Universidad de la República
Facultad de Ingeniería

Predicción genómica multimodal: Integración de datos genómicos y ambientales para estimar el rendimiento en maíz mediante Aprendizaje Automático

Memoria de proyecto presentada a la Facultad de Ingeniería de la Universidad de la República por

Iván Abatte, Joaquín Ledesma, Santiago Sarachu

en cumplimiento parcial de los requerimientos para la obtención del título de Ingeniero Electricista.

Tutor

María Inés Fariello ..... Universidad de la República
Federico Lecumberry ..... Universidad de la República

Tribunal

María Inés Berro ..... Universidad de la República
José Crossa ..... Colegio de Postgraduados de Montecillo
Federico Lecumberry ..... Universidad de la República

Montevideo
sábado 28 diciembre, 2024

Predicción genómica multimodal: Integración de datos genómicos y ambientales para estimar el rendimiento en maíz mediante Aprendizaje Automático, Iván Abatte, Joaquín Ledesma, Santiago Sarachu.

Esta tesis fue preparada en LaTeX usando la clase iietesis (v1.1).
Contiene un total de 143 páginas.
Compilada el sábado 28 diciembre, 2024.
http://iie.fing.edu.uy/

No es la especie más fuerte la que sobrevive, ni la más inteligente, sino la que mejor responde al cambio.

Charles Darwin

Si no sabías ya sabes.

La Deskarga

Agradecimientos

Queremos agradecer profundamente a todas las personas que de alguna u otra manera nos ayudaron a llegar hasta acá. Agradecer a nuestras familias y amigos por siempre acompañarnos y brindarnos todo su apoyo. También mencionar a todos los compañeros que conocimos a lo largo de este camino, junto a los que pasamos incontables horas estudiando, anhelando algún día llegar a este momento.

A nuestros tutores, Maine y Fefo, quienes nos acompañaron académica y humanamente a lo largo de esta última etapa de nuestra formación. También a los autores del proyecto *PredGenIA*, quienes nunca dudaron en brindarnos todo su conocimiento, sin el cual no hubiese sido posible este proyecto. A todos los docentes con los que nos cruzamos en estos años, que no solo nos formaron como profesionales, si no también como personas.

Finalmente, a la Universidad de la República y a todos sus funcionarios, que le permitió a quien escribe este último párrafo, ser el primer universitario de su familia.

Esta página ha sido intencionalmente dejada en blanco.

Resumen

La predicción genómica consiste en predecir distintos fenotipos, es decir, las características observables de un organismo, a partir de su información genómica. Esta información viene dada por el ADN y se representa por la secuencia de bases que lo componen. Esto hace que el problema de la predicción genómica, sea un problema interesante de abordar con algoritmos de Aprendizaje Automático, ya que los modelos pueden ser entrenados para aprender a mapear secuencias de bases a valores de los fenotipos de interés.

Si bien los datos genómicos son útiles, factores ambientales como el clima y las condiciones del suelo influyen significativamente en la expresión de los fenotipos. En este proyecto exploraremos la integración de datos genómicos y ambientales para mejorar la predicción de fenotipos, utilizando un enfoque multimodal. Para esto, utilizaremos una base de datos de maíz que contiene secuencias genómicas y variables ambientales, como datos históricos de clima, datos del suelo, etc.

Al igual que en los proyectos de fin de carrera "PredGenIA" y "DNAI", los cuales inspiran esta investigación, utilizaremos modelos estadísticos y de Aprendizaje Automático para la tarea de predicción genómica. En este caso utilizaremos modelos lineales mixtos, gradient boosting y redes neuronales, con una arquitectura multimodal para integrar los datos genómicos con los ambientales.

El objetivo principal de este proyecto es implementar y comparar estos modelos para evaluar el impacto de los datos ambientales en la predicción genómica. Los resultados muestran cómo la integración de datos multimodales mejora la capacidad de capturar las interacciones genoma-ambiente, ofreciendo predicciones de fenotipos más robustas y realistas.

Esta página ha sido intencionalmente dejada en blanco.

Tabla de contenidos

Agradecimientos III

Resumen V

1. Introducción 1

2. Marco Teórico Genómica 3
2.1. ADN 101 . . . 3
2.2. Secuenciación y genotipado . . . 5
2.3. Fenotipo . . . 6
2.4. Predicción genómica . . . 6
2.5. Contexto de los datos a utilizar . . . 7

3. Marco Teórico Aprendizaje Automático 9
3.1. Introducción . . . 9
3.2. Aprendizaje supervisado vs. no-supervisado . . . 9
3.3. Clasificación vs. regresión . . . 10
3.4. Métricas de desempeño . . . 10
3.5. Modelos . . . 12
3.5.1. Modelos Lineales Mixtos . . . 12
3.5.2. Árboles de decisión . . . 14
3.5.3. *Gradient Boosting* . . . 15
3.5.4. Redes neuronales . . . 16
3.6. Entrenamiento . . . 20
3.7. Evaluación . . . 23

4. Estado del arte 27

5. Análisis y preprocesamiento de datos 29
5.1. Descripción de los datos originales . . . 29
5.2. Geografía de las *locaciones* . . . 31
5.3. Tratamiento de los datos ambientales . . . 33
5.4. Codificación de *SNPs* . . . 34
5.5. Selección de *SNPs* . . . 35
5.5.1. Criterio de correlación individual . . . 35
5.5.2. Criterio de valores nulos . . . 37

5.5.3. Criterio de correlación mutua . . . 40
5.6. Procesamiento de los datos fenotípicos . . . 41
5.7. Construcción de la base de datos . . . 42
5.8. Análisis de los datos combinados . . . 44
5.9. Análisis exploratorio mediante *PCA* . . . 48
5.9.1. Sobre los datos genómicos . . . 48
5.9.2. Sobre los datos ambientales . . . 50
5.10. Resumen . . . 50

6. Esquemas de validación **53**
6.1. Esquemas a utilizar . . . 53
6.1.1. Híbridos desconocidos (*HD*) . . . 53
6.1.2. Ambientes desconocidos (*AD*) . . . 55
6.2. Otros posibles esquemas de validación . . . 58
6.3. Representación mediante matrices . . . 59

7. Modelos a utilizar **61**
7.1. Modelos Lineales Mixtos . . . 61
7.2. *Gradient Boosting* . . . 64
7.3. Redes neuronales . . . 69

8. Análisis de los datos multimodales, integrando genoma y ambiente **73**
8.1. Métricas de comparación . . . 73
8.2. Modelos Lineales Mixtos . . . 74
8.3. *Gradient Boosting* . . . 78
8.4. Redes neuronales . . . 84
8.5. Comparación entre modelos . . . 88

9. Importancia de variables **91**
9.1. Valores SHAP . . . 91
9.2. Importancia de las componentes . . . 91
9.3. Recuperación de variables más importantes . . . 93
9.4. Variables más importantes . . . 95

10. Conclusiones **99**
10.1. Importancia de la información ambiental . . . 99
10.2. Modelos . . . 99
10.3. Trabajo futuro . . . 100

A. Variables de la base de datos **103**

B. Entrenamiento de los modelos **107**
B.1. Calculo de métricas . . . 107
B.2. Hiperparámetros de los modelos . . . 108
B.3. Implementación de bloques de redes neuronales . . . 109

C. Análisis utilizando solo datos genómicos **113**
C.1. Reformulación del problema 113
C.2. Modelado 113
C.3. Análisis de resultados 114

Referencias **117**

Glosario **121**

Índice de tablas **122**

Índice de figuras **125**

Esta página ha sido intencionalmente dejada en blanco.

Capítulo 1

Introducción

La predicción genómica consiste en predecir diferentes fenotipos (cualquier característica o rasgo observable de un organismo) presentes en un individuo a partir de su información genómica (información dada por la secuencia de ADN). Este enfoque aprovecha datos derivados de la secuenciación de ADN para modelar y entender la relación entre variaciones genéticas y rasgos de interés, como el rendimiento de cultivos, la resistencia a enfermedades, etc. Gracias a los avances en algoritmos de Aprendizaje Automático, se han logrado importantes avances en la capacidad para capturar patrones complejos en los datos genéticos.

Sin embargo, el genoma por sí solo no determina completamente el fenotipo, el ambiente juega un papel crucial en la expresión de características observables. Factores como las condiciones climáticas, la calidad del suelo o las interacciones ecológicas influyen de manera determinante en el desarrollo de un organismo. La integración de datos ambientales junto con los datos genómicos representa un enfoque multimodal que permite modelar no solo los efectos individuales de cada componente, sino también sus interacciones.

Este proyecto surge como continuación de los proyectos de grado anteriores, *PredGenIA* [7] y *DNAI* [11], donde se utilizan distintas técnicas de Aprendizaje Automático para la predicción de distintos fenotipos haciendo uso únicamente de la información genómica. Para seguir con la investigación en este campo, nuestro objetivo es combinar ambas fuentes de información (genómica y ambiental) para obtener predicciones más robustas y realistas, abordando uno de los principales desafíos de la predicción genómica: capturar y modelar la interacción genoma-ambiente.

Para este proyecto, nos proponemos implementar y comparar diferentes tipos de modelos de Aprendizaje Automático y estadísticos. Entre ellos, destacan los modelos lineales mixtos, ampliamente utilizados en predicción genómica debido a su capacidad para modelar efectos fijos y aleatorios. También exploraremos técnicas de Aprendizaje Automático como el *Gradient Boosting*, que ha demostrado ser efectivo en problemas tabulares complejos, y redes neuronales profundas, que ofrecen la flexibilidad de capturar relaciones no lineales en datos multimodales. El objetivo principal es evaluar el desempeño de cada modelo en escenarios con datos multimodales y ver el impacto de agregar información ambiental para la predicción genómica, en lugar de utilizar solamente datos genómicos.

Esta página ha sido intencionalmente dejada en blanco.

Capítulo 2

Marco Teórico Genómica

2.1. ADN 101

Conceptos básicos

El ácido desoxirribonucleico, conocido popularmente por sus siglas ADN, es una sustancia compuesta por grandes moléculas (*polímeros*), formada por la repetición de nucleótidos (unión de una base nitrogenada, un azúcar simple y un grupo fosfato). Las bases nitrogenadas que componen los nucleótidos del ADN pueden ser de cuatro tipos: adenina (A), timina (T), guanina (G) o citosina (C).

El ADN contiene la información necesaria para el desarrollo, funcionamiento y reproducción de todos los seres vivos. Además es el encargado de la transmisión hereditaria, es decir, la transmisión de las características de los progenitores a sus descendientes.

El ADN se encuentra principalmente en el núcleo de las células, aunque también aparece en otras partes de las mismas. Por ejemplo, en células eucariotas, también aparece ADN en las mitocondrias; y en células vegetales, en los cloroplastos.

Cromosomas

El ADN se organiza en cromosomas, que son estructuras más grandes y complejas. Estas estructuras están formadas por largas cadenas de ADN enrolladas y compactadas alrededor de proteínas llamadas histonas.

En base a como se organizan los cromosomas dentro de las células las podemos clasificar en haploides o diploides. Las haploides son aquellas en que los cromosomas aparecen individualmente, mientras que en las diploides aparecen de a pares.

Los animales, por ejemplo, son organismos mayoritariamente diploides, pues todas sus células, salvo los gametos, son diploides, mientras que, por ejemplo, los hongos son predominantemente haploides. En general, para el resto de seres vivos, la condición mayoritaria cambia a lo largo de su desarrollo, o incluso entre sexos.

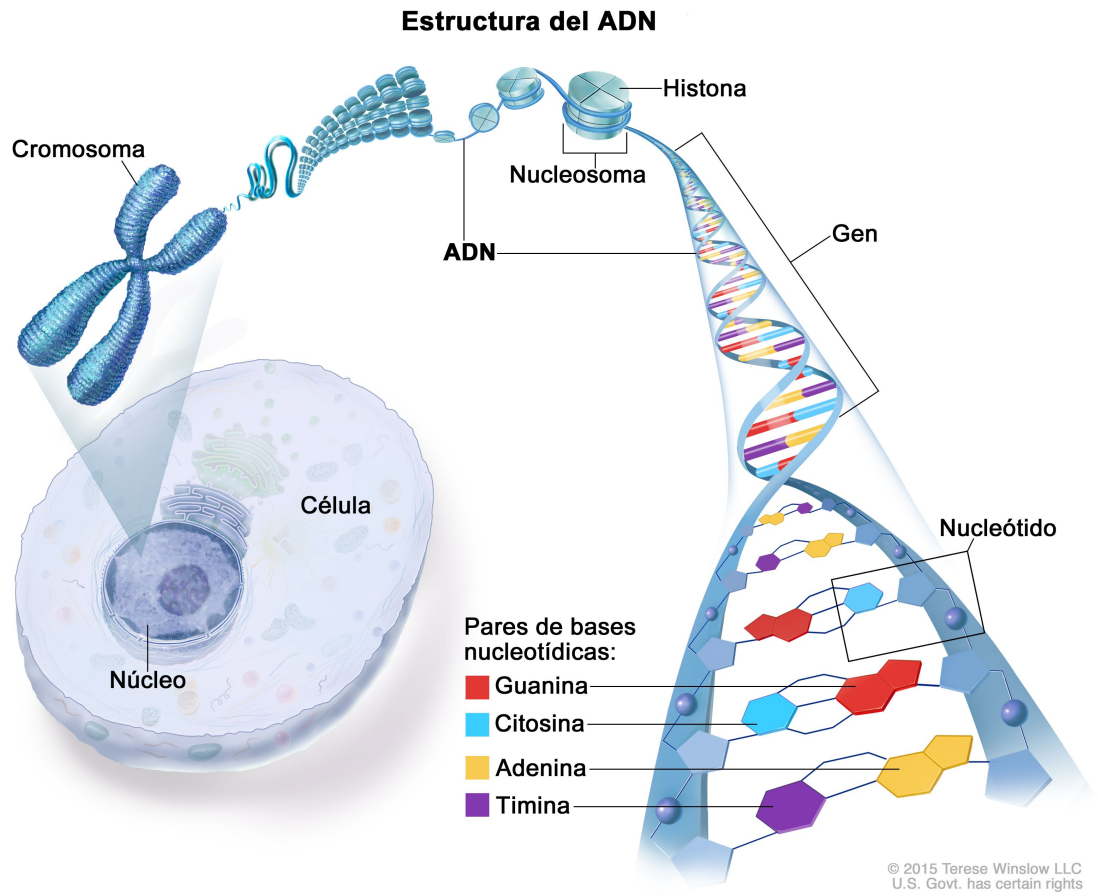


Figura 2.1: Resumen de la estructura del ADN [1].

Estructura de doble hélice

El ADN tiene una estructura de doble hélice formada por dos cadenas de nucleótidos. Las dos cadenas se unen mediante enlaces de hidrógeno entre las bases nitrogenadas de cada nucleótido, siguiendo un criterio de complementariedad en el cual las adeninas se unen con las timinas (A-T), y las guaninas con las citosinas (G-C). Este criterio se debe al tamaño de las mismas, puesto que la primera y la tercera tienen un tamaño mayor al de la segunda y cuarta, por lo cual este criterio permite mantener una uniformidad a lo largo de la doble cadena.

Gracias a esta complementariedad es que podemos describir al ADN mediante la sucesión de bases de una única de sus cadenas. Esta sucesión es entonces la que contiene toda la información genética de un individuo. Se llama genoma a todo el conjunto de ADN de una célula de una determinada especie.

Genes

Se le llama gen a una sección específica del ADN, la cual contiene la información necesaria para sintetizar una proteína. Dentro de los genes a su vez se distinguen dos

regiones: los exones, segmentos codificantes (es decir que codifican la síntesis de las proteínas) e intrones, segmentos no codificantes. Junto a él, también existen otras regiones reguladoras, que indican cuándo, dónde y en qué cantidad debe expresarse el gen.

Es importante aclarar que solo una pequeña fracción del ADN está compuesta por genes. El resto del ADN tiene otras funciones más variadas, siendo algunas aún desconocidas, pudiendo llegar a ser incluso restos evolutivos sin utilidad.

Como en los organismos diploides los cromosomas se encuentran de a pares, cada gen está duplicado, ocupando el mismo *locus*[2] en cromosomas homólogos.

A cada una de las versiones del gen (que no tienen por qué ser exactamente iguales), se le llama alelo. Si un individuo tiene ambos alelos iguales se dice que es homocigoto para ese gen. En cambio, si son diferentes se dice que es heterocigoto para el mismo.

El concepto de alelo se puede extender (por lo tanto también el de homocigoto o heterocigoto) a una base nitrogenada específica en vez de a un gen entero.

En la figura 2.1 se pueden observar las diversas partes del ADN descritas previamente.

SNPs

Se definen los polimorfismos de nucleótido simples, o por sus siglas en inglés *SNP* (*Single nucleotid polymorphisms*), como una variación de una única base nitrogenada respecto a un genoma de referencia. Es decir, una posición fija del genoma que varía respecto al común de la especie. Para que una variación sea considerada *SNP* debe ocurrir en al menos el 1 % de la población de la especie. Este umbral separa a un polimorfismo (variación relativamente frecuente) de una mutación rara.

Los *SNPs* pueden aparecer tanto en regiones codificantes del genoma como en partes no codificantes. Cuando un *SNP* aparece dentro de un gen, o en sus zonas reguladoras, puede generar alteraciones observables en el individuo, pudiendo llegar incluso a producir enfermedades.

Existe un fenómeno conocido como desequilibrio de ligamiento, que ocurre cuando en una determinada población, ciertos *SNPs* están correlacionados entre sí. Es decir, que hay alteraciones que se dan conjuntamente más frecuente de lo que sería esperable por probabilidad. Esto suele darse por lo general en *SNPs* que se encuentran físicamente cercanos en el genoma.

2.2. Secuenciación y genotipado

Junto con el ambiente, el genoma cumple un rol fundamental en el desarrollo de un individuo. Por esta razón es que resulta de interés obtener la información contenida en el mismo, tanto para un individuo concreto como para una población.

Como se mencionó previamente, la información del genoma está almacenada en el orden en que aparecen las bases nitrogenadas de la cadena.

La determinación del orden de aparición de cada una de las bases del genoma se denomina secuenciación. Si bien todas las especies de seres vivos tienen un genoma de distinto largo, este siempre es muy extenso. Por ejemplo, en animales o plantas, el

[2]Posición de un gen dentro de un cromosoma.

genoma tiene de cientos a miles de millones de bases. Por esto es que la secuenciación completa del genoma en algunas especies es un proceso complicado y costoso.

Más allá del extenso largo del genoma, la mayoría de posiciones del mismo se mantienen constantes a lo largo de individuos de la misma especie. Por ejemplo, en humanos, el 99.9 % del genoma es idéntico entre personas. Es por eso que resulta interesante observar solo las posiciones en que se sabe que hay variaciones frecuentes, es decir los *SNPs*. A este proceso se lo denomina genotipado, y resulta mucho menos costoso y complejo. Sin embargo, requiere conocer de antemano los denominados marcadores genéticos, que son las posiciones del genoma donde se ubican los *SNPs*.

Se define la densidad de un proceso de genotipado como la cantidad de marcadores que analiza el mismo en relación a la longitud total del genoma. A mayor densidad, más costoso resulta el genotipado, pero también proporciona una mejor cobertura del genoma.

2.3. Fenotipo

Se define el fenotipo como el conjunto de características observables o medibles de un individuo. El fenotipo es el resultado de la interacción del genoma con el ambiente. Es decir que dos individuos con la misma información genómica (clones), pueden poseer un fenotipo diferente si estuvieron expuestos a distintas condiciones ambientales. Entendiéndose condiciones ambientales como todas las interacciones del individuo con su entorno, por ejemplo, la alimentación del mismo, el clima del lugar en que habita, el aire que respira, etc.

Al porcentaje de la varianza del fenotipo que si se puede explicar únicamente por la genética se lo denomina *heredabilidad*. De esta manera, si un fenotipo tiene una *heredabilidad* alta significa que su variación está más influida por la genética que por el ambiente.

2.4. Predicción genómica

La predicción genómica consiste en predecir un fenotipo de un individuo, en base a su genoma, utilizando métodos estadísticos o de Aprendizaje Automático.

Usualmente no se cuenta con el genoma entero, si no que se genotipa el mismo, obteniendo una serie de *SNPs*. Para hacer este proceso, como ya mencionamos, se debe analizar previamente el genoma de la especie para identificar en qué posiciones aparecen los *SNPs*.

La principal dificultad reside en que a priori no se sabe qué partes del genoma, o qué mutaciones del mismo, pueden afectar al desarrollo del individuo. Menos aún cuáles de estas están asociadas al fenotipo que se desea predecir. Por lo que podría pasar que los *SNPs* asociados a estas mutaciones ni siquiera hayan sido genotipados. Incluso pudiendo identificar y genotipar estos *SNPs*, sigue existiendo la dificultad de entender de qué manera interactúan entre sí para afectar al fenotipo.

Por otro lado, como para muchos fenotipos la *heredabilidad* está lejos de ser total, la información genómica no alcanza para predecir a los mismos, pues en estos casos existe

una influencia ambiental. Es por eso que usualmente se añade información ambiental para intentar obtener predicciones más precisas. A los modelos que incorporan tanto información genómica como ambiental se les llama multimodales.

Otra dificultad se encuentra en que, si bien una componente del fenotipo se puede explicar únicamente con la información genómica, y otra únicamente con la ambiental, la parte restante es la resultante de la interacción entre el genoma y el ambiente. Esta interacción es la más difícil de entender, pues implica no solo comprender cómo el genoma afecta directamente al fenotipo, sino también cómo lo hace indirectamente mediante la interacción con el ambiente.

Finalmente, nunca se dispone de la información suficiente como para describir completamente la influencia del genoma ni la de las condiciones ambientales. Por lo tanto, hay una componente que inevitablemente se va a manifestar, aparentemente, aleatoria.

Suponiendo entonces que como información genómica disponemos de un conjunto de variables G, y de otro conjunto de variables E para la información ambiental, entonces podemos modelar cualitativamente el fenotipo mediante:

$$F(G, E) = F_G(G) + F_E(E) + F_{GE}(G, E) + \epsilon. \tag{2.1}$$

En este modelo el fenotipo F se compone de cuatro factores. Primero F_G que representa los efectos puramente genéticos, y que por lo tanto depende solo de G. Análogamente F_E representa los efectos puramente ambientales y por tanto depende solo de E. Luego F_{GE} es la resultante de la interacción entre genoma y ambiente, por lo que depende tanto de G como de E. Por último ϵ representa el error de este modelo, que lo suponemos como un efecto aleatorio.

2.5. Contexto de los datos a utilizar

En nuestro caso trabajamos con datos de plantas de maíz, la cual es una planta que se reproduce sexualmente a través de polinización cruzada. El maíz tiene además la peculiaridad de tener los dos sexos en una misma planta, poseyendo tanto flor masculina como femenina.

El maíz es una de las pocas especies diploides de los cultivos alimenticios, es decir, que tiene dos juegos de cromosomas. Esta especie concreta cuenta con diez pares de cromosomas.

Denominamos *clones* a las plantas (también llamadas individuos) que comparten genoma. A cada una de las variaciones genéticas de una planta, las denominamos *híbridos*.

Como estamos trabajando con plantas, los factores ambientales que afectan a su desarrollo son principalmente el clima y el estado del suelo en que fue plantada. En este contexto, se entiende clima como un amplio conjunto de condiciones atmosféricas.

Estas plantas fueron cultivadas en distintos lugares de los Estados Unidos y en diferentes años. A cada uno de estos lugares físicos se les denomina *locaciones*. En cambio, a la combinación de locación y año, se la denomina *ambiente*. Esto viene del hecho de que sabiendo dónde y cuándo fue plantada se puede saber a qué condiciones ambientales estuvo expuesta.

Se le llama *experimento* a la combinación de plantar un determinado híbrido en un determinado ambiente, para luego registrar sus fenotipos a lo largo de su desarrollo. Un

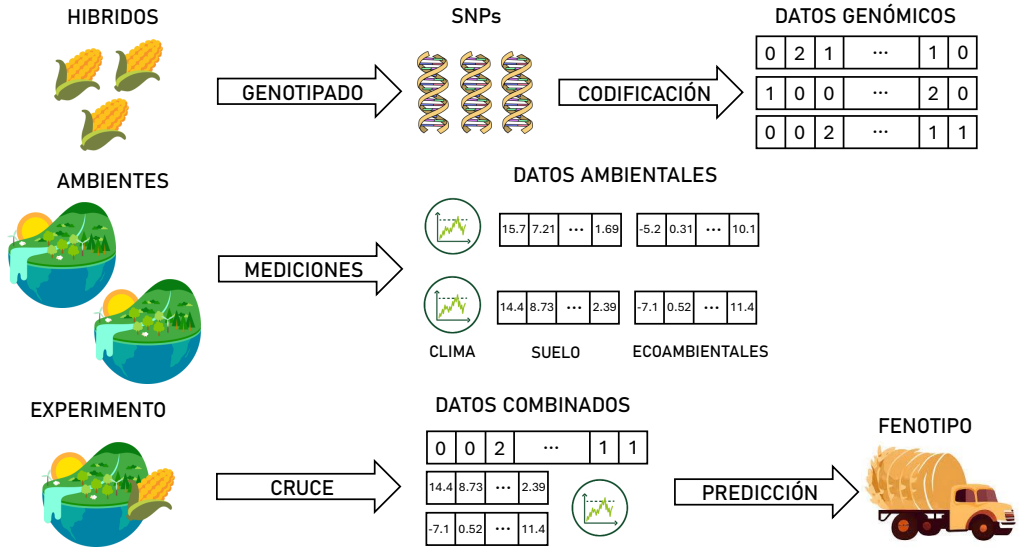


Figura 2.2: Esquema de los componentes del problema. Se distingue como se genotipan los *híbridos* y se registran variables de los *ambientes*. Luego para cada experimento se juntan ambas componentes y se usan para predecir el fenotipo.

mismo experimento se puede repetir y obtener resultados distintos, por la componente aleatoria. Al conjunto de experimentos, repetidos o no, junto a sus resultados, se les llama muestras o registros.

Como estamos trabajando con un cultivo alimenticio, es común medir los fenotipos asociados a la calidad del alimento, o sea del grano. Como el grano es un alimento que se comercializa en grandes cantidades, también es común medir el rendimiento productivo, es decir, qué cantidad de grano se produce por hectárea sembrada.

En la figura 2.2 se pueden visualizar varios de los conceptos mencionados, y cómo forman parte del problema a resolver.

Capítulo 3

Marco Teórico Aprendizaje Automático

3.1. Introducción

El Aprendizaje Automático es una rama de la inteligencia artificial enfocada en estudiar y desarrollar algoritmos estadísticos que permitan a un modelo matemático aprender patrones a partir de los datos y que éstos sean generalizables a datos nuevos.

Para el desarrollo de los modelos matemáticos y los algoritmos de entrenamiento se emplean herramientas de estadística y de optimización matemática respectivamente. Por otro lado, el análisis teórico del desempeño de los algoritmos se realiza en el marco del *"Aprendizaje Probablemente Aproximadamente Correcto"*, mediante el cual se trata de dar estimaciones para intervalos de confianza probables para los resultados en lugar de garantizar la exactitud de los mismos.

Esta disciplina tiene aplicaciones en numerosas áreas de la ciencia y la tecnología. Algunos ejemplos son: procesamiento de lenguaje natural, procesamiento de imágenes, procesamiento de audio, física computacional, química, genética, robótica y redes de datos.

3.2. Aprendizaje supervisado vs. no-supervisado

Dentro de los diferentes paradigmas del Aprendizaje Automático pueden destacarse dos paradigmas contrapuestos: el aprendizaje supervisado y el aprendizaje no-supervisado. A continuación se brinda una breve descripción de ambos paradigmas.

Por un lado, el aprendizaje supervisado trata con datos representados como pares (X, y), donde X es un vector que representa la entrada al sistema bajo estudio e y es un vector que representa la salida del sistema para la entrada X correspondiente. De esta forma, en el aprendizaje supervisado se busca entrenar un modelo de Aprendizaje Automático que prediga las salidas y a partir de las entradas X. El adjetivo "supervisado"se usa porque las predicciones que realiza el modelo son comparadas con un conjunto de salidas de referencia, y la medida de desempeño del modelo surge de esta comparación. El problema de predicción del rendimiento de una cosecha en base a datos genómicos y ambientales entra dentro de esta categoría.

Por otro lado, en el aprendizaje no-supervisado se trabajan con datos representados por un vector X, pero sin una etiqueta o salida correspondiente. Entonces, bajo este paradigma se busca entrenar modelos para que descubran relaciones o estructuras complejas en los datos. En este caso, la medida de desempeño del modelo se basa únicamente en las predicciones que realiza, pero no existen etiquetas de referencia contra las cuales comparar; de ahí que se emplee el termino "no-supervisado".

3.3. Clasificación vs. regresión

Bajo el paradigma de aprendizaje supervisado, existen dos tipos de problemas de predicción: clasificación y regresión. En el caso de los problemas de clasificación, la etiqueta a predecir es una variable que toma valores discretos, por ejemplo la raza de un perro presente en una imagen o el sentimiento que transmite un texto. Por otra parte, en los problemas de regresión se busca predecir una etiqueta que toma valores continuos, como por ejemplo el precio de un inmueble o los ingresos de una persona. El problema tratado en este proyecto pertenece a esta última categoría de problemas, pues los fenotipos usualmente son medidas continuas.

3.4. Métricas de desempeño

En esta sección se presentan algunas métricas de desempeño comúnmente utilizadas en problemas de regresión para medir el desempeño de los modelos.

Error Cuadrático Medio y Raíz de Error Cuadrático Medio

El *Error Cuadrático Medio* (*ECM* o *MSE* por sus siglas en ingles) es una métrica frecuentemente utilizada para evaluar el desempeño de un modelo predictivo. Se calcula como el promedio de los cuadrados de las diferencias entre los valores reales y los valores predichos por el modelo. Matemáticamente se expresa como:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2,$$

donde N es el número total de observaciones, $\{y_i\}_{i=1}^{N}$ los valores reales e $\{\hat{y}_i\}_{i=1}^{N}$ los valores predichos por el modelo. En general, un *MSE* más bajo indica un mejor ajuste del modelo a los datos, aunque puede ser complementado con otras métricas para evaluar el desempeño global.

El hecho de que estemos midiendo el cuadrado de los errores, facilita la detección de *outliers*[1], pues amplifica el error de estos. Sin embargo, tiene la contraparte de que atenúa los errores pequeños.

Adicionalmente se define la *Raíz del Error Cuadratico Medio* (*RMSE* por sus siglas en inglés), que no más que la raíz cuadrada del *MSE*. Se suele preferir esta medida de

[1]Valores de un conjunto que se desvían significativamente del resto.

error pues tiene las mismas unidades que la magnitud a predecir[2], lo que facilita su interpretación.

Coeficiente de correlación de *Pearson*

El Coeficiente de correlación de *Pearson* mide la magnitud y el signo de la relación lineal entre dos variables. En el contexto de las métricas de desempeño, mide el grado de linealidad de la relación entre los valores reales y las predicciones del modelo. Se denota comúnmente como r y toma valores en el rango $[-1, 1]$. Matemáticamente se define como:

$$r = \frac{\sum_{i=1}^{N}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2 \sum_{i=1}^{N}(\hat{y}_i - \bar{\hat{y}})^2}},$$

donde:

- N es el número total de observaciones,
- $\{y_i\}_{i=0}^{N}$ son los valores reales,
- $\{\hat{y}_i\}_{i=0}^{N}$ son las predicciones del modelo,
- $\bar{y}$ es la media empírica de los valores reales,
- $\bar{\hat{y}}$ es la media empírica de las predicciones del modelo.

Es importante mencionar que, aunque una dependencia lineal proporcional entre las variables implica valores de r cercanos a 1 (o -1 en el caso de una dependencia inversamente proporcional), el recíproco no se cumple necesariamente. Prueba de esto es el clásico ejemplo del cuarteto de *Anscombe* (ver figura 3.1).

[2]El uso del *RMSE* en lugar del *MSE* es análogo al uso de la desviación estándar por sobre la varianza.

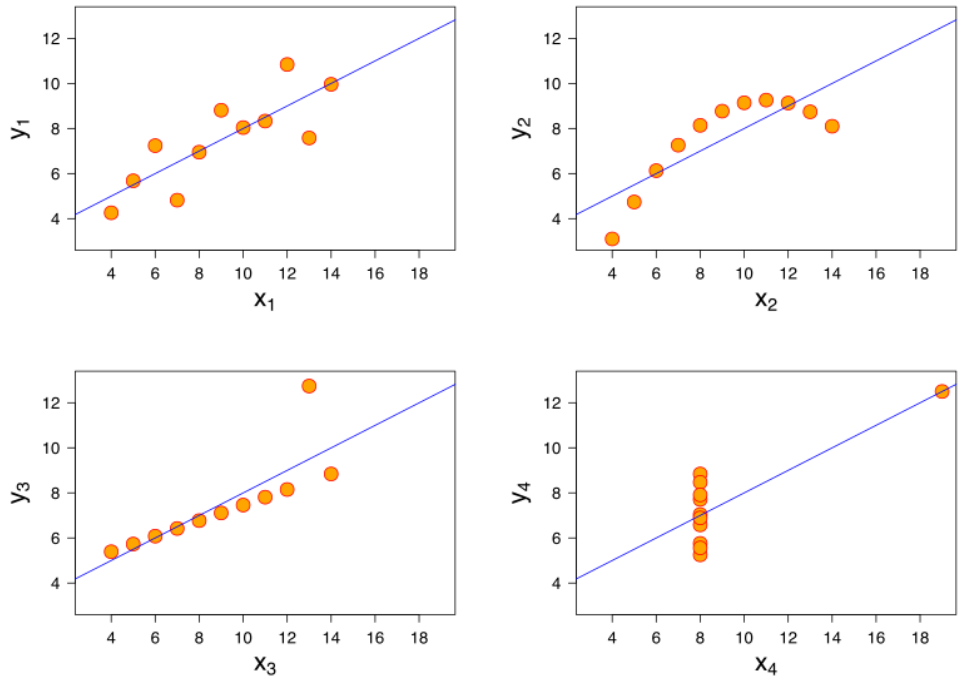


Figura 3.1: Cuarteto de *Anscombe* [44]. En la figura se muestran cuatro conjuntos de datos con estadísticos muy similares hasta segundo orden (medias y varianzas), pero que en su estructura son muy diferentes. En particular, los cuatro conjuntos de datos tienen $r \approx 0{,}816$, sin embargo el modelo lineal basado en media y correlación solo es apropiado para uno de los cuatro conjuntos de datos.

3.5. Modelos

3.5.1. Modelos Lineales Mixtos

A continuación haremos una breve introducción a los modelos lineales mixtos que son unos de los modelos más utilizados en predicción genómica y nos servirán como referencia para comparar los demás modelos de Aprendizaje Automático.

Los modelos lineales mixtos (*LMM* por sus siglas en inglés) son extensiones de los modelos lineales que se utilizan para modelar datos que presentan tanto efectos fijos como aleatorios. Se introducen los efectos aleatorios para capturar la correlación o variación aleatoria entre observaciones de un mismo grupo de individuos [34].

El modelo lineal mixto general viene dado por la ecuación:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{b} + \epsilon,$$

donde:

- $\mathbf{y} \in \mathbb{R}^{n\times 1}$ es el vector de respuesta aleatoria.
- $\mathbf{X} \in \mathbb{R}^{n\times(p+1)}$ es la matriz de diseño para los efectos fijos.
- $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^T \in \mathbb{R}^{(p+1)\times 1}$ es el vector de coeficientes de efectos fijos.

- $\mathbf{Z} \in \mathbb{R}^{n \times q}$ es la matriz de diseño para los efectos aleatorios.
- $\mathbf{b} \in \mathbb{R}^{q \times 1}$ es el vector de efectos aleatorios, tiene distribución $\mathcal{N}(0, D)$ con D una matriz de varianza-covarianza $\in \mathbb{R}^{q \times q}$.
- $\epsilon \in \mathbb{R}^{n \times 1}$ es el vector de errores aleatorios, tiene distribución $\mathcal{N}(0, R)$ con R una matriz de varianza-covarianza.

En aplicaciones genómicas, por lo general $\mathbf{b}$ incluye efectos de genotipo y efectos de interacción genotipo-ambiente, mientras que $\mathbf{X}$ puede contener información sobre covariables ambientales y otra información relacionada.

Surge de este modelo que $E(\mathbf{y}) = \mathbf{X}\beta$, y que la matriz de varianza-covarianza del vector de respuesta es $\text{Var}(\mathbf{y}) = \mathbf{Z}D\mathbf{Z}^T + R$.

Un método para estimar los efectos fijos y aleatorios es el del mejor estimador lineal insesgado (*BLUE, Best Linear Unbiased Estimator*) y el del mejor predictor lineal insesgado (*BLUP, Best Linear Unbiased Predictor*), desarrollado originalmente por Henderson en 1975 [16]. Este método consiste en resolver las ecuaciones del modelo mixto (*MME, Mixed Model Equations*):

$$\begin{bmatrix} \mathbf{X}^T R^{-1}\mathbf{X} & \mathbf{X}^T R^{-1}\mathbf{Z} \\ \mathbf{Z}^T R^{-1}\mathbf{X} & \mathbf{Z}^T R^{-1}\mathbf{Z} + D^{-1} \end{bmatrix} \begin{bmatrix} \tilde{\beta} \\ \tilde{\mathbf{b}^*} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^T R^{-1}\mathbf{y} \\ \mathbf{Z}^T R^{-1}\mathbf{y} \end{bmatrix},$$

para obtener $\tilde{\beta}$ y $\tilde{\mathbf{b}^*}$, el mejor estimador insesgado para efectos fijos y el mejor predictor insesgado para efectos aleatorios respectivamente. Si se conocen las componentes de varianza D y R se obtiene:

$$\tilde{\mathbf{b}^*} = D\mathbf{Z}^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\tilde{\beta}),$$

donde $\mathbf{V} = \mathbf{Z}^T D\mathbf{Z} + R$ es la varianza marginal de $\mathbf{y}$, y $\tilde{\beta} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}$ es el estimador generalizado de mínimos cuadrados (*GLS* por sus siglas en inglés) de β.

GBLUP

En predicción genómica, generalmente $p \gg n$ (más *SNPs* que individuos), por lo que el modelo anterior no puede aproximarse por mínimos cuadrados. Una solución es tratarlo como un modelo de efectos aleatorios y, con con la información de los marcadores construir una matriz de parentesco G [34]:

$$Y_j = \mu + \sum_{k=1}^{p} x_{jk}\beta_k + \varepsilon_j \implies Y_j = \mu + g_j + \varepsilon_j,$$

donde $g = \left[g_1, \ldots, g_J\right]^T \sim \mathcal{N}\left(0, \sigma_g^2 G\right)$ y $G = \frac{1}{p}XX^T$ con $\sigma_g^2 = p\sigma_\beta^2$.

Por lo tanto, cuando se tiene la matriz de parentesco genómico (G) de las líneas, un modelo muy útil es el siguiente (GBLUP):

$$Y = \mathbf{1}_n\mu + Z_L\mathbf{b} + \varepsilon,$$

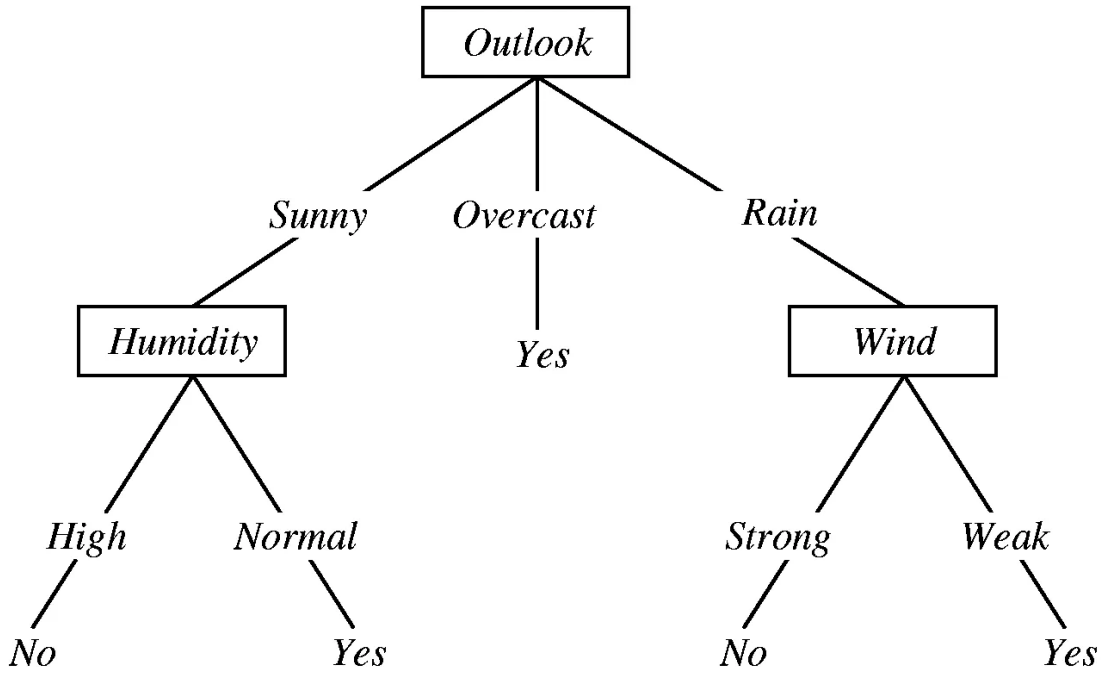


Figura 3.2: Ejemplo de árbol de decisión para decidir si jugar al tenis. Fuente: [35]

donde μ denota la media general, Z_L es la matriz diseño (o matriz de incidencia) de las líneas, $\mathbf{b} \sim \mathcal{N}_J\left(0, \sigma_g^2 G\right)$ denota el vector de efectos aleatorios de las J líneas, y $\varepsilon \sim \mathcal{N}_n\left(0, \sigma^2 I_n\right)$.

En el modelo lineal mixto general:

$$X = \mathbf{1}_n, \quad Z = Z_L, \quad \beta = \mu, \quad D = \sigma_g^2 G, \quad R = \sigma^2 I_n.$$

3.5.2. Árboles de decisión

Los árboles de decisión son una técnica de Aprendizaje Automático utilizada tanto para problemas de clasificación como de regresión. Se basa en un modelo de decisiones jerárquicas que descompone un problema complejo en subconjuntos más pequeños y manejables, facilitando el análisis y la predicción.

Los nodos internos de un árbol de decisión representan preguntas o condiciones basadas en los valores de una característica, mientras que los nodos hoja contienen el resultado final o la categoría a la que pertenece un ejemplo de datos (ejemplo: ver figura 3.2).

El entrenamiento de un árbol de decisión implica encontrar divisiones óptimas en los datos que maximicen la ganancia de información (para clasificación) o minimicen la variación (para regresión). Por ejemplo, en clasificación, métricas como la entropía y el índice de *Gini* son comunes para decidir la calidad de las divisiones.

Algunas ventajas de los arboles de decisión son su simplicidad e interpretabilidad. Pueden manejar tanto datos numéricos como categóricos [21], sin necesidad de procesamiento previo, e incluso aceptando datos faltantes en muchos casos. Por otro lado, los

arboles de decisión corren riesgo de sobreajuste si se permite que se vuelvan muy profundos. Además suelen ser poco robustos, ya que pequeñas variaciones en los datos de entrada pueden resultar en un árbol con una estructura completamente diferente [21]. Técnicas como la poda o el uso de modelos en conjunto, como los bosques aleatorios (*Random Forests*), pueden mitigar estos problemas y mejorar la precisión y generalización del modelo.

3.5.3. *Gradient Boosting*

El *boosting* es una técnica avanzada de Aprendizaje Automático que busca mejorar la precisión de modelos simples combinándolos de manera secuencial para corregir los errores de predicción de cada iteración. En esencia, crea un modelo fuerte a partir de múltiples modelos débiles (normalmente árboles de decisión de baja profundidad). El *boosting* se basa en entrenar varios modelos de manera secuencial, donde cada modelo se entrena para corregir los errores del modelo anterior, asignándole mayor peso a las observaciones mal clasificadas o con errores grandes. Por ultimo, la combinación ponderada de todos los modelos da como resultado el modelo final, que alcanza un rendimiento superior.

El *Gradient Boosting* es una mejora del *Boosting* estándar que utiliza métodos de optimización basados en el gradiente para minimizar una función objetivo específica (como el error cuadrático para regresión o la entropía para clasificación). En el caso de *Gradient Boosting*, el modelo base más típicamente empleado son los arboles de decisión.

La técnica de *Gradient Boosting* consiste en los siguientes pasos:

1. **Inicialización del modelo:** Comienza con un modelo inicial simple, que suele ser una predicción constante basada en la media o mediana del objetivo.

2. **Cálculo del *pseudo-residuo*:** Se evalúan los *pseudo-residuos*, correspondientes al gradiente de la función objetivo respecto del predictor.

   - Para el caso de regresión empleando *MSE* como función objetivo, el *pseudo-residuo* es directamente proporcional al residuo usual, entendido como la diferencia entre el valor real y la predicción.

3. **Entrenamiento secuencial:** Se entrena un nuevo modelo para predecir los *pseudo-residuos*.

4. **Actualización del modelo:** Las predicciones del residuo del nuevo modelo se suman a las predicciones anteriores, ajustadas por un factor de aprendizaje (*Learning Rate*), para mejorar gradualmente la predicción global.

La técnica de *Gradient Boosting* tiene la ventaja de permitir optimizar diversas funciones de pérdida, por lo que puede adaptarse a problemas de regresión, clasificación y *ranking*. Además, controlando los hiperparámetros de los modelos base, como el número de árboles y la profundidad máxima en el caso de arboles de decisión, puede evitarse el sobreajuste. Por último, ha demostrado tener un desempeño superior a otras técnicas en tareas complejas con datos tabulares.

Aunque tiene numerosas ventajas, entrenar multiples modelos secuencialmente puede ser computacionalmente costoso. Afortunadamente, existen librerías como *XGBoost* [8] y *LightGBM* [26] que optimizan el proceso. Por otro lado, alcanzar un buen desempeño depende de una buena configuración de los hiperparámetros, por lo que técnicas de búsqueda automatizada como búsqueda de grilla u optimización *bayesiana* son frecuentemente empleadas para encontrar configuraciones óptimas.

3.5.4. Redes neuronales

Introducción

Las redes neuronales son modelos de Aprendizaje Automático que están inspirados en el funcionamiento del cerebro animal. Están compuestas por nodos básicos llamados neuronas, que implementan funciones lineales y se interconectan formando una estructura en capas.

Mediante composiciones de operaciones lineales y activaciones no-lineales, las redes neuronales son capaces de aprender patrones complejos a partir de los datos. Esta capacidad le ha permitido a las redes neuronales resolver problemas que otros modelos de Aprendizaje Automático no han sido capaces de resolver, como clasificación y generación de imágenes y video, reconocimiento de voz o procesamiento de lenguaje natural.

Neuronas

La unidad básica en una red neuronal se denomina neurona. Esta toma un vector de entrada $x = (x_1, x_2, \ldots, x_m) \in \mathbb{R}^m$ y produce una salida escalar $y \in \mathbb{R}$ de la siguiente forma:

$$(x_1, x_2, \ldots, x_m) \mapsto y := f(w_1 x_1 + w_2 x_2 + \cdots + w_m x_m + b)$$

donde $w_i \in \mathbb{R} \; \forall \; i \in \{0, 1, \ldots, m\}$ y $b \in \mathbb{R}$ son los parámetros de la neurona, y $f : \mathbb{R} \longrightarrow \mathbb{R}$ es una función no lineal. La operación se ejecuta en dos pasos: primero se aplica una transformación afín al vector de entrada y luego se aplica una función de activación no lineal. En la figura 3.3 se muestra un ejemplo con $m = 3$, $b = 0$ y $f = H$ la función escalón unitario.

Existen muchos tipos de funciones de activación no lineales, entre las cuales se encuentran:

- Escalón unitario: Se define simplemente como:

$$H(x) = \begin{cases} 0 & \text{si } x < 0, \\ 1 & \text{si } x \geq 0. \end{cases}$$

- *Sigmoide*: Se define como $f(x) = \frac{1}{1+e^{-x}}$ y toma valores entre 0 y 1.
- Tangente hiperbólica: Se define como $f(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$ y toma valores entre -1 y 1.
- *ReLU*: La función ReLU o *Rectified Linear Unit* es una función a trozos definida como:

$$ReLU(x) = \begin{cases} 0 & \text{si } x \leq 0, \\ x & \text{si } x > 0. \end{cases}$$

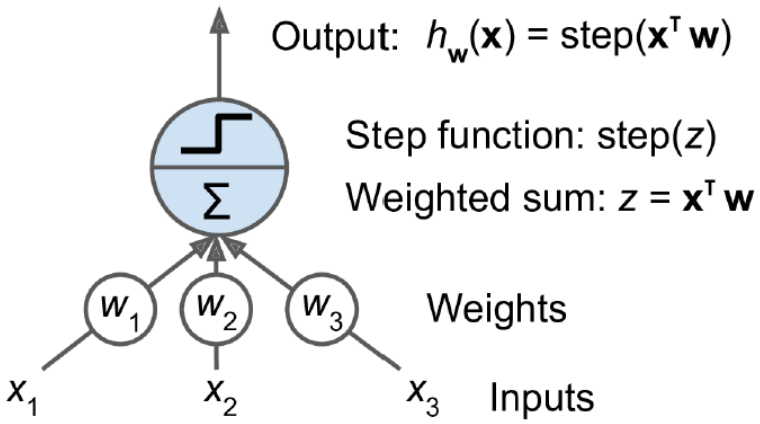


Figura 3.3: Diagrama de una neurona con un vector de dimensión 3 y escalón unitario como función de activación. Fuente: [14].

Capas completamente conectadas

Las capas completamente conectadas toman como entrada un vector $x = (x_1, x_2, \ldots, x_m) \in \mathbb{R}^m$ y producen un vector de salida $(y_1, y_2, \ldots, y_n) \in \mathbb{R}^n$ de la siguiente forma:

$$\forall i \in \{1, \ldots, n\} : (x_1, x_2, \ldots, x_m) \mapsto y_i := f(w_{i1}x_1 + w_{i2}x_2 + \cdots + w_{im}x_m + b_i)$$

donde $w_{ij} \in \mathbb{R}$ y $b_i \in \mathbb{R}$ $\forall$ $i \in \{0, 1, \ldots, m\}$, $j \in \{1, \ldots, n\}$ son los parámetros de la capa, y $f : \mathbb{R} \longrightarrow \mathbb{R}$ es una función no lineal. Es decir, la red completamente conectada es la concatenación de múltiples neuronas con sus parámetros únicos pero empleando la misma función de activación no lineal.

El teorema de aproximación universal de redes neuronales indica que las capas completamente conectadas son capaces de aprender relaciones arbitrariamente complejas entre las entradas, dada la cantidad suficiente de neuronas [19]. Sin embargo, el hecho de que cada neurona esté conectada a cada una de las entradas implica que la cantidad de parámetros es $n \times (m+1)$, por lo que depende tanto de la cantidad de neuronas como del tamaño de la entrada. Esto puede representar un problema para procesar datos de dimensión alta como imágenes, señales de audio o documentos de texto.

Capas convolucionales

Las capas convolucionales unidimensionales son un tipo de capa de neuronas que se diferencian de las completamente conectadas en dos aspectos importantes. Primero, en que las neuronas en las capas convolucionales en general no están conectadas a todas las entradas, sino a un subconjunto reducido de estas. Segundo, las neuronas en las capas convolucionales no suelen tener parámetros propios para cada una, sino que suelen compartir parámetros.

Las capas convolucionales son los componentes principales de las redes convolucionales, que han sido ampliamente exitosas atacando problemas de procesamiento de imágenes, audio y video. Esto se debe a que las redes convolucionales suelen tener una es-

tructura jerárquica, procesando características locales de las secuencias en las primeras capas y características globales de alto nivel en las últimas [45].

Como su nombre lo indica, las capas convolucionales unidimensionales implementan la operación de convolución de una secuencia unidimensional, que puede expresarse como:

$$y_i = \sum_{j=1}^{k} x_{i+j} \cdot w_j + b$$

donde:

- x_i es el valor de la secuencia de entrada en el índice i,
- w_j es el valor del filtro (o *kernel*) en el índice j,
- b es el término de sesgo y
- k es el tamaño del filtro.

Capas de Pooling

Las capas de *Pooling*, también conocidas como capas de submuestreo, son componentes fundamentales en las arquitecturas de redes convolucionales, diseñadas para reducir la dimensionalidad de los datos y extraer características clave. Estas capas aplican una operación de reducción o agregación sobre los datos de entrada, dividiendo el espacio en áreas pequeñas y seleccionando un valor representativo de cada área (por ejemplo, el valor máximo o el promedio). La reducción en la cantidad de datos permite que las redes sean más eficientes y menos propensas a sobreajuste, además de hacerlas menos sensibles a pequeñas variaciones o traslaciones en las imágenes de entrada, lo cual es especialmente útil en tareas de reconocimiento de imágenes y visión por computadora.

Existen varios tipos de capas de *Pooling*, entre las cuales destacan el *Max Pooling* y el *Average Pooling*. En el *Max Pooling*, se toma el valor máximo de cada región de la señal, lo cual es útil para resaltar las características más prominentes de la entrada (ejemplo en figura 3.4). En el *Average Pooling*, en cambio, se calcula el promedio de los valores de cada región, proporcionando una representación más suavizada y menos agresiva en términos de reducción de datos. Además, han surgido variantes como el *Global Pooling*, que aplica *Pooling* en la señal entrada completa para obtener una representación única de la misma, y el *Stochastic Pooling*, que selecciona valores de manera probabilística para añadir una ligera aleatoriedad, promoviendo la generalización.

Batch Normalization

La técnica de *Batch Normalization* fue introducida en el trabajo *Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift* [20]. En este trabajo se menciona que se propone como forma de contrarrestar el hecho de que la distribución de las entradas de cada capa cambia a lo largo del entrenamiento, enlenteciéndolo y haciéndolo más difícil cuando se emplean no linealidades saturantes. Para contrarrestar esto se propone estandarizar cada *mini-batch* de datos a la entrada de cada capa (ver figura 3.5), y hacer un re-escalado y re-centrado con parámetros aprendidos

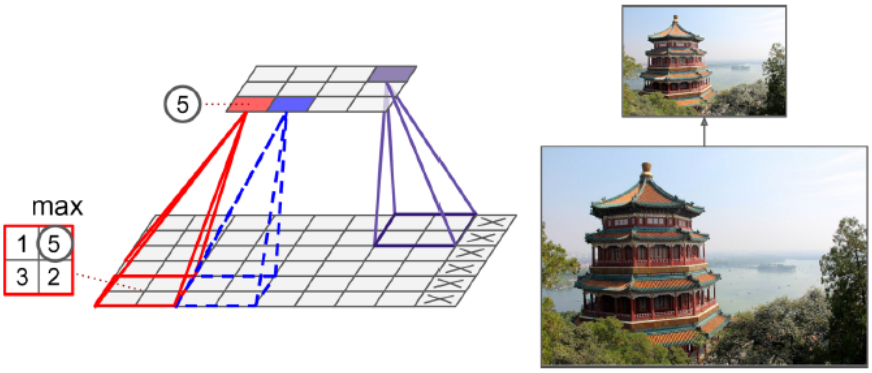


Figura 3.4: Ejemplo de *Max Pooling* en una imagen sobre áreas de 2×2. Fuente: [14].

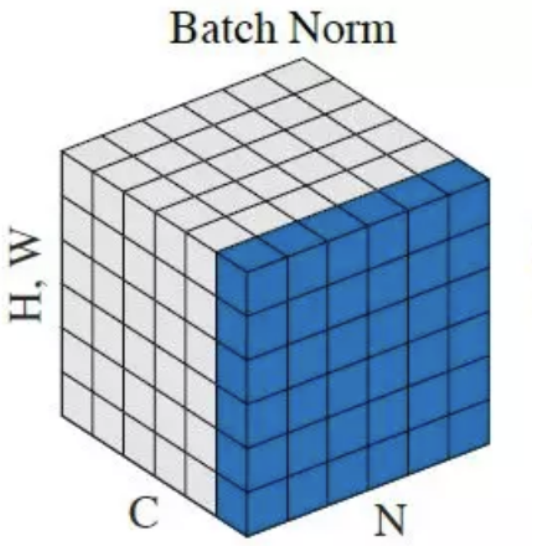


Figura 3.5: Visualización de la técnica de *Batch Normalization*. Si se considera el *batch* de datos como un arreglo con dimensiones `tamaño_de_batch` $\times$ `alto` $\times$ `ancho` $\times$ `canales`, la normalización se hace en las dimensiones `tamaño_de_batch`, `alto` y `ancho`. Es decir, se normaliza cada canal del *batch* por separado.

.

durante el entrenamiento. Esto permite utilizar tasas de aprendizaje mayores y robustecer el entrenamiento frente a la variabilidad causada por la inicialización aleatoria de los parámetros.

Dropout

La técnica de *Dropout* surgió en el artículo *Dropout: A Simple Way to Prevent Neural Networks from Overfitting* [40] como una alternativa para lidiar con el problema del sobreajuste en redes neuronales. La idea principal consiste en descartar neuronas y sus

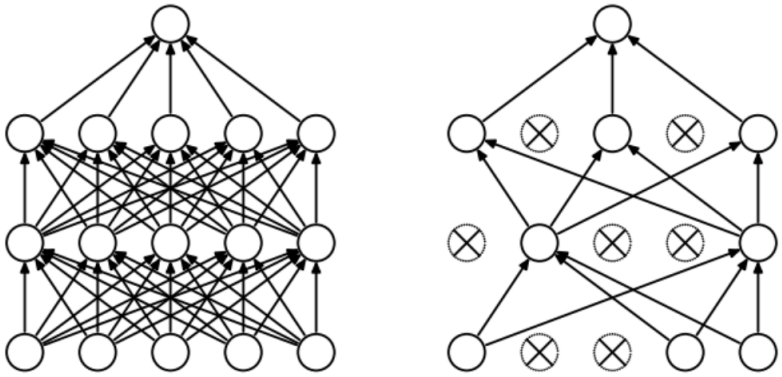


Figura 3.6: Visualización de la técnica de *Dropout* [4]

.

conexiones de forma aleatoria durante el entrenamiento (ver figura 3.6). Esto induce el efecto de muestrear predicciones de un número exponencial de redes neuronales. Mientras tanto, en tiempo de inferencia la predicción se realiza usando una red neuronal que tiene todas sus conexiones activas pero con sus pesos reescalados.

En el trabajo mencionado se muestra cómo el empleo de *Dropout* reduce el sobreajuste de manera significativa y mejora el desempeño de redes neuronales para visión artificial, reconocimiento de voz, clasificación de documentos y biología computacional.

Para entender por qué esta técnica puede resultar beneficiosa, se puede realizar la siguiente analogía: *"En un banco se capacita a los empleados para que sepan realizar las tareas de los demás, de manera que si un empleado se ausenta, los demás puedan realizar la tarea pese a su ausencia."* En esta analogía el banco sería la red entera, y cada empleado una neurona.

3.6. Entrenamiento

Algoritmos de optimización

En el ámbito del Aprendizaje Automático los algoritmos de optimización permiten seleccionar, dentro de una familia de modelos estadísticos, el modelo que minimiza una función de pérdida determinada. En otras palabras, permiten encontrar el modelo que mejor se ajusta a los datos de entrenamiento.

Dependiendo del tipo de modelo, las estrategias de optimización varían en complejidad y enfoque. Por ejemplo, para el ajuste de un modelo de regresión lineal la ecuación normal resultante puede resolverse de forma analítica mediante Mínimos Cuadrados Ordinarios [10]. Sin embargo, cuando el cálculo de la pseudo-inversa es ineficiente (por ejemplo, por tratarse de una cantidad muy grande de datos), puede hallarse una solución que tenga un buen desempeño empleando alguna variante de descenso por gradiente [2]. Por otro lado, cuando se trata de modelos no-paramétricos se emplean otro tipo de algoritmos de optimización. Por ejemplo, para el entrenamiento de arboles de decisión el

algoritmo más usado es una búsqueda voraz (*greedy*) que selecciona divisiones minimizando una métrica como la entropía en cada nodo [3].

El entrenamiento de redes neuronales es más complejo debido a la naturaleza no convexa de la función de pérdida y el gran número de parámetros a optimizar. En este caso suelen utilizarse variantes de descenso por gradiente, algunas de las cuales son:

- Descenso de Gradiente Estocástico (SGD) [38]: Actualiza los pesos del modelo usando pequeños lotes (*mini-batches*) de datos. Esto tiene dos efectos: reducir el costo computacional en comparación con el gradiente estándar, e introducir ruido en el ajuste, lo que puede evitar que el algoritmo se estanque en un mínimo local de la función de perdida.
- *SGD* con *Momentum* [37]: Los autores proponen acelerar el entrenamiento acumulando gradientes pasados, lo que permite superar mínimos locales.
- *RMSProp*: Presentado por Geoffrey Hinton en el curso *Neural Networks for Machine Learning* [17], se propone ajustar dinámicamente la tasa de aprendizaje para cada parámetro del modelo según una media móvil del cuadrado de los gradientes recientes.
- *Adam* (*Adaptive Moment Estimation*) [28]: Combina las ventajas de *Momentum* y *RMSProp*, siendo uno de los métodos más populares debido a su rápida convergencia.

Generalización y sobreajuste

La generalización se refiere a la capacidad de un modelo para desempeñarse bien en datos no vistos durante el entrenamiento. Un modelo que generaliza adecuadamente logra capturar los patrones subyacentes en los datos de entrenamiento, permitiéndole hacer predicciones confiables sobre nuevos ejemplos. Esto es el objetivo último en el Aprendizaje Automático: no solo encontrar un modelo que se ajuste a los datos (dentro de la muestra de entrenamiento), sino que también generalice a datos no vistos (fuera de muestra). Además, se entiende que si un modelo generaliza es porque se encontró un equilibrio entre complejidad y ajuste a los datos.

Por otro lado, el sobreajuste es un síntoma que refleja la incapacidad del modelo para generalizar. Este ocurre cuando un modelo se ajusta demasiado a los datos de entrenamiento, capturando no solo los patrones generales, sino también el ruido o las anomalías específicas del conjunto de entrenamiento (ver figura 3.7). Esto resulta en un modelo que tiene un desempeño bueno en los datos de entrenamiento, pero pobre cuando se aplica sobre datos nuevos.

Regularización

La regularización comprende a una familia de métodos diseñados para combatir el sobreajuste basados en imponer restricciones o penalizaciones durante el entrenamiento de los modelos para evitar que se ajusten demasiado a los datos de entrenamiento. De esta forma se incentiva al algoritmo de entrenamiento a encontrar modelos más

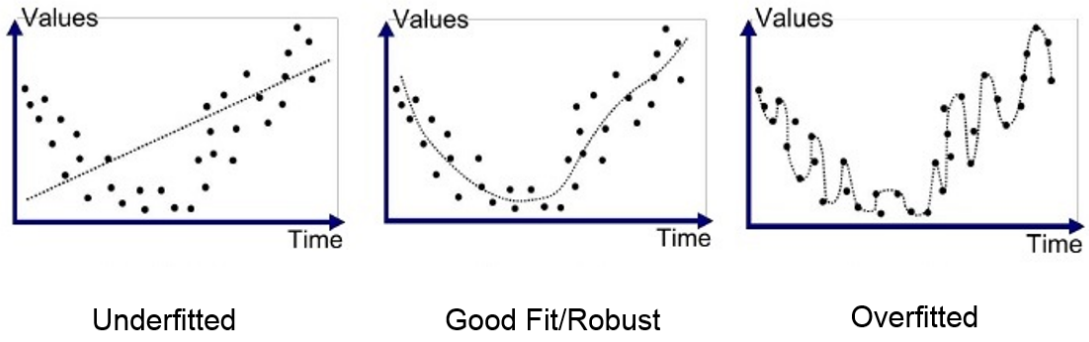


Figura 3.7: Visualización del fenómeno de subajuste (*underfitting*) y sobreajuste (overfitting). Izquierda: el modelo no logra ajustarse a los datos de entrenamiento. Centro: el modelo se ajusta adecuadamente a los datos de entrenamiento, manteniendo un balance entre desempeño y generalización. Derecha: el modelo se ajusta al ruido presente en los datos de entrenamiento, perjudicando su capacidad de generalizar a nuevos datos. Fuente: [18]

simples, priorizando patrones generales sobre detalles específicos. Algunos tipos comunes de regularización son:

- Regularización L_2: A la función objetivo se agrega un termino de penalización proporcional a la suma de los cuadrados de los parámetros del modelo. Tiende a reducir la magnitud de los parámetros sin llevarlos a cero, evitando que algunos parámetros dominen sobre otros.

- Regularización L_1: Similar al anterior, agrega una penalización proporcional a la suma de los valores absolutos de los parámetros. En este caso se promueve la esparsidad[3], llevando a cero parámetros que contribuyen poco al desempeño.

- *ElasticNet* [46]: Combina regularización L_1 y L_2, permitiendo controlar tanto la esparsidad como la magnitud de los parámetros.

- *Dropout*: Visto en la sección 3.5.4, se considera una forma de regularización ya que impone restricciones sobre el modelo durante el entrenamiento y ayuda a prevenir el sobreajuste.

- *Early Stopping*: Implica evaluar el modelo en un conjunto separado de validación durante el entrenamiento, y detenerlo cuando el desempeño en este conjunto deja de mejorar.

[3]Cualidad de esparso, adaptación de la palabra anglosajona *sparse*, que significa disperso o poco denso.

3.7. Evaluación

Conjuntos de entrenamiento, validación y test

En el desarrollo y entrenamiento de modelos de Aprendizaje Automático, los datos disponibles suelen dividirse en tres conjuntos diferentes con propósitos específicos: entrenamiento, validación y test. Estos conjuntos diferenciados contribuyen a garantizar que el modelo no solo sea capaz de ajustarse a los datos utilizados para entrenamiento, sino que además pueda generalizar a datos nuevos.

Conjunto de entrenamiento

El conjunto de entrenamiento es el subconjunto de los datos disponibles directamente utilizado para ajustar los parámetros del modelo mediante algoritmos de optimización. Por ejemplo, en un modelo de regresión, se ajustan los coeficientes; en redes neuronales, se actualizan los pesos a través de *propagación hacia atrás*. Es fundamental que este conjunto sea suficientemente representativo del problema a resolver, y suele comprender la mayor parte del total de datos disponibles.

Conjunto de validación

El conjunto de validación se utiliza en paralelo al conjunto de entrenamiento durante el proceso para evaluar el desempeño del modelo obteniendo una estimación del error fuera de muestra, es decir, de cómo se comportará el modelo con datos que no ha visto durante el entrenamiento. Esto ayuda a evitar problemas como el sobreajuste donde el modelo se ajusta demasiado a los datos de entrenamiento y pierde capacidad de generalización. También es la base de algoritmos de búsqueda de hiperparámetros que permiten encontrar valores para los hiperparámetros del modelo (ej.: cantidad de neuronas en una capa de una red neuronal) que logren un buen desempeño para una estimación del error fuera de muestra.

Conjunto de test

Por último, una vez que el modelo ha sido entrenado y refinado, el conjunto de test se emplea para evaluar su desempeño final. Sirve como una medida objetiva y no sesgada de cómo el modelo funcionará en el mundo real. Este conjunto no debe haber sido utilizado en ninguna etapa del entrenamiento o ajuste del modelo para garantizar que los resultados sean válidos.

Validación cruzada

Un problema que surge al dividir el conjunto de datos disponible en entrenamiento, validación y test es que la cantidad de datos que quedan disponibles para el ajuste de los parámetros del modelo se ve reducida, lo que puede provocar que no se aproveche plenamente la capacidad expresiva del modelo (subajuste). Una técnica comúnmente utilizada para compensar esto, es conocida como validación cruzada (*Cross Validation* o *CV* por sus siglas en inglés). En términos generales, esta se adhiere al siguiente procedimiento:

1. Se divide el conjunto de datos disponible en entrenamiento y prueba.
2. Sobre el conjunto de entrenamiento se aplican múltiples iteraciones de:
    a) Subdividir el conjunto en un subconjunto de entrenamiento y un subconjunto de validación en base a algún criterio determinado.
    b) Entrenar el modelo usando el subconjunto de entrenamiento, usando el subconjunto de validación para la estimación del error fuera de muestra.
    c) Registrar el error fuera de muestra obtenido sobre el subconjunto de validación.
3. Se obtiene la estimación final del error fuera de muestra promediando las estimaciones obtenidas en cada iteración sobre el conjunto de entrenamiento.

La técnica de validación cruzada tiene la ventaja de que todas las muestras disponibles en el conjunto de entrenamiento se utilizan también para validación, aspecto útil cuando se dispone de pocos datos. Además, como se evalúa sobre múltiples particiones, es menos probable que se ajuste excesivamente a un subconjunto específico de los datos. Por otro lado, entrenar un modelo múltiples veces para obtener una estimación del error fuera de muestra puede ser costoso, sobre todo si se trata de modelos complejos y grandes volúmenes de datos.

Una variante frecuentemente utilizada de validación cruzada es *k-folds Cross Validation*, donde el conjunto de entrenamiento se divide en k subconjuntos (disjuntos) llamados *folds* y se entrena el modelo k veces con $k-1$ *folds*, evaluando siempre sobre el *fold* restante. Otra variante utilizada es ejecutar múltiples iteraciones de sortear aleatoriamente una proporción dada de muestras del conjunto de entrenamiento para que formen parte del subconjunto de validación, dejando el resto de las muestras como el subconjunto de entrenamiento.

Búsqueda de hiperparámetros

La búsqueda de hiperparámetros es un proceso fundamental en el Aprendizaje Automático que consiste en encontrar los valores óptimos para los hiperparámetros de un modelo. Los hiperparámetros son configuraciones específicas del modelo o del proceso de entrenamiento que no se aprenden durante el mismo (como los coeficientes de un modelo lineal o los pesos de una red neuronal) sino que deben ser fijados previamente, como la tasa de aprendizaje para el descenso por gradiente, el número de capas en una red neuronal, la profundidad de un árbol de decisión o el número de vecinos en un clasificador de vecinos más cercanos.

Elegir los valores adecuados para los hiperparámetros es crucial, ya que afecta directamente el rendimiento del modelo. Un modelo con hiperparámetros bien ajustados puede mejorar su desempeño respecto a configuraciones predeterminadas de parámetros. Además puede reducir el costo computacional mediante configuraciones más eficientes.

Algunas estrategias para llevar a cabo la búsqueda de hiperparámetros son la búsqueda en cuadrícula (o *Grid Search*), donde se explora sistemáticamente todas las combinaciones posibles de un conjunto predefinido de valores para los hiperparámetros, o la

búsqueda aleatoria, donde se selecciona un número fijo de combinaciones de forma aleatoria. También se emplean técnicas más sofisticadas como la optimización *bayesiana*, que utiliza modelos probabilísticos para estimar qué combinaciones de hiperparámetros son más prometedoras, enfocando la búsqueda en esas áreas.

Esta página ha sido intencionalmente dejada en blanco.

Capítulo 4

Estado del arte

En esta sección realizaremos una revisión de la literatura acerca de la predicción genómica y los distintos métodos para incorporar información ambiental.

La predicción genómica ha involucrado una amplia gama de modelos de aprendizaje automático y métodos estadísticos. Sin embargo, ningún modelo se destaca por sobre los demás, para cualquier conjunto de datos, debido a las variaciones en la estructura genética entre las distintas poblaciones, la complejidad de los rasgos, el número de muestras y la cantidad de SNPs. Estos factores afectan el rendimiento de los modelos y dificultan las comparaciones, especialmente en ausencia de un conjunto de datos ampliamente aceptado para evaluar y comparar los métodos.

Dentro de los modelos que más se han utilizado se encuentran los modelos lineales mixtos y las regularizaciones *bayesianas* [1]. Estos modelos han obtenido buenos resultados en varios conjuntos de datos pero tienen la limitación de asumir relaciones lineales entre las características. Es por eso que con el auge del aprendizaje automático surgen distintos modelos como los árboles de decisión o *Random Forest*, las máquinas de vectores de soporte (o *SVM* por sus siglas en inglés), redes neuronales [29], *transformers* [25], etc.

En el estudio de la predicción genómica, la interacción del genotipo con el ambiente juega un rol esencial. Hay estudios que muestran que si no se considera, se tiene un peor desempeño en ensayos multiambientes [5] [23] [15].

Al igual que vimos anteriormente, una estrategia comúnmente utilizada en los ensayos multiambientes son los modelos lineales mixtos. Una forma de incorporar la información ambiental es utilizando modelos lineales con estructura *factor analytic* para modelar la varianza entre ambientes [9]. Otra alternativa utilizando modelos mixtos es incorporar los datos de ambiente con un modelo de norma de reacción que utilice estructuras de covarianza que tengan en cuenta la similitud genética entre genotipos y la similitud entre las condiciones ambientales [22].

En el análisis multimodal también están ganando fuerza los modelos de Aprendizaje Automático, aunque todavía no hay un consenso sobre cuáles son los mejores modelos [33] [24]. Hay estudios con redes neuronales [36] [32], redes neuronales convolucionales (*CNNs* [42]), redes neuronales profundas (*DNN* [27]), árboles de decisión (*Random Forest*,*XGBoost*, y *LightGBM* [43]). Estos modelos por lo general usan *embeddings* para características genómicas, ambientales y también para la interacción entre ambas, y fi-

nalmente los concatenan para pasar por una capa completamente conectada y obtener la salida.

El problema de los modelos de aprendizaje automático es que para la mayoría de los conjuntos de datos se tienen muchísimos *SNPs* para pocas observaciones, lo que dificulta el ajuste de hiperparámetros. Algunos estudios usan técnicas de reducción de dimensionalidad como PCA [42] [27] aunque esto resulta en pérdida de trazabilidad en las variantes genéticas, por lo que a veces se prefiere hacer una preselección.

Para nuestra investigación, utilizamos los datos de la competencia *Genomes to Fields (G2F) Genotype by Environment Prediction Competition* [1], sobre la cual se hicieron varias publicaciones interesantes. Entre ellas destacamos una de S. Sharma et al. [39] donde diseñan una red neuronal basada en procesamiento del lenguaje natural para manejar datos genómicos y de ambiente. Muestran que modelando el ADN como lenguaje natural se obtiene un mejor desempeño que otras técnicas, prediciendo para nuevos ambientes y, similar a otras, prediciendo híbridos desconocidos. Otra investigación que tomamos de referencia es la de *I. K. Fernandes et al.* [12], el segundo lugar de la competencia en el año 2022, donde incorporan información ambiental a modelos de *Gradient Boosting* para predicción genómica mediante ingeniería de características, logrando un enfoque eficiente para modelar interacciones entre el genotipo y el ambiente.

En general, las estrategias de modelado de los datos genómicos consisten en interpretar a los SNPs como un conjunto, donde en cada elemento se indica ya sea la base nitrogenada o si hay o no una alteración respecto al genoma de referencia, y donde se codifica empleando algún mecanismo de vectorización, por ejemplo representándolo como números naturales o One-Hot Encoding. En principio, no se suele asumir a priori una relación particular entre los diferentes SNPs. Sin embargo, algúnas técnicas que modelan el genoma bajo el enfoque del procesamiento de lenguaje natural interpretan a los diferentes nucleótidos como palabras en un vocabulario, y buscan obtener contexto adicional en cada SNP a partir de su entorno local [39].

[1] https://www.genomes2fields.org/

Capítulo 5

Análisis y preprocesamiento de datos

5.1. Descripción de los datos originales

Origen y contenido

Utilizamos los datos provenientes de la competencia *Genomes to Fields (G2F) Genotype by Environment Prediction Competition*. Estos son una colección de datos genómicos, ambientales y fenotípicos de plantas de maíz. El objetivo original de la competencia es, a partir de los datos de 2014 a 2021, predecir el desempeño de las plantaciones de 2022, es decir, predecir los distintos fenotipos en base a los datos genómicos y ambientales de las plantaciones de dicho año.

Originalmente contamos con una base de datos compuesta por tres conjuntos. El primero, de datos fenotípicos y con 136012 registros, donde cada uno corresponde a un experimento. Cada uno de estos registros contiene, además de los fenotipos medidos, cuál fue el híbrido sembrado y en qué ambiente (o sea en qué locación y en qué año). En segundo lugar uno de datos genómicos, que contiene el genotipado de los distintos híbridos. Por último, en tercer lugar, uno de datos ambientales, que contiene diversas mediciones de cada ambiente. El primer conjunto puede ser cruzado con los últimos dos, utilizando el nombre del híbrido para los datos genómicos, y el del ambiente para los ambientales.

Primeras observaciones

Dentro de los registros del primer conjunto, hay 4683 híbridos únicos, en 217 ambientes (ubicados en 50 locaciones a lo largo de 9 años), y con 87412 combinaciones únicas de ambos. De esto se concluyen tres cosas:

- Como $87412 < 136012$, significa que necesariamente hay registros que comparten híbrido y ambiente. Esto quiere decir que puede haber varias plantaciones del mismo híbrido en el mismo tiempo y lugar, pero con fenotipos diferentes.
- Por otro lado $50 \times 9 > 217$, por lo que no todas las locaciones están presentes todos los años.

- Finalmente, al ser $4683 \times 217 > 87412$, concluimos que no todos los híbridos están presentes en todos los ambientes (ni siquiera en todas las locaciones, pues $4683 \times 50 > 87412$).

Desglosado de los datos

Por un lado, tenemos los datos genómicos, que se representan con una serie de 437214 *SNPs*. Por otro lado, los datos que llamamos ambientales se dividen en tres grupos: datos climáticos, datos del suelo y covariables ambientales. Los datos climáticos son 16 variables medidas diariamente a lo largo de un año, conformando así un conjunto de series temporales. Los datos del suelo son 27 variables asociadas a mediciones realizadas al suelo de la locación, en un determinado año. Finalmente, las covariables ambientales son un conjunto ya procesado de 765 variables, obtenidas a partir de distintas mediciones del ambiente. En el apéndice A se encuentra una descripción más detallada de todas las variables de la base de datos.

En cuanto a los datos fenotípicos, se presenta una serie de mediciones de distintos rasgos, en distintas etapas del desarrollo de la planta. Nosotros vamos a trabajar únicamente con la variable denominada *Yield_Mg_Ha*, que no es más que el *tonelaje*[1] de la producción que tuvo esa plantación (en una plantación se planta un único híbrido).

En resumen, podemos decir que los datos genómicos son una matriz de dimensión:

$$\underbrace{N_H}_{\text{Híbridos}} \times \underbrace{437214}_{SNPs},$$

los ambientales una de dimensión:

$$\underbrace{N_E}_{\text{Ambientes}} \times \left(\underbrace{365}_{\text{Días}} \cdot \underbrace{16}_{\text{Series}} + \underbrace{27}_{\text{Suelo}} + \underbrace{765}_{\text{Cov.}} \right),$$

y los fenotípicos un vector de largo 136012. Donde N_H es la cantidad de híbridos genotipados y N_E la de ambientes de los que se tienen datos.

Además de la información mencionada, también se tienen *metadatos* con información acerca del contexto de la plantación a nivel de ambiente. Por ejemplo, en qué fecha fue la siembra y cuándo la cosecha, qué tipo de regado se utilizó, posición geográfica de la plantación, extensión de la misma, distancia a la estación meteorológica más cercana (de donde se obtienen los datos climáticos), entre otras. Nosotros decidimos no utilizar esta información para reducir la complejidad del problema.

Consideraciones acerca de los datos

Los mismos organizadores de la competencia informan que los datos no están totalmente verificados, ni tan siquiera disponibles en algunos casos, por lo que recomiendan evaluar la integridad y usabilidad de los mismos. Como no disponemos del conocimiento de la materia como para evaluar si los datos son verosímiles (sobre todo los genómicos), supusimos que todos los datos disponibles eran utilizables y nos limitamos a realizar un análisis de datos faltantes, y en base a eso, realizar nuestro curado de los datos.

[1] Medido en toneladas por hectárea sembrada.

Distancia máxima	**Porcentaje de locaciones**
100 m	22.22 %
1 km	55.55 %
10 km	80.55 %
100 km	88.88 %
395 km	100.00 %

Tabla 5.1: Distribución acumulada de la distancia máxima que hubo entre las plantaciones de una misma locación a lo largo de los años. Dada la distancia de la izquierda, se muestra a la derecha que porcentaje de locaciones tienen una distancia máxima de variación menor a esa distancia.

5.2. Geografía de las *locaciones*

Con el fin de tener un mejor entendimiento sobre el contexto de los datos, decidimos analizar la geografía de las plantaciones. Para eso utilizamos los *metadatos* para ubicar a las distintas locaciones en el mapa.

Desplazamientos y ubicación de las plantaciones

Si bien las locaciones hacen referencia a un lugar geográfico, la ubicación exacta en que se realizan las plantaciones va cambiando año a año. Por lo tanto, decidimos evaluar si había grandes desplazamientos entre los cultivos realizados en una misma locación a lo largo de los años.

Lo anterior lo hicimos calculando la máxima distancia entre plantaciones de una misma locación a lo largo de los años.

La tabla 5.1 muestra qué porcentaje de locaciones tienen una distancia máxima entre plantaciones menor a cierto umbral (la columna de la derecha). Se puede apreciar que las plantaciones no suelen moverse demasiado, pues el 80 % de las plantaciones no se movieron más de 10 km a lo largo de los años. Estos movimientos no son suficientes como para cambiar las condiciones climáticas de la zona, aunque es factible que modifiquen las cualidades del suelo.

En cuanto al 20 % de plantaciones que se movieron más de 10 km, sí pueden haber experimentado cambios climáticos. Por lo que habría que dividir a estas locaciones en dos, o más, de manera que la distancia máxima de cada una de estas sea más acotada. Sin embargo, cuando analizamos estos grandes desplazamientos, vemos que corresponden a un único año en que las plantaciones se realizaron fuera de la zona habitual. Por lo que por simplicidad, y para no añadir locaciones que apareciesen en un único año, decidimos no hacer esta reasignación.

Para complementar el análisis, decidimos ubicar las locaciones en el mapa de los Estados Unidos, para entender qué tan lejos estaban entre sí, y poder identificar zonas de mejor o peor desempeño.

En la figura 5.1 se muestran las distintas locaciones ubicadas en el mapa por las coordenadas de las plantaciones del último año en que aparece la locación. Estas locaciones se encuentran coloreadas según el desempeño promedio de la locación y dimensionadas según la cantidad de registros disponibles en el último año en que aparece. Se puede

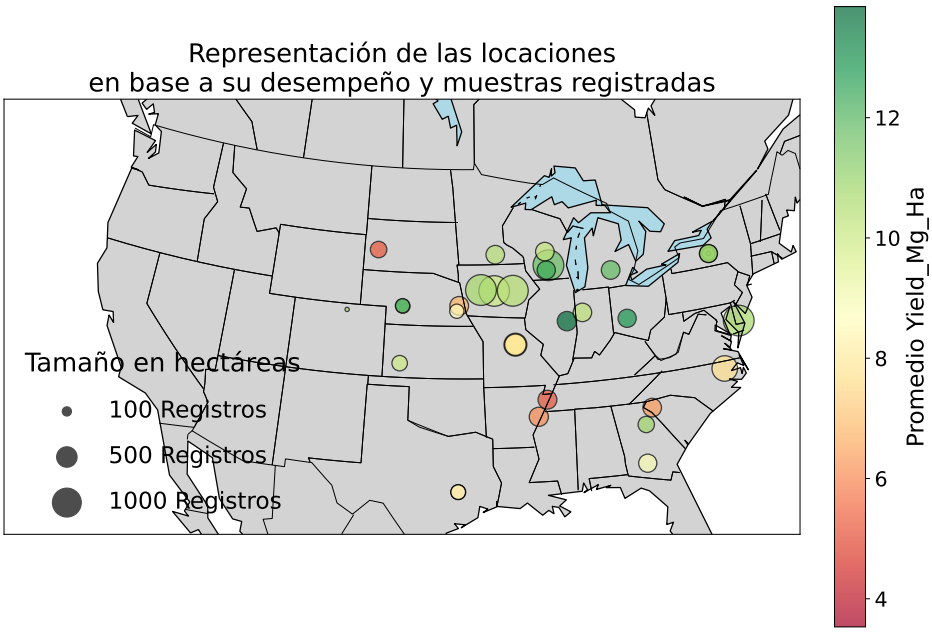


Figura 5.1: Locaciones ubicadas en el mapa según las coordenadas de las plantaciones del último año en que aparece la locación. Coloreado según desempeño promedio de la locación y dimensionado según cantidad de registros disponibles, en el último año en que aparece.

ver que la mayoría de las locaciones se encuentran en la zona este del país, y no hay plantaciones ubicadas en el oeste ya que es una zona árida.

Respecto al desempeño de las locaciones, se observa como las que están ubicadas en la zona del noreste, tienen en promedio un mejor rendimiento que las que están más al sur.

Elevación y tamaño de las plantaciones

Finalmente, considerando que Estados Unidos es un país que posee una gran variedad topográfica, decidimos analizar a qué altura estaban las plantaciones. En la *metadata* solo tenemos información de latitud y longitud, por lo que para obtener la altitud utilizamos la *API* de *Open-Elevation*[2].

Por otro lado, también consideramos interesante analizar qué tan grandes eran las plantaciones. En la *metadata* están las coordenadas de cada uno de los vértices de la plantación. Para calcular el área a partir de estos datos, hicimos una aproximación plana (lo cual es razonable pues ninguna plantación es tan extensa).

[2]https://github.com/Jorl17/open-elevation

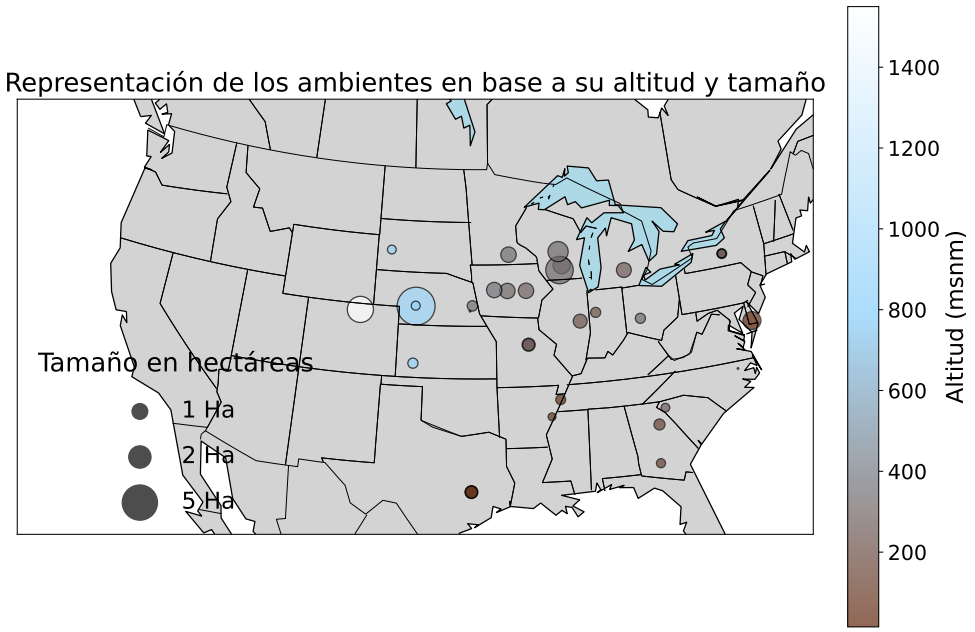


Figura 5.2: Locaciones ubicadas en el mapa (coordenadas de las plantaciones del último año en que aparece la locación). Coloreado según altitud y dimensionado según la extensión de las plantaciones, en el último año en que aparece.

En el mapa de la figura 5.2 se muestran las distintas locaciones ubicadas geográficamente, coloreadas según la altitud y dimensionadas según el tamaño de las plantaciones. Se puede ver que hay una variedad considerable de alturas, habiendo plantaciones a nivel del mar y otras a más de mil metros por encima del mismo. Sobre todo se ve que las plantaciones con mayor elevación son las que están más hacia el centro del país, sobre la zona de las montañas rocosas.

En cuanto a la extensión de las plantaciones, identificamos de la figura 5.2 que no hay plantaciones demasiado grandes, teniendo la mayoría menos de cinco hectáreas.

5.3. Tratamiento de los datos ambientales

Los datos climáticos y las covariables ambientales ya venían íntegros totalmente, por lo que no precisaron ninguna limpieza previa. En cuanto a los datos del suelo, se realizaron algunos procesamientos asociados a imputar valores faltantes, codificar columnas categóricas (*One Hot Encoder*) y codificar valores de las mismas con baja aparición.

Resumen estadístico de los datos climáticos

Si bien, como explicaremos más adelante, utilizamos modelos que consumen la serie temporal entera, también diseñamos modelos que solo operan con un conjunto de variables climáticas que son un resumen estadístico de estas series. El motivo de este tratamiento fue en pos de reducir el poder de cómputo requerido, y considerando que la información relevante para el desarrollo de la planta podía ser capturada en un resumen estadístico de las distintas series.

Para construir este resumen partimos de las 16 series temporales asociadas a las mediciones climáticas disponibles. Cada una de estas series fue segmentada en cuatro periodos estacionales, en función de las estaciones del hemisferio norte, correspondiente a la región donde se realizaron las plantaciones.

Sobre cada una de estas particiones estacionales, calculamos los siguientes estadísticos descriptivos:

- Media: promedio de los valores dentro de la partición.
- Desviación estándar: medida de dispersión que refleja la variabilidad de los datos.
- *Skewness*: medida de la asimetría de la distribución de los datos.
- Tercer cuartil: valor por debajo del cual se encuentra el 75 % de los datos.
- Primer cuartil: valor por debajo del cual se encuentra el 25 % de los datos.
- Máximo suave: valor por debajo del cual se encuentra el 95 % de los datos.
- Mínimo suave: valor por debajo del cual se encuentra el 5 % de los datos.

Estos estadísticos incluyen una combinación de momentos y percentiles que permiten caracterizar de forma robusta las distribuciones climáticas en cada período.

Dado que para cada ambiente contamos inicialmente con 16 series climáticas, y cada serie fue dividida en cuatro periodos estacionales, calculando siete estadísticos por período, el resultado final consiste en $16 \times 4 \times 7 = 448$ variables por ambiente.

Este enfoque garantiza una representación detallada y comprensible de las condiciones climáticas para cada ambiente evaluado.

5.4. Codificación de *SNPs*

En contraposición a los datos ambientales, la información genómica sí necesitó un mayor preprocesamiento. En primer lugar, como son datos genómicos de especies diploides, al momento de codificar un marcador asociado a un *SNP*, existen tres posibilidades:

- Homocigoto para el alelo de referencia: Ambos alelos presentan la misma base que el alelo considerado de referencia. En los archivos originales, estos casos están representados por el símbolo 0/0. En nuestro sistema, los codificamos con un valor de 0.

Valor	Porcentaje
0 (Homocigoto para el alelo de referencia)	65.68 %
1 (Heterocigoto)	30.52 %
2 (Homocigoto para el alelo alternativo)	3.80 %

Tabla 5.2: Distribución del valor más frecuente (entre individuos) de la codificación de los SNPs.

- Heterocigoto: Un alelo como el de referencia y otro alternativo. Se representan con 1/0 o con 0/1. Los codificamos con un valor de 1.
- Homocigoto para el alelo alternativo: Ambos alelos tienen la misma base que el alelo alternativo. En el archivo original, están representados como 1/1 y en nuestra codificación se asigna un valor de 2.
- Casos atípicos: Existe un porcentaje reducido (menor al 0.01 %) de situaciones en las que la base registrada en uno o ambos alelos no coincide ni con la del alelo de referencia ni con la alternativa. Esto puede ocurrir porque el alelo alternativo representa la variación más común y no abarca la totalidad de las posibles variantes. Para evitar complejidades adicionales, estos casos fueron codificados como valores nulos[3].

Al construir un genoma de referencia es importante considerar que existen posiciones que, de forma natural, pueden ser heterocigotas en una población. Esto significa que en estas posiciones es común encontrar individuos con alelos distintos (uno de referencia y otro alternativo) en lugar de homocigotos para el alelo de referencia.

Por este motivo, el valor cero (que representa individuos homocigotos para el alelo de referencia) no siempre es el más frecuente en los datos codificados de los individuos de una población. Esta situación se refleja claramente en los resultados presentados en la tabla 5.2, donde se observa que la distribución de frecuencias no está dominada por el valor cero en ciertas posiciones genómicas.

5.5. Selección de *SNPs*

Debido a la gran cantidad de *SNPs* disponibles, con el objetivo de aligerar los costos computacionales decidimos reducir la dimensionalidad de los datos genómicos, haciendo una preselección de SNPs basada en diferentes criterio presentados a continuación.

5.5.1. Criterio de correlación individual

En primer lugar decidimos filtrar *SNPs* en base a la correlación que tienen con el fenotipo a predecir. Al hacer esto es verdad que podemos estar perdiendo aquellos *SNPs* que no tengan una influencia directa pero si en conjunto con otros, o bien que tengan una influencia no lineal (si bien esto último es difícil que pase pues mayoritariamente codificamos los *SNPs* con solo dos valores).

[3]Valores faltantes.

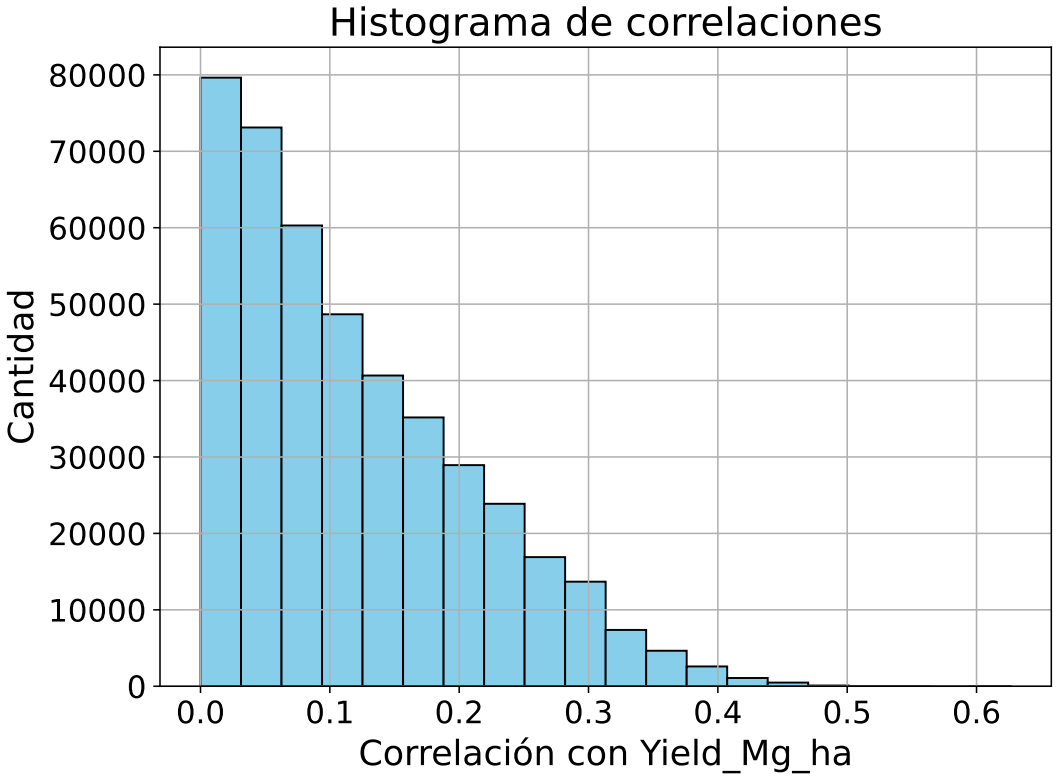


Figura 5.3: Histograma de correlaciones de los *SNPs* con el fenotipo a predecir.

En la figura 5.3, se presenta el histograma de correlaciones entre los *SNPs* y el fenotipo que se desea predecir. Este gráfico evidencia que la mayoría de los *SNPs* poseen una correlación muy baja con el fenotipo (el 49 % de ellos tiene una correlación menor a 0,1). Este comportamiento sugiere que es posible optimizar el modelo reduciendo la cantidad de *SNPs* utilizados, seleccionando únicamente aquellos que muestran una correlación significativa con la variable objetivo.

Por su parte, la figura 5.4 ilustra el análisis de *Pareto*[4], graficando el porcentaje de correlación total acumulada frente al porcentaje de *SNPs*. Este análisis permite identificar la distribución desigual de la correlación total entre los *SNPs*. En concreto, se observa que con tan solo la mitad de los *SNPs* más correlacionados, se acumula el 80 % de la correlación total. Esto último respalda la viabilidad de una selección eficiente de marcadores, para reducir su cantidad sin perder demasiada información.

En base a este análisis elegimos como umbral el valor de 0,3 de correlación mínima entre el *SNP* y el fenotipo, quedándonos así con 20763 de los 437214 *SNPs* originales.

Una observación importante de este proceso de reducción es que también permitió eliminar SNPs con poca varianza. En la figura 5.5, se muestra un histograma de la

[4]En un gráfico de *Pareto* se muestra que porcentaje del efecto total es provocado por un determinado porcentaje de las causas más significativas. En particular se marca que porcentaje de estas causas provoca el 80 % del efecto total.

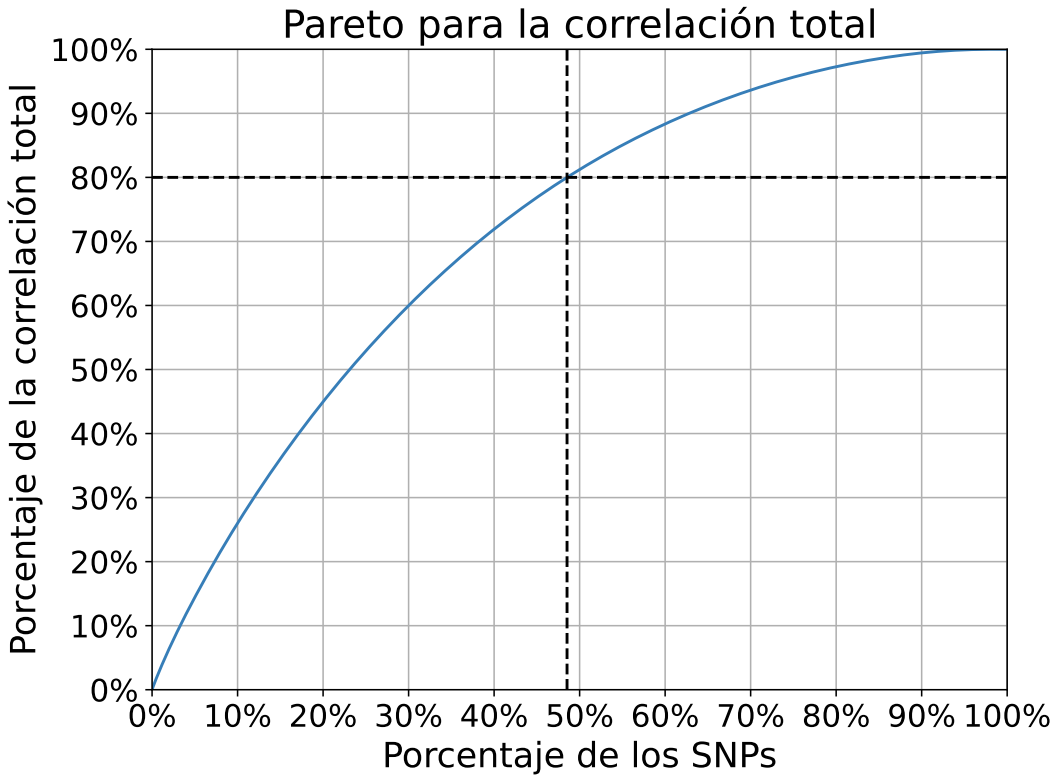


Figura 5.4: Gráfico de *Pareto*. Correlación total acumulada según el porcentaje de *SNPs* más correlacionados con fenotipo.

proporción del valor mayoritario para cada SNP, calculada como la frecuencia del valor más común dividido entre el total de valores (individuos genotipados). Es importante recordar que el mínimo valor posible en este análisis es $1/3$, ya que los *SNPs* fueron codificados utilizando tres valores (0, 1 y 2). Como se ilustra en la figura, existen *SNPs* con muy poca varianza, llegando a ser casi constantes en todos los individuos.

5.5.2. Criterio de valores nulos

Además del análisis basado en la correlación, se llevó a cabo un segundo análisis con el objetivo de identificar y descartar los SNPs con una alta proporción de datos nulos. Este análisis permitió optimizar aún más el conjunto de datos al reducir la presencia de información incompleta.

La figura 5.6 muestra la distribución de porcentajes de valores nulos por SNP. Se observa que la gran mayoría de los SNPs tienen pocos o ningún dato faltante. En particular, el 91 % de los SNPs presenta menos del 5 % de valores nulos. Esto sugiere que la mayor parte de los datos genómicos disponibles es íntegra.

Para reforzar esta conclusión, la figura 5.7 ilustra un análisis de *Pareto*, en el cual se relaciona el porcentaje acumulado de valores nulos con el porcentaje de SNPs. Este

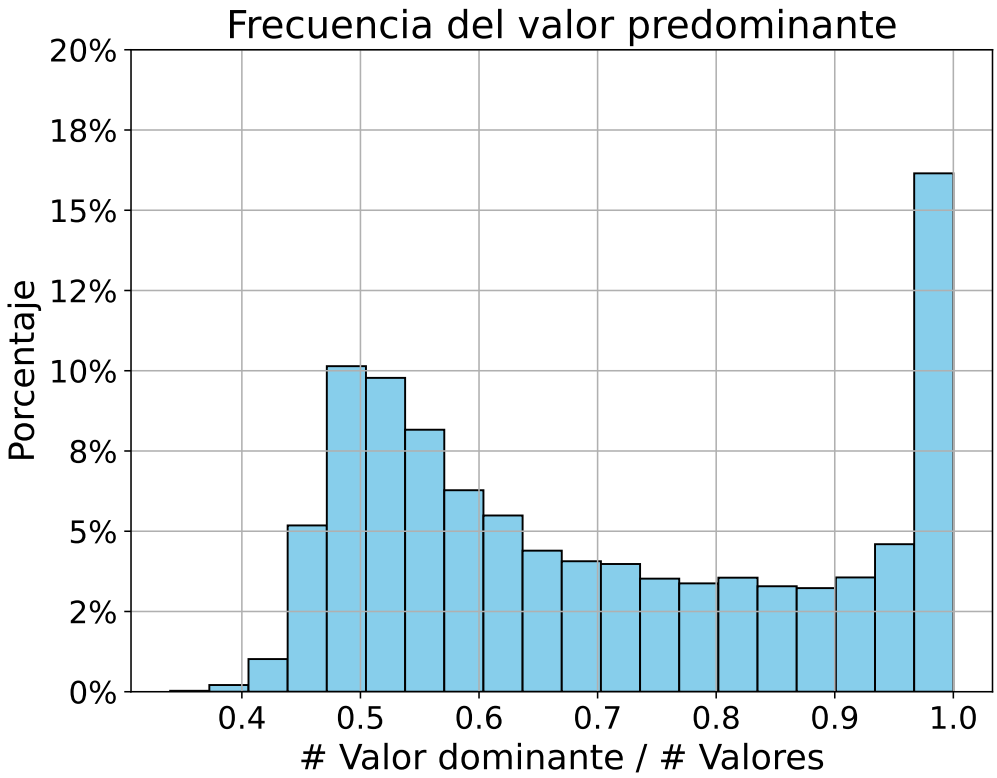


Figura 5.5: Histograma de la proporción del valor mayoritario en los *SNPs*. Cantidad de veces que aparece el valor más frecuente, entre la cantidad total de valores, o sea entre la cantidad de individuos genotipados.

gráfico revela que el 80 % de los datos faltantes totales se concentran en tan solo el 15 % de los SNPs. Por lo tanto, al descartar esta pequeña proporción de SNPs, se logra eliminar la mayoría de los datos nulos sin afectar significativamente la cantidad de marcadores disponibles.

Finalmente, la figura 5.8 muestra cómo varía el porcentaje de *SNPs* utilizables en función del umbral máximo permitido de valores nulos. Este gráfico demuestra que, incluso si limitamos la selección a los *SNPs* con una proporción muy baja de datos faltantes, aún disponemos de la mayoría de los marcadores originales.

En base a estos resultados y considerando que el principal recorte aplicado fue el basado en la correlación, se decidió establecer un umbral máximo del 10 % de valores nulos por *SNP*. Esta decisión permitió retener 407800 *SNPs* de los 437214 originales. Los valores faltantes en los *SNPs* seleccionados (aunque en baja proporción) fueron imputados utilizando el valor más frecuente para cada marcador.

Al aplicar los dos filtros, reducimos la cantidad a 18902 *SNPs*, lo que significa una reducción al 4,3 % de la cantidad original.

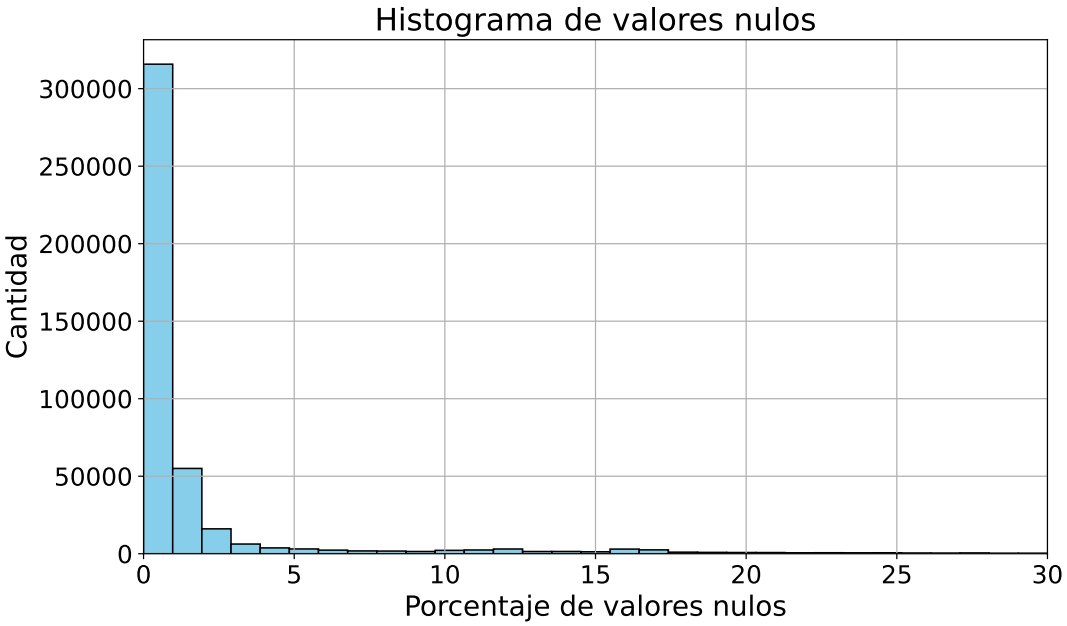


Figura 5.6: Histograma de porcentajes de valores nulos de cada *SNP*.

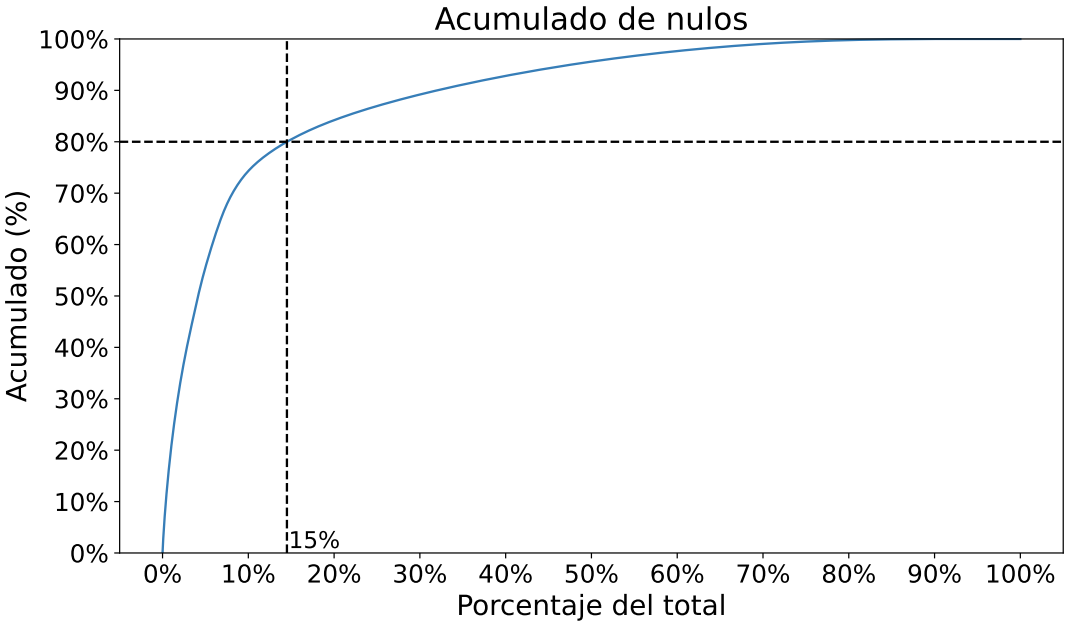


Figura 5.7: Gráfico de *Pareto*. Muestra qué porcentaje del total de datos faltantes totales, alcanzamos con un determinado porcentaje de los *SNPs* con más datos faltantes.

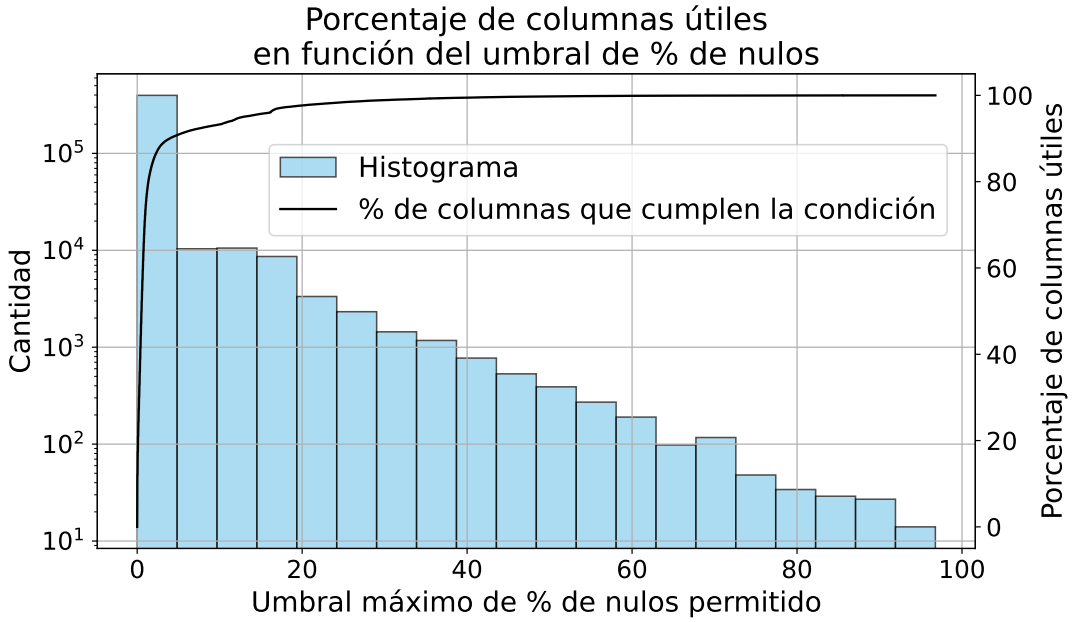


Figura 5.8: En el eje derecho y con linea negra, vemos el porcentaje de *SNPs* que podemos usar en base al umbral máximo de valores nulos que permitimos. En el eje izquierdo y con barras celestes está el histograma de porcentajes de valores nulos de cada *SNPs*, pero esta vez en escala logarítmica.

5.5.3. Criterio de correlación mutua

Aunque la reducción inicial fue suficiente para aliviar significativamente los cómputos, persistía un problema importante: la alta correlación entre algunos *SNPs*. Esto se traduce en la presencia de marcadores genómicos que son muy similares, incluso idénticos en algunos casos (de los 18902 seleccionados, solo hay 12481 únicos). Aún removiendo los duplicados, sigue habiendo correlaciones muy altas. Esto probablemente esté asociado con un desequilibro de ligamiento en la población con la que estamos trabajando.

Podemos sustentar estas afirmaciones con la figura 5.9. A la izquierda, se grafica el histograma de las correlaciones máximas para cada *SNP*. Los resultados reflejan que para casi todos los *SNPs*, existe al menos otro marcador que es casi idéntico, con correlaciones cercanas a uno. A la derecha graficamos el histograma de las entradas de la matriz de correlación entre *SNPs* sin contar los unos de la diagonal ni los *SNPs* repetidos. Se puede ver como en general los *SNPs* tienden a tener correlaciones significativas (mayores a 0,2) entre ellos.

Inspirados en este análisis, se tomó la decisión de establecer un umbral de correlación máxima de 0,90. Esto significa que cualquier SNP con una correlación superior a este umbral con otro marcador ya incluido fue descartado. Este criterio redujo la cantidad final de *SNPs* a 5144, lo que equivale a solo el 1,18 % del tamaño original de la base de datos.

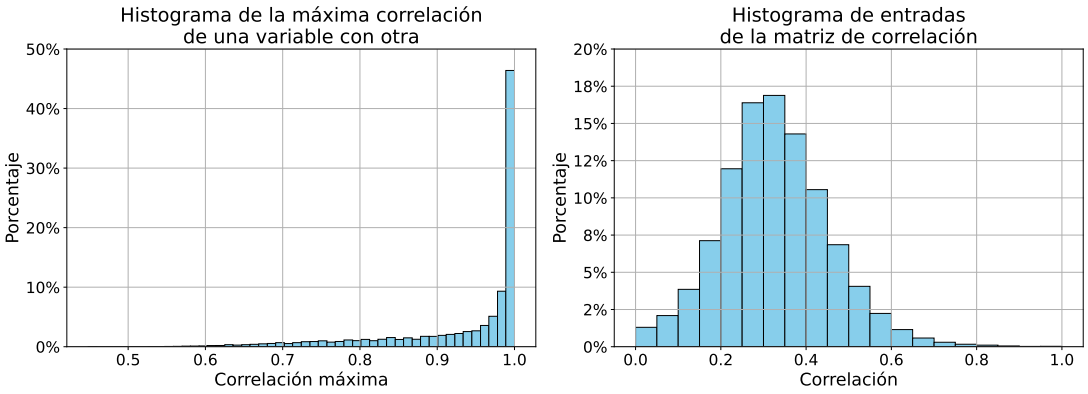


Figura 5.9: A la izquierda, histograma de las correlaciones máximas para cada *SNP*, es decir que el conteo se realiza sobre el vector resultante de obtener el máximo de cada columna de la matriz de correlaciones, sin contar los unos de la diagonal, ni los *SNPs* repetidos. A la derecha, histograma de las entradas de la matriz de correlación entre *SNPs* (nuevamente sin contar los unos de la diagonal).

5.6. Procesamiento de los datos fenotípicos

El conjunto de datos fenotípicos también requirió un breve preprocesamiento que detallaremos a continuación.

Fenotipos faltantes

En primer lugar, descartamos aquellos registros donde no teníamos disponible la medición del fenotipo de *Yield_Mg_Ha*. Pasando así de los 136012 registros originales a 128167 (lo cual no fue una pérdida significativa).

Híbridos repetidos en el mismo ambiente

En segundo lugar, decidimos promediar el desempeño de los clones que estuviesen plantados en el mismo ambiente, haciendo que el par híbrido-ambiente sea único. Esto lo hicimos porque dos clones plantados en el mismo ambiente tienen los mismos datos genómicos y los mismos datos ambientales. Por lo tanto, frente a nuestros modelos quedarían dos muestras idénticas con distinto valor a predecir, lo que puede introducir ruido y dificultar la capacidad de generalización. De esta manera pasamos de tener 128167 a 84527 registros (aproximadamente una reducción a dos tercios).

Lo anterior, lo podemos observar en la figura 5.10 donde se grafican las ocurrencias de híbridos repetidos en un mismo ambiente, por ejemplo, la primera columna del histograma refleja que hay más de 10^4 experimentos únicos, mientras que la última columna muestra que hay un experimento que se repite 20 veces. En general, la mayoría de los experimentos son únicos (55 %), luego hay un 41 % que están duplicados, y solo un 4 % que aparece más de dos veces.

Como promediamos a nivel de experimento, es interesante analizar la variabilidad que hay entre los resultados de experimentos repetidos, para evaluar la posible pérdida

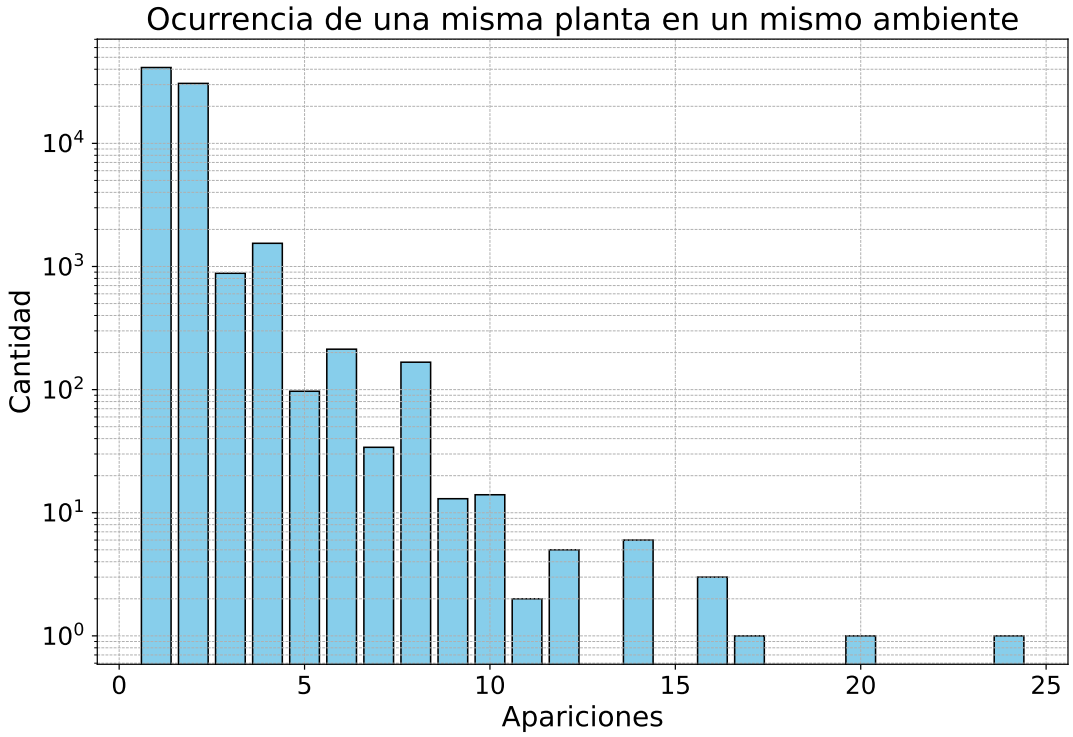


Figura 5.10: Histograma de la ocurrencia de experimentos. Con el eje vertical en escala logarítmica.

debido al promediado.. En la figura 5.11 se grafica un histograma de las desviaciones relativas de cada uno de los conjuntos de experimentos, excluyendo a los experimentos únicos. Se puede observar que no hay una desviación significativa en el desempeño de los fenotipos. Esto tiene sentido, ya que al plantar los mismos clones en el mismo lugar su producción debería ser similar.

5.7. Construcción de la base de datos

Luego de haber hecho los preprocesamientos correspondientes para cada fuente de datos, creamos un único conjunto con todos los datos disponibles. Para crear este conjunto final, se tomaron como base los registros de fenotipos, los cuales se cruzaron con los datos genómicos utilizando el nombre del híbrido, y con los ambientales según el ambiente en que estuvo la plantación. Sin embargo, no todos los híbridos estaban genotipados, ni todos los ambientes tenían datos disponibles de clima, suelo y covariables ambientales.

En el caso de los híbridos no genotipados, no quedó otra opción más que descartar sus datos asociados. En cuanto a los datos ambientales, se adoptó la siguiente estrategia: cuando un ambiente no contaba con información disponible, se examinaba la locación correspondiente (es decir, el mismo lugar geográfico pero en años distintos). Si esa locación tenía datos para al menos la mitad de los años en que se cultivó, se imputaba el

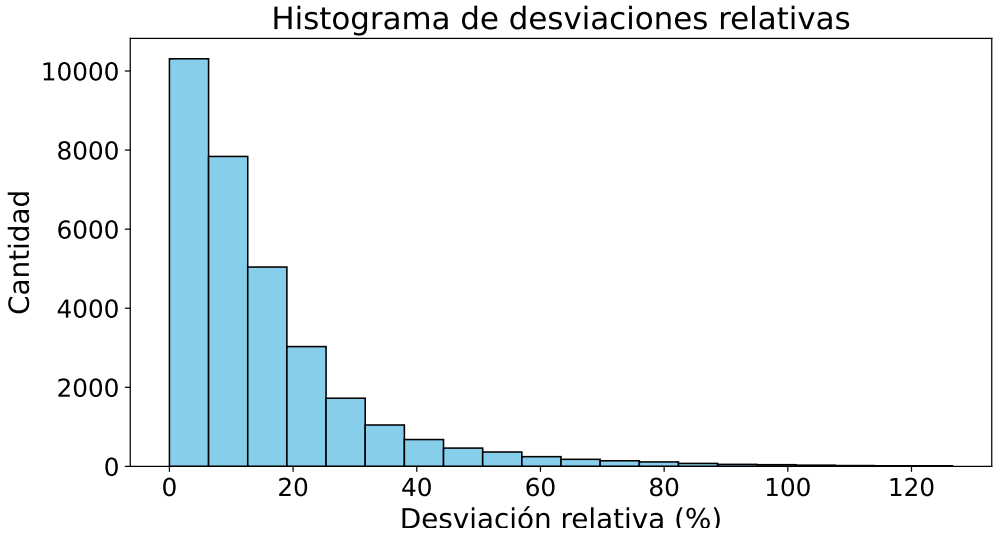


Figura 5.11: Histograma de las desviaciones relativas de cada uno de los conjuntos de experimentos. Sin tener en cuenta los experimentos únicos.

dato faltante del año correspondiente utilizando la información del año más cercano que sí estuviera disponible. Esta imputación se fundamenta en la hipótesis de que las condiciones ambientales no presentan variaciones significativas entre años próximos, hecho que verificamos posteriormente.

En las figuras 5.12 y 5.13 se muestra la proporción de datos faltantes segmentando por locación, año y fuente. De las mismas se puede ver que hay locaciones que no disponen de datos para casi ningún año. Por ejemplo, *IAH1a*, *IAH1b* y *IAH1c* en la figura 5.12 de la izquierda se ve que tienen una proporción de datos faltantes cercana a uno para los datos de suelo. Por lo tanto, para este tipo de locaciones no se pudieron imputar los datos faltantes y hubo que descartarlas.

Una vez depurada la base de datos, nos quedamos con 82534 registros[5], dentro de los cuales hay 4920 híbridos únicos y 210 ambientes, ubicados en 34 locaciones a lo largo de 9 años. Además del fenotipo, cada registro tiene asociadas 6385[6] variables, que las podemos desagregar en:

$$\underbrace{5144}_{SNPs} + \overbrace{\underbrace{488}_{\text{Clima}} + \underbrace{28}_{\text{Suelo}} + \underbrace{765}_{\text{Cov.}}}^{\text{Variables ambientales}}.$$

Cabe aclarar que tras cruzar los distintos conjuntos, aquellos registros correspondientes a un mismo híbrido, comparten datos genómicos, y por otro lado, aquellos correspondientes a un mismo ambiente, comparten datos ambientales.

[5]Que ahora si son una combinación única de híbrido y ambiente.

[6]Suponiendo que como datos climáticos usamos el resumen estadístico de las series.

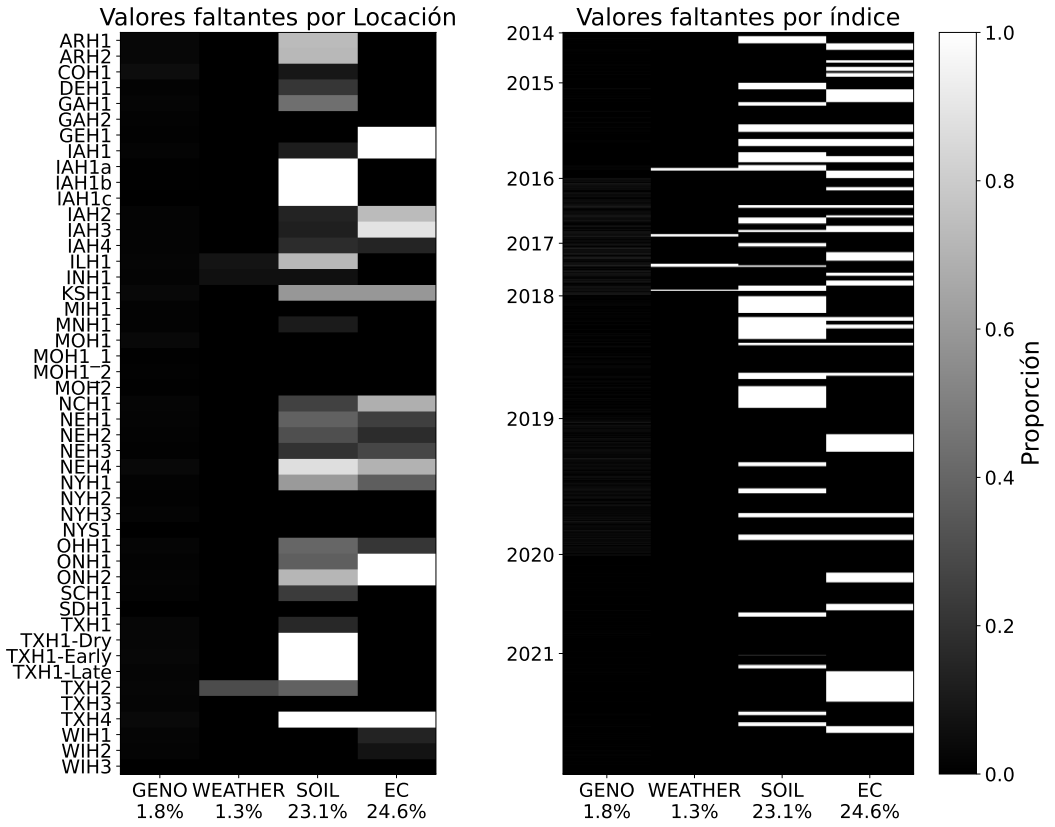


Figura 5.12: A la izquierda, proporción de datos faltantes (escala de color) por locación y por fuente; a la derecha lo mismo pero ordenando los datos por índice y marcando el año en el eje vertical. En ambas, se muestra bajo el nombre de cada fuente, el porcentaje de datos faltantes de estas, considerando la totalidad de la base de datos.

5.8. Análisis de los datos combinados

Conteos básicos

Como primera aproximación a entender los datos con los que íbamos a trabajar se observó la cantidad de muestras por año y por locación. En la figuras 5.14 y 5.15 se presentan la cantidad de registros disponibles por año, y por locación, respectivamente.

Análisis de híbridos y locaciones únicas

Debido a que cada muestra de los datos filtrados es una combinación única de híbrido y ambiente, se analizó cómo evolucionan las locaciones e híbridos usados para las plantaciones a lo largo de los años.

Primero analizamos en que años fueron usadas las distintas las locaciones. En la figura 5.16, se muestra una matriz de aparición de locaciones a lo largo de los años, en blanco se indica si una locación se usó en determinado año, y en negro si no. Se puede

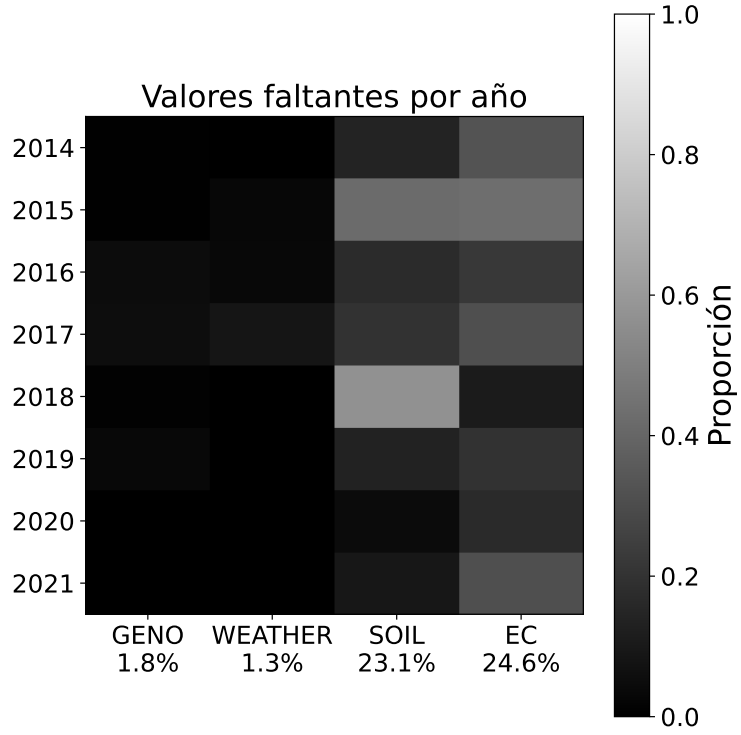


Figura 5.13: Figura análoga a 5.12 pero segmentando por año.

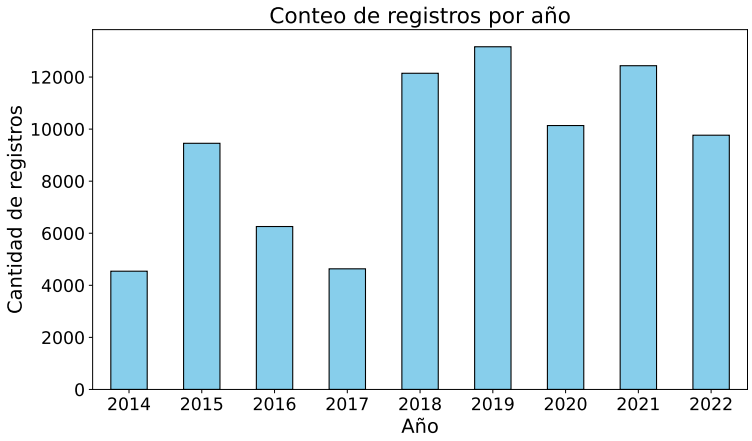


Figura 5.14: Cantidad de registros disponibles por año.

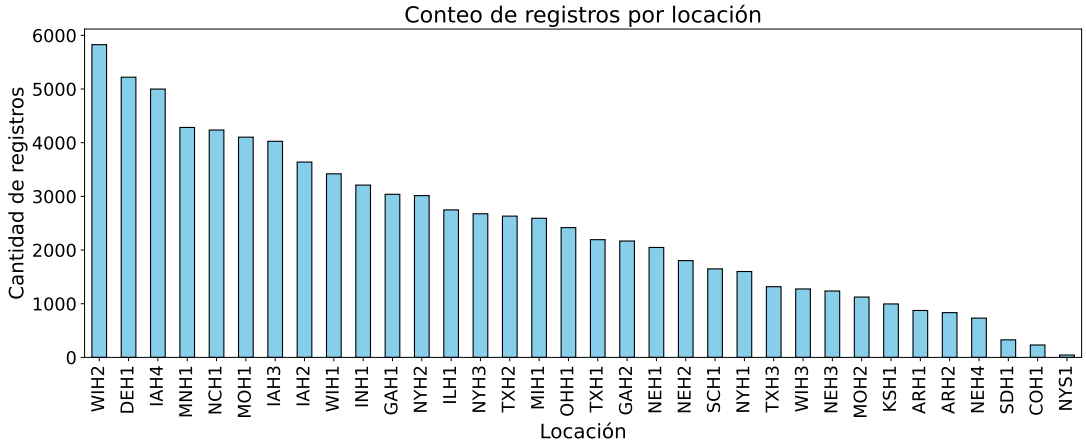


Figura 5.15: Cantidad de registros por locación.

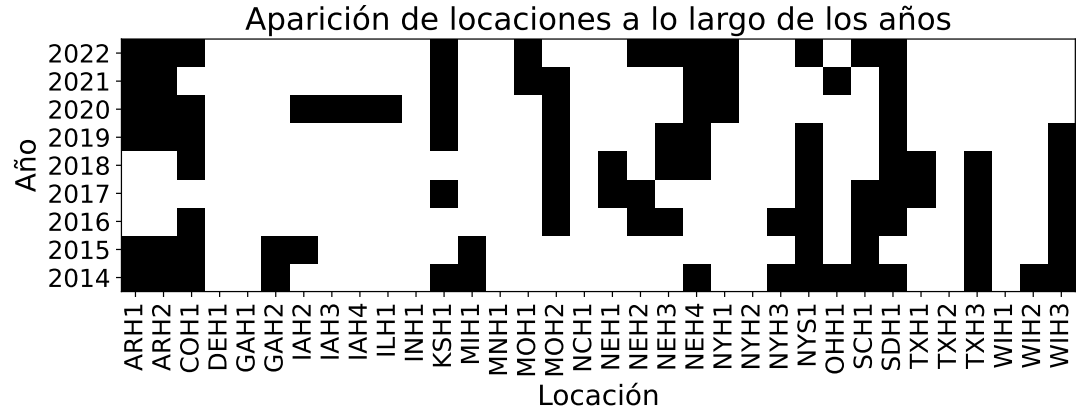


Figura 5.16: Matriz de aparición de las locaciones. Se marca en blanco si una locación se usó en un determinado año, y en negro si no.

ver que son pocas las locaciones en las que se sembró todos los años. La mayoría de locaciones dejan de ser usadas en ciertos años, aunque pueden volver a ser utilizadas luego.

A partir de este análisis, resultó interesante examinar de manera más cuantitativa cuáles locaciones se compartían entre los diferentes años. En particular, para evaluar la estabilidad en el uso de locaciones a lo largo de años consecutivos

En la figura 5.17 se muestra a la izquierda una matriz de intersección que muestra la cantidad de locaciones en común entre dos años y a la derecha la matriz normalizada. Se observa que si bien hay variación de locaciones a lo largo de los años, no hay gran variación de ellas, sobre todo entre años cercanos (se puede ver en la matriz de la derecha, que tiene valores cercanos a uno cerca de la diagonal). Además se ve que la cantidad de locaciones por año se mantiene siempre en el mismo orden.

Al realizar el mismo análisis para los híbridos se descubrió un comportamiento clave para el entendimiento de los datos. En la figura 5.18 se muestran las mismas matrices que

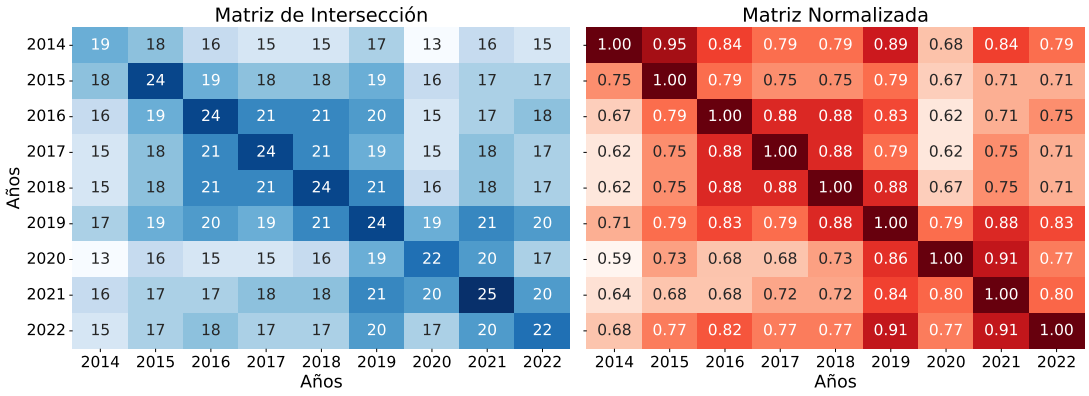


Figura 5.17: A la izquierda, tabla cruzada que muestra la cantidad de locaciones en común entre dos años. En la diagonal se muestra la cantidad de locaciones utilizadas en un determinado año. A la derecha tabla cruzada que muestra la proporción de locaciones de un año (eje horizontal) que aparecen en otro (eje vertical). A diferencia de la primera, esta no es necesariamente simétrica.

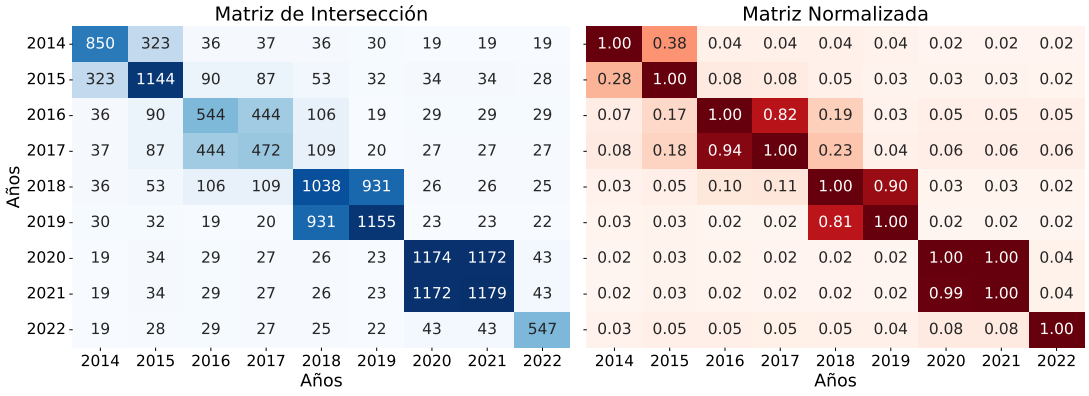


Figura 5.18: Mismas tablas que en 5.17 pero analizando híbridos en lugar de locaciones.

antes pero para los híbridos en vez de las locaciones. Se puede apreciar que cada dos años se renuevan la mayoría de híbridos con los que se experimenta. Esto se ve claramente en la tabla de la derecha donde aparecen bloques de largo dos sobre la diagonal. Más precisamente se puede ver que en los años pares se cambian los híbridos, y ademas una vez se remplazan no se vuelven a usar en años posteriores. A su vez, se ve que la cantidad de híbridos distintos usados en cada generación (año par y su consecutivo impar), tampoco es constante.

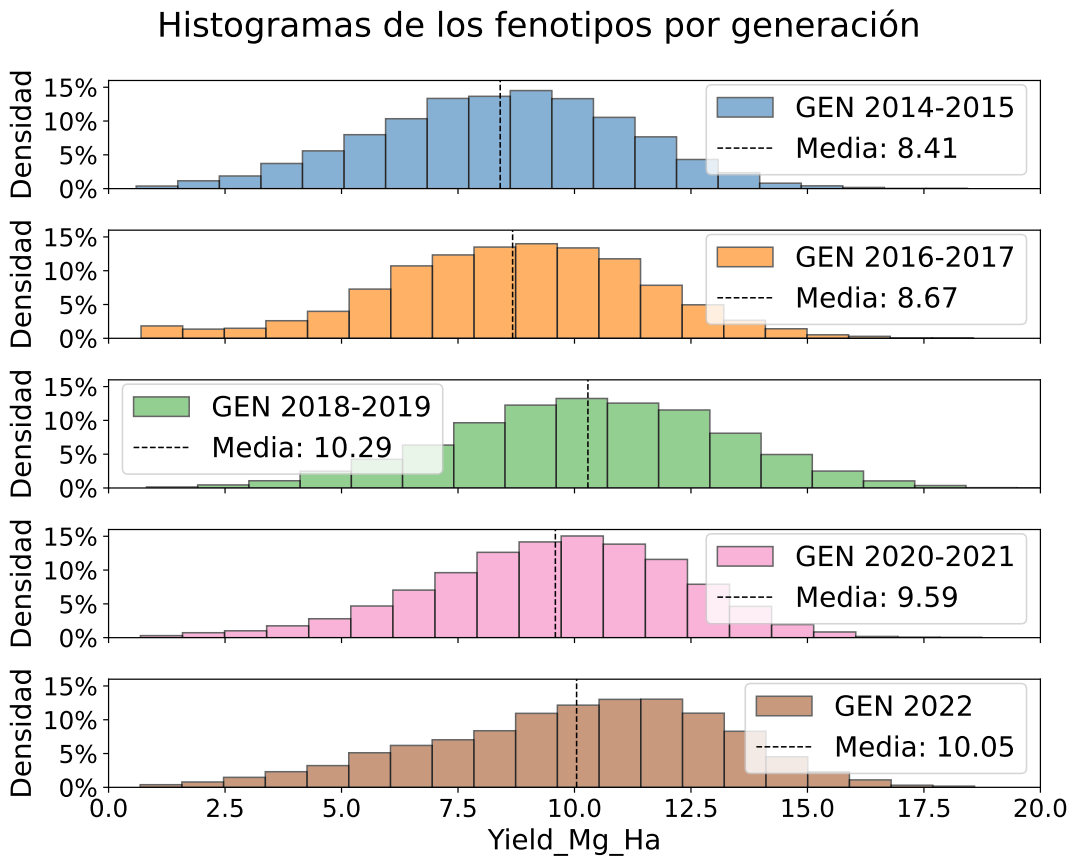


Figura 5.19: Histograma de los tonelajes de cada experimento, separado por generación. Se marca con una linea punteada la media en cada caso.

Fenotipo por generación

Al descubrir la existencia de las generaciones de híbridos, decidimos investigar si el desempeño de las generaciones variaba sustancialmente.

Al observar los histogramas de la figura 5.19 se observa como a partir de la tercera generación (a partir del recambio de híbridos realizado en 2018) el desempeño de las plantaciones tuvo un incremento, llegando a subir más de tonelada y media por hectárea. Esto da la pauta de que si quisiéramos predecir a partir de 2018, no sería conveniente usar años previos al mismo para modelar, pues poseen distribuciones distintas.

5.9. Análisis exploratorio mediante *PCA*

5.9.1. Sobre los datos genómicos

A modo de análisis exploratorio, y con el objetivo de buscar patrones, decidimos hacer una representación gráfica de los híbridos a partir de información genómica. Para esto tomamos los datos genómicos ya filtrados y aplicamos una reducción de dimensionalidad

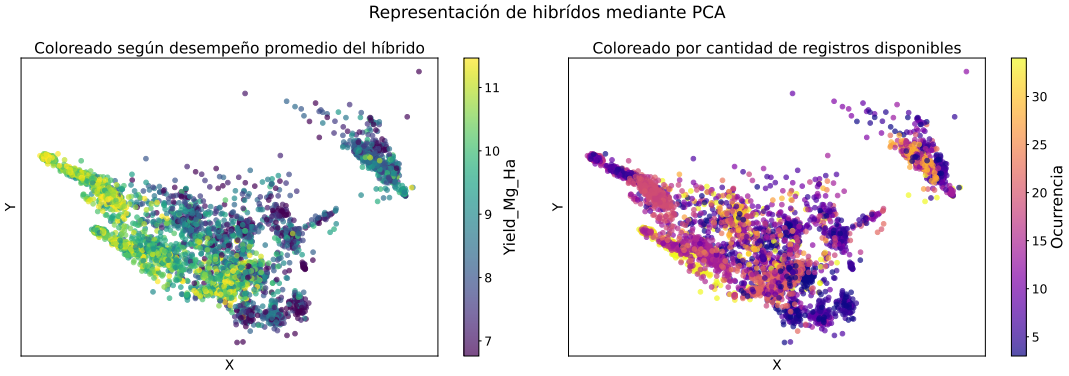


Figura 5.20: Representación en dos dimensiones de los híbridos (mediante sus datos genómicos), donde cada punto representa el genoma de un híbrido, coloreados según distintos criterios; izquierda, desempeño promedio del híbrido (fenotipo a predecir); derecha, cantidad de muestras disponibles de ese híbrido.

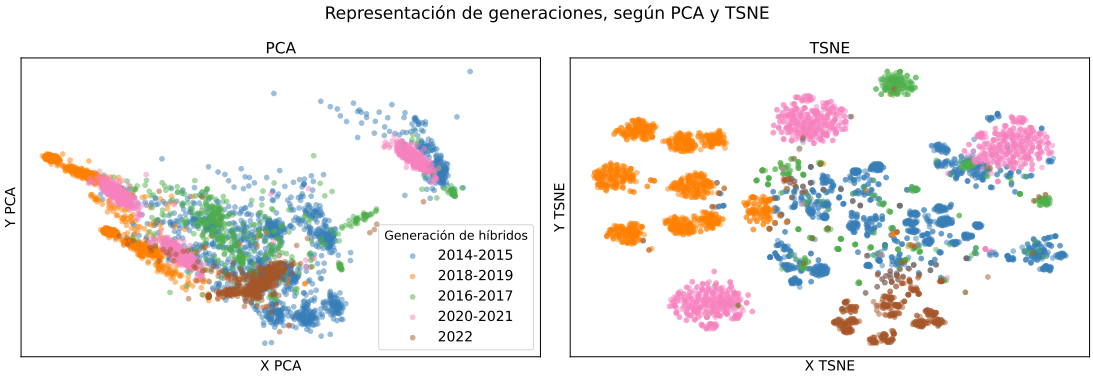


Figura 5.21: Representación en dos dimensiones de los híbridos (mediante sus datos genómicos), utilizando *PCA* (izquierda) y *TSNE* (derecha), coloreando según generación.

a dos dimensiones mediante *PCA* [13], capturando así el 34 % de la varianza total.

En la figura 5.20 se muestra la representación en dos dimensiones de los híbridos donde cada punto representa el genoma de un híbrido. Se ve como existen dos grupos bien definidos, aunque uno con muchos más híbridos que el otro. También vemos que los que tienen mejor desempeño (gráfico de la izquierda), están en el grupo mayoritario, y orientados sobre un extremo del mismo. Esta última observación nos alienta a pensar que los datos genómicos contienen información útil para predecir el desempeño. Respecto a la cantidad de muestras de los híbridos (gráfico de la derecha), no parece tener correlación ni con los grupos ni con el desempeño.

Otro análisis que hicimos fue, utilizando la misma técnica, analizar qué tan diferentes eran las generaciones de híbridos.

Como se ve en 5.21, claramente se forman *clusters* en base a las generaciones. El *PCA* que es una técnica lineal ya puede identificarlos, pero utilizando *TSNE* [41] que es capaz de detectar relaciones no lineales, quedan aún más expuestos, incluso separando a la primera generación del resto, algo que *PCA* no hacía.

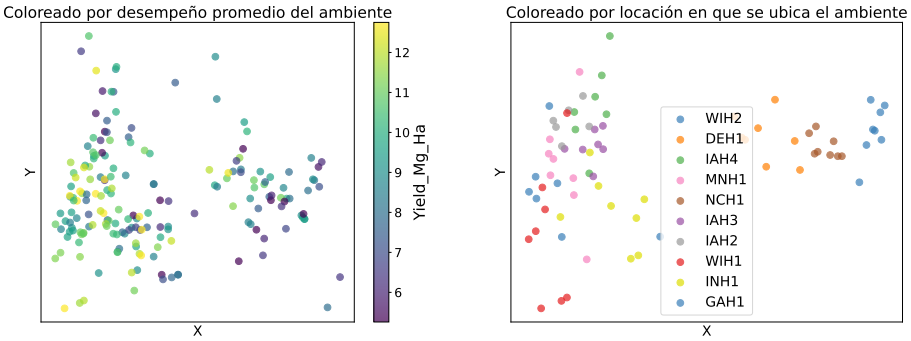


Figura 5.22: Representación en dos dimensiones de los datos genómicos, donde cada punto representa el genoma de un híbrido, coloreados según distintos criterios; izquierda, desempeño promedio del ambiente; derecha, locación en que se ubicó el ambiente. Considerando solo las diez locaciones con más muestras (las cuales acumulan la mitad de las muestras).

Este comportamiento significa que las generaciones de híbridos son sustancialmente diferentes entre si. Esto hace que el trabajo de comprender como afecta el genoma al fenotipo, sea aún más difcil, porque los *SNPs* de cada generación se estructuran de una manera diferente.

5.9.2. Sobre los datos ambientales

Se replicó el anterior análisis, pero ahora representando cada ambiente mediante sus datos ambientales. Ahora la reducción capturó solo un 17 % de la varianza total. La idea del análisis era verificar si los ambientes más productivos se ubicaban cercanos en el espacio reducido.

Sin embargo de la gráfica de la izquierda de 5.22 se ve que no hay una clara correlación entre posición y rendimiento. Aunque esto claramente no desvalida la utilidad de los datos ambientales.

En cuanto a la gráfica de la derecha, tenía como objetivo verificar si las condiciones ambientales de una misma locación se mantenían similares a lo largo de los años, lo cual nos serviría para validar nuestro método de imputado descrito anteriormente. De la misma vemos como efectivamente las locaciones tienden a mantenerse cercanas en el espacio reducido, lo cual significa que las condiciones ambientales no varían demasiado entre años. O dicho de otra manera, afecta más a la componente ambiental variar entre locaciones en un mismo año, que variar años en una misma locación.

5.10. Resumen

Debido a la extensión de esta sección, decidimos hacer un resumen de la misma. A continuación, se listan los puntos más importantes para entender la estructura de los datos, así como los descubrimientos más relevantes que surgieron del análisis de los mismos.

- Después del preprocesamiento, obtuvimos 82534 registros, cada uno de los cuales corresponde a una plantación de un determinado híbrido, en un determinado ambiente.
- El fenotipo asociado a cada registro, que es el que intentaremos predecir luego, es el tonelaje de la plantación medido en toneladas por hectárea (*Yield_Mg_Ha*).
- Los ambientes se sitúan temporalmente entre 2014 y 2022, y geográficamente en la zona este de Estados Unidos, teniendo elevaciones desde los cero a los mil metros por sobre el nivel del mar.
- Si había dos plantaciones en el mismo ambiente y con el mismo híbrido, la convertimos en una única plantación promediando su tonelaje, de modo que la combinación ambiente e híbrido sea única.
- Para cada ambiente corresponde una serie de datos ambientales, mientras que para cada híbrido una de datos genómicos. Por lo tanto, cada registro tenemos una combinación única de datos ambientales y genómicos.
- Los datos genómicos originalmente eran un conjunto de 437214 *SNPs* que codificamos como 0, 1 o 2. Luego de filtrarlos según porcentaje de nulos, correlación mutua y correlación con el fenotipo, quedaron 5144 SNPs.
- Los datos ambientales se dividen en tres, los del suelo, las covariables ambientales y las series temporales climáticas. Además, creamos un resumen estadístico de las series, tomando diversos estadísticos de las mismas para cada estación del año.
- Estos datos ambientales son en general más sensibles al cambio geográfico que al temporal. Esto lo usamos para imputar valores faltantes utilizando los de años anteriores.
- Las ubicaciones en que se realizan las plantaciones no suelen cambiar entre años. Sin embargo, cada dos años se cambia el conjunto de híbridos a sembrar. Es decir que existen generaciones de híbridos. Además, estas generaciones son muy diferentes genéticamente.

Esta página ha sido intencionalmente dejada en blanco.

Capítulo 6

Esquemas de validación

En el contexto de la predicción genómica, es especialmente importante el diseño de los conjuntos de validación, pues la forma en que se arman estos conjuntos afecta directamente a los resultados obtenidos.

Cuando se separa un conjunto para validación, se hace con el objetivo de ver si el modelo es capaz de generalizar sobre datos no vistos durante el entrenamiento. Sin embargo, los datos están compuestos de una componente genética y una ambiental. Esto da matices a la separación, pues puede que el modelo no haya visto la combinación de ambas, pero sí alguna de ellas. Es razonable entonces, que los modelos desempeñen mejor si ya conocen alguna de las componentes de los datos de validación.

6.1. Esquemas a utilizar

Decidimos utilizar dos esquemas de validación diferentes para nuestros experimentos, cada uno de los cuales intenta responder a una pregunta diferente.

6.1.1. Híbridos desconocidos (*HD*)

Esta separación consiste en predecir híbridos desconocidos, en ambientes conocidos. Es decir, que el modelo, al momento de predecir, ya va a haber visto la componente ambiental, pero no la genómica.

Esta forma de entrenar al modelo busca que el mismo sea capaz de responder la pregunta de ¿Cómo le hubiese ido a este nuevo híbrido si lo hubiese plantado en alguna plantación pasada? El modelo observa cómo se desempeñaron algunos híbridos en determinados ambientes, y busca predecir cómo se hubiesen desempeñado nuevos híbridos en esos ambientes ya conocidos.

Sin embargo, en este esquema donde buscamos ver cómo el modelo generaliza sobre genomas no conocidos, es importante recordar la existencia de las generaciones de híbridos. Si un modelo es capaz de generalizar, lo va a lograr sobre la distribución de datos con los que se entrenó. Es por eso que si se entrena con híbridos de una generación no se debería esperar que el modelo sea capaz de generalizar sobre los de otra generación.

Esto es debido a que los híbridos son muy distintos entre generaciones, teniendo cada generación su propia distribución, como mencionamos en el capítulo anterior.

Por estas consideraciones es que decidimos utilizar, para este esquema de validación, solamente los años 2020 y 2021. Para crear los conjuntos de validación y entrenamiento, procedimos de la siguiente manera:

- Creamos dos conjuntos primarios de entrenamiento y validación, T y V respectivamente. T está formado con los registros de 2020 y 2021, y V solo con los de 2021.
- Llamémosle H al conjunto de híbridos únicos de V (1179 híbridos). De manera aleatoria, partimos H en cinco subconjuntos $H_{i=1,\dots,5}$ de aproximadamente el mismo tamaño.
- Para cada uno de los H_i, creamos un *fold* de validación V_i y otro de entrenamiento T_i.
    - Cada V_i se obtuvo seleccionando los registros de V en que aparecían los híbridos de H_i.
    - Cada T_i, se obtuvo eliminando los registros de T en que aparecían los híbridos de H_i.

Este esquema de validación, implementado de esta manera, tiene algunas consideraciones a tener en cuenta:

- Los *folds* de entrenamiento están formados por muestras de 2020 y 2021, mientras que los *folds* de validación, solo con muestras de 2021.
- Los V_i forman una partición de V.
- Si bien cada *fold* de validación tiene el mismo número de híbridos únicos, no necesariamente tienen el mismo número de muestras, pues no todos los híbridos aparecen en la misma cantidad de locaciones. Como los híbridos de cada conjunto fueron seleccionados al azar, estadísticamente, sería esperable que cada conjunto tenga aproximadamente el mismo tamaño.
- Eventualmente podría pasar que, al elegir al azar los híbridos de cada partición, haya alguna que tome todos los híbridos que había en una locación en 2021. Esto haría que el respectivo *fold* de entrenamiento no tenga muestras de esa locación en 2021, lo cual significaría que en validación habrían híbridos en ambientes desconocidos. Esto rompería con las premisas del esquema de validación y constituiría un problema más complejo.
- Sin embargo, lo anterior es poco probable, pues en el 20 % de híbridos de un H_i, deberíamos seleccionar (aleatoriamente) toda la variedad de híbridos que aparece en una locación. Dado que la locación de 2021 con menos variedad, tiene 21 híbridos, es poco probable que se genere un desbalance. De todas maneras, verificamos al momento de la generación de conjuntos que esto no pasase.

	# Entrenamiento	# Validación	% Validación
Fold 1	17983	2534	12.35
Fold 2	18159	2428	11.79
Fold 3	17984	2524	12.31
Fold 4	18109	2453	11.93
Fold 5	18062	2496	12.14

Tabla 6.1: Cantidad de muestras para cada *fold* de entrenamiento y validación, junto al porcentaje del total que representan las muestras de validación.

Después de generar los conjuntos, los tamaños (en términos de muestras) fueron los de la tabla 6.1.

De 6.1 se ve que, confirmando la hipótesis, la cantidad de muestras por *fold* se mantiene aproximadamente constante, así como la proporción de las mismas entre validación y entrenamiento.

6.1.2. Ambientes desconocidos (*AD*)

En contraposición al esquema anterior, esta separación consiste en predecir híbridos conocidos, en ambientes desconocidos. Esta vez el modelo, al momento de predecir, va a haber visto la componente genómica, pero no la ambiental.

Como la componente ambiental depende de la locación y el año, hay varias formas de realizar este esquema, y cada forma intenta responder a una pregunta diferente.

- Usar para validación locaciones conocidas pero en años desconocidos. Es decir, ¿Cómo desempeñará este híbrido, que ya planté previamente en este lugar, el año que viene?
- Usar locaciones desconocidas, pero en años conocidos. Es decir, ¿Cómo hubiese rendido esté híbrido, que este año cultivé en determinados lugares, si lo hubiese plantado en otros distintos?
- Usar locaciones desconocidas en años desconocidos. Lo que equivale a ¿Cómo desempeñará este híbrido, que ya planté previamente en determinados lugares, el año que viene, en otros lugares distintos?

Se observa que siempre que se habla de predecir en años futuros, se asume que se conocen las variables ambientales de los mismos. Esto no es realista en la práctica, donde las mismas son inciertas. Esta es una desventaja de usar la componente ambiental, pues si quisiéramos hacer predicciones para años próximos, deberíamos hacer un trabajo aparte para inferir las condiciones ambientales futuras. Esto último queda totalmente fuera del alcance de este proyecto.

Si bien para una determinada locación no hay variables ambientales que sean constantes entre años, conocer la componente ambiental de la misma para años previos puede brindar algún tipo de información de cómo se comporta la locación. De la misma manera, aunque dado un año, no hay variables ambientales que sean constantes entre locaciones,

conocer la componente ambiental del mismo para otras locaciones ayuda a comprender cómo se comportó el año, sobre todo climáticamente.

Es por eso que las dos primeras variantes mencionadas resultan menos desafiantes que la última. A su vez, como mencionamos en el capítulo anterior, la componente ambiental se ve más afectada por el cambio de locación que por el de año. Por lo tanto, es esperable que la primera variante resulte más sencilla que la segunda. En este proyecto, por simplicidad y porque era la más alineada con la competición de la cual obtuvimos los datos, elegimos implementar la primera variante del esquema.

En este caso, para crear los conjuntos de validación y entrenamiento, procedimos de la siguiente manera.

- Creamos dos conjuntos primarios, de entrenamiento y validación V y T. Esta vez formamos T a partir de los registros de 2019 y 2020, y V con los de 2021.
- Filtramos el conjunto V, de manera que no aparezcan híbridos ni locaciones que no estén en T. De esta manera nos aseguramos que el único factor desconocido en validación es el año.
- Nuevamente definimos el conjunto H, formado por los híbridos únicos de V (ahora 1172 híbridos).
- Al igual que para el esquema anterior, obtuvimos los $H_{i=1,\dots,5}$ partiendo H de manera aleatoria en subconjuntos del mismo tamaño.
- Para cada uno de los H_i, volvimos a crear un *fold* de validación V_i y otro de entrenamiento T_i.
    - Cada V_i se obtuvo seleccionando los registros de V, en que aparecían los híbridos de H_i.
    - Cada T_i se obtuvo seleccionando los registros de T, en que aparecían los híbridos de H_i; y se complementó con un 40 % de muestras aleatorias de lo que sobró de T.

Por la forma en que se implementó este esquema, resulta pertinente detallar algunas consideraciones a tener presente.

- Los *folds* de entrenamiento estan formados por muestras de 2019 y 2020, mientras que los *folds* de validación con muestras de 2021.
- Los conjuntos de validación V_i forman una partición de V.
- No necesariamente todas las muestras de T aparecen en un T_i[1].
- Cada T_i esencialmente se conforma de dos partes, una que comparte híbridos con V_i, y otra que complementa con otros híbridos que no están en V_i. A estas partes las nombramos S_i y C_i respectivamente, es decir, $T_i = S_i \cup C_i$.

[1]Esta casuística se puede dar en híbridos que aparezcan en T y no en V, y no hayan sido sorteados en ninguno de los complementos.

	$\#S_i$			$\#C_i$			$\#V_i$
	2019	*2020*	*Total*	*2019*	*2020*	*Total*	*Total*
Fold 1	79	2035	2114	5267	3207	8474	2473
Fold 2	67	2011	2078	5261	3228	8489	2463
Fold 3	51	2044	2095	5302	3180	8482	2511
Fold 4	65	1926	1991	5277	3247	8524	2389
Fold 5	114	2112	2226	5281	3149	8430	2548

Tabla 6.2: Cantidad de muestras para cada *fold* de entrenamiento, segregado por año y separando en el conjunto seleccionado (el que comparte híbridos con validación) y el complemento. Junto a ello también se reportan la cantidad de muestras para los *folds* de validación, las cuales son todas de 2021.

- Como los híbridos de los H_i son de la generación 2020-2021, cuando seleccionemos los S_i, la mayoría de las muestras van a ser de 2020. Por lo tanto, las muestras de 2019 de los *folds* de entrenamiento aparecen principalmente en los complementos C_i.

- Incluir datos de 2019 en entrenamiento hace que tengamos dos distribuciones genómicas diferentes (una por cada generación). Pero por otro lado también da mayor riqueza a la distribución de variables ambientales, lo que ayuda a generalizar sobre ambientes desconocidos.

- Si no complementásemos con el 40 % extra, cada *fold* de validación tendría aproximadamente las mismas muestras que su par de entrenamiento. Se elige la cantidad de 40 % para lograr una proporción aproximada de uno a cuatro entre *folds* de validación y entrenamiento.

En las tablas 6.2 y 6.3 se ejemplifican, en base a los conjuntos generados, las distintas consideraciones que mencionamos previamente. Por ejemplo, cómo se logra tener aproximadamente un 20 % de muestras para validación, o cómo las muestras de 2019 aparecen principalmente en el complemento.

	# Entrenamiento	**# Validación**	**% Validación**
Fold 1	10588	2473	18.93
Fold 2	10567	2463	18.90
Fold 3	10577	2511	19.19
Fold 4	10515	2389	18.51
Fold 5	10656	2548	19.30

Tabla 6.3: Cantidad de muestras para cada *Fold* de entrenamiento y validación, junto al porcentaje del total que representan las muestras de validación.

Conjunto de test

Para este esquema, también creamos un conjunto de test, utilizando las muestras de 2022. Para mantener el paradigma del esquema deberíamos haber eliminado los híbridos o locaciones que no apareciesen en T. Para las locaciones fue posible hacer este filtro, sin embargo, para los híbridos no. Esto debido a que en 2022 hubo un cambio de generación, es decir, que de la mayoría de híbridos presentes en ese año, no teníamos histórico.

Creamos este conjunto con la idea de simular el funcionamiento de la competencia, en la cual se predice sobre el próximo año. Además, predecir sobre el año entrante es una práctica común en la industria, por lo que resultó interesante ver qué tan bien realizaban esta labor nuestros modelos.

La idea de utilizar estos dos esquemas de validación, y la forma en que los generamos, estuvo inspirada en el trabajo realizado por *I. K. Fernandes et al.* [12], el segundo puesto de la competencia. En la literatura, es normal encontrar el nombre *CV1* para referirse al primer esquema descrito, y *CV0* para el segundo.

6.2. Otros posibles esquemas de validación

Decidimos utilizar los dos esquemas previamente descritos porque con ellos abarcamos las dos vertientes del proyecto. Si bien en ambos esquemas está presente la interacción genoma-ambiente, con *HD* nos centramos en la influencia del genoma, y con *AD* en la del ambiente. Sin embargo, existen otros posibles esquemas de validación, que se centran en otras aplicaciones prácticas.

ND (Ninguno desconocido)

En este esquema, se busca predecir sobre híbridos conocidos, en ambientes conocidos, pero sin haber visto nunca la combinación. Es decir, conocemos cómo le fue a un híbrido en ciertas locaciones, pero queremos saber cómo le hubiera ido en otras donde no se plantó, pero sí se plantaron otras variedades.

Esta variante es la más sencilla, pues el modelo ya conoce la componente ambiental y la genómica por separado, y lo que tiene que generalizar es la interacción entre ambas.

HAD (Híbridos y ambientes desconocidos)

En este caso, se busca predecir híbridos desconocidos en ambientes desconocidos.

Nuevamente, al igual que con el esquema *AD* estándar, existen varias maneras de desconocer el ambiente. Generalmente se desconoce el año, es decir, se conoce el desempeño de las locaciones pero para años anteriores. De esta manera, la pregunta que se busca responder sería: ¿Cómo va a desempeñar este híbrido, que voy a plantar por primera vez, el año que viene?

En contraposición a *ND*, esta variante es la más complicada, pues el modelo desconoce tanto la componente ambiental como la genómica. Teniendo por lo tanto que ser capaz de generalizar sobre ambas a la vez, además de entender su interacción.

	H1	H2	H3	H4	H5	H6
A1	T	T	T	V	V	V
A2	T	T	T	V	V	V
A3	T	T	T	V	V	V
A4	T	T	T	V	V	V
A5	T	T	T	V	V	V
A6	T	T	T	V	V	V

HD

	H1	H2	H3	H4	H5	H6
A1	T	T	T	T	T	T
A2	T	T	T	T	T	T
A3	T	T	T	T	T	T
A4	V	V	V	V	V	V
A5	V	V	V	V	V	V
A6	V	V	V	V	V	V

AD

	H1	H2	H3	H4	H5	H6
A1	T	T	T	V	V	V
A2	T	T	T	V	V	V
A3	T	T	T	V	V	V
A4	V	V	V	T	T	T
A5	V	V	V	T	T	T
A6	V	V	V	T	T	T

ND

	H1	H2	H3	H4	H5	H6
A1	T	T	T			
A2	T	T	T			
A3	T	T	T			
A4				V	V	V
A5				V	V	V
A6				V	V	V

HAD

Figura 6.1: Representación matricial de los distintos esquemas de validación. Cada fila representa un ambiente y cada columna un híbrido. Las celdas azules representan las combinaciones que son usadas para el entrenamiento, las naranjas las que son usadas para la validación y las blancas las que no son usadas.

6.3. Representación mediante matrices

Una forma de mostrar cómo funciona cada esquema, es mediante una matriz de ambientes contra híbridos. En las matrices de la figura 6.1, cada fila representa un ambiente y cada columna un híbrido, por lo que cada entrada de la matriz sería una muestra. En ellas se representan, conceptualmente, particiones de entrenamiento y validación para cada esquema descrito.

Esta página ha sido intencionalmente dejada en blanco.

Capítulo 7

Modelos a utilizar

En este capítulo explicaremos los distintos modelos que utilizamos para la predicción genómica. Introduciremos las estructuras de cada modelo, explicando cómo transforman y utilizan las distintas fuentes de datos. También explicaremos, para cada uno de los modelos, la forma en que los entrenamos, qué métricas monitoreamos, la forma en que buscamos los hiperparámetros óptimos y cómo evitamos el sobreajuste.

7.1. Modelos Lineales Mixtos

En el capítulo 4 acerca del estado del arte se menciona que los modelos lineales son ampliamente utilizados en el área de predicción genómica, obteniendo muy buenos resultados. Es por eso que decidimos utilizarlos en nuestro proyecto con el fin de tener modelos que sirvan de referencia para compararlos con modelos más complejos de aprendizaje automático, como lo son los árboles de decisión y las redes neuronales.

Variantes del modelo

Para evaluar el rendimiento de un modelo entrenado utilizando datos multimodales, es fundamental explorar cómo las distintas fuentes de información contribuyen al desempeño de los modelos predictivos. Para eso, definimos las siguientes configuraciones basadas en las combinaciones de datos disponibles. Estas variantes permiten evaluar de manera sistemática el impacto de cada tipo de dato al hacer las predicciones.

- `GENO`: Se utilizan únicamente datos genómicos. Su objetivo es capturar la variación genética que influye directamente en los fenotipos.
- `WEATHER`: Se basa únicamente en los datos climáticos. Esta variante se centra en cómo las condiciones ambientales temporales afectan el rendimiento.
- `ENV`: Incluye solo datos ambientales, un conjunto más amplio que abarca información climática junto con las características del suelo y las covariables ambientales, proporcionando un panorama completo del entorno de crecimiento..

- `GENO_WEATHER`: Datos genómicos más datos climáticos. Permite modelar interacciones entre el genotipo y el clima.
- `GENO_ENV`: Datos genómicos más todos los datos ambientales. Esta configuración busca capturar interacciones complejas entre los genotipos y el entorno de crecimiento

Notar que en la primera variante, al utilizar solo datos genómicos, todos los registros correspondientes a un mismo híbrido van a tener las mismas variables predicadoras pero distintos fenotipos. Podríamos haber promediado a nivel híbrido, para tener tantos registros como híbridos y así evitar este problema. Sin embargo, hacer esto reduciría la varianza de los fenotipos, por lo que bajaría el orden del *RMSE* de las predicciones. Así que para mantener la comparabilidad entre variantes, decidimos no hacerlo.

Se puede razonar análogamente para las variantes que solo usan información ambiental y los registros que comparten ambiente.

GBLUP

En primera instancia, antes de realizar el análisis multimodal, decidimos hacer una serie de pruebas utilizando únicamente la información genómica. El objetivo de estas pruebas era ver si los *SNPs* seleccionados contenían información relevante para la predicción del fenotipo.

Para lo anterior utilizamos las librerías *sommer* y *rrBLUP* de R para implementar el modelo *GBLUP*. En el apéndice C, se muestra el análisis detallado donde se concluye que los datos genómicos tienen información relevante para la predicción del fenotipo. En este análisis se promedian los fenotipos por híbrido teniendo únicamente 4415 registros (la cantidad de híbridos únicos). Por lo tanto, no se incluye en el análisis multimodal al no ser comparables.

Para la variante `GENO` de los modelos lineales, utilizamos la librería *rrBLUP* de R que implementa *GBLUP* ya que en el análisis mencionado resultó mejor y más rápida que *sommer*.

Factor Analyitic Multiplicative Mixed Model

Para incorporar datos ambientales a los modelos lineales, nos basamos en la estructura de *Factor Analytic* para modelar la varianza entre ambientes motivados por algunos estudios mencionados en el capítulo 4 [9] [12].

El *Factor Analytic Multiplicative Mixed Model* (*FA*) es una extensión de los modelos lineales mixtos que permite modelar de manera eficiente la interacción genotipo-ambiente ($G \times E$) al aprovechar información ambiental y genómica simultáneamente. Este enfoque es particularmente útil en estudios de predicción genómica donde se desea capturar patrones complejos en los datos.

La estructura general de los modelos *FA* viene dada por:

$$Y = \mu + E + h(E, G) + \epsilon,$$

donde:

- Y_{ij}: Valor esperado ($BLUE$) del híbrido j en el ambiente i.
- μ: Promedio global o intercepto general.
- E_i: Efecto fijo del ambiente i, modela diferencias entre ambientes.
- $h(E)_{ij}$: Efecto aleatorio del híbrido j en el ambiente i, modela la interacción $G \times E$.
- e_{ij}: Término residual, modela las variaciones no explicadas asumiendo normalidad e independencia. Tiene distribución $\mathcal{N}(0, \mathbf{\Sigma}_e^2)$

El efecto aleatorio de la interacción $G \times E$, $h(E)_{ij}$, tiene distribución $\mathcal{N}(0, \mathbf{\Sigma}_g)$, donde $\mathbf{\Sigma}_g$ es la matriz de covarianza entre híbridos en distintos ambientes, definida como:

$$\mathbf{\Sigma}_g = FA_1 \otimes G,$$

siendo:

- FA_1: Una matriz *Factor Analytic* de orden uno. Captura la estructura de correlaciones entre los ambientes y permite reducir la dimensionalidad al modelar la interacción $G \times E$ mediante factores latentes.
- G: La matriz de relaciones genómicas (GRM). Representa las similitudes genéticas entre híbridos basada en marcadores genómicos.
- $\otimes$: El producto de *Kronecker*, que combina las estructuras ambientales con las genómicas en un solo modelo.

Para implementar el modelo anteriormente descrito, utilizamos la librería *ASReml-R* 4 de R [6].

Datos a utilizar

Para este modelo decidimos utilizar todos los datos ambientales: los climáticos (`WEATHER`), los del suelo (`SOIL`) y las covariables ambientales (`EC`). De todos modos, hicimos variantes de este modelo donde se suprimían algunas fuentes de datos (`GENO`, `WEATHER`, `GENO_WEATHER` y `GENO_ENV`). Para los datos climáticos, en lugar de utilizar las series temporales completas, empleamos el resumen estadístico para armar una matriz de ambientes con las distintas características.

Entrenamiento

El entrenamiento de los modelos *Factor Analytic* con la librería *ASReml-R 4* se basa en la estimación de los parámetros a través de métodos de máxima verosimilitud restringida ($REML$). Esto significa que no se utiliza directamente una función de pérdida como en muchos algoritmos de aprendizaje automático modernos, sino que se optimiza una función de verosimilitud que maximiza la probabilidad de los datos dados los parámetros del modelo.

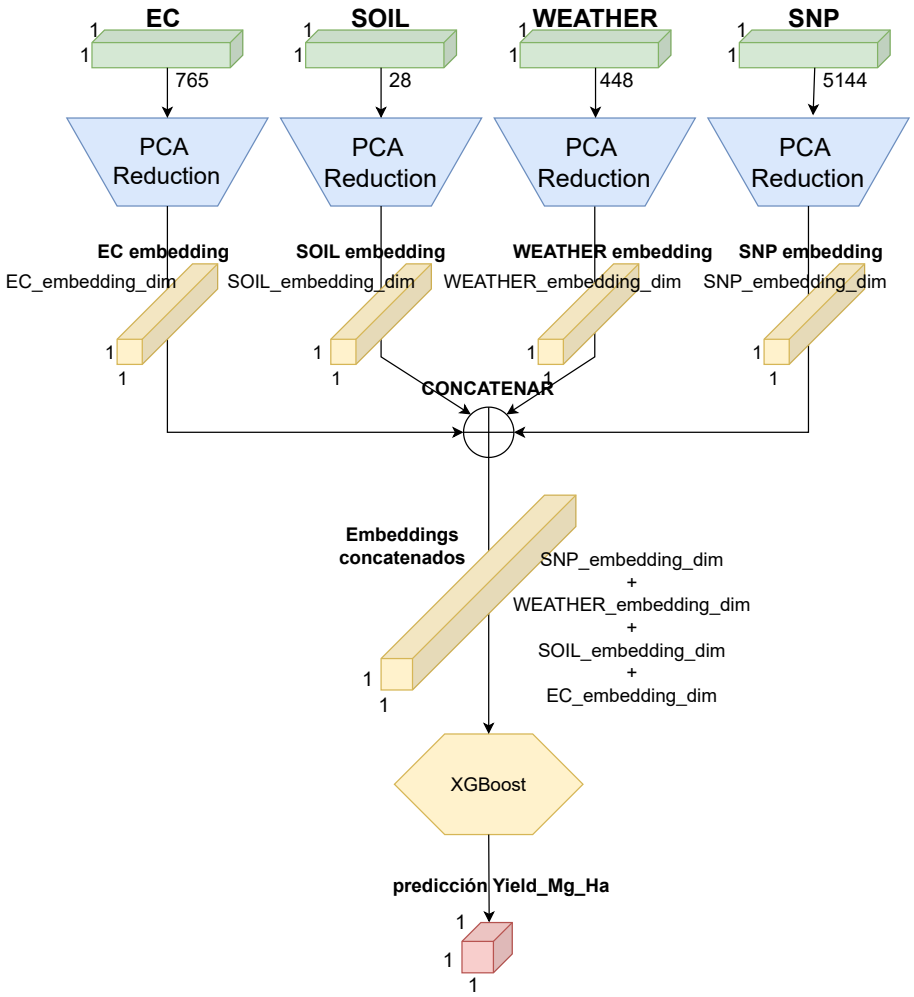


Figura 7.1: Canalización de los datos para el *Gradient Boosting*.

7.2. *Gradient Boosting*

Idea general

Para este modelado canalizamos los datos como se muestra en 7.1, para luego introducirlos en un modelo de regresión. Debido a la alta dimensionalidad, decidimos utilizar la técnica de *PCA* [13] a modo de *embedding*, sobre los datos de cada fuente, para luego concatenarlos y usar el vector resultante para entrenar un modelo de *Gradient Boosting* que estime el valor de *Yield_Mg_Ha*.

Datos a utilizar

Al igual que con los modelos lineales, en este caso también utilizamos todos los datos ambientales (`WEATHER`, `SOIL` y `EC`). De la misma manera que antes, hicimos variantes de este modelo donde se suprimían algunas fuentes de datos.

Otra consideración es que como datos climáticos, en lugar de las series temporales, utilizamos nuevamente el resumen estadístico. Principalmente para poder aplicar la reducción de dimensionalidad sin tener que hacer procesamientos previos más que normalizar los datos. Además, de esta manera nos aseguramos que lo que llega al modelo de *Gradient Boosting* es simplemente un conjunto de datos tabulares, sin una relación temporal entre ellos. Esto no solo baja la complejidad, sino que brinda una mejor interpretabilidad de los resultados.

Variantes del modelo

Al igual que para los modelos lineales, con la motivación de entender cómo influencia cada fuente de datos, en cada esquema de validación, creamos diversas variantes del modelo, alternando las fuentes de datos que se utilizan. Esto fue posible gracias a que este modelo era el más rápido de entrenar, por lo que pudimos realizar gran cantidad de experimentos con él.

- `GENO`: Solo datos genómicos.
- `WEATHER`: Solo datos climáticos.
- `ENV`: Solo datos ambientales.
- `GENO_OH_L`: Datos genómicos más un indicador de en que locación fue plantada la muestra.
- `GENO_OH_E`: Datos genómicos más un indicador de en que ambiente fue plantada la muestra. (Solo aplicable a *HD*, pues en *AD* predecimos en ambientes desconocidos)
- `GENO_WEATHER`: Datos genómicos más datos climáticos.
- `GENO_ENV`: Datos genómicos más todos los datos ambientales.

Los indicadores introducidos para las variantes `GENO_OH_L` y `GENO_OH_E` fueron creados utilizando *One Hot Encoding* y sin aplicarles *PCA*, para luego ser concatenados a los datos genómicos ya reducidos en dimensionalidad.

Motivación de usar *PCA*

Variables correlacionadas

Si bien procesamos los *SNPs* para eliminar aquellos con poca varianza o alta correlación entre sí, sigue habiendo información redundante entre ellos. Lo mismo pasa con las demás fuentes de datos, como se ve en 7.2. En los datos climáticos, esto pasa porque

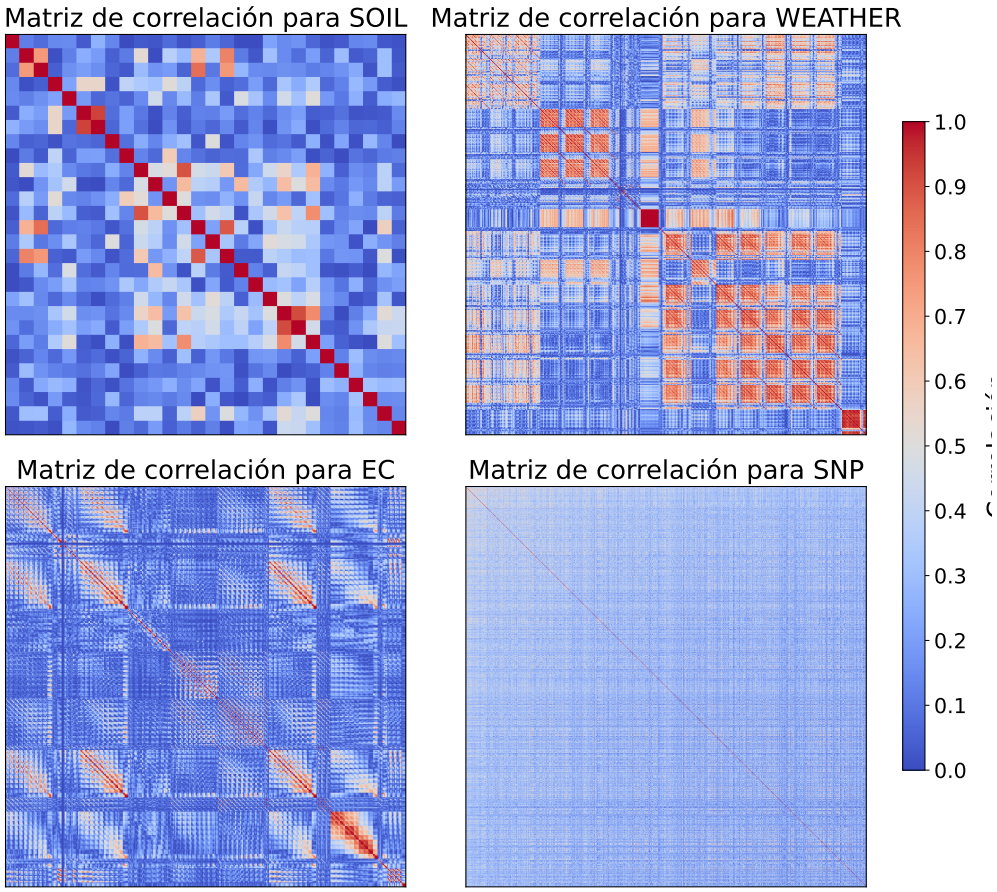


Figura 7.2: Matriz de correlación para cada fuente de datos. Las correlaciones se calculan a nivel ambiente para las fuentes ambientales, y a nivel híbrido para los datos genómicos. En `WEATHER`, las variables (descriptores estadísticos) aparecen agrupadas en base a la serie temporal de la que provienen, de ahí la estructura en bloques de la matriz.

los distintos estadísticos obtenidos de una misma serie temporal tienden a estar correlacionados entre sí, sobre todo los asociados a percentiles y promedios. Además, puede haber series temporales que midan variables que naturalmente estén correlacionadas, lo que genera que sus descriptores también lo estén.

Por estos motivos es que consideramos prometedora la idea de reducir la dimensionalidad para eliminar la información redundante aún existente en los datos.

Redundancia matemática

Si analizamos los años 2019, 2020 y 2021 (que son los que vamos a usar para el modelado), hay $n = 35735$ muestras, en las que aparecen $r = 2313$ híbridos diferentes. Los datos genómicos tienen un total de $m = 5144$ variables, una por cada *SNP*.

Por lo tanto, la matriz G que posee unicamente los datos genómicos para cada muestra, tiene dimensión $n \times m$. Como los datos genómicos son a nivel híbrido, la matriz

G tiene solo r filas distintas. Por lo que $\text{rank}(G) \leq r < m$. De esto último surge que la matriz de covarianza $\Sigma_G \in \mathcal{M}_{m\times m}$, tiene como mucho r valores propios no nulos. Lo que quiere decir que se puede aplicar *PCA* sobre los datos genómicos, reduciendo a al menos r el número de variables, mientras se conserva el total de la varianza.

En general podemos decir que la dimensión mínima del espacio al que podemos reducir los datos sin perder varianza, está acotada superiormente por la cantidad de valores únicos que tengan los mismos.

Se puede seguir un razonamiento similar para cada una de las fuentes de datos ambientales, donde ahora la cantidad de filas distintas va a ser $r = 72$, correspondiente a la cantidad de ambientes únicos.

Motivos prácticos

Además existen motivaciones prácticas para reducir la dimensionalidad de los datos. La principal de ellas es la reducción del tiempo de entrenamiento, pero también está la reducción de los recursos computacionales requeridos. Esto último nos permitió ejecutar los entrenamientos en computadoras convencionales sin tener que recurrir al *cluster*.

Análisis *PCA*

Si bien se puede comprimir cada fuente hasta llegar a tener tantas variables como valores únicos de la misma, la existencia de variables correlacionadas indicaba que podíamos reducir aún más la dimensionalidad (sin perder demasiada varianza acumulada).

De 7.3 se nota como la cantidad de variables de `SNP`, `WEATHER` y `EC`, se puede comprimir a aproximadamente un 5 %, perdiendo solo un 5 % de la varianza original. En cambio para `SOIL`, no podemos más que comprimir su cantidad de variables a la mitad, para lograr una perdida no mayor al 5 %.

Fuente	**# Var. orig.**	**# Var. *PCA***	**% Reducción**	r
`SNP`	5144	285	5.54	2313
`WEATHER`	448	19	4.24	72
`EC`	765	29	3.79	72
`SOIL`	28	14	50.0	72
Total	6385	347	5.43	-

Tabla 7.1: Tabla con cantidad original de variables por fuente, la cantidad luego de aplicar *PCA* al 95 %, mostrada también como porcentaje de la cantidad original. Además se muestra la cantidad de valores únicos para cada fuente en la columna r.

En 7.1 se puede ver como efectivamente pudimos reducir la cantidad de variables por debajo del umbral de valores únicos, lo cual significa que efectivamente había variables correlacionadas.

Finalmente notar que la reducción total que en este caso fue aproximadamente al 5 %, depende de los datos utilizados, pues al aplicar *PCA* seleccionamos el porcentaje de varianza a preservar (95 %) y no la cantidad de componentes a utilizar. Esto tiene la

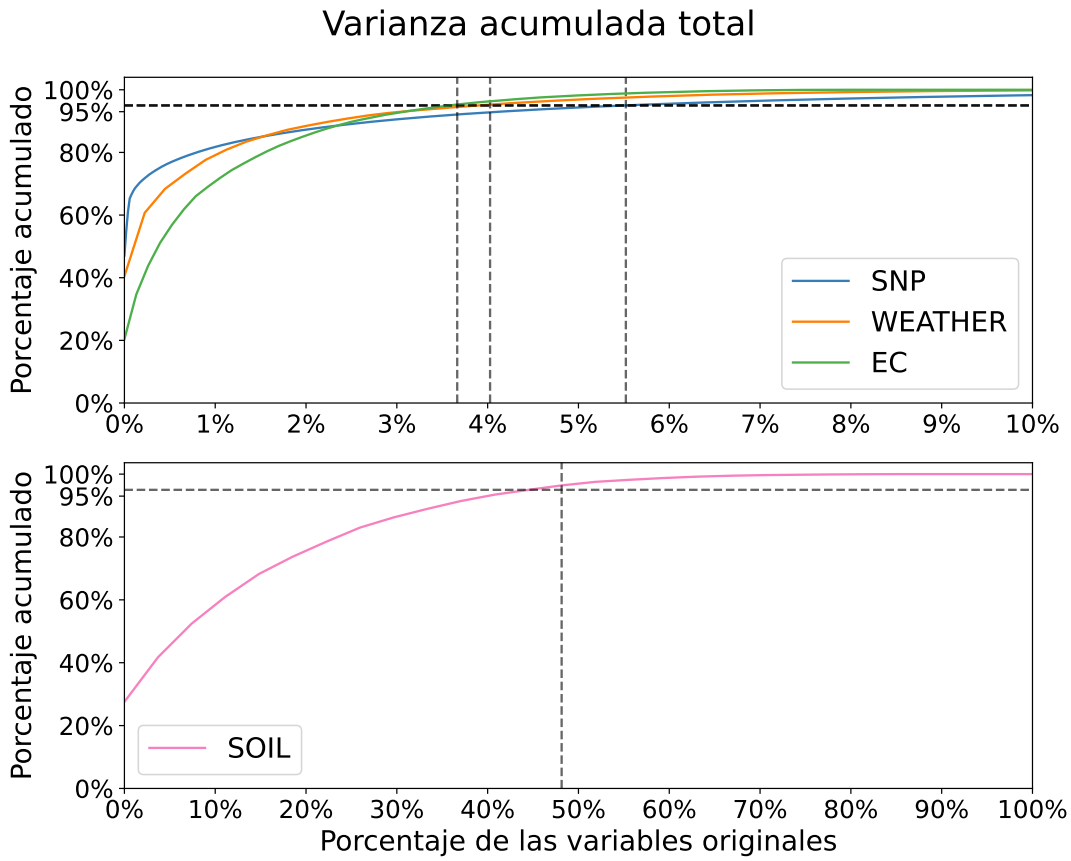


Figura 7.3: Porcentaje de la varianza acumulada (medida de que tanta información estamos preservando) en función de la cantidad de componentes principales utilizadas, expresada como porcentaje de la cantidad de variables originales. Con lineas punteadas se marca el porcentaje en el que se alcanza el $95\,\%$ de la varianza acumulada.

desventaja de que la cantidad de variables que ve el modelo puede ser ligeramente distinta en cada entrenamiento. Por otro lado, tiene la ventaja de que es capaz de adaptarse mejor a los datos de entrada, pues siempre se van a elegir la cantidad de componentes necesarios para alcanzar la varianza deseada.

Entrenamiento

Configuración

Para el entrenamiento del modelo se utilizó *MSE* (*Mean Squared Error*) como función de costo, que es una elección natural tratándose de un problema de regresión. También se evaluó la posibilidad de utilizar la correlación de *Pearson* como función de ganancia, sin embargo, esto no fue posible debido a que la implementación de *XGBoost* exige que

la función de costo sea separable[1].

Además, durante el entrenamiento se monitoreó el Coeficiente de Correlación de *Pearson* y el *RMSE* (*Root Mean Squared Error*), tanto en el conjunto de entrenamiento como en el de validación. Ambas métricas son de interés en problemas de regresión en general, pero particularmente en los de predicción genómica.

Para evitar el sobreajuste se utilizó la técnica de *Early Stopping*. Monitoreando la función de costo en el conjunto de validación, y teniendo una tolerancia de hasta diez épocas sin mejoras.

Procedimiento

El procedimiento de entrenamiento consistió en realizar una búsqueda de hiperparámetros mediante validación cruzada para cada esquema de validación, utilizando los *folds* descritos previamente.

Para encontrar los valores óptimos de los hiperparámetros (ver apéndice B.2), definimos un conjunto de valores para cada hiperparámetro y evaluamos el modelo con todas las posibles combinaciones de estos, buscando cuál alcanzaba la mejor correlación promedio en los *folds*. Pudimos realizar la búsqueda de esta manera tan directa pues el tiempo de entrenamiento era menor a un minuto. Cabe aclarar que para encontrar con precisión los hiperparámetros, refinamos iterativamente los conjuntos sobre los que buscábamos. Llegando así a hacer tres búsquedas distintas, cada una más específica que la anterior.

Para cada *fold* y para cada combinación de hiperparámetros, se entrenó el modelo durante un máximo de 500 épocas. Esta cantidad resultó más que suficiente pues la mayoría de entrenamientos se detenían antes de las 100, debido al *Early Stopping*.

7.3. Redes neuronales

Arquitectura general

Para el modelado utilizando redes neuronales empleamos la arquitectura presente en la figura 7.4. Esta vez decidimos utilizar como datos ambientales únicamente los climáticos, y empleando las series temporales completas. La arquitectura de la red consiste en una rama que procesa los datos climáticos, otra encargada de procesar los datos genómicos y una cabeza final encargada de procesar de manera conjunta los vectores de características generados por ambas ramas y estimar el valor de *Yield_Mg_Ha*.

Datos de entrada

Datos ambientales de clima

La rama encargada de procesar los datos climáticos recibe como entrada un arreglo bidimensional de tamaño `n_weather_features` $\times$ `sequence_length` (ignorando la dimen-

[1] Que la función de costo sea separable significa que se puede escribir como la suma de los errores de cada muestra, o sea $E(y, \hat{y}) = \sum_i E(y_i, \hat{y}_i)$.

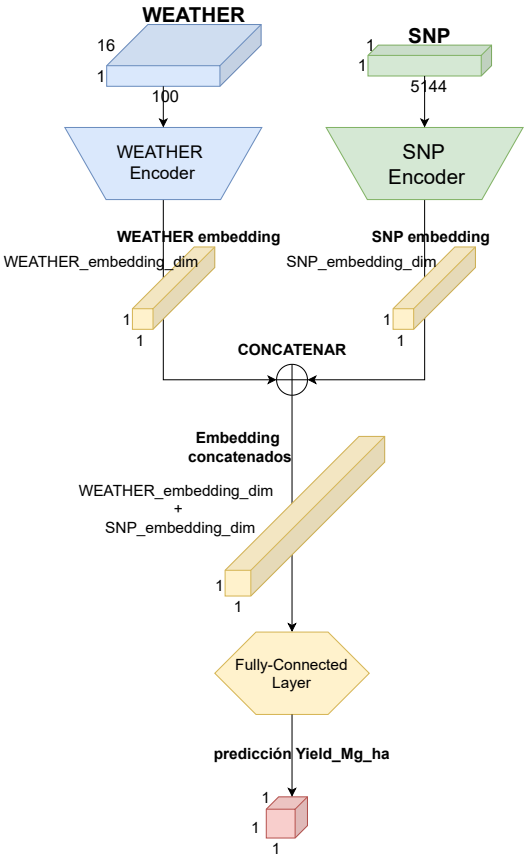


Figura 7.4: Arquitectura de red neuronal multimodal.

sión correspondiente al *batch*), siendo `n_weather_features` $= 16$ la cantidad de variables climáticas medidas y `sequence_length` $= 100$ el largo de las secuencias a procesar. Para generar este arreglo, de los datos climáticos [2], se seleccionan solo aquellos correspondientes a los días entre la fecha de plantado y la fecha de cosechado del experimento en cuestión. Luego se realiza una interpolación en la dimensión temporal a tamaño 100 con método *nearest* para garantizar el mismo largo de secuencia para todas las muestras.

Datos genómicos

La rama encargada de procesar los datos genómicos recibe como entrada un arreglo bidimensional de tamaño $1 \times$ `sequence_length` (ignorando la dimensión correspondiente al *batch*), siendo `sequence_length` $= 5144$ el largo de la secuencia de *SNPs* filtrada descrita en la sección. Aunque especificar el 1 correspondiente a la dimensión de canales

[2] Cada serie temporal es previamente reescalada para tener un rango de variación entre $-\frac{1}{2}$ y $\frac{1}{2}$.

parece redundante, esto lo hicimos por motivos de la implementación de *Pytorch*, cuyos módulos suelen requerir una cantidad de dimensiones específica.

Variantes del modelo

Al igual que con los modelos anteriores, evaluamos diferentes variantes de la arquitectura de red neuronal propuesta. En particular, probamos disponibilizando diferentes combinaciones de datos para el entrenamiento.

- `GENO`: Para esta variante usamos solamente los datos genómicos, ignorando la rama de la arquitectura que procesa datos ambientales de clima y alimentando la cabeza de regresión unicamente con el *embedding* producido por el *encoder* de datos genómicos.
- `WEATHER`: Similar a `GENO`, pero usando solamente los datos ambientales de clima.
- `MULTI`: En este caso usamos ambas modalidades de datos, entrenando tanto el *encoder* de datos genómicos como el *encoder* de datos ambientales de clima y la cabeza de regresión.

Implementación de la arquitectura

En esta sección se da una descripción de como se compone cada una de las ramas de la arquitectura y la cabeza final. Para más detalle pueden consultarse los diagramas de los diferentes bloques presentes en el apéndice B.3.

Encoder de datos climáticos

El *Encoder* de datos climáticos consiste de dos componentes principales: un extractor de características y una capa completamente conectada con activación no lineal. El extractor de características está compuesto por una serie de capas convolucionales y activaciones no lineales. Este se encarga de extraer información útil de las series temporales, contrayendo la dimensión temporal y dilatando la dimensión de los canales. De esta forma, al aplicar una operación de *GlobalAveragePooling* en la dimensión temporal (ver sección 3.5.4), al final del proceso se obtiene un vector que caracteriza a la serie temporal.

Por otro lado, la capa completamente conectada toma como entrada el vector de características y le aplica una serie de transformaciones lineales y activaciones no lineales, aumentando la capacidad expresiva del modelo y permitiendo capturar relaciones no lineales entre las diferentes características.

Encoder de datos genómicos

El *Encoder* de datos genómicos es similar al *Encoder* de datos ambientales, ya que contiene un extractor de características basado en capas convolucionales y una red completamente conectada con transformaciones no lineales a la salida. Sin embargo, el *Encoder* de datos genómicos también incluye una capa de embedding en la entrada. Esta

tiene la tarea de transformar la etiqueta categórica del alelo codificado como un entero $\{0, 1, 2\}$, a un vector de valores reales en un espacio de embedding de dimensión mayor, que corresponde a una codificación más apta para ser procesada por una red neuronal.

Cabeza de regresión

La cabeza de regresión a la salida del modelo corresponde a una red completamente conectada idéntica a las que se encuentran a la salida de los *Encoders* para ambas modalidades de datos. Su estructura consiste en:

- Una capa lineal a la entrada, que transforma el embedding conjunto de entrada a una dimensión mayor.
- Una función de activación *ReLU*, que introduce una no linealidad.
- Una capa de *Dropout* como mecanismo de regularización para prevenir el sobreajuste del modelo a los datos de entrenamiento.
- Una capa lineal a la salida, que proyecta los datos a una dimensión diferente. En el caso de la cabeza de regresión, se proyecta a un escalar correspondiente a la predicción de *Yield_Mg_Ha*.

Entrenamiento

Configuración

Para el entrenamiento del modelo volvimos a utilizar el *MSE* como función de costo. Ademas volvimos a monitorear los valores del *RMSE* y de la correlación de *Pearson*.

Por otro lado, durante el entrenamiento empleamos *batches* de 128 muestras tanto para el conjunto de entrenamiento como para el de validación. Como optimizador utilizamos *Adam* con `lr` $= 10^{-6}$, especificando además una regularización L_2 con `weight_decay` $= 0{,}1$.

Procedimiento

Para la búsqueda de hiperparámetros procedimos de manera muy similar al modelo anterior, utilizando validación cruzada sobre ambos esquemas. La única diferencia radicó en que, como los entrenamientos de este modelo eran mucho más costosos computacionalmente y por lo tanto requerían varias horas, hubo que optimizar la búsqueda de hiperparámetros. Para esto utilizamos el *framework* de optimización *Optuna*, que hace una búsqueda eficiente de los parámetros que maximizan (o minimizan) una función. En nuestro caso la función a minimizar era el promedio de los *RMSE* en cada *fold* de validación, y los parámetros justamente los hiperparámetros del modelo (ver anexo).

Para cada *fold* y para cada combinación de hiperparámetros se entrenó el modelo durante un máximo de 1000 épocas, guardando un *checkpoint* con el modelo que obtuvo el mejor desempeño en validación hasta el momento, y finalizando el entrenamiento si transcurrían 100 épocas sin mejoras en el desempeño. Finalmente utilizamos como resultado del *fold* la correlación registrada en el último *checkpoint*, que corresponde al modelo que menor error alcanzó.

Capítulo 8

Análisis de los datos multimodales, integrando genoma y ambiente

8.1. Métricas de comparación

A modo de comparar resultados y evaluar si los modelos son capaces de aprender de los datos, definimos una serie de predicciones triviales, sobre las cuales calculamos métricas de comparación.

- E_G: Consiste en predecir utilizando el promedio del fenotipo, pues es la constante que minimiza el *RMSE* [1]. Es lo mejor que se podría hacer si no se conoce nada de la muestra a predecir. El *RMSE* asociado a esta predicción se puede interpretar como una desviación estándar del fenotipo.

- E_L: Implica utilizar como predicción la media del fenotipo en la locación que fue sembrada la planta. Ahora la predicción ya no es constante si no que depende de la muestra. Este sería el mejor rendimiento que podría tener un modelo que solo utilice la locación como dato.

- E_E: Igual que la métrica anterior, pero utilizando el promedio por ambiente, que es más especifico que por locación. Sería el mejor rendimiento que podría tener un modelo que además de conocer la locación de la planta, conoce también el año en que fue plantada.

- E_H: Análoga a las anteriores pero utilizando el promedio a nivel híbrido. Es la mejor predicción posible si solo se conoce a qué variedad pertenece la planta.

Notar que para el esquema de validación *HD* (sección 6.1.1) no podemos calcular E_H, pues los híbridos que hay en los *folds* de validación no están en entrenamiento.

[1] En realidad como el error se calcula en validación, la constante que lo minimiza es el promedio del fenotipo en validación, y no entrenamiento, que es el que podemos calcular. Sin embargo, si suponemos que el fenotipo tiene una distribución similar en ambos conjuntos, la media en entrenamiento resulta una buena aproximación de la media en validación.

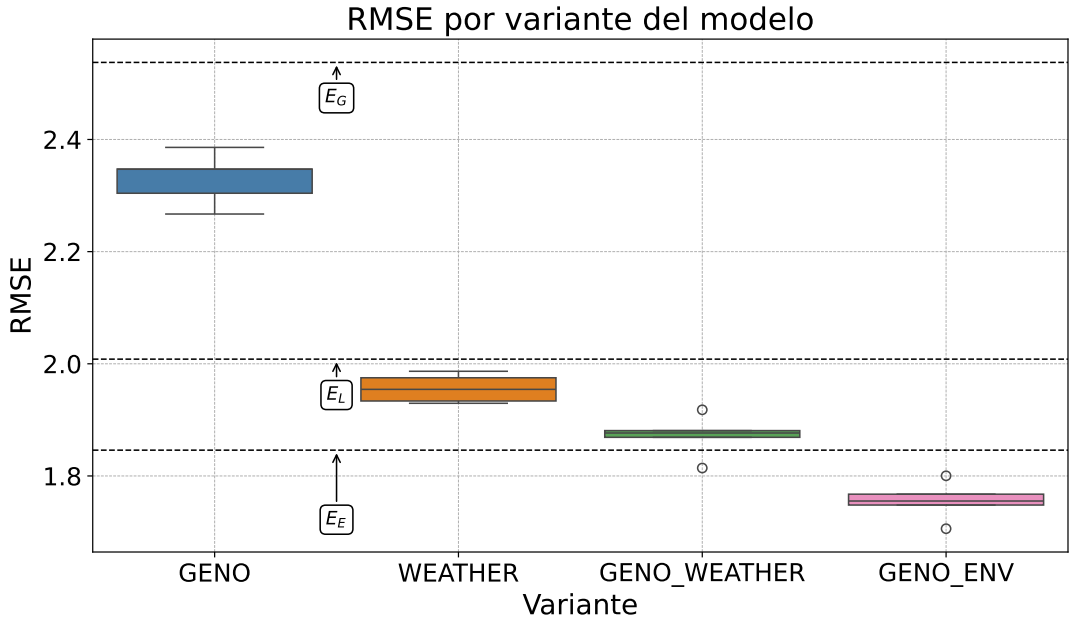


Figura 8.1: *Boxplots* del *RMSE* para cada *fold*, agrupados por variantes del modelo para el esquema de validación *HD*. En líneas punteadas se marcan los errores de comparación.

Análogamente no podemos calcular E_E en AD, pues en los *folds* de validación hay ambientes que no están en entrenamiento (aunque como lo que varía es el año y no las locaciones, si podemos calcular E_L).

8.2. Modelos Lineales Mixtos

HD: Híbridos desconocidos

Tras entrenar los modelos lineales para cada variante, siguiendo el procedimiento descrito previamente en la sección 7.1, evaluamos las mismas utilizando validación cruzada con los *folds* del esquema *HD*.

Resultados por *fold*

Se puede observar en las figuras 8.1 y 8.2 que al utilizar todas las fuentes de datos, como en la variante `GENO_ENV`, mejora el desempeño superando todas las métricas de comparación tanto para el *RMSE* como para la correlación.

Este esquema de validación busca predecir híbridos desconocidos en ambientes conocidos, por lo que es esperable que la variante `GENO` que utiliza solamente los datos genómicos tenga un peor desempeño que las demás variantes. Sin embargo, se puede observar en la figura 8.1 que supera a la métrica de comparación E_G dando a entender que el modelo es capaz de generalizar en híbridos nunca antes vistos.

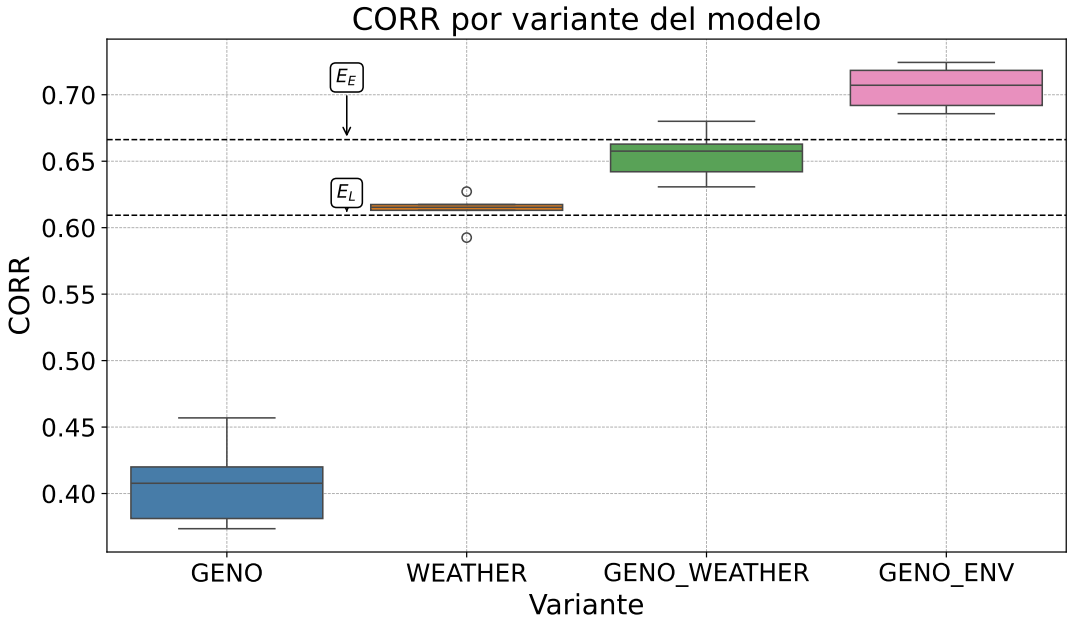


Figura 8.2: *Boxplots* de la correlación de *Pearson* para cada *Fold*, agrupados por variantes del modelo para el esquema de validación *HD*. En líneas punteadas se marcan las correlaciones de comparación.

Al considerar solamente los datos climáticos (`WEATHER`), mejoran considerablemente los resultados ya que se están haciendo predicciones en ambientes conocidos. Al combinar la información genómica con la ambiental se tiene una leve mejora que puede estar explicada por la interacción entre ambos tipos de datos.

Resumiendo, al utilizar este esquema de validación, los datos ambientales aportan más que los genómicos, aunque los mejores resultados se obtienen cuando se combinan ambas fuentes de información.

Resultados promediados

Finalmente, en la tabla 8.1 presentamos los resultados pero promediando en los distintos *folds*.

AD: Ambientes desconocidos

Luego de entrenar los modelos lineales para cada variante, siguiendo el procedimiento descrito previamente en la sección 7.1, evaluamos las mismas utilizando validación cruzada con los *folds* del esquema *AD*.

Resultados por *fold*

Nuevamente podemos ver en las figuras 8.3 y 8.4 que la variante `GENO_ENV` que utiliza todas las fuentes de datos, obtiene los mejores resultados tanto en *RMSE* como en correlación.

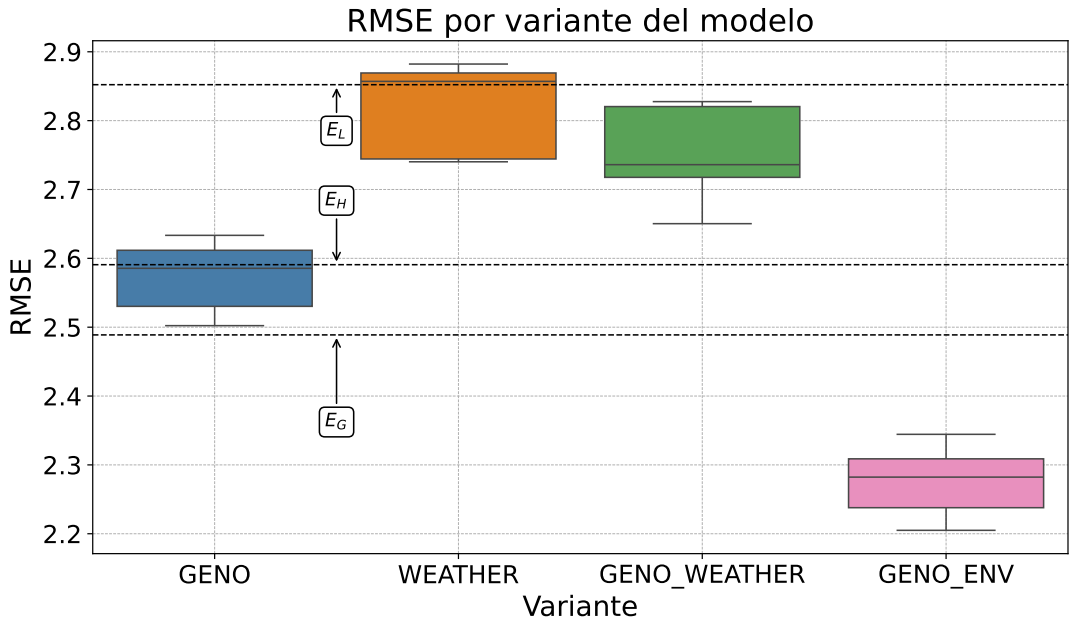


Figura 8.3: *Boxplots* del *RMSE* para cada *fold*, agrupados por variantes del modelo para el esquema de validación *AD*. En líneas punteadas se marcan los errores de comparación.

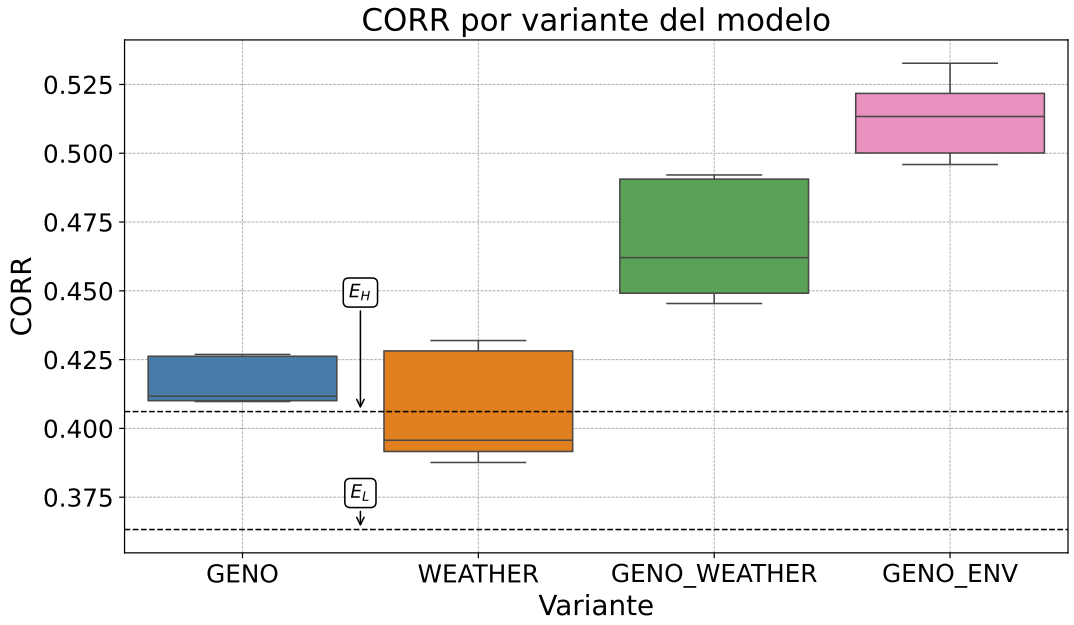


Figura 8.4: *Boxplots* de la correlación de *Pearson* para cada *Fold*, agrupados por variantes del modelo para el esquema de validación *AD*. En líneas punteadas se marcan las correlaciones de comparación.

Variante	RMSE	CORR
E_G	2.537	-
E_L	2.008	0.609
E_E	1.846	0.666
GENO	2.33	0.41
WEATHER	1.96	0.61
GENO_WEATHER	1.87	0.66
GENO_ENV	**1.76**	**0.71**

Tabla 8.1: Promedio entre *folds* de distintas métricas, para cada variante del modelo. Las tres primeras filas corresponden a las métricas de comparación, las cuales también están promediadas a nivel de *folds*. Se marca en negrita el mejor resultado de cada métrica. Ver anexo para consultar la definición de cada métrica.

En este caso, a diferencia del esquema anterior, las variantes que utilizan solo un tipo de datos, GENO y WEATHER, tienen un desempeño similar en términos de correlación, pero la variante WEATHER tiene peor *RMSE*. Que ni la variante GENO sea capaz de superar a E_G, ni la variante WEATHER a E_L, significa que ninguna de estas está siendo capaz de generalizar. Esto muestra la necesidad de utilizar ambas fuentes de información para generalizar sobre ambientes no conocidos.

Resultados promediados

Nuevamente para resumir, en la tabla 8.2 presentamos los mismos resultados pero promediando entre *folds*.

Variante	RMSE	CORR
E_G	2.489	-
E_H	2.591	0.406
E_L	2.852	0.363
GENO	2.57	0.42
WEATHER	2.82	0.41
GENO_WEATHER	2.75	0.47
GENO_ENV	**2.27**	**0.52**

Tabla 8.2: Promedio entre *folds* de distintas métricas, para cada variante del modelo. Las tres primeras filas corresponden a las métricas de comparación, las cuales también están promediadas a nivel de *folds*. Se marca en negrita el mejor resultado de cada métrica. Ver anexo para consultar la definición de cada métrica.

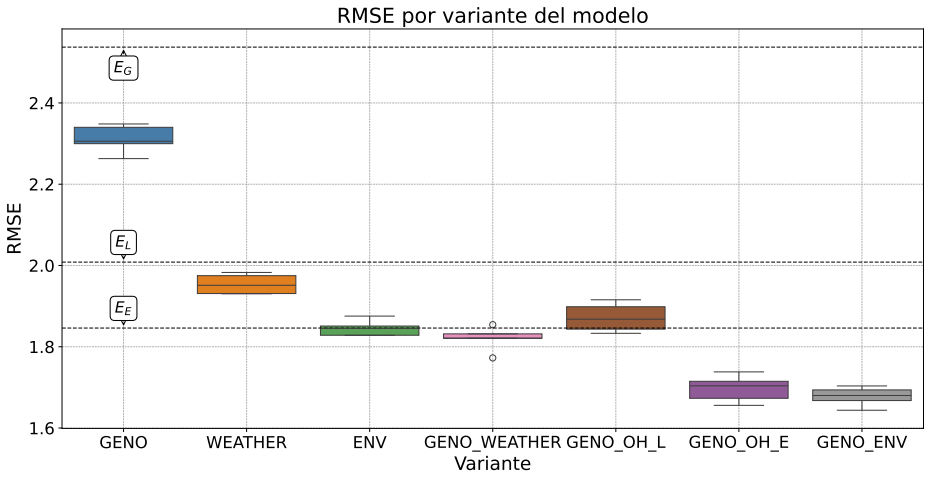


Figura 8.5: *Boxplots* del *RMSE* para cada *fold*, agrupados por variantes del modelo. Notar que como son cinco *folds*, cada linea del *boxplot* representa el error en un *fold*. En lineas punteadas se marcan los errores de comparación.

Comparación entre ambos esquemas

En base a los resultados obtenidos utilizando ambos esquemas de validación, con las distintas variantes del modelo, podemos concluir que la componente ambiental aporta más información que la genómica al hacer predicciones. Esto hace que generalizar sobre híbridos desconocidos sea más fácil que sobre ambientes desconocidos, por lo cual las métricas obtenidas bajo el esquema AD son siempre peores que bajo HD.

Comparando las gráficas 8.1 y 8.2 que utilizan el esquema de validación HD, con las del esquema AD 8.3 y 8.4, las primeras tienen menor varianza entre *folds*. Esto se debe a que en AD la componente desconocida a predecir es la ambiental, que como mencionamos anteriormente, es la que aporta más información para la predicción, entonces tiene un mayor rango de variación.

8.3. *Gradient Boosting*

HD: Híbridos desconocidos

Tras obtener los hiperparámetros óptimos para cada variante, siguiendo el procedimiento descrito previamente en la sección 7.2, evaluamos las mismas utilizando validación cruzada con los *folds* del esquema *HD*.

Resultados por *fold*

El modelo que incorpora toda la información es el que obtiene mejores resultados para ambas métricas, como se ve en las figuras 8.5 y 8.6. Por otro lado, vemos que `GENO_OH_E`

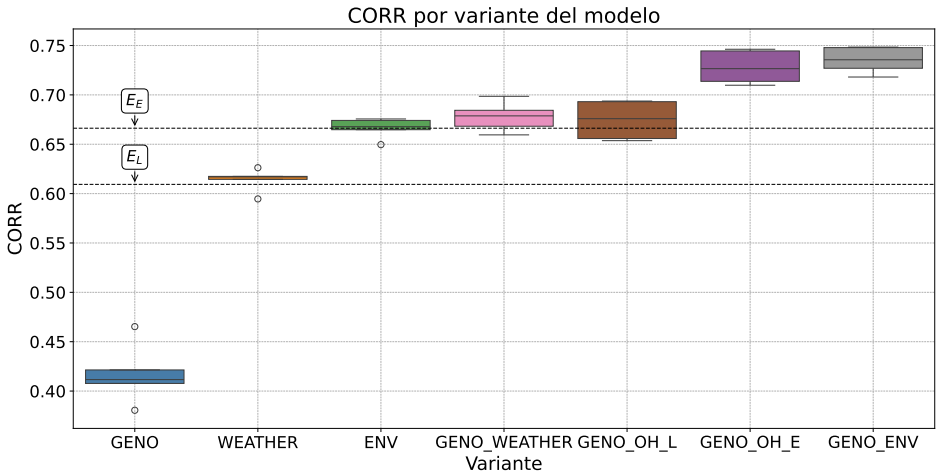


Figura 8.6: *Boxplots* de la correlación de *Pearson* para cada *Fold*, agrupados por variantes del modelo. En lineas punteadas se marcan las correlaciones de comparación.

y `GENO_ENV` son las únicas variantes que logran superar las métricas de comparación, en todos los *folds*.

Las variantes mencionadas logran resultados muy similares debido al esquema de validación utilizado. Recordando la ecuación 2.1, ambos modelos van a predecir F_G de igual manera, pues ambos disponen de la misma información genética. Por otro lado, como estamos prediciendo en ambientes conocidos, la componente F_E ya fue vista en entrenamiento, por lo que el modelo no debe generalizar sobre la misma. Es decir, que alcanza con que el modelo pueda identificar el ambiente para conocer F_E. Por lo que en lo único que se diferencian las predicciones de estos dos modelos, es en el término F_{GE}. Entonces, o bien F_{GE} no es significativo frente a $F_G + F_E$, o bien los modelos no están siendo capaces de aprender esta interacción conjunta.

Que `ENV` tenga prácticamente las mismas métricas que E_E tiene sentido, pues al no tener que generalizar sobre otros ambientes, lo que termina aprendiendo el modelo es el promedio de cada ambiente. Sin embargo, `WEATHER`, que utiliza solo datos climáticos, no alcanza el mismo desempeño. Esto podría deberse a que los datos climáticos no son suficientes para distinguir completamente a los ambientes.

Por otro lado, que `ENV` o `WEATHER` tengan mejores resultados que `GENO`, indica que la componente ambiental aporta más información que la genómica.

Al igual que para los modelos lineales, con este esquema de validación, nos podemos quedar con que los datos ambientales aportan más que los genómicos, aunque los mejores resultados se obtienen cuando se combinan ambas fuentes. Además, utilizar las variables ambientales en lugar de un simple indicador del ambiente, no parece tener resultados significativamente mejores.

Resultados promediados

Por último, presentamos en la tabla 8.3 el promedio entre *folds* de las métricas evaluadas previamente, junto a otras adicionales.

Variante	MAE	MAPE	RMSE	NRMSE	CORR	R^2
E_G	1.998	26.80	2.537	0.406	-	-
E_L	1.560	18.92	2.008	0.156	0.609	0.341
E_E	1.385	16.94	1.846	0.128	0.666	0.444
GENO	1.786	23.90	2.311	0.331	0.417	0.128
WEATHER	1.482	19.90	1.954	0.213	0.614	0.377
ENV	1.386	16.97	1.846	0.129	0.666	0.444
GENO_OH_L	1.449	17.67	1.872	0.135	0.674	0.428
GENO_OH_E	1.275	15.60	1.697	0.114	0.728	0.529
GENO_WEATHER	1.365	18.54	1.820	0.195	0.678	0.459
GENO_ENV	**1.256**	**15.31**	**1.678**	**0.109**	**0.735**	**0.540**

Tabla 8.3: Promedio entre *folds* de distintas métricas, para cada variante del modelo. Las tres primeras filas corresponden a las métricas de comparación, las cuales también están promediadas a nivel de *folds*. Se marca en negrita el mejor resultado de cada métrica. Ver B.1 para consultar la definición de las métricas adicionales.

AD: Ambientes desconocidos

Nuevamente, siguiendo el procedimiento de la sección 7.2, obtuvimos los hiperparámetros óptimos para cada variante, tras lo cual evaluamos las mismas utilizando validación cruzada con los *folds* del esquema *AD*.

Resultados por *folds*

Vemos en las figuras 8.7 y 8.8 como en ambas métricas, incluida la correlación, el modelo que mejor desempeño tuvo, fue nuevamente el multimodal GENO_ENV.

En contraste con el esquema anterior, esta vez todos los modelos tienden a superar las métricas de comparación. Esto puede significar que los modelos están generalizando mejor que antes. Otra razón para explicar este comportamiento, es que en este esquema, como en validación varía la componente ambiental, las predicciones que se basan en utilizar promedios a nivel locación o ambiente, normalmente van a desempeñar peor. Esto último es lógico, pues ahora ya no se cumple la hipótesis de que el fenotipo tenga la misma distribución en entrenamiento que en validación. Pues como ya habíamos visto, la componente ambiental, que es la que ahora cambia entre conjuntos, es la que más influencia tiene en el fenotipo.

Resulta interesante ver como ahora que los modelos deben generalizar sobre los datos ambientales, ENV vuelve a tener un rendimiento superior a WEATHER. Esto, junto a que

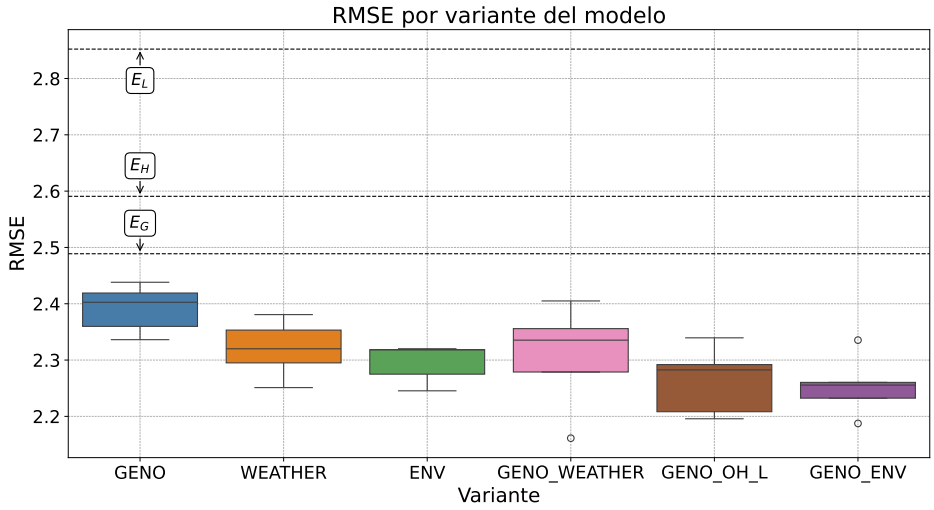


Figura 8.7: *Boxplots* del RMSE para cada *fold*, agrupados por variantes del modelo. Notar que nuevamente como son cinco *folds*, cada linea del *boxplot* representa el error en un *fold*. En lineas punteadas se marcan los errores de comparación.

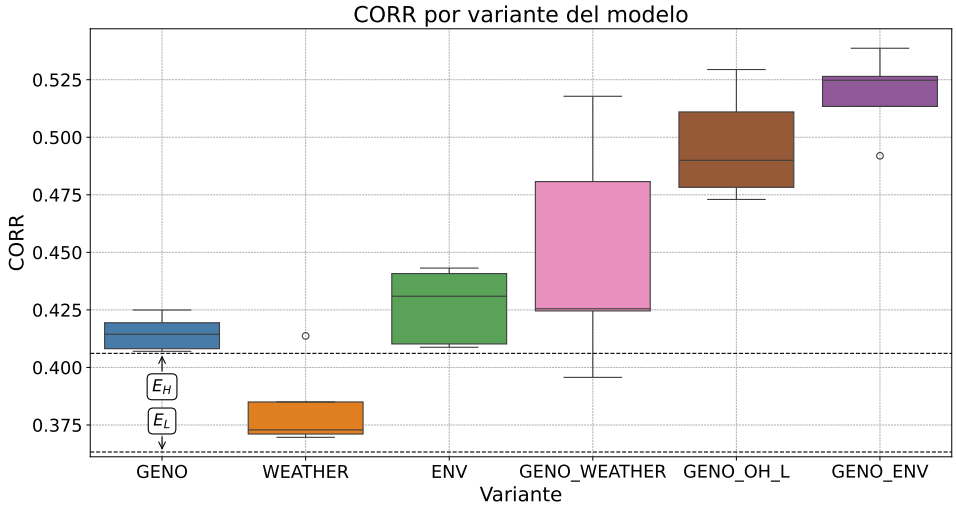


Figura 8.8: *Boxplots* de la correlación de *Pearson* para cada *Fold*, agrupados por variantes del modelo. En lineas punteadas se marcan las correlaciones de comparación.

GENO_ENV supera a GENO_WEATHER nos confirma la importancia de las variables de SOIL y las EC para describir a los ambientes.

Por otro lado, que GENO_ENV supere a GENO_OH_L significa que la primera variante está siendo capaz de generalizar sobre los datos ambientales, cosa que la segunda naturalmente no puede hacer.

Como conclusión de este esquema, podemos decir que los datos ambientales aportan más que los genómicos, aunque nuevamente los mejores resultados se obtienen al utilizar las variantes multimodales. Además, esta vez se mostró que los datos ambientales aportaron más que el indicador de locación, lo que significa que el modelo pudo generalizar sobre la componente ambiental.

Resultados promediados

Para dar cierre a la evaluación del modelo en este esquema, presentamos en la tabla 8.4 los resultados promediados entre *folds*, para las dos métricas utilizadas y para otras adicionales.

Variante	MAE	MAPE	RMSE	NRMSE	CORR	R^2
E_G	1.926	26.96	2.489	0.440	-	-
E_H	2.070	25.10	2.591	0.274	0.406	-0.095
E_L	2.320	29.23	2.852	0.330	0.363	-0.326
GENO	1.872	24.80	2.391	0.351	0.415	0.067
WEATHER	1.770	24.11	2.320	0.346	0.382	0.122
ENV	1.785	23.70	2.295	0.309	0.427	0.140
GENO_OH_L	1.759	22.89	2.264	0.287	0.496	0.164
GENO_WEATHER	1.783	22.75	2.307	**0.253**	0.449	0.131
GENO_ENV	**1.759**	**22.45**	**2.254**	0.272	**0.519**	**0.171**

Tabla 8.4: Promedio entre *folds* de distintas métricas, para cada variante del modelo. Las tres primeras filas corresponden a las métricas de comparación, las cuales también estan promediadas a nivel de *folds*. Se marca en negrita el mejor resultado de cada métrica. Ver B.1 para consultar la definición de las métricas adicionales.

Conjunto de test

Finalmente, utilizando como conjunto de entrenamiento los datos de 2019 y 2020, y los de 2021 como validación para monitorear el *Early Stopping*, entrenamos GENO_ENV para predecir sobre el conjunto de test, compuesto por los datos de 2022.

En la tabla 8.5 se ve como en el conjunto de test las predicciones del modelo no resultan mejores que las triviales por locación, aunque sí lo son en validación. Este comportamiento indica que los datos de test no poseen la misma distribución que los de validación. Por otro lado, que las predicciones en el conjunto de entrenamiento tampoco

Conjunto	Variante	MAE	MAPE	RMSE	NRMSE	CORR
Entrenamiento (2019-2020)	E_L	1.698	**20.27**	2.160	**0.118**	0.708
	GENO_ENV	**1.555**	20.30	**1.982**	0.141	**0.837**
Validación (2021)	E_L	2.320	29.42	2.857	0.340	0.358
	GENO_ENV	**1.649**	**21.74**	**2.129**	**0.256**	**0.529**
Test (2022)	E_L	**2.284**	32.99	2.915	0.578	**0.422**
	GENO_ENV	2.383	**31.56**	**2.905**	**0.469**	0.350

Tabla 8.5: Resultados del modelo para cada conjunto de datos. Además se agregan las métricas de comparación de E_L, que resultó ser la mejor predicción trivial.

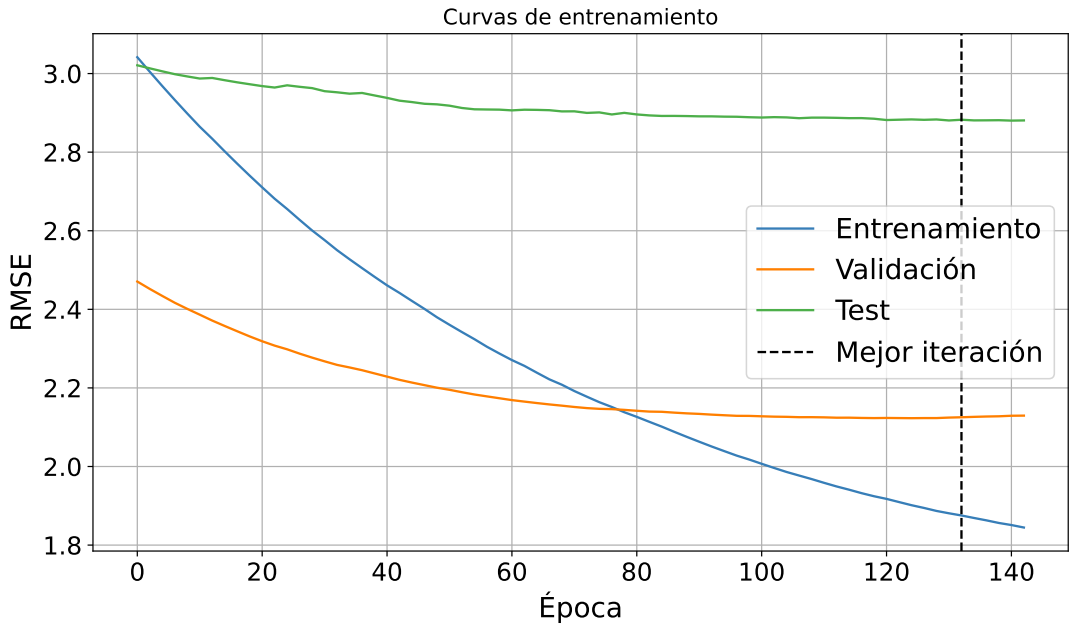


Figura 8.9: *RMSE* (función de perdida) en cada conjunto, en función de la época. Se muestra además con una linea punteada la época en que se obtuvo el menor error en validación.

superen a las triviales, hace sentido, pues estas últimas están sobreajustadas a este conjunto [2], mientras que el modelo busca generalizar.

Si observamos las curvas de entrenamiento en la figura 8.9, se ve que las de validación y entrenamiento siguen un comportamiento esperado, terminando la última por debajo de la primera. Sin embargo, la curva de test tiene un comportamiento muy distinto a la de validación. Esto último confirma que ambos conjuntos son muy distintos, por lo que no es posible estimar el desempeño en test a partir del de validación. Si bien no lo podemos afirmar con certeza, es muy probable que la causa de esto sea que casi todos

[2]Pues las predicciones triviales E_L se obtienen directamente de los promedios en entrenamiento.

los híbridos que aparecen en test, no están ni en validación ni en entrenamiento, por tratarse estos de una nueva generación.

Comparación entre ambos esquemas

En base a los resultados obtenidos utilizando ambos esquemas de validación, con las distintas variantes del modelo, obtuvimos una serie de conclusiones, las cuales listamos a continuación:

- La componente del fenotipo F_E, es decir la ambiental, predomina por sobre F_G, la genética. Esto hace que generalizar sobre híbridos desconocidos sea más fácil que sobre ambientes desconocidos. Este es el motivo por el que las métricas al utilizar AD, son siempre peores que al usar HD.
- Como se ve comparando las figuras 8.5 con 8.7, y 8.6 con 8.8, los resultados de AD tienen mayor varianza entre *folds* que los de HD. Esto se desprende de lo anterior, pues en AD la componente desconocida a predecir es F_E, que al ser la componente predominante del fenotipo, tiene mayor rango de variación.
- Usar los datos ambientales completos, en lugar de solo los climáticos, siempre da mejores resultados. Esto sucede tanto en AD como en HD, donde estos datos juegan un rol distinto.

8.4. Redes neuronales

HD: Híbridos desconocidos

Tras obtener los hiperparámetros óptimos para cada variante, siguiendo el procedimiento descrito anteriormente en la sección 7.3, evaluamos las mismas utilizando validación cruzada con los *folds* del esquema HD.

Resultados por *fold*

En las figuras 8.10 y 8.11 se aprecia que la variante `MULTI` es la que mejor desempeño alcanza para ambas métricas evaluadas, aunque la variante `WEATHER` alcanza un desempeño marginalmente inferior, y superando ambas las tres medidas de desempeño de referencia. También se observa que ambas variantes mantienen un desempeño similar al considerar todos los *folds*. Mientras tanto, la variante `GENO` no logra acercarse a los niveles de desempeño del resto, superando solo la referencia E_G.

El pobre desempeño de la variante `GENO` en comparación a `WEATHER` y `MULTI` puede explicarse recordando que en el esquema de validación HD se busca predecir el desempeño de híbridos desconocidos en ambientes conocidos. Por lo tanto, al solo considerar los datos genómicos el conjunto de datos de validación es completamente desconocido para el modelo, a diferencia de las otras variantes donde se dispone de datos ambientales de clima ya vistos durante el entrenamiento.

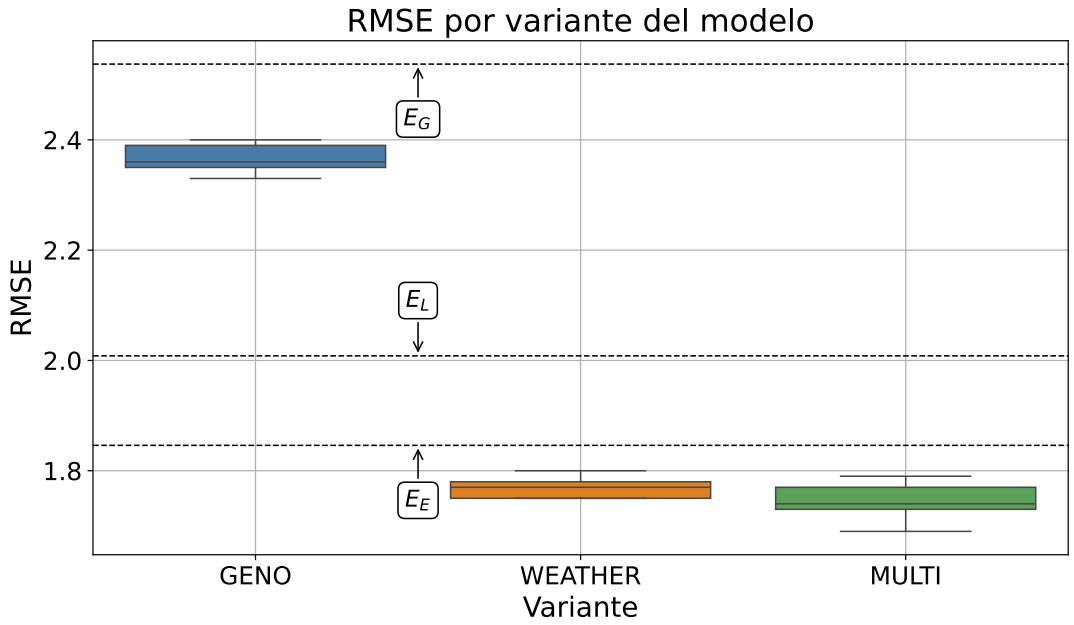


Figura 8.10: *Boxplots* de *RMSE* para cada *fold*, agrupados por variantes del modelo. En lineas punteadas se marcan los errores de comparación.

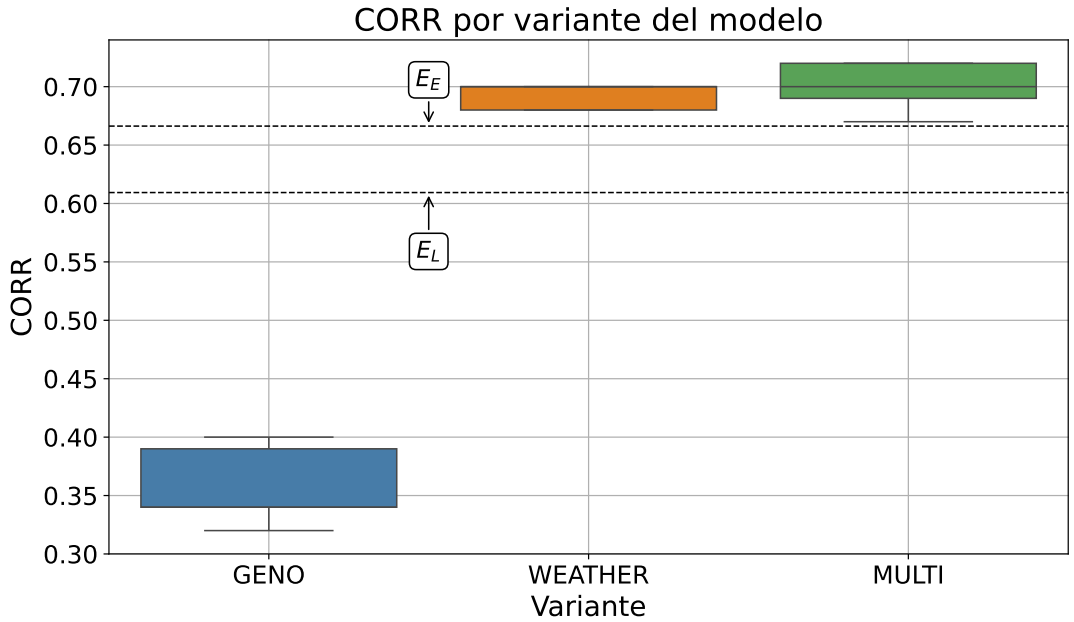


Figura 8.11: *Boxplots* de *CORR* para cada *fold*, agrupados por variantes del modelo. En lineas punteadas se marcan los errores de comparación.

Por otro lado, el hecho de que `WEATHER` tenga un desempeño marginalmente inferior a `MULTI` indica que conocer la información genómica de un individuo contribuye poco a la predicción del fenotipo y que la contribución principal proviene de conocer los datos ambientales de clima correspondientes al entorno donde fue plantado el individuo.

Resultados promediados

A modo de resumen, en la tabla 8.6 presentamos los resultados promediados entre *folds*.

Variante	RMSE	CORR
E_G	2.537	-
E_L	2.008	0.609
E_E	1.846	0.666
`GENO`	2.36	0.34
`WEATHER`	1.77	0.70
`MULTI`	**1.74**	**0.70**

Tabla 8.6: Promedio entre *folds* de ambas métricas, para cada variante del modelo. Las tres primeras filas corresponden a las métricas de comparación, las cuales también están promediadas a nivel de *folds*. Se marca en negrita el mejor resultado de cada métrica.

AD: Ambientes desconocidos

Nuevamente, tras obtener los hiperparámetros óptimos para cada variante, siguiendo el procedimiento descrito previamente en la sección 7.3, evaluamos las mismas utilizando validación cruzada con los *folds* del esquema *HD*.

Resultados por *fold*

Para este esquema de validación se observa en las figuras 8.12 y 8.13 que la variante `MULTI` es la que mejor desempeño tiene para ambas métricas, superando con un margen considerable las medidas de desempeño de referencia. Por otro lado, la variante `GENO` obtiene el segundo lugar en *RMSE*, superando las tres medidas de referencia, y el tercer lugar en correlación, donde no alcanza a superar ninguna de las medidas de correlación de referencia disponibles. Por último, la variante `WEATHER` tuvo un mal desempeño en general, no superando ninguna de las medidas de referencia.

El pobre desempeño de la variante `WEATHER` puede deberse a que en términos generales la predicción del fenotipo es más sensible a los datos ambientales, lo que ocasiona que las predicciones para los ambientes donde el clima se encuentra fuera de la distribución vista durante el entrenamiento tengan un error mayor. Esto también puede explicar la gran variabilidad que existe en los resultados en los diferentes *folds* para esta variante.

A su vez, la diferencia que existe entre el desempeño alcanzado por `MULTI` y las variantes que emplean una sola modalidad de datos, indica que ambas modalidades

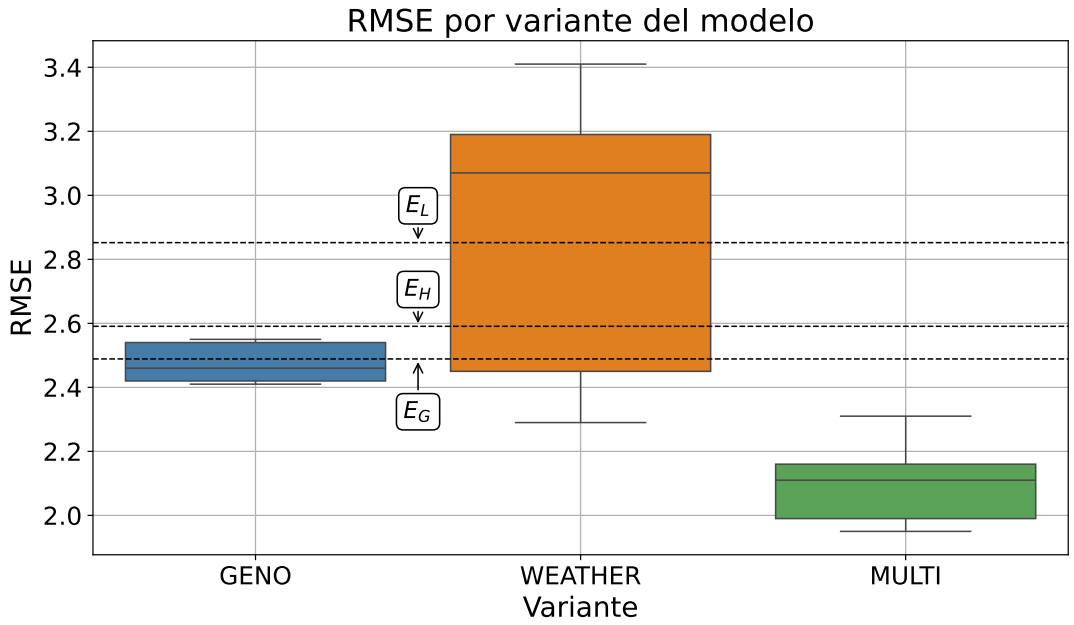


Figura 8.12: *Boxplots* de *RMSE* para cada *fold*, agrupados por variantes del modelo. En lineas punteadas se marcan los errores de comparación.

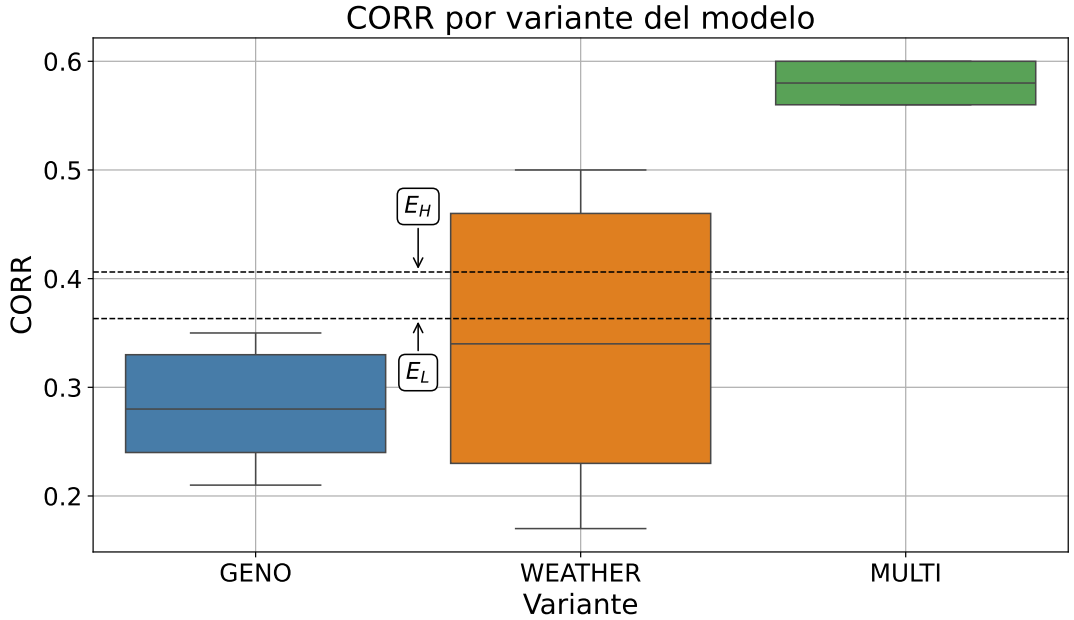


Figura 8.13: *Boxplots* de *CORR* para cada *fold*, agrupados por variantes del modelo. En lineas punteadas se marcan los errores de comparación.

contribuyen de manera no trivial a la predicción del fenotipo para individuos de híbridos conocidos que son plantados en ambientes desconocidos.

Resultados promediados

Por último y a modo de resumen, en la tabla 8.7 presentamos los resultados promediados entre *folds*.

Variante	RMSE	CORR
E_G	2.489	-
E_H	2.591	0.406
E_L	2.852	0.363
`GENO`	2.46	0.28
`WEATHER`	3.07	0.34
`MULTI`	**2.11**	**0.58**

Tabla 8.7: Promedio entre *folds* de ambas métricas, para cada variante del modelo. Las tres primeras filas corresponden a las métricas de comparación, las cuales también están promediadas a nivel de *folds*. Se marca en negrita el mejor resultado de cada métrica.

Comparación entre ambos esquemas

Las conclusiones que sacamos de las pruebas realizadas evaluando las diferentes variantes de la arquitectura de red neuronal desarrollada empleando las modalidades de validación cruzada *HD* y *AD* son similares a las concluidas en la sección anterior. En primer lugar, en estos experimentos se repite el hecho de que el desempeño alcanzado al evaluar con *HD* es superior al alcanzado evaluando con *AD*, lo que sirve como evidencia de que generalizar sobre híbridos desconocidos es más fácil que generalizar sobre ambientes desconocidos también para las redes neuronales. Luego, la varianza en los desempeños entre los diferentes folds también es superior al evaluar con *AD* que con *HD*, lo que indica que la variabilidad de los datos ambientales de clima no vistos durante el entrenamiento en *AD*, introduce mayor ruido en la predicción que la variabilidad de los datos genómicos no vistos en *HD*. Por último, se cumple en todos los casos que la variante `MULTI` es la que mejor desempeño alcanza, lo que indica que siempre es beneficioso contar con ambas modalidades de datos.

8.5. Comparación entre modelos

Tras realizar estos experimentos, seleccionamos para cada esquema y modelo las variantes que mejores resultados obtuvieron para compararlas entre sí.

En primer lugar, notar de las figuras 8.14 y de 8.15 que, como ya habíamos visto en cada sección, las variantes de cada modelo que mejores resultados obtuvieron fueron siempre las multimodales. Otro comportamiento general que se puede observar fue que

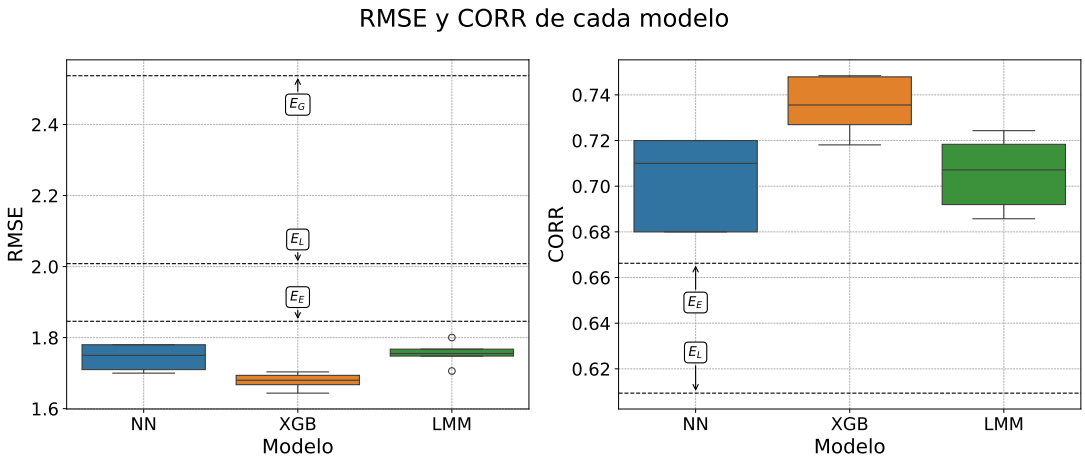


Figura 8.14: *Boxplots* de *CORR* y *RMSE* para para cada *fold*, de la mejor variante de cada modelo, utilizando el esquema de validación *HD*. En lineas punteadas se marcan los errores y correlaciones de comparación.

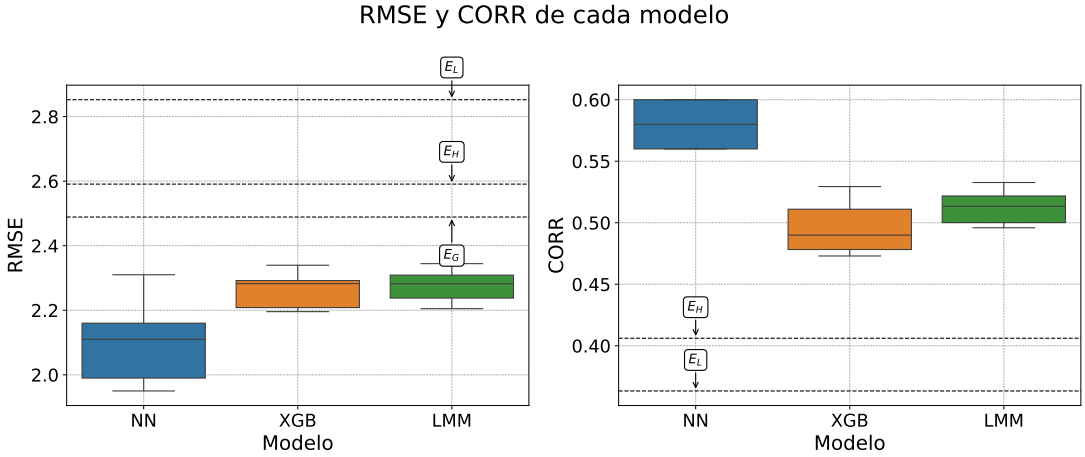


Figura 8.15: Análogo a la figura previa pero utilizando el esquema de validación *AD*.

todos los modelos alcanzaron mejores resultados en el esquema *HD* que en *AD*. Además, cabe destacar que, mientras en *HD* los mejores resultados los obtuvo *XGBoost*, en *AD* fueron las redes neuronales las que lograron los mejores resultados.

En cuanto a los modelos lineales mixtos, en *HD* desempeñaron prácticamente igual que las redes neuronales, teniendo los primeros un grado mucho menor de complejidad. Sin embargo, en *AD*, que resultó ser el esquema de mayor dificultad, sí fueron superados por las redes neuronales.

Resulta interesante que los modelos más simples fueron los que mejor pudieron utilizar la información genómica para generalizar sobre genomas desconocidos. En contraparte, las redes, mucho más complejas, fueron las que mejor entendieron el efecto del ambiente al momento de generalizar sobre condiciones ambientales desconocidas.

En particular, que al momento de generalizar sobre híbridos, los modelos lineales, que solo son capaces de captar relaciones de este tipo, tuviesen resultados similares a las redes, que pueden aprender cualquier tipo de relación, indica que el efecto del genoma sobre el fenotipo es mayormente aditivo [3]. También podría haber pasado que los efectos más complejos del genoma, no los estemos siendo capaces de aprender ni siquiera con las redes. Otra opción es que al momento de hacer la selección de los *SNPs*, hayamos descartado posiciones del genoma que servían para entender estas interacciones más complejas.

Como resumen de todos los análisis realizados en esta sección, presentamos en la tabla 8.8 los resultados correspondientes a la mejor variante de cada modelo, en cada esquema, promediados entre *folds*.

Esquema	Modelo	Variante	CORR	RMSE
HD	NN	`GENO_WEATHER`	0.702	1.744
	XGB	`GENO_ENV`	**0.735**	**1.678**
	LMM	`GENO_WEATHER`	0.705	1.755
AD	NN	`GENO_WEATHER`	**0.580**	**2.104**
	XGB	`GENO_ENV`	0.496	2.264
	LMM	`GENO_WEATHER`	0.513	2.276

Tabla 8.8: Promedio entre *folds* de ambas métricas, para la mejor variante de cada modelo y en cada esquema de validación. Se marca en negrita, para cada esquema, el mejor resultado de cada métrica.

[3]Que el efecto del genoma sea aditivo significa que el fenotipo se puede escribir como una combinación lineal de los *SNPs* ya codificados. O sea $F_G = \sum_i a_i g_i$, donde los g_i representan los *SNPs* codificados y los a_i los coeficientes de la combinación.

Capítulo 9

Importancia de variables

En este capítulo vamos a proponer e implementar un método que permite estimar la importancia que el modelo le da a cada variable, incluso habiendo utilizado *PCA*. Vamos a aplicarlo únicamente al segundo modelo, el del *XGBoost*, pues es justamente el que utiliza *PCA* en su arquitectura. Además, este modelo fue el que mejores resultados obtuvo para el esquema *HD* de híbridos desconocidos, por lo que fue el que mejor aprendió el efecto del genoma en el fenotipo. Por esto, resulta especialmente interesante entender sobre qué partes del genoma está prestando mayor atención este modelo.

9.1. Valores SHAP

Para obtener la importancia que el modelo le da a cada una de sus entradas, utilizamos la técnica *SHAP* [31] (*SHapley Additive exPlanations*). Está técnica posee varias ventajas, una de ellas es que mide la importancia en las mismas unidades que la variable a predecir. Lo que hace esta técnica es estimar qué tanto pueden modificar las predicciones cada una de las variables. Esto facilita la interpretabilidad, pues hace que las importancias tengan sentido individualmente, más allá de la comparación.

Otra de las ventajas es que permite ver las importancias a nivel modelo o a nivel muestra. Esto quiere decir que, además de entender a qué variables da más importancia el modelo, podemos también entender cómo el modelo llega a una predicción concreta dada.

Finalmente, otra de las ventajas es que la técnica, si bien está pensada para funcionar de manera agnóstica al modelo, tiene un procedimiento especializado para explicar el comportamiento de árboles de decisión [30], que es el tipo de modelo que estamos utilizando.

9.2. Importancia de las componentes

Preliminarmente decidimos utilizar *SHAP* para analizar que importancia le daban los modelos de *XGBoost* a cada una de sus entradas, es decir a las componentes *PCA*. Para hacer este análisis, seleccionamos de ambos esquemas, el modelo correspondiente a la variante que mejor rendimiento obtuvo. Aunque en este caso ambos modelos pertenecen

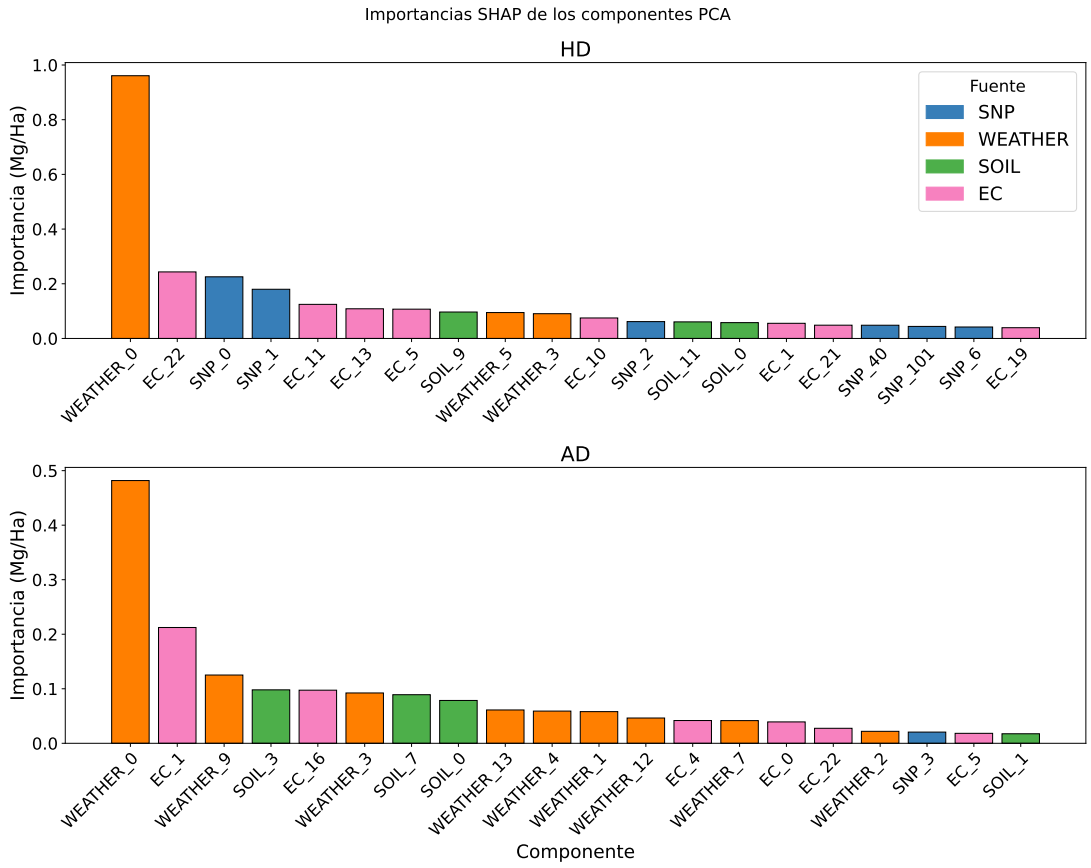


Figura 9.1: Veinte componentes con mayor importancia, para los modelos de cada esquema. Las componentes se nombran primero con la fuente de la que provienen, y luego con un número que indica el orden de la componente, siendo la número cero la que mayor varianza captura.

a la misma variante (`GENO_ENV`), son diferentes entre si, pues fueron entrenados bajo distintos paradigmas y con distintos datos.

Si bien en la sección previa ya habíamos analizado de manera cualitativa qué tan importante era cada fuente de datos, este análisis nos permitió hacerlo también de una manera cuantitativa.

Componentes más importantes

Para hacer el análisis con el modelo de *AD* evaluamos el mismo sobre el conjunto de test. Por otro lado, como en *HD* no teníamos conjunto de test, usamos tres de los cinco *folds* para entrenar, otro para monitorear el *Early Stopping* y sobre el *fold* restante evaluamos el modelo.

En la figura 9.1 se puede observar que en ambos modelos la componente más relevante es la primera de los datos climáticos, lo cual nos indica de la importancia de las condiciones climáticas para el desarrollo del fenotipo. Otro detalle a tener en cuenta es que, en *HD*, dentro de las componentes más importantes aparecen varias provenientes

de los *SNPs*, mientras que en *AD* apenas aparece una en el top veinte. Esto último se puede asociar a que, como en *HD* hay que generalizar sobre híbridos, es natural que el modelo de este esquema preste más atención a la información genómica.

Luego en la figura 9.2 agrupamos las importancias por fuente y las agregamos de distintas maneras.

Al acumular importancia por fuente vemos como para *HD* la fuente más importante es la genómica, aunque no es más relevante que el conjunto de todas las ambientales. En cambio, en *AD* la información genómica pasa a tener muy poca importancia, tanto si se compara con el resto con fuentes, como si se compara con la que tenía en *HD*. Esto hace sentido con lo que comentamos previamente de que al no tener que generalizar sobre híbridos en *AD*, la información genómica pierde valor. Más en general vemos como las importancias son menores en *AD* que en *HD*, esto indica que la salida del modelo de *HD* tiene más varianza que la de *AD*.

Después, al promediar la importancia por cantidad de componentes en cada fuente vemos que tanto en *HD* y *AD* las componentes que más aportan individualmente son las que provienen de los datos climáticos. En cambio, al promediar por la cantidad de variable originales, es decir, antes de aplicar *PCA*, vemos como las variables que resultan más útiles son las asociadas a los suelos. Esto surge de que las variables del suelo, siendo solo un 0,439 % del total, capturan un 7,138 % de la importancia final en *HD* y un 16,27 % en *AD*.

Otro detalle no menor es que en *HD* necesitamos 171 componentes para capturar el 95 % de la varianza de los datos genómicos, mientras que en *AD* necesitamos 315, casi el doble. Esto habla de que los híbridos utilizados en los conjuntos de entrenamiento de *AD*, tienen mucha más variabilidad que los de los conjuntos de *HD*. La explicación de esto está en que en los conjuntos de entrenamiento de *HD* aparece solo una generación (2020-2021), mientras que los de *AD* aparecen dos (2020-2021 y 2018-2019).

9.3. Recuperación de variables más importantes

Previamente mostramos cómo se podía utilizar la técnica *SHAP* para obtener las importancias asociadas a cada componente de los *PCA*, que son las que utiliza el modelo para predecir. Sin embargo, para interpretar realmente qué estaba viendo el modelo era necesario obtener las importancias de cada variable previas a aplicar *PCA*.

Si bien la reducción de dimensionalidad mediante *PCA* es un proceso no reversible, pues se pierde información, es posible redistribuir la importancia de las componentes entre las variables que las formaron.

Proceso de redistribución de importancias

Cada fuente de datos se puede representar como una matriz $X \in \mathbb{R}^{n \times m}$, que vamos a suponer centrada, donde n es la cantidad de muestras y m la de variables. Entonces, al aplicar *PCA* lo que estamos haciendo es multiplicar la matriz X, por una matriz $Q \in \mathbb{R}^{m \times k}$, donde k es la cantidad de componentes. Esta matriz Q se compone de los primeros k autovectores de la matriz de covarianza.

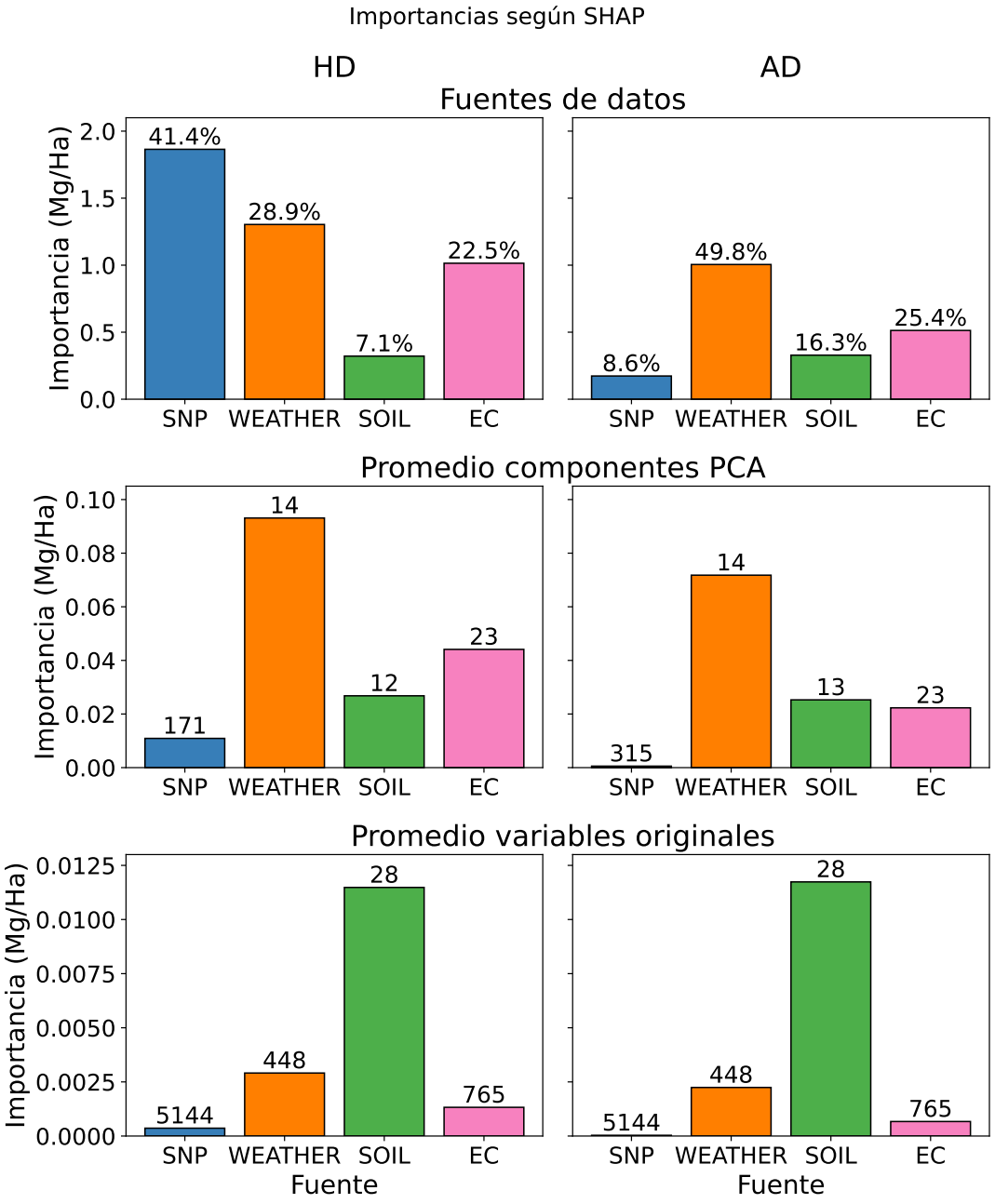


Figura 9.2: En la primera fila, importancia acumulada por fuente de información. En la segunda, importancia promedio de cada componentes. En la tercera, importancia promedio de cada variable original. En la primera columna para *HD* y en la segunda para *AD*. En la primera y segunda fila se marca la cantidad de componentes por fuente, y en la tercera la cantidad de variables.

$$Q = \begin{bmatrix} v_1 & v_2 & \cdots & v_i & \cdots & v_k \end{bmatrix}, \quad v_i \in \mathbb{R}^m.$$

Así obtenemos los datos reducidos $Y \in \mathbb{R}^{n \times k}$, donde $Y = XQ$. Por lo que la componente i-ésima se puede escribir como

$$Y_i = Xv_i = \sum_{j=1}^{m} X_j v_{ij}.$$

Es decir que los valores de los v_i indican en que proporción se combinan las distintas variables para crear la i-ésima componente.

Por lo que si a la i-ésima componente, se le asigna una importancia I_{Y_i}, es razonable dividir esta importancia usando las proporciones de v_i. Por lo que para saber la importancia de una variable, hay que recorrer todas las componentes e ir sumando la proporción de importancia que le corresponde.

Sin embargo hay una serie de detalles a tener en cuenta, pues los v_i no son exactamente vectores de proporciones. El primero es que como los v_i indican direcciones, sus valores pueden ser positivos o negativos. O sea que si queremos usar sus valores como proporciones, hay que tomar el valor absoluto de estos. El segundo es que buscamos que la importancia se conserve, es decir que si las componentes tenían una importancia total $I_{Y_T} = \sum_{j=1}^{k} I_{Y_j}$, queremos que las variables originales tengan la misma, o sea $I_{X_T} = \sum_{i=1}^{m} I_{X_i} = I_{Y_T}$. La forma de lograr esto es que los valores absolutos de los v_i estén normalizados, es decir que $\sum_{j=1}^{m} |v_{ij}| = 1$.

Por lo que si definimos

$$(\tilde{Q})_{ij} = \frac{|Q_{ij}|}{\sum_{j=1}^{m} |Q_{ij}|},$$

que no es más que Q adaptada para que sus columnas sí sean proporciones, podemos recuperar las importancias originales

$$I_{X_i} = \sum_{j=1}^{k} I_{Y_j} \tilde{Q}_{ij}.$$

Donde I_{X_i} es la importancia de la i-ésima variable original. También podemos escribirlo en forma matricial como:

$$I_X = I_Y \tilde{Q}^T.$$

9.4. Variables más importantes

Al redistribuir las importancias que habíamos calculado antes entre las variables originales, pudimos calcular las variables más importantes para cada modelo, como se ve en 9.1.

Todas las variables que aparecían dentro de la diez más importantes, tanto para *HD* como para *AD*, eran del suelo. Si bien esto concordaba con los análisis previos, es posible que para *HD* dichas variables solo estuviesen siendo usadas para identificar el ambiente.

Top	AD	HD	HD (SNPs)
1	WDRF Buffer pH	Is Sandy Loam	S3_181064661
2	%H Sat	%K Sat	S10_10451292
3	%Ca Sat	%Silt	S7_165046947
4	1:1 Soil pH	Is Silty Clay Loam	S2_205343430
5	Nitrate-N ppm N	%Sand	S4_211555188
6	lbs N/A	Potassium ppm K	S1_302246496
7	Calcium ppm Ca	%Mg Sat	S7_14190560
8	%Silt	%H Sat	S7_161184307
9	%Mg Sat	Nitrate-N ppm N	S5_223681991
10	1:1 S Salts mmho/cm	E Depth	S6_53031607

Tabla 9.1: Top diez de variables más importantes para los modelos de cada esquema. Para *HD* se muestra además, en la última columna, los diez marcadores más importantes.

Por eso para *HD* mostramos además los marcadores más importantes, que son los que de verdad usa el modelo para generalizar.

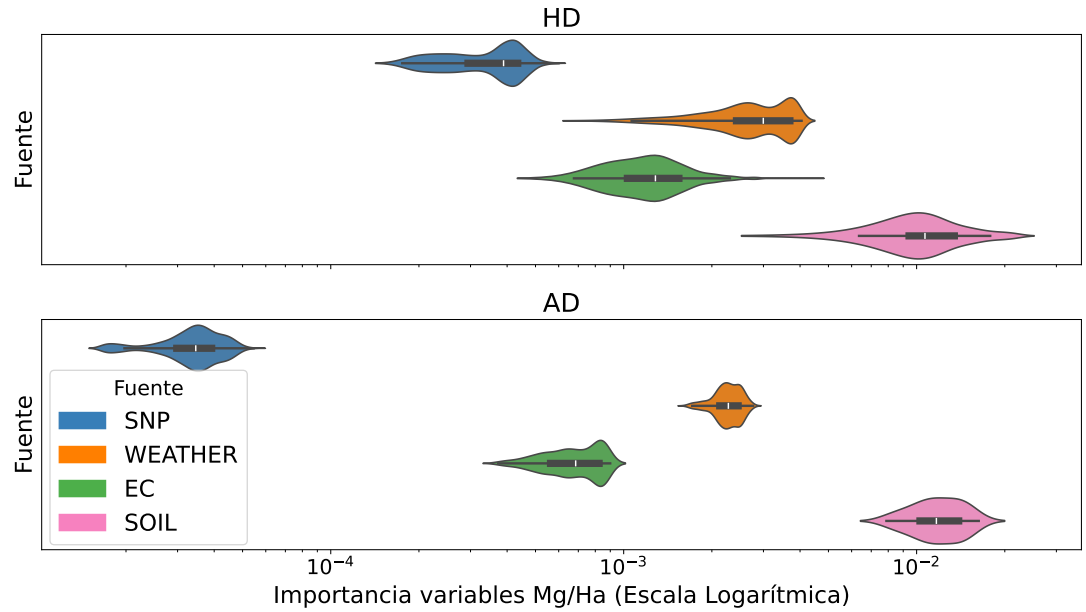


Figura 9.3: Gráfico de violín para las importancias de cada variable, segmentado por fuente. Arriba para *HD* y abajo para *AD*.

Finalmente analizamos la distribución de las importancias de cada variable, como se ve en 9.3. Estas distribuciones nos confirman que la información genómica es más importante en *HD* que en *AD*. Pues incluso los marcadores menos relevantes en *HD*,

siguen siendo más importantes que cualquier otro en *AD*. Por otro lado, en cuanto a las variables ambientales, las importancias tienen un comportamiento similar en ambos esquemas, salvando que hay menor variabilidad de las mismas en *AD*. Esto último se puede interpretar como que en *AD* el modelo presta atención a las variables ambientales de forma más equitativa que en *HD*.

Esta página ha sido intencionalmente dejada en blanco.

Capítulo 10

Conclusiones

10.1. Importancia de la información ambiental

El objetivo principal de este proyecto era investigar si en el contexto de la predicción genómica, la información ambiental es capaz de complementar a la genómica para obtener mejores resultados. En base a todos los resultados obtenidos, para los distintos modelos y esquemas de validación, concluimos que aquellas predicciones que tienen en cuenta la información ambiental, siempre resultan mejores que las que no lo hacen. En particular, observamos que en algunos casos la información ambiental aporta más que la genómica. De todas maneras, los mejores resultados siempre se obtuvieron al utilizar ambas fuentes de información.

Por otro lado, se mostró la importancia de la selección del esquema de validación. Observamos que el problema de predecir sobre híbridos desconocidos es totalmente diferente al de predecir en ambientes desconocidos. Siendo el último un problema mucho más difícil de abordar, donde las predicciones resultan menos precisas.

En base a los comportamientos descritos previamente y a los resultados obtenidos, concluimos que la componente ambiental juega un rol más determinante que la componente genómica en el problema abordado. Por lo tanto, estamos en condiciones de afirmar que el fenotipo, en las condiciones dadas, es más susceptible a factores ambientales que genéticos.

10.2. Modelos

En cuanto a los modelos utilizados, obtuvimos resultados variados. Por un lado, cuando hubo que generalizar sobre híbridos desconocidos, el *XGBoost* fue el modelo que mejor resultados obtuvo. En cuanto a las redes, pese a su capacidad de aprender patrones, no pudieron superar a los modelos lineales mixtos. Esto último, indica que modelar los efectos del genoma como aditivos, sigue siendo una buena aproximación del problema.

Sin embargo, cuando hubo que generalizar sobre condiciones ambientales desconocidas, las redes fueron las que mejores resultados obtuvieron. Esto muestra que la complejidad de las redes, especialmente en la forma en que tratan las series temporales, resulta

útil para entender los efectos climatológicos en el desarrollo de las plantas.

Más allá de lo realizado durante este proyecto, hay algunas pruebas que por motivos de tiempo no pudimos realizar, y que podrían haber mejorado los resultados. A continuación presentamos un lista de las mismas.

- Cuando diseñamos la arquitectura del *XGBoost* decidimos quedarnos con el 95 % de la varianza en el *PCA*. Seleccionamos este número porque apenas perdíamos un 5 % de la varianza y reducíamos significativamente la dimensionalidad de los datos. Sin embargo, pruebas aisladas demostraron que reduciendo aún más la dimensionalidad, se lograban resultados mejores. Por esto es que resultaría interesante analizar el porcentaje de varianza a conservar como un hiperparámetro más del modelo.

- Otra posible mejora que merece ser investigada es incorporar el resto de datos ambientales a las redes neuronales, y no solo los climáticos. La motivación de esto surge de que el resto de modelos tuvieron un mejor desempeño cuando utilizaron todos los datos ambientales, en comparación a cuando usaron solo los climáticos.

- Finalmente, mencionar que al momento de optimizar los hiperparámetros que definieron la arquitectura de la red, nos vimos limitados por la capacidad computacional. Es por esto que si hubiésemos dispuesto de más capacidad computacional, o en su defecto, más tiempo, hubiéramos refinado más la búsqueda de estos hiperparámetros.

10.3. Trabajo futuro

Pensando en trabajos futuros, este proyecto deja varios frentes de investigación abiertos. En los siguientes puntos detallamos algunos de ellos.

- En este proyecto utilizamos dos esquemas de validación, sin embargo, cuando fuimos a evaluar los datos de test, nos dimos cuenta que estos no caían en el paradigma de ninguno de los dos. Es por eso que creemos importante el utilizar otros esquemas de validación, sobre todo el que describimos en 6.2.

- Como ya mencionamos previamente, el estar recortando la cantidad de *SNPs* a utilizar, puede estar haciendo que perdamos información valiosa. Sobretodo, al momento de entender los efectos más complejos del genoma y la interacción del mismo con el ambiente. Por esto es que resulta interesante el implementar modelos que utilicen un mayor número de *SNPs*, o incluso la totalidad de ellos.

- En 5.1 hablamos de los *metadatos*, una serie de registros que describen las condiciones en que se realizaron los cultivos. Estos datos no los utilizamos pero podrían haber aportado información, sobretodo habiendo visto la importancia que tienen los factores externos por sobre los genéticos.

- En los datos de la competencia, además del tonelaje, había muchos otros fenotipos registrados para cada experimento. Sería interesante por tanto ver como desempeñan los distintos modelos al intentar predecir otros fenotipos. En particular se podría analizar que partes del genoma afectan a cada fenotipo.

Esta página ha sido intencionalmente dejada en blanco.

Apéndice A

Variables de la base de datos

Variables fenotípicas

Nombre	**Descripción**
Pollen_DAP_days	Días después de la siembra hasta que el 50 % de las plantas liberan polen.
Silk_DAP_days	Días después de la siembra hasta que el 50 % de las plantas tienen sedas visibles.
Plant_Height_cm	Altura de la planta desde la base hasta el lígula de la hoja bandera (cm).
Ear_Height_cm	Altura desde el suelo hasta el nodo principal que lleva mazorca (cm).
Root_Lodging_plants	Número de plantas por parcela con encamado de raíces.
Stalk_Lodging_plants	Número de plantas por parcela rotas entre el nivel del suelo y el nodo de la mazorca superior.
Yield_Mg_ha	Rendimiento de grano en toneladas por hectárea al 15.5 % de humedad en el grano.
Grain_Moisture	Contenido de agua en el grano al momento de la cosecha (%).
Twt_kg_m3	Peso de prueba del grano desgranado (kg/m^3).

Variables del suelo

Variable	**Descripción**
E Depth	Profundidad de recolección de la muestra de suelo.
1:1 Soil pH	Nivel de pH del suelo en una mezcla, en peso, una parte de suelo por una parte de agua destilada (H_2O).
WDRF Buffer pH	Método de Woodruff para medir la acidez total del suelo.
1:1 S Salts mmho/cm	Concentración de sales solubles en el suelo.
Texture No	Análisis del tamaño de partículas (PSA) con componentes minerales menores de 2 mm.
Organic Matter LOI %	Porcentaje de materia orgánica en el suelo.
Nitrate-N ppm N	Nitratos disponibles en partes por millón (ppm).
lbs N/A	Cantidad de nitrógeno en libras por acre.
Potassium ppm K	Potasio disponible en ppm.
Sulfate-S ppm S	Sulfato disponible en ppm.
Calcium ppm Ca	Calcio disponible en ppm.
Magnesium ppm Mg	Magnesio disponible en ppm.
Sodium ppm Na	Sodio disponible en ppm.
CEC/Sum of Cations me/100g	Capacidad de intercambio catiónico.
%H Sat	Porcentaje de hidrógeno.
%K Sat	Porcentaje de potasio.
%Ca Sat	Porcentaje de calcio.
%Mg Sat	Porcentaje de magnesio.
%Na Sat	Porcentaje de sodio.
Mehlich P-III ppm P	Extracción de fósforo con fluoruro ácido diluido.
% Sand	Porcentaje de composición de arena en la muestra de suelo.
% Silt	Porcentaje de composición de limo en la muestra de suelo.
% Clay	Porcentaje de composición de arcilla en la muestra de suelo.
Texture	Proporción de partículas de tamaño arena, limo y arcilla que componen el suelo.
BpH	Método de Sikora para medir la acidez del suelo.
Zinc ppm Zn	Zinc disponible en ppm.
Iron ppm Fe	Hierro disponible en ppm.
Manganese ppm Mn	Manganeso disponible en ppm.
Copper ppm Cu	Cobre disponible en ppm.
Boron ppm B	Boro disponible en ppm.

Series temporales climáticas

Parámetro	**Descripción**
ALLSKY_SFC_PAR_TOT	Radiación Fotosintéticamente Activa (PAR) en Superficie Total bajo Cielo Abierto (W/m^2).
T2MWET	Temperatura de Bulbo Húmedo a 2 Metros (°C).
QV2M	Humedad Específica a 2 Metros (g/kg).
RH2M	Humedad Relativa a 2 Metros (%).
T2M_MAX	Temperatura Máxima a 2 Metros (°C).
ALLSKY_SFC_SW_DWN	Irradiancia Descendente de Onda Corta en Superficie bajo Cielo Abierto (MJ/m^2/día).
PS	Presión en la Superficie (kPa).
GWETPROF	Humedad del Suelo en el Perfil.
T2MDEW	Punto de Rocío/Escarcha a 2 Metros (°C).
GWETTOP	Humedad Superficial del Suelo.
WS2M	Velocidad del Viento a 2 Metros (m/s).
T2M_MIN	Temperatura Mínima a 2 Metros (°C).
T2M	Temperatura a 2 Metros (°C).
GWETROOT	Humedad del Suelo en la Zona de Raíces.
PRECTOTCORR	Precipitación Corregida (mm/día).
ALLSKY_SFC_SW_DNI	Irradiancia Directa Normal Descendente de Onda Corta en la Superficie para Todo el Cielo

Covariables ambientales

Las covariables ambientales son una combinación de una de las siguientes variables:

Variable	**Descripción**
SDR	Relación Oferta-Demanda de Agua.
TT	Tiempo Térmico (Celsius).
biomass	Biomasa aérea.
yield	Rendimiento de grano.
Eo	Evapotranspiración potencial del sistema suelo-planta completo (mm).
Eos	Evaporación potencial desde la superficie del suelo (mm).
Es	Evaporación real de agua del suelo (mm).
ESW	Agua extraíble del suelo relativo a LL15 (mm).
Flow	Agua moviéndose hacia arriba (mm).
FlowNO3	Pérdida de nitrógeno como NO3 de cada capa del suelo (kg/ha).
Flux	Agua moviéndose hacia abajo (mm).
Infiltration	Infiltración (mm).
PAWmm	Agua disponible para las plantas (SW-LL15) (mm).

PotInf	Infiltración potencial: precipitación menos agua interceptada por el dosel y los residuos.
PotRunoff	Escorrentía potencial (mm).
Runoff	Escorrentía (mm).
SWmm	Contenido de agua en el suelo (mm).
CoverGreen	Cobertura total de planta verde de todos los órganos.
CoverTotal	Cobertura total de planta de todos los órganos.
LAI	Índice de área foliar.
AccumulatedTT	Tiempo Térmico Acumulado (Celsius).

, con uno de los siguientes periodos fenológicos:

Variable	**Descripción**
pGerEme	Germinación a emergencia
pEmeEnJ	Emergencia al fin de la etapa juvenil
pEnJFlo	Fin de la etapa juvenil a la iniciación floral
pFloFla	Iniciación floral a la hoja bandera
pFlaFlw	Hoja bandera a floración
pFlwStG	Floración al inicio del llenado de grano
pStGEnG	Inicio del llenado de grano al fin del llenado de grano
pEnGMat	Fin del llenado de grano a madurez
pMatHar	Madurez a la cosecha/madurez comercial

y con una capa del suelo; representada con un número del uno al diez, donde uno es la capa más superficial y diez la más profunda.

Apéndice B

Entrenamiento de los modelos

B.1. Calculo de métricas

A continuación se detalla el cálculo de las métricas usadas para evaluar los modelos de *XGBoost* que no se utilizaron en los otros modelos.

Sean $\{y_i\}_{i=1}^N$ el conjunto de valores reales e $\{\hat{y_i}\}_{i=1}^N$ el de los predichos, entonces se definen:

- MAE (*Mean Absolute Error*):

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|.$$

- $MAPE$ (*Mean Absolute Percentage Error*):

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\frac{|y_i - \hat{y}_i|}{|y_i|}.$$

- $NRMSE$ (*Normalized Root Mean Squared Error*):

$$MAPE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\frac{y_i - \hat{y}_i}{y_i})^2}.$$

- R^2:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(\hat{y}_i - \tilde{y})^2} \quad \text{con} \quad \tilde{y} = \frac{1}{N}\sum_{i=1}^{N} y_i.$$

B.2. Hiperparámetros de los modelos

XGBoost

En B.1 y B.2 se pueden ver los hiperparámetros óptimos para cada variante en cada esquema. Los hiperparámetros que se optimizaron fueron:

- `max_depth` (**MD**): Máxima profundidad que puede alcanzar un árbol.
- `eta` (**E**): *Learning Rate*.
- `subsample` (**Ss**): Proporción de submuestreo durante el entrenamiento.
- `colsample_bytree` (**CsBt**): Proporción de columnas utilizadas para la construcción de cada árbol.
- `gamma` (**G**): Mejora mínima requerida para realizar una partición de un nodo hoja del árbol.
- `reg_lambda` (**RL**): Termino de la regularización *L2*.
- `reg_alpha` (**RA**): Termino de la regularización *L1*.

Para más detalle sobre cada uno de estos hiperparámetros consultar [8].

HD

Variante	MD	E	Ss	CsBt	G
GENO	5	0.01	0.8	0.3	1.0
WEATHER	10	0.1	0.8	0.3	5.0
ENV	3	0.05	0.8	1.0	5.0
GENO_WEATHER	4	0.05	1.0	1.0	0.0
GENO_OH_L	5	0.1	0.9	0.3	1.0
GENO_OH_E	5	0.05	0.9	0.3	5.0
GENO_ENV	5	0.03	0.9	0.9	0.1

Tabla B.1: Hiperparámetros óptimos de cada variante, utilizando el esquema *HD*.

AD

Redes Neuronales

En B.3 y B.4 se pueden ver los hiperparámetros óptimos para cada variante en cada esquema. Esta vez los hiperparámetros que se optimizaron fueron los que definen la arquitectura de la red. En este caso fueron:

- `genotype_encoder/input_embedding/embedding_dim` (**GED**)

Variante	MD	E	Ss	CsBt	G	RL	RA
GENO	10	0.01	1.0	0.3	1.0	-	-
WEATHER	2	0.1	1.0	1.0	1.0	-	-
ENV	3	0.1	0.9	0.8	0.0	-	-
GENO_WEATHER	3	0.1	1.0	1.0	5.0	-	-
GENO_OH_L	15	0.05	1.0	0.5	0.0	100.0	0.0
GENO_ENV	9	0.01	1.0	0.6	0.0	250.0	250.0

Tabla B.2: Hiperparámetros óptimos de cada variante, utilizando el esquema *AD*.

- genotype_encoder/cnn_feature_extractor/cnn_stack/n_cnn_blocks (**GnCB**)
- genotype_encoder/embedding_layer/hidden_features" (**GEHF**)
- regressor/hidden_features (**RHF**)
- weather_encoder/cnn_feature_extractor/cnn_stack/n_cnn_blocks (**WnCB**)
- weather_encode/embedding_layer/hidden_features (**WEHF**)

HD

Variante	GED	GnCB	GEHF	RHF	WnCB	WEHF
GENO	4	5	256	256	-	-
WEATHER	-	-	-	128	4	256
MULTI	2	6	256	512	5	256

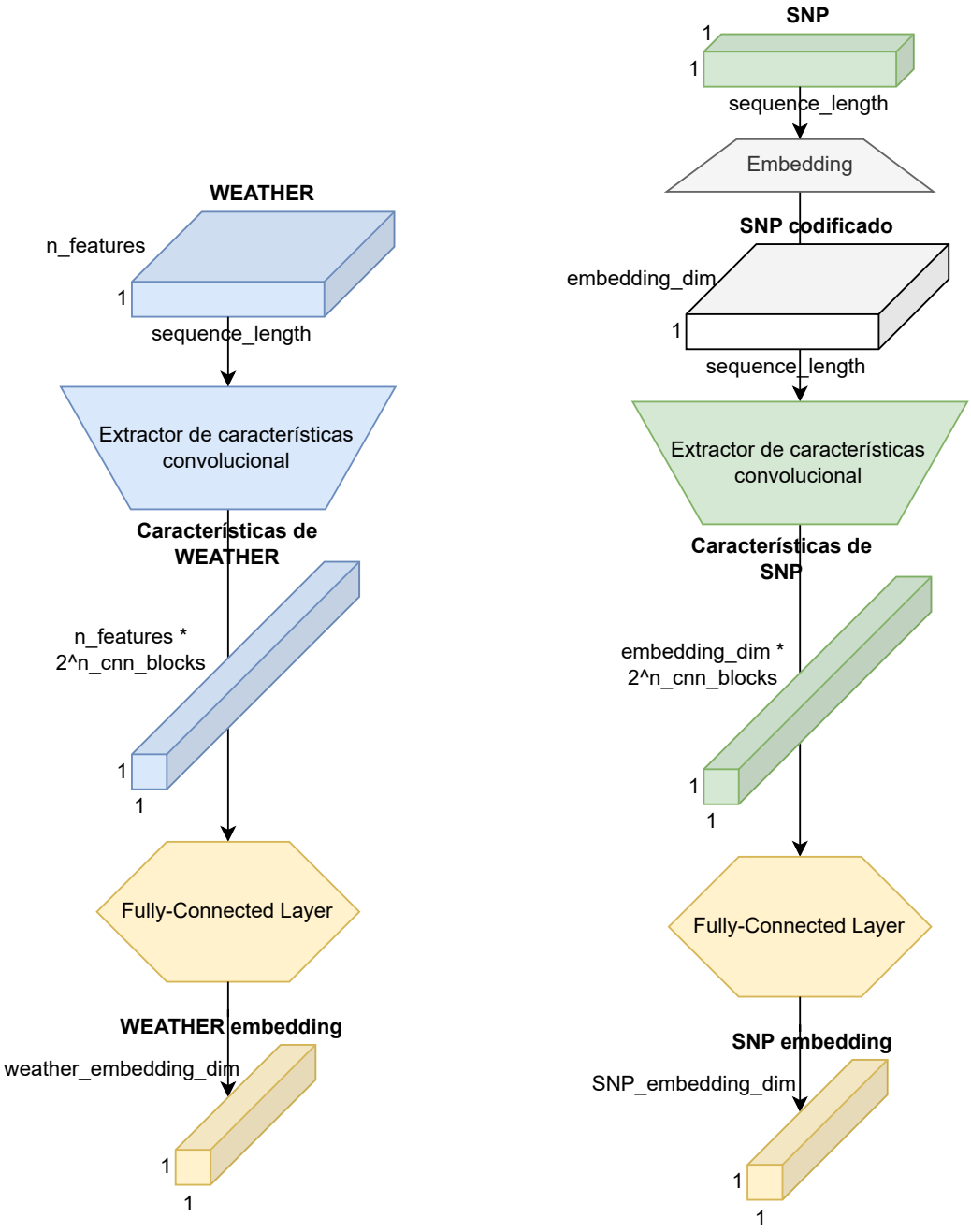
Tabla B.3: Hiperparámetros óptimos de cada variante, utilizando el esquema *HD*.

AD

Variante	GED	GnCB	GEHF	RHF	WnCB	WEHF
GENO	2	3	128	512	-	-
WEATHER	-	-	-	256	3	256
MULTI	4	6	128	512	3	256

Tabla B.4: Hiperparámetros óptimos de cada variante, utilizando el esquema *AD*.

B.3. Implementación de bloques de redes neuronales



(a) Diagrama del *Encoder* de datos ambientales de clima.

(b) Diagrama del *Encoder* de SNPs.

Figura B.1: Diagramas para los *Encoder*.

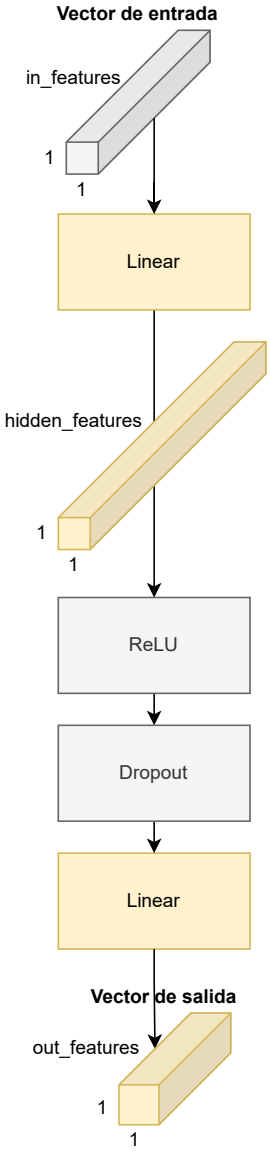
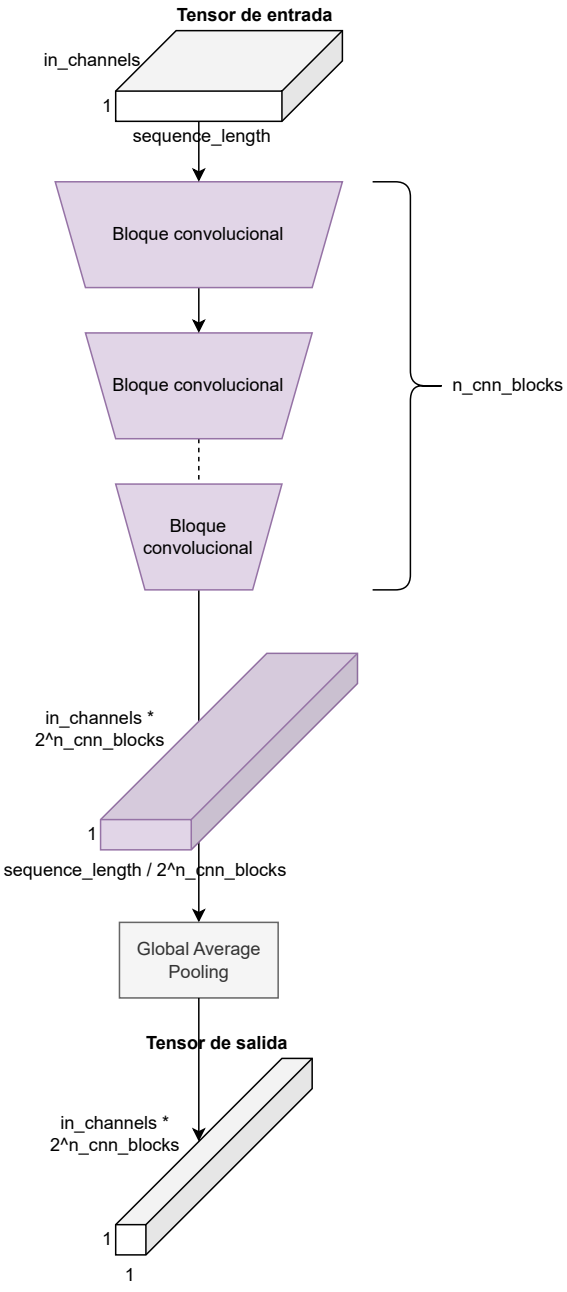
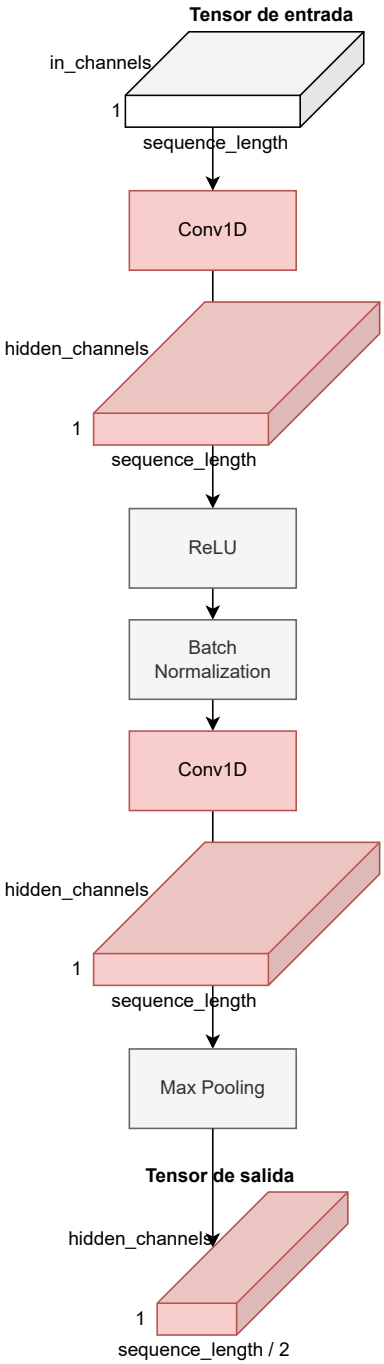


Figura B.2: Diagrama del bloque completamente conectado.



(a) Diagrama del extractor de características convolucional.



(b) Diagrama del bloque convolucional.

Figura B.3: Diagramas de los módulos basados en redes convolucionales.

Apéndice C

Análisis utilizando solo datos genómicos

C.1. Reformulación del problema

Todos aquellos registros correspondientes a un mismo híbrido plantado en distintos ambientes, ahora que no vamos a utilizar la componente ambiental, van a tener la misma información total. Es decir, que tenemos muestras con los mismos datos pero distinto valor de fenotipo a predecir. Fue por esto que para este análisis decidimos promediar el desempeño de los híbridos en los distintos ambientes en que fueron plantados. De esta manera obtuvimos tantos registros como híbridos únicos en el *dataset*, es decir 4415, cada uno con sus 5144 *SNPs*.

Notar que este es un problema sustancialmente distinto al original, pues ahora no queremos predecir el desempeño de una plantación concreta, sino el de un híbrido históricamente.

C.2. Modelado

Para estas pruebas vamos a utilizar solamente los modelos *GBLUP* y *XGBoost*. Para realizar las mismas se entrenaron los modelos haciendo validación cruzada con diez *folds*, utilizando el 80 % de los híbridos para entrenamiento y el 20 % para test (esto implica que los *folds* de validación no fueron disjuntos). En la figura C.1 se muestran los resultados de *RMSE* y correlación de *Pearson* para los distintos modelos.

Además, a modo de comparación, se calculó para cada *fold* un *RMSE* base (E_G), correspondiente a predecir utilizando el promedio del fenotipo en el conjunto de entrenamiento [1]. Como en términos de error cuadrático la mejor predicción constante es el promedio, cualquier resultado que obtengamos por debajo de E_G significa que el modelo está aprendiendo de la información genómica.

Modelo	RMSE	CORR
E_G	1.471	-
`XGBoost`	**0.899**	**0.791**
`GBLUP Sommer`	1.012	0.758
`GBLUP rrBLUP`	0.943	0.767

Tabla C.1: Promedio entre *folds* de *RMSE* y correlación, para cada modelo. La primera fila corresponde al error de comparación E_G, el cual también está promediado a nivel de *fold*. Se marca en negrita el mejor resultado de cada métrica.

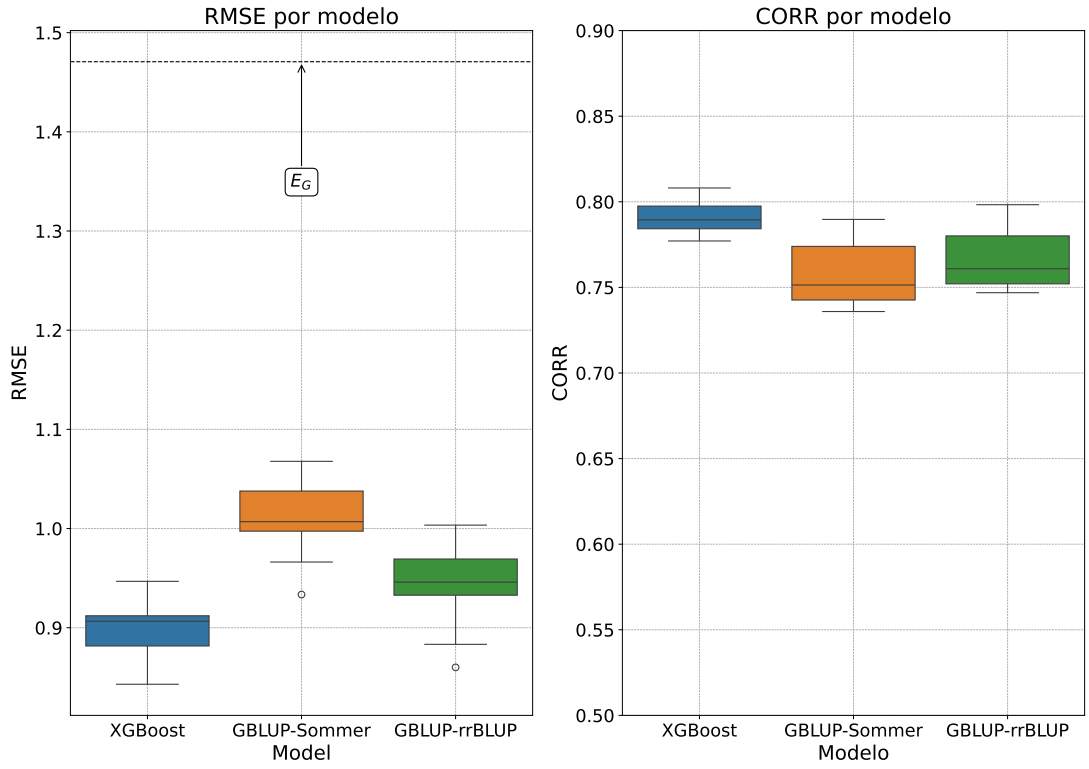


Figura C.1: *RMSE* y correlación a nivel *fold* para cada modelo. Además se muestra el valor promedio de los E_G con una linea punteada.

C.3. Análisis de resultados

De C.1 se puede observar que todos los modelos mejoran el error base, por lo que todos ellos están siendo capaces de aprender de la información genómica. Respecto a la comparación de modelos, en C.1 se observa que el *XGBoost* supera levemente el rendimiento de los modelos lineales, aunque sus *boxplots* se intersectan.

[1]Se puede interpretar a E_G como la desviación estándar del fenotipo en el conjunto de entrenamiento.

Observar que se utilizaron dos librerías distintas para implementar el modelo *GBLUP*, *sommer* y *rrblup*. Se obtuvieron desempeños similares con ambas librerías, pero *rrblup* tuvo un menor tiempo de ejecución por lo que fue la que utilizamos de aquí en adelante para aplicar el modelo *GBLUP*.

De estas pruebas concluimos que la información genómica de un híbrido puede ser usada para predecir su desempeño histórico. Esto lo podíamos intuir, pues cuando seleccionamos los *SNPs* observamos que algunos de ellos tenían una correlación considerable con el fenotipo.

Por último, recordar que estos resultados no son comparables a los del análisis multimodal, donde se aborda el problema original. En particular, podemos decir que en este enfoque, como estamos promediando fenotipos, estamos reduciendo la varianza de los mismos, por lo cual es esperable tener un error menor.

Esta página ha sido intencionalmente dejada en blanco.

Referencias

[1] Christina B Azodi, Emily Bolger, Andrew McCarren, Mark Roantree, Gustavo de Los Campos, and Shin-Han Shiu. Benchmarking parametric and machine learning models for genomic prediction of complex traits. *G3: Genes, Genomes, Genetics*, 9(11):3691–3702, 2019.

[2] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

[3] Leo Breiman. *Classification and regression trees*. Routledge, 2017.

[4] Amar Budhiraja. Learning less to learn better — dropout in (deep) machine learning. https://medium.com/@amarbudhiraja/https-medium-com-amarbudhiraja-learning-less-to-learn-better-dropout-in-deep-machine-learning-74334da4bfc5, 2016.

[5] Juan Burgueño, Gustavo de los Campos, Kent Weigel, and José Crossa. Genomic prediction of breeding values when modeling genotype× environment interaction using pedigree and dense molecular markers. *Crop Science*, 52(2):707–719, 2012.

[6] DG Butler, BR Cullis, AR Gilmour, BJ Gogel, and R Thompson. Asreml-r reference manual version 4. *VSN International Ltd, Hemel Hempstead, HP1 1ES, UK*, 2017.

[7] Graciana Castro, Romina Hoffman, and Mateo Musitelli. Predgenia: Transformers para predicción genómica. Master's thesis, Universidad de la República, 2023.

[8] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, 2016. ACM.

[9] Gustavo de Los Campos and Daniel Gianola. Factor analysis models for structuring covariance matrices of additive genetic effects: a bayesian implementation. *Genetics Selection Evolution*, 39:1–14, 2007.

[10] Clara Dismuke and Richard Lindrooth. Ordinary least squares. *Methods and designs for outcomes research*, 93(1):93–104, 2006.

[11] Juan Elenter, Guillermo Etchebarne, and Ignacio Hounie. Dnai: Machine learning for genome enabled prediction of complex traits in agriculture. Master's thesis, Universidad de la República, 2021.

[12] Igor K Fernandes, Caio Canella Vieira, Kaio Olimpio das Gracas Dias, and Samuel B Fernandes. Using machine learning to integrate genetic and environmental data to model genotype-by-environment interactions. *bioRxiv*, pages 2024–02, 2024.

[13] Karl Pearson F.R.S. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.

[14] Aurélien Géron. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. .ºReilly Media, Inc.", 2022.

[15] Jussi Gillberg, Pekka Marttinen, Hiroshi Mamitsuka, and Samuel Kaski. Modelling g× e with historical weather information improves genomic prediction in new environments. *Bioinformatics*, 35(20):4045–4052, 2019.

[16] Charles R Henderson. Best linear unbiased estimation and prediction under a selection model. *Biometrics*, pages 423–447, 1975.

[17] Geoffrey Hinton, Nitish Srivastava, and Kevin Swersky. Overview of mini-batch gradient descent. *Neural Networks for Machine Learning*, 575(8), 2012.

[18] Ken Hoffman. Machine Learning: How to Prevent Overfitting — medium.com. `https://medium.com/swlh/machine-learning-how-to-prevent-overfitting-fdf759cc00a9`, 2021. [Accedido el 18-11-2024].

[19] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989.

[20] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift, 2015.

[21] Gareth James. *An introduction to statistical learning*. springer, 2013.

[22] Diego Jarquín, José Crossa, Xavier Lacaze, Philippe Du Cheyron, Joëlle Daucourt, Josiane Lorgeou, François Piraux, Laurent Guerreiro, Paulino Pérez, Mario Calus, et al. A reaction norm model for genomic selection using high-dimensional genomic and environmental data. *Theoretical and applied genetics*, 127:595–607, 2014.

[23] Diego Jarquín, Cristiano Lemes da Silva, R Chris Gaynor, Jesse Poland, Allan Fritz, Reka Howard, Sarah Battenfield, and Jose Crossa. Increasing genomic-enabled prediction accuracy by modeling genotype× environment interactions in kansas wheat. *The plant genome*, 10(2):plantgenome2016–12, 2017.

[24] Sheikh Jubair, Olivier Tremblay-Savard, and Mike Domaratzki. Gxenet: Novel fully connected neural network based approaches to incorporate gxe for predicting wheat yield. *Artificial Intelligence in Agriculture*, 8:60–76, 2023.

[25] Sheikh Jubair, James R Tucker, Nathan Henderson, Colin W Hiebert, Ana Badea, Michael Domaratzki, and WG Dilantha Fernando. Gptransformer: a transformer-based deep learning method for predicting fusarium related traits in barley. *Frontiers in plant science*, 12:761402, 2021.

[26] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30:3146–3154, 2017.

[27] Daniel R Kick, Jason G Wallace, James C Schnable, Judith M Kolkman, Barış Alaca, Timothy M Beissinger, Jode Edwards, David Ertl, Sherry Flint-Garcia, Joseph L Gage, et al. Yield prediction through integration of genetic, environment, and management data through deep learning. *G3: Genes, Genomes, Genetics*, 13(4):jkad006, 2023.

[28] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

[29] Y Liu, D Wang, F He, J Wang, T Joshi, and D Xu. Phenotype prediction and genome-wide association study using deep convolutional neural network of soybean. front genet 10: 1091, 2019.

[30] Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable ai for trees. *Nature Machine Intelligence*, 2(1):2522–5839, 2020.

[31] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017.

[32] Abelardo Montesinos-López, Carolina Rivera, Francisco Pinto, Francisco Piñera, David Gonzalez, Mathew Reynolds, Paulino Pérez-Rodríguez, Huihui Li, Osval A Montesinos-López, and Jose Crossa. Multimodal deep learning methods enhance genomic prediction of wheat breeding. *G3: Genes, Genomes, Genetics*, 13(5):jkad045, 2023.

[33] Osval A Montesinos-López, Abelardo Montesinos-López, Roberto Tuberosa, Marco Maccaferri, Giuseppe Sciara, Karim Ammar, and José Crossa. Multi-trait, multi-environment genomic prediction of durum wheat with genomic best linear unbiased predictor and deep learning methods. *Frontiers in Plant Science*, 10:1311, 2019.

[34] Osval Antonio Montesinos López, Abelardo Montesinos López, and José Crossa. *Multivariate statistical machine learning methods for genomic prediction*. Springer Nature, 2022.

[35] Prince Yadav. Decision tree in machine learning, 2018. [Online; accessed 24-November-2024].

[36] Susweta Ray, Diego Jarquin, and Reka Howard. Comparing artificial-intelligence techniques with state-of-the-art parametric prediction models for predicting soybean traits. *The Plant Genome*, 16(1):e20263, 2023.

[37] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

[38] David Saad. *On-line learning in neural networks*. Cambridge University Press, 2009.

[39] Swati Sharma, Aditi Partap, Maria Angels de Luis Balaguer, Sara Malvar, and Ranveer Chandra. Deepg2p: fusing multi-modal data to improve crop production. *arXiv preprint arXiv:2211.05986*, 2022.

[40] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

[41] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[42] Jacob D Washburn, Emre Cimen, Guillaume Ramstein, Timothy Reeves, Patrick O'Briant, Greg McLean, Mark Cooper, Graeme Hammer, and Edward S Buckler. Predicting phenotypes from genetic, environment, management, and historical data using cnns. *Theoretical and Applied Genetics*, 134(12):3997–4011, 2021.

[43] Cathy C Westhues, Gregory S Mahone, Sofia da Silva, Patrick Thorwarth, Malthe Schmidt, Jan-Christoph Richter, Henner Simianer, and Timothy M Beissinger. Prediction of maize phenotypic traits with genomic and environmental predictors using gradient boosting frameworks. *Frontiers in plant science*, 12:699589, 2021.

[44] Wikipedia contributors. Anscombe's quartet — Wikipedia, the free encyclopedia, 2024. [Online; accessed 18-November-2024].

[45] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks, 2013.

[46] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2):301–320, 2005.

Glosario

ambiente Combinación de locación y año. Representa una ubicación geográfica en un momento dado. 7, 8, 125

clon Planta con la misma información genética que otra. 7

experimento Combinación de híbrido y ambiente. Es la experiencia de plantar una variedad genética en unas condiciones ambientales concretas. 7, 41

heredabilidad Porcentaje de la varianza de un fenotipo que se debe únicamente al genoma. 6

híbrido Variedad genética de una planta. 7, 8, 125

locación Ubicación geográfica donde se realizan plantaciones. 7

SNP Variación con una frecuencia mayor al 1 % de una base nitrogenada concreta respecto a un genoma de referencia. 5, 6

tonelaje Peso de la cosecha de una plantación. Usualmente medido en toneladas por hectárea sembrada. 30

Esta página ha sido intencionalmente dejada en blanco.

Índice de tablas

5.1. Distribución acumulada de la distancia máxima que hubo entre las plantaciones de una misma locación a lo largo de los años. Dada la distancia de la izquierda, se muestra a la derecha que porcentaje de locaciones tienen una distancia máxima de variación menor a esa distancia. 31

5.2. Distribución del valor más frecuente (entre individuos) de la codificación de los SNPs. 35

6.1. Cantidad de muestras para cada *fold* de entrenamiento y validación, junto al porcentaje del total que representan las muestras de validación. 55

6.2. Cantidad de muestras para cada *fold* de entrenamiento, segregado por año y separando en el conjunto seleccionado (el que comparte híbridos con validación) y el complemento. Junto a ello también se reportan la cantidad de muestras para los *folds* de validación, las cuales son todas de 2021. 57

6.3. Cantidad de muestras para cada *Fold* de entrenamiento y validación, junto al porcentaje del total que representan las muestras de validación. 57

7.1. Tabla con cantidad original de variables por fuente, la cantidad luego de aplicar *PCA* al 95 %, mostrada también como porcentaje de la cantidad original. Además se muestra la cantidad de valores únicos para cada fuente en la columna r. 67

8.1. Promedio entre *folds* de distintas métricas, para cada variante del modelo. Las tres primeras filas corresponden a las métricas de comparación, las cuales también están promediadas a nivel de *folds*. Se marca en negrita el mejor resultado de cada métrica. Ver anexo para consultar la definición de cada métrica. 77

8.2. Promedio entre *folds* de distintas métricas, para cada variante del modelo. Las tres primeras filas corresponden a las métricas de comparación, las cuales también están promediadas a nivel de *folds*. Se marca en negrita el mejor resultado de cada métrica. Ver anexo para consultar la definición de cada métrica. 77

8.3. Promedio entre *folds* de distintas métricas, para cada variante del modelo. Las tres primeras filas corresponden a las métricas de comparación, las cuales también están promediadas a nivel de *folds*. Se marca en negrita el mejor resultado de cada métrica. Ver B.1 para consultar la definición de las métricas adicionales. 80
8.4. Promedio entre *folds* de distintas métricas, para cada variante del modelo. Las tres primeras filas corresponden a las métricas de comparación, las cuales también estan promediadas a nivel de *folds*. Se marca en negrita el mejor resultado de cada métrica. Ver B.1 para consultar la definición de las métricas adicionales. 82
8.5. Resultados del modelo para cada conjunto de datos. Además se agregan las métricas de comparación de E_L, que resultó ser la mejor predicción trivial. 83
8.6. Promedio entre *folds* de ambas métricas, para cada variante del modelo. Las tres primeras filas corresponden a las métricas de comparación, las cuales también están promediadas a nivel de *folds*. Se marca en negrita el mejor resultado de cada métrica. 86
8.7. Promedio entre *folds* de ambas métricas, para cada variante del modelo. Las tres primeras filas corresponden a las métricas de comparación, las cuales también están promediadas a nivel de *folds*. Se marca en negrita el mejor resultado de cada métrica. 88
8.8. Promedio entre *folds* de ambas métricas, para la mejor variante de cada modelo y en cada esquema de validación. Se marca en negrita, para cada esquema, el mejor resultado de cada métrica. 90

9.1. Top diez de variables más importantes para los modelos de cada esquema. Para *HD* se muestra además, en la última columna, los diez marcadores más importantes. 96

B.1. Hiperparámetros óptimos de cada variante, utilizando el esquema *HD*. . . 108
B.2. Hiperparámetros óptimos de cada variante, utilizando el esquema *AD*. . . 109
B.3. Hiperparámetros óptimos de cada variante, utilizando el esquema *HD*. . . 109
B.4. Hiperparámetros óptimos de cada variante, utilizando el esquema *AD*. . . 109

C.1. Promedio entre *folds* de *RMSE* y correlación, para cada modelo. La primera fila corresponde al error de comparación E_G, el cual también está promediado a nivel de *fold*. Se marca en negrita el mejor resultado de cada métrica. 114

Índice de figuras

4figure.caption.5

2.2. Esquema de los componentes del problema. Se distingue como se genotipan los *híbridos* y se registran variables de los *ambientes*. Luego para cada experimento se juntan ambas componentes y se usan para predecir el fenotipo. 8

3.1. Cuarteto de *Anscombe* [44]. En la figura se muestran cuatro conjuntos de datos con estadísticos muy similares hasta segundo orden (medias y varianzas), pero que en su estructura son muy diferentes. En particular, los cuatro conjuntos de datos tienen $r \approx 0{,}816$, sin embargo el modelo lineal basado en media y correlación solo es apropiado para uno de los cuatro conjuntos de datos. 12

3.2. Ejemplo de árbol de decisión para decidir si jugar al tenis. Fuente: [35] . . 14

3.3. Diagrama de una neurona con un vector de dimensión 3 y escalón unitario como función de activación. Fuente: [14]. 17

3.4. Ejemplo de *Max Pooling* en una imagen sobre áreas de 2×2. Fuente: [14]. 19

3.5. Visualización de la técnica de *Batch Normalization*. Si se considera el *batch* de datos como un arreglo con dimensiones `tamaño_de_batch`×`alto`× `ancho`×`canales`, la normalización se hace en las dimensiones `tamaño_de_batch`, `alto` y `ancho`. Es decir, se normaliza cada canal del *batch* por separado. . 19

3.6. Visualización de la técnica de *Dropout* [4] 20

3.7. Visualización del fenómeno de subajuste (*underfitting*) y sobreajuste (overfitting). Izquierda: el modelo no logra ajustarse a los datos de entrenamiento. Centro: el modelo se ajusta adecuadamente a los datos de entrenamiento, manteniendo un balance entre desempeño y generalización. Derecha: el modelo se ajusta al ruido presente en los datos de entrenamiento, perjudicando su capacidad de generalizar a nuevos datos. Fuente: [18] . . 22

5.1. Locaciones ubicadas en el mapa según las coordenadas de las plantaciones del último año en que aparece la locación. Coloreado según desempeño promedio de la locación y dimensionado según cantidad de registros disponibles, en el último año en que aparece. 32

5.2. Locaciones ubicadas en el mapa (coordenadas de las plantaciones del último año en que aparece la locación). Coloreado según altitud y dimensionado según la extensión de las plantaciones, en el último año en que aparece. 33
5.3. Histograma de correlaciones de los *SNPs* con el fenotipo a predecir. . . . 36
5.4. Gráfico de *Pareto*. Correlación total acumulada según el porcentaje de *SNPs* más correlacionados con fenotipo. 37
5.5. Histograma de la proporción del valor mayoritario en los *SNPs*. Cantidad de veces que aparece el valor más frecuente, entre la cantidad total de valores, o sea entre la cantidad de individuos genotipados. 38
5.6. Histograma de porcentajes de valores nulos de cada *SNP*. 39
5.7. Gráfico de *Pareto*. Muestra qué porcentaje del total de datos faltantes totales, alcanzamos con un determinado porcentaje de los *SNPs* con más datos faltantes. 39
5.8. En el eje derecho y con linea negra, vemos el porcentaje de *SNPs* que podemos usar en base al umbral máximo de valores nulos que permitimos. En el eje izquierdo y con barras celestes está el histograma de porcentajes de valores nulos de cada *SNPs*, pero esta vez en escala logarítmica. 40
5.9. A la izquierda, histograma de las correlaciones máximas para cada *SNP*, es decir que el conteo se realiza sobre el vector resultante de obtener el máximo de cada columna de la matriz de correlaciones, sin contar los unos de la diagonal, ni los *SNPs* repetidos. A la derecha, histograma de las entradas de la matriz de correlación entre *SNPs* (nuevamente sin contar los unos de la diagonal). 41
5.10. Histograma de la ocurrencia de experimentos. Con el eje vertical en escala logarítmica. 42
5.11. Histograma de las desviaciones relativas de cada uno de los conjuntos de experimentos. Sin tener en cuenta los experimentos únicos. 43
5.12. A la izquierda, proporción de datos faltantes (escala de color) por locación y por fuente; a la derecha lo mismo pero ordenando los datos por índice y marcando el año en el eje vertical. En ambas, se muestra bajo el nombre de cada fuente, el porcentaje de datos faltantes de estas, considerando la totalidad de la base de datos. 44
5.13. Figura análoga a 5.12 pero segmentando por año. 45
5.14. Cantidad de registros disponibles por año. 45
5.15. Cantidad de registros por locación. 46
5.16. Matriz de aparición de las locaciones. Se marca en blanco si una locación se usó en un determinado año, y en negro si no. 46
5.17. A la izquierda, tabla cruzada que muestra la cantidad de locaciones en común entre dos años. En la diagonal se muestra la cantidad de locaciones utilizadas en un determinado año. A la derecha tabla cruzada que muestra la proporción de locaciones de un año (eje horizontal) que aparecen en otro (eje vertical). A diferencia de la primera, esta no es necesariamente simétrica. 47
5.18. Mismas tablas que en 5.17 pero analizando híbridos en lugar de locaciones. 47

5.19. Histograma de los tonelajes de cada experimento, separado por generación. Se marca con una linea punteada la media en cada caso. 48
5.20. Representación en dos dimensiones de los híbridos (mediante sus datos genómicos), donde cada punto representa el genoma de un híbrido, coloreados según distintos criterios; izquierda, desempeño promedio del híbrido (fenotipo a predecir); derecha, cantidad de muestras disponibles de ese híbrido. 49
5.21. Representación en dos dimensiones de los híbridos (mediante sus datos genómicos), utilizando *PCA* (izquierda) y *TSNE* (derecha), coloreando según generación. 49
5.22. Representación en dos dimensiones de los datos genómicos, donde cada punto representa el genoma de un híbrido, coloreados según distintos criterios; izquierda, desempeño promedio del ambiente; derecha, locación en que se ubicó el ambiente. Considerando solo las diez locaciones con más muestras (las cuales acumulan la mitad de las muestras). 50

6.1. Representación matricial de los distintos esquemas de validación. Cada fila representa un ambiente y cada columna un híbrido. Las celdas azules representan las combinaciones que son usadas para el entrenamiento, las naranjas las que son usadas para la validación y las blancas las que no son usadas. 59

7.1. Canalización de los datos para el *Gradient Boosting*. 64
7.2. Matriz de correlación para cada fuente de datos. Las correlaciones se calculan a nivel ambiente para las fuentes ambientales, y a nivel híbrido para los datos genómicos. En `WEATHER`, las variables (descriptores estadísticos) aparecen agrupadas en base a la serie temporal de la que provienen, de ahí la estructura en bloques de la matriz. 66
7.3. Porcentaje de la varianza acumulada (medida de que tanta información estamos preservando) en función de la cantidad de componentes principales utilizadas, expresada como porcentaje de la cantidad de variables originales. Con lineas punteadas se marca el porcentaje en el que se alcanza el 95 % de la varianza acumulada. 68
7.4. Arquitectura de red neuronal multimodal. 70

8.1. *Boxplots* del *RMSE* para cada *fold*, agrupados por variantes del modelo para el esquema de validación *HD*. En líneas punteadas se marcan los errores de comparación. 74
8.2. *Boxplots* de la correlación de *Pearson* para cada *Fold*, agrupados por variantes del modelo para el esquema de validación *HD*. En líneas punteadas se marcan las correlaciones de comparación. 75
8.3. *Boxplots* del *RMSE* para cada *fold*, agrupados por variantes del modelo para el esquema de validación *AD*. En líneas punteadas se marcan los errores de comparación. 76

8.4. *Boxplots* de la correlación de *Pearson* para cada *Fold*, agrupados por variantes del modelo para el esquema de validación *AD*. En líneas punteadas se marcan las correlaciones de comparación. 76
8.5. *Boxplots* del *RMSE* para cada *fold*, agrupados por variantes del modelo. Notar que como son cinco *folds*, cada linea del *boxplot* representa el error en un *fold*. En lineas punteadas se marcan los errores de comparación. . . 78
8.6. *Boxplots* de la correlación de *Pearson* para cada *Fold*, agrupados por variantes del modelo. En lineas punteadas se marcan las correlaciones de comparación. 79
8.7. *Boxplots* del RMSE para cada *fold*, agrupados por variantes del modelo. Notar que nuevamente como son cinco *folds*, cada linea del *boxplot* representa el error en un *fold*. En lineas punteadas se marcan los errores de comparación. 81
8.8. *Boxplots* de la correlación de *Pearson* para cada *Fold*, agrupados por variantes del modelo. En lineas punteadas se marcan las correlaciones de comparación. 81
8.9. *RMSE* (función de perdida) en cada conjunto, en función de la época. Se muestra además con una linea punteada la época en que se obtuvo el menor error en validación. 83
8.10. *Boxplots* de *RMSE* para cada *fold*, agrupados por variantes del modelo. En lineas punteadas se marcan los errores de comparación. 85
8.11. *Boxplots* de *CORR* para cada *fold*, agrupados por variantes del modelo. En lineas punteadas se marcan los errores de comparación. 85
8.12. *Boxplots* de *RMSE* para cada *fold*, agrupados por variantes del modelo. En lineas punteadas se marcan los errores de comparación. 87
8.13. *Boxplots* de *CORR* para cada *fold*, agrupados por variantes del modelo. En lineas punteadas se marcan los errores de comparación. 87
8.14. *Boxplots* de *CORR* y *RMSE* para para cada *fold*, de la mejor variante de cada modelo, utilizando el esquema de validación *HD*. En lineas punteadas se marcan los errores y correlaciones de comparación. 89
8.15. Análogo a la figura previa pero utilizando el esquema de validación *AD*. . 89

9.1. Veinte componentes con mayor importancia, para los modelos de cada esquema. Las componentes se nombran primero con la fuente de la que provienen, y luego con un número que indica el orden de la componente, siendo la número cero la que mayor varianza captura. 92
9.2. En la primera fila, importancia acumulada por fuente de información. En la segunda, importancia promedio de cada componentes. En la tercera, importancia promedio de cada variable original. En la primera columna para *HD* y en la segunda para *AD*. En la primera y segunda fila se marca la cantidad de componentes por fuente, y en la tercera la cantidad de variables. 94
9.3. Gráfico de violín para las importancias de cada variable, segmentado por fuente. Arriba para *HD* y abajo para *AD*. 96

B.1. Diagramas para los *Encoder*. 110

B.2. Diagrama del bloque completamente conectado. 111
B.3. Diagramas de los módulos basados en redes convolucionales. 112

C.1. *RMSE* y correlación a nivel *fold* para cada modelo. Además se muestra el valor promedio de los E_G con una linea punteada. 114

Esta es la última página.
Compilado el sábado 28 diciembre, 2024.
http://iie.fing.edu.uy/