CLASIFICACIÓN DE NIVELES DE OBESIDAD MEDIANTE MODELOS DE APRENDIZAJE AUTOMÁTICO: UN ENFOQUE PREDICTIVO BASADO EN DATOS

Simón Olivieri
Leandro Urturi
Camila Muñoz

Facultad de Ciencias Exactas y Naturales y Agrimensura
UNIVERSIDAD NACIONAL DEL NORDESTE
FaCENA - UNNE

DIPLOMATURA EN
CIENCIAS DE DATOS

Facultad de Ciencias Exactas
y Naturales y Agrimensura
UNIVERSIDAD NACIONAL DEL NORDESTE

INTRODUCCIÓN

100% AUMENTO DE LA **OBESIDAD**
1990 - 2022

IMC
≥ 25 **SOBREPESO**
≥ 30 **OBESIDAD**

Facultad de Ciencias Exactas y Naturales y Agrimensura
UNIVERSIDAD NACIONAL DEL NORDESTE

METODOLOGÍA

OBJETIVOS

Desarrollar modelos de clasificación del nivel de obesidad basados en el Aprendizaje Automático utilizando variables características.

Comparar el rendimiento de los distintos modelos propuestos para esta clasificación.

E1 RECOPILACIÓN DE DATOS

DATASET Estimation of Obesity Levels Based On Eating Habits and Physical Condition

2111 REGISTROS

17 VARIABLES

23% **DATOS REALES**

77% **DATOS SINTÉTICOS**

Facultad de Ciencias Exactas y Naturales y Agrimensura
UNIVERSIDAD NACIONAL DEL NORDESTE

METODOLOGÍA

E2
PREPROCESAMIENTO DE DATOS

ESPERADO
3 **VARIABLES NUMÉRICAS**

14 **VARIABLES CATEGÓRICAS**

REAL
8 **NUMÉRICAS**

9 **CATEGÓRICAS**

FINAL
17 **NUMÉRICAS**

Train_set
70%
1477
REGISTROS

Val_set
15%
317
REGISTROS

Test_set
15%
317
REGISTROS

Facultad de Ciencias Exactas y Naturales y Agrimensura
UNIVERSIDAD NACIONAL DEL NORDESTE

METODOLOGÍA

E3
SELECCIÓN DE CARACTERÍSTICAS IMPORTANTES

scikit learn

RANDOM FOREST CLASSIFIER
SELECT FROM MODEL

E4
MODELADO

RANDOM FOREST

HIPERPARÁMETROS	
n_estimators	bootstrap
max_features	max_depth

OPTIMIZADORES

- OPTUNA — Optuna
- scikit learn — RandomizerSearchCV

E4 MODELADO

MODELO DEEP LEARNING

ARQUITECTURA DE LA RED

Modelo Red neuronal	Número de capas ocultas	cantidad de nodos de capas ocultas	Ecuación de Agregación	Ecuación de activación capas de entrada y ocultas	Ecuación de activación capa de salida	DropOut por capa
Modelo 1	2	64	Lineal	Relu	Softmax	No tiene
Modelo 2	2	64	Lineal	Relu	Softmax	0.2
Modelo 3	2	64	Lineal	Tanh	Softmax	0.2
Modelo 4	2	64	Lineal	Tanh	Softmax	0.2
Modelo 5	1	64	Lineal	Tanh	Softmax	No tiene

E4 MODELADO

MODELO DEEP LEARNING

FUN. COSTO – OPTIMIZADORES- HIPERPARÁMETROS

Modelo Red neuronal	Función de coste	Optimizador				
		Optimizador	learning rate	beta	Regularización L2	amsgrad
Modelo 1	CrossEntropyLoss	Adam	0.001	Valor por defecto: (0.9, 0.999)	0	False
Modelo 2	CrossEntropyLoss	Adam	0.0005	Valor por defecto: (0.9, 0.999)	0	False
Modelo 3	CrossEntropyLoss	Adam	0.001	Valor por defecto: (0.9, 0.999)	0	False
Modelo 4	CrossEntropyLoss	Adam	0.0005	Valor por defecto: (0.9, 0.999)	0	False
Modelo 5	CrossEntropyLoss	Adam	0.0005	Valor por defecto: (0.9, 0.999)	0	False

Facultad de Ciencias Exactas y Naturales y Agrimensura
UNIVERSIDAD NACIONAL DEL NORDESTE

RESULTADOS

E1
RECOPILACIÓN DE DATOS

E2
PREPROCESAMIENTO DE DATOS

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2111 entries, 0 to 2110
Data columns (total 17 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   Gender                          2111 non-null   object
 1   Age                             2111 non-null   float64
 2   Height                          2111 non-null   float64
 3   Weight                          2111 non-null   float64
 4   family_history_with_overweight  2111 non-null   object
 5   FAVC                            2111 non-null   object
 6   FCVC                            2111 non-null   float64
 7   NCP                             2111 non-null   float64
 8   CAEC                            2111 non-null   object
 9   SMOKE                           2111 non-null   object
 10  CH2O                            2111 non-null   float64
 11  SCC                             2111 non-null   object
 12  FAF                             2111 non-null   float64
 13  TUE                             2111 non-null   float64
 14  CALC                            2111 non-null   object
 15  MTRANS                          2111 non-null   object
 16  NObeyesdad                      2111 non-null   object
dtypes: float64(8), object(9)
memory usage: 280.5+ KB
```

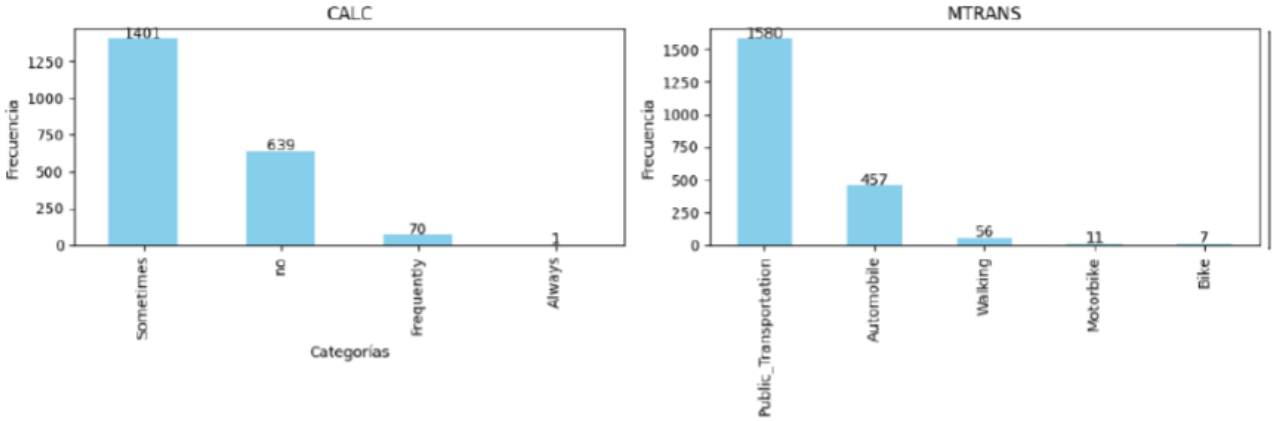


Figura 2. Variables MTRANS y CALC antes del preprocesamiento.

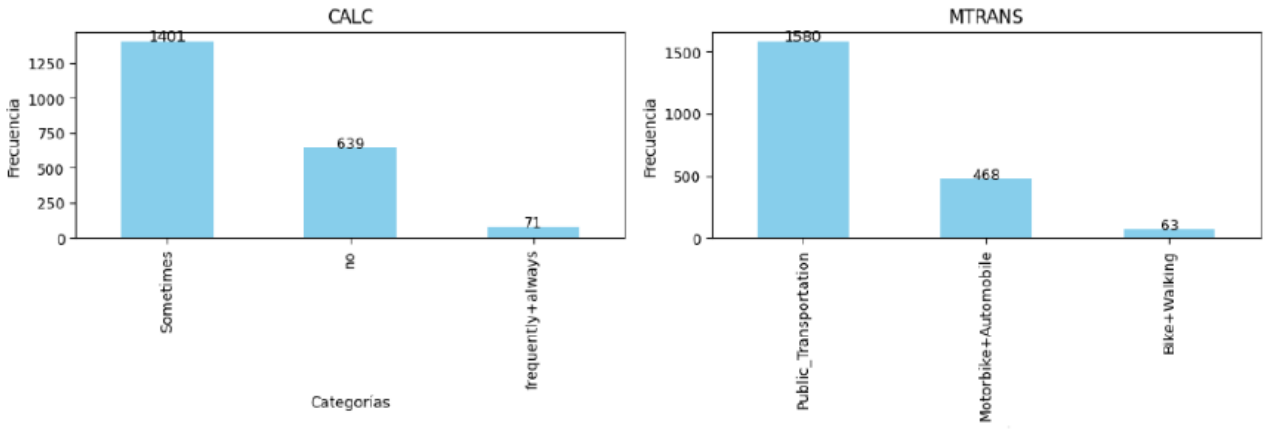


Figura 3. Variables MTRANS y CALC después del preprocesamiento.

Facultad de Ciencias Exactas
y Naturales y Agrimensura
UNIVERSIDAD NACIONAL DEL NORDESTE

RESULTADOS

E2
PREPROCESAMIENTO DE DATOS

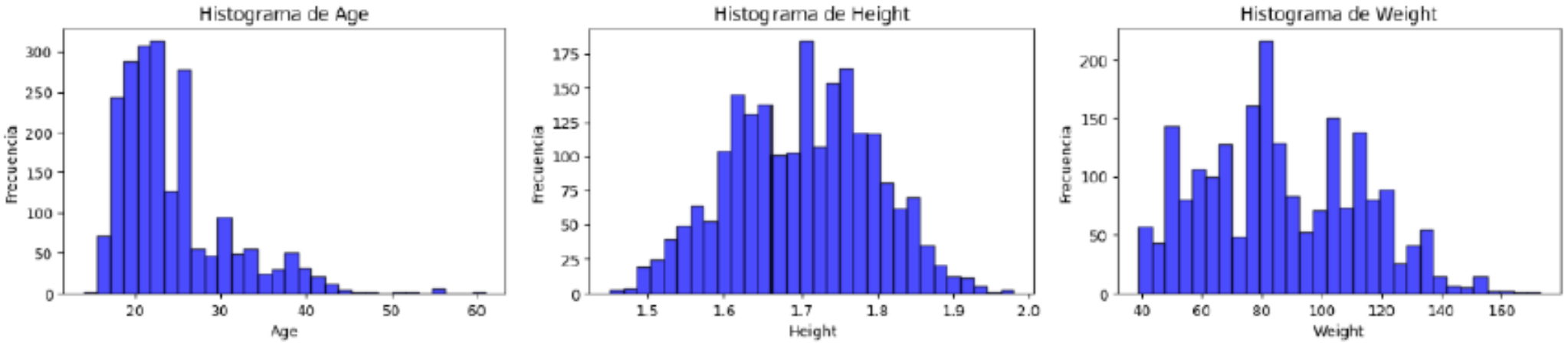


Figura 4. Histograma de las variables numéricas.

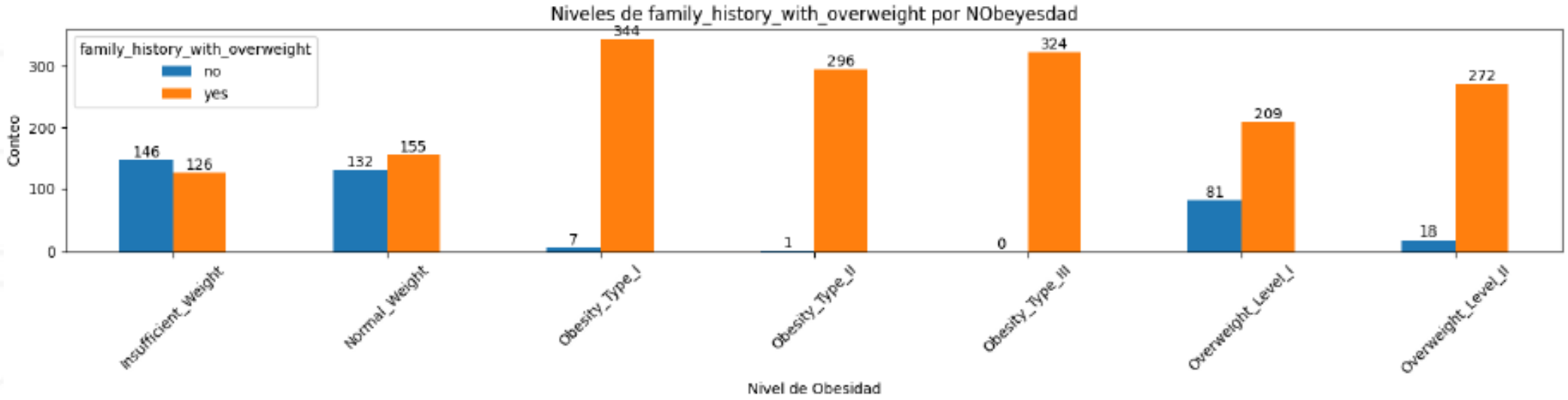


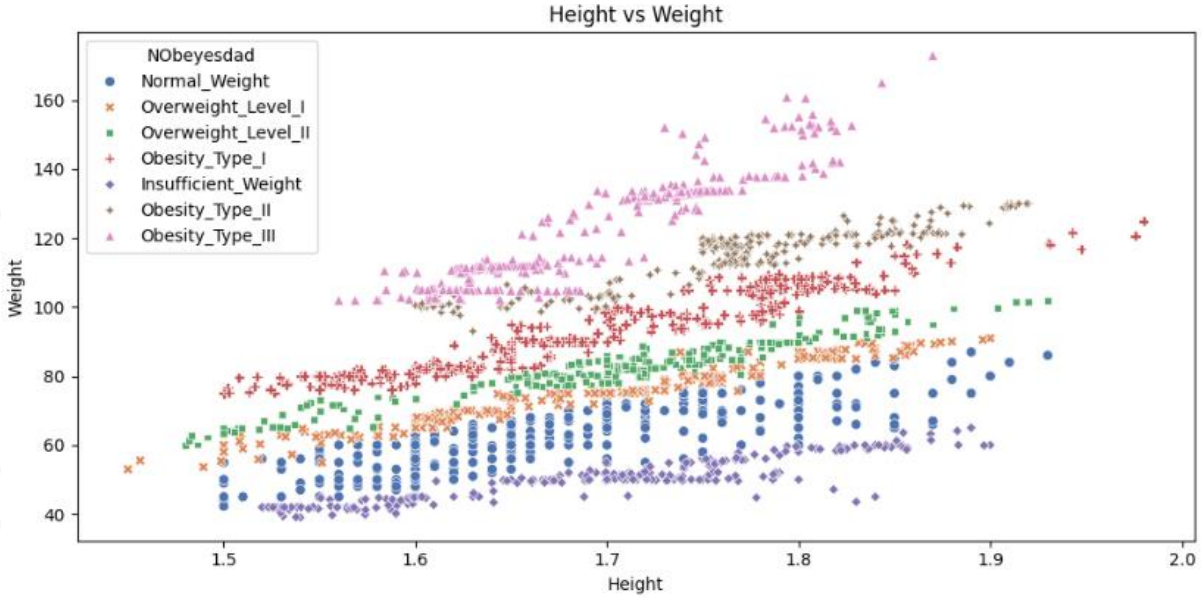
Figura 5. Relación entre la variable de historial familiar de obesidad y los niveles de obesidad.

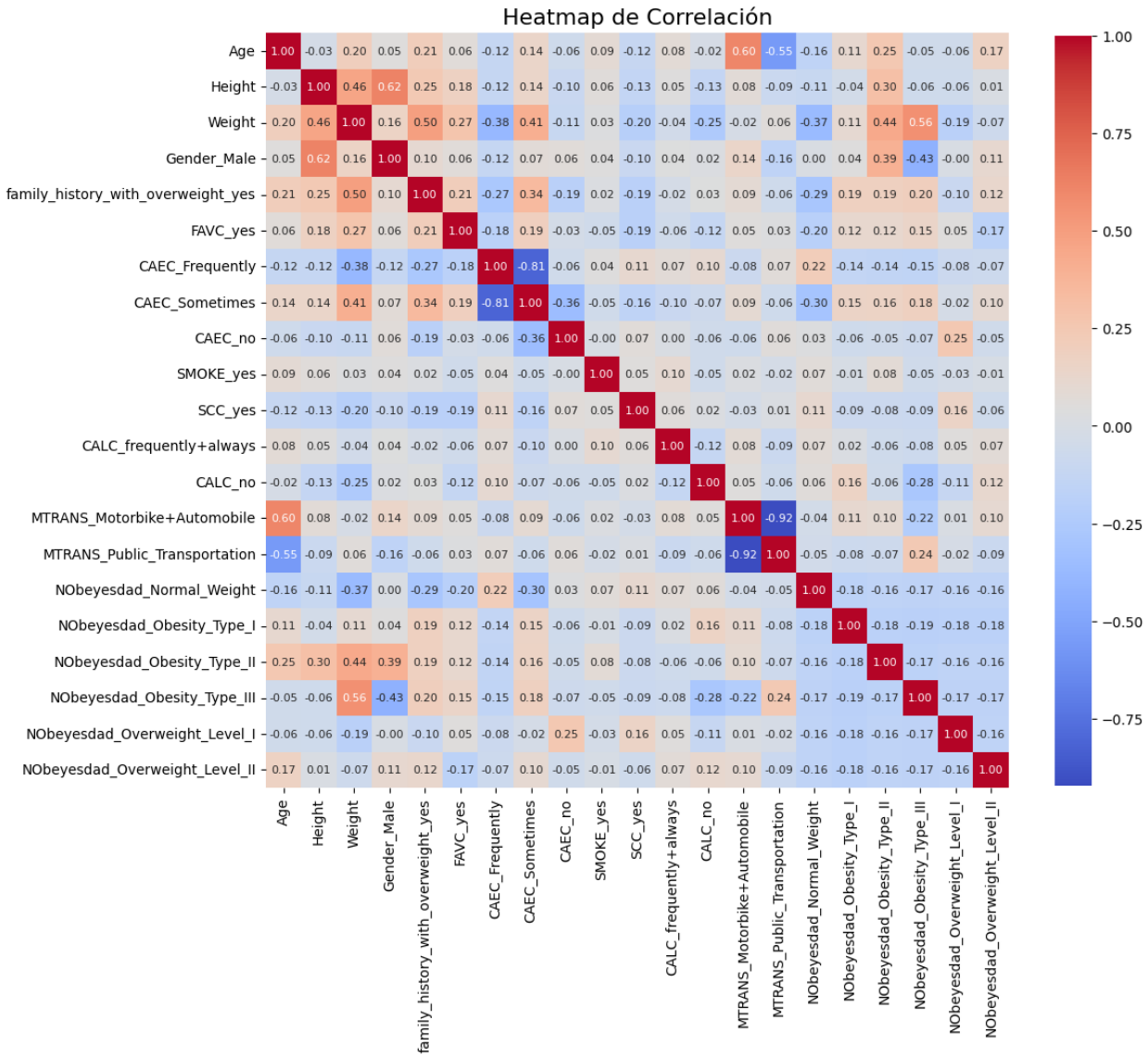
Facultad de Ciencias Exactas
y Naturales y Agrimensura
UNIVERSIDAD NACIONAL DEL NORDESTE

RESULTADOS

E2

PREPROCESAMIENTO DE DATOS





Facultad de Ciencias Exactas y Naturales y Agrimensura
UNIVERSIDAD NACIONAL DEL NORDESTE

RESULTADOS

E3
SELECCIÓN DE CARACTERÍSTICAS IMPORTANTES

SELECCIÓN DE CARACTERÍSTICAS DEL MODELO

Optimización de Hiperparámetros	max_features							
	1	2	3	4	5	6	7	8
RandomizedSearch CV	Weight	Height	Age	Gender_ Male	family_history_with_ overweight_yes	CALC_ no	FAVC_y es	
Optuna	Weight	Height	Age	Gender_ Male	family_history_with_ overweight_yes	CALC_ no	FAVC_y es	
Select model y RandomizedSearch CV	Weight	Height	Age	Gender_ Male	family_history_with_ overweight_yes	CALC_ no	FAVC_y es	CAEC_Someti mes

E4
MODELADO

RESULTADOS DE LOS MODELOS CON RANDOMFORESTCLASSIFIER

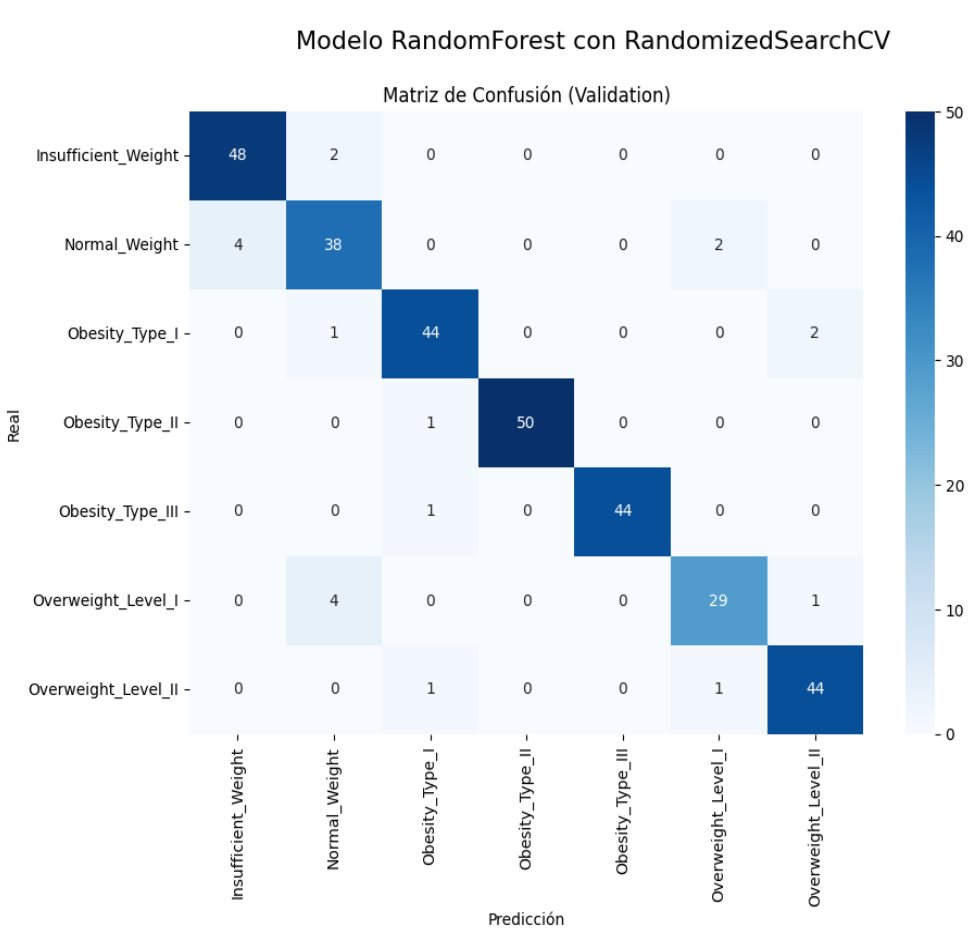
Optimización de Hiperparámetros	Mejores hiperparametros				Muestra de validación		Muestra de testeo	
	n_estimators	max_depth	bootstrap	max_ features	Acurracy	F1-score	Acurracy	F1-score
RandomizedSearch CV	120	58	False	7	0.952	0.952	0.927	0.927
Optuna	203	84	False	7	0.952	0.95	0.946	0.946

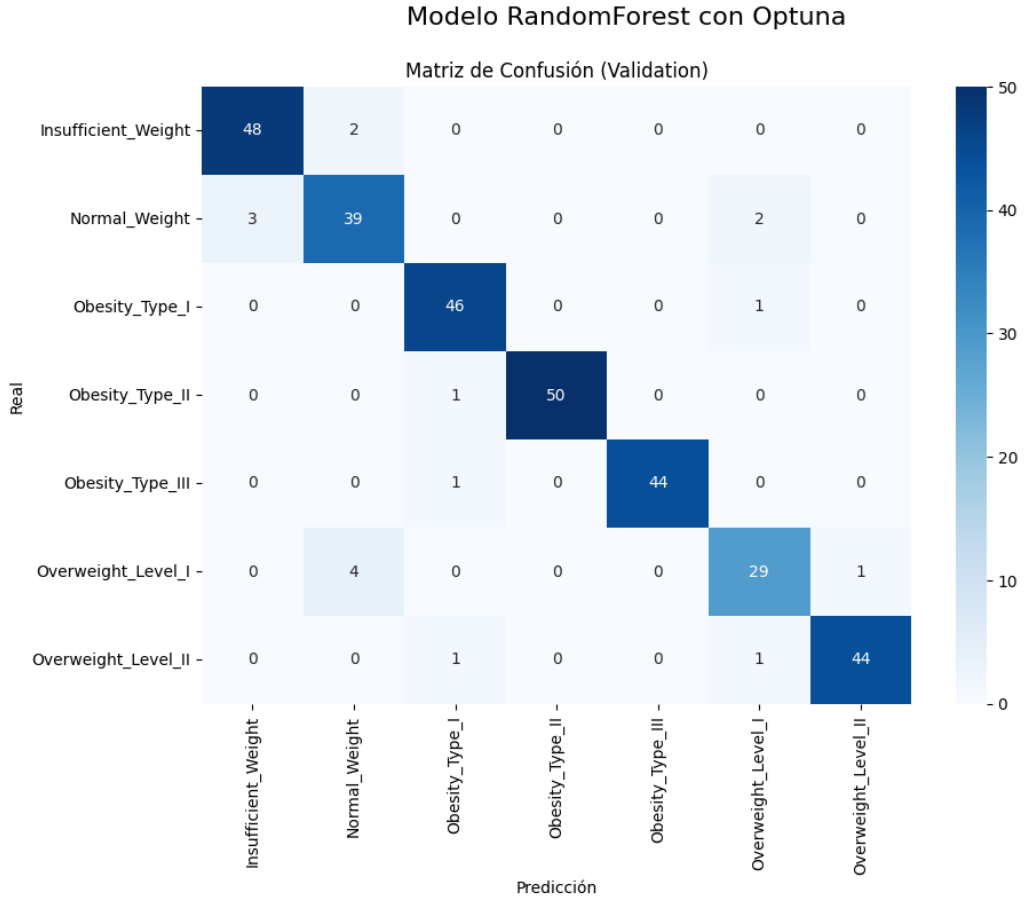
Facultad de Ciencias Exactas y Naturales y Agrimensura
UNIVERSIDAD NACIONAL DEL NORDESTE

RESULTADOS

E4
MODELADO

COMPARACIÓN ENTRE MATRICES DE CONFUSIÓN ENTRE AMBOS MODELOS





Facultad de Ciencias Exactas y Naturales y Agrimensura
UNIVERSIDAD NACIONAL DEL NORDESTE

RESULTADOS

E4
MODELADO

REDES NEURONALES

Modelo Red neuronal	mini batch	epochs	Muestra de validación			Muestra de testeo	
			Acurracy	F1-score	Error	Acurracy	F1-score
Modelo 1	32	1250	0.959	0.958	0.30	0.981	0.981
Modelo 2	32	1250	0.962	0.961	0.30	0.974	0.974
Modelo 3	32	1250	0.962	0.962	0.12	0.981	0.981
Modelo 4	64	1250	0.962	0.962	0.09	0.977	0.978
Modelo 5	32	1250	0.971	0.971	0.12	0.971	0.971

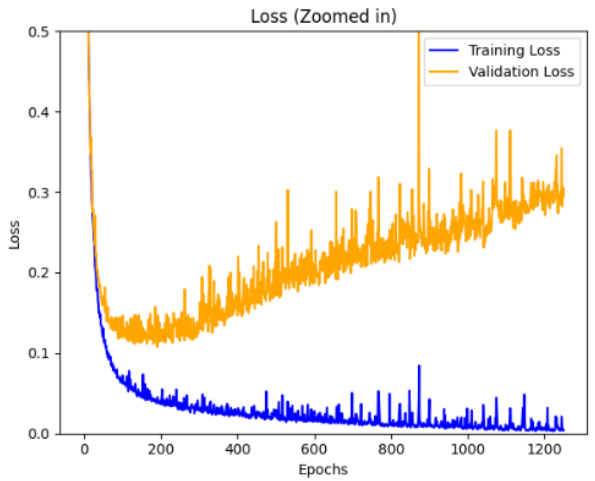
MODELO 5

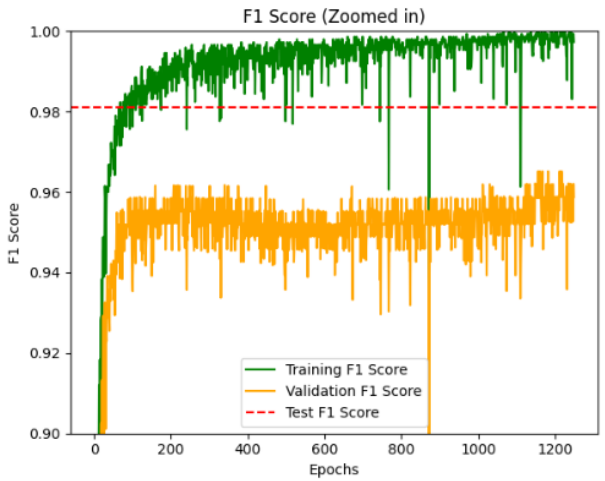
Facultad de Ciencias Exactas y Naturales y Agrimensura
UNIVERSIDAD NACIONAL DEL NORDESTE

RESULTADOS

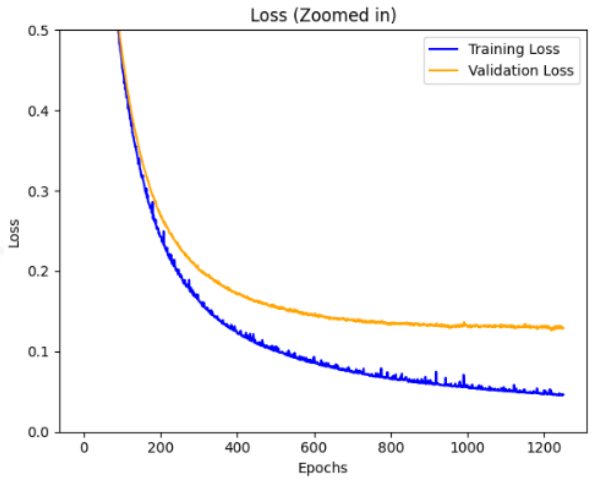
E4
MODELADO

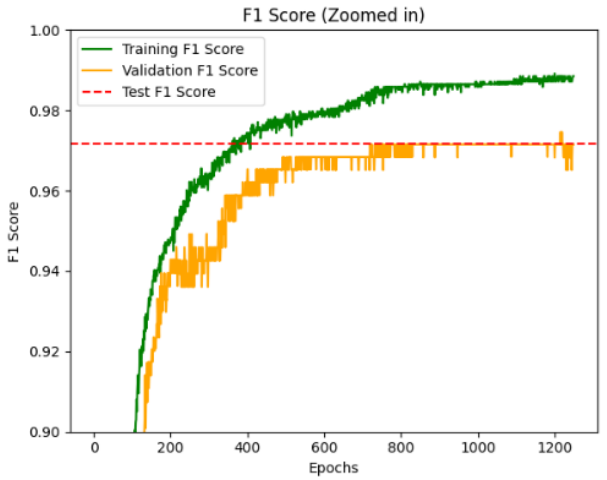
MODELO 1



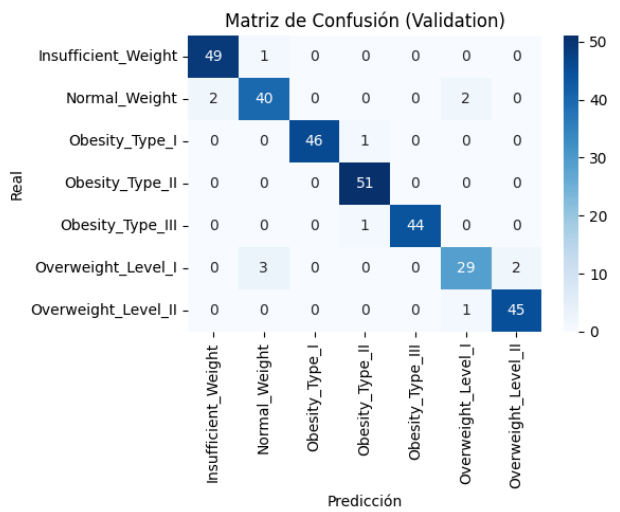


MODELO 5

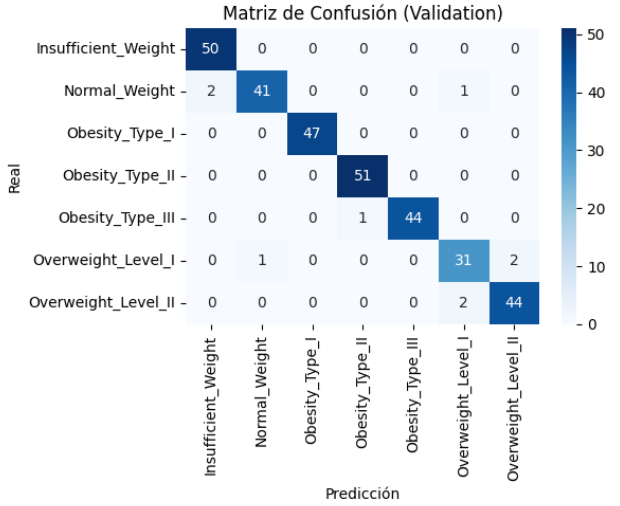




MODELO 1



MODELO 5



Facultad de Ciencias Exactas
y Naturales y Agrimensura
UNIVERSIDAD NACIONAL DEL NORDESTE

CONCLUSIONES

OPTUNA
RANDOM FOREST

94,6%

F1-score

MODELO 5
REDES NEURONALES

97,1%

F1-score