UNIVERSIDAD DE COSTA RICA
SISTEMA DE ESTUDIOS DE POSGRADO

IDENTIFICACIÓN DE BIOMARCADORES ASOCIADOS A LOS SÍNDROMES TALASÉMICOS MEDIANTE EL USO DE ALGORITMOS DE APRENDIZAJE AUTOMÁTICO

Tesis sometida a la consideración del Programa de Estudios de Posgrado en Computación e Informática para optar al grado y título de Maestría Académica en Computación e Informática

LUIS DIEGO MORA JIMÉNEZ

Ciudad Universitaria Rodrigo Facio, Costa Rica
2024

Esta tesis fue aceptada por la Comisión del Programa de Estudios de Posgrado en Computación e Informática de la Universidad de Costa Rica, como requisito parcial para optar al grado y título de Maestría Académica en Computación e Informática.

Dr. Ignacio Díaz Oreiro
Representante de la Decana
Sistema de Estudios de Posgrado

Dr. José Andrés Guevara Coto
Director de Tesis

Dra. Kryscia Daviana Ramírez Benavides
Asesora

Dr. Luis José Quesada Quirós
Asesor

Dr. Gustavo López Herrera
Director
Programa de Posgrado en Computación e Informática

Luis Diego Mora Jiménez
Candidato

Nota: La autorización del Dr. José Andrés Guevara Coto se encuentra en resguardo en el programa de posgrado.

Tabla de Contenido

Hoja de Aprobación

Resumen en Español v

Abstract vi

Lista de Tablas vii

Lista de Figuras viii

Lista de Abreviaturas ix

1. Introducción **1**

1.1. Pregunta de Investigación 3

1.2. Justificación de la Investigación 4

1.3. Objetivos 5

1.4. Limitaciones 6

1.5. Estructura del Documento 6

2. Marco Teórico **7**

2.1. Aspectos Biológicos Asociados a la Talasemia 7

2.2. Aprendizaje Automático y su Aplicación en Biología y Medicina 11

3. Antecedentes **17**

4. Metodología **21**

4.1. Enfoque de Clasificación 21

4.2. Enfoque de Detección de Anomalías 26

4.3. Herramientas 32

5. Resultados **33**

5.1. Enfoque de Clasificación 33

5.2. Enfoque de Detección de Anomalías 37

6. Discusión **43**

6.1. Enfoque de Clasificación 43

6.2. Enfoque de Detección de Anomalías 45

7. Conclusiones **50**

Bibliografía **56**

Resumen en Español

La talasemia es un trastorno sanguíneo hereditario que afecta la producción de hemoglobina, lo que conlleva niveles anormalmente bajos de esta proteína. En 2015, esta enfermedad fue responsable de 16800 muertes y afecta aproximadamente al 1,5 % de la población mundial. El diagnóstico se realiza mediante análisis de sangre y pruebas genéticas. Sin embargo, la falta de registros y tamizajes adecuados impide la detección de numerosos casos graves, lo cual contribuye a una elevada tasa de mortalidad. Aunque su prevalencia es baja, la talasemia tiene un impacto significativo en la calidad de vida de los pacientes, quienes requieren tratamiento de por vida.

Este estudio propone el uso de perfiles de expresión génica y métodos de aprendizaje automático para identificar biomarcadores asociados con la talasemia. Se emplearon modelos de clasificación y un algoritmo de detección de anomalías, observándose que los métodos de clasificación no fueron efectivos para abordar el problema. Sin embargo, el modelo basado en bosque de aislamiento permitió identificar 72 genes anómalos. La validación funcional de estos genes destacó términos biológicos relevantes, como traducción citoplasmática y apoptosis, sugiriendo posibles vías moleculares implicadas en la talasemia. Además, se identificaron genes relacionados con la homeostasis del hierro, estableciendo un vínculo entre el estrés oxidativo, la apoptosis y esta enfermedad.

La comparación con estudios previos reveló procesos biológicos comunes, lo que resalta el potencial del aprendizaje automático para mejorar el diagnóstico y profundizar en la comprensión de las vías moleculares, con el objetivo de optimizar los tratamientos para los pacientes.

Abstract

Thalassemia is a hereditary blood disorder that affects hemoglobin production, resulting in abnormally low levels of this protein. In 2015, this disease was responsible for 16,800 deaths and affects approximately 1.5 % of the global population. Diagnosis involves blood tests and genetic screening; however, the lack of adequate records and screening programs hinders the detection of many severe cases, contributing to high mortality rates. Although its prevalence is low, thalassemia significantly impacts the quality of life of patients, who require lifelong treatment.

This study proposes the use of gene expression profiles and machine learning methods to identify biomarkers associated with thalassemia. Classification models and an anomaly detection algorithm were applied, revealing that classification methods were not effective for this problem. However, the isolation forest based model identified 72 anomalous genes. Functional validation of these genes highlighted relevant biological terms such as cytoplasmic translation and apoptosis, suggesting potential molecular pathways involved in thalassemia. Additionally, genes related to iron homeostasis were identified, establishing a link between oxidative stress, apoptosis, and the disease.

Comparison with previous studies revealed common biological processes, underscoring the potential of machine learning to enhance diagnosis and deepen the understanding of molecular pathways, aiming to optimize patient treatments.

Lista de Tablas

2.1. Métricas de desempeño de clasificación comúnmente utilizadas [29]. . . 12
2.2. Algoritmos de clasificación. 13

4.1. Herramientas de software utilizadas. 32

5.1. Configuración de hiperparámetros para los distintos modelos construidos. 35
5.2. Métricas de desempeño obtenidas para los distintos modelos de aprendizaje automático. 36
5.3. Estadística descriptiva para los valores de expresión genética de acuerdo a la condición. 37
5.4. Descripción de los genes en el conjunto de entrenamiento con niveles de expresión identificados como anómalos [61]. 40
5.5. Genes asociados con los principales procesos biológicos identificados. . . 42

Lista de Figuras

2.1. Comparación de la forma normal de los glóbulos rojos y la forma de los glóbulos rojos de una persona con talasemia. Adaptada de [19]. 8
2.2. Flujo de trabajo de un proyecto de aprendizaje automático. 15

3.1. Distribución de los estudios relacionados de acuerdo al período de años en el que fueron publicados. 18
3.2. Usos de cada algoritmo identificado por período de años. 19

4.1. Flujo de trabajo de las actividades llevadas a cabo en la implementación del enfoque de clasificación. 27
4.2. Flujo de trabajo del enfoque de detección de anomalías [61]. 30

5.1. Diagrama de cajas del nivel de genética promedio de los genes de acuerdo a su asociación con la talasemia. 34
5.2. Histograma de la diferencia en los valores de expresión genética (afectado - control) [61]. 38
5.3. Mapa de calor de los valores de expresión genética (control vs. afectado) [61]. 39
5.4. Gráfico de barras del análisis de enriquecimiento para los genes candidatos. 41

Lista de Abreviaturas

ADN Ácido desoxirribonucleico. 8

CAD Diagnóstico Asistido por Computadora (*Computer-aided Diagnosis*). 4, 25

FN Falsos Negativos (*False Negatives*). 12

FP Falsos Positivos (*False Positives*). 12

GEO *Gene Expression Omnibus*. 22

GPU Unidad de Procesamiento Gráfico *(Graphics Processing Unit)*. 23, 24

GSEA Análisis de Enriquecimiento de Conjuntos de Genes (*Gene Set Enrichment Analysis*). 15, 20, 29

PU positivo-sin etiqueta (*positive-unlabeled*). 44

RNA-seq Secuenciación de ARN (*RNA sequencing*). 10, 20, 22, 28, 50, 51

ROS Especies Reactivas de Oxígeno (*Reactive Oxygen Species*). 46, 47

SVM Máquinas de Soporte Vectorial (*Support Vector Machine*). 13, 23, 24, 34–36

TN Verdaderos Negativos (*True Negatives*). 12

TP Verdaderos Positivos (*True Positives*). 12

TPOT *Tree Based Pipeline Optimization Tool*. 13, 23, 24, 34–36

XGBoost *eXtreme Gradient Boosting*. 13, 23, 24, 34–36, 43

Capítulo I. Introducción

A lo largo de los años, las técnicas de aprendizaje automático han ganado importancia en la realización de tareas como la clasificación y la predicción en diversos campos disciplinarios. Uno de estos campos es la biología computacional, donde se ha empleado el aprendizaje automático para entender mejor el funcionamiento de ciertas enfermedades y asistir en su diagnóstico [1]. Un ejemplo notable es el estudio de las talasemias, en el cual se han usado algoritmos de aprendizaje automático para distinguir entre portadores y no portadores de la enfermedad. Por lo tanto, resulta plausible considerar que el aprendizaje automático puede posicionarse como una herramienta valiosa para incrementar el conocimiento y comprender la patogénesis de las enfermedades.

Las talasemias constituyen un grupo de trastornos sanguíneos hereditarios caracterizados por la producción anormal de hemoglobina [2]. La hemoglobina es la proteína presente en los glóbulos rojos que se une al oxígeno en los pulmones y lo transporta a los tejidos y órganos de todo el cuerpo. Los médicos dividen la enfermedad en i) α- y ii) β-talasemia, dependiendo de la cadena de globina deficiente en la síntesis de hemoglobina [3]. La β-talasemia implica una reducción o ausencia de cadenas de beta-globina, lo que lleva a una producción disminuida de hemoglobina, mientras que la α-talasemia implica una reducción o ausencia de cadenas de alfa-globina. Además, la α-talasemia puede tener consecuencias más graves. Esta condición incluso puede causar condiciones que no son compatibles con la vida [2], [3]

El diagnóstico de la enfermedad normalmente se realiza mediante conteos sanguíneos completos, pruebas especializadas de hemoglobina o pruebas genéticas, como la prueba de detección genética del recién nacido. Aunque esta prueba no siempre se utiliza para detectar talasemia a pesar de estar disponible en muchos países. La enfermedad también puede ser identificada antes del nacimiento a través de exámenes prenatales [2]. Sin embargo, en pocos países existe un registro de pacientes y, en muchos otros, los niños

con los casos más graves fallecen sin ser diagnosticados [4]. En 2015, la enfermedad causó 16.800 muertes, y actualmente, el 1,5 % de la población mundial son portadores, con aproximadamente 60.000 diagnósticos anuales en recién nacidos [3], [4].

Además de las limitaciones en la infraestructura para el manejo de enfermedades, los métodos y enfoques de diagnóstico enfrentan desafíos actuales. Un ejemplo de esto es la clasificación errónea de personas con β-talasemia. A menudo, los individuos afectados por la enfermedad pueden ser confundidos con casos de anemia por deficiencia de hierro, una condición mucho más común [2]. Esto ocurre porque los parámetros necesarios para distinguir entre ambas enfermedades no se obtienen en exámenes de laboratorio rutinarios [5]. Además, las técnicas necesarias para realizar un diagnóstico diferencial son costosas y tienen implicaciones clínicas significativas [6].

En términos generales, los síntomas de la talasemia pueden incluir sobrecarga de hierro, deformidades óseas, agrandamiento del bazo, retraso en el crecimiento y problemas cardíacos. No obstante, la manifestación de estos síntomas varía según la gravedad de la enfermedad. Dado su origen genético, las personas con talasemia requieren atención y manejo continuos a lo largo de su vida. Dependiendo de la severidad, el tratamiento puede incluir desde terapia de quelación hasta múltiples transfusiones de sangre e incluso trasplante de células madre [2]. El continuo requerimiento de monitoreo, tratamiento y atención especializada representa una carga significativa para los sistemas de salud pública, especialmente en economías emergentes que a menudo carecen de la infraestructura necesaria o los recursos económicos para atender adecuadamente a los pacientes afectados [4].

La carga económica de la talasemia ha sido identificada, cuantificada y analizada, lo que subraya la importancia de investigar mejoras en los métodos de diagnóstico para aumentar la calidad de vida de los pacientes. De acuerdo a una evaluación económica realizada en 2021, el costo promedio para prevenir el nacimiento de un paciente con talasemia mayor fue de 32.624 USD, mientras que el costo del tratamiento de por vida para un paciente con talasemia mayor ascendió a 136.532 USD [7].

Un tratamiento temprano de la talasemia puede reducir significativamente la tasa de mortalidad. Esto implica una gran responsabilidad para los profesionales de la salud,

quienes deben tomar decisiones acertadas y utilizar procedimientos precisos y confiables que permitan diferenciar adecuadamente entre diversos diagnósticos [8]-[10]. En este contexto de pruebas clínicas con limitaciones conocidas, el aprendizaje automático puede desempeñar un papel crucial tanto en el diagnóstico como en la detección, ofreciendo una alternativa a las restricciones de los métodos moleculares. Por ejemplo, se puede mejorar la clasificación de pacientes con talasemia mediante sistemas de diagnóstico basados en aprendizaje automático [2], [11]. Además, estos métodos se han vuelto cada vez más confiables para el descubrimiento y comprensión de grandes volúmenes de datos biológicos complejos [1].

Por lo anteriormente descrito, esta investigación propone el uso de técnicas de aprendizaje automático para identificar nuevos biomarcadores[1] asociados con la talasemia. Al descubrir nuevos biomarcadores, es posible obtener tanto mayor conocimiento acerca de la enfermedad como nuevas dianas terapéuticas. Esto permitirá mejorar la capacidad de detección de la enfermedad y potencialmente reducir los costos de diagnóstico y tratamiento. Para realizar esta tarea, se utilizará un conjunto de datos de perfiles de expresión genética.

1.1. Pregunta de Investigación

Como se describirá posteriormente en el Capítulo 3, sobre los antecedentes, se han implementado numerosas técnicas de aprendizaje automático para abordar el problema de la clasificación en la talasemia. Sin embargo, estas técnicas no se han explorado en profundidad para el descubrimiento de nuevos biomarcadores asociados a la enfermedad.

Con el fin de proveer una solución para este problema, se plantea la pregunta: ¿Qué técnicas de aprendizaje automático son capaces de brindar nueva información sobre potenciales biomarcadores asociados a la talasemia?

Asimismo, se plantea la pregunta complementaria: ¿Cuáles modelos presentan los mejores valores para las métricas de desempeño, definidas en el ámbito (Sensibilidad y Especificidad), para la identificación de biomarcadores?

[1] Posibles indicadores de procesos biológicos normales o patogénicos [12].

1.2. Justificación de la Investigación

Como se ha mencionado, mientras que la utilidad de las técnicas de aprendizaje automático ya ha sido demostrada, los métodos moleculares tradicionales presentan limitaciones, exacerbadas por problemas en la infraestructura de manejo de enfermedades. Un ejemplo es lo ocurrido en China durante la primera ola de la pandemia de COVID-19. Uno de los requisitos para dar de alta a un paciente era obtener dos pruebas RT-PCR negativas para SARS-CoV-2 en menos de 24 horas. Sin embargo, el 14 % de los pacientes dados de alta en la provincia de Guangdong volvieron a dar positivo entre 5 y 13 días después. Esta situación se atribuye a diversos factores, como los procedimientos de muestreo con hisopos faríngeos, la calidad del tubo de muestreo, el almacenamiento y transporte de las muestras, así como la calidad de los kits de detección [13].

La relevancia del aprendizaje automático en el campo de la salud y la biomedicina se destaca en el desarrollo de sistemas de diagnóstico asistidos por computadora (CAD, por sus siglas en inglés), diseñados para mejorar el proceso de toma de decisiones de los profesionales de la salud. Un CAD puede utilizar imágenes médicas como radiografías, tomografías computarizadas o imágenes histopatológicas para su análisis. No obstante, estos sistemas no se limitan solo a imágenes, también pueden emplear datos clínicos obtenidos de ensayos multiómicos, como datos genómicos, proteómicos, inmunológicos o metabolómicos. Por ejemplo, se han desarrollado sistemas CAD basados en aprendizaje automático para detectar precursores de patologías asociadas con el cáncer a partir de datos de citometría de flujo [14], [15]. En este contexto, dichos sistemas pueden utilizarse para prevenir enfermedades e incluso determinar el estado potencial de una patología en un individuo, lo que permite ajustar el régimen de tratamiento según la condición específica del paciente [16].

El uso de perfiles de expresión genética se propone debido a su disponibilidad a un costo relativamente bajo en la actualidad. Además, han demostrado ser valiosos para comprender los mecanismos de producción de enfermedades, así como en los diagnósticos, pronósticos y la selección de planes de tratamiento [17].

En los últimos años, la investigación sobre la aplicación del aprendizaje automático

en la talasemia ha ganado relevancia. Una revisión de la literatura determinó que la mitad de los estudios en esta área desde 2001 se realizaron entre 2016 y 2020. Además, el 80 % de estos estudios utilizan métodos basados en redes neuronales artificiales. Esto sugiere la oportunidad de desarrollar modelos basados en diferentes algoritmos para identificar biomarcadores asociados con la talasemia y mejorar la capacidad de detección y diagnóstico.

Por otro lado, los estudios sobre talasemia se han realizado principalmente en Tailandia e Italia, países con alta prevalencia de la enfermedad. Sin embargo, la talasemia afecta a nivel mundial. En Costa Rica, por ejemplo, se identificó una frecuencia del 0,25 % del rasgo de β-talasemia menor en una muestra representativa de 12.000 niños en 2006. Por lo tanto, esta enfermedad y su tratamiento no son ajenos al contexto costarricense.

1.3. Objetivos

En esta sección se presentan los objetivos de la investigación propuesta.

1.3.1. Objetivo General

Identificar biomarcadores asociados a los síndromes talasémicos mediante el análisis de perfiles de expresión genética y el uso de modelos predictivos y estadísticos.

1.3.2. Objetivos Específicos

Los objetivos específicos de la investigación incluyen:

1. Caracterizar el conjunto de datos de expresión genética donde se encuentran los genes asociados a la talasemia.

2. Construir modelos de aprendizaje automático con el fin de identificar la asociación de genes con la talasemia.

3. Evaluar los modelos de aprendizaje automático construidos.

4. Identificar un subconjunto de potenciales biomarcadores asociados a la talasemia.

1.4. Limitaciones

El estudio propuesto se centrará en identificar un subconjunto de biomarcadores potencialmente asociados con la talasemia. Sin embargo, la validación clínica de estos biomarcadores no se llevará a cabo, ya que se considera fuera del alcance de este trabajo.

Para concretar el impacto de este trabajo en la detección y el diagnóstico de la talasemia, los resultados obtenidos deben ser revisados por investigadores de otras disciplinas. Serán estos investigadores los responsables de llevar a cabo un proceso de validación experimental y clínica.

1.5. Estructura del Documento

Esta sección detalla la estructura del presente documento. El Capítulo 2 comprende el marco conceptual con las definiciones necesarias para comprender esta investigación. El Capítulo 3 detalla los trabajos similares más relevantes que se han realizado previamente en el área. El Capítulo 4 detalla los métodos y procedimientos planteados para realizar esta investigación. El Capítulo 5 presenta los principales resultados obtenidos, los cuales son posteriormente discutidos en el Capítulo 6. Finalmente, en el Capítulo 7 se presentan las principales conclusiones de este trabajo.

Capítulo II. Marco Teórico

En este capítulo se describen los principales términos necesarios para comprender esta investigación. Para ello se abordarán primero los conceptos relacionados al campo de la biología y luego aquellos relacionados al campo del aprendizaje automático.

2.1. Aspectos Biológicos Asociados a la Talasemia

Inicialmente, se profundizará en algunos aspectos de la talasemia, incluyendo detalles relacionados con la hemoglobina, la proteína donde las mutaciones causan la talasemia. Dado que es un trastorno genético, también es importante abordar conceptos clave de genética como el ADN y los genes, ya que estos últimos tienen una relación causal con la enfermedad. Finalmente, se explicará con más detalle qué son los biomarcadores y de dónde provienen.

2.1.1. Talasemia

Como ya se mencionó en el Capítulo 1, las talasemias corresponden a un grupo de trastornos hereditarios de la sangre en los que se da una producción anormal de hemoglobina y son causadas cuando ocurre una mutación en los genes que controlan la producción de las proteínas globina alfa y globina beta [2], [4]. En la Figura 2.1 se ejemplifica el efecto de la talasemia en las células rojas de la sangre.

La talasemia es solamente el resultado de eventos que ocurren a un nivel molecular. Para entender mejor qué la produce, es necesario ahondar en detalles acerca de las proteínas y los genes. Estos se van a hilar a través de esta sección.

2.1.2. Proteínas

Las proteínas son biomoléculas que se sintetizan a partir de bloques de construcción fundamentales llamados aminoácidos. Estos aminoácidos son conjuntos de moléculas que comparten un núcleo común, pero tienen diferentes cadenas laterales unidas. Es-

tas diferentes cadenas laterales alteran el comportamiento y las características físico-químicas de la proteína, como el tamaño y la estructura. Las proteínas son cadenas de aminoácidos unidos uno al lado del otro [18].

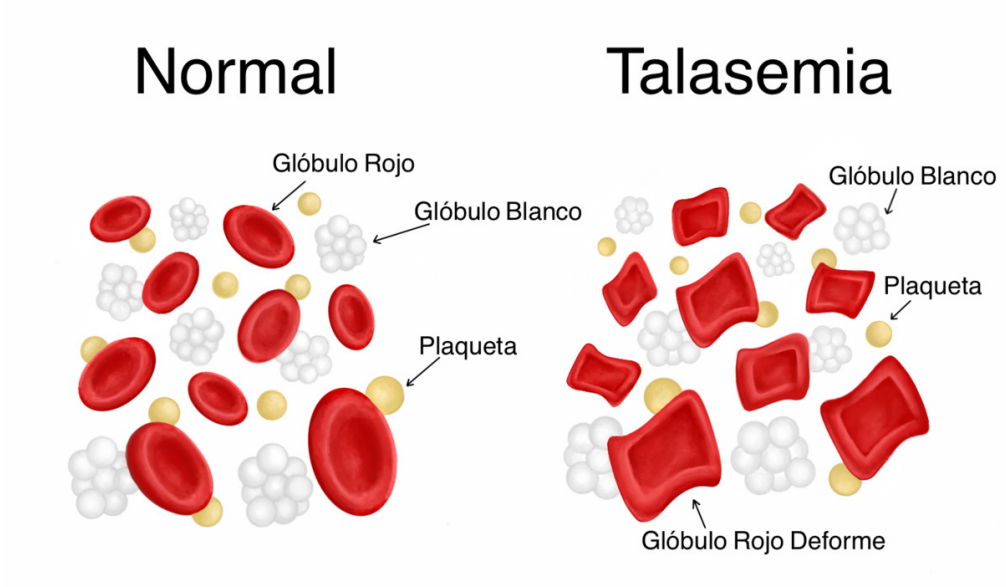

Figura 2.1: Comparación de la forma normal de los glóbulos rojos y la forma de los glóbulos rojos de una persona con talasemia. Adaptada de [19].

La hemoglobina, la cual se produce inadecuadamente en la talasemia, es una proteína de los glóbulos rojos que se une al oxígeno. Esta proteína está conformada por cuatro polipéptidos[1]: dos alfa globinas y dos beta globinas [20].

2.1.3. El ADN y los Genes

El ADN es como un libro que contiene todas las instrucciones para formar un individuo. El ADN es un polímero: una larga cadena de unidades repetitivas unidas entre sí. Cada hebra de ADN está formada por una cadena de cuatro unidades (denominadas bases nucleotídicas): adenina, timina, guanina y citosina. Casi toda la información sobre cómo hacer un organismo vivo está codificada en el patrón específico de estas cuatro unidades que componen su genoma [18]. La información genética es el orden lineal, o secuencia, de las cuatro bases en las hebras de ADN. Estas secuencias están divididas en subconjuntos denominados genes [20].

[1]Cadenas largas de aminoácidos involucrados en la síntesis de proteínas [20].

Una de las principales funciones del ADN es registrar las secuencias de aminoácidos de las proteínas de un organismo. Tramos particulares de ADN corresponden directamente a proteínas particulares. Sin embargo, pasar del ADN a una proteína involucra otra molécula, el ARN, que sirve como una representación intermedia para llevar información de una parte de la célula a otra. El ARN es otro polímero y es químicamente muy similar al ADN. Para crear una proteína, la información debe copiarse dos veces. Primero, la secuencia de ADN se transcribe en una secuencia de ARN equivalente y luego la molécula de ARN se traduce en una molécula de proteína [18].

La codificación de proteínas es uno de los trabajos del ADN. Tramos de ADN (llamados genes codificantes) codifican proteínas mediante un código simple y bien definido. El ADN se convierte en ARN, que sirve solo como portador de información. Luego, el ARN se convierte en proteínas, que hacen todo el trabajo real [18].

Todas las células de un mismo organismo contienen el mismo ADN y los mismos genes. Algunos de esos genes son transcritos por todas las células, dichos genes están relacionados con características estructurales y rutas metabólicas comunes en todas las células. Sin embargo, casi todas las células del cuerpo están especializadas. Por ejemplo, sólo los glóbulos rojos inmaduros utilizan los genes que codifican la globina. La identidad de los genes expresados en un tiempo dado depende de diversos factores, como las condiciones del citoplasma y fluido extracelular, así como del tipo de célula. Estos factores influyen sobre los mecanismos de regulación que rigen la expresión génica [20].

Cuando ocurre un cambio permanente en la secuencia de bases del ADN, se da una mutación. Esto puede ocurrir debido a errores en la duplicación del ADN, a ciertas sustancias químicas o a algunos tipos de radiación. Si una mutación cambia las instrucciones genéticas codificadas en un ADN, puede resultar un producto genético alterado. Algunas mutaciones no tienen ninguna consecuencia. Sin embargo, otras pueden tener consecuencias importantes. Por ejemplo, algunas mutaciones benéficas pueden verse favorecidas por la selección natural. Por el otro lado, la deleción de un nucleótido en particular del ADN del gen de la beta globina es lo que provoca la β-talasemia y reduce la capacidad del cuerpo para transportar oxígeno [20], [21].

Una manera en la que se puede estudiar el comportamiento de los genes y los efectos

de las mutaciones sobre estos es a través de los **perfiles de expresión genética**. Los perfiles de expresión genética miden la expresión de los genes en una muestra en particular [22]. Estos han demostrado ser útiles para determinar la respuesta de cada gen a un cambio en el estado celular, como lo podría ser una enfermedad o la respuesta a un químico [23]. De esta manera se han logrado caracterizar complejas circunstancias biológicas y enfermedades [17].

Existen varios métodos para obtener perfiles de expresión genética, cada uno con sus propias ventajas y limitaciones. Entre estos se encuentran el método de microarreglos de ADN y el de Secuenciación de ARN (*RNA Sequencing* o RNA-seq).

Algunos tipos de **microarreglos** miden la abundancia de un conjunto definido de transcritos mediante el uso de sondas, que corresponden a genes definidos. Por lo tanto, esta técnica requiere un conocimiento previo del organismo de interés [24].

La **secuenciación de ARN** es una técnica de biología molecular utilizada para medir la abundancia y diversidad de moléculas de ARN en una muestra biológica. Proporciona una visión global de la expresión génica mediante la secuenciación de las moléculas de ARN presentes en la muestra. Este método permite analizar el transcriptoma. Esta técnica ha permitido a los científicos descubrir conocimientos sobre la regulación génica, identificar genes expresados diferencialmente, descubrir nuevos transcritos (por ejemplo, ARN no codificantes) y comprender procesos biológicos complejos a nivel molecular [24], [25].

2.1.4. Biomarcadores

Un marcador biológico (**biomarcador**) corresponde a características biológicas que pueden ser objetivamente medidas y evaluadas como un indicador de procesos biológicos normales, procesos patogénicos o respuesta a una intervención terapéutica [12]. Un biomarcador puede ayudar a determinar la predicción, causa, diagnóstico, progresión, regresión, o resultado del tratamiento de una enfermedad. Esto permite identificar individuos que están destinados a padecer una enfermedad u objetivos para ensayos clínicos. Los biomarcadores que muestran signos iniciales de una enfermedad permiten un diagnóstico más temprano o permiten determinar el resultado de interés en una eta-

pa más primitiva de la enfermedad [12]. En el caso de la talasemia, los biomarcadores identificados representarían indicadores de la presencia de las patologías relacionadas a la enfermedad.

2.2. Aprendizaje Automático y su Aplicación en Biología y Medicina

El **aprendizaje automático** se refiere a un conjunto de temas que implican la creación y evaluación de algoritmos que facilitan el reconocimiento de patrones, clasificación y predicción, basados en modelos derivados de datos existentes [1]. El aprendizaje automático se ha utilizado y sigue siendo útil para la interpretación de grandes conjuntos de datos genómicos [26].

Hay diversos tipos diferentes de sistemas de aprendizaje automático, por lo que es útil clasificarlos en categorías amplias según algunos criterios. Uno de estos criterios corresponde a la cantidad y el tipo de supervisión que reciben los sistemas durante el entrenamiento. De acuerdo con este, hay cuatro categorías principales: aprendizaje supervisado, aprendizaje no supervisado, aprendizaje semi supervisado y aprendizaje por refuerzo.

Para esta investigación, resulta pertinente profundizar en el aprendizaje supervisado y el aprendizaje no supervisado. En el caso del **aprendizaje supervisado**, el conjunto de entrenamiento que se le muestra al algoritmo incluye las soluciones deseadas, llamadas etiquetas. Se habla de que esta categoría consiste en un proceso de utilizar la experiencia para ganar pericia [27]. Dependiendo del tipo de salida, el problema puede ser categorizado como de clasificación o de regresión [28]. En los problemas de **clasificación**, como el que se pretendía estudiar, el algoritmo se entrena con instancias y sus respectivas clases y debe aprender a clasificar nuevas instancias [29].

Dado que en un modelo de aprendizaje supervisado se disponen de las etiquetas reales de cada instancia, es posible evaluar el desempeño del modelo. Una manera sencilla de hacerlo es contar cuántas veces las instancias de una clase específica fueron clasificadas correctamente (verdaderos positivos y verdaderos negativos) y cuántas ve-

ces fueron clasificadas incorrectamente (falsos negativos y falsos positivos) [29]. En el contexto de esta investigación, un falso positivo sería una instancia no relacionada con la talasemia que fue clasificada erróneamente como asociada a la talasemia.

Sin embargo, es habitual preferir el uso de métricas más concisas para evaluar el desempeño de un clasificador. En [29, Tabla 2.1] se presenta una descripción de las métricas de desempeño de clasificación más comúnmente utilizadas.

Tabla 2.1: Métricas de desempeño de clasificación comúnmente utilizadas [29].

Métrica	**Descripción**	**Cálculo**
Exactitud	Proporción de las instancias clasificadas correctamente.	$\frac{TP+TN}{TP+FN+TN+FP}$
Precisión	Proporción de las instancias clasificadas como positivas que fueron clasificadas correctamente.	$\frac{TP}{TP+FP}$
Sensibilidad (*Recall*)	Proporción de las instancias positivas que fueron clasificadas correctamente. También se conoce como tasa de verdaderos positivos (TPR).	$\frac{TP}{TP+FN}$
Puntuación F_1	Media harmónica de la precisión y la sensibilidad.	$F_1 = 2 \times \frac{\text{precisión} \times \text{sensibilidad}}{\text{precisión} + \text{sensibilidad}} = \frac{TP}{TP+\frac{FN+FP}{2}}$
AUC-ROC	Indicador de la calidad de la clasificación. Un valor de 1 corresponde a un clasificador perfecto, mientras que un valor de 0,5 equivale a un clasificador aleatorio.	No se puede calcular mediante una ecuación.

Una técnica para llevar a cabo la evaluación de un modelo de clasificación consiste en la **validación cruzada** con k particiones. En esta técnica, el conjunto de datos se divide en k particiones, donde $k-1$ particiones se utilizan para entrenar el modelo mientras que la partición restante se utiliza para evaluarlo. Así sucesivamente hasta haber evaluado el modelo con cada una de las k particiones mientras que realiza el entrenamiento con las $k-1$ restantes. Al utilizar esta técnica se obtienen k puntuaciones de las que se puede obtener un promedio y una desviación estándar [29]. Los valores más comunes para k son $k=3$, $k=5$ y $k=10$. A pesar de ello, la elección del valor de k usualmente depende del tamaño del conjunto de datos, en donde, para conjuntos más grandes, el uso de $k=3$ o $k=5$ es adecuado, pero para datos de mediana y pequeña escala, $k=10$

es ampliamente recomendado.

En la Tabla 2.2 se describen algunos algoritmos de clasificación que ya han demostrado su utilidad en contextos biológicos.

Tabla 2.2: Algoritmos de clasificación.

Algoritmo	**Descripción**	**Principales parámetros**
Regresión logística	Modelo lineal para clasificación binaria y multinomial [30], [31].	Tasa de aprendizaje, regularización (L1, L2), C (regularización inversa)
Máquinas de Soporte Vectorial (SVM)	Clasificador que encuentra el hiperplano óptimo que separa las clases [32], [33].	Kernel (lineal, rbf, etc.), C, gamma
Bosque aleatorio	Conjunto de árboles de decisión que agregan sus resultados para tomar una decisión global [34].	Número de árboles, Características mínimas por nodo, características máximas, *bootstrap*, criterio, profundidad
eXtreme Gradient Boosting (XGBoost)	Algoritmo de refuerzo de gradiente que optimiza los aprendices débiles [35].	Tasa de aprendizaje, número de árboles, profundidad del árbol, Gamma, regularización
Tree Based Pipeline Optimization Tool (TPOT)	Herramienta automatizada de aprendizaje para optimizar *pipelines* [36].	Generaciones, tamaño de población, tasa de cruzamiento, tasa de mutación

En el caso del **aprendizaje no supervisado** los datos de entrenamiento no están etiquetados. Por lo tanto, el sistema intenta aprender sin una guía. Para implementar este tipo de sistema se propuso el uso de técnicas de detección de anomalías. Se considera este enfoque ya que los algoritmos de detección de anomalías pueden ayudar a identificar mutaciones raras que podrían no ser evidentes de inmediato mediante métodos tradicionales. Estos algoritmos están diseñados para detectar desviaciones de la norma, lo que los hace efectivos para identificar variaciones genéticas poco comunes que podrían ser de gran interés para la investigación o propósitos clínicos [37].

Para esta investigación se aborda el uso de un modelo predictivo basado en bosque de aislamiento. La idea principal detrás de este algoritmo es aislar anomalías (valores atípicos) en un conjunto de datos mediante la creación de un conjunto de árboles de aislamiento. Estos árboles se construyen seleccionando aleatoriamente una característica y un valor de división aleatorio para esa característica en cada nodo del árbol. Las

anomalías tienen más probabilidades de ser aisladas cerca de la raíz del árbol, mientras que los puntos de datos normales tienden a ser aislados en niveles más profundos del árbol. Los usuarios pueden medir el puntaje de anomalía de un punto de datos contando el número de divisiones necesarias para aislarlo. Las anomalías tendrán puntajes más bajos, mientras que los puntos de datos normales tendrán puntajes más altos. El bosque de aislamiento también es relativamente sencillo de implementar y tiene diversas aplicaciones, como la detección de fraudes, la seguridad de redes y la detección de valores atípicos en el preprocesamiento de datos [29], [38].

El aprendizaje puede ser ajustado mediante **hiperparámetros**, los cuales son parámetros del algoritmo y no del modelo en sí. De esta manera, un hiperparámetro no es influenciado por el algoritmo, sino que debe ser definido antes del entrenamiento y mantenerse constante durante todo el proceso [29].

Existen formas de encontrar conjuntos de hiperparámetros que proporcionan un rendimiento sobresaliente para un modelo de aprendizaje automático. Este proceso recibe el nombre de **optimización de hiperparámetros**. Dos distintos enfoques corresponden a la búsqueda en rejilla y a la búsqueda aleatoria. La **búsqueda en rejilla** consiste en evaluar un modelo utilizando todas las combinaciones posibles de hiperparámetros de una rejilla predefinida. Este método garantiza que se encuentren los mejores hiperparámetros dentro de los conjuntos de valores especificados, pero puede ser costoso desde el punto de vista computacional. La principal desventaja de este método es que únicamente prueba las combinaciones especificadas, pudiendo omitir valores mejores que se encuentren fuera de la rejilla. La **búsqueda aleatoria** implica el muestreo de un número fijo de combinaciones de hiperparámetros a partir de una distribución especificada. Este método no evalúa todas las combinaciones posibles, sino que selecciona un subconjunto aleatorio, lo que lo hace más rápido y escalable para modelos con muchos hiperparámetros. Sin embargo, debido a su carácter estocástico, no hay garantía de encontrar los hiperparámetros óptimos [29].

En un proyecto de aprendizaje automático, siguiendo el modelo de referencia **CRISP-DM**, el primer paso es comprender el problema y definir los objetivos. Es crucial determinar qué tipo de datos se pueden recolectar si es necesario. A continuación, se debe

realizar un preprocesamiento de los datos para que los algoritmos de aprendizaje automático puedan generar un modelo. La preparación de los datos y el modelado suelen ir de la mano, ya que los resultados obtenidos permiten evaluar el impacto de las técnicas de preprocesamiento seleccionadas [39].

Una vez que se ha generado un modelo, el siguiente paso es evaluarlo. Es crucial verificar si la representación que el modelo hace de los datos posee valor predictivo. Si la evaluación muestra resultados deficientes, podría ser necesario retroceder y ajustar el proceso [39]. Estos pasos pueden complementarse con la selección de herramientas antes del modelado y un paso adicional de interpretación de resultados [40]. La Figura 2.2 resume estos pasos.

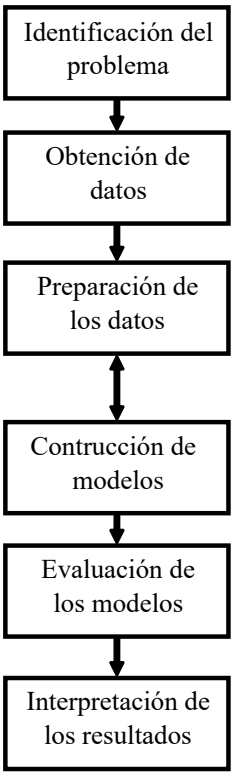

Figura 2.2: Flujo de trabajo de un proyecto de aprendizaje automático.

Es posible realizar un **Análisis de Enriquecimiento de Conjuntos de Genes** (Gene Set Enrichment Analysis o GSEA) sobre un grupo de genes. Este análisis tiene como objetivo identificar si conjuntos predefinidos de genes, a menudo relacionados con vías o funciones biológicas específicas, están estadísticamente enriquecidos o sobrerrepresentados en una lista clasificada. Este también evalúa si los genes de un determinado

conjunto aparecen en la parte superior o inferior de la lista clasificada con más frecuencia de lo esperado por casualidad, proporcionando información sobre los procesos biológicos que podrían ser relevantes en un contexto analítico particular [41].

Para medir el nivel de significancia de un análisis de enriquecimiento se utiliza el **valor *p* ajustado**. El valor p ajustado es una medida estadística utilizada en las pruebas de hipótesis múltiples para controlar la tasa global de error de tipo I (falsos positivos). Al realizar comparaciones múltiples, la probabilidad de encontrar un resultado estadísticamente significativo por azar aumenta con el número de pruebas. El valor p ajustado lo corrige modificando los valores p para tener en cuenta el número de comparaciones realizadas [42].

En el próximo capítulo, se presentarán en detalle los hallazgos obtenidos a partir de una revisión de la literatura sobre esta temática. Además, se mencionarán algunos de los estudios más relevantes que se han identificado y que son pertinentes para esta investigación.

Capítulo III. Antecedentes

Antes de llevar a cabo esta investigación, se realizó una revisión de la literatura para explorar el papel que el aprendizaje automático ha desempeñado en la investigación sobre la talasemia.

En esta revisión de literatura se evaluaron 26 artículos académicos provenientes de las bases de datos IEEE Xplore, ScienceDirect y Scopus. En total, 21 de estos estudios fueron seleccionados para la síntesis. Entre los principales hallazgos, se encontró:

- La cantidad de estudios del 2016 al 2020 corresponde a la misma que los estudios llevados a cabo desde el 2001 hasta el 2020 (Figura 3.1).
- Los algoritmos de aprendizaje automático más utilizados corresponden a aquellos basados en redes neuronales artificiales (ANN) y máquinas de soporte vectorial (SVM). En la Figura 3.2 se pueden observar los diversos algoritmos implementados en el estudio de la talasemia y las tendencias recientes de aplicar métodos basados en ANN a este problema.

Estos hallazgos indican que si bien se han planteado soluciones para el diagnóstico de síndromes talasémicos, no se ha explorado mucho la identificación de biomarcadores. Lo cual, sin embargo, representa un punto de partida para llevar a cabo esta investigación.

En cuanto a los estudios con mejores métricas de desempeño, en [43] se implementó árbol de decisión C4.5, bosque aleatorio y perceptrón multicapa (algoritmo basado en ANN) para investigar la posibilidad de utilizar datos de tipificación de hemoglobina para clasificación automática de talasemia utilizando datos de cromatografía líquida. El algoritmo con mejor desempeño fue árbol de decisión C4.5 con métricas de 97,24 % de exactitud, 99,78 % especificidad y 97,24 % sensibilidad. Además, se implementó selección de variables por correlación.

A continuación se mencionan algunas investigaciones que ejemplifican la aplicación de las técnicas de aprendizaje automático en el estudio de la talasemia.

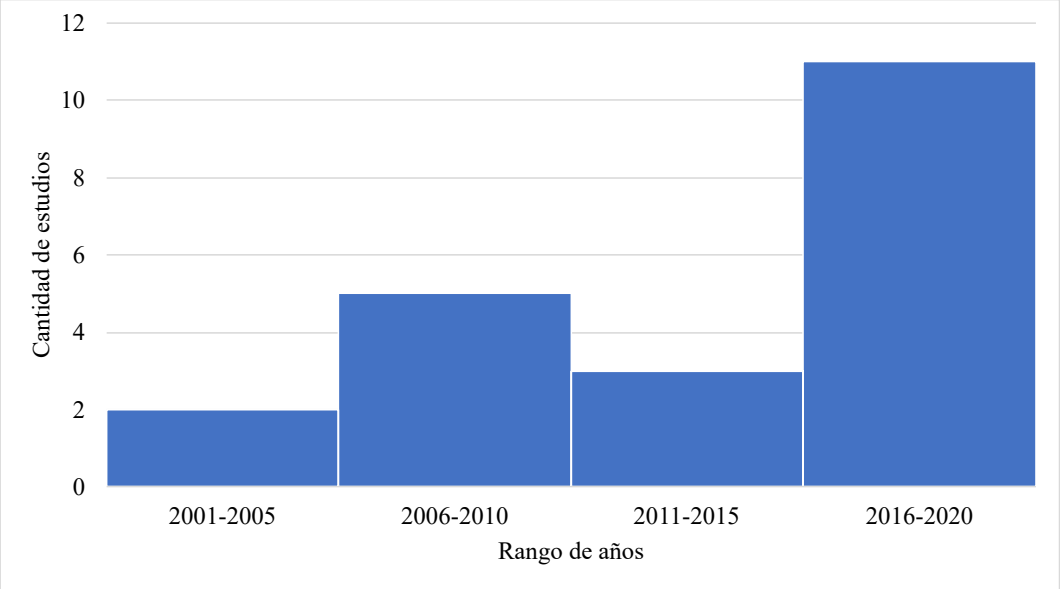

Figura 3.1: Distribución de los estudios relacionados de acuerdo al período de años en el que fueron publicados.

En [44] se implementó un modelo basado en redes neuronales artificiales (ANN) para separar personas portadoras de la enfermedad de la población general utilizando datos de conteo sanguíneo completo. El modelo presentó métricas de sensibilidad de 100 % y especificidad de 96,7 %.

En [9] se implementó k vecinos más cercanos (KNN), perceptrón multicapa (método basado en ANN), clasificador bayesiano ingenuo y árbol de decisión J48 para la proposición de un modelo de minería de datos para la detección de β-talasemia utilizando exámenes simples de laboratorio (conteo sanguíneo completo). El mejor algoritmo fue clasificador bayesiano ingenuo y presentó métricas de exactitud de 99,47 %, especificidad de 99,48 %, sensibilidad de 98,81 % y AUC-ROC de 0,9978. Además, se utilizó SMOTE como técnica de balanceo de datos.

En [45] se implementó máquina de soporte vectorial (SVM) con el fin de hallar una fórmula confiable para la identificación de portadores de β-talasemia utilizando datos de conteo sanguíneo completo y datos de cromatografía líquida. El modelo presentó una sensibilidad de 98,88 % y especificidad de 88,18 %.

En [46] se implementó un modelo basado en máquina de soporte vectorial (SVM) y k vecinos más cercanos (KNN) para la diferenciación entre anemia por deficiencia de

hierro y β-talasemia usando datos de Conteo Sanguíneo Completo y sexo biológico. El algoritmo con mejor desempeño fue k vecinos más cercanos (KNN) con una exactitud de 95,3 %, especificidad de 94,1 %, sensibilidad de 96,2 % para KNN y AUC-ROC de 0,97. Además, en el estudio se implementó análisis de componentes de vecindario (NCA) para selección de variables.

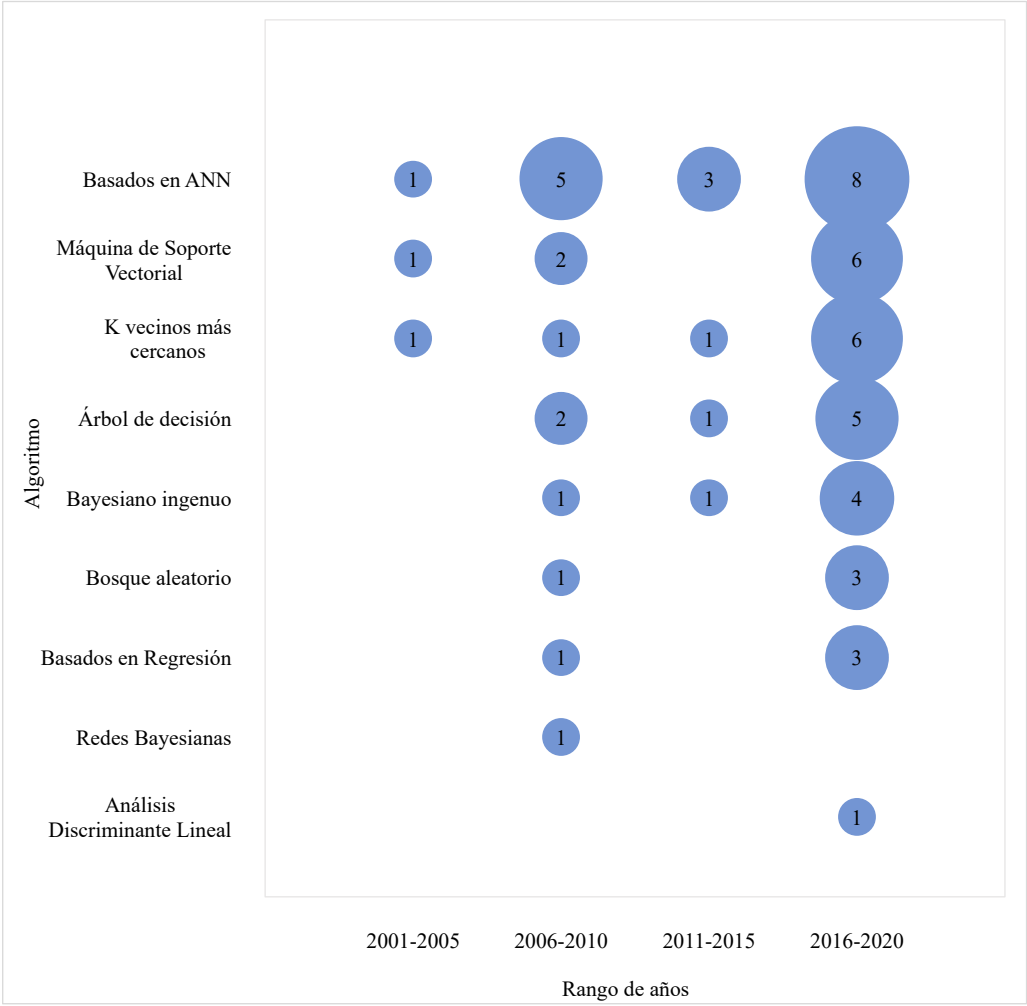

Figura 3.2: Usos de cada algoritmo identificado por período de años.

Uno de los estudios más relevantes en el contexto de esta investigación, es el realizado por Sikandar, Sohail, Saeed et al. [47]. En este estudio, se aborda la relación gen-enfermedad y cómo una mutación en un solo gen es capaz de alterar procesos biológicos, conduciendo a una enfermedad. Sin embargo, al interactuar los genes con otros, se pueden alterar no solo uno, sino múltiples procesos biológicos causando múltiples enfermedades. Es por eso que en su estudio, Sikandar, Sohail, Saeed et al. [47]

plantean la exploración de nuevos métodos computacionales para estudiar la relación gen-enfermedad con el fin de identificar o priorizar genes asociados con una enfermedad basados en secuencias o interacciones de genes. Esta tarea se basa en calcular la similitud entre genes conocidos de una enfermedad y un gen dado. En el estudio, se trabaja con características biológicas avanzadas que no habían sido utilizadas previamente, como interacciones proteína-proteína. Con esto, entrenaron diversos modelos computacionales para clasificar los genes en cuatro diferentes enfermedades, siendo la talasemia una de ellas.

En su estudio, Sikandar, Sohail, Saeed et al. [47] implementaron una máquina de aprendizaje extremo profundo (algoritmo basado en ANN), bosque aleatorio, clasificación vía regresión y árbol de clasificación y regresión. El mejor modelo implementado fue el de bosque aleatorio con un AUC-ROC de 0,991.

La principal diferencia entre esta investigación y la realizada por Sikandar, Sohail, Saeed et al. [47] es que su estudio se enfoca en evaluar los métodos computacionales propuestos mientras que la finalidad de esta investigación es identificar un subconjunto de potenciales biomarcadores sin enfocarse en el método. Asimismo, el conjunto de datos propuesto es distinto, en esta investigación no se plantea el uso de características topológicas de los genes. En este caso se plantea utilizar únicamente la expresión diferencial.

Taghavifar, Hamid y Shariati [48] intentan mejorar la comprensión de la enfermedad realizando análisis de expresión génica diferencial. Para ello, utilizan datos de expresión génica de un individuo que tiene el fenotipo de la enfermedad (hija/afectada), un individuo con mutaciones que no tiene el fenotipo de la enfermedad (madre) y un individuo que no tiene mutaciones relacionadas y no tiene el fenotipo de la enfermedad (control/no afectado). Taghavifar et al. obtuvieron estos datos a partir de RNA-seq. Este trabajo integró aprendizaje automático y GSEA para identificar posibles marcadores candidatos asociados a posibles ontologías génicas relacionadas con la talasemia.

Considerando los trabajos anteriores que fueron encontrados, en el siguiente capítulo se establecerá la metodología seguida para la identificación de biomarcadores asociados a síndromes talasémicos utilizando algoritmos de aprendizaje automático.

Capítulo IV. Metodología

En este capítulo se detallan tanto el diseño de la investigación como las actividades y los pasos seguidos para cumplir con los objetivos planteados anteriormente. Mediante estas actividades se esperaba obtener los resultados necesarios para plantear la pregunta de investigación. Se abordarán por separado dos distintos enfoques utilizados, los cuales corresponden a i) clasificación y ii) detección de anomalías.

Para esta investigación se propuso un enfoque positivista en el que se buscó explicar de forma objetiva los datos obtenidos. Esto se basa en un análisis cuantitativo, ya que los datos que se recolectarán están asociados a las métricas de desempeño de los modelos de aprendizaje automático construidos.

Para el planteamiento de las actividades se tomaron como referencia los pasos del modelo CRISP-DM [39] y se complementan con los pasos planteados por François [40], todos estos previamente detallados en el Capítulo 2.

4.1. Enfoque de Clasificación

Inicialmente se abarcará el planteamiento de la investigación para el enfoque de clasificación.

4.1.1. Diseño de la Investigación

Se diseñó y ejecutó un experimento donde las unidades experimentales corresponden a los modelos de aprendizaje automático y entre los factores se encuentran los algoritmos escogidos, los hiperparámetros y las distintas técnicas necesarias de preprocesamiento de los datos.

4.1.2. Actividades

A continuación se describen las actividades realizada para el enfoque de clasificación.

Actividades para el Objetivo Específico 1

Esta etapa inicial se enfocó en el conjunto de datos y su procesamiento. Anteriormente se realizaron pruebas de concepto utilizando distintas conformaciones y procesamientos del conjunto de datos [49]. Para la conformación del conjunto de datos se utilizaron diferentes fuentes. Una de estas corresponde a la base de datos DisGeNET [50] (https://www.disgenet.org/), de la cual se recuperó una lista de genes asociados tanto a la α-talasemia como a la β-talasemia. Sin embargo, debido a que la α-talasemia es usualmente fatal [2], se mantuvieron únicamente los genes asociados a la β-talasemia. Esta lista contiene 198 genes asociados a la enfermedad. También se recuperó de esta base de datos una lista de 294 genes asociados a la vasculitis. Estos genes se utilizaron como genes negativos, es decir, genes no asociados a la talasemia. Fue necesario remover 25 genes de cada una de estas listas debido a que estaban asociados a ambas enfermedades.

En cuanto a los datos de expresión genética, se recuperó un conjunto de datos de un experimento de expresión génica de células mononucleares de sangre periférica en adultos sanos transportados rápidamente a gran altitud obtenidos mediante la tecnología de microarreglos (GSE46480) [51]. Estos datos se recuperaron del *Gene Expression Omnibus* (GEO) [52] y únicamente se utilizaron los datos obtenidos antes de transportar a los individuos.

Este conjunto fue elegido debido a que, tras realizar una búsqueda en el GEO, este resultó ser el único que cumplía con los criterios específicos requeridos para este estudio (células sanguíneas y pacientes sanos). Cabe destacar que la generación de perfiles de expresión génica de células sanguíneas desde cero implica un proceso experimental complejo, que incluye la extracción y purificación de ARN, su posterior cuantificación, y la realización de experimentos de microarreglos o RNA-seq, todos ellos procedimientos que requieren equipamiento especializado, estrictos controles de calidad y experiencia técnica en biología molecular [25]. Por esta razón, el uso de datos públicos disponibles no solo optimiza los recursos y el tiempo de investigación, sino que también permite focalizar los esfuerzos en el análisis computacional y la interpretación biológica de los resultados [53].

Una vez obtenidos todos los datos necesarios, se llevó a cabo el procesamiento del conjunto. Esto incluye la remoción de pseudogenes [54]. De esta manera se preparó así un archivo con los genes etiquetados para entrenamiento y otro archivo con los genes no etiquetados. Asimismo, se aplicó una transformación log_2 a los valores de expresión, en la cuál todos los valores se reemplazan por su logaritmo base 2. Esta transformación ha sido útil en el contexto biológico para reducir los efectos de las diferencias entre muestras y dentro de las muestras [55]. El conjunto de datos de entrenamiento resultó conformado por 160 genes asociados a la talasemia y 253 asociados a la vasculitis.

Actividades para el Objetivo Específico 2

En esta etapa se construyeron los diversos modelos de clasificación con los cuales se llevaría a cabo la clasificación de los genes no etiquetados. Los modelos que se construyeron se basaron en los siguientes algoritmos:

- Regresión logística
- Bosque aleatorio
- Máquina de soporte vectorial SVM
- Votación por conjunto (empleando los modelos basados en los tres anteriores)
- XGBoost
- Modelo obtenido mediante TPOT
- Modelo obtenido mediante TPOT empleando Torch para habilitar el uso de GPU

La escogencia de los modelos se basó en los hallazgos presentados previamente en el Capítulo 3, referente a los antecedentes. Es debido a esto que se utilizó una vez más el algoritmo de máquinas de soporte vectorial. En el caso de los árboles de decisión, se decidió utilizar en su lugar los algoritmos de bosque aleatorio y XGBoost, los cuales representan una versión más robusta [34], [35]. El algoritmo de regresión logística fue seleccionado debido a su facilidad de implementación, interpretabilidad y su frecuente uso como base de partida en casos de clasificación binaria [30]. No se utilizó un modelo

en particular basado en redes neuronales ya que, dos modelos basados en este método pueden tener topologías muy distintas [56]. Por eso, en su lugar se habilitó el uso de GPU en la configuración de TPOT para que este considerara posibles soluciones empleando redes neuronales [36].

En el proceso de construcción de los modelos de regresión logística, bosque aleatorio, SVM y XGBoost, se utilizó la búsqueda aleatoria para la optimización de hiperparámetros.

Actividades para el Objetivo Específico 3

Como parte de la construcción de los modelos se implementó una estrategia de validación cruzada estratificada con repeticiones. En este caso, se utilizó concretamente validación cruzada estratificada con 3 particiones y 4 repeticiones. Este número de particiones se escogió con el fin de disminuir la variabilidad en las métricas de desempeño, ya que así, la partición de validación contiene una mayor cantidad de instancias si el conjunto se divide en 3 que al dividirlo, por ejemplo, en 10 [57].

La escogencia de esta técnica responde al tamaño del conjunto de datos. Al tener una cantidad limitada de genes, el uso de validación cruzada representa un uso más eficiente de los datos disponibles en comparación a estrategias como dividir el conjunto en una proporción 80:20. A su vez, el uso de repeticiones agrega una capa de robustez, principalmente teniendo en cuenta el tamaño del conjunto, al realizar nuevas particiones de manera aleatoria en cada iteración. Asimismo, el hecho de validar el modelo usando múltiples particiones distintas para entrenamiento y múltiples particiones distintas para validación reduce el sobreajuste en contraste con utilizar una única partición de entrenamiento y de validación para todo el proceso [30].

La validación se realizó de forma integrada como parte del ajuste de hiperparámetros mediante búsqueda aleatoria. De esta manera, cada combinación de valores fue evaluada para seleccionar la que brindara mejores resultados. La métrica que se buscó maximizar fue la puntuación F_1. En este caso, la puntuación F_1 ayuda a garantizar que el modelo no esté sesgado hacia la predicción de la clase mayoritaria con demasiada frecuencia (lo que podría inflar la precisión pero conducir a un rendimiento pobre en la identificación

de la clase minoritaria) [29]. Maximizar la puntuación F_1 responde al interés de obtener un modelo que funcione bien tanto en la detección de verdaderos positivos como en la minimización de falsos negativos, lo que es importante a la hora de identificar genes asociados a una enfermedad.

Posteriormente, los modelos deben ser aceptados de acuerdo al valor alcanzado para las métricas de desempeño. Tradicionalmente, en los sistemas CAD se prioriza tener una baja tasa de falsos negativos debido a la importancia de identificar correctamente las condiciones críticas o enfermedades. Esto ocurre cuando el sistema no detecta una enfermedad o condición que realmente está presente. Si el sistema no detecta una enfermedad que está presente, el paciente podría no recibir el tratamiento adecuado o a tiempo, lo que podría agravar su condición, reducir las posibilidades de recuperación o incluso poner en peligro su vida [58]. De igual manera, en muchas enfermedades, como el cáncer, la detección temprana es crucial para mejorar los resultados del tratamiento. Un falso negativo podría retrasar el diagnóstico, reduciendo la efectividad de las intervenciones [59]. Es por ello que, debido a la importancia de esta métrica en el diagnóstico de enfermedades, la aceptación de los modelos se basó en el valor de la sensibilidad. Como se mostró anteriormente, esta métrica indica la proporción de instancias positivas clasificadas correctamente. Las demás métricas de desempeño se utilizaron para obtener información adicional de estos, pero no para definir si serían utilizados en la obtención de genes candidatos.

En cuanto al umbral de aceptación, este puede variar en función del ámbito y del tipo de problema. Tomando en cuenta este contexto, en el que los hallazgos no tienen una repercusión directa en la salud de los pacientes, se propone utilizar un valor del 70 %. Este es un valor que se adapta a múltiples contextos [30] y, en el ámbito de este estudio, una puntuación superior a $0,70$ implica que el modelo logra, con un éxito al menos moderado, predecir correctamente la mayoría de las instancias positivas [29]. Adicionalmente, es un valor cercano a los resultados obtenidos en estudios similares con otras enfermedades [60]. Por lo tanto, este fue el valor mínimo definido y esperado para esta investigación.

Actividades para el Objetivo Específico 4

Una vez construidos los modelos, se tomarían el desempeño alcanzado (Tabla 2.1) y la complejidad del modelo para escoger uno y realizar una predicción sobre el conjunto de datos no etiquetado. Sin embargo, ya que posteriormente las métricas de desempeño no alcanzaron los valores mínimos esperados, se decidió continuar con un enfoque de detección de anomalías, el cual será cubierto en la siguiente sección.

4.1.3. Flujo de Trabajo

En la Figura 4.1 se presenta el diagrama de flujo de las actividades llevadas a cabo en la implementación del enfoque de clasificación.

Las actividades para el objetivo específico 1 se enfocan en la creación de un conjunto etiquetado con los genes asociados a la talasemia y los genes asociados a la vasculitis. Esto implica la limpieza y procesamiento del conjunto de datos, como lo es la aplicación de $\log_2$, la eliminación de pseudogenes y el etiquetado.

Las actividades para los objetivos específicos 2 y 3 se enfocan en la construcción de los modelos de clasificación y la obtención de las métricas de desempeño. Esto involucra el proceso de entrenamiento y validación. Este se representa de forma cíclica debido a que se construyen modelos con diferentes combinaciones de valores para los hiperparámetros para cada algoritmo.

Las actividades para el objetivo específico 4 se asocian a la predicción de genes candidatos y al análisis de enriquecimiento. Sin embargo, debido a que los valores para las métricas de desempeño no alcanzaron el umbral de aceptación, se procedió a cambiar a una estrategia de detección de anomalías.

4.2. Enfoque de Detección de Anomalías

Ahora se cubrirá el planteamiento de la investigación para el enfoque de detección de anomalías.

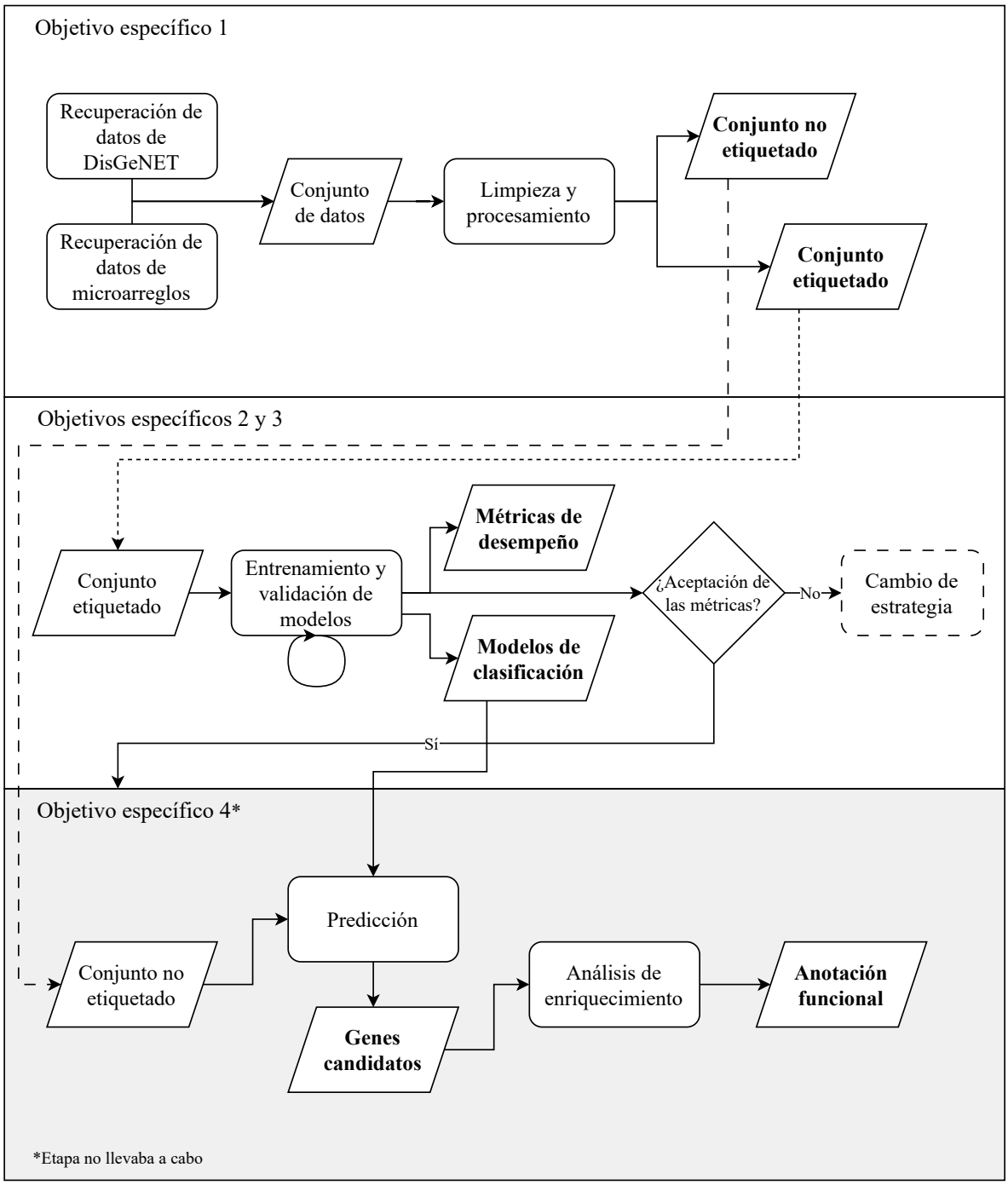

Figura 4.1: Flujo de trabajo de las actividades llevadas a cabo en la implementación del enfoque de clasificación.

4.2.1. Diseño de la Investigación

Se diseñó y ejecutó un experimento donde las unidades experimentales corresponden a los genes y entre los factores se encuentran el método de detección de anomalías y los

hiperparámetros.

4.2.2. Actividades

A continuación se describen las actividades realizadas para el enfoque de detección de anomalías.

Actividades para el Objetivo Específico 1

Para este enfoque, se utilizaron nuevamente los mismos datos recuperados de DisGeNET [50] relacionados a β-talasemia.

Los datos de expresión génica corresponden a los publicados en el estudio de Taghavifar, Hamid y Shariati [48]. En su artículo, los autores intentan mejorar la comprensión de la enfermedad realizando análisis de expresión génica diferencial. Para ello, utilizan datos de expresión génica de un individuo que tiene el fenotipo de la enfermedad (hija/afectada), un individuo con mutaciones que no tiene el fenotipo de la enfermedad (madre), y un individuo que no tiene mutaciones relacionadas y no tiene el fenotipo de la enfermedad (control/no afectado). Taghavifar, Hamid y Shariati [48] obtuvieron estos datos a partir de RNA-seq. Para efectos de este trabajo, fueron utilizados únicamente los datos de la hija y del individuo de control. Se utilizaron estos datos de expresión génica, tanto de control como afectados, como las características, es decir, las columnas del conjunto de datos. De este modo fue posible obtener un valor de expresión génica para dos condiciones diferentes [61].

Para formar el conjunto de datos no etiquetados, se utilizaron genes que no han sido notificados a DisGeNET por ningún investigador. Por lo tanto, se considera que ningún estudio ha asociado estos genes con la talasemia [61].

Actividades para el Objetivo Específico 2

Para el desarrollo de este enfoque, se utilizó un modelo predictivo basado en la detección de anomalías. Sin embargo, inicialmente se utilizó un método estadístico para validar este modelo. Para ello, se calculó la diferencia en los valores de expresión génica del paciente de control frente al paciente afectado. Luego, se tomó el 5 % de los genes (percentil 95) con la mayor diferencia entre las dos condiciones [62].

El motivo de utilizar el percentil 95, es que, en una distribución normal, aproximadamente el 95 % de los datos caen dentro de dos desviaciones estándar ($\pm 2\sigma$) de la media. Los valores que se sitúan fuera de este intervalo se consideran inusuales o extremos porque, estadísticamente, son raros (representan sólo el 5 % de los datos). En resumen, el umbral del 95 % se utiliza habitualmente porque identifica eficazmente los valores que se desvían significativamente de la mayoría de los datos, lo que lo convierte en una regla empírica útil en muchos análisis estadísticos [62].

Posteriormente, se generó un modelo predictivo basado en el algoritmo de bosque de aislamiento. Se consideró este método porque los algoritmos de detección de anomalías pueden ayudar a identificar mutaciones raras que pueden no ser inmediatamente evidentes a través de los métodos tradicionales. Estos algoritmos están diseñados para detectar desviaciones de la norma, lo que los hace eficaces en la detección de variaciones genéticas raras que podrían ser de gran interés para la investigación o con fines clínicos [37].

Con este algoritmo, se generó un modelo entrenado con los genes previamente recuperados de DisGeNET (conjunto de entrenamiento). Para su construcción, se definió un valor de contaminación del 5 %. La escogencia de este valor responde al mismo motivo de haber utilizado el percentil 95 para el modelo estadístico.

Actividades para el Objetivo Específico 3

Para validar el método de bosque de aislamiento, se compararon los genes catalogados como anómalos mediante este modelo con los genes catalogados de la misma forma por el modelo estadístico. De esta manera, se utiliza un modelo para respaldar la validez de otro al obtener resultados similares [61].

Actividades para el Objetivo Específico 4

Se realizó una predicción sobre el conjunto no etiquetado utilizando el modelo de bosque de aislamiento para catalogar los genes como anómalos o no. A continuación, se validaron los genes candidatos obtenidos a partir de los resultados del modelo realizando un GSEA [41]. De esta manera se verificó si hay alguna relación entre la talasemia y los procesos biológicos en los que participan estos genes.

4.2.3. Flujo de Trabajo

En la Figura 4.2 se presenta al diagrama de flujo de las actividades llevadas a cabo en la implementación del enfoque de detección de anomalías.

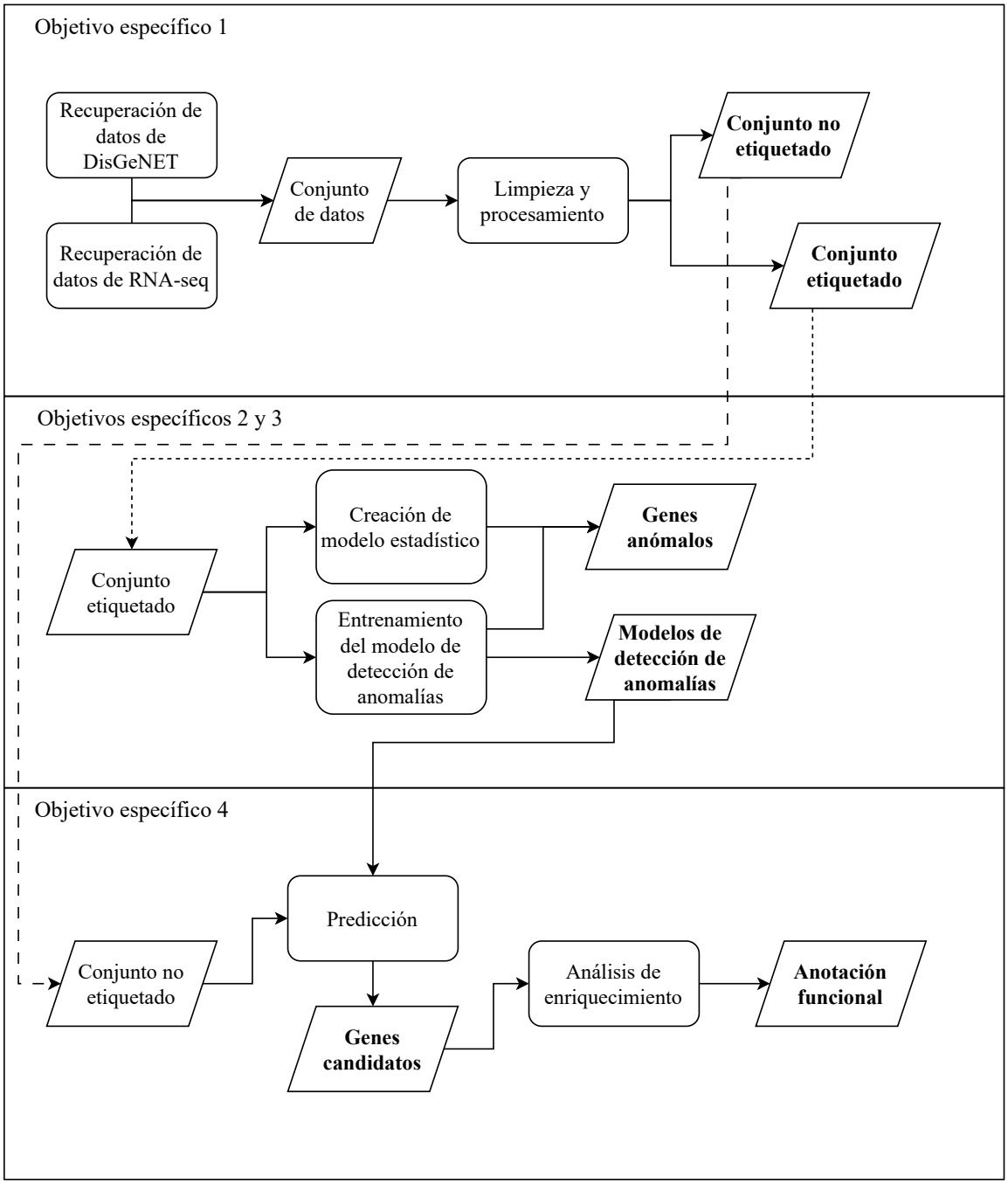

Figura 4.2: Flujo de trabajo del enfoque de detección de anomalías [61].

Nuevamente, las actividades para el objetivo específico 1 se enfocan en la creación

de un conjunto de datos. Sin embargo, en esta ocasión el conjunto únicamente contiene los genes asociados a la β-talasemia y su nivel de expresión génica. De igual manera, en esta etapa se realiza el procesamiento del conjunto de datos.

Las actividades para los objetivos específicos 2 y 3 se enfocan en la construcción y validación de un modelo estadístico y de un modelo de detección de anomalías. Como resultado de la creación de ambos modelos, se obtuvo un subconjunto de genes potencialmente asociados a la talasemia los cuales presentan un nivel de expresión génica anómalo.

Finalmente, las actividades para el objetivo específico 4 se asocian a la predicción de genes candidatos y al análisis de enriquecimiento. Mediante al análisis de enriquecimiento, es posible estudiar si los procesos biológicos en los que participan estos genes tienen alguna relación latente con la talasemia.

4.3. Herramientas

Para llevar a cabo esta investigación se utilizaron tanto las plataformas R [63], en su versión 4.1.2, como Python [64], en su versión 3.10. En la tabla 4.1 se presentan los distintos paquetes utilizados de acuerdo a su funcionalidad y a la plataforma.

Tabla 4.1: Herramientas de software utilizadas.

Funcionalidad	**Paquete**	**Plataforma**
Manipulación de datos y ciencia de datos	`dplyr` 1.1.4 [65] `tidyverse` 2.0.0 [66]	R
	`Pandas` 2.0.2 [67]	Python
Visualizaciones y gráficos	`ggplot2` 3.5.1 [68] `viridis` 0.6.5 [69]	R
	`Matplotlib` 3.8.0 [70] `Seaborn` 0.12.2 [71]	Python
Herramientas de aprendizaje automático	`scikit-learn` 1.2.2 [72]	Python
Herramientas de bioinformática	`BiocManager` 1.30.23 [73]	R
	`GSEApy` 1.0.6 [74]	Python
Funciones matemáticas	`NumPy` 1.24.3 [75]	Python

Capítulo V. Resultados

En este capítulo se detallan los resultados obtenidos en cada etapa distinta de esta investigación y de acuerdo a cada estrategia empleada.

5.1. Enfoque de Clasificación

Inicialmente, se presentarán en esta sección los resultados obtenidos mediante el método de clasificación.

5.1.1. Análisis Exploratorio de Datos

Para el cumplimiento del primer objetivo específico, caracterizar el conjunto de datos, se utilizó la estadística descriptiva para contrastar el comportamiento de los genes asociados a la talasemia contra aquellos asociados a la vasculitis. La distribución para ambos grupos se puede observar en la Figura 5.1. Es posible apreciar que ambos grupos tienen una distribución muy similar, con la particularidad de que los genes asociados a la talasemia tienen una presencia mayor de valores extremos. Sin embargo, es importante recalcar que este tipo de observaciones no son concluyentes. Es por ello por lo que se buscó un patrón común entre los genes asociados a la talasemia que puede no ser fácilmente identificable, mediante el uso de métodos de aprendizaje automático.

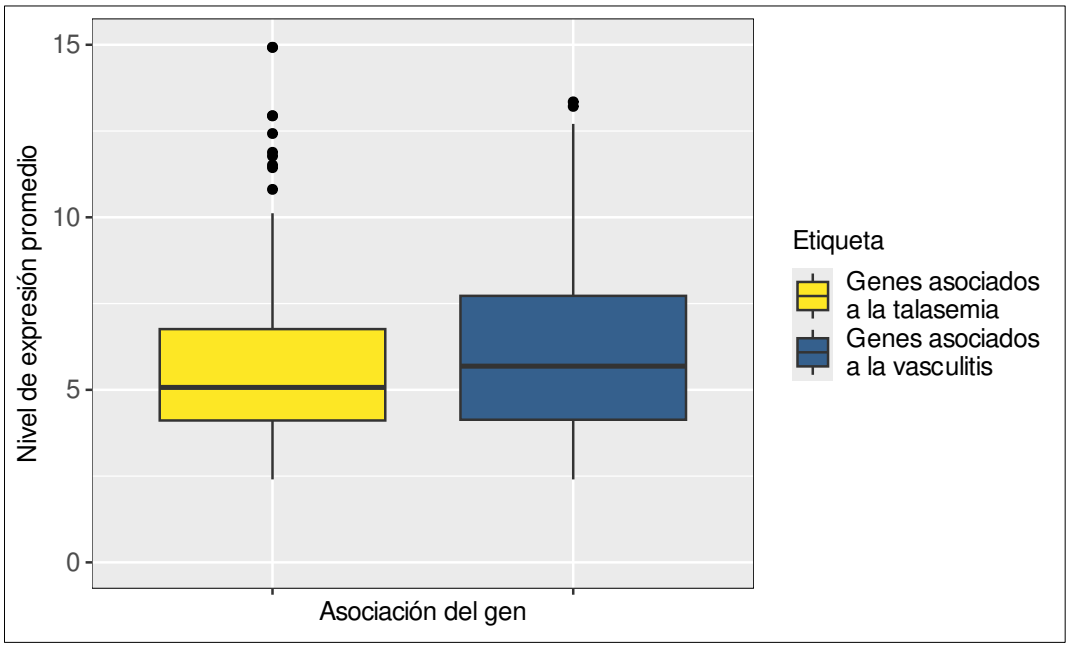

Figura 5.1: Diagrama de cajas del nivel de genética promedio de los genes de acuerdo a su asociación con la talasemia.

5.1.2. Mejores Conjuntos de Hiperparámetros Encontrados

Una vez procesado y etiquetado el conjunto de datos, este fue utilizado para llevar a cabo el entrenamiento de los diversos modelos de aprendizaje automático, tarea correspondiente al segundo objetivo específico. El proceso de optimización de hiperparámetros se llevó a cabo de manera integrada al entrenamiento. De esta manera, para cada modelo se probaron distintas combinaciones de valores para sus hiperparámetros y finalmente se escogió la combinación que brindó el mejor resultado, para la métrica F_1. De esta manera culminó el proceso de construcción de los modelos de clasificación. Para todos los modelos en los que se utilizó la búsqueda aleatoria, la cantidad de iteraciones, es decir, de combinaciones distintas de valores, fue de 200.

En la Tabla 5.1 se presentan los hiperparámetros seleccionados para cada distinto algoritmo de clasificación utilizado. En el caso de regresión logística, bosque aleatorio, SVM y XGBoost, estos hiperparámetros se encontraron mediante un método de optimización, ya fuera búsqueda en rejilla o búsqueda aleatoria. En el caso de TPOT, este es intrínsecamente un algoritmo de optimización, por lo que no es necesario agregar otro método.

Tabla 5.1: Configuración de hiperparámetros para los distintos modelos construidos.

Modelo	Hiperparámetros	Método de optimización
Regresión Logística	Regularización: L1 C: 23,60 Optimización: `liblinear`	Búsqueda aleatoria
Bosque aleatorio	Criterio: Gini Características máximas: 9 Características mín. por nodo: 3 Número de árboles: 100	Búsqueda aleatoria
SVM	C: 3,1446 Kernel: polinomial Grado: 3 Gamma: 0,1210	Búsqueda aleatoria
Votación en conjunto	Votación: suave	Ninguno[1]
XGBoost	Características máximas: 40,82 % Máxima profundidad: 9 Número de árboles: 100	Búsqueda aleatoria
TPOT	Generaciones: 100 Tamaño de población: 100 Tasa de mutación: 85 % Tasa de cruce: 15 % Tiempo de evaluación por modelo: 15 minutos Parada temprana: 25 generaciones	TPOT[2]
TPOT + Torch	Generaciones: 100 Tamaño de población: 100 Tasa de mutación: 85 % Tasa de cruce: 15 % Tiempo de evaluación por modelo: 20 minutos Parada temprana: 20 generaciones	TPOT[2]

[1] Los modelos utilizados para la votación en conjunto fueron optimizados mediante búsqueda aleatoria.

[2] Los hiperparámetros de TPOT fueron definidos manualmente, sin embargo, TPOT se encarga de optimizar el *pipeline*.

5.1.3. Métricas de Desempeño de los Modelos de Clasificación

Con el objetivo de evaluar los modelos de clasificación construidos (objetivo específico 3), se calcularon las métricas de desempeño utilizando la mejor combinación de hiperparámetros encontrada. Nuevamente, se aplicó la técnica de validación cruzada con 3 particiones y 4 repeticiones, sin embargo, se cambió el estado aleatorio para que las particiones empleadas en esta ocasión fueran distintas. La Tabla 5.2 muestra las distintas métricas de desempeño alcanzadas por cada distinto modelo creado durante esta evaluación.

Tabla 5.2: Métricas de desempeño obtenidas para los distintos modelos de aprendizaje automático.

Modelo	Exactitud	Precisión	Sensibilidad	F1	AUC-ROC
Regresión Logística	0,619	0,508	0,514	0,510	0,615
Bosque aleatorio	0,627	0,530	0,391	0,445	0,595
SVM	0,616	0,505	0,541	0,521	0,620
Votación en conjunto	0,628	0,526	0,377	0,438	0,637
XGBoost	0,610	0,499	0,399	0,440	0,594
TPOT	0,631	0,525	0,666	0,582	0,679
TPOT + Torch	0,631	0,525	0,666	0,582	0,679

Como se puede observar, ningún modelo alcanzó valores del 70 % para ninguna de las distintas métricas de desempeño, incluso después del ajuste de hiperparámetros. Debido a esto, no fue posible cumplir el objetivo de identificar un subconjunto de potenciales biomarcadores asociados a la talasemia (objetivo específico 4) mediante el enfoque de clasificación. Esto se debe a lo poco confiables que podrían ser los resultados. De esta manera, se decidió continuar con un enfoque de detección de anomalías, ya que, el de clasificación no resultó una estrategia adecuada para abordar este problema.

5.2. Enfoque de Detección de Anomalías

En esta sección se presentan los resultados obtenidos mediante el método de detección de anomalías.

5.2.1. Estadística Descriptiva

Nuevamente se utilizó estadística descriptiva para visualizar la diferencia en los valores de expresión génica y caracterizar de esta manera el conjunto de datos, en concordancia con el primer objetivo específico. Fueron utilizadas medidas estadísticas para comparar el comportamiento entre las dos condiciones distintas, como se muestra en la Tabla 5.3 [61]. Es posible observar valores similares para los cuartiles, sin embargo, hay una diferencia considerable en el valor medio de la expresión génica, potencialmente influenciada por los valores extremos [62]. Además, hay una variabilidad mucho mayor para la condición afectada.

Tabla 5.3: Estadística descriptiva para los valores de expresión genética de acuerdo a la condición.

Medida	**Afectado**	**Control**
Q1	0.10	0.08
Mediana	4.24	2.95
Q3	24.25	20.01
Media	11574.49	7202.22
Desv. Estándar	103885.6	63386.42

Posteriormente, se realizó un histograma de la diferencia en el valor de expresión génica (afectado - control) para el mismo gen [61, Fig. 5.2]. Se puede observar cómo, para la gran mayoría de los casos, las diferencias se acumulan en torno a 0. Esto indica que, para la mayoría de los genes, el valor de expresión génica no varía significativamente entre condiciones [61]. Es importante mencionar que, para mejorar la interpretabilidad del gráfico, se omitieron 18 valores que se encontraban fuera de los límites.

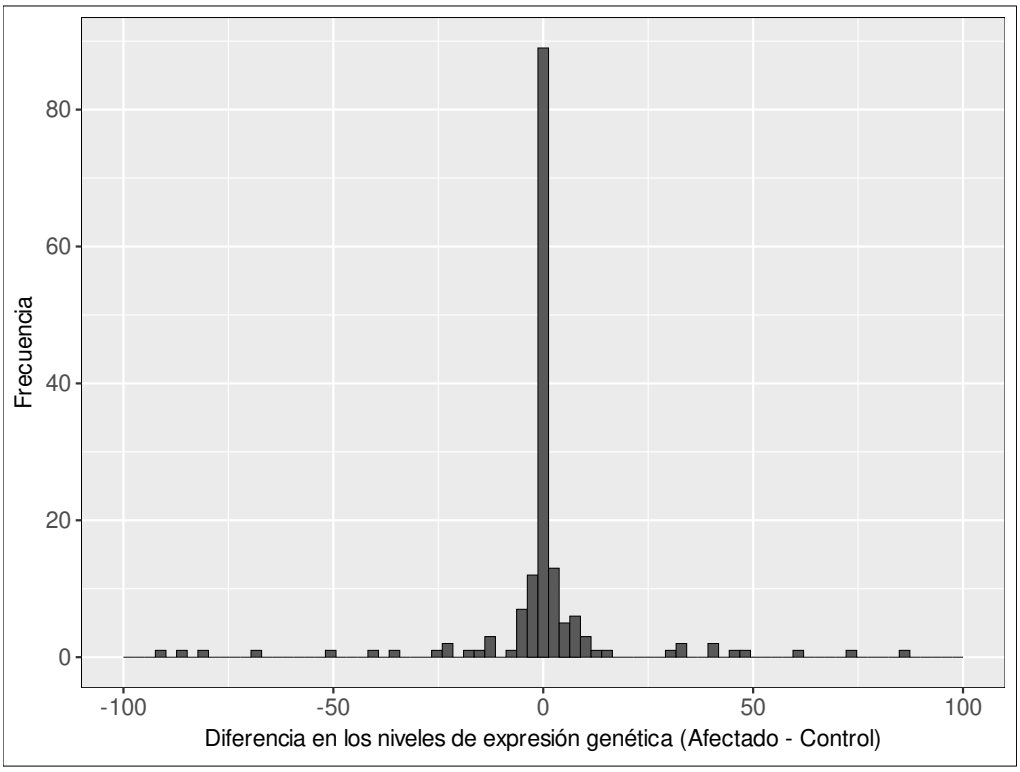

Figura 5.2: Histograma de la diferencia en los valores de expresión genética (afectado - control) [61].

5.2.2. Análisis Estadístico y Descriptivo

Se generó un mapa de calor para identificar el grupo de genes cuyo valor tiene una diferencia considerable entre condiciones, como se muestra en [61, Fig. 5.3]. En esta figura no se incluyen todos los genes candidatos, ya que la propósito es mostrar que existen genes cuyo valor de expresión génica difiere considerablemente entre condiciones.

Una vez construidos ambos modelos (objetivo específico 2), se obtuvo quince genes definiendo el umbral en el percentil 95 y tomando las instancias con mayor diferencia entre las dos condiciones. Posteriormente, se ejecutó el algoritmo de bosque de aislamiento para construir un modelo de detección de anomalías para realizar la detección de anomalías. Este modelo fue capaz de detectar nueve genes anómalos. Para obtener una evaluación del modelo (objetivo específico 3), se analizó si existía una intersección entre los genes detectados por ambos métodos. El resultado fue que todos los genes identificados por el bosque de aislamiento fueron también identificados por el método estadístico [61].

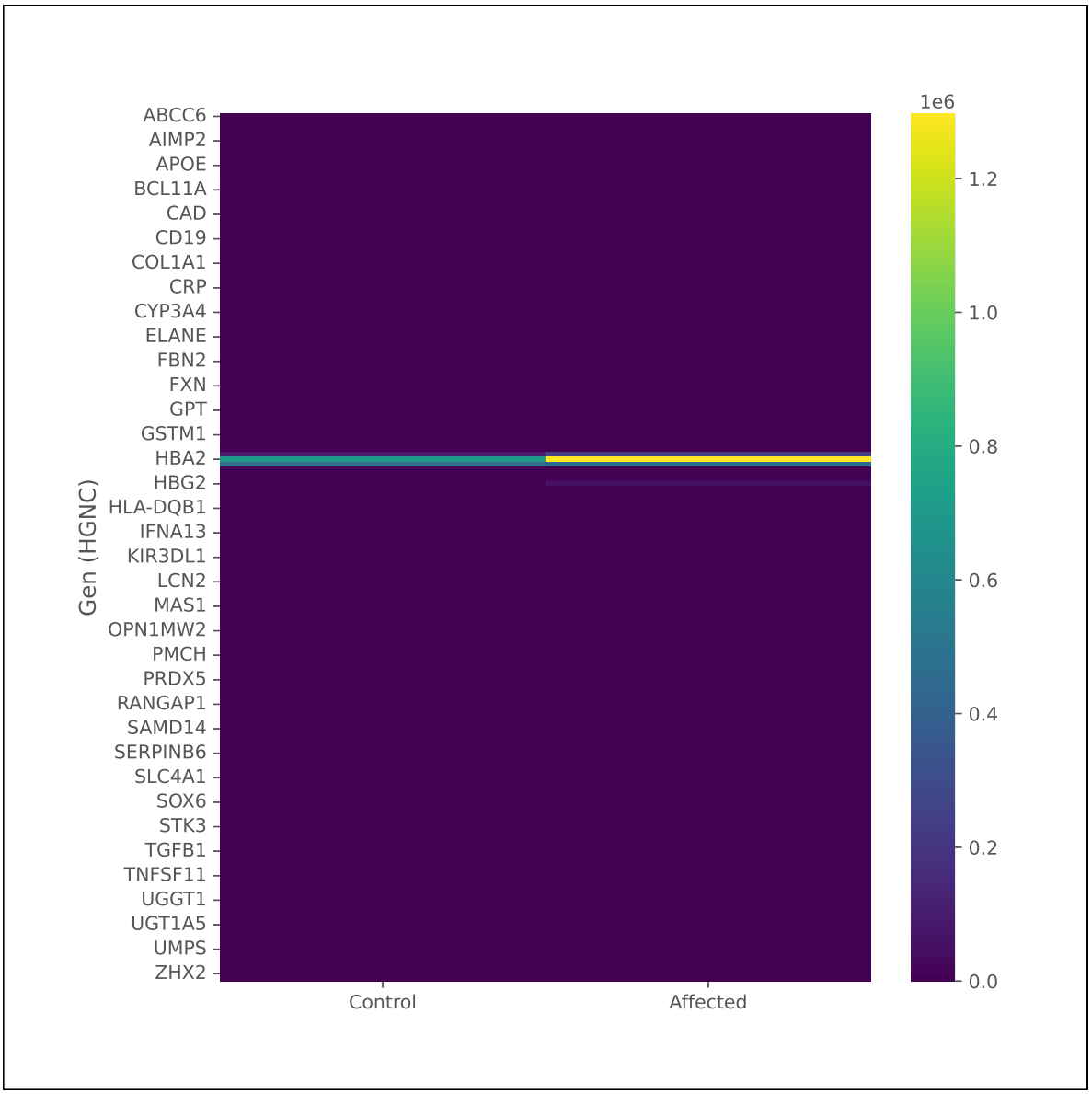

Figura 5.3: Mapa de calor de los valores de expresión genética (control vs. afectado) [61].

La Tabla 5.4 contiene la lista de los quince genes y su descripción. Entre ellos se encuentran las subunidades de hemoglobina (HBA1, HBA2, HBB, HBD, HBG1, HBG2) [61].

Tabla 5.4: Descripción de los genes en el conjunto de entrenamiento con niveles de expresión identificados como anómalos [61].

Gen (HGNC)	**Descripción**	**Método de identificación**
AHSP	La AHSP actúa como chaperona para estabilidad la alfa-globina y evitar su precipitación cuando la beta-globina está ausente o escasea [76].	Estadístico Bosque de aislamiento
B2M	La B2M es un componente de las moléculas del complejo mayor de histocompatibilidad (CMH) de clase I, que intervienen en la respuesta inmunitaria [77].	Estadístico Bosque de aislamiento
CA2	La C2A es una enzima que cataliza la hidratación reversible del dióxido de carbono, crucial para diversos procesos fisiológicos [78].	Estadístico
CAP1	CAP1 interviene en la organización del citoesqueleto de actina y la motilidad celular [79].	Estadístico
CD37	CD37 es una proteína de la superficie celular implicada en la transducción de señales y la respuesta inmunitaria [80].	Estadístico
HBA1, HBA2, HBB, HBD, HBG1, HBG2 (Subunidades de hemoglobina)	Estos genes codifican diferentes subunidades de la hemoglobina, una proteína responsable del transporte de oxígeno en las células sanguíneas [81].	Estadístico Bosque de aislamiento
PRDX1	PRDX1 es una enzima antioxidante que desempeña un papel en la protección celular contra el estrés oxidativo [82].	Estadístico
RPS19	RPS19 es un componente del ribosoma y participa en la síntesis de proteínas [83].	Estadístico
SLC4A1	SLC4A1 participa en el transporte de bicarbonato a través de las membranas celulares. Mantiene el equilibrio ácido-base en los glóbulos rojos [84].	Estadístico
TGFB1	TGFB1 es una citocina que interviene en el crecimiento celular, la diferenciación, la regulación de la respuesta inmunitaria y la apoptosis [85].	Estadístico

5.2.3. Análisis de Enriquecimiento

Una vez aplicado el modelo construido a los datos no etiquetados (genes que ningún estudio publicado ha asociado con la talasemia), se obtuvo 72 genes con comportamiento anómalo. Posteriormente, con el fin de obtener un conjunto de potenciales genes biomarcadores (objetivo específico 4), se realizó un análisis de enriquecimiento con estos genes y se extrajeron los diez términos con mayor significancia, tal y como se muestra en la Figura 5.4 [61].

El proceso con mayor significancia y, simultáneamente, mayor número de unidades corresponde a la traducción citoplasmática [61].

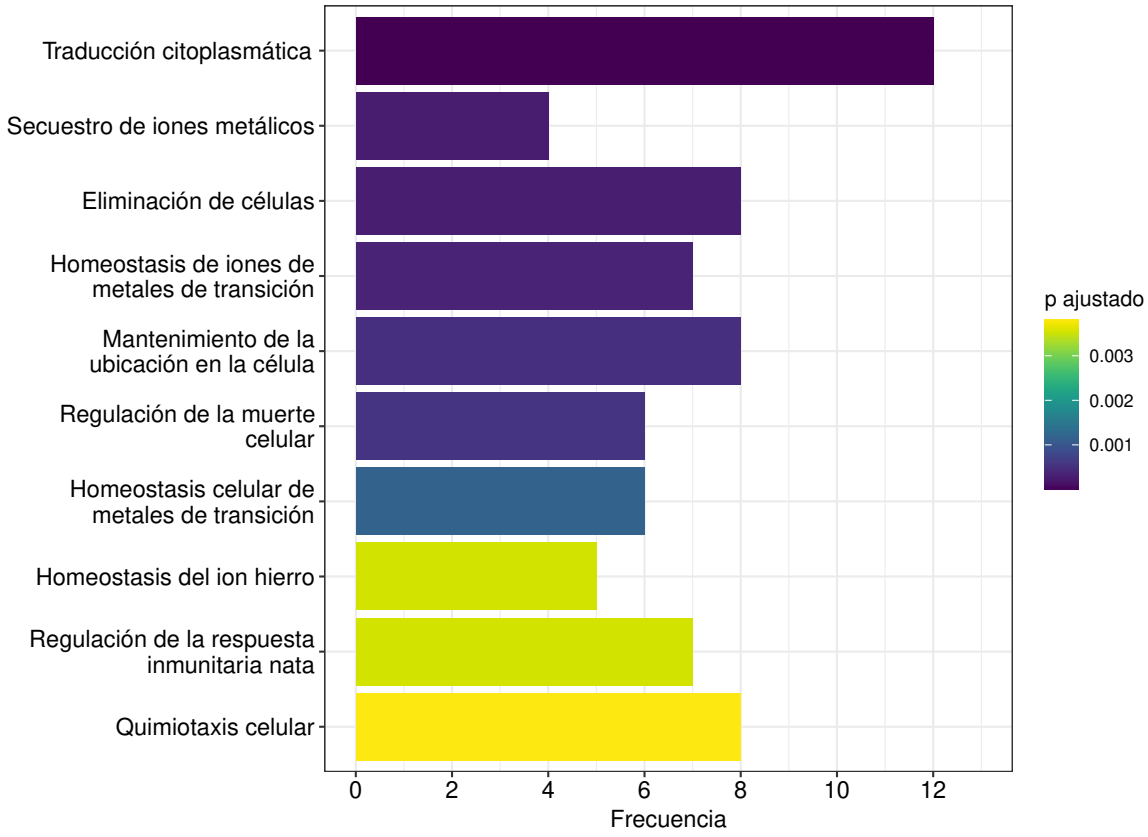

Figura 5.4: Gráfico de barras del análisis de enriquecimiento para los genes candidatos.

5.2.4. Potenciales Genes Biomarcadores

La Tabla 5.5 contiene los genes identificados como anómalos mediante el modelo basado en bosque de aislamiento. También se asociaron estos genes al proceso biológico en el que están involucrados.

Tabla 5.5: Genes asociados con los principales procesos biológicos identificados.

Términos ontológicos	**Cantidad de genes**	**Valor p ajustado**	**Genes**
Traducción citoplasmática	12	$5{,}56 \times 10^{-10}$	RPL19, CORO1A, CSF3R, HSPB1, S100A8, SPI1, TMSB4X
Secuestro de iones metálicos	4	$2{,}85 \times 10^{-4}$	FTH1, FTL, TRGJP2, S100A9
Eliminación de células	8	$3{,}15 \times 10^{-4}$	BCL2L1, CORO1A, HLA-A, HLA-B HSP90AB1, LYZ, SPI1, TYROBP
Homeostasis de iones de metales de transición	7	$3{,}89 \times 10^{-4}$	ALAS2, FTH1, FTL, NCOA4, S100A8, SLC25A37
Mantenimiento de la ubicación en la célula	8	$5{,}01 \times 10^{-4}$	CORO1A, CYBA, FTH1, FTL, S100A8, S100A9, SRGN, TMSB4X
Regulación de la apoptosis	6	$5{,}80 \times 10^{-4}$	BCL2L1, HLA-A, HLA-B, HSP90B1, SPI1, TYROBP
Homeostasis celular de metales de transición	6	$1{,}19 \times 10^{-3}$	ALAS2, FTH1, FTL, NCOA4, S100A8, S100A9
Homeostasis del ion hierro	5	$3{,}54 \times 10^{-3}$	ALAS2, FTH1, FTL, NCOA4, SLC25A37
Regulación de la respuesta inmunitaria innata	7	$3{,}54 \times 10^{-3}$	HLA-A, HLA-B, HSP90A1, HSP90AB1, MNDA, SPI1, TYROBP
Quimiotaxis celular	8	$3{,}80 \times 10^{-3}$	BSG, CORO1A, CSF3R, HSPB1, S100A8, S100A9, SPI1, TMSB4X

Capítulo VI. Discusión

En este capítulo se discute el porqué de los resultados obtenidos de acuerdo a cada enfoque empleado en la investigación.

6.1. Enfoque de Clasificación

Es posible analizar el bajo desempeño de los modelos de clasificación desde diversas aristas. En este caso, es posible identificar potenciales causas desde la conformación del conjunto de datos hasta la naturaleza genética de la enfermedad.

Partiendo desde el conjunto de datos, se mencionó que los valores de expresión genética fueron obtenidos mediante la tecnología de microarreglos. Los datos de microarreglos suelen contener un gran número de genes, pero muchos de ellos pueden no ser relevantes para la enfermedad de interés. Las características sin importancia o ruidosas pueden ocultar la verdadera señal de los genes asociados a la talasemia, especialmente en conjuntos de datos de alta dimensión [86]. Una forma de lidiar con esto es mediante métodos de selección de características. Sin embargo, estos métodos no fueron directamente implementados ya que, tanto el algoritmo de bosque aleatorio [34] como el de XGBoost [35], implementan estás técnicas intrínsecamente y los modelos no presentaron un desempeño sobresaliente en comparación a los demás.

Es posible también que, la falta de integración de conocimientos biológicos puede estar obstaculizando el rendimiento de los modelos construidos. Los modelos de aprendizaje automático sin conocimientos específicos del ámbito podrían no captar algunas interacciones biológicas clave, las vías reguladoras o las redes de coexpresión génica. La incorporación de datos específicos del ámbito o de conocimientos biológicos previos podría conducir a una mejor ingeniería de las características o a la elaboración de modelos a priori informados, lo que mejoraría la precisión de la predicción [87]. Sin embargo, implementar modelos que tengan esta capacidad entra en conflicto con el ob-

jetivo general de descubrir genes novedosos asociados a la enfermedad. Esto se debe a que verificar si existe un camino que asocie el gen con procesos biológicos relacionados a la enfermedad es la manera en la que se propone validar los hallazgos. Incorporar esta información a un modelo, podría inducir un sesgo en las predicciones que actualmente se basan únicamente en la expresión genética.

Es importante destacar que el método de validación puede también haber tenido un impacto sobre las métricas de desempeño. Escoger $k = 3$ conlleva a que el conjunto de entrenamiento sea más pequeño que al utilizar valores como $k = 5$ o $k = 10$. Esto implica que el modelo recibe una menor exposición a todo el conocimiento disponible, lo que puede afectar su capacidad de predicción [57]. Por el otro lado, al aumentar este valor, puede aumentar también la variabilidad de las métricas de desempeño, ya que hay cantidad limitada de instancias sobre las que realizar la predicción y el valor promedio de las métricas se puede ver influenciado fácilmente por clasificaciones erróneas [62]

Finalmente, hay características intrínsecas de la enfermedad que pueden afectar el rendimiento de los modelos de aprendizaje automático. Los genes con mutaciones deletéreas pueden estar regulados a la baja o mostrar patrones de expresión aberrantes, lo que dificulta que los modelos convencionales de aprendizaje automático los distingan del ruido [88]. Particularmente, estas mutaciones podrían alterar la transcripción normal o el procesamiento del ARN, lo que complicaría la detección de diferencias de expresión significativas. Asimismo, en algunos casos, las mutaciones en un gen pueden desencadenar un aumento compensatorio de otros genes que desempeñan funciones similares o coincidentes [89]. Esta redundancia genética puede disminuir la brecha entre la asociación entre la expresión de un gen y su expresión observable. Además, las mutaciones deletéreas, especialmente en enfermedades graves como la β-talasemia, suelen inducir vías de estrés celular (por ejemplo, estrés oxidativo [90]). La activación de estas respuestas generales al estrés puede ocultar firmas específicas de expresión génica directamente relacionadas con la mutación [91].

Una posible dirección futura para mejorar el rendimiento de la clasificación de genes es reformular el problema utilizando un marco de aprendizaje positivo-sin etiqueta (*positive-unlabeled* o PU), en el que sólo los genes que se sabe que están asociados con

la beta talasemia se tratan como positivos, y todos los demás genes permanecen sin etiquetar [92], [93]. Este enfoque evita depender de etiquetas negativas potencialmente engañosas y refleja mejor la naturaleza incompleta de las anotaciones actuales de genes de enfermedades. Para mejorar el rendimiento predictivo, las características basadas en la expresión pueden complementarse con características derivadas de redes biológicas, como métricas de interacción proteína-proteína y anotaciones de vías u ontologías. Esta estrategia híbrida integraría datos transcriptómicos con el contexto funcional y topológico de los genes, en consonancia con pruebas recientes de que las características biológicas de múltiples fuentes mejoran la priorización de los genes [94]. El empleo de algoritmos de aprendizaje de PU [95] en este espacio de características enriquecido podría conducir a una identificación más sólida de genes candidatos asociados a la β-talasemia.

6.2. Enfoque de Detección de Anomalías

La diferencia presente en el valor medio de la expresión génica entre ambas condiciones (Tabla 5.3) indica una alta influencia de los valores extremos [62]. En este caso, se puede concluir que algunos genes del paciente afectado tienen niveles mucho más altos de expresión genética.

Entre los genes previamente asociados a la enfermedad que fueron identificados como anómalos, se encontraron las subunidades de hemoglobina. Esto es un resultado esperable debido a que, como se mencionó anteriormente, la talasemia se produce debido a la producción anormal de hemoglobina [3]. El hecho de que estos genes fueran identificados como anómalos indica que, en el paciente afectado, el nivel de expresión genética de estos genes se encuentra fuera de la norma, lo cual apoya la validez del modelo.

También es relevante la presencia del gen AHSP. Este gen está implicado en la estabilización de la alfa-globina, un componente de la hemoglobina. La proteína estabilizadora de la hemoglobina alfa (AHSP) actúa como una chaperona para estabilizar la alfa-globina y evitar su precipitación cuando la beta-globina está ausente o escasea [96], [97]. De nuevo, este proceso está altamente correlacionado con una producción anormal

de hemoglobina.

Varios de los términos ontológicos (procesos biológicos) encontrados (Figura 5.4) para los genes candidatos son relevantes en el contexto de la talasemia, como lo es el caso de la traducción citoplasmática. Los errores en la traducción citoplasmática, que es el proceso de síntesis de proteínas en el citoplasma de una célula, pueden tener diversas consecuencias, de leves a graves. La fidelidad de la traducción es crucial para mantener la estructura y función adecuadas de las proteínas. Pueden producirse errores en diferentes etapas de este proceso y dar lugar a la producción de proteínas anormales o no funcionales. Los errores graves de traducción pueden activar la apoptosis, un mecanismo de muerte celular programada, otro término obtenido en el análisis. La apoptosis es un mecanismo de protección para eliminar células con daños irreparables o disfunciones [98]. Dado que la síntesis de las cadenas de globina (componente de la hemoglobina) se produce en el citoplasma, las alteraciones en este proceso podrían ser relevantes en la talasemia. Además, los estudios apoyan que las mutaciones en la beta-globina (HBB) pueden afectar a la eficiencia de la transcripción en el citoplasma [99].

Otro término importante es la homeostasis de los iones de metales de transición. El hierro es un componente crucial de la hemoglobina, y las alteraciones en la homeostasis del hierro pueden afectar a la síntesis de hemoglobina y contribuir a la talasemia [100]. Asimismo, ya se ha mencionado que la sobrecarga de hierro es un síntoma común de la talasemia [2]. Además, existen proteínas encargadas de exportar el hierro al plasma para llevarlo al resto del organismo. Cuando son defectuosas, inducen la retención celular de hierro en tipos celulares específicos [101]. El hierro también desempeña un papel crucial en el transporte de oxígeno en el cuerpo humano, principalmente a través de su asociación con la hemoglobina. El grupo hemo de la hemoglobina, que contiene hierro, es esencial para su función de transporte de oxígeno [102].

Las mutaciones en los genes responsables de la homeostasis del hierro pueden producir especies reactivas del oxígeno (ROS) [100]. Las ROS son moléculas químicamente reactivas que contienen oxígeno [90]. La alteración de la homeostasis del hierro por mutaciones en genes implicados en la regulación del hierro (como los genes que codifican la ferritina, la transferrina o las proteínas transportadoras de hierro) puede producir

una acumulación excesiva de hierro. Este exceso de hierro puede catalizar la producción de ROS [100]. Estas especies altamente reactivas pueden causar daño oxidativo a los componentes celulares, incluidos los lípidos, las proteínas y el ADN. La sobreproducción de ROS puede provocar estrés oxidativo, una situación en la que el equilibrio entre antioxidantes y ROS se ve afectado. El estrés oxidativo prolongado puede contribuir al daño celular y está implicado en varias enfermedades, como los trastornos neurodegenerativos, las enfermedades cardiovasculares y el cáncer.

La producción de especies reactivas del oxígeno (ROS) también puede conducir a la apoptosis. El estrés oxidativo puede activar vías de señalización específicas que regulan la apoptosis. Uno de los actores clave en este proceso es la mitocondria. Las ROS pueden dañar directamente las mitocondrias, provocando su disfunción. Esta disfunción puede provocar la liberación de moléculas proapoptóticas de las mitocondrias al citoplasma. La regulación de la apoptosis es crucial para mantener la homeostasis celular normal y eliminar las células dañadas o innecesarias. Sin embargo, la desregulación de la apoptosis ya sea debida a una producción excesiva de ROS o a otros factores, puede contribuir a diversas enfermedades [90].

Estos resultados permiten establecer una posible vía común para diferentes procesos biológicos entre los genes candidatos. La talasemia se debe a una producción inadecuada de hemoglobina. La hemoglobina, junto con el hierro, es esencial en el transporte de oxígeno. Un síntoma común de la talasemia es la sobrecarga de hierro. Como indican los estudios, la sobrecarga de hierro puede conducir a la formación de ROS. Estas pueden causar daños oxidativos en las células y, finalmente, provocar la muerte celular. Por lo tanto, podemos relacionar los procesos biológicos obtenidos mediante el análisis de enriquecimiento con procesos asociados a la hemoglobina, como el transporte de oxígeno.

Se compararon los resultados obtenidos en el análisis de enriquecimiento con los obtenidos por Taghavifar, Hamid y Shariati [48]. Se encontró un término común, que corresponde a la homeostasis de iones de hierro, y un término similar, regulación de la apoptosis. Se realizó una consulta a AmiGO [103] y se encontró que cuatro de los cinco genes (ALAS2, FTH1, FTL, NCOA4) relacionados con este proceso biológico

estaban presentes en esta aplicación. El único gen que no estaba presente corresponde a SLC25A37. Este resultado es inesperado porque este gen está estrechamente relacionado con el transporte de hierro en las mitocondrias.

También se toma en cuenta el proceso de regulación de la apoptosis. Aunque Taghavifar, Hamid y Shariati [48] no reportaron este término, encontraron la regulación de la apoptosis, que es similar. Se volvió a realizar la búsqueda AmiGO [103] y se obtuvo el gen BCL2L1. Este gen podría ser relevante ya que está asociado a la expresión de gamma-globina y a procesos anti-apoptóticos [104].

Los hallazgos de este estudio refuerzan el conocimiento actual sobre la fisiopatología de la talasemia al identificar genes y procesos biológicos ya bien establecidos, como la desregulación de la hemoglobina y la homeostasis del hierro. Sin embargo, también aportan evidencia adicional al sugerir que ciertos genes implicados en la apoptosis y el estrés oxidativo, como BCL2L1 y NCOA4, podrían desempeñar un papel más central en la progresión de la enfermedad de lo que se ha reportado previamente. En especial, la identificación de procesos como la traducción citoplasmática y la regulación mitocondrial sugiere nuevas interacciones funcionales relevantes para el fenotipo talasémico, lo cual amplía las fronteras del conocimiento actual sobre los mecanismos moleculares de la enfermedad.

Desde la perspectiva del aprendizaje automático, este estudio demuestra que los modelos de detección de anomalías pueden ser útiles para revelar patrones atípicos de expresión génica asociados a enfermedades hereditarias como la talasemia, incluso cuando la señal biológica está parcialmente oculta por la variabilidad interindividual. Esto contrasta con enfoques supervisados tradicionales que requieren etiquetas confiables, las cuales pueden no estar disponibles en el contexto de enfermedades raras o poco caracterizadas. La utilidad del análisis no supervisado para redescubrir genes clave (como HBB, AHSP y ALAS2) y para proponer nuevos candidatos funcionales destaca el valor de estos métodos para la priorización de biomarcadores en estudios transcriptómicos [26].

No obstante, nuestros hallazgos también ponen en evidencia las limitaciones actuales de las bases de datos funcionales. Por ejemplo, la ausencia de SLC25A37 en AmiGO,

a pesar de su conocida implicación en el transporte mitocondrial de hierro [105], revela brechas en la anotación ontológica y destaca la necesidad de una actualización constante de estos recursos para mejorar la interpretación de resultados generados mediante enfoques computacionales.

Capítulo VII. Conclusiones

En este trabajo se llevó a cabo una investigación orientada a identificar nuevos biomarcadores asociados a la enfermedad de la talasemia mediante el análisis de perfiles de expresión genética y el empleo de modelos predictivos y estadísticos. Se implementaron dos enfoques principales: uno basado en clasificación y otro en detección de anomalías.

En cuanto al enfoque de clasificación, se construyó un conjunto de datos a partir de perfiles de expresión genética obtenidos mediante microarreglos. Además, aunque se desarrollaron múltiples modelos, los valores obtenidos en las métricas de desempeño fueron bajos, lo que limitó la viabilidad de este enfoque en el contexto del estudio.

Para el enfoque basado en detección de anomalías se generó un conjunto de datos a partir de perfiles de expresión genética obtenidos mediante RNA-seq. Este permitió construir un modelo de detección de anomalías basado en bosques de aislamiento. Este modelo identificó un subconjunto de 72 genes candidatos, obteniendo una alta coincidencia al compararlo con un análisis estadístico de los genes con mayor diferencia en nivel de expresión. Los genes candidatos se asociaron a procesos biológicos relevantes en el contexto de la talasemia, como traducción citoplasmática, apoptosis y homeostasis de metales de transición, señalando una posible relación de estos con la talasemia.

La implicación de la apoptosis sugiere que la muerte celular programada puede desempeñar un papel hasta ahora subestimado en la patogénesis de la talasemia, contribuyendo potencialmente a la destrucción de precursores eritroides (células precursoras de los glóbulos rojos) o glóbulos rojos defectuosos y exacerbando así la anemia. Este hallazgo subraya la necesidad de seguir explorando las vías apoptóticas como posibles dianas terapéuticas.

Del mismo modo, la identificación de genes relacionados con la homeostasis del hierro amplía el conocimiento del metabolismo desregulado del hierro en la talasemia, una característica distintiva de la enfermedad. Aunque la sobrecarga de hierro está bien

documentada en pacientes con talasemia, el descubrimiento de nuevos genes implicados en este proceso ofrece nuevas perspectivas sobre los mecanismos que conducen a una absorción y almacenamiento anormales del hierro. En conjunto, estos hallazgos no sólo mejoran el conocimiento actual de las bases moleculares de la talasemia, sino que también abren nuevas vías de investigación dirigidas a desarrollar terapias específicas y mejorar las estrategias de tratamiento de la enfermedad.

Desde el marco teórico, se abordó la talasemia como una enfermedad genética cuya complejidad y variabilidad clínica requieren enfoques diagnósticos más precisos y eficientes. En ese contexto, el uso de técnicas de aprendizaje automático, especialmente los modelos de detección de anomalías, permitió identificar patrones sutiles en perfiles de expresión genética que no habrían sido evidentes mediante métodos moleculares convencionales. Esta integración de conceptos permitió traducir datos en hipótesis biológicamente relevantes, generando una lista de genes candidatos con potencial valor diagnóstico y terapéutico. Así, el estudio demuestra cómo las herramientas computacionales pueden complementar y potenciar la investigación biomédica, proporcionando nuevas rutas para abordar problemas clínicos complejos como los que plantea la talasemia.

En particular, el uso de perfiles de expresión genética como fuente primaria de información permitió aprovechar uno de los niveles más dinámicos de la regulación biológica para inferir la posible implicación de genes en la fisiopatología de la talasemia. Tal como se discutió en el marco teórico, las tecnologías como los microarreglos y el RNA-seq proporcionan mediciones cuantitativas del nivel de expresión de miles de genes de forma simultánea, lo que habilita el análisis de patrones globales asociados a estados patológicos. Si bien los modelos de clasificación supervisada fueron inicialmente considerados por su capacidad para aprender reglas a partir de ejemplos etiquetados, su efectividad se vio limitada por el desequilibrio de clases y la escasez de etiquetas confiables. En contraste, la detección de anomalías ofreció una alternativa no supervisada que permitió identificar genes con comportamientos atípicos, potencialmente asociados a funciones alteradas en la enfermedad. Este enfoque, al centrarse en desviaciones en el perfil transcriptómico, facilitó la identificación de candidatos que podrían actuar como

biomarcadores, es decir, genes cuya expresión diferencial podría estar reflejando alteraciones funcionales, incluyendo aquellas relacionadas con la síntesis de proteínas, el estrés oxidativo o la homeostasis del hierro, todos procesos estrechamente vinculados a la talasemia. De esta manera, los resultados del estudio no solo están alineados con los fundamentos teóricos revisados, sino que demuestran cómo su aplicación práctica puede traducirse en hallazgos biológicamente significativos.

Aunque gran parte de los resultados obtenidos en este trabajo coinciden con hallazgos previamente descritos en la literatura, lo cual respalda la solidez biológica del enfoque utilizado, este estudio también aporta elementos novedosos. En particular, la identificación de genes no tradicionalmente asociados a la talasemia, pero implicados en procesos relevantes como la traducción citoplasmática, el estrés oxidativo o la regulación mitocondrial del hierro, sugiere posibles nuevas rutas patológicas o genes modificadores de la enfermedad. Además, la aplicación de modelos de aprendizaje automático no supervisado para detectar estas señales en datos de expresión génica refuerza el valor de las herramientas computacionales en la investigación biomédica y abre la puerta al descubrimiento de biomarcadores alternativos o complementarios. Así, este trabajo no solo corrobora conocimientos existentes, sino que también contribuye a ampliar el panorama actual sobre la biología subyacente de la talasemia.

El modelo propuesto complementa estudios previos al identificar no solo genes canónicos relacionados con la talasemia, como HBB y AHSP, sino también al resaltar la relevancia de procesos biológicos adicionales, como la regulación de la apoptosis y la homeostasis del hierro, que han sido menos enfatizados en análisis tradicionales. Asimismo, a diferencia de enfoques centrados únicamente en variantes genéticas conocidas o perfiles clínicos, este modelo se basa en patrones de expresión génica y aprendizaje automático, lo que permite detectar alteraciones sistémicas más amplias. En este sentido, el modelo desafía implícitamente estudios previos al proponer que el impacto funcional de la talasemia puede extenderse más allá de los genes estructurales de la hemoglobina, abarcando vías celulares relacionadas con el estrés oxidativo y la muerte celular. Esta perspectiva más integradora permite generar nuevas hipótesis sobre los mecanismos compensatorios o secundarios de la enfermedad, contribuyendo así a una

visión más compleja y dinámica de su fisiopatología.

Desde la perspectiva de la bioinformática médica, los hallazgos de este estudio contribuyen significativamente al desarrollo de metodologías computacionales aplicadas a la identificación de biomarcadores genéticos en enfermedades complejas como la talasemia. Esta investigación cubre una brecha importante al aplicar técnicas de detección de anomalías, comúnmente utilizadas en otros dominios, al análisis de perfiles de expresión génica, lo cual representa una alternativa viable frente a los métodos de clasificación tradicionales que suelen verse limitados por el desbalance de clases y la falta de etiquetas confiables. La relevancia científica y académica de este enfoque radica en su capacidad para generar conocimiento biológicamente significativo a partir de datos ómicos, al tiempo que propone estrategias reproducibles que pueden adaptarse al estudio de otras enfermedades genéticas. Así, el trabajo no solo aporta evidencia empírica, sino también una contribución metodológica valiosa para la comunidad de investigación en bioinformática y medicina traslacional.

Los resultados obtenidos en esta investigación ofrecen múltiples posibilidades para avanzar en el estudio de la talasemia y la identificación de biomarcadores genéticos asociados. Como parte de ello, es fundamental realizar estudios experimentales en entornos clínicos para validar la relación de los 72 genes identificados con la talasemia. Esto podría incluir análisis funcionales en líneas celulares, así como pruebas en cohortes de pacientes con la enfermedad. Una vez validados, los genes identificados pueden servir como base para desarrollar paneles genéticos específicos. Estos paneles permitirían un diagnóstico más rápido y económico de la talasemia, especialmente en regiones con alta prevalencia de la enfermedad [106].

La detección temprana de la talasemia mediante el uso de un panel clínicamente validado de genes asociados a la enfermedad puede mejorar la precisión diagnóstica y facilitar las intervenciones oportunas, lo que se traduce en mejores resultados para los pacientes. Por ejemplo, la identificación de las mutaciones genéticas que predisponen a los individuos a la sobrecarga de hierro cardiaco permite la monitorización temprana mediante herramientas como la resonancia magnética cardiaca (CMR) T2*, reduciendo significativamente el riesgo de complicaciones cardiacas graves [107]. Del mismo modo,

la incorporación de paneles de genes a los programas de tamizaje neonatal permite un diagnóstico preciso y precoz, garantizando que los individuos afectados reciban una atención rápida, como la terapia transfusional y el seguimiento de los retrasos en el desarrollo. La integración de paneles de genes en los marcos de diagnóstico y tamizaje no sólo personaliza la atención al paciente, sino que también optimiza las intervenciones tempranas, minimizando las complicaciones y mejorando la prognosis a largo plazo [108].

Asimismo, comprender las funciones de los genes asociados y su papel en los procesos biológicos relacionados con la talasemia podría abrir nuevas vías para el diseño de terapias personalizadas. Esto incluye la identificación de dianas terapéuticas para tratamientos más efectivos [109]. Es posible también integrar otras tecnologías ómicas. Combinar datos de expresión génica con otras tecnologías ómicas, como proteómica y metabolómica, podría proporcionar una visión más integral de las alteraciones moleculares asociadas a la talasemia.

Una forma de continuar la línea de trabajo planteada en esta investigación corresponde a desarrollar nuevos modelos de detección de anomalías utilizando técnicas avanzadas de aprendizaje automático, como redes neuronales profundas o modelos híbridos, para mejorar la precisión en la identificación de genes candidatos. Por otro lado, queda también abierta de realizar una validación clínica de los resultados obtenidos.

Similarmente, probar un enfoque basado en aprendizaje positivo-sin etiqueta (PU) podría ser valioso. Aunque los métodos de detección de anomalías como los bosques de aislamiento pueden identificar patrones de expresión génica raros o inusuales, no están supervisados y asumen que las anomalías son fundamentalmente diferentes de la clase mayoritaria. Esto contrasta con el aprendizaje (PU), que aprovecha los genes asociados a enfermedades conocidas como positivos y trata el resto como no etiquetados, con el objetivo de identificar genes similares a pesar de las anotaciones incompletas. Dado que los genes asociados a enfermedades no son necesariamente valores estadísticos atípicos, el aprendizaje PU ofrece un marco de clasificación relevante desde el punto de vista biológico y específico para las tareas de asociación gen-enfermedad.

En el contexto de esta investigación, es también importante considerar la colaboración con grupos de investigación clínica, bioinformática y biología molecular. Esto

permite validar experimentalmente los resultados y proponer nuevas hipótesis de investigación. Así como establecer alianzas con instituciones médicas para acelerar la transferencia del conocimiento generado hacia aplicaciones clínicas.

En síntesis, esta investigación no solo aporta un enfoque novedoso para la identificación de biomarcadores asociados a la talasemia, sino que también destaca el potencial de combinar herramientas computacionales avanzadas con análisis biológicos para abordar preguntas complejas en el ámbito biomédico. Aunque aún quedan pasos importantes para validar y aplicar clínicamente los hallazgos, los resultados obtenidos sientan las bases para futuras investigaciones que podrían transformar tanto el diagnóstico como el tratamiento de la talasemia. Este trabajo refuerza el valor de la interdisciplinariedad en la búsqueda de soluciones innovadoras para problemas de salud global y abre la puerta a nuevas posibilidades en el estudio de enfermedades genéticas.

Bibliografía

[1] A. L. Tarca, V. J. Carey, X. wen Chen, R. Romero y S. Drăghici, «Machine learning and its applications to biology.,» *PLoS computational biology*, vol. 3, n.º 6, 2007, ISSN: 15537358. DOI: 10.1371/journal.pcbi.0030116.

[2] R. Unissa, B. Monica, S. Konakanchi, R. Darak, S. L. Keerthana y S. A. Kumar, «Thalassemia: A Review,» *Asian Journal of Pharmaceutical Research*, vol. 8, n.º 3, pág. 195, 2018, ISSN: 2231-5683. DOI: 10.5958/2231-5691.2018.00034.5.

[3] K. Tari, P. Valizadeh Ardalan, M. Abbaszadehdibavar, A. Atashi, A. Jalili y M. Gheidishahran, «Thalassemia an update: molecular basis, clinical features and treatment,» *International Journal of Biomedicine and Public Health*, vol. 1, n.º 1, págs. 48-58, 2018. DOI: 10.22631/ijbmph.2018.56102.

[4] M. Angastiniotis y S. Lobitz, «Thalassemias: An overview,» *International Journal of Neonatal Screening*, vol. 5, n.º 1, págs. 1-11, 2019, ISSN: 2409515X. DOI: 10.3390/ijns5010016.

[5] G. Sáenz-Renauld y W. Rodríguez-Romero, «Síndromes talasémicos: nuevos conceptos y estado actual del conocimiento en Costa Rica,» *Acta méd. costarric*, vol. 48, n.º 4, págs. 172-178, 2006, ISSN: 0001-6002.

[6] L. Kabootarizadeh, A. Jamshidnezhad, Z. Koohmareh y A. Ghamchili, «Differential diagnosis of iron-deficiency anemia from β-thalassemia trait using an intelligent model in comparison with discriminant indexes,» *Acta Informatica Medica*, vol. 27, n.º 2, págs. 78-84, 2019, ISSN: 19865988. DOI: 10.5455/aim.2019.27.78-84.

[7] F. Esmaeilzadeh, B. Ahmadi, S. Vahedi, S. Barzegari y A. Rajabi, «Major Thalassemia, Screening or Treatment: An Economic Evaluation Study in Iran,» *International Journal of Health Policy and Management*, vol. 0, págs. 1-8, feb. de

2021, ISSN: 2322-5939. DOI: 10.34172/IJHPM.2021.04. dirección: https://www.ijhpm.com/article_4008.html.

[8] E. A. El-Sebakhy y M. A. Elshafei, «Thalassemia screening using unconstrained functional networks classifier,» *ICSPC 2007 Proceedings - 2007 IEEE International Conference on Signal Processing and Communications*, n.º November, págs. 1027-1030, 2007. DOI: 10.1109/ICSPC.2007.4728497.

[9] A. S. AlAgha, H. Faris, B. H. Hammo y A. M. Al-Zoubi, «Identifying β-thalassemia carriers using a data mining approach: The case of the Gaza Strip, Palestine,» *Artificial Intelligence in Medicine*, vol. 88, n.º July 2017, págs. 70-83, 2018, ISSN: 18732860. DOI: 10.1016/j.artmed.2018.04.009. dirección: https://doi.org/10.1016/j.artmed.2018.04.009.

[10] Z. Rustam, A. Kamalia, R. Hidayat, F. Subroto y A. S. S, «Comparison of Fuzzy C-Means , Fuzzy Kernel C-Means , and Fuzzy Kernel Robust C-Means to Classify Thalassemia Data,» *International Journal on Advanced Science Engineering and Information Technology*, vol. 9, n.º 4, págs. 1205-1210, 2019. DOI: 10.18517/ijaseit.9.4.9580.

[11] B. Çil, H. Ayyıldız y T. Tuncer, «Discrimination of β-thalassemia and iron deficiency anemia through extreme learning machine and regularized extreme learning machine based decision support system,» *Medical Hypotheses*, vol. 138, n.º December 2019, 2020, ISSN: 15322777. DOI: 10.1016/j.mehy.2020.109611.

[12] R. Mayeux, «Biomarkers: Potential Uses and Limitations,» *NeuroRx*, vol. 1, n.º 2, págs. 182-188, 2004, ISSN: 15455343. DOI: 10.1602/neurorx.1.2.182.

[13] H. Kang, Y. Wang, Z. Tong y X. Liu, «Retest positive for SARS-CoV-2 RNA of "recovered" patients with COVID-19: Persistence, sampling issues, or re-infection?» *Journal of Medical Virology*, vol. 92, n.º 11, págs. 2263-2265, 2020, ISSN: 10969071. DOI: 10.1002/jmv.26114.

[14] S. R. Alanee, Z. Roumayah, M. Deebajah et al., «Adaptive genetic algorithms combined with high sensitivity single cell-based technology derived urine-based score to differentiate between high-grade and low-grade transitional cell carcino-

ma of the bladder.,» vol. 38, n.º 6_suppl, 2020, ISSN: 0732-183X. DOI: 10.1200/jco.2020.38.6_suppl.572.

[15] S. Alanee, M. Deebajah, P. I. Chen et al., «Using adaptive genetic algorithms combined with high sensitivity single cell-based technology to detect bladder cancer in urine and provide a potential noninvasive marker for response to anti-PD1 immunotherapy,» *Urologic Oncology: Seminars and Original Investigations*, vol. 38, n.º 3, 2020, ISSN: 18732496. DOI: 10.1016/j.urolonc.2019.08.019.

[16] H. R. Roth, L. Lu, J. Liu et al., «Improving Computer-Aided Detection Using Convolutional Neural Networks and Random View Aggregation,» *IEEE Transactions on Medical Imaging*, vol. 35, n.º 5, págs. 1170-1181, 2016, ISSN: 1558254X. DOI: 10.1109/TMI.2015.2482920. arXiv: 1505.03046.

[17] M. Asyali, D. Colak, O. Demirkaya y M. Inan, «Gene Expression Profile Classification: A Review,» *Current Bioinformatics*, vol. 1, n.º 1, págs. 55-73, 2008, ISSN: 15748936. DOI: 10.2174/157489306775330615.

[18] B. Ramsundar, P. Eastman, P. Walters y V. Pande, *Deep Learning for the Life Sciences: Applying Deep Learning to Genomics, Microscopy, Drug Discovery, and More*, 1.ª ed. O'Reilly Media, Inc., 2019, ISBN: 978-1-492-03983-9.

[19] IVF Turkey, *Síntomas y tratamiento de la talasemia*. dirección: https://ivfturkey.com/es/thalassemia/ (visitado 01-07-2022).

[20] C. Starr, C. A. Evers y L. Starr, *Biology: concepts and applications*, 8.ª ed. Cengage Learning, 2011, ISBN: 978-1-4390-4673-9.

[21] T. Audesirk, G. Audesirk y B. E. Byers, *Biología: La vida en la tierra*. Pearson, 2017, vol. 0, ISBN: 978-607-32-1526-8.

[22] R. Breitling, «Biological microarray interpretation: The rules of engagement,» *Biochimica et Biophysica Acta - Gene Structure and Expression*, vol. 1759, n.º 7, págs. 319-327, 2006, ISSN: 01674781. DOI: 10.1016/j.bbaexp.2006.06.003.

[23] «Functional Discovery via a Compendium of Expression Profiles,» *Cell*, vol. 102, n.º 1, págs. 109-126, jul. de 2000, ISSN: 0092-8674. DOI: 10.1016/S0092-8674(00)

00015-5.

[24] R. Lowe, N. Shirley, M. Bleackley, S. Dolan y T. Shafee, «Transcriptomics technologies,» *PLoS Computational Biology*, vol. 13, 5 mayo de 2017, ISSN: 15537358. DOI: 10.1371/JOURNAL.PCBI.1005457. dirección: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5436640/.

[25] Z. Wang, M. Gerstein y M. Snyder, «RNA-Seq: a revolutionary tool for transcriptomics,» *Nature reviews. Genetics*, vol. 10, págs. 57-63, 1 ene. de 2009, ISSN: 1471-0064. DOI: 10.1038/NRG2484. dirección: https://pubmed.ncbi.nlm.nih.gov/19015660/.

[26] M. W. Libbrecht y W. S. Noble, «Machine learning applications in genetics and genomics,» *Nature Reviews Genetics*, vol. 16, n.º 6, págs. 321-332, 2015, ISSN: 14710064. DOI: 10.1038/nrg3920.

[27] S. Shalev-Shwartz y S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, mayo de 2014, ISBN: 978-1-107-05713-5.

[28] B. Remeseiro y V. Bolon-Canedo, «A review of feature selection methods in medical applications,» *Computers in Biology and Medicine*, vol. 112, n.º February, pág. 103 375, 2019, ISSN: 18790534. DOI: 10.1016/j.compbiomed.2019.103375. dirección: https://doi.org/10.1016/j.compbiomed.2019.103375.

[29] A. Geron, *Hands-On Machine Learning With Scikit-Learn & Tensor Flow*. O'Reilly Media, Inc., 2019, ISBN: 978-1-492-03264-9.

[30] T. Hastie, R. Tibshirani y J. Friedman, *The elements of statistical learning: data mining, inference and prediction*, 2.ª ed. Springer, 2009.

[31] C. Y. J. Peng, K. L. Lee y G. M. Ingersoll, «An introduction to logistic regression analysis and reporting,» *Journal of Educational Research*, vol. 96, 1 2002, ISSN: 19400675. DOI: 10.1080/00220670209598786.

[32] A. Ben-Hur, C. S. Ong, S. Sonnenburg, B. Schölkopf y G. Rätsch, «Support vector machines and kernels for computational biology,» *PLoS Computational*

Biology, vol. 4, 10 2008, ISSN: 1553734X. DOI: 10.1371/journal.pcbi.1000173.

[33] C. Cortes, V. Vapnik y L. Saitta, «Support-vector networks,» *Machine Learning 1995 20:3*, vol. 20, págs. 273-297, 3 sep. de 1995, ISSN: 1573-0565. DOI: 10.1007/BF00994018. dirección: https://link.springer.com/article/10.1007/BF00994018.

[34] L. Breiman, «Random forests,» *Machine Learning*, vol. 45, págs. 5-32, 1 oct. de 2001, ISSN: 08856125. DOI: 10.1023/A:1010933404324/METRICS. dirección: https://link.springer.com/article/10.1023/A:1010933404324.

[35] T. Chen y C. Guestrin, «XGBoost: A scalable tree boosting system,» en *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, vol. 13-17-August-2016, Association for Computing Machinery, ago. de 2016, págs. 785-794, ISBN: 9781450342322. DOI: 10.1145/2939672.2939785. dirección: http://dx.doi.org/10.1145/2939672.2939785.

[36] R. S. Olson, N. Bartley, R. J. Urbanowicz y J. H. Moore, «Evaluation of a tree-based pipeline optimization tool for automating data science,» en *GECCO 2016 - Proceedings of the 2016 Genetic and Evolutionary Computation Conference*, Association for Computing Machinery, Inc, jul. de 2016, págs. 485-492, ISBN: 9781450342063. DOI: 10.1145/2908812.2908918. dirección: https://dl.acm.org/doi/10.1145/2908812.2908918.

[37] L. Selicato, F. Esposito, G. Gargano et al., «A new ensemble method for detecting anomalies in gene expression matrices,» *Mathematics*, vol. 9, pág. 882, 8 abr. de 2021, ISSN: 22277390. DOI: 10.3390/MATH9080882/S1. dirección: https://www.mdpi.com/2227-7390/9/8/882.

[38] F. T. Liu, K. M. Ting y Z. H. Zhou, «Isolation forest,» en *Proceedings - IEEE International Conference on Data Mining, ICDM*, 2008, págs. 413-422, ISBN: 9780769535029. DOI: 10.1109/ICDM.2008.17.

[39] I. H. Witten, E. Frank, M. A. Hall y C. J. Pal, *Data Mining: Practical Machine Learning Tools and Techniques*. Elsevier, 2016, ISBN: 978-0123748560. DOI: 10.1016/c2009-0-19715-5.

[40] D. François, «Methodology and standards for data analysis with machine learning tools,» en *ESANN 2008 Proceedings, 16th European Symposium on Artificial Neural Networks - Advances in Computational Intelligence and Learning*, 2008.

[41] A. Subramanian, P. Tamayo, V. K. Mootha et al., «Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles,» en *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, National Academy of Sciences, oct. de 2005, págs. 15 545-15 550. DOI: 10.1073/PNAS.0506580102/SUPPL_FILE/06580FIG7.JPG. dirección: https://www.pnas.org/doi/abs/10.1073/pnas.0506580102.

[42] Y. Benjamini e Y. Hochberg, «Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing,» *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 57, págs. 289-300, 1 ene. de 1995, ISSN: 2517-6161. DOI: 10.1111/J.2517-6161.1995.TB02031.X. dirección: https://onlinelibrary.wiley.com/doi/full/10.1111/j.2517-6161.1995.tb02031.x%20https://onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1995.tb02031.x%20https://rss.onlinelibrary.wiley.com/doi/10.1111/j.2517-6161.1995.tb02031.x.

[43] T. Piroonratana, W. Wongseree, A. Assawamakin et al., «Classification of haemoglobin typing chromatograms by neural networks and decision trees for thalassaemia screening,» *Chemometrics and Intelligent Laboratory Systems*, vol. 99, n.º 2, págs. 101-110, 2009, ISSN: 01697439. DOI: 10.1016/j.chemolab.2009.07.014. dirección: http://dx.doi.org/10.1016/j.chemolab.2009.07.014.

[44] G. Barnhart-magen, V. Gotlib, R. Marilus e Y. Einav, «Differential Diagnostics of Thalassemia Minor by Artificial Neural Networks Model,» vol. 486, n.º May, págs. 481-486, 2013. DOI: 10.1002/jcla.21631.

[45] I. L. Roth, B. Lachover, G. Koren, C. Levin, L. Zalman y A. Koren, «Detection of β -Thalassemia Carriers by Red Cell Parameters Obtained from Automatic Counters using Mathematical Formulas,» 2018. DOI: 10.4084/MJHID.2018.008.

[46] H. Ayyıldız y S. Arslan Tuncer, «Determination of the effect of red blood cell parameters in the discrimination of iron deficiency anemia and beta thalassemia via Neighborhood Component Analysis Feature Selection-Based machine learning,» *Chemometrics and Intelligent Laboratory Systems*, vol. 196, n.º November 2019, 2020, ISSN: 18733239. DOI: 10.1016/j.chemolab.2019.103886.

[47] M. Sikandar, R. Sohail, Y. Saeed et al., «Analysis for Disease Gene Association Using Machine Learning,» *IEEE Access*, vol. 8, págs. 160 616-160 626, 2020, ISSN: 21693536. DOI: 10.1109/access.2020.3020592.

[48] F. Taghavifar, M. Hamid y G. Shariati, «Gene expression in blood from an individual with β-thalassemia: An RNA sequence analysis,» *Molecular Genetics and Genomic Medicine*, vol. 7, 7 jul. de 2019, ISSN: 23249269. DOI: 10.1002/MGG3.740.

[49] L. D. Mora-Jimenez, J. A. Guevara-Coto y A. Berrocal-Rojas, «AutoML approaches to the identification of novel biomarkers associated with thalassemia,» en *2023 VI Jornadas Costarricenses de Investigación en Computación e Informática (JoCICI)*, IEEE, in press.

[50] J. Piñero, J. M. Ramírez-Anguita, J. Saüch-Pitarch et al., «The DisGeNET knowledge platform for disease genomics: 2019 update,» *Nucleic Acids Research*, vol. 48, n.º D1, págs. D845-D855, ene. de 2020, ISSN: 0305-1048. DOI: 10.1093/NAR/GKZ1021. dirección: https://academic.oup.com/nar/article/48/D1/D845/5611674.

[51] N. Herman, D. Grill, P. Anderson et al., *Peripheral blood mononuclear cell (PBMC) gene expression in healthy adults rapidly transported to high altitude*. dirección: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=gse46480 (visitado 07-07-2023).

[52] T. Barrett, S. E. Wilhite, P. Ledoux et al., «NCBI GEO: archive for functional genomics data sets—update,» *Nucleic Acids Research*, vol. 41, n.º D1, págs. D991-D995, ene. de 2013, ISSN: 0305-1048. DOI: 10.1093/NAR/GKS1193. dirección: https://academic.oup.com/nar/article/41/D1/D991/1067995.

[53] R. Edgar, M. Domrachev y A. E. Lash, «Gene Expression Omnibus: NCBI gene expression and hybridization array data repository,» *Nucleic Acids Research*, vol. 30, págs. 207-210, 1 ene. de 2002, ISSN: 0305-1048. DOI: 10.1093/NAR/30.1.207. dirección: https://dx.doi.org/10.1093/nar/30.1.207.

[54] E. S. Balakirev y F. J. Ayala, «Pseudogenes: are they "junk.ºr functional DNA?» *Annual review of genetics*, vol. 37, págs. 123-151, 2003, ISSN: 0066-4197. DOI: 10.1146/ANNUREV.GENET.37.040103.103949. dirección: https://pubmed.ncbi.nlm.nih.gov/14616058/.

[55] J. A. Thompson, J. Tan y C. S. Greene, «Cross-platform normalization of microarray and RNA-seq data for machine learning applications,» *PeerJ*, vol. 4, ene. de 2016. DOI: 10.7717/peerj.1621.

[56] E. Stevens y L. Antiga, *Deep Learning with PyTorch Essential Excerpts*. 2019.

[57] M. Kuhn y K. Johnson, *Applied predictive modeling*. Springer New York, ene. de 2013, págs. 1-600, ISBN: 9781461468493. DOI: 10.1007/978-1-4614-6849-3/COVER.

[58] M. Petticrew, A. Sowden y D. Lister-Sharp, «False-negative results in screening programs: Medical, psychological, and other implications,» *International Journal of Technology Assessment in Health Care*, vol. 17, págs. 164-170, 2 2001, ISSN: 02664623. DOI: 10.1017/S0266462300105021.

[59] E. C. Bartlett, M. Silva, M. E. Callister y A. Devaraj, «False-Negative Results in Lung Cancer Screening-Evidence and Controversies,» *Journal of thoracic oncology : official publication of the International Association for the Study of Lung Cancer*, vol. 16, págs. 912-921, 6 jun. de 2021, ISSN: 1556-1380. DOI: 10.1016/J.JTHO.2021.01.1607. dirección: https://pubmed.ncbi.nlm.nih.gov/33545386/.

[60] M. Asif, H. F. Martiniano, A. M. Vicente y F. M. Couto, «Identifying disease genes using machine learning and gene functional similarities, assessed through Gene Ontology,» *PLoS ONE*, vol. 13, e0208626, 12 dic. de 2018, ISSN: 19326203. DOI:

10.1371/JOURNAL.PONE.0208626. dirección: https://pmc.ncbi.nlm.nih.gov/articles/PMC6287949/.

[61] L. D. Mora-Jimenez, K. Ramírez-Benavides, L. Quesada y J. A. Guevara-Coto, «Identification and Functional Annotation of Potential Biomarkers Associated with Thalassemia Using Machine Learning-Based Knowledge Discovery,» en *ICT for Intelligent Systems*, ép. Smart Innovation, Systems and Technologies, Springer, Singapore, 2024, págs. 191-201, ISBN: 978-981-97-5799-2. DOI: 10.1007/978-981-97-5799-2_17. dirección: https://link.springer.com/chapter/10.1007/978-981-97-5799-2%5C_17.

[62] R. E. Walpole, R. H. Myers, S. L. Myers y K. Ye, *Probability and Statistics for Engineers & Scientists*, 9.ª ed. Pearson, 2012.

[63] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2021. dirección: https://www.R-project.org/.

[64] G. Van Rossum y F. L. Drake Jr, *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam, 1995.

[65] H. Wickham, R. François, L. Henry, K. Müller y D. Vaughan, *dplyr: A Grammar of Data Manipulation*, R package version 1.1.4, https://github.com/tidyverse/dplyr, 2023. dirección: https://dplyr.tidyverse.org.

[66] H. Wickham, M. Averick, J. Bryan et al., «Welcome to the tidyverse,» *Journal of Open Source Software*, vol. 4, n.º 43, pág. 1686, 2019. DOI: 10.21105/joss.01686.

[67] T. pandas development team, *pandas-dev/pandas: Pandas*, ver. latest, feb. de 2020. DOI: 10.5281/zenodo.3509134. dirección: https://doi.org/10.5281/zenodo.3509134.

[68] H. Wickham, *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016, ISBN: 978-3-319-24277-4. dirección: https://ggplot2.tidyverse.org.

[69] S. Garnier, N. Ross, R. Rudis, A. P. Camargo y M. S. C. Scherer, *viridis(Lite) - Colorblind-Friendly Color Maps for R*, viridis package version 0.6.5, 2024. DOI: 10.5281/zenodo.4679423. dirección: https://sjmgarnier.github.io/viridis/.

[70] J. D. Hunter, «Matplotlib: A 2D graphics environment,» *Computing in Science & Engineering*, vol. 9, n.º 3, págs. 90-95, 2007. DOI: 10.1109/MCSE.2007.55.

[71] M. L. Waskom, «seaborn: statistical data visualization,» *Journal of Open Source Software*, vol. 6, n.º 60, pág. 3021, 2021. DOI: 10.21105/joss.03021. dirección: https://doi.org/10.21105/joss.03021.

[72] F. Pedregosa, G. Varoquaux, A. Gramfort et al., «Scikit-learn: Machine Learning in Python,» *Journal of Machine Learning Research*, vol. 12, págs. 2825-2830, 2011.

[73] M. Morgan y M. Ramos, *BiocManager: Access the Bioconductor Project Package Repository*, R package version 1.30.23.1, 2024. dirección: https://github.com/bioconductor/biocmanager.

[74] Z. Fang, X. Liu y G. Peltz, «GSEApy: a comprehensive package for performing gene set enrichment analysis in Python,» *Bioinformatics*, vol. 39, 1 ene. de 2023, ISSN: 1367-4803. DOI: 10.1093/BIOINFORMATICS/BTAC757. dirección: https://dx.doi.org/10.1093/bioinformatics/btac757.

[75] C. R. Harris, K. J. Millman, S. J. van der Walt et al., «Array programming with NumPy,» *Nature*, vol. 585, n.º 7825, págs. 357-362, sep. de 2020. DOI: 10.1038/s41586-020-2649-2. dirección: https://doi.org/10.1038/s41586-020-2649-2.

[76] M. J. Weiss, S. Zhou, L. Feng et al., «Role of Alpha Hemoglobin-Stabilizing Protein in Normal Erythropoiesis and β-Thalassemia,» *Annals of the New York Academy of Sciences*, vol. 1054, págs. 103-117, 1 nov. de 2005, ISSN: 1749-6632. DOI: 10.1196/ANNALS.1345.013. dirección: https://onlinelibrary.wiley.com/doi/full/10.1196/annals.1345.013.

[77] R. Maya-Martinez, Y. Xu, N. Guthertz et al., «Dimers of D76N-β_2-microglobulin

display potent antiamyloid aggregation activity,» *The Journal of Biological Chemistry*, vol. 298, pág. 102 659, 12 dic. de 2022, ISSN: 1083351X. DOI: 10.1016/J.JBC.2022.102659. dirección: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9712992/.

[78] M. J. Strowitzki, R. Nelson, M. P. Garcia et al., «Carbon Dioxide Sensing by Immune Cells Occurs through Carbonic Anhydrase 2–Dependent Changes in Intracellular pH,» *The Journal of Immunology*, vol. 208, págs. 2363-2375, 10 mayo de 2022, ISSN: 0022-1767. DOI: 10.4049/JIMMUNOL.2100665. dirección: https://dx.doi.org/10.4049/jimmunol.2100665.

[79] Q. Zhang, Q. Tang, W. Liu et al., «Novel role of CAP1 in regulation RNA polymerase II-mediated transcription elongation depends on its actin-depolymerization activity in nucleoplasm,» *Oncogene*, vol. 40, págs. 3492-3509, abr. de 2021, ISSN: 1476-5594. DOI: 10.1038/S41388-021-01789-3. dirección: https://www-nature-com.ezproxy.sibdi.ucr.ac.cr/articles/s41388-021-01789-3.

[80] Q. Zhang, Q. Han, J. Zi, C. Song y Z. Ge, «CD37 high expression as a potential biomarker and association with poor outcome in acute myeloid leukemia,» *Bioscience reports*, vol. 40, 5 mayo de 2020, ISSN: 1573-4935. DOI: 10.1042/BSR20200008. dirección: https://pubmed.ncbi.nlm.nih.gov/32400873/.

[81] B. G. Forget y H. F. Bunn, «Classification of the Disorders of Hemoglobin,» *Cold Spring Harbor Perspectives in Medicine*, vol. 3, 2 feb. de 2013, ISSN: 21571422. DOI: 10.1101/CSHPERSPECT.A011684. dirección: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3552344/.

[82] C. A. Neumann, J. Cao e Y. Manevich, «Peroxiredoxin 1 and its role in cell signaling,» *Cell cycle*, vol. 8, págs. 4072-4078, 24 dic. de 2009, ISSN: 15514005. DOI: 10.4161/CC.8.24.10242. dirección: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7161701/.

[83] P. Quarello, E. Garelli, A. Carando et al., «Diamond-Blackfan anemia: genotype-phenotype correlations in Italian patients with RPL5 and RPL11 mutations,» *Haematologica*, vol. 95, págs. 206-213, 2 feb. de 2010, ISSN: 1592-8721. DOI:

10.3324/HAEMATOL.2009.011783. dirección: https://pubmed.ncbi.nlm.nih.gov/19773262/.

[84] N. Deejai, N. Sawasdee, C. Nettuwakul et al., «Impaired trafficking and instability of mutant kidney anion exchanger 1 proteins associated with autosomal recessive distal renal tubular acidosis,» *BMC Medical Genomics*, vol. 15, 1 dic. de 2022, ISSN: 17558794. DOI: 10.1186/S12920-022-01381-Y. dirección: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9623938/.

[85] P. Y. Mantel y C. B. Schmidt-Weber, «Transforming growth factor-beta: recent advances on its role in immune tolerance.,» *Methods in molecular biology (Clifton, N.J.)*, vol. 677, págs. 303-338, 2011, ISSN: 19406029. DOI: 10.1007/978-1-60761-869-0_21/FIGURES/2. dirección: https://link-springer-com.ezproxy.sibdi.ucr.ac.cr/protocol/10.1007/978-1-60761-869-0%5C_21.

[86] I. Guyon, J. Weston, S. Barnhill y V. Vapnik, «Gene selection for cancer classification using support vector machines,» *Machine Learning*, vol. 46, págs. 389-422, 1-3 2002, ISSN: 08856125. DOI: 10.1023/A:1012487302797/METRICS. dirección: https://springerlink.proxyucr.elogim.com/article/10.1023/A:1012487302797.

[87] B. Zhang y S. Horvath, «A general framework for weighted gene co-expression network analysis,» *Statistical Applications in Genetics and Molecular Biology*, vol. 4, 1 ago. de 2005, ISSN: 15446115. DOI: 10.2202/1544-6115.1128/MACHINEREADABLECITATRIS. dirección: https://www.degruyter.com/document/doi/10.2202/1544-6115.1128/html.

[88] J. Majewski y T. Pastinen, «The study of eQTL variations by RNA-seq: From SNPs to phenotypes,» *Trends in Genetics*, vol. 27, págs. 72-79, 2 feb. de 2011, ISSN: 01689525. DOI: 10.1016/J.TIG.2010.10.006/ASSET/1DABEC02-B9D0-48F4-A14D-E4CA8MAIN.ASSETS/GR1.SML. dirección: http://www.cell.com/article/S016895251000212X/fulltext%20http://www.cell.com/article/S016895251000212X/abstract%20https://www.cell.com/trends/genetics/abstract/S0168-9525(10)00212-X.

[89] M. A. El-Brolosy y D. Y. Stainier, «Genetic compensation: A phenomenon in search of mechanisms,» *PLOS Genetics*, vol. 13, e1006780, 7 jul. de 2017,

ISSN: 1553-7404. DOI: 10.1371/JOURNAL.PGEN.1006780. dirección: https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1006780.

[90] G. E. Villalpando-Rodriguez y S. B. Gibson, «Reactive Oxygen Species (ROS) Regulates Different Types of Cell Death by Acting as a Rheostat,» *Oxidative Medicine and Cellular Longevity*, vol. 2021, 2021, ISSN: 19420994. DOI: 10.1155/2021/9912436. dirección: /pmc/articles/PMC8380163/%20/pmc/articles/PMC8380163/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8380163/.

[91] P. Walter y D. Ron, «The unfolded protein response: From stress pathway to homeostatic regulation,» *Science*, vol. 334, págs. 1081-1086, 6059 nov. de 2011, ISSN: 10959203. DOI: 10.1126/SCIENCE.1209038/ASSET/238BD530-CE66-4C19-BA32-FC1364ASSETS/GRAPHIC/334_1081_F3.JPEG. dirección: https://science.proxyucr.elogim.com/doi/10.1126/science.1209038.

[92] J. Bekker y J. Davis, «Learning from positive and unlabeled data: a survey,» *Machine Learning*, vol. 109, 4 2020, ISSN: 15730565. DOI: 10.1007/s10994-020-05877-5.

[93] P. Yang, X. L. Li, J. P. Mei, C. K. Kwoh y S. K. Ng, «Positive-unlabeled learning for disease gene identification,» *Bioinformatics*, vol. 28, 20 2012, ISSN: 13674803. DOI: 10.1093/bioinformatics/bts504.

[94] L. Cowen, T. Ideker, B. J. Raphael y R. Sharan, *Network propagation: A universal amplifier of genetic associations*, 2017. DOI: 10.1038/nrg.2017.38.

[95] R. Kiryo, G. Niu, M. C. D. Plessis y M. Sugiyama, «Positive-unlabeled learning with non-negative risk estimator,» en *Advances in Neural Information Processing Systems*, vol. 2017-December, 2017.

[96] M. E. Favero y F. F. Costa, «Alpha-Hemoglobin-Stabilizing Protein: An Erythroid Molecular Chaperone,» *Biochemistry Research International*, vol. 2011, mar. de 2011, ISSN: 20902247. DOI: 10.1155/2011/373859. dirección: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3070166/.

[97] T. L. Mollan, X. Yu, M. J. Weiss y J. S. Olson, «The Role of Alpha-Hemoglobin Stabilizing Protein in Redox Chemistry, Denaturation, and Hemoglobin As-

sembly,» *Antioxidants & Redox Signaling*, vol. 12, pág. 231, 2 ene. de 2010, ISSN: 15230864. DOI: 10.1089/ARS.2009.2780. dirección: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2821148/.

[98] B. Alberts, A. Johnson, J. Lewis et al., *Molecular Biology of the Cell*, 6.ª ed. Garland Science, 2017. DOI: 10.1201/9781315735368.

[99] I. Peixeiro, A. L. Silva y L. Romão, «Control of human β-globin mRNA stability and its impact on beta-thalassemia phenotype,» *Haematologica*, vol. 96, pág. 913, 6 feb. de 2011, ISSN: 03906078. DOI: 10.3324/HAEMATOL.2010.039206. dirección: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3105653/.

[100] A. U. Steinbicker y M. U. Muckenthaler, «Out of Balance—Systemic Iron Homeostasis in Iron-Related Disorders,» *Nutrients*, vol. 5, págs. 3034-3061, 8 ago. de 2013, ISSN: 2072-6643. DOI: 10.3390/NU5083034. dirección: https://www.mdpi.com/2072-6643/5/8/3034/htm%20https://www.mdpi.com/2072-6643/5/8/3034.

[101] R. Mariani, P. Trombini, M. Pozzi y A. Piperno, «Iron Metabolism in Thalassemia and Sickle Cell Disease.,» *Mediterranean Journal of Hematology and Infectious Diseases*, vol. 1, 1 oct. de 2009. DOI: 10.4084/MJHID.2009.006. dirección: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3033158/.

[102] J. M. Berg, J. L. Tymockzko y L. Stryer, *Biochemistry*, 7.ª ed. W. H. Freeman, 2010, ISBN: 1429229365.

[103] S. Carbon, A. Ireland, C. J. Mungall et al., «AmiGO: online access to ontology and annotation data,» *Bioinformatics*, vol. 25, pág. 289, 2 ene. de 2009, ISSN: 13674811. DOI: 10.1093/BIOINFORMATICS/BTN615. dirección: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2639003/.

[104] L. N. Moore, D. L. Holmes, A. Sharma, J. L. Vinueza y M. Lagunoff, «Bcl-xL is required to protect endothelial cells latently infected with KSHV from virus induced intrinsic apoptosis,» *PLoS pathogens*, vol. 19, 5 mayo de 2023, ISSN: 1553-7374. DOI: 10.1371/JOURNAL.PPAT.1011385. dirección: https://pubmed.ncbi.nlm.nih.gov/37163552/.

[105] G. C. Shaw, J. J. Cope, L. Li et al., «Mitoferrin is essential for erythroid iron assimilation,» *Nature*, vol. 440, 7080 2006, ISSN: 14764687. DOI: 10.1038/nature04512.

[106] D. L. Veenstra, J. Mandelblatt, P. Neumann, A. Basu, J. F. Peterson y S. D. Ramsey, «Health economics tools and precision medicine: Opportunities and challenges,» *Forum for Health Economics and Policy*, vol. 23, 1 jun. de 2020, ISSN: 15589544. DOI: 10.1515/FHEP-2019-0013/MACHINEREADABLECITATION/RIS. dirección: https://www.degruyter.com/document/doi/10.1515/fhep-2019-0013/html.

[107] N. Koonrungsesomboon, S. C. Chattipakorn, S. Fucharoen y N. Chattipakorn, «Early detection of cardiac involvement in thalassemia: From bench to bedside perspective,» *World Journal of Cardiology*, vol. 5, pág. 270, 8 2013, ISSN: 1949-8462. DOI: 10.4330/WJC.V5.I8.270. dirección: https://pmc.ncbi.nlm.nih.gov/articles/PMC3761180/.

[108] M. A. Bender, M. Hulihan, M. C. Dorley, M. D. P. Aguinaga, J. Ojodu y C. Yusuf, «Newborn screening practices for beta-thalassemia in the United States,» *International Journal of Neonatal Screening*, vol. 7, pág. 83, 4 dic. de 2021, ISSN: 2409515X. DOI: 10.3390/IJNS7040083/S1. dirección: https://pmc.ncbi.nlm.nih.gov/articles/PMC8703506/.

[109] T. A. Manolio, F. S. Collins, N. J. Cox et al., «Finding the missing heritability of complex diseases,» *Nature 2009 461:7265*, vol. 461, págs. 747-753, 7265 oct. de 2009, ISSN: 1476-4687. DOI: 10.1038/NATURE08494. dirección: https://nature.proxyucr.elogim.com/articles/nature08494.