CLASIFICACIÓN DE PACIENTES CON LEISHMANIASIS BASADO EN MUTACIONES GENÉTICAS POR POLIMORFISMO DE NUCLEÓTIDO ÚNICO (SNP) USANDO TÉCNICAS DE MACHINE LEARNING.

CARLOS ANDRÉS GÓMEZ VASCO

Nota de Aceptación

Certificamos que el presente Trabajo de Grado Satisface, en alcances y calidad, todos los requisitos que demanda un Trabajo de Grado de Maestría.

Director

Jurado

Jurado

Aprobado en cumplimiento de los requisitos exigidos por la Pontificia Universidad Javeriana Cali, para optar el título de Magister en Ciencia de Datos.

HERNÁN CAMILO ROCHA NIÑO Ph. D.
Decano Facultad de Ingeniería y Ciencias

JUAN CARLOS MARTÍNEZ ARIAS
Director Posgrados de Ingeniería y Ciencias

Santiago de Cali, 15, 08, 2023

Acta de Correcciones al Documento de Trabajo de Grado

Santiago de Cali, 03 de agosto de 2023

Autor: Carlos Andrés Gómez Vasco

Título del Trabajo de Grado: "Clasificación de pacientes con Leishmaniasis basado en mutaciones genéticas por Polimorfismo de Nucleótido Único (SNP) usando Técnicas de Machine Learning"

Director: Gloria Inés Álvarez Vargas

Como indica el artículo 2.13 de las Directrices para Trabajo de Grado de Maestría, he verificado que el estudiante indicado arriba ha implementado todas las correcciones que los Jurados del Proyecto de Trabajo de Grado definieron que se efectuaran, como consta en el Acta de Evaluación correspondiente.

Firma del Director del Trabajo de Grado

Santiago de Cali, 27 de junio del 2023

Doctora
Gloría Inés Alvarez V.
Directora Maestría en Ciencia de Datos
Facultad de Ingeniería y Ciencias
Pontificia Universidad Javeriana de Cali

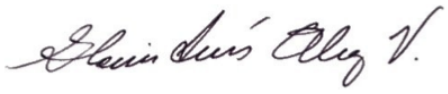
Asunto: Presentación para evaluación del proyecto aplicado

Cordial Saludo,

Con el fin de cumplir con los requisitos exigidos por la Universidad para optar por el título de Magíster en Ciencia de Datos, nos permitimos presentar a su consideración el proyecto denominado "Clasificación de pacientes con leishmaniasis basado en mutaciones genéticas por polimorfismo de nucleótido único (SNP) usando técnicas de Machine Learning", el cual fue realizado por el estudiante Carlos Andrés Gómez Vasco con código 80.897.240 perteneciente a la Maestría en Ciencia de Datos, bajo la dirección de la Dra. Gloria Inés Álvarez Vargas y codirección del Dr. Diego Luis Linares Ospina.

El suscrito director del Proyecto Aplicado autoriza para que se proceda a hacer la evaluación de este proyecto, toda vez que ha revisado cuidadosamente el documento y avala que ya se encuentra listo para ser presentado y sustentado oficialmente.

Atentamente,



Carlos Andrés Gómez Vasco

C.C. 80897240 de Bogotá

Gloria Inés Álvarez Vargas

C.C. 30306105 de Manizales

Documentación anexa:
Resumen del Proyecto Aplicado en formato digital (máximo 1 página).
Una copia digital (PDF) del documento del proyecto aplicado

FICHA RESUMEN

PROYECTO APLICADO – MAESTRÍA EN CIENCIA DE DATOS

TÍTULO: Clasificación de pacientes con leishmaniasis basado en mutaciones genéticas por polimorfismo de nucleótido único (SNP) usando técnicas de Machine Learning.

1. ÁREA DE TRABAJO: Machine Learning
2. TIPO DE PROYECTO: Aplicado
3. ESTUDIANTE (S): Carlos Andrés Gómez Vasco
4. CORREO ELECTRÓNICO: cagv@javerianacali.edu.co
5. DIRECCIÓN Y TELEFONO: Carrera 72 Bis 56 D 08 Sur (Barrio Olarte - Bogotá) - 321 345 43 57
6. DIRECTOR: Gloria Inés Álvarez Vargas
7. VINCULACIÓN DEL DIRECTOR: Profesor de tiempo completo Facultad Ingeniería y Ciencias.
8. CORREO ELECTRÓNICO DEL DIRECTOR: galvarez@javerianacali.edu.co
9. CO-DIRECTOR: Diego Luis Linares Ospina
10. GRUPO O EMPRESA QUE LO AVALA: Centro Internacional de Entrenamiento e Investigaciones Médicas CIDEIM.
11. OTROS GRUPOS O EMPRESAS: Grupo de Investigación Destino. Pontificia Universidad Javeriana, Cali.
12. PALABRAS CLAVE: Leishmaniasis, SNP, Machine Learning, ADN, Genómica, Ciencia, Datos.
13. FECHA DE INICIO: 27 de Julio de 2022.
14. FECHA DE FINALIZACIÓN: 27 de Junio de 2023.
15. RESUMEN: La leishmaniasis es una enfermedad tropical transmitida mediante la picadura de insectos que son los vectores de la enfermedad, se considera una endemia en más de 88 países de diferentes geografías. Las tasas reales de incidencia son sustancialmente altas y con una alta prevalencia en países de América Latina. Aunque existen diferentes tratamientos terapéuticos, son muy complicados para los pacientes

y suelen ser bastante tóxicos para otros órganos del cuerpo, y, en general tienen altos índices porcentuales de fallo, es decir, cumplido de tratamiento los pacientes no se recuperan. Actualmente no existe una herramienta clínica que le permita a un médico tratante determinar la probabilidad a priori de que un tratamiento sea efectivo, por el contrario, de manera indiscriminada se aplica a los pacientes las terapias bajo la premisa del ensayo y error. En este proyecto aplicado, se realiza un estudio basado en mutaciones genéticas producidas por polimorfismo de nucleótido único (SNP) a un conjunto de setenta y dos (72) pacientes tratados con las técnicas terapéuticas existentes, a estos pacientes se les realizó una secuenciación genética consiguiendo 618872 SNPs para cada uno y la información clínica del grupo étnico, así como la respuesta al tratamiento después de aplicado, etiquetado como cura o falla. Esta información es suficiente para generar un dataset que fue analizado mediante GWAS (Estudio de asociación de genoma completo) consiguiendo tres datset denominados COMPLETO, AFRO DESCENDIENTES y NO-AFRO DESCENDIENTES con 41, 14 y 36 SNPs correspondientemente. Mediante técnicas de reducción de dimensionalidad, como el análisis de componentes principales (PCA), eliminación recursiva de características y regresión LASSO se reduce el número de variables a aquellas mutaciones genéticas más relevantes para la respuesta inmune al tratamiento consiguiendo 69 subconjuntos de características. Mediante técnicas de aprendizaje automático se construyen 483 clasificadores basados en algoritmos de Regresión Lineal (**RL**), Stochastic Gradient Descent (**SGD**), Support Vector Machine (**SVM**), Decision Tree (**DT**), Random Forest (**RF**), Boosting (**BT**) y Gradient Boosting (**GB**) de los 69 subconjuntos, para clasificar con precisión las mutaciones genéticas relacionadas con la respuesta inmune al tratamiento terapéutico contra la leishmaniasis. Se utilizaron métricas de evaluación, como accuracy, precision, recall y F1 score para medir el rendimiento de los clasificadores. Estas métricas proporcionaron una visión detallada de la capacidad de los modelos para identificar correctamente las mutaciones relevantes. Después de la evaluación inicial de los 683 experimentos, se realizó la optimización de los hiperparámetros de los modelos mediante una búsqueda por cuadricula explorando diferentes combinaciones y configuraciones, lo que permitió refinar los modelos, y nuevamente estimar su desempeño permitiendo evaluar y comparar los resultados antes y después de la optimización, confirmando la mejora significativa en la capacidad de los clasificadores para identificar con precisión las mutaciones genéticas relacionadas con la respuesta inmune al tratamiento terapéutico contra la leishmaniasis. Al final, se consiguió una selección de 22 SNPs ubicados en genes con funciones biológicas altamente relacionadas con movimiento, transcripción, estructura y transporte celular, así, como el transporte de metales, respuesta inmune y cicatrización. Evidenciando que las técnicas aplicadas son eficientes en la identificación de biomarcadores asociados con la respuesta al tratamiento contra la leishmaniasis.

Pontificia Universidad
JAVERIANA
Cali

CLASIFICACIÓN DE PACIENTES CON LEISHMANIASIS BASADO EN MUTACIONES GENÉTICAS POR POLIMORFISMO DE NUCLEÓTIDO ÚNICO (SNP) USANDO TÉCNICAS DE MACHINE LEARNING

Carlos Andrés Gómez Vasco

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

Director (a)
PhD Gloria Inés Alvarez Vargas

Codirector (a)
PhD Diego Luis Linares

FACULTAD DE INGENIERÍA Y CIENCIAS
MAESTRÍA EN CIENCIA DE DATOS
SANTIAGO DE CALI, JUNIO 29 DE 2023

Pontificia Universidad JAVERIANA Cali

TABLA DE CONTENIDO

Pontificia Universidad JAVERIANA Cali

LISTA DE FIGURAS

Pontificia Universidad JAVERIANA Cali

LISTA DE TABLAS

INTRODUCCIÓN

Las enfermedades tropicales suelen afectar mayoritariamente a las poblaciones más pobres y marginadas constituyendo un grave problema de salud pública en los países en vías de desarrollo. El grueso de estas enfermedades son transmitidas por insectos vectores, que se encuentran con frecuencia en regiones tropicales y subtropicales, donde el clima facilita su establecimiento y proliferación.

La leishmanianis es una de esas enfermedades tropicales que aqueja a una considerable parte de esta población en diferentes regiones del mundo. El problema principal de la lucha clínica contra la leishmaniasis pude resumirse en: *"es peor el remedio que la enfermedad"*, por lo menos en el 30 % ó 40 % de los pacientes el tratamiento que es altamente tóxico no surge efecto alguno. Sin embargo, es la única herramienta existente para combatir la enfermedad.

El objetivo central de este trabajo es la construcción de un clasificador mediante aprendizaje automático de un dataset con mutaciones genéticas generadas por polimorfismos de un solo nucleótido o SNP del ingles ***Single Nucleotide Polymorphism*** y que se pronuncia **-snip-**, así, poder clasificar pacientes a partir de la estimación de la probabilidad de falla o éxito a priori del tratamiento terapéutico tradicional[1].

Este documento ilustra inicialmente la definición del problema dentro del contexto de la enfermedad de la leishmaniasis. En un segundo capitulo están los objetivos del trabajo. En el tercer apartado, una revisión de los conceptos teóricos y el estado del arte. En el capitulo cuatro la preparación de los datos y la aplicación de las primeras técnicas especializadas para la selección de características usando un GWAS[2], en el capitulo cinco la aplicación de algoritmos de aprendizaje de maquina para la reducción de la dimensionalidad y selección de características.

En los capítulos seis, siete y ocho se muestra la construcción de los diversos datasets, el entrenamiento y aplicación de siete algoritmos a cada set de datos, su respectiva evaluación y posterior ajuste de hiperparametros para la optimización de modelos, obteniendo los subconjuntos de características con las mejores métricas de evaluación, así, como la revisión de los SNPs que se perfilan como posibles biomarcadores de la respuesta al tratamiento contra la leishmaniasis. Por último, el análisis de resultados, las conclusiones y las perspectivas respecto a trabajos futuros que se deseen realizar en el área.

[1] Con tradicional se hace referencia al tratamiento que posee actualmente el sistema sanitario Colombiano.

[2] Estudio de asociación de genoma completo.

1. DEFINICIÓN DEL PROBLEMA

La leishmaniasis es una enfermedad transmitida por la picadura de un vector que se encuentre infectado, generalmente un mosquito (ver figura 1a). Son múltiples las formas y afecciones que aparecen como consecuencia de la adquisición de la enfermedad. Las dos mas comunes son; las lesiones cutáneas caracterizadas por afectar la piel y las membranas mucosas formándose llagas en la piel que comienzan en la zona donde se produce la picadura (figura 1b) y se va expandiendo, incluso a otras regiones del cuerpo como se observa en la figura 1c.

(a) Mosquito flebotomus. Vector transmisor de la enfermedad

(b) Lesiones cutáneas iniciales.

(c) Difusión de las lesiones cutáneas a lo largo del rostro. Tomada de [1]

Figura 1: La leishmaniasis se produce por el parásito que recibe el nombre de Leishmania. Se transmite desde los animales al ser humano a través de la picadura de la hembra de unos insectos llamados flebotomus. Se estima que en todo el mundo hay entre 12 y 14 millones de personas afectadas y cada año se diagnostican 1,5-2 millones de nuevos casos. Tomado de [2].

La segunda se denomina leishmaniasis visceral, ésta se caracteriza por afectar órganos internos como el bazo, el hígado e incluso la médula ósea. Son múltiples y variados los síntomas de la enfermedad[3]. La enfermedad actualmente es catalogada como una endemia y esta presente en regiones parcializadas de al menos 88 países [3]. Latinoamerica es una de las regiones más afectadas por la enfermedad a nivel mundial dado que 17 de los 18 países presentan la enfermedad.

El informe presentado por el Ministerio Nacional de Salud -MNS- para la OPS/OMS muestra un total de 6362 nuevos casos en el periodo trianual 2016-2018. Es decir, por cada

[3]Dificultad para respirar, llagas en la piel, congestión, dificultad para deglutir. Úlceras y desgaste en la boca, la lengua, las encías, los labios, la nariz, pérdida de peso, aumento de tamaño del bazo y el hígado.

100.000 habitantes 26,2 habitantes contrajeron la enfermedad[4] [4]. Cifras que categorizan la transmisión en el país como intensa[5]. En el mismo informe se presentan cifras poco alentadoras respecto a la asistencia que se les ofrecen a estos pacientes por parte de la institucionalidad y del sistema de salud, a pesar que el 100 % de los casos de la infección cutánea fueron diagnosticados por pruebas de laboratorio, no existen datos sobre el porcentaje de curados después de realizar algún tratamiento terapéutico. Y peor aún, no se reportan datos ni información sobre la detección clínica ni tratamiento de los 16 casos del tipo visceral.

1.1. Planteamiento del problema

La leishmaniasis no es una enfermedad letal, sin embargo reduce considerablemente la calidad de vida. Adicionalmente, dada las características necesarias para que exista el vector transmisor de la enfermedad, suele afectar a personas que tienen difíciles condiciones socio-económicas.

En Colombia el tratamiento consiste en proporcionar medicinas que contienen antimonio, un tipo de metal o potentes antibióticos, estando disponible; Antimoniato de Meglumina, Isetionato de Pentamidina y Anfotericina B Liposomal[5]. El primer procedimiento terapéutico contra la enfermedad son las inyecciones de antimonio, aplicadas diariamente durante veinte días, implicando que el paciente debe tener acceso al medicamento y a un tercero que lo suministre, es decir, tener acceso a servicios de salud, lo que resulta en la mayoría de los casos poco probable, dadas las ubicaciones geográficas donde se presenta la enfermedad. Por otra parte, el porcentaje de fallo en el tratamiento es de entre 30 % y 40 % y su alta toxicidad genera riesgos para el hígado y el corazón.

El segundo medicamento usado obedece a la miltefosina, suministrado de forma oral durante veintiocho días, no genera riesgos hepático ni al corazón, pero si presenta toxicidad para las vías gástricas y altos riesgos para mujeres gestantes. Este procedimiento posee una probabilidad de falla de entre un 10 % y 15 % para niños y un 10 % y 40 % en adultos[5].

Con todo lo anterior el panorama se resume en tratamientos altamente tóxicos, dolorosos y con porcentajes muy altos de falla para una enfermedad no letal. Muchos pacientes sometidos a estos tratamientos no logran completarlos por no soportar las peripecias que conlleva para su cotidianidad, por otra parte un alto porcentaje de los que lo terminan, no consiguen erradicar o curar la enfermedad y si exponer diferentes órganos a altas dosis

[4]El 98.6 % de los nuevos casos fueron leishmaniasis cutánea mientras que el 1.4 % fueron mucosas. Por otra parte, en el mismo periodo ocurrieron 16 nuevos casos (2.65 personas por cada 100.000 habitantes) para leishmaniasis visceral

[5]En una escala que se define por categorías de: bajo, medio, alto, intenso y muy intenso.

de contaminación que podrían tener graves efectos secundarios.

1.2. Formulación del problema

Someter a todos los pacientes de manera generalizada a los diferentes tratamientos, es llevar a este grupo poblacional a una ruleta rusa, que resulta mal para un porcentaje considerable de los mismos. La necesidad impetuosa es conseguir y/o clarificar estrategias de estratificación de la población de pacientes, que puedan ser métricas fiables para cuantificar la probabilidad de fallas o éxitos a priori del tratamiento terapéutico, de esa forma mejorar el accionar científico del médico tratante. Y es preciso aquí, donde la ciencia de datos entra en acción obteniendo relevancia en la búsqueda de soluciones a este problema.

Desde la biología y la bioquímica se han realizado intentos estadísticos descriptivos estudiando diferentes dataset sobre información biológica de pacientes con leishmaniasis identificando marcadores en cadenas de ARN, proteínicas e incluso llegando a rastrear metabolitos que puedan facilitar la identificación de características de clasificación antes mencionadas.

Aún inexplorado se encuentra la influencia de las mutaciones genéticas presentes en pacientes con éxito o fracaso en el tratamiento. Las mutaciones suelen estudiarse mediante los **SNPs**. Son múltiples las explicaciones que se encuentran en la literatura sobre lo que es un SNP, no todas simples ni aptas para publico no especializado. Para fines prácticos trataremos una explicación gráfica que permita acercarnos a la esencia de su significado.

La estructura del ADN se encuentra ubicada en el interior del núcleo de una célula como se observa en la parte izquierda de la figura 2, formando los cromosomas (estructura en forma de X de color azul en la figura 2) que a su vez contienen proteínas llamadas histonas, unidas al ADN. Las dos cadenas enroscadas que forman un espiral parecido a una escalera de caracol se llama hélice (parte derecha de la figura 2) y están compuestas por cuatro elementos básicos llamados nucleótidos: adenina (A), timina (T), guanina (G) y citosina (C). Estos nucleótidos se unen entre sí (A con T y G con C) mediante enlaces químicos formando pares de bases que conectan las dos cadenas de ADN. Entonces, los genes son pequeñas piezas de ADN que tienen información genética específica.

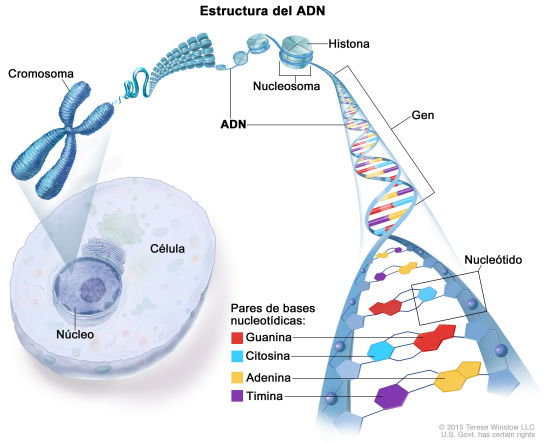


Figura 2: Estructura del ADN. Tomado de [6].

Ahora bien, de la experiencia cotidiana basta con observar un grupo de personas y ver que cada uno presenta diferencias en su aspecto físico. Diferencias que están relacionadas con la composición individual del ADN de cada individuo. Un SNP es un cambio genético que solo afecta a un nucleótido dentro de la secuencia del ADN de una persona, en la figura 3 se observan dos tramos de la misma cadena de ADN, 1 parte superior y 2 parte inferior, en 2 existe un cambio genético que solo afecta a un nucleótido [6]. Estas mutaciones pueden provocar cambios sutiles en el funcionamiento de una proteína, sin embargo, en algunas ocasiones pueden dar lugar a cambios en la forma en la que la célula se comporta [7], por ello, se suelen utilizar los SNPs para estudiar como un individuo puede predisponerse a sufrir una determinada enfermedad, o influir en su respuesta a medicamentos.

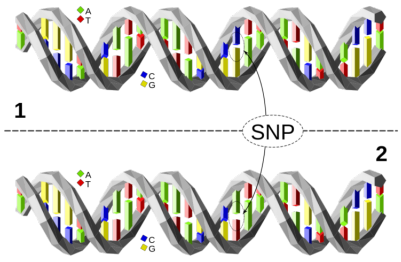


Figura 3: Un polimorfismo de nucleótido único (SNP) es un cambio genético que solo afecta a un nucleótido dentro de la secuencia del ADN de una persona. En promedio hay un SNP con variación por cada 100-300 bases (o nucleótidos) de entre las 3 millones de bases del genoma humano. Tomado de [7].

[6]Una mutación genética no solamente es producto de los SNPs, pero si representan un 90 % de las variaciones genéticas humanas

El CIDEIM [7] ha desarrollado investigaciones previas sobre las principales causas de la falla del tratamiento en pacientes con leishmaniasis, su preocupación en esencia ha sido establecer a lo largo de las diferentes escalas biológicas (celular, cadenas proteínicas y escalas estructurales de ARN y ADN) donde puede estar la razón más probable para que los tratamientos existentes contra la enfermedad fallen. Actualmente posee un conjunto de datos con información bastante robusta sobre mutaciones genéticas producidas por los SNPs en un grupo de 72 pacientes y un aproximado de 618,872 SNPs. Con el fin de establecer un estudio de descubrimiento, se pretende aplicar técnicas de asociación estadística buscando una reducción de dimensionalidad inicial, que permita avanzar en una segunda etapa de reducción dimensional bajo técnicas de aprendizaje automático, y por último, construir un clasificador mediante aprendizaje supervisado. Todo lo anterior, enmarcado en aportar nuevo conocimiento para estudios futuros de exploración experimental sobre la influencia de las mutaciones en la probabilidad a priori de éxito o fracaso terapéutico.

Para acotar el problema de estudio y bajo el panorama ya definido, este proyecto aplicado se enfoca en la evaluación de posibles respuestas a:

¿Es factible la definición de criterios mediante técnicas de aprendizaje automático para la reducción de la dimensionalidad y clasificación de mutaciones genéticas generadas por SNPs que estén altamente asociadas con la probabilidad de fracaso o éxito terapéutico en la leishmanisis?

[7] Centro Internacional de Entrenamiento e Investigaciones Médicas

2. OBJETIVOS DEL PROYECTO

2.1. Objetivo General

Clasificar mediante técnicas de aprendizaje automático mutaciones genéticas producidas por SNPs que tengan una alta probabilidad de estar relacionadas con la respuesta inmune al tratamiento terapéutico contra la leishmaniasis.

2.2. Objetivos Específicos

- Preparar y procesar los datos del dataset seleccionado para su posterior análisis.
- Realizar un análisis descriptivo exhaustivo de la información contenida en el dataset, enfocándose en comprender su naturaleza y características.
- Aplicar técnicas de reducción de dimensionalidad para identificar las mutaciones de mayor interés en los datos.
- Evaluar experimentalmente las mutaciones o conjuntos de mutaciones identificadas, empleándolas como base para la construcción de un clasificador eficiente y preciso.
- Analizar y comparar los resultados obtenidos por el clasificador desarrollado, considerando su rendimiento y su capacidad para identificar patrones y tendencias relevantes.

3. MARCO TEÓRICO Y ANTECEDENTES

3.1. Etapas de un proyecto de Ciencia de Datos

Un proyecto en ciencia de datos tiene un ciclo de vida, donde no necesariamente los algoritmos de aprendizaje de maquina (Machine Learning) son la única etapa. Es muy importante conocer a profundidad los algoritmos, las técnicas de modelamiento y tener una visión holística del proyecto, guardando las proporciones dada la amplia y vasta aplicación de la ciencia de datos.

Las etapas que se encuentran involucradas son; entendimiento del problema, la localización y análisis de la información, preprocesamiento y calidad de los datos, la modelización y el despliegue final[8]. Dentro de la primera etapa, realizar la limpieza y el preprocesamiento de los datos se resume en un ciclo como el mostrado en la figura 4 donde continuamente se definen las estrategias y formas en que se van a importar y exportar los datos, el manejo de datos no disponibles, la redundancia, estandarización, normalización, estimación del tiempo de la información y sus posibles cambios [9].

Figura 4: Ciclo de limpieza de los datos. Tomado de [10]

La siguiente etapa, puede involucrar en algunos casos una reducción de características denominada reducción de dimensionalidad; usando métodos de transformación de variables para encontrar características significativas que faciliten la representación mas adecuada de los datos para los objetivos del proyecto. En la etapa posterior se debe entrenar los datos después de seleccionar los algoritmos adecuados, para este proyecto aplicado se usan algoritmos de clasificación [9]. Luego, la etapa de búsqueda de hiperparámetros que se refiere a la tarea de encontrar los valores óptimos para los hiper-parámetros del modelo [11], La importancia de esta etapa radica en que una elección incorrecta de estos puede llevar a un modelo de aprendizaje automático poco eficiente o incluso no funcional. Por último, un entrenamiento con aquellos mejores modelos.

3.2. La maldición de la dimensionalidad

Muchas situaciones en el mundo del aprendizaje automático están basadas en miles e inclusive millones de variables por cada set de datos de entrenamiento. Lo que hace que entrenar modelos sea una tarea demasiado lenta, adicionalmente, se genera un aumento de la complejidad y costo computacional impidiendo encontrar una analítica predictiva. Al aumentar el número de características o dimensiones en un conjunto de datos, la cantidad de datos para obtener una precisión adecuada también aumenta, y lo hace de manera exponencial. Ello implica la necesidad de más y más datos para obtener un rendimiento óptimo, a este inconveniente operativo se le denomina la **maldición de la dimensionalidad** [12].

Existen formas de reducir el número de variables, cambiando de un set de datos imposible de entrenar a uno con capacidad de ser entrenado. Es frecuente que existan variables que se correlacionan, entonces, se facilita la disminución de dimensiones. Sin embargo, el beneficio de reducir dimensiones no es magia pura, se debe pagar un costo por la reducción que facilite el entrenamiento que es la perdida de información relevante. Por ello se debe hacer una reducción de la dimensionalidad solo cuando sea estrictamente necesario.

Algunas de la técnicas más utilizadas para realizar reducción de la dimensionalidad están clasificadas en métodos de filtro y en métodos de envoltura [10].

3.2.1. Métodos de Filtro

Los métodos de filtro funcionan mediante la selección de un subconjunto de características relevantes del conjunto completo. La selección se logra mediante la aplicación de una métrica de relevancia que evalúa la importancia de cada característica en relación a la variable objetivo. Las características que se consideran más relevantes según la métrica elegida se mantienen y las características menos relevantes se eliminan.

Un par de ejemplos de métodos de filtro comunes son:

- **Correlación:** se calcula la correlación entre cada característica y la variable objetivo y se seleccionan las características que tienen la correlación más alta.
- **Análisis de componentes principales (PCA):** se utiliza para transformar el conjunto de datos en un conjunto de características no correlacionadas que explican la mayor cantidad posible de la varianza en los datos.

Los métodos de filtro son relativamente simples y rápidos de aplicar, pero pueden tener limitaciones en la capacidad de seleccionar las características más importantes en un conjunto de datos complejo. En algunos casos, se pueden utilizar junto con otros métodos de reducción de dimensionalidad más avanzados para mejorar su efectividad.

3.2.2. Métodos de Envoltura

Los métodos de envoltura a diferencia de los métodos de filtro, seleccionan características basándose en una métrica de relevancia, los métodos de envoltura buscan encontrar el subconjunto óptimo de características que maximizan el rendimiento de un modelo de aprendizaje automático. Funcionan aplicando iterativamente un algoritmo de aprendizaje automático a diferentes subconjuntos de características y seleccionando el subconjunto que produce el mejor rendimiento en términos de la métrica de evaluación.

Un par de ejemplos de métodos de envoltura son:

- **Recursive Feature Elimination (RFE):** utiliza un modelo de aprendizaje automático para seleccionar iterativamente el subconjunto óptimo de características, comenzando con todas las características y eliminando las menos importantes en cada iteración.
- **Regresión LASSO:** Least Absolute Shrinkage and Selection Operator es un método de regresión lineal que utiliza la regularización para reducir la complejidad del modelo y evitar el sobreajuste. La regularización (L1) implica agregar una penalización a la función de costo que resulta en la eliminación de características irrelevantes y la reducción de características redundantes. El método reduce los coeficientes de regresión de las características menos importantes a cero, eliminando esas características del modelo. Esto ayuda a seleccionar las características más importantes y reducir la complejidad del modelo.

Los métodos de envoltura pueden producir mejores resultados que los métodos de filtro porque tienen en cuenta las interacciones entre las características y pueden seleccionar un conjunto de características óptimo para un modelo de aprendizaje automático específico.

3.3. Estudio de asociación de genoma completo -GWAS-

En general dos personas del mismo sexo comparten alrededor del $99,9\,\%$ de su secuencia de ADN, por otra parte $0{,}1\,\%$ restante es la parte que posee variaciones genéticas que influyen en el fenotipo de los individuos. La identificación de las variantes que contribuyen al riesgo de una enfermedad común ofrece una de las mejores oportunidades para la comprensión de las causas de dicha enfermedad. Las variantes del genoma se pueden clasificar en; variantes comunes y variantes raras. Las primeras son polimorfismos definidos como aquella variante genética cuya frecuencia del alelo menor (MAF) es de por lo menos un $1\,\%$ en la población, las segundas se presentan en menos del $1\,\%$. De igual forma, las variantes en el genoma se pueden clasificar de acuerdo con su composición de nucleótidos en variantes de un solo nucleótido SNPs y variantes estructurales.

Un estudio de asociación del genoma completo conocido como GWAS (*"Genome-Wide Association Study"* por sus siglas en ingles) se trata de un enfoque para buscar relaciones entre variantes genéticas (como SNPs) y enfermedades o características fenotípicas en poblaciones humanas[13]. Analizando miles de variantes genéticas simultáneamente dentro de un gran conjunto de individuos, en busqueda de identificar variantes genéticas asociadas con un riesgo aumentado de una enfermedad o característica específica. Estos estudios son utilizados para entender la base genética de una variedad de trastornos y enfermedades, algunos ejemplos que ya se han realizado son estudios de diabetes [14], cáncer [15] y enfermedades cardíacas.

Un GWAS realiza principalmente pruebas estadísticas de asociación para buscar relaciones entre variantes genéticas y características fenotípicas. Dentro de estas estadísticas se encuentran comúnmente el test de chi-cuadrado para comparar la frecuencia de una variante genética en individuos con y sin la característica fenotípica en cuestión. Este test se basa en la hipótesis de que la variante genética y la característica fenotípica están relacionadas entre sí, se realiza mediante la construcción de una tabla de contingencia que compara la frecuencia observada de una variante genética en individuos con y sin la característica fenotípica con la frecuencia esperada, si no hubiera relación entre la variante genética y la característica fenotípica [16]. Definiendo como hipótesis nula (H0) la NO relación entre la variante genética y la característica fenotípica, es decir, la frecuencia observada es igual a la frecuencia esperada. Si el p_{value} obtenido es menor que el nivel de significancia establecido (generalmente se utiliza un valor de 0.05), se rechaza la hipótesis nula y se concluye que hay una relación estadísticamente significativa entre la variante genética y la característica fenotípica [17].

3.4. Algoritmos de Clasificación

- ***Regresión Logística:*** Es un algoritmo utilizado para predecir la probabilidad de una variable de resultado categórico binario, como por ejemplo, sí/no, éxito/fracaso, verdadero/falso, entre otras. El algoritmo funciona mediante la estimación de los coeficientes de regresión que relacionan las variables predictoras con la variable de resultado categórica. Estos coeficientes se utilizan para calcular la probabilidad estimada de que la variable de resultado categórica sea igual a 1 o 0 dadas las variables predictoras. Con la función logística de la ecuación 1 se transforma el resultado de la ecuación lineal (resultado de la suma ponderada de las variables predictoras y los coeficientes) en una probabilidad entre 0 y 1.

$$p = \frac{1}{1 + \exp(-z)} \tag{1}$$

En la ecuación 1 p es la probabilidad estimada de que la variable de resultado sea igual a 1, exp es la constante matemática de euler con valor aproximado a 2,718, y z es el resultado de la ecuación lineal. Para estimar los coeficientes de regresión en la regresión logística, se utiliza el algoritmo de optimización, basado en el método de máxima verosimilitud, que busca maximizar la probabilidad conjunta de observar los resultados de la muestra dada una configuración de los coeficientes [18].

- ***Descenso de Gradiente Estocástico (SGD):*** Este algoritmo funciona mediante una optimización iterativa para encontrar el mínimo de una función de costo. En cada iteración se ajustan parámetros del modelo en función del gradiente de la función de costo encontrando parámetros del modelo que la minimizan. Este método es muy eficiente dado que en cada iteración, se toma una muestra aleatoria de los datos de entrenamiento, se calcula el gradiente de la función de costo con respecto a los parámetros del modelo utilizando solo para esa muestra, se actualizan los parámetros del modelo en la dirección opuesta al gradiente multiplicando el gradiente por una tasa de aprendizaje que determina la longitud del paso que se da en la dirección opuesta al gradiente [19].

- ***Kernel Support Vector Machine (SVM) Classifier:*** Generan clasificaciones de datos no lineales a partir de la transformación de los datos de entrada a otros espacios de mayores dimensiones en busca encontrar aquella curva que sea capaz de separar y clasificar los datos de entrenamiento garantizando que la separación entre ésta y ciertas observaciones del conjunto de entrenamiento sea la mayor posible. Al mapear los datos a un espacio dimensional superior, es más probable que se puedan separar las diferentes clases con un hiperplano lineal. Una de sus mas grandes ventajas esta en la capacidad de manejar datos con alta dimensionalidad, es decir, funciona bastante bien en problemas que involucran muchos atributos, sin embargo, es computacionalmente costoso [20].

- ***Decision Tree Classifier:*** Su funcionamiento se basa en la construcción de reglas lógicas (divisiones de los datos entre rangos o condiciones) a partir de los datos de entrada. El entrenamiento de los árboles de decisión se centra principalmente en la maximización de la ganancia de información al momento de realizar las reglas lógicas que forman el árbol, cada nodo del árbol representa una pregunta sobre una característica particular del conjunto de datos, así, y tomando la respuesta a cada pregunta, se va ramificando el árbol en diversas direcciones hasta conseguir una hoja del árbol que representa la predicción de la clase. La pregunta más importante se ubica en el nodo raíz, desde aquí se deben dividir los datos, es clave elegir la pregunta que haga la división inicial de los datos. Para ello se utilizan criterios de división, como la ganancia de información, el índice Gini o la reducción de impureza. Luego de elegir la pregunta, se crean dos nuevos nodos con las posibles respuestas a la

pregunta inicial. Se repite el proceso para cada uno de los nuevos nodos, se divide el conjunto de datos, nuevamente en subconjuntos más reducidos para cada nueva iteración, hasta que se alcanzan las hojas del árbol [21].

- ***Ramdon Forest:*** Este algoritmo se basa en múltiples árboles de decisión para realizar sus predicciones, en principio, cada uno de los árboles en el bosque se construye de forma independiente, seleccionado una muestra aleatoria de los datos de entrenamiento y de características para dividir cada nodo del árbol. Se repite este procedimiento para construir varios arboles, luego para hacer una predicción toma un nuevo conjunto de datos y los hace pasar a través de cada uno de los árboles de decisión del bosque, por último, para la predicción final deja la media o la moda de las predicciones de cada árbol. La mas grande de las bondades y ventajas del método es el uso de múltiples árboles de decisión y la aleatoriedad en la selección de muestras y características, evitando el sobreajuste (overfitting) [22].
- ***Boosting:*** Con este método se realiza una combinación de varios modelos de aprendizaje débiles para crear un modelo más fuerte y preciso. Se realizan múltiples e iterativas repeticiones. Primero, se entrena un conjunto de datos con un modelo de aprendizaje débil, por ejemplo, un árbol de decisión. Después, se evalúa el desempeño del modelo en el conjunto de entrenamiento y se otorga peso según su la métrica obtenida. Así, los datos mal clasificados tienen un peso mayor, mientras los que se han clasificados adecuadamente tienen un menor peso. Para una siguiente iteración, se entrena otro modelo de aprendizaje débil centrando una mayor atención a los datos mal clasificados en la iteración inmediatamente anterior, para ello se ajustan los pesos de los datos antes de entrenar el modelo, dándoles un peso mayor en el conjunto de entrenamiento para la siguiente iteración, ello significa que el nuevo entrenamiento estará más enfocado en corregir los errores cometidos en la iteración anterior. Este procedimiento se repite varias veces, con cada modelo de aprendizaje débil tratando de corregir los errores cometidos por los modelos anteriores. Por último, se combinan todos los modelos de aprendizaje débiles en un modelo final más fuerte y preciso, mediante la asignación de pesos a cada modelo, de acuerdo a su desempeño en las iteraciones previas. Este procedimiento entrega claramente mejoras de rendimiento, flexibilización, control de datos desequilibrados, reducción de sesgo y una considerable reducción del ruido [23].
- ***Gradient Boost:*** Al igual que el algoritmo anterior, este utiliza un conjunto de modelos de aprendizaje débiles, usualmente árboles de decisión para construir un modelo más fuerte y preciso. Sin embargo, ahora utiliza el gradiente descendente con el fin de minimizar la función de pérdida, que mide la diferencia entre la predicción del modelo y el valor real observado. Ajustando un modelo inicial, logra tener predicciones utilizando datos de entrenamiento. A continuación, se calcula el gradiente de la función de pérdida con respecto a la predicción del modelo inicial. Y

así, indicar la dirección y la magnitud del cambio que se debe realizar a la predicción inicial para minimizar la función de pérdida. Nuevamente, se entrena un nuevo modelo débil para predecir la diferencia entre la predicción inicial y el valor real. Este nuevo modelo se agrega al modelo inicial, y se obtiene una nueva predicción para el objetivo. Este proceso se repite varias veces, y en cada iteración se ajusta un nuevo modelo débil para predecir la diferencia entre la predicción actual y el valor real, y se suma al modelo anterior. A medida que se agregan más y más modelos débiles, se consigue un aumento considerable en la precisión, dado que se corrigen las deficiencias de modelos anteriores [24].

3.4.1. Métricas de Evaluación

Las métricas de evaluación de un modelo de aprendizaje automático se usan para establecer la calidad de un entrenamiento realizado. Las métricas seleccionadas deben reflejar el objetivo del modelo y las necesidades del problema que se esta estudiando. Para definir estas métricas de evaluación se utiliza la matriz de confusión, esta herramienta permite analizar la clasificación para evaluar el rendimiento de un modelo. Muestra la frecuencia con la que el modelo clasifica correctamente o incorrectamente las instancias de cada clase.

La matriz de confusión tiene una estructura de tabla con dos dimensiones como se observa en la figura 5, una que representa las clases reales y otra que representa las clases predichas por el modelo. La matriz tiene cuatro casillas que representan la frecuencia con la que se clasifica correctamente o incorrectamente cada clase. Las cuatro casillas se denominan Verdadero Positivo (VP), Falso Positivo (FP), Verdadero Negativo (VN) y Falso Negativo (FN) [25].

Figura 5: Estructura de una matriz de confusión. (Tomado de [26])

La forma de leer la matriz de confusión es la siguiente; Verdadero Positivo (VP) es la cantidad de instancias que pertenecen a la clase en cuestión y que fueron clasificadas correctamente por el modelo. Falso Positivo (FP) es la cantidad de instancias que no pertenecen a la clase en cuestión pero que fueron clasificadas por el modelo como si lo fuesen.

Verdadero Negativo (VN) es la cantidad de instancias que no pertenecen a la clase en cuestión y que fueron clasificadas correctamente por el modelo, por último, Falso Negativo (FN) es la cantidad de instancias que pertenecen a la clase en cuestión pero que fueron clasificadas incorrectamente por el modelo.

La lectura de la matriz de confusión puede ayudar a entender cómo está funcionando el modelo de clasificación e identificar posibles errores [27]. Algunas de las métricas comúnmente utilizadas en la evaluación de algoritmos de aprendizaje automático son [25]:

- **Precisión (Accuracy):** Es la proporción de predicciones correctas sobre el número total de predicciones. Se calcula como:

$$Precision = \frac{VP + VN}{VP + FP + VN + FN} \tag{2}$$

- **Sensibilidad (Recall o Sensitivity):** Es la proporción de verdaderos positivos en relación con todos los valores verdaderos en los datos de prueba. Se calcula como:

$$Recall = \frac{VP}{VP + FN} \tag{3}$$

- **Especificidad (Specificity):** Es la proporción de verdaderos negativos en relación con todos los valores falsos en los datos de prueba. Se calcula como:

$$Specificity = \frac{VN}{VN + FP} \tag{4}$$

- **Valor F1 (F1 Score):** Es la media armónica de la precisión y la sensibilidad, proporciona una medida de la precisión general del modelo. Se calcula como:

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{5}$$

- **Área bajo la curva ROC (AUC-ROC):** Es una medida de la capacidad del modelo para distinguir entre las clases, representa la probabilidad de que el modelo clasifique una instancia positiva más alta que una instancia negativa. Se calcula utilizando la curva ROC (Receiver Operating Characteristic).

La elección de la métrica de evaluación adecuada depende del problema y los requisitos propios que se enmarquen en cada contexto.

3.5. Validación Cruzada

La validación cruzada es una técnica estadística utilizada en el aprendizaje automático para evaluar la capacidad predictiva de un modelo. El método tradicional implica dividir un conjunto de datos en dos partes: un conjunto de entrenamiento y un conjunto de prueba. El modelo se entrena en el conjunto de entrenamiento y luego se evalúa su capacidad de generalización en el conjunto de prueba. Sin embargo, la validación cruzada va más allá de este enfoque simple y divide el conjunto de datos en múltiples subconjuntos de nominados k-folds en lugar de solo dos. El modelo se entrena en k-1 de estos subconjuntos y se evalúa en el subconjunto restante. El proceso se repite k veces, utilizando cada subconjunto una vez como conjunto de prueba, en la figura 6 se observa una ilustración del método.

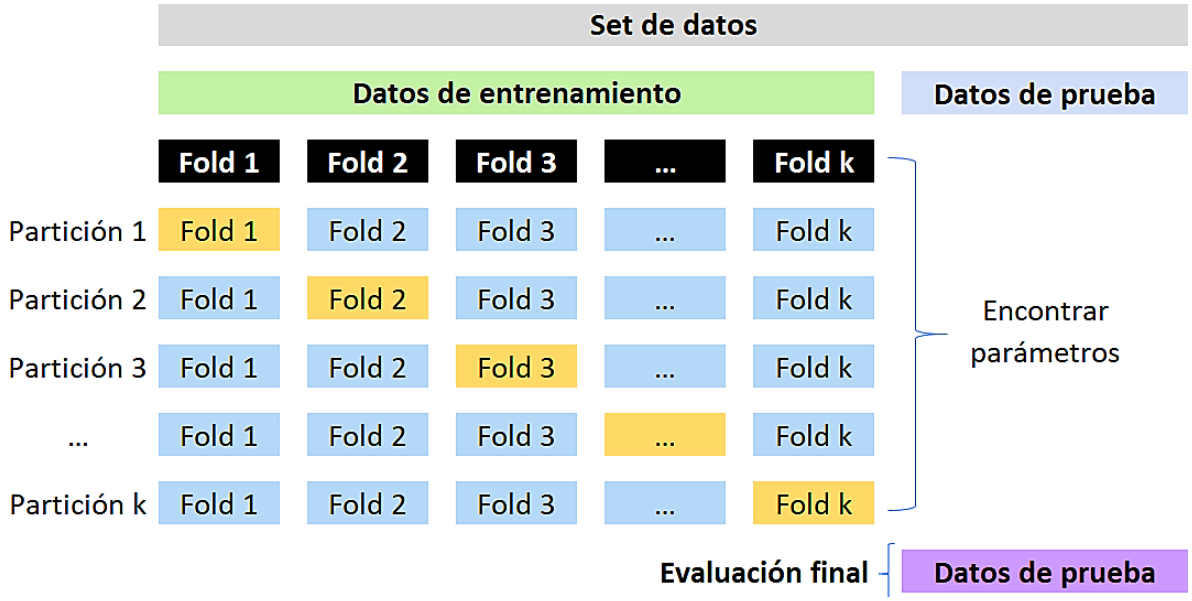


Figura 6: Método de validación cruzada. (Tomado de [28].)

Al final de cada iteración, se calcula una métrica de rendimiento y se promedian para obtener una medida general del rendimiento del modelo. Esto proporciona una evaluación más precisa del modelo, ya que utiliza todo el conjunto de datos para el entrenamiento y la evaluación. Esta técnica es útil porque cada vez que se entrena el modelo, se utilizan diferentes datos de entrenamiento y prueba. De esta manera, el modelo puede aprender patrones más generales en lugar de memorizar datos específicos. Si el modelo puede rendir bien en diferentes conjuntos de prueba, es probable que esté generalizando bien y no esté sobreajustando.

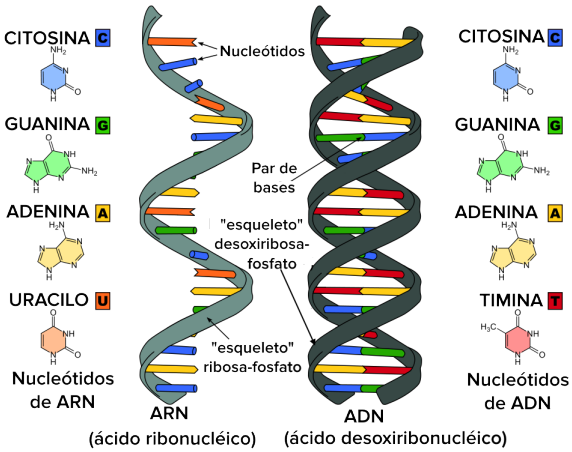
3.6. Búsqueda y ajuste de hiperparámetros

El ajuste de hiperparámetros en el aprendizaje automático busca los parámetros que no son aprendidos directamente por el modelo durante el proceso de entrenamiento, sino que son ajustados manualmente antes de que se inicie el proceso de entrenamiento. El objetivo del ajuste de hiperparámetros es encontrar la combinación óptima de valores de hiperparámetros que maximicen el rendimiento del modelo en el conjunto de datos de prueba o validación. Encontrar los mejores parámetros para cada modelo son una tarea de exploración sistemática en diferentes valores de hiperparámetros y la comparación de su rendimiento en una métrica de evaluación. Existen varias técnicas para ajustar los hiperparámetros de un modelo, entre ellas se encuentran:

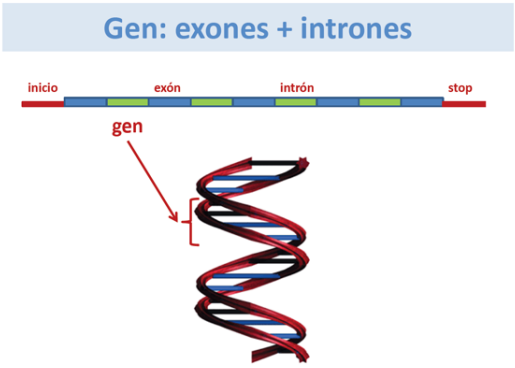
- **Búsqueda de cuadrícula:** La búsqueda de cuadrícula es un enfoque simple pero efectivo, implica definir un conjunto de valores para cada hiperparámetro que se desea ajustar, luego probar todas las combinaciones posibles de valores para determinar la combinación óptima. Es un enfoque costoso computacionalmente [29].
- **Búsqueda aleatoria:** La búsqueda aleatoria implica muestrear valores de los hiperparámetros de manera aleatoria dentro de un rango especificado. Este enfoque puede ser más eficiente que la búsqueda de cuadrícula, ya que se pueden probar más combinaciones de hiperparámetros en menos tiempo [29].
- **Optimización bayesiana:** La optimización bayesiana utiliza una función de evaluación para evaluar el rendimiento del modelo para diferentes combinaciones de hiperparámetros. Al evalúar más combinaciones, se utiliza un modelo probabilístico para aprender sobre la función de evaluación y determinar qué combinaciones probar a continuación. Es un enfoque más eficiente que la búsqueda aleatoria y la búsqueda de cuadrícula porque puede aprender de las evaluaciones previas y enfocar la búsqueda en las áreas más prometedoras del espacio de hiperparámetros [30].

3.7. La estructura del ADN y la generación de las proteínas

El funcionamiento de la estructura del ADN es altamente complejo, y su explicación sobrepasa los limites de esté trabajo. Sin embargo, un intento por dar una explicación sencilla puede ser el siguiente. Existen cinco bases químicas nitrogenadas la adenina (A), guanina (G), timina (T), la citosina (C) y el uracilo (U), que son compuestos orgánicos cíclicos, con dos o más átomos de nitrógeno, su representación química se puede observar en los extremos izquierdo y derecho de la figura 7a. Cuando estas bases se juntan en pares, generalmente una A se junta a una T y una C adhiere a una G, conforman un "escalon"(cuyo nombre técnico es **nucleótido**), de la gran escalera con forma de espiral que se observa en la figura 7a.



(a) Estructura química del ADN y ARN.



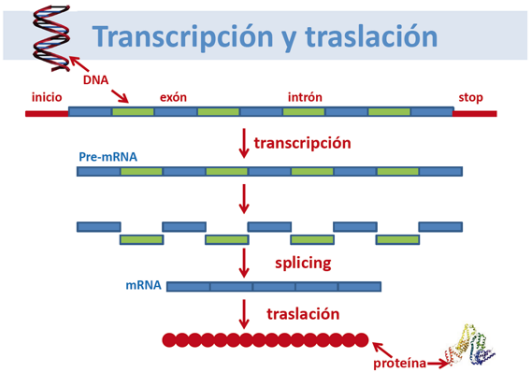
(b) Estructura elemental de un Gen.

Figura 7: Proceso de generación de proteínas y sus funcionalidades.

En la estructura del ADN los lados son cadenas de azúcares y fosfatos conectadas por "escalones", es decir, las bases nitrogenadas. Dichos "escalones"se van combinando unos con otros, dando lugar a lo que sería un código. El código genético de cada persona. Ahora bien, una larga cadena de nucleótidos consecutivos se denomina Gen, que es la unidad de material genético que consigue entregar información necesaria para llegar a sintetizar una proteína. El Gen se caracteriza por tener exones e intrones. Los exones son las regiones codificantes, es decir, aquellas que proporcionar información para la síntesis de una proteína, por otra parte, están los intrones, regiones no codificantes y que están intercaladas entre los exones a lo largo del Gen como se observa en la figura 7b.

La información contenida por los nucleótidos en las regiones codificantes (exones) se extrae del ADN mediante dos fases: la transcripción y la traslación. La transcripción es el proceso por el cual las secuencias de los exones del DNA se transcriben y forman el ARN mensajero y la traslación se realiza en los ribosomas celulares, capaces de descifrar el código de la secuencia que trae en ARN mensajero y traducirla a la cadena de aminoácidos de una determinada proteína para su formación, este procedimiento que es mucho mas complejo de lo que aquí pueda detallarse se esquematiza en la figura 8a.

Las proteínas tienen diferentes funciones en el organismo: algunas son enzimas, transportadoras, hormonas, proteínas estructurales o de membrana, anticuerpos, entre otras funciones que se pueden ver el la figura 8b.



(a) Esquema de los procesos de transcripción y traslación.

(b) Funciones de las proteínas.

Figura 8: Proceso de generación de proteínas y sus funcionalidades.

3.8. Antecedentes

3.8.1. Locales

A nivel Local el CIDEIM ha trabajado durante varios años en investigación biomédica en diferentes frentes contra la leishmaniasis, su trabajo arduo ha sido reconocido al punto que es Centro Colaborador de la Organización Mundial de la Salud en el campo de la leishmaniasis [31]. La PhD. Maria Adelaida Gómez en convenio con el grupo de investigación DESTINO de la Universidad Javeriana Cali, viene desarrollando acercamientos significativos de la aplicación de la ciencia de datos en las causas puntuales que pudiesen ser fuentes concluyentes del fallo del tratamiento terapéutico aplicado a pacientes con esta enfermedad, consiguiendo ligeros avances en la identificación de cadenas proteínicas con alta probabilidad de incidencia en este ítem.

3.8.2. Nacionales

A nivel nacional ya han existido diferentes trabajos con respecto a la aplicación de la ciencia de datos y el machine learning en la construcción de modelos predictivos que puedan determinar la ocurrencia de leishmaniasis cutánea en el país, a partir de datos ambientales y socio-económicas. En estos trabajos se han caracterizados poblaciones que se encuentran en regiones donde se es altamente susceptible de contraer la enfermedad. En [32] se recolectaron datos, se realizó el preprocesamiento, se transformaron e implementaron técnicas de reducción de dimensionalidad. Luego se aplicaron técnicas de aprendizaje de máquina de clasificación[8] y de regresión[9]. Generando así el modelo con una buena métrica

[8] naive bayes, redes neuronales (perceptrón multicapa), árboles de decisión y redes bayesianas
[9] redes neuronales y XGBoost

que facilita la predicción basándose en la población, precipitación, temperatura, índice de vegetación mejorado.

Otros estudios realizados por el grupo de investigación en bioprospección de la Facultad de Ingeniería de la Universidad de la Sabana han abonado esfuerzos mucho mas cercanos a los propósitos de la presente propuesta. Dado los inconvenientes que producen los fármacos antileishmaniacos tradicionales para los pacientes se han buscado tratamientos de compuestos naturales y sintéticos contra promastigotes y/o amastigotes. Los resultados de tales estudios han demostrado por separado éxitos importantes y un potencial razonable, sin embargo, aun no existe una visión holística de estos resultados, y es precisamente allí donde la ciencia de datos entró en escena realizando una revisión de los datos construida durante treinta y dos años para el parásito leishmania panamensis, causante de leishmaniasis cutánea mas incidente en el territorio nacional. Usando un análisis basados en huellas dactilares y bajo el uso de algoritmos de aprendizaje automático convencionales y algunos otros métodos de agrupamiento, se encontraron algunas características moleculares que determinan la aplicación de esta investigación, así como una simplificación y agrupamiento de aciertos isofuncionales [33].

3.8.3. Internacionales

En el ámbito internacional, se han desarrollado actividades de investigación con intereses específicos de usar el modelado de machine learning en el estudio, diagnóstico y evaluación de tratamientos terapéuticos contra la leishmaniasis. Grupos de investigación de centros médicos Iraníes en la ciudad de Shiraz, desarrollaron un sistema de diagnóstico para leishmania por medio de un algoritmo de machine learning aplicado a imágenes tomadas mediante microscopia. Usando el algoritmo de Viola-Jones[10] se realizó la extracción de características, la creación de imágenes integrales y la clasificación mediante adaBoost permite seleccionar las características discriminadas y entrenar al clasificador. Obteniendo un sistema preciso, rápido, fácil de usar y rentable que mostró un 65 % de recuerdo y 50 % de precisión en la detección de macrófagos infectados con el leishma-parásitonia. Además, estos números fueron 52 % y 71 %, respectivamente, relacionados con amastigotes fuera de los macrófagos [34].

En el 2018 el grupo de procesamiento de imágenes y vídeo de la Universidad de Catalunya, también reporto avances significativos en el desarrollo de un procedimiento de detección haciendo uso del modelo U-NET, que es una red neuronal dedicada a tareas de visión artificial, concretamente a problemas de segmentación semántica como alternativa a la ob-

[10]Es un método de detección de objetos que se usa ampliamente en la detección de caras en imágenes y vídeo. Basado en la comparación entre las intensidades luminosas de regiones rectangulares de las imágenes.

servación manual de los parásitos de la leishmania, se entrenaron los datos considerando un desequilibrio de clases entre las regiones de fondo y las regiones parásitas. Encontrando en los mapas de segmentación de imágenes de prueba, regiones que estaban bastante cerca para una fácil definición, no sólo para las células y su núcleo, sino también para sus tres formas de parásitos. Los resultados cuantitativos de diferentes métricas mostraron resultados muy prometedores, podrían mejorar considerablemente, utilizando bases de datos más grandes, siendo el desequilibrio de las clases el inconvenientes a solucionar [35].

Entre 2019 y 2020 se encuentran algunos avances respecto a la estudios con enfoques en algoritmos de aprendizaje automático que pretenden establecer predicción de la susceptibilidad de un paciente a la adquisición de la enfermedad. Bhunia, Kesari, Jeyaram, Kumar, Das y Tandon muestran diferentes técnicas de aprendizaje automático; árboles de decisión, redes neuronales artificiales, máquinas de vectores de soporte, algunos algoritmos de regresión y ensamblado de modelos [36]. También se puede observar en [37] y [38] la identificación de polimorfismos de un solo nucleótido (SNP) asociados con la susceptibilidad a la leishmaniasis visceral, con datos obtenidos de una cohorte de pacientes en India y utilizando una combinación de análisis de asociación de genoma completo (GWAS) y técnicas de aprendizaje automático; árboles de decisión, regresión logística y ensamblado de modelos identificaron SNPs asociados con la susceptibilidad a la leishmaniasis visceral.

En otras referencias como [39] se usaron regresión logística y análisis de componentes principales, para identificar los SNPs más importantes y desarrollar modelos predictivos de la susceptibilidad a la enfermedad. En todos estos estudios los modelos de aprendizaje automático pudieron predecir con precisión la susceptibilidad a la enfermedad en un conjunto de datos de validación no visto antes en el entrenamiento de los modelos. Y, en cada uno de los estudios anteriores se logro concluir que los modelos de aprendizaje automático son herramientas útiles para predecir la susceptibilidad a la leishmaniasis y que su uso puede mejorar la eficacia de la detección temprana y el control de la enfermedad.

Por otra parte en [40] incluyeron información sobre la edad, el género, el estado socio económico y los antecedentes de enfermedades previas de los participantes demostrando que mediante un modelo de regresión logística y un modelo de árbol de decisión la información genética esta íntimamente relacionada con la susceptibilidad a contraer la enfermedad. Los resultados mostraron que el modelo de árbol de decisión tuvo un mejor rendimiento con una precisión del 82 % en la predicción de la susceptibilidad.

4. PREPARACIÓN DE LOS DATOS

4.1. Descripción de los datos

La información proporcionada por el CIDEIM[41] tiene originalmente dos bases de datos de interés. La primera obedece a una secuenciación genómica de 618,872 SNPs en 72 pacientes en dos grupos étnicos diferentes; afro descendientes y no-afro descendientes, como se observa en la figura 9, adicional al nombre del SNP y los fenotipos para cada paciente, se tiene información del cromosoma, la posición y el strand. Cada SNP esta presentado por dos alelos, una combinación de adenina (A), timina (T), citosina (C) o guanina (G).

name	chr	pos	strand	X101	X102	X103	X104
rs1000000	12	126890980	-	CC	TC	CC	CC
rs1000002	3	183635768	-	GG	AG	AG	GG
rs10000023	4	95733906	+	TT	TG	TT	TG
rs1000003	3	98342907	-	AA	AG	AG	AA
rs10000030	4	103374154	-	GG	AG	GG	GG
rs10000037	4	38924330	-	GG	GG	GG	GG
rs10000041	4	165621955	-	TG	TT	TG	TT

Figura 9: Base datos de genotipo para 618,872 SNPs en 72 pacientes. Las primera columna corresponde al nombre del SNP, chr es el cromosoma al que pertenece, pos la posición, strand la cadena de azucares y fosfatos + o -, y las columnas X101 a X171 corresponden al par de bases de cada paciente. Datos obtenidos de [41].

La segunda base de datos se observa en la figura 10, esta contiene información clínica de los 72 pacientes, allí está contenida la información del fenotipo que nos interesa (la respuesta obtenida después de realizar el tratamiento terapéutico aplicado a los pacientes, cura ó falla), la raza a la que pertenecen (afro descendientes o no-afro descendientes (otro en la figura) entre otras de menor interés para los fines de este trabajo como el sexo, peso, la edad, el grupo clínico, número de lesiones, etc.

Para realizar de manera adecuada el análisis de los datos de interés para este proyecto, se generó una tabla que contiene la estructura mostrada en la figura 11, con un total de 72 registros que obedecen a los diferentes pacientes y 618,874 características (columnas), de las cuales 618,872 obedecen a los SNPs de cada paciente, una columna adicional para al grupo étnico (afro descendientes ó no-afro descendiente), y, una última columna correspondiente al fenotipo que en este caso corresponde a la respuesta al tratamiento, cura etiquetada con un 1 ó falla etiquetada con un 0.

ID MACRG	id	afro	respuestatto
101	MT1010	Otro	FALLA
102	MT2013	Afrocolombi	FALLA
103	MT2019	Afrocolombi	FALLA
104	MT2004	Afrocolombi	FALLA
105	MT1009	Afrocolombi	CURA
106	MT1013	Afrocolombi	CURA
107	MT1018	Afrocolombi	CURA
108	MT1029	Afrocolombi	CURA

Figura 10: Base datos clínica de los 72 pacientes. Las primeras dos columnas son identificadores de pacientes, la tercera es la raza y la última la respuesta al tratamiento. Datos obtenidos de [41].

	name	rs10127827	rs1017484	rs10177008	...	rs996204	Grupo_Etnico	Respuesta
0		-0.40161	1.347738	-0.664765	...	0.377964	-1.183216	0
1		-0.40161	-0.857651	1.313048	...	-2.645751	0.845154	0
2		-0.40161	1.347738	-0.664765	...	0.377964	0.845154	0
3		-0.40161	-0.857651	1.313048	...	0.377964	0.845154	0
4		-0.40161	-0.857651	-0.664765	...	0.377964	0.845154	1
...		...	...	...	...	...	...	...
67		-0.40161	-0.857651	1.313048	...	-2.645751	-1.183216	1
68		2.48998	1.347738	1.313048	...	0.377964	-1.183216	1
69		-0.40161	1.347738	1.313048	...	0.377964	-1.183216	1
70		-0.40161	-0.348715	-0.664765	...	0.377964	0.845154	1
71		-0.40161	-0.348715	-0.664765	...	0.377964	0.845154	1

72 rows × 618874 columns

Figura 11: Conjunto de datos del proyecto. Los SNPs y la raza se estandarizaron y normalizaron en el dataframe.

Adicionalmente, hay que tener en cuenta que en los estudios genéticos, el grupo étnico es un factor importante porque a menudo presentan diferencias genéticas únicas que pueden ser relevantes e influir en la predisposición a ciertas enfermedades, la eficacia de

ciertos medicamentos y la respuesta a tratamientos específicos [42]. Además, la diversidad étnica es importante para garantizar que los resultados sean generalizables a toda la población. Si un estudio solo se enfoca en una población específica, los resultados pueden no ser aplicables a otras poblaciones debido a las diferencias genéticas y ambientales. Tambien es importante tener en cuenta que el concepto de grupo étnico no se refiere solo a la ascendencia genética, sino también incluye factores culturales, lingüísticos y/o geográficos.

En este proyecto aplicado se realizan estudios con el dataset de la figura 11 denominado de aquí en adelante como **COMPLETO** y en donde el grupo étnico es una característica predictora mas. En procedimientos posteriores se realiza una división en dos dataset, a partir del grupo étnico; **AFRO DESCENDIENTES** Y **NO-AFRO DESCENDIENTES**. En la tabla 1 se puede observar una descripción de los datasets, mostrando en detalle la cantidad de registros que tienen, así como la distribución de curas y fallas en la respuesta al tratamiento.

Tabla 1: Descripción de los dataset conseguidos.

DATASET	DESCRIPCIÓN
COMPLETO	Tiene 72 pacientes. 42 afrodescendientes. 30 no – afrodescendientes. 47 curas. 25 fallas.
AFRO DESCENDIENTES	Tiene 42 pacientes. 25 curas. 17 fallas.
NO-AFRO DESCENDIENTES	Tiene 30 pacientes. 22 curas. 8 fallas.

4.2. Análisis estadístico de los datos

Inicialmente se realiza una extracción de características iniciales usando el software especializado PLINK [43]. Este software posee un conjunto de herramientas de análisis genético de asociación, funciona desde la línea de comandos y fue desarrollado por Shaun Purcell [11] en el laboratorio de genética poblacional del Hospital Universitario de Harvard.

PLINK se desarrolló en lenguaje C/C++, es de código abierto y se ejecuta en diferentes sistemas operativos (Windows, Linux y macOS). El software está diseñado de forma flexible para realizar una amplia gama de análisis genéticos básicos a gran escala. Permite la manipulación y análisis de datos en formato de archivo de texto plano (como archivos

[11] Epidemiólogo genético y genetista estadístico británico. Desarrolló el software de genética PLINK durante su trabajo postdoctoral con Mark Daly en el Instituto Whitehead.

de genotipo y archivos de información de posición de los SNPs) proporcionando una variedad de herramientas de asociación genética; regresión, prueba de tendencia e inferencia estadística [44].

El análisis genético de asociación en línea de comando que facilita PLINK tiene múltiples funciones, algunas se detallan a continuación:

- **Manipulación de datos genéticos:** permite la conversión de datos genéticos en diferentes formatos, la eliminación de individuos o marcas, la selección de individuos o marcas, la creación de archivos de submuestra y la creación de archivos de parejas.
- **Asociación genética:** proporciona herramientas para realizar análisis de asociación de un solo marcador y de múltiples marcadores, así como para realizar análisis de asociación de parejas.
- **Regresión:** permite realizar regresiones lineales y logísticas, tanto con covariables como sin ellas.
- **Pruebas estadísticas:** ofrece una variedad de pruebas estadísticas para analizar los datos genéticos, como pruebas de tendencia, pruebas de chi-cuadrado y pruebas de exactitud de dos etapas.
- **Inferencia estadística:** proporciona herramientas para la inferencia estadística, como la inferencia de efectos aleatorios y la inferencia bayesiana.

4.2.1. ¿Cómo hacer un GWAS con PLINK?

Para realizar la reducción de la dimensionalidad inicial de los datos genómicos, se realizó un estudio GWAS explicado en la sección 3.3 usando las herramientas proporcionadas por PLINK en su versión 1.07.

Primero es necesario preparar dos archivos de ingreso a la herramienta. El primero, con extensión **.ped** corresponde a un archivo de texto plano en columnas delimitado por espacios en blanco (espacio ó tabulación) con la estructura que se muestra en la tabla 2. En la última columna se agrega toda la información asociada con los alelos para cada SNP y es la única columna donde se pueden encontrar datos faltantes, presto a que en todas las otras columnas se tiene el 100 % de la información. Para el manejo de dichos datos se etiqueta los datos faltantes con el marcador "- -". Cuando PLINK encuentra este marcador, reconoce que se trata de un dato faltante y lo trata adecuadamente durante el análisis; la imputación de datos o la exclusión si se presentan demasiados datos faltantes.

El segundo archivo debe tener extensión **.map**, igualmente es un archivo de texto plano en columnas delimitado por espacios en blanco con la estructura que se muestra en la tabla 3. En este archivo no tenemos datos faltantes, toda la información es conocida.

Una vez construidos los dos archivos, se inicia la ejecución de las diferentes herramientas. Estas herramientas se ejecutan desde una terminal de comandos Unix o DOS, en sistemas de 32 o 64 bits[12]. En la tabla 4 se observan algunos de los comando para la ejecución de las herramientas.

Tabla 2: Estructura del archivo .ped para GWAS con PLINK.

Columna	Solicitada por PLINK	Información Consignada
Family ID	Un identificación familiar	IDMARGEN (Ver figura 10)
Individuo ID	Un identificación individual	IDMARGEN (Ver figura 10)
Paternal ID	Identificador paterno.	Sin información se completó con 0.
Maternal ID	Identificador materno.	Sin información se completó con 0.
Sexo	1=Mas - 2=Fem - otro=desconocido	Se completó a c/u según sexo.
Fenotipo	Fenotipo de interés	0=Falla 1=Cura
Genotipo	Información genotípica a analizar	Se registro alelos de cada SNP´s

Tabla 3: Estructura del archivo .map para GWAS con PLINK.

Columna	Solicitada por PLINK	Información Consignada
Cromosoma	1-22, X, Y ó 0 si no se sabe	Se completo para cada uno.
rs	identificador de cada SNP	Según información. (ver figura 9)
Genet distance	Distancia genética	Sin información se completó con 0.
Position	Posición del SNP	Según información. (ver figura 9).

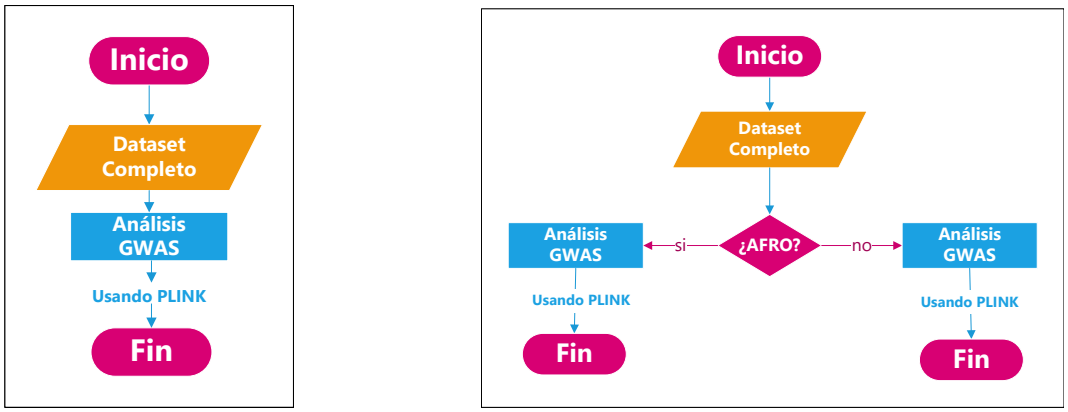
Tabla 4: Comandos de instrucción específica en PLINK para operar sobre los datos genómicos.

Instrucción Especifica	Acción
–make-bed	Construye archivos binarios para el análisis PLINK.
–check-sex	verificar coincidencia génetica con el sexo.
–missing	encontrar demasiada información perdida en los genotipos.
–hardy	que no cumplan con el Hardy-Weinberg Equilibrium.
–freq	revisión de la frecuencia de aparición alelica.
–maf 0.1	una aparición de frecuencia alelica inferior al 10 %.
–het	un estudio de heterocigosidad.
–assoc	realización de la asociación.

[12]En la pagina oficial del proyecto están las múltiples versiones e incluso el código de desarrollo disponible para su descarga. Sitio Web oficial del proyecto PLINK: `https://zzz.bwh.harvard.edu/plink/download.shtml`

4.3. Procedimiento

El procedimiento seguido para aplicar GWAS a los datos disponibles de este proyecto se detalla en la figura 12. En la figura 12a se muestra el proceso en el que se utiliza el dataset **COMPLETO**, se aplica el procedimiento detallado en la sección anterior usando PLINK.



(a) Izquierda (b) Derecha

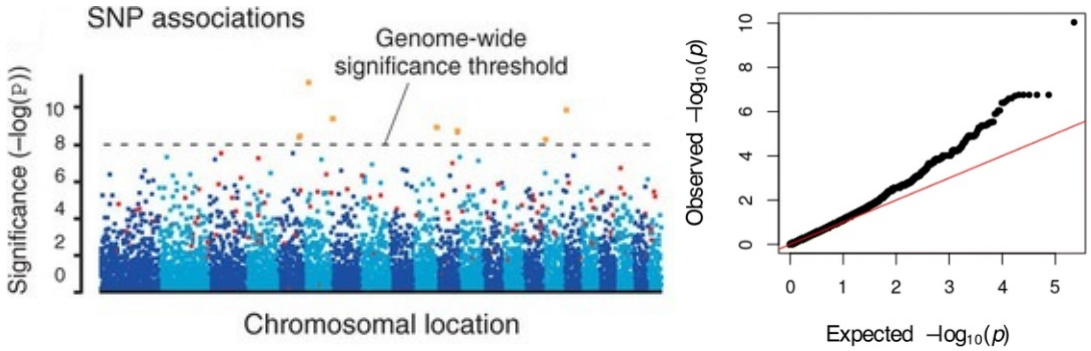
Figura 12: Procedimientos para la realización GWAS. **Izquierda:** Dataset COMPLETO. **Derecha:** Dataset dividido en AFRO DESCENDIENTES y NO-AFRO DESCENDIENTES

En un segundo procedimiento se toma el dataset **COMPLETO** y se subdivide en dos dataset **AFRO DESCENDIENTES** y **NO-AFRO DESCENDIENTES**, a partir de la característica Grupo_Etnico y posteriormente se aplica el GWAS a cada uno como se ilustra en la figura 12b. Una vez terminado el filtro de los SNPs en cada uno de los procedimientos detallados en la figura 12 se realiza el estudio de asociación (instrucción **--assoc**) explicado en detalle en la sección 3.3.

El resultado de la asociación se representa de manera gráfica con un Manhattan Plot[13], como el que se muestra en la figura 13a. Cada punto en el gráfico representa un SNP en el genoma, el valor de la coordenada X indica la posición en el cromosoma correspondiente. El valor en la coordenada Y=$-log_{10}(p)$ (Debido a que las asociaciones más fuertes tienen los valores p_{value} más pequeños, se extraen los logaritmos multiplicados por menos uno para obtener los valores mas altos en el eje vertical) indica la significancia estadística de la asociación entre cada SNP y el fenotipo, cuánto mas arriba en el eje Y se encuentre el

[13]El nombre "Manhattan plot" proviene de la forma que toma el gráfico al representar varias asociaciones en él. Los puntos que representan las asociaciones se agrupan en"rascacielos" que parecen los edificios de un horizonte urbano, como el de Manhattan en Nueva York.

SNP más significativo será su asociación con el fenotipo. Es así como en la figura 13a los SNPs representados en color amarillo son los que presentan una mayor asociación, en este ejemplo el p_{value} tomado como limite es igual a 10^{-8}, un valor común en este tipo de estudios de asociación. Esta representación es una herramienta visual y útil para interpretar la información de GWAS e identificar las regiones del genoma que están significativamente asociadas con la respuesta al tratamiento terapéutico para los fines de este trabajo.



(a) Izquierda (b) Derecha

Figura 13: Procedimientos para la realización GWAS. **Izquierda:** Ejemplo de un Manhattan Plot. **Derecha:** Ejemplo de un qqline. (Tomado de [45].)

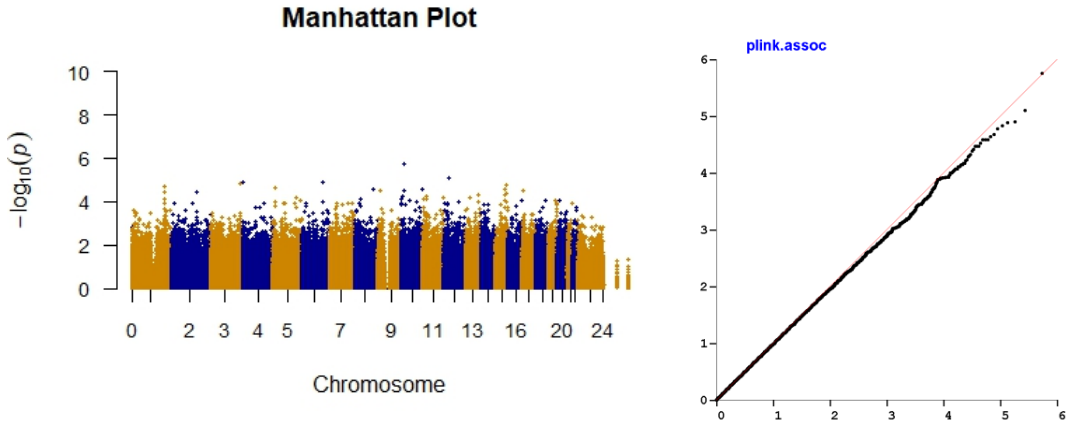
Otra representación importante es el gráfico qqline, este se utiliza para evaluar si la distribución observada de los valores de puntuación estadística o p-values sigue la distribución esperada bajo la hipótesis nula de NO asociación. Cada SNP se somete a una prueba de asociación y se calcula un valor de puntuación estadística (como el p-value) que indica la evidencia de asociación entre el SNP y el fenotipo, en este caso cura o falla.

En el gráfico 13b, se representan los valores observados de p-values en el eje vertical y los valores esperados de p-values (calculados a partir de una distribución teórica) en el eje horizontal. Si los valores observados de p-values se ajustan bien a la distribución esperada bajo la hipótesis nula, los puntos en el gráfico qq se alinearán aproximadamente en una línea recta. Sin embargo, si hay un exceso de valores pequeños de p-values (lo que indica una asociación significativa), los puntos del gráfico qq se desviarán de la línea recta.

Es así como un qqplot en un análisis de GWAS proporciona una visualización de la estructura de asociación entre los SNPs y el fenotipo de interés. Si el gráfico qq muestra una desviación significativa de la línea esperada, puede sugerir la presencia de asociaciones genéticas y señalar SNPs candidatos para un análisis más detallado o validación experimental.

4.4. Resultados Obtenidos

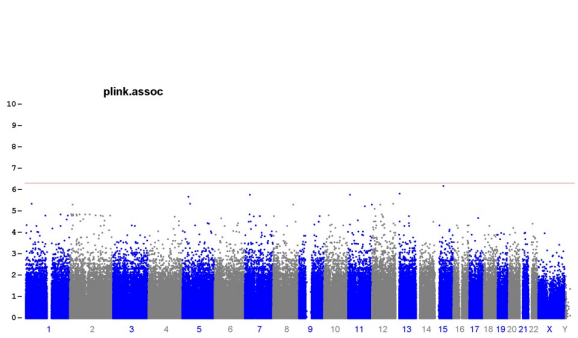
Una vez terminados los GWAS para el datset **COMPLETO** se obtienen 549,483 SNPs, evidenciandose una disminución del número de original de atributos. Al realizar el estudio de asociación se consiguen para este dataset los resultados que se ilustran en la figura 14. El Manhattan Plot se puede ver en la figura 14a y el qqline en la figura 14b se observa una ligera desviación de la linea roja para SNPs con valores superiores a $-log_{10}[p_{value}] = 4$, es decir los SNPs con un $p_{value} < 10^{-4}$ que podrían representar algún tipo de asociación con el fenotipo. Es importante analizar de este gráfico que la asociación con el fenotipo no necesariamente es fuerte, dado por la ligera desviación hacia la parte inferior de la línea recta. Lo que podría indicar que hay problemas en el análisis GWAS, como una inflación excesiva de las pruebas estadísticas u otros factores que terminan incorporando ruido, por ejemplo, la no diferenciación del grupo étnico. Los 41 SNPs obtenidos se pueden consultar en detalle en la tabla 5.



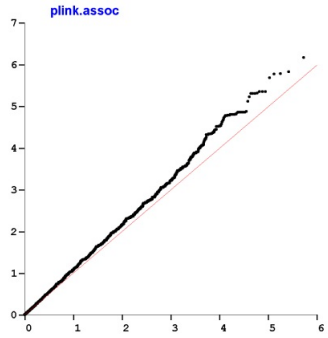
(a) Diagrama Manhattan Plot (b) Diagrama qqline obtenido.

Figura 14: Diagramas de los resultados obtenidos al realizar GWAS para el dataset **COMPLETO**.

El resultado anterior muestra que la división del grupo étnico puede ser una alternativa para explorar y realizar un GWAS por separado. En el caso del dataset **AFRO DESCENDIENTES** se obtienen 539,330 SNPs y luego de la asociación se culmina con un grupo de 14 SNPs que tiene $p_{value} < 10^{-4}$. En el gráfico qqline de la figura 15b se observa una separación de algunos SNPs hacia la parte superior de la linea roja lo que significa que hay más valores de p-values pequeños de lo esperado bajo la hipótesis nula, indicando la presencia de asociaciones genéticas significativas entre los marcadores y el fenotipo de interés



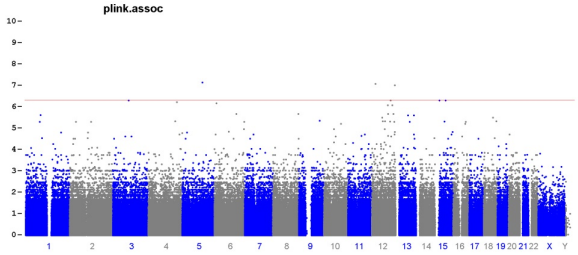
(a) Diagrama Manhattan Plot



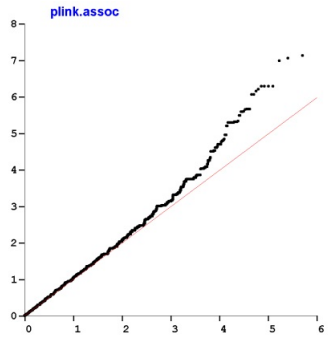
(b) Diagrama qqline obtenido.

Figura 15: Diagramas de los resultados obtenidos al realizar GWAS para el dataset **AFRO DESCENDIENTES**.

Para el dataset **NO-AFRO DESCENDIENTES** los resultados conseguidos se observan en la figura 16, el gráfico qqline ilustra una desviación mucho mas significativa de algunos SNPs donde el $p_{value} < 10^{-5}$ fue el criterio establecido para la selección y con el cual se obtuvieron 36 SNPs con asociaciones genéticas muy significativas al fenotipo.



(a) Diagrama Manhattan Plot



(b) Diagrama qqline obtenido.

Figura 16: Diagramas de los resultados obtenidos al realizar GWAS para el dataset **NO-AFRO DESCENDIENTES**.

El punto de corte comúnmente utilizado para determinar la significancia estadística es un valor p (p-value) menor que 0.05, al utilizar un valor de corte más estricto como un p-value menor que 10^{-4} ó 10^{-5} permite identificar asociaciones aún más fuertes y confiables implicando con esto que los resultados hallados son altamente improbables de haber ocurrido por azar y aunque este no es el único criterio a tener en cuenta para asegurar que las asociaciones conseguidas son altamente eficientes, si es una herramienta de partida que funciona apropiadamente para obtener una considerable reducción en el espacio de características. En la tabla 5 se pueden observar en detalle los resultados conseguidos.

Tabla 5: SNPs selecionados para los diferentes dataset al concluir GWAS.

Experimento	p_{value}	**Número de SNPs seleccionados**	**SNPs Seleccionados**
Dataset COMPLETO.	$< 10^{-4}$	41	rs10800745, rs2486939, rs11120748, rs12996110, rs4859120, rs12644497, rs7733771, rs10061385, rs1480149, rs13173842, rs3813355, rs2929079, rs12548168, rs573057, rs7917410, rs11259260, rs12761922, rs2764804, rs2174257, rs11101913, rs9804548, rs11026669, rs12283577, rs502566, rs10771313, rs12817202, rs1198329, rs9513426, rs9518363, rs10851935, rs2379991, rs12910606, rs8042254, rs720680, rs922533, rs17598590, rs4074608, rs8079726, rs1381548, rs11667789, rs11087039
Dataset AFRO DESCENDIENTES.	$< 10^{-4}$	14	rs12023541, rs1529833, rs6886361, rs9292869, rs41351, rs6999746, rs6484689, rs669776, rs6590735, rs2302688, rs3497, rs4576883, rs9510008, rs2289578
Dataset NO-AFRO DESCENDIENTES.	$< 10^{-5}$	36	rs1906281, rs818524, rs7595485, rs17804991, rs2880961, rs10031902, rs7442549, rs6885608, rs2022016, rs530160, rs4708676, rs9801760, rs12551529, rs11187729, rs2731594, rs4075325, rs273814, rs2731582, rs2579088, rs2579127, rs4337094, rs4075503, rs2948167, rs2942076, rs4559740, rs4075243, rs9317764, rs9318541, rs4408399, rs4028683, rs2440327, rs3825776, rs1106304, rs7499876, rs12457718, rs9636030

5. REDUCCIÓN DE LA DIMENSIONALIDAD

Como se detalló en la sección 3.2, la dimensionalidad de los datos es un tópico que tiene gran impacto en cuanto al costo computacional. Los tres dataset detallados en la tabla 1 tienen demasiadas características y muy pocos registros, aun después de realizar GWAS. Es necesario aplicar técnicas para la reducción de las dimensiones en cada uno de ellos.

En este capítulo se muestran los resultados conseguidos después en la reducción de la dimensionalidad aplicando técnicas de aprendizaje de máquina. Hasta el momento hemos construido tres dataset denominados COMPLETO, AFRO DESCENDIENTES, NO-AFRODESCENDIENTES y en los resultados de GWAS encontramos que la separación de la información a partir de la raza genera resultados significativos. Por tal motivo, el dataset COMPLETO se divide en dos dataset adicionales, el primero, solo con pacientes afrodescendientes y el segundo, con los pacientes no afrodescendientes.

Es así como conseguimos en esta instancia cinco dataset para aplicar técnicas de reducción de la dimensionalidad.

Tabla 6: Características de los datasets para la aplicación de técnicas de aprendizaje automático con fines de reducción de la dimensionalidad.

N°	**DATASET**	**DESCRIPCIÓN**	**# SNPs**
1	COMPLETO	Tiene 72 pacientes. 42 afrodescendientes. 30 no – afrodescendientes. 47 curas. 25 fallas.	41
2	COMPLETO AFRO	Tiene 42 pacientes. 25 curas. 17 fallas.	41
3	COMPLETO NO AFRO	Tiene 30 pacientes. 22 curas. 8 fallas.	41
4	AFRO DESCENDIENTES	Tiene 42 pacientes. 25 curas. 17 fallas.	14
5	NO-AFRO DESCENDIENTES	Tiene 30 pacientes. 22 curas. 8 fallas.	36

5.1. Procedimiento

En la figura 17 se muestra el diagrama de flujo del procedimiento realizado a cada **DATASET** de los detallados en la tabla 6:

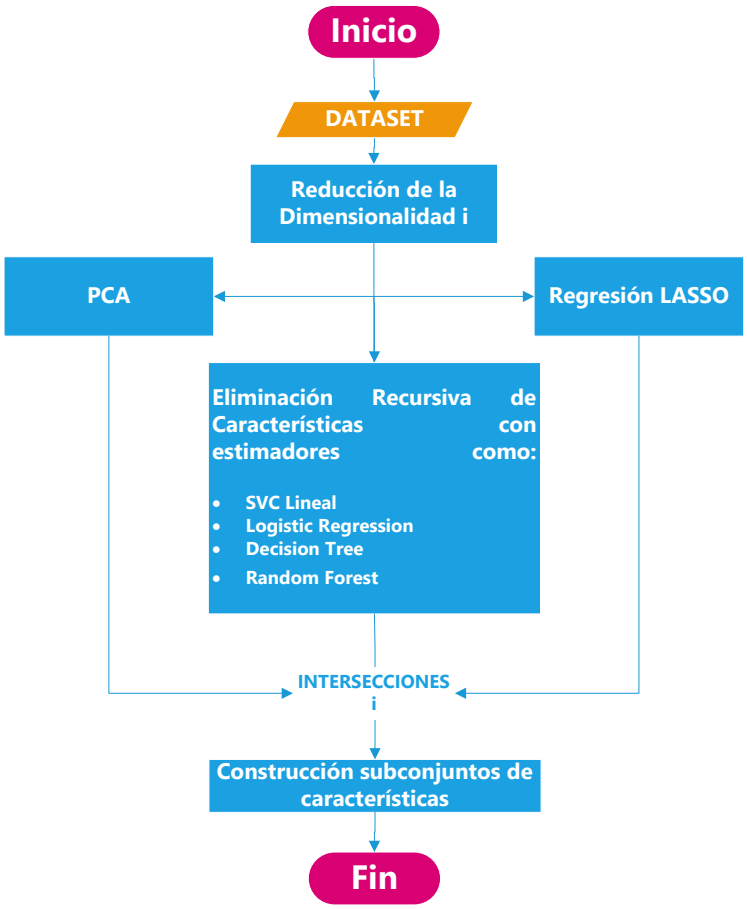


Figura 17: Diagrama de flujo reducción de la dimensionalidad.

1. Se toma la totalidad de los datos y se realiza un GWAS usando PLINK. Este procedimiento ya se explico en detalle en la sección 4.2.
2. Se realiza la división en un 70 % para entrenamiento y un 30 % para la evaluación.
3. Se aplican algoritmos de reducción de la dimensionalidad:
   - Análisis de Componentes Principales.
   - Eliminación recursiva utilizando estimadores: SVC (support vector classifier), árbol de decisión (DT), random forest (RF), regresión logística (RL).
   - Regresión LASSO.

4. Los subconjuntos de características obtenidas obedecerán a cuatro, dados por los estimadores SVC, DT, RF, RL y un quinto obtenido de la regresión LASSO.

5. Usando los subconjuntos obtenidos en el numeral anterior se toman las diferentes intersecciones no vacías, y se construyen subconjuntos de características adicionales.

5.2. Resultados Obtenidos

Después de aplicar las técnicas PCA, eliminación recursiva de características y regresión LASSO a los datasets **COMPLETO**, **COMPLETO AFRO**, **COMPLETO NO AFRO**, **AFRO DESCENDIENTES** y **NO-AFRO DESCENDIENTES**. Se obtuvieron los resultados que se muestran en las siguientes secciones.

5.2.1. Análisis de Componentes Principales PCA

El análisis de los componentes principales es una técnica usada en la reducción de la dimensionalidad. PCA consiste en la representación de combinaciones lineales que no se encuentren correlacionadas entre sí de las variables originales y que consigan maximizar la varianza de las observaciones. La primera componente principal captura la mayor cantidad de la varianza de los datos, la segunda componente captura la segunda mayor cantidad de la varianza, y así sucesivamente.

Al aplicar el algoritmo sobre el dataset **COMPLETO** se exploraron 9 componentes principales debido a que el número de pacientes es de solo 72, estas 9 componentes analizadas entregan la varianza explicada que se observa en la tabla 7.

Tabla 7: Resultado de la aplicación de PCA en el dataset COMPLETO.

Componente Principal	**Varianza Explicada**
Componente 1	0.10654991
Componente 2	0.08902017
Componente 3	0.07531905
Componente 4	0.05868237
Componente 5	0.05610378
Componente 6	0.05350632
Componente 7	0.04765263
Componente 8	0.04557642
Componente 9	0.04149562

La varianza explicada en la tabla 7 no entrega información relevante para explicar ni siquiera el 60 % de los datos, lo que muestra un resultado que no aporta un significado relevante sobre las correlaciones existente entre los SNPs y el fenotipo de cura o falla.

Incluso seleccionando un número mayor de características la varianza no aumento significativamente.

5.2.2. Eliminación recursiva de características

Se aplica la eliminación recursiva de características utilizando cuatro estimadores externos (Support Vector Classifier **SVC**, Regresión Logística **RL**, Árbol de Decisión **DT** y Random Forest **RF**) para asignar pesos a las características, en cada nueva iteración se consideran conjuntos de características cada vez más pequeños, entrenando un conjunto inicial de características y eliminando las menos representativas cada vez que se repite el entrenamiento, hasta alcanzar siete características que corresponderían alrededor del 10 % de los 72 pacientes que tenemos en nuestra base de datos. Dado que son cinco datasets en la tablas 8 a la 12 se observan los resultados conseguidos con los cuatro estimadores.

Tabla 8: Resultados conseguidos para el dataset **COMPLETO**

Estimador	Subconjunto	SNPs
SVC	1	rs10800745, rs12644497, rs10061385 rs1198329, rs922533, rs4074608 rs8079726
RL	2	rs10800745, rs2486939, rs2929079 rs7917410, rs1198329, rs9518363 rs8079726
DT	3	rs10800745, rs2486939, rs10061385 rs573057, rs10771313, rs12910606 rs8079726
RF	4	rs10800745, rs4859120, rs10061385 rs573057, rs11101913, rs2379991 rs8042254

Tabla 9: Resultados conseguidos para el dataset **COMPLETO AFRO**

Estimador	Subconjunto	SNPs
SVC	5	rs3813355, rs11259260, rs9518363, rs10851935, rs720680, rs4074608, rs11087039
RL	6	rs10800745, rs3813355, rs1198329, rs9518363, rs10851935,rs4074608, rs11667789
DT	7	rs10061385, rs11259260, rs1198329, rs9513426, rs9518363,rs8042254, rs720680
RF	8	rs4859120, rs10061385, rs3813355, rs11101913, rs1198329,rs9518363, rs2379991

Tabla 10: Resultados conseguidos para el dataset **COMPLETO NO AFRO**

Estimador	Subconjunto	SNPs
SVC	9	rs12644497, rs1480149, rs3813355, rs11101913, rs9513426,rs4074608, rs8079726
RL	10	rs2486939, rs4859120, rs12644497, rs2929079, rs9804548,rs12283577, rs8079726
DT	11	rs4859120, rs11101913, rs10851935, rs922533, rs17598590,rs11667789, rs11087039
RF	12	rs10800745, rs2486939, rs4859120, rs12644497, rs11101913,rs11026669, rs8079726

Tabla 11: Resultados conseguidos para el dataset **AFRO DESCENDIENTES**

Estimador	Subconjunto	SNPs
SVC	13	rs12023541, rs1529833, rs6886361, rs6590735, rs2302688,rs3497, rs4576883
RL	14	rs12023541, rs6886361, rs9292869, rs6590735, rs2302688, rs3497, rs9510008
DT	15	rs12023541, rs1529833, rs6886361, rs6999746, rs6590735,rs2302688, rs9510008
RF	16	rs12023541, rs1529833, rs6886361, rs6590735, rs2302688,rs4576883, rs9510008

Tabla 12: Resultados conseguidos para el dataset **NO-AFRO DESCENDIENTES**

Estimador	Subconjunto	SNPs
SVC	17	rs17804991, rs12551529, rs2948167, rs2942076, rs4408399,rs4028683, rs2440327
RL	18	rs17804991, rs10031902, rs12551529, rs2948167, rs2942076, rs4028683, rs2440327
DT	19	rs818524, rs4708676, rs2948167, rs2942076, rs4075243,rs9318541, rs4028683
RF	20	rs818524, rs530160, rs12551529, rs11187729, rs2948167,rs2942076, rs4028683

5.2.3. Regresión LASSO

La reducción dimensional usando una Regresión LASSO entrega un subconjunto de SNPs para cada uno de los datasets. En la tabla 13 se observan los resultados conseguidos en esta proceso de reducción de características. En el datset COMPLETO se consiguen 6 SNPs, en COMPLETO AFRO 7, en COMPLETO NO AFRO 6, en AFRO DESCENDIENTES 2 y en NO-AFRO DESCENDIENTES 5.

Tabla 13: Obtención de características para cada uno de los Datasets mediante la regresión LASSO.

DATASET	Subconjunto	SNPs
COMPLETO	21	rs10800745, rs2486939 rs4859120, rs10061385 rs11259260, rs1198329
COMPLETO AFRO	22	rs11120748, rs4859120, rs10061385, rs3813355, rs11259260, rs1198329, rs922533
COMPLETO NO AFRO	23	rs2486939, rs4859120, rs12644497, rs2929079, rs12283577, rs8079726
AFRO DESCENDIENTES	24	rs6590735, rs2302688
NO-AFRO DESCENDIENTES	25	rs17804991, rs10031902, rs12551529, rs2942076, rs4028683

5.3. Intersecciones

En varios dataset existen características comunes entre diferentes subconjuntos de la reducción de la dimensionalidad mediante eliminación recursiva y regresión LASSO. Por ello, se calculan intersecciones de los resultados conseguidos con ambas técnicas y que resultaron en cinco subconjuntos para cada dataset. La selección de características se realiza bajo el criterio de que se encuentren intersectas en los cinco, cuatro, tres o dos de los subconjuntos.

Estas intersecciones generan nuevos subconjuntos de características que se van a utilizar posteriormente en el entrenamiento de algoritmos de aprendizaje de maquina, dado que del resultado del proceso de reducción de la dimensionalidad son los SNPs que parecen tener una mayor capacidad de discriminación entre cura y falla.

En la tabla 14 se observa los resultados de las intersecciones realizadas para el dataset COMPLETO, en la primera columna se encuentra la numeración del subconjunto, la segunda ilustra los subconjuntos de donde se toman las intersecciones y la tercera muestra los SNPs en detalle.

Tabla 14: Resultados conseguidos de las intersecciones para el dataset **COMPLETO**.

Subconjunto	Intersecciones	SNPs
26	SVC, RL, DT, RF, LASSO	rs10800745
27	SVC, DT, RF, LASSO	rs10800745, rs10061385
28	SVC, RL, DT	rs8079726, rs10800745
29	RL, DT, LASSO	rs10800745, rs2486939
30	SVC, RL, LASSO	rs10800745, rs1198329
31	RL, DT	rs8079726, rs10800745, rs2486939
32	SVC, RL	rs8079726, rs10800745, rs1198329
33	DT, RF	rs10800745, rs10061385, rs573057
34	SVC, DT	rs8079726, rs10800745, rs10061385
35	DT, LASSO	rs10800745, rs2486939, rs10061385
36	SVC, RF	rs10800745, rs10061385
37	RF, LASSO	rs10800745, rs4859120, rs10061385
38	SVC, LASSO	rs10800745, rs10061385, rs1198329

Para el dataset COMPLETO se consiguen 18 subconjuntos en total, 4 de la eliminación recursiva, 1 de la regresión LASSO y 13 de las intersecciones. El SNP rs10800745 en el subconjunto número 26 es la única característica que tienen en común los cinco subgrupos obtenidos con las técnicas de reducción de la dimensionalidad estimadores (subconjuntos 1 al 4 y el 21), lo que indica que esta característica tiene una fuerte influencia en la clasificación que se desea realizar, adicionalmente se puede considerar como relevante para la descripción del fenotipo de cura o falla. Otros SNPs relevantes están en los subconjuntos 27 al 29, el rs10061385, rs8079726, rs2486939 y rs1198329 resultan de la intersección de las características obtenidas con cuatro y tres estimadores en la eliminación recursiva, así como en la regresión LASSO.

Para el caso del dataset COMPLEO AFRO los resultados se pueden ver en la tabla 15, aquí se encuentran 13 subconjuntos en total ningún SNP es común en los cinco subconjuntos obtenidos de la reducción de la dimensionalidad, solamente los SNPs de los subconjuntos 39, 40 y 41 dados por el rs9518363, rs1198329 y rs3813355 se encuentran en común dentro de cuatro de los cinco subconjuntos (5, 6, 7, 8 y 22), en el caso de los dos últimos se obtienen tanto con eliminación recursiva como con regresión LASSO lo que refuerza aún más la importancia de estas características comunes en el problema de la clasificación que estamos abordando.

Tabla 15: Resultados conseguidos de las intersecciones para el dataset **COMPLETO AFRO**.

Subconjunto	Intersecciones	SNPs
39	SVC, RL, DT, RF	rs9518363
40	RL, DT, RF, LASSO	rs1198329
41	SVC, RL, RF, LASSO	rs3813355
42	RL, DT, RF	rs9518363, rs1198329
43	SVC, RF	rs9518363, rs3813355
44	RL, RF, LASSO	rs3813355, rs1198329
45	DT, RF, LASSO	rs10061385, rs1198329
46	SVC, DT, LASSO	rs11259260

Los subconjuntos comprendidos del 42 al 46 son obtenidos con SNPs en común en tres estimadores y tres de estos subconjuntos tienen elementos en común con los conseguidos en la regresión LASSO (subconjunto 22), adicionalmente son combinaciones de los SNPs de los subconjuntos 39 al 41.

En el dataset COMPLETO NO AFRO se obtienen en total 12 subconjuntos (9 al 12, 23 y 47 al 53). Ningún SNP resulto común en los cinco subconjuntos resultantes de las técnicas de reducción de la dimensionalidad, mientras el SNP rs4859120 es común a cuatro subconjuntos entre ellos LASSO, igual comportamiento tienen los SNPs rs8079726 y rs12644497 conformando los subconjuntos 47 y 48. Por otra parte, los subconjuntos 49 y 50 tienen SNPs comunes en tres subconjuntos de la tabla 10 y los subconjuntos 51 al 53 tiene SNPs comunes en solo dos.

Tabla 16: Resultados conseguidos de las intersecciones para el dataset **COMPLETO NO AFRO**.

Subconjunto	Intersecciones	SNPs
47	RL, DT, RF, LASSO	rs4859120
48	SVC, RL, RF, LASSO	rs8079726, rs12644497
49	RL, RF, LASSO	rs8079726, rs2486939, rs4859120, rs12644497
50	SVC, DT, RF	rs11101913
51	RL, LASSO	rs12283577, rs2486939, rs12644497, rs8079726, rs2929079, rs4859120
52	DT, RF	rs4859120, rs11101913
53	SVC, RF	rs8079726, rs12644497, rs11101913

Los subconjuntos obtenidos con las intersecciones para el dataset AFRO DESCENDIENTES se observan en la tabla 17 y junto con los subconjuntos conseguidos con la reducción de la dimensionalidad se tienen 12 subconjuntos en total, los SNPs rs6590735 y rs230268

son comunes en los cinco subconjuntos de la reducción de la dimensionalidad; subconjuntos 13 al 16 y el subconjunto 23. Lo que indica que son los SNPs más representativos de este set de datos y obedecen a la reducción de características obtenidos después de realizar el GWAS solamente con la población afrodescendiente.

Tabla 17: Resultados conseguidos de las intersecciones para el dataset **AFRO DESCENDIENTES**

Subconjunto	Intersecciones	SNPs
54	SVC, RL, DT, RF, LASSO	rs6590735, rs2302688
55	SVC, RL, DT, RF	rs12023541, rs6590735, rs6886361, rs2302688
56	RL, DT, RF	rs12023541, rs9510008, rs6590735, rs2302688, rs6886361
57	SVC, DT	rs12023541, rs1529833, rs6590735, rs2302688, rs6886361
58	SVC, RL	rs12023541, rs6590735, rs2302688, rs3497, rs6886361
59	DT, RF	rs12023541, rs9510008, rs1529833, s6590735, rs2302688, rs6886361
60	SVC, RF	rs12023541, rs1529833, rs6590735, rs2302688, rs4576883, rs6886361

El subconjunto 54 termina teniendo los mismos SNPs que el subconjunto 24, es decir, las características obtenidas por LASSO están presentes en los cuatro subconjuntos de eliminación recursiva. Los subconjuntos 55 al 60 son construidos con SNPs comunes en cuatro, tres y dos estimadores. A diferencia de los tres datasets anteriores, en ninguno de estos últimos subconjuntos la regresión LASSO entrega información adicional a la que proporciona la eliminación recursiva.

Para el dataset NO-AFRO DESCENDIENTES se obtienen 15 subconjuntos en total desde el 17 al 20 obtenidos con eliminación recursiva, el subconjunto 25 con regresión LASSO y sus intersecciones en los subconjuntos 61 al 70. Las intersecciones se pueden consultar en la tabla 18, la mayoría de los subconjuntos construidos con las intersecciones poseen un número alto de SNPs en común lo que se puede interpretar como la existencia de una convergencia en los diferentes modelos de selección de características e indicaría una alta correlación de estas características con la clasificación de curas y fallas.

Tabla 18: Resultados conseguidos de las intersecciones para el dataset **NO-AFRO DESCENDIENTES**

Subconjunto	Intersecciones	SNPs
61	SVC, RL, DT, RF, LASSO	rs2942076, rs4028683

62	SVC, RL, DT, RF	rs2942076, rs4028683, rs2948167
63	SVC, RL, RF, LASSO	rs2942076, rs4028683, rs12551529
64	SVC, RL, RF	rs12551529, rs2942076, rs4028683, rs2948167
65	SVC, RL, LASSO	rs17804991, rs2942076, rs4028683, rs12551529
66	SVC, DT, RF	rs4028683, rs2942076, rs2948167
67	SVC, RF	rs4028683, rs2942076, rs2948167, rs12551529
68	SVC, RL	rs17804991, rs2942076, rs2948167, rs12551529, rs2440327, rs4028683
69	RL, LASSO	rs10031902, rs17804991, rs2942076, rs12551529, rs4028683
70	DT, RF	rs4028683, rs2942076, rs2948167, rs818524

Los resultados conseguidos con esta reducción de la dimensionalidad permiten crear 69 subconjuntos de características ó SNPs que tienen asociación directa con el fenotipo de cura o falla en el resultado del tratamiento contra la leishmaniasis. Estos resultados muestran las bondades que presenta la búsqueda de SNPs y biomarcadores al combinar GWAS con técnicas de reducción de dimensionalidad. Si solo se realiza GWAS, debido a la gran cantidad de variables genéticas analizadas en todo el genoma, pueden terminar con problemas de alta dimensionalidad, dificultando el análisis y la interpretación de los resultados. En el capitulo anterior se definieron criterios diversos de p_{value} para la selección de características, terminando con tres datasets; COMPLETO, AFRO DESCENDIENTES Y NO AFRO DESCENDIENTES que en comparación con la dimensión inicial son supremamente pequeños pero aun muy grandes para la cantidad de registros que obedece al orden de los 72 pacientes o menos. Los 69 nuevos dataset, cada uno de ellos con un máximo de 7 características, presentan una reducción del número de variables genéticas consideradas con una mayor relevancia y facilitando el entrenamiento con algoritmos de aprendizaje automático. Dado que es mas fácil entrenar datsets con 7 características que con 600 mil.

6. CONSTRUCCIÓN Y EVALUACIÓN DE MODELOS

6.1. Entrenamiento y evaluación de modelos

En este capítulo se muestran los resultados conseguidos en los experimentos de aprendizaje automático realizados sobre los 69 dataset en el proceso de reducción de la dimensionalidad. En la figura 18 se muestra el procedimiento a seguir para realizar los experimentos de entrenamientos de los diversos datasets que se consiguieron en el proceso de reducción de la dimensionalidad. Se realizan experimentos con 7 algoritmos; Regresión Logistica (**RL**), Gradiente Descendiente Estocástico (**SGD**), Support Vector Machine (**SVM**), Arbol de Decisión (**DT**), Random Forest (**RF**), Boosting (**BT**) y Gradient Boost (**GB**).

Realizando este procedimiento se consiguen en total 490 modelos diferentes donde metodológicamente se exploran el total de subconjuntos de SNPs que se extrajeron de la reducción de dimensionalidad sobre cada dataset (COMPLETO, COMPLETO AFRO, COMPLETO NO-AFRO, AFRO DESCENDIENTES y NO-AFRO DESCENDIENTES). Se les procesa con los siete métodos de aprendizaje automático supervisado con el fin de buscar modelos de predicción de desenlace terapéutico que mejor se ajusten a los datos. Para poder diferenciarlos se realiza la siguiente distribución:

- Conjunto de Experimentos 1: Tiene 18 subconjuntos obtenidos de la reducción de la dimensionalidad e intersecciones del dataset COMPLETO, al entrenar con 7 algoritmos se consiguen 126 experimentos de modelos diferentes.
- Conjunto de Experimentos 2: Tiene 13 subconjuntos obtenidos de la reducción de la dimensionalidad e intersecciones del dataset COMPLETO AFRO, al entrenar con 7 algoritmos se consiguen 91 experimentos de modelos diferentes.
- Conjunto de Experimentos 3: Tiene 12 subconjuntos obtenidos de la reducción de la dimensionalidad e intersecciones del dataset COMPLETO NO-AFRO, al entrenar con 7 algoritmos se consiguen 84 experimentos de modelos diferentes.
- Conjunto de Experimentos 4: Tiene 11 subconjuntos obtenidos de la reducción de la dimensionalidad e intersecciones del dataset AFRO DESCENDIENTES, al entrenar con 7 algoritmos se consiguen 77 experimentos de modelos diferentes.
- Conjunto de Experimentos 5: Tiene 15 subconjuntos obtenidos de la reducción de la dimensionalidad e intersecciones del dataset NO-AFRO DESCENDIENTES, al entrenar con 7 algoritmos se consiguen 105 experimentos de modelos diferentes.

En la figura 18 se puede observar un resumen gráfico de la definición de los conjuntos de experimentos que se realizaron.

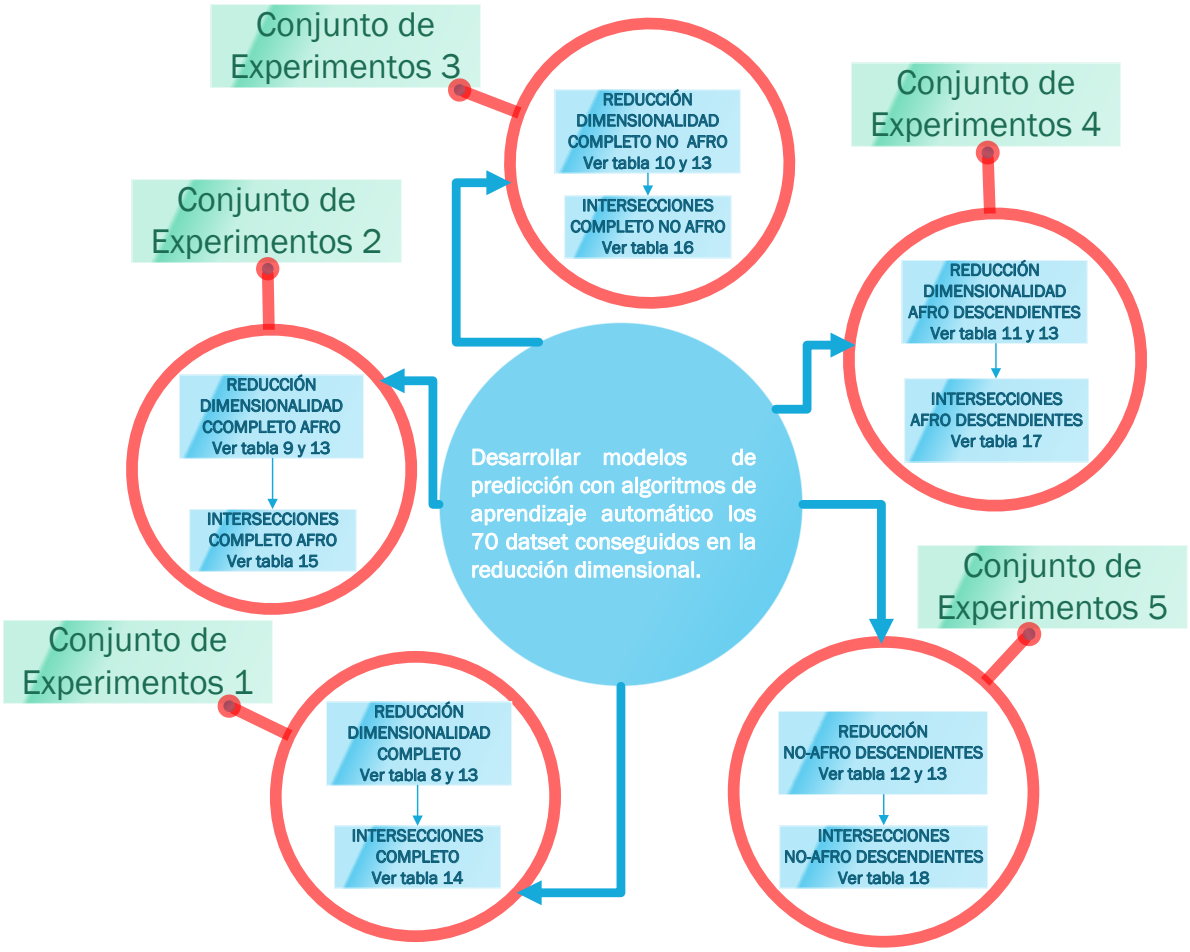


Figura 18: Definición de los conjuntos de experimentos para la realización de entrenamientos con algoritmos de aprendizaje automático.

6.2. Resultados Obtenidos

Para mostrar todos los resultados conseguidos de los diferentes entrenamientos realizados para cada conjunto de experimentos, se usa la métrica de desempeño **F1 score** que combina tanto la precisión como el recall del modelo. Esta métrica es efectiva cuando ambas (precisión y recall) son importantes. La precisión mide la proporción de resultados positivos que son realmente positivos, mientras que el recall mide la proporción de resultados

positivos identificados correctamente respecto al total de positivos reales. Al combinar ambas métricas se consigue proporcionar un equilibrio en la evaluación del modelo. Adicionalmente, **F1 score** es sensible al desbalance de clases, mientras, la precisión y el recall pueden dar una visión sesgada del rendimiento del modelo. Esto lo hace una opción valiosa para evaluar el rendimiento del modelo en situaciones donde se deben predecir categorías binarias.

Todas las métricas en detalle asociadas a cada experimento se puede consultar en el anexo A.

6.2.1. Conjunto de Experimentos 1

El resultado de los 126 experimentos se resume en la tabla 19, mostrando buenos desempeños en la clasificación tanto de las curas como de las fallas. De la división de los datos realizada antes de la reducción de dimensiones se tiene que el conjunto de prueba que consta de 22 registros con 16 curas y 6 fallas. El conjunto de entrenamiento cuenta con 50 registros en total, 31 curas y 19 fallas, las curas son 1.6 veces más que las fallas.

Los subconjuntos de características con muy buenos desempeños en seis de sus siete modelos son el número 1, 2, 30 y 38. Teniendo en común los SNPs rs10800745 y rs1198329. Para el conjunto de características número 1 los modelos construidos con el algoritmos **RL**, **SVM**, y **RF** son los modelos que mejor aprendieron clasificar curas y fallas. Mientras para el subconjunto número 2 los mejores modelos fueron los construidos con los algoritmos **DT**, **RF** y **BT**. De igual manera susede con el subconjunto 30 los modelos construidos con los algoritmos **DT**, **RF** y **BT** son los que presentan mejores desempeños, incluso mucho mejores que los de los dos subconjuntos anteriores.

Tabla 19: Resultado de la métrica de desempeño **F1 score** obtenidas para el conjunto de experimentos 1.

Subconjunto	SNPs	RL	SGD	SVM	DT	RF	BT	GB
1	rs10800745 rs12644497 rs10061385 rs1198329 rs922533 rs4074608 rs8079726	0.85	0.77	0.85	0.81	0.85	0.81	0.82
2	rs10800745 rs2486939 rs2929079 rs7917410 rs1198329 rs9518363 rs8079726	0.77	0.67	0.77	0.81	0.89	0.82	0.75
3	rs10800745 rs2486939 rs10061385 rs573057 rs10771313 rs12910606 rs8079726	0.75	0.71	0.71	0.81	0.79	0.77	0.75
4	rs10800745 rs4859120 rs10061385 rs573057 rs11101913 rs2379991 rs8042254	0.89	0.82	0.89	0.81	0.89	0.81	0.89

21	rs10800745 rs2486939 rs4859120 rs10061385 rs11259260 rs1198329	0.89	0.89	0.89	0.91	0.89	0.91	0.89
26	rs10800745	0.69	0.84	0.64	0.69	0.69	0.69	0.69
27	rs10800745 rs10061385	0.86	0.84	0.86	0.85	0.85	0.85	0.85
28	rs8079726 rs10800745	0.65	0.64	0.65	0.69	0.69	0.69	0.65
29	rs10800745 rs2486939	0.71	0.84	0.71	0.75	0.75	0.75	0.75
30	rs10800745 rs1198329	0.86	0.48	0.86	0.91	0.91	0.91	0.86
31	rs8079726 rs10800745 rs2486939	0.71	0.84	0.71	0.62	0.75	0.62	0.75
32	rs8079726 rs10800745 rs1198329	0.67	0.52	0.65	0.62	0.89	0.62	0.65
33	rs10800745 rs10061385 rs573057	0.88	0.85	0.86	0.85	0.88	0.85	0.85
34	rs8079726 rs10800745 rs10061385	0.67	0.64	0.65	0.59	0.85	0.59	0.85
35	rs10800745 rs2486939 rs10061385	0.71	0.75	0.71	0.75	0.82	0.75	0.82
36	rs10800745 rs10061385	0.86	0.84	0.86	0.85	0.85	0.85	0.85
37	rs10800745 rs4859120 rs10061385	0.89	0.86	0.89	0.89	0.89	0.89	0.89
38	rs10800745 rs10061385 rs1198329	0.88	0.73	0.86	0.84	0.88	0.88	0.85

Por otra parte, los subconjuntos de características 4, 21, 27, 33, 36 y 37 presentan muy buenos desempeños en los siete modelos construidos con cada subconjunto de características, los SNPs rs10800745 y rs10061385 son comunes en estos subconjuntos. A pesar de su altas métricas de F1 score, para estos datsets es notorio un desbalance entre las diferentes métricas; accuracy, precision, recall y F1 score, por ejemplo, para el subconjunto 36 con el algoritmo SGD las métricas de accuracy son de 0.73, precision de 0.73, recall de 1.00 y un F1 score de 0.84. Indicando que el 73 % de las clasificaciones se realizaron correctamente, 73 % que obedece a curas y que son realmente positivas. Por ello el recall del modelo es del 100 %. Es decir, que este modelo aprendió significativamente a clasificar las curas pero no tan bien la clasificación de fallas. En general, la mayoría de los modelos presentan estas variaciones en sus métricas para estos subconjuntos de características, lo que se puede ver en detalle en el anexo A. Entonces, el desbalance de las clases influye considerablemente en los resultados de estos modelos. Lo que puede subsanarse en la etapa de optimización de los modelos.
Es importante no perder de vista para la fase de optimización el hecho que los modelos con las métricas de F1 score mas altas (0.91) tienen en común los SNPs rs10800745 y rs1198329, así como los algoritmos **DT**, **RF** y **BT**.

6.2.2. Conjunto de Experimentos 2

Al realizar los 91 experimentos con solo los pacientes afro descendientes, se consiguen los resultados que se observan en la tabla 20. El 50 % de los experimentos tienen un F1 score igual o superior a 0.80 y sin un balance muy adecuado con las métricas como accuracy, precision y recall (véase el anexo A), debido a un preferente aprendizaje de clasificación de las curas. En el conjunto de entrenamiento hay 17 curas 12 fallas dejando una relación de 1.4 veces mas curas que fallas, mientras, en el conjunto de prueba existen 7 curas y 5 fallas.

En el subconjunto de características número 5, número 6 y número 22 los siete modelos desarrollados presentan buenas métricas de desempeño. Para el caso del conjunto 5 los modelos construidos con **DT**, **RF**, **BT** y **GB** tiene un balance adecuado en sus métricas, mientras los modelos desarrollados con **RL**, **SGD** y **SVM** presentan una precisión igual a la unidad mostrando que estos tres modelos predicen todas las etiquetas, fallas y curas.

En el subconjunto número 6 los modelos generados con los algoritmos **SGD** y **SVM** obtienen un recall igual a la unidad, mostrando que estos modelos predicen adecuadamente todas las curas en el conjunto de datos de prueba. Para el subconjunto número 22 sucede lo mismo exclusivamente en el modelo construido con la **RL**.

Tabla 20: Resultado de la métrica de desempeño **F1 score** obtenidas para el conjunto de experimentos 2.

Subconjunto	SNPs	RL	SGD	SVM	DT	RF	BT	GB
5	rs3813355 rs11259260 rs9518363 rs10851935 rs720680 rs4074608 rs11087039	0.82	0.82	0.88	0.67	0.82	0.67	0.67
6	rs10800745 rs3813355 rs1198329 rs9518363 rs10851935 rs4074608 rs11667789	0.94	0.93	0.93	0.82	0.88	0.82	0.82
7	rs10061385 rs11259260 rs1198329 rs9513426 rs9518363 rs8042254 rs720680	0.78	0.78	0.82	0.82	0.78	0.78	0.78
8	rs4859120 rs10061385 rs3813355 rs11101913 rs1198329 rs9518363 rs2379991	0.84	0.84	0.84	0.78	0.89	0.78	0.78
22	rs11120748 rs4859120 rs10061385 rs3813355 rs11259260 rs1198329 rs922533	0.82	0.94	0.89	0.82	0.82	0.82	0.82
39	rs9518363	0.67	0.67	0.67	0.78	0.78	0.78	0.78
40	rs1198329	0.89	0.76	0.89	0.93	0.93	0.93	0.78
41	rs3813355	0.82	0.82	0.82	0.71	0.71	0.71	0.71
42	rs9518363 rs1198329	0.89	0.67	0.89	0.93	0.89	0.93	0.89
43	rs9518363 rs3813355	0.84	0.84	0.84	0.75	0.75	0.75	0.84
44	rs3813355 rs1198329	0.89	0.88	0.82	0.88	0.88	0.88	0.88
45	rs10061385 rs1198329	0.89	0.76	0.89	0.93	0.93	0.93	0.93
46	rs11259260	0.74	0.74	0.74	0.74	0.74	0.74	0.74

En los subconjuntos de características número 22, 42, 44 y 45 se consigue varios modelos

con métricas altamente balanceadas entre la totalidad de las clases predichas y las que obedecen exclusivamente a las curas, los modelos **RL**,**SGD**, **SVM** y **RF**.

Por último, el subconjunto 46 tiene un SNP que no presenta buenas métricas en ningún algoritmo implementado, esto indica que no contribuyen a explicar adecuadamente el fenotipo. Los dos mejores modelos obtenidos en este conjunto de experimentos son el construido con **RL** en el subconjunto número 6 y el del subconjunto 22 con el algoritmo **SGD**.

6.2.3. Conjunto de Experimentos 3

El resultado de los 84 experimentos con pacientes no afro descendientes se observan en la tabla 21, aquí se tiene una distribución de 16 curas y 5 fallas en el conjunto de entrenamiento, 6 curas y 3 fallas en el conjunto de prueba. Se obtienen buenas métricas de desempeño en más del 60 % de los experimentos. El desbalance en las clases del conjunto de entrenamiento, en este caso hay 3,2 veces mas curas que fallas facilita que se predigan de manera preferente las curas. En un 8 % del total de experimentos se predice adecuadamente el total de las categorías en los datos de prueba lo que puede representar un buen aprendizaje de los patrones en los datos.

El modelo construido con el subconjunto número 9 y el algoritmo **DT** muestra un buen balance entre las diferentes métricas calculadas. Así mismo, el modelo de este subconjunto de características con el algoritmo **SVM** presenta una de las métricas mas altas (F1 score de 0.91) y de mejor comportamiento en la explicación de las curas y fallas.

Tabla 21: Resultado de la métrica de desempeño **F1 score** obtenidas para el conjunto de experimentos 3.

Subconjunto	SNPs	RL	SGD	SVM	DT	RF	BT	GB
9	rs12644497 rs1480149 rs3813355 rs11101913 rs9513426 rs4074608 rs8079726	0.86	0.86	0.91	0.83	0.86	0.86	0.80
10	rs2486939 rs4859120 rs12644497 rs2929079 rs9804548 rs12283577 rs8079726	1.00	1.00	1.00	1.00	1.00	1.00	0.80
11	rs4859120 rs11101913 rs10851935 rs922533 rs17598590 rs11667789 rs11087039	0.77	0.77	0.77	0.86	0.80	0.86	0.86

12	rs10800745 rs2486939 rs4859120 rs12644497 rs11101913 rs11026669 rs8079726	0.86	1.00	0.91	0.91	0.92	0.91	0.86
23	rs2486939 rs4859120 rs12644497 rs2929079 rs12283577 rs8079726	1.00	1.00	1.00	1.00	1.00	1.00	1.00
47	rs4859120	0.80	0.00	0.80	0.91	0.80	0.91	0.80
48	rs8079726 rs12644497	0.80	0.80	0.80	0.92	0.92	0.92	0.80
49	rs8079726 rs2486939 rs4859120 rs12644497	1.00	1.00	1.00	0.91	1.00	1.00	0.92
50	rs11101913	0.80	0.67	0.80	0.86	0.86	0.86	0.80
51	rs12283577 rs2486939 rs12644497 rs8079726 rs2929079 rs4859120	1.00	1.00	1.00	1.00	1.00	1.00	1.00
52	rs4859120 rs11101913	0.77	0.77	0.77	0.86	0.86	0.86	0.86
53	rs8079726 rs12644497 rs11101913	0.86	0.86	0.80	0.80	0.80	0.80	0.80

Seis de los siete modelos construidos con el subconjunto de características número 10 presentan métricas perfectas en la predicción de las curas, lo que resulta bastante positivo teniendo presente que todas estas métricas corresponden al conjunto de prueba. El séptimo modelo que obedece concretamente al algoritmo **GB** también presenta buenas métricas en la predicción del fenotipo. De igual forma el subconjunto de características número 23 y 51 presentaron una predicción de las clases de manera perfecta en los siete modelos construidos con él. Estos subconjuntos tiene en común los SNPs rs4859120, rs12644497, rs2929079, rs12283577, rs8079726 y rs2486939, algunos de ellos se encuentran también en subconjuntos con buenas métricas lo que podría llevar a pensar que los que solo se encuentran en estos subconjuntos están estrechamente asociados con el fenotipo dado por el alto grado de de desempeño.

También el subconjunto número 49 tiene una predicción perfecta en cinco de los siete modelos. En estos modelos podría existir una muy buena predicción, dado por el buen aprendizaje de los modelos a los nuevos datos, o incluso un posible sobreajuste que tendría que ser revisado posterior a una optimización y estimación de los modelos.

Otros modelos con buen desempeño fueron construidos con los subconjuntos de características 12, los mejores resultaron al usar los algoritmos **SGD** y **RF**, y vale resaltar que los restantes cinco modelos tienen métricas igualmente altas. Los tres modelos del subconjunto número 48 con los algoritmos **DT**, **RF** y **BT** tienen métricas bastante altas. En todos estos modelos fueron comunes los SNPs rs12644497, rs11101913, rs8079726, rs2486939 y rs4859120.

Todos los modelos del subconjunto número 53 y varios del número 52, 50, 47 y 11 presentan buenos desempeños, hay varios modelos con un recall igual a 1.0, evidenciando el aprendizaje preferencial por la clasificación de las curas.

Este conjunto de experimentos es el mejor comportado hasta el momento, la gran mayoría de los modelos predicen bastante bien los fenotipos, un buen conjunto de modelos lo hace con exactitud y lo restantes modelos los hacen con buena precisión.

6.2.4. Conjunto de Experimentos 4

El resultado de los 77 experimentos con pacientes afro descendientes en un conjunto de entrenamiento que consta de 17 curas y 12 fallas, es decir, 1.4 veces mas curas que fallas, y, el conjunto de prueba tiene 7 curas y 5 falla se observan en la tabla 22. A diferencia de los resultados de conjuntos de experimentos anteriores las métricas de desempeño son bastante regulares.

Al rededor de un 10 % de los modelos construidos con estos subconjuntos de características presentan una métrica de F1 score superiores al 80 %. Los mejores modelos se consiguen en los subconjuntos de características número 15, 16, 57 y 59, todos con el algoritmo **RF**. Los SNPs en común para estos subconjuntos son el rs12023541, rs1529833, rs6590735, rs2302688 y rs6886361, aunque son los modelos con las métricas de desempeño mas altas tienen una tendencia a clasificar de manera mas apropiada las curas y presentan grandes falencias para la predicción de fallas, metricas como accuracy, precision y recall pueden verse en detalle en el anexo A.

Tabla 22: Resultado de la métrica de desempeño **F1 score** obtenidas para el conjunto de experimentos 4.

Subconjunto	SNPs	RL	SGD	SVM	DT	RF	BT	GB
13	rs12023541 rs1529833 rs6886361 rs6590735 rs2302688 rs3497 rs4576883	0.50	0.46	0.46	0.63	0.74	0.63	0.74
14	rs12023541 rs6886361 rs9292869 rs6590735 rs2302688 rs3497 rs9510008	0.43	0.31	0.31	0.63	0.63	0.63	0.71
15	rs12023541 rs1529833 rs6886361 rs6999746 rs6590735 rs2302688 rs9510008	0.67	0.57	0.67	0.71	0.80	0.71	0.80
16	rs12023541 rs1529833 rs6886361 rs6590735 rs2302688 rs4576883 rs9510008	0.67	0.46	0.67	0.71	0.84	0.63	0.74
24	rs6590735 rs2302688	0.67	0.67	0.67	0.78	0.78	0.78	0.78

55	rs12023541 rs6590735 rs6886361 rs2302688	0.75	0.67	0.67	0.67	0.71	0.67	0.75
56	rs12023541 rs9510008 rs6590735 rs2302688 rs6886361	0.75	0.50	0.67	0.57	0.71	0.57	0.75
57	rs12023541 rs1529833 rs6590735 rs2302688 rs6886361	0.75	0.57	0.67	0.63	0.84	0.71	0.80
58	rs12023541 rs6590735 rs2302688 rs3497 rs6886361	0.63	0.67	0.67	0.71	0.71	0.71	0.71
59	rs12023541 rs9510008 rs1529833 rs6590735 rs2302688 rs6886361	0.75	0.63	0.67	0.71	0.80	0.71	0.80
60	rs12023541 rs1529833 rs6590735 rs2302688 rs4576883 rs6886361	0.67	0.59	0.67	0.71	0.84	0.63	0.74

Modelos con métricas mas moderadas y balanceados en la clasificación de ambas clases son los construidos con el subconjunto de características número 24 compuesto por los SNPs rs6590735 y rs2302688 que usando los algoritmos **DT**, **RF**, **BT** y **GB** consiguen un F1 score de 0.78, recall de 0.88 y precisión de 0.70. Otros modelos balanceados se encuentran en los subconjuntos número 55 y 57 con los SNPs rs12023541, rs6590735 y rs2302688 en común que obtienen 0.75 tanto en F1 score, recall y precisión.

Los modelos de los subconjuntos número 13, 14 y 58 presentan desempeños muy bajos en los diferentes modelos lo que permite descartar que estos tengas algún tipo correlación con la explicación del fenotipo.

6.2.5. Conjunto de Experimentos 5

En este conjunto de experimentos se realizan 105 entrenamientos con datasets de no afro descendientes, en cada uno de ellos el conjunto de entrenamientos tenia 16 curas y 5 fallas, 3,2 veces mas curas que fallas y el conjunto de prueba cuenta con 6 curas y 3 fallas. Los resultados que se consiguen se pueden observar en la tabla 23, se encontraron buenos desempeños en mas del 87 % de los modelos construidos. En su mayoría todos consiguen clasificar muy bien las curas y no tanto las fallas. La métricas de recall aparecen en la mayoría de los modelos estimadas en la unidad, mientras la precisión resulta tener valores bajos lo que evidencia la dificultad de estos modelos por predecir en igual proporción las fallas, como lo hace con las curas.

Los mejores modelos pueden ascender a un total de 25 con un F1 score de 0.92 y un recall igual a 1.0, los SNPs en común para todos estos modelos son el rs17804991, rs10031902,

rs12551529, rs2948167, rs2942076, rs4028683 y rs2440327 y están entre los subconjuntos de características número 18, 25, 63, 65, 68 y 69.

Tabla 23: Resultado de la métrica de desempeño **F1 score** obtenidas para el conjunto de experimentos 5.

Subconjunto	SNPs	RL	SGD	SVM	DT	RF	BT	GB
17	rs17804991 rs12551529 rs2948167 rs2942076 rs4408399 rs4028683 rs2440327	1.00	0.67	0.80	0.86	0.80	0.86	0.86
18	rs17804991 rs10031902 rs12551529 rs2948167 rs2942076 rs4028683 rs2440327	0.92	1.00	0.80	0.92	0.86	0.83	0.86
19	rs818524 rs4708676 rs2948167 rs2942076 rs4075243 rs9318541 rs4028683	0.86	0.77	0.86	0.86	0.86	0.86	0.86
20	rs818524 rs530160 rs12551529 rs11187729 rs2948167 rs2942076 rs4028683	0.86	0.86	0.86	0.77	0.80	0.77	0.80
25	rs17804991 rs10031902 rs12551529 rs2942076 rs4028683	0.92	0.92	0.83	0.92	0.92	0.92	0.86
61	rs2942076 rs4028683	0.86	0.83	0.86	0.83	0.86	0.83	0.86
62	rs2942076 rs4028683 rs2948167	0.86	0.83	0.86	0.83	0.86	0.83	0.80
63	rs2942076 rs4028683 rs12551529	0.92	0.86	0.86	0.92	0.86	0.92	0.86
64	rs12551529 rs2942076 rs4028683 rs2948167	0.86	0.86	0.86	0.92	0.86	0.92	0.86
65	rs17804991 rs2942076 rs4028683 rs12551529	0.92	0.83	0.83	0.92	0.92	0.92	0.86
66	rs4028683 rs2942076 rs2948167	0.86	0.77	0.86	0.83	0.86	0.83	0.80
67	rs4028683 rs2942076 rs2948167 rs12551529	0.86	0.86	0.86	0.92	0.86	0.92	0.86
68	rs17804991 rs2942076 rs2948167 rs12551529 rs2440327 rs4028683	0.92	0.80	0.80	0.92	0.86	0.92	0.86
69	rs10031902 rs17804991 rs2942076 rs12551529 rs4028683	0.92	0.80	0.83	0.92	0.92	0.92	0.86
70	rs4028683 rs2942076 rs2948167 rs818524	0.86	0.86	0.86	0.86	0.80	0.86	0.80

Los modelos de los subconjuntos 17, 19, 20, 61, 62, 64, 66, 67 y 70 cuentan también con buenas métricas de desempeño que rondan un F1 score entre 0.80 - 0.86, lo que sin duda los perfila como modelos candidatos a mejorar substancialmente en el proceso de optimización y convertirse en modelos que consigan explicar adecuadamente la relación de estos SNPs y el fenotipo de cura o falla.

Este conjunto de experimentos es por mucho el mas exitoso de los cinco conjuntos de experimentos que se realizaron. Dos modelos del subconjunto de características número 17 con el algoritmo de **RL** y número 18 con el algoritmo **SGD** mostrando que todas sus métricas obedecen a 1.0 haciendo clasificaciones exactas de todas las clases.

Por otra parte, el conjunto completo de experimentos se ejecuto en un tiempo aproximado de 70 minutos en donde se entrenaron y evaluaron los 483 modelos usando Google Colab en su versión gratuita.

7. OPTIMIZACIÓN DE MODELOS

7.1. Definición de Hiperparámetros

En este capítulo, se presentan los resultados obtenidos a partir de la búsqueda exhaustiva de hiperparámetros por cuadricula, el método explicado en la sección 3.6, se aplico a los 470 modelos conseguidos en los experimentos de base. Después de un riguroso proceso de experimentación y evaluación, se encontraron combinaciones óptimas de parámetros que deben ser evaluados en búsqueda de mejoras significativas de los modelos.

Para la búsqueda de los hiperparámetros a explorar se construyeron diccionarios que permiten realizar el ajuste de los mismos en cada algoritmo, luego se aplica búsqueda por grid search con una validación cruzada de 5 particiones.

- **Regresión Logística**

```
solvers = ['newton-cg', 'lbfgs', 'liblinear', 'sag', 'saga']
penalty = ['l1','l2', 'elasticnet', 'none']
c_values = [100, 10, 1.0, 0.1, 0.01]
lr_grid = dict(solver=solvers,penalty=penalty,C=c_values)
```

  Con un total de 100 combinaciones de parámetros a explorar.

- **Descenso de Gradientes Estocástico - SGD**

  Dado que SGDClassifier es un módulo de scikit-learn que se basa en una implementación de algoritmos de optimización usando SGD y el parámetro loss corresponde a la función de pérdida que se utiliza durante el entrenamiento del modelo. En esta búsqueda de hiperparámetros se usaran los siguientes algoritmos para la función de perdida:

  - hinge: Algoritmo lineal (SVC).
  - log: Algoritmo regresión logística. Busca maximizar la probabilidad logarítmica de la clase correcta.
  - modified_huber: Es menos sensible a los valores atípicos que hinge y log. Combina elementos de la función de pérdida hinge y la regresión logística. Es muy útil cuando los datos contienen valores atípicos.
  - squared_hinge: Es similar a hinge, pero utiliza una versión cuadrática.
  - perceptron: Busca minimizar la cantidad de errores de clasificación utilizando el modelo de Perceptrón.

De esta manera se puede observar que los modelos que aquí se quieren construir no se repiten en las demás configuraciones.

```
loss = ['hinge', 'log', 'modified_huber', 'squared_hinge', '
    perceptron']
penalty = ['l1','l2','elasticnet']
alpha= [0.0001,0.001,0.01,0.1]
l1_ratio= [0.15,0.05,.025]
max_iter = [1,5,10,100,1000,10000]
sgd_grid = dict(penalty=penalty,max_iter=max_iter,alpha=alpha,
    l1_ratio=l1_ratio)
```

Con un total de 1080 combinaciones de parámetros a explorar.

- **Support Vector Classifier**

```
kernel = ['poly', 'rbf', 'sigmoid']
C = [50, 10, 1.0, 0.1, 0.01]
svc_grid = dict(kernel=kernel,C=C)
```

Con un total de 15 combinaciones de parámetros a explorar.

- **Árbol de Decisión**

```
splitter=["random",'best']
max_depth = [0, 1, 2, 3, 4]
criterion = ["gini", "entropy"]
ccp_alpha = [0.0, 0.1, 0.2]
dt_grid = dict(splitter=splitter, max_depth=max_depth, criterion=
    criterion, ccp_alpha=ccp_alpha)
```

Con un total de 60 combinaciones de parámetros a explorar.

- **Random Forest**

```
n_estimators = [10, 100, 1000, 10000]
max_features = ['sqrt', 'log2']
rf_grid = dict(n_estimators=n_estimators,max_features=max_features)
```

Con un total de 8 combinaciones de parámetros a explorar.

- **Boosting**

```
n_estimators=[50, 100, 200]
learning_rate=[0.1, 0.5, 1.0]
b_grid=dict(learning_rate=learning_rate, n_estimators=n_estimators)
```

Con un total de 9 combinaciones de parámetros a explorar.

- **Gradient Boost**

```
learning_rate = [0.01, 0.05, 0.1]
max_depth = [0, 1, 2, 3, 4]
n_estimators=[100, 200, 300]
gb_grid=dict(learning_rate=learning_rate, max_depth = max_depth,
    n_estimators=n_estimators)
```

Con un total de 45 combinaciones de parámetros a explorar.

7.2. Resultados Obtenidos

Después de realizar la búsqueda de hiperparámetros para el conjunto de experimentos número 1 se obtienen los resultados mostrados en la tabla 24. En los 126 experimentos optimizados no se encuentra una tendencia especifica de los mejores hiperparámetros encontrados. Algunos subconjuntos que tienen en común varios SNPs terminaron con hiperparámetros muy similares, lo que es de esperar dado que los dataset tendrían una distribución similar de los datos.

Tabla 24: Resultados obtenidos en la búsqueda de hiperparámetros para el conjunto de experimentos 1.

Subconjunto	SNPs	Algoritmo	Mejor Modelo
1	rs10800745 rs12644497 rs10061385 rs1198329 rs922533 rs4074608 rs8079726	REGRESION LOGISTICA	C 10 penalty l2 solver liblinear
		SGD	alpha 0,0001 l1_ratio 0,05 max_iter 10000 penalty l2
		SVM	C 1,0 gamma scale kernel sigmoid
		ARBOL DE DECISIÓN	ccp_alpha 0,0 criterion entropy max_depth 3 splitter best
		RANDOM FOREST	max_features sqrt n_estimators100
		BOOSTING	learning_rate 0,5 n_estimators100
		GRADIENTE BOOST	learning_rate 0,1 max_depth 2 n_estimators200
2	rs10800745 rs2486939 rs2929079 rs7917410 rs1198329 rs9518363 rs8079726	REGRESION LOGISTICA	C 100 penalty l1 solver liblinear
		SGD	alpha 0,0001 l1_ratio 0,05 max_iter 100 penalty l2
		SVM	C 10 gamma scale kernel rbf
		ARBOL DE DECISIÓN	ccp_alpha 0,0 criterion gini max_depth 3 splitter random
		RANDOM FOREST	max_features sqrt n_estimators100
		BOOSTING	learning_rate 0,1 n_estimators200
		GRADIENTE BOOST	learning_rate 0,01 max_depth 3 n_estimators200
3	rs10800745 rs2486939 rs10061385 rs573057 rs10771313 rs12910606 rs8079726	REGRESION LOGISTICA	C 0,01 penalty l2 solver liblinear
		SGD	alpha 0,01 l1_ratio 0,025 max_iter 1 penalty l1
		SVM	C 1,0 gamma scale kernel rbf
		ARBOL DE DECISIÓN	ccp_alpha 0,0 criterion gini max_depth 4 splitter random
		RANDOM FOREST	max_features sqrt n_estimators10000
		BOOSTING	learning_rate 0,1 n_estimators200
		GRADIENTE BOOST	learning_rate 0,01 max_depth 2 n_estimators200
4	rs10800745 rs4859120 rs10061385 rs573057 rs11101913 rs2379991 rs8042254	REGRESION LOGISTICA	C 1,0 penalty l2 solver newton-cg
		SGD	alpha 0,01 l1_ratio 0,05 max_iter 10000 penalty elasticnet
		SVM	C 10 gamma scale kernel poly
		ARBOL DE DECISIÓN	ccp_alpha 0,0 criterion gini max_depth 1 splitter random
		RANDOM FOREST	max_features log2 n_estimators10
		BOOSTING	learning_rate 0,1 n_estimators200
		GRADIENTE BOOST	learning_rate 0,1 max_depth 1 n_estimators200
21	rs10800745 rs2486939 rs4859120 rs10061385 rs11259260 rs1198329	REGRESION LOGISTICA	C 1,0 penalty l2 solver liblinear
		SGD	alpha 0,0001 l1_ratio 0,05 max_iter 10 penalty elasticnet
		SVM	C 1,0 gamma scale kernel sigmoid
		ARBOL DE DECISIÓN	ccp_alpha 0,0 criterion gini max_depth 1 splitter random
		RANDOM FOREST	max_features sqrt n_estimators10
		BOOSTING	learning_rate 0,1 n_estimators50
		GRADIENTE BOOST	learning_rate 0,05 max_depth 1 n_estimators200
		REGRESION LOGISTICA	C 100 penalty l1 solver liblinear
		SGD	alpha 0,01 l1_ratio 0,05 max_iter 100 penalty l2
		SVM	C 50 gamma scale kernel rbf
		ARBOL DE DECISIÓN	ccp_alpha 0,0 criterion gini max_depth 1 splitter best
		RANDOM FOREST	max_features sqrt n_estimators10

26	rs10800745		

Tabla 24 continua de la pagina anterior.

		BOOSTING	learning_rate 0,1 n_estimators50
		GRADIENTE BOOST	learning_rate 0,01 max_depth 1 n_estimators200
27	rs10800745 rs10061385	REGRESION LOGISTICA	C 1,0 penalty l1 solver liblinear
		SGD	alpha 0,001 l1_ratio 0,025 max_iter 1000 penalty elasticnet
		SVM	C 50 gamma scale kernel poly
		ARBOL DE DECISIÓN	ccp_alpha 0,0 criterion gini max_depth 2 splitter best
		RANDOM FOREST	max_features sqrt n_estimators100
		BOOSTING	learning_rate 0,1 n_estimators50
		GRADIENTE BOOST	learning_rate 0,01 max_depth 1 n_estimators100
28	rs8079726 rs10800745	REGRESION LOGISTICA	C 0,1 penalty l2 solver newton-cg
		SGD	alpha 0,001 l1_ratio 0,15 max_iter 10 penalty l1
		SVM	C 1,0 gamma scale kernel rbf
		ARBOL DE DECISIÓN	ccp_alpha 0,0 criterion entropy max_depth 4 splitter random
		RANDOM FOREST	max_features sqrt n_estimators100
		BOOSTING	learning_rate 0,1 n_estimators50
		GRADIENTE BOOST	learning_rate 0,05 max_depth 1 n_estimators100
29	rs10800745 rs2486939	REGRESION LOGISTICA	C 100 penalty l1 solver liblinear
		SGD	alpha 0,0001 l1_ratio 0,025 max_iter 100 penalty l1
		SVM	C 50 gamma scale kernel poly
		ARBOL DE DECISIÓN	ccp_alpha 0,0 criterion gini max_depth 2 splitter random
		RANDOM FOREST	max_features sqrt n_estimators10
		BOOSTING	learning_rate 0,1 n_estimators50
		GRADIENTE BOOST	learning_rate 0,01 max_depth 1 n_estimators200
30	rs10800745 rs1198329	REGRESION LOGISTICA	C 100 penalty l1 solver liblinear
		SGD	alpha 0,0001 l1_ratio 0,15 max_iter 5 penalty l2
		SVM	C 50 gamma scale kernel sigmoid
		ARBOL DE DECISIÓN	ccp_alpha 0,0 criterion gini max_depth 1 splitter best
		RANDOM FOREST	max_features sqrt n_estimators10
		BOOSTING	learning_rate 0,1 n_estimators50
		GRADIENTE BOOST	learning_rate 0,01 max_depth 1 n_estimators100
31	rs8079726 rs10800745 rs2486939	REGRESION LOGISTICA	C 0,1 penalty l2 solver newton-cg
		SGD	alpha 0,0001 l1_ratio 0,025 max_iter 100 penalty elasticnet
		SVM	C 50 gamma scale kernel poly
		ARBOL DE DECISIÓN	ccp_alpha 0,0 criterion entropy max_depth 3 splitter random
		RANDOM FOREST	max_features sqrt n_estimators100
		BOOSTING	learning_rate 1,0 n_estimators100
		GRADIENTE BOOST	learning_rate 0,1 max_depth 2 n_estimators300
32	rs8079726 rs10800745 rs1198329	REGRESION LOGISTICA	C 1,0 penalty l1 solver liblinear
		SGD	alpha 0,0001 l1_ratio 0,15 max_iter 1000 penalty l2
		SVM	C 10 gamma scale kernel rbf
		ARBOL DE DECISIÓN	ccp_alpha 0,0 criterion gini max_depth 1 splitter best
		RANDOM FOREST	max_features sqrt n_estimators10
		BOOSTING	learning_rate 0,1 n_estimators50
		GRADIENTE BOOST	learning_rate 0,01 max_depth 3 n_estimators200
33	rs10800745 rs10061385 rs573057	REGRESION LOGISTICA	C 0,01 penalty l2 solver liblinear
		SGD	alpha 0,01 l1_ratio 0,05 max_iter 1000 penalty l1
		SVM	C 50 gamma scale kernel rbf
		ARBOL DE DECISIÓN	ccp_alpha 0,0 criterion gini max_depth 4 splitter best
		RANDOM FOREST	max_features sqrt n_estimators100
		BOOSTING	learning_rate 1,0 n_estimators50
		GRADIENTE BOOST	learning_rate 0,01 max_depth 3 n_estimators100
34	rs8079726 rs10800745 rs10061385	REGRESION LOGISTICA	C 1,0 penalty l2 solver liblinear
		SGD	alpha 0,0001 l1_ratio 0,05 max_iter 1000 penalty l2
		SVM	C 1,0 gamma scale kernel rbf
		ARBOL DE DECISIÓN	ccp_alpha 0,0 criterion gini max_depth 3 splitter random
		RANDOM FOREST	max_features log2 n_estimators10
		BOOSTING	learning_rate 0,1 n_estimators50
		GRADIENTE BOOST	learning_rate 0,01 max_depth 1 n_estimators300
35	rs10800745 rs2486939 rs10061385	REGRESION LOGISTICA	C 10 penalty l2 solver liblinear
		SGD	alpha 0,0001 l1_ratio 0,05 max_iter 1000 penalty l2
		SVM	C 10 gamma scale kernel sigmoid
		ARBOL DE DECISIÓN	ccp_alpha 0,0 criterion gini max_depth 2 splitter best
		RANDOM FOREST	max_features log2 n_estimators10
		BOOSTING	learning_rate 0,1 n_estimators50
		GRADIENTE BOOST	learning_rate 0,01 max_depth 1 n_estimators100
36	rs10800745 rs10061385	REGRESION LOGISTICA	C 1,0 penalty l1 solver liblinear
		SGD	alpha 0,001 l1_ratio 0,025 max_iter 1 penalty l2
		SVM	C 50 gamma scale kernel poly
		ARBOL DE DECISIÓN	ccp_alpha 0,0 criterion gini max_depth 2 splitter best
		RANDOM FOREST	max_features sqrt n_estimators100
		BOOSTING	learning_rate 0,1 n_estimators50
		GRADIENTE BOOST	learning_rate 0,01 max_depth 1 n_estimators100
37	rs10800745 rs4859120 rs10061385	REGRESION LOGISTICA	C 1,0 penalty l1 solver saga
		SGD	alpha 0,01 l1_ratio 0,05 max_iter 10000 penalty elasticnet
		SVM	C 1,0 gamma scale kernel poly
		ARBOL DE DECISIÓN	ccp_alpha 0,0 criterion gini max_depth 2 splitter best
		RANDOM FOREST	max_features sqrt n_estimators10
		BOOSTING	learning_rate 0,1 n_estimators100
		GRADIENTE BOOST	learning_rate 0,01 max_depth 2 n_estimators100
		REGRESION LOGISTICA	C 0,01 penalty l2 solver liblinear

Tabla 24 continua de la pagina anterior.

38	rs10800745 rs10061385 rs1198329	SGD	alpha 0,001 l1_ratio 0,025 max_iter 100 penalty l1
		SVM	C 50 gamma scale kernel poly
		ARBOL DE DECISIÓN	ccp_alpha 0,0 criterion gini max_depth 4 splitter random
		RANDOM FOREST	max_features log2 n_estimators10
		BOOSTING	learning_rate 0,1 n_estimators50
		GRADIENTE BOOST	learning_rate 0,05 max_depth 1 n_estimators200

En la tabla 25 se presentan los resultados conseguidos en la optimización de los 91 experimentos del conjunto de experimentos número 2. De acuerdo a los resultados conseguidos los algoritmos para los diferentes experimentos muestran solver lineal, arboles de decisión con poca profundidad, bosques aleatorios con numero de estimadores pequeños en la mayoría de los casos y tasas de aprendizaje de 0.1 para Boosting y 0.01 para Gradient Boost.

Tabla 25: Resultados obtenidos en la búsqueda de hiperparámetros para el conjunto de experimentos 2.

Subconjunto	SNPs	Algoritmo	Mejor Modelo
5	rs3813355 rs11259260 rs9518363 rs10851935 rs720680 rs4074608 rs11087039	REGRESION LOGISTICA	C 0,1 penalty none solver sag
		SGD	alpha 0,01 l1_ratio 0,15 max_iter 100 penalty l2
		SVM	C 1,0 gamma scale kernel rbf
		ARBOL DE DECISIÓN	ccp_alpha 0,1 criterion entropy max_depth 4 splitter random
		RANDOM FOREST	max_features sqrt n_estimators100
		BOOSTING	learning_rate 0,1 n_estimators50
		GRADIENTE BOOST	learning_rate 0,01 max_depth 4 n_estimators100
6	rs10800745 rs3813355 rs1198329 rs9518363 rs10851935 rs4074608 rs11667789	REGRESION LOGISTICA	C 1,0 penalty l2 solver liblinear
		SGD	alpha 0,1 l1_ratio 0,15 max_iter 1 penalty l1
		SVM	C 10 gamma scale kernel sigmoid
		ARBOL DE DECISIÓN	ccp_alpha 0,0 criterion gini max_depth 4 splitter random
		RANDOM FOREST	max_features log2 n_estimators10
		BOOSTING	learning_rate 0,1 n_estimators200
		GRADIENTE BOOST	learning_rate 0,01 max_depth 3 n_estimators100
7	rs10061385 rs11259260 rs1198329 rs9513426 rs9518363 rs8042254 rs720680	REGRESION LOGISTICA	C 1,0 penalty l1 solver liblinear
		SGD	alpha 0,01 l1_ratio 0,025 max_iter 1 penalty l1
		SVM	C 10 gamma scale kernel sigmoid
		ARBOL DE DECISIÓN	ccp_alpha 0,0 criterion entropy max_depth 4 splitter best
		RANDOM FOREST	max_features sqrt n_estimators10
		BOOSTING	learning_rate 0,1 n_estimators100
		GRADIENTE BOOST	learning_rate 0,01 max_depth 4 n_estimators100
8	rs4859120 rs10061385 rs3813355 rs11101913 rs1198329 rs9518363 rs2379991	REGRESION LOGISTICA	C 1,0 penalty l1 solver liblinear
		SGD	alpha 0,1 l1_ratio 0,05 max_iter 1000 penalty elasticnet
		SVM	C 1,0 gamma scale kernel rbf
		ARBOL DE DECISIÓN	ccp_alpha 0,0 criterion gini max_depth 4 splitter best
		RANDOM FOREST	max_features sqrt n_estimators10
		BOOSTING	learning_rate 0,5 n_estimators200
		GRADIENTE BOOST	learning_rate 0,05 max_depth 2 n_estimators100
22	rs11120748 rs4859120 rs10061385 rs3813355 rs11259260 rs1198329 rs922533	REGRESION LOGISTICA	C 0,1 penalty l2 solver liblinear
		SGD	alpha 0,01 l1_ratio 0,025 max_iter 1 penalty l2
		SVM	C 1,0 gamma scale kernel sigmoid
		ARBOL DE DECISIÓN	ccp_alpha 0,0 criterion entropy max_depth 3 splitter random
		RANDOM FOREST	max_features log2 n_estimators10
		BOOSTING	learning_rate 0,1 n_estimators50
		GRADIENTE BOOST	learning_rate 0,05 max_depth 1 n_estimators200
39	rs9518363	REGRESION LOGISTICA	C 100 penalty l1 solver liblinear
		SGD	alpha 0,001 l1_ratio 0,15 max_iter 100 penalty l1
		SVM	C 50 gamma scale kernel rbf
		ARBOL DE DECISIÓN	ccp_alpha 0,0 criterion gini max_depth 1 splitter best
		RANDOM FOREST	max_features sqrt n_estimators10
		BOOSTING	learning_rate 0,1 n_estimators50
		GRADIENTE BOOST	learning_rate 0,01 max_depth 1 n_estimators100
40	rs1198329	REGRESION LOGISTICA	C 100 penalty l1 solver liblinear
		SGD	alpha 0,0001 l1_ratio 0,05 max_iter 1 penalty elasticnet
		SVM	C 50 gamma scale kernel rbf
		ARBOL DE DECISIÓN	ccp_alpha 0,0 criterion gini max_depth 1 splitter best
		RANDOM FOREST	max_features sqrt n_estimators10
		BOOSTING	learning_rate 0,1 n_estimators50
		GRADIENTE BOOST	learning_rate 0,01 max_depth 1 n_estimators300
		REGRESION LOGISTICA	C 100 penalty l1 solver liblinear
		SGD	alpha 0,0001 l1_ratio 0,15 max_iter 10000 penalty elasticnet
		SVM	C 50 gamma scale kernel poly

Tabla 25 continua de la pagina anterior.

41	rs3813355	ARBOL DE DECISIÓN	ccp_alpha 0,0 criterion entropy max_depth 3 splitter random
		RANDOM FOREST	max_features sqrt n_estimators100
		BOOSTING	learning_rate 0,1 n_estimators50
		GRADIENTE BOOST	learning_rate 0,01 max_depth 1 n_estimators100
42	rs9518363 rs1198329	REGRESION LOGISTICA	C 100 penalty l1 solver liblinear
		SGD	alpha 0,01 l1_ratio 0,15 max_iter 10000 penalty l2
		SVM	C 50 gamma scale kernel rbf
		ARBOL DE DECISIÓN	ccp_alpha 0,0 criterion entropy max_depth 2 splitter random
		RANDOM FOREST	max_features sqrt n_estimators1000
		BOOSTING	learning_rate 0,1 n_estimators50
		GRADIENTE BOOST	learning_rate 0,01 max_depth 3 n_estimators200
43	rs9518363 rs3813355	REGRESION LOGISTICA	C 0,1 penalty l2 solver newton-cg
		SGD	alpha 0,001 l1_ratio 0,025 max_iter 10000 penalty elasticnet
		SVM	C 50 gamma scale kernel poly
		ARBOL DE DECISIÓN	ccp_alpha 0,0 criterion gini max_depth 3 splitter random
		RANDOM FOREST	max_features sqrt n_estimators10
		BOOSTING	learning_rate 0,1 n_estimators200
		GRADIENTE BOOST	learning_rate 0,01 max_depth 2 n_estimators100
44	rs3813355 rs1198329	REGRESION LOGISTICA	C 100 penalty l1 solver liblinear
		SGD	alpha 0,0001 l1_ratio 0,05 max_iter 5 penalty l2
		SVM	C 1,0 gamma scale kernel sigmoid
		ARBOL DE DECISIÓN	ccp_alpha 0,0 criterion gini max_depth 1 splitter random
		RANDOM FOREST	max_features sqrt n_estimators1000
		BOOSTING	learning_rate 0,1 n_estimators50
		GRADIENTE BOOST	learning_rate 0,01 max_depth 1 n_estimators100
45	rs10061385 rs1198329	REGRESION LOGISTICA	C 100 penalty l1 solver liblinear
		SGD	alpha 0,01 l1_ratio 0,025 max_iter 1000 penalty l1
		SVM	C 50 gamma scale kernel rbf
		ARBOL DE DECISIÓN	ccp_alpha 0,0 criterion gini max_depth 2 splitter best
		RANDOM FOREST	max_features sqrt n_estimators10
		BOOSTING	learning_rate 0,1 n_estimators50
		GRADIENTE BOOST	learning_rate 0,01 max_depth 1 n_estimators200
46	rs11259260	REGRESION LOGISTICA	C 100 penalty none solver newton-cg
		SGD	alpha 0,0001 l1_ratio 0,15 max_iter 10000 penalty l1
		SVM	C 50 gamma scale kernel poly
		ARBOL DE DECISIÓN	ccp_alpha 0,0 criterion gini max_depth 2 splitter random
		RANDOM FOREST	max_features sqrt n_estimators10
		BOOSTING	learning_rate 0,1 n_estimators50
		GRADIENTE BOOST	learning_rate 0,01 max_depth 1 n_estimators100

En la tabla 26 se presentan los resultados conseguidos en la optimización de los 84 experimentos del conjunto de experimentos número 3. Se observa que en la mayoria de los subconjunto el solver liblinear es común para la regresión logística en varios subconjuntos. Lo que se debe principalmente a que muchos SNPs son comunes en los múltiples subconjuntos. Para los algoritmos de ensamble Random Forest, Boosting y Gradient Boost en los diferentes subconjuntos, el número de estimadores esta entre 100 y 1000 y las tasa de aprendizaje son de 0.1 y 0.01.

Tabla 26: Resultados obtenidos en la búsqueda de hiperparámetros para el conjunto de experimentos 3.

Subconjunto	SNPs	Algoritmo	Mejor Modelo
9	rs12644497 rs1480149 rs3813355 rs11101913 rs9513426 rs4074608 rs8079726	REGRESION LOGISTICA	C 100 penalty l1 solver saga
		SGD	alpha 0,001 l1_ratio 0,025 max_iter 10000 penalty l2
		SVM	C 50 gamma scale kernel rbf
		ARBOL DE DECISIÓN	ccp_alpha 0,0 criterion gini max_depth 2 splitter random
		RANDOM FOREST	max_features log2 n_estimators100
		BOOSTING	learning_rate 0,1 n_estimators50
		GRADIENTE BOOST	learning_rate 0,05 max_depth 2 n_estimators200
10	rs2486939 rs4859120 rs12644497 rs2929079 rs9804548 rs12283577 rs8079726	REGRESION LOGISTICA	C 100 penalty l2 solver newton-cg
		SGD	alpha 0,0001 l1_ratio 0,15 max_iter 100 penalty l2
		SVM	C 50 gamma scale kernel sigmoid
		ARBOL DE DECISIÓN	ccp_alpha 0,0 criterion entropy max_depth 4 splitter random
		RANDOM FOREST	max_features log2 n_estimators100
		BOOSTING	learning_rate 1,0 n_estimators50
		GRADIENTE BOOST	learning_rate 0,1 max_depth 1 n_estimators300
		REGRESION LOGISTICA	C 10 penalty l1 solver saga

Tabla 26 continua de la pagina anterior.

11	rs4859120 rs11101913 rs10851935 rs922533 rs17598590 rs11667789 rs11087039	SGD	alpha 0,0001 l1_ratio 0,05 max_iter 10 penalty l2
		SVM	C 10 gamma scale kernel sigmoid
		ARBOL DE DECISIÓN	ccp_alpha 0,0 criterion entropy max_depth 4 splitter random
		RANDOM FOREST	max_features sqrt n_estimators1000
		BOOSTING	learning_rate 0,1 n_estimators100
		GRADIENTE BOOST	learning_rate 0,1 max_depth 4 n_estimators100
12	rs10800745 rs2486939 rs4859120 rs12644497 rs11101913 rs11026669 rs8079726	REGRESION LOGISTICA	C 1,0 penalty l2 solver liblinear
		SGD	alpha 0,0001 l1_ratio 0,025 max_iter 1 penalty l1
		SVM	C 10 gamma scale kernel sigmoid
		ARBOL DE DECISIÓN	ccp_alpha 0,0 criterion entropy max_depth 4 splitter random
		RANDOM FOREST	max_features sqrt n_estimators100
		BOOSTING	learning_rate 0,1 n_estimators100
		GRADIENTE BOOST	learning_rate 0,1 max_depth 4 n_estimators100
23	rs2486939 rs4859120 rs12644497 rs2929079 rs12283577 rs8079726	REGRESION LOGISTICA	C 100 penalty l1 solver liblinear
		SGD	alpha 0,0001 l1_ratio 0,05 max_iter 10 penalty l1
		SVM	C 50 gamma scale kernel sigmoid
		ARBOL DE DECISIÓN	ccp_alpha 0,0 criterion entropy max_depth 2 splitter random
		RANDOM FOREST	max_features sqrt n_estimators10
		BOOSTING	learning_rate 0,1 n_estimators100
		GRADIENTE BOOST	learning_rate 0,01 max_depth 2 n_estimators200
47	rs4859120	REGRESION LOGISTICA	C 100 penalty l2 solver saga
		SGD	alpha 0,001 l1_ratio 0,025 max_iter 10000 penalty l2
		SVM	C 0,1 gamma scale kernel rbf
		ARBOL DE DECISIÓN	ccp_alpha 0,0 criterion gini max_depth 1 splitter random
		RANDOM FOREST	max_features sqrt n_estimators10
		BOOSTING	learning_rate 0,1 n_estimators50
		GRADIENTE BOOST	learning_rate 0,01 max_depth 1 n_estimators100
48	rs8079726 rs12644497	REGRESION LOGISTICA	C 1,0 penalty l1 solver liblinear
		SGD	alpha 0,0001 l1_ratio 0,05 max_iter 1 penalty l2
		SVM	C 50 gamma scale kernel rbf
		ARBOL DE DECISIÓN	ccp_alpha 0,0 criterion gini max_depth 2 splitter random
		RANDOM FOREST	max_features sqrt n_estimators100
		BOOSTING	learning_rate 0,1 n_estimators50
		GRADIENTE BOOST	learning_rate 0,01 max_depth 3 n_estimators100
49	rs8079726 rs2486939 rs4859120 rs12644497	REGRESION LOGISTICA	C 0,1 penalty l2 solver liblinear
		SGD	alpha 0,0001 l1_ratio 0,05 max_iter 10000 penalty l1
		SVM	C 50 gamma scale kernel sigmoid
		ARBOL DE DECISIÓN	ccp_alpha 0,0 criterion entropy max_depth 2 splitter random
		RANDOM FOREST	max_features log2 n_estimators10
		BOOSTING	learning_rate 0,1 n_estimators50
		GRADIENTE BOOST	learning_rate 0,01 max_depth 1 n_estimators300
50	rs11101913	REGRESION LOGISTICA	C 100 penalty none solver newton-cg
		SGD	alpha 0,0001 l1_ratio 0,05 max_iter 10 penalty l2
		SVM	C 50 gamma scale kernel poly
		ARBOL DE DECISIÓN	ccp_alpha 0,0 criterion gini max_depth 1 splitter best
		RANDOM FOREST	max_features sqrt n_estimators10
		BOOSTING	learning_rate 1,0 n_estimators50
		GRADIENTE BOOST	learning_rate 0,01 max_depth 1 n_estimators200
51	rs12283577 rs2486939 rs12644497 rs8079726 rs2929079 rs4859120	REGRESION LOGISTICA	C 100 penalty l1 solver liblinear
		SGD	alpha 0,0001 l1_ratio 0,05 max_iter 5 penalty l2
		SVM	C 50 gamma scale kernel sigmoid
		ARBOL DE DECISIÓN	ccp_alpha 0,0 criterion gini max_depth 2 splitter random
		RANDOM FOREST	max_features sqrt n_estimators10
		BOOSTING	learning_rate 0,1 n_estimators100
		GRADIENTE BOOST	learning_rate 0,01 max_depth 2 n_estimators200
52	rs4859120 rs11101913	REGRESION LOGISTICA	C 100 penalty l1 solver liblinear
		SGD	alpha 0,001 l1_ratio 0,05 max_iter 100 penalty l1
		SVM	C 50 gamma scale kernel poly
		ARBOL DE DECISIÓN	ccp_alpha 0,0 criterion gini max_depth 2 splitter best
		RANDOM FOREST	max_features sqrt n_estimators10
		BOOSTING	learning_rate 0,1 n_estimators50
		GRADIENTE BOOST	learning_rate 0,05 max_depth 3 n_estimators200
53	rs8079726 rs12644497 rs11101913	REGRESION LOGISTICA	C 100 penalty l1 solver liblinear
		SGD	alpha 0,0001 l1_ratio 0,05 max_iter 100 penalty l1
		SVM	C 50 gamma scale kernel sigmoid
		ARBOL DE DECISIÓN	ccp_alpha 0,0 criterion entropy max_depth 3 splitter random
		RANDOM FOREST	max_features log2 n_estimators10
		BOOSTING	learning_rate 0,1 n_estimators50
		GRADIENTE BOOST	learning_rate 0,05 max_depth 1 n_estimators300

En la tabla 27 se presentan los resultados conseguidos en la optimización de los 84 experimentos del conjunto de experimentos número 4.

Tabla 27: Resultados obtenidos en la búsqueda de hiperparámetros para el conjunto de experimentos 4.

Subconjunto	SNPs	Algoritmo	Mejor Modelo
13	rs12023541 rs1529833 rs6886361 rs6590735 rs2302688 rs3497 rs4576883	REGRESION LOGISTICA	C 0,1 penalty l2 solver liblinear
		SGD	alpha 0,001 l1_ratio 0,025 max_iter 1 penalty l2
		SVM	C 10 gamma scale kernel sigmoid
		ARBOL DE DECISIÓN	ccp_alpha 0,0 criterion entropy max_depth 4 splitter best
		RANDOM FOREST	max_features sqrt n_estimators10000
		BOOSTING	learning_rate 0,1 n_estimators100
		GRADIENTE BOOST	learning_rate 0,1 max_depth 4 n_estimators300
14	rs12023541 rs6886361 rs9292869 rs6590735 rs2302688 rs3497 rs9510008	REGRESION LOGISTICA	C 100 penalty l1 solver liblinear
		SGD	alpha 0,0001 l1_ratio 0,025 max_iter 10 penalty l1
		SVM	C 10 gamma scale kernel sigmoid
		ARBOL DE DECISIÓN	ccp_alpha 0,0 criterion entropy max_depth 4 splitter best
		RANDOM FOREST	max_features sqrt n_estimators10
		BOOSTING	learning_rate 0,1 n_estimators200
		GRADIENTE BOOST	learning_rate 0,05 max_depth 2 n_estimators300
15	rs12023541 rs1529833 rs6886361 rs6999746 rs6590735 rs2302688 rs9510008	REGRESION LOGISTICA	C 100 penalty l1 solver liblinear
		SGD	alpha 0,0001 l1_ratio 0,05 max_iter 5 penalty l2
		SVM	C 50 gamma scale kernel poly
		ARBOL DE DECISIÓN	ccp_alpha 0,0 criterion entropy max_depth 4 splitter best
		RANDOM FOREST	max_features log2 n_estimators10
		BOOSTING	learning_rate 1,0 n_estimators50
		GRADIENTE BOOST	learning_rate 0,05 max_depth 2 n_estimators100
16	rs12023541 rs1529833 rs6886361 rs6590735 rs2302688 rs4576883 rs9510008	REGRESION LOGISTICA	C 0,1 penalty l2 solver newton-cg
		SGD	alpha 0,0001 l1_ratio 0,15 max_iter 1 penalty l2
		SVM	C 10 gamma scale kernel poly
		ARBOL DE DECISIÓN	ccp_alpha 0,0 criterion entropy max_depth 4 splitter best
		RANDOM FOREST	max_features log2 n_estimators100
		BOOSTING	learning_rate 0,1 n_estimators100
		GRADIENTE BOOST	learning_rate 0,01 max_depth 4 n_estimators100
24	rs6590735 rs2302688	REGRESION LOGISTICA	C 100 penalty l1 solver liblinear
		SGD	alpha 0,0001 l1_ratio 0,05 max_iter 10000 penalty elasticnet
		SVM	C 1,0 gamma scale kernel sigmoid
		ARBOL DE DECISIÓN	ccp_alpha 0,0 criterion gini max_depth 1 splitter random
		RANDOM FOREST	max_features log2 n_estimators10
		BOOSTING	learning_rate 0,1 n_estimators200
		GRADIENTE BOOST	learning_rate 0,01 max_depth 4 n_estimators100
55	rs12023541 rs6590735 rs6886361 rs2302688	REGRESION LOGISTICA	C 100 penalty l1 solver liblinear
		SGD	alpha 0,0001 l1_ratio 0,05 max_iter 1000 penalty l1
		SVM	C 50 gamma scale kernel rbf
		ARBOL DE DECISIÓN	ccp_alpha 0,0 criterion entropy max_depth 4 splitter best
		RANDOM FOREST	max_features sqrt n_estimators1000
		BOOSTING	learning_rate 0,5 n_estimators100
		GRADIENTE BOOST	learning_rate 0,05 max_depth 3 n_estimators200
56	rs12023541 rs9510008 rs6590735 rs2302688 rs6886361	REGRESION LOGISTICA	C 100 penalty l1 solver liblinear
		SGD	alpha 0,001 l1_ratio 0,05 max_iter 100 penalty l2
		SVM	C 50 gamma scale kernel rbf
		ARBOL DE DECISIÓN	ccp_alpha 0,0 criterion entropy max_depth 4 splitter random
		RANDOM FOREST	max_features log2 n_estimators10
		BOOSTING	learning_rate 1,0 n_estimators200
		GRADIENTE BOOST	learning_rate 0,05 max_depth 2 n_estimators300
57	rs12023541 rs1529833 rs6590735 rs2302688 rs6886361	REGRESION LOGISTICA	C 100 penalty l1 solver liblinear
		SGD	alpha 0,0001 l1_ratio 0,05 max_iter 5 penalty elasticnet
		SVM	C 10 gamma scale kernel sigmoid
		ARBOL DE DECISIÓN	ccp_alpha 0,0 criterion gini max_depth 4 splitter best
		RANDOM FOREST	max_features sqrt n_estimators10
		BOOSTING	learning_rate 1,0 n_estimators200
		GRADIENTE BOOST	learning_rate 0,1 max_depth 3 n_estimators100
58	rs12023541 rs6590735 rs2302688 rs3497 rs6886361	REGRESION LOGISTICA	C 100 penalty l1 solver liblinear
		SGD	alpha 0,001 l1_ratio 0,15 max_iter 5 penalty elasticnet
		SVM	C 10 gamma scale kernel sigmoid
		ARBOL DE DECISIÓN	ccp_alpha 0,0 criterion gini max_depth 3 splitter random
		RANDOM FOREST	max_features log2 n_estimators10
		BOOSTING	learning_rate 0,1 n_estimators100
		GRADIENTE BOOST	learning_rate 0,05 max_depth 2 n_estimators100
59	rs12023541 rs9510008 rs1529833 rs6590735 rs2302688 rs6886361	REGRESION LOGISTICA	C 0,1 penalty l2 solver newton-cg
		SGD	alpha 0,001 l1_ratio 0,05 max_iter 5 penalty elasticnet
		SVM	C 50 gamma scale kernel poly
		ARBOL DE DECISIÓN	ccp_alpha 0,0 criterion entropy max_depth 4 splitter best
		RANDOM FOREST	max_features log2 n_estimators10
		BOOSTING	learning_rate 1,0 n_estimators200
		GRADIENTE BOOST	learning_rate 0,1 max_depth 3 n_estimators100
		REGRESION LOGISTICA	C 1,0 penalty l2 solver newton-cg
		SGD	alpha 0,001 l1_ratio 0,15 max_iter 1000 penalty l2
		SVM	C 50 gamma scale kernel rbf
		ARBOL DE DECISIÓN	ccp_alpha 0,0 criterion entropy max_depth 4 splitter best
		RANDOM FOREST	max_features log2 n_estimators1000
		BOOSTING	learning_rate 0,1 n_estimators100

60	rs12023541 rs1529833 rs6590735 rs2302688 rs4576883 rs6886361		

Tabla 27 continua de la pagina anterior.

		GRADIENTE BOOST	learning_rate 0,05 max_depth 3 n_estimators100

En la tabla 28 se presentan los resultados conseguidos en la optimización de los 105 experimentos del conjunto de experimentos número 5.

Tabla 28: Resultados obtenidos en la búsqueda de hiperparámetros para el conjunto de experimentos 5.

Subconjunto	SNPs	Algoritmo	Mejor Modelo
17	rs17804991 rs12551529 rs2948167 rs2942076 rs4408399 rs4028683 rs2440327	REGRESION LOGISTICA	C 100 penalty l1 solver saga
		SGD	alpha 0,001 l1_ratio 0,05 max_iter 100 penalty l1
		SVM	C 50 gamma scale kernel sigmoid
		ARBOL DE DECISIÓN	ccp_alpha 0,0 criterion gini max_depth 4 splitter best
		RANDOM FOREST	max_features sqrt n_estimators100
		BOOSTING	learning_rate 0,5 n_estimators200
		GRADIENTE BOOST	learning_rate 0,1 max_depth 4 n_estimators100
18	rs17804991 rs10031902 rs12551529 rs2948167 rs2942076 rs4028683 rs2440327	REGRESION LOGISTICA	C 100 penalty l2 solver liblinear
		SGD	alpha 0,0001 l1_ratio 0,025 max_iter 10000 penalty l1
		SVM	C 50 gamma scale kernel sigmoid
		ARBOL DE DECISIÓN	ccp_alpha 0,0 criterion gini max_depth 3 splitter best
		RANDOM FOREST	max_features sqrt n_estimators100
		BOOSTING	learning_rate 0,5 n_estimators200
		GRADIENTE BOOST	learning_rate 0,1 max_depth 1 n_estimators300
19	rs818524 rs4708676 rs2948167 rs2942076 rs4075243 rs9318541 rs4028683	REGRESION LOGISTICA	C 100 penalty none solver sag
		SGD	alpha 0,001 l1_ratio 0,05 max_iter 10000 penalty l1
		SVM	C 10 gamma scale kernel sigmoid
		ARBOL DE DECISIÓN	ccp_alpha 0,0 criterion entropy max_depth 3 splitter random
		RANDOM FOREST	max_features log2 n_estimators10
		BOOSTING	learning_rate 0,1 n_estimators50
		GRADIENTE BOOST	learning_rate 0,01 max_depth 3 n_estimators200
20	rs818524 rs530160 rs12551529 rs11187729 rs2948167 rs2942076 rs4028683	REGRESION LOGISTICA	C 10 penalty l2 solver liblinear
		SGD	alpha 0,001 l1_ratio 0,15 max_iter 10000 penalty l2
		SVM	C 10 gamma scale kernel sigmoid
		ARBOL DE DECISIÓN	ccp_alpha 0,0 criterion gini max_depth 4 splitter best
		RANDOM FOREST	max_features sqrt n_estimators1000
		BOOSTING	learning_rate 0,1 n_estimators50
		GRADIENTE BOOST	learning_rate 0,1 max_depth 3 n_estimators300
25	rs17804991 rs10031902 rs12551529 rs2942076 rs4028683	REGRESION LOGISTICA	C 10 penalty l2 solver newton-cg
		SGD	alpha 0,0001 l1_ratio 0,15 max_iter 1 penalty l2
		SVM	C 10 gamma scale kernel rbf
		ARBOL DE DECISIÓN	ccp_alpha 0,0 criterion gini max_depth 2 splitter random
		RANDOM FOREST	max_features sqrt n_estimators10000
		BOOSTING	learning_rate 0,1 n_estimators100
		GRADIENTE BOOST	learning_rate 0,01 max_depth 2 n_estimators200
61	rs2942076 rs4028683	REGRESION LOGISTICA	C 0,01 penalty none solver sag
		SGD	alpha 0,0001 l1_ratio 0,15 max_iter 10 penalty l1
		SVM	C 0,1 gamma scale kernel rbf
		ARBOL DE DECISIÓN	ccp_alpha 0,0 criterion gini max_depth 2 splitter random
		RANDOM FOREST	max_features log2 n_estimators10
		BOOSTING	learning_rate 0,5 n_estimators200
		GRADIENTE BOOST	learning_rate 0,01 max_depth 1 n_estimators100
62	rs2942076 rs4028683 rs2948167	REGRESION LOGISTICA	C 100 penalty l1 solver saga
		SGD	alpha 0,0001 l1_ratio 0,15 max_iter 5 penalty l1
		SVM	C 50 gamma scale kernel rbf
		ARBOL DE DECISIÓN	ccp_alpha 0,0 criterion gini max_depth 3 splitter best
		RANDOM FOREST	max_features log2 n_estimators10
		BOOSTING	learning_rate 1,0 n_estimators100
		GRADIENTE BOOST	learning_rate 0,01 max_depth 1 n_estimators100
63	rs2942076 rs4028683 rs12551529	REGRESION LOGISTICA	C 100 penalty l1 solver liblinear
		SGD	alpha 0,0001 l1_ratio 0,025 max_iter 5 penalty l1
		SVM	C 50 gamma scale kernel sigmoid
		ARBOL DE DECISIÓN	ccp_alpha 0,0 criterion entropy max_depth 2 splitter random
		RANDOM FOREST	max_features sqrt n_estimators10
		BOOSTING	learning_rate 0,1 n_estimators50
		GRADIENTE BOOST	learning_rate 0,05 max_depth 1 n_estimators200
64	rs12551529 rs2942076 rs4028683 rs2948167	REGRESION LOGISTICA	C 100 penalty l1 solver saga
		SGD	alpha 0,01 l1_ratio 0,025 max_iter 1 penalty elasticnet
		SVM	C 50 gamma scale kernel poly
		ARBOL DE DECISIÓN	ccp_alpha 0,0 criterion gini max_depth 3 splitter best
		RANDOM FOREST	max_features sqrt n_estimators10000
		BOOSTING	learning_rate 1,0 n_estimators50
		GRADIENTE BOOST	learning_rate 0,01 max_depth 2 n_estimators200

Tabla 28 continua de la pagina anterior.

65	rs17804991 rs2942076 rs4028683 rs12551529	REGRESION LOGISTICA	C 100 penalty l1 solver liblinear
		SGD	alpha 0,0001 l1_ratio 0,025 max_iter 5 penalty l2
		SVM	C 50 gamma scale kernel sigmoid
		ARBOL DE DECISIÓN	ccp_alpha 0,0 criterion gini max_depth 2 splitter random
		RANDOM FOREST	max_features log2 n_estimators10
		BOOSTING	learning_rate 0,1 n_estimators50
		GRADIENTE BOOST	learning_rate 0,01 max_depth 2 n_estimators200
66	rs4028683 rs2942076 rs2948167	REGRESION LOGISTICA	C 100 penalty l1 solver saga
		SGD	alpha 0,001 l1_ratio 0,025 max_iter 5 penalty l2
		SVM	C 50 gamma scale kernel rbf
		ARBOL DE DECISIÓN	ccp_alpha 0,0 criterion gini max_depth 3 splitter best
		RANDOM FOREST	max_features log2 n_estimators10
		BOOSTING	learning_rate 1,0 n_estimators50
		GRADIENTE BOOST	learning_rate 0,01 max_depth 1 n_estimators100
67	rs4028683 rs2942076 rs2948167 rs12551529	REGRESION LOGISTICA	C 100 penalty l1 solver saga
		SGD	alpha 0,001 l1_ratio 0,025 max_iter 5 penalty elasticnet
		SVM	C 50 gamma scale kernel poly
		ARBOL DE DECISIÓN	ccp_alpha 0,0 criterion gini max_depth 3 splitter best
		RANDOM FOREST	max_features sqrt n_estimators10
		BOOSTING	learning_rate 1,0 n_estimators50
		GRADIENTE BOOST	learning_rate 0,01 max_depth 2 n_estimators200
68	rs17804991 rs2942076 rs2948167 rs12551529 rs2440327 rs4028683	REGRESION LOGISTICA	C 100 penalty l1 solver saga
		SGD	alpha 0,0001 l1_ratio 0,15 max_iter 10000 penalty elasticnet
		SVM	C 50 gamma scale kernel sigmoid
		ARBOL DE DECISIÓN	ccp_alpha 0,0 criterion gini max_depth 3 splitter best
		RANDOM FOREST	max_features sqrt n_estimators100
		BOOSTING	learning_rate 0,5 n_estimators200
		GRADIENTE BOOST	learning_rate 0,1 max_depth 1 n_estimators300
69	rs10031902 rs17804991 rs2942076 rs12551529 rs4028683	REGRESION LOGISTICA	C 10 penalty l2 solver newton-cg
		SGD	alpha 0,0001 l1_ratio 0,15 max_iter 1 penalty l2
		SVM	C 10 gamma scale kernel rbf
		ARBOL DE DECISIÓN	ccp_alpha 0,0 criterion gini max_depth 2 splitter random
		RANDOM FOREST	max_features log2 n_estimators10000
		BOOSTING	learning_rate 0,1 n_estimators100
		GRADIENTE BOOST	learning_rate 0,01 max_depth 2 n_estimators200
70	rs4028683 rs2942076 rs2948167 rs818524	REGRESION LOGISTICA	C 100 penalty l1 solver liblinear
		SGD	alpha 0,0001 l1_ratio 0,15 max_iter 100 penalty l2
		SVM	C 50 gamma scale kernel sigmoid
		ARBOL DE DECISIÓN	ccp_alpha 0,0 criterion gini max_depth 3 splitter best
		RANDOM FOREST	max_features sqrt n_estimators10
		BOOSTING	learning_rate 1,0 n_estimators50
		GRADIENTE BOOST	learning_rate 0,01 max_depth 1 n_estimators100

Los diferentes conjuntos de datos terminaron requiriendo diferentes hiperparámetros para obtener el mejor rendimiento de un algoritmo de aprendizaje automático. Subconjuntos de datos con SNPs similares terminaron con los mismos hiperparámetros. La búsqueda de los hiperarámetros para los 470 modelos tardo 1 hora y 50 minutos usando la herramienta de Google Colab gratuita.

8. ESTIMACIÓN DE LOS MEJORES MODELOS

8.1. Entrenamiento de los modelos optimizados

En este capítulo se muestra los experimentos realizados luego de la identificación de los hiperparámetros más influyentes y la evaluación del rendimiento de los modelos implementando los mismos siete algoritmos usados en los experimento iniciales que se detallaron en el capitulo 6, en el entrenamiento de estos modelos se realiza una validación cruzada de 5 divisiones para una estimación más precisa del rendimiento. En los resultados se muestra la métrica **F1 score**. El entrenamiento y validación de los 490 modelos tarda un tiempo total de 55 minutos usando la herramienta de Google Colab en su versión gratuita.

8.1.1. Conjunto de Experimentos 1

En la tabla 29 se observan las métrica **F1 score** de todos los modelos del conjunto de experimentos 1. Todos los dataframe analizados cuentan con 50 registros en el conjunto de entrenamiento de los cuales 31 son curas y 19 fallas, así como 22 registros en el conjunto de prueba con 16 curas y 6 fallas. Al realizar la estimación con los mejores parámetros encontrados y con validación cruzada en cinco divisiones. De 42 modelos que tuvieron un buen desempeño en los experimentos de base detallados en el capitulo 6 en 6 modelos persiste el buen desempeño después de la optimización de hiperparámetros. Sin embargo, en todos los modelos se evidencia una mejora considerable en la predicción balanceadas tanto de curas como de fallas. Para observar todas las métricas de desempeño de estos modelos puede remitirse al anexo C.

Tabla 29: Resultado de la métrica de desempeño **F1 score** obtenidas para el conjunto de experimentos 1 con los mejores hiperparámetros.

Subconjunto	SNPs	RL	SGD	SVM	DT	RF	BT	GB
1	rs10800745 rs12644497 rs10061385 rs1198329 rs922533 rs4074608 rs8079726	0.70	0.61	0.66	0.44	0.82	0.74	0.70
2	rs10800745 rs2486939 rs2929079 rs7917410 rs1198329 rs9518363 rs8079726	0.62	0.43	0.51	0.48	0.41	0.58	0.48
3	rs10800745 rs2486939 rs10061385 rs573057 rs10771313 rs12910606 rs8079726	0.58	0.51	0.54	0.61	0.58	0.66	0.54
4	rs10800745 rs4859120 rs10061385 rs573057 rs11101913 rs2379991 rs8042254	0.69	0.69	0.79	0.74	0.69	0.69	0.69
21	rs10800745 rs2486939 rs4859120 rs10061385 rs11259260 rs1198329	0.69	0.65	0.69	0.74	0.74	0.69	0.69

26	rs10800745	0.54	0.42	0.54	0.54	0.54	0.54	0.54
27	rs10800745 rs10061385	0.65	0.54	0.70	0.70	0.70	0.70	0.70
28	rs8079726 rs10800745	0.40	0.40	0.40	0.54	0.54	0.54	0.54
29	rs10800745 rs2486939	0.51	0.51	0.58	0.54	0.54	0.54	0.54
30	rs10800745 rs1198329	0.58	0.42	0.40	0.54	0.79	0.79	0.79
31	rs8079726 rs10800745 rs2486939	0.45	0.44	0.54	0.58	0.54	0.54	0.51
32	rs8079726 rs10800745 rs1198329	0.48	0.54	0.48	0.54	0.44	0.79	0.48
33	rs10800745 rs10061385 rs573057	0.70	0.70	0.66	0.70	0.70	0.70	0.70
34	rs8079726 rs10800745 rs10061385	0.48	0.66	0.40	0.34	0.54	0.70	0.70
35	rs10800745 rs2486939 rs10061385	0.51	0.47	0.54	0.54	0.61	0.61	0.61
36	rs10800745 rs10061385	0.65	0.54	0.70	0.70	0.70	0.70	0.70
37	rs10800745 rs4859120 rs10061385	0.69	0.69	0.69	0.69	0.66	0.70	0.69
38	rs10800745 rs10061385 rs1198329	0.74	0.65	0.61	0.77	0.77	0.79	0.70

Los subconjuntos con los mejores desempeños son los número 1, 4, 30, 32 y 38 en siete modelos resaltados en color rosa de la tabla 29. En su mayoría fueron construidos con algoritmos basado en técnicas de ensamblaje que combinan múltiples modelos base para generar un modelo final **RF**, **BT** y **GB**. Así como un modelo con **SVM** en el subconjunto número 4. El mejor modelo se consigue con el subconjunto 1 y el algoritmo **RF** con parámetros de 100 arboles y una elección de características de la raíz cuadrada del número de SNPs del subconjunto, consiguiendo un F1 score de 0.82, una precisión de 0.84 y un recall de 0.80. Evidenciando que la optimización del modelo permite conseguir un mejor balance entre la capacidad que tiene el modelo para clasificar tanto fallas como curas en nuevos datos. Este subconjunto de características se obtiene de la reducción de la dimensionalidad por eliminación recursiva usando un estimador SVC.

En los restantes seis modelos dados en los subconjuntos 4, 30, 32 y 38 resaltados en la tabla 29 aunque el F1 score es de 0.79 se presenta un desbalance entre la predicción de curas y fallas; la precisión ronda el 0.92 mientras el recall un 0.75 mostrando que disminuye considerablemente la predicción de las curas respecto a los resultados encontrados en los experimentos de base.

8.1.2. Conjunto de Experimentos 2

En la tabla 30 se observan resaltados en color rosa 14 modelos que presentan una mejora substancial en el desempeño respecto a los experimentos de base realizados en el capitulo 6. Los subconjuntos que tienen los mejores modelos son el número 6, 8, 40, y 45. El desempeño en F1 score igual a 1 ocurre con el modelo que se construye en el subconjunto 6 y algoritmo **SVM**, en este modelo todas las curas y fallas se predicen de manera acertada. En los restantes 13 modelos el F1 score es igual a 0.92, 2 de ellos en el subconjunto de características número 6, 1 en el subconjunto de características número 40 y 5 modelos

en el subconjunto de características número 45. Adicionalmente al F1 score de 0.92 la precisión y el recall en los modelos se encuentra entre 0.92 y 0.94 lo que permite evidenciar que en estos modelos existe un equilibrio para clasificar tanto fallas como curas en nuevos datos. Los resultados de estas métricas pueden verse en detalle en el apéndice C.

Aunque estos modelos presentan la métrica mas alta de desempeño, otros 22 modelos también presentan mejoras considerables en su desempeño aunque no consiguieron métricas tan altas.

Tabla 30: Resultado de la métrica de desempeño **F1 score** obtenidas para el conjunto de experimentos 2 con los mejores hiperparámetros.

Subconjunto	SNPs	RL	SGD	SVM	DT	RF	BT	GB
5	rs3813355 rs11259260 rs9518363 rs10851935 rs720680 rs4074608 rs11087039	0.84	0.84	0.64	0.61	0.84	0.61	0.61
6	rs10800745 rs3813355 rs1198329 rs9518363 rs10851935 rs4074608 rs11667789	0.92	0.82	1.00	0.54	0.85	0.92	0.75
7	rs10061385 rs11259260 rs1198329 rs9513426 rs9518363 rs8042254 rs720680	0.64	0.64	0.75	0.75	0.64	0.75	0.64
8	rs4859120 rs10061385 rs3813355 rs11101913 rs1198329 rs9518363 rs2379991	0.71	0.71	0.71	0.64	0.92	0.58	0.64
22	rs11120748 rs4859120 rs10061385 rs3813355 rs11259260 rs1198329 rs922533	0.75	0.64	0.82	0.75	0.84	0.75	0.84
39	rs9518363	0.61	0.38	0.64	0.64	0.64	0.64	0.64
40	rs1198329	0.82	0.82	0.92	0.92	0.92	0.92	0.92
41	rs3813355	0.75	0.75	0.75	0.69	0.69	0.69	0.69
42	rs9518363 rs1198329	0.82	0.82	0.82	0.82	0.82	0.64	0.92
43	rs9518363 rs3813355	0.71	0.71	0.71	0.68	0.68	0.68	0.71
44	rs3813355 rs1198329	0.82	0.82	0.75	0.75	0.84	0.84	0.84
45	rs10061385 rs1198329	0.82	0.85	0.92	0.92	0.92	0.92	0.92
46	rs11259260	0.51	0.51	0.51	0.51	0.51	0.51	0.51

De los 14 modelos resaltados, 8 obedecen a técnicas combinan múltiples modelos base para generar un modelo final **RF**, **BT** y **GB**. 2 modelos con **DT**, 3 con **SVM** y un último con **RL**.

El modelo obtenido con **SMV** y el subconjunto número 6 tiene un C = 10, un valor alto que se uso para penalizar más los errores de clasificación y que resulta en un modelo más complejo, probablemente más ajustado a los datos de entrenamiento. Así como, un kernel sigmoid usado para mapear los datos de entrada a un espacio de características de mayor dimensión, lo que sugiere que se ha realizado una transformación no lineal de los datos facilitando la captura de relaciones más complejas entre las características. Debido al alto valor en las métricas para este modelo, se puede apreciar que ha aprendido bien los patrones y es capaz de hacer predicciones precisas. Este subconjunto de características número 6 se obtuvo de la reducción de la dimensionalidad por eliminación recursiva con un estimador de RL.

El modelo obtenido con **RL** del subconjunto número 6 fue optimizado con una regularización de los coeficiente de manera moderada obligándolos hacer pequeños y usando una función iterativa de descenso de gradiente para minimizar la función de costo del modelo. Este fue uno de los dos mejores modelos obtenidos en la sección 6.2.2, posterior a la optimización tiene un balance entre la predicción de curas y fallas mas adecuado para nuevos datos.

Los restantes dos modelos **SVM** se consiguen con el subconjunto 40 y el subconjunto 45, su optimización muestra una moderada regularización en la penalización de los errores mas grandes. En este mismo par de subconjunto de características se obtienen dos modelos de **DT** optimizados de manera sencilla en lo que no fue necesario aplicar poda, tomando la mejor división en cada nodo del árbol basada en criterios probabilísticos y que en generar tuvo muy poca profundidad.

De los restantes 8 modelos, 3 de ellos se construyen con el algoritmo **RF** en los subconjuntos 8, 40 y 45. En los tres modelos es común el SNP rs1198329 y el valor de los parámetros de optimización que se utiliza; se tiene que el número máximo de características que se deben considerar en cada división de árbol es la raíz cuadrada del número total de características en cada división y que el número de 10 arboles es suficientemente alto para proporcionar una buena precisión sin incurrir en un costo computacional excesivo.

Los mismo tres subconjuntos con el algoritmo **BT** entregan tres modelos adicionales con una optimización donde cada modelo previo contribuye en una fracción de 0,1 a la predicción final. En cuanto a la cantidad de modelos se puede observar que entre mas características se tiene se deben plantear mas modelos en la secuencia; para la característica que tiene el segundo subgrupo se usan 50 modelos, mientras para el tercer subconjunto con dos características se usan 100 modelos y para el primero con siete características se deben usar 200 modelos.

Los últimos dos modelos se construyen con el subconjunto 40 y el subconjunto 45. En ambos se implementa un algoritmo **GB** donde cada nuevo modelo se enfoca en corregir el 1 % de los errores del modelo anterior, en cada árbol se plantea una sola rama y 200 ó 300 modelos dependiendo si es el subconjunto 40 o 45.

En todo estos subconjuntos están presentes los SNPs rs10800745, rs10851935, rs4074608, rs11667789, rs4859120, rs10061385, rs3813355, rs11101913, rs1198329, rs9518363 y rs2379991. Es muy relevante el resultado conseguido con el SNP rs1198329 (subconjunto 40) que presenta excelentes desempeños en todos los siete modelos construidos. Para revisar en detalle todas la métricas de desempeño se puede consultar el apéndice C.

8.1.3. Conjunto de Experimentos 3

Los modelos construidos en los experimentos iniciales del capitulo 6 para este conjunto de experimentos presentaban en su gran mayoría muy buenos desempeños. Después de la optimización de hiperparámetros los modelos continúan con un excelente desempeño. Se consiguen 40 modelos en 8 subconjuntos de características con buenos desempeños. Los modelos se resaltan en color rosa en la tabla 31.

Los mejores modelo conseguidos en los experimentos de base mostrados en la tabla 21 del capitulo 6 se consiguieron con el subconjunto de características número 10, 23, 49 y 51 que tenían demasiados SNPs en común como el rs2486939, rs4859120, rs12644497, rs2929079, rs12283577 y rs8079726 y lo que explicaría el porque la mayoría de estos modelos tienen métricas tan cercanas. El subconjunto de características 10 se obtuvo con una eliminación recursiva y un estimador de RL, el 23 se obtuvo mediante regresión LASSO, el 49 y 51 se debe a la intersección de los subconjuntos 10, 12 y 23. Después de la optimización de los modelos estos subconjunto continuaron entregando los mejores modelos, mostrando consistencia con resultados obtenidos en lo experimentos de base.

Tabla 31: Resultado de la métrica de desempeño **F1 score** obtenidas para el conjunto de experimentos 3 con los mejores hiperparámetros.

Subconjunto	SNPs	RL	SGD	SVM	DT	RF	BT	GB
9	rs12644497 rs1480149 rs3813355 rs11101913 rs9513426 rs4074608 rs8079726	0.86	0.68	0.68	0.58	0.68	0.40	0.58
10	rs2486939 rs4859120 rs12644497 rs2929079 rs9804548 rs12283577 rs8079726	1.00	1.00	1.00	1.00	1.00	1.00	1.00

11	rs4859120 rs11101913 rs10851935 rs922533 rs17598590 rs11667789 rs11087039	0.58	0.58	0.58	0.68	0.40	0.58	0.68
12	rs10800745 rs2486939 rs4859120 rs12644497 rs11101913 rs11026669 rs8079726	1.00	0.88	0.88	0.68	0.86	0.68	0.68
23	rs2486939 rs4859120 rs12644497 rs2929079 rs12283577 rs8079726	1.00	1.00	1.00	0.88	1.00	1.00	1.00
47	rs4859120	0.40	0.88	0.40	0.88	0.88	0.40	0.40
48	rs8079726 rs12644497	0.40	0.88	0.86	0.86	0.86	0.86	0.86
49	rs8079726 rs2486939 rs4859120 rs12644497	1.00	1.00	0.40	0.88	1.00	1.00	1.00
50	rs11101913	0.68	0.68	0.40	0.68	0.40	0.68	0.40
51	rs12283577 rs2486939 rs12644497 rs8079726 rs2929079 rs4859120	1.00	1.00	1.00	0.77	1.00	1.00	1.00
52	rs4859120 rs11101913	0.58	0.58	0.68	0.68	0.68	0.58	0.68
53	rs8079726 rs12644497 rs11101913	0.68	0.68	0.68	0.40	0.40	0.40	0.40

El subconjunto 47 que obedece al SNP rs4859120, intersección de los subconjuntos 10, 11, 12 y23 presentaba buenos desempeños en los experimentos de base mostrados en la sección 6.2.3, luego de la optimización resultan tres modelos con una mejora substancial en el desempeño, construidos con **SGD**, **DT** y **RF**. Especialmente los últimos dos que presentan parámetros con valores similares a los ya conseguidos en otros subconjuntos, lo que no implica un mayor costo computacional. En cuanto al modelo construido con **SGD**, su desempeño inicial mostraba que fue el peor modelo de todos los 84 desarrollados en este conjunto de experimentos, en la optimización sus métricas mejoraron bastante, sin embargo, al centrarse en el valor máximo del número de iteraciones que corresponde a 10000 puede presentarse un inconveniente a considerar dado que este modelo podría tomar un gran tiempo de entrenamiento y un costo computacional considerable, si bien el conjunto actual de datos no es muy numeroso, en un posible despliegue e implementación el modelo podría reentrenarse con conjuntos mas numerosos y esto implicaría un posible inconveniente.

Los SNPs rs2486939, rs4859120, rs12644497, rs2929079, rs12283577 y rs807972 comunes en muchos de estos subconjuntos tendrian la capacidad de proporcionar una explicación sólida y precisa de los datos, capturando patrones esenciales y brindando una comprensión profunda y completa de su naturaleza.

8.1.4. Conjunto de Experimentos 4

Los resultados de los experimentos de base mostrados en el capitulo 6 detallaban cuatro modelos con buenos desempeños, todos basado en un algoritmo **RF** y en los subconjuntos de características 15, 16, 57 y 59. En este conjunto de experimentos existe un desbalance de clases, por lo que la clasificación de las curas fue aprendida de manera preferente por los modelos.

Después de realizado el ajuste de hiperparámetros y la estimación de los modelos con los mejores hiperparámetros, los cuatro modelos de los experimentos iniciales con buen desempeño presentan un desempeño menor del presentado en los experimentos de base. Ahora, el mejor modelo conseguido obedece al conformado por el subconjunto 15 y el algoritmo de **DT** como se observa en la tabla 32.

Tabla 32: Resultado de la métrica de desempeño **F1 score** obtenidas para el conjunto de experimentos 4 con los mejores hiperparámetros.

Subconjunto	SNPs	RL	SGD	SVM	DT	RF	BT	GB
13	rs12023541 rs1529833 rs6886361 rs6590735 rs2302688 rs3497 rs4576883	0.45	0.38	0.54	0.64	0.58	0.35	0.64
14	rs12023541 rs6886361 rs9292869 rs6590735 rs2302688 rs3497 rs9510008	0.31	0.31	0.31	0.61	0.45	0.62	0.51
15	rs12023541 rs1529833 rs6886361 rs6999746 rs6590735 rs2302688 rs9510008	0.61	0.68	0.61	0.75	0.54	0.57	0.57
16	rs12023541 rs1529833 rs6886361 rs6590735 rs2302688 rs4576883 rs9510008	0.51	0.69	0.68	0.64	0.61	0.35	0.51
24	rs6590735 rs2302688	0.61	0.61	0.61	0.61	0.64	0.61	0.64
55	rs12023541 rs6590735 rs6886361 rs2302688	0.68	0.62	0.58	0.61	0.61	0.69	0.61
56	rs12023541 rs9510008 rs6590735 rs2302688 rs6886361	0.68	0.58	0.51	0.51	0.68	0.45	0.69
57	rs12023541 rs1529833 rs6590735 rs2302688 rs6886361	0.68	0.46	0.35	0.58	0.71	0.41	0.64
58	rs12023541 rs6590735 rs2302688 rs3497 rs6886361	0.51	0.51	0.46	0.71	0.58	0.62	0.58
59	rs12023541 rs9510008 rs1529833 rs6590735 rs2302688 rs6886361	0.71	0.64	0.68	0.64	0.61	0.41	0.54

60	rs12023541 rs1529833 rs6590735 rs2302688 rs4576883 rs6886361	0.61	0.46	0.68	0.64	0.64	0.35	0.64

El modelo optimizado resulta con un F1 score de 0.75, una precisión de 0.76 y un recall de 0.74, los resultados en detalle se pueden observar en el anexo C. Estas métricas evidencian un balance adecuado para la clasificación de ambas clases. Este modelo es uno de los mas complejos conseguidos hasta el momento dado que se tienen 4 niveles en el árbol lo que permite capturar relaciones más complejas en los datos. El subconjunto número 15 se obtuvo de una reducción dela dimensionalidad mediante una eliminación recursiva con un estimador DT.

Los restantes tres modelos resaltados en color rosa son modelos con métricas altas pero que no muestran el mismo grado de balance para la clasificación de ambas clases, sus métricas altas pueden obedecer a que tienen en común los SNPs rs12023541, rs1529833, rs6886361, rs6590735 y rs2302688 con el del mejor modelo, pero sin dudas los SNPs adicionales en cada subconjunto terminan incorporando ruido al set de datos. El subconjunto número 15, es entonces, el conjunto de características con la mas alta capacidad de explicar y representar adecuadamente la estructura y las propiedades de los datos.

8.1.5. Conjunto de Experimentos 5

En la tabla 33 se resaltan 24 modelos entrenados después de la optimización de hiperparámetros. Los modelos se encuentran distribuidos en siete subconjuntos de características. Los mejores resultados de los entrenamientos iniciales mostrados en el capitulo 6 también se encontraban distribuidos en estos siete subconjuntos. El subconjunto de características número 17 junto con el algoritmo **SGD** determinan el mejor modelo conseguido. Este modelo itera en un máximo de 100 veces antes de detenerse y aplica una penalización L1 que forzó a los coeficientes hacer más dispersos en busca de una reducción del sobreajuste. Un segundo modelo, presenta un desempeño muy similar, esta construido con el subconjunto de características número 18 y el algoritmo **BT** que fue optimizado con 200 modelos base para la posterior consecución del modelo final y una tasa de aprendizaje relativamente rápida del 0.5.

Los SNPs del subconjunto número 17 y número 18 son muy similares lo que permite concluir que estos SNPs permiten explicar en buen porcentaje el fenotipo de interés y que estas características proporcionan una explicación sólida y precisa de la naturaleza de los datos. Tanto el subconjunto número 17, como el número 18 se obtienen de una reducción de la dimensionalidad usando estimadores SVC y RL respectivamente.

Tabla 33: Resultado de la métrica de desempeño **F1 score** obtenidas para el conjunto de experimentos 5 con los mejores hiperparámetros.

Subconjunto	SNPs	RL	SGD	SVM	DT	RF	BT	GB
17	rs17804991 rs12551529 rs2948167 rs2942076 rs4408399 rs4028683 rs2440327	0.77	0.88	0.75	0.68	0.68	0.86	0.68
18	rs17804991 rs10031902 rs12551529 rs2948167 rs2942076 rs4028683 rs2440327	0.77	0.77	0.65	0.75	0.86	0.88	0.75
19	rs818524 rs4708676 rs2948167 rs2942076 rs4075243 rs9318541 rs4028683	0.50	0.50	0.50	0.42	0.58	0.68	0.68
20	rs818524 rs530160 rs12551529 rs11187729 rs2948167 rs2942076 rs4028683	0.68	0.68	0.40	0.58	0.40	0.40	0.36
25	rs17804991 rs10031902 rs12551529 rs2942076 rs4028683	0.86	0.75	0.86	0.75	0.86	0.86	0.86
61	rs2942076 rs4028683	0.75	0.40	0.40	0.75	0.68	0.68	0.40
62	rs2942076 rs4028683 rs2948167	0.68	0.68	0.75	0.75	0.75	0.68	0.40
63	rs2942076 rs4028683 rs12551529	0.86	0.86	0.86	0.75	0.86	0.86	0.86
64	rs12551529 rs2942076 rs4028683 rs2948167	0.68	0.68	0.68	0.75	0.68	0.68	0.68
65	rs17804991 rs2942076 rs4028683 rs12551529	0.75	0.86	0.68	0.75	0.68	0.86	0.86
66	rs4028683 rs2942076 rs2948167	0.68	0.68	0.75	0.75	0.68	0.68	0.40
67	rs4028683 rs2942076 rs2948167 rs12551529	0.68	0.68	0.68	0.75	0.68	0.68	0.68
68	rs17804991 rs2942076 rs2948167 rs12551529 rs2440327 rs4028683	0.77	0.77	0.75	0.75	0.86	0.77	0.75
69	rs10031902 rs17804991 rs2942076 rs12551529 rs4028683	0.86	0.68	0.86	0.75	0.86	0.86	0.86
70	rs4028683 rs2942076 rs2948167 rs818524	0.68	0.68	0.68	0.75	0.40	0.68	0.40

Los subconjuntos número 25, 63, 65, 68 y 69 que aportan los restantes 22 modelos tiene SNPs muy similares entre si e incluso los SNPs rs17804991 y rs10031902 que también

están presentes en los subconjuntos 17 y 18. Sin embargo, aunque en los modelos que estos subconjuntos entregan el F1 score es de 0.86 el balance entre la capacidad de clasificar curas y fallas es mucho menos equilibrado que el de los dos primeros modelos. Estos 22 modelos tiene una variedad de valores en los parámetros que fueron optimizados, hay **RL** con algoritmos lineales y no lineales usados para la separación de los datos, así como regularización tanto L1 como L2. Hay modelos basados en **BT**, **GB** y **RF** desde los 10 hasta los 10000 estimadores, con tasas de aprendizaje muy variadas entre 0,001 y 0,1, así como profundidades en los modelos entre 1 y 2. Cinco creados con algoritmos **SGD** y **SVM** que tienen parámetros ajustados como el kernel sigmoide o rbf que se suelen usar para aumentar la dimensionalidad de los datos y conseguir separar y clasificar las clases. En los algoritmos iterados se encuentran valores de iteraciones entre 1 y 5 repeticiones. Todos las métricas para todos los modelos se pueden observar en el anexo C.

Los dos (subconjunto 17 **SGD** y subconjunto 18 **BT**) mejores modelos que se consiguieron aquí presentaron una mejora considerable en el desempeño respecto a los resultados obtenidos en los experimentos de base mostrados en la sección 6.2.5, sin embargo, estos dos mejores modelos no eran los mejores comportados en esos experimentos, lo que indica que no es necesariamente obligatorio que los modelos iniciales con una buena proyección terminen mejorando de manera considerable al realizar una búsqueda de hiperparámetros.

9. ANÁLISIS DE RESULTADOS

9.1. Análisis de resultados sobre la reducción de la dimensionalidad

Una análisis de GWAS logro filtrar los SNPs más representativos tomados mediante dos caminos diferentes, uno realizar GWAS al conjunto de datos completo y separar por raza, y otro, separar por raza para posteriormente realizar GWAS. Consiguiendo los tres datsets COMPLETO, AFRO DESCENDIENTES y NO-AFRO DESCENDIENTES con 41, 14 y 36 SNPs respectivamente, este análisis entrego resultados que se pueden interpretar como muy prometedores para el dataset NO-AFRODESCENDIENTES puesto a que los SNPs aquí seleccionados tuviesen una fuerte correlación con el fenotipo y seguidos de los SNPs conseguidos en el dataset AFRODESCENDIENTES. Un poco mas modestos fueron los resultados de los 41 SNPs seleccionados en el dataset COMPLETO y con algunas dudas sobre el posible ruido que pudiesen tener.

Una vez conseguidas estas características se asumen cinco caminos en la exploración de los datos y posterior aplicación de técnicas de aprendizaje automático para culminar con la reducción de la dimensionalidad. Para comenzar se aplicó el método de PCA, el cual no consiguió capturar la estructura de los datos, por tanto le fue imposible explicar su variabilidad, probablemente debido a una tendencia no lineal de los datos en los tres datasets o incluso a uno no correlación entre las variables.

La eliminación recursiva de características y la regresión LASSO resultaron ser mucho más efectivas para la reducción de la dimensionalidad, lo que confirma la complejidad de los diferentes dataset. En el COMPLETO el estimador de la eliminación recursiva que resultó más relevante fue el SVC y la regresión LASSO entregando una buena cantidad de los SNPs para la construcción de los diferentes subconjuntos de intersección. De los 13 subconjuntos, estas dos técnicas estuvieron presentes en siete subconjuntos.

En el dataset COMPLETO AFRO el estimador más relevante fue RF y la regresión LASSO, aportando varios SNPs en siete y cinco subconjuntos correspondientemente de ocho subconjuntos construidos con las intersecciones. En COMPLETO NO-AFRO, múltiples SNPs del estimador RF se encontraron en seis subconjuntos de los siete construidos por las intersecciones y el estimador de RL junto con la regresión LASSO aportan SNPs en cuatro de los siete subconjuntos obtenidos de las intersecciones.

Para el dataset AFRO DESCENDIENTES los estimadores mas relevantes son SVC y RF que tienen presentes SNPs en cinco de los siete subconjuntos construidos. En este caso la regresión LASSO no aporta mayor información para la construcción de estas intersecciones. Por otra parte, se puede apreciar que los SNPs son bastante parecidos en los cuatro

estimadores de la eliminación recursiva, lo que hace muy reducida las diferentes intersecciones que se pueden construir.

Para el dataset NO-AFRO DESCENDIENTES existe una mayor cantidad de subconjuntos construidos dados por las intersecciones, en total son nueve subconjuntos donde los resultados de los estimadores SVC y RL tienen en común varios SNPs.

9.2. Análisis de los resultados desde el punto de vista biológico

Una revisión de los diferentes resultados encontrados en los diversos conjuntos de experimentos por parte de la PhD. Maria Adelaida Gómez actual coordinadora e investigadora del Laboratorio de Bioquímica y Biología Molecular del CIDEIM, concluye en un subgrupo de 22 SNPs que se pueden observar en la tabla 34, las columnas etiquetadas entre 1 y 5 obedecen a los respectivos conjuntos de experimentos y el color gris de fondo en algunas celdas representa la presencia de dicho SNP en el conjunto de experimentos correspondiente.

Los criterios de selección establecidos por la Dra. Maria Adelaida para la selección de los SNPs se presentan a continuación:

- **Conjunto de Experimentos 1:** SNPs que se encontrarán en subconjuntos de características con una $F1 >= 0{,}7$.
- **Conjunto de Experimentos 2:** SNPs que se encontrarán en subconjuntos de características con una $F1 >= 0{,}7$ con un *accuracy* $>= 0{,}8$ y que estuviesen presentes en más de tres modelos de la totalidad de los entrenados.
- **Conjunto de Experimentos 3:** SNPs que se encontrarán en subconjuntos de características con una $F1 >= 0{,}7$ con un *accuracy* $>= 0{,}8$ y que estuviesen presentes en más de tres modelos de la totalidad de los entrenados.
- **Conjunto de Experimentos 4:** SNPs que se encontrarán en subconjuntos de características con una $F1 >= 0{,}7$.
- **Conjunto de Experimentos 5:** SNPs que se encontrarán en subconjuntos de características con una $F1 >= 0{,}7$ con un *accuracy* $>= 0{,}8$ y que estuviesen presentes en más de tres modelos de la totalidad de los entrenados.

N°	SNP	1	2	3	4	5	Categoría General
1	rs10031902					■	
2	rs10800745	■	■	■			Desconocida
3	rs10851935		■				Transcripción celular en respuestaa metales o xenobioticos
4	rs11101913	■		■			Respuesta inmune
5	rs11120748		■				Respuesta inmune
6	rs12551529		■			■	Transporte de proteínas
7	rs11259260	■	■				Estructura celular
8	rs12023541				■		Respuesta inmune
9	rs12644497	■		■			Transporte celular
10	rs17804991					■	Transporte de metales
11	rs2302688				■		Muerte celular
12	rs2379991	■					Cicatrización
13	rs3813355		■				Señalización celular
14	rs4074608	■	■				Expresión genica
15	rs4859120	■	■	■			Desarrollo celular
16	rs573057	■	■	■			Transcripción
17	rs6590735				■		Replicación celular
18	rs6886361				■		Muerte celular
19	rs8042254	■					Cicatrización, estructura celular
20	rs8079726	■		■			Moviemiento celular
21	rs9518363		■				Activación celular
22	rs9804548			■			División celular

Tabla 34: SNPs con interés biológico para potenciales explicaciones a la respuesta del tratamiento terapéutico.

La columna denominada Categoría General describe la función biológica del gen donde se encuentra ubicado el SNP correspondiente. La función biológica del gen se refiere al papel que desempeñan los genes en la transmisión de información y en la determinación de las características y rasgos del organismo. Es decir, son las instrucciones que se encuentran en el ADN y que guían el desarrollo y funcionamiento de la producción de proteínas, que realizan diversas funciones, como estructurar y mantener las células, participar en reacciones químicas y actuar como mensajeros en las vías de comunicación celular.

Los sistemas biológicos son multifuncionales porque están compuestos por diversos componentes y estructuras interconectadas que realizan múltiples funciones para el funcionamien-

to y la supervivencia de un organismo. Estas funciones están estrechamente relacionadas y se complementan entre sí, lo que permite un equilibrio y una eficiencia en el sistema. Hacer un análisis de conjuntos de SNPs puede ser altamente complejo e incompleto por la falta de información adicional de otro sin número de aspectos que no se encuentran incluidos en los datasets elaborados. Lo que sugiere realizar una selección de SNPs individuales a partir de los criterios definidos por la experta.

Dentro de las funcionalidades de estos 22 SNPs se encuentran los SNPs rs17804991 y rs10851935 alojados en genes asociados con el transporte de metales y la transcripción celular en respuesta a metales, bastante interesante si recordamos que los tratamientos suministrados para combatir la enfermedad son el glucantime y el pentostam, basados en antimonio pentavalente, un tipo de metaloide que actúa interfiriendo con el metabolismo de los parásitos y su capacidad de multiplicarse en el organismo humano.

Por otra parte los SNPs rs11101913, rs11120748 y rs12023541 están asociados directamente con genes cuya funcionalidad es la respuesta inmune, lo que resulta de interés dado que la respuesta inmune es la primera línea de defensa del organismo contra infecciones y otras enfermedades. En el caso de la leishmania el sistema inmunológico empieza a combatir los agentes patógenos del parásito y como resulta insuficiente en la mayoria de los casos termina comprometido, y es allí donde se hace necesario el uso del glucantime ó el pentostam para ayudar al sistema inmunológico a combatir la enfermedad. Vale aclarar que es importante que exista una sinergia entre la respuesta inmune y el medicamento. Es decir, debe complementar y potenciar la respuesta inmune, para lograr una respuesta efectiva contra la infección.

Los SNP rs12551529 y rs12644497 asociados con el transporte de proteínas y transporte celular, también resulta de interés dado que algunas proteínas actúan como enzimas; moléculas especializadas que aceleran las reacciones químicas en el organismo, funciones defensivas como anticuerpos, transporte de moléculas y sustancias dentro y fuera de las células.

El SNPs asociados con genes que tienen funciones de desarrollo, replicación, señalización, transcripción, transporte, estructura, movimiento y muerte celular (rs4859120, rs6590735, rs3813355, rs573057, rs12644497, rs11259260, rs8079726, rs2302688) resultan de interés puesto que es conocido que el parásito vive en el fagolisosoma que es un compartimento específico de las células humanas, el transporte celular hace referencia al transporte entre compartimentos de la célula y cambios en el transporte podrían estar reflejados en cambios del transporte del medicamento al compartimento en el que vive el parásito. De igual manera, cambios en la muerte celular podría indicar que mueren células donde el parásito se hospeda para evitar su multiplicación. Así como cambios en la división, activación y re-

plicación celular estarían estrechamente relacionadas con ataques directos hacia el parásito.

Por otra parte, las lecciones cutáneas son un síntoma evidente de enfermedad, por tanto, el proceso de cura de la misma debe llevar a un proceso de cicatrización y es aquí donde los SNPs rs2379991 y rs8042254 ubicados en genes con funciones biológicas de cicatrización se convierten en un resultado relevante de este estudio.

Los SNPs rs10031902 y rs10800745 no presentan una ubicación en genes con funciones biológicas conocidas, sin embargo, podrían explorarse algunos SNPs en regiones vecinas ó volcar el interés por una búsqueda de la función biológica con la que estén relacionados.

Los resultados conseguidos en los SNPs identificados con en el análisis propuesto, muestran una relación directa con implicaciones muy importantes en varias funciones biológicas relacionadas con la leishmaniasis. Sin duda, esto muestra que algoritmos de aprendizaje automático y la ciencia de datos es una herramienta muy útil en la investigación y el estudio de enfermedades como la leishmaniasis. Desde los datos genéticos y clínicos se pueden identificar patrones, correlaciones y relaciones que de otra manera pasan desapercibidos. Probablemente, la ciencia de datos puede ayudar a identificar biomarcadores o perfiles genéticos que sean predictivos de la respuesta al tratamiento, facilitando el personalizar enfoques terapéuticos.

9.3. Análisis de resultados en la etapa de predicción

El conjunto de experimento número 1 tuvo mas de 42 modelos que en los experimentos de base que presentaban buenos desempeños, luego de la optimización de los parámetros y de la estimación de los modelos con dichos hiperparámetros se obtiene el mejor modelo, en la tabla 35 se observan las métricas del modelo antes y después de la optimización y estimación de hiperparámetros para el modelo con el mejor comportamiento. Y es precisamente con el subconjunto de características número 1 que se encuentran seis modelos con buen desempeño en los experimentos de base. Los parámetros del mejor modelo resultan ser 100 arboles y un sqrt para determinar la cantidad máxima de características que se consideran al buscar la mejor división en cada nodo de un árbol.

Tabla 35: Resultados del mejor modelo conseguido en el conjunto de experimentos 1.

Subconjunto	SNPs	Algoritmo	Resultados de Métricas Experimentos de Base				Resultados de Métricas Experimentos Optimizados			
			Accuracy	Precisión	Recall	F1 Score	Accuracy	Precisión	Recall	F1 Score
1	rs10800745 rs12644497 rs10061385 rs1198329 rs922533 rs4074608 rs8079726	RF	0.77	0.82	0.88	0.85	0.86	0.84	0.80	0.82

Se observa una mejora en la precisión después de la optimización mostrando que el modelo está obteniendo más predicciones correctas en comparación con el experimento de base. Lo que muestra que la optimización le facilitó al modelo aprender patrones más precisos y consigue hacer predicciones más confiables. En cuanto a la disminución en el recall se puede afirmar que el modelo final aumento la efectividad en la clasificación de fallas lo que lo hace más eficiente en comparación con el modelo obtenido en el experimento de base. El F1 Score disminuye un poco pero muestra un mejor equilibrio entre la precisión y el recall. Este subconjunto se obtuvo después de realizar una reducción de la dimensionalidad mediante eliminación recursiva utilizando como estimador un algoritmo **SVC**.

Los SNPs rs12644497, rs4074608 y rs8079726 están ubicados en genes de transporte celular. Respecto a los SNPs rs10800745, rs10061385, rs1198329 y rs922533 no existe información clínica relevante. En los restantes seis modelos conseguidos con la optimización, cinco son subconjuntos del mejor modelo. Los SNPs adicionales en su mayoría se encuentran en regiones no codificantes o intrones exceptuando el SNP rs8042254 que se encuentra en un gen que codifica una proteína fundamental para la organización del citoesqueleto y la dinámica de los microtúbulos y filamentos de actina en las células.

En la tabla 36 se observan los resultados de los mejores modelos para el conjunto de experimentos 2 donde se construyen los datasets exclusivamente con los 42 pacientes afro descendientes y en los experimentos de base se obtuvieron mas de 45 modelos prometedores, luego de la optimización de parámetros se reducen a solo 14 modelos con altos desempeños.

Tabla 36: Resultados del mejor modelo conseguido en el conjunto de experimentos 2.

Subconjunto	SNPs	Algoritmo	Resultados de Métricas Experimentos de Base				Resultados de Métricas Experimentos Optimizados			
			Accuracy	Precisión	Recall	F1 Score	Accuracy	Precisión	Recall	F1 Score
6	rs10800745 rs12644497 rs10061385 rs1198329 rs922533 rs4074608 rs8079726	SVM	0.92	1.00	0.88	0.93	1.00	1.00	1.00	1.00
		RL	0.92	0.89	1.00	0,94	0.92	0.94	0.90	0.92
		BT	0.77	0.78	0.88	0.82	0.92	0.92	0.94	0.92
8	rs4859120 rs10061385 rs3813355 rs11101913 rs1198329 rs9518363 rs2379991	RF	0.85	0.80	1.00	0.89	0.92	0.94	0.90	0.92
40	rs1198329	SVM	0.85	0.80	1.00	0.89	0.92	0.92	0.94	0.92
		DT	0.92	1.00	0.88	0.93	0.92	0.92	0.94	0.92
		RF	0.92	1.00	0.88	0.93	0.92	0.92	0.94	0.92
		BT	0.92	1.00	0.88	0.93	0.92	0.92	0.94	0.92
		GB	0.69	0.70	0.88	0.78	0.92	0.92	0.94	0.92
45	rs10061385 rs1198329	SVM	0.77	0.78	0.88	0.82	0.92	0.92	0.94	0.92
		DT	0.85	0.88	0.88	0.88	0.92	0.92	0.94	0.92
		RF	0.85	0.88	0.88	0.88	0.92	0.92	0.94	0.92
		BT	0.85	0.88	0.88	0.88	0.92	0.92	0.94	0.92
		GB	0.85	0.88	0.88	0.88	0.92	0.92	0.94	0.92

El modelo resaltado en color rosa en la tabla 36 muestra el mejor desempeño para este

conjunto de experimentos. El subconjunto número 6 con el algoritmo **SVM** en los experimentos de base del capitulo 6 mostraron una buena precisión en la clasificación de las curas y fallas, sin embargo, las métricas como recall no se encontraban con tan buenos desempeños. Probablemente un sesgo del modelo, le impedía reconocer patrones específicos asociados con las curas ó el desequilibrio de clases. Luego de la optimización, este modelo alcanza en todas sus métricas el valor de la unidad, mostrado un excelente rendimiento en la clasificación de los SNPs para este subconjunto en particular e ilustrando una mejora considerable en el desempeño del modelo. Los dos modelos adicionales obtenidos con el subconjunto número 6 y los algoritmos **RL** y **BT** despues de la optimización presentan de forma análoga métricas; accuracy de 0.92, precisión de 0.92, recall de 0.94 y F1 score de 0.92 todas mejores a las conseguidas en los experimentos de base.

Para el modelo optimizado y construido con el subconjunto de características número 8 y el algoritmo **RF** se consiguen métricas altas y mucho mejores que las encontradas en los experimentos de base, con un accuracy de 0.92, precisión de 0.94, recall de 0.90 y F1 score de 0.92. Vale resaltar que varios de los SNPs de este subconjunto están también en el subconjunto número 6 como el rs10800745, rs1198329, rs4074608 y rs10061385 de los que se detallo antes que se encuentran ubicados en genes con función biológica asociada a la respuesta del fenotipo. La mayoría de los restantes SNPs se encuentran regiones no codificantes (intrones).

En este conjunto de experimentos es importante resaltar que el SNP rs1198329 (subconjunto 40) tiene buenos desempeños en cinco modelos, con los algoritmos **SVM**, **DT**, **RF**, **BT** y **GB**. En todos ellos la optimización presenta métricas de desempeño bastante buenas y todas mejoraron respecto a los resultados de los experimentos de base. También son evidentes las mejoras en los cinco modelos con iguales algoritmos pero en el subconjunto número 45 (SNPs rs1198329 y rs10061385). Sin embargo, estos dos SNPs no tienen un reporte clínico de la funcionalidad biológica del gen donde se encuentran ubicados, aunque, muestran una gran contribución a la explicación a la respuesta de cura o falla.

El subconjunto número 6 y 8 se obtienen de la reducción de la dimensionalidad con la eliminación recursiva de característica usando los estimadores **RL** y **RF** respectivamente. Mientras el subconjuntos 40 se consigue de las intersección de la **RL**, **DT**, **RF**, **Regresión LASSO** y el 45 de la intersección de **DT**, **RF**, **Regresión LASSO** afirmando que los SNPs rs1198329 y rs10061385 estan estrechamente relacionados con la explicación del fenotipo de interés.

Lo que respecta al conjunto de experimentos 3 donde se construyen todos los dataset con los 30 pacientes no afro descendientes el subconjunto de características número 10 con los SNPs rs2486939, rs4859120, rs12644497, rs2929079, rs9804548, rs12283577 yrs8079726

generan los mejores modelos. En los siete modelos creados la exactitud es común a la hora de clasificar curas y fallas. Desde los experimentos de base seis de los siete modelos presentaba un alto desempeño, tendencia que persiste luego de la optimización, mostrando que ahora los siete modelos clasifican de manera exacta el fenotipo de interés. En la tabla 37 se observan todas las métricas conseguidas antes y después de la optimización para este subconjunto de características.

Tabla 37: Resultados del mejor modelo conseguido en el conjunto de experimentos 3.

Subconjunto	SNPs	Algoritmo	Resultados de Métricas Experimentos de Base				Resultados de Métricas Experimentos Optimizados			
			Accuracy	Precisión	Recall	F1 Score	Accuracy	Precisión	Recall	F1 Score
10	rs2486939 rs4859120 rs12644497 rs2929079 rs9804548 rs12283577 rs8079726	RL	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
		SGD	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
		SVM	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
		DT	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
		RF	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
		BT	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
		GB	0,67	0,67	1,00	0,80	1,00	1,00	1,00	1,00

De este conjunto de SNPs la funcionalidad biológica del rs4859120 muestra que se encuentra en genes que sintetizan proteínas involucradas en el desarrollo celular, mientras el rs12644497 contribuye en el transporte celular, y el rs8079726 esta asociado con el movimiento celular. Es decir, todos están asociados con funciones que están involucradas con el transporte del medicamento que se aplica para combatir la enfermedad. Este subconjunto de características se obtuvo de la reducción directa de la dimensionalidad mediante eliminación recursiva usando el estimador de RL. El alto desempeño de estos modelos puede se base en que el conjunto de características utilizado demuestra una alta capacidad de explicar y representar adecuadamente la estructura y las propiedades de los datos. Estos SNPs fueron recurrentes en los cuatro estimadores de la reducción de diemsionalidad por eliminación recursiva y también por la Regresión LASSO lo que establece una alta presencia de los mismos en los subconjuntos sintéticos que se formaron con las diferentes intersecciones, y que termina en un alto número de subconjuntos con altos desempeños.

Estos primeros tres conjuntos de experimentos obedecían todos a una reducción de la dimensionalidad que se planteo haciendo GWAS al conjunto completo de los 72 pacientes y procediendo con las técnicas de aprendizaje automático en primera instancia al conjunto completo y luego a la división hecha a partir de la raza. En los tres experimentos son comunes los hallazgos de SNPs que se encuentran ubicados en genes con funciones de movilidad, traslación, división y transición celular lo que es un resultado interesante y se encuentra en concordancia con el hecho que el parásito causante de la leishmaniasis vive en las células del sistema inmunológico.

Para el análisis del resultados del conjunto de experimentos 4 y 5, es importante recordar que la reducción de la dimensionalidad empezó después de dividir los datos bajo el criterio

de raza en afro descendientes y no afro descendientes lo que llevo a la consecución de algunos SNPs diferentes. En el caso del conjunto de experimentos numero 4 en todos los experimentos de base explicados en el capitulo 6, las métricas resultaron bastante regulares, sin conseguir de manera óptima un modelo que pudiese ser un firme candidato a explicar los fenotipos de cura o falla. Después de realizar la optimización el mejor modelo se obtiene con el subconjunto número 15 basado en un algoritmo **DT**. Este subconjunto de características se obtuvo después de realizar una reducción de la dimensionalidad mediante eliminacón recursiva usando un estimador DT. Las métricas de desempeño logran un buen balance como se observa en la tabla 38.

Tabla 38: Resultados del mejor modelo conseguido en el conjunto de experimentos 4.

Subconjunto	SNPs	Algoritmo	Resultados de Métricas Experimentos de Base				Resultados de Métricas Experimentos Optimizados			
			Accuracy	Precisión	Recall	F1 Score	Accuracy	Precisión	Recall	F1 Score
15	rs12023541 rs1529833 rs6886361 rs6999746 rs6590735 rs2302688 rs9510008	DT	0.62	0.67	0.75	0.71	0.77	0.76	0.74	0.75

Se puede apreciar una mejora considerable en todas las métricas antes y después de la optimización del modelo. En cuanto a la funcionalidad biológica de los SNPs, el rs12023541 esta ubicado en un gen con funciones en la respuesta inmune. El rs6886361 y rs2302688 están ubicados en genes asociados con muerte celular. El rs6590735 con la transcripción celular, los restantes tres no tienen un concepto clínico de funcionalidad biológica.

Los dos mejores modelo conseguido con el conjunto de experimentos número 5 obedece al subconjunto 17 y 18 con los algoritmo **SGD** y **BT** correspondientemente, en la tabla 39 se observan los resultados conseguidos para las métricas del modelo antes y después de la optimización. Estos dos subconjuntos se obtuvieron con la eliminación recursiva usando estimadores SVC y RL.

En los experimentos de base este par de subconjuntos no obtuvieron buenos desempeños, sin embargo, en la optimización su desempeño mejoro obteniendo modelos que detectan las clases con un buen grado de confiabilidad. La función biológica asociada a los SNPs también resulta de interés, el rs17804991 esta ubicado en un gen que sintetiza proteínas involucradas en el transporte de metales, y precisamente, los medicamentos usados contra la enfermedad se basan en metaloides. El SNP rs12551529 asociado con el transporte de proteínas, el rs2948167 ubicado en un gen que codifica el desarrollo embrionario y la formación de los miembros superiores, como las manos y los brazos. El rs2440327 ubicado en un gen regulador de otros genes. De los SNPs rs2942076, rs4408399 y rs4028683 no tienen información clínica de la función biológica del gen donde se encuentran.

Tabla 39: Resultados del mejor modelo conseguido en el conjunto de experimentos 5.

Subconjunto	SNPs	Algoritmo	Resultados de Métricas Experimentos de Base				Resultados de Métricas Experimentos Optimizados			
			Accuracy	Precisión	Recall	F1 Score	Accuracy	Precisión	Recall	F1 Score
17	rs17804991 rs12551529 rs2948167 rs2942076 rs4408399 rs4028683 rs2440327	SGD	0.67	1.00	0.50	0.67	0.89	0.88	0.92	0.88
18	rs17804991 rs10031902 rs12551529 rs2948167 rs2942076 rs4028683 rs2440327	BT	0.78	0.83	0.83	0.83	0.89	0.88	0.92	0.88

A diferencia de los tres primeros conjuntos de experimentos donde los mejores modelos conseguidos con los experimentos iniciales al ser optimizados continuaron siendo los mejores modelos, con mejoras considerables. En los conjuntos de experimentos 4 y 5 los experimentos de base fueron bastante bajos en su desempeño, al realizar la optimización de los modelos, casi todos mejoraron, incluso los modelos mas irregulares consiguieron mejorar de manera considerable y entregar información importante para la explicación de los fenotipos de cura y falla como los SNPs ubicados en genes cuyas funciones son la muerte celular y el transporte de metales.

En las sección 9.2 cuando se realizó el análisis biológicos se encontraron algunos SNPs asociados con la cicatrización, ninguno de estos SNPs se encuentran en los dataset de los mejores modelos pero ambos están en un único dataset que fue el segundo mejor modelo en el conjunto de experimentos numero 1 y que están en compañía de algunos otros SNPs ubicados en genes asociados a la respuesta inmune.

En general no se consigue un mejor modelo ó un único modelo como aquel que logra explicar los fenotipos desde el punto de vista biológico, la información conseguida es extraída de la variedad de modelos construidos, entrenados y optimizados. Exclusivamente desde la ciencia de datos el modelo ganador sin duda puede ser cualquiera de los modelos conseguidos con el conjunto de experimentos 3, dado que con un único SNPs el rs1198329 se logra predecir el 92 % de los nuevos datos, adicionalmente lo hace con algoritmos relativamente sencillos como **SVM** no lineal y un **DT** de una profundidad igual a 1, como con modelos de ensamble **RF**, **BT** y **GB** que tienen entre 10 y 300 estimadores. Tanto los modelos simples como los complejos logran un rendimiento excepcional en todas las métricas evaluadas, de lo que podríamos afirmar que ambos tipos de modelos fueron capaces de aprender de manera efectiva los patrones presentes en los datos. Es decir, los más sencillos no se quedaron cortos y capturaron las principales características del dataset, y por otro lado, los más complejos aprovecharon su capacidad para aprender patrones quizás más sutiles o detalles específicos. Sin embargo, desde el contexto biológico un único SNP no representan la suficiente información para dar explicaciones. Lo que podría dejar como mejor modelo ganador al conjunto 6 con el algoritmo **SVM**,,con sus SNPs rs10800745, rs12644497, rs10061385, rs1198329, rs922533, rs4074608, rs8079726 y cuyas métricas permitieron clasificar con toda exactitud las curas y fallas.

10. CONCLUSIONES

El uso de técnicas de aprendizaje de maquina como algoritmos de clasificación y selección de características sobre el conjunto de datos que contenía información genética relevante y registros de respuesta inmune al tratamiento en pacientes con leishmaniasis, facilitó la identificación de las mutaciones genéticas más significativas relacionadas con la respuesta inmune al tratamiento terapéutico contra la leishmaniasis. Los resultados obtenidos mostraron que el enfoque propuesto fue efectivo para clasificar y detectar las mutaciones más relevantes y permitir clasificar los SNPs mas influyentes en la respuesta inmune al tratamiento.

Se realizó una exhaustiva preparación de los datos para garantizar su calidad y coherencia. Incluyendo la eliminación de datos faltantes o inconsistentes, la normalización de las características y la codificación adecuada de las variables categóricas desde técnicas especializadas en el contexto biológico como GWAS y desde la propia ciencia de datos; análisis de componentes principales, eliminación recursiva de características, regresión LASSO e intersecciones de características de los resultados de estas técnicas. Logrando reducir el número de variables a aquellas mutaciones genéticas más relevantes para la respuesta inmune al tratamiento. Esto facilitó la interpretación de los resultados y mejoró la eficiencia de los algoritmos de clasificación utilizados.

Los mejores modelos resultaron de los subconjuntos conseguidos con la eliminación recursiva de características bajo estimadores como SVC, DT y RL. La regresión LASSO en un par de oportunidades entrego subconjuntos iguales a los entregados por la eliminación recursiva. Por otra parte, los subconjuntos sintéticos construidos con las intersecciones de las diferentes técnicas de reducción de la dimensionalidad no resultaron aportando mejores modelos. En cuanto al análisis de componentes principales (PCA) no capturó de manera efectiva la variabilidad de los datos.

A través de una evaluación experimental exhaustiva de conjuntos y mutaciones identificadas como relevantes, se logró clasificar con precisión mutaciones genéticas asociadas con la respuesta inmune al tratamiento terapéutico contra la leishmaniasis. Este estudio involucró un total de 483 experimentos de base, evaluando diversas configuraciones y modelos de clasificadores, utilizando métricas de evaluación como accuracy, precisión, recall y F1 score para estimar de manera precisa su desempeño. La búsqueda de hiperparámetros mejoró aún más el rendimiento de los clasificadores, refinando los modelos y permitiendo obtener resultados más precisos y confiables. Al comparar los resultados antes y después de la optimización, se demuestra de manera concluyente la capacidad de los modelos para clasificar con precisión las mutaciones genéticas relevantes en relación con la respuesta inmune al tratamiento terapéutico.

Debido a la complejidad y multifuncionalidad de los sistemas biológicos intentar explicar la respuesta al tratamiento con un solo modelo es desconocer el hecho que el mecanismo de cura de la enfermedad obedezca a múltiples factores, complejos de por sí. Es así, como se deben definir criterios para la búsqueda y selección de SNPs individuales en los que se pueda explorar la función biológica relacionada y evaluar su responsabilidad en el éxito o fracaso del tratamiento.

La selección de los 22 SNPs mostrados en la tabla 34 es un avance significativo en la búsqueda de biomarcadores para la explicación del éxito o fracaso del tratamiento contra la leishmaniasis, si consideramos que el 90 % de ellos están asociados con actividades de respuesta inmune, acciones sobre las células de activación, transporte, muerte, división, estructura y alteraciones sobre procesos de cicatrización. Estos hallazgos respaldan la utilidad de las técnicas de aprendizaje automático en la identificación de biomarcadores genéticos para mejorar el tratamiento de esta enfermedad.

REFERENCIAS BIBLIOGRÁFICAS

[1] *Leishmaniasis*, ene. de 2022. dirección: https://www.who.int/es/news-room/fact-sheets/detail/leishmaniasis.

[2] *¿Cómo se adquiere la leishmaniasis?* Dirección: https://www.redaccionmedica.com/recursos-salud/diccionario-enfermedades/leishmaniasis.

[3] OPS/OMS, «Informe de Leishmaniasis N° 8,» OPS/OMS, inf. téc., 2018.

[4] MINSALUD, «Informe Leishmaniasis cutánea y mucosa 2018,» OPS/OMS, inf. téc., 2018. dirección: https://www.panaftosa.org/leish/inf2018_es/INFO_COL_2018_ESP.pdf.

[5] M. E. Ana Nilce S., *Manual de procedimientospara vigilancia y control de lasleishmaniasis en las Américas*, OPS/OMS, ed. OPS, 2019.

[6] *Diccionario de cáncer del NCI*. dirección: https://www.cancer.gov/espanol/publicaciones/diccionarios/diccionario-cancer/def/nucleotido.

[7] E. Prbb, *SNPs: variaciones de un tema*, jul. de 2019. dirección: https://ellipse.prbb.org/es/snps-variaciones-de-un-tema/.

[8] Alvaro, *El ciclo de vida de la ciencia de datos*, jun. de 2020. dirección: https://machinelearningparatodos.com/ciclo-vida-ciencia-de-datos/.

[9] J. G. et al., *Ciencia de datos técnicas analíticas y aprendizaje estadístico. Un enfoque práctico*, Publicaciones Altaria, Alfaomega, ed. Altaria, 2018, ISBN: 978-958-778-425-1.

[10] A. Carranza, *Data Cleansing: averigua cómo limpiar datos erróneos y conservar información valiosa*, mayo de 2022. dirección: https://www.crehana.com/blog/data-analitica/data-cleansing/.

[11] J. Bergstra, R. Bardenet, Y. Bengio y B. Kégl, «Algorithms for hyper-parameter optimization,» en *Advances in neural information processing systems*, 2011, págs. 2546-2554.

[12] P. Domingos, «A few useful things to know about machine learning,» *Communications of the ACM*, vol. 55, n.º 10, págs. 78-87, 2012. DOI: 10.1145/2347736.2347755.

[13] R. Iniesta, E. Guino y V. Moreno, «Análisis estadıstico de polimorfismos genéticos en estudios epidemiológicos,» *Gaceta sanitaria*, vol. 24, n.º 2, págs. 164-172, 2010.

[14] A. P. Morris, B. F. Voight, T. M. Teslovich et al., «Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes,» *Nature Genetics*, vol. 44, n.º 9, págs. 981-990, 2012. DOI: 10.1038/ng.2383.

[15] T. J. Eisen, R. S. Houlston, R. A. Eeles et al., «Genetic associations with lung cancer risk: meta-analyses of 15 GWASs in populations of European and East Asian ancestry,» *The Lancet Oncology*, vol. 21, n.º 8, págs. 1141-1151, 2020. DOI: 10.1016/S1470-2045(20)30317-6.

[16] C. Cotsapas, B. F. Voight, E. Rossin et al., «Pervasive sharing of genetic effects in autoimmune disease,» *PLoS Genetics*, vol. 7, n.º 8, e1002254, 2011. DOI: 10.1371/journal.pgen.1002254.

[17] J. Yang, S. H. Lee, M. E. Goddard y P. M. Visscher, «GCTA: a tool for genome-wide complex trait analysis,» *The American Journal of Human Genetics*, vol. 88, n.º 1, págs. 76-82, 2011. DOI: 10.1016/j.ajhg.2010.11.011.

[18] A. Ng, K. Li y A. Lin, *Machine Learning Yearning*. Independently published, 2017.

[19] L. Bottou, «Large-scale machine learning with stochastic gradient descent,» en *Proceedings of COMPSTAT'2010*, 2010.

[20] C. Cortes y V. Vapnik, «Support-vector networks,» *Machine Learning*, vol. 20, n.º 3, págs. 273-297, 1995.

[21] *RPubs - IntroducciÃ³n a los Modelos de ClasificaciÃ³n en R*, jun. de 2018. dirección: https://rpubs.com/rdelgado/397838.

[22] L. Breiman, «Random forests,» *Machine Learning*, vol. 45, n.º 1, págs. 5-32, 2001.

[23] Y. Freund y R. E. Schapire, «Experiments with a new boosting algorithm,» en *International Conference on Machine Learning*, 1996, págs. 148-156.

[24] J. H. Friedman, «Greedy function approximation: A gradient boosting machine,» *The Annals of Statistics*, vol. 29, n.º 5, págs. 1189-1232, 2001.

[25] M. Sokolova y G. Lapalme, «A systematic analysis of performance measures for classification tasks,» *Information Processing & Management*, vol. 45, n.º 4, págs. 427-437, 2009.

[26] J. I. B. Arce, *La matriz de confusión y sus métricas*, 2023. dirección: https://www.juanbarrios.com/la-matriz-de-confusion-y-sus-metricas/.

[27] F. Provost y T. Fawcett, «Robust classification for imprecise environments,» *Machine Learning*, vol. 42, n.º 3, págs. 203-231, 2001.

[28] *Cross-Validation en Python*, ago. de 2022. dirección: https://deepnote.com/@a_mas/Cross-Validation-en-Python-685fa851-b5b2-4c5b-b5fb-3dc5ae64838f.

[29] J. Bergstra e Y. Bengio, «Random search for hyper-parameter optimization,» *Journal of Machine Learning Research*, vol. 13, págs. 281-305, 2012.

[30] J. Snoek, H. Larochelle y R. P. Adams, «Practical bayesian optimization of machine learning algorithms,» *Advances in neural information processing systems*, vol. 25, págs. 2951-2959, 2012.

[31] *LEISHMANIASIS « cideim*. dirección: https://www.cideim.org.co/historia_visual/entendimiento-para-mitigar-el-impacto-de-la-leishmaniasis/#:%7E:text=CIDEIM%20encuentra%20m%C3%BAltiples%20evidencias%20de,el%20campo%20de%20la%20leishmaniasis..

[32] J. D. S. Mora, «Modelo predictivo para la ocurrenciade LC en Colombia, a partir devariables ambientales ysocioeconómicas,» Tesis de mtría., Universidad Nacional de Colombia, 2021. dirección: https://repositorio.unal.edu.co/bitstream/handle/unal/80442/1022385505.2021.pdf?sequence=2&isAllowed=y.

[33] C.-B. E. Sánchez-Suárez J Bernal FA, «Colombian Contributions Fighting Leishmaniasis: A Systematic Review on Antileishmanials Combined with Chemoinformatics Analysis.,» *Molecules*, vol. 25, n.º 23, dic. de 2020. DOI: 10.3390/molecules25235704. dirección: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7730898/pdf/molecules-25-05704.pdf.

[34] Z. et al., «A machine learning-based systemfor detecting leishmaniasis in microscopicimages,» *BMC Infectious Disease*, 2022. DOI: https://doi.org/10.1186/s12879-022-07029-7.

[35] M. G. at el., «Leishmaniasis Parasite Segmentation and Classification using Deep Learning,» *Springer International Publishing AG*, vol. 10945, 2018.

[36] G. S. Bhunia, S. Kesari, A. Jeyaram, V. Kumar, P. Das y R. Tandon, «Machine learning approaches for the prediction of Leishmaniasis disease susceptibility: A review,» *Acta tropica*, vol. 181, págs. 52-60, 2018. DOI: 10.1016/j.actatropica.2018.02.011.

[37] S. Das, A. Chattopadhyay y M. K. Basu, «Identification of SNPs associated with susceptibility to visceral leishmaniasis using machine learning approaches,» *BMC genomics*, vol. 19, n.º Suppl 8, pág. 797, 2018. DOI: 10.1186/s12864-018-5165-5.

[38] P. Roy, A. Dhara y C. Chakraborty, «Exploring the correlation between SNPs and leishmaniasis using machine learning techniques,» *Journal of Medical Systems*, vol. 44, n.º 4, pág. 85, 2020. DOI: 10.1007/s10916-020-01514-x.

[39] S. Shilpi, M. Majumder y C. Chakraborty, «A comprehensive analysis of SNPs associated with visceral leishmaniasis using machine learning techniques,» *PloS one*, vol. 14, n.º 5, e0217007, 2019. DOI: 10.1371/journal.pone.0217007.

[40] P. K. Singh, A. Chattopadhyay, A. Prasad y S. Sundar, «Machine learning approach for the prediction of susceptibility to visceral leishmaniasis,» *Journal of infectious diseases*, vol. 215, n.º 12, págs. 1984-1988, 2017.

[41] M. A. Gómez, *DATASET Genotypes and Clinical*, 2017. dirección: https://www.cideim.org.co/cideim/es/nosotros/nuestro-equipo/41-mariaadelaidagomez.html.

[42] C. D. Royal, «The Use of Racial, Ethnic, and Ancestral Categories in Human Genetics Research,» *American Journal of Human Genetics*, vol. 77, n.º 4, págs. 519-532, 2005. DOI: 10.1086/499346.

[43] S. Purcell. «PLINK.» English, Hospital de la Universidad de Harvard. (2007), dirección: https://zzz.bwh.harvard.edu/plink/.

[44] P. S. N. B. T.-B. K. T. L. F. M. B. D. M. J. S. P. de Bakker PIW; Daly MJ; Sham PC, «PLINK: a toolset for whole-genome association and population-based linkage analysis,» *American Journal of Human Genetics.*, vol. 81, n.º 3, págs. 559-575, 2007. DOI: 10.1086/519795.

[45] F. R. Villatoro, *La genética, la estatura y el problema de la herencia perdida - La Ciencia de la Mula Francis*, abr. de 2015. dirección: https://francis.naukas.com/2010/07/03/la-genetica-la-estatura-y-el-problema-de-la-herencia-perdida/.

A. ANEXO A

RESULTADOS CONSEGUIDOS DE TODOS LOS MODELOS ENTRENADOS

Tabla 40: Resultados de todas las metricas para el conjunto de experimentos número 1.

Subconjunto	SNPs	Algoritmo	Accuracy	Precisión	Recall	F1 Score
1	rs10800745 rs12644497 rs10061385 rs1198329 rs922533 rs4074608 rs8079726	REGRESION LOGISTICA	0.77	0.82	0.88	0.85
		SGD	0.68	0.80	0.75	0.77
		SVM	0.77	0.82	0.88	0.85
		ARBOL DE DECISIÓN	0.73	0.81	0.81	0.81
		RANDOM FOREST	0.77	0.82	0.88	0.85
		BOOSTING	0.73	0.81	0.81	0.81
		GRADIENTE BOOST	0.73	0.78	0.88	0.82
2	rs10800745 rs2486939 rs2929079 rs7917410 rs1198329 rs9518363 rs8079726	REGRESION LOGISTICA	0.68	0.80	0.75	0.77
		SGD	0.55	0.71	0.63	0.67
		SVM	0.68	0.80	0.75	0.77
		ARBOL DE DECISIÓN	0.73	0.81	0.81	0.81
		RANDOM FOREST	0.82	0.80	1.00	0.89
		BOOSTING	0.73	0.78	0.88	0.82
		GRADIENTE BOOST	0.64	0.75	0.75	0.75
3	rs10800745 rs2486939 rs10061385 rs573057 rs10771313 rs12910606 rs8079726	REGRESION LOGISTICA	0.64	0.75	0.75	0.75
		SGD	0.59	0.73	0.69	0.71
		SVM	0.59	0.73	0.69	0.71
		ARBOL DE DECISIÓN	0.73	0.81	0.81	0.81
		RANDOM FOREST	0.68	0.76	0.81	0.79
		BOOSTING	0.68	0.80	0.75	0.77
		GRADIENTE BOOST	0.64	0.75	0.75	0.75
4	rs10800745 rs4859120 rs10061385 rs573057 rs11101913 rs2379991 rs8042254	REGRESION LOGISTICA	0.82	0.80	1.00	0.89
		SGD	0.73	0.78	0.88	0.82
		SVM	0.82	0.80	1.00	0.89
		ARBOL DE DECISIÓN	0.73	0.81	0.81	0.81
		RANDOM FOREST	0.82	0.80	1.00	0.89
		BOOSTING	0.73	0.81	0.81	0.81
		GRADIENTE BOOST	0.82	0.80	1.00	0.89
21	rs10800745 rs2486939 rs4859120 rs10061385 rs11259260 rs1198329	REGRESION LOGISTICA	0.82	0.80	1.00	0.89
		SGD	0.82	0.80	1.00	0.89
		SVM	0.82	0.80	1.00	0.89
		ARBOL DE DECISIÓN	0.86	0.84	1.00	0.91
		RANDOM FOREST	0.82	0.80	1.00	0.89
		BOOSTING	0.86	0.84	1.00	0.91
		GRADIENTE BOOST	0.82	0.80	1.00	0.89
26	rs10800745	REGRESION LOGISTICA	0.59	0.77	0.63	0.69
		SGD	0.73	0.73	1.00	0.84
		SVM	0.55	0.75	0.56	0.64
		ARBOL DE DECISIÓN	0.59	0.77	0.63	0.69
		RANDOM FOREST	0.59	0.77	0.63	0.69
		BOOSTING	0.59	0.77	0.63	0.69
		GRADIENTE BOOST	0.59	0.77	0.63	0.69
		REGRESION LOGISTICA	0.77	0.79	0.94	0.86
		SGD	0.73	0.73	1.00	0.84
		SVM	0.77	0.79	0.94	0.86
		ARBOL DE DECISIÓN	0.77	0.82	0.88	0.85

Tabla 40: Resultados de todas las metricas para el conjunto de experimentos número 1.

Subconjunto	SNPs	Algoritmo	Accuracy	Precisión	Recall	F1 Score
27	rs10800745 rs10061385	RANDOM FOREST	0.77	0.82	0.88	0.85
		BOOSTING	0.77	0.82	0.88	0.85
		GRADIENTE BOOST	0.77	0.82	0.88	0.85
28	rs8079726 rs10800745	REGRESION LOGISTICA	0.50	0.67	0.63	0.65
		SGD	0.55	0.75	0.56	0.64
		SVM	0.50	0.67	0.63	0.65
		ARBOL DE DECISIÓN	0.59	0.77	0.63	0.69
		RANDOM FOREST	0.59	0.77	0.63	0.69
		BOOSTING	0.59	0.77	0.63	0.69
		GRADIENTE BOOST	0.50	0.67	0.63	0.65
29	rs10800745 rs2486939	REGRESION LOGISTICA	0.59	0.73	0.69	0.71
		SGD	0.73	0.73	1.00	0.84
		SVM	0.59	0.73	0.69	0.71
		ARBOL DE DECISIÓN	0.64	0.75	0.75	0.75
		RANDOM FOREST	0.64	0.75	0.75	0.75
		BOOSTING	0.64	0.75	0.75	0.75
		GRADIENTE BOOST	0.64	0.75	0.75	0.75
30	rs10800745 rs1198329	REGRESION LOGISTICA	0.77	0.76	1.00	0.86
		SGD	0.50	1.00	0.31	0.48
		SVM	0.77	0.76	1.00	0.86
		ARBOL DE DECISIÓN	0.86	0.84	1.00	0.91
		RANDOM FOREST	0.86	0.84	1.00	0.91
		BOOSTING	0.86	0.84	1.00	0.91
		GRADIENTE BOOST	0.77	0.76	1.00	0.86
31	rs8079726 rs10800745 rs2486939	REGRESION LOGISTICA	0.59	0.73	0.69	0.71
		SGD	0.73	0.73	1.00	0.84
		SVM	0.59	0.73	0.69	0.71
		ARBOL DE DECISIÓN	0.50	0.69	0.56	0.62
		RANDOM FOREST	0.64	0.75	0.75	0.75
		BOOSTING	0.50	0.69	0.56	0.62
		GRADIENTE BOOST	0.64	0.75	0.75	0.75
32	rs8079726 rs10800745 rs1198329	REGRESION LOGISTICA	0.55	0.71	0.63	0.67
		SGD	0.41	0.64	0.44	0.52
		SVM	0.50	0.67	0.63	0.65
		ARBOL DE DECISIÓN	0.50	0.69	0.56	0.62
		RANDOM FOREST	0.82	0.80	1.00	0.89
		BOOSTING	0.50	0.69	0.56	0.62
		GRADIENTE BOOST	0.50	0.67	0.63	0.65
33	rs10800745 rs10061385 rs573057	REGRESION LOGISTICA	0.82	0.83	0.94	0.88
		SGD	0.77	0.82	0.88	0.85
		SVM	0.77	0.79	0.94	0.86
		ARBOL DE DECISIÓN	0.77	0.82	0.88	0.85
		RANDOM FOREST	0.82	0.83	0.94	0.88
		BOOSTING	0.77	0.82	0.88	0.85
		GRADIENTE BOOST	0.77	0.82	0.88	0.85
34	rs8079726 rs10800745 rs10061385	REGRESION LOGISTICA	0.55	0.71	0.63	0.67
		SGD	0.55	0.75	0.56	0.64
		SVM	0.50	0.67	0.63	0.65
		ARBOL DE DECISIÓN	0.50	0.73	0.50	0.59
		RANDOM FOREST	0.77	0.82	0.88	0.85
		BOOSTING	0.50	0.73	0.50	0.59
		GRADIENTE BOOST	0.77	0.82	0.88	0.85
		REGRESION LOGISTICA	0.59	0.73	0.69	0.71
		SGD	0.64	0.75	0.75	0.75
		SVM	0.59	0.73	0.69	0.71
		ARBOL DE DECISIÓN	0.64	0.75	0.75	0.75
		RANDOM FOREST	0.73	0.78	0.88	0.82

Tabla 40: Resultados de todas las metricas para el conjunto de experimentos número 1.

Subconjunto	SNPs	Algoritmo	Accuracy	Precisión	Recall	F1 Score
35	rs10800745 rs2486939 rs10061385	BOOSTING	0.64	0.75	0.75	0.75
		GRADIENTE BOOST	0.73	0.78	0.88	0.82
36	rs10800745 rs10061385	REGRESION LOGISTICA	0.77	0.79	0.94	0.86
		SGD	0.73	0.73	1.00	0.84
		SVM	0.77	0.79	0.94	0.86
		ARBOL DE DECISIÓN	0.77	0.82	0.88	0.85
		RANDOM FOREST	0.77	0.82	0.88	0.85
		BOOSTING	0.77	0.82	0.88	0.85
		GRADIENTE BOOST	0.77	0.82	0.88	0.85
37	rs10800745 rs4859120 rs10061385	REGRESION LOGISTICA	0.82	0.80	1.00	0.89
		SGD	0.77	0.79	0.94	0.86
		SVM	0.82	0.80	1.00	0.89
		ARBOL DE DECISIÓN	0.82	0.80	1.00	0.89
		RANDOM FOREST	0.82	0.80	1.00	0.89
		BOOSTING	0.82	0.80	1.00	0.89
		GRADIENTE BOOST	0.82	0.80	1.00	0.89
38	rs10800745 rs10061385 rs1198329	REGRESION LOGISTICA	0.82	0.83	0.94	0.88
		SGD	0.64	0.79	0.69	0.73
		SVM	0.77	0.76	1.00	0.86
		ARBOL DE DECISIÓN	0.77	0.87	0.81	0.84
		RANDOM FOREST	0.82	0.83	0.94	0.88
		BOOSTING	0.82	0.88	0.88	0.88
		GRADIENTE BOOST	0.77	0.82	0.88	0.85

Tabla 41: Resultados de todas las métricas para el conjunto de experimentos número 2.

Subconjunto	SNPs	Algoritmo	Accuracy	Precisión	Recall	F1 Score
5	rs3813355 rs11259260 rs9518363 rs10851935 rs720680 rs4074608 rs11087039	REGRESION LOGISTICA	0.77	0.78	0.88	0.82
		SGD	0.77	0.78	0.88	0.82
		SVM	0.85	0.88	0.88	0.88
		ARBOL DE DECISIÓN	0.62	0.71	0.63	0.67
		RANDOM FOREST	0.77	0.78	0.88	0.82
		BOOSTING	0.62	0.71	0.63	0.67
		GRADIENTE BOOST	0.62	0.71	0.63	0.67
6	rs10800745 rs3813355 rs1198329 rs9518363 rs10851935 rs4074608 rs11667789	REGRESION LOGISTICA	0.92	0.89	1.00	0.94
		SGD	0.92	1.00	0.88	0.93
		SVM	0.92	1.00	0.88	0.93
		ARBOL DE DECISIÓN	0.77	0.78	0.88	0.82
		RANDOM FOREST	0.85	0.88	0.88	0.88
		BOOSTING	0.77	0.78	0.88	0.82
		GRADIENTE BOOST	0.77	0.78	0.88	0.82
7	rs10061385 rs11259260 rs1198329 rs9513426 rs9518363 rs8042254 rs720680	REGRESION LOGISTICA	0.69	0.70	0.88	0.78
		SGD	0.69	0.70	0.88	0.78
		SVM	0.77	0.78	0.88	0.82
		ARBOL DE DECISIÓN	0.77	0.78	0.88	0.82
		RANDOM FOREST	0.69	0.70	0.88	0.78
		BOOSTING	0.69	0.70	0.88	0.78
		GRADIENTE BOOST	0.69	0.70	0.88	0.78
8	rs4859120 rs10061385 rs3813355 rs11101913 rs1198329 rs9518363 rs2379991	REGRESION LOGISTICA	0.77	0.73	1.00	0.84
		SGD	0.77	0.73	1.00	0.84
		SVM	0.77	0.73	1.00	0.84
		ARBOL DE DECISIÓN	0.69	0.70	0.88	0.78
		RANDOM FOREST	0.85	0.80	1.00	0.89
		BOOSTING	0.69	0.70	0.88	0.78
		GRADIENTE BOOST	0.69	0.70	0.88	0.78
		REGRESION LOGISTICA	0.77	0.78	0.88	0.82

Tabla 41: Resultados de todas las métricas para el conjunto de experimentos número 2.

Subconjunto	SNPs	Algoritmo	Accuracy	Precisión	Recall	F1 Score
22	rs11120748 rs4859120 rs10061385 rs3813355 rs11259260 rs1198329 rs922533	SGD	0.92	0.89	1.00	0.94
		SVM	0.85	0.80	1.00	0.89
		ARBOL DE DECISIÓN	0.77	0.78	0.88	0.82
		RANDOM FOREST	0.77	0.78	0.88	0.82
		BOOSTING	0.77	0.78	0.88	0.82
		GRADIENTE BOOST	0.77	0.78	0.88	0.82
39	rs9518363	REGRESION LOGISTICA	0.62	0.71	0.63	0.67
		SGD	0.62	0.71	0.63	0.67
		SVM	0.62	0.71	0.63	0.67
		ARBOL DE DECISIÓN	0.69	0.70	0.88	0.78
		RANDOM FOREST	0.69	0.70	0.88	0.78
		BOOSTING	0.69	0.70	0.88	0.78
		GRADIENTE BOOST	0.69	0.70	0.88	0.78
40	rs1198329	REGRESION LOGISTICA	0.85	0.80	1.00	0.89
		SGD	0.62	0.62	1.00	0.76
		SVM	0.85	0.80	1.00	0.89
		ARBOL DE DECISIÓN	0.92	1.00	0.88	0.93
		RANDOM FOREST	0.92	1.00	0.88	0.93
		BOOSTING	0.92	1.00	0.88	0.93
		GRADIENTE BOOST	0.69	0.70	0.88	0.78
41	rs3813355	REGRESION LOGISTICA	0.77	0.78	0.88	0.82
		SGD	0.77	0.78	0.88	0.82
		SVM	0.77	0.78	0.88	0.82
		ARBOL DE DECISIÓN	0.69	0.83	0.63	0.71
		RANDOM FOREST	0.69	0.83	0.63	0.71
		BOOSTING	0.69	0.83	0.63	0.71
		GRADIENTE BOOST	0.69	0.83	0.63	0.71
42	rs9518363 rs1198329	REGRESION LOGISTICA	0.85	0.80	1.00	0.89
		SGD	0.62	0.71	0.63	0.67
		SVM	0.85	0.80	1.00	0.89
		ARBOL DE DECISIÓN	0.92	1.00	0.88	0.93
		RANDOM FOREST	0.85	0.80	1.00	0.89
		BOOSTING	0.92	1.00	0.88	0.93
		GRADIENTE BOOST	0.85	0.80	1.00	0.89
43	rs9518363 rs3813355	REGRESION LOGISTICA	0.77	0.73	1.00	0.84
		SGD	0.77	0.73	1.00	0.84
		SVM	0.77	0.73	1.00	0.84
		ARBOL DE DECISIÓN	0.69	0.75	0.75	0.75
		RANDOM FOREST	0.69	0.75	0.75	0.75
		BOOSTING	0.69	0.75	0.75	0.75
		GRADIENTE BOOST	0.77	0.73	1.00	0.84
44	rs3813355 rs1198329	REGRESION LOGISTICA	0.85	0.80	1.00	0.89
		SGD	0.85	0.88	0.88	0.88
		SVM	0.77	0.78	0.88	0.82
		ARBOL DE DECISIÓN	0.85	0.88	0.88	0.88
		RANDOM FOREST	0.85	0.88	0.88	0.88
		BOOSTING	0.85	0.88	0.88	0.88
		GRADIENTE BOOST	0.85	0.88	0.88	0.88
45	rs10061385 rs1198329	REGRESION LOGISTICA	0.85	0.80	1.00	0.89
		SGD	0.62	0.62	1.00	0.76
		SVM	0.85	0.80	1.00	0.89
		ARBOL DE DECISIÓN	0.92	1.00	0.88	0.93
		RANDOM FOREST	0.92	1.00	0.88	0.93
		BOOSTING	0.92	1.00	0.88	0.93
		GRADIENTE BOOST	0.92	1.00	0.88	0.93
		REGRESION LOGISTICA	0.62	0.64	0.88	0.74
		SGD	0.62	0.64	0.88	0.74

Tabla 41: Resultados de todas las métricas para el conjunto de experimentos número 2.

Subconjunto	SNPs	Algoritmo	Accuracy	Precisión	Recall	F1 Score
46	rs11259260	SVM	0.62	0.64	0.88	0.74
		ARBOL DE DECISIÓN	0.62	0.64	0.88	0.74
		RANDOM FOREST	0.62	0.64	0.88	0.74
		BOOSTING	0.62	0.64	0.88	0.74
		GRADIENTE BOOST	0.62	0.64	0.88	0.74

Tabla 42: Resultados de todas las métricas para el conjunto de experimentos número 3.

Subconjunto	SNPs	Algoritmo	Accuracy	Precisión	Recall	F1 Score
9	rs12644497 rs1480149 rs3813355 rs11101913 rs9513426 rs4074608 rs8079726	REGRESION LOGISTICA	0.78	0.75	1.00	0.86
		SGD	0.78	0.75	1.00	0.86
		SVM	0.89	1.00	0.83	0.91
		ARBOL DE DECISIÓN	0.78	0.83	0.83	0.83
		RANDOM FOREST	0.78	0.75	1.00	0.86
		BOOSTING	0.78	0.75	1.00	0.86
		GRADIENTE BOOST	0.67	0.67	1.00	0.80
10	rs2486939 rs4859120 rs12644497 rs2929079 rs9804548 rs12283577 rs8079726	REGRESION LOGISTICA	1.00	1.00	1.00	1.00
		SGD	1.00	1.00	1.00	1.00
		SVM	1.00	1.00	1.00	1.00
		ARBOL DE DECISIÓN	1.00	1.00	1.00	1.00
		RANDOM FOREST	1.00	1.00	1.00	1.00
		BOOSTING	1.00	1.00	1.00	1.00
		GRADIENTE BOOST	0.67	0.67	1.00	0.80
11	rs4859120 rs11101913 rs10851935 rs922533 rs17598590 rs11667789 rs11087039	REGRESION LOGISTICA	0.67	0.71	0.83	0.77
		SGD	0.67	0.71	0.83	0.77
		SVM	0.67	0.71	0.83	0.77
		ARBOL DE DECISIÓN	0.78	0.75	1.00	0.86
		RANDOM FOREST	0.67	0.67	1.00	0.80
		BOOSTING	0.78	0.75	1.00	0.86
		GRADIENTE BOOST	0.78	0.75	1.00	0.86
12	rs10800745 rs2486939 rs4859120 rs12644497 rs11101913 rs11026669 rs8079726	REGRESION LOGISTICA	0.78	0.75	1.00	0.86
		SGD	1.00	1.00	1.00	1.00
		SVM	0.89	1.00	0.83	0.91
		ARBOL DE DECISIÓN	0.89	1.00	0.83	0.91
		RANDOM FOREST	0.89	0.86	1.00	0.92
		BOOSTING	0.89	1.00	0.83	0.91
		GRADIENTE BOOST	0.78	0.75	1.00	0.86
23	rs2486939 rs4859120 rs12644497 rs2929079 rs12283577 rs8079726	REGRESION LOGISTICA	1.00	1.00	1.00	1.00
		SGD	1.00	1.00	1.00	1.00
		SVM	1.00	1.00	1.00	1.00
		ARBOL DE DECISIÓN	1.00	1.00	1.00	1.00
		RANDOM FOREST	1.00	1.00	1.00	1.00
		BOOSTING	1.00	1.00	1.00	1.00
		GRADIENTE BOOST	1.00	1.00	1.00	1.00
47	rs4859120	REGRESION LOGISTICA	0.67	0.67	1.00	0.80
		SGD	0.33	0.00	0.00	0.00
		SVM	0.67	0.67	1.00	0.80
		ARBOL DE DECISIÓN	0.89	1.00	0.83	0.91
		RANDOM FOREST	0.67	0.67	1.00	0.80
		BOOSTING	0.89	1.00	0.83	0.91
		GRADIENTE BOOST	0.67	0.67	1.00	0.80
		REGRESION LOGISTICA	0.67	0.67	1.00	0.80
		SGD	0.67	0.67	1.00	0.80
		SVM	0.67	0.67	1.00	0.80
		ARBOL DE DECISIÓN	0.89	0.86	1.00	0.92

Tabla 42: Resultados de todas las métricas para el conjunto de experimentos número 3.

Subconjunto	SNPs	Algoritmo	Accuracy	Precisión	Recall	F1 Score
48	rs8079726 rs12644497	RANDOM FOREST	0.89	0.86	1.00	0.92
		BOOSTING	0.89	0.86	1.00	0.92
		GRADIENTE BOOST	0.67	0.67	1.00	0.80
49	rs8079726 rs2486939 rs4859120 rs12644497	REGRESION LOGISTICA	1.00	1.00	1.00	1.00
		SGD	1.00	1.00	1.00	1.00
		SVM	1.00	1.00	1.00	1.00
		ARBOL DE DECISIÓN	0.89	1.00	0.83	0.91
		RANDOM FOREST	1.00	1.00	1.00	1.00
		BOOSTING	1.00	1.00	1.00	1.00
		GRADIENTE BOOST	0.89	0.86	1.00	0.92
50	rs11101913	REGRESION LOGISTICA	0.67	0.67	1.00	0.80
		SGD	0.56	0.67	0.67	0.67
		SVM	0.67	0.67	1.00	0.80
		ARBOL DE DECISIÓN	0.78	0.75	1.00	0.86
		RANDOM FOREST	0.78	0.75	1.00	0.86
		BOOSTING	0.78	0.75	1.00	0.86
		GRADIENTE BOOST	0.67	0.67	1.00	0.80
51	rs12283577 rs2486939 rs12644497 rs8079726 rs2929079 rs4859120	REGRESION LOGISTICA	1.00	1.00	1.00	1.00
		SGD	1.00	1.00	1.00	1.00
		SVM	1.00	1.00	1.00	1.00
		ARBOL DE DECISIÓN	1.00	1.00	1.00	1.00
		RANDOM FOREST	1.00	1.00	1.00	1.00
		BOOSTING	1.00	1.00	1.00	1.00
		GRADIENTE BOOST	1.00	1.00	1.00	1.00
52	rs4859120 rs11101913	REGRESION LOGISTICA	0.67	0.71	0.83	0.77
		SGD	0.67	0.71	0.83	0.77
		SVM	0.67	0.71	0.83	0.77
		ARBOL DE DECISIÓN	0.78	0.75	1.00	0.86
		RANDOM FOREST	0.78	0.75	1.00	0.86
		BOOSTING	0.78	0.75	1.00	0.86
		GRADIENTE BOOST	0.78	0.75	1.00	0.86
53	rs8079726 rs12644497 rs11101913	REGRESION LOGISTICA	0.78	0.75	1.00	0.86
		SGD	0.78	0.75	1.00	0.86
		SVM	0.67	0.67	1.00	0.80
		ARBOL DE DECISIÓN	0.67	0.67	1.00	0.80
		RANDOM FOREST	0.67	0.67	1.00	0.80
		BOOSTING	0.67	0.67	1.00	0.80
		GRADIENTE BOOST	0.67	0.67	1.00	0.80

Tabla 43: Resultados de todas las métricas para el conjunto de experimentos número 4.

Subconjunto	SNPs	Algoritmo	Accuracy	Precisión	Recall	F1 Score
13	rs12023541 rs1529833 rs6886361 rs6590735 rs2302688 rs3497 rs4576883	REGRESION LOGISTICA	0.38	0.50	0.50	0.50
		SGD	0.46	0.60	0.38	0.46
		SVM	0.46	0.60	0.38	0.46
		ARBOL DE DECISIÓN	0.54	0.63	0.63	0.63
		RANDOM FOREST	0.62	0.64	0.88	0.74
		BOOSTING	0.54	0.63	0.63	0.63
		GRADIENTE BOOST	0.62	0.64	0.88	0.74
		REGRESION LOGISTICA	0.38	0.50	0.38	0.43
		SGD	0.31	0.40	0.25	0.31
		SVM	0.31	0.40	0.25	0.31
		ARBOL DE DECISIÓN	0.54	0.63	0.63	0.63
		RANDOM FOREST	0.54	0.63	0.63	0.63
		BOOSTING	0.54	0.63	0.63	0.63

Tabla 43: Resultados de todas las métricas para el conjunto de experimentos número 4.

Subconjunto	SNPs	Algoritmo	Accuracy	Precisión	Recall	F1 Score
14	rs12023541 rs6886361 rs2292869 rs6590735 rs2302688 rs3497 rs9510008	GRADIENTE BOOST	0.62	0.67	0.75	0.71
15	rs12023541 rs1529833 rs6886361 rs6999746 rs6590735 rs2302688 rs9510008	REGRESION LOGISTICA	0.62	0.71	0.63	0.67
		SGD	0.54	0.67	0.50	0.57
		SVM	0.62	0.71	0.63	0.67
		ARBOL DE DECISIÓN	0.62	0.67	0.75	0.71
		RANDOM FOREST	0.69	0.67	1.00	0.80
		BOOSTING	0.62	0.67	0.75	0.71
		GRADIENTE BOOST	0.69	0.67	1.00	0.80
16	rs12023541 rs1529833 rs6886361 rs6590735 rs2302688 rs4576883 rs9510008	REGRESION LOGISTICA	0.62	0.71	0.63	0.67
		SGD	0.46	0.60	0.38	0.46
		SVM	0.62	0.71	0.63	0.67
		ARBOL DE DECISIÓN	0.62	0.67	0.75	0.71
		RANDOM FOREST	0.77	0.73	1.00	0.84
		BOOSTING	0.54	0.63	0.63	0.63
		GRADIENTE BOOST	0.62	0.64	0.88	0.74
24	rs6590735 rs2302688	REGRESION LOGISTICA	0.62	0.71	0.63	0.67
		SGD	0.62	0.71	0.63	0.67
		SVM	0.62	0.71	0.63	0.67
		ARBOL DE DECISIÓN	0.69	0.70	0.88	0.78
		RANDOM FOREST	0.69	0.70	0.88	0.78
		BOOSTING	0.69	0.70	0.88	0.78
		GRADIENTE BOOST	0.69	0.70	0.88	0.78
55	rs12023541 rs6590735 rs6886361 rs2302688	REGRESION LOGISTICA	0.69	0.75	0.75	0.75
		SGD	0.62	0.71	0.63	0.67
		SVM	0.62	0.71	0.63	0.67
		ARBOL DE DECISIÓN	0.62	0.71	0.63	0.67
		RANDOM FOREST	0.62	0.67	0.75	0.71
		BOOSTING	0.62	0.71	0.63	0.67
		GRADIENTE BOOST	0.69	0.75	0.75	0.75
56	rs12023541 rs9510008 rs6590735 rs2302688 rs6886361	REGRESION LOGISTICA	0.69	0.75	0.75	0.75
		SGD	0.54	0.75	0.38	0.50
		SVM	0.62	0.71	0.63	0.67
		ARBOL DE DECISIÓN	0.54	0.67	0.50	0.57
		RANDOM FOREST	0.62	0.67	0.75	0.71
		BOOSTING	0.54	0.67	0.50	0.57
		GRADIENTE BOOST	0.69	0.75	0.75	0.75
57	rs12023541 rs1529833 rs6590735 rs2302688 rs6886361	REGRESION LOGISTICA	0.69	0.75	0.75	0.75
		SGD	0.54	0.67	0.50	0.57
		SVM	0.62	0.71	0.63	0.67
		ARBOL DE DECISIÓN	0.54	0.63	0.63	0.63
		RANDOM FOREST	0.77	0.73	1.00	0.84
		BOOSTING	0.62	0.67	0.75	0.71
		GRADIENTE BOOST	0.69	0.67	1.00	0.80
58	rs12023541 rs6590735 rs2302688 rs3497 rs6886361	REGRESION LOGISTICA	0.54	0.63	0.63	0.63
		SGD	0.54	0.60	0.75	0.67
		SVM	0.62	0.71	0.63	0.67
		ARBOL DE DECISIÓN	0.62	0.67	0.75	0.71
		RANDOM FOREST	0.62	0.67	0.75	0.71
		BOOSTING	0.62	0.67	0.75	0.71
		GRADIENTE BOOST	0.62	0.67	0.75	0.71
59	rs12023541 rs9510008 rs1529833 rs6590735 rs2302688 rs6886361	REGRESION LOGISTICA	0.69	0.75	0.75	0.75
		SGD	0.54	0.63	0.63	0.63
		SVM	0.62	0.71	0.63	0.67
		ARBOL DE DECISIÓN	0.62	0.67	0.75	0.71
		RANDOM FOREST	0.69	0.67	1.00	0.80
		BOOSTING	0.62	0.67	0.75	0.71
		GRADIENTE BOOST	0.69	0.67	1.00	0.80

Tabla 43: Resultados de todas las métricas para el conjunto de experimentos número 4.

Subconjunto	SNPs	Algoritmo	Accuracy	Precisión	Recall	F1 Score
60	rs12023541 rs1529833 rs6590735 rs2302688 rs4576883 rs6886361	REGRESION LOGISTICA	0.62	0.71	0.63	0.67
		SGD	0.46	0.56	0.63	0.59
		SVM	0.62	0.71	0.63	0.67
		ARBOL DE DECISIÓN	0.62	0.67	0.75	0.71
		RANDOM FOREST	0.77	0.73	1.00	0.84
		BOOSTING	0.54	0.63	0.63	0.63
		GRADIENTE BOOST	0.62	0.64	0.88	0.74

Tabla 44: Resultados de todas las métricas para el conjunto de experimentos número 5.

Subconjunto	SNPs	Algoritmo	Accuracy	Precisión	Recall	F1 Score
17	rs17804991 rs12551529 rs2948167 rs2942076 rs4408399 rs4028683 rs2440327	REGRESION LOGISTICA	1.00	1.00	1.00	1.00
		SGD	0.67	1.00	0.50	0.67
		SVM	0.78	1.00	0.67	0.80
		ARBOL DE DECISIÓN	0.78	0.75	1.00	0.86
		RANDOM FOREST	0.67	0.67	1.00	0.80
		BOOSTING	0.78	0.75	1.00	0.86
		GRADIENTE BOOST	0.78	0.75	1.00	0.86
18	rs17804991 rs10031902 rs12551529 rs2948167 rs2942076 rs4028683 rs2440327	REGRESION LOGISTICA	0.89	0.86	1.00	0.92
		SGD	1.00	1.00	1.00	1.00
		SVM	0.78	1.00	0.67	0.80
		ARBOL DE DECISIÓN	0.89	0.86	1.00	0.92
		RANDOM FOREST	0.78	0.75	1.00	0.86
		BOOSTING	0.78	0.83	0.83	0.83
		GRADIENTE BOOST	0.78	0.75	1.00	0.86
19	rs818524 rs4708676 rs2948167 rs2942076 rs4075243 rs9318541 rs4028683	REGRESION LOGISTICA	0.78	0.75	1.00	0.86
		SGD	0.67	0.71	0.83	0.77
		SVM	0.78	0.75	1.00	0.86
		ARBOL DE DECISIÓN	0.78	0.75	1.00	0.86
		RANDOM FOREST	0.78	0.75	1.00	0.86
		BOOSTING	0.78	0.75	1.00	0.86
		GRADIENTE BOOST	0.78	0.75	1.00	0.86
20	rs818524 rs530160 rs12551529 rs11187729 rs2948167 rs2942076 rs4028683	REGRESION LOGISTICA	0.78	0.75	1.00	0.86
		SGD	0.78	0.75	1.00	0.86
		SVM	0.78	0.75	1.00	0.86
		ARBOL DE DECISIÓN	0.67	0.71	0.83	0.77
		RANDOM FOREST	0.67	0.67	1.00	0.80
		BOOSTING	0.67	0.71	0.83	0.77
		GRADIENTE BOOST	0.67	0.67	1.00	0.80
25	rs17804991 rs10031902 rs12551529 rs2942076 rs4028683	REGRESION LOGISTICA	0.89	0.86	1.00	0.92
		SGD	0.89	0.86	1.00	0.92
		SVM	0.78	0.83	0.83	0.83
		ARBOL DE DECISIÓN	0.89	0.86	1.00	0.92
		RANDOM FOREST	0.89	0.86	1.00	0.92
		BOOSTING	0.89	0.86	1.00	0.92
		GRADIENTE BOOST	0.78	0.75	1.00	0.86
61	rs2942076 rs4028683	REGRESION LOGISTICA	0.78	0.75	1.00	0.86
		SGD	0.78	0.83	0.83	0.83
		SVM	0.78	0.75	1.00	0.86
		ARBOL DE DECISIÓN	0.78	0.83	0.83	0.83
		RANDOM FOREST	0.78	0.75	1.00	0.86
		BOOSTING	0.78	0.83	0.83	0.83
		GRADIENTE BOOST	0.78	0.75	1.00	0.86
		REGRESION LOGISTICA	0.78	0.75	1.00	0.86
		SGD	0.78	0.83	0.83	0.83
		SVM	0.78	0.75	1.00	0.86

Tabla 44: Resultados de todas las métricas para el conjunto de experimentos número 5.

Subconjunto	SNPs	Algoritmo	Accuracy	Precisión	Recall	F1 Score
62	rs2942076 rs4028683 rs2948167	ARBOL DE DECISIÓN	0.78	0.83	0.83	0.83
		RANDOM FOREST	0.78	0.75	1.00	0.86
		BOOSTING	0.78	0.83	0.83	0.83
		GRADIENTE BOOST	0.67	0.67	1.00	0.80
63	rs2942076 rs4028683 rs12551529	REGRESION LOGISTICA	0.89	0.86	1.00	0.92
		SGD	0.78	0.75	1.00	0.86
		SVM	0.78	0.75	1.00	0.86
		ARBOL DE DECISIÓN	0.89	0.86	1.00	0.92
		RANDOM FOREST	0.78	0.75	1.00	0.86
		BOOSTING	0.89	0.86	1.00	0.92
		GRADIENTE BOOST	0.78	0.75	1.00	0.86
64	rs12551529 rs2942076 rs4028683 rs2948167	REGRESION LOGISTICA	0.78	0.75	1.00	0.86
		SGD	0.78	0.75	1.00	0.86
		SVM	0.78	0.75	1.00	0.86
		ARBOL DE DECISIÓN	0.89	0.86	1.00	0.92
		RANDOM FOREST	0.78	0.75	1.00	0.86
		BOOSTING	0.89	0.86	1.00	0.92
		GRADIENTE BOOST	0.78	0.75	1.00	0.86
65	rs17804991 rs2942076 rs4028683 rs12551529	REGRESION LOGISTICA	0.89	0.86	1.00	0.92
		SGD	0.78	0.83	0.83	0.83
		SVM	0.78	0.83	0.83	0.83
		ARBOL DE DECISIÓN	0.89	0.86	1.00	0.92
		RANDOM FOREST	0.89	0.86	1.00	0.92
		BOOSTING	0.89	0.86	1.00	0.92
		GRADIENTE BOOST	0.78	0.75	1.00	0.86
66	rs4028683 rs2942076 rs2948167	REGRESION LOGISTICA	0.78	0.75	1.00	0.86
		SGD	0.67	0.71	0.83	0.77
		SVM	0.78	0.75	1.00	0.86
		ARBOL DE DECISIÓN	0.78	0.83	0.83	0.83
		RANDOM FOREST	0.78	0.75	1.00	0.86
		BOOSTING	0.78	0.83	0.83	0.83
		GRADIENTE BOOST	0.67	0.67	1.00	0.80
67	rs4028683 rs2942076 rs2948167 rs12551529	REGRESION LOGISTICA	0.78	0.75	1.00	0.86
		SGD	0.78	0.75	1.00	0.86
		SVM	0.78	0.75	1.00	0.86
		ARBOL DE DECISIÓN	0.89	0.86	1.00	0.92
		RANDOM FOREST	0.78	0.75	1.00	0.86
		BOOSTING	0.89	0.86	1.00	0.92
		GRADIENTE BOOST	0.78	0.75	1.00	0.86
68	rs17804991 rs2942076 rs2948167 rs12551529 rs2440327 rs4028683	REGRESION LOGISTICA	0.89	0.86	1.00	0.92
		SGD	0.78	1.00	0.67	0.80
		SVM	0.78	1.00	0.67	0.80
		ARBOL DE DECISIÓN	0.89	0.86	1.00	0.92
		RANDOM FOREST	0.78	0.75	1.00	0.86
		BOOSTING	0.89	0.86	1.00	0.92
		GRADIENTE BOOST	0.78	0.75	1.00	0.86
69	rs10031902 rs17804991 rs2942076 rs12551529 rs4028683	REGRESION LOGISTICA	0.89	0.86	1.00	0.92
		SGD	0.78	1.00	0.67	0.80
		SVM	0.78	0.83	0.83	0.83
		ARBOL DE DECISIÓN	0.89	0.86	1.00	0.92
		RANDOM FOREST	0.89	0.86	1.00	0.92
		BOOSTING	0.89	0.86	1.00	0.92
		GRADIENTE BOOST	0.78	0.75	1.00	0.86
		REGRESION LOGISTICA	0.78	0.75	1.00	0.86
		SGD	0.78	0.75	1.00	0.86
		SVM	0.78	0.75	1.00	0.86
		ARBOL DE DECISIÓN	0.78	0.75	1.00	0.86

Tabla 44: Resultados de todas las métricas para el conjunto de experimentos número 5.

Subconjunto	SNPs	Algoritmo	Accuracy	Precisión	Recall	F1 Score
5	rs4028683 rs2942076 rs2948167 rs818524	RANDOM FOREST	0.67	0.67	1.00	0.80
		BOOSTING	0.78	0.75	1.00	0.86
		GRADIENTE BOOST	0.67	0.67	1.00	0.80

B. ANEXO B

HIPERPARÁMETROS AJUSTADOS

B.1. Regresión Logística

Solver: Ajusta el algoritmo utilizado para optimizar la función de costo. Los solvers más comunes son liblinear, newton-cg y lbfgs.

Penalty: Es el tipo de penalización utilizada para evitar el sobreajuste en el modelo. Las opciones más comunes son **l1** y **l2**. La penalización l1 ó l2 aplica una restricción a los pesos del modelo, lo que favorece la aparición de características esparsas ó la aparición de características distribuidas.

C_{values}: Es el parámetro de regularización utilizado para controlar la fuerza de la penalización. Un valor más grande significa una penalización más débil, mientras que un valor más pequeño significa una penalización más fuerte. Con un valor demasiado pequeño, el modelo puede sobreajustar, mientras que si es demasiado grande, el modelo puede subajustar.

B.2. Descenso de Gradientes Estocástico - SGD

En este algoritmo se ajustaron los siguientes parámetros.

Loss: Es la función de costo utilizada para optimizar el modelo. Las opciones más comunes son hinge, log y squared_loss.

Alpha: Es el parámetro de regularización utilizado para controlar la fuerza de la penalización. Un valor más grande de alpha significa una penalización más fuerte, mientras que un valor más pequeño de alpha significa una penalización más débil. Si el valor de alpha es demasiado pequeño, el modelo puede sobreajustar, mientras que si es demasiado grande, el modelo puede subajustar.

L1_ratio: Es el parámetro utilizado para controlar la proporción de penalización L1 y L2 en la regularización Elastic Net. Un valor de 1 significa que se aplica solo la penalización L1, un valor de 0 significa que se aplica solo la penalización L2 y un valor intermedio significa que se aplica una combinación de ambas penalizaciones.

Max_iter: Es el número máximo de iteraciones permitido para que el modelo converja. Si el modelo no converge antes de alcanzar el número máximo de iteraciones, el entrenamiento se detendrá.

B.3. Support Vector Classifier

Kernel: Es una función matemática que mapea los datos de entrada en un espacio de mayor dimensión para poder separarlos de manera más efectiva. Los kernel más comunes son lineales, polinómicos y RBF (Función de Base Radial). El kernel se selecciona en función del tipo de datos que se están clasificando y de la complejidad de la relación entre ellos.

C y gamma: Se utilizan en la fórmula de la función de costo que se optimiza durante el entrenamiento del modelo. Por un parte C controla la penalización por errores de clasificación. Un valor alto implica que el modelo tratará de clasificar todos los puntos correctamente, incluso si eso significa crear una frontera de decisión más compleja y menos generalizable. Un valor bajo permitirá que algunos puntos se clasifiquen incorrectamente si eso significa tener una frontera de decisión más simple y generalizable. Por otro lado, Gamma controla el ancho del kernel RBF. Un valor alto significa que el modelo dará más peso a los puntos cercanos para la clasificación, lo que puede llevar a una frontera de decisión más compleja. Un valor bajo de gamma significa que el modelo dará menos peso a los puntos cercanos y se enfocará más en la estructura general de los datos.

B.4. Árbol de Decisión

max_depth: Profundidad máxima del árbol, que obedece al número máximo de divisiones que puede tener. Un árbol con una profundidad muy grande puede sobreajustarse a los datos de entrenamiento y no generalizar bien para nuevos datos, mientras que un árbol con una profundidad muy pequeña puede ser demasiado simple y no capturar suficientemente la complejidad de los datos.

criterion: Es una función que mide la calidad de una división en el árbol. Los más comunes son gini y entropy. Gini mide la impureza de una división, la mezcla de clases en un grupo, mientras que entropía mide la información necesaria para describir la distribución de clases. La elección del criterio depende del conjunto de datos y del problema en cuestión.

ccp_alpha: Es un parámetro que controla la complejidad del árbol. Cuanto mayor sea el valor de ccp_alpha, más se penaliza la complejidad del árbol y, por lo tanto, se reduce el sobreajuste. Si el valor de ccp_alpha es muy alto, el árbol será muy pequeño y poco

complejo, lo que puede llevar a un subajuste.

B.5. Random Forest

n_estimators: Es el número de árboles de decisión que se construyen en el modelo. Cada árbol de decisión se entrena con una muestra aleatoria del conjunto de datos original y de las características disponibles, el resultado final es una combinación de las predicciones de todos los árboles. Un mayor número de árboles puede mejorar la precisión del modelo, pero también aumenta el tiempo de entrenamiento y la complejidad del modelo.

max_features: Es el número máximo de características que se consideran al dividir un nodo en un árbol de decisión. La elección de las características para cada árbol se realiza de forma aleatoria, lo que ayuda a reducir la correlación entre los árboles y mejorar la precisión del modelo. Un valor más alto de max_features puede aumentar la precisión del modelo, pero también puede aumentar el riesgo de sobreajuste.

B.6. Boosting y Gradient Boost

learning_rate: Se utiliza para reducir la contribución de cada árbol a la predicción final del modelo. Un valor más bajo de learning_rate significa que cada árbol contribuye menos a la predicción final y que se necesita un mayor número de árboles para ajustar el modelo. Sin embargo, también reduce el riesgo de sobreajuste y puede ayudar a mejorar la generalización del modelo.

Por otro lado, un valor más alto de learning_rate puede hacer que el modelo se ajuste más rápidamente y requiera menos árboles, pero también puede aumentar el riesgo de sobreajuste. Es importante ajustar adecuadamente el learning_rate junto con otros hiperparámetros para obtener un modelo de Gradient Boosting bien ajustado y generalizable.

C. ANEXO C

RESULTADOS CONSEGUIDOS DESPUÉS DE LA ESTIMACIÓN DE LOS MEJORES MODELOS

Tabla 45: Resultados de todas las métricas después de la optimización y estimación realizada en el conjunto de experimentos número 1.

Subconjunto	SNPs	Algoritmo	Accuracy	Precisión	Recall	F1 Score
1	rs10800745 rs12644497 rs10061385 rs1198329 rs922533 rs4074608 rs8079726	REGRESION LOGISTICA	0.77	0.71	0.69	0.70
		SGD	0.73	0.64	0.60	0.61
		SVM	0.73	0.66	0.66	0.66
		ARBOL DE DECISIÓN	0.50	0.46	0.45	0.44
		RANDOM FOREST	0.86	0.84	0.80	0.82
		BOOSTING	0.82	0.79	0.72	0.74
		GRADIENTE BOOST	0.77	0.71	0.69	0.70
2	rs10800745 rs2486939 rs2929079 rs7917410 rs1198329 rs9518363 rs8079726	REGRESION LOGISTICA	0.68	0.61	0.63	0.62
		SGD	0.55	0.43	0.43	0.43
		SVM	0.68	0.54	0.52	0.51
		ARBOL DE DECISIÓN	0.64	0.49	0.49	0.48
		RANDOM FOREST	0.68	0.36	0.47	0.41
		BOOSTING	0.68	0.58	0.57	0.58
		GRADIENTE BOOST	0.64	0.49	0.49	0.48
3	rs10800745 rs2486939 rs10061385 rs573057 rs10771313 rs12910606 rs8079726	REGRESION LOGISTICA	0.64	0.58	0.59	0.58
		SGD	0.55	0.53	0.53	0.51
		SVM	0.64	0.54	0.54	0.54
		ARBOL DE DECISIÓN	0.73	0.64	0.60	0.61
		RANDOM FOREST	0.68	0.58	0.57	0.58
		BOOSTING	0.73	0.66	0.66	0.66
		GRADIENTE BOOST	0.64	0.54	0.54	0.54
4	rs10800745 rs4859120 rs10061385 rs573057 rs11101913 rs2379991 rs8042254	REGRESION LOGISTICA	0.82	0.90	0.67	0.69
		SGD	0.82	0.90	0.67	0.69
		SVM	0.86	0.92	0.75	0.79
		ARBOL DE DECISIÓN	0.82	0.79	0.72	0.74
		RANDOM FOREST	0.82	0.90	0.67	0.69
		BOOSTING	0.82	0.90	0.67	0.69
		GRADIENTE BOOST	0.82	0.90	0.67	0.69
21	rs10800745 rs2486939 rs4859120 rs10061385 rs11259260 rs1198329	REGRESION LOGISTICA	0.82	0.90	0.67	0.69
		SGD	0.77	0.73	0.64	0.65
		SVM	0.82	0.90	0.67	0.69
		ARBOL DE DECISIÓN	0.82	0.79	0.72	0.74
		RANDOM FOREST	0.82	0.79	0.72	0.74
		BOOSTING	0.82	0.90	0.67	0.69
		GRADIENTE BOOST	0.82	0.90	0.67	0.69
26	rs10800745	REGRESION LOGISTICA	0.59	0.55	0.56	0.54
		SGD	0.73	0.36	0.50	0.42
		SVM	0.59	0.55	0.56	0.54
		ARBOL DE DECISIÓN	0.59	0.55	0.56	0.54
		RANDOM FOREST	0.59	0.55	0.56	0.54
		BOOSTING	0.59	0.55	0.56	0.54
		GRADIENTE BOOST	0.59	0.55	0.56	0.54
		REGRESION LOGISTICA	0.77	0.73	0.64	0.65
		SGD	0.59	0.55	0.56	0.54
		SVM	0.77	0.71	0.69	0.70
		ARBOL DE DECISIÓN	0.77	0.71	0.69	0.70

Tabla 45: Resultados de todas las métricas después de la optimización y estimación realizada en el conjunto de experimentos número 1.

Subconjunto	SNPs	Algoritmo	Accuracy	Precisión	Recall	F1 Score
27	rs10800745 rs10061385	RANDOM FOREST	0.77	0.71	0.69	0.70
		BOOSTING	0.77	0.71	0.69	0.70
		GRADIENTE BOOST	0.77	0.71	0.69	0.70
28	rs8079726 rs10800745	REGRESION LOGISTICA	0.50	0.40	0.40	0.40
		SGD	0.50	0.40	0.40	0.40
		SVM	0.50	0.40	0.40	0.40
		ARBOL DE DECISIÓN	0.59	0.55	0.56	0.54
		RANDOM FOREST	0.59	0.55	0.56	0.54
		BOOSTING	0.59	0.55	0.56	0.54
		GRADIENTE BOOST	0.59	0.55	0.56	0.54
29	rs10800745 rs2486939	REGRESION LOGISTICA	0.59	0.51	0.51	0.51
		SGD	0.59	0.51	0.51	0.51
		SVM	0.68	0.58	0.57	0.58
		ARBOL DE DECISIÓN	0.64	0.54	0.54	0.54
		RANDOM FOREST	0.64	0.54	0.54	0.54
		BOOSTING	0.64	0.54	0.54	0.54
		GRADIENTE BOOST	0.64	0.54	0.54	0.54
30	rs10800745 rs1198329	REGRESION LOGISTICA	0.77	0.88	0.58	0.58
		SGD	0.73	0.36	0.50	0.42
		SVM	0.50	0.40	0.40	0.40
		ARBOL DE DECISIÓN	0.59	0.55	0.56	0.54
		RANDOM FOREST	0.86	0.92	0.75	0.79
		BOOSTING	0.86	0.92	0.75	0.79
		GRADIENTE BOOST	0.86	0.92	0.75	0.79
31	rs8079726 rs10800745 rs2486939	REGRESION LOGISTICA	0.59	0.45	0.46	0.45
		SGD	0.50	0.46	0.45	0.44
		SVM	0.59	0.55	0.56	0.54
		ARBOL DE DECISIÓN	0.68	0.58	0.57	0.58
		RANDOM FOREST	0.64	0.54	0.54	0.54
		BOOSTING	0.64	0.54	0.54	0.54
		GRADIENTE BOOST	0.59	0.51	0.51	0.51
32	rs8079726 rs10800745 rs1198329	REGRESION LOGISTICA	0.55	0.48	0.48	0.48
		SGD	0.59	0.55	0.56	0.54
		SVM	0.55	0.48	0.48	0.48
		ARBOL DE DECISIÓN	0.59	0.55	0.56	0.54
		RANDOM FOREST	0.50	0.46	0.45	0.44
		BOOSTING	0.86	0.92	0.75	0.79
		GRADIENTE BOOST	0.55	0.48	0.48	0.48
33	rs10800745 rs10061385 rs573057	REGRESION LOGISTICA	0.77	0.71	0.69	0.70
		SGD	0.77	0.71	0.69	0.70
		SVM	0.73	0.66	0.66	0.66
		ARBOL DE DECISIÓN	0.77	0.71	0.69	0.70
		RANDOM FOREST	0.77	0.71	0.69	0.70
		BOOSTING	0.77	0.71	0.69	0.70
		GRADIENTE BOOST	0.77	0.71	0.69	0.70
34	rs8079726 rs10800745 rs10061385	REGRESION LOGISTICA	0.55	0.48	0.48	0.48
		SGD	0.73	0.66	0.66	0.66
		SVM	0.50	0.40	0.40	0.40
		ARBOL DE DECISIÓN	0.41	0.36	0.33	0.34
		RANDOM FOREST	0.59	0.55	0.56	0.54
		BOOSTING	0.77	0.71	0.69	0.70
		GRADIENTE BOOST	0.77	0.71	0.69	0.70
		REGRESION LOGISTICA	0.59	0.51	0.51	0.51
		SGD	0.50	0.50	0.50	0.47
		SVM	0.64	0.54	0.54	0.54
		ARBOL DE DECISIÓN	0.64	0.54	0.54	0.54

Tabla 45: Resultados de todas las métricas después de la optimización y estimación realizada en el conjunto de experimentos número 1.

Subconjunto	SNPs	Algoritmo	Accuracy	Precisión	Recall	F1 Score
35	rs10800745 rs2486939 rs10061385	RANDOM FOREST	0.73	0.64	0.60	0.61
		BOOSTING	0.73	0.64	0.60	0.61
		GRADIENTE BOOST	0.73	0.64	0.60	0.61
36	rs10800745 rs10061385	REGRESION LOGISTICA	0.77	0.73	0.64	0.65
		SGD	0.59	0.55	0.56	0.54
		SVM	0.77	0.71	0.69	0.70
		ARBOL DE DECISIÓN	0.77	0.71	0.69	0.70
		RANDOM FOREST	0.77	0.71	0.69	0.70
		BOOSTING	0.77	0.71	0.69	0.70
		GRADIENTE BOOST	0.77	0.71	0.69	0.70
37	rs10800745 rs4859120 rs10061385	REGRESION LOGISTICA	0.82	0.90	0.67	0.69
		SGD	0.82	0.90	0.67	0.69
		SVM	0.82	0.90	0.67	0.69
		ARBOL DE DECISIÓN	0.82	0.90	0.67	0.69
		RANDOM FOREST	0.73	0.66	0.66	0.66
		BOOSTING	0.77	0.71	0.69	0.70
		GRADIENTE BOOST	0.82	0.90	0.67	0.69
38	rs10800745 rs10061385 rs1198329	REGRESION LOGISTICA	0.82	0.79	0.72	0.74
		SGD	0.77	0.73	0.64	0.65
		SVM	0.73	0.64	0.60	0.61
		ARBOL DE DECISIÓN	0.82	0.77	0.77	0.77
		RANDOM FOREST	0.82	0.77	0.77	0.77
		BOOSTING	0.86	0.92	0.75	0.79
		GRADIENTE BOOST	0.77	0.71	0.69	0.70

Tabla 46: Resultados de todas las métricas después de la optimización y estimación realizada en el conjunto de experimentos número 2.

Subconjunto	SNPs	Algoritmo	Accuracy	Precisión	Recall	F1 Score
5	rs3813355 rs11259260 rs9518363 rs10851935 rs720680 rs4074608 rs11087039	REGRESION LOGISTICA	0.85	0.84	0.84	0.84
		SGD	0.85	0.84	0.84	0.84
		SVM	0.69	0.68	0.64	0.64
		ARBOL DE DECISIÓN	0.62	0.61	0.61	0.61
		RANDOM FOREST	0.85	0.84	0.84	0.84
		BOOSTING	0.62	0.61	0.61	0.61
		GRADIENTE BOOST	0.62	0.61	0.61	0.61
6	rs10800745 rs3813355 rs1198329 rs9518363 rs10851935 rs4074608 rs11667789	REGRESION LOGISTICA	0.92	0.94	0.90	0.92
		SGD	0.85	0.90	0.80	0.82
		SVM	1.00	1.00	1.00	1.00
		ARBOL DE DECISIÓN	0.54	0.55	0.55	0.54
		RANDOM FOREST	0.85	0.86	0.88	0.85
		BOOSTING	0.92	0.92	0.94	0.92
		GRADIENTE BOOST	0.77	0.76	0.74	0.75
7	rs10061385 rs11259260 rs1198329 rs9513426 rs9518363 rs8042254 rs720680	REGRESION LOGISTICA	0.69	0.68	0.64	0.64
		SGD	0.69	0.68	0.64	0.64
		SVM	0.77	0.76	0.74	0.75
		ARBOL DE DECISIÓN	0.77	0.76	0.74	0.75
		RANDOM FOREST	0.69	0.68	0.64	0.64
		BOOSTING	0.77	0.76	0.74	0.75
		GRADIENTE BOOST	0.69	0.68	0.64	0.64
		REGRESION LOGISTICA	0.77	0.86	0.70	0.71
		SGD	0.77	0.86	0.70	0.71
		SVM	0.77	0.86	0.70	0.71

Tabla 46: Resultados de todas las métricas después de la optimización y estimación realizada en el conjunto de experimentos número 2.

Subconjunto	SNPs	Algoritmo	Accuracy	Precisión	Recall	F1 Score
8	rs4859120 rs10061385 rs3813355 rs11101913 rs1198329 rs9518363 rs2379991	ARBOL DE DECISIÓN	0.69	0.68	0.64	0.64
		RANDOM FOREST	0.92	0.94	0.90	0.92
		BOOSTING	0.62	0.58	0.58	0.58
		GRADIENTE BOOST	0.69	0.68	0.64	0.64
22	rs11120748 rs4859120 rs10061385 rs3813355 rs11259260 rs1198329 rs922533	REGRESION LOGISTICA	0.77	0.76	0.74	0.75
		SGD	0.69	0.68	0.64	0.64
		SVM	0.85	0.90	0.80	0.82
		ARBOL DE DECISIÓN	0.77	0.76	0.74	0.75
		RANDOM FOREST	0.85	0.84	0.84	0.84
		BOOSTING	0.77	0.76	0.74	0.75
		GRADIENTE BOOST	0.85	0.84	0.84	0.84
39	rs9518363	REGRESION LOGISTICA	0.62	0.61	0.61	0.61
		SGD	0.62	0.31	0.50	0.38
		SVM	0.69	0.68	0.64	0.64
		ARBOL DE DECISIÓN	0.69	0.68	0.64	0.64
		RANDOM FOREST	0.69	0.68	0.64	0.64
		BOOSTING	0.69	0.68	0.64	0.64
		GRADIENTE BOOST	0.69	0.68	0.64	0.64
40	rs1198329	REGRESION LOGISTICA	0.85	0.90	0.80	0.82
		SGD	0.85	0.90	0.80	0.82
		SVM	0.92	0.92	0.94	0.92
		ARBOL DE DECISIÓN	0.92	0.92	0.94	0.92
		RANDOM FOREST	0.92	0.92	0.94	0.92
		BOOSTING	0.92	0.92	0.94	0.92
		GRADIENTE BOOST	0.92	0.92	0.94	0.92
41	rs3813355	REGRESION LOGISTICA	0.77	0.76	0.74	0.75
		SGD	0.77	0.76	0.74	0.75
		SVM	0.77	0.76	0.74	0.75
		ARBOL DE DECISIÓN	0.69	0.70	0.71	0.69
		RANDOM FOREST	0.69	0.70	0.71	0.69
		BOOSTING	0.69	0.70	0.71	0.69
		GRADIENTE BOOST	0.69	0.70	0.71	0.69
42	rs9518363 rs1198329	REGRESION LOGISTICA	0.85	0.90	0.80	0.82
		SGD	0.85	0.90	0.80	0.82
		SVM	0.85	0.90	0.80	0.82
		ARBOL DE DECISIÓN	0.85	0.90	0.80	0.82
		RANDOM FOREST	0.85	0.90	0.80	0.82
		BOOSTING	0.69	0.68	0.64	0.64
		GRADIENTE BOOST	0.92	0.92	0.94	0.92
43	rs9518363 rs3813355	REGRESION LOGISTICA	0.77	0.86	0.70	0.71
		SGD	0.77	0.86	0.70	0.71
		SVM	0.77	0.86	0.70	0.71
		ARBOL DE DECISIÓN	0.69	0.68	0.68	0.68
		RANDOM FOREST	0.69	0.68	0.68	0.68
		BOOSTING	0.69	0.68	0.68	0.68
		GRADIENTE BOOST	0.77	0.86	0.70	0.71
44	rs3813355 rs1198329	REGRESION LOGISTICA	0.85	0.90	0.80	0.82
		SGD	0.85	0.90	0.80	0.82
		SVM	0.77	0.76	0.74	0.75
		ARBOL DE DECISIÓN	0.77	0.76	0.74	0.75
		RANDOM FOREST	0.85	0.84	0.84	0.84
		BOOSTING	0.85	0.84	0.84	0.84
		GRADIENTE BOOST	0.85	0.84	0.84	0.84
		REGRESION LOGISTICA	0.85	0.90	0.80	0.82
		SGD	0.85	0.86	0.88	0.85
		SVM	0.92	0.92	0.94	0.92

Tabla 46: Resultados de todas las métricas después de la optimización y estimación realizada en el conjunto de experimentos número 2.

Subconjunto	SNPs	Algoritmo	Accuracy	Precisión	Recall	F1 Score
45	rs10061385 rs1198329	ARBOL DE DECISIÓN	0.92	0.92	0.94	0.92
		RANDOM FOREST	0.92	0.92	0.94	0.92
		BOOSTING	0.92	0.92	0.94	0.92
		GRADIENTE BOOST	0.92	0.92	0.94	0.92
46	rs11259260	REGRESION LOGISTICA	0.62	0.57	0.54	0.51
		SGD	0.62	0.57	0.54	0.51
		SVM	0.62	0.57	0.54	0.51
		ARBOL DE DECISIÓN	0.62	0.57	0.54	0.51
		RANDOM FOREST	0.62	0.57	0.54	0.51
		BOOSTING	0.62	0.57	0.54	0.51
		GRADIENTE BOOST	0.62	0.57	0.54	0.51

Tabla 47: Resultados de todas las métricas después de la optimización y estimación realizada en el conjunto de experimentos número 3.

Subconjunto	SNPs	Algoritmo	Accuracy	Precisión	Recall	F1 Score
9	rs12644497 rs1480149 rs3813355 rs11101913 rs9513426 rs4074608 rs8079726	REGRESION LOGISTICA	0.89	0.93	0.83	0.86
		SGD	0.78	0.88	0.67	0.68
		SVM	0.78	0.88	0.67	0.68
		ARBOL DE DECISIÓN	0.67	0.61	0.58	0.58
		RANDOM FOREST	0.78	0.88	0.67	0.68
		BOOSTING	0.67	0.33	0.50	0.40
		GRADIENTE BOOST	0.67	0.61	0.58	0.58
10	rs2486939 rs4859120 rs12644497 rs2929079 rs9804548 rs12283577 rs8079726	REGRESION LOGISTICA	1.00	1.00	1.00	1.00
		SGD	1.00	1.00	1.00	1.00
		SVM	1.00	1.00	1.00	1.00
		ARBOL DE DECISIÓN	1.00	1.00	1.00	1.00
		RANDOM FOREST	1.00	1.00	1.00	1.00
		BOOSTING	1.00	1.00	1.00	1.00
		GRADIENTE BOOST	1.00	1.00	1.00	1.00
11	rs4859120 rs11101913 rs10851935 rs922533 rs17598590 rs11667789 rs11087039	REGRESION LOGISTICA	0.67	0.61	0.58	0.58
		SGD	0.67	0.61	0.58	0.58
		SVM	0.67	0.61	0.58	0.58
		ARBOL DE DECISIÓN	0.78	0.88	0.67	0.68
		RANDOM FOREST	0.67	0.33	0.50	0.40
		BOOSTING	0.67	0.61	0.58	0.58
		GRADIENTE BOOST	0.78	0.88	0.67	0.68
12	rs10800745 rs2486939 rs4859120 rs12644497 rs11101913 rs11026669 rs8079726	REGRESION LOGISTICA	1.00	1.00	1.00	1.00
		SGD	0.89	0.88	0.92	0.88
		SVM	0.89	0.88	0.92	0.88
		ARBOL DE DECISIÓN	0.78	0.88	0.67	0.68
		RANDOM FOREST	0.89	0.93	0.83	0.86
		BOOSTING	0.78	0.88	0.67	0.68
		GRADIENTE BOOST	0.78	0.88	0.67	0.68
23	rs2486939 rs4859120 rs12644497 rs2929079 rs12283577 rs8079726	REGRESION LOGISTICA	1.00	1.00	1.00	1.00
		SGD	1.00	1.00	1.00	1.00
		SVM	1.00	1.00	1.00	1.00
		ARBOL DE DECISIÓN	0.89	0.88	0.92	0.88
		RANDOM FOREST	1.00	1.00	1.00	1.00
		BOOSTING	1.00	1.00	1.00	1.00
		GRADIENTE BOOST	1.00	1.00	1.00	1.00
		REGRESION LOGISTICA	0.67	0.33	0.50	0.40
		SGD	0.89	0.88	0.92	0.88

Tabla 47: Resultados de todas las métricas después de la optimización y estimación realizada en el conjunto de experimentos número 3.

Subconjunto	SNPs	Algoritmo	Accuracy	Precisión	Recall	F1 Score
47	rs4859120	SVM	0.67	0.33	0.50	0.40
		ARBOL DE DECISIÓN	0.89	0.88	0.92	0.88
		RANDOM FOREST	0.89	0.88	0.92	0.88
		BOOSTING	0.67	0.33	0.50	0.40
		GRADIENTE BOOST	0.67	0.33	0.50	0.40
48	rs8079726 rs12644497	REGRESION LOGISTICA	0.67	0.33	0.50	0.40
		SGD	0.89	0.88	0.92	0.88
		SVM	0.89	0.93	0.83	0.86
		ARBOL DE DECISIÓN	0.89	0.93	0.83	0.86
		RANDOM FOREST	0.89	0.93	0.83	0.86
		BOOSTING	0.89	0.93	0.83	0.86
		GRADIENTE BOOST	0.89	0.93	0.83	0.86
49	rs8079726 rs2486939 rs4859120 rs12644497	REGRESION LOGISTICA	1.00	1.00	1.00	1.00
		SGD	1.00	1.00	1.00	1.00
		SVM	0.67	0.33	0.50	0.40
		ARBOL DE DECISIÓN	0.89	0.88	0.92	0.88
		RANDOM FOREST	1.00	1.00	1.00	1.00
		BOOSTING	1.00	1.00	1.00	1.00
		GRADIENTE BOOST	1.00	1.00	1.00	1.00
50	rs11101913	REGRESION LOGISTICA	0.78	0.88	0.67	0.68
		SGD	0.78	0.88	0.67	0.68
		SVM	0.67	0.33	0.50	0.40
		ARBOL DE DECISIÓN	0.78	0.88	0.67	0.68
		RANDOM FOREST	0.67	0.33	0.50	0.40
		BOOSTING	0.78	0.88	0.67	0.68
		GRADIENTE BOOST	0.67	0.33	0.50	0.40
51	rs12283577 rs2486939 rs12644497 rs8079726 rs2929079 rs4859120	REGRESION LOGISTICA	1.00	1.00	1.00	1.00
		SGD	1.00	1.00	1.00	1.00
		SVM	1.00	1.00	1.00	1.00
		ARBOL DE DECISIÓN	0.78	0.80	0.83	0.77
		RANDOM FOREST	1.00	1.00	1.00	1.00
		BOOSTING	1.00	1.00	1.00	1.00
		GRADIENTE BOOST	1.00	1.00	1.00	1.00
52	rs4859120 rs11101913	REGRESION LOGISTICA	0.67	0.61	0.58	0.58
		SGD	0.67	0.61	0.58	0.58
		SVM	0.78	0.88	0.67	0.68
		ARBOL DE DECISIÓN	0.78	0.88	0.67	0.68
		RANDOM FOREST	0.78	0.88	0.67	0.68
		BOOSTING	0.67	0.61	0.58	0.58
		GRADIENTE BOOST	0.78	0.88	0.67	0.68
53	rs8079726 rs12644497 rs11101913	REGRESION LOGISTICA	0.78	0.88	0.67	0.68
		SGD	0.78	0.88	0.67	0.68
		SVM	0.78	0.88	0.67	0.68
		ARBOL DE DECISIÓN	0.67	0.33	0.50	0.40
		RANDOM FOREST	0.67	0.33	0.50	0.40
		BOOSTING	0.67	0.33	0.50	0.40
		GRADIENTE BOOST	0.67	0.33	0.50	0.40

Tabla 48: Resultados de todas las métricas después de la optimización y estimación realizada en el conjunto de experimentos número 4.

Subconjunto	SNPs	Algoritmo	Accuracy	Precisión	Recall	F1 Score
		REGRESION LOGISTICA	0.46	0.45	0.45	0.45
		SGD	0.62	0.31	0.50	0.38

Tabla 48: Resultados de todas las métricas después de la optimización y estimación realizada en el conjunto de experimentos número 4.

Subconjunto	SNPs	Algoritmo	Accuracy	Precisión	Recall	F1 Score
13	rs12023541 rs1529833 rs6886361 rs6590735 rs2302688 rs3497 rs4576883	SVM	0.54	0.55	0.55	0.54
		ARBOL DE DECISIÓN	0.69	0.68	0.64	0.64
		RANDOM FOREST	0.62	0.58	0.58	0.58
		BOOSTING	0.38	0.35	0.35	0.35
		GRADIENTE BOOST	0.69	0.68	0.64	0.64
14	rs12023541 rs6886361 rs9292869 rs6590735 rs2302688 rs3497 rs9510008	REGRESION LOGISTICA	0.31	0.33	0.33	0.31
		SGD	0.31	0.33	0.33	0.31
		SVM	0.31	0.33	0.33	0.31
		ARBOL DE DECISIÓN	0.62	0.61	0.61	0.61
		RANDOM FOREST	0.46	0.45	0.45	0.45
		BOOSTING	0.62	0.65	0.65	0.62
		GRADIENTE BOOST	0.54	0.51	0.51	0.51
15	rs12023541 rs1529833 rs6886361 rs6999746 rs6590735 rs2302688 rs9510008	REGRESION LOGISTICA	0.62	0.61	0.61	0.61
		SGD	0.69	0.68	0.68	0.68
		SVM	0.62	0.61	0.61	0.61
		ARBOL DE DECISIÓN	0.77	0.76	0.74	0.75
		RANDOM FOREST	0.54	0.55	0.55	0.54
		BOOSTING	0.69	0.83	0.60	0.57
		GRADIENTE BOOST	0.69	0.83	0.60	0.57
16	rs12023541 rs1529833 rs6886361 rs6590735 rs2302688 rs4576883 rs9510008	REGRESION LOGISTICA	0.62	0.57	0.54	0.51
		SGD	0.69	0.70	0.71	0.69
		SVM	0.69	0.68	0.68	0.68
		ARBOL DE DECISIÓN	0.69	0.68	0.64	0.64
		RANDOM FOREST	0.62	0.61	0.61	0.61
		BOOSTING	0.38	0.35	0.35	0.35
		GRADIENTE BOOST	0.54	0.51	0.51	0.51
24	rs6590735 rs2302688	REGRESION LOGISTICA	0.62	0.61	0.61	0.61
		SGD	0.62	0.61	0.61	0.61
		SVM	0.62	0.61	0.61	0.61
		ARBOL DE DECISIÓN	0.62	0.61	0.61	0.61
		RANDOM FOREST	0.69	0.68	0.64	0.64
		BOOSTING	0.62	0.61	0.61	0.61
		GRADIENTE BOOST	0.69	0.68	0.64	0.64
55	rs12023541 rs6590735 rs6886361 rs2302688	REGRESION LOGISTICA	0.69	0.68	0.68	0.68
		SGD	0.62	0.65	0.65	0.62
		SVM	0.62	0.58	0.58	0.58
		ARBOL DE DECISIÓN	0.62	0.61	0.61	0.61
		RANDOM FOREST	0.62	0.61	0.61	0.61
		BOOSTING	0.69	0.70	0.71	0.69
		GRADIENTE BOOST	0.62	0.61	0.61	0.61
56	rs12023541 rs9510008 rs6590735 rs2302688 rs6886361	REGRESION LOGISTICA	0.69	0.68	0.68	0.68
		SGD	0.62	0.58	0.58	0.58
		SVM	0.54	0.51	0.51	0.51
		ARBOL DE DECISIÓN	0.54	0.51	0.51	0.51
		RANDOM FOREST	0.69	0.68	0.68	0.68
		BOOSTING	0.46	0.45	0.45	0.45
		GRADIENTE BOOST	0.69	0.70	0.71	0.69
57	rs12023541 rs1529833 rs6590735 rs2302688 rs6886361	REGRESION LOGISTICA	0.69	0.68	0.68	0.68
		SGD	0.46	0.49	0.49	0.46
		SVM	0.54	0.29	0.44	0.35
		ARBOL DE DECISIÓN	0.62	0.58	0.58	0.58
		RANDOM FOREST	0.77	0.86	0.70	0.71
		BOOSTING	0.46	0.40	0.41	0.41
		GRADIENTE BOOST	0.69	0.68	0.64	0.64
		REGRESION LOGISTICA	0.54	0.51	0.51	0.51
		SGD	0.54	0.51	0.51	0.51

Tabla 48: Resultados de todas las métricas después de la optimización y estimación realizada en el conjunto de experimentos número 4.

Subconjunto	SNPs	Algoritmo	Accuracy	Precisión	Recall	F1 Score
58	rs12023541 rs6590735 rs2302688 rs3497 rs6886361	SVM	0.46	0.49	0.49	0.46
		ARBOL DE DECISIÓN	0.77	0.86	0.70	0.71
		RANDOM FOREST	0.62	0.58	0.58	0.58
		BOOSTING	0.62	0.65	0.65	0.62
		GRADIENTE BOOST	0.62	0.58	0.58	0.58
59	rs12023541 rs9510008 rs1529833 rs6590735 rs2302688 rs6886361	REGRESION LOGISTICA	0.77	0.86	0.70	0.71
		SGD	0.69	0.68	0.64	0.64
		SVM	0.69	0.68	0.68	0.68
		ARBOL DE DECISIÓN	0.69	0.68	0.64	0.64
		RANDOM FOREST	0.62	0.61	0.61	0.61
		BOOSTING	0.46	0.40	0.41	0.41
		GRADIENTE BOOST	0.54	0.55	0.55	0.54
60	rs12023541 rs1529833 rs6590735 rs2302688 rs4576883 rs6886361	REGRESION LOGISTICA	0.62	0.61	0.61	0.61
		SGD	0.54	0.47	0.48	0.46
		SVM	0.69	0.68	0.68	0.68
		ARBOL DE DECISIÓN	0.69	0.68	0.64	0.64
		RANDOM FOREST	0.69	0.68	0.64	0.64
		BOOSTING	0.38	0.35	0.35	0.35
		GRADIENTE BOOST	0.69	0.68	0.64	0.64

Tabla 49: Resultados de todas las métricas después de la optimización y estimación realizada en el conjunto de experimentos número 5.

Subconjunto	SNPs	Algoritmo	Accuracy	Precisión	Recall	F1 Score
17	rs17804991 rs12551529 rs2948167 rs2942076 rs4408399 rs4028683 rs2440327	REGRESION LOGISTICA	0.78	0.80	0.83	0.77
		SGD	0.89	0.88	0.92	0.88
		SVM	0.78	0.75	0.75	0.75
		ARBOL DE DECISIÓN	0.78	0.88	0.67	0.68
		RANDOM FOREST	0.78	0.88	0.67	0.68
		BOOSTING	0.89	0.93	0.83	0.86
		GRADIENTE BOOST	0.78	0.88	0.67	0.68
18	rs17804991 rs10031902 rs12551529 rs2948167 rs2942076 rs4028683 rs2440327	REGRESION LOGISTICA	0.78	0.80	0.83	0.77
		SGD	0.78	0.80	0.83	0.77
		SVM	0.67	0.65	0.67	0.65
		ARBOL DE DECISIÓN	0.78	0.75	0.75	0.75
		RANDOM FOREST	0.89	0.93	0.83	0.86
		BOOSTING	0.89	0.88	0.92	0.88
		GRADIENTE BOOST	0.78	0.75	0.75	0.75
19	rs818524 rs4708676 rs2948167 rs2942076 rs4075243 rs9318541 rs4028683	REGRESION LOGISTICA	0.56	0.50	0.50	0.50
		SGD	0.56	0.50	0.50	0.50
		SVM	0.56	0.50	0.50	0.50
		ARBOL DE DECISIÓN	0.44	0.43	0.42	0.42
		RANDOM FOREST	0.67	0.61	0.58	0.58
		BOOSTING	0.78	0.88	0.67	0.68
		GRADIENTE BOOST	0.78	0.88	0.67	0.68
20	rs818524 rs530160 rs12551529 rs11187729 rs2948167 rs2942076 rs4028683	REGRESION LOGISTICA	0.78	0.88	0.67	0.68
		SGD	0.78	0.88	0.67	0.68
		SVM	0.67	0.33	0.50	0.40
		ARBOL DE DECISIÓN	0.67	0.61	0.58	0.58
		RANDOM FOREST	0.67	0.33	0.50	0.40
		BOOSTING	0.67	0.33	0.50	0.40
		GRADIENTE BOOST	0.56	0.31	0.42	0.36
		REGRESION LOGISTICA	0.89	0.93	0.83	0.86

25	rs17804991 rs10031902 rs12551529 rs2942076 rs4028683	SGD	0.78	0.75	0.75	0.75
		SVM	0.89	0.93	0.83	0.86
		ARBOL DE DECISIÓN	0.78	0.75	0.75	0.75
		RANDOM FOREST	0.89	0.93	0.83	0.86
		BOOSTING	0.89	0.93	0.83	0.86
		GRADIENTE BOOST	0.89	0.93	0.83	0.86
61	rs2942076 rs4028683	REGRESION LOGISTICA	0.78	0.75	0.75	0.75
		SGD	0.67	0.33	0.50	0.40
		SVM	0.67	0.33	0.50	0.40
		ARBOL DE DECISIÓN	0.78	0.75	0.75	0.75
		RANDOM FOREST	0.78	0.88	0.67	0.68
		BOOSTING	0.78	0.88	0.67	0.68
		GRADIENTE BOOST	0.67	0.33	0.50	0.40
62	rs2942076 rs4028683 rs2948167	REGRESION LOGISTICA	0.78	0.88	0.67	0.68
		SGD	0.78	0.88	0.67	0.68
		SVM	0.78	0.75	0.75	0.75
		ARBOL DE DECISIÓN	0.78	0.75	0.75	0.75
		RANDOM FOREST	0.78	0.75	0.75	0.75
		BOOSTING	0.78	0.88	0.67	0.68
		GRADIENTE BOOST	0.67	0.33	0.50	0.40
63	rs2942076 rs4028683 rs12551529	REGRESION LOGISTICA	0.89	0.93	0.83	0.86
		SGD	0.89	0.93	0.83	0.86
		SVM	0.89	0.93	0.83	0.86
		ARBOL DE DECISIÓN	0.78	0.75	0.75	0.75
		RANDOM FOREST	0.89	0.93	0.83	0.86
		BOOSTING	0.89	0.93	0.83	0.86
		GRADIENTE BOOST	0.89	0.93	0.83	0.86
64	rs12551529 rs2942076 rs4028683 rs2948167	REGRESION LOGISTICA	0.78	0.88	0.67	0.68
		SGD	0.78	0.88	0.67	0.68
		SVM	0.78	0.88	0.67	0.68
		ARBOL DE DECISIÓN	0.78	0.75	0.75	0.75
		RANDOM FOREST	0.78	0.88	0.67	0.68
		BOOSTING	0.78	0.88	0.67	0.68
		GRADIENTE BOOST	0.78	0.88	0.67	0.68
65	rs17804991 rs2942076 rs4028683 rs12551529	REGRESION LOGISTICA	0.78	0.75	0.75	0.75
		SGD	0.89	0.93	0.83	0.86
		SVM	0.78	0.88	0.67	0.68
		ARBOL DE DECISIÓN	0.78	0.75	0.75	0.75
		RANDOM FOREST	0.78	0.88	0.67	0.68
		BOOSTING	0.89	0.93	0.83	0.86
		GRADIENTE BOOST	0.89	0.93	0.83	0.86
66	rs4028683 rs2942076 rs2948167	REGRESION LOGISTICA	0.78	0.88	0.67	0.68
		SGD	0.78	0.88	0.67	0.68
		SVM	0.78	0.75	0.75	0.75
		ARBOL DE DECISIÓN	0.78	0.75	0.75	0.75
		RANDOM FOREST	0.78	0.88	0.67	0.68
		BOOSTING	0.78	0.88	0.67	0.68
		GRADIENTE BOOST	0.67	0.33	0.50	0.40
67	rs4028683 rs2942076 rs2948167 rs12551529	REGRESION LOGISTICA	0.78	0.88	0.67	0.68
		SGD	0.78	0.88	0.67	0.68
		SVM	0.78	0.88	0.67	0.68
		ARBOL DE DECISIÓN	0.78	0.75	0.75	0.75
		RANDOM FOREST	0.78	0.88	0.67	0.68
		BOOSTING	0.78	0.88	0.67	0.68
		GRADIENTE BOOST	0.78	0.88	0.67	0.68
		REGRESION LOGISTICA	0.78	0.80	0.83	0.77
		SGD	0.78	0.80	0.83	0.77
		SVM	0.78	0.75	0.75	0.75
		ARBOL DE DECISIÓN	0.78	0.75	0.75	0.75

68	rs17804991 rs2942076 rs2948167 rs12551529 rs2440327 rs4028683	RANDOM FOREST	0.89	0.93	0.83	0.86
		BOOSTING	0.78	0.80	0.83	0.77
		GRADIENTE BOOST	0.78	0.75	0.75	0.75
69	rs10031902 rs17804991 rs2942076 rs12551529 rs4028683	REGRESION LOGISTICA	0.89	0.93	0.83	0.86
		SGD	0.78	0.88	0.67	0.68
		SVM	0.89	0.93	0.83	0.86
		ARBOL DE DECISIÓN	0.78	0.75	0.75	0.75
		RANDOM FOREST	0.89	0.93	0.83	0.86
		BOOSTING	0.89	0.93	0.83	0.86
		GRADIENTE BOOST	0.89	0.93	0.83	0.86
70	rs4028683 rs2942076 rs2948167 rs818524	REGRESION LOGISTICA	0.78	0.88	0.67	0.68
		SGD	0.78	0.88	0.67	0.68
		SVM	0.78	0.88	0.67	0.68
		ARBOL DE DECISIÓN	0.78	0.75	0.75	0.75
		RANDOM FOREST	0.67	0.33	0.50	0.40
		BOOSTING	0.78	0.88	0.67	0.68
		GRADIENTE BOOST	0.67	0.33	0.50	0.40