

# Tabla de Contenido

<b>1. Introducción</b>	<b>1</b>
1.1. Hipótesis . . . . .	2
1.2. Objetivos . . . . .	3
1.2.1. Objetivo General . . . . .	3
1.2.2. Objetivos Específicos . . . . .	3
<b>2. Marco Teórico</b>	<b>4</b>
2.1. Marco Conceptual . . . . .	4
2.1.1. Genes como texto: Funcionamiento del genoma . . . . .	4
2.1.2. Trabajando con variantes genéticas . . . . .	6
2.1.2.1. Secuencia completa . . . . .	7
2.1.2.2. Polimorfismo de Nucleódo Único, SNP . . . . .	8
2.1.2.3. GWAS, Estudio de asociación del genoma completo . . . . .	9
2.1.3. Aprendizaje Profundo . . . . .	10
2.1.4. Interpretabilidad . . . . .	11
2.1.4.1. <i>Feature Selection</i> . . . . .	11
2.1.4.2. Saliencia . . . . .	12
2.2. Trabajos Relacionados . . . . .	13
2.2.1. Aprendizaje de Máquinas en datos genéticos . . . . .	13
2.2.2. Aprendizaje Profundo en bioinformática y biomedicina . . . . .	13
2.2.2.1. Redes convolucionales . . . . .	13
2.2.2.2. Redes recurrentes . . . . .	14
2.2.2.3. <i>Self-Attention</i> sobre secuencias de DNA . . . . .	15
2.2.3. Aplicaciones en Covid-19 . . . . .	16
2.3. Antecedentes . . . . .	19
2.3.1. Análisis proyecto COVID19hg . . . . .	19
2.3.2. Datos genéticos disponibles . . . . .	19
2.3.2.1. Significancia reportada de las variantes genéticas . . . . .	20
<b>3. Metodología</b>	<b>24</b>
3.1. Datos clínicos autoreportados . . . . .	24
3.1.1. Descripción de los datos . . . . .	24

3.1.2.	Severidad . . . . .	25
3.1.2.1.	Cálculo de fenotipo para análisis . . . . .	26
3.1.3.	Imputación . . . . .	28
3.1.4.	Análisis estadístico . . . . .	28
3.2.	Datos Genéticos a nivel de Genoma . . . . .	29
3.2.1.	Preprocesamiento . . . . .	30
3.2.2.	Determinación de significancia . . . . .	31
3.3.	Modelos . . . . .	32
3.3.1.	Modelos de aprendizaje de máquina . . . . .	32
3.3.2.	Redes Neuronales . . . . .	33
3.3.2.1.	FNN, <i>Fully Connected Neural Networks</i> . . . . .	33
3.3.2.2.	CNN, <i>Convolutional Neural Networks</i> . . . . .	35
3.3.2.3.	Modelos adaptados . . . . .	37
3.3.3.	Métricas . . . . .	39
<b>4.</b>	<b>Resultados</b>	<b>41</b>
4.1.	Imputación y Análisis Estadístico . . . . .	42
4.1.1.	Análisis estadísticos . . . . .	48
4.2.	Selección de variables clínicas . . . . .	48
4.3.	Modelos sobre variables clínicas seleccionadas . . . . .	51
4.4.	Modelos sobre datos genéticos . . . . .	52
4.4.1.	Selección de hiperparámetros . . . . .	53
4.4.2.	Desbalance . . . . .	53
4.4.3.	Métricas . . . . .	54
4.4.4.	Saliencia . . . . .	54
4.5.	Agregando datos genéticos . . . . .	58
4.5.1.	Modelos adaptados . . . . .	59
4.5.1.1.	Hiperparámetros . . . . .	59
4.5.1.2.	Métricas y Saliencia . . . . .	60
4.5.2.	Añadiendo variantes más significativas . . . . .	67
<b>5.</b>	<b>Discusión y Conclusiones</b>	<b>71</b>
5.1.	Discusión . . . . .	71
5.1.1.	Datos clínicos autoreportados . . . . .	71
5.1.2.	Datos genéticos a nivel de genoma . . . . .	72
5.1.3.	Contribución de variantes genéticas a los modelos sobre variables clínicas . . . . .	72
5.2.	Conclusiones . . . . .	74
	<b>Bibliografía</b>	<b>75</b>
	<b>Anexo A. Glosario</b>	<b>81</b>

<b>Anexo B. Tablas datos clínicos</b>	<b>84</b>
B.1. Descripción de Encuesta y CRF . . . . .	84
B.2. Análisis Estadístico . . . . .	90
B.2.1. Resumen de Variables . . . . .	90
B.2.2. Análisis Univariado . . . . .	102
B.2.3. Análisis Multivariado . . . . .	106
<b>Anexo C. Resultados de experimentos</b>	<b>110</b>
C.1. Datos Clínicos . . . . .	110
C.1.1. Algoritmos de ML sobre datos imputados y selección de variables . .	110
C.1.2. Algoritmos de ML sobre datos seleccionados . . . . .	115
C.1.3. Redes Neuronales Artificiales (ANN) . . . . .	118
C.1.3.1. Sobre todos los datos clínicos . . . . .	118
C.1.3.2. Sobre datos clínicos seleccionados . . . . .	120
C.2. GWAS . . . . .	122
C.2.1. Resultados Iniciativa . . . . .	122
C.2.2. Todas las variantes . . . . .	125
C.2.3. Cromosoma 3 . . . . .	127
C.2.4. SNPs Genotipificados por microarreglo . . . . .	133
C.3. Datos Genéticos . . . . .	139
C.3.1. Modelos de ML . . . . .	139
C.3.2. Dual-stream CNN . . . . .	142
C.3.2.1. Selección de Hiperparámetros . . . . .	142
C.3.2.1.1 SNPs Genotipificados por microarreglo . . . . .	142
C.3.2.1.2 SNPs Seleccionados . . . . .	144
C.3.2.2. Experimentos con desbalance . . . . .	146
C.3.2.2.1 SNPs Genotipificados por Microarreglo . . . . .	146
C.3.2.2.2 Cromosoma 3 . . . . .	149
C.3.2.2.3 SNPs Seleccionados . . . . .	152
C.3.2.3. Métricas . . . . .	155
C.3.2.4. Saliencia . . . . .	158
C.4. Arquitectura Dual-stream CNN Extendida sobre Datos clínicos y genéticos .	162
C.4.1. Selección de Hiperparámetros . . . . .	162
C.4.2. Experimentos con desbalance . . . . .	164
C.4.2.1. SNPs Genotipificados por Microarreglo . . . . .	164
C.4.2.2. Cromosoma 3 . . . . .	167
C.4.2.3. SNPs Seleccionados . . . . .	170
C.4.3. Métricas . . . . .	173
C.4.4. Saliencia . . . . .	176