# DESARROLLO DE MODELO DE *MACHINE LEARNING* PARA LA IDENTIFICACIÓN DE CORRELACIONES ENTRE GENOTIPO Y FENOTIPO DE INDIVIDUOS CON SÍNDROME DE PRADER-WILLI

*Daniel Felipe Romero Bernal*
*Luis Alberto Tafur Jiménez*

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

Director(a)
H. Fabian Tobar Tosse, PhD

FACULTAD DE INGENIERÍA Y CIENCIAS
MAESTRÍA EN CIENCIA DE DATOS
SANTIAGO DE CALI
2025

**RESUMEN:**

El presente proyecto aborda el Síndrome de Prader-Willi (SPW), un trastorno genético complejo asociado a alteraciones en la impronta genómica de la región cromosómica 15q11-q13, que se caracteriza por un amplio espectro de manifestaciones clínicas, incluida la obesidad severa. A pesar de los avances en diagnóstico genético, persisten limitaciones significativas en la comprensión de cómo las modificaciones genéticas y epigenéticas contribuyen a las características fenotípicas del SPW.

El objetivo principal fue desarrollar un modelo de *Machine Learning* (*ML*) para identificar correlaciones entre genotipo y fenotipo, utilizando datos genéticos y epigenéticos. Para ello, se integraron diversas fuentes de datos públicos, creando un conjunto de datos consolidado que permitió representar mejor las manifestaciones clínicas asociadas al síndrome. Se construyeron y evaluaron tres modelos de *ML*, seleccionados por su capacidad para manejar relaciones complejas entre variables y garantizar interpretabilidad. Las métricas de evaluación, como precisión, sensibilidad y *f1-score*, fueron ajustadas mediante la optimización de parámetros y la mejora del procesamiento de datos.

A pesar de las limitaciones inherentes al tamaño y calidad de la base de datos, los resultados del proyecto muestran que el enfoque propuesto es prometedor para inferir el fenotipo dado por cambios en los perfiles de metilación, a partir de las características genómicas en pacientes con SPW. Estos hallazgos podrían facilitar tanto el desarrollo de tratamientos personalizados como la identificación temprana del síndrome. En última instancia, la identificación precisa de correlaciones genotípicas y fenotípicas contribuye significativamente a una mejor comprensión de los mecanismos moleculares subyacentes del SPW y sus posibles implicaciones terapéuticas.

**Palabras clave:** epigenética, genética, modelamiento.

# TABLA DE CONTENIDO

# LISTA DE FIGURAS

# LISTA DE TABLAS

# INTRODUCCIÓN

El Síndrome de Prader-Willi (SPW) es un trastorno genético complejo y poco frecuente que representa un desafío significativo para la comunidad científica y médica debido a la diversidad y severidad de sus manifestaciones clínicas. Este síndrome, reconocido como la principal causa genética de obesidad en humanos, se asocia con alteraciones en la región cromosómica 15q11-q13, principalmente por deleciones paternas o disomías uniparentales maternas. Las personas afectadas presentan una variedad de características multisistémicas, incluidas hipotonía, hiperfagia, obesidad, problemas cognitivos, trastornos conductuales y disfunciones endocrinas.

El diagnóstico molecular del SPW ha evolucionado significativamente mediante el análisis de metilación del ADN en regiones críticas como el gen SNRPN. Estas metodologías tradicionales, aunque efectivas, tienen limitaciones en la integración de grandes volúmenes de datos genéticos y epigenéticos, necesarios para comprender la relación entre características genómicas y fenotípicas. En este proyecto, se desarrolló una base de datos unificada que integró datos genéticos y epigenéticos de fuentes abiertas para analizar las correlaciones genotipo-fenotipo en el SPW mediante modelos de aprendizaje automático (Machine Learning, ML). La creación de esta base de datos permitió superar las barreras relacionadas con la fragmentación de la información, proporcionando un insumo único para el análisis.

A pesar de las limitaciones encontradas en los datos, como su baja calidad y volumen insuficiente, los modelos de ML desarrollados lograron identificar patrones significativos, especialmente en casos donde las diferencias entre clases eran más marcadas. Los algoritmos interpretables empleados, como Random Forest y XGBoost, permitieron relacionar características genéticas, como SNPs y repeticiones genómicas, con las expresiones fenotípicas. Esto subraya su potencial para apoyar investigaciones futuras en la identificación temprana y caracterización de este síndrome.

Los resultados obtenidos reflejan la relevancia de las variaciones genómicas puntuales y las estructuras repetitivas en la región crítica del cromosoma 15, destacando su utilidad para guiar investigaciones posteriores. Este enfoque interdisciplinario no solo contribuye al entendimiento del SPW, sino que también sienta las bases para mejorar estrategias diagnósticas y terapéuticas en enfermedades genéticas complejas.

# 1. DEFINICIÓN DEL PROBLEMA

## 1.1. PLANTEAMIENTO DEL PROBLEMA

El Síndrome de Prader-Willi (SPW) es un raro y complejo trastorno genético multisistémico reconocido como la causa genética más comúnmente conocida de obesidad potencialmente mortal en los seres humanos. El SPW surge de errores en la impresión genómica con la falta de expresión de genes improntados heredados paternamente en la región del cromosoma 15q11-q13, generalmente causada por una deleción paterna o una disomía materna 15, en la cual ambos cromosomas 15 son heredados de la madre. Igualmente, este se caracteriza por una variedad de características clínicas, que incluyen hipotonía, hiperfagia que conduce a la obesidad, problemas cognitivos y de comportamiento, y anomalías endocrinas [1].

La isla CpG1 que flanquea la región promotora de SNRPN en el cromosoma 15q11.2 contiene sitios CpG que están completamente metilados en el alelo de origen materno y no metilados en el alelo de origen paterno. Se observan tanto alelos no metilados como metilados en individuos normales. Solo se observa el alelo metilado en pacientes con el síndrome de Prader-Willi, por lo tanto, la detección de la metilación aberrante en la región diferencialmente metilada es fundamental para el diagnóstico molecular de SPW. Tradicionalmente, se han utilizado el tratamiento con bisulfito y el tratamiento con enzimas de restricción sensibles a la metilación, o la amplificación de sondas dependiente de la ligación multiplex específica de metilación (MS-MLPA, por sus siglas en inglés) [2].

El diagnóstico y la causa molecular pueden identificarse mediante el análisis simultáneo de metilación del ADN y un arreglo combinado de oligos-SNP (OSA). El análisis de metilación del ADN identifica una impronta exclusivamente materna dentro de la región PWCR. El OSA puede identificar la causa molecular en aquellos con una deleción 15q11.2-q13, una deleción del centro de impronta, y una isodisomía uniparental y una isodisomía segmentaria. En individuos con impronta exclusivamente materna identificada en el análisis de metilación del ADN y un OSA normal, el análisis de polimorfismos de ADN puede utilizarse para distinguir entre una heterodisomía uniparental y un defecto de impronta mediante epimutación [3].

La etiología genética del SPW, junto con el amplio espectro fenotípico observado en individuos afectados, plantea un desafío significativo para comprender los mecanismos moleculares subyacentes al síndrome. El defecto molecular específico que subyace al SPW brinda la oportunidad de explorar la terapia epigenética para reactivar la expresión de los genes PWS reprimidos heredados del cromosoma materno. Aunque la comprensión de la base molecular del SPW ha cambiado fundamentalmente, ha habido poco progreso en la terapia epigenética del SPW que apunta a sus defectos genéticos subyacentes [4].

A pesar de los avances en diagnósticos genéticos, la correlación precisa entre la extensión de la deleción en el cromosoma 15 y las manifestaciones clínicas del SPW sigue siendo esquiva.

[1] Una isla CpG se define como una región de ADN de 200 pares de bases con un contenido de GC superior al 50% y una proporción observada de CpG frente a CpG esperado mayor o igual a 0.6 [5].

Además, el impacto de las modificaciones epigenéticas, en particular los patrones anómalos de metilación del ADN en la región crítica del SPW, no se comprende completamente. Enfoques analíticos tradicionales han contribuido sustancialmente a desentrañar la base genética y epigenética del SPW. Sin embargo, la integración de métodos de aprendizaje automático tiene el potencial de mejorar nuestra comprensión de las relaciones intrincadas entre el genotipo y el fenotipo. Las investigaciones sobre la genética del SPW buscan, por tanto, determinar los genes específicos en esta región que están involucrados en causar el síndrome y cómo su estado impreso contribuye a las características y síntomas del SPW.

### 1.2. FORMULACIÓN DEL PROBLEMA

Dada la complejidad en la correlación entre genotipo y fenotipo en el Síndrome de Prader-Willi (SPW), surgió el siguiente interrogante problemático: ¿Cómo desarrollar un modelo de *Machine Learning (ML)* que, a partir de datos de análisis genéticos y epigenéticos de individuos con SPW, permita la identificación de correlaciones entre genotipo y fenotipo?

**Preguntas de sistematización:**

1. ¿Cómo preparar adecuadamente el conjunto de datos de análisis genéticos y epigenéticos, considerando las necesidades de comprensión de las manifestaciones clínicas del SPW?
2. ¿Cuáles son las técnicas de *ML* más apropiadas para construir un modelo que capture las complejas relaciones entre genotipo y fenotipo en el SPW?
3. ¿Cómo evaluar de manera exhaustiva el rendimiento del modelo, ajustando técnicas, parámetros e hiperparámetros con datos de prueba en términos de precisión, *recall* y *f1-score*?
4. ¿En qué medida el sistema desarrollado puede medir con precisión las correlaciones entre genotipo y fenotipo, contribuyendo a una mejor comprensión del Síndrome de Prader-Willi?

# 2. OBJETIVOS DEL PROYECTO

## 2.1 OBJETIVO GENERAL

Desarrollar un modelo de *Machine Learning* a partir de datos de análisis genéticos y epigenéticos de individuos con Síndrome de Prader-Willi para la identificación de correlaciones entre genotipo y fenotipo.

## 2.2 OBJETIVOS ESPECÍFICOS

1 Preparar el conjunto de datos de análisis genéticos y epigenéticos de individuos con Síndrome de Prader-Willi, teniendo en cuenta las necesidades de entendimiento de las manifestaciones clínicas, asegurando la calidad y coherencia de los datos para su posterior análisis.

2 Construir un modelo a partir de la aplicación de las técnicas de *machine learning* seleccionadas sobre el conjunto de datos de análisis genéticos y epigenéticos.

3 Evaluar el rendimiento del modelo con datos de prueba en términos de precisión, *recall* y *f1-score*, mediante la modificación de técnicas, parámetros e hiperpárametros del mismo.

4 Medir el grado de precisión en la identificación de correlaciones entre genotipo y fenotipo, como apoyo a la comprensión del Síndrome de Prader-Willi.

# 3. MARCO DE REFERENCIA

## 3.1 MARCO TEÓRICO

A continuación, se presentan los temas que se relacionan con el desarrollo del proyecto, teniendo en cuenta que para el presente proyecto se tienen dos fundamentos, por un lado, las técnicas de *Machine Learning*, en este caso aplicadas a datos genéticos, pero, por otro lado, es necesario entender la patología de base que origina el problema, para poder interpretar correctamente los resultados obtenidos.

### 3.1.1. SÍNDROME DE PRADER-WILLI: SINTOMATOLOGÍA CLÍNICA

El síndrome de Prader-Willi (SPW) es un complejo trastorno genético multisistémico con una ocurrencia de 1 por cada 15000 personas [1]. Este síndrome se caracteriza por una alimentación deficiente e hipotonía en la infancia, además, una alimentación excesiva y obesidad después de la niñez temprana [2]. El SPW surge de errores en la impresión genómica con la falta de expresión de genes improntados heredados paternamente en la región del cromosoma 15q11-q13, generalmente causada por una deleción paterna o una disomía materna 15, en la cual ambos cromosomas 15 son heredados de la madre [1].

De esta manera se destaca cómo el SPW se expresa de forma comportamental, a pesar de que el diagnostico se realiza a partir de análisis genético. Actualmente el SPW está relacionado también con arrebatos de temperamento, rasgos obsesivos y terquedad, y características clínicas como la adiposidad central, trastornos del sueño, anormalidades de temperatura y percepción del dolor. Gracias a la identificación de estos patrones, el factor comportamental del SPW se evalúa según los Criterios Diagnósticos de Consenso, los cuales funcionan a partir de verificación de síntomas. Un puntaje ponderado de 8 o más para edades > 3 (5 o más para edades < 4), basado en la presencia de ocho síntomas mayores (puntaje 1) y 11 síntomas menores (puntaje 0.5), se considera suficiente para un diagnóstico clínico de PWS [6].

A partir de los criterios menores y mayores diagnósticos de consenso clínicos, consignados en la tabla 1, se entiende que el SPW se caracteriza por manifestaciones clínicas que varían con la edad, desde hipotonía prenatal que puede resultar en parto asistido, hasta problemas conductuales y obesidad severa en la infancia y adolescencia. La hipotonía, marcada por letargia y succión deficiente, persiste de manera leve a lo largo de la vida, afectando hitos físicos y sociales que se alcanzan a aproximadamente el doble de la edad normal. Los desafíos incluyen hiperfagia, falta de saciedad, problemas endocrinos como deficiencia de hormona del crecimiento e hipogonadismo, y complicaciones graves como insuficiencia respiratoria, problemas cardiovasculares y diabetes tipo 2, lo que contribuye a tasas de mortalidad entre el 1.25% y el 3% anual [7]. Aun así, muchos de estos síntomas aún no se comprende la causa genética.

Tabla 1 Criterios Diagnósticos de Consenso

| | | |
|---|---|---|
| **Criterios menores** | - Hipotonía central neonatal e infantil con succión deficiente, que mejora gradualmente con la edad.<br>- Problemas de alimentación en la infancia con necesidad de técnicas especiales de alimentación y escaso aumento de peso/fallo en el crecimiento.<br>- Ganancia de peso excesiva o rápida en la gráfica de peso para longitud (definida como cruzar dos canales centiles) después de los 12 meses, pero antes de los 6 años de edad; obesidad central en ausencia de intervención.<br>- Rasgos faciales característicos con dolicocefalia en la infancia, rostro estrecho o diámetro bifrontal reducido, ojos en forma de almendra, boca pequeña con labio superior delgado, esquinas de la boca hacia abajo (se requieren 3 o más).<br>- Retraso global en el desarrollo en un niño menor de 6 años; retraso mental leve a moderado o problemas de aprendizaje en niños mayores. | - Hipogonadismo con cualquiera de los siguientes, según la edad:<br>- (A) Hipoplasia genital, masculina: hipoplasia escrotal, criptorquidia, pene y/o testículos pequeños para la edad (<percentil 5); femenina: ausencia o hipoplasia severa de los labios menores y/o clítoris.<br>- (B) Maduración gonadal retrasada o incompleta con signos puberales retrasados en ausencia de intervención después de los 16 años (masculino: gónadas pequeñas, disminución del vello facial y corporal, falta de cambio de voz; femenino: amenorrea/oligomenorrea después de los 16 años).<br>- Hiperfagia/exploración de alimentos/obsesión con la comida.<br>- Deleción 15q11-13 en alta resolución (>650 bandas) u otra anormalidad citogenética/molecular de la región cromosómica de SPW, incluyendo la disomía materna. |
| **Criterios mayores** | - Movimiento fetal disminuido o letargia infantil o llanto débil en la infancia, que mejora con la edad.<br>- Problemas de comportamiento característicos: rabietas, arrebatos violentos y comportamiento obsesivo-compulsivo, tendencia a ser argumentativo, opositor, rígido, manipulador, posesivo y terco; perseverancia, robo y mentira (se requieren 5 o más de estos síntomas).<br>- Trastorno del sueño o apnea del sueño.<br>- Baja estatura para el fondo genético a los 15 años (en ausencia de intervención con hormona de crecimiento).<br>- Hipopigmentación: piel y cabello claros en comparación con la familia.<br>- Manos pequeñas (<percentil 25) y/o pies (<percentil 10) para la edad.<br>- Manos estrechas con borde ulnar recto. | - Anomalías oculares (esotropía, miopía).<br>- Saliva gruesa y viscosa con costras en las comisuras de la boca.<br>- Defectos en la articulación del habla.<br>- Arrancarse la piel.<br>- Hallazgos de apoyo.<br>- Umbral de dolor elevado.<br>- Vómitos disminuidos.<br>- Inestabilidad térmica en la infancia o sensibilidad térmica alterada en niños mayores y adultos.<br>- Escoliosis y/o cifosis.<br>- Adrenarquia temprana.<br>- Osteoporosis, huesos delgados, por ejemplo, fácilmente fracturables.<br>- Habilidad inusual con rompecabezas.<br>- Estudios neuromusculares normales. |

Fuente: Tomado de Whittington [6].

Nota: La tabla presenta los criterios de diagnostico usado para la identificación del SPW en pacientes, así como su categorización, según su importancia.

### 3.1.2. SÍNDROME DE PRADER-WILLI: GENÉTICA MOLECULAR

En el caso del proyecto se busca contrastar la información clínica sintomática, mayormente como expresión del fenotipo, con estudios genéticos, por lo que se hace necesario comprender genéticamente el SPW. La región de PWS abarca aproximadamente ~6 Mb en el brazo largo del cromosoma 15 (Figura 1). Al menos 2.5 Mb de esta región comprenden genes con expresión diferencial, según el origen parental. Este locus contiene genes codificadores de proteínas y varias

moléculas de ARN no codificantes, involucradas en la regulación del empalme alternativo, principalmente en el cerebro. El gen bicistrónico SNURF-SNRPN es central para la región de SPW y es crucial para entender el patrón de metilación en el síndrome. La isla CpG en el extremo 5' de SNURF-SNRPN (que abarca la región promotora, el exón 1 y el intrón 1) está diferencialmente impresa según el origen parental: el alelo paterno no metilado se expresa mientras que el alelo materno metilado se reprime [7].

El centro de impronta de PWS (IC en la Figura 1) involucra la isla CpG y el exón 1 dentro de la región de solapamiento más pequeña de 4.3 Kb. Además, la expresión de SNURF-SNRPN produce un transcripto largo que también incluye PWS-IC, seis genes snoRNA, IPW y UBE3A antisentido, lo que reprime UBE3A paterno. La mayoría de los pacientes con PWS (65–75%) presentan una deleción de 5–6 Mb en 15q11-q13 de origen paterno. La Disomía Uniparental Materna (mDUP) ocurre cuando ambos cromosomas 15 se heredan de la madre y representa aproximadamente el 20–30% de los casos, asociándose con la edad materna avanzada. Adicionalmente, los defectos de impronta causados por epimutaciones o microdeleciones en el PWS-IC en el representan el 1–3% de los casos de SPW, teniendo herencia biparental de alelos, pero un patrón de metilación de ADN solo materno [7].

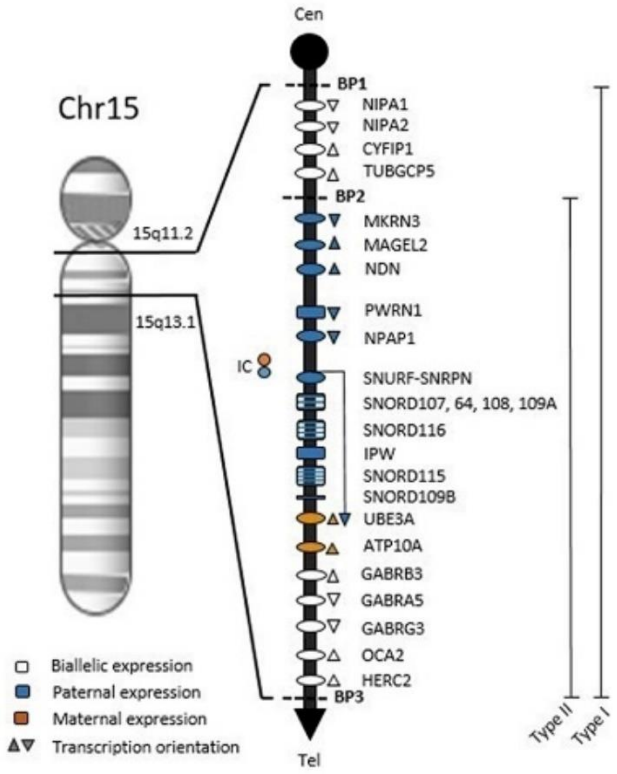


Figura 1: Mapa del cromosoma 15 en la región 15q11.2-q13.1.

Fuente: Tomado de Costa et al [7].

Nota: La figura presenta la localización de la región de interés para el SPW dentro del cromosoma 15, así como el origen de la expresión génica (paterna o materna).

La complejidad clínica y molecular en el SPW destaca la importancia del diagnóstico genético en la

definición terapéutica y el asesoramiento genético. Actualmente, hay tres ensayos con esta capacidad de detección: la PCR específica de metilación (MS-PCR, el estándar de oro), la amplificación dependiente de sonda específica de metilación multiplex (MS-MLPA) y la fusión de alta resolución sensible a la metilación (MS-HRM) [7].

### 3.1.3. SÍNDROME DE PRADER-WILLI: CORRELACIÓN GENOTIPO - FENOTIPO

Ninguno de los errores genéticos asociados con el SPW está vinculado exclusivamente a síntomas específicos, por lo que el análisis de la correlación no es identificar un sentido de causalidad, sino escenarios genéticos más prevalentes en el fenotipo observado. Las clases moleculares más prevalentes (deleción y mUPD) muestran diferencias estadísticas en la frecuencia o gravedad de algunas características clínicas. Los pacientes con deleción paterna estaban más relacionados con problemas de alimentación, trastornos del sueño, hipopigmentación y déficits en el habla y el lenguaje. Igualmente, las personas con la deleción tipo 1 más grande tenían un mejor rendimiento académico, así como más compulsiones en comparación con los pacientes con deleción tipo 2. Varios otros rasgos son más comunes en individuos con mUPD, como parto posérmino, coeficiente intelectual verbal más alto, psicosis y trastorno del espectro autista. Por otro lado, los pacientes con mUPD tienen menos probabilidad de tener la apariencia facial típica de SPW o hipopigmentación [7]. Es decir, actualmente existe una serie de aproximaciones estadísticas a la identificación de correlaciones entre escenarios genéticos y la expresión fenotípica, sin embargo, estas correlaciones no logran ser altamente precisas.

### 3.1.4. MACHINE LEARNING (ML): APRENDIZAJE SUPERVISADO Y NO SUPERVISADO

Los métodos de *ML* son enfoques computacionales utilizados para identificar patrones significativos en los datos, en el caso del proyecto, estos enfoques permitirían predecir la expresión fenotípica a partir de su perfil de expresión génica. Hay dos subtipos principales de *ML*: supervisado y no supervisado. En el supervisado, el algoritmo toma un conjunto de datos de ejemplos etiquetados (por ejemplo, expresado vs. no expresado) y la tarea es predecir etiquetas en nuevos ejemplos (no vistos). En el aprendizaje no supervisado, la tarea es identificar la estructura en los datos sin etiquetas previas. También existe una clase de métodos de *ML* conocidos como enfoques semisupervisados, que se pueden aplicar cuando los datos disponibles solo están parcialmente etiquetados, generalmente incluyendo algunos ejemplos etiquetados y muchos ejemplos no etiquetados [8].

Dos tipos de problemas de aprendizaje supervisado comúnmente encontrados en entornos biológicos son problemas de clasificación y regresión. En un problema de clasificación, como la predicción de resultados o la identificación de subgrupos, el objetivo es asignar cada ejemplo a uno de un conjunto de clases distintas. En un problema de regresión, la tarea es predecir un valor continuo (por ejemplo, predecir el nivel de un biomarcador o la tasa de filtración glomerular estimada dados datos de expresión génica) [8].

En todo problema de *ML* se tiene una entrada, la cual es una lista de atributos para cada objetivo de estudio, por ejemplo, cada paciente, la cual se suministra al algoritmo, llamada conjunto de características (*feature set*). En el caso de un problema de aprendizaje supervisado, el algoritmo también recibe un conjunto de etiquetas de clase, especificando a qué categoría pertenece cada

ejemplo. Un clasificador es el modelo que desarrolla el algoritmo de *ML* para llevar a cabo la tarea especificada dado el conjunto de características proporcionadas. Por otro lado, es común que haya uno o más hiperparámetros asociados con el clasificador, estos son inputs que alteran el comportamiento del clasificador. También suele haber parámetros asociados con el clasificador, que son configuraciones del clasificador que se optimizan automáticamente por el algoritmo de *ML* durante el entrenamiento del modelo [8].

Entre los posibles algoritmos de clasificación usados en *ML*, algunos presentan desempeños altos, sin embargo, con poca interpretabilidad, debido a la complejidad del propio algoritmo. En este caso, la interpretabilidad del modelo es una prioridad debido a la necesidad no sólo de clasificar, sino de observar correlaciones a nivel practico entre el genotipo y fenotipo en el síndrome. Uno de los algoritmos con alta interpretabilidad se encuentra *Random Forest* el cual es un algoritmo que combina múltiples modelos, basados en árboles de decisión, para obtener un modelo más preciso que con sólo un modelo.

De manera simplificada *Random Forest* genera diferentes árboles de decisión con subconjuntos aleatorios de los datos, y cada uno de estos modelos prioriza una clase determinada. Así, este algoritmo tiene una robustez adecuada en la medida que reduce el sobreajuste a trabajar aleatoriamente con múltiples modelos internamente, así como mejora la precisión general [9]. Otro algoritmo altamente usado por su interpretabilidad es *XGBoost* el cual, a diferencia de *Random Forest* no crea modelos aleatorios, sino que prioriza un enfoque secuencial, es decir, cada nuevo modelo se construye corrigiendo los errores del modelo anterior [10].

El enfoque de *XGBoost* permite una mayor precisión en muchos casos, además de incorporar mecanismos de regularización para evitar sobreajuste. Ahora bien, a pesar de tender a presentar una mayor precisión, su naturaleza lo hace menos interpretable que *Random Forest*, sin embargo, ambos algoritmos proporcionan una medida de la importancia de cada característica en el modelo, medida que ayuda a entender qué variables influyen más en la tarea de clasificación del modelo [10].

Uno de los métodos de *ML* que más relevancia ha tomado últimamente es el *deep learning*, lo cual hace referencia realmente a una familia de métodos novedosos para *ML*. En el *deep learning*, las características de entrada se someten a múltiples capas de transformaciones, en las cuales las salidas de cada capa son funciones de subconjuntos de la entrada a esa capa. Estos son más potentes cuando hay grandes cantidades de puntos de datos de entrenamiento disponibles [8].

Uno de los algoritmos que hace uso de esta arquitectura es el *Multilayer Perceptron* el cual es una red neuronal compuesta por nodos interconectados. En tareas de clasificación su precisión es alta siempre que se cuente con una suficiente cantidad de puntos de entrenamiento que le permitan identificar patrones complejos, además, su flexibilidad y adaptabilidad a los datos lo destacan como un algoritmo para *ML* altamente eficiente. Su mayor desventaja es tener una interpretatividad muy baja al ser considerada una caja negra, pues la complejidad de las redes neuronales y gran cantidad de parámetros hace difícil de entender cómo llega a decisiones [11].

En el proceso de implementación de métodos de *ML* es crucial la validación del clasificador para asegurarse de que haga predicciones precisas en datos no vistos. Para evaluar su rendimiento en

datos que no se utilizaron para construir el clasificador, a menudo se dividen los datos de entrada en conjuntos de entrenamiento, validación y prueba. El conjunto de entrenamiento se utiliza para construir el clasificador y aprender parámetros; el conjunto de validación se utiliza para evaluar el rendimiento del clasificador y ajustar hiperparámetros; y el conjunto de prueba, que se examina solo en la etapa final, se utiliza para evaluar el rendimiento del clasificador.

Existen diversas formas de validar el clasificador, una de esas es la validación cruzada (*k-fold*). En esta, los datos de entrada se dividen en k partes. Luego, el clasificador se entrena en todas las partes excepto una, y se valida en la parte final. Este procedimiento se repite hasta que cada parte se ha dejado fuera una vez. Se evalúa el rendimiento de un clasificador en función del número de errores y ejemplos clasificados correctamente. Al construir clasificadores, es importante asegurarse de que los datos utilizados para entrenar y evaluar el clasificador sean similares a los datos a los que se aplicará el clasificador. También es importante, al evaluar el rendimiento de un clasificador, asegurarse de que el conjunto de prueba no se haya utilizado para entrenar parámetros o hiperparámetros [8].

Finalmente, una vez el modelo es desarrollado, es vital evaluar correctamente su rendimiento. Al respecto se encuentran unas métricas básicas de desempeño del modelo, siempre que sea un modelo supervisado [12]. Las métricas más destacadas son: sensibilidad, especificidad, matriz de confusión (falsos positivos, falsos negativos, precisión y predicciones verdaderos negativos), exactitud y F1 *score* como está consignado en la Tabla 2. A pesar de que estas métricas estás generalmente asociadas a clasificaciones binarias, pueden ser extendidas a clasificaciones multiclases.

Algunas de las estrategias que pueden usarse para usarse en clasificaciones multiclases son: observación micro, donde se calculan las mismas métricas para cada clase; observación macro, donde se computa como un todo; promediado, en donde las métricas reportadas corresponden al promedio de las métricas de la observación micro; y finalmente, se puede realizar por muestra, donde es calculado por cada instancia y posteriormente promediado [12].

Tabla 2: Métricas de evaluación en *Machine Learning*

| **Métrica** | **Significado** | **Formula** |
|---|---|---|
| **Sensibilidad** | Fracción de casos positivos predichos como positivos | $TP/(TP+FN)$ |
| **Especificidad** | Fracción de casos negativos predichos como negativos | $TN/(TN+FP)$ |
| **Falsos positivos** | Fracción de casos predichos como positivos que eran realmente negativos | $FP/(TN+FP)$ |

| | | |
|---|---|---|
| **Falsos negativos** | Fracción de casos predichos como negativos que eran realmente positivos | $FN/(TP+FN)$ |
| **Verdaderos positivos (precisión)** | Fracción de casos positivos que sí lo eran, respecto al total de casos predichos positivos. | $TP/(TP+FP)$ |
| **Verdaderos negativos** | Fracción de casos negativos que sí lo eran, respecto al total de casos predichos negativos. | $TN/(TN+FN)$ |
| **Exactitud** | Fracción de casos correctamente predichos | $(TP+TN)/(TN+FN+TN+FP)$ |
| ***F1 score*** | Media armónica de casos positivos predichos y la sensibilidad | $(2TP)/(2TP+FP+FN)$ |
| **FN:** Falso negativo; **FP:** Falso positivo; **TN:** Verdadero negativo; **TP:** Verdadero positivo. | | |

Fuente: Tomado de Erikson y otros [12].

Nota: La tabla presenta las métricas de evaluación más usadas para modelos de clasificación en *Machine Learning*.

## 3.2 ANTECEDENTES

Saber la causa genética de un rasgo permite predecir si un organismo tiene el potencial de desarrollar un fenotipo esperado. Los métodos de mapeo de asociación rasgo-marcador basados en regresión han identificado numerosas regiones genómicas o genes, para los cuales las variaciones en las secuencias de ADN o las expresiones génicas están asociadas con los rasgos de interés [13].

En el caso del SPW no se encuentran trabajos previos en el área de *ML*. Esta asociación entre genotipo y fenotipo es comúnmente realizada para plantas pues ha permitido la selección de rasgos asociados de interés. Los algoritmos avanzados de *ML*, junto con grandes cantidades de datos adquiridos de experimentos estratégicamente diseñados, han fortalecido significativamente los esfuerzos de identificación de genes. La aplicación del aprendizaje profundo para establecer la asociación entre rasgos y genes es prometedora, pero puede requerir modelos interpretables [13].

### *3.2.1. XRARE: A MACHINE LEARNING METHOD JOINTLY MODELING PHENOTYPES AND GENETIC EVIDENCE FOR RARE DISEASE DIAGNOSIS*

Sin embargo, sí se encuentran trabajos en los que se ha usado el *ML* como herramienta para analizar datos genéticos. Por ejemplo, Li et al [14] desarrollaron *Xrare*, el cual es un enfoque de aprendizaje automático para la priorización de variantes causantes de enfermedades basado en un conjunto completo de características fenotípicas y genéticas, esto se realizó teniendo como foco la categoría de enfermedades raras. En resumen, hay 51 características, que incluyen 6 características relacionadas con la frecuencia alélica en la población, 5 puntuaciones de similitud gen-fenotipo, 15 características basadas en pautas de *ACMG (Medical Genetics and Genomics)*

*/AMP (Association for Molecular Pathology)*, 9 puntuaciones de restricción a nivel de gen, 12 puntuaciones existentes de predicción in silico de la patogenicidad, 2 características de impacto funcional de las variantes y 2 características relacionadas con bases de datos a nivel de gen.

De esta manera, Li et al [14] desarrollaron una nueva medida de similitud de fenotipo para manejar la naturaleza imprecisa y ruidosa de los fenotipos clínicos, y utilizaron un enfoque de impulso de árbol de gradiente para ampliar las anotaciones gen-fenotipo a genes que aún no están anotados en *HPO (Human Phenotype Ontology)*. Además, el uso de impulso de árbol de gradiente, y no un árbol de decisión fijo, para combinar todas las características genotípicas y fenotípicas para predecir la patogenicidad de una variante mostró un rendimiento significativamente mejor, ya sea que los genes que contribuyen a la enfermedad genética sean conocidos o novedosos.

### *3.2.2. MACHINE LEARNING, THE KIDNEY, AND GENOTYPE–PHENOTYPE ANALYSIS*

Por otro lado, en trabajos previos de investigación también se ha usado la aproximación por *ML* para la interpretación de datos genotípicos, y su relación con la expresión en diferentes tejidos. Un ejemplo de enfoque integrador que combina múltiples tipos de datos para facilitar la interpretación de los datos genotípicos es el algoritmo *NetWAS*, que vuelve a priorizar genes para identificar genes probablemente causales a partir de estudios de asociación a nivel del genoma *(GWAS)*. *NetWAS* utiliza información de redes funcionales específicas de tejidos que cuantifican la probabilidad de que cualquier par de genes esté funcionalmente relacionado en un tejido específico (por ejemplo, riñón) mediante la integración de las relaciones entre genes en miles de experimentos. Los patrones de conectividad de la red para los genes de interés se pueden utilizar como características para el algoritmo de aprendizaje automático de *NetWAS*, utilizando genes con resultados de *GWAS* marginalmente significativos como ejemplos positivos y reorganizando todos los genes en el genoma por asociación probable con el fenotipo estudiado [8].

### *3.2.3. ADULTS WITH PRADER–WILLI SYNDROME EXHIBIT A UNIQUE MICROBIOTA PROFILE*

Ahora bien, al enfocar la búsqueda de trabajos previos sólo en el SPW donde hayan usado técnicas de *ML*. Dahl et al [15] determinaron si la composición de la microbiota fecal en adultos con SPW difiere de la de adultos no afectados, para ello se utilizaron muestras de dieta habitual/no intervencionista, se analizó la composición de la microbiota fecal mediante secuenciación de amplicones del gen 16S ARNr. En este proyecto se emplearon algoritmos de aprendizaje automático con el clasificador de muestras QIIME 2™ para entrenar cruzadamente las muestras y predecir a qué conjunto de datos pertenecen los perfiles taxonómicos, también se extrajeron los taxones que más diferenciaban entre todos los conjuntos de datos y se realizó una inspección visual con la librería R PiratePlots para seleccionar los taxones que diferían en abundancia específicamente en SPW.

### 3.2.4. *THE GUT MICROBIOTA PROFILE IN CHILDREN WITH PRADER–WILLI SYNDROME*

Siguiendo la misma hipótesis de relación entre microbiota fecal y el SPW, Peng et al [16] caracterizaron las comunidades bacterianas y fúngicas intestinales de niños con SPW y determinar las asociaciones con la hiperfagia. Para ello, recopilaron muestras de heces de 25 niños con SPW y 25 controles emparejados por edad, sexo e índice de masa corporal. También se obtuvieron datos

de ingesta dietética, puntuaciones de hiperfagia e información clínica relevante. Las comunidades bacterianas y fúngicas fecales fueron caracterizadas mediante secuenciación de ARNr 16S e ITS2, respectivamente. En este estudio no se realizó uso de técnicas de *ML*, sino un análisis bivariado, para determinar la correlación entre variables.

### *3.2.5. COMPUTER-AIDED FACIAL ANALYSIS AS A TOOL TO IDENTIFY PATIENTS WITH SILVER–RUSSELL SYNDROME AND PRADER–WILLI SYNDROME*

Un estudio reciente evaluó el rendimiento de la aplicación *Face2Gene*, en pacientes con el síndrome de Silver-Russell (SRS) y el síndrome de Prader-Willi (PWS). Esta aplicación se apoya en la hipótesis de que los síndromes genéticos a menudo muestran rasgos faciales que proporcionan pistas para el diagnóstico, por lo que la aplicación analiza características detectadas en una o más imágenes faciales de individuos afectado para aproximar el diagnostico de síndromes genéticos. En este proyecto se analizaron 23 pacientes pediátricos con SRS diagnosticados clínica o genéticamente y 29 pacientes pediátricos con SPW confirmado genéticamente, y se adquirió una foto frontal de cada paciente para investigar la correlación con el diagnóstico genético específico [17].

La aplicación *Face2Gene* genera un listado de 30 posibles síndromes a los que puede estar asociado, en ese sentido en el grupo de PWS, las sensibilidades de 1, 5 y 10 principales fueron del 76%, 97% y 100%, respectivamente. PWS se sugirió como el primero en el 83% de los pacientes diagnosticados genéticamente con deleción paternal del cromosoma 15q11-13 y en el 60% de los pacientes con disomía uniparental materna del cromosoma 15. También se resalta que, aunque los rasgos faciales típicos en SPW se vuelven más evidentes en pacientes mayores, el rendimiento obtenido con la aplicación fue homogéneamente satisfactorio en todo el rango de edades probado (1–15 años) [17]. En este proyecto se usaron los datos genéticos para realizar el etiquetado de los datos, y las técnicas de *ML* fue exclusivamente la aplicación de *Face2Gene,* que usa *deep learning*, para predecir los síndromes a los que se podría asociar los rasgos faciales.

### *3.2.6. GENERATION OF HYPOTHALAMIC ARCUATE ORGANOIDS FROM HUMAN INDUCED PLURIPOTENT STEM CELLS*

En una investigación realizada por Huang et al [18] se desarrolló un método para generar organoides tridimensionales como sistema experimental para caracterizar la heterogeneidad molecular detallada que subyace a subpoblaciones específicas de células hipotalámicas. El método se usó para generar organoides arqueados hipotalámicos a partir de células madre pluripotentes inducidas por humanos, las cuales exhiben diversidad de subtipos neuronales y firmas moleculares del núcleo arqueado humano que pueden usarse para modelar el síndrome de Prader-Willi. Los investigadores exploraron organoides arqueados generados a partir de células madre de pacientes con el PWS, encontrando que dichos organoides exhibían desregulación transcriptómica similar al hipotálamo posnatal de pacientes con PWS.

Con el fin de analizar con mayor detalle las similitudes en la desregulación de la expresión génica encontrada en los organoides arqueados modelados y el hipotálamo de lo pacientes, los investigadores compararon conjuntos de datos publicados de pacientes con y sin el PWS hallando una superposición de 1.867 genes regulados negativamente y 158 genes regulados positivamente.

Los investigadores destacan la importancia de sus modelos para revelar las causas de trastornos genéticos, afirmando que la superposición de la expresión génica desregulada encontrada sugiere que las firmas de la enfermedad a nivel transcripcional pueden conservarse hasta cierto punto desde el desarrollo embrionario hasta el postnatal.

### *3.2.7. A MACHINE LEARNING PIPELINE FOR QUANTITATIVE PHENOTYPE PREDICTION FROM GENOTYPE DATA*

El uso de modelos de *ML* para la predicción cuantitativa de datos fenotipo a partir de datos genotipo ha sido documentada por Guzzetta et al [19]. En el conjunto de elementos de procesamiento de datos conectados en serie - comúnmente denominado “pipeline” - propuesto por los autores, se resalta la importancia de la aplicación de protocolos de análisis de datos bien documentados para controlar las fuentes de variabilidad y garantizar reproducibilidad de los resultados. Al respecto, los autores indican que el uso de estos protocolos se complementa de manera adecuada con la selección óptima de características en un entorno multivariado como el que se da en los estudios de asociación del genoma completo.

Los autores aplicaron el pipeline de *ML* propuesto en el problema de adaptar rasgos fenotípicos complejos de ratones heterogéneos a partir de polimorfimos de un solo nucleótido. El elemento principal del pipeline fue el método de regularización L1L2, el cual proporciona simultáneamente un modelo de regresión y un procedimiento de reducción de dimensionalidad adecuado para características correlacionadas. El método mostró su efectividad en la selección de marcadores y en la precisión de las predicciones. El estudio indicó que las técnicas de ML son adecuadas en la predicción cuantitativa del fenotipo, siempre que protocolos de análisis sean empleados para evitar el sesgo en la selección del modelo.

# 4. PREPARACIÓN DEL CONJUNTO DE DATOS

En esta sección, se detallan las etapas clave para la adquisición, evaluación y transformación de los datos utilizados en este proyecto. Desde la recopilación de información genética y epigenética de individuos con Síndrome de Prader-Willi (SPW) hasta el preprocesamiento necesario para asegurar la coherencia y precisión de los datos, cada actividad se ha diseñado para optimizar la fiabilidad del análisis posterior.

## 4.1 RECOPILACIÓN DE DATOS

Para el proyecto la información se recopiló de dos fuentes de datos. El conjunto de datos inicial corresponde a perfiles de metilación en individuos con disomía uniparental en regiones metiladas del cromosoma 15 [20]. Los datos se subdividen en 3 pacientes con disomía uniparental materna, y otros 3 pacientes con disomía uniparental paterna, y clasificados a su vez en dos posibles categorías: a) pacientes con Síndrome de Prader-Willi y b) pacientes con Síndrome de Angelman (SA). Adicionalmente, para cada paciente, se tienen muestras en duplicada para cada ADN. El conjunto de datos se encuentra almacenado en la plataforma Gene Expression Omnibus (GEO), producto de un trabajo de investigación externo y de acceso libre, de esta forma también se asegura su representatividad y calidad.

La segunda fuente de datos corresponde a una serie de archivos de texto con datos de anotación y variación genética. Esta base de datos proporciona datos genómicos diversos, incluidos polimorfismos de un solo nucleótido (SNP), variantes estructurales, anotaciones genéticas, elementos repetitivos y elementos potenciadores, todos los cuales se utilizan para anotar y analizar características genómicas. A continuación, se describe cada uno de los archivos de variación genética:

- dbSNP155: dbSNP es un recurso integral para polimorfismos de un solo nucleótido (SNP) y otros tipos de variación genética. Este archivo contiene información sobre variantes genómicas, específicamente SNP, y se utiliza para vincular variaciones genéticas con sus consecuencias biológicas, como la asociación con enfermedades o la variabilidad de rasgos [21].

- dbVarCommon: dbVar es una base de datos de variaciones estructurales genómicas humanas, que incluye inserciones, deleciones, inversiones y duplicaciones de grandes secciones del genoma. El archivo dbVarCommon rastrea específicamente las variantes estructurales comunes, lo que ayuda al estudio de la diversidad genética y el riesgo de enfermedades [22].

- geneid: Este archivo contiene identificadores de genes (como GeneID), que asignan regiones genómicas a genes específicos. Es útil para asociar variantes genéticas con genes funcionales y sus funciones conocidas en las vías biológicas, lo que proporciona una base para estudios genómicos funcionales.

- Omim: OMIM (por sus siglas en inglés: Online Mendelian Inheritance in Man) es una base de datos completa de genes y trastornos genéticos humanos. El archivo Omim conecta genes y regiones

genéticas con rasgos y trastornos fenotípicos conocidos, lo que permite identificar contribuyentes genéticos a enfermedades hereditarias [23].

- RefSeq se utiliza para anotar genes y proteínas, lo que proporciona un marco estandarizado para identificar genes, exones y regiones codificantes dentro del genoma [24].

- Repeticiones: El archivo Repeats contiene información sobre elementos repetitivos en el genoma, como SINE, LINE, satélites y otras repeticiones de secuencias. Estos elementos son esenciales para comprender la estructura, la evolución y la regulación del genoma, ya que a menudo desempeñan funciones en los reordenamientos cromosómicos y el control de la expresión génica [25].

- Duplicaciones segmentarias: El archivo SegmentalDups documenta segmentos grandes y duplicados del genoma, conocidos como duplicaciones segmentarias. Estas duplicaciones son cruciales para estudiar los reordenamientos genómicos y sus efectos sobre la diversidad genética, la evolución y los mecanismos de las enfermedades [25].

- SimpleRepeats: Este archivo cataloga secuencias cortas de repetición en tándem, a menudo llamadas microsatélites. Las repeticiones simples son altamente polimórficas y son valiosas para el mapeo genético, los estudios evolutivos y la ciencia forense, ya que varían mucho entre individuos [26].

- VistaEnhancers: Este archivo proporciona información sobre los elementos reguladores cis (potenciadores) dentro del genoma. Estos potenciadores regulan la expresión genética y se prueban para determinar su actividad específica en los tejidos en modelos de ratones transgénicos. Este archivo ayuda a vincular las regiones genómicas con sus efectos reguladores, lo que es crucial para comprender la regulación genética y los procesos de desarrollo [27].

## 4.2 EVALUACIÓN Y PREPROCESAMIENTO DE DATOS

En este apartado se describen los procedimientos realizados para la estimación de la etiqueta o nivel de significancia, la definición de los rangos de exploración a partir de los marcadores en los datos de perfiles de metilación, la preparación y construcción de los datos crudos para las matrices general y detallada, y el análisis exploratorio de dichas matrices.

### 4.2.1 ESTIMACIÓN DE LA ETIQUETA

El uso de la plataforma GEO, permite una exploración de los datos, ya que la misma plataforma cuenta con un analizador ejecutado en lenguaje R, el cual identifica los genes expresados que son diferenciados en las condiciones experimentales. Es decir, en este caso, identifica los genes que son significativamente diferentes entre los pacientes con SPW y SA. La identificación de los genes diferenciadores se logra por medio de los niveles de metilación, los cuales son cuantificados usando las proporciones de fluorescencia $\log_2$. Estos niveles de metilación fueron normalizados por cuantiles y posteriormente se trataron para eliminar los valores atípicos, siendo reemplazados por la intensidad media de los vecinos metilados.

El procesamiento de los datos consistió en la normalización por cuantiles a los datos de expresión, para que todas las muestras tengan una distribución de valores idéntica. Posteriormente, se aplicó un modelado de la varianza a nivel de observación (*vooma*, en inglés) para eliminar el efecto de que los genes con diferentes niveles de expresión pueden tener diferentes niveles de variabilidad. Así, al tomar en cuenta las diferencias en la varianza de los datos, los resultados son menos susceptibles a errores derivados de genes con mayor varianza.

Los resultados de significancia (*p-value*) fueron ajustados para corregir la ocurrencia de falsos positivos, usando el método de Bemjamini & Hochberg, el cual corrige estas ocurrencias sin ser tan estricto para mantener un volumen de datos aceptable, permitiendo identificar significancias aceptando la aparición de algunos falsos positivos. Finalmente, el nivel de significancia ajustado para identificar las secuencias genéticas significativamente diferentes se estableció en 0.1 para obtener un total de 1,266 genes expresados de forma diferencial, señalados en azul en la Figura 2.

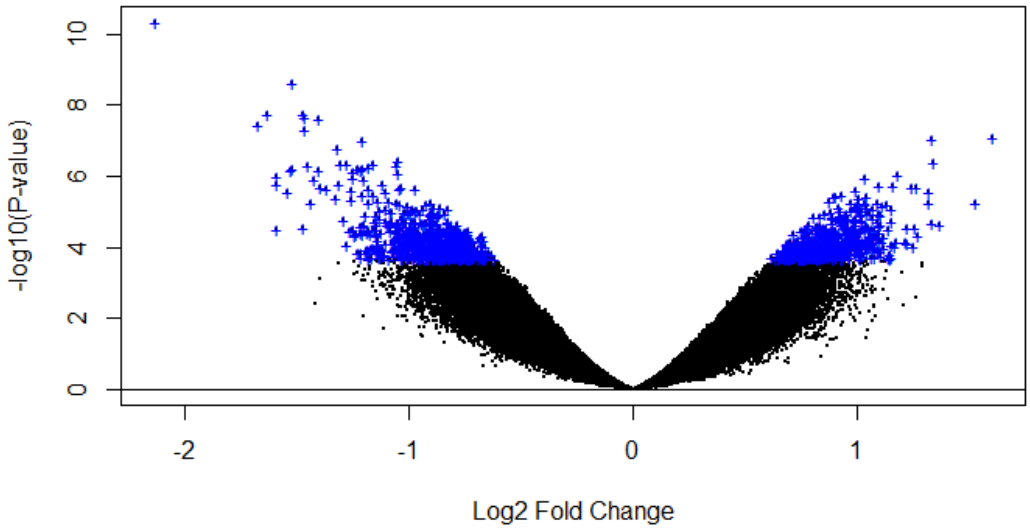


Figura 2: Identificación de marcadores genéticos con expresión diferencial.

Nota: La figura presenta la distribución (*volcano plot*) de los marcadores genéticos con expresión diferencial (color azul), posterior a la cuantificación de las proporciones de fluorescencia $\log_2$.

Teniendo en cuenta la reducida cantidad de genes expresados de forma diferencial, se optó mantener tanto el *p-value* ajustado (*p-value adjusted*) como el *p-value* original (*p-value raw*), ya que una de las limitaciones de los modelos de *ML* es el número de datos con los que se pueda entrenar un modelo, obteniendo dos tipos de categoría de *p value* posibles (Tabla 3).

Tabla 3: Cantidad de genes significativamente diferenciados según p-value

| **Categoría *p-value*** | **Genes diferenciados** | **Porcentaje (%)** |
|---|---|---|
| ***Raw*** | 163,999 | 26.7 |
| ***Adjusted*** | 1,266 | 0.2 |

Nota: La tabla presenta la cantidad de genes diferenciados, según el tipo de *p value* usado, como *raw* (sin ajuste estadístico) y ad*justed* (con ajuste estadístico), así como el porcentaje respecto al total de genes evaluados.

### 4.2.2 DEFINICIÓN DE RANGOS DE EXPLORACIÓN A PARTIR DE LOS MARCADORES

En la fuente de datos de perfiles de metilación en individuos con disomía uniparental en el cromosoma 15, se tienen 612834 identificadores de secuencia de ADN. Estos identificadores, tienen una estructura del tipo 'CHR15P018260026', la cual indica una ubicación específica en el cromosoma. Para estas secuencias se tiene la letra 'P' que denota el brazo corto del cromosoma con la ubicación (18260026), número que permite definir los rangos de exploración para el cruce de información con la serie de archivos de texto de datos de anotación y variación genética. El análisis de las distancias entre marcadores nos indica que la distancia media es de 133.9, la distancia máxima es de 108,408 y la distancia mínima es de 75 nucleótidos. Al tener distancias inconsistentes entre marcadores se genera un reto para la definición de los rangos de exploración, ya que al abordarse un rango fijo da como resultado que marcadores cercanos tengan rangos superpuestos, lo cual daría como consecuencia que se incrementara la presencia de variables en determinados rangos y se introdujeran datos redundantes.

Para abordar la problemática de las distancias inconsistentes entre marcadores, se implementó un rango dinámico, en el cual los límites de rango entre marcadores consecutivos eran acordes a su distancia. Para esto, el final de cada rango se determinó calculando el punto medio entre el marcador actual y el siguiente. De esta manera, se aseguró que los rangos se ajustaran a la distribución genómica real de los marcadores. Cuando la distancia entre marcadores consecutivos excedía los 2,000 nucleótidos, el rango se limitó a 2,000, evitando que los rangos se volvieran demasiado grandes, lo cual hubiera podido diluir la información relevante y hacer que el análisis hubiera sido menos preciso. Para los marcadores vecinos ubicados en los limites inferior y superior, el inicio del rango se ajustó para que comenzara un nucleótido después de que terminara el rango anterior, de manera que se garantizara una cobertura continua que no se superpusiera. Con este procedimiento, se garantizó que los rangos de exploración se distribuyeran de manera uniforme según la ubicación de los marcadores, sin que existiera superposición entre rangos y con rangos no superiores a 2,000 nucleótidos. La media para todos los rangos fue de 132.9np, el rango mínimo fue de 42 np y el máximo fue de 2,000 np (Figura 3).

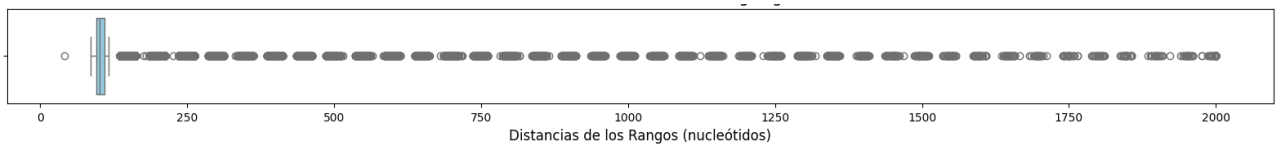


Figura 3: Distribución de las distancias de los rangos genómicos de exploración.

Nota: La figura presenta la distribución de las distancias entre los identificadores de secuencia de ADN de los, medida en número de nucleótidos (moléculas orgánicas que forman los ácidos nucleicos). Se observa una distribución no homogénea.

### 4.2.3 CONSTRUCCIÓN Y PREPARACIÓN DE LOS DATOS

#### 4.2.3.1 CONSTRUCCIÓN DE LAS MATRICES

Antes de avanzar hacia la construcción del modelo de *ML* es necesario una etapa de preparación de los datos, y en este caso se hizo necesario también construir una capa de datos con la cual

abordar el problema de clasificación planteado, como se observa en el diagrama de flujo de la Figura 4. La etiqueta de expresión se definió a partir del nivel de significancia (*p-value raw)* aplicando el procedimiento estadístico descrito en el numeral 4.2.1 sobre los perfiles de metilación. El conjunto de datos de rangos de exploración se construyó según lo descrito en el apartado 4.2.2. Con estos insumos se construyeron dos matrices, una para análisis general de los datos de anotación y variación genética y una segunda matriz con el detalle de los atributos presentes en los archivos con variables genéticas.

Para la categorización de los niveles de significancia se utilizó la siguiente lógica:

*p < 0.01*: *Highly Significant*

*0.01 ≤ p < 0.05*: *Significant*

*0.05 ≤ p < 0.1*: *Moderately Significant*

*p ≥ 0.1*: *Non-Significant*

A la categorización de los niveles de significancia se añadió una columna binaria para cada archivo de variación genética, es decir, a cada archivo de la base de datos de anotaciones y variaciones genéticas se le asignó una columna, en la cual se determinó la presencia de la variación genética con un valor de '1' cuando al cruzar con los rangos de exploración usando el marcador genético de inicio, al menos un atributo tuviera un valor no nulo, en caso contrario se asignaba el valor '0'.

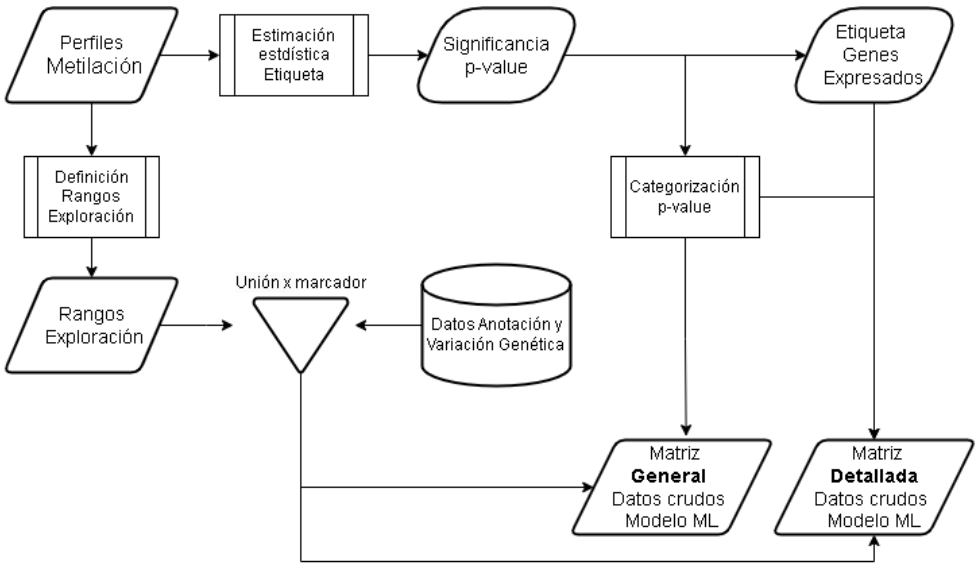


Figura 4: Diagrama de flujo del proceso usado para la construcción de la capa de datos crudos de los modelos de clasificación de ML.

Nota: La figura presenta el diagrama de flujo seguido en la construcción de la matriz de datos usada para implementar los modelos de *ML*.

El proceso general de incorporación de atributos de la base de datos de anotaciones y variaciones genéticas consistió en tres pasos: el filtrado por rangos, la función de agregación y la unión en la

matriz consolidada. En el primer paso, para cada archivo se seleccionaron los registros donde el marcador inicial se encontraba en los rangos de exploración. Para definir la función de agregación se usó un análisis exploratorio básico de los atributos de la base de datos, sea esta suma o promedio, aplicada a los registros filtrados en el paso anterior. Para la unión de los resultados agregados, se realizó la incorporación de nuevas columnas a la matriz consolidada para enriquecerla con la información genética significativa de cada fuente. En la siguiente tabla, se realiza una descripción de los campos utilizados de cada archivo de la base de datos.

Tabla 4: Atributos de los archivos de variación y anotación genética con su descripción y función de agregación usada.

| Archivo Variación Genética | Atributo Numérico | Función de Agregación | Descripción |
|---|---|---|---|
| **GeneID** | exonCountGeneId | Suma | Conteo de exones en el archivo GeneID |
| | GeneId | Binario (0/1) | Indica la presencia de exonCountGeneId |
| **Omim** | omimCount | Conteo | Cuenta las coincidencias donde 'chromStart' está dentro de algún rango |
| **RefSeq** | exonCount | Suma | Conteo de exones en el archivo RefSeq |
| | RefSeq | Binario (0/1) | Indica la presencia de exonCount |
| **SegmentalDups** | otherSize | Suma | Tamaño de otras duplicaciones |
| | posBasesHit | Suma | Número de bases afectadas en la alineación |
| | indelN | Suma | Número de inserciones/eliminaciones (N) |
| | indelS | Suma | Número de inserciones/eliminaciones (S) |
| | alignL | Suma | Longitud de la secuencia alineada |
| | alignB | Suma | Bases alineadas |
| | matchB | Suma | Bases coincidentes |
| | mismatchB | Suma | Bases no coincidentes |
| | transitionsB | Suma | Número de transiciones |
| | transversionsB | Suma | Número de transversiones |
| | fracMatch | Promedio | Número de transversiones en bases alineadas |
| | fracMatchIndel | Promedio | Proporción de bases coincidentes |
| | jck | Promedio | Atributo específico de duplicación |
| | k2k | Promedio | Atributo específico de duplicación |
| **SimpleRepeats** | period | Suma | Longitud de la secuencia repetida |
| | copyNum | Suma | Número de veces que se repite la secuencia |
| | consensusSize | Suma | Tamaño de la secuencia consenso |
| | perMatch | Promedio | Porcentaje de coincidencia entre secuencias |

| | | | |
|---|---|---|---|
| | perIndel | Promedio | Porcentaje de inserciones/eliminaciones |
| | score | Suma | Puntaje de la fuerza de la repetición |
| **dbSNP155** | altCount | Suma | Número de alelos alternativos |
| | freqSourceCount | Suma | Número de fuentes que reportan frecuencias |
| **dbVarCommon** | score | Promedio | Puntaje de confianza para las variaciones |
| **VistaEnhancers** | score | Suma | Puntaje basado en la actividad del potenciador |
| | experimentId | Conteo | Identificador para el experimento que prueba el potenciador |
| **Repeats** | swScore | Suma | Puntuación de alineación para la región repetida |
| | milliDiv | Suma | Divergencia porcentual respecto del consenso |
| | milliDel | Suma | Divergencia porcentual de bases eliminadas respecto al consenso |
| | milliIns | Suma | Divergencia porcentual de bases insertadas respecto al consenso |

Nota: La tabla relaciona los atributos numéricos como *child dependency* de una variación genética definida, así como la definición biológica de cada atributo.

De esta manera se consolidó la base de datos con la que se construyó posteriormente el modelo de *ML*, usando como variable independiente la expresión genética, categorizada en 3 clases: sobreexpresada, subexpresada o no diferencial, y por otro lado las variables presumiblemente explicativas, provenientes del cruce entre los perfiles de metilación con las bases de datos de anotación y variación genética.

#### 4.2.3.2 ESTRUCTURA FINAL DE LA BASE DE DATOS

La estructura final de la base de datos se conformó con los campos de la variable objetivo o etiqueta y todos los atributos de anotación o variación genética. Para la variable objetivo se tiene el nivel de significancia (*p-value raw*), la identificación de genes significativamente diferenciados y la categorización de los niveles de significancia, cuyos procedimientos se explicaron en los apartados 4.2.1 y 4.2.3. Para la conformación de los atributos se realizó una unión de cada uno de los archivos de variación genética con los perfiles de metilación, de manera que se crearon dos grupos de atributos según lo explicado en la sección 4.2.3. El primero consistió en la identificación binaria de los archivos de variaciones genéticas, quedando de esta manera 9 campos, cada uno de ellos asociado a la presencia binaria de la variación genética según la unión realizada con los perfiles de metilación a través de los rangos genéticos de exploración explicados en la sección 4.2.2. Con la identificación de los atributos genéticos en los rangos de exploración se conformó el segundo grupo de atributos, los cuales consistieron en los atributos numéricos presentes en cada archivo genético estimados con una determinada función de agregación según el caso (ver Tabla 4). A continuación, se detallan los campos de la variable objetivo y de los atributos genéticos conformados en la estructura final de la base de datos.

**Campos de la variable objetivo en estructura final**

A continuación, se describen los campos utilizados:

Tabla 5: Estructura de variable objetivo usadas en la construcción de la matriz de datos.

| Campo | Descripción | Tipo de dato y rango |
|---|---|---|
| **p-value raw** | Valor de significancia estadística obtenido en análisis de expresión diferencial de genes, sin ningún tipo de ajuste. | float (0 a 1) |
| **Identificación de genes significativamente diferenciados** | A partir del *p-value adjusted* se asignó una etiqueta a cada gen basado en su nivel de expresión en comparación con la condición control: | integer (-1, 0, 1) |
| | **-1:** Gen subexpresado | |
| | **0:** Gen no significativo | |
| | **1:** Gen sobreexpresado | |
| **Categorización de la significancia estadística** | Clasificación del nivel de significancia de acuerdo con el *p-value raw*: | string ("Highly Significant", "Significant", "Moderately Significant", "Non-Significant") |
| | **Highly Significant:** $p < 0.01$ | |
| | **Significant:** $0.01 \leq p < 0.05$ | |
| | **Moderately Significant:** $0.05 \leq p < 0.1$ | |
| | **Non-Significant:** $p \geq 0.1$ | |

Nota: La tabla presenta los distintos tipos de variable objetivo usadas en la construcción de la matriz de datos, así como los valores que tomará.

**Campos de los atributos en estructura final**

El primer grupo de atributos consistió en la identificación binaria de los archivos de anotaciones y variaciones genéticas a partir de la unión con los perfiles de metilación y los rangos de exploración genética. A continuación, se presenta la descripción de los campos.

Tabla 6: Atributos binarios obtenidos de las variaciones genéticas y su unión con los perfiles de metilación.

| Atributo o Variación Genética | Tipo de Dato |
|---|---|
| GeneID | Binario (1,0) |
| Omim | Binario (1,0) |
| RefSeq | Binario (1,0) |
| SegmentalDups | Binario (1,0) |
| SimpleRepeats | Binario (1,0) |
| dbSNP155 | Binario (1,0) |
| dbVarCommon | Binario (1,0) |
| VistaEnhancers | Binario (1,0) |

| | |
|---|---|
| Repeats | Binario (1,0) |

Nota: La tabla presenta los distintos tipos de variación genética usados para la construcción de la matriz de datos general, así como el tipo de dato usado para representar la presencia (1) o ausencia (0), después de la unión con los perfiles de metilación.

El segundo grupo de campos consistió en la aplicación de una función de agregación a cada uno de los campos numéricos presente en cada archivo de variación genética. A continuación, se presenta la descripción de los atributos numéricos.

Tabla 7: Atributos genéticos numéricos obtenidos para cada fuente de variación genética, su tipo y función de agregación utilizada.

| Variación Genética Fuente | Atributo Numérico estimado | Función de Agregación | Tipo de dato |
|---|---|---|---|
| **GeneID** | exonCountGeneId | Suma | Flotante |
| **Omim** | omimCount | Conteo | Entero |
| **RefSeq** | exonCount | Suma | Flotante |
| **SegmentalDups** | otherSize | Suma | Flotante |
| | posBasesHit | Suma | Flotante |
| | indelN | Suma | Flotante |
| | indelS | Suma | Flotante |
| | alignL | Suma | Flotante |
| | alignB | Suma | Flotante |
| | matchB | Suma | Flotante |
| | mismatchB | Suma | Flotante |
| | transitionsB | Suma | Flotante |
| | transversionsB | Suma | Flotante |
| | fracMatch | Promedio | Flotante |
| | fracMatchIndel | Promedio | Flotante |
| | jck | Promedio | Flotante |
| | k2k | Promedio | Flotante |
| **SimpleRepeats** | period | Suma | Flotante |
| | copyNum | Suma | Flotante |
| | consensusSize | Suma | Flotante |
| | perMatch | Promedio | Flotante |
| | perIndel | Promedio | Flotante |
| | score | Suma | Flotante |
| **dbSNP155** | altCount | Suma | Flotante |
| | freqSourceCount | Suma | Flotante |
| **dbVarCommon** | score | Promedio | Flotante |
| **VistaEnhancers** | score | Suma | Flotante |
| | experimentId | Conteo | Flotante |
| **Repeats** | swScore | Suma | Flotante |

| | | | |
|---|---|---|---|
| | milliDiv | Suma | Flotante |
| | milliDel | Suma | Flotante |
| | milliIns | Suma | Flotante |

Nota: La tabla presenta todos los atributos usados en la matriz de datos detallada, así como el tipo de función de agregación usada para su estimación. Además, cada atributo es un *child dependency* de una variación genética definida.

**Construcción de matriz general y matriz detallada**

En la construcción de la matriz general se utilizó para el análisis de la variable objetivo el campo de Categorización de la significancia estadística (ver Tabla 5) y los atributos binarios de variaciones genéticas detallados en la Tabla 6. La matriz detallada se construyó para el análisis de la variable objetivo con el campo de Identificación de genes significativamente diferenciados (ver Tabla 5) y los atributos numéricos genéticos descritos en la Tabla 7.

### 4.2.4 ANÁLISIS EXPLORATORIO DE DATOS DE LAS MATRICES

Para el análisis exploratorio de las matrices de datos crudos general y detallada se utilizó la librería de Python 'Autoviz', con la cual es posible generar gráficas para cualquier tipo y tamaño de conjunto de datos considerando características y etiqueta [28].

#### 4.2.4.1 ANÁLISIS EXPLORATORIO DE DATOS DE LA MATRIZ GENERAL

En la Figura 5 se presenta la distribución de la etiqueta o nivel de significancia *p* estimado según procedimiento descrito en numeral 4.2.1.

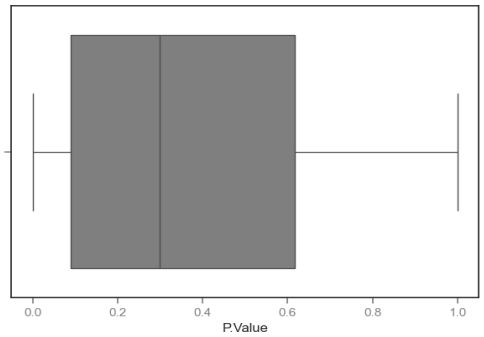


Figura 5: Distribución del nivel de significancia *p-value raw*.

Nota: La figura presenta la distribución por *boxplot* del *p value* sin ningún ajuste estadístico adicional. El 50% de los datos se ubica en aproximadamente 0.3.

En el análisis exploratorio se pudo determinar que el único archivo de variación o anotación genética que no presentó datos al cruzarse con los rangos de exploración fue "dbVarCommon", las demás variaciones genéticas tuvieron presencia de al menos uno de sus atributos, siendo la variación genética "SNP155" la que tuvo mayor participación en los rangos definidos a partir de los marcadores iniciales. La mayoría de las variaciones genéticas estuvieron presente en los rangos de exploración en una proporción muy pequeña respecto a la cantidad de marcadores, razón por la

cual en la Figura 6 la distribución de estas variables pareciera no presentar valores en ‘1’, dado esto también por la escala de la gráfica.

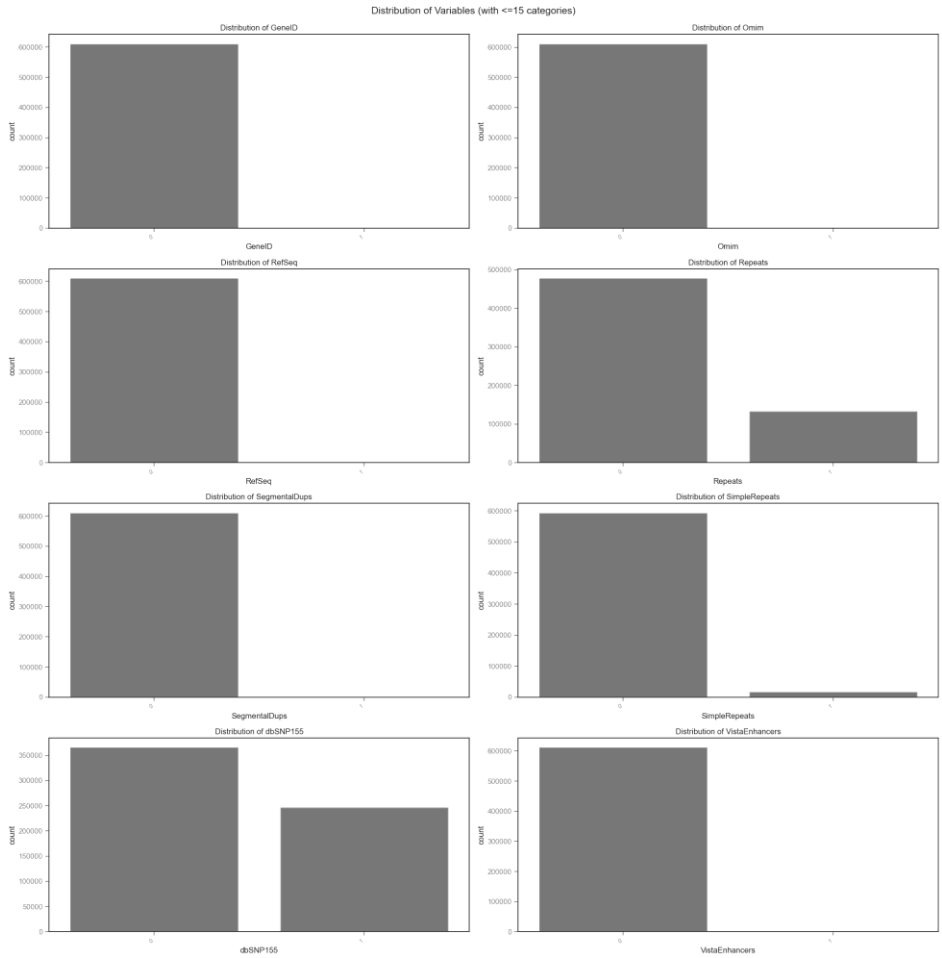


Figura 6: Distribución de las variables de anotación o variación genética posterior al cruce con los rangos de exploración.

Nota: La grafica muestra la distribución para 8 variables de variación genética, según la presencia, o no, de esa variable en los rangos de exploración de las regiones cromosómicas. Se observa poca presencia (1) en la mayoría de las variables, en comparación con la ausencia (0).

En este análisis exploratorio también se examinó la correlación existente entre todas las variables de variación genética y la variable objetivo de nivel de significancia, encontrando que ninguno de los archivos de anotación genética tiene correlación estadística con el *p-value adjusted* (ver Figura 7).

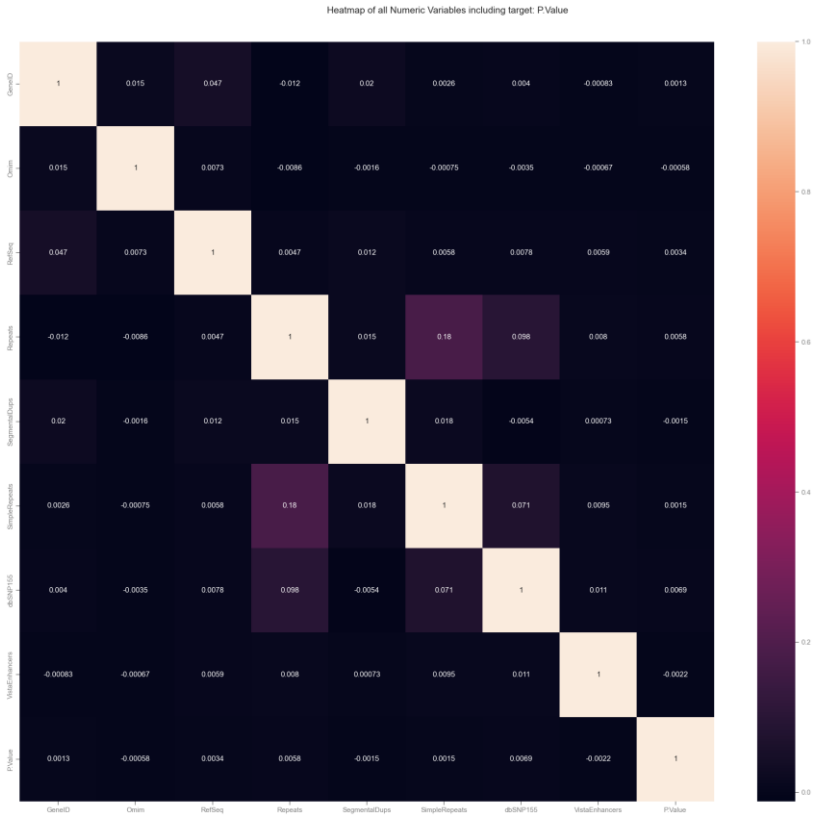


Figura 7: Mapa de calor de las correlaciones de las variables de variación genética y la variable objetivo *p-value adjusted* en la matriz general.

Nota: La figura presenta las correlaciones entre las variables de variación genética y la variable objetivo. No se encuentran correlaciones fuertes (colores más claros) ni con la variable objetiva ni entre variables de variación genética.

### 4.2.4.2 ANÁLISIS EXPLORATORIO DE DATOS DE LA MATRIZ DETALLADA

La exploración de los datos de la matriz detallada indica para la variable objetiva o etiqueta la siguiente distribución: genes con no significancia etiqueta ‘0’ presente en 611,568 marcadores, genes sub expresados etiqueta ‘-1’ en 718 marcadores y genes sobre expresados etiqueta ‘1’ en 548 registros. Por lo anterior, en todos los atributos de las variaciones o anotaciones genéticas se presenta que, para los más de 610 mil registros la presencia de genes expresados y sobre expresados es considerablemente menor que los registros con genes sin significancia o valor 0. Esto quiere decir que las gráficas de distribución de cualquiera de los atributos para representar la cantidad de registros con genes sub expresados, sobre expresados y sin significancia mostraría, por la cantidad de registros, sólo a este último. A continuación, en la Figura 8, se presentan algunos ejemplos de distribución separando la categoría dominante (‘expression’ = 0) de las otras, de manera que se pueda visualizar la distribución de las etiquetas menos frecuentes.

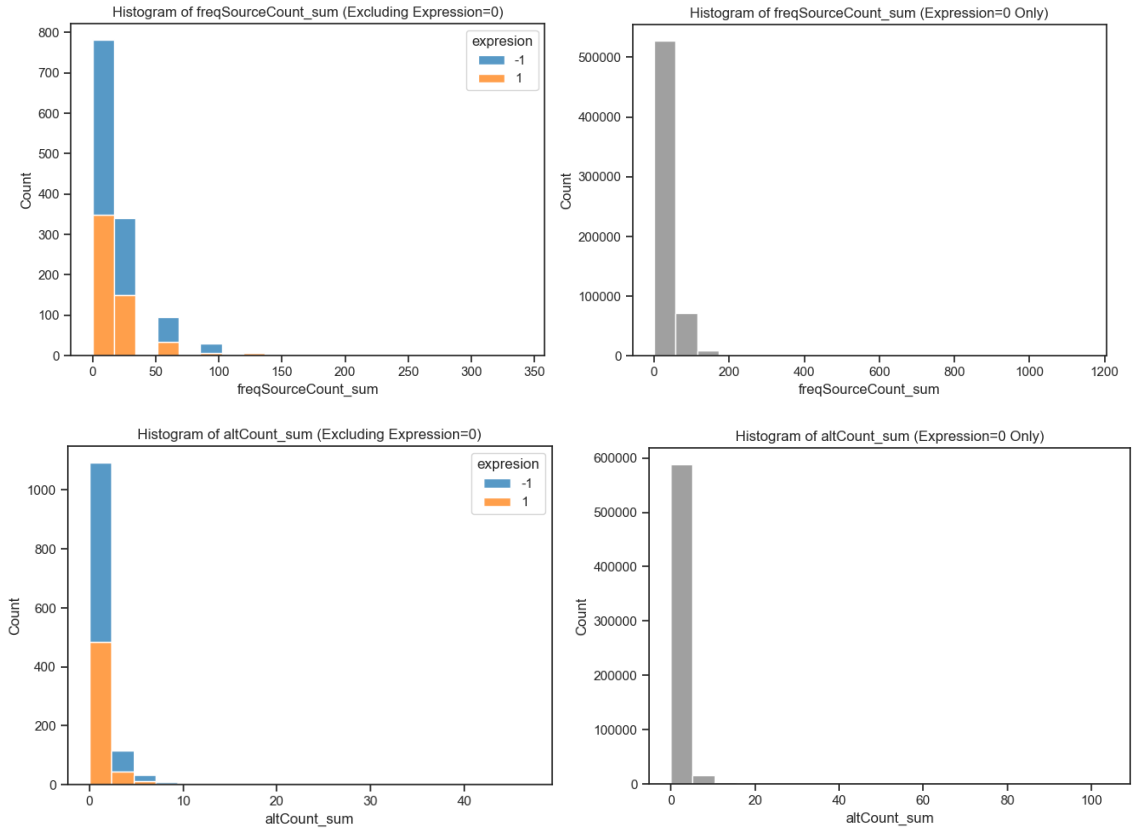


Figura 8: Distribución de los datos de atributos excluyendo genes no significantes para los atributos genéticos de 'freqSourceCount_sum' y 'altCount_sum'.

Nota: La figura presenta al lado izquierdo los datos de atributos comparando sub expresados con sobre expresados, y al lado derecho se presentan solamente los datos no significativos en la expresión del gen.

Al realizarse el análisis exploratorio de la correlación existente entre todos los atributos de las variaciones genéticas y la variable objetivo de genes expresados, se puede apreciar que ninguno de los atributos tiene correlación estadística con la etiqueta o campo 'expresion' (ver Figura 9). En la gráfica de mapa de calor de correlaciones (Figura 9) se visualiza que, aunque no existe una correlación que sea relevante entre variable objetivo y ninguno de los atributos, si existe correlación estadística entre atributos, lo cual indica la presencia de campos redundantes que deben ser eliminados en la construcción de los conjuntos de datos de entrenamiento y prueba de los modelos de clasificación.

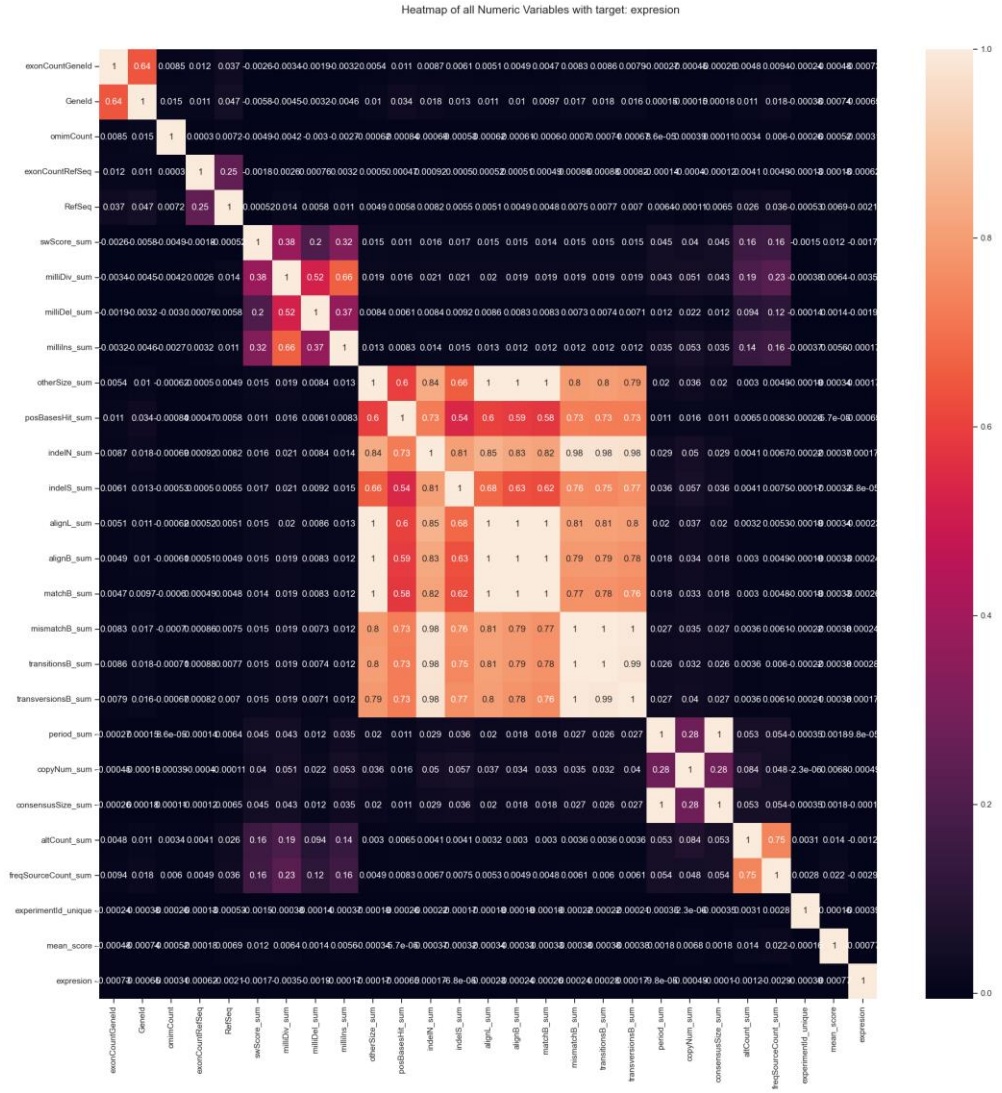


Figura 9: Mapa de calor de las correlaciones de los atributos genéticos y la variable objetivo '*expresion*' en la matriz detallada.

Nota: La figura presenta las correlaciones entre los atributos genéticos y la variable objetivo, usando la matriz de datos detallada. Se presentan datos altamente correlacionados (colores más claros) hacía los atributos ubicados en la mitad de la matriz de correlación.

# 5. CONSTRUCCIÓN DE LOS MODELOS DE ML

En este capítulo se describen los pasos efectuados para la construcción de tres modelos de clasificación de *ML* con el fin de analizar los datos genéticos y epigenéticos relacionados con el SPW. El primer modelo de *ML* es el general, el cual tiene el propósito de identificar asociaciones entre las anotaciones o variaciones genéticas binarias y las categorías de significancia de los *p-value*. Un segundo modelo de *ML* detallado a nivel de atributos numéricos de las variaciones genéticas, con el fin de explorar la relación de estos atributos con las categorías de significancia. El tercer modelo detallado también respecto a los atributos numéricos de las variaciones genéticas, pero con la etiqueta definida en esta oportunidad según los niveles de expresión genética. Para los tres modelos los algoritmos de clasificación explorados fueron: *Random Forest*, *Multilayer Perceptron* y *XGBoost*. La evaluación e interpretación de los resultados de la implementación de los modelos es motivo de discusión de los siguientes capítulos.

## 5.1 MODELO GENERAL DE ML

El modelo general de *ML* se construyó a partir de la matriz general, cuyo procedimiento está descrito en el apartado 4.2.3.1. El insumo para la definición de las características o atributos de la matriz general fue la identificación binaria de la variación genética, determinada esta según los rangos de exploración definidos a partir de los perfiles de metilación. La etiqueta se definió categorizando los valores de significancia *p* según lo descrito en el mismo apartado.

El procesamiento de datos para la implementación del modelo de *ML* general consistió en la identificación de campos no significativos, la categorización de la etiqueta, el balanceo de datos, la división del conjunto de datos en conjuntos de prueba y entrenamiento, y la búsqueda de la mejor combinación de hiperparámetros.

### Identificación de campos no significativos

Tabla 8: Descripción estadística de los atributos de variación genética presentes en los rangos de exploración de los perfiles de metilación.

| | P.Value | GeneID | Omim | RefSeq | Repeats | SegmentalDups | SimpleRepeats | dbSNP155 | VistaEnhancers |
|---|---|---|---|---|---|---|---|---|---|
| Conteo | 612,834 | 612,834 | 612,834 | 612,834 | 612,834 | 612,834 | 612,834 | 612,834 | 612,834 |
| Promedio | 0.369 | 0.002 | 0.001 | 0.003 | 0.219 | 0.002 | 0.030 | 0.403 | 0.000 |
| Desviación Estándar | 0.304 | 0.042 | 0.034 | 0.058 | 0.414 | 0.048 | 0.170 | 0.490 | 0.020 |
| Mínimo | 4.41E-11 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Q25 | 8.91E-02 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Q50 | 3.00E-01 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Q75 | 6.18E-01 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| Máximo | 9.99E-01 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |

Nota: La tabla presenta los estadísticos básicos para las variables cuantitativas de los perfiles de metilación.

Del análisis exploratorio de datos de la matriz general se identificó que, de los 9 archivos de variación o anotación genética, el único en el que no se presentó presencia en los rangos de exploración de los perfiles de metilación fue 'dbVarCommon'. Para los 8 atributos de variación

genética presentes en los rangos de exploración de los marcadores genéticos se obtuvieron medidas estadísticas básicas (Tabla 8).

**Categorización de la etiqueta**

En ejercicios preliminares, se comprobó que la poca representación de los genes significativamente expresados a partir del *p-value adjusted* generaba dificultades en la implementación de los modelos de *ML*, ya que aún con el uso de diferentes técnicas (reducción de dimensionalidad, diferentes algoritmos, balanceo, normalización), no era posible identificar patrones por la poca correlación entre las variables explicativas y la variable objetivo, así como el reducido volumen de datos de aprendizaje. A raíz de esto, se evaluó continuar los modelos asumiendo el *p-value raw* como variable objetivo. La categorización de la *p-value raw* se realizó siguiendo el procedimiento descrito en 4.2.3, obteniendo las 4 categorías presentadas en la Tabla 9.

Tabla 9: Cantidad de registros por categoría de la variable objetivo *p-value* y su porcentaje respecto al total de los datos.

| **Categoría *p-value*** | **Cantidad registros** | **Porcentaje** |
|---|---|---|
| *Non-Significant* | 448,835 | 73% |
| *Significant* | 71,285 | 12% |
| *Moderately-Significant* | 56,618 | 9% |
| *Highly-Significant* | 36,096 | 6% |

Nota: La tabla presenta la cantidad de registros en números absolutos y relativos, según la categoría definida del *p value*. Se obtiene un total de 27% de datos significativos.

**División de los datos en conjuntos de prueba y entrenamiento**

Teniendo en consideración el tamaño del conjunto de datos los modelos de *ML* se construyeron con dos alternativas de división para entrenamiento y prueba, la primera siendo 70-30% y la segunda con 80-20%.

**Balanceo de los datos**

Posterior a la división de los datos en conjuntos de prueba y entrenamiento, se realizó un balanceo de clases eliminando registros de las clases mayoritarias. Tomando como ejemplo la división del conjunto de datos en entrenamiento-prueba de 80-20%, el procedimiento buscó que todas las clases quedaran con el mismo número de registros de la clase minoritaria, la cual tenía en este caso 28842 registros, dejando en este escenario un total de 115,368 registros en el conjunto de datos.

**Búsqueda de la mejor combinación de hiperparámetros**

Para la definición de la mejor combinación de hiperparámetros se implementó la técnica de búsqueda por grilla en cada uno de los algoritmos de *ML*. El procedimiento consistió en la definición del espacio de búsqueda, creación del objeto para la búsqueda por grilla usando la función *GridSearchCV()*, adaptación de la grilla a los datos de entrenamiento y la obtención de los

mejores parámetros.

La búsqueda por grilla con la función *GridSearch()* se configuró en todos los casos de la siguiente manera:

```
gs = GridSearchCV(estimator = mlp_classifier, # El clasificador elegido
                  param_grid = param_grid, # la grilla de parámetros
                  scoring = 'accuracy', # la medida de desempeño
                  cv = 5, # para la búsqueda, el conjunto de entrenamiento se divide
                        # en dos: Entrenamiento y validación. cv indica cuántas
                        # veces se repetirá dicha división
                  n_jobs = -1,
                  verbose = 2,
                  error_score='raise') # Se incluye para rastrear errores
```

El parámetro 'estimator' fue el único que se modificaba según el modelo de clasificación que se estuviera validando y que se hubiera iniciado previamente.

**Algoritmo de clasificación *Random Forest***

En el modelo de clasificación *Random Forest* el espacio de búsqueda se definió de la siguiente manera:

```
param_grid = {
    'n_estimators': [10, 50, 100, 250],
    'max_depth': [5, 10, 20],
    'class_weight': [None]}
```

La mejor opción de hiperparámetros obtenida para el modelo de clasificación *Random Forest* fue:

{'max_depth': 10, 'n_estimators': 10}

**Algoritmo de clasificación *Multilayer Perceptron***

En el algoritmo de clasificación de *Multilayer Perceptron* el espacio de búsqueda se definió de la siguiente manera:

```
param_grid = {
    'hidden_layer_sizes': [(64,), (32, 32), (64, 32), (64, 64)],
    'activation': ['relu', 'tanh'],
    'alpha': [0.0001, 0.001, 0.01],
    'learning_rate_init': [0.001, 0.01, 0.1],
    'max_iter': [50, 100, 200]
    }
```

La mejor opción de hiperparámetros obtenida para el modelo de clasificación *Multilayer Perceptron* fue:

{'activation': 'tanh', 'alpha': 0.01, 'hidden_layer_sizes': (32, 32), 'learning_rate_init': 0.001, 'max_iter': 50}

**Algoritmo de clasificación *XGBoost***

En el modelo de clasificación *XGBoost* el espacio de búsqueda se definió de la siguiente manera:

```
param_grid = {
    'n_estimators': [50, 100, 200],          # Number of boosting rounds
    'max_depth': [3, 5, 7],                  # Maximum depth of trees
    'learning_rate': [0.01, 0.1, 0.2],       # Step size shrinkage
    'subsample': [0.6, 0.8, 1.0],            # Subsample ratio of training
instances
    'colsample_bytree': [0.6, 0.8, 1.0],     # Subsample ratio of columns
    'gamma': [0, 0.1, 0.2],                  # Minimum loss reduction
    'reg_alpha': [0, 0.01, 0.1],             # L1 regularization
    'reg_lambda': [1, 1.5, 2.0]              # L2 regularization
}
```

La mejor opción de hiperparámetros obtenida para el modelo de clasificación XGBoost fue:

{'colsample_bytree': 1.0, 'gamma': 0, 'learning_rate': 0.01, 'max_depth': 5, 'n_estimators': 50, 'reg_alpha': 0, 'reg_lambda': 1.5, 'subsample': 1.0}

## 5.2 MODELO DETALLADO DE ML CON VARIABLE OBJETIVO CATEGORIZADA

El modelo detallado de *ML* con variable objetivo categorizada se construyó a partir de la matriz general, cuyo procedimiento está descrito en el apartado 4.2.3. El insumo para la definición de las características de la matriz detallada fueron los atributos numéricos de las variaciones genéticas detallados en la Tabla 3 de la sección 4.2.3.1, determinados a partir de los rangos de exploración. La etiqueta se definió de manera similar al modelo general, categorizando los valores de significancia.

El procesamiento de datos para la implementación del modelo de *ML* detallado con variable objetivo categorizada consistió en la identificación de campos no significativos, la categorización de la etiqueta, el tratamiento de datos faltantes, el filtrado de datos considerando los registros categorizados con algún grado de significancia, la división del conjunto de datos en conjuntos de prueba y entrenamiento, el balanceo de datos, y la búsqueda de la mejor combinación de hiperparámetros. Todos los pasos a excepción de la identificación de atributos no significativos, el tratamiento de datos faltantes y el filtrado de datos considerando los registros categorizados con algún grado de significancia, se realizaron de manera similar a lo efectuado en el modelo general explicado en la sección anterior.

**Identificación de campos no significativos**

El número total de atributos numéricos de las variaciones genéticas que se encontraron a partir de los rangos de exploración fue de 33. En el análisis exploratorio de la sección 4.2.4.2 se pudo identificar en el Mapa de calor de correlaciones de la Figura 9 que una cantidad importante de atributos redundantes. Para la identificación de estos atributos se utilizó el siguiente procedimiento:

- Creación de matriz de correlación
- Búsqueda de valores mayores 0.9 o menores a -0.9
- Borrado de los registros nulos y armado de matriz con combinaciones de correlación encontradas según la condición anterior
- Creación de *clusters* o grupos con combinaciones que excedieran el umbral de 0.9
- Conversión de los grupos a listas con atributos únicos
- Selección de una característica representativa por grupo y borrado de los atributos redundantes manteniendo los representativos.

Con este procedimiento se redujeron los atributos a 22 campos asegurando que no se perdiera información.

**Tratamiento de datos faltantes**

Debido al procedimiento utilizado en la construcción de la estructura final de la base de datos crudos (ver apartado 4.2.3), la presencia de datos faltantes fue mínima. De los 33 atributos numéricos conformados en la base de datos, tan sólo 3 atributos tuvieron presencia de datos faltantes, estos atributos fueron 'jck', 'perMatch' y 'perIndel'. De los 612829 registros que conforman la matriz detallada, estos atributos tuvieron 384, 4853 y 4853 registros nulos, respectivamente. Esta situación se presentó al momento de unir los perfiles de metilación y los atributos numéricos mencionados, los cuales se incluyeron con la función de agregación de promedio según los rangos de exploración genética coincidentes.

Debido a que estos atributos fueron construidos a partir de una función de agregación, surgen como una suma o un promedio, es decir, teniendo en cuenta la lógica de la función de agregación, estos datos que inicialmente son datos faltantes, corresponden a un valor '0' que significa suma cero o un promedio cero.

**Filtrado de datos considerando registros categorizados con algún grado de significancia**

El conjunto de datos fue filtrado antes de la implementación de los modelos de clasificación excluyendo la clase "No Significativa" (*p-value raw* > 0.1) con el fin de analizar las relaciones entre características genéticas y epigenéticas. Las ventajas de realizar este filtrado consisten en que el análisis se centra en la data que más contribuye al entendimiento del síndrome, el rendimiento y métricas de los modelos mejora, se reduce el tamaño y complejidad del conjunto de datos, mejorando la capacidad de generalización del modelo al quitar los genes sin significancia.

## 5.3 MODELO DETALLADO DE *ML* CON VARIABLE OBJETIVO DEFINIDA SEGÚN NIVELES DE EXPRESIÓN

El modelo detallado de *ML* con la variable objetivo, definida según niveles de expresión, se construyó a partir de la matriz general, cuyo procedimiento está descrito en el apartado 4.2.3. El insumo para la definición de las características de la matriz detallada fueron los atributos numéricos de las variaciones genéticas detallados en la Tabla 3 de la sección 4.2.3.1, determinados a partir de los rangos de exploración. La etiqueta se definió según lo expuesto en el apartado 4.2.1,

categorizando la etiqueta según su significancia en 3 clases: sobre expresado, sub expresado o no diferencial.

El procesamiento de datos fue similar al modelo anterior expuesto en sección 5.2.1, con la diferencia en que la variable objetivo tiene la siguiente distribución según su significancia genética:

- Etiqueta ‘0’: genes no significativos con 611,568 registros.
- Etiqueta ‘-1’: genes sub expresados con 718 registros.
- Etiqueta ‘1’: genes sobre expresados con 548 registros.

Esto quiere decir que al momento de realizar el balanceo de datos tomando la clase minoritaria posterior al entrenamiento, el conjunto de datos termina teniendo 1,266 registros.

# 6. EVALUACIÓN DEL RENDIMIENTO DE LOS MODELOS

La evaluación del modelo se realizó principalmente con las métricas expuestas en la Tabla 2, principalmente con relación a la precisión, sensibilidad y *f1 score*. El enfocarse en estas métricas responde a los tipos de errores que estas métricas pueden revelar del modelo desarrollado.

En este caso, la precisión indica qué proporción de los casos clasificados como positivos son realmente positivos, es decir, revelará si el modelo comete errores de Tipo I, al clasificar incorrectamente como positivos casos que en realidad son negativos; por otro lado, la sensibilidad mide qué proporción de los casos positivos son correctamente identificados como tales, por lo que se dice que comete pocos errores Tipo II al no clasificar incorrectamente como negativos casos que en realidad son positivos. Por otro lado, representa el equilibrio entre los otros dos errores. En las 3 métricas, cuando más cerca de 1 sea el puntaje, es mejor.

Luego de una evaluación preliminar, se descartó el algoritmo de *Multilayer Perceptron* debido a su dificultad para generar explicaciones interpretables de los resultados. A diferencia de los modelos *Random Forest* y *XGBoost*, que permiten identificar las características más importantes para la predicción (conocido como "*Feature Importance*"), el *Multilayer Perceptron* produce modelos que son más difíciles de comprender y analizar. Además, su rendimiento en términos de precisión y F1-score fue inferior al de los otros dos algoritmos.

## 6.1 EVALUACIÓN PRELIMINAR

En la evaluación preliminar (Tabla 10) de los modelos se evaluaron características como la división de *train*, *validation* y *test*, así como el uso de división tripartita, o bipartita, dado el volumen de datos, y selección de algoritmos e hiperparámetros a evaluar. Adicionalmente, se plantearon dos escenarios, usando un conjunto grande de datos significativos y no significativos, en comparación con la propuesta de clasificación adoptada (capitulo 5.1).

Este paso se realizó a modo de *funnel* para ir seleccionando y descartando variables, lo cual permite la optimización de recursos tanto de procesamiento como de tiempo humano.

Tabla 10: Evaluación preliminar de algoritmos.

| Escenario etiqueta | {1, 0} | | | {1, -1, 0} | | |
|---|---|---|---|---|---|---|
| Algoritmo | *RF* | *MLP* | *XGBoost* | *RF* | *MLP* | *XGBoost* |
| *Mean Accuracy* | 0.508 | **0.518** | 0.511 | 0.327 | 0.331 | **0.343** |
| *Mean Precision* | 0.504 | **0.515** | 0.504 | 0.314 | 0.333 | **0.355** |
| *Mean Recall* | **0.931** | 0.608 | 0.672 | 0.331 | 0.339 | **0.347** |
| *Mean F1-score* | **0.654** | 0.555 | 0.561 | 0.231 | 0.276 | **0.277** |
| *Mean specificity* | 0.075 | 0.248 | **0.376** | 0.666 | 0.669 | **0.673** |
| *AUC* | **0.530** | 0.490 | 0.510 | 0.499 | **0.569** | 0.563 |

Nota: Se presentan dos escenarios de uso de la etiqueta, {1,0} corresponde a la categorización de la etiqueta como diferencial y no diferencial, mientras {1, -1, 0} es la categorización como sobre expresado, sub expresado y no expresado. Se evaluaron 3 algoritmos de clasificación (*Random Forest*, *Multilayer Perceptron* y *XGBoost*)

Los resultados obtenidos para las métricas de manera general son bajos para un modelo de clasificación. Aunque el escenario con etiqueta {1,0} tiene resultados más altos en las métricas, no necesariamente significa que sea la mejor alternativa, ya que en el escenario con etiqueta {1,-1, 0} las métricas están promediando lo que se obtiene cuando *expresión* = 0, lo cual hace que baje significativamente el promedio.

Adicionalmente, al evaluar el rendimiento de los algoritmos, bajo el supuesto de *Multilayer Perceptron* iba a tener un rendimiento mucho mejor que los otros dos algoritmos dado su capacidad de abstraer características complejas, no sucedió. Lo que se notó al comparar preliminarmente los modelos es que este algoritmo sufre con el bajo volumen de datos usado, ya que la premisa no se cumple, y tiene un rendimiento muy similar a los otros dos modelos, sacrificando la interpretabilidad.

Respecto a la separación tripartita o bipartita de los datos, es decir, entre usar *train*, *validation* y *test*, o sólo usar *train* y *test*. El bajo volumen de datos obligó a usar una separación bipartita para tener un conjunto adecuado de información, así como la decisión de usar una proporción 80-20, ya que demostró un mejor ajuste al modelo.

### 6.2 EVALUACIÓN MODELO GENERAL CON VARIABLE OBJETIVO CATEGORIZADA

El desempeño general de los modelos de clasificación *XGBoost* y *Random Forest* es limitado, como lo reflejan los bajos valores promedio del *F1*-score, la sensibilidad y la precisión. Esto sugiere que ambos modelos tienen dificultades para aprender patrones robustos en los datos y pueden estar incurriendo en errores tanto de Tipo I (falsos positivos) como de Tipo II (falsos negativos).

Al evaluar el desempeño por clase, se observa que ambos modelos tienen un rendimiento particularmente deficiente en la clasificación de las clases *Moderately-Significant* y *Non-Significant*, mientras que la clase *Highly-Significant* logra los valores más altos de sensibilidad. Esto indica que, aunque el modelo comete errores en general, es más confiable cuando se trata de identificar correctamente los casos altamente significativos, lo que implica que hay menos falsos negativos en esta categoría.

La incapacidad del modelo para distinguir adecuadamente entre los niveles intermedios de significancia puede deberse a características poco discriminativas en los datos de entrada, como se había identificado en la fase de exploración de datos. Es decir, las variables explicativas pueden no contener suficiente información para diferenciar entre clases con límites más difusos.

Comparando los modelos, *XGBoost* muestra un rendimiento más consistente entre clases, aunque no ofrece mejoras significativas en términos absolutos respecto a *Random Forest*. Esto sugiere que, si bien *XGBoost* es más estable, no logra una mejora sustancial en la clasificación de los casos moderadamente significativos y no significativos.

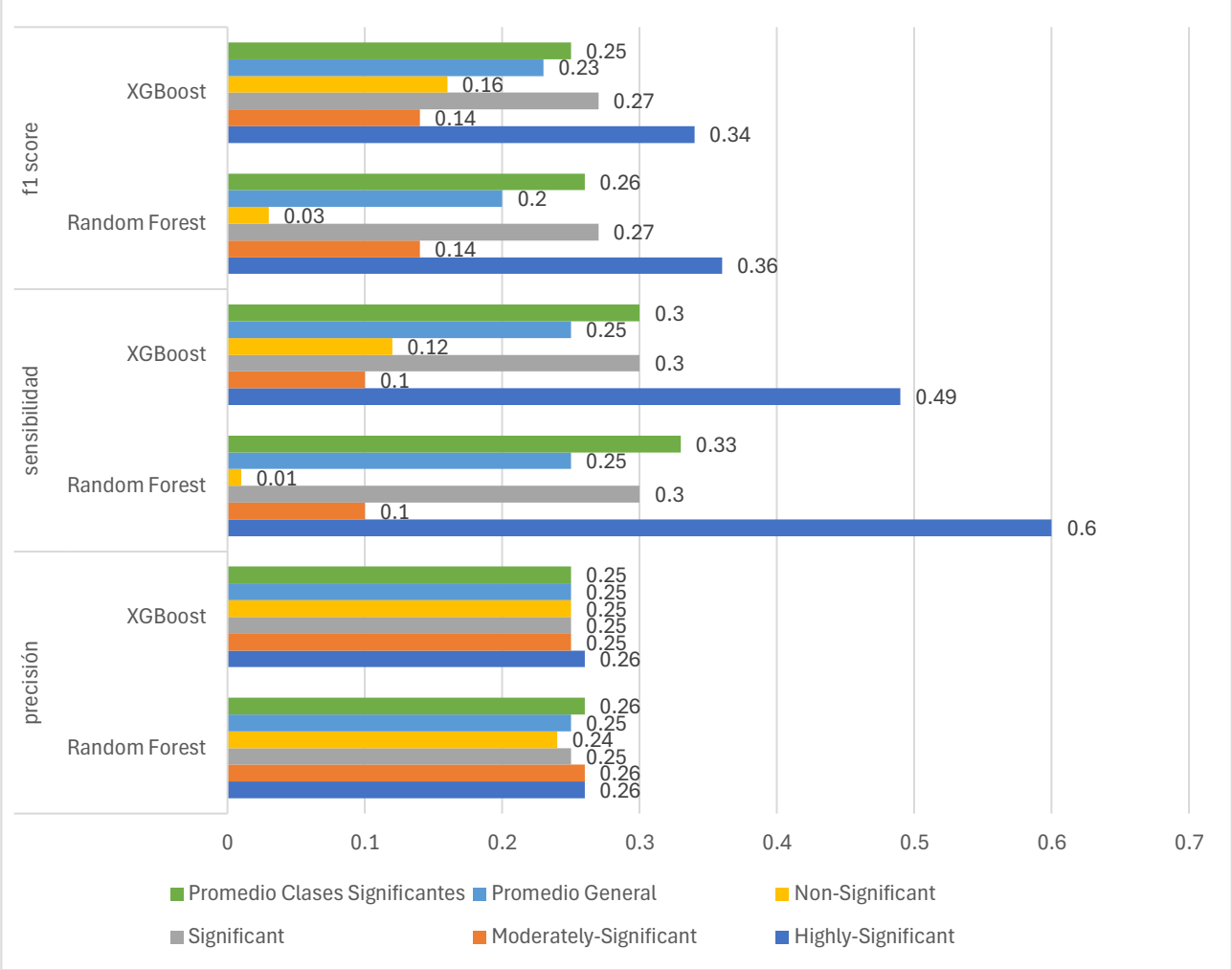


Figura 10: Métricas de desempeño del modelo de *ML* 1.

Nota: La figura muestra las métricas de rendimiento para cada algoritmo usado, según la categoría de la expresión, aplicado al modelo general.

### 6.3 EVALUACIÓN MODELO DETALLADO CON VARIABLE OBJETIVO CATEGORIZADA

En el modelo 2, al eliminar la clase no significativa, el desempeño del modelo varia con respecto al modelo de *ML* 1. Esta consideración sobre la clase de significancia nula incrementó notablemente el rendimiento global, además, ambos modelos tienen un desempeño más equilibrado interclase. Aun así, el modelo sigue presentando limitaciones con la clase *Moderately-Significant*.

A nivel general, se logró una mejora de cerca del 50% con respecto al modelo previo con cada uno de los algoritmos. La sensibilidad baja del modelo, en particular con la clase *Moderately-Significant* sugiere que el modelo no logra encontrar ejemplos de esta clase, siendo más grave con el algoritmo de XGBoost.

Así, de manera análoga a la evaluación de desempeño del modelo 1, se priorizó un desempeño más equilibrado entre clases, por lo que *Random Forest*, es el algoritmo con mejor rendimiento, a

pesar de que sea *XGBoost* el algoritmo con mejor rendimiento en la clase *Highly-Significant*.

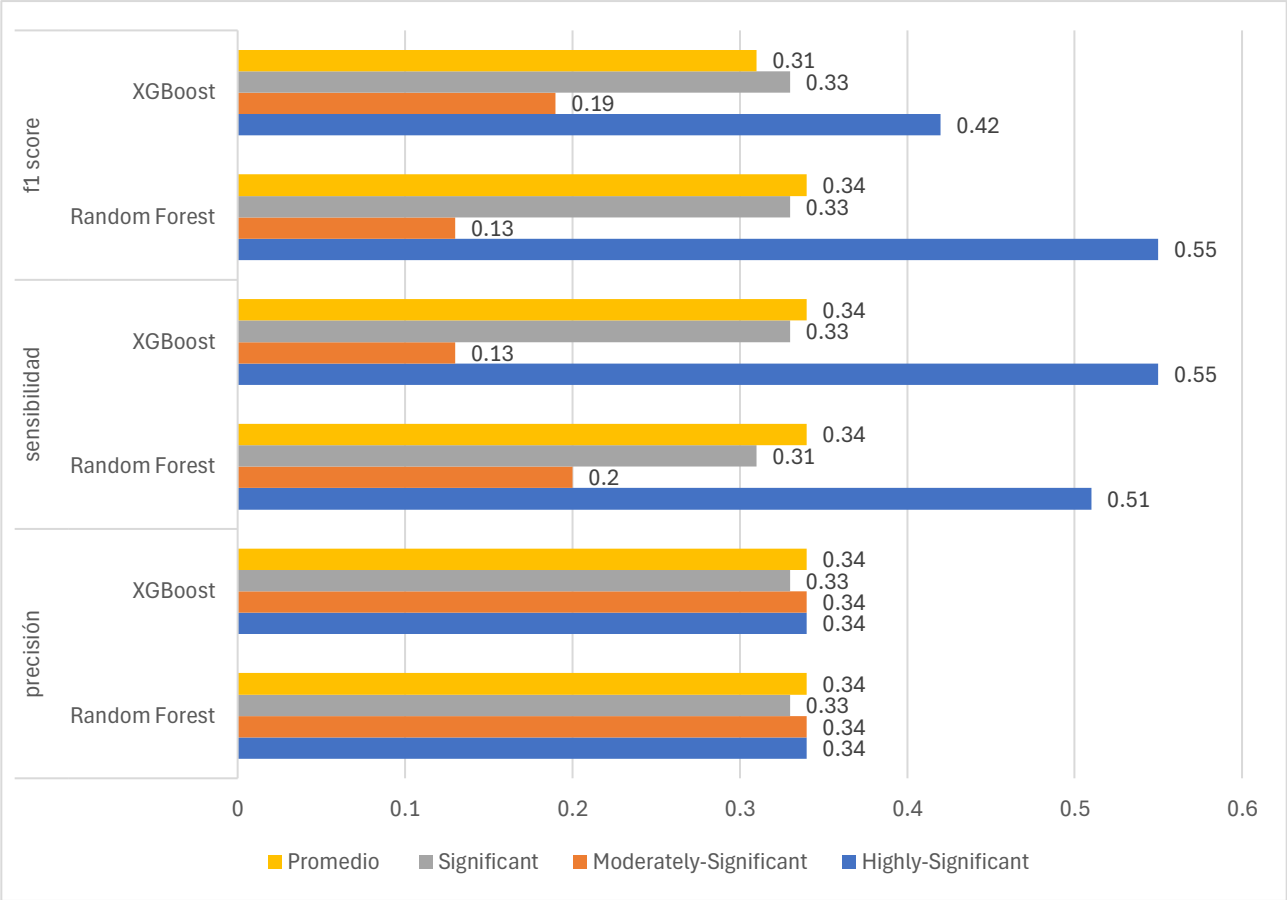


Figura 11: Métricas de desempeño del modelo de *ML* 2.

Nota: La figura muestra las métricas de rendimiento para cada algoritmo usado, según la categoría de la expresión, aplicado al modelo detallado.

## 6.4 EVALUACIÓN MODELO DETALLADO CON VARIABLE OBJETIVO DEFINIDA SEGÚN NIVELES DE EXPRESIÓN

Respecto al modelo de *ML* 3, el cual su clasificación está basada entre la sobre expresión o sub expresión de genes, se encuentra que el modelo es menos propenso a cometer errores de tipo II, debido a su alta sensibilidad, en especial con la clase de Sobre expresados. A nivel general, la precisión del modelo sigue siendo baja, pero comparable con los modelos anteriores, siendo más constante el algoritmo de *Random Forest* al comparar entre clases.

Adicionalmente, en la clase Sin significancia, los resultados sugieren que el modelo puede estar confundiendo esta clase con casos de otras clases, especialmente con el algoritmo de *XGBoost*, siendo ligeramente superior en con el algoritmo de *Random Forest*.

De manera general, el modelo presenta limitaciones mayormente con la clase de Sub expresados. Esto indica que el modelo está bien calibrado para identificar los genes que se sobre expresan en el SPW, mientras que no logra identificar características distintivas en los genes que no se expresan

en el síndrome.

Con esta redefinición de las clases, se logró una mejora en la distinción de una de las clases, clase de alta importancia biológica para ser explicativa de las correlaciones entre la característica epigenética y caracteristicas genéticas en el Síndrome de Prader-Willi. Así, el algoritmo con mejor desempeño general es *Random Forest*, especialmente en la clase de sobre expresados la que, en principio, permitiría distinguir las características fenotípicas presentes en un paciente que presente el síndrome.

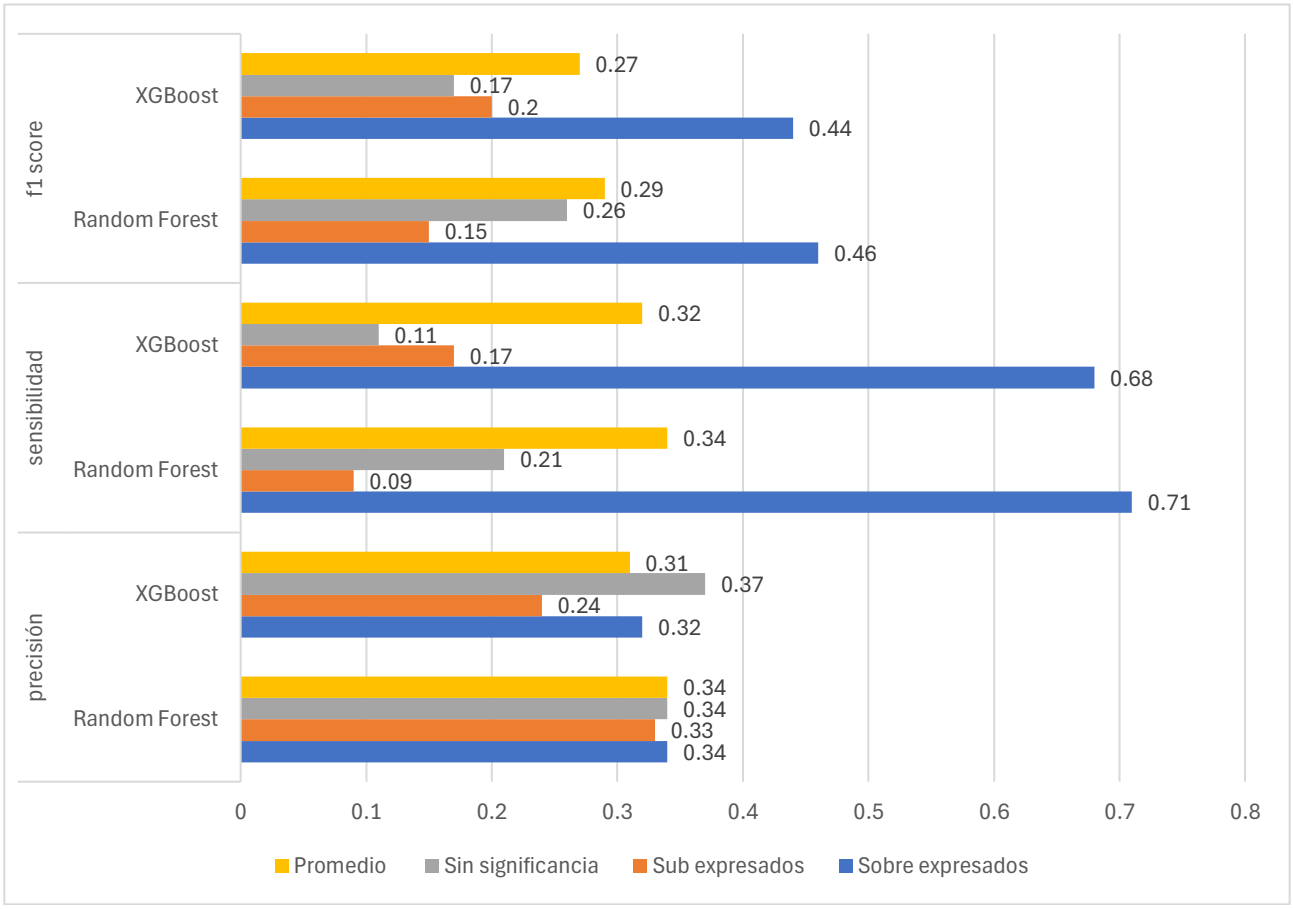


Figura 12: Métricas de desempeño del modelo de *ML* 3.

Nota: La figura muestra las métricas de rendimiento para cada algoritmo usado, según la categoría de la expresión, aplicado al modelo detallado.

La mejoría en los resultados al usar el algoritmo de *Random Forest* puede deberse a que este es más estable en problemas con mucho ruido, o volumen menor de datos. Mientras que *XGBoost* es posible que consiga identificar relaciones complejas entre las variables con mejor desempeño que *Random Forest*, requeriría de unos datos de entramiento de mayor volumen y calidad que le permita aprender estos patrones complejos.

# 7. MEDICIÓN DEL GRADO DE PRECISIÓN EN LA IDENTIFICACIÓN DE CORRELACIONES

La correlación entre la expresión genómica y la fenotípica se evaluó por medio de la relevancia de las características en la toma de decisiones de un modelo, denominada como *feature importance*. Tanto en el algoritmo de *Random Forest* como en *XGBoost*, estas medidas se calculan a partir del impacto que tiene cada característica en el proceso de construcción de los árboles de decisión. A pesar de que estas características no impliquen una causalidad, para el objetivo del presente proyecto, es útil su identificación al aportar una forma de identificar las características genéticas asociadas a una característica epigenética como lo es la metilación del ADN.

## 7.1 RESULTADOS IMPORTANCIA DE ATRIBUTOS MODELO GENERAL CON VARIABLE OBJETIVO CATEGORIZADA

En la figura 13 se presentan las importancias de características para el modelo de *ML* 1, estas reflejan el impacto relativo de cada una en el modelo predictivo usado.

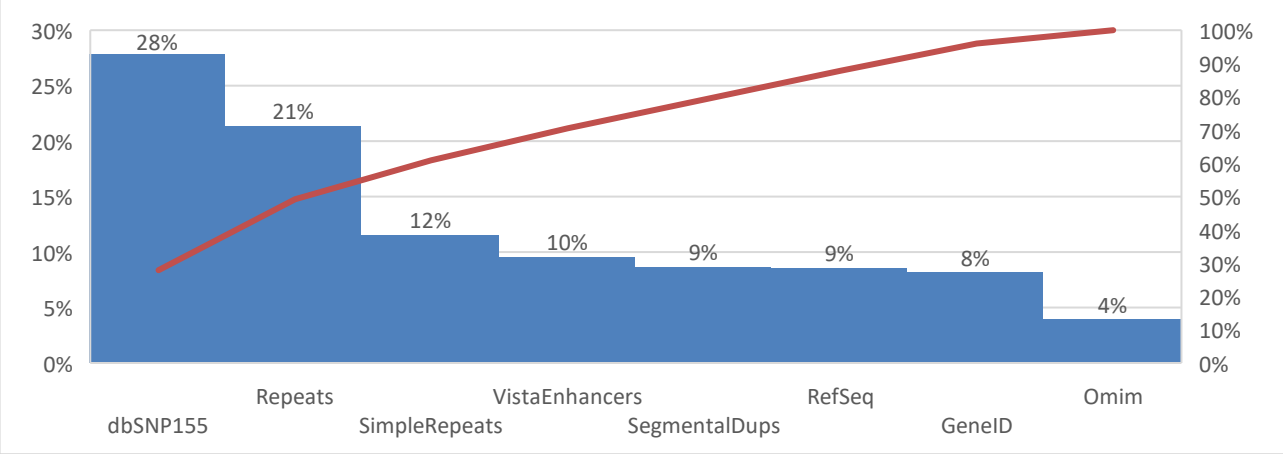


Figura 13: Importancia de atributos del modelo de *ML* 1.

Nota: La figura muestra el *feature importance* de cada característica (datos de entrada) en las columnas, mientras que la línea es el acumulado del *feature importance* (eje derecho), aplicado al modelo general.

La figura 13 sugiere que las variaciones en los polimorfismos de un solo nucleótido (*dbSNP155*) tienen mayor importancia, probablemente porque están asociados directamente con cambios puntuales en el ADN que pueden alterar expresión proteica y génica, por lo que en el caso del SPW ayuda a identificar regiones cromosómicas específicas afectadas.

Los elementos repetitivos del genoma (*Repeats*) son importantes también, lo cual se relaciona con la regulación génica, la estabilidad estructural y su influencia en procesos de reordenamiento genómicos, este resultado sugiere que las secuencias repetitivas influyen en la arquitectura del genoma en las improntas genómicas en el cromosoma 15q11-q13. Los microsatélites (*SimpleRepeats*) tienen importancia moderada, posiblemente debido a su polimorfismo y refinamiento en las regiones asociadas al SPW [29].

Los potenciadores genéticos (*VistaEnhancers*) tienen un impacto menor pero significativo. Esto puede indicar que las regiones reguladoras están asociadas con la modulación de la expresión génica, aunque su efecto sea indirecto, es decir, los genes regulados por potenciadores en las regiones improntadas podrían estar involucrados en los rasgos fenotípicos.

Luego, las anotaciones funcionales o estructurales tienen un menor peso, por lo que posiblemente su influencia es indirecta o ya está relacionada con las anteriores características más relevantes.

## 7.2 RESULTADOS IMPORTANCIA DE ATRIBUTOS MODELO DETALLADO CON VARIABLE OBJETIVO CATEGORIZADA

La figura 14 indica una alta relevancia de los atributos estructurales (*Repeats*), confirmando lo encontrado en el modelo de *ML* 1, pues las repeticiones y sus variaciones tienen un papel en la regulación y estabilidad genómica del SPW. Además, mientras que en el modelo de *ML* 1 se sugería una importancia mayor de las variaciones en los polimorfismos de un sólo nucleótido (*dbSNP155*), con este resultado se muestra que existe una contribución moderada de algunas variantes puntuales. Así mismo, se confirma la poca relevancia que tienen atributos como *omimCount* y *exonCountGeneId*, es decir, el SPW está más relacionado con mecanismos epigenéticos y estructurales que en alteraciones codificantes.

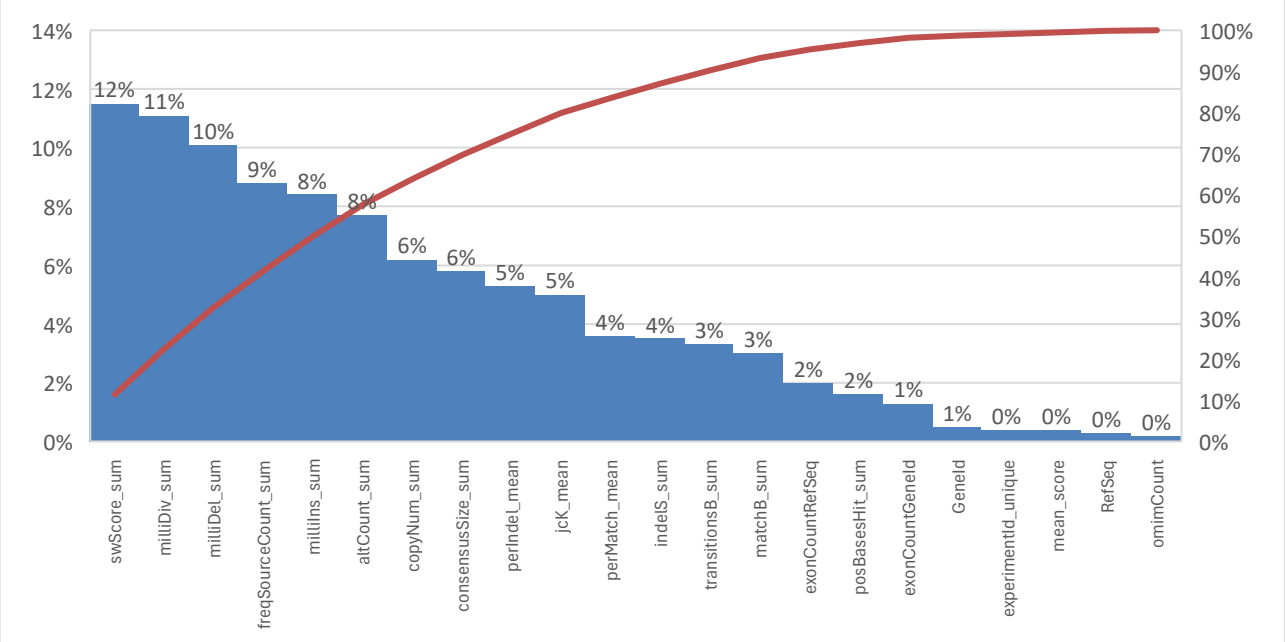


Figura 14: Importancia de atributos del modelo de *ML* 2.

Nota: La figura muestra el *feature importance* de cada característica (datos de entrada) en las columnas, mientras que la línea es el acumulado del *feature importance* (eje derecho), aplicado al modelo detallado con variable objetivo categorizada según *p value*.

La alta importancia de *swScore_sum* sugiere que la calidad y estabilidad de las repeticiones genómicas impacta en el SPW. En adición, las métricas que cuantifican divergencias, deleciones e

inserciones en secuencias repetitivas también son altamente relevantes para el modelo (*milliDiv_sum*, *milliDel_sum*, *milliIns_sum*) [30].

Respecto a *dbSNP155*, atributos como *freqSourceCount_sum* y *altCount_sum* son los más relevantes entre las variaciones puntuales. Esto indica que la frecuencia de alelos alternativos y el número de fuentes reportadas tienen un peso notable en la predicción de correlaciones entre la característica epigenética y caracteristicas genéticas.

Por otro lado, es importante destacar que este modelo surge posterior a una reducción de las dimensiones de este, por la alta correlación entre las variables explicativas, por lo que la alta importancia de algunos atributos es posible extrapolar para otros atributos. Tal es el caso de las variaciones en microsatélites (*SimpleRepeats*) con los atributos *period_sum* y *consensusSize_sum*.

Por un lado, *period_sum* representa la longitud de las secuencias repetidas, mientras que *consensusSize_sum* describe el tamaño de la secuencia consenso de dichas repeticiones. Dada la alta correlación entre los atributos, sugiere que las regiones con repeticiones largas también tienen un tamaño consenso grande, lo que puede implicar que las características de las repeticiones en estas regiones son altamente conservadas o estructuralmente coherentes, es decir, pueden estar involucradas en mecanismos regulatorios, como la impronta genómica, que depende de la integridad de secuencias repetitivas.

La importancia de esta identificación de atributos radica en que las repeticiones simples pueden influir en el silenciamiento de genes a través de la metilación del ADN, mientras que las alteraciones en el número y tamaño de las repeticiones pueden generar reordenamientos cromosómicos o alterar la transcripción de genes en el SPW [29].

### 7.3 RESULTADOS IMPORTANCIA DE ATRIBUTOS MODELO DETALLADO CON VARIABLE OBJETIVO SEGÚN NIVELES DE EXPRESIÓN

Al evaluar el modelo de *ML* 3, es decir, con la categorización de la variable de respuesta entre sobre expresados, sub expresados y no significativos, la importancia de los atributos presenta variaciones en relación modelo *ML* 2, pues confirma nuevamente que el SPW está fuertemente influenciado por variaciones en regiones repetitivas y variantes genéticas específicas, siendo su importancia aún más alta que en los modelos anteriores [29].

A partir de la figura 15 es posible identificar 3 grupos de atributos en función de su relevancia. Los atributos más importantes para la predicción del modelo están relacionados con las funciones regulatorias como la impronta genética en 15q11-q13 (*swScore_sum*), así como la regulación epigenética (*milliDiv_sum*). En el grupo de relevancia media se encuentran las deleciones e inserciones dentro de las repeticiones (*milliDel_sum* y *milliIns_sum*), lo que está relacionado con la estabilidad de las regiones del cromosoma, así como el número de alelos alternativos (*altCount_sum*), lo que refuerza la hipótesis del papel de las variaciones en polimorfismos en la modulación de los fenotipos observados en el SPW [30].

Los demás atributos tienen un impacto limitado o probablemente sean efectos residuales, por lo

que la importancia como identificador de relación entre entre la característica epigenética y caracteristicas genéticas es casi despreciable.

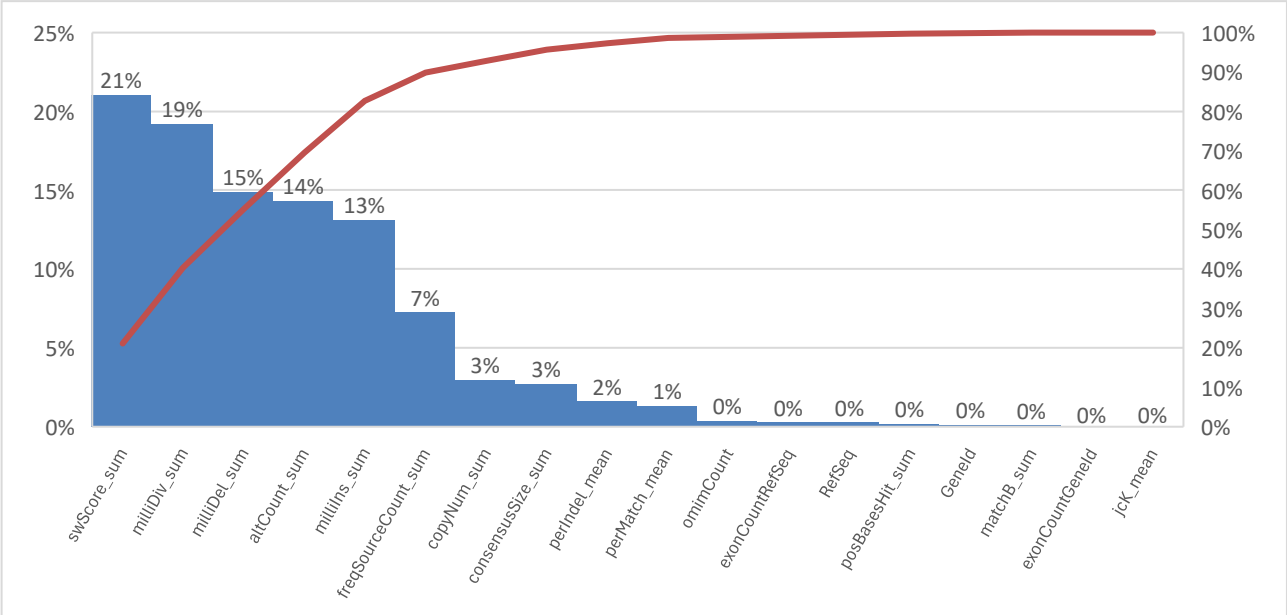


Figura 15: Importancia de atributos del modelo de *ML* 3.

Nota: La figura muestra el *feature importance* de cada característica (datos de entrada) en las columnas, mientras que la línea es el acumulado del *feature importance* (eje derecho), aplicado al modelo detallado con variable objetivo categorizada entre sobre expresados y sub expresados.

Finalmente, teniendo en cuenta que este modelo también tuvo una reducción de dimensiones a partir de la identificación de correlaciones entre las variables explicativas, ninguno de los atributos correlacionados presenta una importancia alta o media con el modelo, por lo que se puede suponer que los atributos sustraídos tampoco representarían una importancia relevante para el modelo.

# 8. DISCUSIÓN Y ANÁLISIS

La construcción de una base de datos general a partir de bases de datos de acceso libre permitió relacionar características epigenéticas (metilación del ADN), con información genética de pacientes con confirmación de SPW, creando un puente que permita caracterizar las expresiones fenotípicas. La metodología usada para construcción de esta matriz de insumo para los modelos de *ML* permitió rescatar una información que no estaba preparada para ser procesada. En este sentido, la transformación de los datos realizada consiguió reconstruir un relacionamiento entre bases de datos de distintas fuentes en una sola matriz.

A pesar de ello, se comprobó desde los primeros instantes que la información carecía de la calidad necesaria, así como el volumen debido para ser analizado con métodos robustos de clasificación. En la exploración de los datos se observó como la matriz consolidada de información carecía de una mínima correlación con nuestra variable objetivo, así como la limitación en el número de datos objetivo, es decir, de una etiqueta positiva, siendo que la mayoría de los datos eran no significativos, dificultaría etapas posteriores del análisis.

Aún con las dificultades encontradas, se construyeron diferentes modelos de *ML*, que permitieron identificar diferentes comportamientos y características de los datos. Además, en su construcción, dado el objetivo final del proyecto, se prefirieron modelos más interpretables para brindar una explicación biológica a los resultados obtenidos.

Posteriormente, se evaluó el rendimiento de cada uno de los modelos realizado, comparando sus métricas de precisión, sensibilidad (*recall*) y *f1-score*, para identificar cuál de ellos, a pesar de las limitaciones de la base de datos, lograba aprender correctamente los patrones no lineales de las interacciones genéticas-fenotípicas.

A partir de estos resultados se obtuvo que, en el modelo general, los datos no significativamente expresados estaban interfiriendo en el desempeño del modelo, por lo que se presentó una mejora conforme se afinaba del modelo *ML* 1 hacia el modelo *ML* 2. Igualmente, se obtuvo mejor desempeño al analizar la clase más altamente significante, lo que sugiere que el modelo desarrollado logró aprender en cuánto más diferenciados los datos eran, esto quiere decir, que se con una base de datos con un volumen de datos mayor, y mejor diferenciación de los genes, el modelo lograría capturar y aprender las diferencias entre las diferentes clases sin cometer tantos errores tipo I y tipo II. Así mismo, en el modelo *ML* 3 se comprobó que el modelo mejora con estas clases más diferenciadas, con la gran dificultad de contar con modelos cada vez menores en el tamaño de datos de entrenamiento.

Con relación a los algoritmos usados, ambos tanto en su estructura como en su rendimiento fueron similares, sin embargo, de manera general *Random Forest* logró capturar mejor las diferencias interclases, lo cual puede deberse a su comportamiento aleatorio, mientras que *XGBoost* al realizarlo de forma secuencial, requiere de un volumen mayor de datos que le permita ir corrigiendo cada uno de los modelos secuenciales, factor que no afectó a *Random Forest*.

El rendimiento de los algoritmos usados, a pesar de presentar bajos niveles de desempeño, presenta un comportamiento constante entre modelos desarrollados, es decir, la mejora secuencial del

modelo 1 hacia el modelo 3 presentó una evolución constante de las métricas de evaluación de los algoritmos usados, sin esto representar un desempeño general elevado.

Finalmente, gracias a la interpretatividad de los algoritmos trabajados, se identificaron las relaciones entre las variables explicativas y el genotipo significante, es decir, se obtuvo una medida de la precisión e importancia de un factor fenotípico en la clasificación del Síndrome de Prader-Willi, lo que, a futuro, se espera ayude a la compresión de este y su identificación temprana.

Este aspecto es sumamente importante, aún con los bajos niveles de desempeño que presentaron los algoritmos usados, pues permite identificar adecuadamente cuales características de los datos de entrada son los que tienen una mayor importancia para lograr la diferenciación. Indiferentemente si el desempeño global fue bajo, tener en consideración la importancia de las características permite reconocer, en términos fenotípicos, cuáles son esas características que logran diferenciar un síndrome del otro. Adicionalmente, es necesario tener en cuenta que *SPW* es un síndrome que ya es difícil de caracterizar y diferenciar, entonces la diferenciación lograda a partir del *feature importance* ya es un paso hacia adelante en reconocer características diferenciadas del *SPW*.

Respecto a la importancia de las características, las más relevantes están asociadas con variaciones puntuales (SNP) y estructuras genómicas repetitivas, lo que subraya la importancia de estas en la identificación de correlaciones entre la característica epigenética y caracteristicas genéticas para el SPW. La alta relevancia de los SNP y las repeticiones refleja su utilidad en localizar y caracterizar variantes genómicas en regiones críticas para el SPW. Este análisis puede guiar los esfuerzos hacia la priorización de variantes SNP y el estudio de repeticiones genómicas en las regiones críticas del cromosoma 15.

# 9. CONCLUSIONES Y TRABAJOS FUTUROS

## 9.1. CONCLUSIONES

Se logró desarrollar un modelo de *Machine Learning* a partir de datos genéticos, es decir, los datos analizados, y epigenéticos, que corresponden a la metilación del ADN, de individuos con síndrome de Prader-Willi, identificando correlaciones clave entre entre la característica epigenética y caracteristicas genéticas. A pesar de las limitaciones en la calidad y el volumen de los datos disponibles, los modelos interpretables, como Random Forest, demostraron su capacidad para capturar patrones significativos relacionados con variaciones genómicas críticas, como SNP y estructuras repetitivas. Este enfoque destaca la importancia de estas características en la identificación de correlaciones entre la característica epigenética y caracteristicas genéticas, subrayando su potencial para mejorar la comprensión y detección temprana del síndrome.

## 9.2. TRABAJOS FUTUROS

A partir de las conclusiones obtenidas en este estudio, se identifican varias oportunidades para continuar el trabajo y mejorar la comprensión del Síndrome de Prader-Willi (SPW) a través de la integración de datos clínicos usando *machine learning*:

**Ampliación y enriquecimiento de la base de datos**:

- Realizar esfuerzos para aumentar el volumen de datos, integrando nuevas bases de datos genómicas y epigenómicas relevantes.
- Mejorar la calidad de los datos recopilados, con énfasis en la verificación y normalización de la información para reducir el ruido y las inconsistencias.
- Priorizar la incorporación de etiquetas positivas más representativas para equilibrar las clases y mejorar el desempeño de los modelos.

**Análisis específico del cromosoma 15**:

- Enfocar futuros análisis en las regiones críticas del cromosoma 15, particularmente las implicadas en el SPW, para identificar variantes genéticas y estructurales específicas.
- Realizar estudios de asociación más detallados entre los SNP y las repeticiones genómicas en esta región para descubrir marcadores predictivos de relevancia clínica.

**Integración de datos multi-ómicos**:

- Ampliar el análisis para incluir datos transcriptómicos y proteómicos que complementen la información genómica y epigenómica, creando un modelo más completo del SPW.
- Evaluar cómo las variaciones genómicas y epigenómicas se reflejan en la expresión génica y

en los fenotipos clínicos.

Una vez logados estos primeros pasos relacionados con una construcción más robusta de la base de datos, principalmente con un volumen mayor de datos, e información precisa de las variables, es posible avanzar hacia una construcción de modelos de *ML*, como sigue:

**Exploración de modelos avanzados**:

- Probar arquitecturas más complejas, como redes neuronales profundas, que puedan manejar mejor la alta dimensionalidad de los datos y capturar relaciones no lineales, que logren clasificar adecuadamente los datos.

- Implementar técnicas de aumento de datos (*data augmentation*) y balanceo de clases para abordar el problema de datos escasos y desbalanceados.

- Explorar técnicas de interpretación de cajas negras para robustecer el análisis con modelos avanzados, pero poco interpretativos.

- Optimizar hiperparámetros tanto en Random Forest como en XGBoost como modelos altamente interpretativos, buscando un mejor rendimiento, utilizando métodos automatizados como búsqueda Bayesiana para maximizar el rendimiento.

Finalmente, al concluir los pasos anteriores y tener una base de datos robusta, enriquecida, que permita explorar modelos más avanzados, se propone avanzar hacia la construcción de modelos predictivos, no sólo clasificatorios, así como explorar el rendimiento en otros síndromes genéticos similares al SPW:

**Desarrollo de herramientas predictivas**:

- Crear modelos predictivos clínicos basados en las características más relevantes identificadas, que puedan utilizarse en el diagnóstico temprano del SPW.

- Desarrollar una plataforma de análisis accesible para investigadores y médicos, integrando los hallazgos de este estudio.

**Extensión a otros síndromes genéticos**:

- Aplicar la metodología empleada en este proyecto para el análisis de otros trastornos genéticos con componentes epigenómicos importantes.

- Comparar los patrones genético-fenotípicos del SPW con los de síndromes relacionados para identificar mecanismos compartidos.

# 10. REFERENCIAS BIBLIOGRÁFICAS


1. M. G. Butler, J. L. Miller, and J. L. Forster, "Prader-Willi Syndrome - Clinical Genetics, Diagnosis and Treatment Approaches: An Update," *Current Pediatric Reviews*, vol. 15, no. 4, pp. 207–244, 2019. [Online]. Available: https://doi.org/10.2174/1573396315666190716120925.
2. M. Yamada, H. Okuno, N. Okamoto, H. Suzuki, F. Miya, T. Takenouchi, and K. Kosaki, "Diagnosis of Prader-Willi syndrome and Angelman syndrome by targeted nanopore long-read sequencing," *European Journal of Medical Genetics*, vol. 66, no. 2, p. 104690, 2023. [Online]. Available: https://doi.org/10.1016/j.ejmg.2022.104690.
3. D. J. Driscoll, J. L. Miller, and S. B. Cassidy, "Prader-Willi Syndrome," in *GeneReviews®*, M. P. Adam et al., Eds., University of Washington, Seattle, 1998.
4. M.-L. Zhong, Y.-Q. Chao, and C.-C. Zou, "Prader-Willi Syndrome: Molecular Mechanism and Epigenetic Therapy," *Current Gene Therapy*, vol. 20, no. 1, 2020. [Online]. Available: https://dx.doi.org/10.2174/1566523220666200424085336.
5. A. Letourneau and S. E. Antonarakis, "Genomic determinants in the phenotypic variability of Down syndrome," in *Progress in Brain Research*, vol. 197, M. Dierssen and R. De La Torre, Eds., Elsevier, 2012, pp. 15-28. [Online]. Available: https://doi.org/10.1016/B978-0-444-54299-1.00002-9.
6. J. Whittington, A. Holland, T. Webb, J. Butler, D. Clarke, and H. Boer, "Relationship between clinical and genetic diagnosis of Prader-Willi syndrome," *Journal of Medical Genetics*, vol. 39, no. 12, pp. 926-932, 2002. [Online]. Available: https://doi.org/10.1136/jmg.39.12.926.
7. R. A. Costa, I. R. Ferreira, H. A. Cintra, L. H. F. Gomes, and L. C. Guida, "Genotype-Phenotype Relationships and Endocrine Findings in Prader-Willi Syndrome," *Frontiers in Endocrinology*, vol. 10, 2019. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fendo.2019.00864. DOI: 10.3389/fendo.2019.00864. ISSN: 1664-2392.
8. R. S. G. Sealfon, L. H. Mariani, M. Kretzler, and O. G. Troyanskaya, "Machine learning, the kidney, and genotype–phenotype analysis," *Kidney International*, vol. 97, no. 6, pp. 1141-1149, 2020. [Online]. Available: https://doi.org/10.1016/j.kint.2020.02.028. ISSN: 0085-2538.
9. Parmar A, Katariya R and Patel V, “A review on random forest: An ensemble classifier”. LNDECT 26, pp. 758–763, 2019
10. Ramraj s, Nishant U, Sunil R, and Shatadeep B, “Expreimenting XGBoost algorithm for prediction and classification of different datasets”, *International Journal of Control Theory and Applications*, vol. 4, 40, 2016.
11. Popescu M, Perescu L, Balas and Mastorakis, “Multilayer perceptron and neural network”.

*Waseas transactions and circuits and systems*. Vol 8, 7, 2009.

12. B. J. Erickson and F. Kitamura, 'Magician's Corner: 9. Performance Metrics for Machine Learning Models', *Radiology: Artificial Intelligence*, vol. 3, no. 3, p. e200126, May 2021.
13. T. Guo and X. Li, "Machine learning for predicting phenotype from genotype and environment," *Current Opinion in Biotechnology*, vol. 79, pp. 102853, 2023. [Online]. Available: https://doi.org/10.1016/j.copbio.2022.102853. ISSN: 0958-1669.
14. Q. Li, K. Zhao, C. D. Bustamante, X. Ma, and W. H. Wong, "Xrare: a machine learning method jointly modeling phenotypes and genetic evidence for rare disease diagnosis," *Genetics in Medicine*, vol. 21, no. 9, pp. 2126-2134, 2019. [Online]. Available: https://doi.org/10.1038/s41436-019-0439-8. ISSN: 1098-3600.
15. W. J. Dahl, J. Auger, Z. Alyousif, et al., "Adults with Prader–Willi syndrome exhibit a unique microbiota profile," *BMC Research Notes*, vol. 14, p. 51, 2021. [Online]. Available: https://doi.org/10.1186/s13104-021-05470-6.
16. Y. Peng *et al.*, "The Gut Microbiota Profile in Children with Prader–Willi Syndrome," *Genes*, vol. 11, no. 8, p. 904, Aug. 2020, doi: 10.3390/genes11080904. Available: http://dx.doi.org/10.3390/genes11080904.
17. S. Ciancia, W. J. Goedegebuure, L. N. Grootjen, et al., "Computer-aided facial analysis as a tool to identify patients with Silver–Russell syndrome and Prader–Willi syndrome," *European Journal of Pediatrics*, vol. 182, pp. 2607–2614, 2023. [Online]. Available: https://doi.org/10.1007/s00431-023-04937-x.
18. Huang WK, Wong SZH, Pather SR, Nguyen PTT, Zhang F, Zhang DY, Zhang Z, Lu L, Fang W, Chen L, Fernandes A, Su Y, Song H, Ming GL. Generation of hypothalamic arcuate organoids from human induced pluripotent stem cells. Cell Stem Cell. 2021 Sep 2;28(9):1657-1670.e10. doi: 10.1016/j.stem.2021.04.006. Epub 2021 May 6. PMID: 33961804; PMCID: PMC8419002.
19. Guzzetta, G., Jurman, G. & Furlanello, C. A machine learning pipeline for quantitative phenotype prediction from genotype data. BMC Bioinformatics 11 (Suppl 8), S3 (2010). https://doi.org/10.1186/1471-2105-11-S8-S3.
20. Sharp AJ, Migliavacca E, Dupre Y, Stathaki E et al. Methylation profiling in individuals with uniparental disomy identifies novel differentially methylated regions on chromosome 15. *Genome Res* 2010 Sep;20(9):1271-8.
21. Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., & Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. Nucleic Acids Research, 29(1), 308-311. DOI: 10.1093/nar/29.1.308.
22. Lappalainen, I., Lopez, J., Skipper, L., Hefferon, T., Spalding, J. D., Garner, J., ... & Church, D. M. (2013). dbVar and DGVa: public archives for genomic structural variation. Nucleic Acids Research, 41(D1), D936-D941. DOI: 10.1093/nar/gks1213.

23. Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F., & Hamosh, A. (2015). OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. Nucleic Acids Research, 43(D1), D789-D798. DOI: 10.1093/nar/gku1205.
24. O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., ... & Pruitt, K. D. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Research, 44(D1), D733-D745. DOI: 10.1093/nar/gkv1189.
25. Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Research, 27(2), 573-580. DOI: 10.1093/nar/27.2.573.
26. Hannan AJ. Tandem repeats mediating genetic plasticity in health and disease. Nat Rev Genet. 2018 May;19(5):286-298. doi: 10.1038/nrg.2017.115. Epub 2018 Feb 5. PMID: 29398703.
27. Visel, A., Bristow, J., & Pennacchio, L. A. (2007). Enhancer identification through comparative genomics. Seminars in Cell & Developmental Biology, 18(1), 140-152. DOI: 10.1016/j.semcdb.2006.12.007.
28. "Seshadri, Ram (2020). GitHub - AutoViML/AutoViz: Automatically Visualize any dataset, any size with a single line of code. source code: https://github.com/AutoViML/AutoViz".
29. Bisba, M., Malamaki, C., Constantoulakis, P., & Vittas, S. (2024). Chromosome 15q11-q13 Duplication Syndrome: A Review of the Literature and 14 New Cases. *Genes*, *15*(10), 1304. https://doi.org/10.3390/genes15101304
30. Huang, X.; Chen, J.; Hu, W.; Li, L.; He, H.; Guo, H.; Liao, Q.; Ye, M.; Tang, D.; Dai, Y. A report on seven fetal cases associated with 15q11-q13 microdeletion and microduplication. Mol. Genet. Genom. Med. 2021, 9, e1605