UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN

PREDICCIÓN DE SEVERIDAD COVID-19 UTILIZANDO MODELOS DE MACHINE LEARNING APLICADO A SECUENCIAS EN DATOS GENÉTICOS A NIVEL DE GENOMA

TESIS PARA OPTAR AL GRADO DE MAGÍSTER EN CIENCIAS MENCIÓN COMPUTACIÓN

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL EN COMPUTACIÓN

JUAN MANUEL SAEZ HIDALGO

PROFESOR GUÍA:
RICARDO VERDUGO SALGADO
PROFESOR CO-GUÍA:
JORGE PÉREZ ROJAS

MIEMBROS DE LA COMISIÓN:
JOCELYN DUNSTAN ESCUDERO
ANDRÉS ABELIUK KIMELMAN
MARCELO MENDOZA ROCHA

Este trabajo ha sido parcialmente financiado por:
ANID mediante proyects FONDECYT Regular 1191948 y Anillo ACT210085

SANTIAGO DE CHILE
2022

RESUMEN DE LA TESIS PARA OPTAR AL GRADO DE
MAGÍSTER EN CIENCIAS MENCIÓN COMPUTACIÓN
MEMORIA PARA OPTAR A AL TÍTULO
DE INGENIERO CIVIL EN COMPUTACIÓN
POR: JUAN MANUEL SAEZ HIDALGO
FECHA: 2022
PROF. GUÍA: RICARDO VERDUGO SALGADO
PROF. CO-GUÍA: JORGE PÉREZ ROJAS

PREDICCIÓN DE SEVERIDAD COVID-19 UTILIZANDO MODELOS DE MACHINE LEARNING APLICADO A SECUENCIAS EN DATOS GENÉTICOS A NIVEL DE GENOMA

Como parte de la iniciativa internacional Covid19 *host genetics*, el proyecto COVID0961 recopila datos clínicos autoreportados y variantes genéticas a nivel de genoma de participantes infectados por el virus SARS-CoV-2. En el contexto de este proyecto, esta tesis busca utilizar dichos datos para generar modelos de *Machine Learning* (ML) que permitan predecir la severidad del cuadro Covid19, particularmente la hospitalización provocada por el virus.

En este trabajo, se utilizan los datos clínicos obtenidos para entrenar un modelo ML. Este modelo sirve como punto de partida, para comparar si al agregar variantes genéticas a los datos de entrenamiento, las predicciones de severidad mejoran. La implementación del modelo que utiliza tanto datos clínicos como genéticos se realizó mediante dos aproximaciones: (1) una arquitectura de redes neuronales diseñada para este propósito y (2) la selección de variantes genéticas que se agregan al *set* de entrenamiento del modelo *baseline* ML.

Mediante el entrenamiento con datos clínicos de 1872 participantes, se obtiene un modelo XGBoost capaz de predecir la hospitalización con un *accuracy* de 88 % y $f1$-*score* de un 60 %. Utilizando este modelo como referencia se busca obtener un modelo de procesamiento de secuencia que obtenga mejores métricas utilizando tanto datos clínicos como variantes genéticas.

En la primera aproximación, debido a las limitaciones de recursos computacionales, se seleccionaron las variantes según su significancia estadística. Usando estas variantes y los datos clínicos disponibles, el modelo utilizado Dual-stream CNN, no alcanza métricas mayores a la referencia que utiliza solo datos clínicos, alcanzando un $f1$-*score* de 56.9 % y un *accuracy* de 87.6 %. Sin embargo, el perfil de la saliencia de las variantes en la red entrenada se asemeja al perfil obtenido por la técnica actual de genética de poblaciones. Incluso, señala como punto de mayor saliencia una variante genética dentro del gen FOXP4-AS1 no reportado en estudios anteriores.

En la segunda aproximación, se agregan progresivamente variantes genéticas al entrenamiento del modelo XGBoost. Esta selección de variantes se realiza de forma de agregar variantes reportadas como significativas y descartar variantes ligadas entre sí. En este experimento, se evidencia que el $f1$-*score* sí aumenta al agregar entre 100 y 200 variantes no ligadas, que es superior al número de loci reportados. Esto último sugiere que la genética del hospedero sí entrega información relevante para la severidad.

Ambos resultados sugieren que las variantes genéticas entregan información, aunque un modelo capaz de filtrar aquellas que agregan ruido requieren una mayor cantidad de ejemplos al entrenamiento que permitan un modelo generalizable.

Agradezco a mis padres, por su apoyo emocional y económico durante estos años de formación. Por comprender mi cambio de carrera y mantener su confianza en mí. A mi hermano por sus comentarios y aportes durante la carrera y otros proyectos. Y a mí gato, Supes[1], por su compañia y sostén de mi salud mental.

Agradezco a mi mejor amiga, Janahina Bravo, por su contención en los momentos más frustrantes de mi carrera en general y de este trabajo en particular. Por muchas veces recargarme de energías para continuar, sin las cuales siento que no hubiese podido terminar este proyecto. Agradecerle además, a ella y a su rubro, la atención de salud primaria, por todo el trabajo y esfuerzo sobrehumano que vivieron durante estos años de pandemia.

Agradezco a mis compañeras de laboratorio, Stefenia Rosso, Laura Carvajal y Francisca Carvajal, por orientarme en técnicas y conocimientos que no eran de mi área, y por asentir con la cabeza cuando les hablaba de redes neuronales. A Cristian Yañez por su guía en bioinformática y la confección de los GWAS del proyecto. A mis compañeres de carrera y de trabajo de tesis, por sus comentarios y apoyo. A Cristopher Alfaro, Lucas Torrealba, Alonso Reyes, Javiera Bermúdez, Pablo Torres y Sebastián Alfaro.

Le agradezco a mis profesores guía. Al profesor Ricardo Verdugo por todo el conocimiento en genética de esta tesis, por su continua superversión y por jamás dejar de precuparse por sus estudiantes. Sin su estricta e ininterrumpida vigilancia este trabajo no sería posible. A los consejos y orientación del profesor Jorge Pérez, y por sus advertencias en el desarrollo de modelos de inteligencia, a las que de todos modo choqué de frente durante la realización de este trabajo.

Agradecerle a les profesores que influyeron y apoyaron mis proyectos más allá de esta tesis. A les profesores Jorge Mpodozis, Álvaro Olivera Nappa, Francisco Forster, Javier Bustos, Maite Gonzales, Diego Madariaga y Sandra Céspedes. La profesora Karen Oróstica por introducirme a la bioinfórmatica. El profesor Benjamín Bustos por sus conocimientos en visualización de información y por guiarme en la docencia universitaria. A él y al profesor Jérémy Barbay que me mostraron que sí se puede privilegiar la docencia más allá de la evaluación, porque la verdadera misión de un profesor es compartir su conocimiento.

Agradezco también a Sebastián Sepúlveda, del Laboratorio de Escritura Armadillo, por hacer de este escrito legible. A los miembros de mi comisión por sus sugerencias y correciones. A la profesora Nancy Hitschfeld por sus comentarios y retroalimentación como profesora acargo de Taller de Tesis. Y a Sandra Gaez por la información y gestiones administrativas dentro del programa de Magíster.

[1] Se pronuncia “Sups” y es la abreviatura de Superman

Tabla de Contenido

Índice de Tablas

Índice de Ilustraciones

Capítulo 1

Introducción

En los dos últimos años hemos visto nuestra vida afectada por la pandemia viral Covid-19, una crisis sanitaria global que no había ocurrido desde La Pandemia de Influenza de 1918. La cantidad de contagiados al 1 de marzo de 2022 supera los 437 millones a nivel mundial y casi 6 millones de fallecidos [17], de los cuales más de 42 mil son en Chile, dando la cifra de 2176 muertes por millón de habitantes. El desarrollo de varias vacunas ha permitido reducir el número de contagios, la hospitalización y la cantidad de muertes [25]. Sin embargo, aun existe el riesgo de nuevas variantes [6] y la posibilidad de una nueva pandemia zoonótica [39, 50]. En ello radica la importancia de seguir investigando y producir nuevas aproximaciones a la investigación de los mecanismos biológicos de este tipo de enfermedades.

Con el fin de estudiar cómo la genética del hospedero determina la susceptibilidad y severidad de la enfermedad, la iniciativa internacional *COVID-19 host genetics* [12] reúne 46 estudios de 19 países, para compartir y analizar los datos y resultados obtenidos. Dentro de esta iniciativa, se encuentra el proyecto ANID[1] COVID0961 encabezado por Ricardo Verdugo y Andrea Silva, que recopila muestras genéticas y datos clínicos autoreportados de participantes chilenos en cada una de las 5 macrozonas del país [48]. Además de poder determinar las posibles poblaciones de riesgo, este estudio busca contribuir a la iniciativa internacional y al conocimiento de los mecanismos genéticos de la enfermedad.

Los datos recopilados por el proyecto, corresponden a datos clínicos autoreportados y muestras de saliva o sangre para obtener datos genéticos a nivel de genoma. Los datos clíncos autoreportados se consiguen por encuesta o, mediante el acceso, con consentimiento informado, de la ficha clínica. Éstos corresponden a datos demográficos, étnicos, socieconómicos, de severidad del cuadro presentado y antecedentes de salud, como comorbilidades y tratamientos actuales. Los datos genéticos corresponden a las variaciones entre participantes en todo el genoma del individuo.

Como el genotipo de un individuo determina el fenotipo, o características observables, es un proceso complejo que implica varios mecanismos moleculares [2]. Una aproximación actual es determinar la significancia estadística de variantes genéticas en una población, lo

[1] Agencia Nacional de Investigación y Desarrollo

que se conoce como GWAS [55]. Alternativamente, utilizar modelos aplicados a secuencia como XGBoost [8], redes convolucionales (CNN) [33, 34, 38, 68], redes recurrentes (RNN) [23, 26, 66], o, en el último tiempo, transformers [10, 30, 65], buscan determinar estas variantes incluyendo nucleótidos específicos o interacciones entre variantes. Esto último se hace complejo en el caso del genoma humano debido al tamaño de éste [20] y a la baja variación entre individuos [56] repartida a lo largo de toda la secuencia.

La contribución de este trabajo radica en la aplicación de estos modelos en la predicción de la severidad de Covid-19 utilizando participantes chilenos. Con lo que se obtienen modelos de *Machine Leaning* (ML) que predicen la hospitalización utilizando datos clínicos. También se pretende determinar si las variantes genéticas se correlacionan con la severidad, lo que se realiza agregando variantes seleccionadas por otro método a un modelo que utiliza variables clínicas y se evidencia cómo mejoran las métricas de predicción. Por último, encontrar genes específicos, estadísticamente significativos, a través de la interpretación de modelos que procesan secuencias. Lo que no se logra de forma directa, si no que mediante la reducción de variantes genéticas por un método externo.

En el Capítulo 2 se presentan conceptos en genética y bioinformática que son utilizados en esta investigación. Además de los modelos a usar, cómo fueron usados en trabajos relacionados y cómo aquellos que se suelen utilizar sobre secuencias son los adecuados para este tipo de datos. Al final de este capítulo se presentan también algunas aproximaciones que permiten interpretar estos modelos.

En el Capítulo 3 se describe la metodología utilizada en este trabajo. El Capítulo 4 describe los principales resultados obtenidos, lo que incluye la selección de variables clínicas, la predicción de hospitalización mediante datos clínicos autoreportados y la contribución de variantes genéticas a la predicción. Finalmente, las principales discusiones se muestran en el Capítulo 5.

Si bien las arquitecturas utilizadas en este trabajo no obtuvieron mejores métricas al agregar variantes genéticas, la interpretación de estos modelos sí reportan algunos resultados interesantes para la investigaciones futuras. Por otro lado, utilizando modelos más simples de ML, se pudo mejorar las métricas al agregar variantes genéticas seleccionadas, lo que sugiere que estás si se correlacionan con la severidad de la enfermedad. Ambos resultados indican que, si bien existe información relevante en las variaciones genéticas, poder filtrar éstas según un fenotipo requiere más datos de entrenamiento o una selección previa de estas variantes.

1.1. Hipótesis

La siguiente es la hipótesis principal del trabajo:

> Un modelo de aprendizaje automático permite predecir la severidad de un cuadro de Covid-19 utilizando antecedentes clínicos y datos genéticos a nivel de genoma.

1.2. Objetivos

1.2.1. Objetivo General

Implementar un modelo de inteligencia artificial que permita predecir la severidad en pacientes chilenos de Covid-19, utilizando secuencias genómicas y antecedentes clínicos.

1.2.2. Objetivos Específicos

1. Definir un modelo que permita predecir la severidad de un cuadro de Covid-19, utilizando datos clínicos autoreportados.
2. Implementar un modelo de Aprendizaje de Máquinas (ML por sus siglas en inglés *Machine Learning*), que se entrene con datos clínicos y genómicos.
3. Comparar la certeza de la predicción del modelo que utiliza datos clínicos y variantes genéticas, con modelos que solo utilizan datos clínicos.
4. Empaquetar un software que realice visualizaciones, en base al clasificador, que permitan a un experto en el área interpretar posibles zonas del input que sean de mayor importancia en la investigación.

Capítulo 2

Marco Teórico

Debido a que este trabajo incluye datos, métodos y conceptos de genética, en este capítulo se expone una breve conceptualización. Se presenta el funcionamiento del genoma en un organismo vivo en la Sección 2.1.1, realizando una comparación con el texto en lenguaje natural. En la Sección 2.1.2 se presenta como se trabajan estos datos desde el punto de vista de la bioinformática y computación. Como este trabajo se orienta a la ciencias de la computación se incluye la terminología utilizada en genetica en el Anexo A. En la Sección 2.2 se presenta el estado del arte en el área de los algoritmos de aprendizaje sobre datos genéticos, para contextualizar el aporte de este trabajo.

2.1. Marco Conceptual

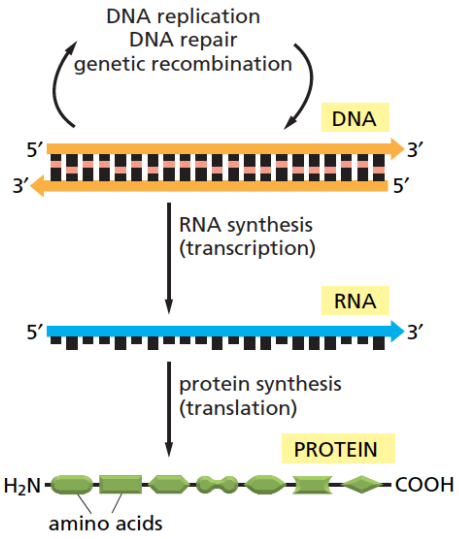


Figura 2.1: Vía en que la información codificada en el DNA que es utilizada para sintetizar proteínas. Fuente: [2]

Esta sección está orientada a la presentación de los conceptos usados en este trabajo. Estos conceptos son principalmente del área de la genética (Secciones 2.1.1-2.1.2) y del aprendiza automático o *Machine Learning* (Secciones 2.1.3-2.1.4).

2.1.1. Genes como texto: Funcionamiento del genoma

El fenotipo de un organismo se define como sus características observables, lo que incluye, características morfológicas, funcionales (como el metabolismo) y conductuales. El fenotipo se determina tanto por sus genes como por la influencia del ambiente [2]. De forma simplificada, el funcionamiento de los genes es una ruta de información que va desde el DNA, como genoma completo, a una molécula de RNA, que contiene la información necesaria, para

finalmente, sintetizar una proteína, que es la molécula que realiza las funciones a nivel celular [2] (**Figura 2.1**).

En otras palabras, las propiedades fisicoquímicas de una proteína vienen determinadas por la secuencia de los aminoácidos que la componen [2]. La secuencia de aminoácidos viene codificada dentro de la molécula de RNA. Ésta, a su vez, se encuentra en la secuencia codificante dentro del DNA en el genoma.

El genoma de un organismo se compone físicamente de una macromolécula de DNA (*Desoxiribonucleic acid*, ácido desoxiribonucléico). Esta molécula se compone de dos cadenas conformadas por nucleótidos, caracterizados por su base nitrogenada que puede corresponder a cuatro posibles moléculas: Adenina (A), Guanina (G), Citocina (C) y Timina (T). El conjunto de 3 nucleótidos, conocido como codón, se corresponde con el aminoácido de la proteína resultante. En el contexto celular, la proteína que se sintetiza, también se regula por otras secuencias dentro del genoma. Estos promotores, activadores y represores pueden estar en cualquier posición del genoma (**Figura 2.2**).

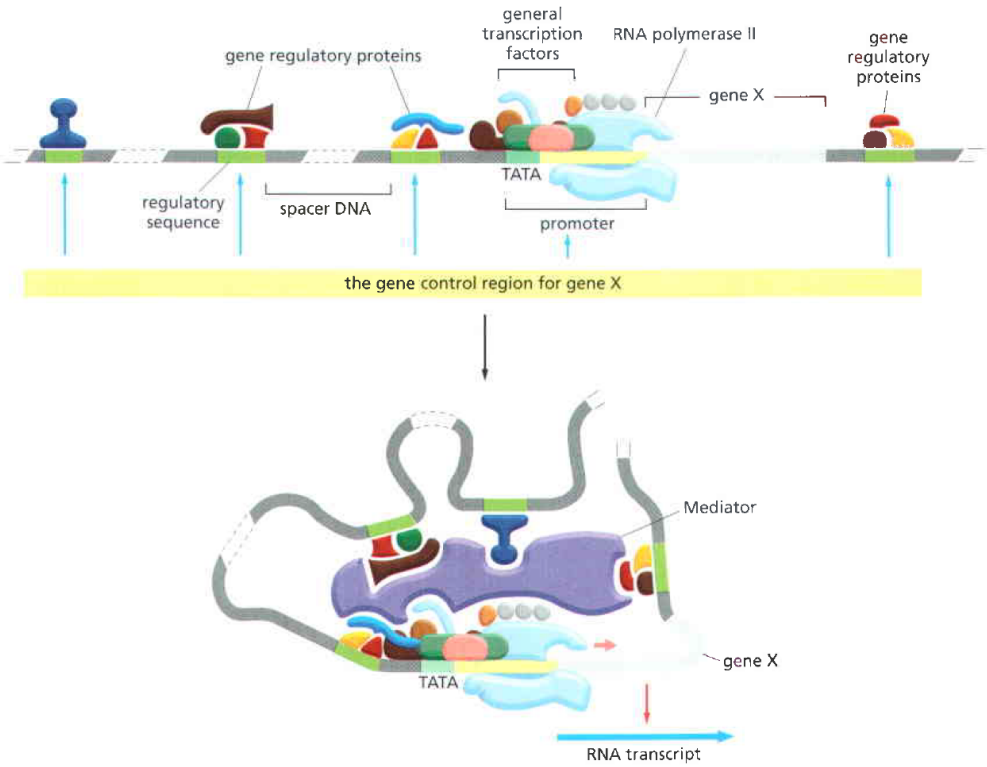


Figura 2.2: Ejemplo de regulación de genes mediante diferentes secuencias dentro del genoma. En la parte superior se muestran las secuencias y proteínas como si el genoma estuviese de forma lineal, en la parte inferior, es un esquema de como estarían distribuidos dentro del núcleo celular. Fuente: [2]

Si bien, el genoma se organiza en subunidades conocidas como cromosomas, y por lo general, se representa como secuencias lineales, en la realidad se encuentran en diferentes parte del núcleo celular interactuando entre sí. El genoma humano, se compone de 46 de estos cromosomas, los que se representan en la **Figura 2.3**, donde la imagen en **A** es una

representación de los cromosomas condensados y **B** es como se disponen dentro del núcleo. Así como en un libro, que un capítulo puede determinar el significado de una oración en otro capítulo, la información de un cromosoma puede depender de otro cromosoma para su regulación.

Otro concepto importante dentro de la genética es el de cromosomas homólogos. Los cromosomas son estructuras formadas por DNA y proteínas que aparecen cuando la célula se divide. Durante la fecundación, se unen el material genético de dos organismos para formar la información del nuevo organismo. Cada *set* de cromosomas equivalente y cada par de ellos se conoce como cromosoma homólogo. En el caso de los humanos, un *set* proviene del padre y otro de la madre. Como cada par de cromosoma es equivalente, las diferentes posiciones dentro de un cromosoma se conocen como loci[2], por lo que cada individuo posee dos versiones para cada locus. Las diferentes versiones de información, ya sea de uno o más nucleótidos, se conoce como alelo. Cada individuo, entonces, posee dos alelos para cada locus en el genoma [2].

En resumen, el genoma se compone de un abecedario de cuatro letras $\{A, C, T, G\}$ que contiene información dependiente de una sintaxis jerárquica y ordenada, dependiente de contexto, como el lenguaje natural. Esto justifica el estudio de la biología, o bioinformática, desde una perspectiva lingüística [47]. Otro ejemplo, es que en 1984, Volker Brendel y Heinrich-Gustav Busse realizaron un análisis de secuencias de nucleótidos desde el punto de vista del lenguaje formal, describiendo Autómatas Finitos de varias secuencias en el genoma [5].

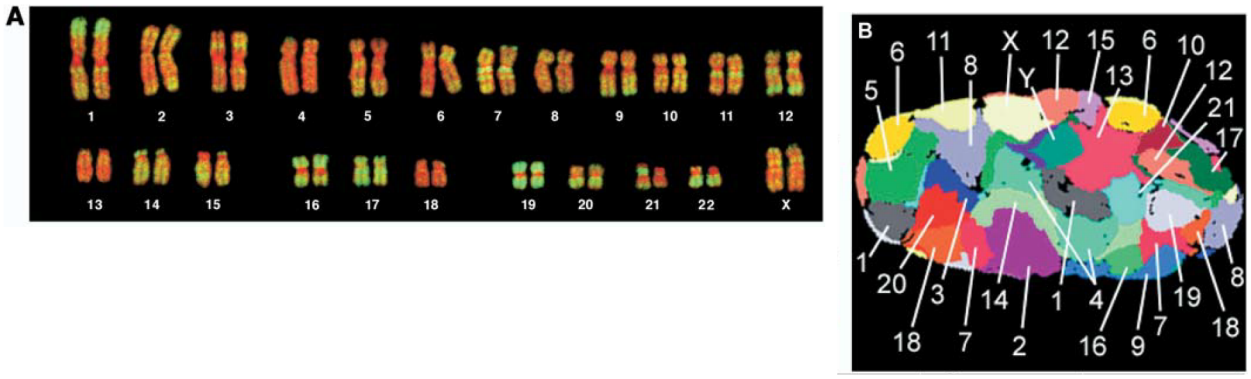


Figura 2.3: Representación del genoma humano como **A** 46 cromosomas y **B** como se encuentra distribuido dentro del núcleo de una célula. Este estudio se realizó en un fibroblasto humano. Fuente: [4]

2.1.2. Trabajando con variantes genéticas

Como se describe en la Sección 2.1.1, el genoma, desde la perspectiva de la información, se compone por una secuencia de nucleótidos (A, G, C, T). Esta secuencia está compuesta por subunidades llamadas cromosomas, que representaremos como una secuencia X_j:

[2] Plural de locus.

$$X_j = b_{1,j}b_{2,j}\cdots b_{m_j,j} = \{b_{i,j}\}, \text{ con } i \in \{1, 2, \cdots, m_j\}$$

Con m_j la cantidad de nucleótido que componen cada cromosoma. Particularmente, el genoma humano se compone de 46 cromosomas, dos conocidos como cromosomas sexuales, y otros 44 llamados **autosómicos**. Los cromosomas se heredan del padre y de la madre, esto pares se conocen como **cromosomas homólogos**, y se enumeran del 1 al 22. El largo de cada par de cromosomas homólogos es conocido y lo denominaremos m_j con $j \in \{1, 2, \cdots, 22\}$. El genoma de un participante x se puede escribir, entonces:

$$\text{Genoma}_x = X_{1_M}X_{1_P}X_{2_M}X_{2_P}\cdots X_{22_M}X_{22_P}X_{\mathcal{X}}X_{\mathcal{Z}} \tag{2.1}$$

Con $\mathcal{X}$ el cromosoma X y $\mathcal{Z}$ puede ser el cromosoma X, en caso de mujeres y cromosoma Y, en caso de hombres. Esta representación por cromosoma se describe para ser usada como identificador, pero no será utilizada directamente en los modelos.

2.1.2.1. Secuencia completa

Una representación más simple del genoma completo, es considerar que las 46 secuencias se encuentran en orden, es decir, la concatenación de todos los cromosomas de la **Ecuación 2.1**, teniendo entonces:

$$\text{Genoma}_x = b_1b_2\cdots b_m, \text{ con } b_i \in \{A, G, C, T\} \tag{2.2}$$

En este caso m corresponden a todos los nucleótidos que se tienen acceso. En caso de tener la secuencia completa m correspondería a:

$$m = 2\cdot m_1 + 2\cdot m_2 + \cdots + 2\cdot m_{22} + m_{\mathcal{X}} + m_{\mathcal{Z}} = \sum_{j\in\text{cromosomas}} m_j$$

En el caso de un humano es aproximadamente $3.2\cdot 10^9$ [20]. Como el genoma completo entre personas se diferencia en tan solo un $0.11\,\%$ [56] y obtener la secuencia completa es difícil en la práctica, por lo general m es mucho menor al calculo mostrado. De esa forma, m es la cantidad de nucleótidos a las que sí se tiene acceso. Para poder replicar los experimentos o utilizar el modelo por otros estudios, los nucleótidos usados tienen identificadores y no son solo una posición relativa.

Los datos disponibles bajo esta representación (**Ecuación 2.2**) se ven como una matriz de la forma:

$$\text{Data} = \begin{pmatrix} \text{Genoma}_1 \\ \text{Genoma}_2 \\ \vdots \\ \text{Genoma}_n \end{pmatrix} = \begin{pmatrix} b_{1,1} & b_{1,2} & \cdots & b_{1,m} \\ b_{2,1} & b_{2,2} & \cdots & b_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ b_{n,1} & b_{n,2} & \cdots & b_{n,m} \end{pmatrix} \tag{2.3}$$

Llamaremos a esta representación como **Secuencia Completa** o **SNP concatenados** (**Sección 2.1.2.2**).

2.1.2.2. Polimorfismo de Nucleódo Único, SNP

Como se mencionó, en la práctica, no se tiene el genoma completo de un individuo. En su lugar, se accede a las diferencia o variaciones genéticas entre individuos, para luego imputar el resto del genoma mediante herramientas estadísticas. La diferencia de un solo nucleótido en el genoma entre un individuo y otro se conoce como SNP (Polimorfismo de Nucleótido único, por sus siglas en inglés *Single Nucleotic Polymorphism*).

Para este trabajo, solo usaremos los **SNP dialélicos**, es decir, que puede tomar solo dos valores entre A, G, C, T. Sea la representación según secuencia completa (**Ecuación 2.2**) de un individuo de referencia (representado con el subíndice REF):

$$\text{Genoma}_{REF} = b_{1_M} b_{1_P} b_{2_M} b_{2_P} \cdots b_{m_M} b_{m_P},\ b_i \in \{A, G, C, T\}$$

Recordando que existe una posición en los cromosomas heredados del padre y otra de la madre. En la práctica esto no se puede saber, así que lo llamaremos 0 y 1. Quedando entonces:

$$\text{Genoma}_{REF} = b_{1,0} b_{1,1} b_{2,0} b_{2,1} \cdots b_{m,0} b_{m,1},\ b_{i,k} \in \{A, G, C, T\},\ k \in \{0, 1\}$$

Si tomamos un individuo (SAMPLE), su representación del genoma a través de SNP será:

$$SNP_{\text{Sample},i} = \begin{cases} 0 & b_{\text{Sample},i,0} = b_{\text{REF},i,0} \wedge b_{\text{Sample},i,1} = b_{\text{REF},i,1} \\ 1 & b_{\text{Sample},i,0} = b_{\text{REF},i,0} \wedge b_{\text{Sample},i,1} \neq b_{\text{REF},i,1} \\ 2 & b_{\text{Sample},i,0} \neq b_{\text{REF},i,0} \wedge b_{\text{Sample},i,1} \neq b_{\text{REF},i,1} \end{cases} \tag{2.4}$$

Se aprecia que esto reduce la dimensión k, que corresponde al cromosoma particular dentro del par homólogo. Nuestros datos se puede representar mediante SNP como sigue:

$$\text{Data} = \begin{pmatrix} SNP_{1,1} & \cdots & SNP_{1,m} \\ \vdots & \ddots & \vdots \\ SNP_{x,1} & \cdots & SNP_{x,m} \\ \vdots & \ddots & \vdots \\ SNP_{n,1} & \cdots & SNP_{n,m} \end{pmatrix} \in \{0, 1, 2\} \times \{0, 1, 2\} \tag{2.5}$$

Cuando estos datos se utilizan de entrada para una red convolucional, se dice que m es la cantidad de *features* de entrada, n es el tamaño del *batch* y tenemos 3 canales. Cuando se utiliza en una RNN u otro algoritmo de texto, se dice que m es el largo de la secuencia y que tenemos 3 letras en el diccionario.

2.1.2.3. GWAS, Estudio de asociación del genoma completo

Actualmente, para encontrar asociaciones entre una variante de un gen en particular se utilizan los Estudios de Asociación del Genoma Completo (o GWAS, por sus siglas en inglés *Genome-Wide Association Study*). Este estudio consiste en realizar una prueba de hipótesis nula sobre cada SNP (**Sección 2.1.2.2**) y obtener el $p - value$ asociado. Esto reduce la necesidad de buscar mecanismos moleculares en todo el genoma de un individuo, reduciendo la experimentación a los genes ubicados alrededor de la zona en la que se encuentra el SNP. Las desventajas de esta aproximación consideran la falsa correlación de la variante, ya que si la población del estudio corresponde a múltiples historias genéticas, se puede encontrar como significativos genes asociados a la variabilidad de la etnia y no al mecanismo molecular del fenómeno estudiado [55].

Por lo general, se utiliza la regresión bayesiana para obtener los coeficientes asociados a cada variante y realizar la prueba de hipótesis nula sobre los coeficientes del modelo ajustado. En nuestro caso, al ser un fenotipo binario (caso o control) se utiliza un modelo logístico:

$$p(y) = \frac{1}{1 + e^{-\hat{y}}}, \text{ con y} \in \mathcal{B}(0, 1) \tag{2.6}$$

$$\hat{y} = \beta_0 + \beta_{SNP} x_{SNP} + \sum_{j \in \text{Covariantes}} \beta_j x_j + \sum_{k \in \text{Covariables}} \beta_k x_k + \epsilon \tag{2.7}$$

Con $\hat{y}$ el puntaje asociado al fenotipo estudiado, x_{SNP} el SNP que se quiere calcular significancia, x_j un valor que indica la covariablidad de todos los SNP obtenido desde un PCA, x_k cada una de las covariables usadas, ϵ el error asociado y β cada uno de los coeficientes estimados de la regresión. Para un modelo que solo considera las variantes genotípicas la regresión se reduce a:

$$\hat{y} = \beta_0 + \beta_{SNP} x_{SNP} + \sum_{j \in \text{PCs}} \beta_j x_j + \epsilon \tag{2.8}$$

El GWAS solo reporta el $p - value$ del coeficiente de la variable en estudio β_{SNP}. Para visualizar estos resultados se producen dos tipos de gráficos. El primero es conocido como *Manhattan Plot* donde se visualiza el $-\log(p)$ para cada SNP, ubicados de forma ordenada según su cromosoma y posición en este. Se puede ver un ejemplo en la **Figura C.18**. El segundo es un gráfico cuantil-cuantil o QQ (*QQplot*) que compara los cuantiles para dos distribuciones. En este caso, las dos distribuciones son la significancia esperada y la significancia observada. Para la significancia esperada se asume que cada variante cumple la hipótesis nula para una distribución χ^2. Se puede ver un ejemplo en la **Figura C.19**.

Para comparar los resultados con el modelo *Dual-stream CNN* (**Sección 3.3.2.2**), usaremos la misma metodología reportada [38], obteniendo el valor de la prueba de Wald sobre el modelo bayesiano que incluye todas las variantes:

$$\hat{y} = \beta_0 + \sum_{k=1}^{m} SNP_k$$

Estos resultados se visualizan utilizando un *Manhattan Plot* del valor estadísitico de Wald.

2.1.3. Aprendizaje Profundo

En este trabajo se utilizaran algoritmos de Aprendizaje Profundo (o *Deep Learning*), los que se basan en la utilización de redes neuronales artificiales (ANN, por sus siglas en inglés *Artificial Neural Network*). Para facilitar la descripción de las arquitecturas utilizadas, se presentan las definiciones y ecuaciones generales de las ANN. Como idea general, la ecuación de una capa de una red neuronal se escribe [21]:

$$h^{(i)} = g^{(i)}\left(W^{(i)^T} h^{(i-1)} + b^{(i)}\right) \tag{2.9}$$

Donde $g(M)$ es la función de activación que añade no linealidad al modelos, $x = h^{(0)}$ será la entrada y la predicción $\hat{y}$ se calcula como:

$$\hat{y} = W^{(l)} = U^T h^{(l-1)} x + b^{(l)} \tag{2.10}$$

Con U, la última capa de tamaño $B \times F \times C$, con B el tamaño del *batch*, F las *features* de salida de la capa anterior C la cantidad clases, en este caso 2. Por lo que la predicción $\hat{y}$ es de tamaño $B \times C$. Utilizamos *CrossEntropy* [42] como la función de pérdida:

$$\mathcal{L}(\hat{y}, y) = \begin{Bmatrix} l_1 & l_2 & \cdots & l_B \end{Bmatrix}^T \tag{2.11}$$

$$l_n = \frac{\sum_{n=1}^{B} l_n^{'}}{B} \tag{2.12}$$

$$l_n^{'} = -\sum_{c=1}^{C} \log\left(\frac{\exp(\hat{y}_{n,c})}{\exp\left(\sum_{i=1}^{C} \hat{y}_{n,i}\right)}\right) y_{n,c} \tag{2.13}$$

Con $l_n^{'}$ la función no reducida. La función de pérdida también se conoce como función de *loss* o simplemente *loss*.

Las capas convoluciones se utilizan para la clasificación de imagánes y texto. Estas capas se basan en la operación matemática conocida como convolución. Esta se define como [21]:

$$s(t) = \int x(a) w(t-a) da \tag{2.14}$$

Esto se escribe como el operador $*$, quedando la **Ecuación 2.14** como sigue:

$$s(t) = (x * w)(t) \tag{2.15}$$

El primer argumento x se le llama entrada o **input** y el segundo w se llama kernel. Como en computación se trabaja con una entreda discreta y no continua, la integral de la **Ecuación 2.14** se reescribe como:

$$s(t) = (x * w)(t) = \sum_{a=-\infty}^{\infty} x(a)w(t-a) \tag{2.16}$$

Lo más común es utilizar la operación de convolución sobre imágenes, lo que es una entrada en dos dimensiones. Sin embargo, en este trabajo se utiliza sobre secuencias de SNP (**Sección 2.1.2.2**), por lo que se usa la convolución según la **Ecuación 2.16**. La operación que se define en una capa convolucional de una dimensión en la librería Pytorch se escribe [42]:

$$h^{(i)}(B, C_{out_j}) = b^{(i)}(C_{out_j}) + \sum_{k=0}^{C_{in}-1} W^{(i)}(C_{out_j}, k) * h^{(i-1)}(B, k) \tag{2.17}$$

En la ecuación anterior se describe el cálculo de cada elemento de la matriz de salida $h^{(i)}$ en la posición k para el canal j.

2.1.4. Interpretabilidad

Obtener un modelo capaz de predecir la severidad es útil para determinar grupos de riesgo en la población y desarrollar mejores planes de prevención. Sin embargo, la efectividad de la vacuna [25] y la aparición de nuevas variantes [6], dejan rápidamente obsoleto este modelo. Una mayor contribución, es ser capaz de determinar las variables que determinan esta predicción. Si bien esta asociación corresponde solo a una correlación, funciona como un filtro para la búsqueda de los mecanismos reales que determinan la severidad y susceptibilidad para esta enfermedad. En este capítulo se describen las diferentes aproximaciones que se utilizan para interpretar los modelos utilizados.

2.1.4.1. *Feature Selection*

Los modelos basados en árboles de decisión, como Random Forest o XGBoost, ajustan un algoritmo basado en decisiones fácilmente interpretable. Incluso, estos modelos asignan un puntaje, conocido como *feature importance*, para cada una de las variables que recibe. En el caso de los clasificadores basados en regresiones, como la regresión logística o SVM, estos ajustan los valores de sus coeficientes (por lo general designados como β_i) para cada variable, los que se suelen interpretar como la influencia que tiene la variable en la predicción. Los modelos de ML implementados en la librería SciKit-Learn [43] entrega algunos de estos puntajes. Esta aproximación se utiliza en los datos clínicos para seleccionar las variables que más afectan o se correlacionan con la severidad predicha.

La desventaja de esta aproximación es que no evalúa las métricas del ajuste realizado, teniendo una idea de la significancia de las variables que solo corresponde a los datos de entrenamiento. Para aumentar la robustes de la determinación de variantes seleccionadas se utiliza *Cross validation* con el fin de obtener valores de significancia para cada todos los datos

del set.

2.1.4.2. Saliencia

Un mapa de saliencia[3] es un concepto usado en el procesamiento de información visual en organismos vivos. Según la hipótesis de saliencia V1 (de Corteza Visual Primaria) [67] la atención que se pone en una locación del campo visual, se relaciona directamente con las neuronas activadas en la corteza visual primaria. A este mapa de neuronas activadas se le llama mapa de saliencia y corresponde a un valor proporcional a la atención que se pone en dicha región del campo visual.

De acuerdo a esta teoría, las capas de neuronas en la corteza visual tiene un funcionamiento similar a las capas de una red convolucional. Es por esto que, para visualizar la clasificación que realiza una red convolucional, Simonyan, Vedaldi y Zisserman propusieron visualizar la saliencia obtenida en la clasificación de imagenes [49]. Esta idea también fue aplicada por Liu y colaboradores, en la secuencia de SNP para determinar, de forma equivalente, las variantes que más se correlacionan con la predicción [38]. Además de utilizar la arquitectura reportada en dicha investigación, también se adoptará este aproximación para interpretar los modelos entrenados.

Según [49] podemos definir la saliencia a partir del valor lineal del modelo para una clase c. Ésta se puede calcular a partir de una red convolucional utilizando la expansión de Taylor de primer orden:

$$S_c(I) \approx w^T I + b \tag{2.18}$$

Donde I es la imagen como una matriz por pixel (3 en caso de una imagen a color) o una secuencia en una dimensión, b es el *bias* y w^T se define como:

$$w^T = \left.\frac{\partial S_c}{\partial I}\right|_{I_0} \tag{2.19}$$

Finalmente, la saliencia (*saliency*) de un punto (o pixel) específico de la imagen se define como:

$$M_{i,j} = \max_c |w_{i,j}| \tag{2.20}$$

Como usaremos una secuencia, utilizamos una versión en una sola dimensión:

$$M_i = \max_c |w_i|$$

Lo que se puede obtener usando Pytorch [42], a partir de la entrada de la red en el *test set* [38]:

$$w = \frac{\partial Y}{\partial X} \Rightarrow M_i = \max_c |w_i| \tag{2.21}$$

[3] Traducción libre del inglés *saliency map*.

2.2. Trabajos Relacionados

En esta sección se justifica la aplicación de algoritmos usados en NLP (*Natural Language Processing*) en la predicción del fenotipo de un organismo utilizando como entrada datos genómicos. Para ello, se ejemplifica con trabajos relacionados la aplicación de estos algoritmos en datos genéticos.

2.2.1. Aprendizaje de Máquinas en datos genéticos

Como una estructura de datos ordenada, el genoma se ha utilizado como entrada de algoritmos de ML. Dejando de lado el Aprendizaje Profundo (**Sección 2.2.2**), los modelos de aprendizaje estadístico son usados de forma rutinaria en estudios genéticos poblacionales (GWAS, en más detalle en la **Sección 2.1.2.3**), donde se utilizan regresiones lineales de la forma [55]:

$$y = \beta_0 + \sum_{j=1}^{m_g} \beta_j x_j + \sum_{j=m_g+1}^{m} \beta_j x_j \tag{2.22}$$

Con m el total de variables, m_g las variables genéticas, $m - m_g$ las covariables no genéticas e y el fenotipo estudiado.

Otros algoritmos de ML usados son los árboles de decisión (*Decision Tree*), incluyendo *ensemblaing* como *Random Forest* o *XGBoost*, *Support Vector Machine* (SVM) y modelos estadísticos bayesianos [27] también han sido utilizados. Si bien, las métricas son menores que las obtenidas con algoritmos más avanzados, como Aprendizaje Profundo, estos algoritmos han permitido identificar las variables genéticas que se correlacionan con el fenotipo estudiado. Lograr obtener estas asociaciones correlacionales entre la genética y el fenotipo en modelos menos interpretables [54] se ha hecho bajo diferentes aproximaciones, a ser exploradas en la **Sección 2.1.4**.

2.2.2. Aprendizaje Profundo en bioinfórmatica y biomedicina

Desde que existe la capacidad computacional para algoritmos de Aprendizaje Profundo, éstos han sido utilizados en múltiples tareas [21], y la biología no es la excepción. Casi todas las arquitecturas de redes neuronales tienen un equivalente en la predicción de enfermedades, características morfológicas, imageneología y diagnóstico [18]. En esta sección se exploran algunos ejemplos para justificar la selección de la arquitectura a utilizar.

2.2.2.1. Redes convolucionales

El principal objetivo de las redes neuronales es el procesamiento de imágenes [21, 36]. En biomedicina, identificar de forma automática radiografías [51], cultivos bacterianos [19] o biopsias [31] es de gran utilidad en el diagnóstico de enfermedades. En ese caso, las arquitecturas de redes neuronales usadas son redes convolucionales en dos dimensiones. Este no es el

caso de este trabajo, en que los datos usados son de una dimensión.

Las redes convolucionales de una dimensión son usadas principalmente en series de tiempo como sonidos o datos meteorológicos [13]. Las secuencias de genes se pueden tratar como este tipo de dato, aunque no sean datos temporales. Son múltiples los ejemplos de redes convolucionales para datos genéticos [1, 23, 33, 34, 68], la parte que interesa para este estudio es la interpretabilidad de esa arquitecturas. La saliencia se utiliza para asignar un puntaje a la influencia de los pixeles en una imagen en la identificación de esta [49] **Sección 2.1.4.2**. En 2019, Yang Liu y colaboradores [38] utilizaron esta aproximación para identificar variantes asociadas a diferentes fenotipos en el haba de soya. Para comprobar si la saliencia de los SNPs asigna valores correctos, se comparan los valores obtenidos con el análisis estadísitico GWAS (**Figura 2.4**).

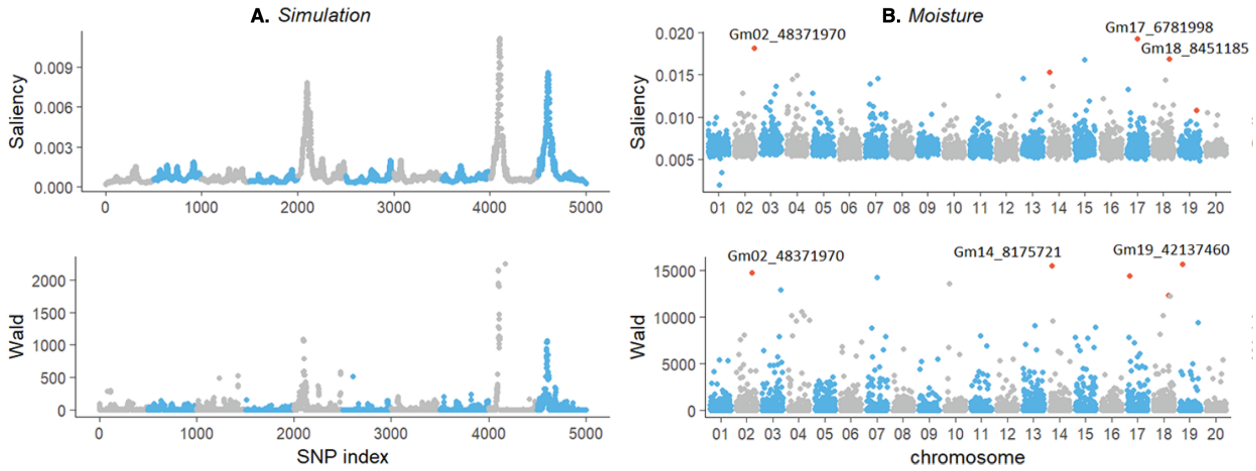


Figura 2.4: Comparación entre la saliencia obtenida utilizando la red Dual-stream CNN (arriba) y el GWAS utilizando prueba estadísitica de Wald (abajo). **A.** Comparación para una simulación de variantes genéticas. Esta simulación genera variantes que se correlacionan en mayor medida con el fenotipo artificial. **B.** Comparación sobre las variantes genéticas de la soya para la predicción de la humedad presente en la haba. Fuente [38]

Otra utilizacación de las redes convolucionales, es como reductores de la dimensionalidad, por ejemplo, para secuencias temporales muy largas [29]. Secuencias muy largas que utilizarían demasiada memoria en GPU o tomarían demasiado tiempo de entrenamiento, se reducen a secuencias más cortas adosando el modelo a utilizar en la clasificación, a una red convolucional de una dimensión. Al adosar ambas arquitecturas, el modelo que realiza la clasificación puede ser más pequeño y la red convolucional adapta sus pesos a la misma clasificiación para mantener la información reduciendo la dimensionalidad.

2.2.2.2. Redes recurrentes

Las redes recurrentes surgen como una alternativa a recuperar información contextual y ser capaces de permitir entradas de largo variable [21]. Siendo usadas principalmente en NLP, han sido utilizadas en bioinformática utilizando como entrada el genotipo nucleótido a nucleótido [66], secuencias de mRNA [26], o utilizando $k-$meros [23]. Esto último se refiere a tomar k

nucleótidos y utilizar este conjunto como palabra, la siguiente palabra se forma moviéndose un nucleótido y tomar los siguiente k. Formalmente, esto es:

$$k - mers(\text{Secuencia} = x_1 x_2 \cdots x_m) = \tag{2.23}$$

$$(x_1 x_2 \cdots x_k)(x_2 x_3 \cdots x_{k+1}) \cdots (x_j x_{j+1} \cdots x_{j+k-1}) \cdots (x_{m-k+1} x_{m-k+2} \cdots x_m)$$

En un ejemplo:

$$\text{Entrada: } AAAATTTCCGCTGT \ \wedge \ k = 4$$

$k-$meros: *AAAA AAAT AATT ATTT TTTC TTCC TCCG CCGC CGCT GCTG CTGT*

Es fácil notar que de una secuencia de largo m, la forma en $k-$meros es de $m - k + 1$.

2.2.2.3. *Self-Attention* sobre secuencias de DNA

Desde la introducción de *Self-Attention* en la publicación de Vaswani y colaboradores en 2017 [59], las arquitecturas conocidas como *transformers* se han convertido en el nuevo estado del arte en el procesamiento de lenguaje natural (NLP). Éstas arquitecturas permiten mantener la ventaja de las redes recurrentes de capturar información contextual, pero además recuperan la capacidad de paralelizar operaciones que las RNN realizan de forma secuencial.

Dado lo computacionalmente costoso de la operación $Multi - head\ Attention$, las redes *transformers* están limitadas en el tamaño de la secuencia de entrada, a$\approx$ 1000 de largo, antes de perder la información contextual. En miras a esta restricción, en 2019, Dai y colaboradores [14] propusieron una arquitectura basada en *transformer* capaz de obtener información contextual mucho más alejadas. Esta arquitectura fue utilizada por Clauwaert, Menschaert y Menschaert en 2021 sobre un genoma procarionte completo, para predecir sitios de unión de la transcriptasa [9, 10]. Otra de las ventajas de estas arquitecturas es que la visualización de la atención permite interpretar la correlación entre la predicción y los nucleótidos usados en la entrada. Este mapeo es de mucha utilidad para los estudios genéticos y permite ser interpretado desde el conocimiento en el área, como lo que hicieron Clauwaert, Menschaert y Menschaert y que se muestra en la **Figura 2.5**.

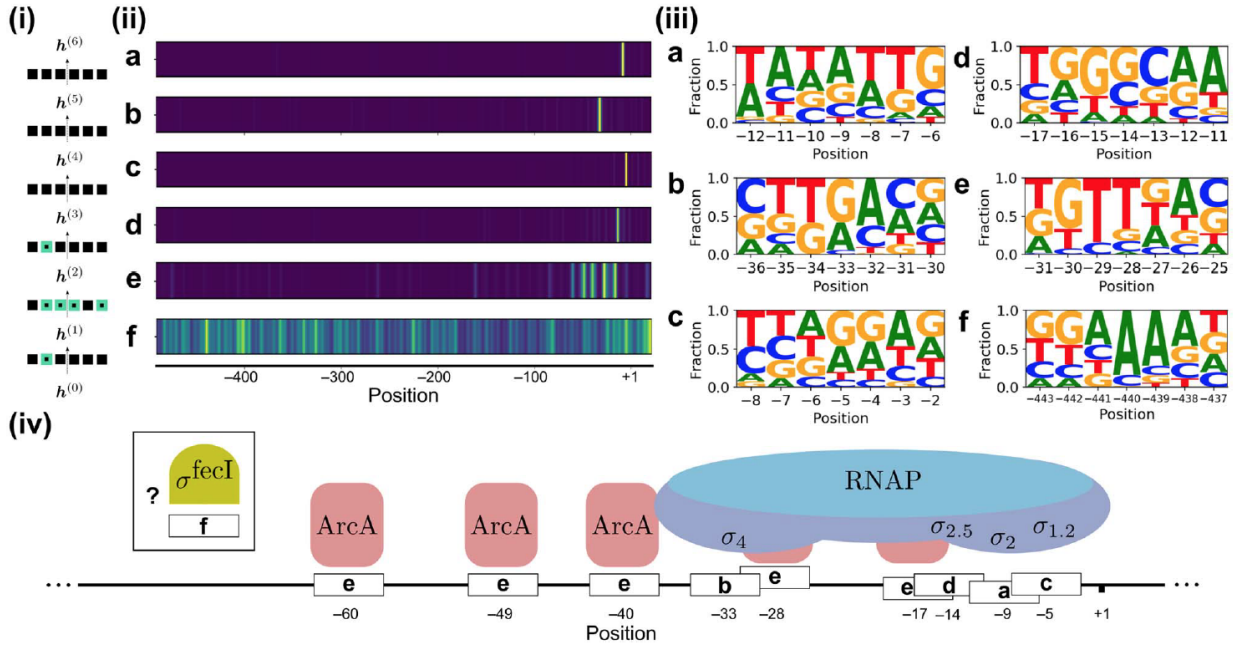


Figura 2.5: Ejemplo de interpretabilidad de la arquitectura usada por Clauwaert, Menschaert y Menschaert **(i)** *Attention heads* usadas para calcular los estados ocultos dentro de la arquitectura. Los cuadrados verdes se muestran en **(ii)**. **(ii)** Puntajes de atención para cada una de las *Attention head* mostradas en verde en **(i)**. **(iii)** Motivos de secuencia de los 50 nucleótidos de mayor puntaje para cada una de las *Attention head* mostradas en **(ii)**. **(iv)** Interpretación de las secuencias de mayor puntaje en el contexto del reconocimiento de los sitios de transcripción por la RNA polimerasa. El diagrama fue obtenido de forma integra desde la publicación original. Fuente: [9]

En 2019 también, Devlin y colaboradores introducen el concepto de pre-entrenamiento en *transformers* [16]. La idea consiste en entrenar el modelo en alguna tarea generalizada, particularmente en el Modelamiento de Lenguaje Enmascarado (MLM, por sus siglas en inglés *Masked Language Model*), para luego realizar *fine tuning* en la tarea en específico. Esta idea fue llevada al genoma por Ji y colaboradores en 2020, sobre promotores, factores de transcripción e identificación de variantes funcionales [30]. Para mostrar la interpretabilidad de los modelos, Ji y colaboradores, desarrollan un software que muestra la atención sobre la secuencia de entrada.

DNABERT aplicado a secuencias de DNA implementado por Ji y colaboradores, posee la desventaja de recibir secuencias de máximo 512 tokens. En vista a esta limitante, investigadores de Google desarrollaron la arquitectura BigBird [65], que admite 4096 tokens. Big Bird también fue utilizado sobre secuencias de DNA para predecir metilizaciones y patrones de cromatina.

2.2.3. Aplicaciones en Covid-19

Luego que el Covid-19 fuese declarado como pandemia el 30 de enero de 2020 por la OMS (Organización Mundial de Salud, o WHO, pos sus siglas en inglés *World Health Organization*)

[52] los algoritmos de ML se han utilizado para detectar nuevos brotes de la enfermedad, analizar tratamientos, realizar trazabilidad de contactos, predicción de las diferentes olas, desarrollo de drogas y vacunas [35]. En 2020, utilizando datos en sintomatología recopilados mediante una aplicación para Smartphone, Zoabi, Deri-Rozov y Shomron, desarrollaron un modelo de regresión lineal para predecir la probabilidad de contagio basado en los síntomas presentados [69]. En el presente año, se utiliza otro modelo de ML, el análisis de discriminante lineal (o LDA, por sus siglas en inglés *Linear discriminant analysis*), en la predicción de dignóstico y severidad utilizando datos clínicos [63].

Más específicamente, dentro de los algoritmos de Aprendizaje Profundo [35], se han desarrollado modelos de NLP para detectar brotes utilizando *tweets*, detección de noticias falsas y búsqueda de publicaciones relacionadas a Covid-19. Análisis de imágenes se han utilizado en diagnóstico y detección de lugares con mayor probabilidad de contagio. También, se desarrollaron una red convolucional capaz de detectar Covid-19 utilizando como entrada imagénes de rayos X del pecho de pacientes [32] y un modelo que predice la incidencia de la enfermedad basada en el sónido de la tos en un tumulto [7].

Dentro de la tarea específica de este trabajo, predicción de severidad mediante datos genéticos, la iniciativa COVID-19 hg (*COVID-19 host genetics*)[4], de la que este trabajo es parte, recopila y publica los estudios de asociación entre variantes genéticas y la severidad del cuadro [12]. La principal herramienta utilizada es el GWAS (**Sección 2.1.2.3**), que utiliza una regresión logística como modelo predictor. Existen estudios que también utilizan GWAS para detectar otras variantes no genéticas [61]. En cuanto a la utilización de algoritmos NLP en genética asociado a Covid-19, existe un análisis de la evolución del virus [24] y una propuesta de proyecto que busca analizar la secuencia del virus [53].

Dado al tamaño del genoma humano y a la dificultad de obtener el número necesario de ejemplos, no hay actualmente investigaciones que utilicen Aprendizaje Profundo en el genoma del hospedero en Covid-19, donde radica la importancia de este trabajo. Se utilizan datos de participantes chilenos, para reducir la variabilidad genética de los participantes.

[4] https://www.covid19hg.org/about/

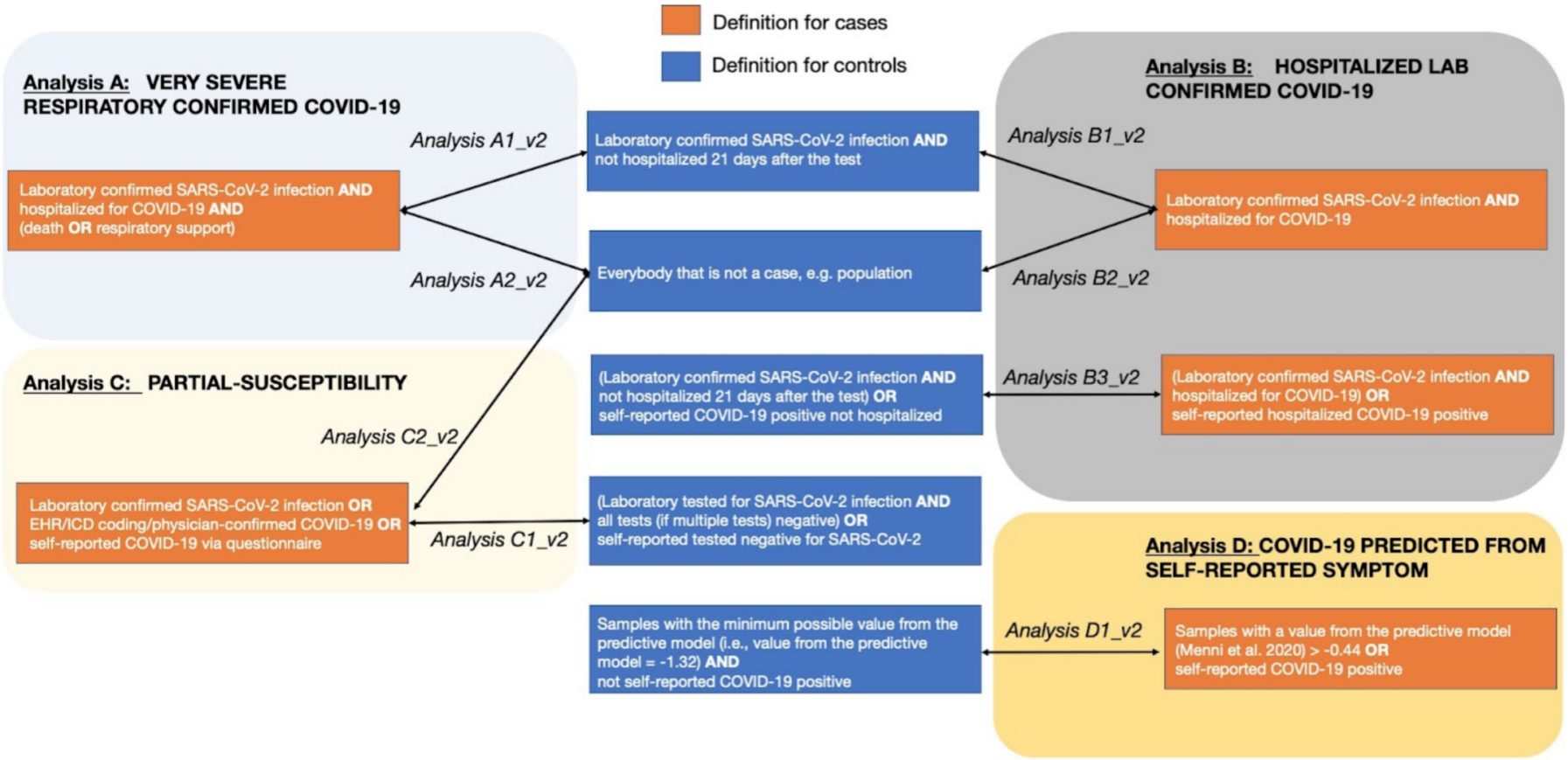


Figura 2.6: Resumen gráfico de los análisis descritos por la iniciativa internacional *COVID19 Host Genetics*. Para cada análisis se describe el fenotipo de los participantes, asignando si estos son controles o casos. Por la característica de los datos recopilados, el proyecto no puede realizar el análisis D1. Fuente: [12, 15]

2.3. Antecedentes

En esta sección se incluye la información sobre la iniciativa COVID-19hg [12]. Este trabajo se encuentra dentro del marco de dicha iniciativa internacional. La definición del fenotipo a predecir por los modelos presentados se encuentra estandarizada por esta iniciativa. También se presenta el formato de los datos genéticos disponibles.

2.3.1. Análisis proyecto COVID19hg

Los análisis a realizar son definidos por el proyecto macro internacional COVID19hg (*COVID19 Host Genetics*) [15]. Estos se abrevian usando una letra, para identificar los casos a analizar, y un número, que definen el control. Si los controles corresponden a participantes contagiados, se trata de análisis de severidad y se identifican con el número 1. Los controles de susceptibilidad parcial, utilizan como grupo control a la población general y se identifican con el número 2. Una imagen resumen de los análisis definidos se muestra en la **Figura 2.6**. Para este trabajo se utiliza el fenotipo B1 de hospitalizados entre la población infectada. Formalmente, se define como "hospitalizados COVID-19 confirmados por laboratorio"[5]. Para este análisis los casos y controles se definen como sigue:

- **Casos**: Hospitalizados confirmados por laboratorio (PCR o prueba serológica) SARS-CoV-2, cuya admisión se relaciona a síntomas por coronavirus.
- **Controles**: Confirmados por laboratorio (PCR o prueba serológica) SARS-CoV-2 y que no fueron hospitalizados después de 21 días de la confirmación por examen de laboratorio.

2.3.2. Datos genéticos disponibles

Para este proyecto, los datos genotipificados corresponden a $\sim 5 \cdot 10^5$ SNPs. Los que fueron obtenidos por los laboratorios Broad Institute[6] y el Finland Institute for Molecular Medicine, FIMM[7]. Los datos son recibidos en formato VCF (por sus siglas en inglés *Variant Call Format*), para ser trabajados en formato Plink [44, 45]. Este formato especifica 3 archivos (1) un archivo `.bim` que contiene información sobre las variantes genéticas; (2) `.fam` que contiene un id de participante, un id de familia, sexo y fenotipo; y (3) `.bed` que contiene una matriz (como se explica en la **Sección 2.1.2.2**) en formato binario, que indica el valor del SNP para cada variante y participante.

Las variantes originales de todos los centros se imputan mediante técnicas estadísticas y genomas de referencia. Realizada la imputación y descartando las variantes monomórficas, iguales en todos los participantes, el total de variantes obtenidas es de $7406860 \approx 7.4 \cdot 10^6$. Estos datos se someten a controles de rutina, para descartar posibles muestras contaminadas.

[5] Traducción libre
[6] https://www.broadinstitute.org/
[7] https://www.fimm.fi/en/

Como los participantes incluidos no son familiares entre sí, una de estos controles consiste en eliminar participantes que se parecen mucho entre ellos, ya que esto puede significar que las muestras se cruzaron. Eliminar diferencias genéticas con el sexo reportado. O eliminar *outliers*, los que se detectan graficando los componentes principales obtenidos por PCA.

2.3.2.1. Significancia reportada de las variantes genéticas

La iniciativa internacional COVID-19hg reporta la significancia realizando un meta-análisis de los proyectos que la conforman. Esta significancia se encuentra disponible para cada variante[8]. Utilizando estos datos se realiza el *Manhattan Plot* que se muestra en la **Figura 2.7**.

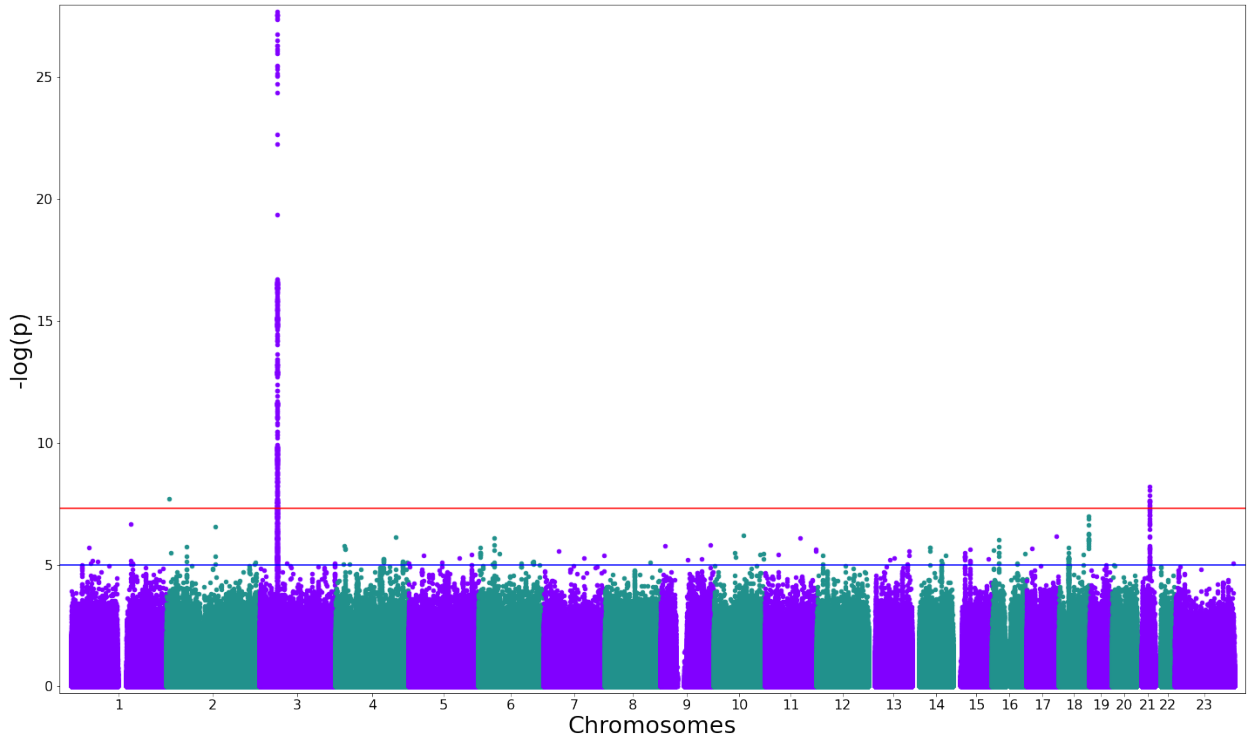


Figura 2.7: *Manhattan Plot* del meta-analisis realizado por la iniciativa internacional COVID-19hg [12]. Se muestran las variantes en orden según el cromosoma y la posición en este. La línea azul se encuentra en $-\log(p) = 5$ o $p = 10^{-5}$. Los SNPs por arriba de esta línea se consideran **sugerentes** de significancia. La línea roja indica $p = 5 \cdot 10^{-7}$. Sobre esta línea, los SNPs se consideras estadísiticamente significativos para el fenotipo estudiado. Una versión más detallada se encuentra en el Anexo en la **Figura C.18**. El *QQplot* se muestra en la **Figura C.19**.

La significancia calculada en la iniciativa corresponde al valor p del coeficiente β_{SNP} de la siguiente regresión [11]:

$$\hat{y} = \beta_0 + \beta_{SNP} x_{SNP} + \beta_{\text{age}}\text{age} + \beta_{\text{age},2}\text{age}^2 + \beta_{\text{sex}}\text{sex} + \beta_{\text{age,sex}}\text{age} \cdot \text{sex} + g$$

[8] De forma interactiva en https://app.covid19hg.org/ y disponible para descarga en https://www.covid19hg.org/results/r6/.

Con g las 20 componentes principales de un PCA sobre todas las variantes. Recordando (**Ecuación 2.6**) que $\hat{y}$ es:

$$p(y) = \frac{1}{1 + e^{-\hat{y}}}, \text{ con y} \in \mathcal{B}(0, 1)$$

Con $p(y) = 0$ para controles y $p(y) = 1$ para los casos. Para obtener la significancia se realiza dicha regresión logísitica para cada variante.

En el caso de este proyecto, se realiza un GWAS parecido con los participantes disponibles para el análisis B1 (**Sección 2.3.1**), o de hospitalización entre los participantes infectados. Este GWAS utiliza la regresión lógisitica:

$$\hat{y} = \beta_0 + \beta_{SNP} x_{SNP} + \beta_{\text{age}} \text{age} + \beta_{\text{sex}} \text{sex} + \beta_g g$$

El *Manhattan Plot* de todas las variantes imputadas disponibles, se muestra en la **Figura 2.8**. La significancia reportada es menor, debido a que la cantidad de participantes es mucho menor para el proyecto (87671 para la iniciativa y 1912 para el proyecto). Sin embargo, los *peaks*, o SNP más significantes, se encuentran aproximadamente en las mismas posiciones.

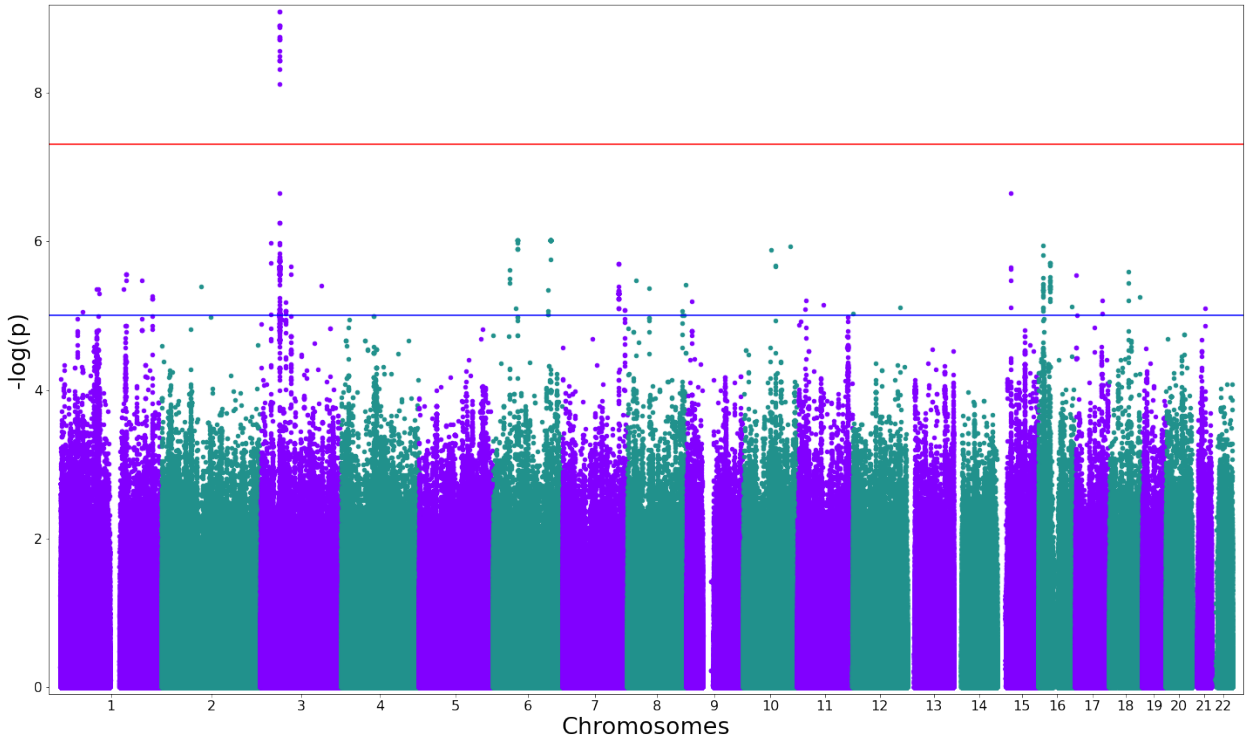


Figura 2.8: *Manhattan Plot* del GWAS realizado para todas las variantes imputadas del proyecto. Se muestran las variantes en orden según el cromosoma y la posición en este. La línea azul se encuentra en $-\log(p) = 5$ o $p = 10^{-5}$. Los SNPs por arriba de esta línea se consideran **sugerentes** de significancia. La línea roja indica $p = 5 \cdot 10^{-7}$. Sobre esta línea, los SNPs se consideras estadísiticamente significativos para el fenotipo estudiado. Una versión más detallada se encuentra en el Anexo en la **Figura C.20**. El *QQplot* se muestra en la **Figura C.21**.

Debido a que el análisis se realiza para cada SNP de forma independiente, podemos extraer la significancia para las variantes originalmente secuencias. Este GWAS se muestra en la **Figura 2.9**, donde se aprecía que ninguna variante pasa el punto de significancia ($p < 5 \cdot 10^{-7}$). Este resultado se repite repitiendo el análisis para otros dos regresiones: (1) considerando solo las variantes y sus PCs (**Figura C.30**), y (2) añadiendo las variables clínicas seleccionadas (**Figura 2.10**). Debido a esto, se utilizan otros dos *sets* de datos de variantes genéticas como entrada para los modelos.

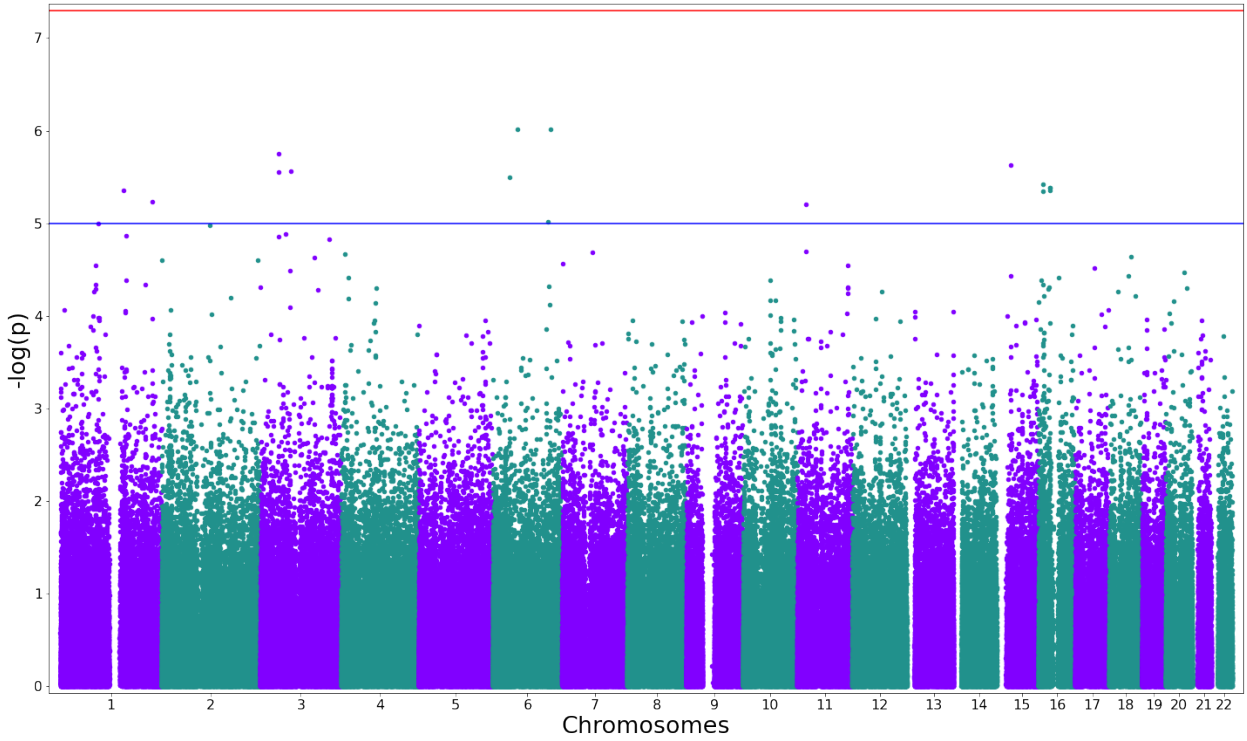


Figura 2.9: *Manhattan Plot* del GWAS realizado solo para las variantes imputadas originalmente genotipificadas por microarreglo del proyecto. Se muestran las variantes en orden según el cromosoma y la posición en éste. La línea azul se encuentra en $-\log(p) = 5$ o $p = 10^{-5}$, los SNPs por arriba de esta línea se consideran **sugerentes** de significancia. Mientras que la línea roja indica $p = 5 \cdot 10^{-7}$. Una versión más detallada se encuentra en el Anexo en la **Figura C.28**. El *QQplot* se muestra en la **Figura C.29**.

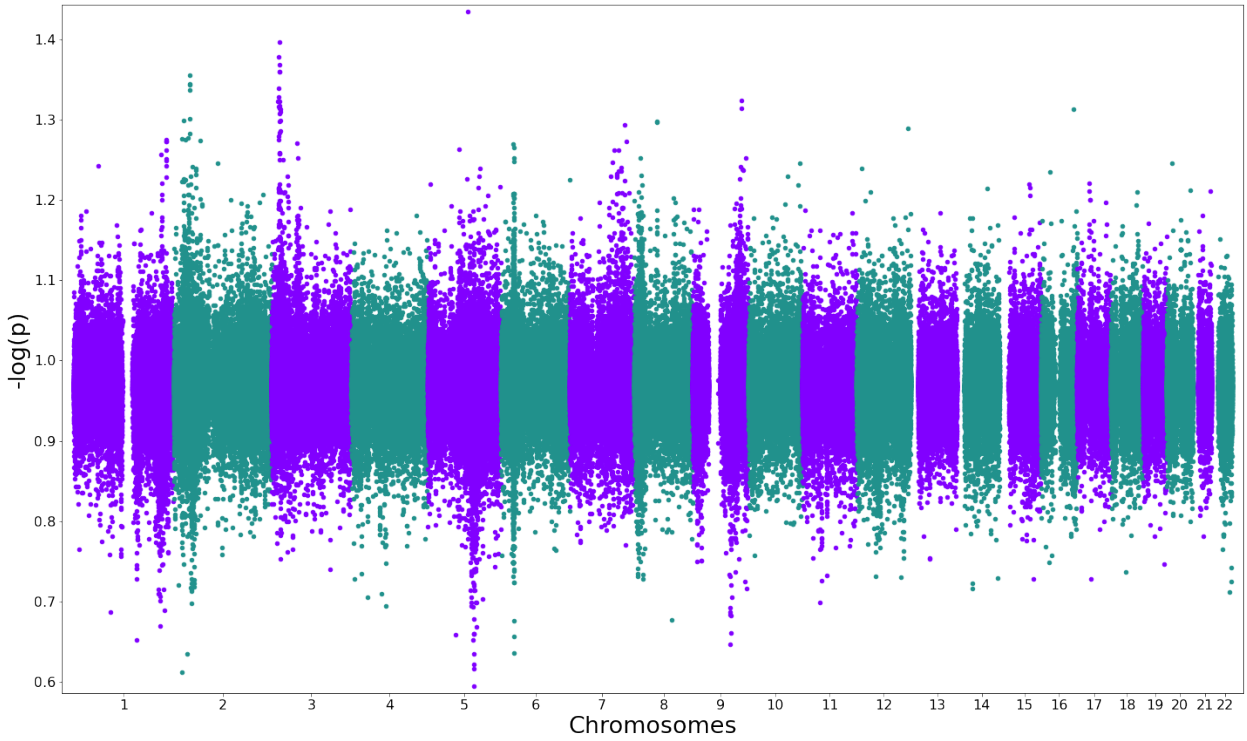


Figura 2.10: *Manhattan Plot* del GWAS realizado para las variantes imputadas originalmente genotipificadas por microarreglo para el proyecto, además de las variables clínicas como covariantes. Se muestran las variantes en orden según el cromosoma y la posición en éste. Sobre esta línea se considera significante para el fenotipo estudiado. Una versión más detallada se encuentra en el Anexo en la **Figura C.32**. El *QQplot* se muestra en la **Figura C.33**.

Capítulo 3

Metodología

En esta sección se describe la metodología a utilizar y el diseño de los experimentos. Con el fin de determinar si los datos genéticos contribuyen en la predicción de la severidad Covid-19, se entrenan modelos utilizando los datos clínicos autoreportados como punto de partida. Estos son comparados con otros modelos que utilizan tanto los datos clínicos como genéticos.

Para realizar el diseño de estos modelos, primero se prueban modelos que utilizan la información genética, aprovechando de aplicar las aproximaciones descritas en la interpretabilidad de estos modelos. Como la cantidad de variables genéticas es mayor que los datos disponibles, es esperable que muchas de estas variantes aporten ruido a la predicción. Por lo que, además, se utilizarán los modelos de datos genéticos, para extraer las variantes que más se correlacionan con la predicción y agregarlas a los modelos de datos clínicos, para constatar si las predicciones mejoran.

3.1. Datos clínicos autoreportados

Como parte de la inciativa COVID-19 hg, el proyecto ANID COVID0961 realizó un reclutamiento de participantes contagiados con el virus SARS-CoV-2 [48] en las cinco macrozonas del país. Los datos utilizados corresponden a la encuesta desarrollada por el proyecto y a la ficha clínica (CRF, por sus siglas en inglés *Case Report Form*) de participantes que consintieron su uso. En este capítulo se describen los datos obtenidos, preprocesados y estandarizados, junto con la definición de severidad definida desde los datos autoreportados según los criterios de la iniciativa internacional [12].

3.1.1. Descripción de los datos

Los datos obtenidos corresponden a datos autoreportados llenados mediante encuesta. Esta encuesta fue diseñada para obtener datos personales, socieconómicos, demográficos y antecedentes de salud de los participantes (ver Tablas B.1-B.5, tercera columna), junto con datos que nos permitan asignar una severidad al cuadro, como síntomas y hospitalización. Junto con la encuesta, CRF recopila estos mismos datos desde la ficha clínica del participante (ver

Tablas B.1-B.5, cuarta columna). Las columnas utilizadas corresponden a las comunes entre estas dos fuentes (ver Tablas B.1-B.5, quinta columna), eliminando aquellas que se utilizan para definir la severidad de la enfermedad.

Tabla 3.1: Resumen de la cantidad de variables según el tipo de dato y la categoría a la que pertenecen. El tipo de dato puede corresponder a categoría, booleano o numérico. La categoría a la que pertenece puede ser datos personales, geográficos, socioconómicos, etnicidad, hábitos de salud, comorbilidades, tipo de cáncer o medicamentos que consume.

	Categóricos	Booleanos	Numéricos	Total categoría
Personal	3	0	4	7
Geográficos	2	1	1	4
Socioeconómicos	3	0	0	3
Etnicidad	1	0	0	1
Hábitos de Salud	3	8	0	11
Comorbilidades	0	25	0	25
Cáncer	0	9	0	9
Drogas	0	16	0	16
Total variables	12	59	5	76

El dataset final a utilizar contiene 76 columnas las cuales se resumen según su tipo y categoría en la **Tabla 3.1**. Los datos a usar son 1912, los que corresponden a los 2665 participantes reclutados para el proyecto, cuyo genotipo se obtuvo y se filtro según criterios rutinarios (**Sección 3.2**). Los datos faltantes se muestran en la **Tabla 3.3**.

3.1.2. Severidad

Para definir un criterio de severidad del cuadro, y hacer más fácil el etiquetado de los datos para los análisis, se utiliza la siguiente definición excluyente de severidad:

1. **No contagiado**: participantes que, mediante encuesta, reportaron no haberse contagiado pese a estar expuesto al virus. Estos participantes se utilizan para otros análisis del proyecto.

2. **Asintomático**: participantes contagiados que no reportaron síntomas. Se le asigna esta categoría si presenta el valor booleano `false` en todos los síntomas.

3. **Leve**: participantes contagiados que reportaron síntomas y que no reportaron haber sido hospitalizados. Si fueron hospitalizados presentan el valor booleano `true` en la variables hospitalización.

4. **Hospitalizados**: participantes que fueron hospitalizados debido a Covid-19, pero que no pasaron a unidad hospitalarias de cuidados intensivos.

5. **Hospitalizados Graves**: participantes que fueron hospitalizados en cuidados intensivos (UCI, UTI, UAC), pero que no requirieron ventilación asistida invasiva (pueden haber requerido ventilación no invasiva).

6. **Crítico con ventilación mecánica:** participantes que requirieron ventilación mecánica invasiva, pero que sobrevivieron.

7. **Letal**: participantes que fallecieron debido a Covid-19.

Para simplificar la predicción de los modelos, en este trabajo se utilizará el análisis B1 de severidad (**Sección 2.3.1**). En la siguiente sección se describe cómo se clasifican los participantes, entre casos y controles, para este análisis, a partir de la escala de severidad descrita.

3.1.2.1. Cálculo de fenotipo para análisis

Para determinar si los casos corresponden a controles o casos para el análisis B1, se utilizaron otros datos del reclutamiento sobre síntomas y hospitalización (todas las variables se encuentra listadas en las Tablas B.1-B.5 en el **Anexo B.1**), el detalle del uso de estas variables se muestra en la **Tabla 3.2**.

Tabla 3.2: Determinación de fenotipo a través de las variables disponibles. En el caso del índice de riesgo, se muestran las categorías en que el participante debe pertenecer para ser designado como caso o control. Además de esta variables se utilizaron los atributos **Confirmado**, **Contagiado**, y **<21 días**. **Confirmado** se determinó desde el método en que fue diagnósticado, asignandose como verdadero si este se realizó por PCR o serología. **Contagiado** se asigna falso a todos los participantes que reportan no haberse contagiado o que corresponden a participantes control. **<21 días** comprueba que la fecha de admisión a alguna unidad hospitalaria sea antes de 21 días de la fecha de aparición de síntomas o examen positivo. Si algún participante no es control o caso para algún análisis, se excluye de dicho análisis.

Variable	Valor	B1	
		Caso	Control
Índice de Riesgo	No contagiado	✗	✗
	Asintomático	✗	✓
	Leve	✗	✓
	Hospitalizado	✓	✗
	Hospitalizado Grave	✓	✗
	Crítico con ventilación mecánica	✓	✗
	Letal	✓	✗
Contagiado	Sí	✓	✓
	No	✗	✗
Confirmado	Sí	✓	✓
	No	✗	✗
<21 días	Sí	✓	✗
	No	✗	✓

Tabla 3.3: Datos faltantes del dataset final. Se muestra la cantidad participantes sin dato para cada columna. El porcentaje que este corresponde, considerando el total de datos, para los controles esto corresponde a 1556 y para los casos 356, con 1912 datos en total.

	Controls	Cases	Totals
Age	0 (0.00 %)	8 (2.47 %)	8 (0.43 %)
Province	1 (0.06 %)	12 (3.70 %)	13 (0.69 %)
Population Density	1 (0.06 %)	15 (4.63 %)	16 (0.85 %)
Blood type	698 (45.09 %)	215 (66.36 %)	913 (48.77 %)
Rh type	684 (44.19 %)	217 (66.98 %)	901 (48.13 %)
Weight	11 (0.71 %)	92 (28.40 %)	103 (5.50 %)
Height	10 (0.65 %)	118 (36.42 %)	128 (6.84 %)
BMI	11 (0.71 %)	119 (36.73 %)	130 (6.94 %)
Country of birth	3 (0.19 %)	2 (0.62 %)	5 (0.27 %)
Educational level	21 (1.36 %)	119 (36.73 %)	140 (7.48 %)
Occupational situation	22 (1.42 %)	128 (39.51 %)	150 (8.01 %)
Tobacco Consumption	28 (1.81 %)	28 (8.64 %)	56 (2.99 %)
Daily cigars	41 (2.65 %)	46 (14.20 %)	87 (4.65 %)
More than 100 cigars	38 (2.45 %)	34 (10.49 %)	72 (3.85 %)
Alcohol Consumption	29 (1.87 %)	37 (11.42 %)	66 (3.53 %)
Diabetes	8 (0.52 %)	2 (0.62 %)	10 (0.53 %)
Coronary atherosclerosis	9 (0.58 %)	2 (0.62 %)	11 (0.59 %)
Hypertension	5 (0.32 %)	1 (0.31 %)	6 (0.32 %)
Vascular accident	10 (0.65 %)	2 (0.62 %)	12 (0.64 %)
Heart problem	20 (1.29 %)	3 (0.93 %)	23 (1.23 %)
Dialysis	9 (0.58 %)	2 (0.62 %)	11 (0.59 %)
Hepatitis	9 (0.58 %)	2 (0.62 %)	11 (0.59 %)
Anemia	10 (0.65 %)	2 (0.62 %)	12 (0.64 %)
Asthma	10 (0.65 %)	2 (0.62 %)	12 (0.64 %)
Cystic fibrosis	20 (1.29 %)	2 (0.62 %)	22 (1.18 %)
Pulmonary fibrosis	19 (1.23 %)	2 (0.62 %)	21 (1.12 %)
Pulmonary condition	9 (0.58 %)	2 (0.62 %)	11 (0.59 %)
Cancer	9 (0.58 %)	2 (0.62 %)	11 (0.59 %)
Condition that affects the brain	9 (0.58 %)	2 (0.62 %)	11 (0.59 %)
HIV	9 (0.58 %)	2 (0.62 %)	11 (0.59 %)
Tuberculosis	9 (0.58 %)	2 (0.62 %)	11 (0.59 %)
Weakened immune system	11 (0.71 %)	3 (0.93 %)	14 (0.75 %)
Mental health problems	10 (0.65 %)	2 (0.62 %)	12 (0.64 %)
High cholesterol	21 (1.36 %)	3 (0.93 %)	24 (1.28 %)
Obesity	20 (1.29 %)	3 (0.93 %)	23 (1.23 %)
Ulcerative colitis or Crohn's disease	20 (1.29 %)	3 (0.93 %)	23 (1.23 %)
Rheumatoid arthritis	20 (1.29 %)	3 (0.93 %)	23 (1.23 %)
Lupus	22 (1.42 %)	3 (0.93 %)	25 (1.34 %)
Other rheumatoid diseases	20 (1.29 %)	3 (0.93 %)	23 (1.23 %)

3.1.3. Imputación

Debido a que la mayoría de los modelos a usar, a excepción de XGBoost, no reciben datos nulos, se opta por realizar la imputación de estas variables. Esto se realiza usando el algoritmo KNN (por sus siglas en inglés *K-nearest neightborhood*, o k vecinos más cercanos) [58], mediante el paquete DMwR [57]. La cantidad de datos nulos por variable se muestra en la **Tabla 3.3**.

La necesidad de imputar datos faltantes se evidencia también en la cantidad de datos completos, es decir, participantes sin ninguna variable faltante. Previo a la imputación los participantes con datos completos eran 871 de los 1912 totales, conformados por 788 controles y 83 casos, en contraste con los 1556 controles y 356 casos del total del set. Se justifica la imputación para evitar la pérdida de 1041 participantes.

3.1.4. Análisis estadístico

Antes de utilizar los datos clínicos, se realizan controles estadísticos para determinar anomalías, porcentaje de datos nulos y variables significativas. Este proceso se realizó de forma iterativa durante el reclutamiento, por lo que algunos resultados, que llevaron a eliminar o cambiar variables, no se encuentran en este documento por simplicidad, mostrando solamente la versión final. Este análisis se realiza en R [46] utilizando librerías de Tidyverse [62] y MASS [60].

Los análisis estadísticos exploratorios incluyen:

1. Para los datos categóricos y booleanos, la cantidad de valores por casos, controles y total.
2. Para los datos numéricos el promedio y desviación estándar.
3. Para todas las variables la cantidad de datos faltantes.

La primera aproximación de la significancia se realiza utilizando una prueba estadística univariada, es decir, las variables se analizan en relación con el fenotipo una a la vez. Para los datos numéricos se realiza la prueba estadística de t-student, que compara el promedio de la variable para cada categoría de una variable respuesta, fenotipo en este caso, y mide la significancia usando la distribución t para $n-1$ grados de libertad. Esto corresponde a realizar la siguiente prueba de hipótesis nula:

- $\mathbf{H_0}$: $\mu_{\text{Label=0}} = \mu_{\text{Label=1}}$
- $\mathbf{H_1}$: $\mu_{\text{Label=0}} \neq \mu_{\text{Label=1}}$

Con μ el promedio poblacional de la variable estudiada. Para esta prueba, el valor estadístico T se calcula como sigue:

$$T = \frac{\overline{X}_{\text{Label=0}} - \overline{X}_{\text{Label=1}}}{\sqrt{\frac{s_0^2}{n_0} + \frac{s_1^2}{n_1}}} \sim t_{n-1} \tag{3.1}$$

Para los datos categóricos y booleanos, se utiliza la prueba de χ^2 de Pearson. El valor estadístico χ^2 se calcula como sigue:

$$X^2 = n \sum_{i=1}^{r} \sum_{j=1}^{c} p_i p_j \left(\frac{O_{i,j} - p_i p_j}{p_i p_j} \right)^2 \sim \chi^2_{n-1} \tag{3.2}$$

Con c categorías por variable para los controles y r para los casos, n es el total de observación, p_i la probabilidad para cada categoría en los controles $\left(\frac{1}{c}\right)$, p_j probabilidad por los casos $\left(\frac{1}{r}\right)$ y $O_{i,j}$ la cantidad observada de cada combinación de categorías.

Para comparar la significancia de las variables entre sí con la variable respuesta, severidad, se realiza una prueba estadrísitica multivariada utilizando el función de R `drop1`. Que utilizando un ajuste logístico compara los coeficientes obtenidos utilizando todas las variables, con los coeficientes de un segundo modelo sin la variable en estudio.

Los análisis estadísticos se realizan sobre los datos no imputados e imputados, para constatar que este proceso no altera la distribución de los datos.

3.2. Datos Genéticos a nivel de Genoma

El genoma se refiere a toda la información genética de un organismo. Por lo que tener datos genéticos a nivel de genoma, significa tener datos genéticos distribuidos en el genoma completo, que en el caso de los humanos, corresponde a $\sim 3.1 \cdot 10^9$ nucleótidos (o letras) [20]. En esta sección se describe el formato de estos datos, algunos preprocesamientos y la codificación que se utilizará para que estos datos sirvan de entrada a los modelos.

Figura 3.1: Representación de la posición de cada variante (SNP) disponible para el proyecto en relación al genoma humano completo por cada cromosoma.

3.2.1. Preprocesamiento

Como se describe en la **Sección 2.3.2**, los datos imputados obtenidos corresponden a $7.4 \cdot 10^6$. Esto es inmanejable por los actuales modelos de secuencia utilizando GPUs de máximo 24 GB. Por ello, se extraen los SNPs comunes a todas las genotipificaciones desde las variantes imputadas. Exactamente, la cantidad de SNPs comunes obtenidos desde los diferentes centros

corresponde a 438307 los que se distribuyen a lo largo de 22 cromosomas autosómicos. Para mostrar esta distribución, en la **Figura 3.1**, se muestra la ubicación de estos cuatrocientos mil SNPs en el genoma. Los espacios vacíos son zonas más díficiles de genotipificar, como los centrómeros. Otras zonas díficiles de secuenciar son los telómeros, que se encuentran en los bordes, pero que no se muestran en la figura, ya que se muestra cada cromosoma desde la posición mínima hasta la máxima. Recordar que aunque no todos los SNPs estén en el mismo cromosomas sí puede haber influencia entre ellos (**Figura 2.2-2.3**).

Debido a que los datos provienen de distintos centros, los SNP reportados no son los mismos. Esta es una de las justificaciones para imputar otros SNP a través de la información entregada en la genotipificación [55] (**Sección 2.3.2**). Otra justificación, es que las variantes obtenidas podrían no estar asociadas con el fenotipo en estudio. Esto tiene sentido si consideramos que la diferencia entre individuos es de $\sim 0.11\,\%$ [56] de un total de $3.2 \cdot 10^9$ nucleótidos [20], siendo $\sim 3.52 \cdot 10^6$ nucleótidos diferentes por participante, una muestra de $5 \cdot 10^5$ es solo el $14.2\,\%$.

En caso que las variantes utilizadas no sean significativas, se recuperarán otros datos desde el *set* imputado, además de los SNPs comunes descritos anteriormente. Las aproximaciones que se utilizan en este trabajo son (1) usar solo un cromosoma de los 22, éste puede ser seleccionado por significancia conocida, por la iniciativa internacional COVIDhg o por análisis de significancia (**Sección 3.2.2**) con los mismos datos en estudio; y (2) extraer las variantes del set imputado que se consideran "sugerentes" de significancia, es decir, que $p < 10^{-5}$ (**Sección 3.2.2**).

3.2.2. Determinación de significancia

La aproximación actual para determinar significancia estadística entre variantes genéticas y un fenotipo, es realizar un GWAS (**Sección 2.1.2.3**) [55]. El fenotipo de un individuo puede estar determiando por su genotipo o por factores ambientales, en el caso de enfermedades humanas, estos factores pueden ser sexo o edad. Los GWAS realizados para cada una de las aproximaciones para seleccionar variantes son:

1. Valor p del coeficiente β de cada variante, utilizando solo las variantes genéticas.
2. Valor p del coeficiente β de cada variante, utilizando las variantes genéticas, la edad y el sexo.
3. Valor p del coeficiente β de cada variante, utilizando las variantes genéticas y los datos clínicos más significativos.
4. Utilizando la prueba estadística de Wald solo para las variantes genéticas. Esto se utiliza para comparar con los mapas de saliencia según el método descrito en [38].

Para decir que una variantes es significante bajo la tasa de falsos positivos de un $5\,\%$, hay que considerar que se realizan múltiples pruebas estadísticas independientes, se asume un

millón para el genoma humano. Para esto, se realiza la corrección de Bonferroni bajo estos supuestos [55]:

$$p < \frac{\alpha}{n} = \frac{0.05}{10^6} \Rightarrow p < 5 \cdot 10^{-8}$$

3.3. Modelos

En esta sección se presentan los modelos utilizados, y cómo se codifican los datos para ser procesados por estos. Los modelos elegidos buscan maximizar las métricas de predicción, además de poder extraer las variables que más se correlacionan con la severidad (Sección 2.1.4).

3.3.1. Modelos de aprendizaje de máquina

Para establecer un punto de comparación para los modelos de Aprendizaje Profundo, se aplicarán modelos de *Machine Learning* (ML) sobre los datos. Estos algoritmos de aprendizaje ya se utilizan tanto en datos autoreportados como en datos genéticos [27] y tiene la ventaja de ser interpretables según el enfoque estudiado por la tesis de magíster de Bernardo Subercaseax [54]. Ya sea por la generación de un árbol de decisión, o mediante los coeficientes de los modelos de clasificación basados en regresiones.

Los modelos de ML a utilizar serán:

1. *Dummy Mode Classifier*: El primer *baseline* es el modelo por probabilidad. Por lo general, se utiliza uno al azar, sin embargo, para este trabajo, se usa uno que siempre clasifica según moda, ya que es un mejor punto de comparación dado el desbalance de los datos.
2. Árbol de Decisión (*Decision Tree*): Ya que se utilizarán *ensambles* de este modelo, Random Forest y XGBoost.
3. Random Forest: Se utiliza, ya que se han obtenido buenas métricas con este algoritmo en problemas similares [27].
4. Regresión Logística: Como se menciona en **Sección 2.1.2.3** es base para el cálculo de significancia en GWAS.
5. XGBoost: Ha dado bueno resultados en problemas similares, inclusive en Covid-19 [8, 37].

Se utiliza la librería Sci-kit learn [43], la que también se utiliza para *Cross validation* y obtención de métricas. Aprovechando que los modelos ya se encuentran implementados, se agregan otros modelos de ML amplicamente usados como SVM, Bayes Ingenuo y Clasificador Gaussiano, en caso de obtener mejores métricas. Como *baseline* para comparar las ANN en datos clínicos, se utilizará el modelo que obtenga mejores métricas utilizando 5-*fold Cross Validation*.

Como estos datos son más fácilmente interpretables [54], se realiza una selección de variables de los datos clínicos utilizando algunos de estos modelos. Estos datos seleccionados se utilizarán como punto de comparación con los modelos que utilicen tanto datos clínicos como genéticos.

3.3.2. Redes Neuronales

Si bien, las redes neuronales son menos interpretables que los algoritmos basados en árboles de decisión o que las regresiones [54], estos obtienen mejores métricas en la clasificación de datos complejos [21]. En esta sección se describen las ecuaciones y modelos que se utilizarán en este trabajo.

Debido a que los modelos de Aprendizaje Profundo toman más tiempo de entrenamiento, solo se utilizará *Holdout Validation* para evaluar los modelos con 70 % entrenamiento, 15 % validación y 15 % de *testing*. Se utilizará optimización de Adam para todos los modelos. La función de pérdida (*loss*) será *Cross Entropy* (**Sección 3.3.2**).

Para los modelos FNN se utiliza *learning rate* (lr) de 10^{-4} y *weight decay* (wd) de 0.0 de forma arbitraria. Para los modelos sobre datos genéticos se realiza una experimentación preliminar para obtener los hiperparámetros. Se realiza una validación *Holdout* para cada $lr = 10^i$ para $i \in \{-10, -9, \cdots, -1\}$. Para las lr seleccionadas con menor sobreajuste se realiza la selección de wd sobre $wd = 10^i$ para $i \in \{-10, -9, \cdots, -1\}$.

Para evitar el sobreajuste sobre los datos de entrenamiento, se utilizara una modificación de *Early Stopping* sobre el *loss* de validación con paciencia (*patience*) de 5. Esta modificación consiste en guardar el modelo cada vez que la *loss* de validación suba o se mantenga y continuar el experimento. Si la *loss* de validación baja más allá de la versión guardad, se vuelve a guardar el modelo. Esto es, para evitar guardar un modelo con menor *performance* al principio del entrenamiento.

Una vez definidos los modelos con mejores métricas usando *holdout validation*, se realiza una nueva validación usando *k-Fold Cross Validation* ($k = 5$ para datos clínicos, $k = 10$ para calcular saliencia, **Sección 3.3.2.2**) la que se compara con los modelos de ML.

3.3.2.1. FNN, *Fully Connected Neural Networks*

Las arquitecturas *fully connected* (FNN) solo se usan en los datos clínicos, ya que en datos menos estructurados, como imágenes y texto, poseen menor capacidad que otras arquitecturas. Además de ser un buen punto de comparación, servirán para incluir los datos clínicos a las arquitecturas que utilizan variantes genéticas. A modo de prueba, se utilizan cuatro arquitecturas simples (1) con una capa escondida (dos capas) (**Figura 3.2**), (2) tres capas (**Figura 3.3**), (3) cinco capas (**Figura 3.4**) y (4) nueves capas (**Figura 3.5**). Esto servirá para definir una arquitectura que tenga la capacidad suficiente y que no se sobreajuste a los datos de entrenamiento. De forma arbitraría se utiliza como función de activación $ReLu$, exceptuando la segunda capa de la arquitectura de tres capas, donde se usa la función de tangente hiperbólica tanh.

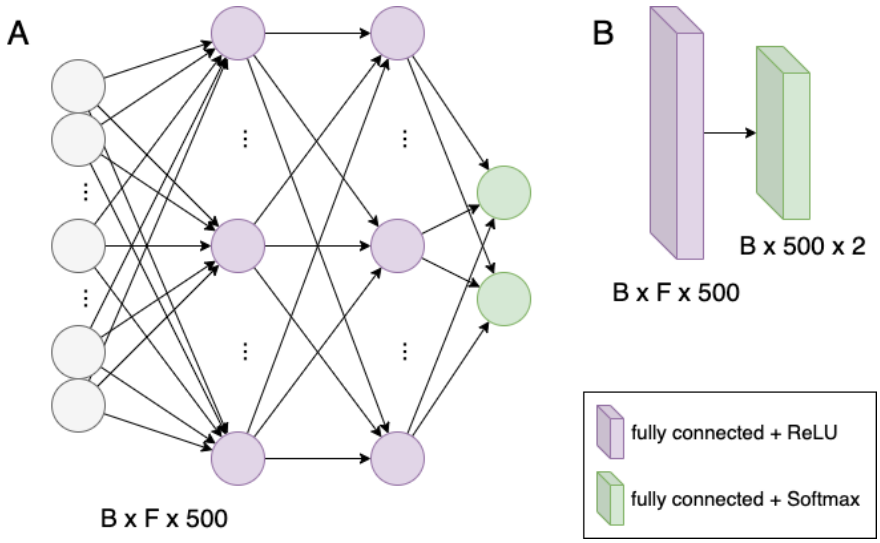


Figura 3.2: Esquema de la arquitectura de red neuronal *fully connected* de dos capas. **A** Se muestran las conexiones neurona a neurona. **B** Esquema simplificado mostrando solo las capas.

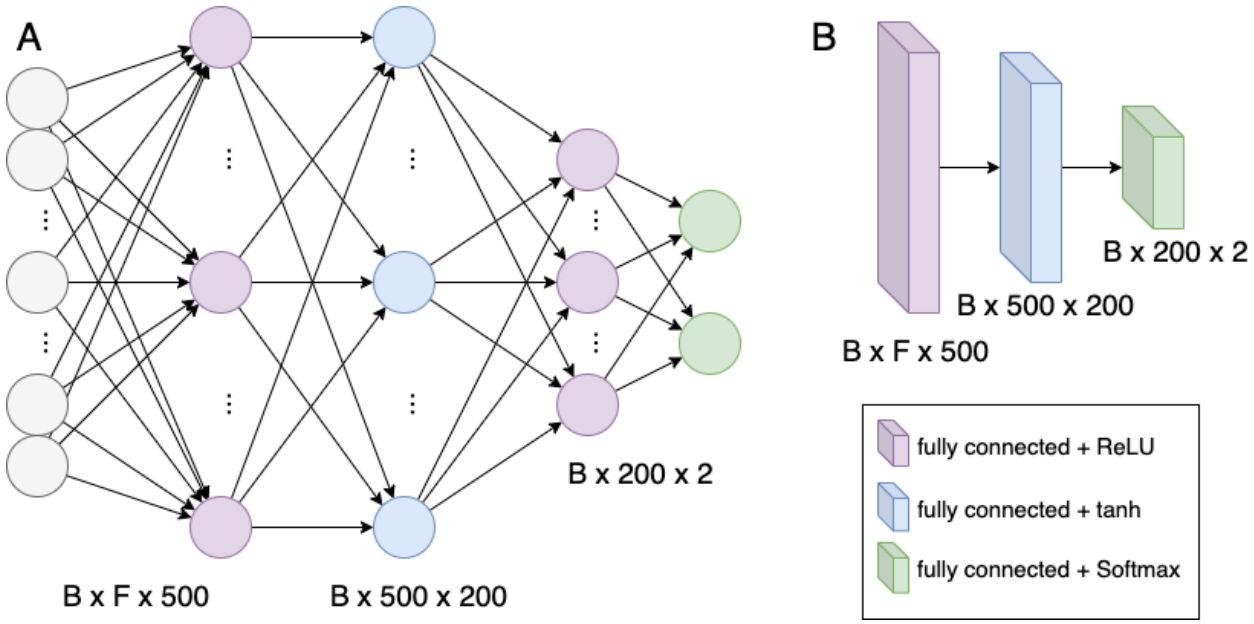


Figura 3.3: Esquema de la arquitectura de red neuronal *fully connected* de tres capas. **A** Se muestran las conexiones neurona a neurona. **B** Esquema simplificado mostrando solo las capas.

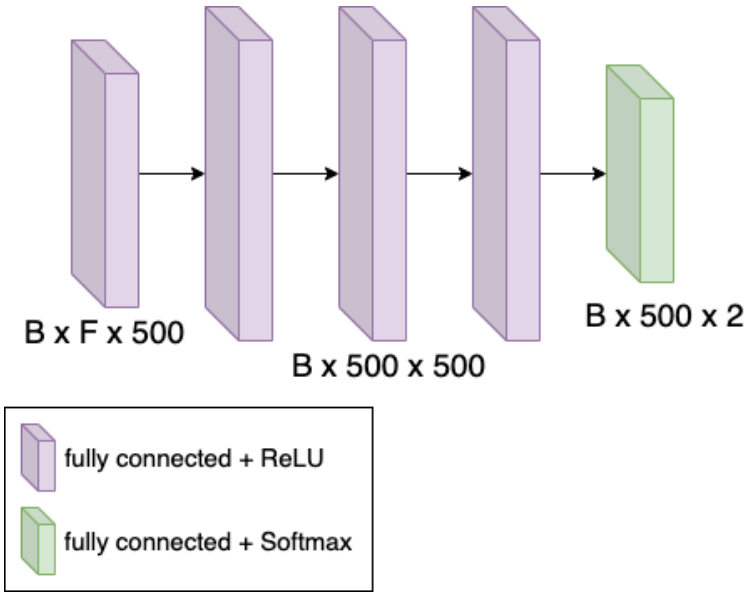


Figura 3.4: Esquema de la arquitectura de red neuronal *fully connected* de cinco capas. Para simplificar, solo se esquematizan las capas.

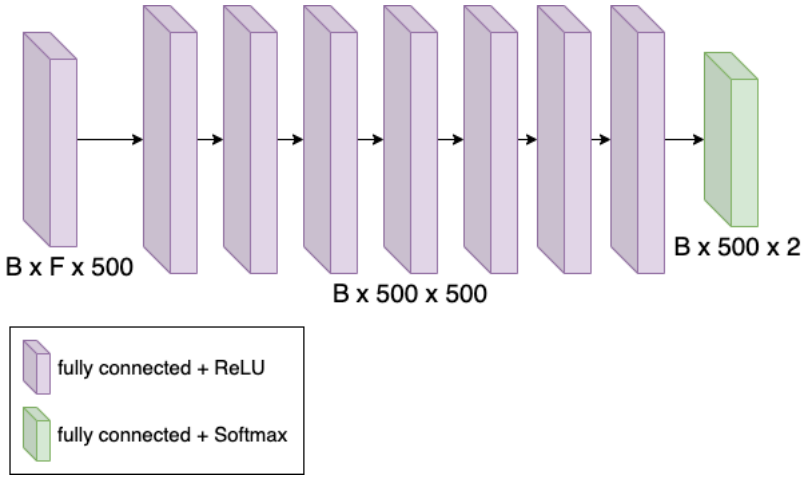


Figura 3.5: Esquema de la arquitectura de red neuronal *fully connected* de nueve capas. Para simplificar, solo se esquematizan las capas.

En las **Figuras 3.2-3.5** se muestran los esquemas de las arquitecturas usadas. Los esquemas simplificados, donde solo se muestran las capas, servirán para esquematizar las arquitecturas adaptadas que utilizan como entrada datos clínicos y genéticos.

3.3.2.2. CNN, *Convolutional Neural Networks*

Las redes convolucionales se utilizan actualmente para el procesamiento de imagenes [21, 36, 49], o para secuencias, como datos temporales [21, 29]. Hay varios ejemplos del uso de estas arquitecturas en secuencias de datos genéticos [1, 23, 33, 34, 38, 68]. Particularmente, la arquitectura desarrollada por Liu y colaboradores [38] se ha utilizado para datos genéticos a nivel de genoma en la haba de soya. Esta arquitectura no solo ha obtenido buenas métricas,

si no que también ha podido capturar la correlación entre las variantes genéticas y el fenotipo predicho, de forma equivalente a un GWAS.

Para describir esta arquitectura, llamada *Dual-stream CNN*, primero recordamos la definición de una capa convolucional [21]:

$$h^{(i)} = \sum_{k=0}^{C_{in}-1} {W^{(i)}}^T \star h^{(i-1)} + b^{(i)}$$

Mantenemos la entrada como x como $h^{(0)}$. *Dual-stream CNN* se compone de dos capas convolucionales apiladas y una tercera que se concatena con su resultado:

$$h^{(1)} = \sum_{k=0}^{C_{in}-1} {W^{(1)}}^T \star x + b^{(1)}$$

$$h^{(2)} = \sum_{k=0}^{10-1} {W^{(2)}}^T \star h^{(1)} + b^{(2)}$$

$$h^{(3)} = \sum_{k=0}^{C_{in}-1} {W^{(3)}}^T \star x + b^{(3)}$$

$$h^{(4)} = Join(h^{(2)}, h^{(3)})$$

$$h^{(5)} = \sum_{k=0}^{10-1} {W^{(5)}}^T \star h^{(4)} + b^{(5)}$$

$$\hat{y} = U^T h^{(5)} + b^{(6)}$$

El esquema de esta arquitectura se muestra en la **Figura 3.6**.

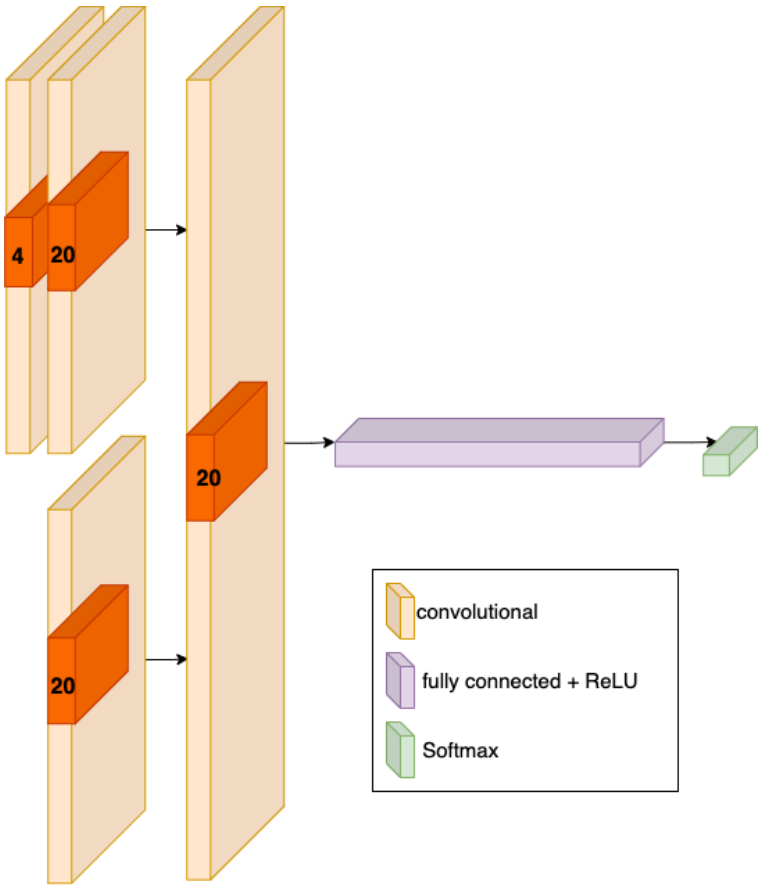


Figura 3.6: Esquema de la arquitectura de red convolucional dual presentada en [38]. La arquitectura corresponde a dos capas convolucionales apiladas y una tercera, cuya salida se concatena y se pasa por una última capa convolucional para obtener los *features* utilizados para la clasificación realizada por una última capa FNN.

3.3.2.3. Modelos adaptados

Para comparar la influencia de los datos genéticos en la predicción de la severidad, utilizaremos un modelo capaz de recibir ambos tipos de datos. Este será una modificación de la arquitectura *Dual-stream CNN*, acoplada al modelo FNN que mejores métricas obtenga con los datos clínicos seleccionados,. La arquitectura *Dual-stream CNN* extendida se muestra en la **Figura 3.7**.

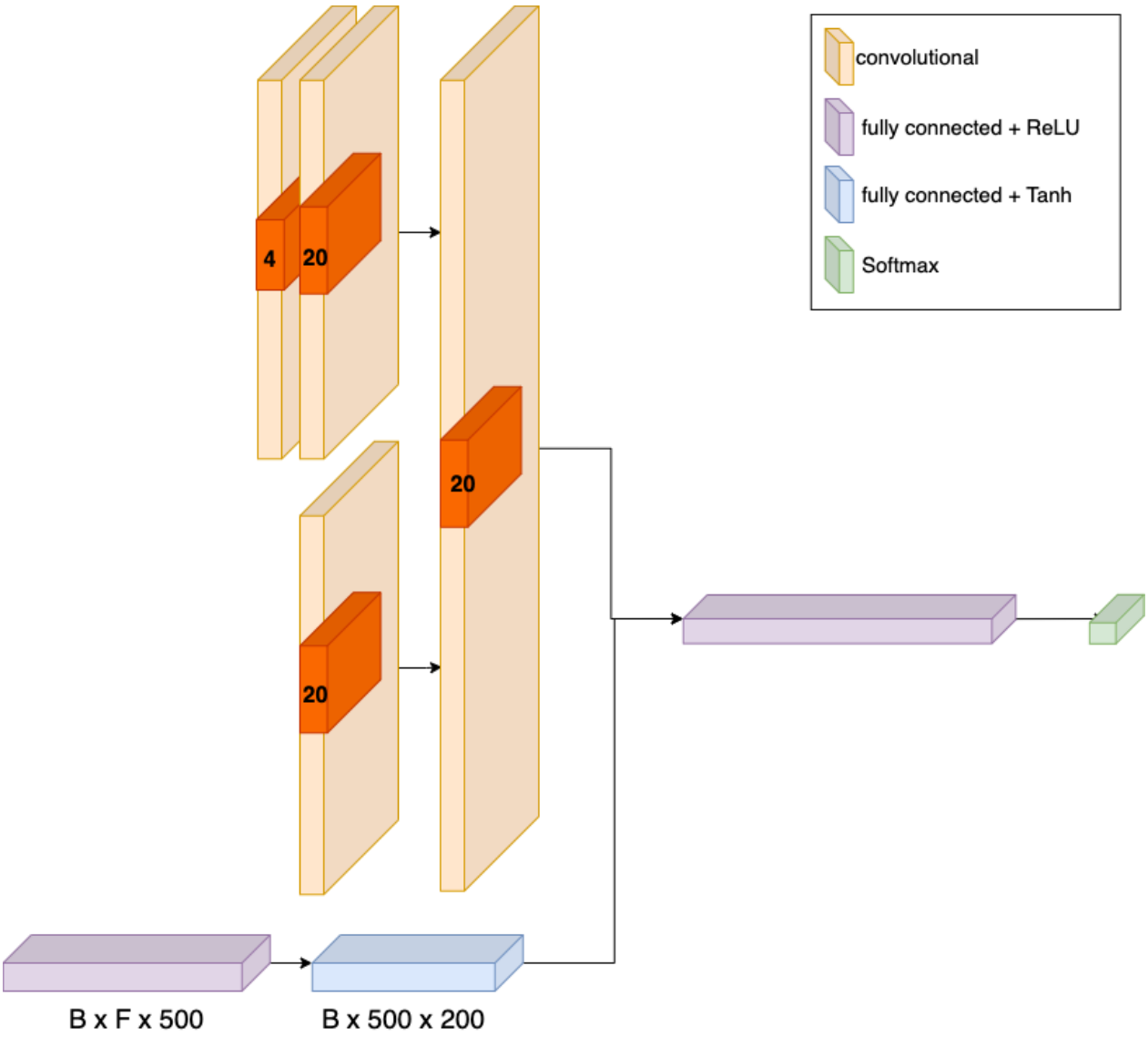


Figura 3.7: Esquema de la arquitectura de red convolucional dual presentada en [38] concatenada con una arquitectura FNN que procesa los datos clínicos. La arquitectura corresponde a la presentada en **Figura 3.6**, la que recibe los datos genéticos en formato SNP. La salida de la última capa convolucional se concatena con la salida de la penúltima capa de la red FNN, para ser pasados por una última capa que realiza la clasificación.

Para describir las ecuaciones de la arquitectura *Dual-stream CNN* extendida, tomamos las últimas capas de cada arquitectura, *Dual-stream CNN* y FNN, como:

$$\text{Última capa Soybean: } h^{(5)} = \sum_{k=0}^{10-1} {W^{(5)}}^T \star h^{(4)} + b^{(5)}$$

$$\text{Última capa FNN: } h^{(k)} = g^{(k)} \left({W^{(k)}}^T h^{(k-1)} + b^{(k)} \right)$$

La predicción final se calcula:

$$\hat{y} = U^T Join\left(h^{(5)}, h^{(k)}\right) + b^{(6)}$$

3.3.3. Métricas

Para comparar la certeza en las predicciones de los modelos, se utilizarán las métricas estándar para modelos de Aprendizaje de Máquinas. Estas métricas se basan en una tabla de contingencia que compara las etiquetas a predecir con las predicciones del modelo.

$$\text{Matriz de confusión} = \tag{3.3}$$

		Predicho	
		Control (Negativo)	Caso (Positivo)
Verdadero	Control (Negativo)	True Negative (TN)	False Positive (FP)
	Caso (Positivo)	False Negative (FN)	True Positive (TP)

$$accuracy = \frac{TP + TN}{TP + FN + FP + TN} = \frac{TP + TN}{P + N} \tag{3.4}$$

$$precision = \frac{TP}{TP + FP} = \frac{TP}{\text{Predicho positivo } (PP)} \tag{3.5}$$

$$recall = \frac{TP}{TP + FN} = \frac{TP}{P} \tag{3.6}$$

$$f1\text{-}score = \frac{2TP}{2TP + FP + FN} = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{3.7}$$

Para comparar se utilizará como métrica $f1$-*score*, ya que el *accuracy* podría llevar a conclusiones érrores dado el desbalance de los datos. Por ejemplo, si tuvieramos un *set* de datos con 1556 controles y 127 casos. El modelo *Dummy* dirá que todos son controles, por lo que tendrá un *accuracy*:

$$accuracy_{Dummy} = \frac{0 + 1556}{0 + 127 + 0 + 1556} = 0.92$$

Un modelo que prediga 500 casos, con todos los casos dentro de estos 500, obtendrá:

$$accuracy = \frac{127 + 1183}{127 + 0 + 373 + 1183} = 0.77$$

Si comparamos $f1$-*score*:

$$f1\text{-}score_{Dummy} = \frac{2 \cdot 0}{2 \cdot 0 + 0 + 127} = 0.00 < f1\text{-}score = \frac{2 \cdot 127}{2 \cdot 127 + 373 + 0} = 0.41$$

Como definición para *precision*, en casos que no haya predichos positivos, se asigna un valor de 0.0.

Además de estás métricas, también se utilizará la curva de ROC (Característica Operativa del Receptor, por sus siglas en inglés *Receiver Operating Characteristic*), que grafica la razón de verdaderos positivos (TPR, por sus siglas en inglés *True Positive Rate*) en función de la razón de falsos positivos (FPR, *False Positive Rate*). El TPR es el *recall* (**Ecuación 3.6**), el FPR se calcula:

$$FPR = \frac{FP}{FP + VN} = \frac{FP}{N} \tag{3.8}$$

En caso de utilizar *holdout validation* las métricas se calculan directamente con el *set* de *testing*. Si se utiliza *Cross Validation*, se obtiene la matriz de confusión con todos los *fold* de *testing* comparado con el respectivo modelo entrenado, la curva de ROC se calculará por *fold* y utilizando este método.

Capítulo 4

Resultados

En esta sección se presentan los principales resultados obtenidos, lo que incluye las métricas de los diferentes modelos para los diferentes *set* de datos usados en este trabajo.

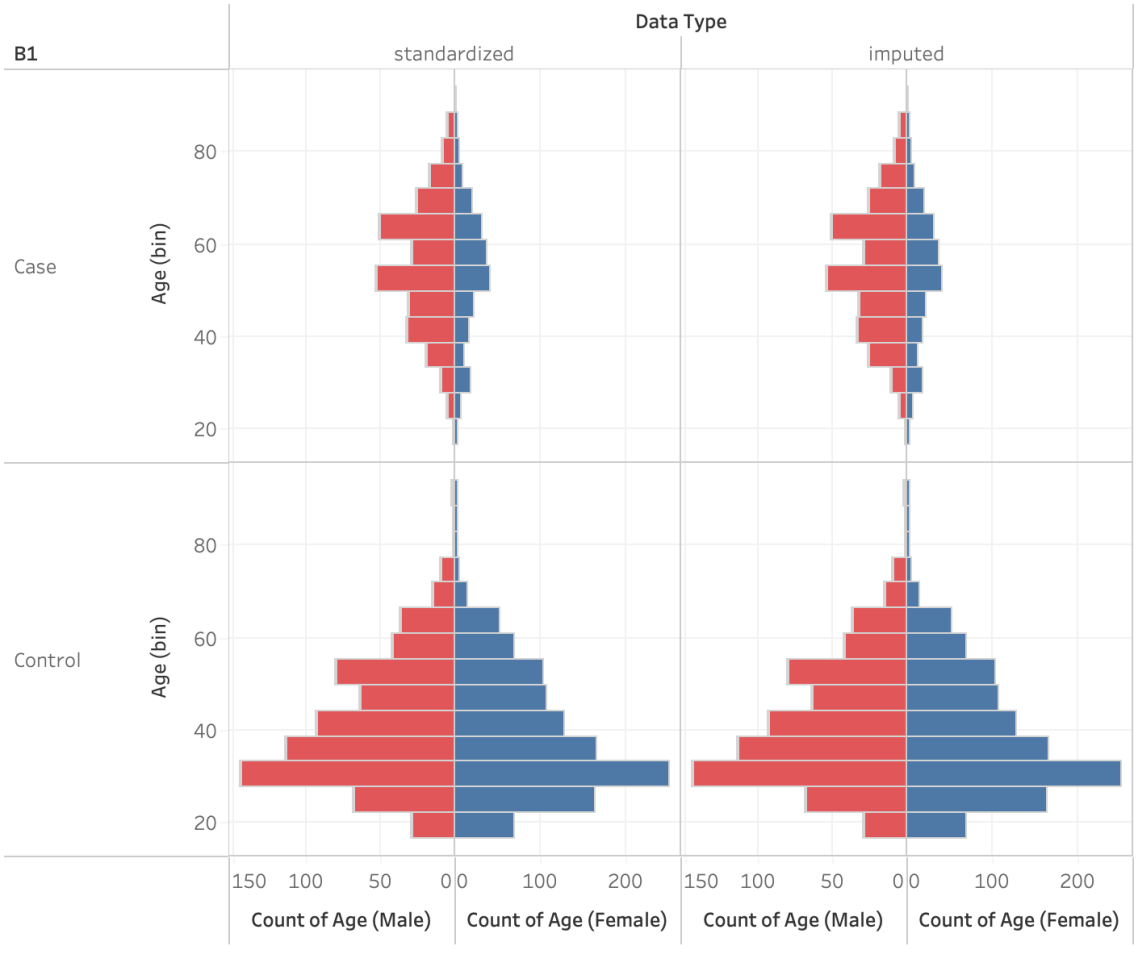


Figura 4.1: Gráficos de pirámide de las edades separadas por caso y control para el análisis de hospitalizados. Se muestra la comparación antes (*standarized*) y después (*imputed*) de la imputación.

4.1. Imputación y Análisis Estadístico

Para comprobar que la imputación no afecta de forma considerable la distribución de los datos clínicos, se comparan los valores de algunos atributos antes y después de la imputación. Las columnas que se comparan son aquellas que resultan más fáciles de visualizar como edad, IMC, densidad poblacional y grupo sanguíneo.

En cuando a la edad, la cantidad de datos imputados corresponde a 0.43 % (8 datos, **Tabla 3.3**), por lo que se espera que la distribución no cambie considerablemente. Esto muestra en la **Figura 4.1**, donde se comparan los gráficos de pirámide de las diferentes edades separadas por caso y control para datos previos (*standarized*) y luego de la imputación (*imputed*). En esta figura también se evidencia que los casos están en un rango etario mayor.

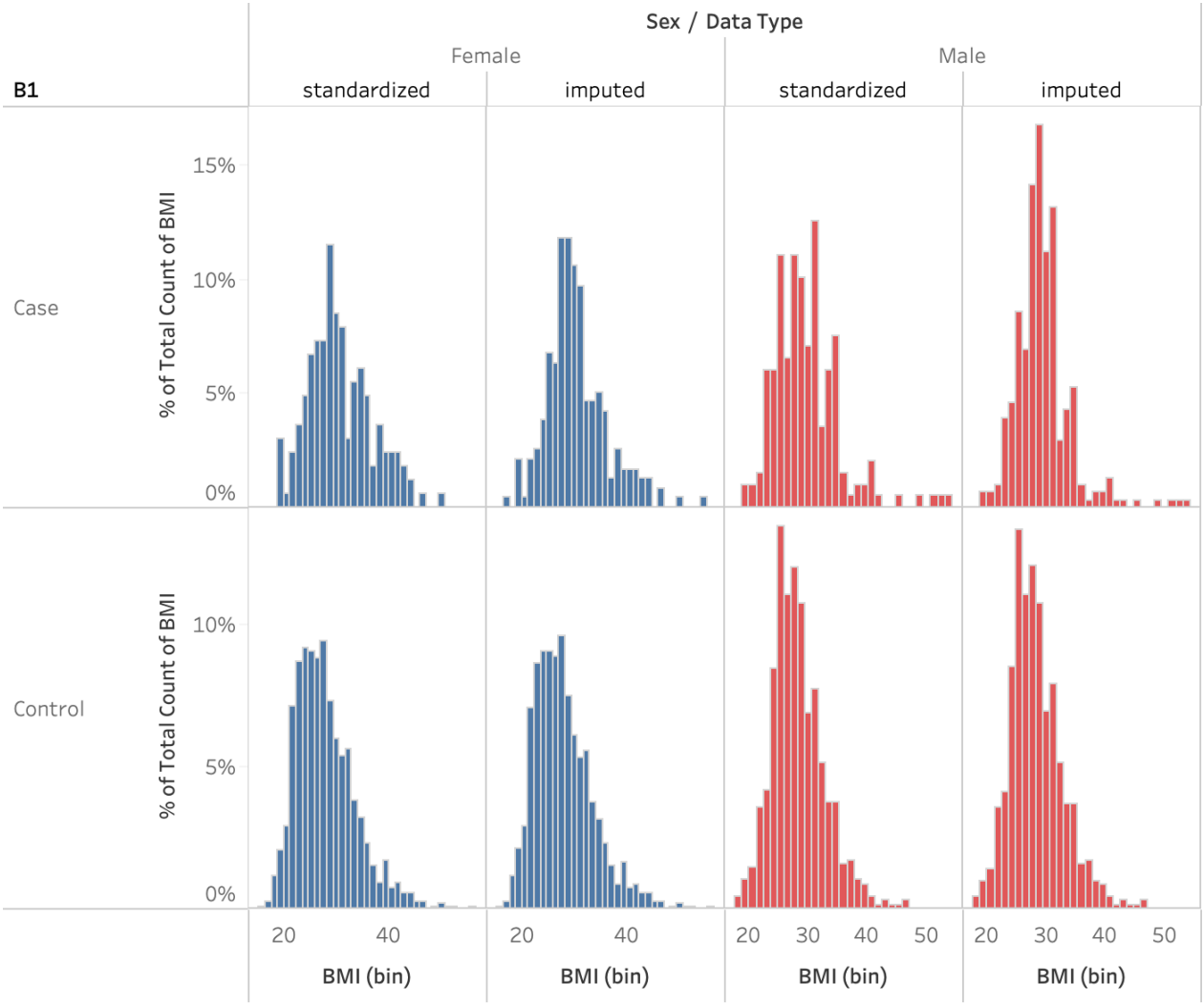


Figura 4.2: Distribución de IMC separada por sexo reportado y severidad de hospitalizados, casos y controles. Se muestra la comparación antes (*standarized*) y después (*imputed*) de la imputación.

El IMC de cada participante se imputa en un 6.94 % (130 participantes, **Tabla 3.3**). Como

se muestra en la **Figura 4.2** en la parte superior, los cambios en la distribución son mayores para los casos, lo que se espera considerando que para estos participantes se imputan 36.73 % (119 de 130 casos, **Tabla 3.3**). Sin embargo, la moda se mantiene aunque se exacerba.

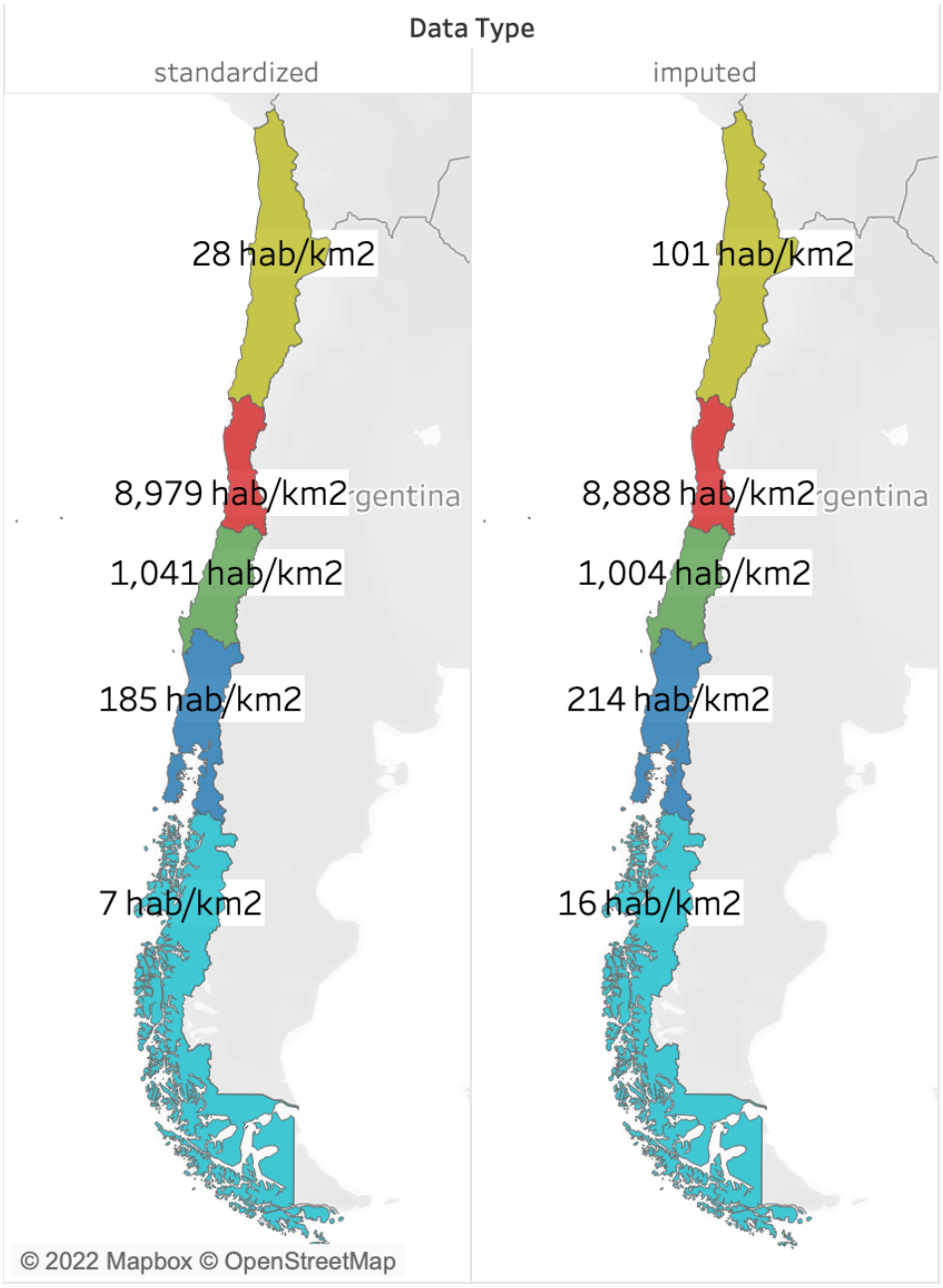


Figura 4.3: Densidad poblacional promedio por macrozona antes (*standardized*) y después (*imputed*) de la imputación.

La densidad poblacional corresponde a la cantidad de habitantes por kilómetro cuadrado de la comuna de residencia reportada por los participantes. Los datos que se imputan corresponden al 0.85 % (16 participantes), los que en su mayoría son casos (4.63 % de estos, correspondientes a 15 participantes, **Tabla 3.3**). Para mostrar si la densidad poblacional promedio por macrozona se altera en la imputación se generan los mapas de la **Figura 4.3**,

donde se evidencian mayores cambios solo en la macrozona austral y norte. Para comparar si se alteran las distribuciones por casos y controles, se muestran las densidades poblacionales promedio por macrozona en la **Figura 4.4**, donde se evidencia que el cambio, por macrozona, es mayor para los casos.

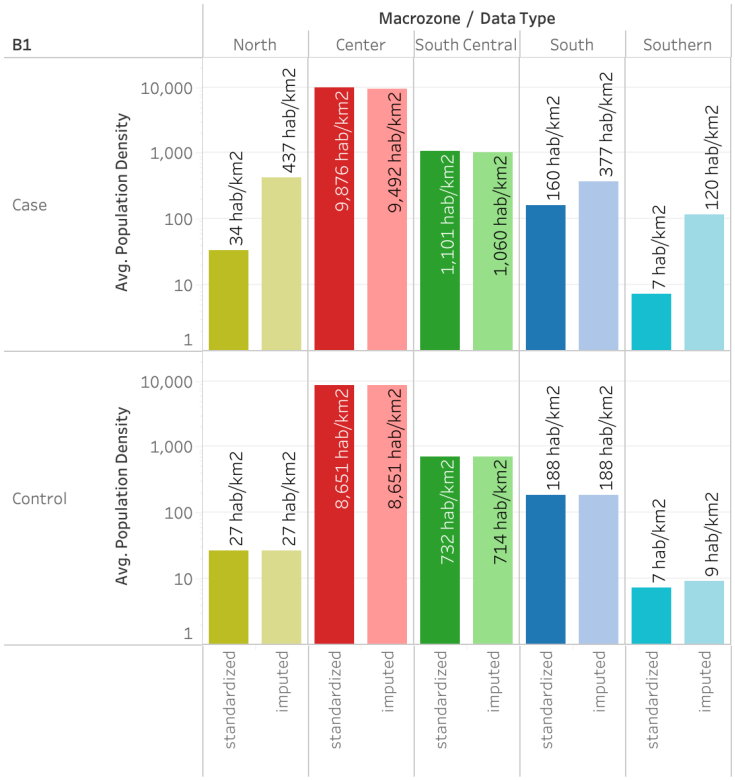


Figura 4.4: Densidad poblacional promedio por macrozona. Se compara los valores antes de la imputación (*standardized*), en oscuro y después (*imputed*) en claro.

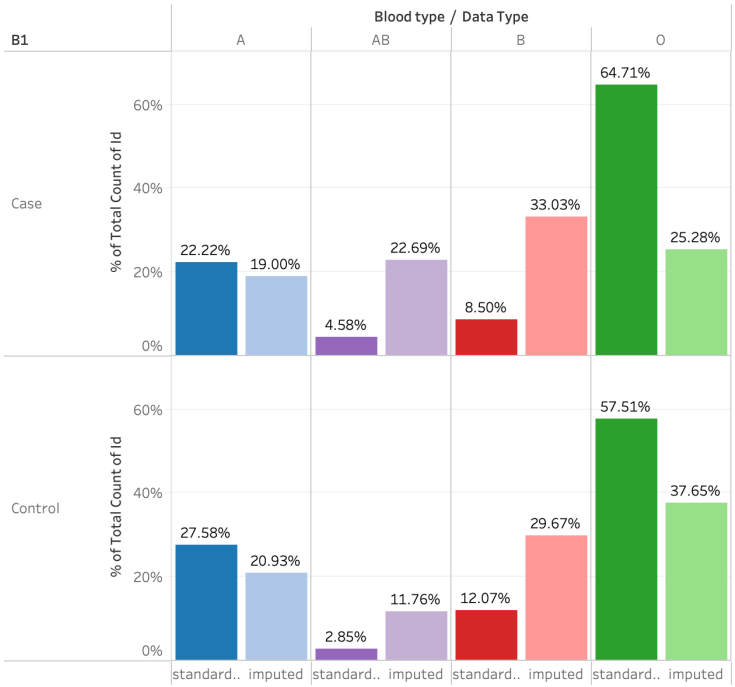


Figura 4.5: Porcentaje de participantes por grupo sanguineo en referencia al total por severidad. Se comparan los datos antes de la imputación (*standardized*) en colores oscuros y después (*imputed*) en colores claros.

El grupo sanguíneo es la variable que más datos faltantes tiene, con un 48.77 % correspondiente a 913 participantes. La distribución de los datos antes y después de la imputación se evidencia en la **Figura 4.5**, donde se aprecia también que no hay mayor diferencia entre los cambios de distribución del grupo sanguíneo entre casos y controles.

Tabla 4.1: Resumen de las variables significativas para análisis de hospitalizados entre población infectada, según prueba estadísitica univariada. Se comparan los resultados entre los datos estandarizados y los datos imputados por KNN.

Standardized					Imputed				
Numerical variables									
Variable	$\bar{x}$ (Cases)	$\bar{x}$ (Controls)	t	p	Variable	$\bar{x}$ (Cases)	$\bar{x}$ (Controls)	t	p
Age	54.46	39.70	-17.48	2.39e-52	Age	54.06	39.70	-17.09	7.28e-51
Population Density	4051.47	2098.47	-5.72	2.09e-08	Population Density	3942.65	2097.72	-5.60	3.90e-08
Weight	82.97	77.03	-4.63	5.51e-06	Weight	82.56	77.02	-5.66	2.66e-08
BMI	30.41	28.23	-5.00	1.10e-06	BMI	30.23	28.23	-6.16	1.62e-09
Categorical variables									
Variable	df	χ^2	p		Variable	df	χ^2	p	
Province	1	16.05	6.17e-05		Province	1	18.42	1.78e-05	
Sex	1	30.22	3.85e-08		Sex	1	30.22	3.85e-08	
Blood type[9]	3	4.21	2.40e-01		Blood type	3	53.30	1.58e-11	
Rh type[9]	1	0.46	4.99e-01		Rh type	1	4.28	3.86e-02	
Country of birth	10	61.17	2.17e-09		Country of birth	10	60.01	3.61e-09	
Educational level	9	262.50	2.28e-51		Educational level	9	187.77	1.20e-35	
Occupational situation	2	20.31	3.88e-05		Occupational situation	2	17.72	1.42e-04	
Health System	1	62.59	2.55e-15		Health System	1	62.59	2.55e-15	
Tobacco Consumption	4	51.73	1.57e-10		Tobacco Consumption	4	42.00	1.67e-08	
Daily cigars	5	54.53	1.63e-10		Daily cigars	5	78.46	1.76e-15	
More than 100 cigars	1	16.55	4.74e-05		More than 100 cigars	1	14.59	1.33e-04	
Alcohol Consumption	4	119.65	6.34e-25		Alcohol Consumption	4	146.56	1.11e-30	
Marijuana	1	62.93	2.14e-15		Marijuana	1	62.93	2.14e-15	
Hallucinogens	1	4.16	4.14e-02		Hallucinogens	1	4.16	4.14e-02	
Diabetes	1	91.62	1.05e-21		Diabetes	1	91.41	1.17e-21	
Hypertension	1	152.61	4.65e-35		Hypertension	1	152.55	4.81e-35	
Vascular accident	1	5.65	1.75e-02		Vascular accident	1	5.65	1.75e-02	
Heart problem	1	25.10	5.44e-07		Heart problem	1	25.26	5.01e-07	
Dialysis	1	46.46	9.37e-12		Dialysis	1	46.44	9.47e-12	
Pulmonary condition	1	7.28	6.98e-03		Pulmonary condition	1	7.27	7.00e-03	
Cancer	1	4.33	3.75e-02		Cancer	1	4.32	3.76e-02	
High cholesterol	1	5.69	1.70e-02		High cholesterol	1	5.82	1.59e-02	
Obesity	1	30.34	3.62e-08		Obesity	1	30.63	3.13e-08	
Rheumatoid arthritis	1	3.84	5.01e-02		Rheumatoid arthritis	1	3.88	4.89e-02	
Prostate cancer	1	9.57	1.98e-03		Prostate cancer	1	9.57	1.98e-03	
Antidiabetic	1	32.69	1.08e-08		Antidiabetic	1	32.69	1.08e-08	
Antihypertensive	1	106.61	5.42e-25		Antihypertensive	1	106.61	5.42e-25	
Aspirin or clopidogrel	1	49.86	1.65e-12		Aspirin or clopidogrel	1	49.86	1.65e-12	
Statin	1	40.44	2.02e-10		Statin	1	40.44	2.02e-10	
Cough medication	1	6.43	1.12e-02		Cough medication	1	6.43	1.12e-02	
Inhaler	1	13.48	2.41e-04		Inhaler	1	13.48	2.41e-04	
Vitamins	1	14.61	1.32e-04		Vitamins	1	14.61	1.32e-04	
No drug	1	116.62	3.47e-27		No drug	1	116.62	3.47e-27	
Medication DK/NA/REF	1	4.33	3.75e-02		Medication DK/NA/REF	1	4.33	3.75e-02	

[9] Variable no significativa utilizando datos estandarizados, se muestra para comparar significancia obtenida por datos imputados

Tabla 4.2: Resumen de las variables significativas para análisis de hospitalizados entre infectados, según prueba estadística multivariada. Se utiliza un modelo de regresión logística. En el caso estandarizado el módelo base obtiene $Deviance = 322.38$ y $AIC = 534.38$. Para el caso imputado por KNN se tiene $Deviance = 981.70$ y $AIC = 1215.70$.

Standardized						Imputed					
Variable	df	Deviance	AIC	LRT	$Pr(>\chi)$	Variable	df	Deviance	AIC	LRT	$Pr(>\chi)$
Age	1	347.26	557.26	24.88	6.10e-07	Age	1	1028.23	1260.23	46.53	9.01e-12
Province	1	326.99	536.99	4.61	3.18e-02	Province[10]	1	981.97	1213.97	0.28	6.00e-01
Population Density[9]	1	323.12	533.12	0.74	3.91e-01	Population Density	1	986.92	1218.92	5.22	2.23e-02
Blood type[9]	3	327.73	533.73	5.35	1.48e-01	Blood type	3	993.05	1221.05	11.35	9.96e-03
Educational level[9]	8	329.86	525.86	7.47	4.87e-01	Educational level	9	1031.98	1247.98	50.29	9.51e-08
Occupational situation	2	334.18	542.18	11.79	2.75e-03	Occupational situation[10]	2	986.76	1216.76	5.06	7.95e-02
Health System[9]	1	326.14	536.14	3.76	5.26e-02	Health System	1	991.79	1223.79	10.10	1.49e-03
Tobacco Consumption	3	344.52	550.52	22.14	6.10e-05	Tobacco Consumption	4	998.65	1224.65	16.96	1.97e-03
Daily cigars[9]	4	324.44	528.44	2.06	7.25e-01	Daily cigars	5	996.12	1220.12	14.43	1.31e-02
Alcohol Consumption	4	333.31	537.31	10.93	2.74e-02	Alcohol Consumption[10]	4	990.88	1216.88	9.19	5.66e-02
Hypertension	1	330.71	540.71	8.33	3.90e-03	Hypertension[10]	1	981.78	1213.78	0.08	7.75e-01
Hepatitis	1	326.95	536.95	4.57	3.26e-02	Hepatitis[10]	1	982.41	1214.41	0.71	3.98e-01
Ulcerative colitis or Crohn's disease	1	327.09	537.09	4.71	3.00e-02	Ulcerative colitis or Crohn's disease[10]	1	983.88	1215.88	2.19	1.39e-01
Other rheumatoid diseases[9]	1	325.89	535.89	3.50	6.12e-02	Other rheumatoid diseases	1	986.00	1218.00	4.30	3.81e-02
Prostate cancer[9]	1	322.38	532.38	0.00	1.00e+00	Prostate cancer	1	986.20	1218.20	4.50	3.38e-02
Antidiabetic[9]	1	322.88	532.88	0.50	4.81e-01	Antidiabetic	1	987.44	1219.44	5.74	1.65e-02
Antihypertensive	1	329.77	539.77	7.39	6.57e-03	Antihypertensive[10]	1	983.57	1215.57	1.88	1.71e-01
Statin	1	327.56	537.56	5.17	2.29e-02	Statin[10]	1	984.14	1216.14	2.45	1.18e-01
Cough medication[9]	1	322.95	532.95	0.57	4.50e-01	Cough medication	1	989.11	1221.11	7.41	6.47e-03
Vitamins[9]	1	324.78	534.78	2.39	1.22e-01	Vitamins	1	1002.22	1234.22	20.53	5.88e-06
No drug[9]	1	322.43	532.43	0.05	8.30e-01	No drug	1	1032.97	1264.97	51.27	8.04e-13
Another medication[9]	1	323.55	533.55	1.17	2.79e-01	Another medication	1	1001.50	1233.50	19.80	8.58e-06
Medication DK/NA/REF[9]	1	322.99	532.99	0.60	4.37e-01	Medication DK/NA/REF	1	1000.93	1232.93	19.23	1.16e-05

[10] Variable no significativa utilizando datos imputados, se muestra para comparar significancia obtenida por datos estandarizados

4.1.1. Análisis estadísticos

Los análisis estadísticos se realizan sobre los datos estandarizados e imputados. El detalle de los análisis se incluye en el **Anexo B.2**. Aquí se encuentra un resumen de las variables obtenidas (**Anexo B.2.1**), el análisis univariado (**Anexo B.2.2**) y multivariado (**Anexo B.2.3**).

En cuanto al análisis univariado los resultados para la prueba estadísitica de *t-student* se muestran en la **Tabla B.24** y para la prueba estadísitica de χ^2 en la **Tablas B.25-B.27**. Se asume que una variable es estadísticamente significativa para estas pruebas, si el valor $p < 5 \cdot 10^{-2}$. Las variables significativas obtenidas para estas pruebas se muestran en la **Tabla 4.1**.

El resultado del análisis multivariado se muestra en las **Tablas B.28-B.31**. Siguiendo el criterio $p < 5 \cdot 10^{-2}$, las variables significativas según este análisis se muestran en la **Tabla 4.2**. La nacionalidad no se encuentra en el análisis utilizando datos previos a la imputación, no porque no existan la categoría, sino que este analisis utiliza datos completos, lo que deja solo los participantes con la misma nacionalidad.

4.2. Selección de variables clínicas

Las primeras aproximaciones a la significancia de las variables clínicas se realizan con las pruebas estadísticas (**Sección 4.1.1**). La selección de variables se realiza por dos razones principalmente: (1) la menor cantidad de variables reduce el sobreajuste en los modelos y (2) permite interpretar la clasifición realizada por éstos. Como se explica en la **Sección 2.1.4.1**, otra forma de seleccionar variables es utilizado ciertos modelos de ML. Junto con entrenar los modelos utilizando todos los datos imputados, se extrae las variables seleccionadas de aquellos modelos que soporten este análisis, estos son: árboles de decisión, *Random Forest*, XGBoost, regresión logística y SVM (*Support Vector Machine*).

Tabla 4.3: Selección de variables clínicas según diferentes análisis realizados. Se muestra con un ticket verde (✅) aquellos que fueron seleccionados y con una equis roja (❌) aquellos que no. Se muestran solo las variables que fueron seleccionadas por algún análisis. El detalle de la selección se presenta en el texto.

	Statistical Test		Model Selection				
Variable	Univariate	Multivariate	Random Forest	Logistic Regression	Decision Tree	XGBoost	SVM
Age	✅	✅	✅	✅	✅	✅	✅
Population Density	✅	✅	✅	✅	✅	✅	✅
Weight	✅	❌	✅	❌	✅	❌	❌
BMI	✅	❌	✅	✅	✅	✅	❌
Province	✅	❌	✅	✅	✅	✅	❌
Sex	✅	❌	✅	✅	❌	❌	❌
Blood type	✅	✅	✅	✅	❌	✅	❌
Rh type	✅	❌	❌	✅	❌	❌	❌
Country of birth	✅	❌	❌	✅	✅	✅	✅
Educational level	✅	✅	✅	✅	✅	✅	✅
Occupational situation	✅	❌	❌	✅	❌	❌	❌
Health System	✅	✅	✅	✅	❌	✅	❌
Tobacco Consumption	✅	✅	✅	✅	❌	✅	❌
Daily cigars	✅	✅	✅	✅	❌	✅	✅
More than 100 cigars	✅	❌	✅	✅	❌	❌	❌
Alcohol Consumption	✅	❌	✅	✅	✅	✅	❌
Marijuana	✅	❌	✅	✅	✅	✅	❌
Hallucinogens	✅	❌	❌	❌	❌	❌	❌
Diabetes	✅	❌	✅	✅	❌	✅	❌
Hypertension	✅	❌	✅	❌	❌	❌	❌
Vascular accident	✅	❌	❌	❌	❌	❌	❌
Heart problem	✅	❌	❌	❌	❌	❌	❌
Dialysis	✅	❌	❌	✅	❌	❌	❌
Pulmonary condition	✅	❌	❌	❌	❌	❌	❌
Cancer	✅	❌	❌	❌	❌	❌	✅
High cholesterol	✅	❌	❌	✅	❌	✅	❌
Obesity	✅	❌	✅	✅	❌	✅	❌
Rheumatoid arthritis	✅	❌	❌	❌	❌	❌	❌
Prostate cancer	✅	✅	❌	❌	❌	❌	✅
Antidiabetic	✅	✅	❌	✅	❌	✅	❌
Antihypertensive	✅	❌	✅	✅	❌	❌	❌
Aspirin or clopidogrel	✅	❌	❌	❌	❌	❌	❌
Statin	✅	❌	❌	✅	❌	❌	❌
Cough medication	✅	✅	❌	✅	❌	❌	✅
Inhaler	✅	❌	❌	❌	❌	❌	❌
Vitamins	✅	✅	❌	✅	❌	✅	✅
No drug	✅	✅	✅	✅	❌	✅	✅
Medication DK/NA/REF	✅	✅	❌	✅	❌	✅	✅
Other rheumatoid diseases	❌	✅	❌	✅	❌	❌	✅
Another medication	❌	✅	✅	✅	❌	✅	✅
Other chronic diseases	❌	❌	✅	✅	❌	❌	❌
Height	❌	❌	✅	❌	✅	❌	❌
Mental health problems	❌	❌	❌	✅	❌	❌	❌
Asthma	❌	❌	❌	✅	❌	❌	❌
Ethnicity	❌	❌	❌	✅	❌	❌	✅
Thyroid cancer	❌	❌	❌	✅	❌	❌	✅
Medicinal herbs	❌	❌	❌	✅	❌	❌	❌
Hepatitis	❌	❌	❌	✅	❌	❌	❌
Coronary atherosclerosis	❌	❌	❌	✅	❌	❌	❌
Ulcerative colitis or Crohn's disease	❌	❌	❌	✅	❌	❌	✅
Cocaine	❌	❌	❌	❌	❌	✅	❌
Anemia	❌	❌	❌	❌	❌	✅	❌
Breast cancer	❌	❌	❌	❌	❌	❌	✅
Tuberculosis	❌	❌	❌	❌	❌	❌	✅
Antimalaric	❌	❌	❌	❌	❌	❌	✅
Anti-TB	❌	❌	❌	❌	❌	❌	✅
Cystic fibrosis	❌	❌	❌	❌	❌	❌	✅

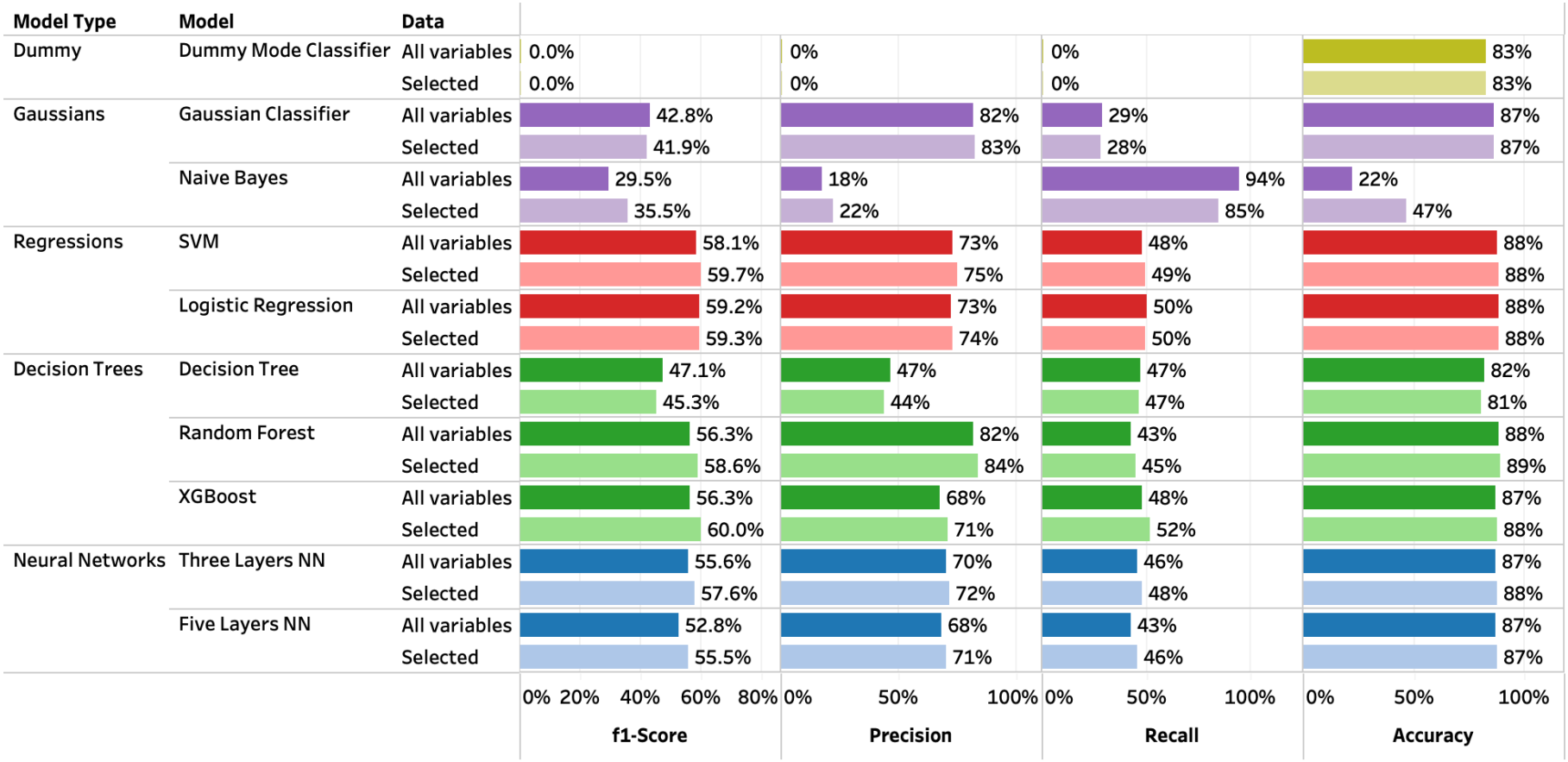


Figura 4.6: Métricas obtenidas para los modelos sobre datos clínicos autoreportados utilizando todas las variables disponibles e imputadas usando el algoritmo KNN y las variables seleccionadas por XGBoost (**Tabla 4.3**). Se presentan las cuatro métricas calculadas sobre los *testsets* utilizando 5-*fold Cross Validation*. En oscuro se muestan las métricas utilizando las variables imputadas y en claro las variables seleccionadas. El modelo *Dummy Mode Classifier* clasifica todos los datos como si fuesen la moda, en este caso controles. Debido a esto las métricas de *precision*, *recall* y $f1$-*score* se indeterminan ya que no existen ni falsos negativos (FN) ni falsos negativos (TN).

Para diferenciar estos modelos, se utiliza el término *Machine Learning* (ML) para referirse a algortimos de aprendizaje que no son de Aprendizaje Profundo (DL, por sus siglas en inglés *Deep Learning*), pese a que las redes neuronales también son algoritmos de aprendizaje de máquina. Los resultados de los modelos ML, siguiendo esta terminología, sobre todos los datos clínicos disponibles se encuentra en el **Anexo C.1.1**. En esta misma sección se encuentran los resultados de la selección de variables (**Figuras C.3-C.7**). El resumen de las variables seleccionadas utilizando estos modelos y las pruebas estadísiticas se muestran en la **Tabla 4.3**.

Para comparar los resultados de estos modelos, las métricas obtenidas se muestran más adelante en la **Figura 4.6**. Particularmente, los modelos sobre todos los datos clínicos disponibles se muestran en barras oscuras. Utilizando como criterio los valores de $f1-score$, el modelo con mejor métrica es la Regresión Logística. Debido a esto las variables seleccionadas corresponden a las obtenidas por este modelo (**Tabla 4.3** columna 4). Estas variables clínicas seleccionadas son usadas para toda la experimentación posterior.

4.3. Modelos sobre variables clínicas seleccionadas

Las variablas seleccionados corresponden a 20 columnas, incluyendo dos variables numéricas edad y densidad poblacional. Este nuevo dataset se utiliza sobre los mismos modelos ML y sobre redes neuronales *feed foward* (FNN). Los modelos ANN también son usados sobre la versión de los datos clínicos que incluyen todas las variables, para verificar que las métricas no disminuyan debido a la selección. Las métricas obtenidas se muestran en la **Figura 4.6**.

Exceptuando el algortimo de Bayes ingenuo gaussiano (*Naive Bayes*) las métricas no varían significativamente utilizando los datos clínicos compppletos o seleccionados. Se verifica además que el *accuracy* es una métrica que no conviene usar para este trabajo ya que el modelo que siempre clasifica la moda obtiene más del 80 %. Las mejores métricas, al igual que con los datos no seleccionados, se obtiene con el modelo XGBoost, aunque, si comparamos el AUC (**Figura C.8**) éste es superado por *Random Forest* y la regresión logística.

Definimos entonces como modelo de comparación, o *baseline*, al modelo XGBoost. El cual es además utilizado para la selección de variables clínicas. Sin embargo, como para poder diseñar un modelo que permita ser entrenado con variables clínicas y variantes genéticas, se requiere una red neuronal, utilizaremos la red de tres capas como otro punto de comparación. La matriz de confusión y la curva de ROC de estos dos modelos se resumen en la **Figura 4.7**, donde se muestra la matriz de confusión calculada usando todos los *set* de *testing* obtenido de la validación 5- *fold* y las curvas de ROC de cada una (la comparación con los demás modelos probados está disponible en **Anexo C.1 Figuras C.8-C.9** y **C.14-C.15**).

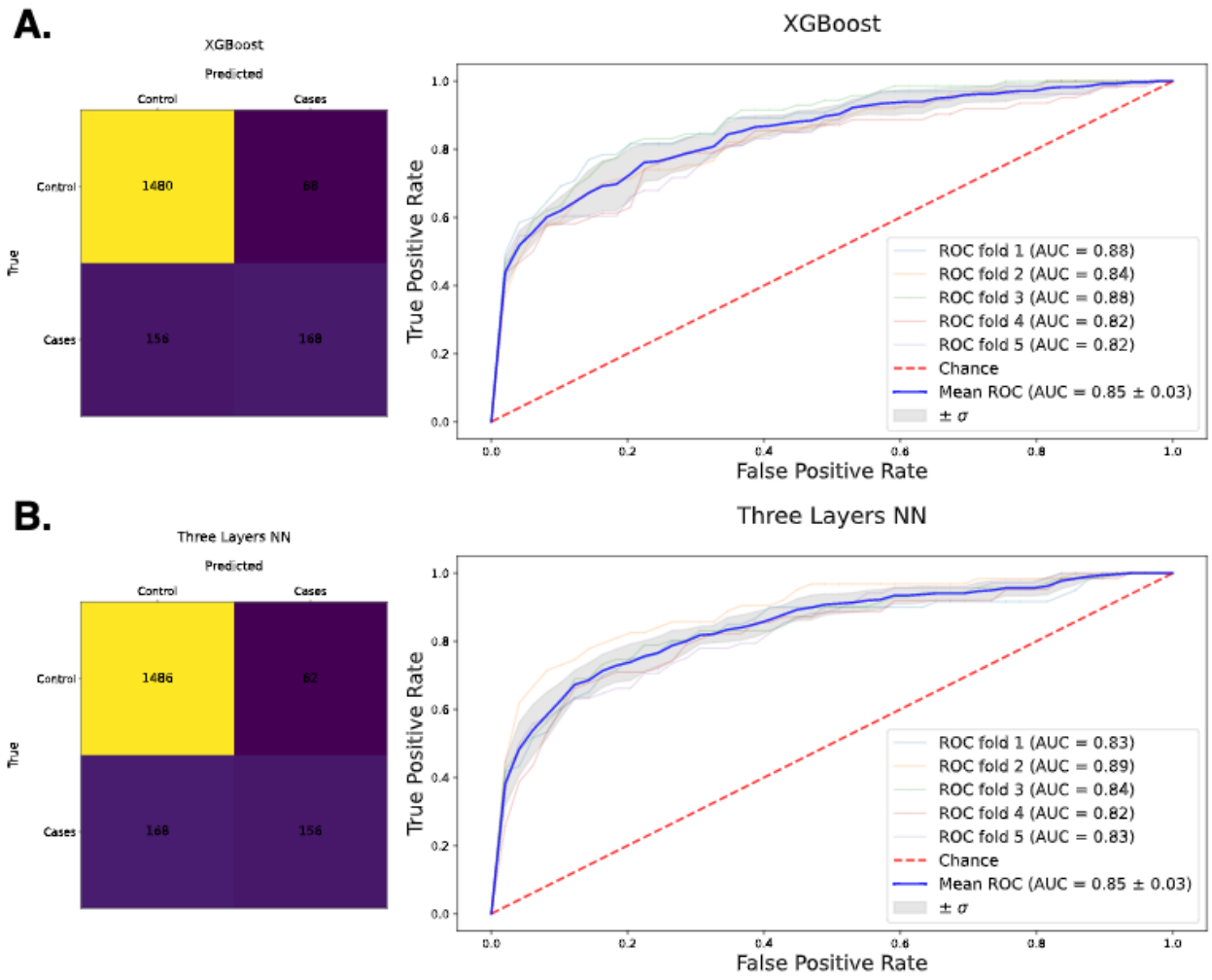


Figura 4.7: Curva de ROC y matriz de confusión de los modelos con mejor $f1$-*score* utilizando como entrada datos clínicos seleccionados. La matriz de confusión muestra el total de las predicciones de todos los *folds* de *testing*. Como se detalla en la **Sección 3.3.3** los cuadrados indican: verdadero negativo (TN, predicho correctamente como control), falso positivo (FP, predicho incorrectamente como caso), falso negativo (FN, predicho incorrectamente como control) y verdadero positivo (TP, predicho correctamente como caso). **A.** XGBoost ($TN = 1480, FP = 68, FN = 156, TP = 168$). **B.** FNN de tres capas ($TN = 1486, FP = 62, FN = 168, TP = 156$).

4.4. Modelos sobre datos genéticos

Si bien, el objetivo de este trabajo es comparar las métricas obtenidas usando variables clínicas con las obtenidas utilizando tanto datos clínicos como variantes genéticas. Antes de esto, se exploran algoritmos de aprendizaje que se entrenan solo con datos genéticos. Esto permite encontrar la mejor aproximación a este tipo de datos, definir hiperpárametros y obtener resultados preliminares de la significancia de estas variantes.

Como se menciona en la **Sección 3.2.1** se utilizan tres *sets* de variantes genéticas:

1. **Microarreglo**: Corresponden a los SNP originalemente genotipificados mediante mi-

croarraglo (o *microarray*) comunes a todos los centros de genetipificado. Éstas son $438307 \approx 4.38 \cdot 10^5$ variantes.

2. **Cromosoma 3**: Corresponden a todos los SNPs disponibles imputados, pero solo del cromosoma 3. Los análisis de asociación de genoma completo (GWAS) se muestran en el **Anexo C.2.3**. La **Figura C.22** corresponde al detalle de solo el cromosoma 3 del *Manhattan Plot* mostrado en la **Figura 2.8**. El total de estas variantes es de $528404 \approx 5.28 \cdot 10^5$.

3. **Seleccionados**: Para asegurar que las variantes usadas sean significativas, se seleccionan las variantes sobre la línea de sugerente ($p < 10^{-5}$) de los datos reportados por la iniciativa internacional (**Figura 2.7** sobre la línea azul). La cantidad de variables seleccionadas por este método corresponde a 649 SNPs.

4.4.1. Selección de hiperparámetros

Para la selección de hiperparámetros se realizan los experimentos descritos en la **Sección 3.3.2**. Debido a que esta selección toma mucho tiempo, solo se realiza para los SNPs originalmente genotipificados (resultados en **Anexo C.3.2.1.1**) y para las variantes seleccionadas (resultados en **Anexo C.3.2.1.2**). Los hiperparámetros para el modelo *Dual-Stream CNN* para los SNPs originalmente genotipificados son $lr = 10^{-7}$ y $wd = 10^{-4}$, estos se utilizan también para el *set* de variantes genéticas imputadas del cromosoma 3, ya que el tamaño de la secuencia es del mismo orden de magnitud. Para el modelo que utiliza los SNPs seleccionados como entrada los hiperparámetros seleccionados son $lr = 10^{-4}$ y $wd = 10^{-8}$.

4.4.2. Desbalance

Durante la búsqueda de hiperparámetros se constata que la métrica de $f1$-*score* se mantienen en 0.0 durante el entrenamiento para el *set* de validación. La matriz de confusión obtenida muestra que la red clasifica todos los participantes del *testset* como controles (**Figura C.41** arriba, izquierda). Esto se puede atribuir al desbalance de las etiquetas. Para abordar este problema se realizan otros tres experimentos: (1) agregando ponderadores a la función de *loss*, (2) realizando *undersampling* (inframuestreo) y (3) *oversampling* (sobremuestreo).

Al agregar ponderadores a la función de *loss* se sobre-penaliza clasificar erróneamente los casos. Esto transforma la función de *loss* (**Ecuación 2.11**), agregando el ponderador w_c al cálculo de la *loss* no reducida (**Ecuación 2.13**).

$$l_n^{'} = -\sum_{c=1}^{C} w_c \log\left(\frac{\exp(\hat{y}_{n,c})}{\exp\left(\sum_{i=1}^{C} \hat{y}_{n,i}\right)}\right) y_{n,c} \tag{4.1}$$

Para este trabajo las clases son $c = 1$ controles y $c = 2$ casos, con sus ponderadores respectivas $w_1 = 0.1$ y $w_2 = 0.9$.

Los resultados para cada *set* se muestran en el **Anexo C.3.2.2**. Para todos los *sets* los experimentos que mayor $f1$-*score* obtienen es realizando *oversampling*.

4.4.3. Métricas

Debido a la cantidad de variantes incluidas en los *sets* del microarreglo ($\sim 4 \cdot 10^5$ SNPs) y para el cromosoma 3 ($\sim 5 \cdot 10^5$ SNPs) no se realizan experimentos utilizando los modelos ML no DL. Sin embargo, se utilizan estos modelos sobre los SNPs seleccionados según sugerencia de significancia ($\sim 6 \cdot 10^2$ SNPs) para comparar si las redes neuronales presentan alguna ventaja utilizando datos genéticos. Las métricas $f1$-*score* y *accuracy* se muestran en la **Figura 4.8** a continuación, en la que se evidencia que al utilizar solo variantes genéticas, los modelos ML solo se ajustan al desbalance de la clase a predecir.

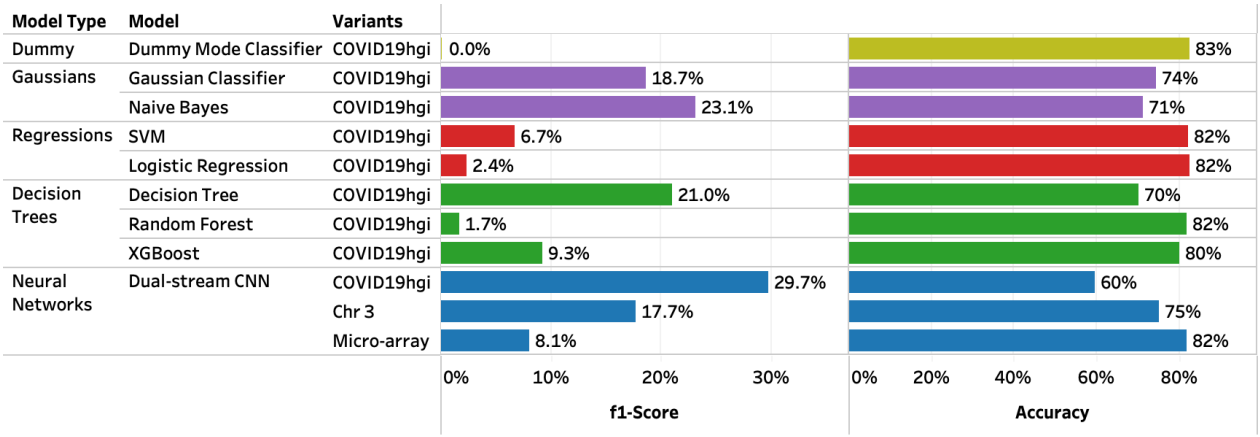


Figura 4.8: Métricas obtenidas para los modelos sobre variantes genéticas seleccionadas según sugerencia de significancia por la iniciativa internacional COVID19hg. Se presentan el $f1$-*score* y el *accuracy* sobre los *testsets* utilizando 10-*fold Cross Validation*. El modelo *Dummy Mode Classifier* clasifica todos los datos como si fuesen la moda, en este caso controles. Debido a esto las métricas de *precision*, *recall* y $f1$-*score* se indeterminan ya que no existen ni falsos negativos (FN) ni falsos negativos (TN).

4.4.4. Saliencia

Para estos modelos preeliminares se extrae la saliencia (**Sección 2.1.4.2**) siguiendo el método propuesto por Liu y colobaradores [38]. El modelo con los hiperparámetros seleccionados para el respectivo *dataset* de variantes genéticos, utilizando *oversampling*, se entrena por 10-*fold Cross Validation* dejando 8 *folds* para entrenamiento, un *fold* para validación y un *fold* para *testing*. La porción de datos dejada para *test* se utiliza para calcular la saliencia. De esta forma se calcula la saliencia para todos los participantes. Al igual que [38] se reporta la mediana de todos los participantes como la **saliencia poblacional**.

Se comparan los resultados de la saliencia obtenida con el GWAS utilizando los mismos datos genéticos. La publicación [38] compara la saliencia con el GWAS que calcula significancia estadística en base a la prueba de Wald, calculado con el método `gwas2` del paquete de R NAM [64] (**Figura 2.4**). Sin embargo, los resultados de estos del GWAS por este método para el fenotipo estudiado en este trabajo, no se corresponde con los resultados obtenidos por el meta-análisis de la iniciativa internacional (**Figura 4.9**). Por ello, la saliencia poblacio-

nal obtenida se compara con el GWAS reportando el valor p descrito previamente (**Sección 3.2.2, método (2)**). Por esta razón, la saliencia se comparará con los GWAS calculados por el proyecto, por la iniciativa internacional y utilizando los datos clínicos seleccionados.

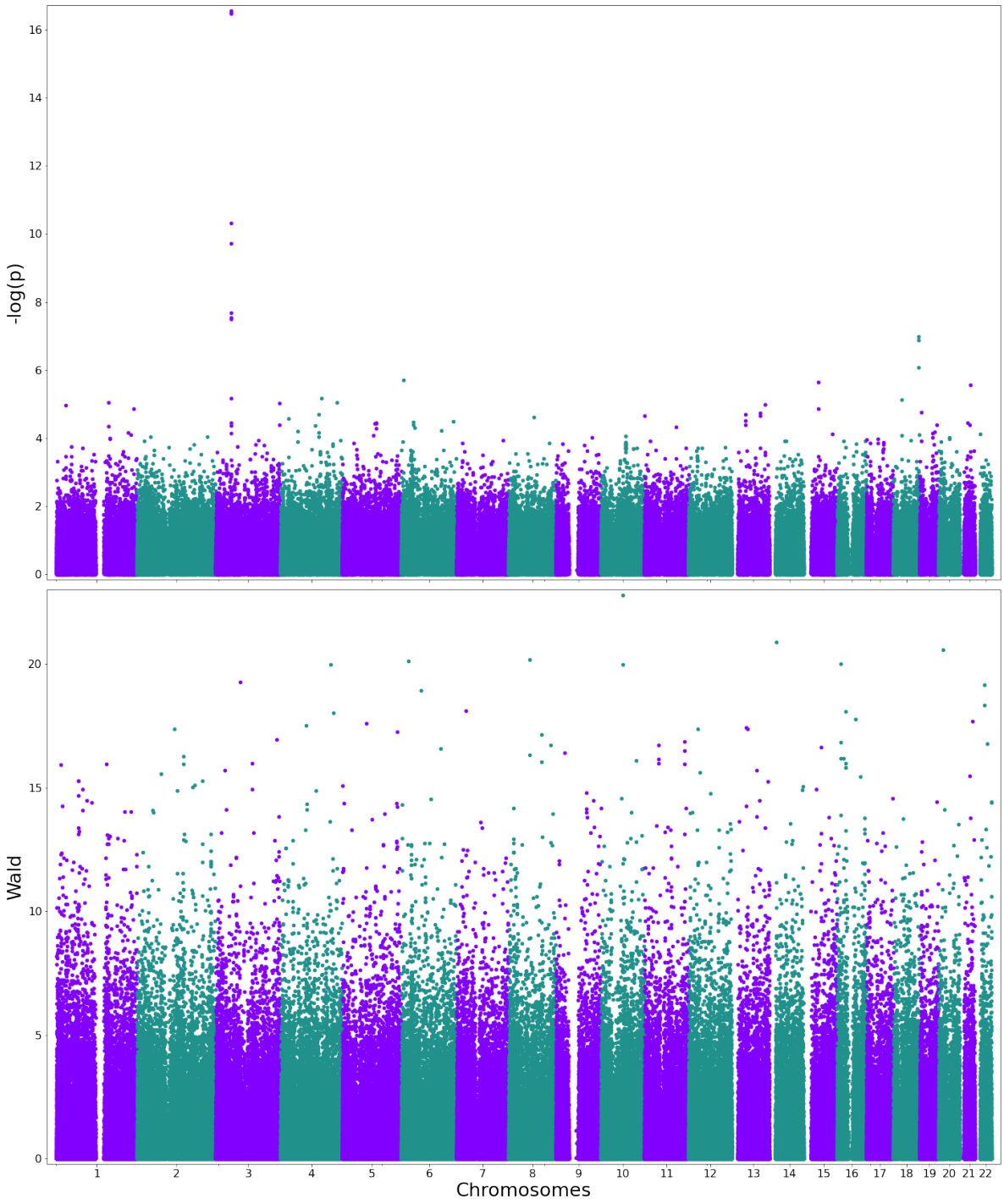


Figura 4.9: Comparación entre los valores del GWAS de la iniciativa internacional COVIDhg (arriba) y el obtenido usando datos del proyecto mediante el método `gwas2` para los SNPs genotipificados (abajo). Se aprecia que que el cromosoma 3 no presenta los *peaks* del proyecto, a diferencia de los GWAS calcudos usando el valor p de las regresiones logísitcas.

La saliencia obtenida que utiliza solo las variantes genéticas obtenidas desde el microarreglo, no muestra ninguna variante interesante (**Figura C.54**). Esto es esperable, ya que las variantes genéticas, originalmente genotipificadas por microarreglo, no presentan significancia estádistica según los análisis previos (**Sección 2.3.2.1**, **Figura 2.9**).

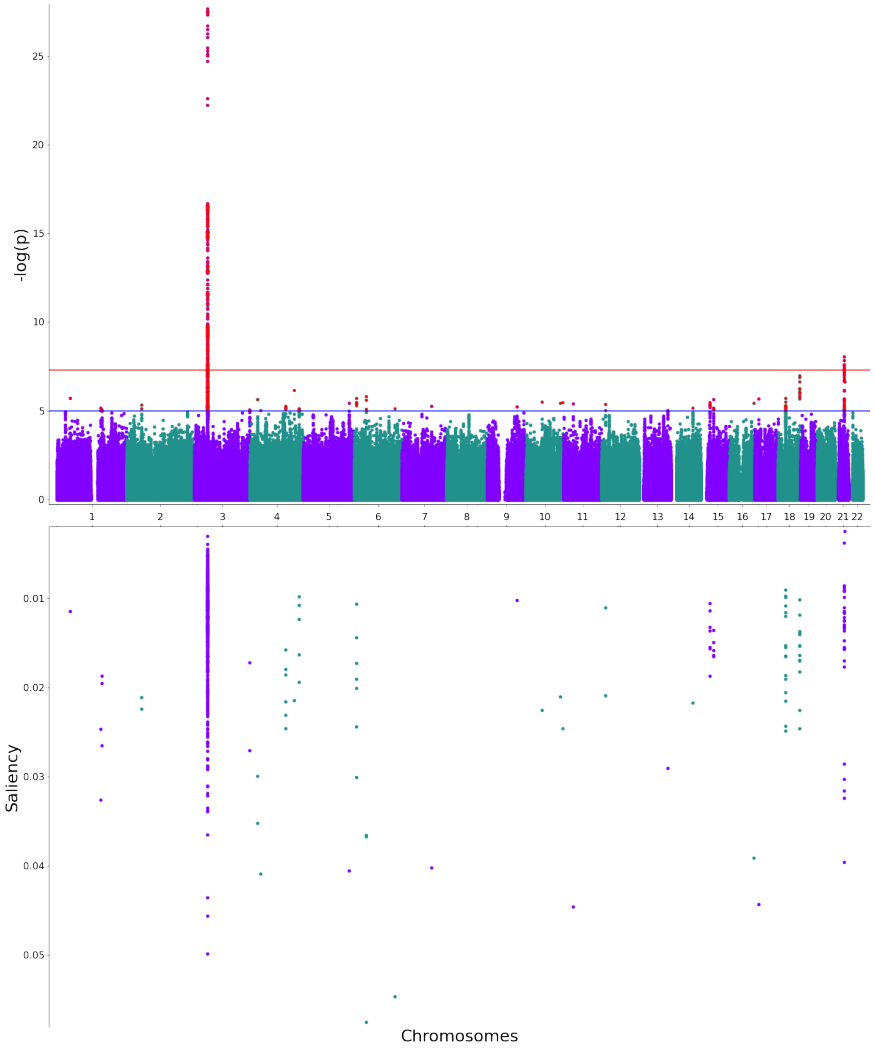


Figura 4.10: *Scatterplot* de doble eje con (1) *Manhattan Plot* de la iniciativa internacional COVID19hg (arriba) y (2) la saliencia obtenida utilizando solo variantes genéticas en el modelo Dual-stream CNN (abajo) con el eje invertido. Las variantes cuya saliencia se muestra en el eje invertido, se encuentran en color rojo en el *scatterplot* del GWAS.

Para comparar los resultados de la saliencia por la red entrenada Dual-stream CNN con los significancia estadísitica por GWAS, se realizan gráficos *Manhattan plots* con el GWAS y la saliencia para las variantes seleccionadas en un *scatterplot* con dos ejes y. En estos gráficos la saliencia se muestra en un eje y invertido para facilitar la ubicación del SNP. Además, las variantes seleccionas se muestran en color rojo en los GWAS.

Al comparar la saliencia con los datos de la iniciativa internacional (**Figura 4.10**) se evidencia el GWAS asigna mayor significancia a un número acotado de variantes, mientras

que la gran mayoría de variantes tienen valores bajos (por debajo de la línea de sugerencia de significancia, $p > 10^{-5}$) en comparacion con los *peaks*. Por el contrario, en la saliencia no se observa dicha preponderancia de solo un par de ubicaciones cromosómicoas, sino que hay un mayor número de variantes que parecen aportar información. Dentro de un mismo locus, hay mayor dispersión de datos, con más variantes con poco aporte de información y muy pocas con altos valores, sugiriendo que son pocas las variantes que entregan información no redundante dentro de un mismo locus. Esto último parece compararse en los mecanismos de redes neuronales, mientras que en GWAS existe la limitación de considerar solo una variante a la vez. Si se realiza este ejercicio con los datos del proyecto **Figura 4.11**) se tiene un parecido similar.

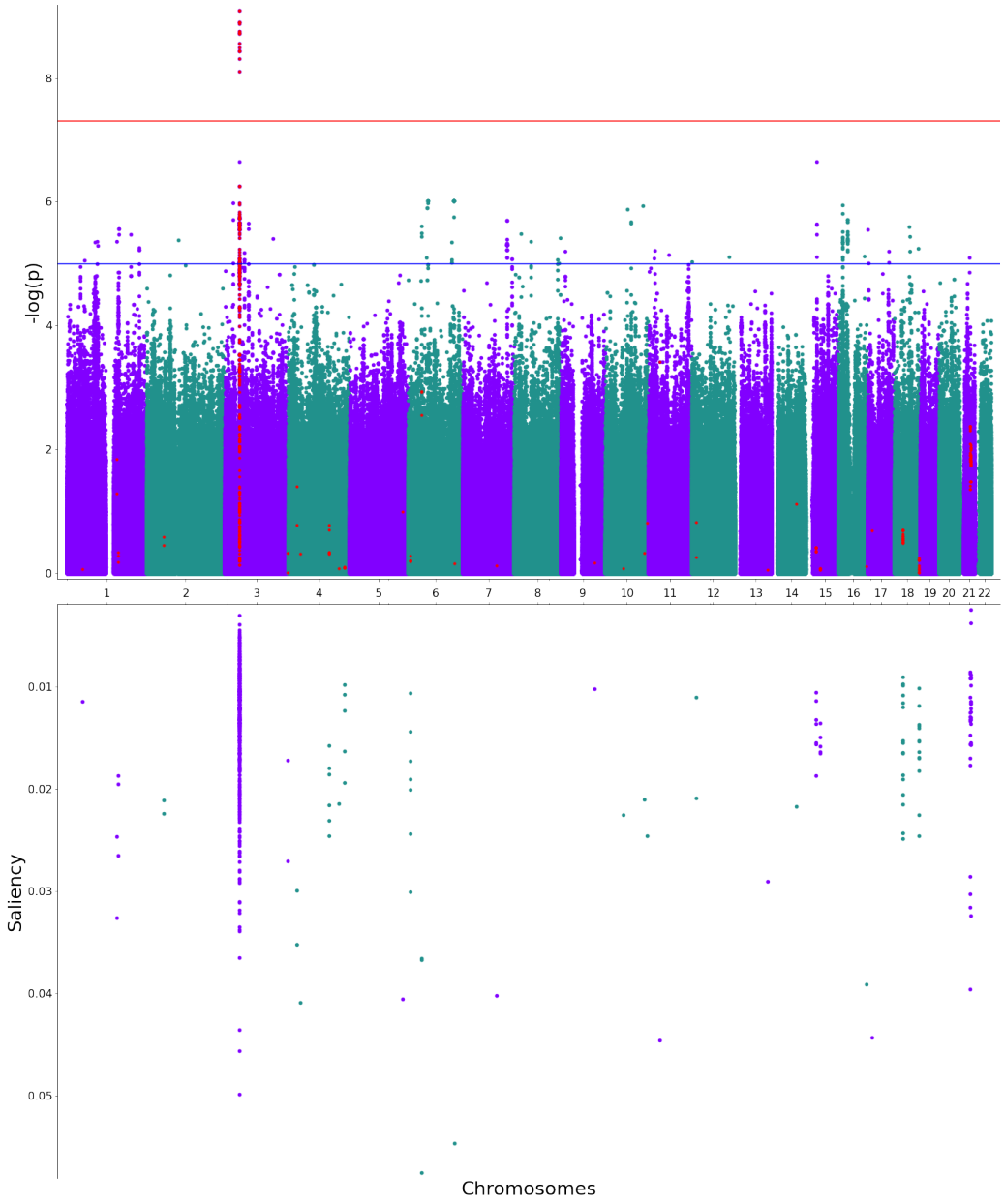


Figura 4.11: *Scatterplot* de doble eje con (1) *Manhattan Plot* de utilizando los datos obtenidos para el proyecto (arriba) y (2) la saliencia obtenida utilizando solo variantes genéticas en el modelo Dual-stream CNN (abajo) con el eje invertido. Las variantes cuya saliencia se muestra en el eje invertido, se encuentran en color rojo en el *scatterplot* del GWAS.

4.5. Agregando datos genéticos

En la **Sección 4.3** el modelo XGBoost obtiene un 60.0 % de $f1$-*score* y un 88 % de *accuracy* siendo entrenado con variables clínicas seleccionadas. El objetivo de este trabajo es verificar que agregar variantes genéticas al entrenamiento del modelo, estas métricas mejoran. Para esto se realizan dos aproximaciones (1) utilizar arquitecturas de red neuronal diseñadas para recibir ambos tipos de datos y entrenarse en conjunto (**Sección 4.5.1**) y (2) añadir algunas

variantes genéticas al *set* de entrenamiento del modelo XGBoost (**Sección 4.5.2**).

Para la primera aproximación, se requiere de: un modelo ANN que se entrene con datos clínicos, y otra ANN que reciba variantes genéticas. Esta nueva arquitectura, puede ser entrenada con ambos tipos de datos a la vez mediante *backpropagation*. En la **Sección 4.3** se obtiene que una FNN de tres capas obtiene $f1$-*score* de 57.6 % y 88 % de *accuracy*. La arquitectura propuesta combina esta FNN con Dual-stream CNN (**Sección 4.4**), a través de una última capa *fully connected* (o perceptrón) (**Figura 3.7**). Se espera que esta nueva arquitectura obtenga métricas más altas que la FNN.

La segunda aproximación, consiste en agregar, progresivamente, variantes genéticas al *set* de entrenamiento con datos clínicos. Cada uno de estos *set* ajustan el modelo XGBoost. Una vez entrenado y validado por *Cross Validation*, se reporta el $f1$-*score* obtenido. Se espera que al añadir variantes genéticas aumenten las métricas de XGBoost.

4.5.1. Modelos adaptados

En esta sección se describen los resultados obtenidos para el modelo descrito en la **Sección 3.3.2.3** (**Figura 3.7**) sobre las variables clínicas seleccionadas por la regresión logística (**Sección 4.2**) y los *set* de variantes genéticas.

4.5.1.1. Hiperparámetros

Debido al tiempo que toma realizar los experimentos sobre los *set* de variantes genéticas del microarreglo ($\sim 4 \cdot 10^5$ variantes) y los SNPs del cromosoma 3 ($\sim 5 \cdot 10^5$ variantes) se reutilizan los hiperparámetros obtenidos en la **Sección 4.4.1** para estos *sets*. Para las variantes seleccionadas según la iniciativa ($\sim 6 \cdot 10^2$) se obtienen nuevo valores (**Anexo C.4.1**). El resumen de estos hiperpárametros se resumen en la **Tabla 4.4**.

Tabla 4.4: Hiperparámetros y método por desbalance sobre la arquitectura Dual-stream CNN adaptada (**Figura 3.7**) para cada *set* de variantes genéticas.

Set	Largo de secuencia	lr	wd	Método
Microarreglo	$438307 \sim 4 \cdot 10^5$	10^{-7}	10^{-4}	*Oversampling*
Cromosoma 3	$528404 \sim 5 \cdot 10^5$	10^{-7}	10^{-4}	*Oversampling*
Seleccionados (COVIDhgi)	$649 \sim 6 \cdot 10^2$	10^{-5}	10^{-7}	*No resampling*

Junto con la selección de hiperpárametros, se realizan experimentos mediante técnicas para ocuparse del desbalance. Estos resultados se muestran en el **Anexo C.4.2**. El método con el que se obtienen mejores métricas para cada *set* de variantes genéticas se incluye en la **Tabla 4.4**. Seleccionados los hiperparámetros se realizan los experimentos utilizando 10-*fold Cross Validation* para obtener métricas robustas y la saliencia de las variantes.

4.5.1.2. Métricas y Saliencia

Una vez definidos los hipérparametros se realizan 10-*folds Cross Validation* para cada *set* de variantes genéticas. Las métricas obtenidas se comparan con los modelos entrenados por variables clínicas. La saliencia, solo del módulo convolucional de la red, se compara con los GWAS realizados.

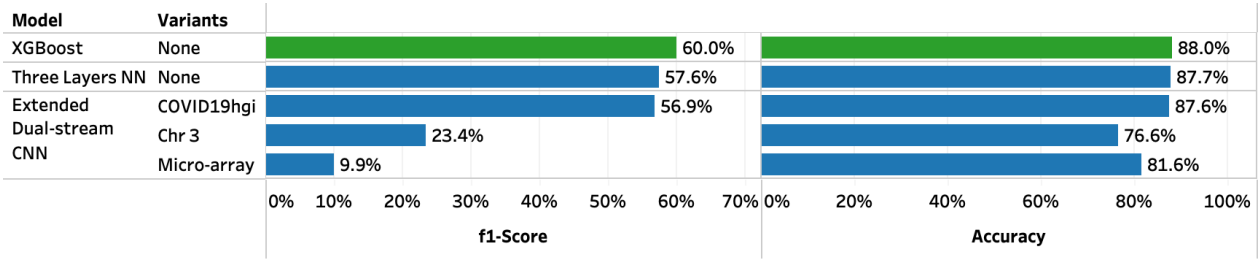


Figura 4.12: Métricas obtenidas para los modelos XGBoost y FNN sobre datos clínicos seleccionadas por la regresión lineal con penalización $l1$ (**Tabla 4.3**). Se presentan las métricas $f1$-*score* y *accuracy* sobre los *testsets* utilizando 5-*fold Cross Validation*. Se comparan con las métricas obtenidas por la red Dual-stream CNN adaptada (**Figura 3.7**) para 10-*fold Cross Validation* entrenada con datos clínicas y con las variantes genéticas de los tres *sets*: microarreglo (*Micro-array*), cromosoma 3 (*chr 3*) y seleccionadas por sugerente de significancia según la iniciativa internacional COVIDhg (*COVID19hgi*).

Los modelos adaptados obtienen métricas muy diferentes dependiendo del *set* de variantes genéticas usadas, lo que se muestra en la **Figura 4.12** donde se aprecia la comparación del modelo Dual-stream CNN adaptado según el *set* de variantes genéticas, junto a los modelos de *baseline* que solo utilizan datos clínicos. Al utilizar las variantes originalmente obtenidas por microarreglo se obtienen muy bajas métricas como es esperado según los GWAS mostrados para los datos de este proyecto (**Figura 2.9**). Pese a tener más variantes, el *set* del cromosma 3 alcanza mejores métricas y mayor generalización que el *set* del microarreglo. Esto parece reafirmar la idea que las variantes presentes en el cromosoma 3 tienen una mayor correlación con la severidad de la enfermedad. Las mejores métricas con variantes genéticas es el modelo que utiliza las variantes seleccionadas por sugerencia de significancia según COVID19hgi, lo que es atribuible a esta misma selección de correlación y a la reducción de variables que aportan ruido. Pese a que este modelo es el que muestra mejor desempeño utilizando variantes genéticas, no se supera las métricas logradas por la red que solo utiliza variables clínicas.

Estos resultados ($f1\text{-}score_{\text{COVID19hgi}} > f1\text{-}score_{\text{Chr 3}} > f1\text{-}score_{\text{microarreglo}}$) se repiten tanto para el modelo Dual-stream CNN como para la versión adaptada (Dual-stream CNN + 3 *layers* FNN) (**Figuras 4.8 y 4.12** respectivamente). Como todos los modelos sí reducen el *loss* y aumentan las métricas en el *set* de entrenamiento, la diferencia radica en la generalización. Vemos que para estos modelos, la selección de variantes mejora la capacidad de generalizar debido a la reducción de variantes que parecen solo aportar ruido. Pese a obtener menor métrica, se puede asignar un valor de la correlación de las variantes genéticas con la severidad por medio de la saliencia del módulo convolucional de la arquitectura adaptada. Esta saliencia para cada variante se compara con los GWAS de la significancia ya reportada

por la iniciativa internacional en la **Figura 4.13**.

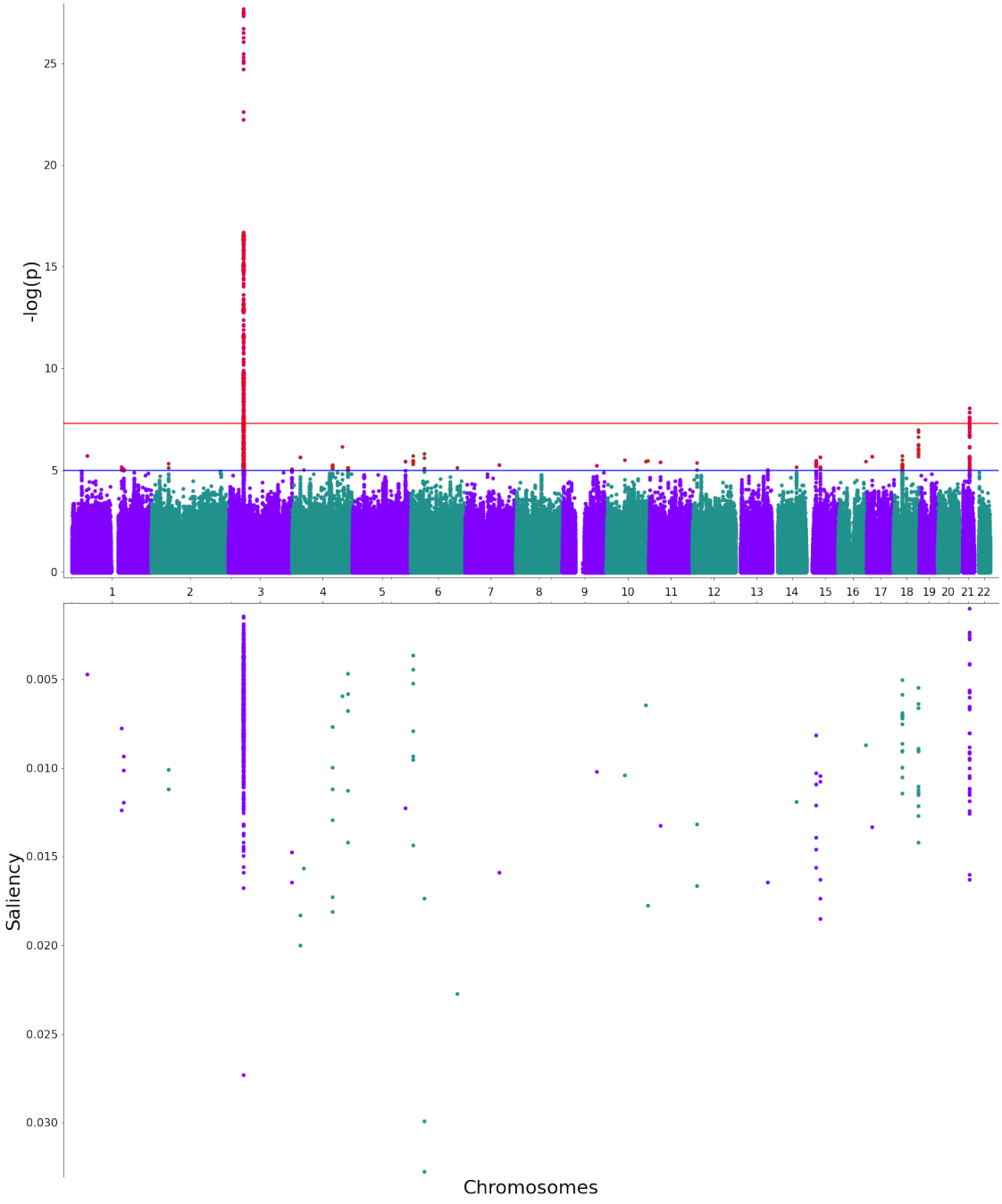


Figura 4.13: *Scatterplot* de doble eje con (1) *Manhattan Plot* de la iniciativa internacional COVID19hg (arriba) y (2) la saliencia obtenida utilizando variantes genéticas y variables clínicas seleccionadas por el modelo Dual-stream CNN adaptado (**Figura 3.7**) (abajo) con el eje invertido. Las variantes cuya saliencia se muestra en el eje invertido, se encuentran en color rojo en el *scatterplot* del GWAS.

Se aprecía que la saliencia asigna nuevos *peaks* en el cromosoma 6 y separa uno de los SNP del cromosoma 3. Para tener una comparación más clara y analizar estos resultados desde la genética, en la **Tabla 4.5** se muestran los SNPs con mayor saliencia para cada cromosoma,

comparado con los menores *p-value* de los GWAS de la iniciativa internacional y los datos del proyecto.

Tabla 4.5: Comparación de la correlación de variantes genéticas con la severidad de Covid19 para cada cromosoma. Se compara la mayor saliencia de cada cromosoma, con las menores *p-values* de los GWAS de la iniciativa (**Figura 2.7**) y con los datos del proyecto (**Figura 2.8**). Debido a que las variantes fueron seleccionadas según sugerencia de significancia por los resultados de la iniciativa internacional COVID19hg, las posiciones con mayor significancia del GWAS del proyecto excluyen aquellas variantes que no fueron consideradas en la selección de SNP descrita.

		Saliency			COVID19hgi			GWAS		
Chromosome	Variants	Position	SNP id	Gene	Position	SNP id	Gene	Position	SNP id	Gene
1	6	155237942	rs28678003	GBA	47731898	rs1436797106	–	155237942	rs28678003	GBA
2	2	52627539	rs1682883996	–	52627429	rs7597967	–	52627539	rs1682883996	–
3	526	45896341	rs3774641	CCR9, LZTFL1	45818159	rs17713054	LOC107986083	45848457	rs35731912	LZTFL1
4	15	25447603	rs7671107	LOC105374536	156999361	rs72683395	–	25447603	rs7671107	LOC105374536
5	1	163759964	rs1893565	–	163759964	rs1893565	–	163759964	rs1893565	–
6	11	41525674	rs1763259718	FOXP4-AS1	41520640	rs12660421	–	41520640	rs12660421	–
7	1	105542599	rs1790504525	RINT1	105542599	rs1790504525	RINT1	105542599	rs1790504525	RINT1
9	1	105927741	rs1468787396	LOC107987108	105927741	rs1468787396	LOC107987108	105927741	rs1468787396	LOC107987108
10	3	129500829	rs11016814	MGMT	55108096	rs1779262	PCDH15	129500829	rs11016814	MGMT
11	1	34507219	rs766826	ELF5	34507219	rs766826	ELF5	34507219	rs766826	ELF5
12	2	14283677	rs74980864	–	14299516	rs140443113	–	14283677	rs74980864	–
13	1	101803461	rs9652125	FGF14	101803461	rs9652125	FGF14	101803461	rs9652125	FGF14
14	1	77125563	rs888071	–	77125563	rs888071	–	77125563	rs888071	–
15	12	45109438	rs1961660	DUOX2	45111868	rs2001616	DUOX2	31229978	rs12101670	LOC283710
16	1	86445090	rs6540293	–	86445090	rs6540293	–	86445090	rs6540293	–
17	1	14408980	rs6502354	LOC107985080	14408980	rs6502354	LOC107985080	14408980	rs6502354	LOC107985080
18	29	77223518	rs12962836	–	77217539	rs7239621	–	26175682	rs9966001	PSMA8
21	35	33239687	rs8127500	IFNAR2	33234329	rs1986288608	IFNAR2	33237291	rs1137779	IFNAR2

En la **Tabla 4.6** se muestran los SNPs con mayor saliencia para cada cromosoma y se compara con los valores de p para los GWAS previos. La selección se hizo por sugerente de significancia ($Pr(|z|) < 10^{-5}$) según los datos de la iniciativa internacional, sin embargo, solo los SNPs del cromosoma 3 y 18 se consideran significativos ($Pr(|z|) < 5 \cdot 10^{-7}$). Para comparar los valores de saliencia y significancia estadística se realiza un *scatterplot* para todas las variantes (**Figura 4.14**). Se aprecia que no existe mayor correlación entre la mayor saliencia y la significancia que asigna el GWAS.

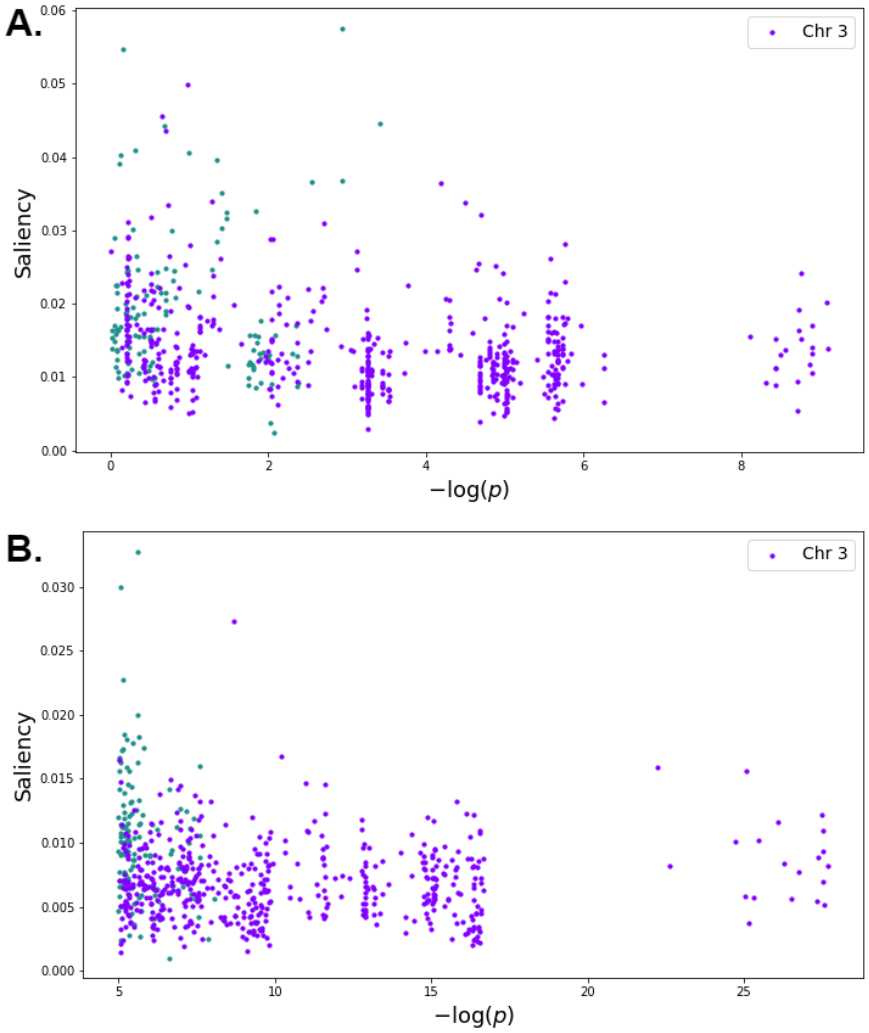


Figura 4.14: *Scatterplot* para cada variante seleccionada según sugerencia de significancia por COVID19hgi. La ordenada contiene la saliencia, mientras que la abscisa mapea el valor de p según GWAS. Las variantes del cromosoma 3 aparecen de color morado, considerando que son las más abundantes y que estás tienen mayores valored de p solo por cercanía a los *peaks*. **A.** Saliencia calculada usando la arquitectura Dual-stream CNN y el valor p utilizando el GWAS del proyecto. **B.** Saliencia calculada usando la arquitectura extendida Dual-stream CNN y el valor p de la iniciativa internacional.

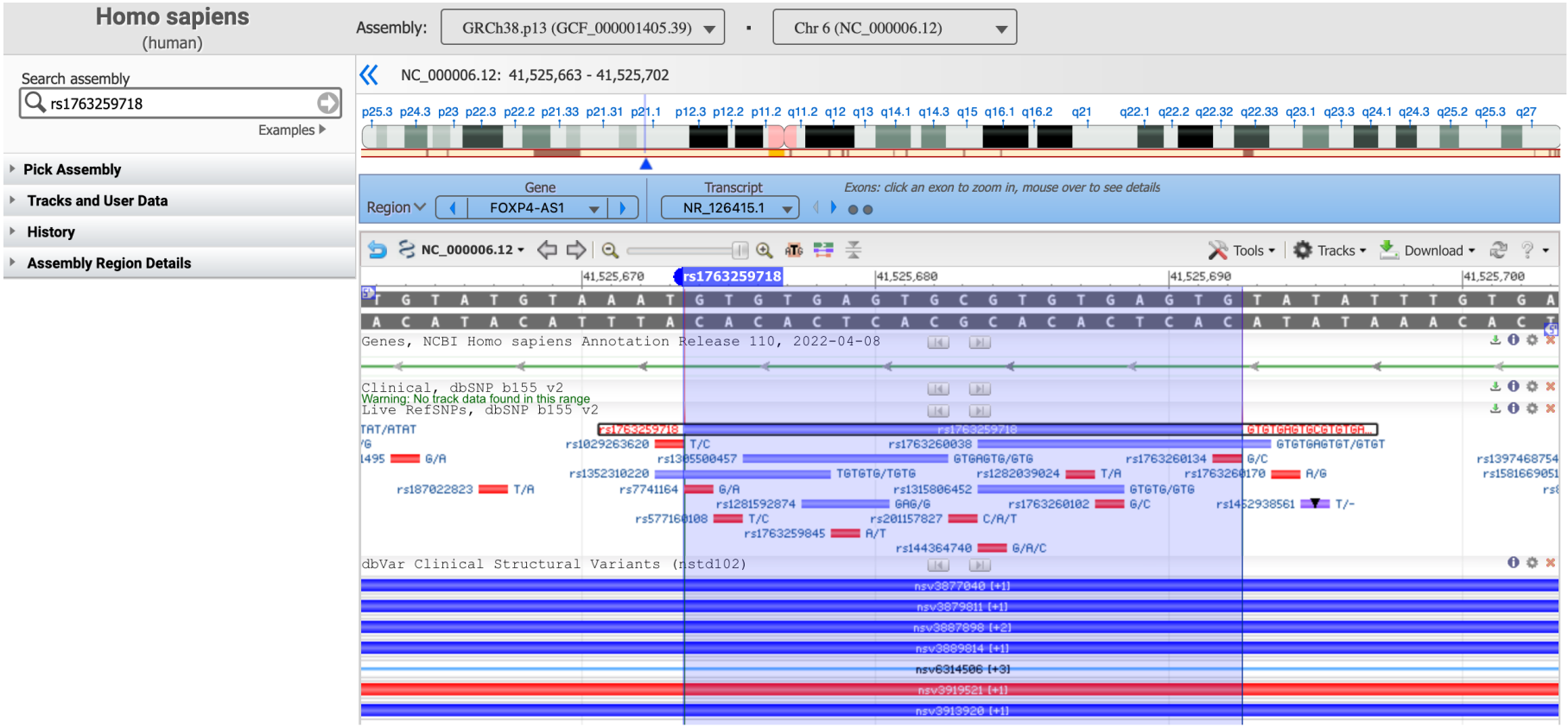


Figura 4.15: Captura de pantalla de la herramienta Variation Viewer para la variante rs1763259718 ubicada en el cromosoma 6. Se aprecia que esta variante se encuentra dentro de la secuencia del gen FOXP4-AS1, el cual es un regulador de RNA antisentido del gen FOXP4, que codifica un factor de transcripción que regula la diferenciación celular. Fuente: https://www.ncbi.nlm.nih.gov/variation/view/ para la variante rs1763259718.

Tabla 4.6: Comparación de la mayor saliencia obtenida de variantes genéticas por cada cromosoma (mismo que en la **Tabla 4.5**). Se comparan con los valores de estas variables con el *p-values* de los GWAS de la iniciativa (**Figura 2.7**) y con los datos del proyecto (**Figura 2.8**).

					COVID19hgi		GWAS	
Chromosome	Position	SNP id	Gene	Saliency	OR	Pr(>|z|)	OR	Pr(>|z|)
1	155237942	rs28678003	GBA	0.0124	0.852	6.87e-06	0.488	1.46e-02
2	52627539	rs1682883996	–	0.0112	1.086	7.43e-06	1.118	2.56e-01
3	45896341	rs3774641	CCR9, LZTFL1	0.0273	1.128	2.14e-09	1.549	3.15e-05
4	25447603	rs7671107	LOC105374536	0.0200	0.918	2.35e-06	1.205	3.93e-02
5	163759964	rs1893565	–	0.0123	1.088	3.72e-06	1.166	1.01e-01
6	41525674	rs1763259718	FOXP4-AS1	0.0328	1.206	2.48e-06	1.352	1.17e-03
7	105542599	rs1790504525	RINT1	0.0159	1.860	5.34e-06	0.918	7.61e-01
9	105927741	rs1468787396	LOC107987108	0.0102	1.282	5.73e-06	0.892	6.75e-01
10	129500829	rs11016814	MGMT	0.0178	1.100	3.45e-06	1.154	1.55e-01
11	34507219	rs766826	ELF5	0.0132	0.925	4.01e-06	0.746	3.82e-04
12	14283677	rs74980864	–	0.0166	0.862	9.33e-06	1.329	1.51e-01
13	101803461	rs9652125	FGF14	0.0164	1.088	9.84e-06	0.987	8.89e-01
14	77125563	rs888071	–	0.0119	0.917	6.84e-06	1.252	7.66e-02
15	45109438	rs1961660	DUOX2	0.0185	1.118	6.82e-06	0.981	8.77e-01
16	86445090	rs6540293	–	0.0087	0.918	3.58e-06	1.026	7.78e-01
17	14408980	rs6502354	LOC107985080	0.0133	1.080	2.19e-06	0.901	2.06e-01
18	77223518	rs12962836	–	0.0142	0.917	2.33e-07	0.968	7.52e-01
21	33239687	rs8127500	IFNAR2	0.0163	1.075	2.45e-06	0.791	4.26e-03

En cuanto a la selección de variantes, el SNPs con mayor saliencia está en el cromosoma 6 y está ubicado en el gen FOXP4-AS1 (**Figura 4.15**). Este gen codifica un RNA antisentido que regula el factor de transcripción FOXP4, el cual a su vez, regula la diferenciación de células en diferentes tejidos. Como un regulador de diferenciación, la mutación de este gen puede llevar al desarrollo de células cancerígenas. Se reporta [41] que el gen FOXP4-AS1 se relaciona con la aparición de carcinoma de células escamosas de esófago. Pese a esto, el cáncer de colón y recto no se encuentra entre las variables clínicas seleccionadas (**Tabla 4.3**). Este es un resultado interesante, ya que el método de selección de variables por saliencia en CNN de una dimensión, propone nuevos candidatos para el estudio de mecanismos moleculares que no se han reportado en estudios anteriores.

4.5.2. Añadiendo variantes más significativas

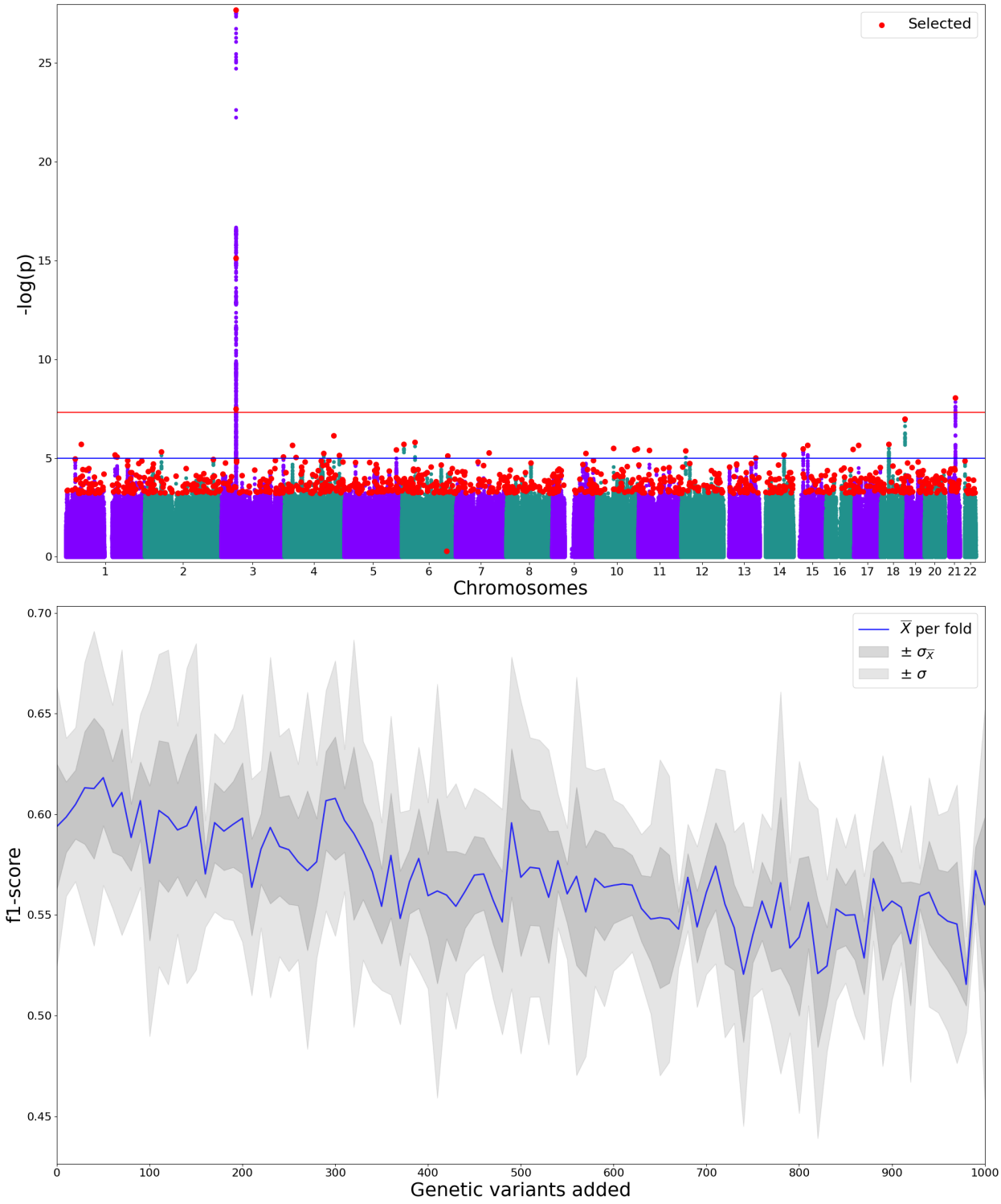


Figura 4.16: F1-score obtenido por el modelo XGBoost utilizando como entrada variables clínicas seleccionadas y una cantidad creciente de variantes genéticas. La salida del modelo es la hospitalización por Covid-19 entre participantes infectados con SARS-CoV-2. La selección de las variantes genéticas se realiza por la significancia reportada por la iniciativa internacional Covid19hg, excluyendo variantes muy cercanas (menores o iguales a 250000 nucleótidos), de forma creciente. En la parte superior se incluye el GWAS de la iniciativa internacional con los SNPs utilizados en rojo.

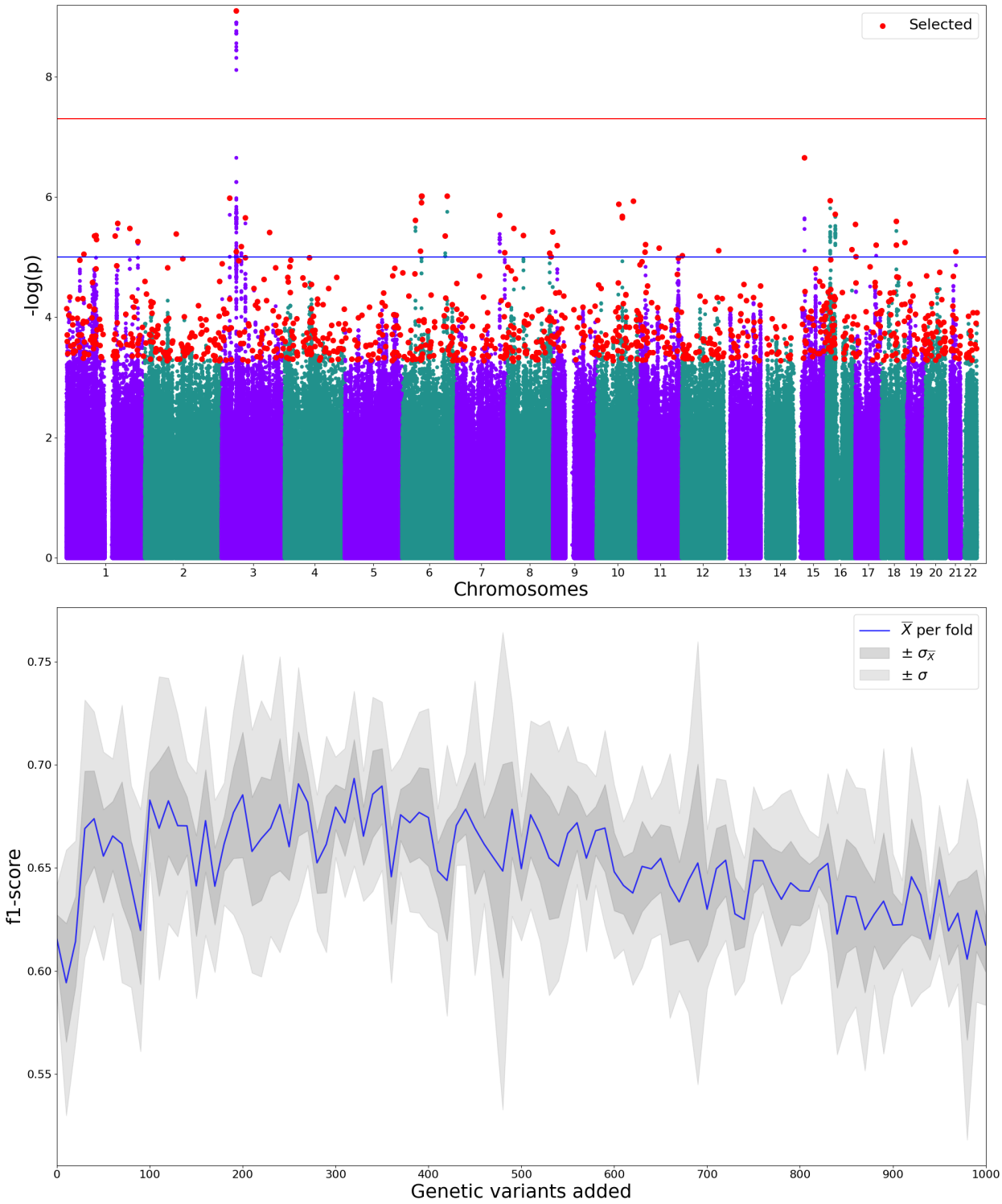


Figura 4.17: F1-score obtenido por el modelo XGBoost utilizando como entrada variables clínicas seleccionadas y una cantidad creciente de variantes genéticas. La salida del modelo es la hospitalización por Covid-19 entre participantes infectados con SARS-CoV-2. La selección de las variantes genéticas se realiza por la significancia obtenida por el GWAS generado a partir de los datos del proyecto, excluyendo variantes muy cercanas (menores o iguales a 250000 nucleótidos), de forma creciente. En la parte superior se incluye el GWAS de la iniciativa internacional con los SNPs utilizados en rojo.

Otra aproximación para determinar si las variantes genéticas se correlacionan con la predicción de hospitalización, es agregar estás variantes al *set* de entrenamiento de un modelo ya validado, en este caso, sobre variables clínicas. Como el módelo que mejor métricas obtiene sobre los datos clínicos autoreportados es XGBoost (**Figura 4.7**), este se entrena agregando las variantes más significativas según la iniciativa y el proyecto. Para que las variantes añadidas no sean de un mismo gen asociado, la selección se hace tomando los menores p, pero descartando los SNPs que se encuentran a menos de $2.5 \cdot 10^5$ nucleótidos de una variante ya seleccionada.

Se reportan los $f1$-*scores* sobre los *testsets* generados por 5-*fold Cross Validation* para XGBoost utilizando solo las variables clínicas seleccionadas, luego utilizando estas variables más 10 SNPs seleccionados, luego 20 y así, de 10 en 10 SNPs hasta las 2000 variantes. Los resultados se muestran en la **Figura 4.16** utilizando los valores de p de la iniciativa internacional y en la **Figura 4.17** utilizando los valores del proyecto.

Como es esperable los valores comienzan $\sim 60.0\,\%$, que es el valor obtenido en los experimentos solo con datos clínicos. Al utilizar las variantes según la significancia reportada en la iniciativa internacional, vemos que no hay mejora en la predicción. Esto puede significar que las variantes solo añaden ruido a la clasificación. Sin embargo, utilizando lo reportado según el GWAS sobre los datos de este proyecto (participantes chilenos), el $f1$-*score* mejora hasta aproximadamente los 200 SNPs.

Para revisar en detalle esta mejora, se realiza un nuevo experimento añadiendo de a una variante genética hasta los 200 SNPs. Los resultados de este experimento se muestran en la **Figura 4.18**.

Esta aproximación entrega resultados interesantes cuando se utiliza como selección de variantes el GWAS obtenido con los mismos datos del proyecto. El punto interesante que se obtiene es que la mejora en la métricas se da incluso después de añadir variantes que no son sugerentes de información, como se muestra en el panel superior de las **Figuras 4.17 y 4.18**, donde las variantes seleccionadas aparecen debajo de la línea azul ($p > 10^{-5}$). Esto sugiere que existe información revelante para la predicción de severidad en variantes que los GWAS no seleccionan.

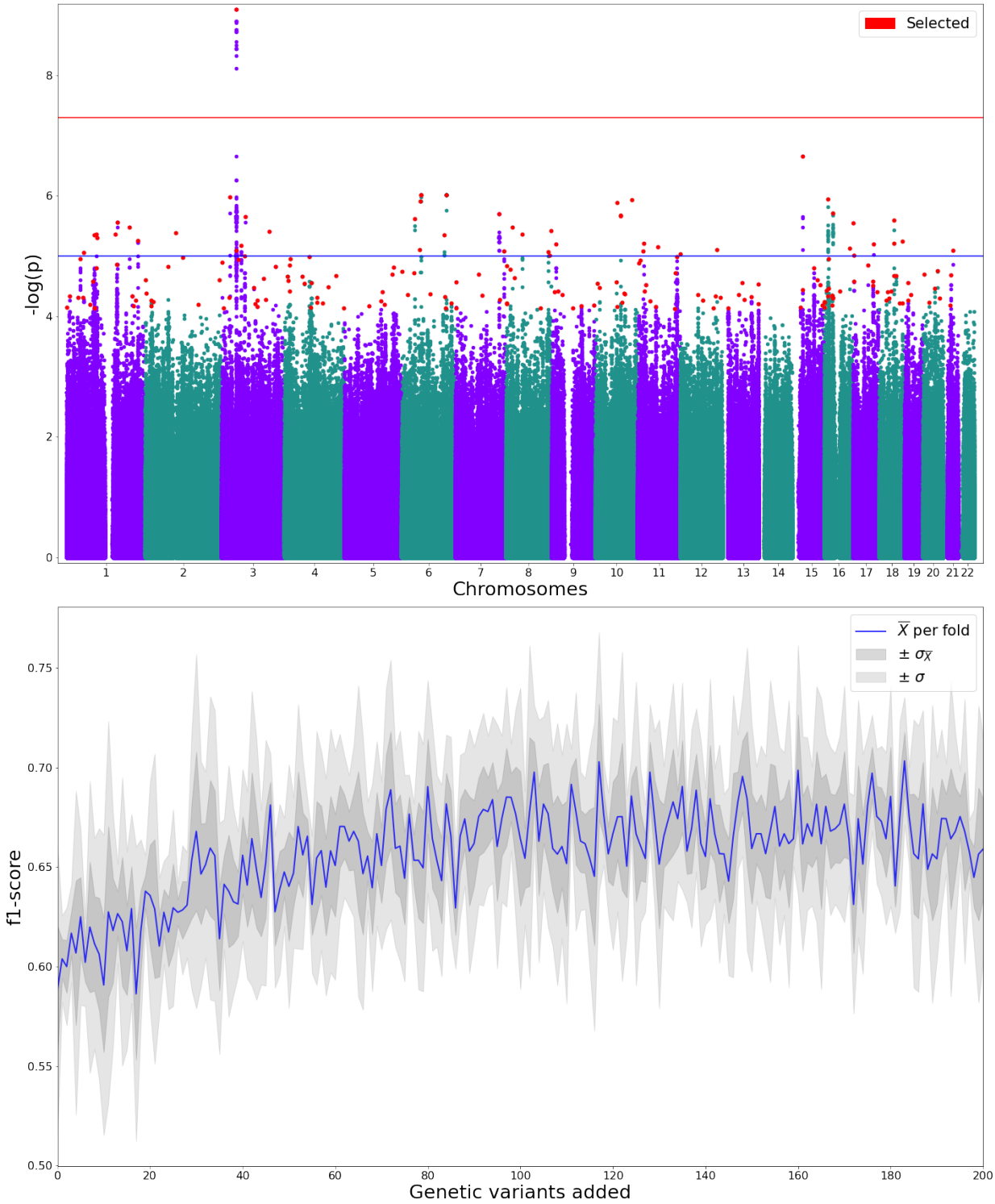


Figura 4.18: F1-score obtenido por el modelo XGBoost utilizando como entrada variables clínicas seleccionadas y una cantidad creciente de variantes genéticas hasta las doscientas. La salida del modelo es la hospitalización por Covid-19 entre participantes infectados con SARS-CoV-2. La selección de las variantes genéticas se realiza por la significancia obtenida por el GWAS generado a partir de los datos del proyecto, excluyendo variantes muy cercanas (menores o iguales a 250000 nucleótidos), de forma creciente. En la parte superior se incluye el GWAS de la iniciativa internacional con los SNPs utilizados en rojo.

Capítulo 5

Discusión y Conclusiones

En estas sección se discuten los resultados obtenidos en el Capítulo 4. Para presentar las conclusiones de este trabajo.

5.1. Discusión

En esta sección se discuten los resultados presentados en el **Capítulo 4**. Esta sección presenta las discusiones dependientes del tipo de datos usados: variables clínicas, variantes genéticas y ambas.

5.1.1. Datos clínicos autoreportados

El primer proceso al que se someten los datos clínicos es la estandarización de los datos por encuesta y por ficha clínica (CRF). Considerando las diferencias de ambas metodologías y los errores humanos de digitalización. Posterior a esto, se imputan los datos faltantes. En cuanto a este último paso (**Sección 4.1**), si bien los cambios son considerables, perder las variables numéricas o los participantes con entradas faltantes, reduce la ya baja cantidad de datos. En caso de los datos categóricos, como el grupo sanguíneo, agregar una categoría extra para datos nulos, agrega sesgo a los modelos [3]. De todos modos, estos cambios en la distribución se tienen presentes durante los análisis de resultados obtenidos. Recordar además que la imputación incluye otros participantes que no son incluídos para la secuenciación o que no entran en la categoría de caso o control para este análisis (**Sección 2.3.1**).

Posterior al procesamiento de imputación, se realiza un analisis estadístico sobre las variables clínicas. Para el test univariado (**Sección 4.1.1**), los diferentes grados de libertad (df) obtenidos usando los datos antes y después de la imputación, muestran que algunas categorías se agregan. Esto se debe a que las categorías se encuentran en el set de datos para otros participantes que no se incluyen entre aquellos que fueron genotipificados, lo que se puede comprabar en el **Anexo B.2.1** (**Tablas B.6 y B.7**). Mediante el test multivariado, se comprueba la importancia de la imputación para la recuperación de información de varias variables. De ambos métodos estadístico, univariado y multivariado, se obtiene un criterio de

selección de variables.

El criterio de selección de variables (**Tabla 4.3**) correspone a los dos test estadísticos y el *feature selection* en base a modelos ML. Como se aprecía en la **Figura 4.1**, la edad se correlaciona de forma notable con la severidad, lo que además se corresponde con la selección de esta variable según todos los criterios trabajados. Pasa lo mismo con la densidad poblacional, lo que ya ha sido reportado como determinante para incidencia y mortalidad del virus en nuestro país [40].

La selección de variables sí mejora las métricas obtenidas utilizando la mayoría de los modelos probados. Esto puede deberse a la reducción de variables que solo aportan ruido a la predicción. Por otro lado, las redes *fully connected* usadas no presentan la mejores métricas (**Figura 4.6**), probablemente por la poca cantidad de datos, ya que si bien la capacidad de estas redes sí capturan la información (**Figuras C.16 y C.17** sobre el *train set*), no generalizan tan bien como los otros modelos de ML.

5.1.2. Datos genéticos a nivel de genoma

Si bien la información genética por si sola no permite predecir la severidad del cuadro, las arquitecturas propuestas sí superan las métricas obtenidas con otros modelos (**Figura 4.8**). Se podrían utilizar otras arquitecturas que sí puedan aprender y generalizar sobre estos datos. La cantidad de variantes que se correlacionan con la severidad deben ser muchas menos que la cantidad de variantes utilizadas, por lo que se requiere una arquitectura con mayor capacidad y más datos para poder filtrar aquellas que aumentan el ruido en la predicción.

Comparando los resultados obtenidos por la arquitectura Dual-stream CNN en su propuesta original [38] con los resultados obtenidos en este trabajo, observamos (1) la similtud entre los mapas de saliencia y GWAS es similar, y (2) que las métricas obtenidas son mucho menores. Esto último, puede deberse a la influencia de otros factores en los fenotipos estudiados. Liu y colaboradores utilizaron un arreglo de SNPs de la haba de soya en la predicción de rendimiento, cantidad de proteína, producción de aceite, humedad y altura. Si bien, estas características son afectadas por el ambiente, los datos obtenidos se realizan sobre una población controlada en hidratación, nutrientes y edad de las plantas, algo que no se puede realizar sobre los participantes de este estudio.

5.1.3. Contribución de variantes genéticas a los modelos sobre variables clínicas

El actual procedimiento para determinar las variantes asociadas a fenotipos multifactoriales son los Estudios de Asociación de Genoma Completo (GWAS). Si bien, éstos no determinan las variantes genéticas que predisponen un fenotipo estudiado, sirve para reducir la cantidad de genes a estudiar en futuros estudios. Como las variantes genéticas corresponden a SNPs, que se pueden codificar como un arreglo de $\{0, 1, 2\}$ en una dimensión, estos pueden servir como entrada (o *input*) para modelos de aprendizaje que se suelen utilizar sobre secuencias.

Este trabajo busca utilizar dichos modelos para predecir la hospitalización en pacientes chilenos infectados con el virus SARS-CoV-2. Si bien, los modelos de redes neuronales empleados (Dual-stream CNN) funcionan mejor que otros modelos de ML, no obtiene mejores métricas que los modelos que se entrenan con variables clínicas. Si se compara el modelo que solo utiliza variantes seleccionadas con los modelos que utilizan el microarreglo o solo el cromosoma 3, parece ser que la cantidad de variantes sobreajusta el modelo, o las variantes no correlacionadas añaden ruido a la predicción.

Otro factor que puede explicar las menores métricas es la cantidad de datos. Las redes neuronales utilizan grandes cantidades de datos [21], sin embargo, para este proyecto, se tienen solo 1872 participantes. Los modelos se entrenan entonces con un *dataset* de pocas instancias y muchas variables, lo que genera *overfitting* en la predicción. Por una parte, seleccionar variantes genéticas, de forma de reducir el tamaño de la secuencia de entrada, mejora las métricas. Por otra parte, obtener más participantes disminuirá el *overfitting* y a la vez, se espera que mejoren las métricas obtenidas.

Este último punto no es prácticamente abordable para este tipo de estudios. La alternativa puede ser aumentar artificialmente la cantidad de participantes, dado que tenemos como antecedente que mediante *oversampling* se obtienen mejores métricas con el mismo modelo y datos de entrada. Otra alternativa es obtener datos de otros países y utilizar *Transfer Learning*, para ajustar el modelo a los participantes chilenos. Este no se realiza, ya que los datos no agregados de los participantes de la iniciativa internacional no se encuentran disponibles, debido a que la iniciativa realiza un meta-análisis.

Para abordar la falta de participantes, también es una opción realizar pre-entrenamiento sobre una tarea menos específica y luego realizar *Fine Tuning* sobre la tarea específica. Esta aproximación es utilizada en NLP mediante *Self-Attention* [16, 59] y ha sido utilizada sobre secuencias completas (**Sección 2.1.2.1**, **Ecuación 2.2**) en el trabajo de Ji y colaboradores [30], donde se genera el modelo preentranado DNABERT. Si bien, con los datos disponibles se puede obtener el genoma como secuencia completa, la cantidad de nucleótidos o letras diferentes es muy baja. DNABERT solo recibe 512 *tokens*, los que en el genoma humano y en los SNPs del proyecto implican máximo 2 nucleótidos diferentes en la secuencia entregada.

Generar una secuencia corta ($\sim$ 512 nucleótidos) en base a los SNPs obtenidos y que tenga efectivamente más *tokens* diferentes, se puede realizar concatenando los nucleótidos de los variantes. Esto genera dos problemas. Por una parte, la selección previa de los SNPs a utilizar, que agrega sesgo al entrenamiento. Por otra parte, la concatenación de los nucleótidos no necesariamente presenta la misma distribución ni patrones que la secuencia completa en el genoma.

Pese a que las métricas obtenidas por las arquitecturas propuestas no superan las de modelos entrenados con variables clínicas, este trabajo reporta algunos resultados útiles, particularmente en la selección de variantes significativas no reportadas. Los principales desafíos que quedan por realizar es aumentar la cantidad de datos a utilizar, ya sea mediante *transfer learning* con *datasets* más grandes o por *data augmentation*. Futuras investigaciones podrían

incorporar la utilización de redes *Transformers*, la cuales requieren secuencias más cortas o mayor poder computacional. Se puede generar un nuevo *Transformer* pre-entrenado utilizando solo las diferencias en el genoma, considerando que existen bases de datos para esto. Utilizar el preentrenamiento de los modelos basados en *Self-Attention*, requiere menos datos para realizar *fine tuning*, que es una de las limitantes de este tipo de trabajo.

5.2. Conclusiones

Un modelo de aprendizaje automático, particularmente XGBoost, permite predecir la severidad de un cuadro de Covid-19 utilizando datos clínicos autoreportados. Incluso, mediante interpretación por *Feature Importance*, se puede seleccionar variables clínicas y mejorar la predicción realizada. Esto sugiere que existen variables como la edad, nivel socieconómico, densidad poblacional y algunas comorbilidades que se correlacionan con la hospitalización causada por el virus SARS-CoV-2.

Modelos de redes neuronales también son capaces de predecir la severidad utilizando el mismo tipo de datos. Sin embargo, el aporte de los datos genéticos no mejora la predicción cuando se utilizan algoritmos de redes neuronales. Una posible explicación de esto es metodológica, ya que la cantidad de participantes es muy poca para que este tipo de algoritmo sea generalizable.

Pese a esto, al calcular las variantes a que la red entrenada pone atención, se encuentra una variante que podría correlacionarse con la hospitalización. Esta variante se encuentra físicamente dentro del gen FOXP4-AS1 que codifica un mRNA que regula de forma antisentido al factor de transcripción FOXP4.

Otra evidencia de que la información genética sí puede contener información asociada a la hospitalización por Covid-19, es que, al añadir variantes previamente seleccionadas según la información de asociación genómica a los datos de entrenamiento de un modelo ML, la predicción mejora. Esta mejora ocurre incluso al agregar variantes genéticas, que según la metodología actual, no son significantes. Esto muestra que este modelo es capaz de capturar información genética que la actual metodología no puede.

Bibliografía

[1] Vikram Agarwal and Jay Shendure. Predicting mRNA Abundance Directly from Genomic Sequence Using Deep Convolutional Neural Networks. *Cell Reports*, 31(7):107663, May 2020.

[2] Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. *Molecular biology of the cell*. Garland Science, New York, 5th ed edition, 2008. OCLC: ocm82473851.

[3] John Barnard and Xiao-Li Meng. Applications of multiple imputation in medical studies: from AIDS to NHANES. *Statistical Methods in Medical Research*, 8(1):17–36, February 1999.

[4] Andreas Bolzer, Gregor Kreth, Irina Solovei, Daniela Koehler, Kaan Saracoglu, Christine Fauth, Stefan Müller, Roland Eils, Christoph Cremer, Michael R Speicher, and Thomas Cremer. Three-Dimensional Maps of All Chromosomes in Human Male Fibroblast Nuclei and Prometaphase Rosettes. *PLoS Biology*, 3(5):e157, April 2005.

[5] V. Brendel and H.G. Busse. Genome structure described by formal languages. *Nucleic Acids Research*, 12(5):2561–2568, 1984.

[6] CDC. Omicron Variant: What You Need to Know, February 2022.

[7] Yi Chang, Xin Jing, Zhao Ren, and Björn W. Schuller. CovNet: A Transfer Learning Framework for Automatic COVID-19 Detection From Crowd-Sourced Cough Sounds. *Frontiers in Digital Health*, 3:799067, January 2022.

[8] Tianqi Chen and Carlos Guestrin. XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, August 2016. arXiv: 1603.02754.

[9] Jim Clauwaert, Gerben Menschaert, and Willem Waegeman. Explainability in transformer models for functional genomics. *Briefings in Bioinformatics*, 22(5):bbab060, September 2021.

[10] Jim Clauwaert and Willem Waegeman. Novel Transformer Networks for Improved Sequence Labeling in genomics. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 19(1):97–106, January 2022.

[11] COVID-19 Host Genetics Initiative. Analysis plan (Version 1.1), August 2020.

[12] COVID-19 Host Genetics Initiative. Mapping the human genetic architecture of COVID-19. *Nature*, 600(7889):472–477, December 2021.

[13] Zhicheng Cui, Wenlin Chen, and Yixin Chen. Multi-Scale Convolutional Neural Networks for Time Series Classification. 2016. Publisher: arXiv Version Number: 4.

[14] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V. Le, and Ruslan Salakhutdinov. Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context. *arXiv:1901.02860 [cs, stat]*, June 2019. arXiv: 1901.02860.

[15] Lea Davis, Andrea Ganna, Sulggi Lee, James Priest, Alessandra Renieri, Vijay Sankaran, David van Heel, Patrick Deelen, Brent Richards, Tomoko Nakanishi, Les Biesecker, and Eric Kerchberger. Phenotype definitions for analyses, COVID19 Host Genetics Initiative (Version 2.0), October 2020.

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*, May 2019. arXiv: 1810.04805.

[17] Ensheng Dong, Hongru Du, and Lauren Gardner. An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious Diseases*, 20(5):533–534, May 2020.

[18] Gökcen Eraslan, Žiga Avsec, Julien Gagneur, and Fabian J. Theis. Deep learning: new computational modelling techniques for genomics. *Nature Reviews Genetics*, 20(7):389–403, July 2019.

[19] Alessandro Ferrari, Stefano Lombardi, and Alberto Signoroni. Bacterial colony counting with Convolutional Neural Networks in Digital Microbiology Imaging. *Pattern Recognition*, 61:629–640, January 2017.

[20] Genome Reference Consortium. GRCh38.p14, February 2019.

[21] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. Adaptive computation and machine learning. The MIT Press, Cambridge, Massachusetts, 2016.

[22] Google Inc. tensorboard: TensorBoard lets you watch Tensors Flow, January 2022.

[23] Gaurav Gupta and Shubhi Saini. DAVI:Deep Learning Based Tool for Alignment and Single Nucleotide Variant identification. preprint, Bioinformatics, September 2019.

[24] Brian Hie, Ellen D. Zhong, Bonnie Berger, and Bryan Bryson. Learning the language of viral evolution and escape. *Science*, 371(6526):284–288, January 2021.

[25] Melissa M Higdon, Brian Wahl, Carli B Jones, Joseph G Rosen, Shaun A Truelove, Anurima Baidya, Anjalika A Nande, Parisa A ShamaeiZadeh, Karoline K Walter, Daniel R Feikin, Minal K Patel, Maria Deloria Knoll, and Alison L Hill. A systematic review of COVID-19 vaccine efficacy and effectiveness against SARS-CoV-2 infection and disease. preprint, Epidemiology, September 2021.

[26] Steven T. Hill, Rachael Kuintzle, Amy Teegarden, Erich Merrill, Padideh Danaee, and

David A. Hendrix. A Deep Recurrent Neural Network Discovers Complex Biological Rules to Decipher RNA Protein-Coding Potential. preprint, Bioinformatics, October 2017.

[27] Daniel Sik Wai Ho, William Schierding, Melissa Wake, Richard Saffery, and Justin O'Sullivan. Machine Learning SNP Based Prediction for Precision Medicine. *Frontiers in Genetics*, 10:267, March 2019.

[28] Kelly Hughes and Stanley R. Maloy. *Brenner's encyclopedia of genetics*. Academic Press, Elsevier Science, London [u.a.], 2. ed edition, 2013.

[29] Kei Ishida, Ali Ercan, Takeyoshi Nagasato, Masato Kiyama, and Motoki Amagasaki. Use of 1D-CNN for input data size reduction of LSTM in Hourly Rainfall-Runoff modeling. *arXiv:2111.04732 [physics]*, November 2021. arXiv: 2111.04732.

[30] Yanrong Ji, Zhihan Zhou, Han Liu, and Ramana V Davuluri. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics*, 37(15):2112–2120, August 2021.

[31] Yoshinori Kabeya, Mariko Okubo, Sho Yonezawa, Hiroki Nakano, Michio Inoue, Masashi Ogasawara, Yoshihiko Saito, Jantima Tanboon, Luh Ari Indrawati, Theerawat Kumutpongpanich, Yen-Lin Chen, Wakako Yoshioka, Shinichiro Hayashi, Toshiya Iwamori, Yusuke Takeuchi, Reitaro Tokumasu, Atsushi Takano, Fumihiko Matsuda, and Ichizo Nishino. Deep convolutional neural network-based algorithm for muscle biopsy diagnosis. *Laboratory Investigation*, 102(3):220–226, March 2022.

[32] Priyansh Kedia, Anjum, and Rahul Katarya. CoVNet-19: A Deep Learning model for the detection and analysis of COVID-19 patients. *Applied Soft Computing*, 104:107184, June 2021.

[33] David R. Kelley. Cross-species regulatory sequence activity prediction. *PLOS Computational Biology*, 16(7):e1008050, July 2020.

[34] David R. Kelley, Yakir A. Reshef, Maxwell Bileschi, David Belanger, Cory Y. McLean, and Jasper Snoek. Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome Research*, 28(5):739–750, May 2018.

[35] Samuel Lalmuanawma, Jamal Hussain, and Lalrinfela Chhakchhuak. Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic: A review. *Chaos, Solitons & Fractals*, 139:110059, October 2020.

[36] Yann LeCun, Koray Kavukcuoglu, and Clement Farabet. Convolutional networks and applications in vision. In *Proceedings of 2010 IEEE International Symposium on Circuits and Systems*, pages 253–256, Paris, France, May 2010. IEEE.

[37] Wei Li, Yanbin Yin, Xiongwen Quan, and Han Zhang. Gene Expression Value Prediction Based on XGBoost Algorithm. *Frontiers in Genetics*, 10:1077, November 2019.

[38] Yang Liu, Duolin Wang, Fei He, Juexin Wang, Trupti Joshi, and Dong Xu. Phenoty-

pe Prediction and Genome-Wide Association Study Using Deep Convolutional Neural Network of Soybean. *Frontiers in Genetics*, 10:1091, November 2019.

[39] Marco Marani, Gabriel G. Katul, William K. Pan, and Anthony J. Parolari. Intensity and frequency of extreme novel epidemics. *Proceedings of the National Academy of Sciences*, 118(35):e2105482118, August 2021.

[40] Gonzalo E. Mena, Pamela P. Martinez, Ayesha S. Mahmud, Pablo A. Marquet, Caroline O. Buckee, and Mauricio Santillana. Socioeconomic status determines COVID-19 incidence and related mortality in Santiago, Chile. *Science*, 372(6545):eabg5298, May 2021.

[41] Yunfeng Niu, Gaoyan Wang, Yan Li, Wei Guo, Yanli Guo, and Zhiming Dong. LncRNA FOXP4-AS1 Promotes the Progression of Esophageal Squamous Cell Carcinoma by Interacting With MLL2/H3K4me3 to Upregulate FOXP4. *Frontiers in Oncology*, 11:773864, December 2021.

[42] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[43] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011.

[44] Shaun Purcell. Plink, October 2009.

[45] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A.R. Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul I.W. de Bakker, Mark J. Daly, and Pak C. Sham. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics*, 81(3):559–575, September 2007.

[46] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021.

[47] David B. Searls. The language of genes. *Nature*, 420(6912):211–217, November 2002.

[48] Paulina Sepúlveda. ¿Existe variación genética de la población en Chile que influya en la gravedad de Covid-19? - La Tercera. *La Tercera*, September 2020. Place: Chile Section: Qué Pasa.

[49] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. *arXiv:1312.6034 [cs]*, April 2014. arXiv: 1312.6034.

[50] Eleni Smitham and Amanda Glassman. The Next Pandemic Could Come Soon and Be Deadlier, August 2021.

[51] Shelly Soffer, Avi Ben-Cohen, Orit Shimon, Michal Marianne Amitai, Hayit Greenspan, and Eyal Klang. Convolutional Neural Networks for Radiologic Images: A Radiologist's Guide. *Radiology*, 290(3):590–606, March 2019.

[52] Catrin Sohrabi, Zaid Alsafi, Niamh O'Neill, Mehdi Khan, Ahmed Kerwan, Ahmed Al-Jabir, Christos Iosifidis, and Riaz Agha. World Health Organization declares global emergency: A review of the 2019 novel coronavirus (COVID-19). *International Journal of Surgery*, 76:71–76, April 2020.

[53] Tri Songz. Coronavirus Sequence Prediction with Transformer Models, 2020.

[54] Bernardo Aníbal Subercaseax Roa. Model Interpretability Through The Lens Of Computational Complexy. Master's thesis, Universidad de Chile, Santiago, Chile, 2020.

[55] Vivian Tam, Nikunj Patel, Michelle Turcotte, Yohan Bossé, Guillaume Paré, and David Meyre. Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics*, 20(8):467–484, August 2019.

[56] The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*, 526(7571):68–74, October 2015.

[57] L. Torgo. *Data Mining with R, learning with case studies*. Chapman and Hall/CRC, 2010.

[58] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman. Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17(6):520–525, June 2001.

[59] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. *arXiv:1706.03762 [cs]*, December 2017. arXiv: 1706.03762.

[60] W. N. Venables, Brian D. Ripley, and W. N. Venables. *Modern applied statistics with S*. Statistics and computing. Springer, New York, 4th ed edition, 2002. OCLC: ocm49312402.

[61] Frank R Wendt, Antonella De Lillo, Gita A Pathak, Flavio De Angelis, COVID-19 Host Genetics Initiative, and Renato Polimanti. Host Genetic Liability for Severe COVID-19 Associates with Alcohol Drinking Behavior and Diabetic Outcomes in Participants of European Descent. *Frontiers in Genetics*, 12:765247, December 2021.

[62] Hadley Wickham, Mara Averick, Jennifer Bryan, Winston Chang, Lucy McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn,

Thomas Pedersen, Evan Miller, Stephan Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani. Welcome to the Tidyverse. *Journal of Open Source Software*, 4(43):1686, November 2019.

[63] Auriel A. Willette, Sara A. Willette, Qian Wang, Colleen Pappas, Brandon S. Klinedinst, Scott Le, Brittany Larsen, Amy Pollpeter, Tianqi Li, Jonathan P. Mochel, Karin Allenspach, Nicole Brenner, and Tim Waterboer. Using machine learning to predict COVID-19 infection and severity risk among 4510 aged adults: a UK Biobank cohort study. *Scientific Reports*, 12(1):7736, December 2022.

[64] Alencar Xavier, Shizhong Xu, William M. Muir, and Katy Martin Rainey. NAM: association studies in multiple populations: Fig. 1. *Bioinformatics*, page btv448, August 2015.

[65] Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. Big Bird: Transformers for Longer Sequences. July 2020.

[66] Jesse M. Zhang. Learning the Language of the Genome using RNNs, 2016.

[67] Li Zhaoping. *Understanding Vision: Theory, Models, and Data*. Oxford University PressOxford, 1 edition, May 2014.

[68] Jian Zhou, Chandra L. Theesfeld, Kevin Yao, Kathleen M. Chen, Aaron K. Wong, and Olga G. Troyanskaya. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nature Genetics*, 50(8):1171–1179, August 2018.

[69] Yazeed Zoabi, Shira Deri-Rozov, and Noam Shomron. Machine learning-based prediction of COVID-19 diagnosis based on symptoms. *npj Digital Medicine*, 4(1):3, December 2021.

Anexo A

Glosario

alelo Una o más versiones de una secuencia de DNA, ya sea de uno o más nucleótidos, en una posición en el genoma. Un individuo herada dos alelos, uno de cada padre. 6

aminoácido es una molécula orgánica formada por un grupo carboxilo, una amina, un hidrógeno y una residuo unido a un carbono central. El residuo puede corresponde a distintos grupos funcionales, los que, en su secuencia, determinan la propiedades fisicoquímicas de la proteína que conforman. 5

base nitrogenada es un compuesto nitrogenado, capaz de funcionar como base y de formar puentes de hidrógeno con otras bases nitrogenadas. El DNA contiene cuatro tipos de bases nitrogenadas: Adenina (A) que forma dos puentes de hidrógeno con Timina (T) y Guanina (G) que forma tres puentes de hidrógeno con Citosina (C). El RNA contiene también cuatro tipos de bases, pero se reemplaza la Timina (T) con el Uracilo (U). 5

centrómero zona de un cromosoma que une al par de cromátidas hermanas durante la división celular. Esta es una región altamente condensada. Durante la mitosis, el huso mitótico se una a los centrómeros de los cromosomas en una estructura proteíca llamada cinetocoro. 31

codón tres nucleótidos que se corresponden con un aminoácido, con la señal de iniciar la traducción de una proteína o de deternela. Esta secuencia de tre nucleótidos se encuentra en la hebra codificante del DNA y en RNA mensajero se usa como molde del ribosoma (intercambiando la T por la U). El codón es la unidad mínima del código genético. 5

cromosoma es una molécula larga de DNA condensada que se forma durante los procesos de mitosis y meiosis para transportar la información genética a las células hijas. Se compone de la molécula de DNA y de proteínas que mantienen la condensación. 5–8

cromosoma homólogo en los organismos diploides, como el humano, son pares de cromosomas de igual largo

aproximado, misma posición de su centrómero y que contienen los mismos loci, que son posiciones en el genoma que codifican los mismos genes. Se identifican porque durante la profase I de la meiosis se encuentran apareados antes de migrar a las dos células resultantes durante la anafase I. 6, 7

densidad poblacional Indicador de la cantidad de personas en una área geográfica determinada. Se expresa como la cantidad de habitantes por kilómetro cuadrado. 42–44

DNA ácido desoxirribonucleico (DNA, por sus siglas en inglés, *Desoxiribonucleic Acid*). Es una molécula formada por dos cadenas de nucleótidos unidos por enlace fosfodiester. Las dos cadenas se encuentran unidas por puentes de hidrógeno. 4, 5

fecundación Compleja secuencia de eventos que involucran la unión de dos gametos haploides [28]. Cada gameto haplode contiene un *set* de cromosomas, de forma que el cigoto posee la información de dos *set* de cromosomas. Cada par de cromosomas equivalente, se conoce como par homólogo. 6

fenotipo Características observables de un organismo, determinadas por su genotipo y ambiente. 2, 4

gen unidad mínima de herencia. Se traspasa de padres a progenie y determina alguna caracterísitica de la progenie. En términos técnicos corresponde a una secuencia que codifica un producto génico que puede ser una proteína o una molécula de RNA. 4

genoma es toda la información genética en un organismo. Tiene toda la información que un organismo requiere para funcionar. 4–9

IMC Índice de Masa Corporal (BMI, en inglés por *Body Mass Index*) es un indicador de la relación peso/altura de una persona. Se calcula como $IMC = {}^{\text{Peso } [kg]}/_{\text{Altura } [m^2]}$. 42

locus Posición física en el genoma, y en un cromosoma en particular, que puede corresponder a un gen o otro segmento de interés. El plural es loci. i, 6, 57

Manhattan plot Visualización de la significancia estadísitica de variantes genéticas. En las ordenadas se mapea la significancia, mientras que las abscisas mapean la posición de cada variante. Los SNPs se ordenan según su cromosoma y posición en el genoma. Por lo general, cuando se visualiza el $p - value$ se grafica el $-\log(p)$, para que la significancia se vea de forma creciente. ix, 9, 20–23, 53, 56, 58, 61

nucleótido es una molécula orgánica compuesta de una base nitrogenada y un fosfato inorgánico unido a un azúcar ribonucleico, en caso de formar una cadena de RNA, o desoxirribonucleico en caso de formar una cadena de DNA. 5–8, 69, 73

proteína es una molécula orgánica nitrogenada formada de una o más cadenas de aminoácidos. En un organismo realiza actividad enzimática como catalizador de reacciones meta-

bólicas o siendo parte de la estructura de la célula. 5

qqplot También llamado gráfico cuantil-cuantil. Se visuliza la comparación de dos distribuciones de probabilidad. En el caso de la significancia estadísitica de variantes genéticas se visualiza el valor obtenido, en orden, con el valor esperado por la hipótesis nula para una distribución χ^2. 9, 20–23

RNA ácido ribonucleico (RNA, por sus siglas en inglés, *Ribonucleic Acid*). Es una molécula formada por una cadena de nucleótidos unidos por enlace fosfodiester. En el funcionamiento celular es un intermediario entre el DNA y una proteína. 4, 5, 65, 66

telómero Zona del cromosoma con secuencias altamente repetidas que se encuentran al final de la secuencia lineal de los cromosomomas, tanto cuando se encuentran descondensada en el núcleo celular como en su forma compacta como cromosoma. 31

variante monomórfica Variantes genéticas que solo aparecen en un estado. Para el caso de SNPs, son aquellas variantes que presentan el mismo valor ($\{0, 1, 2\}$) para todos los participantes. 19

Anexo B

Tablas datos clínicos

B.1. Descripción de Encuesta y CRF

Desde la encuesta se estandarizaron 147 columnas, estas variables, y aquellas disponibles en la ficha clínica (CRF) se muestran en las **Tablas B.1-B.5**.

Tabla B.1: Resumen de variables estandarizadas separadas por categoría, extraídas de la encuesta. La disponibilidad en otros sets de datos se muestra en íconos verdes, mientras que los íconos rojos indican si éstas no están disponibles.

Columna	Tipo	Categorías	Encuesta	CRF	Final
Identificadores					
id	Categórico	2665	✅	✅	✅
Código universal	Categórico	2641	✅	✅	❌
Centro	Categórico	14	✅	✅	❌
Fecha de Aplicación	Fecha	–	✅	✅	❌
Información personal					
Edad	Númerico	–	✅	✅	✅
Sexo	Categórico	2	✅	✅	✅
Grupo Sanguíneo	Categórico	4	✅	✅	✅
Grupo Rh	Categórico	2	✅	✅	✅
Peso	Numérico	–	✅	✅	✅
Altura	Numérico	–	✅	✅	✅
IMC	Numérico	–	✅	✅	✅
Geográficos					
Nacionalidad	Categórico	15	✅	✅	✅
Región	Categórico	16	✅	✅	❌
Provincia	Booleano	2	✅	✅	✅
Macrozona	Categórico	5	✅	✅	❌
Comuna	Categórico	99	✅	✅	❌
Densidad	Númerico	–	✅	✅	✅

Tabla B.2: Continuación **Tabla B.1**

Columna	Tipo	Categorías	Encuesta	CRF	Final
		Geográficos			
País de Nacimiento	Categórico	21	✓	✓	✓
Región de Nacimiento	Categórico	16	✓	✗	✗
Comuna de Nacimiento	Categórico	177	✓	✗	✗
País de Nacimiento Padre	Categórico	22	✓	✗	✗
Región de Nacimiento Padre	Categórico	16	✓	✗	✗
Comuna de Nacimiento Padre	Categórico	231	✓	✗	✗
País de Nacimiento Madre	Categórico	16	✓	✗	✗
Región de Nacimiento Madre	Categórico	16	✓	✗	✗
Comuna de Nacimiento Madre	Categórico	224	✓	✗	✗
		Socieconómico			
Nivel Educacional	Categórico	10	✓	✓	✓
Situación ocupacional	Categórico	3	✓	✓	✓
Sistema Salud	Categórico	2	✓	✓	✓
		Etnicidad			
Ancestría	Categórico	15	✓	✗	✗
Etnicidad	Categórico	14	✓	✓	✓
		Hábitos de Salud			
Consumo Tabaco	Categórico	5	✓	✓	✓
Cigarros diarios	Categórico	6	✓	✓	✓
Más de 100 Cigarros	Booleano	2	✓	✓	✓
Consumo alcohol	Categórico	5	✓	✓	✓
Marihuana	Booleano	2	✓	✓	✓
Cocaína	Booleano	2	✓	✓	✓
Pasta base	Booleano	2	✓	✓	✓
Heroína	Booleano	2	✓	✓	✓
Alucinógenos	Booleano	2	✓	✓	✓
Alguna Otra Droga	Booleano	2	✓	✓	✓
Droga NS/NR	Booleano	2	✓	✓	✓
		Síntomas			
Fiebre	Booleano	2	✓	✓	✗
Escalofríos	Booleano	2	✓	✓	✗
Tos seca persistente	Booleano	2	✓	✓	✗
Tos nueva productiva	Booleano	2	✓	✓	✗
Disnea	Booleano	2	✓	✓	✗
Vómitos	Booleano	2	✓	✓	✗
Diarrea	Booleano	2	✓	✓	✗
Anosmia	Booleano	2	✓	✓	✗
Ageusia	Booleano	2	✓	✓	✗
Opresión en el pecho	Booleano	2	✓	✓	✗
Dolor de garganta	Booleano	2	✓	✓	✗

Tabla B.3: Continuación **Tabla B.2**

Columna	Tipo	Categorías	Encuesta	CRF	Final
		Síntomas			
Cansancio	Booleano	2	✓	✓	✗
Incapacidad de moverse	Booleano	2	✓	✓	✗
Dolor abdominal	Booleano	2	✓	✓	✗
Cefalea	Booleano	2	✓	✓	✗
Mareos	Booleano	2	✓	✓	✗
Dolor en la espalda baja	Booleano	2	✓	✓	✗
Mialgia	Booleano	2	✓	✓	✗
Dolor al respirar	Booleano	2	✓	✓	✗
Dificultad para respirar	Booleano	2	✓	✓	✗
Sensación Calor y Frío	Booleano	2	✓	✗	✗
Entumecimiento	Booleano	2	✓	✓	✗
Decaimiento	Booleano	2	✓	✓	✗
Sensación de extremidades pesadas	Booleano	2	✓	✗	✗
Congestión nasal	Booleano	2	✓	✓	✗
Moquillo	Booleano	2	✓	✗	✗
Ojos rojos	Booleano	2	✓	✓	✗
Estornudos	Booleano	2	✓	✗	✗
Eritema	Booleano	2	✓	✓	✗
Sarpullido	Booleano	2	✓	✗	✗
Decoloración de la piel	Booleano	2	✓	✓	✗
Erupciones en boca	Booleano	2	✓	✓	✗
Erupciones cutáneas	Booleano	2	✓	✓	✗
Otros Síntomas	Booleano	2	✓	✓	✗
Neumonía	Booleano	2	✓	✓	✗
Neumonía (Unilateral)	Booleano	2	✓	✓	✗
Neumonía (Bilateral)	Booleano	2	✓	✓	✗
		Hospitalización			
Hospitalización	Booleano	2	✓	✓	✗
<21 días	Booleano	2	✓	✓	✗
Unidad Hospitalaria	Categórico	6	✓	✓	✗
Respiración asistida	Categórico	4	✓	✓	✗
		Comorbilidades			
Diabetes	Booleano	2	✓	✓	✓
Aterosclerosis coronaria	Booleano	2	✓	✓	✓
Hipertensión	Booleano	2	✓	✓	✓

Tabla B.4: Continuación **Tabla B.3**

Columna	Tipo	Categorías	Encuesta	CRF	Final
Comorbilidades					
Accidente vascular	Booleano	2	✓	✓	✓
Problema al corazón	Booleano	2	✓	✓	✓
Diálisis	Booleano	2	✓	✓	✓
Hepatitis	Booleano	2	✓	✓	✓
Anemia	Booleano	2	✓	✓	✓
Asma	Booleano	2	✓	✓	✓
Fibrosis quística	Booleano	2	✓	✓	✓
Fibrosis pulmonar	Booleano	2	✓	✓	✓
Enfisema	Booleano	2	✓	✓	✗
Afección Pulmonar	Booleano	2	✓	✓	✓
Cáncer	Booleano	2	✓	✓	✓
Condición que afecta el cerebro	Booleano	2	✓	✓	✓
VIH	Booleano	2	✓	✓	✓
Tuberculosis	Booleano	2	✓	✓	✓
Sistema inmunitario debilitado	Booleano	2	✓	✓	✓
Problemas de salud mental	Booleano	2	✓	✓	✓
Colesterol elevado	Booleano	2	✓	✓	✓
Obesidad	Booleano	2	✓	✓	✓
Colitis ulcerosa o enfermedad de Crohn	Booleano	2	✓	✓	✓
Artritis reumatoide	Booleano	2	✓	✓	✓
Lupus	Booleano	2	✓	✓	✓
Otras enfermedades reumatoides	Booleano	2	✓	✓	✓
Tuve trasplante de algún órgano	Booleano	2	✓	✗	✗
Otras enfermedades crónicas	Booleano	2	✓	✓	✓
Cáncer					
Cáncer de colon y recto	Booleano	2	✓	✓	✓
Cáncer de endometrio	Booleano	2	✓	✓	✓
Cáncer de Hígado	Booleano	2	✓	✓	✗
Leucemia	Booleano	2	✓	✓	✗
Linfoma no Hodgkin	Booleano	2	✓	✓	✗
Melanoma	Booleano	2	✓	✓	✗
Cáncer de páncreas	Booleano	2	✓	✓	✓
Cáncer de próstata	Booleano	2	✓	✓	✓
Cáncer de pulmón	Booleano	2	✓	✓	✓
Cáncer de riñón	Booleano	2	✓	✓	✓
Cáncer de mama	Booleano	2	✓	✓	✓
Cáncer de tiroides	Booleano	2	✓	✓	✓
Cáncer de vejiga	Booleano	2	✓	✓	✗
Otro Tipo de Cáncer	Booleano	2	✓	✓	✗
Cáncer de vesícula	Booleano	2	✓	✓	✓
Cáncer NS/NR	Booleano	2	✓	✓	✗

Tabla B.5: Continuación **Tabla B.4**

Columna	Tipo	Categorías	Encuesta	CRF	Final
Actividades					
Actividades de la vida diaria	Booleano	2	✓	✗	✗
Alguien que le ayude	Booleano	2	✓	✗	✗
Quedarse en casa	Booleano	2	✓	✗	✗
Drogas					
Antidiabético	Booleano	2	✓	✓	✓
Anti-TB	Booleano	2	✓	✓	✓
Antihipertensivo	Booleano	2	✓	✓	✓
Aspirina o clopidogrel	Booleano	2	✓	✓	✓
Estatina	Booleano	2	✓	✓	✓
Analgésicos	Booleano	2	✓	✓	✓
Medicación para la tos	Booleano	2	✓	✓	✓
Inhalador	Booleano	2	✓	✓	✓
Anticancerígeno	Booleano	2	✓	✓	✓
Antimaláricos	Booleano	2	✓	✓	✓
Antivirales	Booleano	2	✓	✓	✓
Vitaminas	Booleano	2	✓	✓	✓
Hierbas medicinales	Booleano	2	✓	✓	✓
Ninguna droga	Booleano	2	✓	✓	✓
Otro Medicamento	Booleano	2	✓	✓	✓
Medicamento NS/NR	Booleano	2	✓	✓	✓
Atención Crónica					
Atención crónica	Categórico	4	✓	✗	✗
Confirmación o infección					
Confirmado	Booleano	2	✓	✓	✗
Contagiado	Booleano	2	✓	✓	✗
Índice de Riesgo					
Índice de Riesgo	Categórico	7	✓	✓	✗

B.2. Análisis Estadístico

B.2.1. Resumen de Variables

Para las variables categóricas se reporta la cantidad de participantes obtenidos para cada categoría, separadas por aquellos que se etiquetan como controles (pacientes ambulatorios), casos (pacientes hospitalizados) y el total (detalle del análisis de severidad para hospitalizados en la Sección 2.3.1). Se muestran estos resumenes para los datos antes y después de la imputación por KNN, etiquetados como estandarizados (*standardized*) e imputados (*imputed*) respectivamente. Estás son las **Tablas B.6-B.17**.

Tabla B.6: Cantidad de participantes para cada categoría de la variable Nacionalidad (*Nationality*). Se reportan los participantes separados por caso, control y el total para el análisis de hospitalizados. A la izquierda se muestran los participantes del dataset luego de la estandarización, a la derecha, se encuentran los participantes obtenidos luego de la imputación por KNN.

	Standardized			Imputed		
	Cases	Controls	Total	Cases	Controls	Total
Chile	306	1548	1854	337	1556	1893
Ecuador	1	0	1	1	0	1
Other	1	0	1	1	0	1
Peru	8	0	8	8	0	8
Venezuela	8	0	8	8	0	8
Argentina	-	-	-	1	0	1
NA's	32	8	40	-	-	-

Tabla B.7: Cantidad de participantes para cada categoría de la variable País de Nacimiento (*Country of Birth*). Se reportan los participantes separados por caso, control y el total para el análisis de hospitalizados. A la izquierda se muestran los participantes del dataset luego de la estandarización, a la derecha, se encuentran los participantes obtenidos luego de la imputación por KNN.

	Standardized			Imputed		
	Cases	Controls	Total	Cases	Controls	Total
Argentina	0	5	5	2	8	10
Brazil	0	1	1	0	1	1
Chile	303	1529	1832	303	1529	1832
Ecuador	1	0	1	1	0	1
France	0	1	1	0	1	1
Other	1	0	1	1	0	1
Peru	8	3	11	8	3	11
Russia	0	1	1	0	1	1
Switzerland	0	1	1	0	1	1
Venezuela	8	4	12	8	4	12
EE.UU	1	0	1	1	0	1
NA's	2	3	5	-	-	-

Tabla B.8: Cantidad de participantes para cada categoría de la variable Situación Ocupacional (*Occupational Situation*). Se reportan los participantes separados por caso, control y el total para el análisis de hospitalizados. A la izquierda se muestran los participantes del dataset luego de la estandarización, a la derecha, se encuentran los participantes obtenidos luego de la imputación por KNN.

	Standardized			Imputed		
	Cases	Controls	Total	Cases	Controls	Total
Active	123	1159	1282	223	1181	1404
Inactive	64	288	352	92	288	380
Unemployed	9	79	88	9	79	88
NA's	128	22	150	-	-	-

Tabla B.9: Cantidad de participantes para cada categoría de la variable Sexo (*Sex*). Se reportan los participantes separados por caso, control y el total para el análisis de hospitalizados. A la izquierda se muestran los participantes del dataset luego de la estandarización, a la derecha, se encuentran los participantes obtenidos luego de la imputación por KNN.

	Standardized			Imputed		
	Cases	Controls	Total	Cases	Controls	Total
Female	145	949	1094	145	949	1094
Male	179	599	778	179	599	778
NA's	0	0	0	-	-	-

Tabla B.10: Cantidad de participantes para cada categoría de la variable Grupo Sanguíneo (*Blood type*). Se reportan los participantes separados por caso, control y el total para el análisis de hospitalizados. A la izquierda se muestran los participantes del dataset luego de la estandarización, a la derecha, se encuentran los participantes obtenidos luego de la imputación por KNN.

	Standardized			Imputed		
	Cases	Controls	Total	Cases	Controls	Total
A	24	239	263	45	322	367
AB	5	20	25	83	177	260
B	10	97	107	105	460	565
O	70	494	564	91	589	680
NA's	215	698	913	-	-	-

Tabla B.11: Cantidad de participantes para cada categoría de la variable Tipo Rh (*Rh Type*). Se reportan los participantes separados por caso, control y el total para el análisis de hospitalizados. A la izquierda se muestran los participantes del dataset luego de la estandarización, a la derecha, se encuentran los participantes obtenidos luego de la imputación por KNN.

	Standardized			Imputed		
	Cases	Controls	Total	Cases	Controls	Total
Rh +	100	791	891	317	1475	1792
Rh -	7	73	80	7	73	80
NA's	217	684	901	-	-	-

Tabla B.12: Cantidad de participantes para cada categoría de la variable Nivel Educacional (*Educational Level*). Se reportan los participantes separados por caso, control y el total para el análisis de hospitalizados. A la izquierda se muestran los participantes del dataset luego de la estandarización, a la derecha, se encuentran los participantes obtenidos luego de la imputación por KNN.

	Standardized			Imputed		
	Cases	Controls	Total	Cases	Controls	Total
Complete Career Technical Education	38	308	346	69	317	386
Complete Collage Education	44	641	685	99	650	749
Complete High School	48	283	331	75	285	360
Complete Primary Education	22	43	65	26	44	70
Incomplete Career Technical Education	1	36	37	1	36	37
Incomplete Collage Education	8	153	161	8	153	161
Incomplete High School	9	39	48	9	39	48
Incomplete Primary Education	4	20	24	4	20	24
None	30	4	34	30	4	34
Illiterate	1	0	1	3	0	3
NA's	119	21	140	-	-	-

Tabla B.13: Cantidad de participantes para cada categoría de la variable Sistema de Salud (*Health System*). Se reportan los participantes separados por caso, control y el total para el análisis de hospitalizados. A la izquierda se muestran los participantes del dataset luego de la estandarización, a la derecha, se encuentran los participantes obtenidos luego de la imputación por KNN.

	Standardized			Imputed		
	Cases	Controls	Total	Cases	Controls	Total
Private	51	600	651	51	600	651
Public	273	948	1221	273	948	1221

Tabla B.14: Cantidad de participantes para cada categoría de la variable Etnicidad (*Ethnicity*). Se reportan los participantes separados por caso, control y el total para el análisis de hospitalizados. A la izquierda se muestran los participantes del dataset luego de la estandarización, a la derecha, se encuentran los participantes obtenidos luego de la imputación por KNN.

	Standardized			Imputed		
	Cases	Controls	Total	Cases	Controls	Total
Afro-indigeno	0	1	1	0	1	1
Aphrodescendant	1	3	4	1	3	4
Atacameño	1	3	4	1	3	4
Aymara	7	46	53	7	46	53
Colla	1	3	4	1	3	4
Diaguita	0	10	10	0	10	10
Huilliche	1	4	5	1	4	5
Kawésqar	2	1	3	2	1	3
Mapuche	16	128	144	16	128	144
None	292	1344	1636	292	1344	1636
Others	2	2	4	2	2	4
Quechua	1	2	3	1	2	3
Rapa Nui	0	1	1	0	1	1
Coya	0	0	0	0	0	0
Yámana (Yagán)	0	0	0	0	0	0

Tabla B.15: Cantidad de participantes para cada categoría de la variable Consumo de Tabaco (*Tabacco Consumption*). Se reportan los participantes separados por caso, control y el total para el análisis de hospitalizados. A la izquierda se muestran los participantes del dataset luego de la estandarización, a la derecha, se encuentran los participantes obtenidos luego de la imputación por KNN.

	Standardized			Imputed		
	Cases	Controls	Total	Cases	Controls	Total
Discontinued	90	482	572	97	488	585
Frequent (one or more a day)	11	208	219	21	222	243
He has never smoked	188	690	878	198	698	896
Occasional (less than a cigarette a day)	7	139	146	8	139	147
Other	0	1	1	0	1	1
NA's	28	28	56	-	-	-

Tabla B.16: Cantidad de participantes para cada categoría de la variable Cigarillos diarios (*Daily Cigars*). Se reportan los participantes separados por caso, control y el total para el análisis de hospitalizados. A la izquierda se muestran los participantes del dataset luego de la estandarización, a la derecha, se encuentran los participantes obtenidos luego de la imputación por KNN.

	Standardized			Imputed		
	Cases	Controls	Total	Cases	Controls	Total
0 a day	188	690	878	224	720	944
31 or more	3	6	9	9	11	20
Between 1 and 10	45	363	408	47	368	415
Between 11 and 20	12	57	69	13	58	71
Between 21 and 30	2	17	19	3	17	20
Less than 1 a day (less than 7 a week)	28	374	402	28	374	402
NA's	46	41	87	-	-	-

Tabla B.17: Cantidad de participantes para cada categoría de la variable Consumo de Alcohol (*Alcohol Consumption*). Se reportan los participantes separados por caso, control y el total para el análisis de hospitalizados. A la izquierda se muestran los participantes del dataset luego de la estandarización, a la derecha, se encuentran los participantes obtenidos luego de la imputación por KNN.

	Standardized			Imputed		
	Cases	Controls	Total	Cases	Controls	Total
Four or more times a week	4	18	22	21	24	45
Never drink	170	405	575	181	416	597
Once a month or less	54	513	567	59	521	580
Two or three times a week	10	78	88	14	82	96
Two to four times a month	49	505	554	49	505	554
NA's	37	29	66	-	-	-

Para los datos booleanos se incluye un analisis de la cantidad de participantes para cada valor reportado: verdadero (`True`), falso (`False`) y faltante (`NA's`) (**Tablas B.18-B.19**). También se incluyen un detalle similar al de las variables categóricas, separando los participantes por casos y controles (**Table B.20-B.22**).

Tabla B.18: Cantidad de participantes para cada variable booleana en el análisis de hospitalizados. Se muestran los participantes con un valor positivo (`true`), negativo (`false`) o faltante (`NA`). A la izquierda se muestran los participantes del dataset luego de la estandarización, a la derecha, se encuentran los participantes obtenidos luego de la imputación por KNN.

	Standardized			Imputed		
	True	False	NA's	True	False	NA's
Province	1394	465	13	1401	471	0
More than 100 cigars	740	1060	72	767	1105	0
Marijuana	492	1380	0	492	1380	0
Cocaine	44	1828	0	44	1828	0
Pasta Base	12	1860	0	12	1860	0
Heroin	1	1871	0	1	1871	0
Hallucinogens	30	1842	0	30	1842	0
Some other drug	10	1862	0	10	1862	0
Drug DK/NA/REF	5	1867	0	5	1867	0
Diabetes	179	1683	10	179	1693	0
Coronary atherosclerosis	21	1840	11	21	1851	0
Hypertension	337	1529	6	337	1535	0
Vascular accident	22	1838	12	22	1850	0
Heart problem	57	1792	23	57	1815	0
Dialysis	18	1843	11	18	1854	0
Hepatitis	28	1833	11	28	1844	0
Anemia	70	1790	12	70	1802	0
Asthma	112	1748	12	112	1760	0
Cystic fibrosis	2	1848	22	2	1870	0
Pulmonary fibrosis	4	1847	21	4	1868	0
Pulmonary condition	20	1841	11	20	1852	0
Cancer	45	1816	11	45	1827	0
Condition that affects the brain	11	1850	11	11	1861	0
HIV	9	1852	11	9	1863	0
Tuberculosis	4	1857	11	4	1868	0
Weakened immune system	23	1835	14	23	1849	0
Mental health problems	192	1668	12	193	1679	0
High cholesterol	186	1662	24	186	1686	0
Obesity	272	1577	23	272	1600	0
Ulcerative colitis or Crohn's disease	10	1839	23	10	1862	0
Rheumatoid arthritis	29	1820	23	29	1843	0
Lupus	7	1840	25	7	1865	0
Other rheumatoid diseases	20	1829	23	20	1852	0

Tabla B.19: Continuación **Tabla B.18**

	Standardized			Imputed		
	True	False	NA's	True	False	NA's
Other chronic diseases	619	1253	0	619	1253	0
Colon and straight cancer	3	1869	0	3	1869	0
Endometrial cancer	3	1869	0	3	1869	0
Pancreatic cancer	1	1871	0	1	1871	0
Prostate cancer	2	1870	0	2	1870	0
Lung cancer	1	1871	0	1	1871	0
Kidney cancer	1	1871	0	1	1871	0
Breast cancer	9	1863	0	9	1863	0
Thyroid cancer	14	1858	0	14	1858	0
Vesicle cancer	13	1859	0	13	1859	0
Antidiabetic	202	1670	0	202	1670	0
Anti-TB	1	1871	0	1	1871	0
Antihypertensive	281	1591	0	281	1591	0
Aspirin or clopidogrel	89	1783	0	89	1783	0
Statin	77	1795	0	77	1795	0
Analgesics	122	1750	0	122	1750	0
Cough medication	14	1858	0	14	1858	0
Inhaler	80	1792	0	80	1792	0
Anticancer	7	1865	0	7	1865	0
Antimalaric	2	1870	0	2	1870	0
Antivirals	5	1867	0	5	1867	0
Vitamins	228	1644	0	228	1644	0
Medicinal herbs	70	1802	0	70	1802	0
No drug	689	1183	0	689	1183	0
Another medication	626	1246	0	626	1246	0
Medication DK/NA/REF	40	1832	0	40	1832	0

Tabla B.20: Cantidad de participantes para cada variable booleana en el análisis de hospitalizados, separados por casos y controles. Se muestran los participantes con un valor positivo (`true`), negativo (`false`) o faltante (`NA`) para todos los casos y controles. A la izquierda se muestran los participantes del dataset luego de la estandarización, a la derecha, se encuentran los participantes obtenidos luego de la imputación por KNN.

	Standardized						Imputed					
	Controls			Cases			Controls			Cases		
	True	False	NA's	True	False	NA's	True	False	NA's	True	False	NA's
Province	1188	359	1	206	106	12	1189	359	0	212	112	0
More than 100 cigars	652	858	38	88	202	34	665	883	0	102	222	0
Marijuana	464	1084	0	28	296	0	464	1084	0	28	296	0
Cocaine	37	1511	0	7	317	0	37	1511	0	7	317	0
Pasta Base	10	1538	0	2	322	0	10	1538	0	2	322	0
Heroin	1	1547	0	0	324	0	1	1547	0	0	324	0
Hallucinogens	29	1519	0	1	323	0	29	1519	0	1	323	0
Some other drug	10	1538	0	0	324	0	10	1538	0	0	324	0
Drug DK/NA/REF	3	1545	0	2	322	0	3	1545	0	2	322	0
Diabetes	102	1438	8	77	245	2	102	1446	0	77	247	0
Coronary atherosclerosis	14	1525	9	7	315	2	14	1534	0	7	317	0
Hypertension	201	1342	5	136	187	1	201	1347	0	136	188	0
Vascular accident	14	1524	10	8	314	2	14	1534	0	8	316	0
Heart problem	33	1495	20	24	297	3	33	1515	0	24	300	0
Dialysis	4	1535	9	14	308	2	4	1544	0	14	310	0
Hepatitis	22	1517	9	6	316	2	22	1526	0	6	318	0
Anemia	60	1478	10	10	312	2	60	1488	0	10	314	0
Asthma	90	1448	10	22	300	2	90	1458	0	22	302	0
Cystic fibrosis	2	1526	20	0	322	2	2	1546	0	0	324	0
Pulmonary fibrosis	2	1527	19	2	320	2	2	1546	0	2	322	0
Pulmonary condition	12	1527	9	8	314	2	12	1536	0	8	316	0

Tabla B.21: Continuación **Tabla B.20**

	Standardized						Imputed					
	Controls			Cases			Controls			Cases		
	True	False	NA's	True	False	NA's	True	False	NA's	True	False	NA's
Cancer	32	1507	9	13	309	2	32	1516	0	13	311	0
Condition that affects the brain	7	1532	9	4	318	2	7	1541	0	4	320	0
HIV	6	1533	9	3	319	2	6	1542	0	3	321	0
Tuberculosis	4	1535	9	0	322	2	4	1544	0	0	324	0
Weakened immune system	19	1518	11	4	317	3	19	1529	0	4	320	0
Mental health problems	168	1370	10	24	298	2	169	1379	0	24	300	0
High cholesterol	142	1385	21	44	277	3	142	1406	0	44	280	0
Obesity	193	1335	20	79	242	3	193	1355	0	79	245	0
Ulcerative colitis or Crohn's disease	7	1521	20	3	318	3	7	1541	0	3	321	0
Rheumatoid arthritis	20	1508	20	9	312	3	20	1528	0	9	315	0
Lupus	6	1520	22	1	320	3	6	1542	0	1	323	0
Other rheumatoid diseases	18	1510	20	2	319	3	18	1530	0	2	322	0
Other chronic diseases	506	1042	0	113	211	0	506	1042	0	113	211	0
Colon and straight cancer	2	1546	0	1	323	0	2	1546	0	1	323	0
Endometrial cancer	2	1546	0	1	323	0	2	1546	0	1	323	0
Pancreatic cancer	1	1547	0	0	324	0	1	1547	0	0	324	0
Prostate cancer	0	1548	0	2	322	0	0	1548	0	2	322	0
Lung cancer	1	1547	0	0	324	0	1	1547	0	0	324	0
Kidney cancer	1	1547	0	0	324	0	1	1547	0	0	324	0
Breast cancer	6	1542	0	3	321	0	6	1542	0	3	321	0
Thyroid cancer	9	1539	0	5	319	0	9	1539	0	5	319	0

Tabla B.22: Continuación **Tabla B.21**

	Standardized						Imputed					
	Controls			Cases			Controls			Cases		
	True	False	NA's	True	False	NA's	True	False	NA's	True	False	NA's
Vesicle cancer	11	1537	0	2	322	0	11	1537	0	2	322	0
Antidiabetic	138	1410	0	64	260	0	138	1410	0	64	260	0
Anti-TB	1	1547	0	0	324	0	1	1547	0	0	324	0
Antihypertensive	172	1376	0	109	215	0	172	1376	0	109	215	0
Aspirin or clopidogrel	49	1499	0	40	284	0	49	1499	0	40	284	0
Statin	43	1505	0	34	290	0	43	1505	0	34	290	0
Analgesics	96	1452	0	26	298	0	96	1452	0	26	298	0
Cough medication	8	1540	0	6	318	0	8	1540	0	6	318	0
Inhaler	54	1494	0	26	298	0	54	1494	0	26	298	0
Anticancer	4	1544	0	3	321	0	4	1544	0	3	321	0
Antimalaric	2	1546	0	0	324	0	2	1546	0	0	324	0
Antivirals	4	1544	0	1	323	0	4	1544	0	1	323	0
Vitamins	209	1339	0	19	305	0	209	1339	0	19	305	0
Medicinal herbs	60	1488	0	10	314	0	60	1488	0	10	314	0
No drug	655	893	0	34	290	0	655	893	0	34	290	0
Another medication	530	1018	0	96	228	0	530	1018	0	96	228	0
Medication DK/NA/REF	38	1510	0	2	322	0	38	1510	0	2	322	0

Para las variables numéricas se muestra la cantidad de participantes que sí tienen un valor reportado, la cantidad de datos nulos, el promedio y la desviación estándar para los datos estandarizados e imputados (**Tabla B.23**). Estos mismos análisis separados por casos y controles no se muestran, ya que se realizan durante el análisis univariado (**Anexo B.2.2**).

Tabla B.23: Resumen de variables numéricas para análisis de hospitalizados entre infectados. Se incluye la cantidad de participantes que sí presentan el valor (N), faltante (NA's), promedio ($\overline{X}$) y desviación estándard (σ). A la izquierda se muestran los participantes del dataset luego de la estandarización, a la derecha, se encuentran los participantes obtenidos luego de la imputación por KNN.

	Standardized				Imputed			
	N	NA's	$\overline{x}$	σ	N	NA's	$\overline{x}$	σ
Age	1864	8	42.20	14.35	1872	0	42.18	14.33
Population Density	1856	16	2423.62	4739.36	1872	0	2417.04	4724.03
Weight	1769	103	77.81	16.68	1872	0	77.98	16.30
Height	1744	128	164.92	9.28	1872	0	165.04	9.06
BMI	1742	130	28.49	5.35	1872	0	28.57	5.28

B.2.2. Análisis Univariado

Tabla B.24: Resultados de la prueba estadístico $t-$student para las variables numéricas del análisis para hospitalizados. A la izquierda se muestran los participantes del dataset luego de la estandarización, a la derecha, se encuentran los participantes obtenidos luego de la imputación por KNN.

	Standardized				Imputed			
	$\overline{x}$ (Cases)	$\overline{x}$ (Controls)	t	p	$\overline{x}$ (Cases)	$\overline{x}$ (Controls)	t	p
Age	54.46	39.70	-17.48	2.39e-52	54.06	39.70	-17.09	7.28e-51
Population Density	4051.47	2098.47	-5.72	2.09e-08	3942.65	2097.72	-5.60	3.90e-08
Weight	82.97	77.03	-4.63	5.51e-06	82.56	77.02	-5.66	2.66e-08
Height	164.50	164.98	0.69	4.94e-01	165.23	164.99	-0.47	6.41e-01
BMI	30.41	28.23	-5.00	1.10e-06	30.23	28.23	-6.16	1.62e-09

Tabla B.25: Resultados de la prueba estadístico χ^2 para las variables categóricas y boolenas en el análisis para hospitalizados. A la izquierda se muestran los participantes del dataset luego de la estandarización, a la derecha, se encuentran los participantes obtenidos luego de la imputación por KNN.

	Standardized			Imputed		
	χ^2	df	p	χ^2	df	p
Province	16.05	1	6.17e-05	18.42	1	1.78e-05
Sex	30.22	1	3.85e-08	30.22	1	3.85e-08
Blood type	4.21	3	2.40e-01	53.30	3	1.58e-11
Rh type	0.46	1	4.99e-01	4.28	1	3.86e-02
Country of birth	61.17	10	2.17e-09	60.01	10	3.61e-09
Educational level	262.50	9	2.28e-51	187.77	9	1.20e-35
Occupational situation	20.31	2	3.88e-05	17.72	2	1.42e-04
Health System	62.59	1	2.55e-15	62.59	1	2.55e-15
Ethnicity	16.49	12	1.70e-01	16.49	12	1.70e-01
Tobacco Consumption	51.73	4	1.57e-10	42.00	4	1.67e-08
Daily cigars	54.53	5	1.63e-10	78.46	5	1.76e-15
More than 100 cigars	16.55	1	4.74e-05	14.59	1	1.33e-04
Alcohol Consumption	119.65	4	6.34e-25	146.56	4	1.11e-30
Marijuana	62.93	1	2.14e-15	62.93	1	2.14e-15
Cocaine	0.06	1	8.04e-01	0.06	1	8.04e-01
Pasta Base	0.00	1	9.53e-01	0.00	1	9.53e-01
Heroin	0.21	1	6.47e-01	0.21	1	6.47e-01
Hallucinogens	4.16	1	4.14e-02	4.16	1	4.14e-02
Some other drug	2.10	1	1.47e-01	2.10	1	1.47e-01
Drug DK/NA/REF	1.80	1	1.79e-01	1.80	1	1.79e-01
Diabetes	91.62	1	1.05e-21	91.41	1	1.17e-21
Coronary atherosclerosis	3.81	1	5.08e-02	3.81	1	5.09e-02
Hypertension	152.61	1	4.65e-35	152.55	1	4.81e-35
Vascular accident	5.65	1	1.75e-02	5.65	1	1.75e-02
Heart problem	25.10	1	5.44e-07	25.26	1	5.01e-07
Dialysis	46.46	1	9.37e-12	46.44	1	9.47e-12
Hepatitis	0.34	1	5.61e-01	0.34	1	5.61e-01
Anemia	0.47	1	4.95e-01	0.46	1	4.96e-01
Asthma	0.45	1	5.01e-01	0.45	1	5.00e-01
Cystic fibrosis	0.42	1	5.16e-01	0.42	1	5.17e-01
Pulmonary fibrosis	2.97	1	8.51e-02	2.99	1	8.36e-02
Pulmonary condition	7.28	1	6.98e-03	7.27	1	7.00e-03
Cancer	4.33	1	3.75e-02	4.32	1	3.76e-02

Tabla B.26: Continuación **Tabla B.25**

	Standardized			Imputed		
	χ^2	df	p	χ^2	df	p
Condition that affects the brain	2.81	1	9.37e-02	2.81	1	9.38e-02
HIV	1.62	1	2.02e-01	1.62	1	2.03e-01
Tuberculosis	0.84	1	3.60e-01	0.84	1	3.60e-01
Weakened immune system	0.00	1	9.88e-01	0.00	1	9.91e-01
Mental health problems	3.46	1	6.28e-02	3.57	1	5.89e-02
High cholesterol	5.69	1	1.70e-02	5.82	1	1.59e-02
Obesity	30.34	1	3.62e-08	30.63	1	3.13e-08
Ulcerative colitis or Crohn's disease	1.12	1	2.90e-01	1.13	1	2.87e-01
Rheumatoid arthritis	3.84	1	5.01e-02	3.88	1	4.89e-02
Lupus	0.05	1	8.29e-01	0.04	1	8.32e-01
Other rheumatoid diseases	0.76	1	3.82e-01	0.75	1	3.85e-01
Other chronic diseases	0.58	1	4.46e-01	0.58	1	4.46e-01
Colon and straight cancer	0.54	1	4.63e-01	0.54	1	4.63e-01
Endometrial cancer	0.54	1	4.63e-01	0.54	1	4.63e-01
Pancreatic cancer	0.21	1	6.47e-01	0.21	1	6.47e-01
Prostate cancer	9.57	1	1.98e-03	9.57	1	1.98e-03
Lung cancer	0.21	1	6.47e-01	0.21	1	6.47e-01
Kidney cancer	0.21	1	6.47e-01	0.21	1	6.47e-01
Breast cancer	1.62	1	2.03e-01	1.62	1	2.03e-01
Thyroid cancer	3.34	1	6.77e-02	3.34	1	6.77e-02
Vesicle cancer	0.03	1	8.54e-01	0.03	1	8.54e-01
Antidiabetic	32.69	1	1.08e-08	32.69	1	1.08e-08
Anti-TB	0.21	1	6.47e-01	0.21	1	6.47e-01
Antihypertensive	106.61	1	5.42e-25	106.61	1	5.42e-25
Aspirin or clopidogrel	49.86	1	1.65e-12	49.86	1	1.65e-12
Statin	40.44	1	2.02e-10	40.44	1	2.02e-10
Analgesics	1.46	1	2.27e-01	1.46	1	2.27e-01
Cough medication	6.43	1	1.12e-02	6.43	1	1.12e-02
Inhaler	13.48	1	2.41e-04	13.48	1	2.41e-04
Anticancer	3.20	1	7.34e-02	3.20	1	7.34e-02
Antimalaric	0.42	1	5.17e-01	0.42	1	5.17e-01
Antivirals	0.03	1	8.73e-01	0.03	1	8.73e-01
Vitamins	14.61	1	1.32e-04	14.61	1	1.32e-04

Tabla B.27: Continuación **Tabla B.26**

	Standardized			Imputed		
	χ^2	df	p	χ^2	df	p
Medicinal herbs	0.46	1	4.96e-01	0.46	1	4.96e-01
No drug	116.62	1	3.47e-27	116.62	1	3.47e-27
Another medication	2.56	1	1.10e-01	2.56	1	1.10e-01
Medication DK/NA/REF	4.33	1	3.75e-02	4.33	1	3.75e-02

B.2.3. Análisis Multivariado

Tabla B.28: Resultado de la prueba estadístico multivariado usando la función `drop1` para análisis de hospitalizados. A la izquierda se muestran los participantes del dataset luego de la estandarización, a la derecha, se encuentran los participantes obtenidos luego de la imputación por KNN.

	Standardized					Imputed				
	Df	Deviance	AIC	LRT	$Pr\,(\chi^2)$	Df	Deviance	AIC	LRT	$Pr\,(\chi^2)$
`<none>`	-	322.38	534.38	-	-	-	981.70	1215.70	-	-
Age	1	347.26	557.26	24.88	6.10e-07	1	1028.23	1260.23	46.53	9.01e-12
Province	1	326.99	536.99	4.61	3.18e-02	1	981.97	1213.97	0.28	6.00e-01
Population Density	1	323.12	533.12	0.74	3.91e-01	1	986.92	1218.92	5.22	2.23e-02
Sex	1	322.79	532.79	0.40	5.26e-01	1	985.15	1217.15	3.45	6.31e-02
Blood type	3	327.73	533.73	5.35	1.48e-01	3	993.05	1221.05	11.35	9.96e-03
Rh type	1	323.22	533.22	0.83	3.61e-01	1	983.16	1215.16	1.46	2.26e-01
Weight	1	323.94	533.94	1.55	2.13e-01	1	981.74	1213.74	0.04	8.40e-01
Height	1	324.07	534.07	1.69	1.93e-01	1	981.73	1213.73	0.03	8.64e-01
BMI	1	324.67	534.67	2.29	1.30e-01	1	981.72	1213.72	0.03	8.65e-01
Country of birth	6	327.77	527.77	5.38	4.96e-01	10	989.54	1203.54	7.84	6.44e-01
Educational level	8	329.86	525.86	7.47	4.87e-01	9	1031.98	1247.98	50.29	9.51e-08
Occupational situation	2	334.18	542.18	11.79	2.75e-03	2	986.76	1216.76	5.06	7.95e-02
Health System	1	326.14	536.14	3.76	5.26e-02	1	991.79	1223.79	10.10	1.49e-03
Ethnicity	11	330.45	520.45	8.06	7.08e-01	12	987.79	1197.79	6.09	9.11e-01
Tobacco Consumption	3	344.52	550.52	22.14	6.10e-05	4	998.65	1224.65	16.96	1.97e-03
Daily cigars	4	324.44	528.44	2.06	7.25e-01	5	996.12	1220.12	14.43	1.31e-02
More than 100 cigars	1	322.52	532.52	0.13	7.16e-01	1	981.80	1213.80	0.10	7.51e-01
Alcohol Consumption	4	333.31	537.31	10.93	2.74e-02	4	990.88	1216.88	9.19	5.66e-02
Marijuana	1	322.80	532.80	0.42	5.19e-01	1	982.60	1214.60	0.91	3.41e-01
Cocaine	1	323.70	533.70	1.32	2.50e-01	1	982.82	1214.82	1.12	2.90e-01

Tabla B.29: Continuación **Tabla B.28**

	Standardized					Imputed				
	Df	Deviance	AIC	LRT	$Pr\,(\chi^2)$	Df	Deviance	AIC	LRT	$Pr\,(\chi^2)$
Pasta Base	1	323.09	533.09	0.71	4.00e-01	1	981.70	1213.70	0.01	9.26e-01
Heroin	0	322.38	534.38	0.00	-	1	981.70	1213.70	0.00	1.00e+00
Hallucinogens	1	323.42	533.42	1.04	3.09e-01	1	981.71	1213.71	0.01	9.20e-01
Some other drug	1	322.46	532.46	0.08	7.82e-01	1	982.38	1214.38	0.69	4.07e-01
Drug DK/NA/REF	1	322.47	532.47	0.08	7.74e-01	1	983.62	1215.62	1.92	1.66e-01
Diabetes	1	322.38	532.38	0.00	9.95e-01	1	985.15	1217.15	3.45	6.32e-02
Coronary atherosclerosis	1	322.44	532.44	0.05	8.16e-01	1	983.26	1215.26	1.56	2.11e-01
Hypertension	1	330.71	540.71	8.33	3.90e-03	1	981.78	1213.78	0.08	7.75e-01
Vascular accident	1	322.71	532.71	0.33	5.68e-01	1	982.49	1214.49	0.80	3.72e-01
Heart problem	1	322.82	532.82	0.44	5.06e-01	1	981.93	1213.93	0.23	6.31e-01
Dialysis	1	322.43	532.43	0.05	8.30e-01	1	983.25	1215.25	1.55	2.13e-01
Hepatitis	1	326.95	536.95	4.57	3.26e-02	1	982.41	1214.41	0.71	3.98e-01
Anemia	1	322.51	532.51	0.13	7.21e-01	1	982.05	1214.05	0.35	5.54e-01
Asthma	1	323.26	533.26	0.88	3.47e-01	1	982.15	1214.15	0.45	5.00e-01
Cystic fibrosis	1	322.38	532.38	0.00	1.00e+00	1	984.00	1216.00	2.30	1.29e-01
Pulmonary fibrosis	1	322.75	532.75	0.37	5.42e-01	1	981.95	1213.95	0.26	6.13e-01
Pulmonary condition	1	322.55	532.55	0.17	6.84e-01	1	982.27	1214.27	0.57	4.50e-01
Cancer	1	322.38	532.38	0.00	1.00e+00	1	984.05	1216.05	2.35	1.25e-01
Condition that affects the brain	1	323.70	533.70	1.32	2.51e-01	1	982.61	1214.61	0.92	3.38e-01
HIV	1	322.64	532.64	0.26	6.10e-01	1	982.74	1214.74	1.05	3.06e-01
Tuberculosis	1	322.88	532.88	0.50	4.81e-01	1	982.54	1214.54	0.85	3.58e-01

Tabla B.30: Continuación **Tabla B.29**

	Standardized					Imputed				
	Df	Deviance	AIC	LRT	$Pr\,(\chi^2)$	Df	Deviance	AIC	LRT	$Pr\,(\chi^2)$
Weakened immune system	1	323.06	533.06	0.68	4.11e-01	1	981.89	1213.89	0.20	6.56e-01
Mental health problems	1	322.39	532.39	0.01	9.25e-01	1	981.70	1213.70	0.00	9.72e-01
High cholesterol	1	322.40	532.40	0.02	8.90e-01	1	983.49	1215.49	1.80	1.80e-01
Obesity	1	325.99	535.99	3.61	5.75e-02	1	984.16	1216.16	2.46	1.17e-01
Ulcerative colitis or Crohn's disease	1	327.09	537.09	4.71	3.00e-02	1	983.88	1215.88	2.19	1.39e-01
Rheumatoid arthritis	1	322.38	532.38	0.00	9.82e-01	1	981.76	1213.76	0.06	8.03e-01
Lupus	1	323.65	533.65	1.27	2.60e-01	1	981.86	1213.86	0.16	6.88e-01
Other rheumatoid diseases	1	325.89	535.89	3.50	6.12e-02	1	986.00	1218.00	4.30	3.81e-02
Other chronic diseases	1	324.37	534.37	1.99	1.58e-01	1	983.10	1215.10	1.40	2.36e-01
Colon and straight cancer	1	322.38	532.38	0.00	1.00e+00	1	981.97	1213.97	0.28	5.99e-01
Endometrial cancer	1	322.87	532.87	0.49	4.86e-01	1	981.82	1213.82	0.13	7.20e-01
Pancreatic cancer	0	322.38	534.38	0.00	-	1	981.70	1213.70	0.00	9.63e-01
Prostate cancer	1	322.38	532.38	0.00	1.00e+00	1	986.20	1218.20	4.50	3.38e-02
Lung cancer	1	322.38	532.38	0.00	1.00e+00	1	981.73	1213.73	0.04	8.51e-01
Kidney cancer	0	322.38	534.38	0.00	-	1	981.74	1213.74	0.04	8.42e-01
Breast cancer	1	322.38	532.38	0.00	1.00e+00	1	984.51	1216.51	2.82	9.32e-02
Thyroid cancer	1	322.38	532.38	0.00	1.00e+00	1	984.61	1216.61	2.91	8.80e-02
Vesicle cancer	1	322.38	532.38	0.00	1.00e+00	1	983.54	1215.54	1.85	1.74e-01
Antidiabetic	1	322.88	532.88	0.50	4.81e-01	1	987.44	1219.44	5.74	1.65e-02
Anti-TB	0	322.38	534.38	0.00	-	1	983.36	1215.36	1.66	1.97e-01
Antihypertensive	1	329.77	539.77	7.39	6.57e-03	1	983.57	1215.57	1.88	1.71e-01

Tabla B.31: Continuación **Tabla B.30**

	Standardized					Imputed				
	Df	Deviance	AIC	LRT	$Pr\,(\chi^2)$	Df	Deviance	AIC	LRT	$Pr\,(\chi^2)$
Aspirin or clopidogrel	1	323.50	533.50	1.12	2.90e-01	1	981.79	1213.79	0.09	7.59e-01
Statin	1	327.56	537.56	5.17	2.29e-02	1	984.14	1216.14	2.45	1.18e-01
Analgesics	1	323.28	533.28	0.89	3.45e-01	1	981.98	1213.98	0.28	5.95e-01
Cough medication	1	322.95	532.95	0.57	4.50e-01	1	989.11	1221.11	7.41	6.47e-03
Inhaler	1	323.49	533.49	1.11	2.92e-01	1	981.84	1213.84	0.14	7.08e-01
Anticancer	1	322.41	532.41	0.02	8.79e-01	1	982.39	1214.39	0.70	4.03e-01
Antimalaric	1	322.38	532.38	0.00	1.00e+00	1	982.89	1214.89	1.19	2.75e-01
Antivirals	1	322.38	532.38	0.00	1.00e+00	1	981.73	1213.73	0.04	8.49e-01
Vitamins	1	324.78	534.78	2.39	1.22e-01	1	1002.22	1234.22	20.53	5.88e-06
Medicinal herbs	1	324.84	534.84	2.46	1.17e-01	1	982.62	1214.62	0.93	3.36e-01
No drug	1	322.43	532.43	0.05	8.30e-01	1	1032.97	1264.97	51.27	8.04e-13
Another medication	1	323.55	533.55	1.17	2.79e-01	1	1001.50	1233.50	19.80	8.58e-06
Medication DK/NA/REF	1	322.99	532.99	0.60	4.37e-01	1	1000.93	1232.93	19.23	1.16e-05

Anexo C

Resultados de experimentos

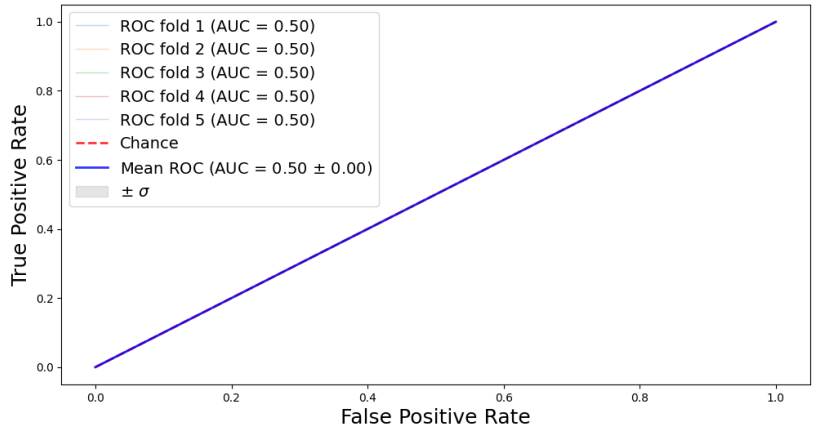
C.1. Datos Clínicos

C.1.1. Algoritmos de ML sobre datos imputados y selección de variables

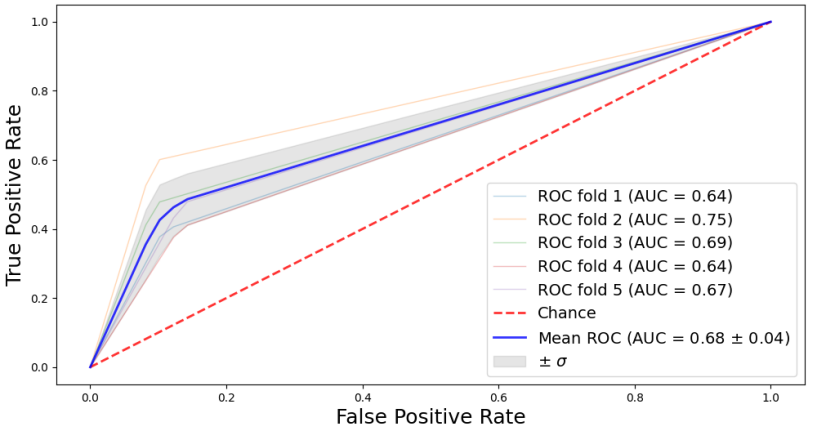
Figura C.1: [Pagina 111] Curva de ROC de los modelos ML sobre todos los datos clínicos imputados disponibles para el análisis de hospitalizados entre la población infectada confirmada.

Figura C.2: [Pagina 112] Matrices de confusión para los modelos ML sobre todos los datos clínicos imputados disponibles para el análisis de hospitalizados entre la población infectada confirmada.

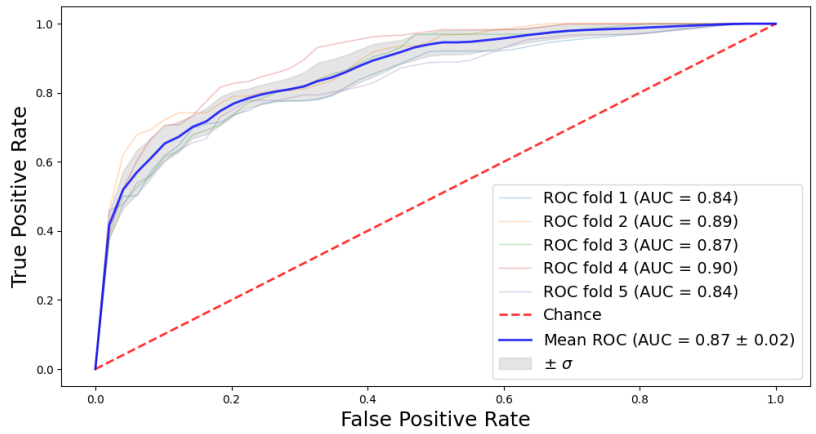
Dummy Mode Classifier



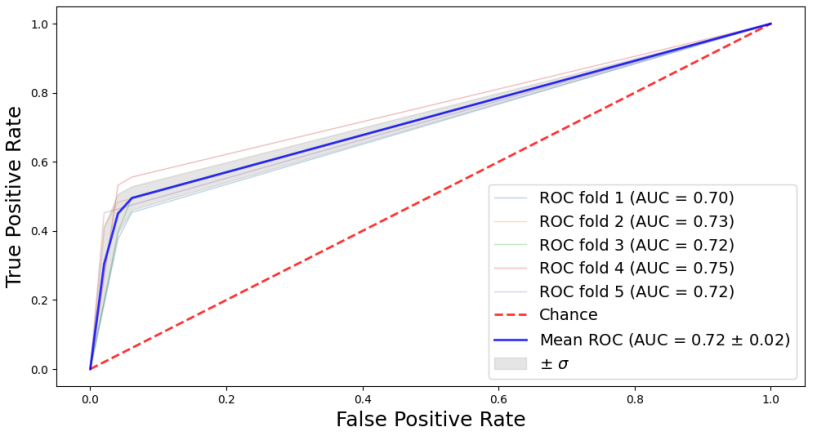
Decision Tree



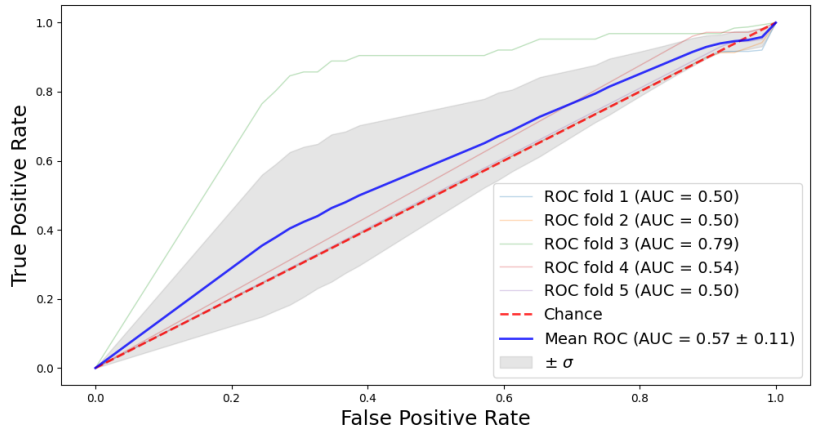
Random Forest



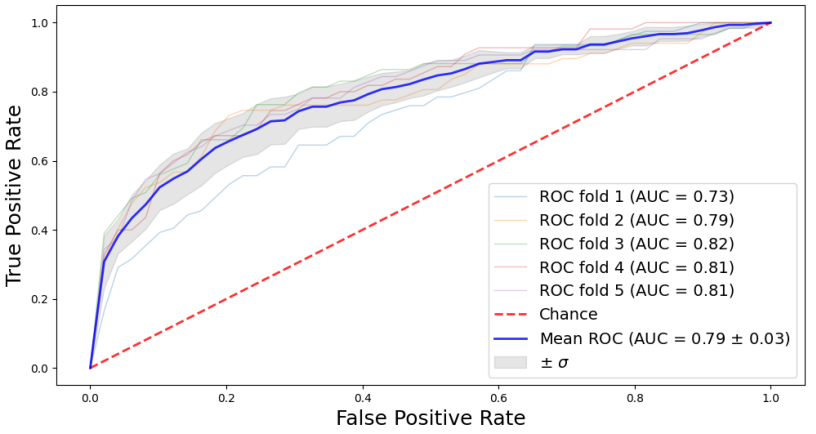
SVM



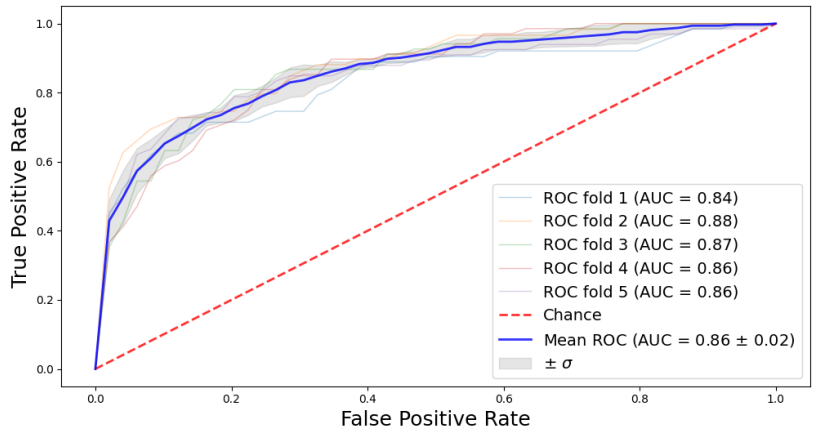
Naive Bayes



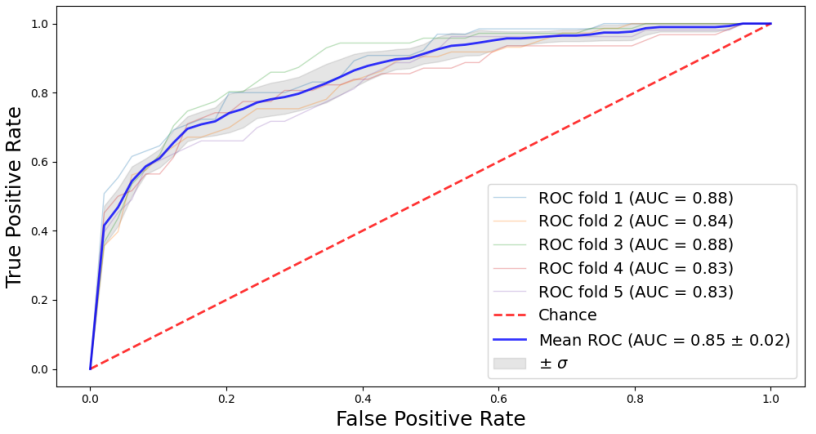
Gaussian Classifier

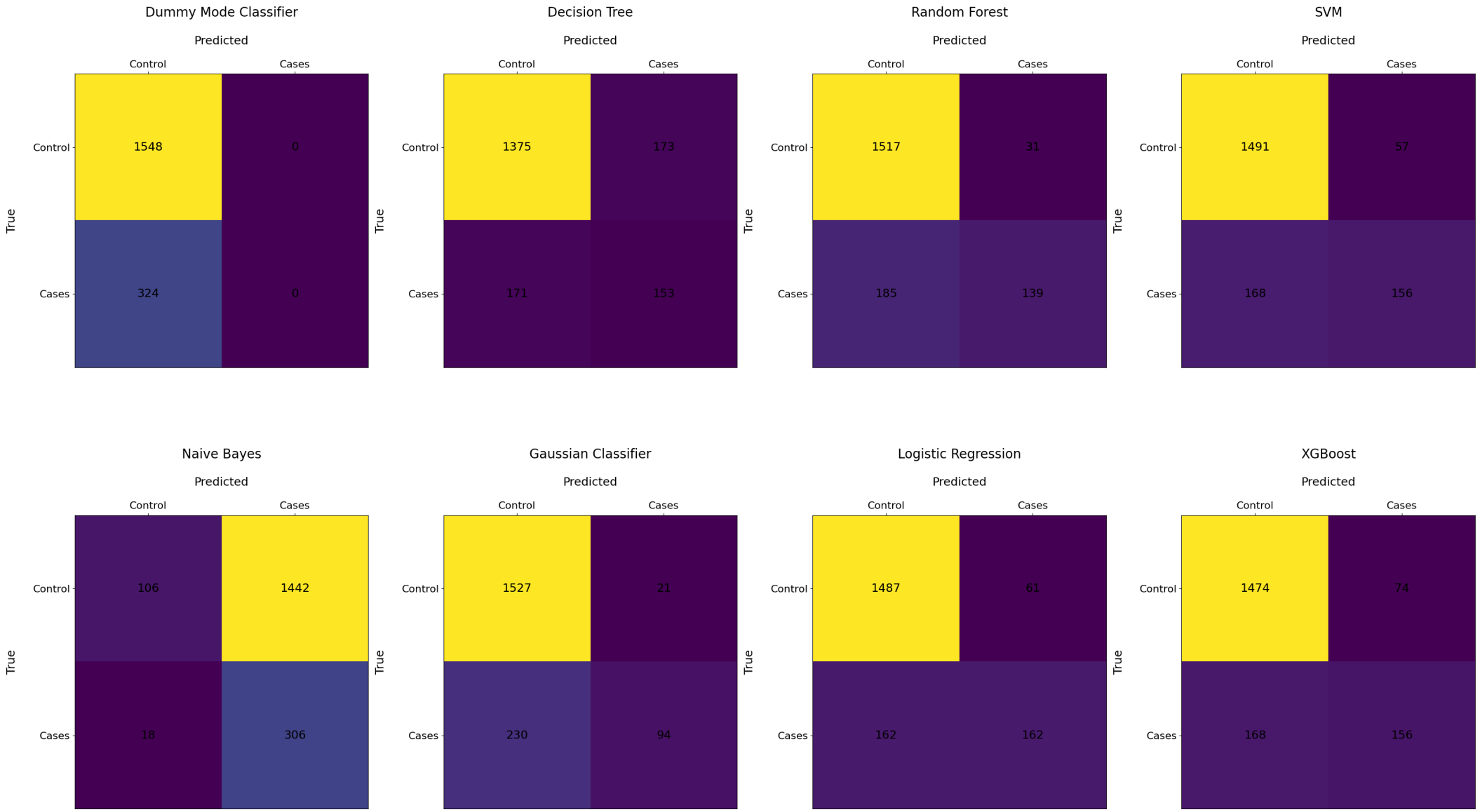


Logistic Regression



XGBoost




Dummy Mode Classifier
Predicted
Control
Cases
True
Control
Cases
1548
0
324
0
Decision Tree
Predicted
Control
Cases
True
Control
Cases
1375
173
171
153
Random Forest
Predicted
Control
Cases
True
Control
Cases
1517
31
185
139
SVM
Predicted
Control
Cases
True
Control
Cases
1491
57
168
156
Naive Bayes
Predicted
Control
Cases
True
Control
Cases
106
1442
18
306
Gaussian Classifier
Predicted
Control
Cases
True
Control
Cases
1527
21
230
94
Logistic Regression
Predicted
Control
Cases
True
Control
Cases
1487
61
162
162
XGBoost
Predicted
Control
Cases
True
Control
Cases
1474
74
168
156

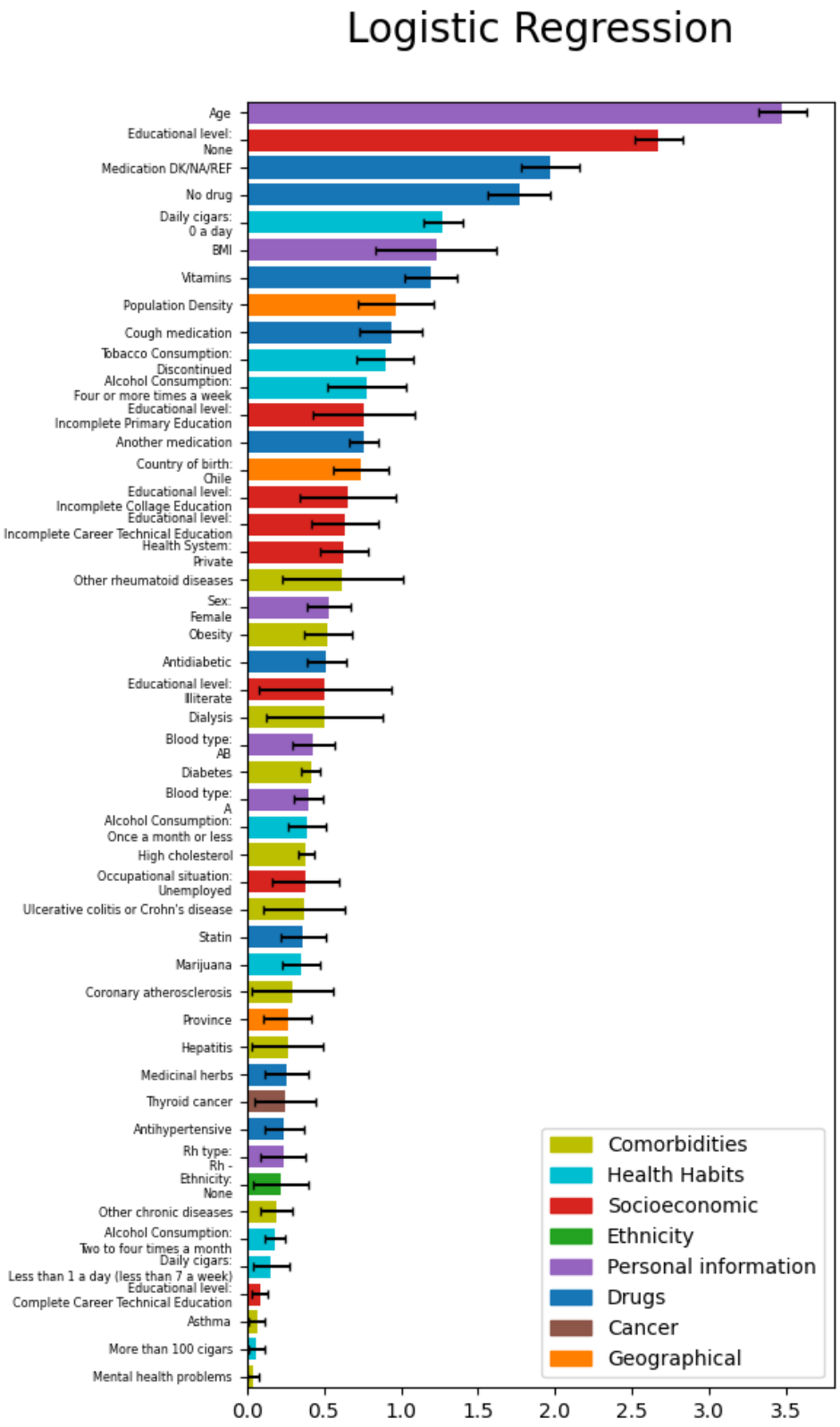


Figura C.3: Selección de variables mediante modelo de regresión logística sobre todos los datos clínicos imputados disponibles. La regresión logística es ajustada utilizando una penalización $l1$. Las variables seleccionadas corresponden a aquellas cuyo valor absoluto del coeficiente promedio para todas las *fold* es mayor a 10^{-5}.

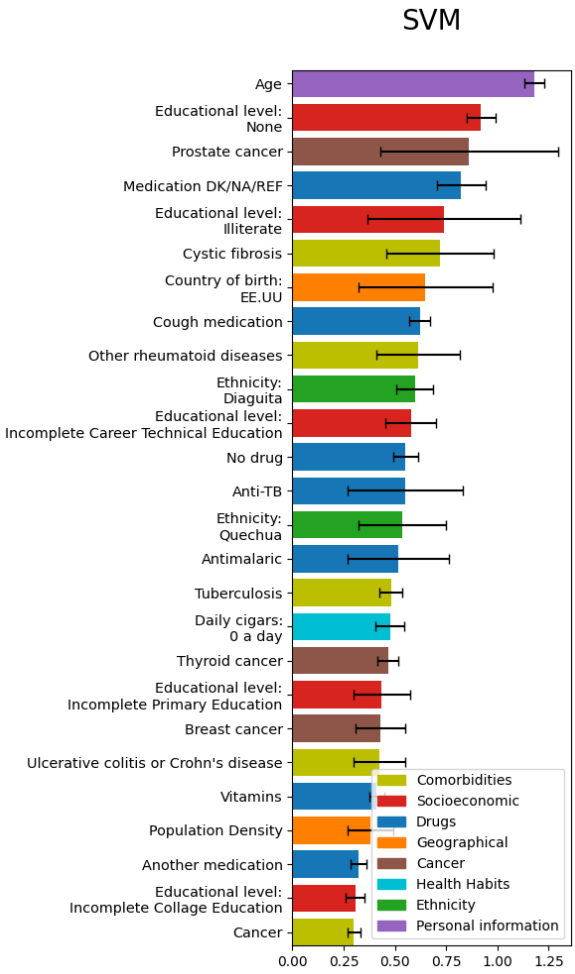


Figura C.4: Selección de variables mediante modelo de SVM sobre todos los datos clínicos imputados disponibles. Las variables seleccionadas corresponden a aquellas cuyo valor absoluto de coeficiente promedio para todas las *fold* es mayor al valor absoluto promedio de todas las variables.

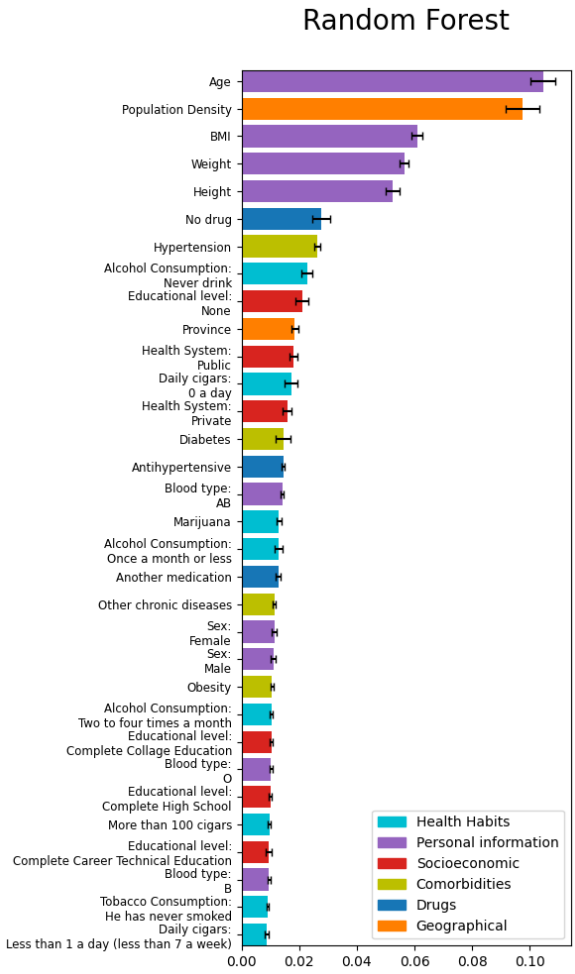


Figura C.5: Selección de variables mediante modelo Random Forest sobre todos los datos clínicos imputados disponibles. El modelo entrega un valor de importancia (*feature importance*) para cada variable. Las variables seleccionadas corresponden a aquellas cuyo valor de importancia promedio para todas las *fold* es mayor a la importancia promedio de todas las variables.

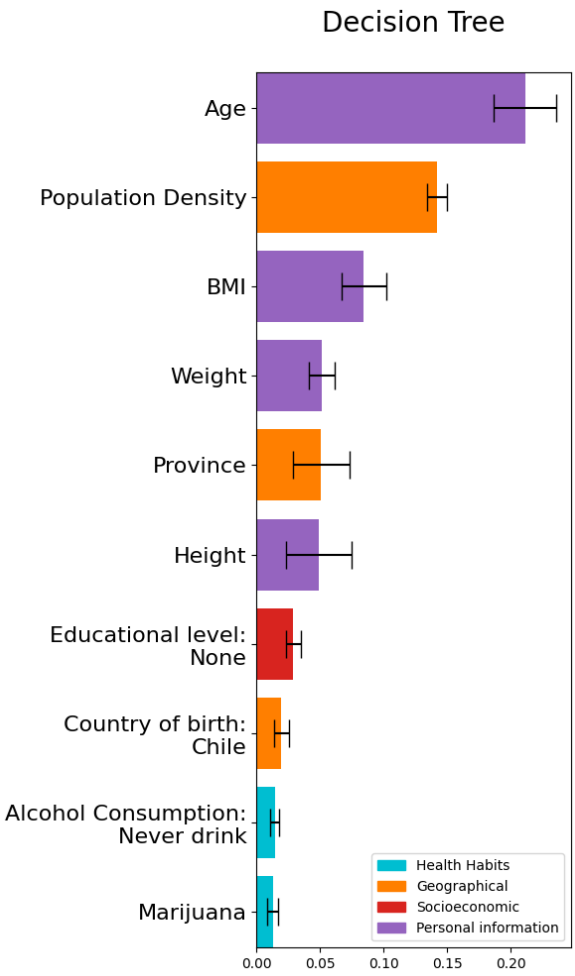


Figura C.6: Selección de variables mediante modelo de árbol de decisión sobre todos los datos clínicos imputados disponibles. El modelo entrega un valor de importancia (*feature importance*) para cada variable. Las variables seleccionadas corresponden a aquellas cuyo valor de importancia promedio para todas las *fold* es mayor a la importancia promedio de todas las variables.

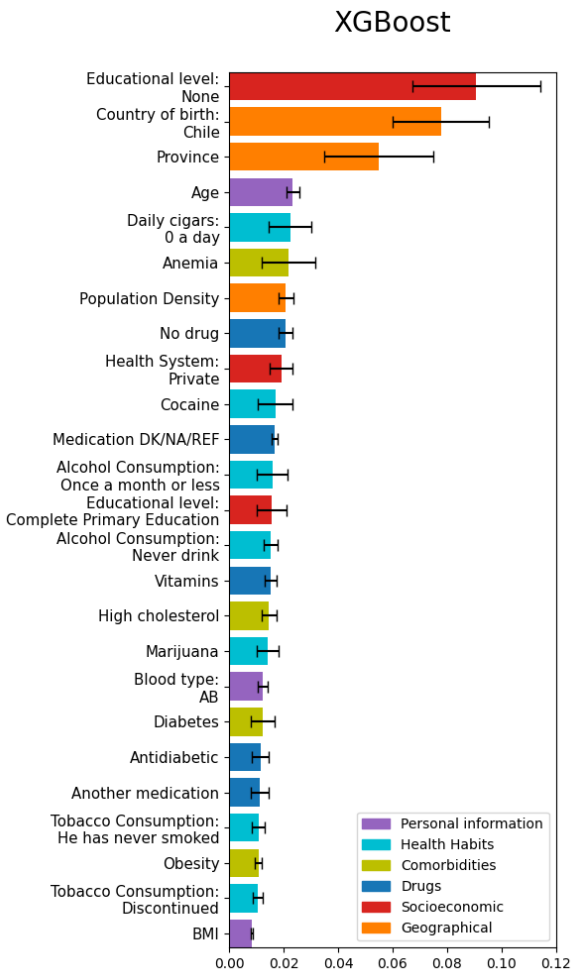


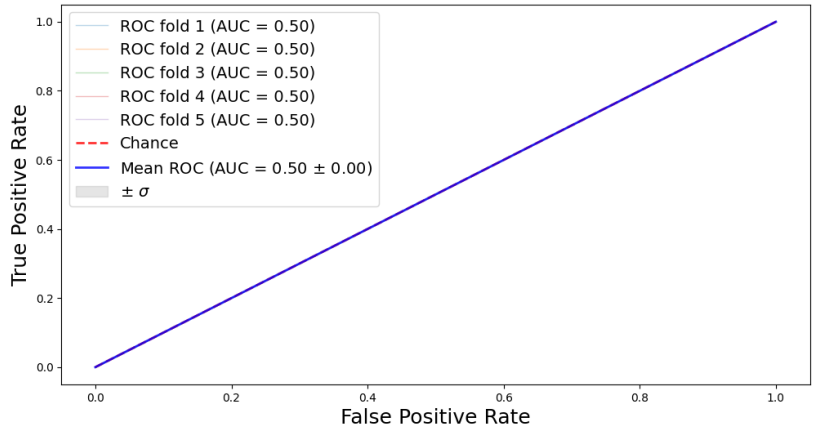
Figura C.7: Selección de variables mediante modelo XGBoost sobre todos los datos clínicos imputados disponibles. El modelo entrega un valor de importancia (*feature importance*) para cada variable. Las variables seleccionadas corresponden a aquellas cuyo valor de importancia promedio para todas las *fold* es mayor a la importancia promedio de todas las variables.

C.1.2. Algoritmos de ML sobre datos seleccionados

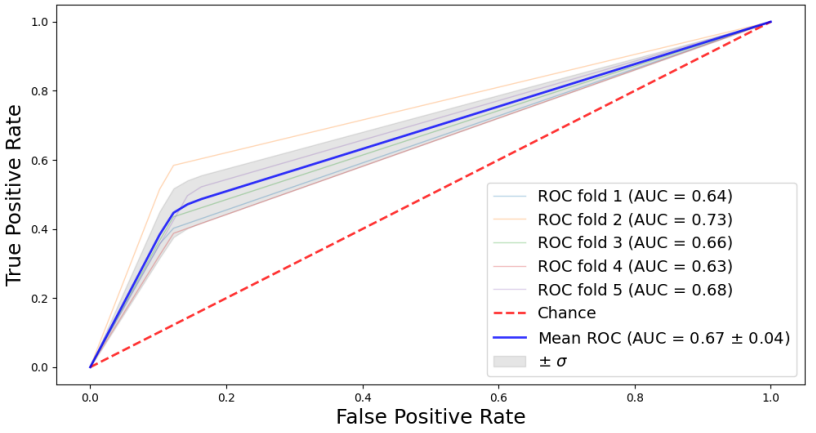
Figura C.8: [Pagina 116] Curva de ROC de los modelos ML sobre los datos clínicos seleccionados para el análisis de hospitalizados entre la población infectada confirmada.

Figura C.9: [Pagina 117] Matrices de confusión para los modelos ML sobre los datos clínicos seleccionados para el análisis de hospitalizados entre la población infectada confirmada.

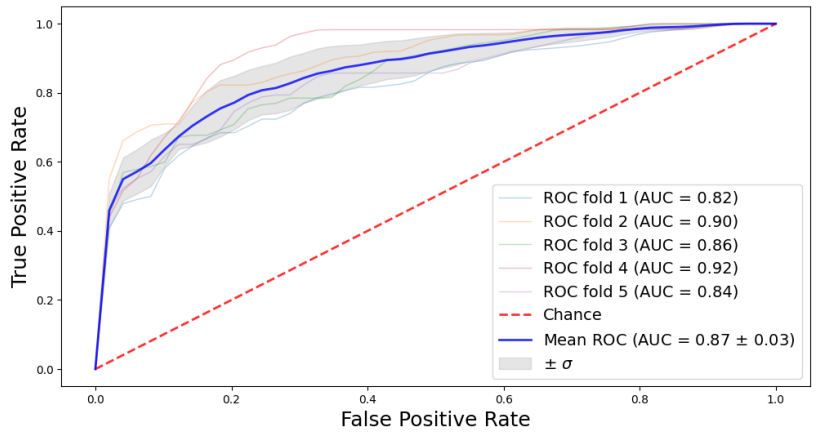
Dummy Mode Classifier



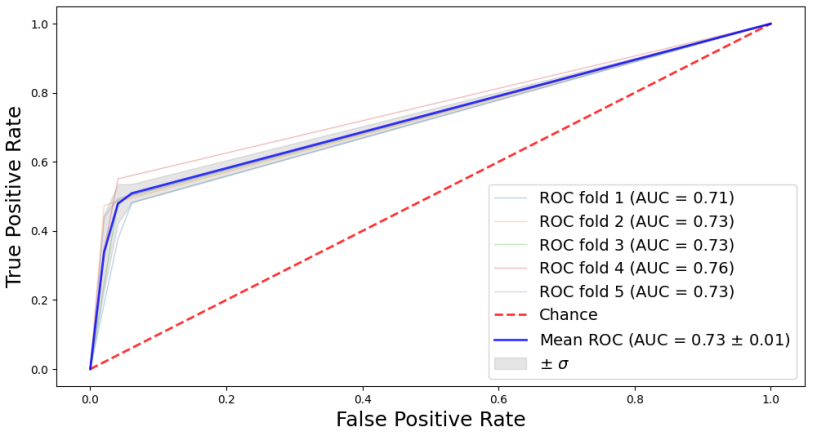
Decision Tree



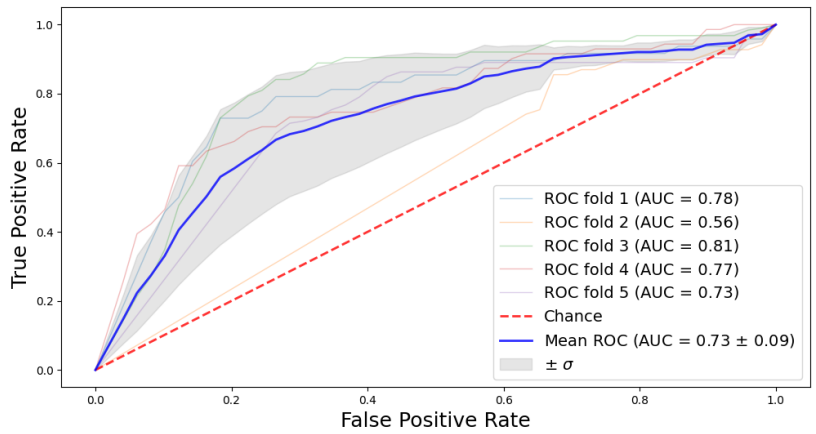
Random Forest



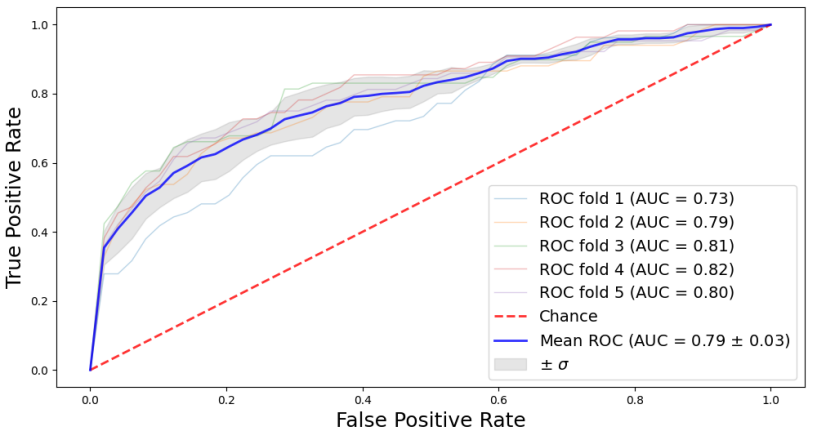
SVM



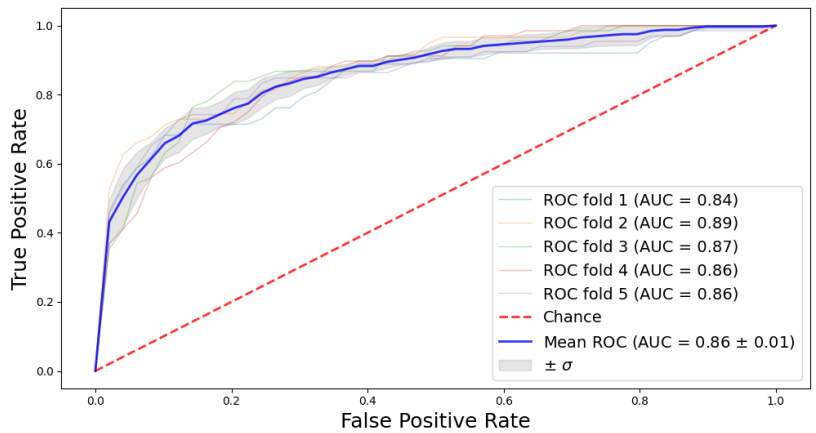
Naive Bayes



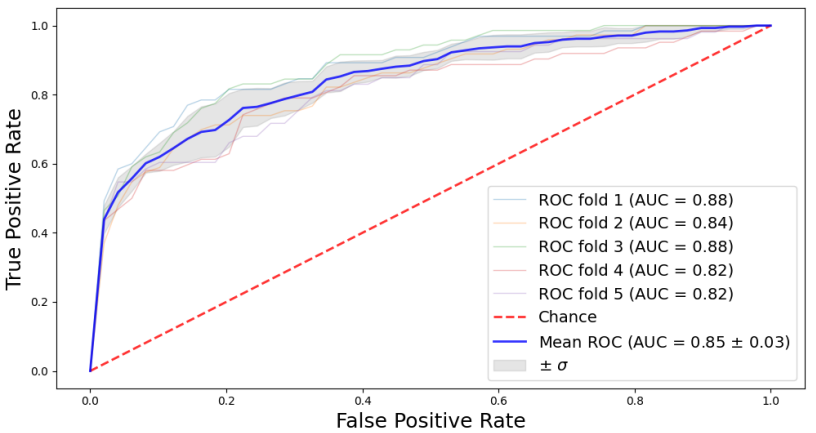
Gaussian Classifier

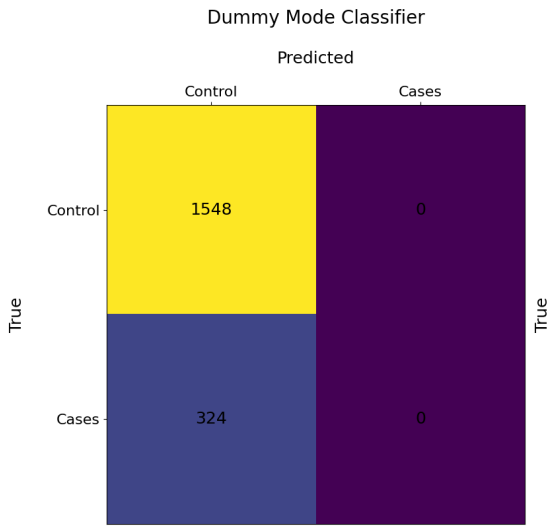


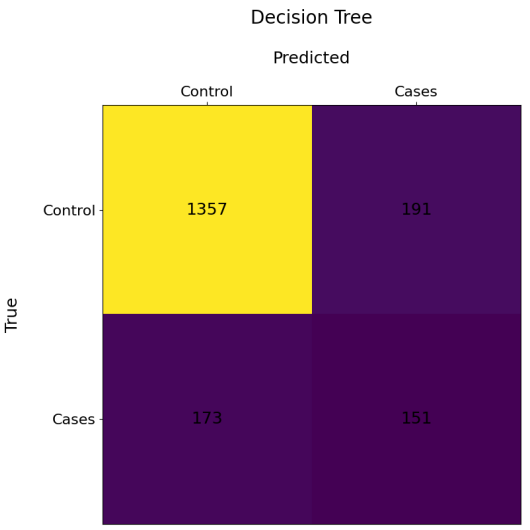
Logistic Regression

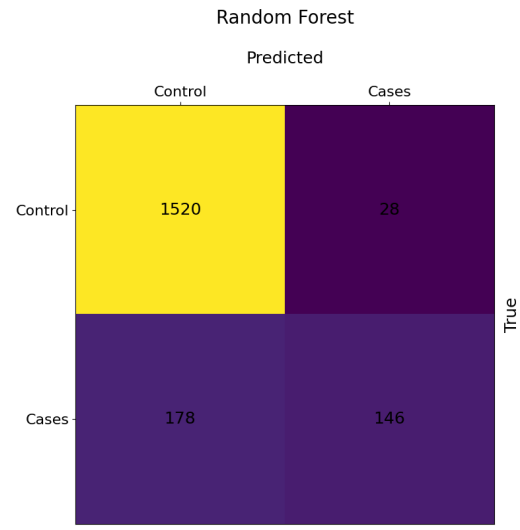


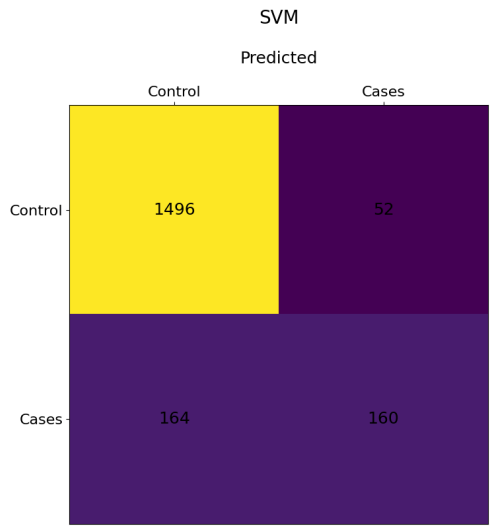
XGBoost

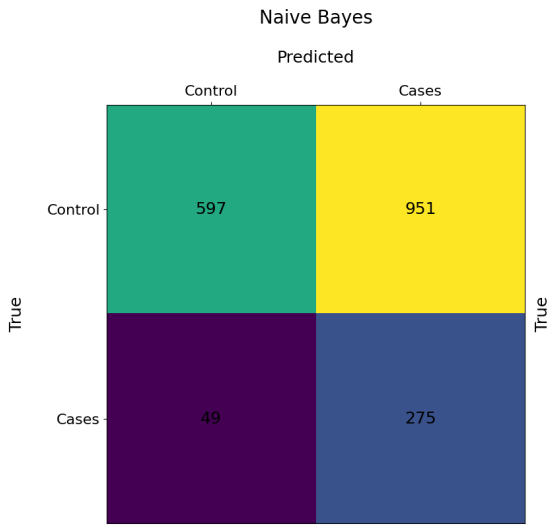


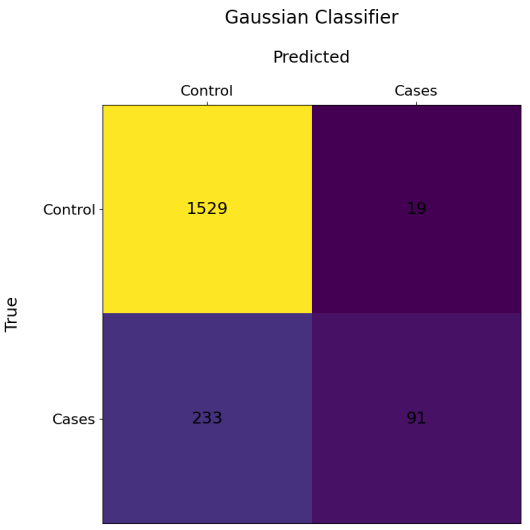

Dummy Mode Classifier
Predicted
Control
Cases
True
Control
Cases
1548
0
324
0

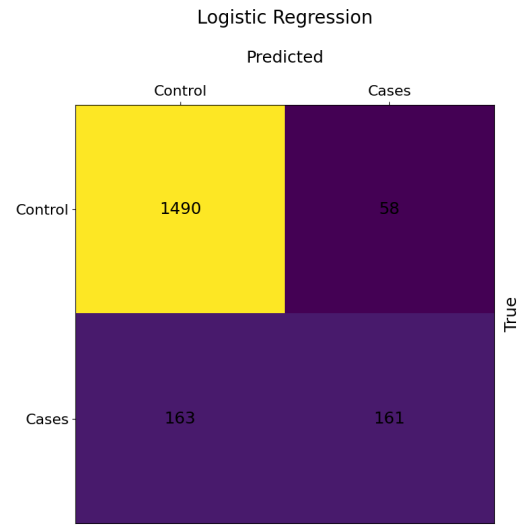

Decision Tree
Predicted
Control
Cases
True
Control
Cases
1357
191
173
151

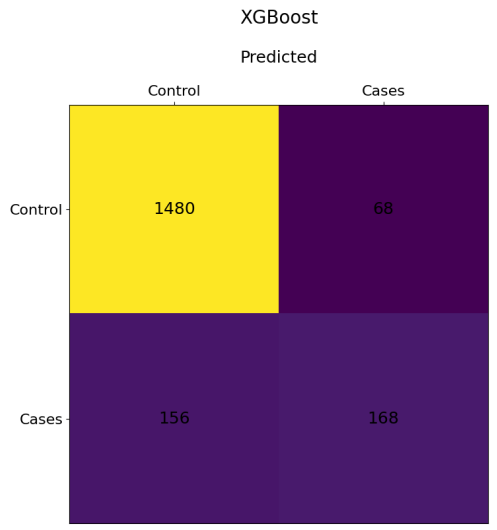

Random Forest
Predicted
Control
Cases
True
Control
Cases
1520
28
178
146


SVM
Predicted
Control
Cases
Control
Cases
1496
52
164
160


Naive Bayes
Predicted
Control
Cases
True
Control
Cases
597
951
49
275


Gaussian Classifier
Predicted
Control
Cases
True
Control
Cases
1529
19
233
91


Logistic Regression
Predicted
Control
Cases
True
Control
Cases
1490
58
163
161


XGBoost
Predicted
Control
Cases
Control
Cases
1480
68
156
168

C.1.3. Redes Neuronales Artificiales (ANN)

C.1.3.1. Sobre todos los datos clínicos

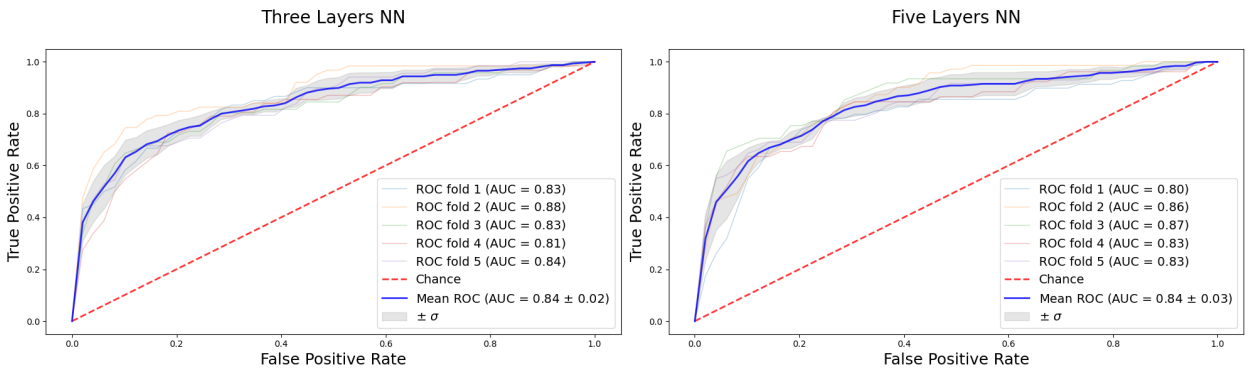


Figura C.10: Curva de ROC de los modelos de redes neuronales *Fast foward* (FNN) sobre todos los datos clínicos imputados disponibles para el análisis de hospitalizados entre la población infectada confirmada.

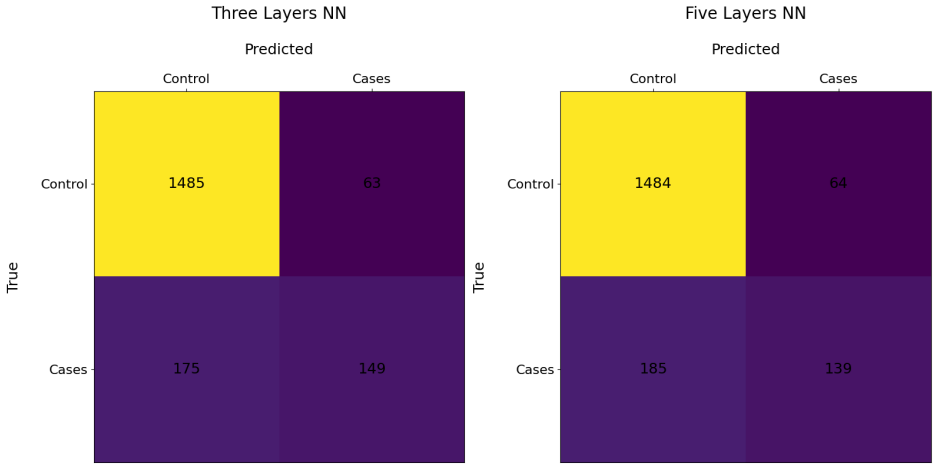


Figura C.11: Matrices de confusión para los modelos de redes neuronales *Fast foward* (FNN) sobre todos los datos clínicos imputados disponibles para el análisis de hospitalizados entre la población infectada confirmada.

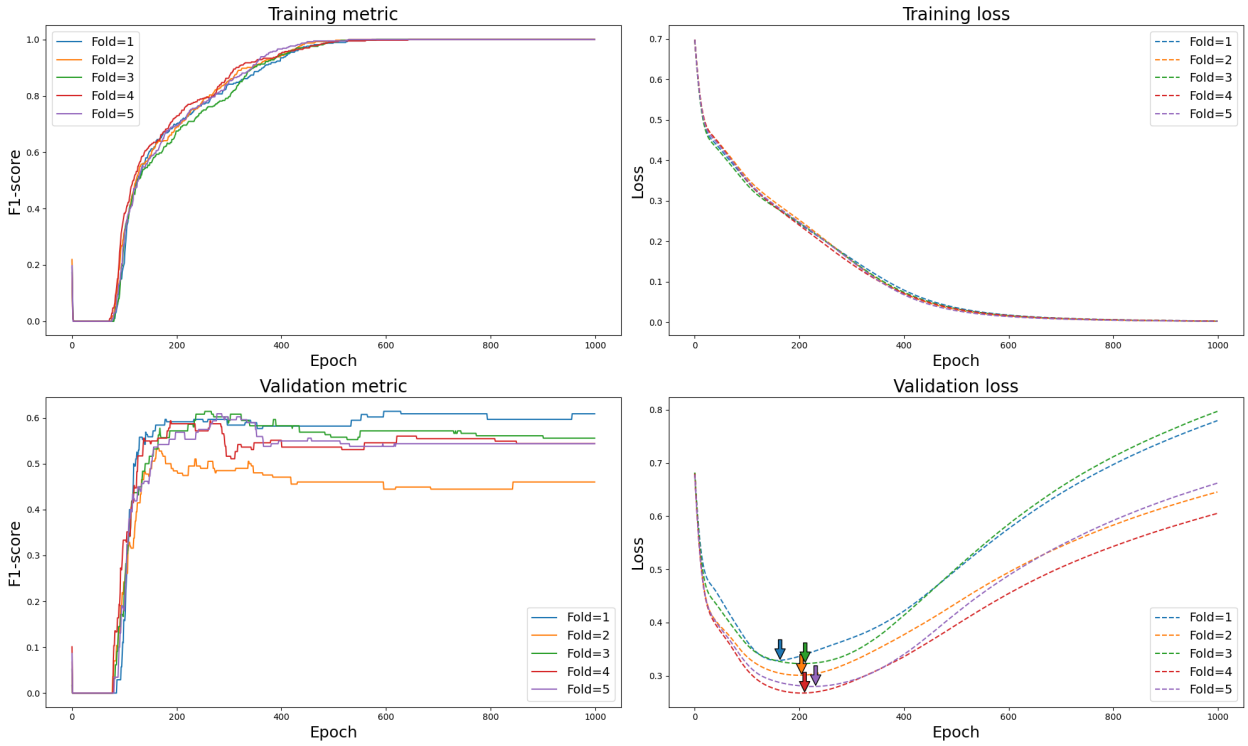


Figura C.12: Curva de entrenamiento de la arquitectura tres capas *Feed foward*, sobre todos los datos clínicos disponibles para el análisis de hospitalizados entre la población infectada. A la esquina superior izquierda se muestra la métrica de entrenamiento, F1-score, a la esquina inferior izquierda la métrica de validación, a la esquina superior derecha la función de pérdida de entrenamiento, y a la esquina inferior derecha la función de pérdida de validación. En esta última se muestran flechas que indican que el modelo se guarda por *Early Stopping* al estancarse o disminuir la función de pérdida.

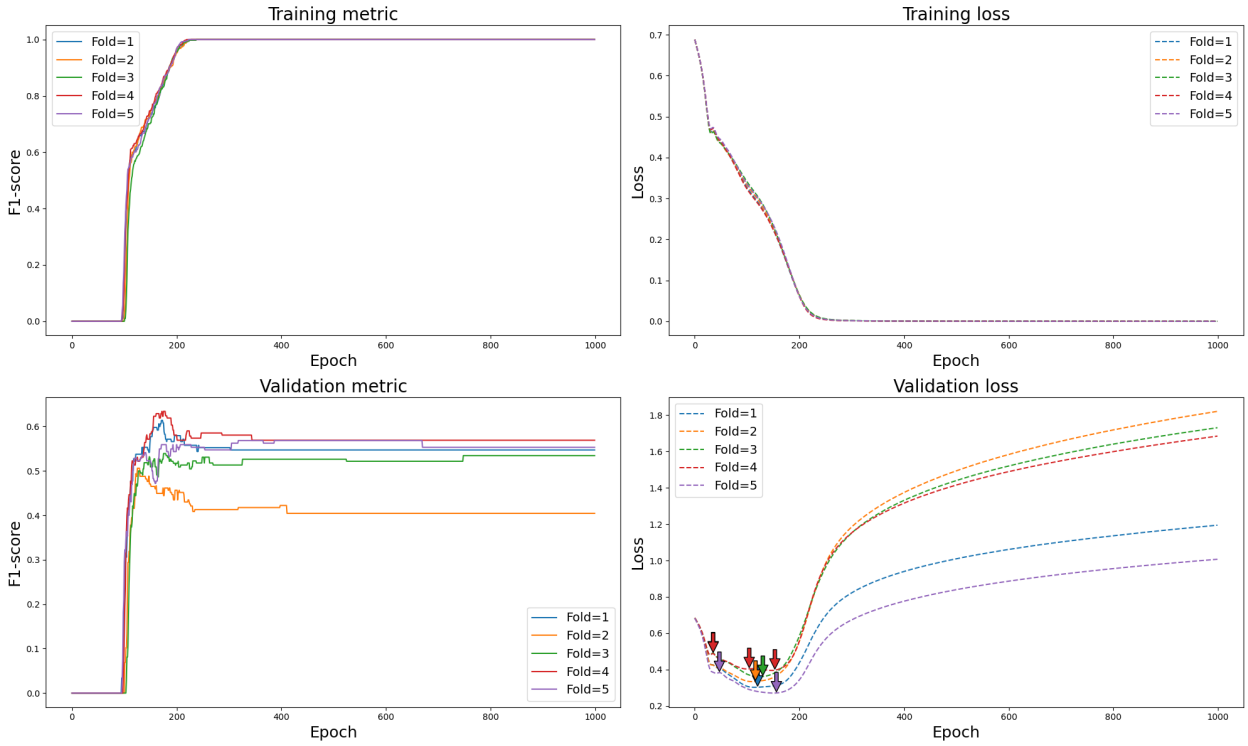


Figura C.13: Curva de entrenamiento de la arquitectura cinco capas *Feed foward*, sobre todos los datos clínicos disponibles para el análisis de hospitalizados entre la población infectada. A la esquina superior izquierda se muestra la métrica de entrenamiento, F1-score, a la esquina inferior izquierda la métrica de validación, a la esquina superior derecha la función de pérdida de entrenamiento, y a la esquina inferior derecha la función de pérdida de validación. En esta última se muestran flechas que indican que el modelo se guarda por *Early Stopping* al estancarse o disminuir la función de pérdida.

C.1.3.2. Sobre datos clínicos seleccionados

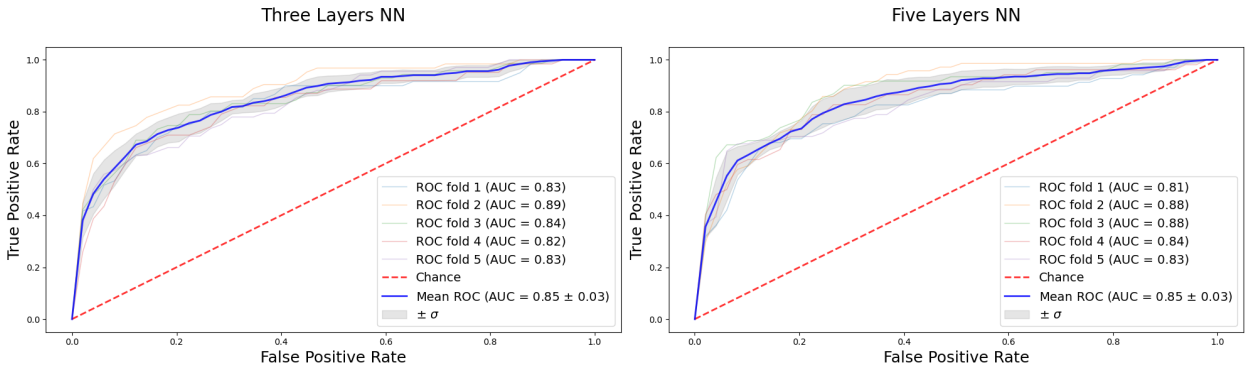


Figura C.14: Curva de ROC de los modelos de redes neuronales *Fast foward* (FNN) sobre los datos clínicos seleccionados para el análisis de hospitalizados entre la población infectada confirmada.

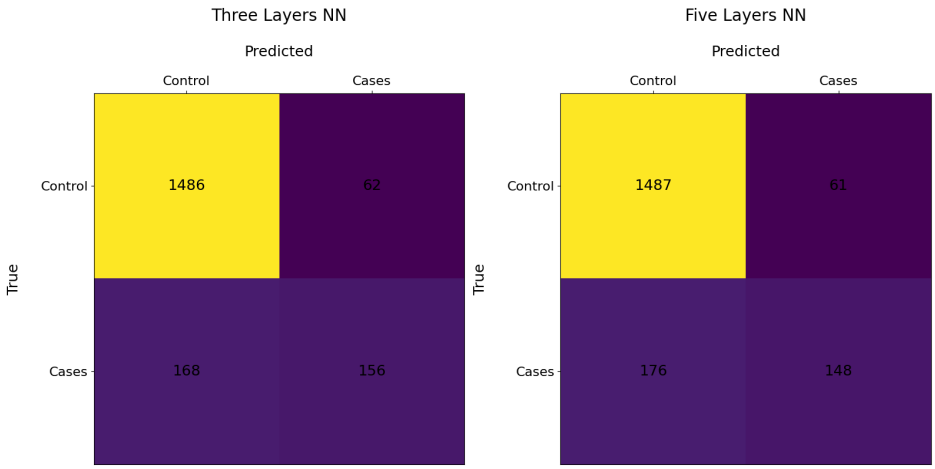


Figura C.15: Matrices de confusión para los modelos de redes neuronales *Fast foward* (FNN) sobre los datos clínicos seleccionados para el análisis de hospitalizados entre la población infectada confirmada.

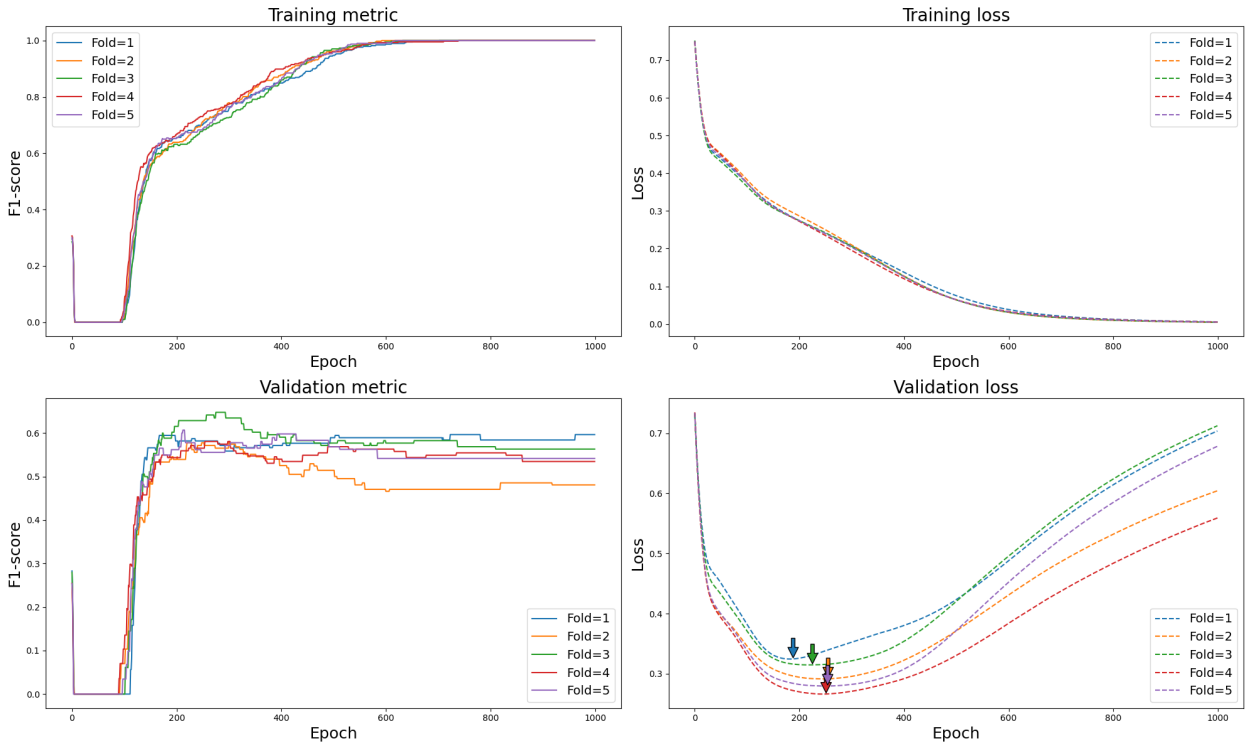


Figura C.16: Curva de entrenamiento de la arquitectura tres capas *Feed foward*, sobre los datos clínicos seleccionados para el análisis de hospitalizados entre la población infectada. A la esquina superior izquierda se muestra la métrica de entrenamiento, F1-score, a la esquina inferior izquierda la métrica de validación, a la esquina superior derecha la función de pérdida de entrenamiento, y a la esquina inferior derecha la función de pérdida de validación. En esta última se muestran flechas que indican que el modelo se guarda por *Early Stopping* al estancarse o disminuir la función de pérdida.

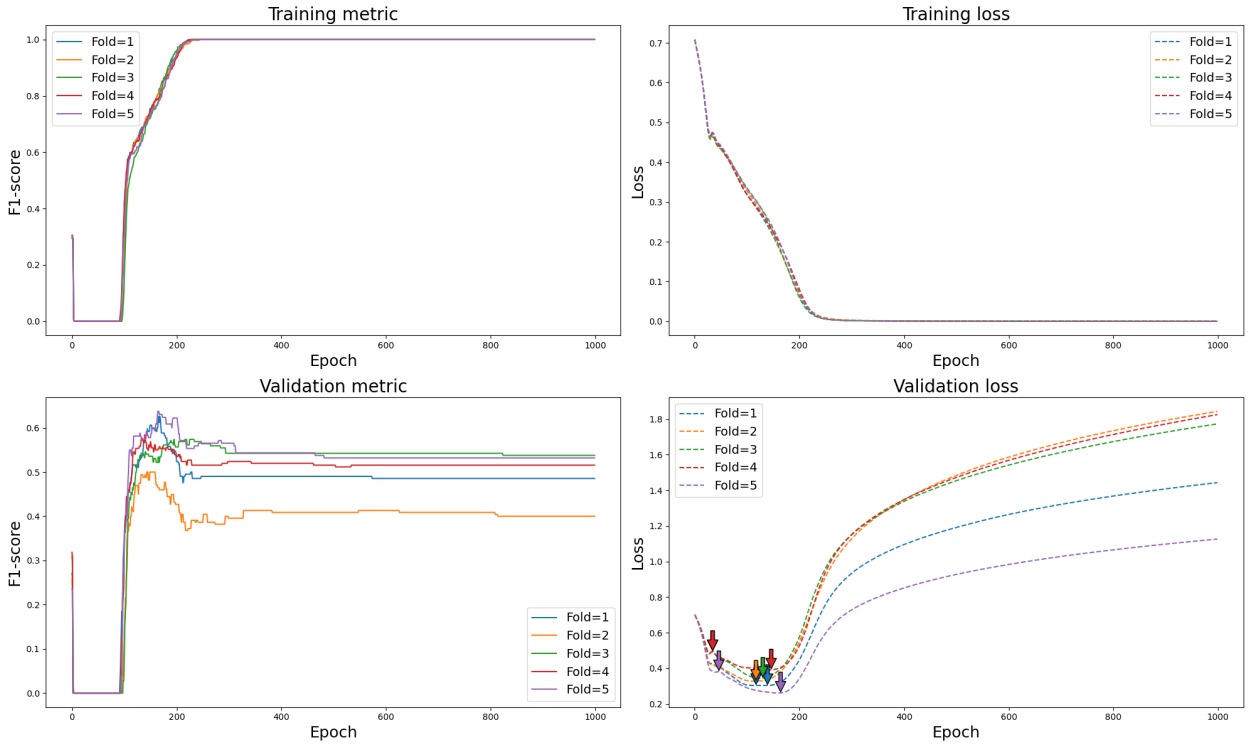


Figura C.17: Curva de entrenamiento de la arquitectura cinco capas *Feed foward*, sobre los datos clínicos seleccionados para el análisis de hospitalizados entre la población infectada. A la esquina superior izquierda se muestra la métrica de entrenamiento, F1-score, a la esquina inferior izquierda la métrica de validación, a la esquina superior derecha la función de pérdida de entrenamiento, y a la esquina inferior derecha la función de pérdida de validación. En esta última se muestran flechas que indican que el modelo se guarda por *Early Stopping* al estancarse o disminuir la función de pérdida.

C.2. GWAS

C.2.1. Resultados Iniciativa

Figura C.18: [Pagina 123] GWAS de la iniciativa internacional Covid19hg. La significancia, expresada como $-\log(p)$, se calcula obteniendo el p-value de los coeficientes del modelo de regresión logística utilizando la variante, el análisis de componentes principales (PCA) de todas las variantes, edad, el cuadrado de la edad, sexo y el producto de sexo y edad de los participantes. Datos obtenidos desde https://app.covid19hg.org/ [12]

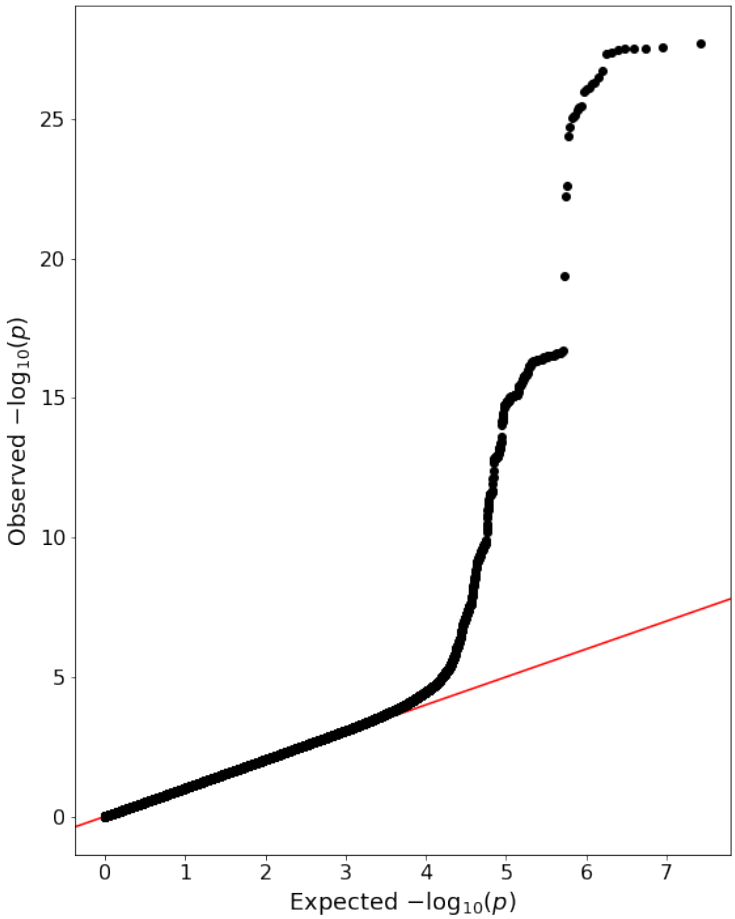
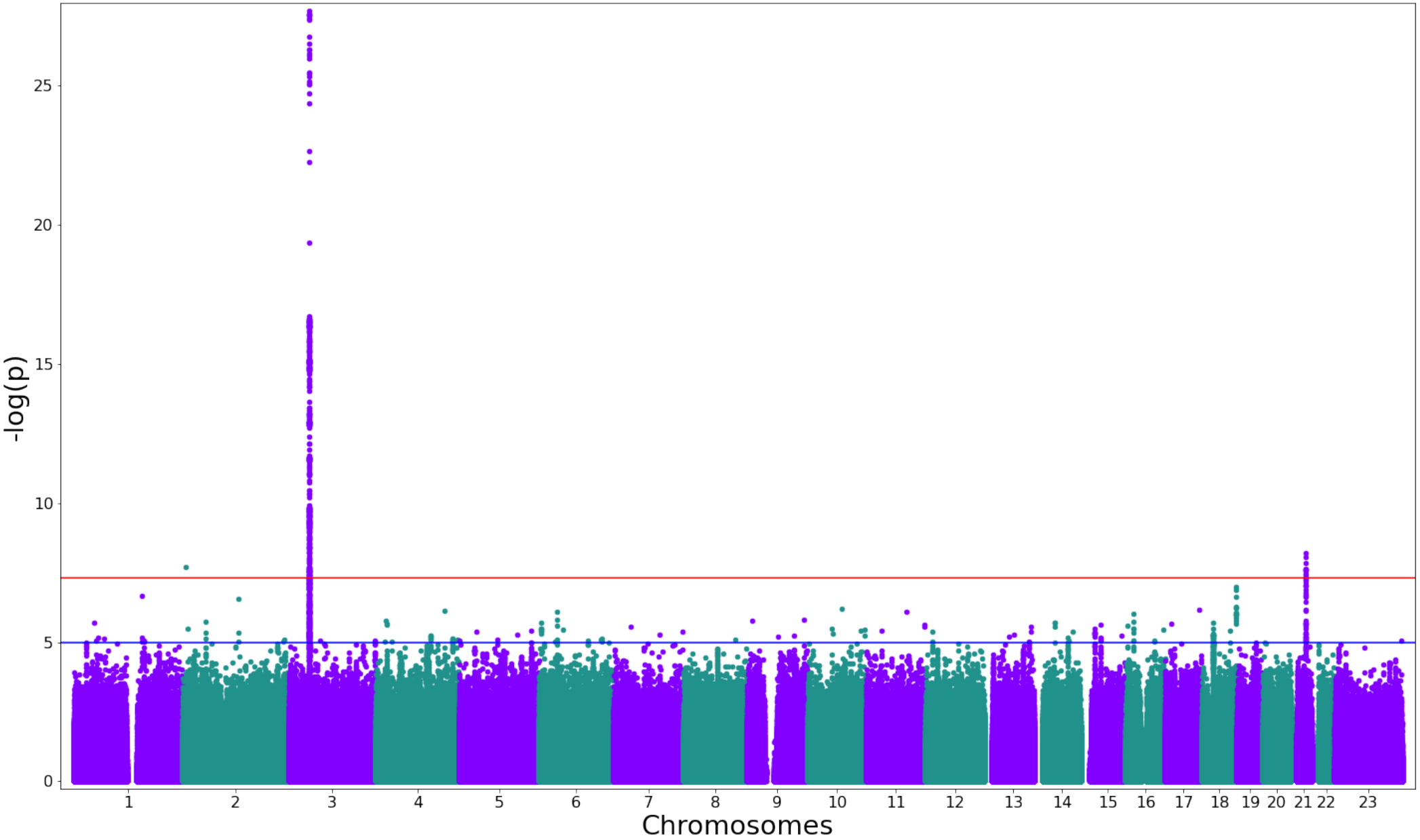


Figura C.19: Gráfico de cuantil-cuantil de la significancia (p-value) de la iniciativa internacional Covid19hg.


-log(p)
25
20
15
10
5
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
Chromosomes

C.2.2. Todas las variantes

Figura C.20: [Pagina 126] GWAS sobre todas las variantes imputadas utilizando datos del proyecto. La significancia, expresada como $-\log(p)$, se calcula obteniendo el p-value de los coeficientes del modelo de regresión logística utilizando la variante, el análisis de componentes principales (PCA) de todas las variantes, edad y sexo de los participantes.

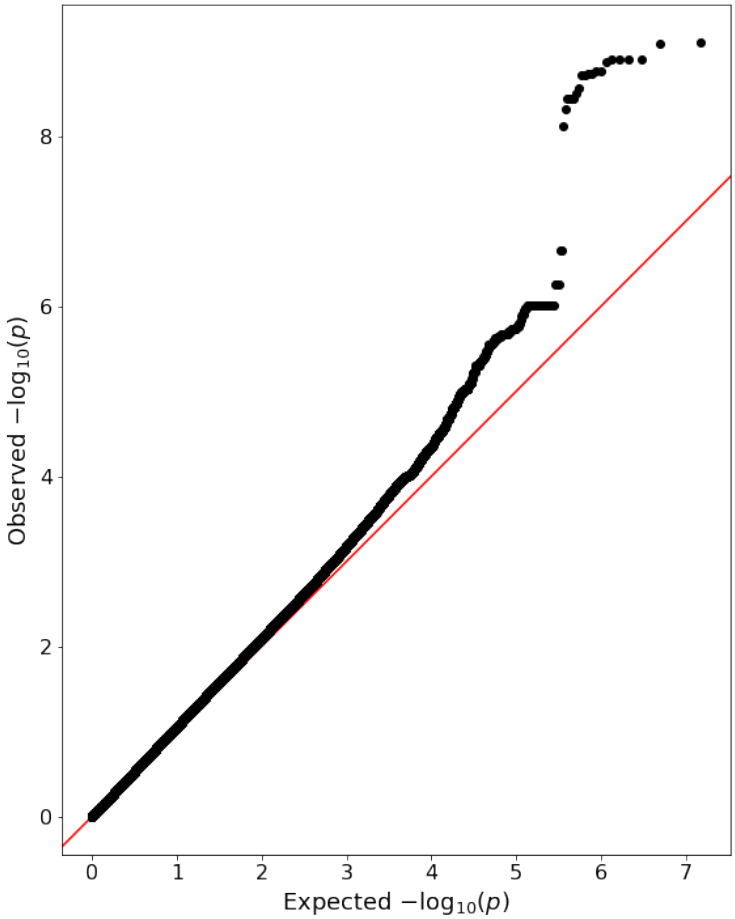
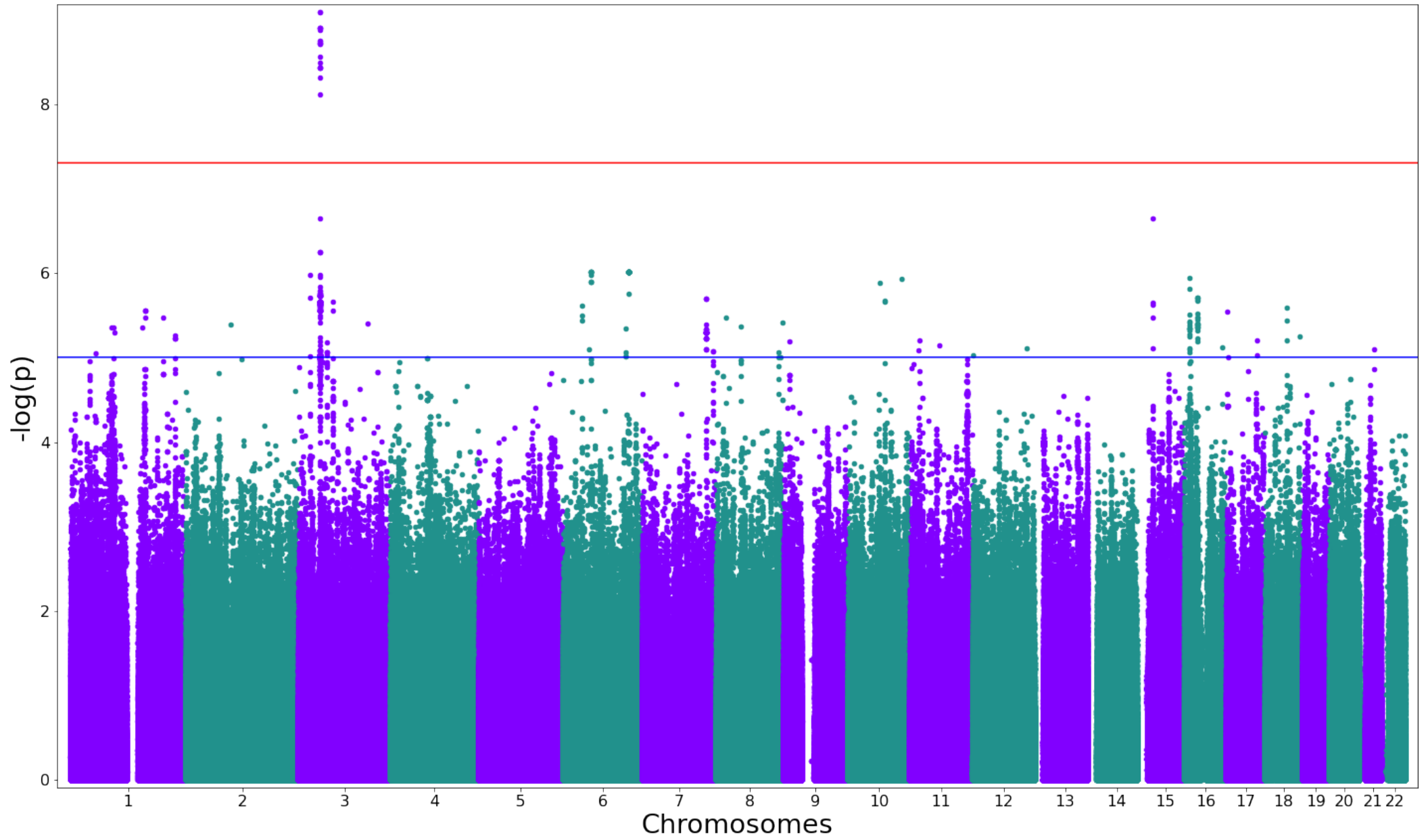


Figura C.21: Gráfico de cuantil-cuantil de la significancia (*p-value*) de todas las variantes imputadas usando datos del proyecto.


-log(p)
0
2
4
6
8
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
Chromosomes

C.2.3. Cromosoma 3

Figura C.22: [Pagina 128] GWAS sobre solo los SNPs del cromosoma 3 de las variantes imputadas utilizando datos del proyecto. La significancia, expresada como $-\log(p)$, se calcula obteniendo el p-value de los coeficientes del modelo de regresión logística utilizando la variante, el análisis de componentes principales (PCA) de todas las variantes, edad y sexo de los participantes. Se extrae desde el GWAS de la **Figura C.20**.

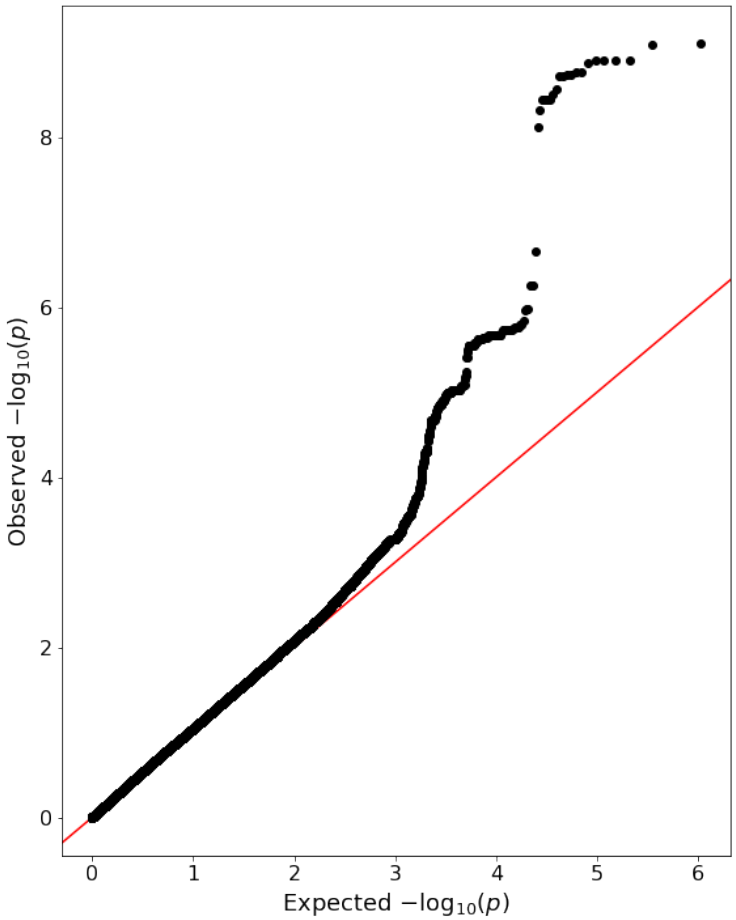


Figura C.23: Gráfico de cuantil-cuantil de la significancia (p-*value*) de las variantes imputadas del cromosoma 3 usando datos del proyecto.

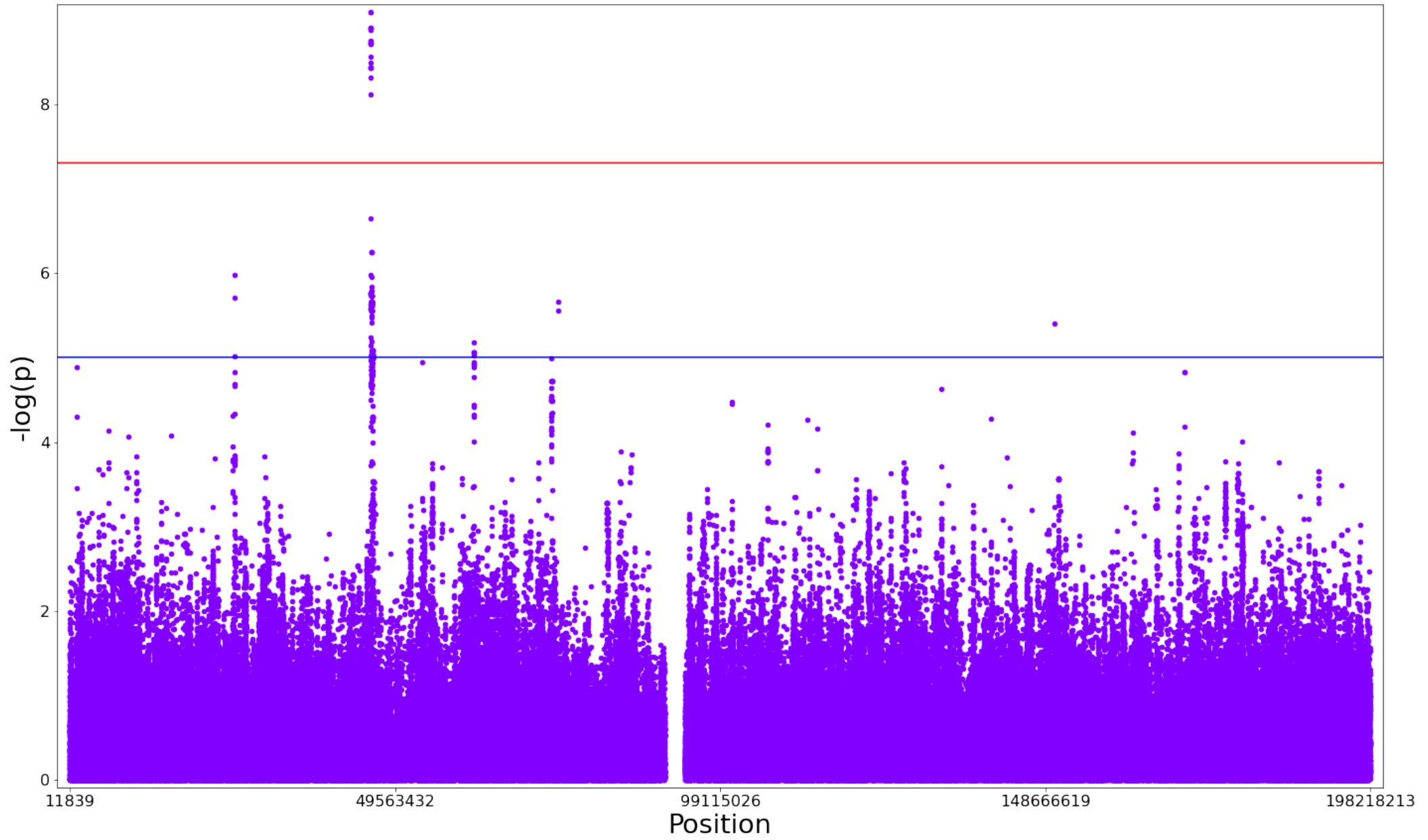

-log(p)
0
2
4
6
8
11839
49563432
99115026
148666619
198218213
Position

Figura C.24: [Pagina 130] GWAS de variantes imputadas utilizando datos del proyecto solo para el cromosoma 3. La significancia, expresada como $-\log(p)$, se calcula obteniendo el p-value de los coeficientes del modelo de regresión logística utilizando la variante y el análisis de componentes principales (PCA) de todas las variantes.

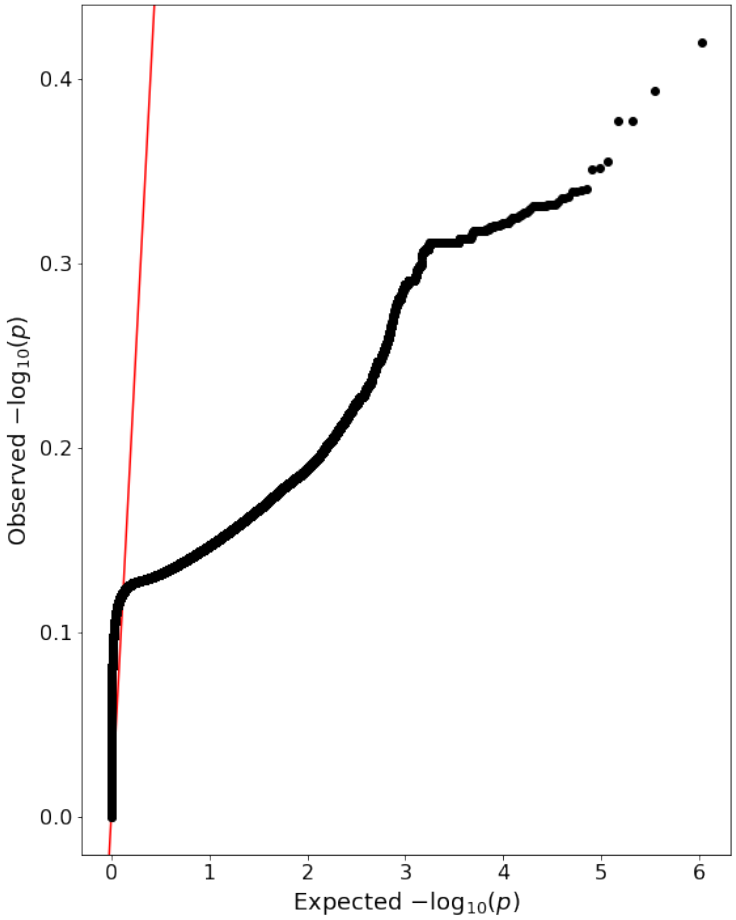


Figura C.25: Gráfico de cuantil-cuantil de la significancia (p-value) de las variantes imputadas usando datos del proyecto solo para el cromosoma 3.

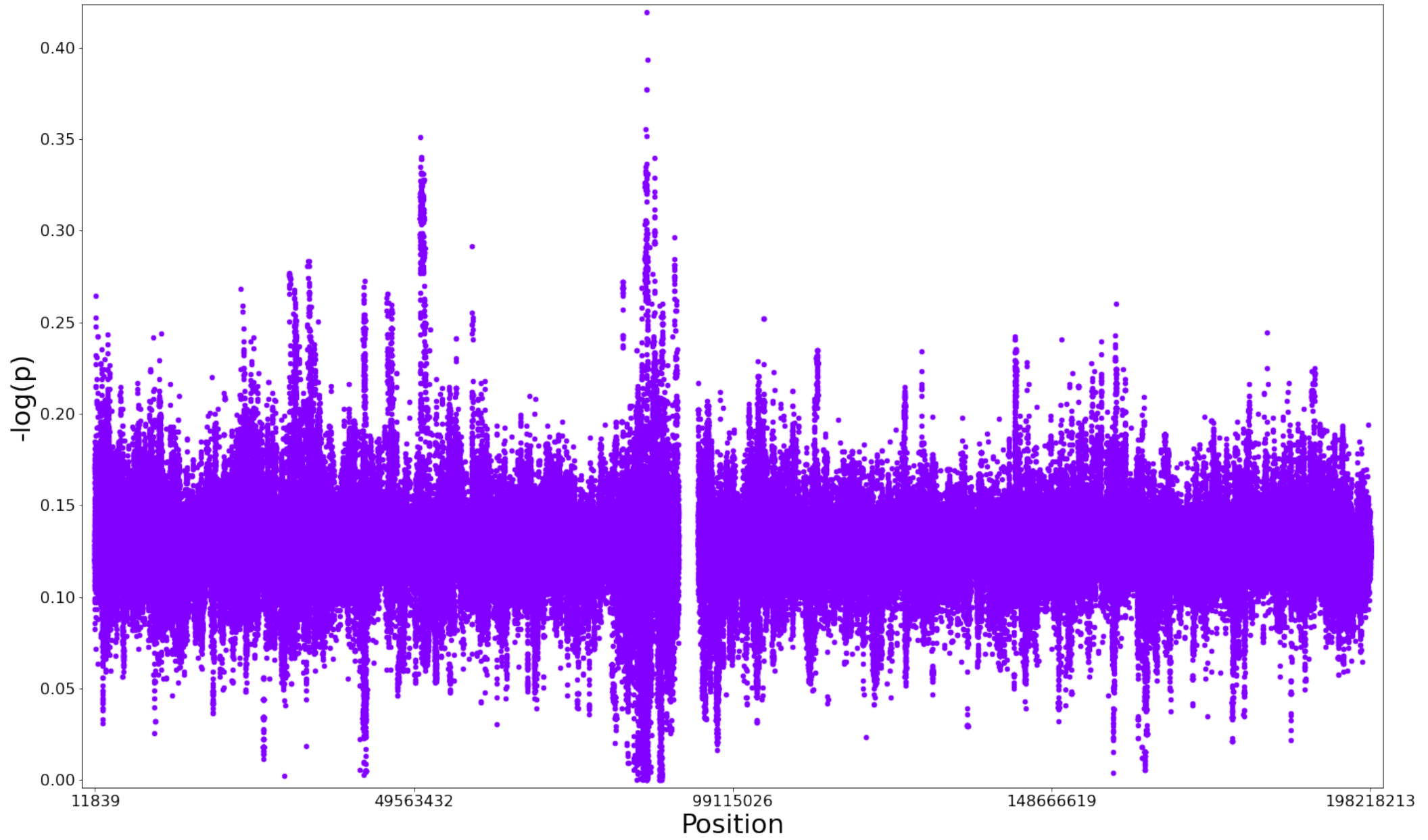

0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
-log(p)
11839
49563432
99115026
148666619
198218213
Position

Figura C.26: [Pagina 132] GWAS de variantes imputadas utilizando datos del proyecto solo para el cromosoma 3. La significancia, expresada como $-\log(p)$, se calcula obteniendo el p-value de los coeficientes del modelo de regresión logística utilizando la variante, el análisis de componentes principales (PCA) de todas las variantes y las variables clínicas seleccionadas.

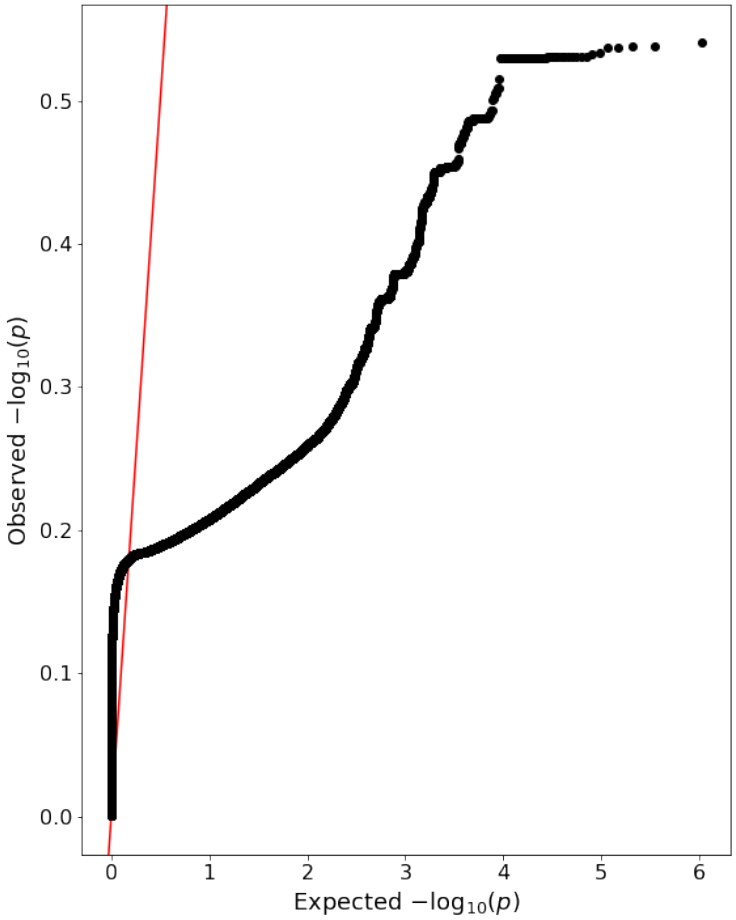
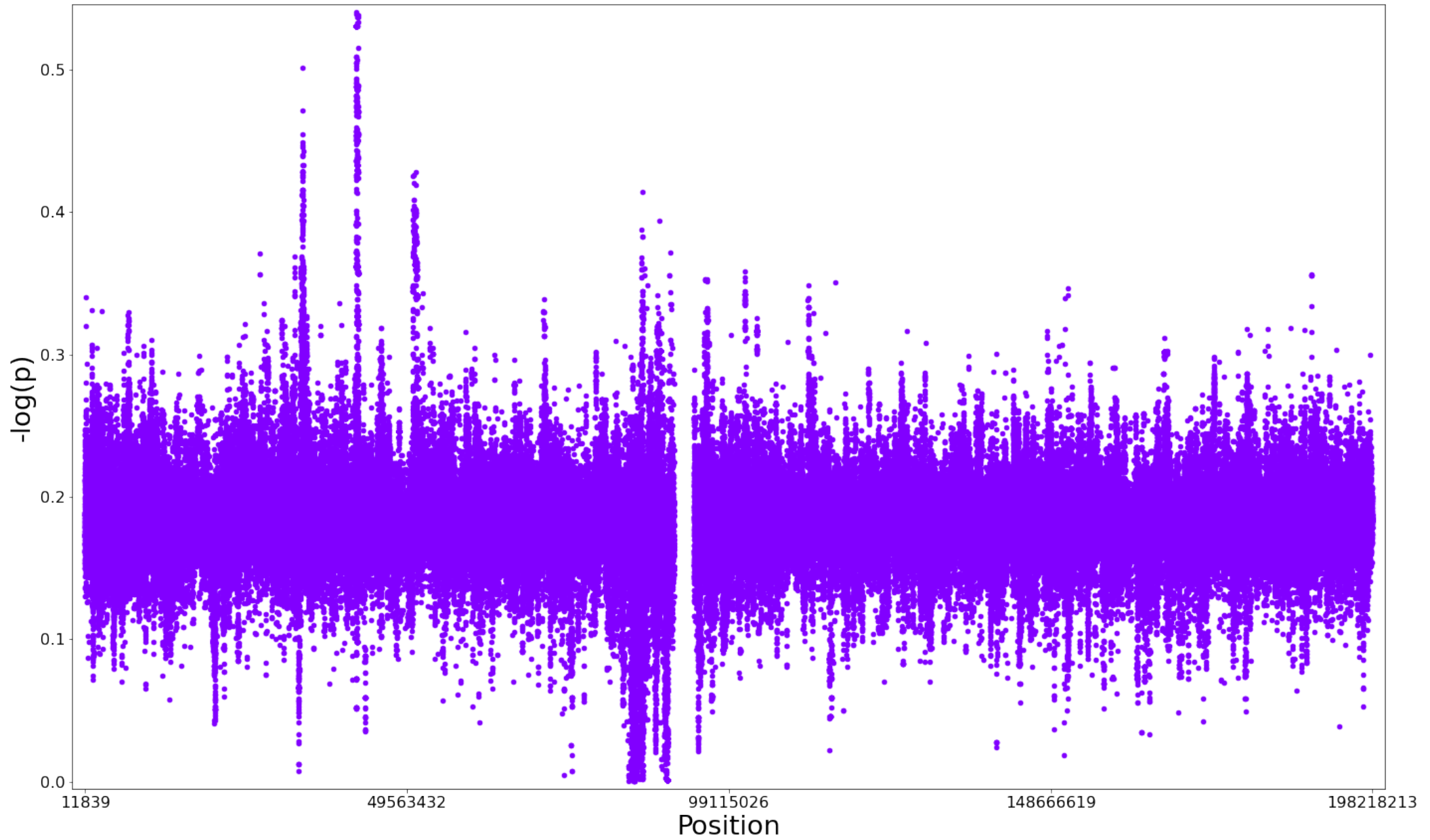


Figura C.27: Gráfico de cuantil-cuantil de la significancia (p-value) de las variantes imputadas usando datos del proyecto solo para el cromosoma 3, utilizando también las variables clínicas seleccionadas


0.5
0.4
0.3
0.2
0.1
0.0
-log(p)
11839
49563432
99115026
148666619
198218213
Position

C.2.4. SNPs Genotipificados por microarreglo

Figura C.28: [Pagina 134] GWAS sobre solo los SNPs originalmente genotipificados para el proyecto. La significancia, expresada como $-\log(p)$, se calcula obteniendo el p-value de los coeficientes del modelo de regresión logística utilizando la variante, el análisis de componentes principales (PCA) de todas las variantes, edad y sexo de los participantes. Se extrae desde el GWAS de la **Figura C.20**.

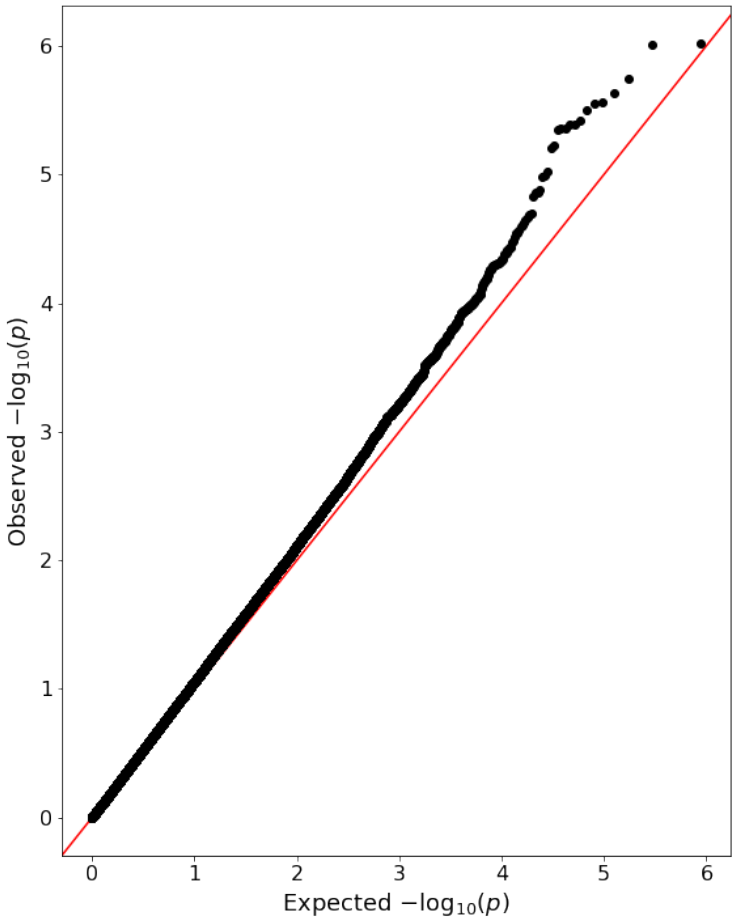


Figura C.29: Gráfico de cuantil-cuantil de la significancia (*p-value*) de las variantes originalmente genotipificadas usando datos del proyecto.

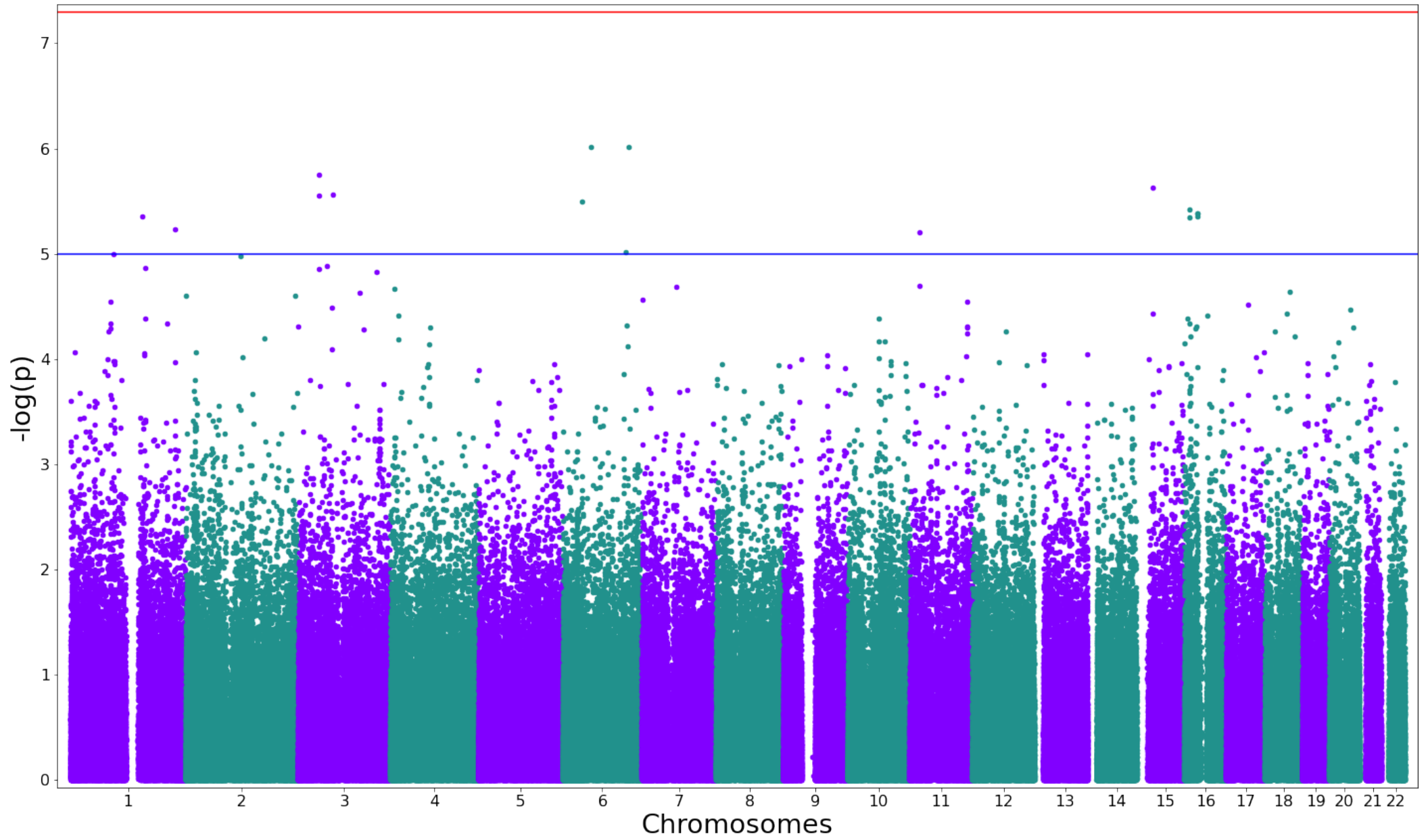

-log(p)
0
1
2
3
4
5
6
7
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
Chromosomes

Figura C.30: [Pagina 136] GWAS de variantes que originalmente se obtienen de la genotipificación utilizando datos del proyecto. La significancia, expresada como $-\log(p)$, se calcula obteniendo el p-value de los coeficientes del modelo de regresión logística utilizando la variante y el análisis de componentes principales (PCA) de todas las variantes.

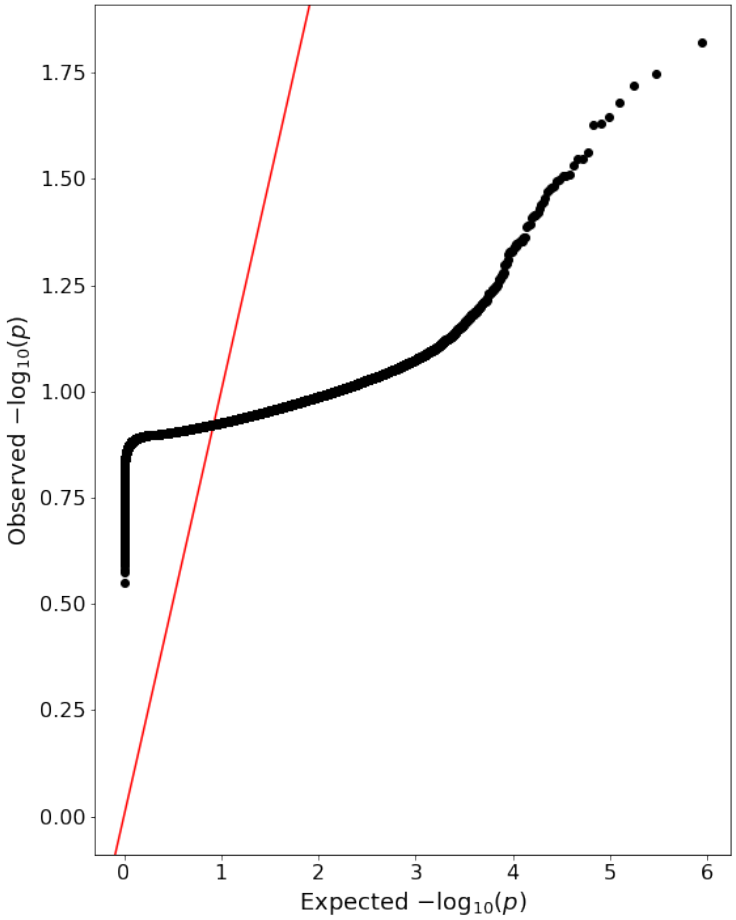


Figura C.31: Gráfico de cuantil-cuantil de la significancia (p-value) de las variantes que originalmente se obtienen de la genotipificación usando datos del proyecto.

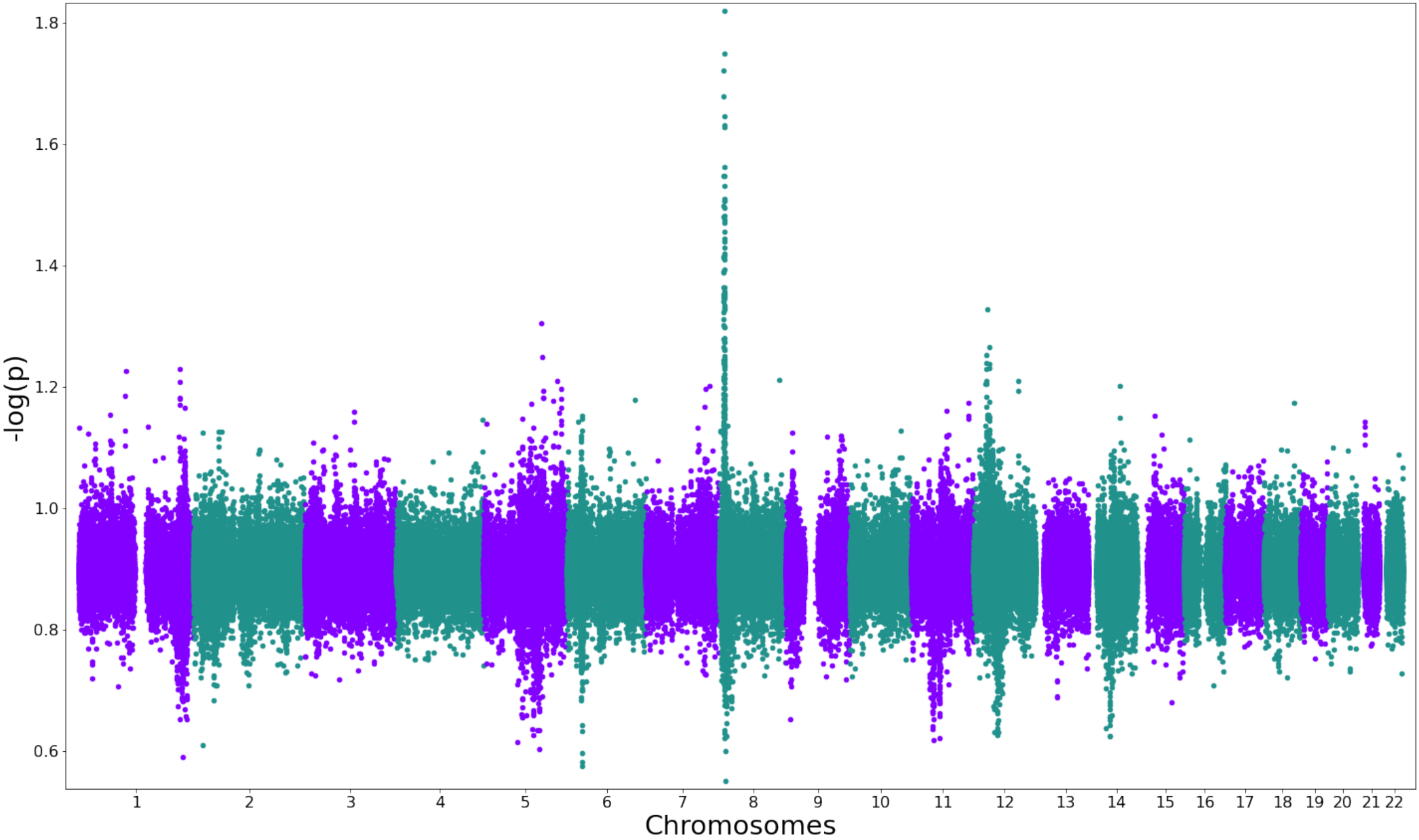

-log(p)
1.8
1.6
1.4
1.2
1.0
0.8
0.6
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
Chromosomes

Figura C.32: [Pagina 138] GWAS de variantes que originalmente se obtienen de la genotipificación utilizando datos del proyecto. La significancia, expresada como $-\log(p)$, se calcula obteniendo el p-value de los coeficientes del modelo de regresión logística utilizando la variante, el análisis de componentes principales (PCA) de todas las variantes y las variables clínicas seleccionadas.

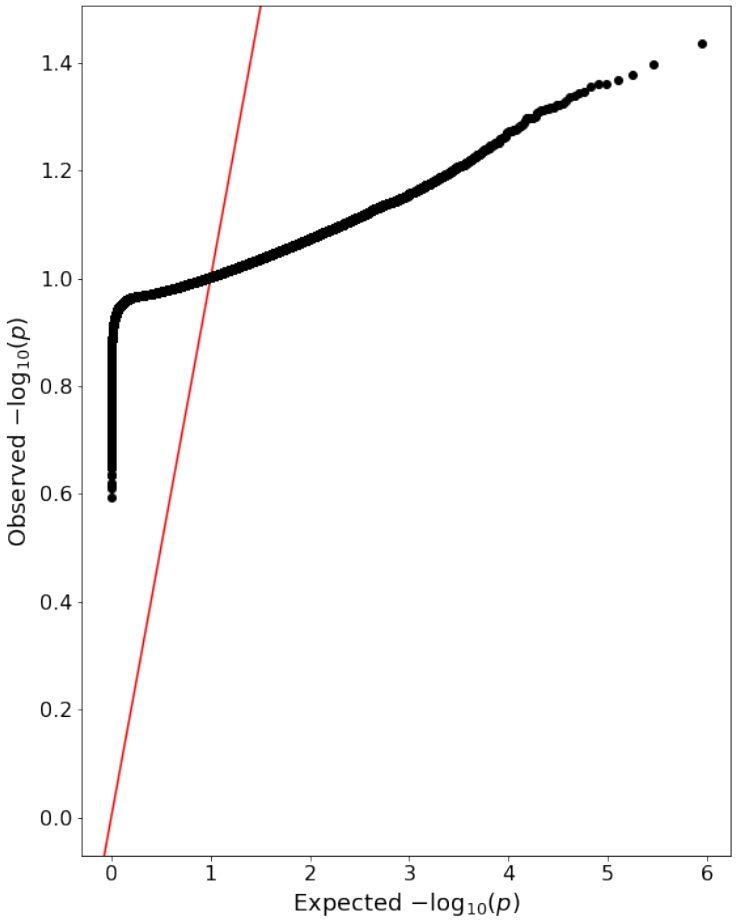
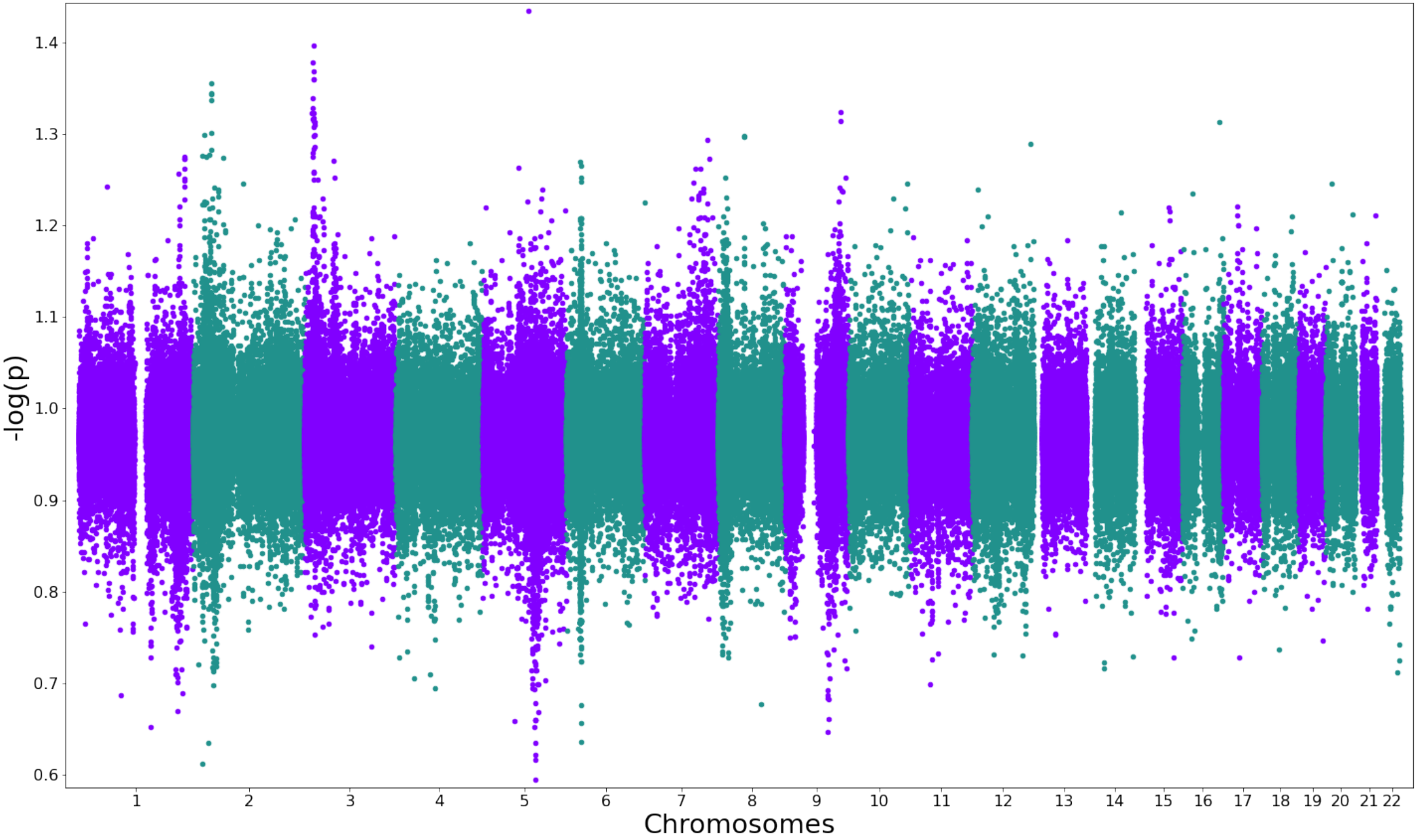


Figura C.33: Gráfico de cuantil-cuantil de la significancia (p-value) de las variantes que originalmente se obtienen de la genotipificación usando datos del proyecto, utilizando también las variables clínicas seleccionadas


-log(p)
1.4
1.3
1.2
1.1
1.0
0.9
0.8
0.7
0.6
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
Chromosomes

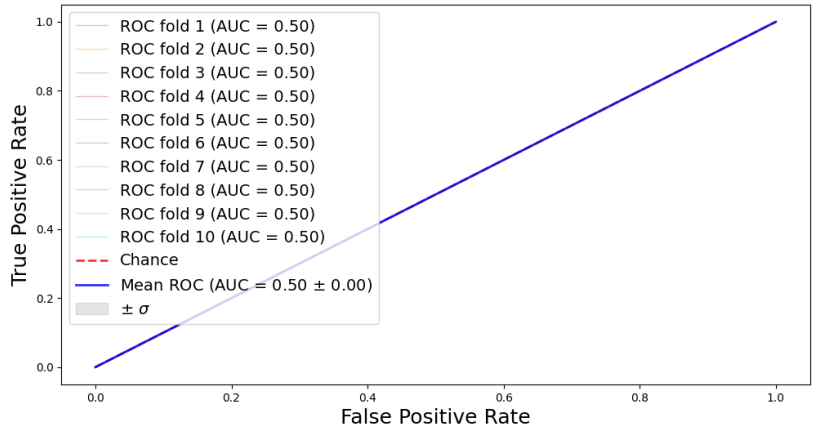
C.3. Datos Genéticos

C.3.1. Modelos de ML

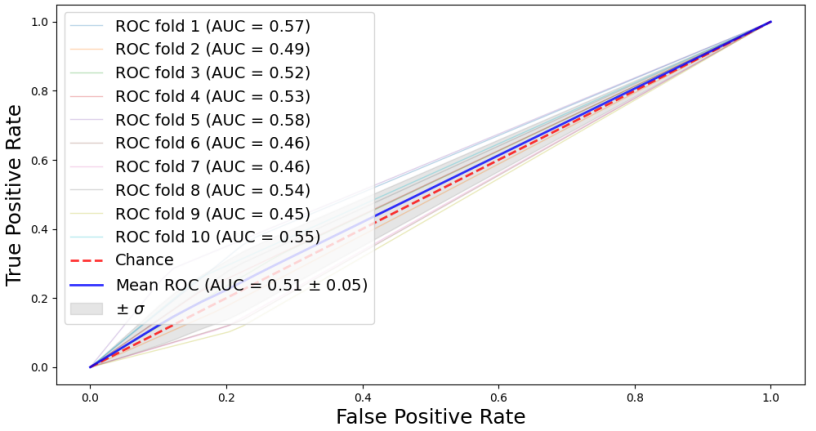
Figura C.34: [Pagina 140] Curva de ROC de los modelos ML sobre las variantes genéticas seleccionadas para el análisis de hospitalizados entre la población infectada confirmada.

Figura C.35: [Pagina 141] Matrices de confusión para los modelos ML sobre las variantes genéticas seleccionadas para el análisis de hospitalizados entre la población infectada confirmada.

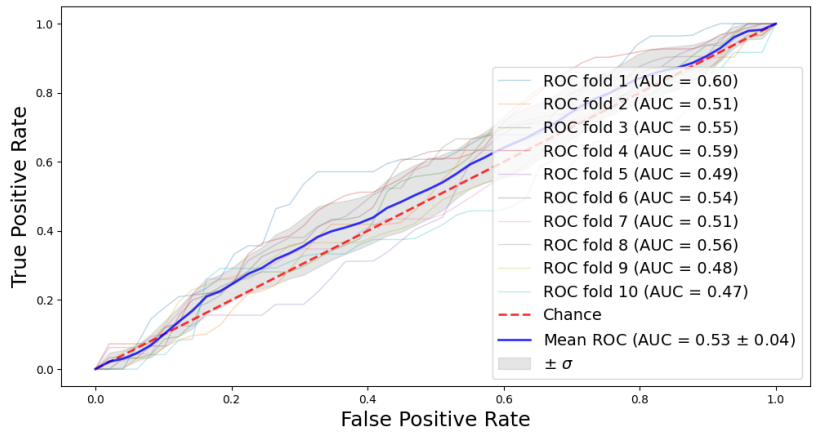
Dummy Mode Classifier



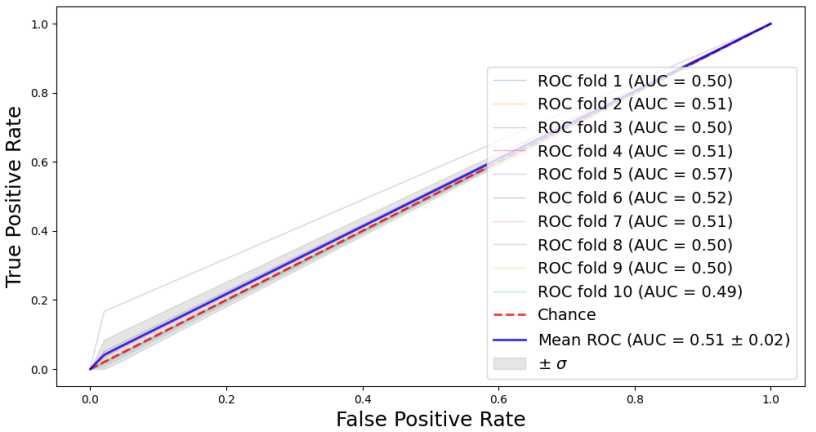
Decision Tree



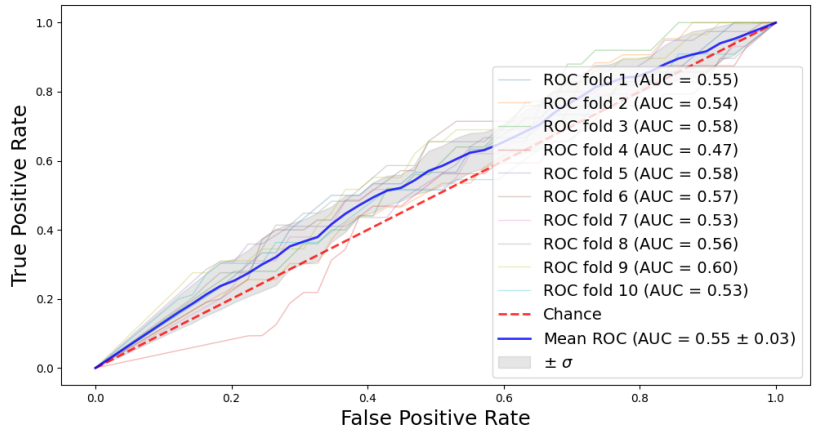
Random Forest



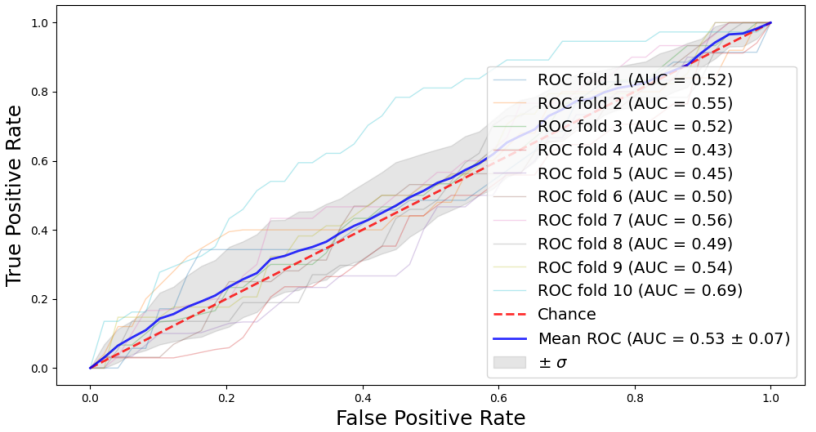
SVM



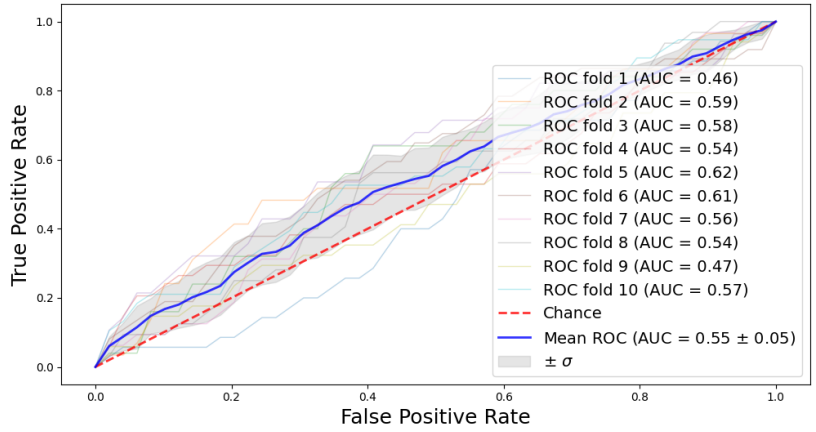
Naive Bayes



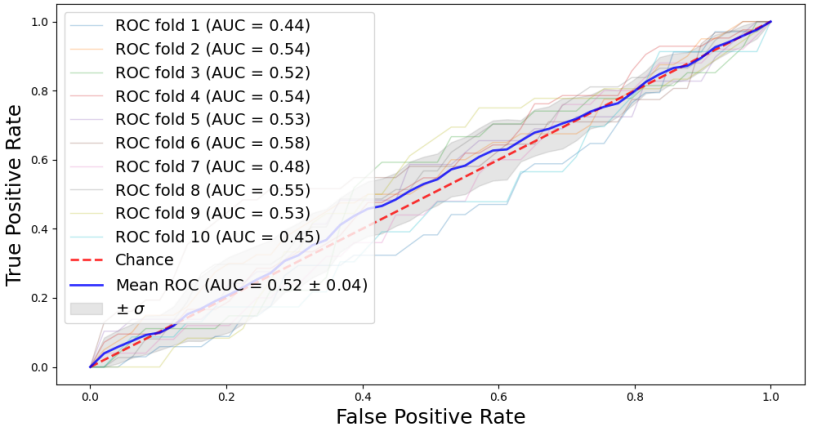
Gaussian Classifier

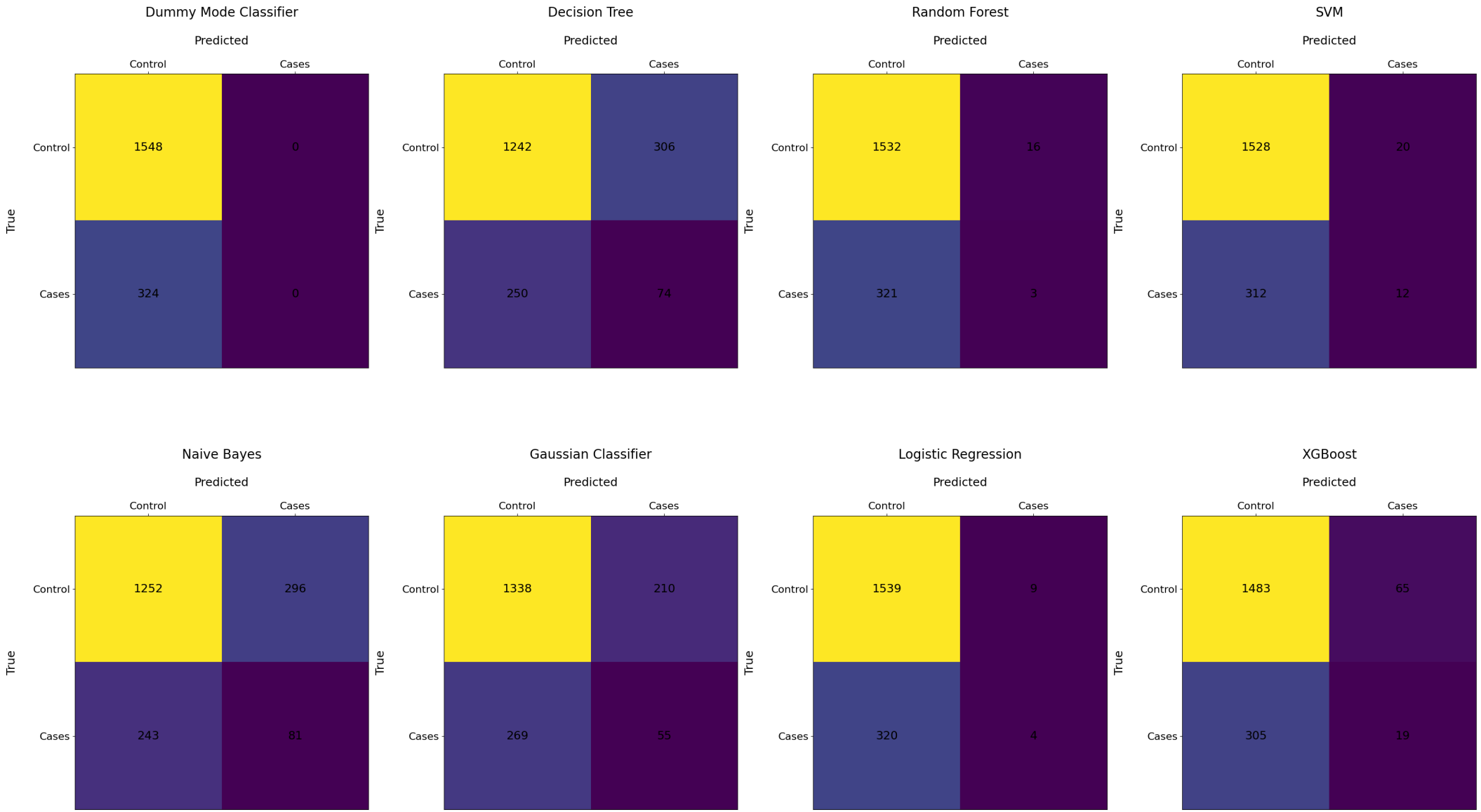


Logistic Regression



XGBoost




Dummy Mode Classifier
Predicted
Control
Cases
True
Control
Cases
1548
0
324
0
Decision Tree
Predicted
Control
Cases
True
Control
Cases
1242
306
250
74
Random Forest
Predicted
Control
Cases
True
Control
Cases
1532
16
321
3
SVM
Predicted
Control
Cases
True
Control
Cases
1528
20
312
12
Naive Bayes
Predicted
Control
Cases
True
Control
Cases
1252
296
243
81
Gaussian Classifier
Predicted
Control
Cases
True
Control
Cases
1338
210
269
55
Logistic Regression
Predicted
Control
Cases
True
Control
Cases
1539
9
320
4
XGBoost
Predicted
Control
Cases
True
Control
Cases
1483
65
305
19

C.3.2. Dual-stream CNN

C.3.2.1. Selección de Hiperparámetros

C.3.2.1.1. SNPs Genotipificados por microarreglo

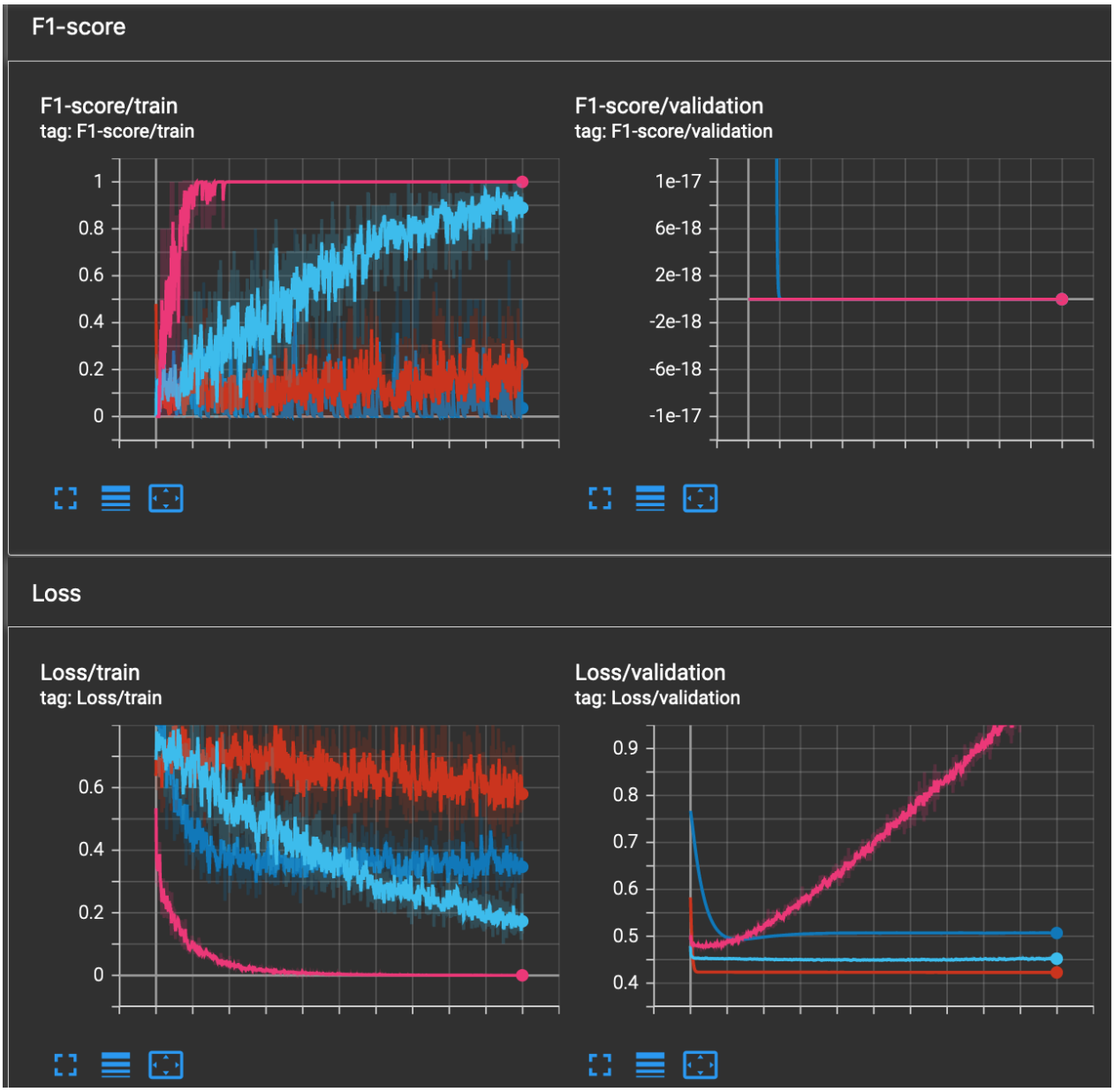


Figura C.36: Experimentos a diferentes *learning rates* con la arquitectura Dual-stream CNN para el análisis de hospitalizados utilizando SNPs originalmente genotipificados por microarreglo. Se realizan con *weight decay* $= wd = 0.0$, la visualización se obtuvo utilizando Tensorboard [22]. Se grafica la métrica $f1$ en la primera fila y la función de perdida (*Cross entropy*) en la segunda. En la primera columna se gráfica los resultados durante el entrenamiento, en la segunda, los resultados sobre el *set* de validación. El experimento se realiza sobre *learning rates* de 10^i con $i \in \{-10, \cdot, -1\}$. Se muestran las mejores métricas obtenidas. En azul: $lr = 10^{-9}$, rojo: $lr = 10^{-8}$, celeste $lr = 10^{-7}$, rosa: $lr = 10^{-6}$.

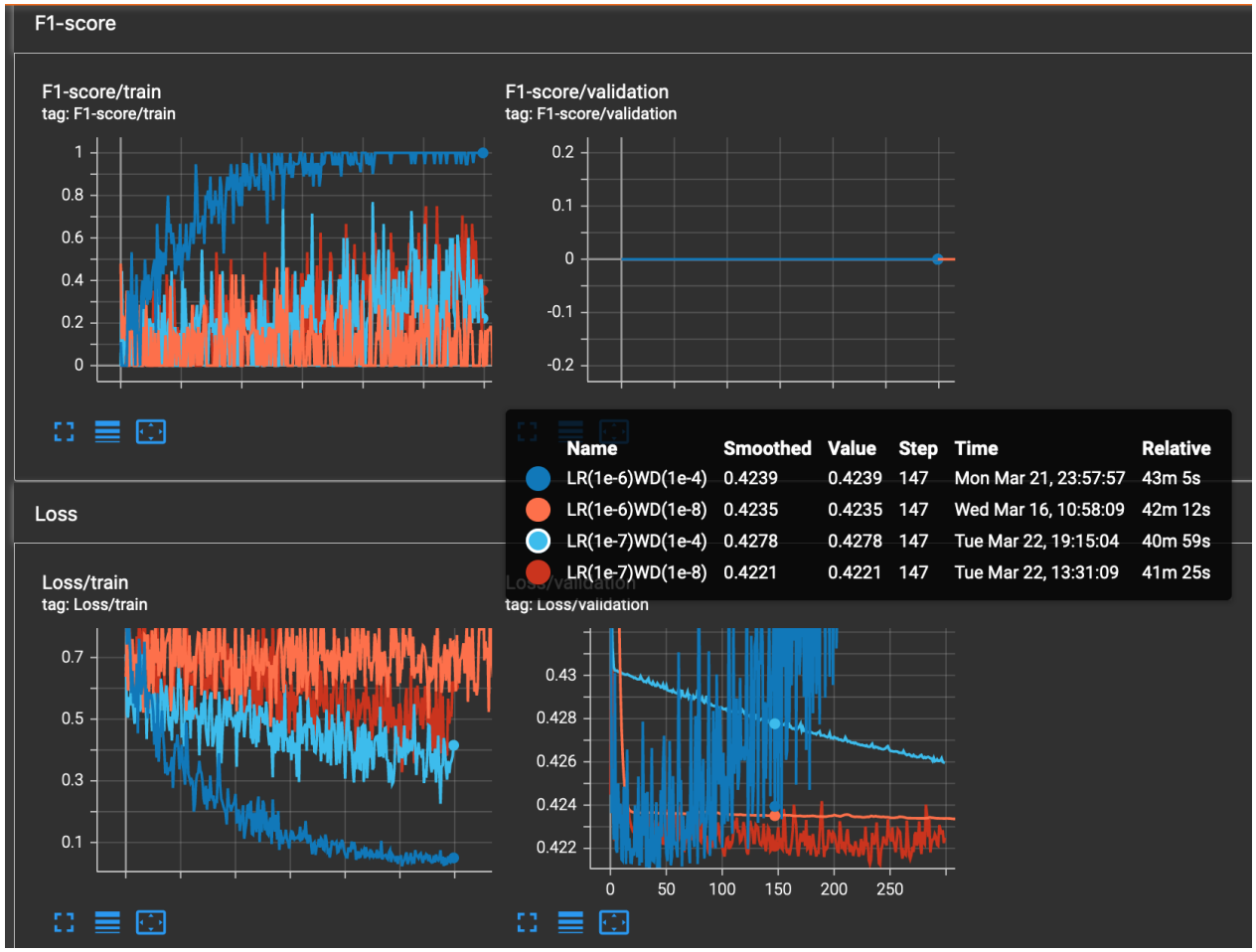


Figura C.37: Experimentos a diferentes *weight decay* con la arquitectura Dual-strem CNN para el análisis de hospitalizados utilizando SNPs originalmente genotipificados por microarreglo. Se realizan con *learning rate* $= lr \in \{10^{-9}, 10^{-8}, 10^{-7}, 10^{-6}\}$ y $wd = 10^i$ con $i \in \{-10, \cdot, -1\}$ para cada lr. La visualización se obtiene utilizando Tensorboard [22]. Se grafica la métrica $f1$ en la primera fila y la función de perdida (*Cross entropy*) en la segunda. En la primera columna se gráfica los resultados durante el entrenamiento, en la segunda, los resultados sobre el *set* de validación. Se muestran las mejores curvas obtenidas. El código color se muestra en el gráfico de función de pérdida para el set de validación.

C.3.2.1.2. SNPs Seleccionados

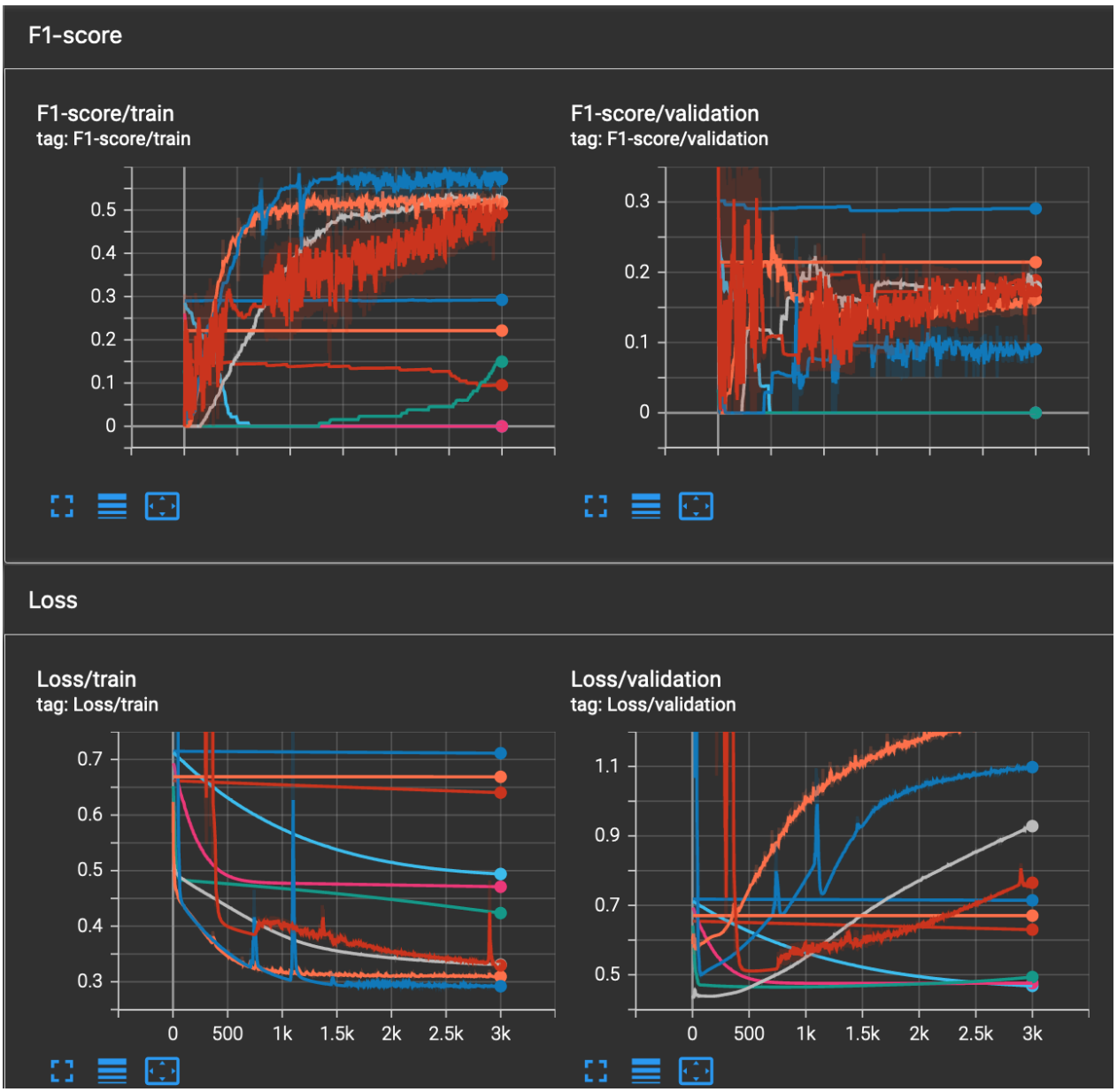


Figura C.38: Experimentos a diferentes *learning rates* con la arquitectura Dual-stream CNN para el análisis de hospitalizados utilizando SNPs seleccionados según COVID19hgi. Se realizan con *weight decay* $= wd = 0.0$, la visualización se obtuvo utilizando Tensorboard [22]. Se grafica la métrica $f1$ en la primera fila y la función de perdida (*Cross entropy*) en la segunda. En la primera columna se gráfica los resultados durante el entrenamiento, en la segunda, los resultados sobre el *set* de validación. El experimento se realiza sobre *learning rates* de 10^i con $i \in \{-10, \cdot, -1\}$.

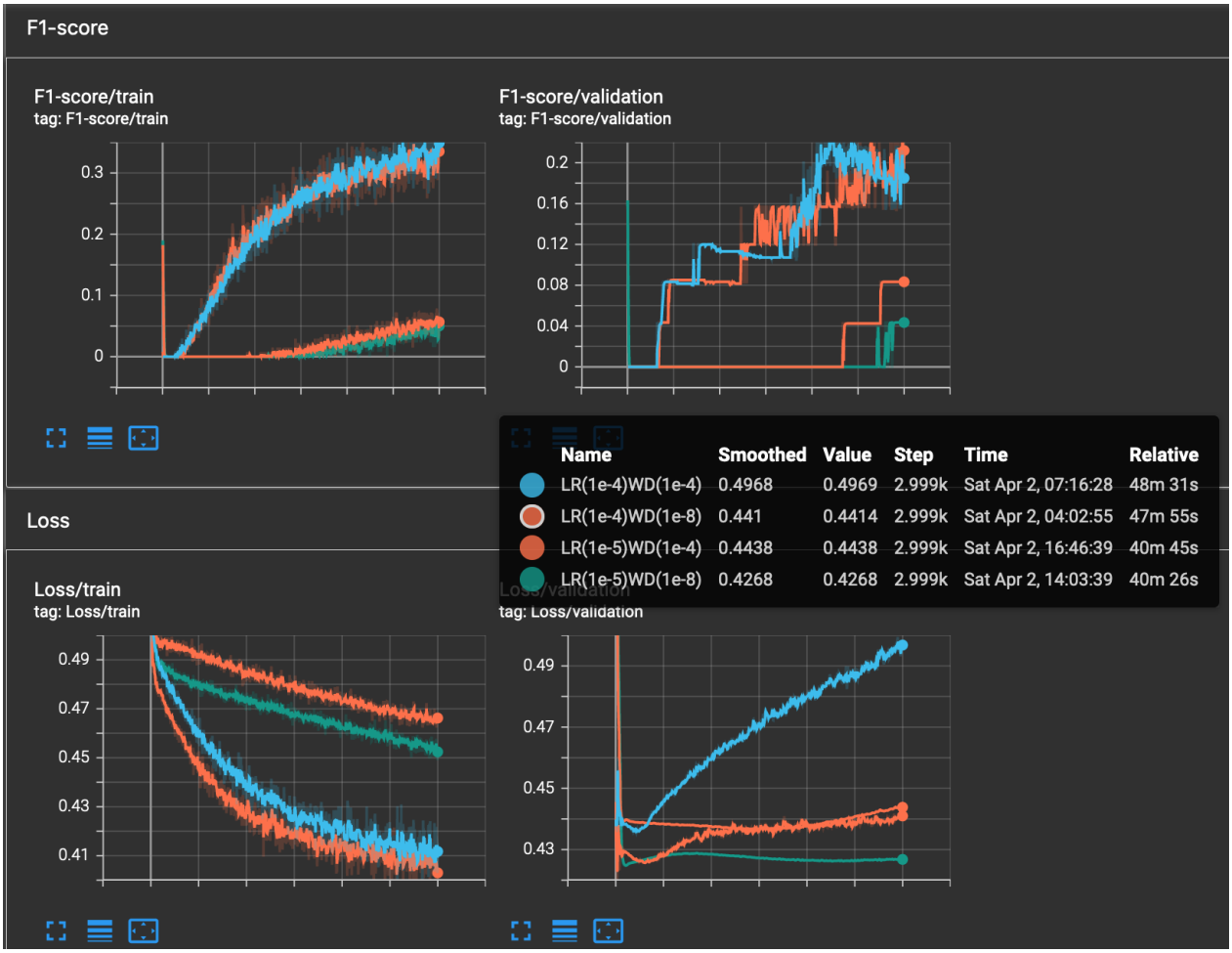


Figura C.39: Experimentos a diferentes *weight decay* con la arquitectura Dual-strem CNN para el análisis de hospitalizados utilizando SNPs seleccionados según COVID19hgi. Se realizan con *learning rate* $= lr \in \{10^{-7}, 10^{-6}, 10^{-5}, 10^{-4}\}$ y $wd = 10^i$ con $i \in \{-10, \cdot, -1\}$ para cada lr. La visualización se obtiene utilizando Tensorboard [22]. Se grafica la métrica $f1$ en la primera fila y la función de perdida (*Cross entropy*) en la segunda. En la primera columna se gráfica los resultados durante el entrenamiento, en la segunda, los resultados sobre el *set* de validación. Se muestran las mejores curvas obtenidas. El código color se muestra en el gráfico de función de pérdida para el set de validación.

C.3.2.2. Experimentos con desbalance

C.3.2.2.1. SNPs Genotipificados por Microarreglo

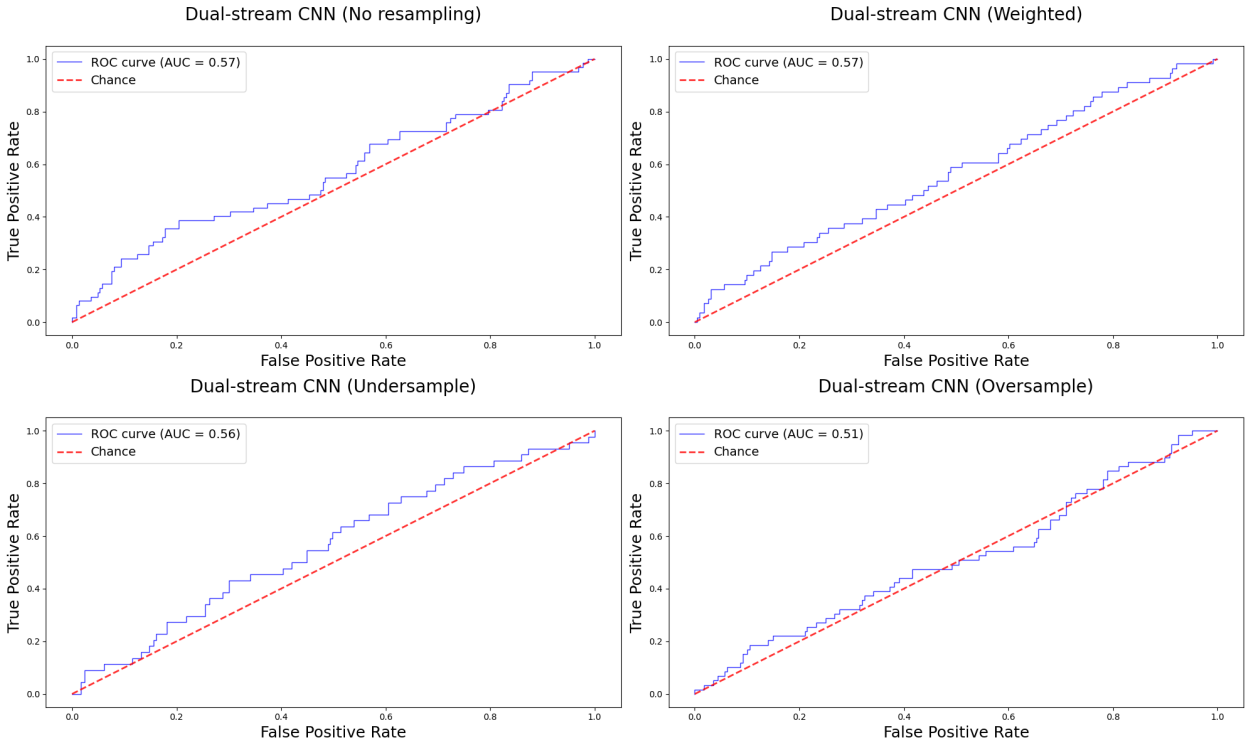


Figura C.40: Curva de ROC del modelo Dual-stream CNN sobre las variantes genéticas originalmente genotipificadas para el análisis de hospitalizados entre la población infectada confirmada. [**Arriba, izquierda**] Se muestran los resultados para ningún tipo de resampleo,[**Arriba, derecha**] los resultados añadiendo un ponderador a la función de *loss*, [**Abajo, izquierda**] utilizando *undersampling* y [**Abajo, derecha**] utilizando *oversampling*.

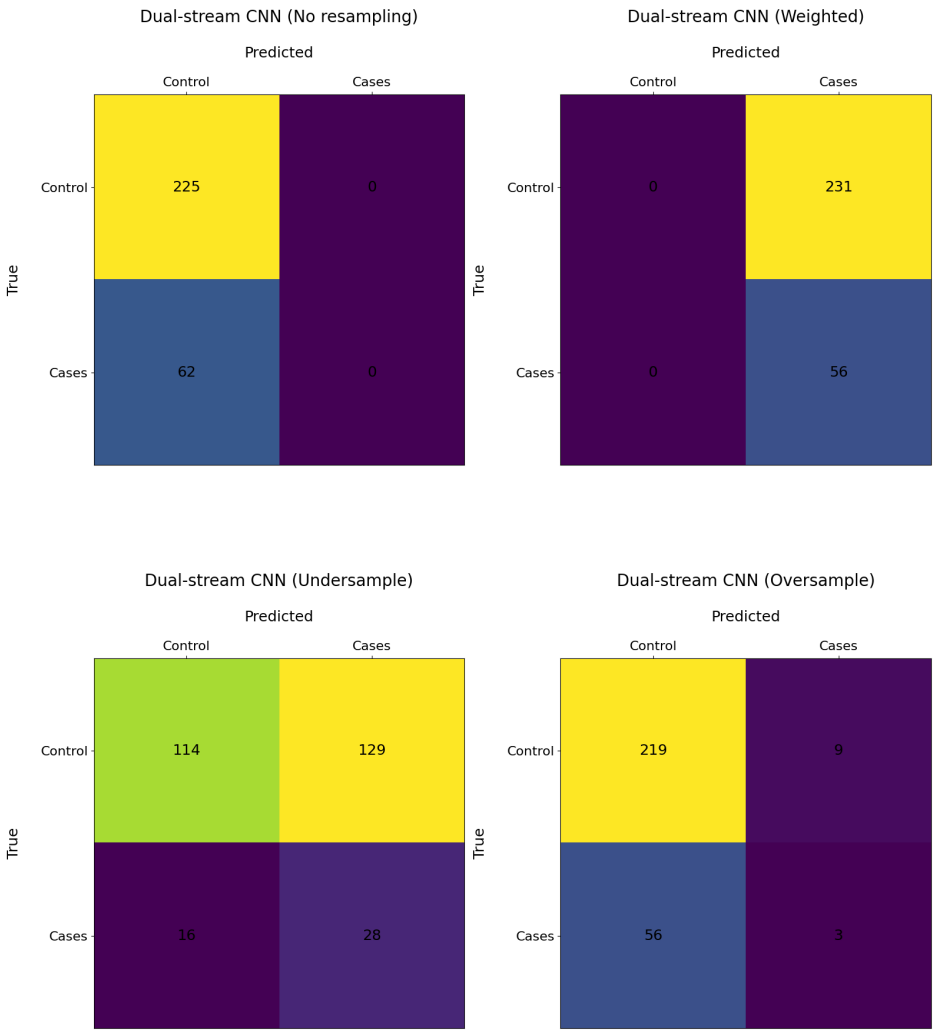


Figura C.41: Matrices de confusión del modelo Dual-stream CNN sobre las variantes genéticas originalmente genotipificadas para el análisis de hospitalizados entre la población infectada confirmada. [**Arriba, izquierda**] Se muestran los resultados para ningún tipo de resampleo,[**Arriba, derecha**] los resultados añadiendo un ponderador a la función de *loss*, [**Abajo, izquierda**] utilizando *undersampling* y [**Abajo, derecha**] utilizando *oversampling*.

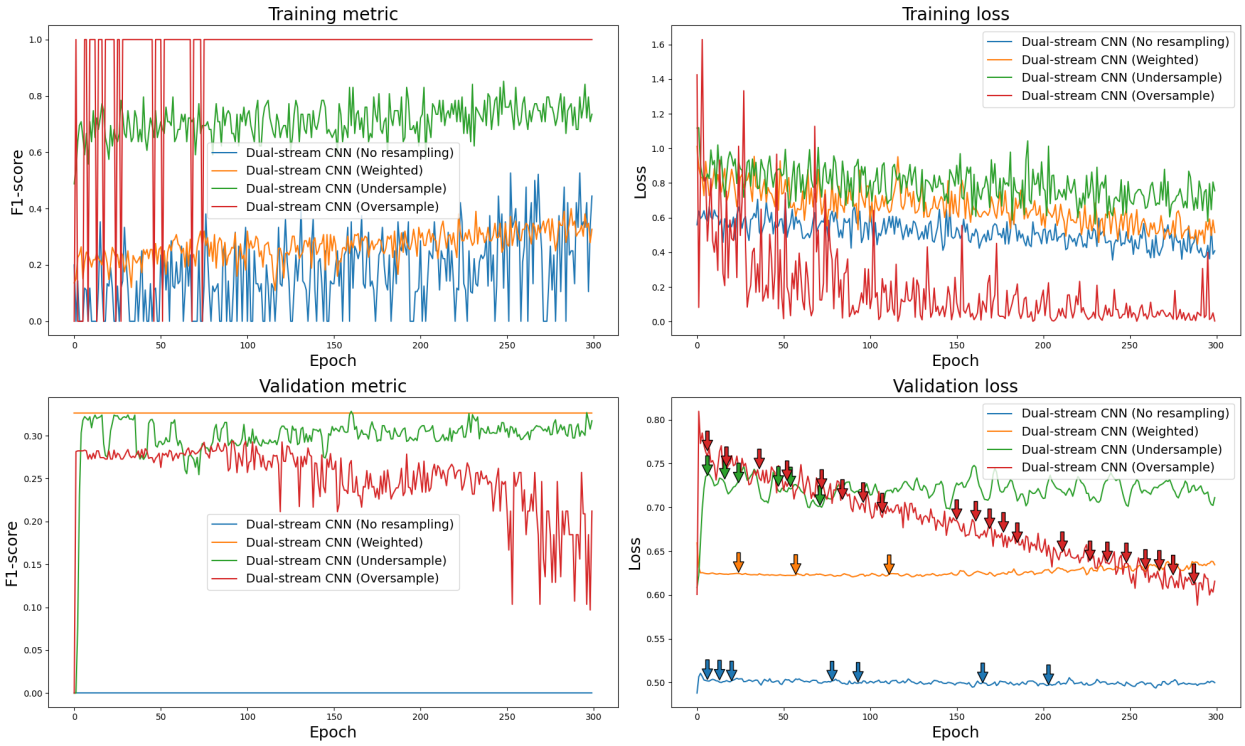


Figura C.42: Curva de entrenamiento de la arquitectura Dual-stream CNN sobre variantes genéticas originalmente genotipificadas para el análisis de hospitalizados entre la población infectada para cada aproximación por desbalance. A la esquina superior izquierda se muestra la métrica de entrenamiento, F1-score, a la esquina inferior izquierda la métrica de validación, a la esquina superior derecha la función de pérdida de entrenamiento, y a la esquina inferior derecha la función de pérdida de validación. En esta última se muestran flechas que indican que el modelo se guarda por *Early Stopping* al estancarse o disminuir la función de pérdida.

C.3.2.2.2. Cromosoma 3

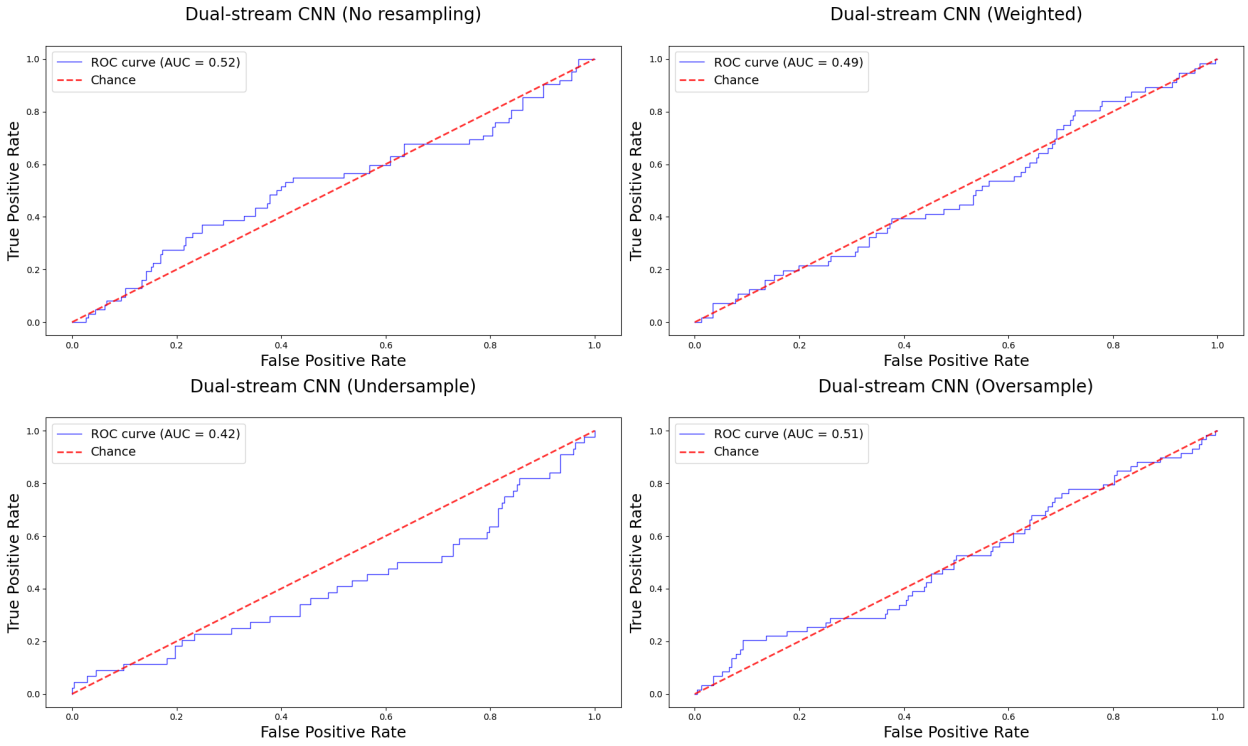


Figura C.43: Curva de ROC del modelo Dual-stream CNN sobre las variantes genéticas imputadas solo en el cromosoma 3 para el análisis de hospitalizados entre la población infectada confirmada. [**Arriba, izquierda**] Se muestran los resultados para ningún tipo de resampleo,[**Arriba, derecha**] los resultados añadiendo un ponderador a la función de *loss*, [**Abajo, izquierda**] utilizando *undersampling* y [**Abajo, derecha**] utilizando *oversampling*.

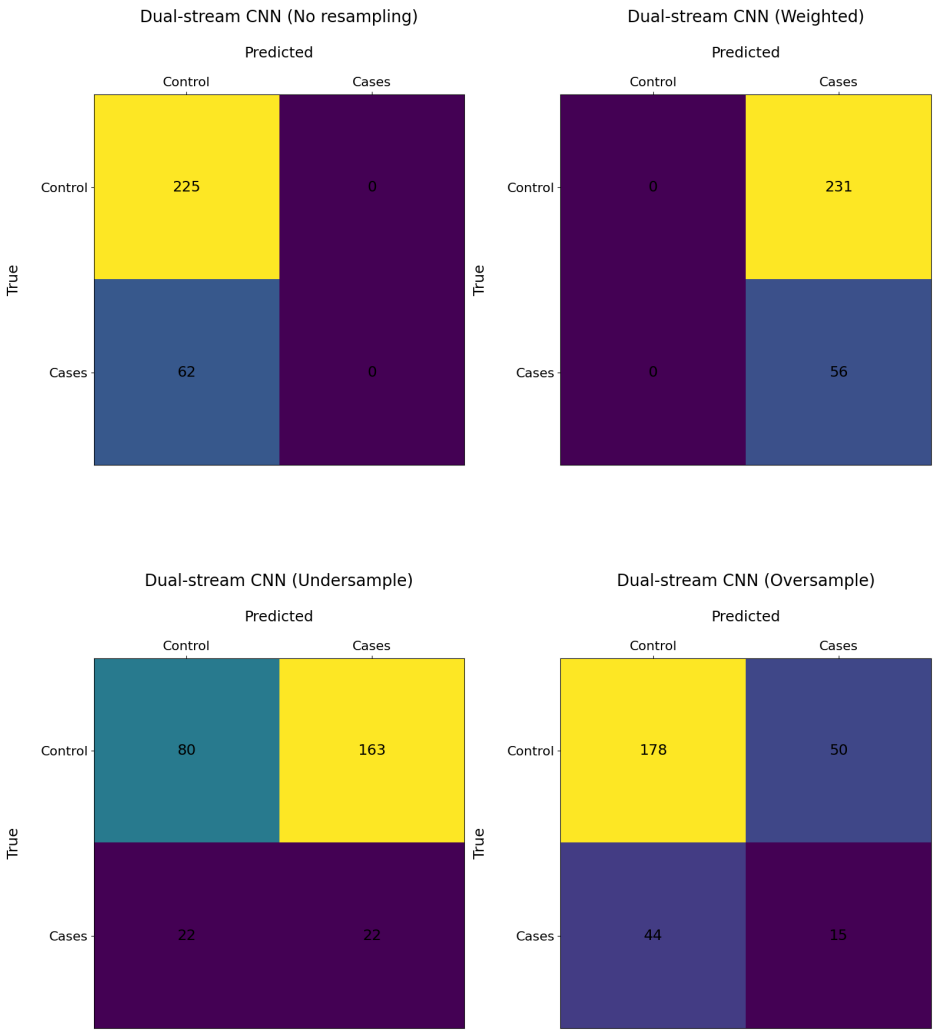


Figura C.44: Matrices de confusión del modelo Dual-stream CNN sobre las variantes genéticas imputadas solo en el cromosoma 3 para el análisis de hospitalizados entre la población infectada confirmada. [**Arriba, izquierda**] Se muestran los resultados para ningún tipo de resampleo,[**Arriba, derecha**] los resultados añadiendo un ponderador a la función de *loss*, [**Abajo, izquierda**] utilizando *undersampling* y [**Abajo, derecha**] utilizando *oversampling*.

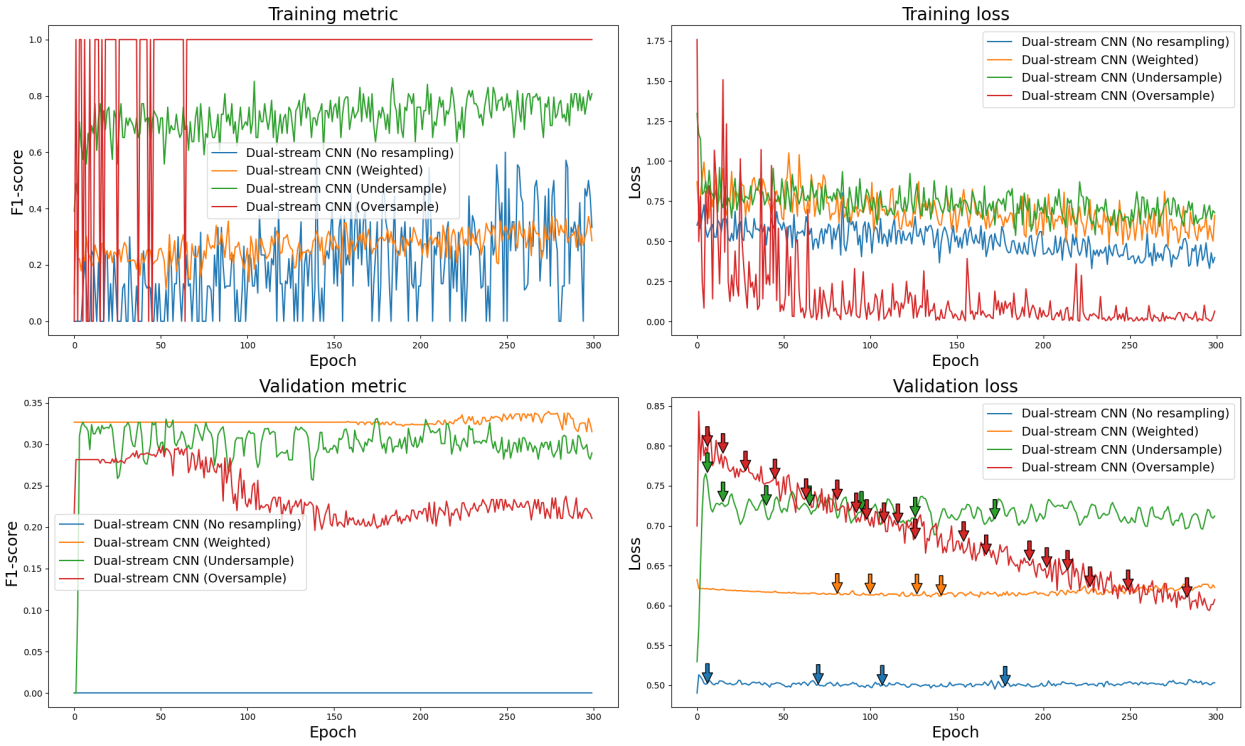


Figura C.45: Curva de entrenamiento de la arquitectura Dual-stream CNN sobre variantes genéticas imputadas solo en el cromosoma 3 para el análisis de hospitalizados entre la población infectada para cada aproximación por desbalance. A la esquina superior izquierda se muestra la métrica de entrenamiento, F1-score, a la esquina inferior izquierda la métrica de validación, a la esquina superior derecha la función de pérdida de entrenamiento, y a la esquina inferior derecha la función de pérdida de validación. En esta última se muestran flechas que indican que el modelo se guarda por *Early Stopping* al estancarse o disminuir la función de pérdida.

C.3.2.2.3. SNPs Seleccionados

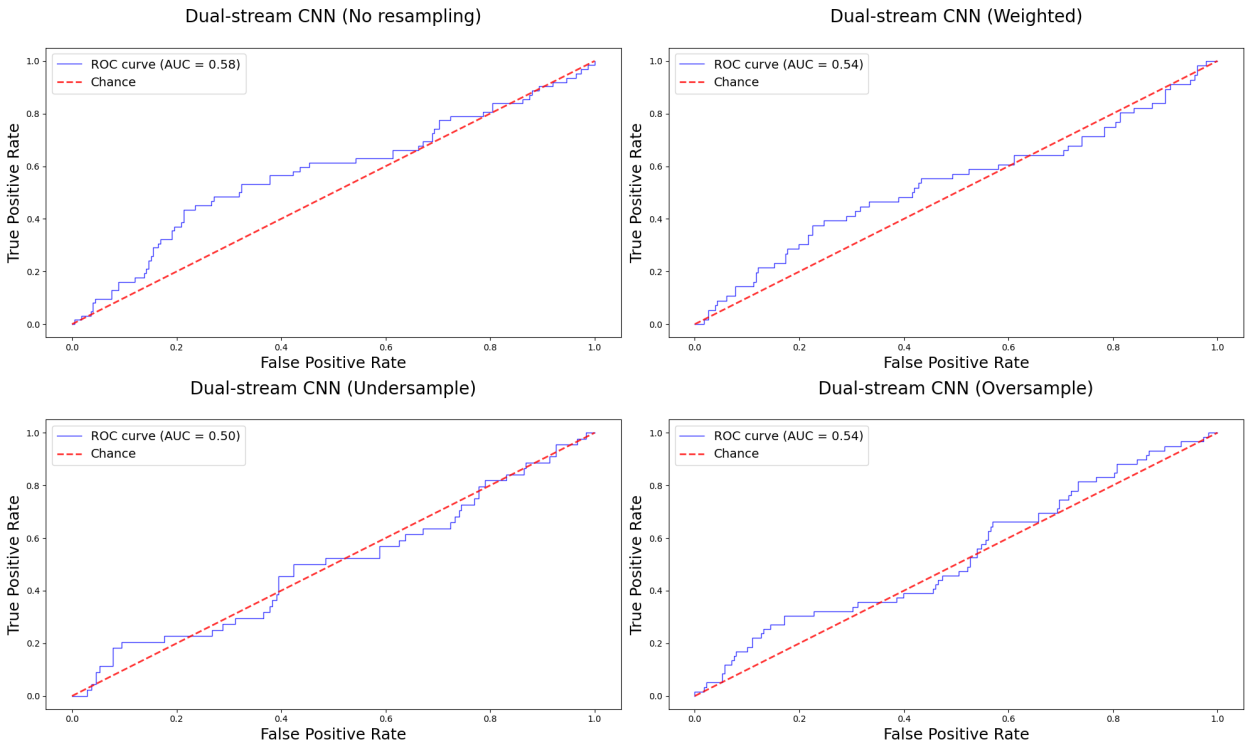


Figura C.46: Curva de ROC del modelo Dual-stream CNN sobre las variantes genéticas seleccionadas para el análisis de hospitalizados entre la población infectada confirmada. [**Arriba, izquierda**] Se muestran los resultados para ningún tipo de resampleo,[**Arriba, derecha**] los resultados añadiendo un ponderador a la función de *loss*, [**Abajo, izquierda**] utilizando *undersampling* y [**Abajo, derecha**] utilizando *oversampling*.

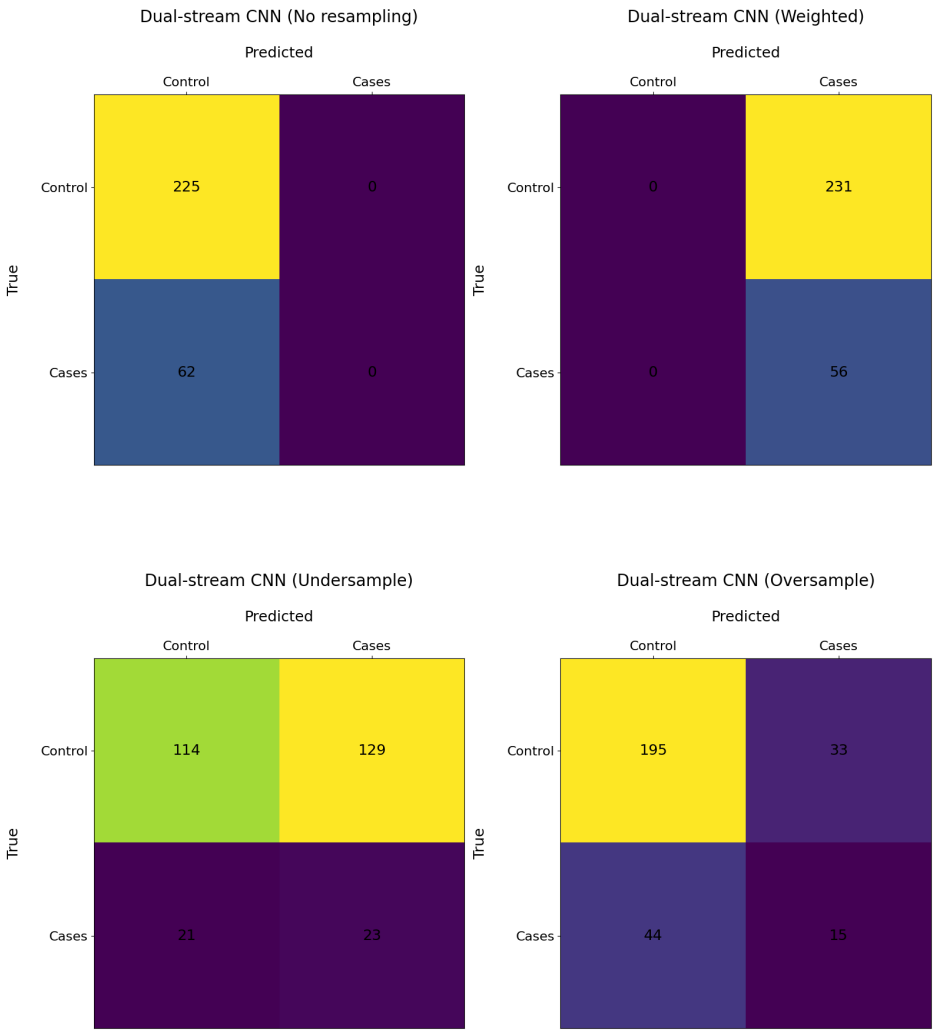


Figura C.47: Matrices de confusión del modelo Dual-stream CNN sobre las variantes genéticas seleccionadas para el análisis de hospitalizados entre la población infectada confirmada. [**Arriba, izquierda**] Se muestran los resultados para ningún tipo de resampleo,[**Arriba, derecha**] los resultados añadiendo un ponderador a la función de *loss*, [**Abajo, izquierda**] utilizando *undersampling* y [**Abajo, derecha**] utilizando *oversampling*.

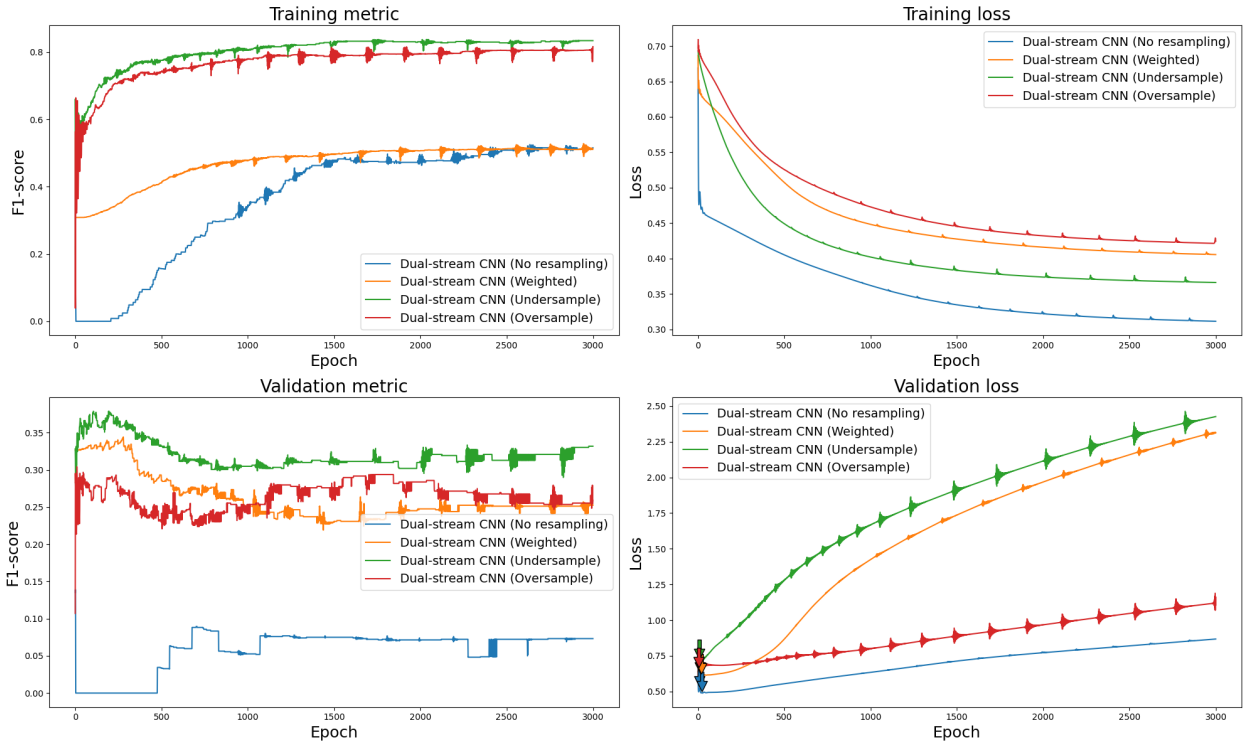


Figura C.48: Curva de entrenamiento de la arquitectura Dual-stream CNN sobre variantes genéticas seleccionadas para el análisis de hospitalizados entre la población infectada para cada aproximación por desbalance. A la esquina superior izquierda se muestra la métrica de entrenamiento, F1-score, a la esquina inferior izquierda la métrica de validación, a la esquina superior derecha la función de pérdida de entrenamiento, y a la esquina inferior derecha la función de pérdida de validación. En esta última se muestran flechas que indican que el modelo se guarda por *Early Stopping* al estancarse o disminuir la función de pérdida.

C.3.2.3. Métricas

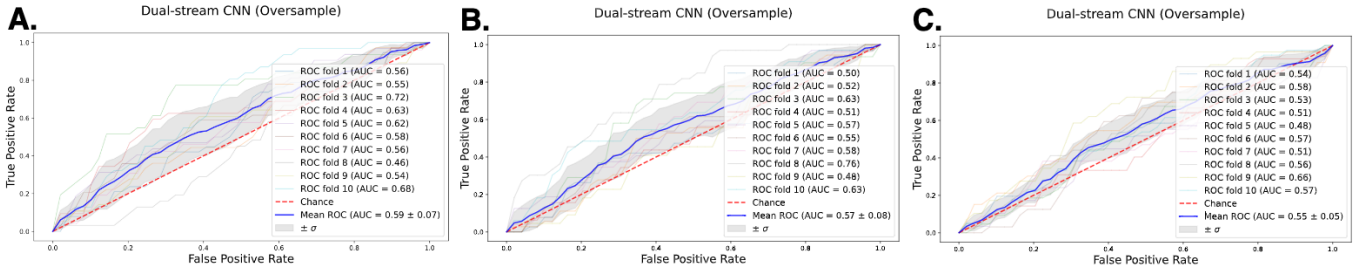


Figura C.49: Curva de ROC del modelo Dual-stream CNN sobre variantes genéticas para el análisis de hospitalizados entre la población infectada confirmada. **A.** Utilizando las variantes que originalmente se obtienen de la genotipificación de las muestras. **B.** Utilizando las variantes imputadas en el cromosoma 3. **C.** Variantes genéticas seleccionadas por la significancia reportada desde la iniciativa internacional Covid19hg.

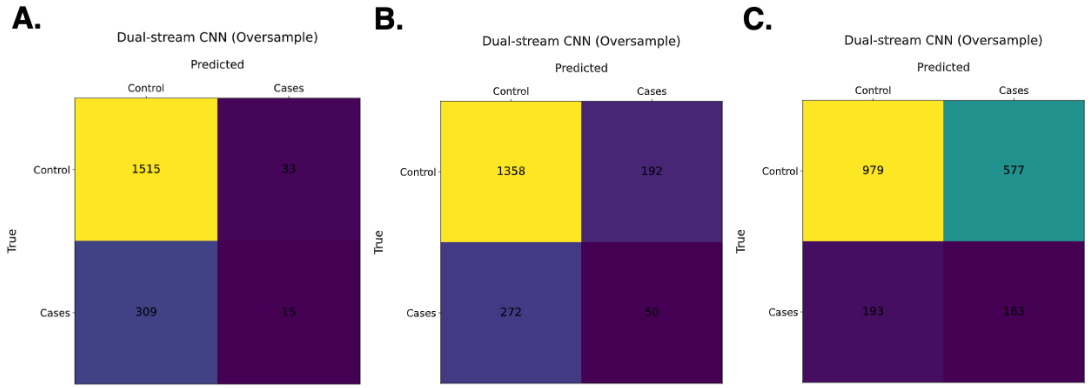


Figura C.50: Matrices de confusión para el modelos de redes neuronales Dual-stream CNN sobre variantes genéticas para el análisis de hospitalizados entre la población infectada confirmada. **A.** Utilizando las variantes que originalmente se obtienen de la genotipificación de las muestras. **B.** Utilizando las variantes imputadas en el cromosoma 3. **C.** Variantes genéticas seleccionadas por la significancia reportada desde la iniciativa internacional Covid19hg.

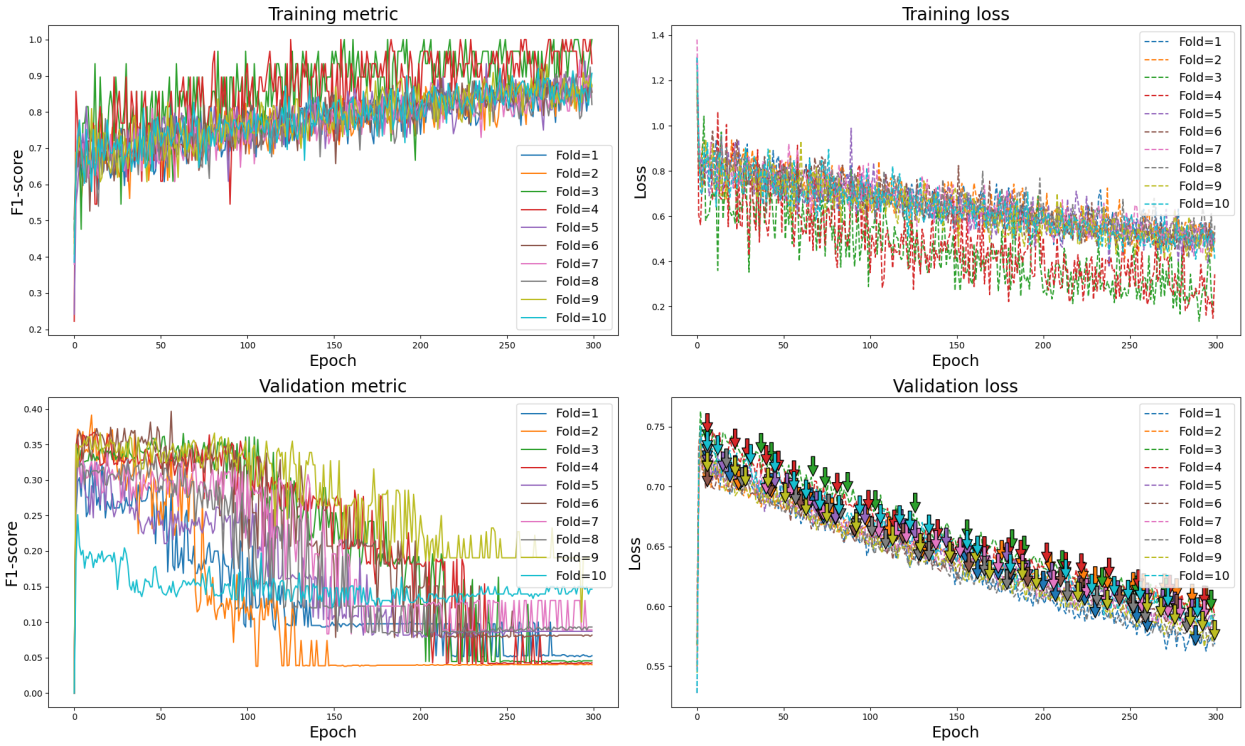


Figura C.51: Curva de entrenamiento de la arquitectura Dual-stream CNN sobre las variantes genéticas obtenidas originalmente por la genotipificación para el análisis de hospitalizados entre la población infectada. A la esquina superior izquierda se muestra la métrica de entrenamiento, F1-score, a la esquina inferior izquierda la métrica de validación, a la esquina superior derecha la función de pérdida de entrenamiento, y a la esquina inferior derecha la función de pérdida de validación. En esta última se muestran flechas que indican que el modelo se guarda por *Early Stopping* al estancarse o disminuir la función de pérdida.

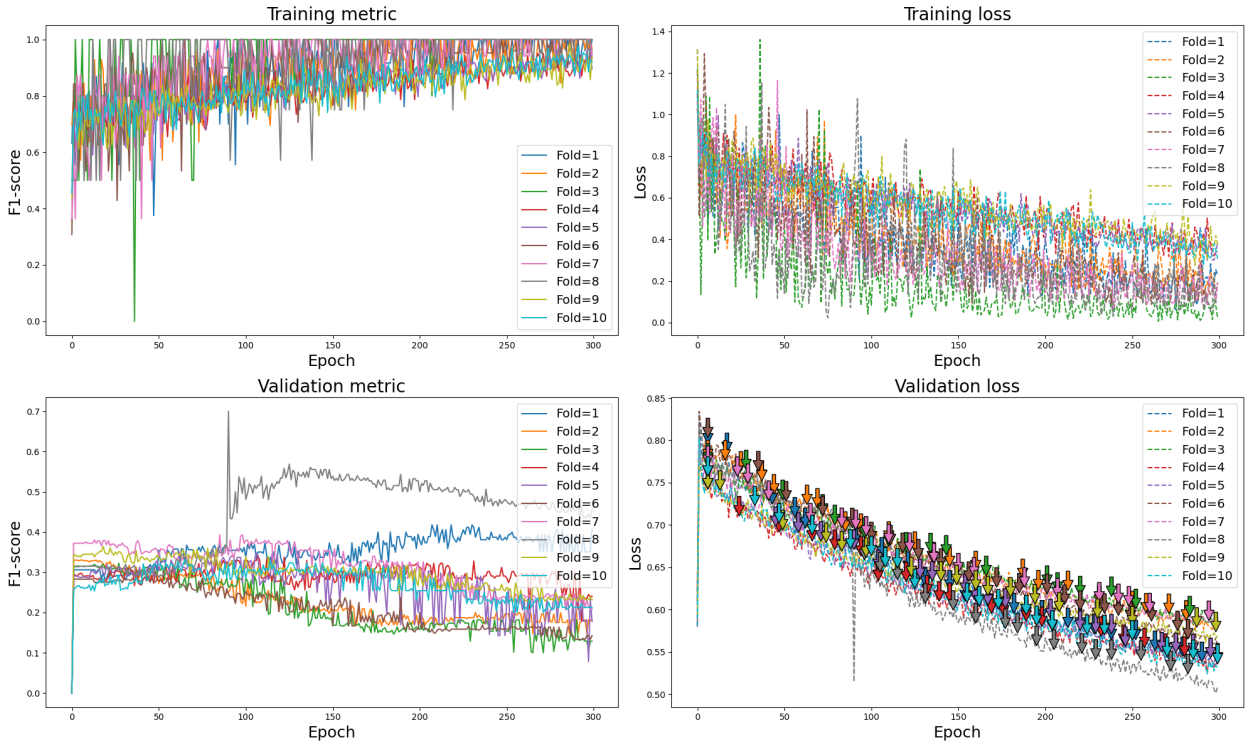


Figura C.52: Curva de entrenamiento de la arquitectura Dual-stream CNN sobre variantes genéticas imputadas solo en el cromosoma 3 para el análisis de hospitalizados entre la población infectada. A la esquina superior izquierda se muestra la métrica de entrenamiento, F1-score, a la esquina inferior izquierda la métrica de validación, a la esquina superior derecha la función de pérdida de entrenamiento, y a la esquina inferior derecha la función de pérdida de validación. En esta última se muestran flechas que indican que el modelo se guarda por *Early Stopping* al estancarse o disminuir la función de pérdida.

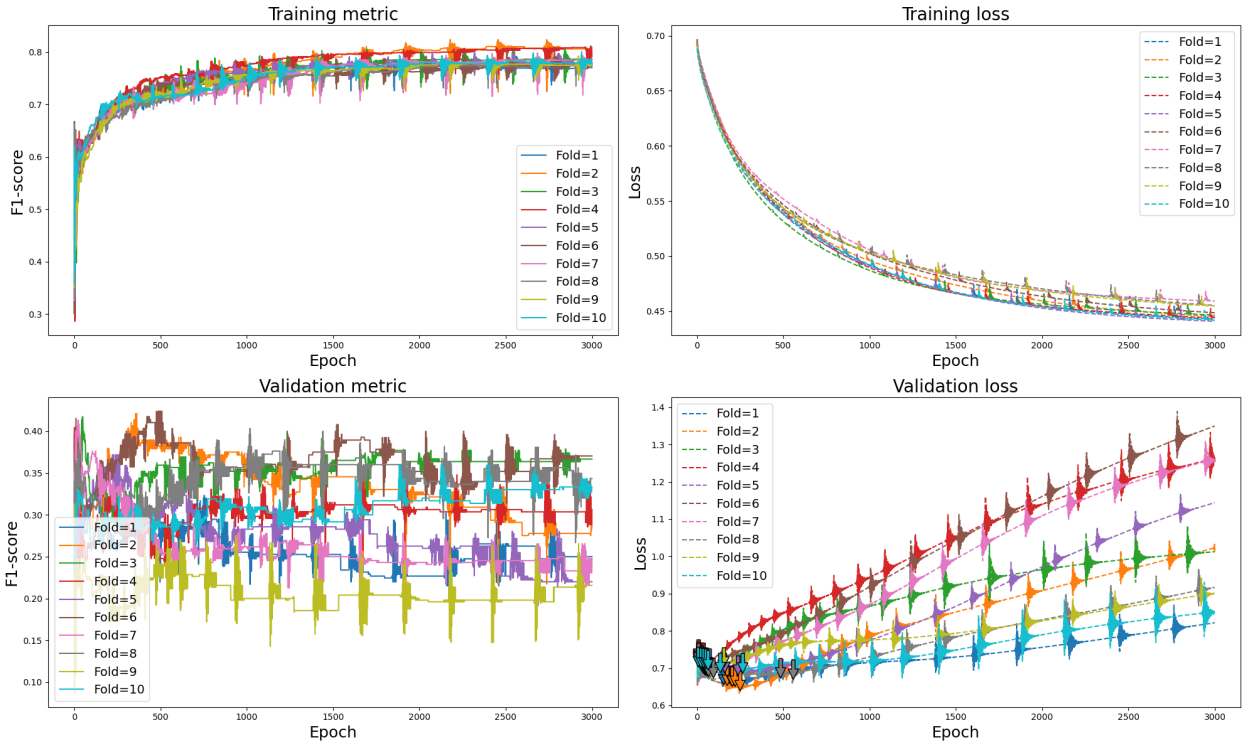


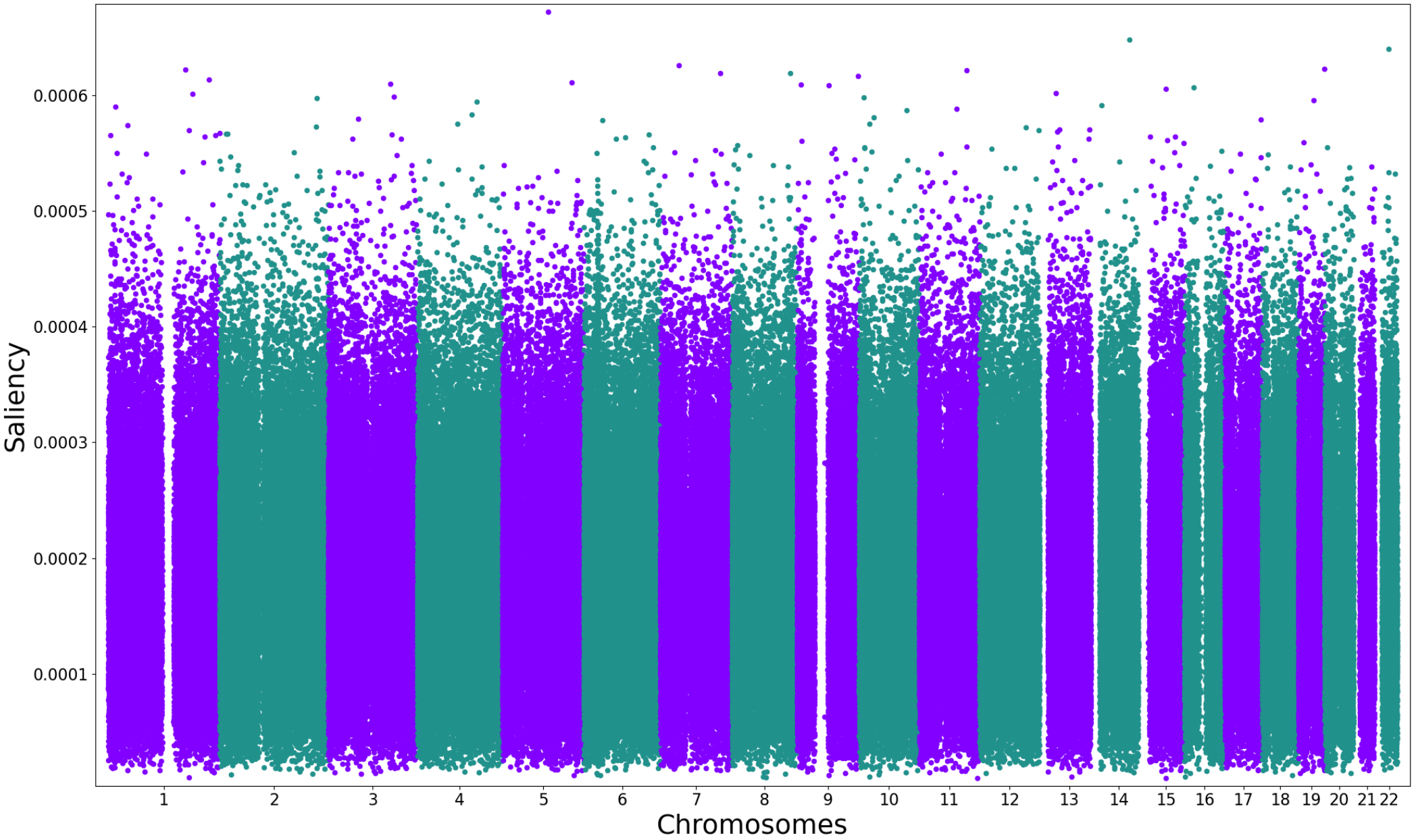
Figura C.53: Curva de entrenamiento de la arquitectura Dual-stream CNN sobre variantes genéticas seleccionadas para el análisis de hospitalizados entre la población infectada. A la esquina superior izquierda se muestra la métrica de entrenamiento, F1-score, a la esquina inferior izquierda la métrica de validación, a la esquina superior derecha la función de pérdida de entrenamiento, y a la esquina inferior derecha la función de pérdida de validación. En esta última se muestran flechas que indican que el modelo se guarda por *Early Stopping* al estancarse o disminuir la función de pérdida.

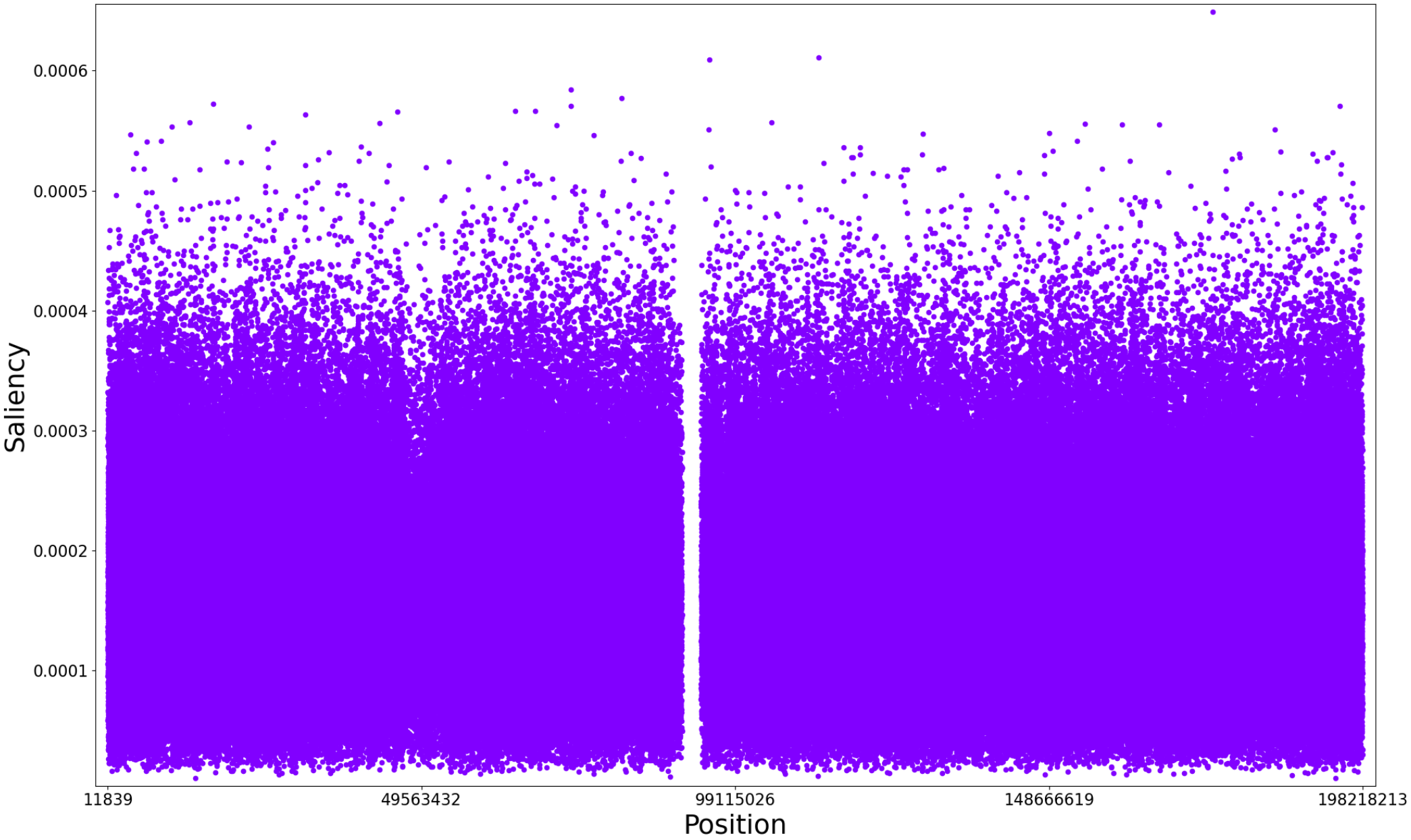
C.3.2.4. Saliencia

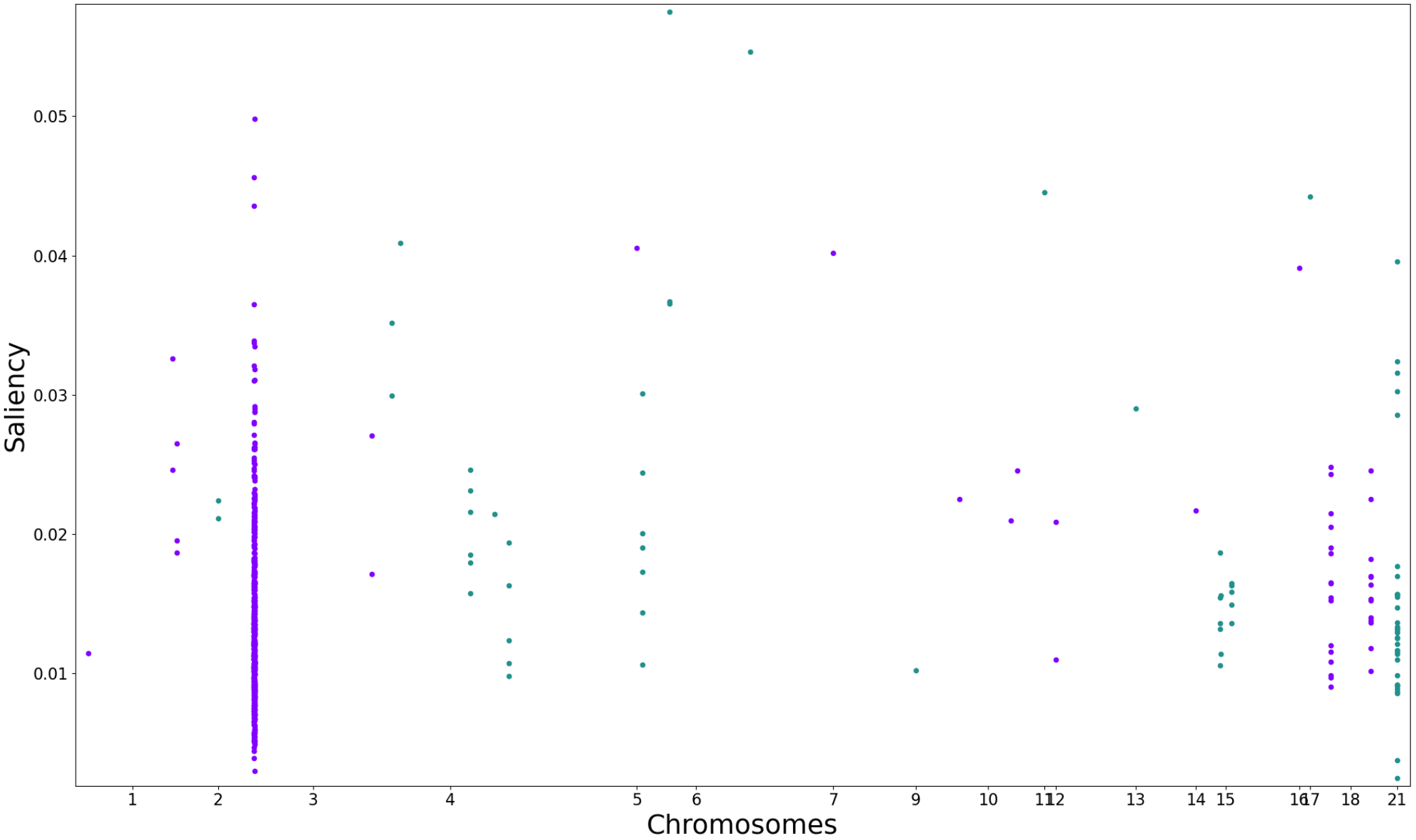
Figura C.54: [Pagina 159] Saliencia obtenida por la red neuronal Dual-stream CNN entrenada con las variantes genéticas obtenidas por el microarreglo ($\sim 4 \cdot 10^5$). El modelo se entrena para predecir hospitalización en participantes contagiados por SARS-Cov-2.

Figura C.55: [Pagina 160] Saliencia obtenida por la red neuronal Dual-stream CNN entrenada con las variantes genéticas imputadas solo del cromosoma 3 ($\sim 5 \cdot 10^5$). El modelo se entrena para predecir hospitalización en participantes contagiados por SARS-Cov-2.

Figura C.56: [Pagina 161] Saliencia obtenida por la red neuronal Dual-stream CNN entrenada con las variantes genéticas obtenidas por el microarreglo ($\sim 6 \cdot 10^2$). El modelo se entrena para predecir hospitalización en participantes contagiados por SARS-Cov-2.


Saliency
0.0006
0.0005
0.0004
0.0003
0.0002
0.0001
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
Chromosomes


0.0006
0.0005
0.0004
0.0003
0.0002
0.0001
Saliency
11839
49563432
99115026
148666619
198218213
Position


Saliency
0.05
0.04
0.03
0.02
0.01
1
2
3
4
5
6
7
9
10
11
12
13
14
15
16
17
18
21
Chromosomes

C.4. Arquitectura Dual-stream CNN Extendida sobre Datos clínicos y genéticos

C.4.1. Selección de Hiperparámetros

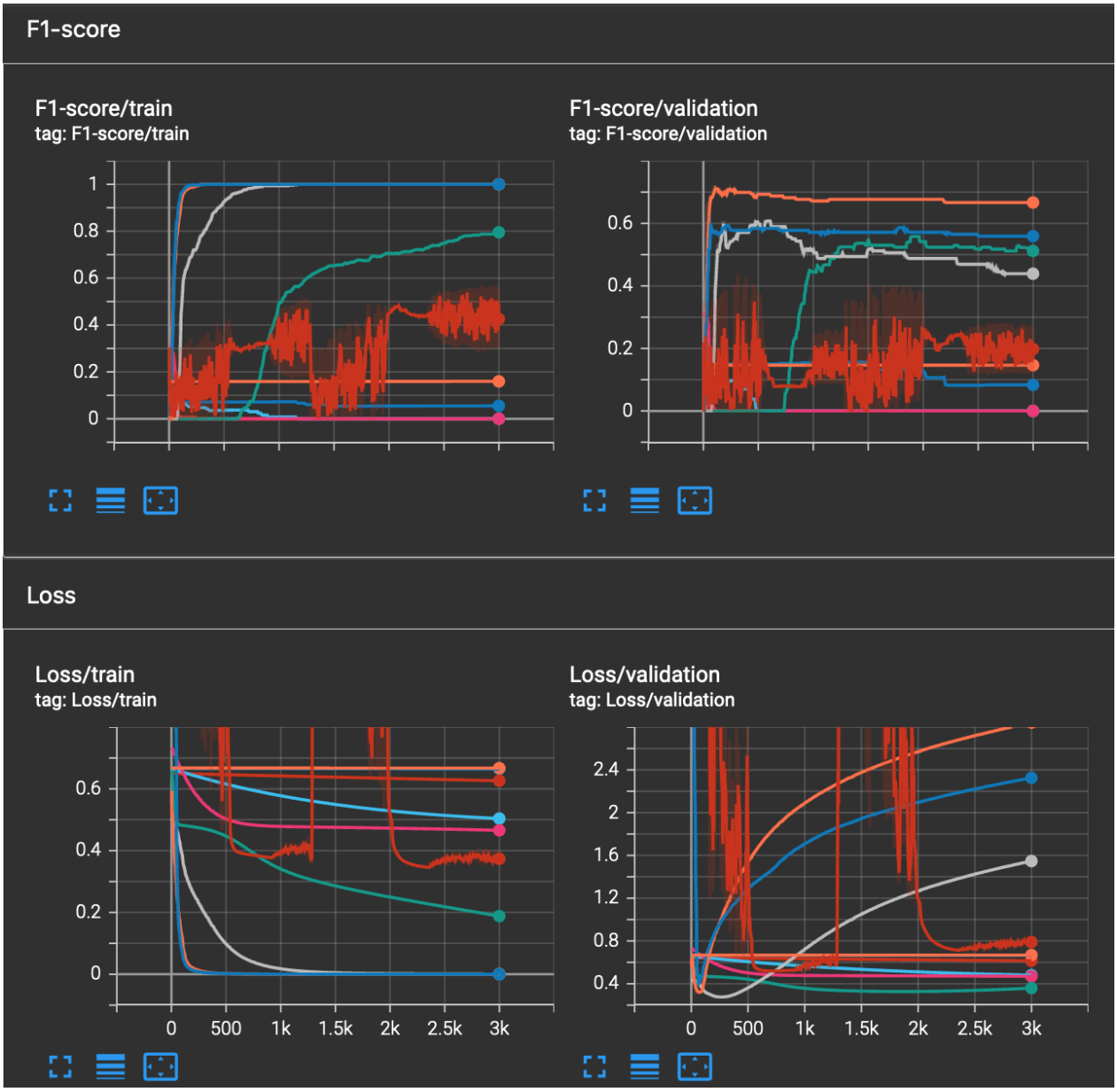


Figura C.57: Experimentos a diferentes *learning rates* con la arquitectura Dual-stream CNN para el análisis de hospitalizados utilizando SNPs seleccionados según COVID19hgi y variables clínicas seleccionadas. Se realizan con *weight decay* $= wd = 0.0$, la visualización se obtuvo utilizando Tensorboard [22]. Se grafica la métrica $f1$ en la primera fila y la función de perdida (*Cross entropy*) en la segunda. En la primera columna se gráfica los resultados durante el entrenamiento, en la segunda, los resultados sobre el *set* de validación. El experimento se realiza sobre *learning rates* de 10^i con $i \in \{-10, \cdot, -1\}$.

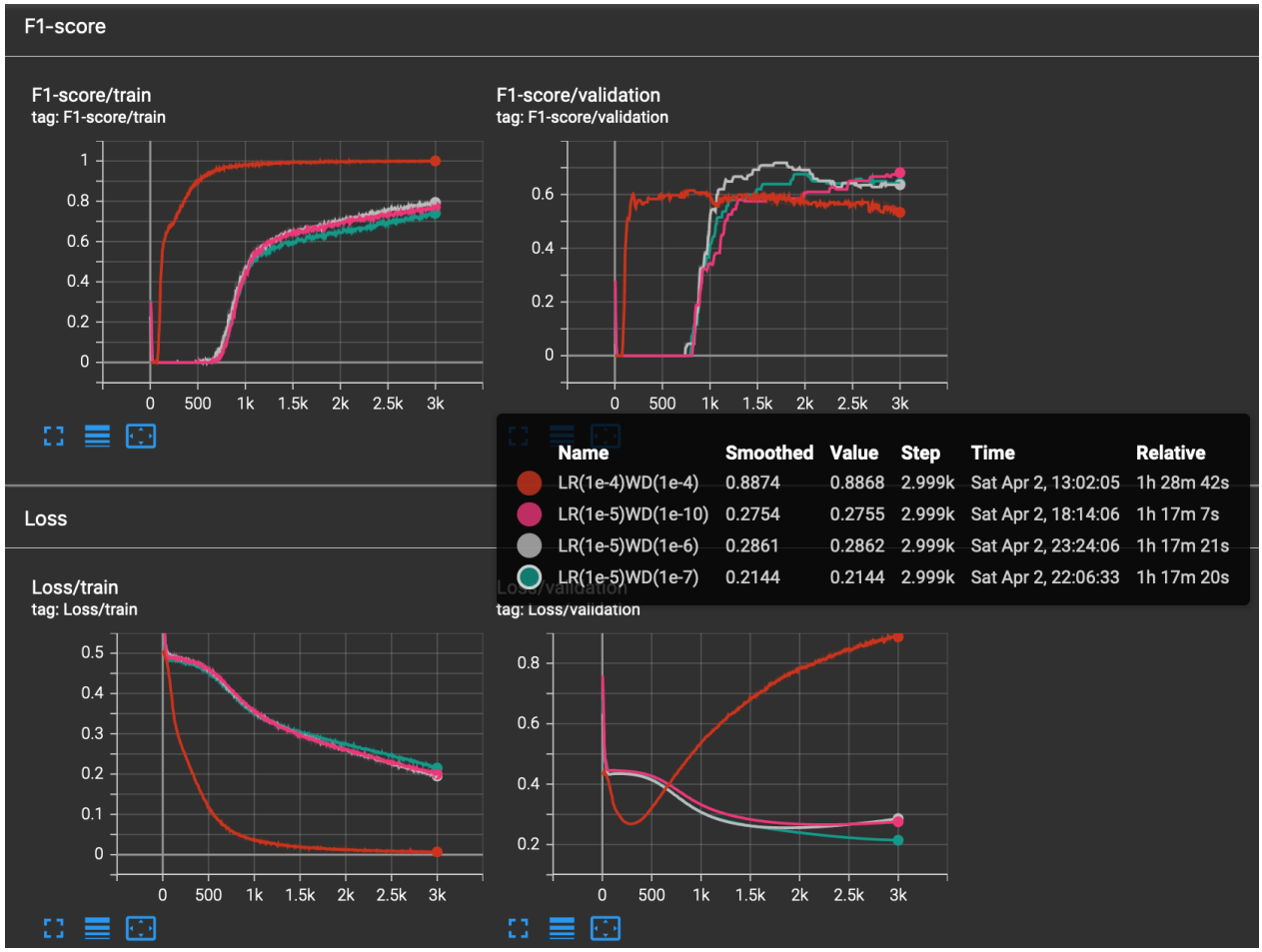


Figura C.58: Experimentos a diferentes *weight decay* con la arquitectura Dual-strem CNN para el análisis de hospitalizados utilizando SNPs seleccionados según COVID19hgi y variables clínicas seleccionadas. Se realizan con *learning rate* $= lr \in \{10^{-7}, 10^{-6}, 10^{-5}, 10^{-4}\}$ y $wd = 10^i$ con $i \in \{-10, \cdot, -1\}$ para cada lr. La visualización se obtiene utilizando Tensorboard [22]. Se grafica la métrica $f1$ en la primera fila y la función de perdida (*Cross entropy*) en la segunda. En la primera columna se gráfica los resultados durante el entrenamiento, en la segunda, los resultados sobre el *set* de validación. Se muestran las mejores curvas obtenidas. El código color se muestra en el gráfico de función de pérdida para el set de validación.

C.4.2. Experimentos con desbalance

C.4.2.1. SNPs Genotipificados por Microarreglo

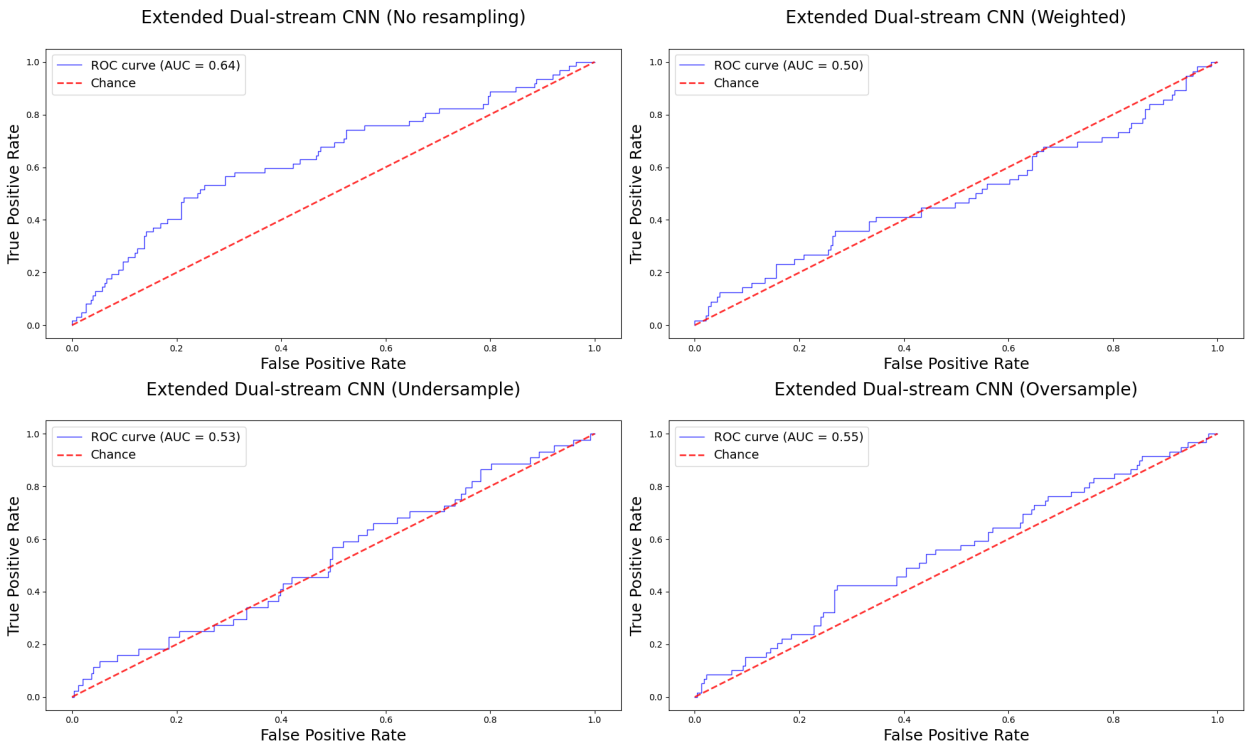


Figura C.59: Curva de ROC del modelo Dual-stream CNN sobre las variantes genéticas originalmente genotipificadas y variables clínicas seleccionadas, para el análisis de hospitalizados entre la población infectada confirmada. [**Arriba, izquierda**] Se muestran los resultados para ningún tipo de resampleo,[**Arriba, derecha**] los resultados añadiendo un ponderador a la función de *loss*, [**Abajo, izquierda**] utilizando *undersampling* y [**Abajo, derecha**] utilizando *oversampling*.

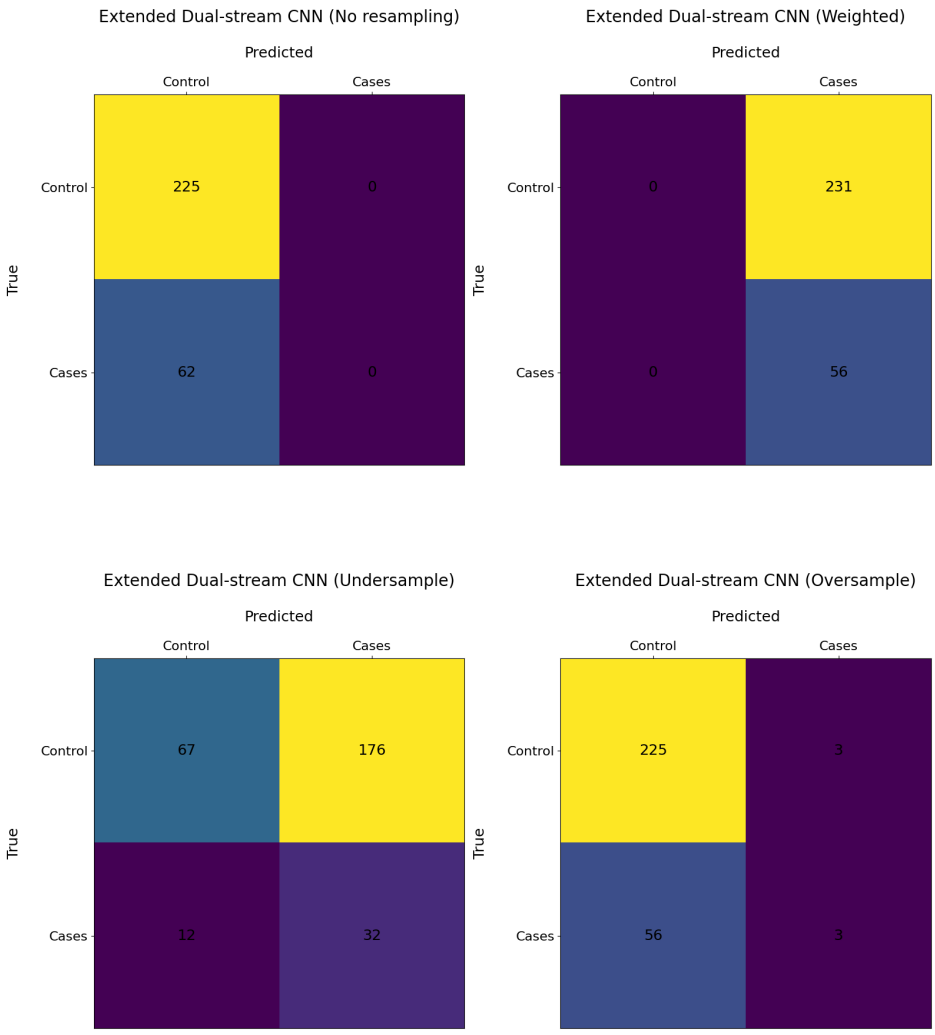


Figura C.60: Matrices de confusión del modelo Dual-stream CNN sobre las variantes genéticas originalmente genotipificadas y variables clínicas seleccionadas, para el análisis de hospitalizados entre la población infectada confirmada. [**Arriba, izquierda**] Se muestran los resultados para ningún tipo de resampleo,[**Arriba, derecha**] los resultados añadiendo un ponderador a la función de *loss*, [**Abajo, izquierda**] utilizando *undersampling* y [**Abajo, derecha**] utilizando *oversampling*.

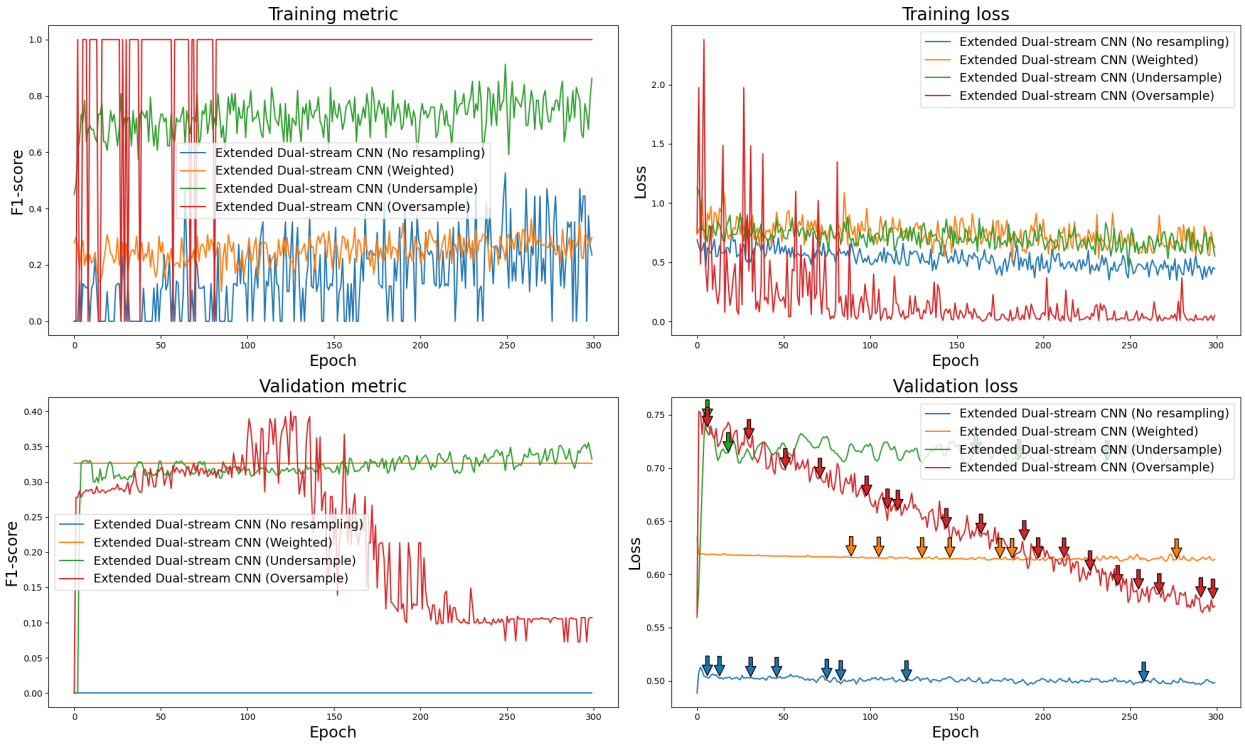


Figura C.61: Curva de entrenamiento de la arquitectura Dual-stream CNN sobre variantes genéticas originalmente genotipificadas y variables clínicas seleccionadas, para el análisis de hospitalizados entre la población infectada para cada aproximación por desbalance. A la esquina superior izquierda se muestra la métrica de entrenamiento, F1-score, a la esquina inferior izquierda la métrica de validación, a la esquina superior derecha la función de pérdida de entrenamiento, y a la esquina inferior derecha la función de pérdida de validación. En esta última se muestran flechas que indican que el modelo se guarda por *Early Stopping* al estancarse o disminuir la función de pérdida.

C.4.2.2. Cromosoma 3

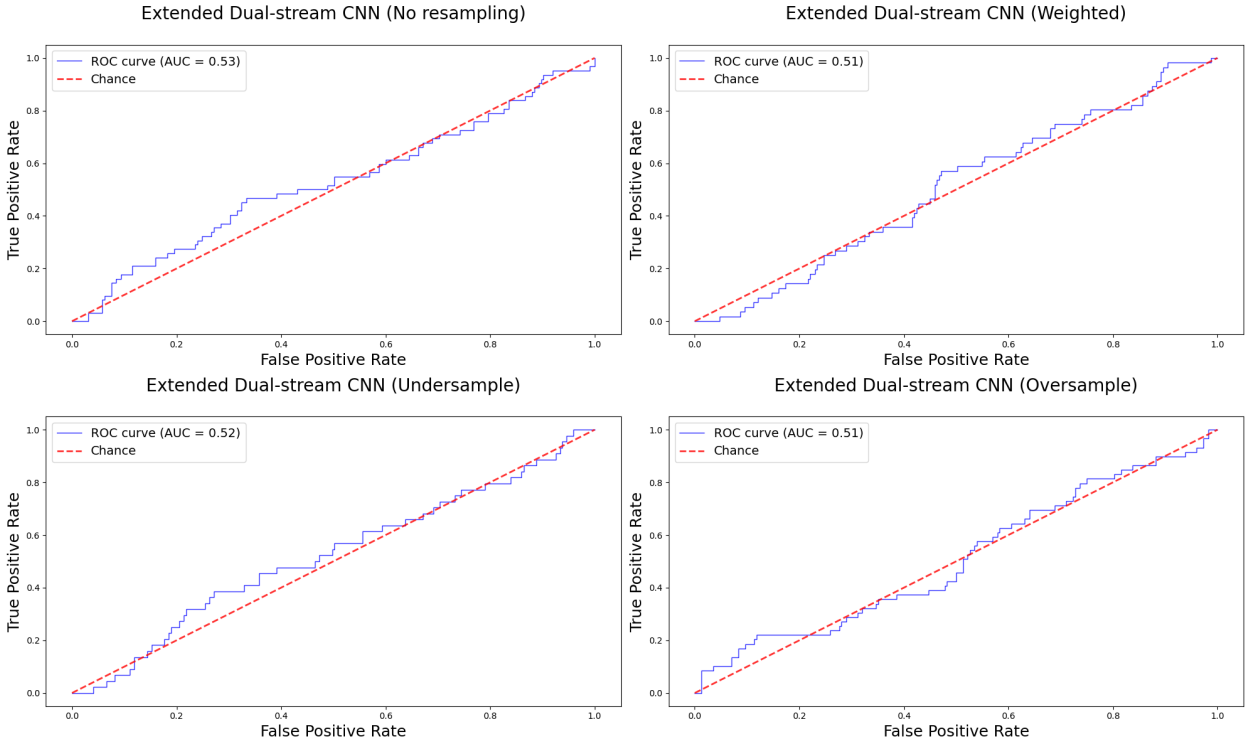


Figura C.62: Curva de ROC del modelo Dual-stream CNN sobre las variantes genéticas imputadas solo en el cromosoma 3 y datos clínicos seleccionados, para el análisis de hospitalizados entre la población infectada confirmada. [**Arriba, izquierda**] Se muestran los resultados para ningún tipo de resampleo,[**Arriba, derecha**] los resultados añadiendo un ponderador a la función de *loss*, [**Abajo, izquierda**] utilizando *undersampling* y [**Abajo, derecha**] utilizando *oversampling*.

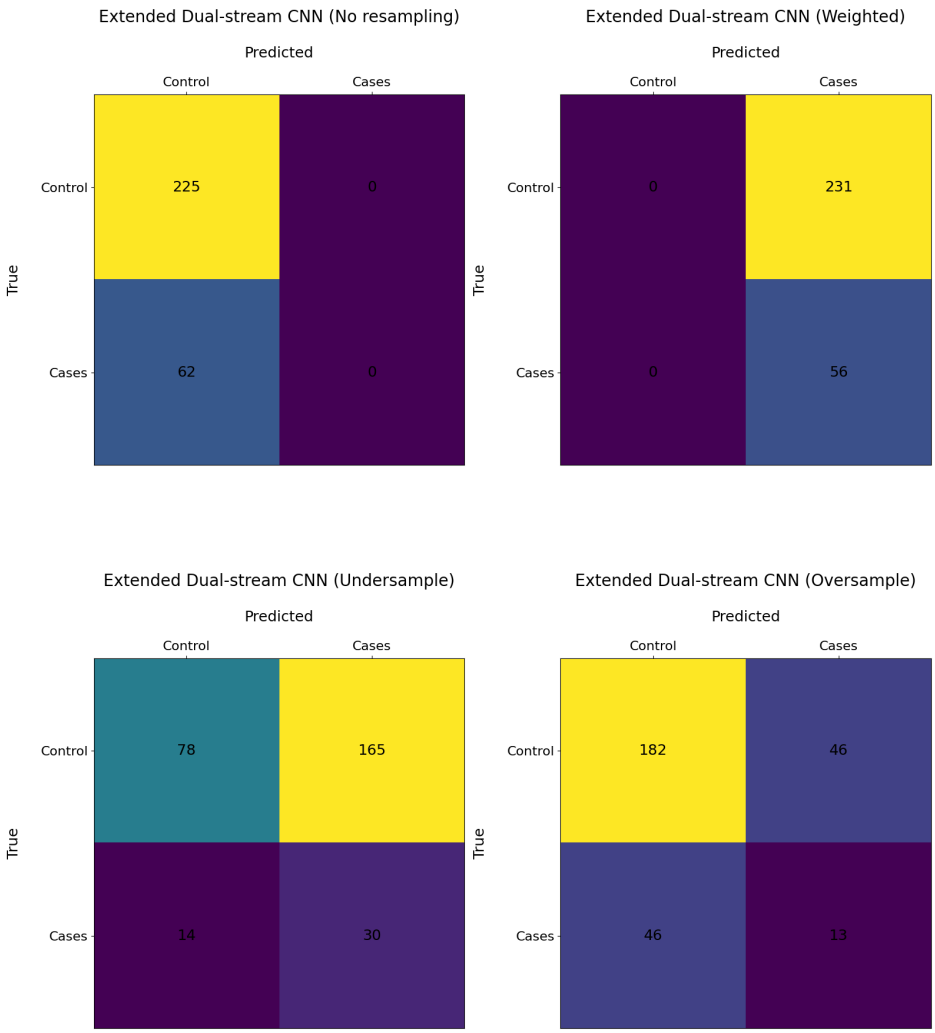


Figura C.63: Matrices de confusión del modelo Dual-stream CNN sobre las variantes genéticas imputadas solo en el cromosoma 3 y datos clínicos seleccionados, para el análisis de hospitalizados entre la población infectada confirmada. [**Arriba, izquierda**] Se muestran los resultados para ningún tipo de resampleo,[**Arriba, derecha**] los resultados añadiendo un ponderador a la función de *loss*, [**Abajo, izquierda**] utilizando *undersampling* y [**Abajo, derecha**] utilizando *oversampling*.

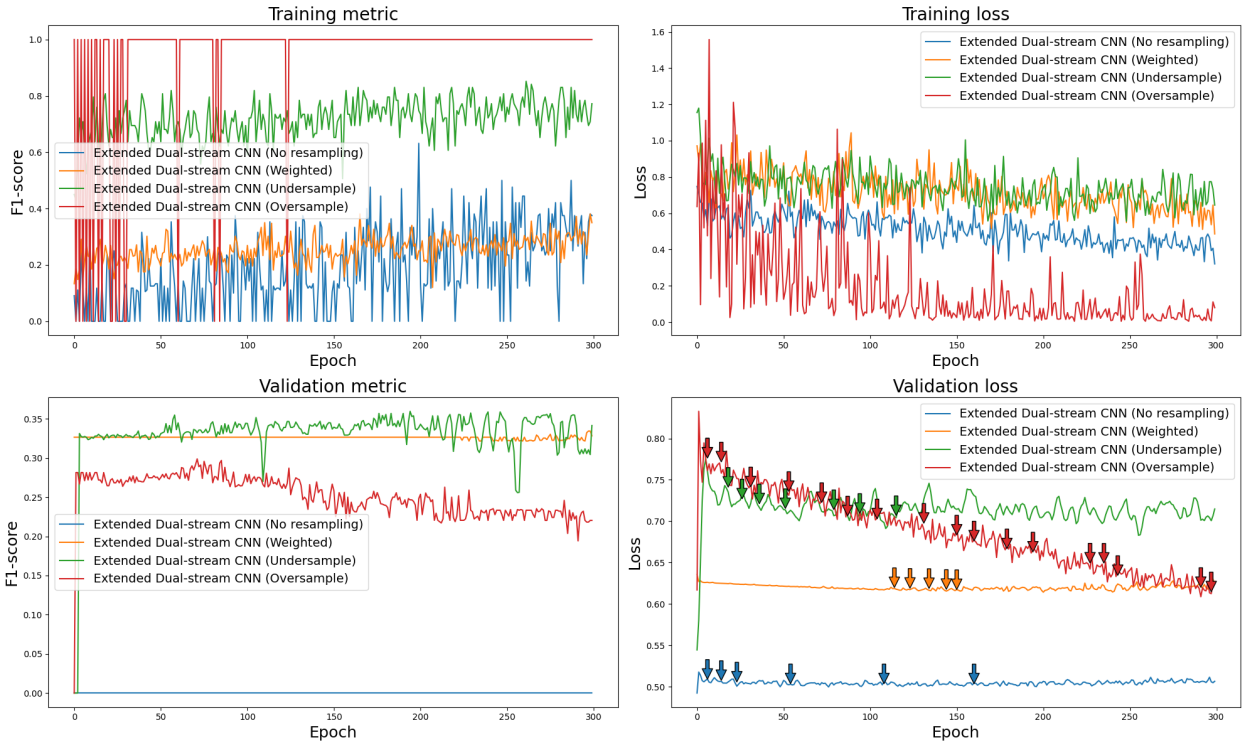


Figura C.64: Curva de entrenamiento de la arquitectura Dual-stream CNN sobre variantes genéticas imputadas solo en el cromosoma 3 y datos clínicos seleccionados, para el análisis de hospitalizados entre la población infectada para cada aproximación por desbalance. A la esquina superior izquierda se muestra la métrica de entrenamiento, F1-score, a la esquina inferior izquierda la métrica de validación, a la esquina superior derecha la función de pérdida de entrenamiento, y a la esquina inferior derecha la función de pérdida de validación. En esta última se muestran flechas que indican que el modelo se guarda por *Early Stopping* al estancarse o disminuir la función de pérdida.

C.4.2.3. SNPs Seleccionados

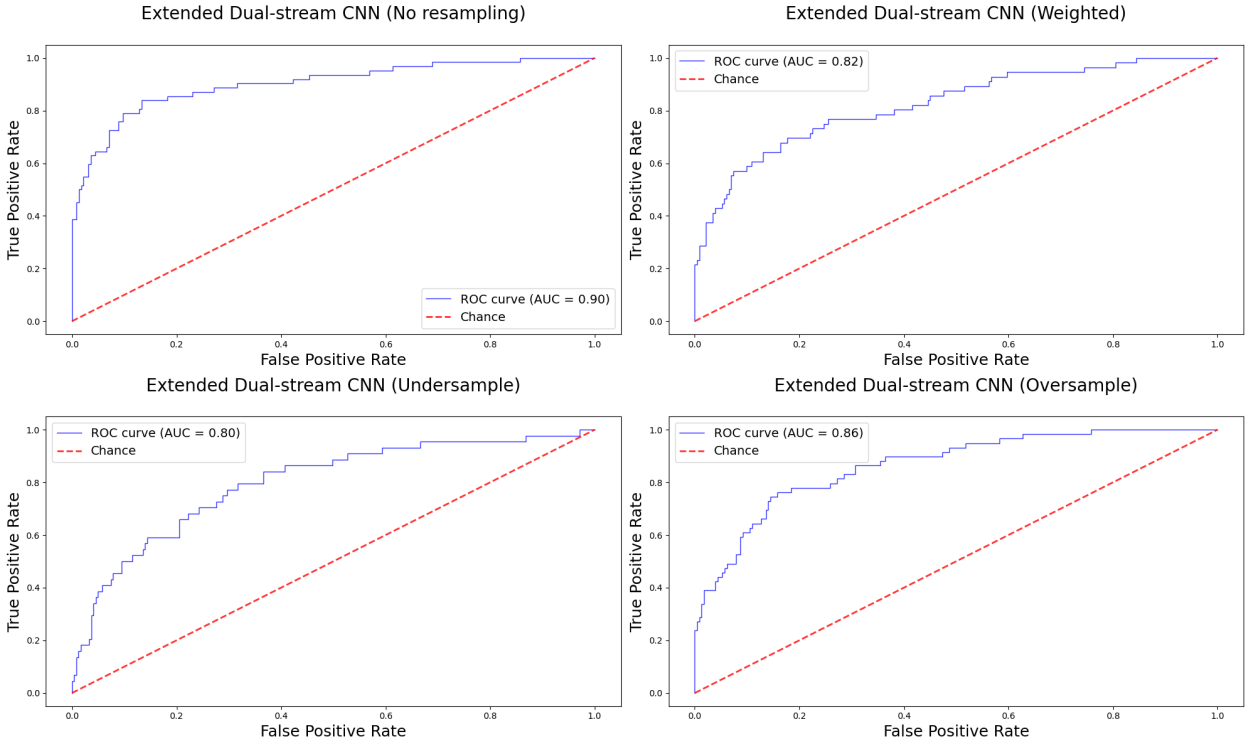


Figura C.65: Curva de ROC del modelo Dual-stream CNN sobre las variantes genéticas seleccionadas y datos clínicos seleccionados, para el análisis de hospitalizados entre la población infectada confirmada. [**Arriba, izquierda**] Se muestran los resultados para ningún tipo de resampleo,[**Arriba, derecha**] los resultados añadiendo un ponderador a la función de *loss*, [**Abajo, izquierda**] utilizando *undersampling* y [**Abajo, derecha**] utilizando *oversampling*.

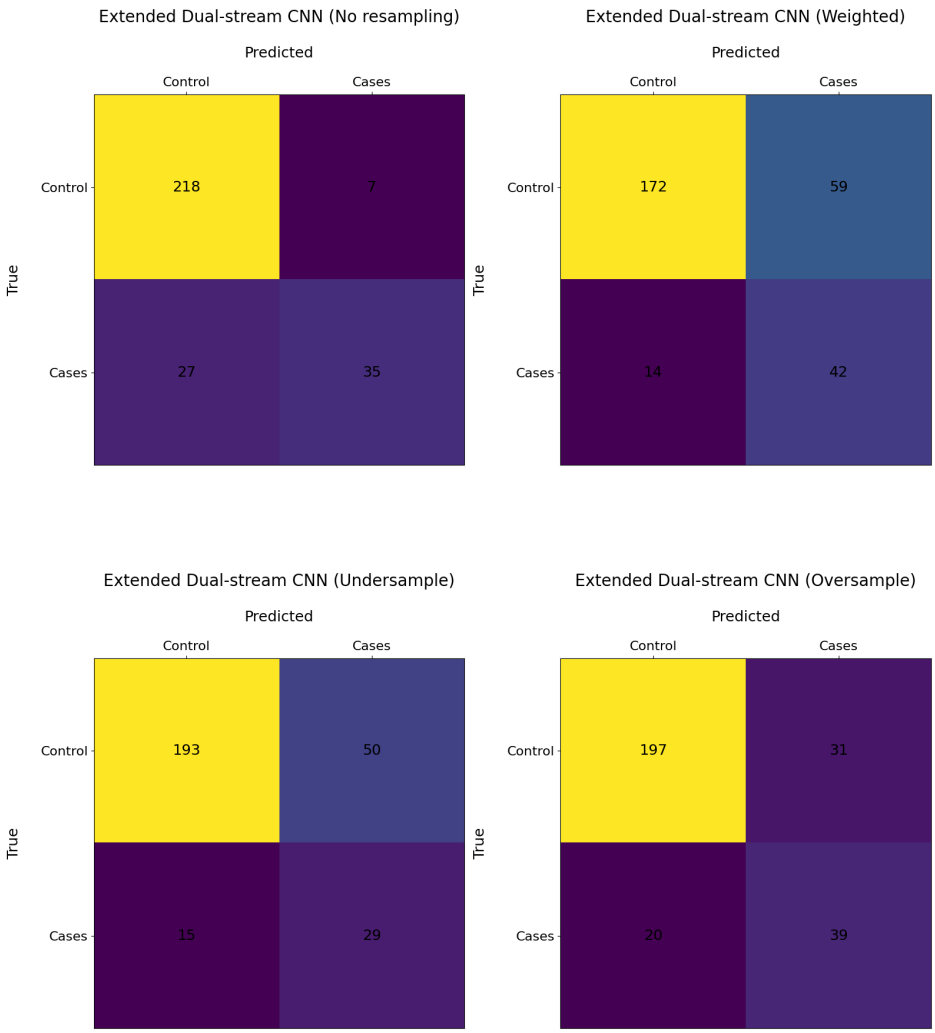


Figura C.66: Matrices de confusión del modelo Dual-stream CNN sobre las variantes genéticas seleccionadas y datos clínicos seleccionados, para el análisis de hospitalizados entre la población infectada confirmada. [**Arriba, izquierda**] Se muestran los resultados para ningún tipo de resampleo,[**Arriba, derecha**] los resultados añadiendo un ponderador a la función de *loss*, [**Abajo, izquierda**] utilizando *undersampling* y [**Abajo, derecha**] utilizando *oversampling*.

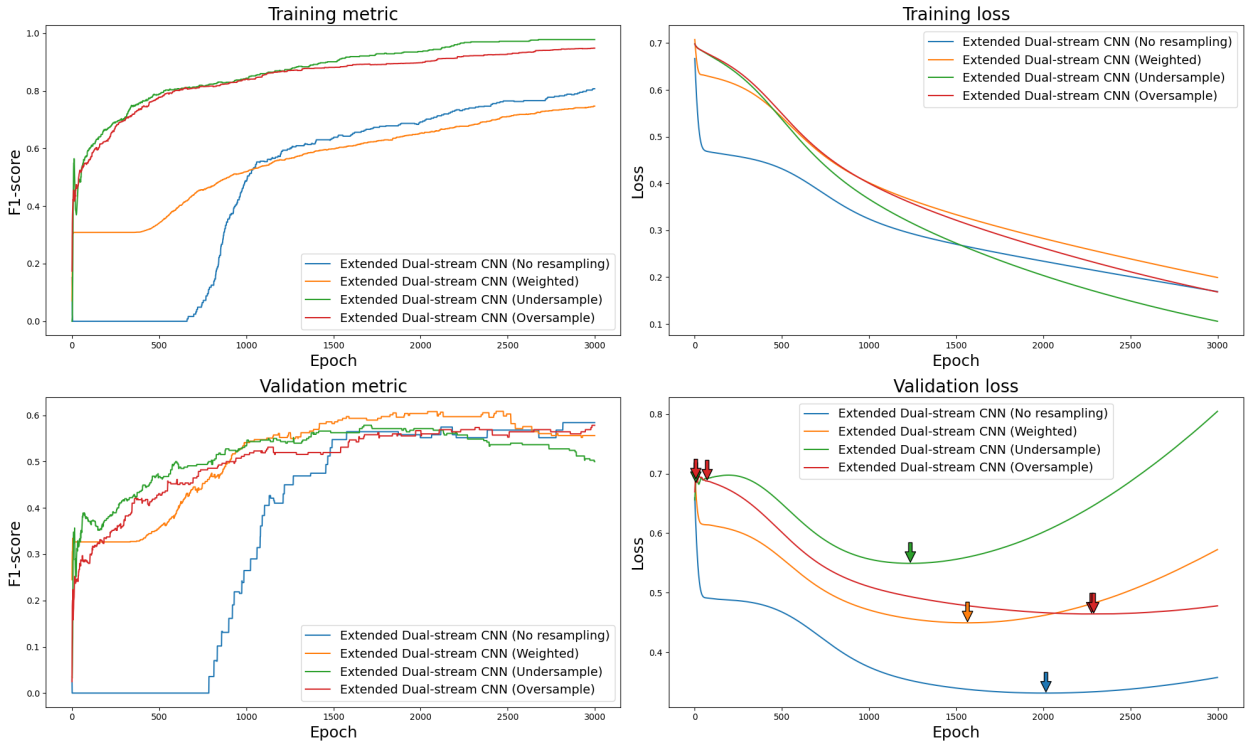


Figura C.67: Curva de entrenamiento de la arquitectura Dual-stream CNN sobre variantes genéticas seleccionadas y datos clínicos seleccionados, para el análisis de hospitalizados entre la población infectada para cada aproximación por desbalance. A la esquina superior izquierda se muestra la métrica de entrenamiento, F1-score, a la esquina inferior izquierda la métrica de validación, a la esquina superior derecha la función de pérdida de entrenamiento, y a la esquina inferior derecha la función de pérdida de validación. En esta última se muestran flechas que indican que el modelo se guarda por *Early Stopping* al estancarse o disminuir la función de pérdida.

C.4.3. Métricas

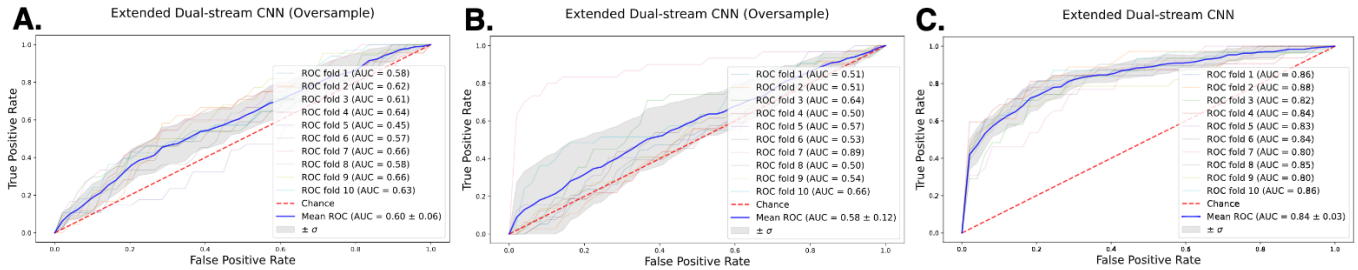


Figura C.68: Curva de ROC del modelo Dual-stream CNN sobre variantes genéticas y variables clínicas seleccionadas, parapara el análisis de hospitalizados entre la población infectada confirmada. **A.** Utilizando las variantes que originalmente se obtienen de la genotipificación de las muestras. **B.** Utilizando las variantes imputadas en el cromosoma 3. **C.** variantes genéticas seleccionadas y datos clínicos seleccionados, por la significancia reportada desde la iniciativa internacional Covid19hg.

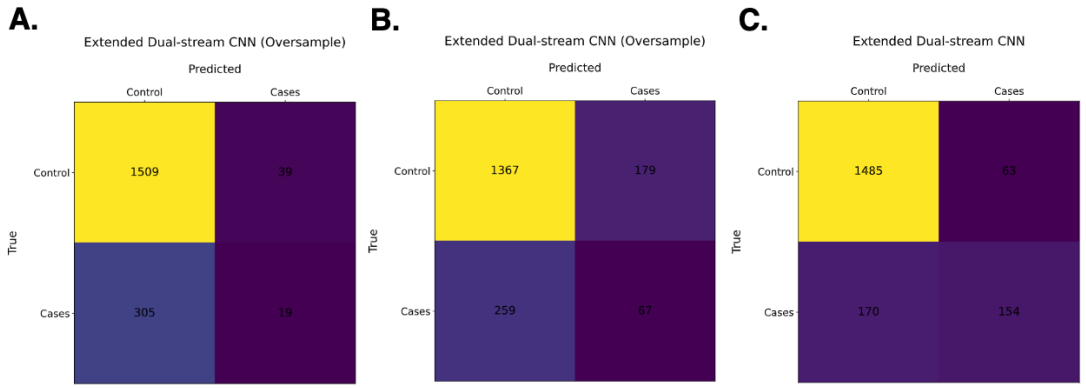


Figura C.69: Matrices de confusión para el modelos de redes neuronales Dual-stream CNN sobre variantes genéticas y variables clínicas seleccionadas, parapara el análisis de hospitalizados entre la población infectada confirmada. **A.** Utilizando las variantes que originalmente se obtienen de la genotipificación de las muestras. **B.** Utilizando las variantes imputadas en el cromosoma 3. **C.** variantes genéticas seleccionadas y datos clínicos seleccionados, por la significancia reportada desde la iniciativa internacional Covid19hg.

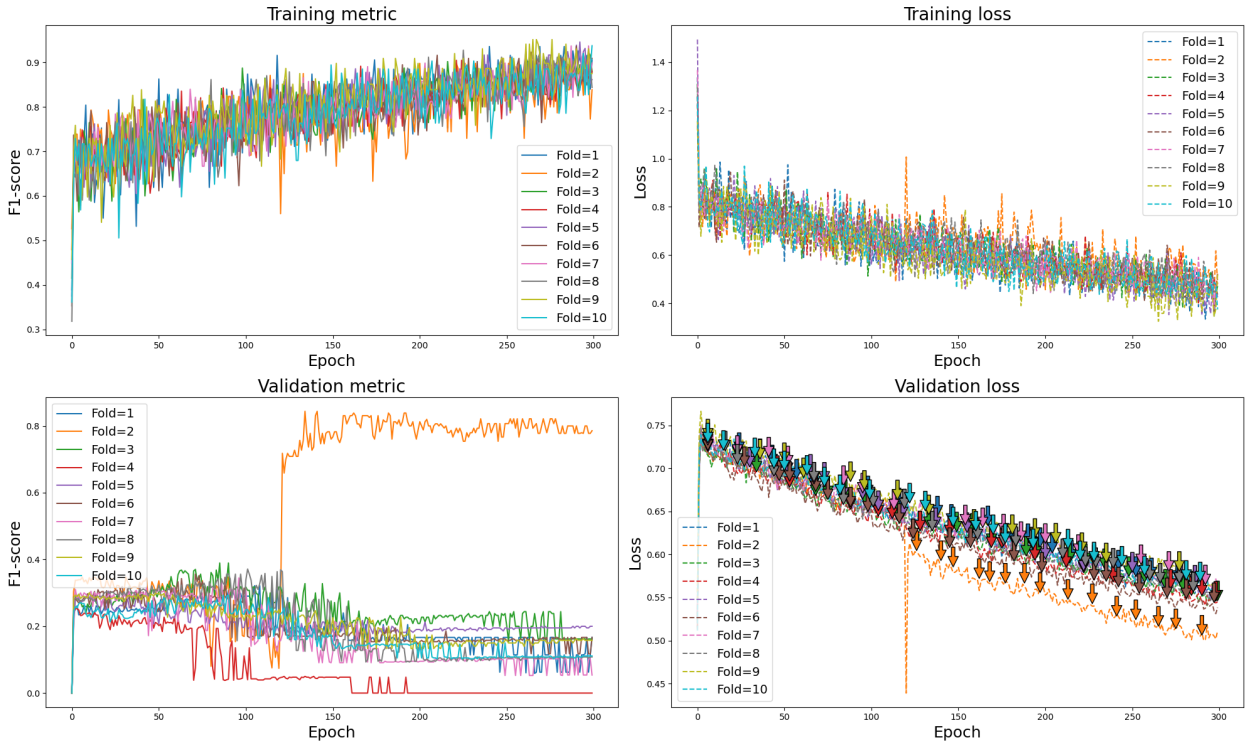


Figura C.70: Curva de entrenamiento de la arquitectura Dual-stream CNN sobre las variantes genéticas obtenidas originalmente por la genotipificación y variables clínicas seleccionadas, para el análisis de hospitalizados entre la población infectada. A la esquina superior izquierda se muestra la métrica de entrenamiento, F1-score, a la esquina inferior izquierda la métrica de validación, a la esquina superior derecha la función de pérdida de entrenamiento, y a la esquina inferior derecha la función de pérdida de validación. En esta última se muestran flechas que indican que el modelo se guarda por *Early Stopping* al estancarse o disminuir la función de pérdida.

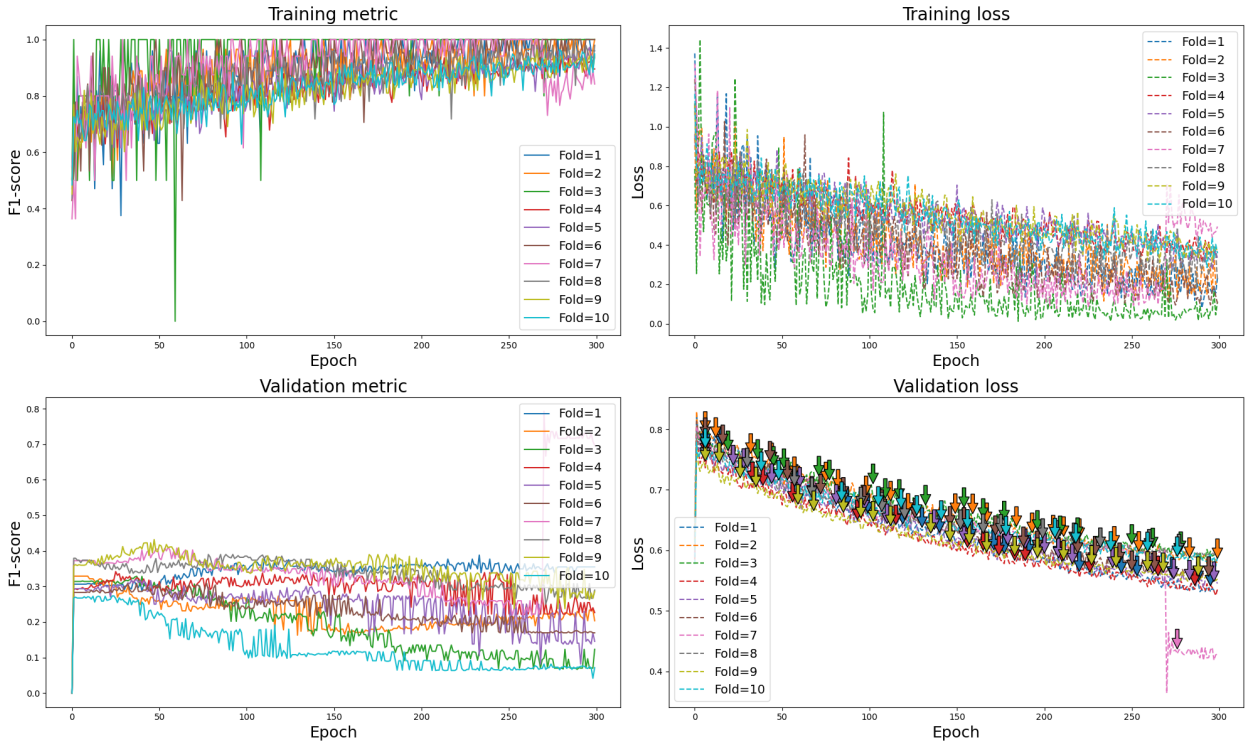


Figura C.71: Curva de entrenamiento de la arquitectura Dual-stream CNN sobre variantes genéticas imputadas solo en el cromosoma 3 y datos clínicos seleccionados, para el análisis de hospitalizados entre la población infectada. A la esquina superior izquierda se muestra la métrica de entrenamiento, F1-score, a la esquina inferior izquierda la métrica de validación, a la esquina superior derecha la función de pérdida de entrenamiento, y a la esquina inferior derecha la función de pérdida de validación. En esta última se muestran flechas que indican que el modelo se guarda por *Early Stopping* al estancarse o disminuir la función de pérdida.

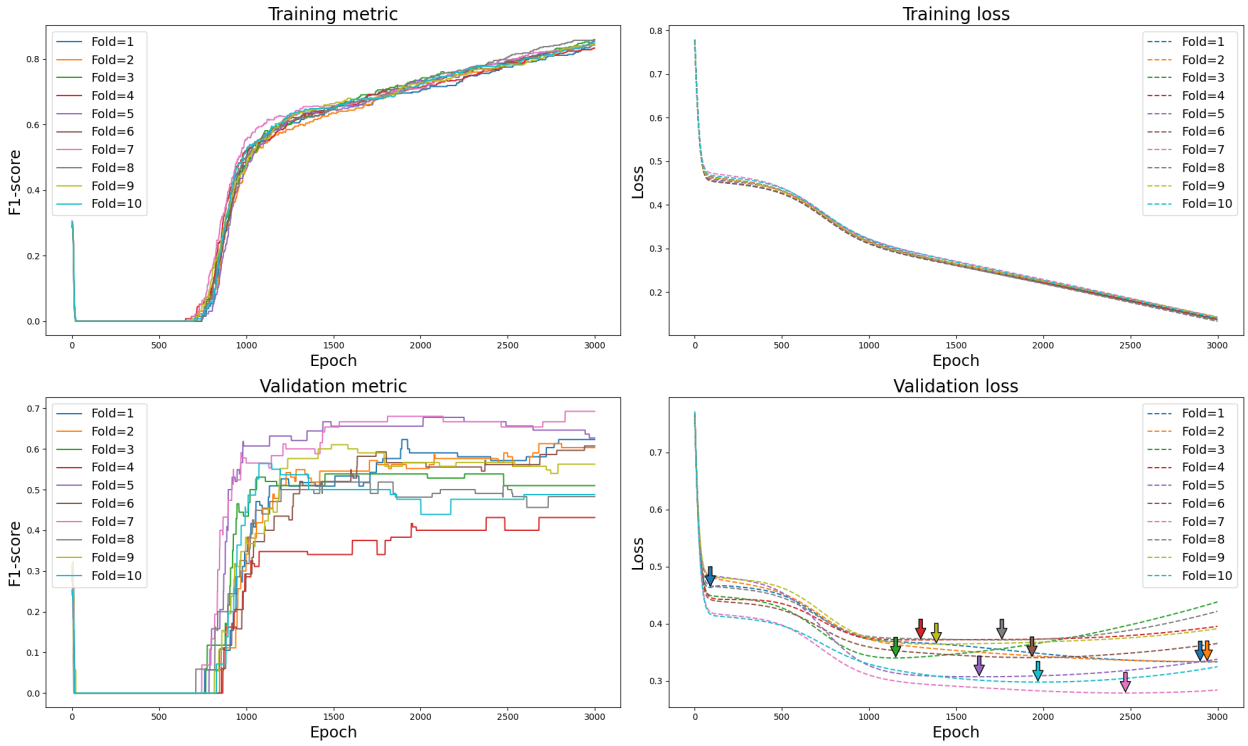


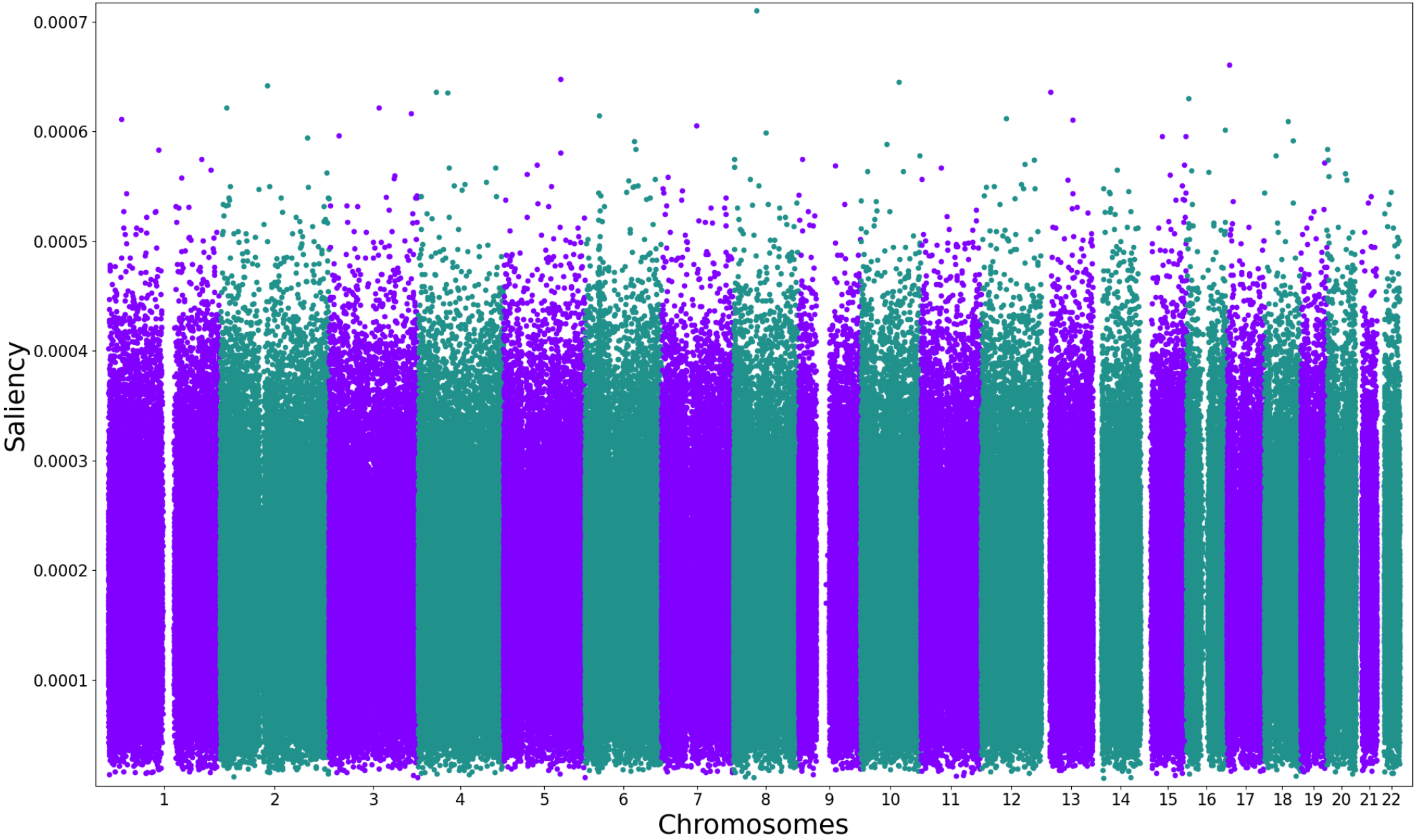
Figura C.72: Curva de entrenamiento de la arquitectura Dual-stream CNN sobre variantes genéticas seleccionadas y datos clínicos seleccionados, para el análisis de hospitalizados entre la población infectada. A la esquina superior izquierda se muestra la métrica de entrenamiento, F1-score, a la esquina inferior izquierda la métrica de validación, a la esquina superior derecha la función de pérdida de entrenamiento, y a la esquina inferior derecha la función de pérdida de validación. En esta última se muestran flechas que indican que el modelo se guarda por *Early Stopping* al estancarse o disminuir la función de pérdida.

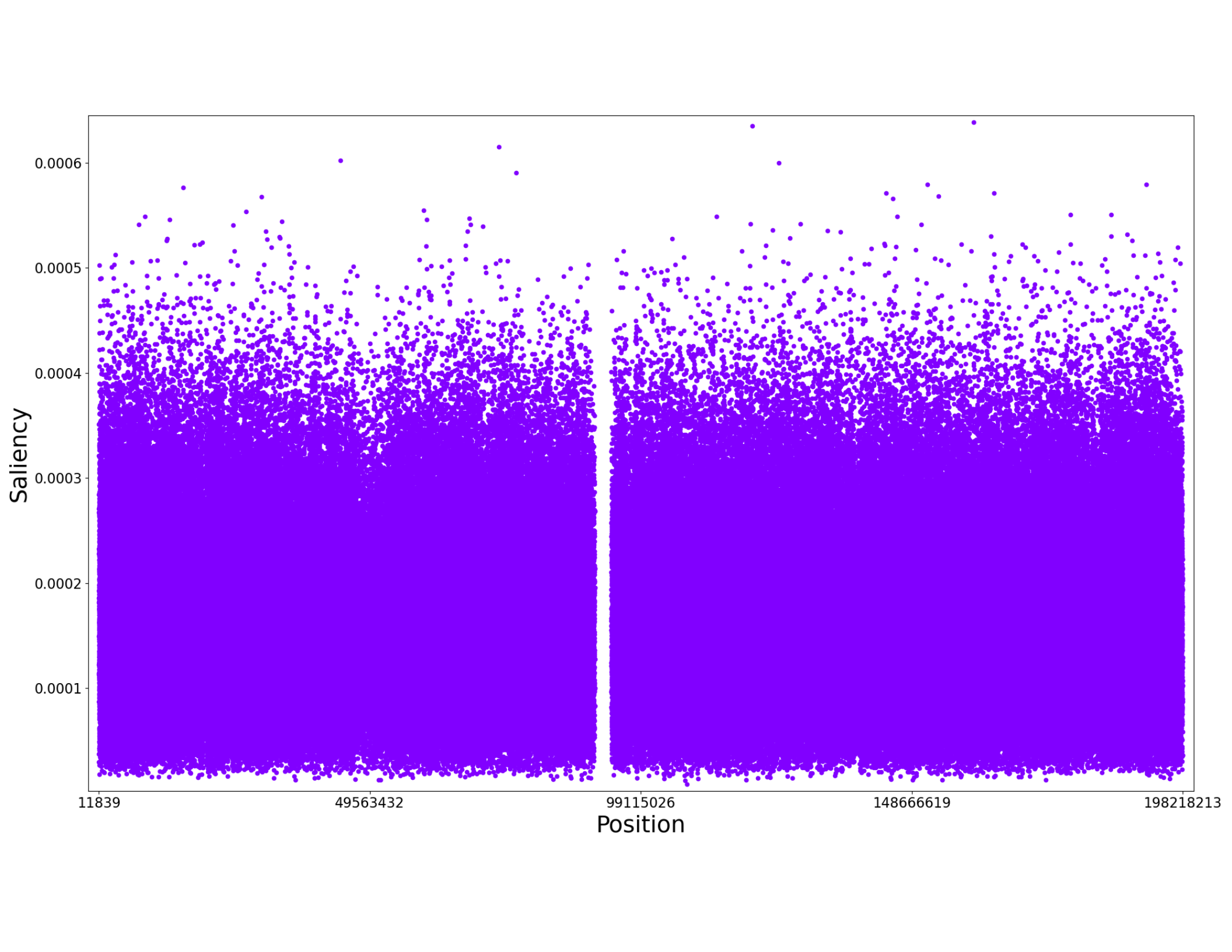
C.4.4. Saliencia

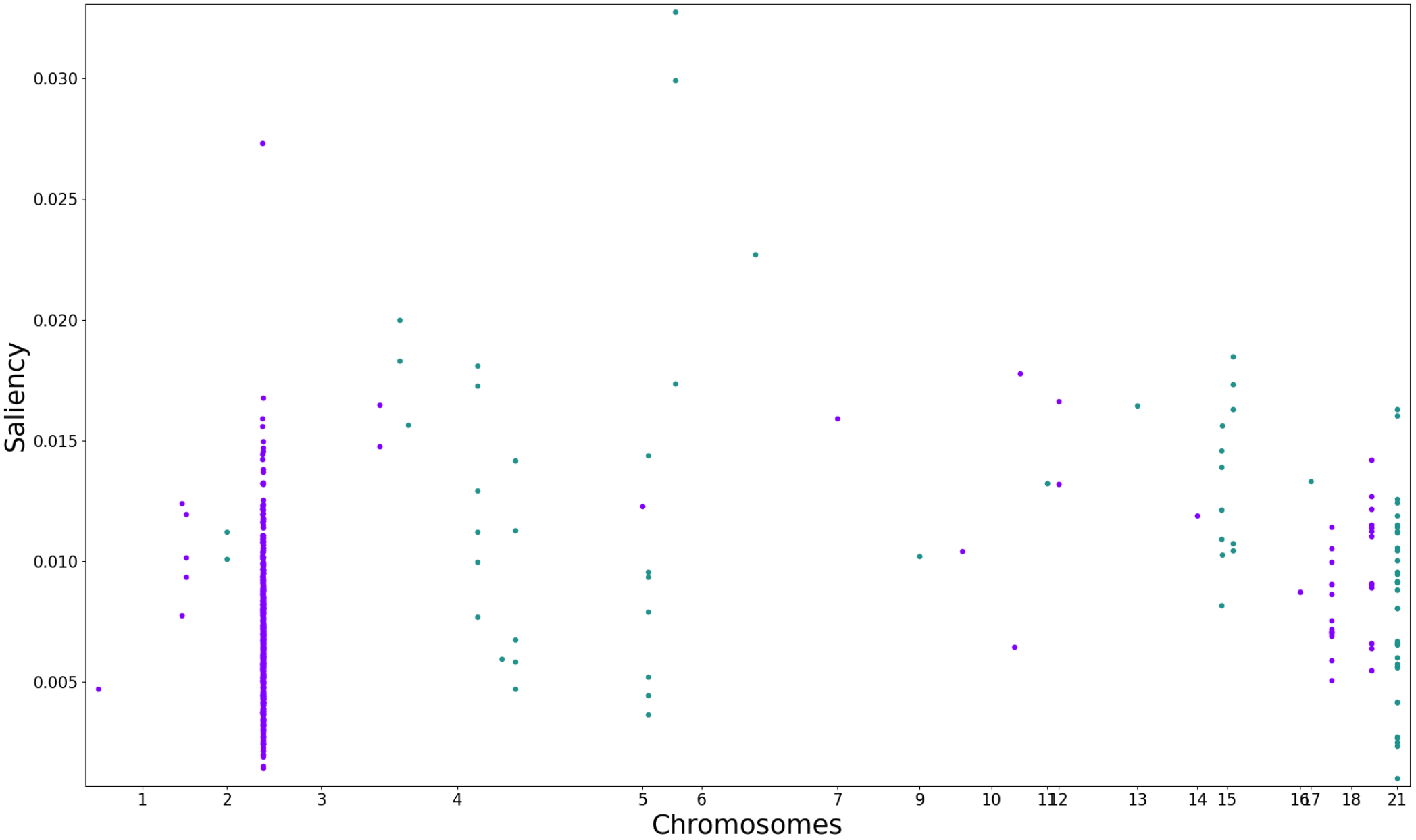
Figura C.73: [Pagina 177] Saliencia obtenida por la red neuronal Dual-stream CNN extendida entrenada con las variantes genéticas obtenidas por el microarreglo ($\sim 4 \cdot 10^5$) y las variables clínicas seleccionadas. El modelo se entrena para predecir hospitalización en participantes contagiados por SARS-Cov-2.

Figura C.74: [Pagina 178] Saliencia obtenida por la red neuronal Dual-stream CNN extendida entrenada con las variantes genéticas imputadas solo del cromosoma 3 ($\sim 5 \cdot 10^5$) y las variables clínicas seleccionadas. El modelo se entrena para predecir hospitalización en participantes contagiados por SARS-Cov-2.

Figura C.75: [Pagina 179] Saliencia obtenida por la red neuronal Dual-stream CNN extendida entrenada con las variantes genéticas obtenidas por el microarreglo ($\sim 6 \cdot 10^2$) y las variables clínicas seleccionadas. El modelo se entrena para predecir hospitalización en participantes contagiados por SARS-Cov-2.


0.0007
0.0006
0.0005
0.0004
0.0003
0.0002
0.0001
Saliency
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
Chromosomes


Saliency
0.0006
0.0005
0.0004
0.0003
0.0002
0.0001
11839
49563432
99115026
148666619
198218213
Position


Saliency
0.030
0.025
0.020
0.015
0.010
0.005
1
2
3
4
5
6
7
9
10
11
12
13
14
15
16
17
18
21
Chromosomes