacción de características mediante XGBoost a través de *xgbfir*, aplicado a estrategias de sobremuestreo (OS, SMOTE y ADASYN), permitió identificar patrones distintivos en la importancia e interrelación de rasgos genómicos para la predicción. En el modelo OS, se observa una topología dominada por las variables healthy_fraction_per_gene y sick_fraction_per_gene, que además establecen conexiones con características epigenéticas. La estructura comunitaria (**Figura 23**) revela que estas últimas se agrupan de forma cohesionada, evidenciando que el sobremuestreo con OS captura potencialmente mecanismos biológicos relacionados con la regulación epigenética, subrepresentados en otros enfoques.



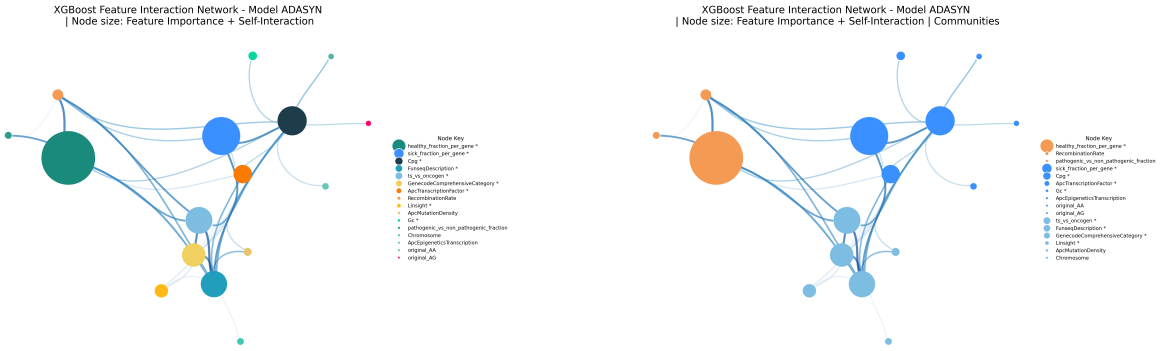
(a) Identidades de Nodos (b) Comunidades

Figura 23: Análisis de red para atributos en OS



(a) Identidades de Nodos (b) Comunidades

Figura 24: Análisis de red para atributos en SMOTE



(a) Identidades de Nodos (b) Comunidades

Figura 25: Análisis de red para atributos en ADASYN

El modelo basado en SMOTE presenta una distribución más equilibrada en la relevancia de sus características. Aunque healthy_fraction_per_gene se mantiene como

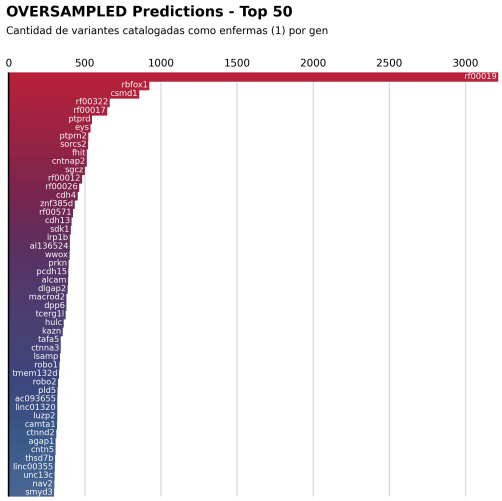
factor prominente, sus interacciones con otros rasgos son especialmente intensas. La estructura de comunidades (**Figura 24b**) se descompone en tres grupos distintos (naranja, amarillo y azul), destacando la capacidad de SMOTE para captar complejas relaciones biológicas, en particular aquellas que involucran la interacción entre factores de transcripción y métricas de salud genómica.

Por último, ADASYN muestra una marcada agrupación funcional de rasgos, con healthy_fraction_per_gene emergiendo como el nodo dominante (**Figura 25a**) y estableciendo conexiones extensas con sick_fraction_per_gene, ts_vs_oncogen y Cpg. La detección de comunidades (**Figura 25b**) evidencia la separación de las características en dos agrupaciones diferenciadas: una compuesta por métricas de salud y otra por rasgos relacionados con CpG y transcripción, lo que sugiere que ADASYN prioriza mecanismos biológicos distintos a los enfatizados por los demás modelos. Esta diferenciación en la topología y las interacciones de características entre los modelos destaca la importancia de la elección de la estrategia de sobremuestreo para el descubrimiento de relaciones biológicas subyacentes, aportando novedosas perspectivas en la modelación predictiva con datos genómicos.

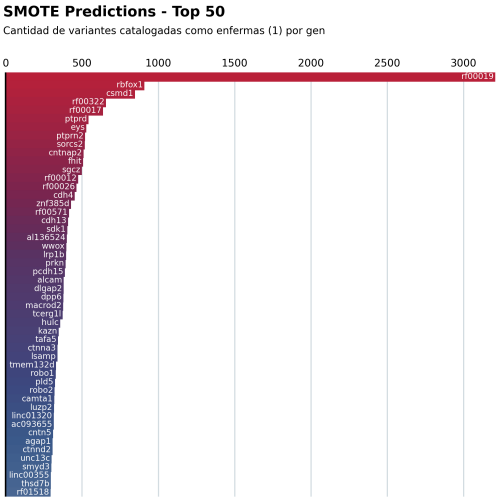
4.6.2.2 Marcadores genéticos asociados con variantes patogénicas

Para profundizar en los determinantes génicos de las variantes clasificadas como enfermas (1) por los modelos de XGBoost sobre ambos datasets pero haciendo énfasis en el *Dataset de variantes (VCF) para pacientes con y sin cáncer de mama familiar*, se analizaron los 50 genes con mayor número de predicciones positivas en cada técnica de balanceo (OS, SMOTE y ADASYN como se muestra en la **Figura 26**). De manera destacable, el análisis identificó un panel de marcadores clave que se presentan de manera consistente y que incluye tanto ARN no codificantes como genes que codifican proteínas:

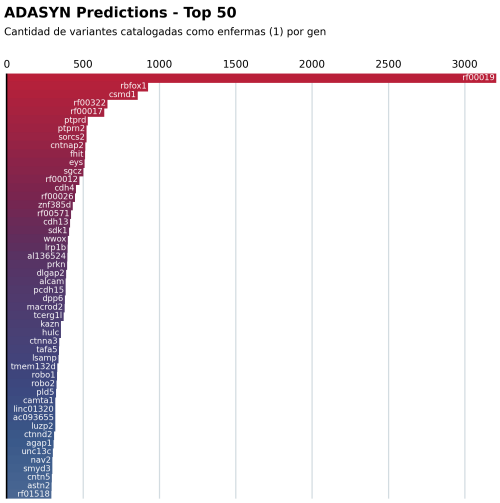
- **ARN no codificantes**
    - *rf00019 (YRNA):* Implicado en la replicación del ADN y el control de calidad del ARN.
    - *rf00322 (snoRNA SNORA31):* Modula modificaciones del ARN ribosomal.
    - *rf00017 (Metazoan signal recognition particle RNA):* Facilita el tráfico y la localización de proteínas.
    - *rf00012 (snoRNA U3):* Esencial para el procesamiento del ARN preribosomal.
    - *rf00026 (spliceosomal RNA U6):* Clave para el empalme del ARN mensajero; su desregulación puede conducir a isoformas proteicas aberrantes.



(a) Conteo de variantes asociadas con cáncer por Gen para OS



(b) Conteo de variantes asociadas con cáncer por Gen para SMOTE



(c) Conteo de variantes asociadas con cáncer por Gen para ADASYN

Figura 26: Genes asociados con variantes cancerígenas por modelo

- **Genes codificantes de proteínas**
  - *RBFOX1:* Factor de empalme alternativo cuyo desajuste promueve la transición epitelio mesénquima.
  - *PTPRD / PTPRN2:* Fosfatasas receptoras implicadas en la transducción de señales; pérdida de función asocia con activación de vías proliferativas.
  - *CNTNAP2:* Proteína de adhesión célula célula, cuya alteración facilita la

invasión tumoral.

- *CDH4 (Cadherin 4) y SDK1 (Sidekick 1):* Moléculas de adhesión celular críticas para la integridad tisular; su disrupción es un rasgo de metástasis.
- *CSMD1:* Posible supresor tumoral con dominios CUB y Sushi,
- *EYS:* Aunque clásicamente asociado a retinopatías, su patrón de expresión en tejido mamario indica un posible rol en la integridad del citoesqueleto y la arquitectura glandular.

- **Comparación de Técnicas de Balanceo**

  Se evidencia que la elección de la técnica de balanceo incide no solo en el rendimiento cuantitativo del modelo, sino también en la perspectiva biológica emergente. En efecto, el oversampling por duplicación (OS) potencia primordialmente los patrones secuenciales de mutación pura (kmers) y los ARN no codificantes, destacándose los elementos RF00019 y RF00322 en la dominancia del grafo de interacción, mientras que, entre los genes codificantes, RBFOX1 y PTPRD presentan una alta centralidad que refleja una especialización en la ocurrencia de eventos mutacionales realmente pertinentes.

  Por su parte, SMOTE otorga visibilidad a atributos funcionales como la clasificación de gen (supresor de tumor u oncogen) y FunseqDescription, propiciando la elevación de la expresión de CNTNAP2 y EYS, lo que sugiere que la generación sintética contribuye a la expansión de rutas de señalización y adhesión; en este contexto, los ARN no codificantes permanecen relevantes, integrándose en comunidades más amplias junto a genes implicados en procesos de reparación y regulación epigenética. Asimismo, la técnica ADASYN no solo conserva la centralidad de RBFOX1, PTPRD y de los Y ARNs, sino que además realza genes pertenecientes a la vía PI3K/AKT y a cascadas de reparación del ADN, reflejando una exploración equilibrada de regiones complejas del espacio de características, lo que se traduce en una mayor ponderación de CSMD1 y CDH4, señalando la posible implicación en rutas de supresión tumoral y adhesión celular en áreas con baja densidad de variantes mapeadas en este dataset.

  Estos hallazgos subrayan la relevancia de la selección meticulosa de la técnica de balanceo, ya que esta determina la “visión biológica” que el modelo adquiere y, en consecuencia, la identificación de potenciales objetivos biomoleculares en el cáncer de mama.

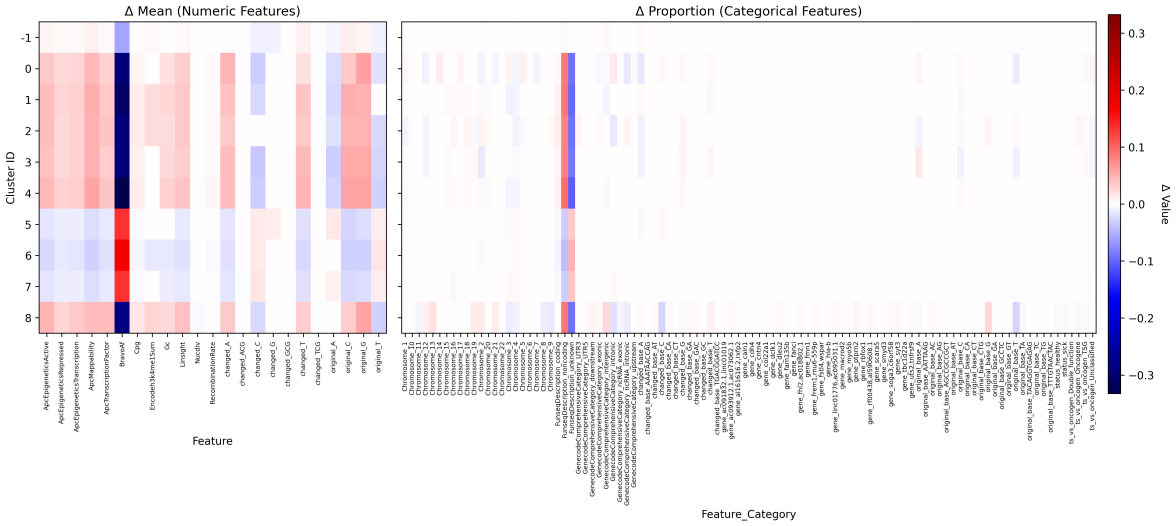


Figura 27: Características de los clústeres obtenidos para el dataset de variantes a partir de archivos de llamado de variantes (VCF) para pacientes con y sin cáncer de mama familiar

4.6.2.3 Caracterización de clústeres

Para evaluar si los clústeres definidos en el espacio tridimensional de UMAP aportan nueva información sobre la relación entre variantes y genes asociados al cáncer de mama, se calculó la diferencia media (Δ) de anotaciones numéricas y proporciones de variables categóricas en cada clúster respecto al conjunto completo (**Figura 27**).

Los clústeres 0-4 y 8 son caracterizados por valores fuertemente negativos de BravoAf (δ = -0,28 a -0,33), sugieren un enriquecimiento de variantes raras o somáticas, mientras que los conglomerados 5-7 (δ = +0,13 a +0,18) reflejan polimorfismos comunes. En los clústeres de alto riesgo (0-4, 8), predominan los estados epigenéticos activos (ApcEpigeneticsActive, δ = +0,03 a+0,05) y la unión de factores de transcripción (ApcTranscriptionFactor, δ = +0,03 a+0,04), lo que implica una disrupción regulatoria. Un ejemplo ilustrativo es PTPRD (δ = -0,00016 en el cluster -1) que se mapea en regiones con actividad epigenética elevada (cluster 8: ApcEpigeneticsActive δ = +0,048), sugiriendo que su papel como supresor tumoral podría estar silenciado mediante mecanismos epigenéticos.

Las puntuaciones de Linsight (δ = +0,03 a+0,04), predictivas de la patogenicidad de variantes no codificantes, están elevadas en los clústers 0-4, alineándose con CNTNAP2 (δ = +0,00022 en el conglomerado 5) y EYS (δ = +0,00017 en el cluster 4), ambos vinculados a la adhesión celular y metástasis del cáncer. Las variantes es-

tructurales y la accesibilidad de la cromatina presentan patrones distintivos. ApcMappability (δ = +0,04 a+0,06 en los clústeres 0-4) destaca regiones de baja complejidad de secuencia, frecuentemente asociadas con reordenamientos estructurales. CDH4 (δ = +0,00103 en el cluster 2), un gen de la familia de las cadherinas, reside en tales regiones, potencialmente explicando su desregulación en la invasión del cáncer de mama. El cluster 8 exhibe una represión extrema de la cromatina (ApcEpigeneticsRepressed δ = +0,027), superponiéndose con PTPRN2 (δ = +0,0068), un gen implicado en la señalización de receptores hormonales.

Los conglomerados de variantes comunes (5-7), marcados por valores positivos de BravoAf (δ = +0,13 a+0,18), se correlacionan con CSMD1 (δ = +0,00014 en el cluster -1), un supresor tumoral frecuentemente eliminado en el cáncer de mama. La relación inversa entre CSMD1 y las variantes comunes sugiere loci de susceptibilidad germinal. Por otro lado, Changed_T (δ = -0,018 a -0,024 en los clústeres 5-7) resalta las transversiones T>G/C, potencialmente dañinas para SDK1 (no observado directamente pero vinculado a conglomerados enriquecidos en ApcTranscriptionFactor), una quinasa involucrada en la migración celular.

Respecto a los ARN no codificantes y factores de splicing, aunque RF00019 y RF00322 (ARNs no codificantes) no fueron explícitamente mapeados para el top 10 de categorías más importantes, los conglomerados con alto EncodeH3K4me1 (δ = +0,023 a +0,028) y ApcEpigeneticsTranscription (δ = +0,03 a+0,04) sugieren potenciadores activos cerca de estos loci, potencialmente modulando el riesgo de cáncer de mama a través de vías mediadas por lncRNA. Estos clústeres delinean distintas vías etiológicas: variantes reguladoras raras (clústeres 0-4, 8) que interrumpen los supresores tumorales (PTPRD, EYS) y cadherinas (CDH4), mientras que las variantes comunes (clústeres 5-7) implican amplios loci de susceptibilidad (CSMD1). La interacción entre la desregulación epigenética, la variación estructural y la actividad de ARN no codificante proporciona un marco para priorizar estudios funcionales en la genómica del cáncer de mama.

5. CONCLUSIONES Y TRABAJOS FUTUROS

5.1 CONCLUSIONES

- **Conclusión general**

  En este estudio se logró implementar exitosamente un modelo de clasificación basado en Machine Learning que identifica marcadores genéticos a partir de patrones estructurales asociados con cáncer de mama, considerando su contexto (loci) para evaluar la probabilidad de generación de factores patogénicos. El modelo desarrollado integra técnicas de aprendizaje supervisado, muestreo de clases, análisis de importancia de características, redes de interacción y métodos de clustering no supervisado, demostrando la capacidad de clasificar variantes genéticas y establecer su asociación con el estado oncológico de los pacientes.

- **Limitaciones de la clasificación de patogenicidad**

  - La asignación de una variante al estado patogénico no garantiza la afectación clínica del paciente, debido a la alta proporción de variantes de significancia desconocida (VUS, 43 % del total).
  - Las SNVs de tipo missense dominan en VUS, mientras que las mutaciones nonsense y frameshift concentran la mayoría de los eventos PG, lo que confirma el valor de la consecuencia molecular (molConseq) como predictor principal en los modelos XGBoost y RF.

- **Eficacia y sesgos de las técnicas de balanceo**

  - El oversampling por duplicación (OS) alcanzó un recall perfecto (FN = 0) y realzó patrones basados en k-mers, lo que sugiere una memorización de mutaciones reales.
  - Las estrategias sintéticas (SMOTE, SMOTENC, ADASYN) equilibraron sensibilidad y precisión, destacando atributos funcionales y epigenéticos.

- **Identificación de marcadores genómicos en cáncer de mama**

  - Se definió un panel robusto que engloba ARN no codificantes (Y-RNA, snoRNAs U3 y U6, SRP RNA) y genes clave de adhesión y señalización (RBFOX1, PTPRD/PTPRN2, CNTNAP2, CDH4, SDK1, CSMD1, EYS).
  - La elección de la técnica de balanceo moduló la visión biológica del modelo, destacando desde patrones mutacionales crudos hasta rutas de señalización y regulación epigenética.

- **Clustering y subestructuras latentes**
    - El optimizador conjunto de UMAP + HDBSCAN reveló ocho clústeres con distinta composición alélica y epigenética: los clústeres de alto riesgo (0 a 4 y clúster 8) se caracterizan por variantes raras y estados epigenéticos activos, mientras que los clústeres 5 a 7 agrupan polimorfismos comunes y supresores germinales como CSMD1.
    - El clustering jerárquico, tras excluir outliers, mejoró notablemente la cohesión (silhouette de 0.6477 vs. 0.4262), permitiendo discernir subgrupos que correlacionan la disrupción de supresores tumorales con modificaciones cromatínicas y splicing.

En conjunto, este trabajo demuestra cómo la conjunción de técnicas de machine learning y análisis de sistemas genómicos puede enriquecer la interpretación de variantes, aunque se subraya la importancia de complementar los hallazgos *in-silico* con estudios experimentales y clínicos para traducirlos en avances diagnósticos y terapéuticos.

5.2 TRABAJOS FUTUROS

Con base en los resultados y limitaciones identificados en este estudio, se propone las siguientes líneas de investigación para profundizar y ampliar el marco de análisis:

- *Modelos multiómicos con Mixtures of Experts y capas de fusión:* La integración simultánea de datos de distintas ómicas (p. ej., genómica, transcriptómica, epigenómica y proteómica) puede mejorar sustancialmente la capacidad de detectar patrones de patogenicidad y susceptibilidad. En trabajos futuros, se explorará el uso de arquitecturas de Mixtures of Experts que asignen de forma dinámica distintos expertos a cada tipo de dato, combinadas con capas de fusión multimodal para aprender representaciones conjuntas. Este enfoque permitirá capturar interacciones no lineales entre modalidades y mitigar el sesgo de cada fuente de información por separado.
- *Análisis de redes de pacientes como nodos:* Más allá de considerar únicamente variantes o genes como entidades en grafos, resulta prometedor modelar cada paciente como un nodo en una red, con aristas definidas por la similitud de su perfil de variantes, características clínicas o patrones de expresión. La aplicación de graph neural networks o medidas de centralidad en esta topología podrá revelar subgrupos de pacientes con mecanismos patogénicos compartidos y mejorar la estratificación clínica.

- *Incorporación de datos longitudinales y clínico-patológicos:* El uso de series temporales de medidas moleculares y registros clínicos (progressión de la enfermedad, respuesta a tratamiento, supervivencia) enriquecerá el modelado, permitiendo analizar cómo evolucionan las firmas genómicas en el tiempo. Métodos de análisis de series y modelos de supervivencia basados en aprendizaje profundo (p. ej., recurrent neural networks) podrían integrarse con los pipelines actuales.

- *Validación experimental y retroalimentación iterativa:* Para cerrar el ciclo in silicoin vitro/in vivo, se planifican ensayos funcionales dirigidos a las VUS priorizadas (p. ej., variantes en supresores de tumores o ARN no codificantes), así como experimentos CRISPR en líneas celulares. Los resultados de laboratorio se incorporarán de nuevo al modelo mediante estrategias de active learning, mejorando la precisión y la interpretabilidad de las predicciones.

- *Escalabilidad y despliegue en entornos clínicos:* Se debe evaluar la escalabilidad de los pipelines propuestos en cohortes más extensas y diversas, así como su integración en sistemas de apoyo a la decisión clínica. La optimización de cómputo (p. ej., inferencia acelerada en GPU/TPU) y la implementación de interfaces de visualización interactivas permitirán una rápida adopción en entornos hospitalarios y de investigación traslacional.

Estas líneas de trabajo no solo ampliarán el alcance del estudio, sino que también contribuirán a la construcción de herramientas más robustas y clínicamente relevantes para la interpretación de variantes genéticas en cáncer de mama y otras patologías.

4. REFERENCIAS BIBLIOGRÁFICAS

[1] N. Instituto Nacional De Cáncer, “Publicaciones: Diccionario de cáncer,” Apr. 5 2016, available: https://www.cancer.gov/espanol/publicaciones/diccionarios/diccionario-cancer.

[2] R. E. Sigala, V. Lagou, A. Shmeliov, S. Atito, S. Kouchaki, M. Awais, I. Prokopenko, A. Mahdi, and A. Demirkan, “Machine learning to advance human genome-wide association studies,” *Genes*, vol. 15, p. 34, 2024.

[3] J. R. MacDonald, R. Ziman, R. K. C. Yuen, L. Feuk, and S. W. Scherer, “The database of genomic variants: a curated collection of structural variation in the human genome,” *Nucleic Acids Research*, vol. 42, no. D1, pp. D986–D992, Jan. 2014.

[4] W. H. Organization, “Breast cancer,” Mar. 13 2024, available: https://www.who.int/es/news-room/fact-sheets/detail/breast-cancer.

[5] L. Wang *et al.*, “Functional regulations between genetic alteration-driven genes and drug target genes acting as prognostic biomarkers in breast cancer,” *Scientific Reports*, vol. 12, no. 1, pp. 1–14, 2022, available: https://doi.org/10.1038/s41598-022-13835-5.

[6] Organización Panamericana de la Salud, “Prevención: Factores de riesgo y prevención del cáncer de mama,” 2015, accessed: 2024-11-03. [Online]. Available: https://www3.paho.org/hq/dmdocuments/2015/prevencion-factores-riesgo.pdf

[7] A. C. Society, “Breast cancer facts & figures 2022-2024,” 2022, available: https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/breast-cancer-facts-and-figures/2022-2024-breast-cancer-fact-figures-acs.pdf.

[8] R. Martínez-Arriaga, *Aportaciones desde la psicooncología en el abordaje del cáncer*, 1st ed. Ciudad de México, México: Universidad Nacional Autónoma de México (UNAM), 2023.

[9] C. A. Isaza and J. Henao B., “Algunas consideraciones acerca de marcadores genéticos,” *Umbral Científico*, vol. 3, pp. 74–81, 2003, available: https://www.redalyc.org/articulo.oa?id=30400310.

[10] E. N. Kontomanolis, A. Koutras, A. Syllaios, D. Schizas, A. Mastoraki, N. Garmpis, M. Diakosavvas, K. Angelou, G. Tsatsaris, A. Pagkalos, T. Ntounis, and Z. Fasoulakis, “Role of oncogenes and tumor-suppressor genes in carcinogenesis: A review,” *Anticancer Research*, vol. 40, pp. 6009–6015, 2020.

[11] Genome.gov, “A brief guide to genomics,” available: https://www.genome.gov/about-genomics/fact-sheets/A-Brief-Guide-to-Genomics.

[12] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning with Applications in Python*. New York: Springer, 2023.

[13] T. Verplancke, S. V. Looy, D. Benoit, G. Vansteelandt, M. Depuydt, P. D. Turck, and F. D. Baets, “Support vector machine versus logistic regression modeling for prediction of hospital mortality in critically ill patients with haematological malignancies,” *BMC Medical Informatics and Decision Making*, vol. 8, no. 1, p. 56, 2008. [Online]. Available: https://doi.org/10.1186/1472-6947-8-56

[14] L. Rokach and O. Maimon, “Top-down induction of decision trees classifiersa survey,” *IEEE Transactions on Systems, Man and Cybernetics, Part C (Applications and Reviews)*, vol. 35, no. 4, pp. 476–487, Nov. 2005.

[15] M. Khalilia, S. Chakraborty, and M. Popescu, “Predicting disease risks from highly imbalanced data using random forest,” *BMC Medical Informatics and Decision Making*, vol. 11, no. 1, p. 51, 2011. [Online]. Available: https://doi.org/10.1186/1472-6947-11-51

[16] XGBoost Documentation, “Xgboost documentation,” available: https://xgboost.readthedocs.io/en/stable.

[17] H. Ahmetoglu and R. Das, “A comprehensive review on detection of cyber-attacks: Data sets, methods, challenges, and future research directions,” *Internet of Things*, vol. 20, p. 100615, Nov. 2022.

[18] X. Guan, Y. Du, R. Ma *et al.*, “Construction of the xgboost model for early lung cancer prediction based on metabolic indices,” *BMC Medical Informatics and Decision Making*, vol. 23, p. 107, 2023. [Online]. Available: https://doi.org/10.1186/s12911-023-02171-x

[19] limexp, “xgbfir: Xgboost feature interactions and importance ranking,” https://github.com/limexp/xgbfir, 2019, gitHub repository.

[20] L. McInnes, J. Healy, and S. Astels, “hdbscan: Hierarchical density based clustering,” *The Journal of Open Source Software*, vol. 2, no. 11, mar 2017. [Online]. Available: https://doi.org/10.21105%2Fjoss.00205

[21] S. Gare, S. Chel, P. D. Pantula, A. Saxena, K. Mitra, R. Sarkar, and L. Giri, “Analytics pipeline for visualization of single cell rna sequencing data from brochoaveolar fluid in covid-19 patients: Assessment of neuro fuzzy-c-means and hdbscan,” in *2022 44th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2022, pp. 1634–1637.

[22] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction," *arXiv preprint arXiv:1802.03426*, 2018, [Online]. Available: https://arxiv.org/abs/1802.03426.

[23] E. Becht, L. McInnes, J. Healy, C. Dutertre, I. W. H. Kwok, L. G. Ng, F. Ginhoux, and E. W. Newell, "Dimensionality reduction for visualizing single-cell data using UMAP," *Nature Biotechnology*, vol. 37, no. 1, pp. 38–44, 2019, [Online]. Available: https://www.nature.com/articles/nbt.4314.

[24] N. A. Giraldo *et al.*, "Spatial umap and image cytometry for topographic immuno-oncology biomarker discovery," *Cancer Immunology Research*, vol. 9, no. 11, pp. 1262–1269, Nov 2021. [Online]. Available: https://doi.org/10.1158/2326-6066.CIR-21-0015

[25] M.-M. Deza and E. Deza, *Encyclopedia of Distances*. Springer-Verlag, 2009.

[26] R. Mohammed, J. Rawashdeh, and M. Abdullah, "Machine learning with oversampling and undersampling techniques: Overview study and experimental results," in *2020 11th International Conference on Information and Communication Systems (ICICS)*. Irbid, Jordan: IEEE, 2020, pp. 243–248.

[27] I. Tomek, "A generalization of the k-nn rule," *IEEE Transactions on Systems, Man, and Cybernetics*, no. 2, pp. 121–126, 1976.

[28] G. Lemaître, F. Nogueira, and C. K. Aridas, "Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 559–563, 2017.

[29] H. Han, W.-Y. Wang, and B.-H. Mao, "Borderline-smote: A new over sampling method in imbalanced data sets learning," in *International Conference on Intelligent Computing*. Springer, 2005, pp. 878–887.

[30] K. W. Bowyer, N. V. Chawla, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *CoRR*, vol. abs/1106.1813, 2011. [Online]. Available: http://arxiv.org/abs/1106.1813

[31] imbalanced-learn contributors, *SMOTENC: Synthetic Minority Oversampling Technique for Nominal and Continuous*, 2025. [Online]. Available: https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.SMOTENC.html

[32] H. He, Y. Bai, E. A. Garcia, and S. Li, "Adasyn: Adaptive synthetic sampling approach for imbalanced learning," in *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, 2008, pp. 1322–1328.

[33] J. Snoek, H. Larochelle, and R. P. Adams, “Practical bayesian optimization of machine learning algorithms,” *Advances in Neural Information Processing Systems*, 2012.

[34] J. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl, “Algorithms for hyper-parameter optimization,” in *Advances in Neural Information Processing Systems*, 2011.

[35] E. Real, A. Aggarwal, Y. Huang, and Q. V. Le, “Large-scale evolution of image classifiers,” in *Proceedings of the 34th International Conference on Machine Learning*, 2017.

[36] L. Li, K. Jamieson, G. DeSalvo, A. Rostamizadeh, and A. Talwalkar, “Hyperband: A novel bandit-based approach to hyperparameter optimization,” in *The Journal of Machine Learning Research*, 2017.

[37] P. J. Rousseeuw, “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis,” *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.

[38] T. Caliński and J. Harabasz, “A dendrite method for cluster analysis,” *Communications in Statistics*, vol. 3, no. 1, pp. 1–27, 1974.

[39] D. L. Davies and D. W. Bouldin, “A cluster separation measure,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, no. 2, pp. 224–227, 1979.

[40] A. Hooshmand, “Naive bayesian machine learning to diagnose breast cancer,” Aug. 2020, doi:10.21203/rs.3.rs-60997/v1.

[41] Y.-J. Lee, J.-H. Park, and S.-H. Lee, “A study on the prediction of cancer using whole-genome data and deep learning,” *International Journal of Molecular Sciences*, vol. 23, Sep. 2022.

[42] M. Darmofal, S. Suman, G. Atwal, M. Toomey, J.-F. Chen, J. C. Chang, E. Vakiani, A. M. Varghese, A. B. Rema, A. Syed, N. Schultz, M. F. Berger, and Q. Morris, “Deep learning model for tumor type prediction using targeted clinical genomic sequencing data,” *Cancer discovery*, vol. 2024, 2024.

[43] Memorial Sloan Kettering Cancer Center, “Oncokb: A precision oncology knowledge base,” 2024, https://www.oncokb.org/.

[44] S. T. Sherry, M. H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and K. Sirotkin, “dbSNP: the NCBI database of genetic variation,” *Nucleic Acids Research*, vol. 29, no. 1, pp. 308–311, 2001. [Online]. Available: https://doi.org/10.1093/nar/29.1.308

[45] UCSC Genome Browser, “Cancer breast clinvar,” 2024, available: https://genome.ucsc.edu/.

[46] H. Zhou, T. Arapoglou, X. Li, Z. Li, X. Zheng, J. Moore, A. Asok, S. Kumar, E. E. Blue, S. Buyske, N. Cox, A. Felsenfeld, M. Gerstein, E. Kenny, B. Li, T. Matise, A. Philippakis, H. Rehm, H. J. Sofia, G. Snyder, Z. Weng, B. Neale, S. R. Sunyaev, and X. Lin, “Favor: Functional annotation of variants online resource and annotator for variation across the human genome,” 2022, available: https://favor.genohub.org/.

[47] J. Oliver, R. Quezada Urban, C. A. Franco Cortés, C. E. Díaz Velásquez, A. L. Montealegre Páez, R. A. Pacheco-Orozco, C. Castro Rojas, R. García-Robles, J. J. López Rivera, S. Gaitán Chaparro, A. M. Gómez, F. Suarez Obando, G. Giraldo, M. I. Maya, P. Hurtado-Villa, A. I. Sanchez, N. Serrano, A. I. Orduz Galvis, S. Aruachan, J. Nuñez Castillo, C. Frecha, C. Riggi, F. Jauk, E. M. Gómez García, C. L. Carranza, V. Zamora, G. Torres Mejía, I. Romieu, C. A. Castañeda, M. Castillo, R. Gitler, A. Antoniano, E. Rojas Jiménez, L. E. Romero Cruz, F. Vallejo Lecuona, I. Delgado Enciso, A. B. Martínez Rizo, A. Flores Carranza, V. Benites Godinez, C. F. Méndez Catalá, L. A. Herrera, Y. I. Chirino, L. I. Terrazas, S. Perdomo, and F. Vaca Paniagua, “Latin american study of hereditary breast and ovarian cancer lacam: A genomic epidemiology approach,” *Frontiers in Oncology*, vol. 9, p. 1429, Dec. 2019.

Anexo 1: Cohort Analysis, Sequencing, and Variant Processing Pipeline

To validate our approach, we analyzed a cohort of 1,098 Latin American women who met the criteria established on National Comprehensive Cancer Network (NCCN) Guidelines, version 2.2018, for Hereditary Breast and Ovarian Cancer (HBOC) syndrome. Peripheral blood sample preparation, genomic DNA extraction, library preparation, and massively parallel sequencing were performed according to the methods described in [47]. Targeted exome sequencing was then performed using a panel of 143 genes associated with susceptibility to various hereditary cancers, with a focus on clinically relevant genes such as BRCA1, BRCA2, TP53, PTEN, and PALB2. Participants are being recruited at 11 centers for the LACAM consortium, including 2 centers from Estado de Mexico and Mexico City in Mexico (Centro Oncológico Estatal de Toluca, Instituto Mexicano del Seguro Social Siglo XXI), 6 centers from different regions in Colombia (Clínica Universitaria Colombia, Hospital Universitario San Ignacio, UPB Clínica Universitaria Bolivariana, IMAT-Oncomédica S.A., Centro Médico Imbanaco, Fundación cardiovascular de Colombia), one center in Peru (Instituto Nacional de Enfermedades Neoplásicas), Argentina (Hospital Italiano de Buenos Aires) and Guatemala (Instituto para la Investigación Científica y la Educación Acerca de las Enfermedades Genéticas y Metabólicas Humanas). The LACAM protocol was approved by the Ethics Committee of each center (COE/UEI/PT/02/2018, INSP-CI:1065, FM-CIE-0409-17, UPB-2018, ONC-CEI-801-2018, INEN-18-06, HI-2730, CE/INVEGEM 1-2017, INSP-341, UEB.471-2018, ISEM 28-09-2015, CEICANCL290515-05GENCMAHER, FCV: CEI-2021-02159, CEI-406) and it is conducted in accordance with the Declaration of Helsinki.

This approach reduced sequencing complexity and cost while maximizing the yield of actionable variants. Raw sequencing data were preprocessed using TRIMMOMATIC (v0.39) to remove adapters, trim low-quality bases, and discard reads shorter than 36 bases. The trimmed reads were aligned to the GRCh38 reference genome using BWA-MEM (v0.7.17), with secondary alignments marked for downstream compatibility. Post-alignment processing included marking duplicate reads with Picards MarkDuplicates (v2.25.0) and adding read groups for sample identification. Base quality scores were recalibrated using GATKs Base Quality Score Recalibration (BQSR) tool (v4.2.0.0), leveraging known variant sites from dbSNP and the 1000 Genomes Project. Variants were called using GATKs HaplotypeCaller in GVCF mode, followed by joint genotyping across the cohort to improve accuracy. The called variants were annotated using ANNOVAR and VEP to predict functional impact and filtered based on quality, functional predictions, and population frequency (MAF <1 % in gnomAD). Finally, the filtered variants were integrated with genomic features retrieved from genes, such as transposable elements, CpG islands, and non-coding RNAs using custom

scripts and databases like UCSC GENOMEBROWSER.

Anexo 2: Búsqueda de Hiperparámetros (TPE)

Dataset de Patogenicidad

A continuación se muestran los hiperparámetros obtenidos por Hyperopt para el diseño experimental de la sección **4.4 Modelamiento** del dataset de Patogenicidad.

Modelo	Hiperparámetro	Valor	Descripción breve
XGBoost (balanceado)	colsample_bytree	0.931	Fracción de columnas seleccionadas aleatoriamente por árbol
	gamma	2.059	Mínima reducción de pérdida para realizar partición (mayor = más conservador)
	learning_rate	0.122	Tasa de aprendizaje (cuánto se ajusta el modelo en cada iteración)
	max_depth	6	Profundidad máxima de los árboles
	n_estimators	150	Número total de árboles
	reg_alpha	2.943	Regularización L1 (penaliza la complejidad del modelo)
	reg_lambda	0.055	Regularización L2
	subsample	0.789	Fracción de filas seleccionadas para cada árbol
Random Forest (balanceado)	bootstrap	True	Muestras con reemplazo en cada árbol
	max_depth	14	Profundidad máxima de los árboles
	min_samples_leaf	1	Mínimo de muestras requeridas en una hoja
	n_estimators	250	Número de árboles en el bosque
SVM (balanceado)	C	6.679	Penalización por errores. Mayor = menos tolerancia a errores
	kernel	rbf	Tipo de kernel (radial, no lineal)
	gamma	0.053	Define la influencia de un solo punto. Bajo = margen más amplio
XGBoost (desbalanceado)	colsample_bytree	0.703	Proporción de columnas seleccionadas por árbol
	gamma	0.411	Reducción mínima de pérdida para particionar
	learning_rate	0.102	Tasa de aprendizaje para cada iteración
	max_depth	6	Profundidad máxima de los árboles
	n_estimators	150	Número total de árboles generados
	reg_alpha	0.096	Término de regularización L1
	reg_lambda	0.245	Término de regularización L2
	subsample	0.829	Proporción de muestras utilizadas en cada árbol
Random Forest (desbalanceado)	bootstrap	True	Uso de muestreo con reemplazo
	max_depth	14	Profundidad máxima permitida de los árboles
	min_samples_leaf	1	Mínimo de muestras por hoja
	n_estimators	110	Número de árboles en el bosque

Bal.: Técnica de balanceo usada. **seed**: valor de random_state usado (2025).

Tabla 7: Hiperparámetros seleccionados por búsqueda bayesiana para combinaciones de modelo y técnicas de balanceo probadas para el dataset de Patogenicidad

Dataset de Variantes a partir de archivos de llamado de variantes (VCF) para pacientes con y sin cáncer de mama familiar

A continuación se muestran los hiperparámetros obtenidos por Hyperopt para el diseño experimental de la sección **4.4 Modelamiento** del dataset de Variantes a partir de archivos de llamado de variantes (VCF) para pacientes con y sin cáncer de mama familiar.

Bal.	max_depth	learning_rate	n_estimators	min_child_weight	subsample	colsample_bytree
Default	10	0.2725	250	1	0.9632	0.7133
Undersampling	7	0.0115	50	3	0.7622	0.7764
Oversampling	10	0.2142	300	1	0.6867	0.7560
SMOTE	10	0.2404	300	4	0.9558	0.6934
SMOTENC	10	0.1754	300	2	0.9287	0.8814
ADASYN	9	0.2417	250	2	0.6988	0.7923

Bal.: Técnica de balanceo usada. **seed**: valor de random_state usado (2025).

Tabla 8: Hiperparámetros seleccionados por búsqueda bayesiana para cada técnica de balanceo para el dataset de Variantes a partir de archivos de llamado de variantes (VCF) para pacientes con y sin cáncer de mama familiar.