

# Comparación de Técnicas de Aprendizaje Automático Supervisado

Autor: Nicolás Seivane

Tutor: Andrea Rey

Fecha

Universidad Nacional de Hurlingham





# Índice General

<b>1</b>	<b>Introducción</b>	<b>11</b>
1.1	Motivación . . . . .	11
1.2	Estado del Arte . . . . .	11
1.3	Conjuntos de datos Utilizados . . . . .	11
1.3.1	Dataset Binario: Insuficiencia Cardíaca Predicción . . . . .	11
1.3.2	Dataset Multiclaso: Cardiotocografía Predicción . . . . .	13
<b>2</b>	<b>Métricas de Rendimiento Utilizadas</b>	<b>17</b>
2.1	Introducción . . . . .	17
2.2	Métricas para caso Binario . . . . .	17
2.2.1	Matriz de Confusión . . . . .	17
2.2.2	<i>Accuracy</i> . . . . .	18
2.2.3	<i>Precision</i> . . . . .	18
2.2.4	<i>Recall</i> . . . . .	18
2.2.5	<i>F-measure</i> . . . . .	19
2.2.6	<i>Recall</i> . . . . .	19
2.2.7	<i>Área Bajo la Curva ROC (ROC AUC)</i> . . . . .	19
2.3	Métricas para caso Multiclaso . . . . .	19
2.3.1	Matriz de Confusión (Multiclaso) . . . . .	20
2.3.2	<i>Precision</i> . . . . .	20
2.3.3	<i>Recall</i> . . . . .	20
2.3.4	<i>F-measure</i> . . . . .	21
2.3.5	<i>Área Bajo la Curva ROC (ROC AUC)</i> . . . . .	21
<b>3</b>	<b>Descripción de los Métodos Utilizados</b>	<b>23</b>
3.1	Regresión Logística . . . . .	23
<b>4</b>	<b>Resultados</b>	<b>25</b>
4.1	Introducción . . . . .	25
4.2	Métricas de Evaluación . . . . .	25
4.2.1	Dataset Binario . . . . .	25
4.2.2	Regresión Logística . . . . .	25
4.2.3	Máquinas de Soporte Vectorial (SVM) . . . . .	25
4.2.4	Naive Bayes Gaussiano . . . . .	26
4.2.5	Random Forest . . . . .	26
4.3	Importancia de las Características . . . . .	26
4.3.1	Random Forest . . . . .	27
4.3.2	Regresión Logística . . . . .	27
4.3.3	SVM . . . . .	27

4.3.4	Naive Bayes Gaussiano . . . . .	27
4.3.5	Importancia de las Características (Coeficientes Absolutos de Regresión Logística)	28
4.3.6	Dataset Multiclasificación . . . . .	28
4.3.7	Regresión Logística . . . . .	28
4.3.8	Máquinas de Soporte Vectorial (SVM) . . . . .	29
4.3.9	Naive Bayes Gaussiano . . . . .	29
4.3.10	Random Forest . . . . .	29
4.4	Importancia de las Características . . . . .	29
4.4.1	Random Forest . . . . .	30
4.4.2	Regresión Logística . . . . .	30
4.4.3	SVM . . . . .	31
4.4.4	Naive Bayes Gaussiano . . . . .	31
4.4.5	Importancia de las Características (Coeficientes Absolutos de Regresión Logística)	32
<b>5</b>	<b>Conclusiones</b>	<b>33</b>

# Índice de Figuras



# Índice de Tablas



# Resumen



# Chapter 1

## Introducción

Describir el problema que se desea resolver

### 1.1 Motivación

Explicar porqué estudiamos este problema, para qué sirve, cuál es el impacto y en qué áreas.

### 1.2 Estado del Arte

En esta sección se realiza una descripción de algunos de los métodos más importantes existentes en la bibliografía describiendo el problema y el método utilizado por cada autor. Se cita la bibliografía.

### 1.3 Conjuntos de datos Utilizados

Se realiza este informe de registros, atributos y métricas relevantes luego de eliminar duplicados, datos faltantes y anormales.

#### 1.3.1 Dataset Binario: Insuficiencia Cardíaca Predicción

**Titulo Original:** Heart Failure Prediction Dataset

**Citación:** fedesoriano. (September 2021). Heart Failure Prediction Dataset. Retrieved [Date Retrieved] from <https://www.kaggle.com/fedesoriano/heart-failure-prediction>

**Descripción:** Enfermedades cardiovasculares son la causa numero uno de muerte globalmente, tomando un estimado de 17.9 millones de vidas cada año, que son aproximadamente 31% de todas las muertes globales.

Este dataset fue creado mediante la combinación de distintos dataset disponibles independientes pero no combinados anteriormente. 5 datasets de información cardíaca están combinados en 11 atributos comunes logrando el dataset mas grande de informacion de enfermedades cardiovasculares utilizado para investigación. Los 5 datasets utilizados para la creación de este son:

- Cleveland: 303 observaciones
- Hungarian: 294 observaciones
- Switzerland: 123 observaciones
- Long Beach VA: 200 observaciones

- Stalog (Heart) Data Set: 270 observaciones

**Cantidad de registros:** 918

**Cantidad de registros valiosos:** 743

**Cantidad de atributos:** 11

**Atributos Categóricos:** 5

**Atributos Numéricos:** 6

Los atributos son (Algunos son numéricos en el dataset pero son codificaciones de categóricos):

Atributo	Tipo de dato	¿Esta codificado?	Unidad
Age	Numérico (int)	No	Años
Sex	Categorico (string)	No	-
ChestPainType	Categorico (string)	No	-
RestingBP	Numérico (int)	No	mm Hg
Cholesterol	Numérico (int)	No	mm/dl
FastingBS	Numérico (int)	Si	"mg/dl"
RestingECG	Categorico (string)	No	-
MaxHR	Numérico (int)	No	-
ExerciseAngina	Categorico (string)	No	-
Oldpeak	Numérico (float)	No	ST en depresión
ST_Slope	Categorico (string)	No	-
HeartDisease	Numérico (int)	Si	-

#### Descripción atributos:

**Age:** Edad de los pacientes. Tiene media: 53 años, valor máximo: 77 y valor mínimo: 28, con proporciones de edad bastante bien distribuidas, siendo la menor de 0.11% para algunas edades y la mayor de 4.14% para otras edades, teniendo otras distribuciones entre estos dos rangos

**Sex:** Sexo de los pacientes; hay 78.98% M (masculinos) y hay 21.02% F (femeninos)

**ChestPainType:** Tipo del dolor en el pecho; Hay 18.85% ATA, hay 22.11% NAP, hay 54.03% ASY, hay 5.01% TA. [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]

**RestingBP:** Presión sanguínea en reposo, donde hay 51.09% de mujeres, codificadas en 1 y 48.91% de hombres, codificados en 0.

**Cholesterol:** Colesterol serico, la medida total de colesterol en sangre; tiene media: 199.02, valor máximo: 603.00 y valor mínimo: 0.00. Miligramos por decilitro

**FastingBS:** Glucosa en sangre en ayuno; hay 76.66% Glucosa en sangre < 120 mg/dl codificado en 0 y hay 23.34% Glucosa en sangre > 120 mg/dl codificado en 1

**RestingECG:** Resultados de electrocardiogramas en reposo; hay 60.09% codificado en Normal, hay 19.41% codificado en ST y hay 20.50% codificado en LVH [Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria]

**MaxHR:** Máximo ritmo cardíaco registrado, tiene media: 136.79, valor maximo: 202.00 y valor minimo: 60.00

**ExerciseAngina:** Angina producido por ejercicio, dolor en el pecho; hay 59.54% No codificado en N y hay 40.46% Si codificado en Y

**Oldpeak:** Valor máximo de depresión del segmento ST (en milímetros) registrado en todas las derivaciones contiguas durante una prueba de esfuerzo. Forma parte del cálculo del riesgo de un paciente de isquemia o infarto de miocardio; valores más altos indican un mayor riesgo de enfermedad coronaria; tiene media: 0.90, valor maximo: 6.20 y valor minimo: -0.10

**ST\_Slope:** The slope of the peak exercise ST segment; hay 43.08%, hay 50.05% Flat y hay 6.87% Down [Up: upsloping, Flat: flat, Down: downsloping]

**HeartDisease:** Variable de salida de si posee una enfermedad cardíaca; hay 44.71% No codificado en 0 y hay 55.29% Si codificado en 1

**Función Objetivo Inicial:** Donde la variable salida es *HeartDisease*, no hay una variable que se use como condición:

$$f(x) = \begin{cases} '1' & \text{si ??} \\ '0' & \text{si ??} \end{cases}$$

### 1.3.2 Dataset Multiclasificación: Cardiotocografía Predicción

**Título Original:** Cardiotocography

**Cita:** Campos, D. & Bernardes, J. (2000). Cardiotocography [Dataset]. UCI Machine Learning Repository. <https://doi.org/10.24432/C51S4N>.

**Descripción:** La cardiotocografía (CTG) es un registro continuo de la frecuencia cardíaca fetal que se obtiene mediante un transductor de ultrasonidos colocado en el abdomen materno. La CTG se utiliza ampliamente durante el embarazo como método para evaluar el bienestar fetal, sobre todo en embarazos con mayor riesgo de complicaciones.

Se procesaron automáticamente 2126 cardiotocogramas fetales (CTG) y se midieron sus características diagnósticas. Tres obstetras expertos clasificaron los CTG y se les asignó una etiqueta de clasificación consensuada. La clasificación se realizó tanto con respecto a un patrón morfológico (A, B, C...) como al estado fetal (N, S, P).

**Cantidad de registros:** 2126

**Cantidad de registros valiosos:** 2115

**Cantidad de atributos:** 21

**Atributos Categóricos:** 0

**Atributos Numéricos:** 21

Los atributos son (Algunos son numéricos en el dataset pero son codificaciones de categóricos):

Atributo	Tipo de dato
LB	Numérico (int)
AC	Numérico
FM	Numérico (float)
UC	Numérico (float)
DL	Numérico (float)
DS	Numérico (float)
DP	Numérico (float)
ASTV	Numérico (int)
MSTV	Numérico (float)
ALTV	Numérico (int)
MLTV	Numérico (float)
Width	Numérico (int)
Min	Numérico (int)
Max	Numérico (int)
Nmax	Numérico (int)
Nzeros	Numérico (int)
Mode	Numérico (int)
Mean	Numérico (int)
Median	Numérico (int)
Variance	Numérico (int)
Tendency	Numérico (int)
NSP	Categórico (string)

**Descripción atributos:**

**LB:** Frecuencia cardíaca fetal basal (latidos por minuto). Tiene media: 133.30, valor máximo: 160.00 y valor mínimo: 106.00

**AC:** Número de aceleraciones por segundo. Tiene media: 0.00, valor máximo: 0.02 y valor mínimo: 0.00

**FM:** Número de movimientos fetales por segundo. Tiene media: 0.01, valor máximo: 0.48 y valor mínimo: 0.00

**UC:** Número de contracciones uterinas por segundo. Tiene media: 0.00, valor máximo: 0.01 y valor mínimo: 0.00

**DL:** Número de desaceleraciones leves por segundo. Tiene media: 0.00, valor máximo: 0.01 y valor mínimo: 0.00

**DS:** Número de desaceleraciones severas por segundo. Hay un 99.67% con valor 0.0 y un 0.33% con un valor 0.001

**DP:** Número de desaceleraciones prolongadas por segundo. Hay un 91.58% con valor 0.0, 3.40% con un valor 0.002, 1.13% con un valor 0.003, 3.31% con un valor 0.001, 0.43% con un valor 0.004 y 0.14% con un valor 0.005

**ASTV:** Porcentaje de tiempo con variabilidad anormal a corto plazo. Tiene media: 46.98, valor máximo: 87.00 y valor mínimo: 12.00

**MSTV:** Valor medio de la variabilidad a corto plazo. Tiene media: 1.34, valor máximo: 7.00 y valor mínimo: 0.20

**ALTV:** Porcentaje de tiempo con variabilidad anormal a largo plazo. Tiene media: 9.79, valor máximo: 91.00 y valor mínimo: 0.00

**MLTV:** Valor medio de la variabilidad a largo plazo. Tiene media: 8.17, valor máximo: 50.70 y valor mínimo: 0.00

**Width:** Ancho del histograma de FCF. Tiene media: 70.51, valor máximo: 180.00 y valor mínimo: 3.00

**Min:** Mínimo del histograma de FCF. Tiene media: 93.57, valor máximo: 159.00 y valor mínimo: 50.00

**Max:** Máximo del histograma de FCF. Tiene media: 164.09, valor máximo: 238.00 y valor mínimo: 122.00

**Nmax:** Número de picos del histograma. Tiene media: 4.08, valor máximo: 18.00 y valor mínimo: 0.00

**Nzeros:** Número de ceros del histograma. Hay un 76.26% con valor 0, 17.30% con un valor 1, 0.99% con un valor 3, 5.11% con un valor 2, 0.09% con un valor 4, 0.05% con un valor 10, 0.09% con un valor 5, 0.05% con un valor 8, y 0.05% con un valor 7.

**Mode:** Moda del histograma. Tiene media: 137.45, valor máximo: 187.00 y valor mínimo: 60.00

**Mean:** Promedio del histograma. Tiene media: 134.60, valor máximo: 182.00 y valor mínimo: 73.00

**Median:** Media del histograma. Tiene media: 138.08, valor máximo: 186.00 y valor mínimo: 77.00

**Variance:** Varianza del histograma. Tiene media: 18.89, valor máximo: 269.00 y valor mínimo: 0.00

**Tendency:** Tendencia del histograma. Hay un 39.67% con valor 1, 52.53% con un valor 0 y 8.27% con un valor -1

**CLASS:** código de clasificación del estado fetal (N=normal; S=sospechoso; P=patológico). Hay un 13.81% con valor Sospechoso, 77.92% con un valor Normal, 8.27% con un valor Patológico.

**Función Objetivo Inicial:** Donde la variable salida es *CLASS*:

$$f(x) = \begin{cases} \text{'Sospechoso'} & \text{si ??} \\ \text{'Normal'} & \text{si ??} \\ \text{'Patológico'} & \text{si ??} \end{cases}$$



# Chapter 2

## Métricas de Rendimiento Utilizadas

### 2.1 Introducción

Dentro del objetivo de este trabajo es evaluar el desempeño calificador de cada modelo de *Machine Learning*. Para alcanzar este objetivo, se utilizaran **métricas de rendimiento** que permiten cuantificar la capacidad del algoritmo de clasificar.

**La importancia de las métricas** se ubica en que el objetivo central de estos algoritmos no es simplemente obtener un buen rendimiento en los datos utilizados para construir el modelo, sino en su **capacidad de generalización**, su habilidad para funcionar correctamente con entradas nuevas y previamente no observadas (no utilizadas en el entrenamiento).

Para la obtención de las métricas y entrenamiento de algoritmo se utilizara la estrategia de **Validación Cruzada  $k$ -fold**, donde el conjunto de datos se divide en  $k$  grupos (o pliegues, en una traducción más fiel) del mismo tamaño, donde en cada iteración un grupo  $k$  es utilizado para entrenar y el resto para evaluar, repitiéndose el proceso  $k$  veces, donde un grupo  $k_i$  es utilizado solo una vez para entrenar. El valor final estimado de la métrica es el promedio de los valores obtenidos de cada grupo.

$$\hat{M} = \frac{1}{k} \sum_{i=1}^k M_i$$

Donde:

- $\hat{M}$  es el valor estimado de la métrica de evaluación.
- $k$  es la cantidad de grupos en los que se dividió el conjunto de datos.
- $M_i$  es el valor de la métrica de evaluación obtenido en el  $i$ -ésimo grupo utilizado como conjunto de prueba.

Dentro de este trabajo no sólo se evaluaran distintos modelos, sino que se utilizaran distintos *datasets* para lograrlos.

### 2.2 Métricas para caso Binario

#### 2.2.1 Matriz de Confusión

Una matriz de confusión es una forma simple de saber de que forma esta clasificando el algoritmo, donde una clase es considerada **positiva  $P$**  y la otra **negativa  $N$** . La matriz de confusión clasifica las predicciones en:

- **Verdaderos Positivos (TP):** Casos positivos clasificados correctamente.

- **Verdaderos Negativos (TN):** Casos negativos clasificados correctamente.
- **Falsos Positivos (FP):** Casos negativos clasificados incorrectamente como positivos.
- **Falsos Negativos (FN):** Casos positivos clasificados incorrectamente como positivos.

		Predicción	
		Positivo	Negativo
Verdad	Positivo	Verdadero Positivo (TP)	Falso Negativo (FN)
	Negativo	Falso Positivo (FP)	Verdadero Negativo (TN)

### 2.2.2 Accuracy

El *Accuracy* es la proporción de instancias clasificadas correctamente, es una medida "ingenua" que puede ser engañosa si existe un gran desbalance entre clases.

En términos de la Matriz de Confusión:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} = \frac{TP + TN}{\text{Total}}$$

En términos del conjunto de predicciones y valores verdaderos:

$$\text{accuracy}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} 1(\hat{y}_i = y_i)$$

Donde:

- $n_{\text{samples}}$ : Representa la cantidad total de ejemplos en la muestra
- $\hat{y}_i$  y  $y_i$ :  $\hat{y}_i$  es el valor predicho del  $i$ -ésimo ejemplo, y  $y_i$  es el valor verdadero correspondiente por lo tanto, calcula:

$$\text{Accuracy} = \frac{\text{Número de predicciones correctas}}{\text{Número total de muestras}}$$

### 2.2.3 Precision

El *Precision* mide la probabilidad de que la predicción positiva del clasificador sea correcta.

En términos de la Matriz de Confusión:

$$\text{Precision} = \frac{TP}{TP + FP}$$

### 2.2.4 Recall

El *Recall* o también conocido como Sensibilidad o Tasa de Verdaderos Negativos (TPR). Mide la probabilidad de que el clasificador detecte un caso positivo cuando en verdad lo es.

En términos de la Matriz de Confusión:

$$\text{Recall} = TPR = \frac{TP}{TP + FN} = \frac{TP}{P}$$

### 2.2.5 *F-measure*

El *F-measure* es la media armónica ponderada de *precision* y *recall*. La versión más común es el **F1-score**, donde el parámetro de ponderación  $\beta$  es igual a 1. Un clasificador perfecto tiene un valor  $F1 = 1$ .

Fórmula General ( $F_\beta$ ):

$$F_\beta = \frac{(1 + \beta^2)\text{precision} \times \text{recall}}{\beta^2\text{precision} + \text{recall}}$$

Fórmula del F1-score ( $\beta = 1$ ) en términos de Precision y Recall:

$$F1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

En términos de la Matriz de Confusión:

$$F1 = \frac{2TP}{2TP + FP + FN}$$

### 2.2.6 *Recall*

El *Recall* o también conocido como Sensibilidad o Tasa de Verdaderos Negativos (TPR). Mide la probabilidad de que el clasificador detecte un caso positivo cuando en verdad lo es.

En términos de la Matriz de Confusión:

$$\text{Recall} = TPR = \frac{TP}{TP + FN} = \frac{TP}{P}$$

### 2.2.7 *Área Bajo la Curva ROC (ROC AUC)*

La métrica *ROC AUC* es un valor que resume la capacidad de un clasificador para distinguir entre clases.

**La Curva ROC** es un gráfico que ilustra el rendimiento de un clasificador binario a medida que se varía su umbral de discriminación. Se crea graficando la **Tasa de Verdaderos Positivos (TPR)** versus la **Tasa de Falsos Positivos (FPR)** en varios umbrales.

Ejes utilizados para el gráfico:

- **Eje Y:** TPR
- **Eje X:** FPR

El **AUC** mide justamente el área debajo de la Curva ROC. El AUC se utiliza para comparar el desempeño de diferentes modelos de clasificación.

Interpretación de valores:

- Un clasificador ideal se ubica en el punto  $(0, 1)$ , donde  $TPR = 1$  y  $FPR = 0$ , lo que resulta en un  $AUC = 1$
- Un clasificador aleatorio se sitúa sobre la línea  $TPR = FPR$ , lo que resulta en un  $AUC = 0.5$
- Un clasificador se considera razonable si  $0.5 < AUC \leq 1$

## 2.3 Métricas para caso Multiclasé

En este caso se utiliza el método "*weighted*", el cual computa el desequilibrio de clases calculando el promedio de métricas binarias en las que la puntuación de cada clase se pondera según su presencia en la muestra de datos reales.

La métrica ponderada por la presencia de la clase,  $M_{\text{weighted}}$ , se calcula como el promedio de la métrica por clase  $M_l$ , donde cada contribución es ponderada por el tamaño de la clase  $|y_l|$ :

$$\hat{M}_{\text{weighted}} = \frac{1}{\sum l \in L |y_l|} \sum_{l \in L} |y_l| \cdot M_l$$

Donde:

- $\hat{M}_{\text{weighted}}$  es el valor estimado de la métrica promedio ponderada.
- $L$  es el conjunto de etiquetas o clases.
- $|y_l|$  es el soporte o cantidad de muestras verdaderas que tienen la etiqueta  $l$ .
- $\sum_{l \in L} |y_l|$  representa el número total de pares (muestra, etiqueta) verdaderos en el conjunto de datos.
- $M_l$  es el valor de la métrica binaria (como Precisión, o  $F\beta$ -score) calculado para la clase individual  $l$

### 2.3.1 Matriz de Confusión (Multiclas)

La matriz de confusión multiclas es una matriz cuadrada de tamaño  $L \times L$ , donde  $L$  es el número de clases. Cada celda  $C_{ij}$  representa la cantidad de muestras verdaderamente pertenecientes a la clase  $i$  que fueron clasificadas como clase  $j$ .

Para cada clase  $l$  se definen los valores:

- $TP_l = C_{ll}$
- $FP_l = \sum_{i \neq l} C_{il}$
- $FN_l = \sum_{j \neq l} C_{lj}$
- $TN_l = N - TP_l - FP_l - FN_l$

		Predicción			
		Clase $C_1$	Clase $C_2$	...	Clase $C_l$
Verdad	Clase $C_1$	$TN_l$	...	$TN_l$	$FP_l$
	Clase $C_2$	$TN_l$	$TN_l$	...	$FP_l$
	:	:	:	:	:
	Clase $C_l$	$FN_l$	...	$FN_l$	$TP_l$

### 2.3.2 Precision

La *Precision* por clase  $l$  mide la proporción de muestras clasificadas como positivas que realmente pertenecen a la clase  $l$ :

En términos de la Matriz de Confusión:

$$\text{Precision}_l = \frac{TP_l}{TP_l + FP_l}$$

### 2.3.3 Recall

El *Recall* por clase  $l$  mide la proporción de muestras verdaderamente positivas de la clase  $l$  que fueron correctamente identificadas:

En términos de la Matriz de Confusión:

$$\text{Recall}_l = \frac{TP_l}{TP_l + FN_l}$$

### 2.3.4 F-measure

El *F-measure* es la media armónica ponderada de *precision* y *recall*. La versión más común es el **F1-score**, donde el parámetro de ponderación  $\beta$  es igual a 1. Un clasificador perfecto tiene un valor  $F1 = 1$ .

Fórmula del F1-score ( $\beta = 1$ ) en términos de Precision y Recall:

$$F1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

El valor global ponderado se obtiene aplicando la fórmula de  $M_{\text{weighted}}$  sobre los  $F_{1,l}$ :

$$F_{1,\text{weighted}} = \sum_{l \in L} w_l F_{1,l}, \quad \text{con } w_l = \frac{n_l}{\sum_{i \in L} n_i}$$

### 2.3.5 Área Bajo la Curva ROC (ROC AUC)

Para extender la métrica ROC AUC a clasificación multiclas se emplea el enfoque **One-vs-Rest (OVR)**:

- Para cada clase  $l$ , se considera la clase  $l$  como positiva y el resto como negativas.
- Se calcula el AUC correspondiente ( $AUC_l$ ) sobre la curva ROC de esa clasificación binaria.
- Finalmente, se obtiene un promedio ponderado por el soporte de cada clase:

$$AUC_{\text{OVR, weighted}} = \sum_{l \in L} w_l AUC_l$$

donde:

$$w_l = \frac{n_l}{\sum_{i \in L} n_i}$$

### 2.3. MÉTRICAS PARA CASO MULTITASKER 2. MÉTRICAS DE RENDIMIENTO UTILIZADAS

## Chapter 3

# Descripción de los Métodos Utilizados

### 3.1 Regresión Logística

El modelo de Regresión Logística (LR, por su equivalente en inglés *Logistic Regression*).....

Se consideraron los siguientes valores para los hiperparámetros....

El mejor rendimiento se obtuvo con... Notamos este modelo como  $RL^{opt}$ .

$CC$



# Chapter 4

## Resultados

Mostrar los resultados obtenidos utilizando gráficos, tablas, figuras, etc

### 4.1 Introducción

Primera aproximación a resultados.

### 4.2 Métricas de Evaluación

A continuación se muestran las mejores métricas obtenidas, ademas del grid utilizado.

#### 4.2.1 Dataset Binario

#### 4.2.2 Regresión Logística

- **Precisión (Acc):** 0.84
- **F1 Score:** 0.84
- **ROC AUC:** 0.90

**Grid de Hiperparámetros:**

- $C = [0, 0.1, 0.01]$
- **Penalty:** None, l1, l2, elasticnet
- **Solver:** lbfgs, saga, newton-s
- **Multiclass:** ovr, multinomial

**Mejor Configuración:**

- $C = 1$ , Penalty = l1, Solver = lbfgs, saga, [*Multiclass = ovr, multinomial*]

#### 4.2.3 Máquinas de Soporte Vectorial (SVM)

- **Precisión (Acc):** 0.861
- **F1 Score:** 0.86
- **Recall:** 0.86

**Grid de Hiperparámetros:**

- $C = [0.001, 0.01, 0.1, 1, 10, 15, 20, 25]$
- **Kernel:** [*linear, poly, rb*", *sigmoid*]
- **Gamma:** [*scale, auto*, 0.001, 0.01, 0.1, 1]
- **Degree:** [2, 3, 4, 5, 6, 7, 8, 9, 10]

**Mejor Configuración:**

- $C = 1$ , kernel = rbf, gamma = scale

#### 4.2.4 Naive Bayes Gaussiano

- **Precisión (Acc):** 0.84
- **F1 Score:** 0.84

**Grid de Hiperparámetros:**

- **Suavizado:** Cualquier suavizado

**Mejor Configuración:**

- Suavizado: Cualquiera

#### 4.2.5 Random Forest

- **Precisión (Acc):** 0.878
- **F1 Score:** 0.877
- **ROC AUC:** 0.92

**Grid de Hiperparámetros:**

- **Criterion:** [*gini, entropy*]
- **Max Depth:** [*None, 3, 5, 7, 9*]
- **Min Samples Split:** [2, 5, 10]
- **Min Samples Leaf:** [1, 2, 4]
- **Max Features:** [*None, sqrt, log2*]

**Mejor Configuración:**

- Criterion: entropy, Maxdepth = 7, min samples split = 5, min samples leaf 1, max features = sqrt, log2.

### 4.3 Importancia de las Características

La importancia de las características de las mejores configuraciones.

### 4.3.1 Random Forest

Característica	Importancia
ST_Slope	0.254265
ChestPainType	0.127319
Oldpeak	0.113156
ExerciseAngina	0.105952
Cholesterol	0.099872
MaxHR	0.088635
Age	0.065807
RestingBP	0.055053
Sex	0.040916
FastingBS	0.030069
RestingECG	0.018956

### 4.3.2 Regresión Logística

Característica	Importancia (Permutación)
Oldpeak	0.045643
ChestPainType	0.036383
MaxHR	0.030174
Cholesterol	0.026797
ST_Slope	0.026580
ExerciseAngina	0.013181
Age	0.008279
Sex	0.002941
RestingECG	0.002179
RestingBP	0.001634
FastingBS	0.001525

### 4.3.3 SVM

Característica	Importancia (Permutación)
MaxHR	0.103704
Cholesterol	0.070915
Age	0.007081
RestingBP	0.002723
Oldpeak	0.000871
ChestPainType	0.000218
Sex	0.000000
RestingECG	0.000000
FastingBS	0.000000
ExerciseAngina	0.000000
ST_Slope	-0.000218

### 4.3.4 Naive Bayes Gaussiano

Característica	Importancia (Permutación)
ST_Slope	0.027015
ExerciseAngina	0.023747
Oldpeak	0.018736
ChestPainType	0.018519
Cholesterol	0.014815
Sex	0.014270
FastingBS	0.004575
RestingBP	0.001852
MaxHR	-0.000218
RestingECG	-0.001198
Age	-0.003595

#### 4.3.5 Importancia de las Características (Coeficientes Absolutos de Regresión Logística)

Característica	Importancia (Coeficientes)
Oldpeak	0.399451
ChestPainType	0.397051
ST_Slope	0.386263
ExerciseAngina	0.268956
Sex	0.150576
FastingBS	0.119947
RestingECG	0.034868
Age	0.024577
MaxHR	0.019045
RestingBP	0.007427
Cholesterol	0.003631

#### 4.3.6 Dataset Multiclasificación

#### 4.3.7 Regresión Logística

- Precisión (Acc): 0.897

- F1 Score: 0.0.8956

Grid de Hiperparámetros:

- $C = [0, 0.1, 0.01]$
- **Penalty:** None, l1, l2, elasticnet
- **Solver:** lbfgs, saga, newton-s
- **Multiclass:** ovr, multinomial

Mejor Configuración:

- $C = [0.01, 0.1, 10]$ , Penalty = [None, l2( $C = 10$ ), elasticnet( $C = 10$ )] , Solver = [lbfgs, saga, newton-s] , [Multiclass = ovr]

### 4.3.8 Máquinas de Soporte Vectorial (SVM)

- **Precisión (Acc):** 0.0.861
- **F1 Score:** 0.86

**Grid de Hiperparámetros:**

- **C:** [0.001, 0.01, 0.1, 1, 10, 15, 20, 25]
- **Kernel:** [*linear, poly, rb*"', *sigmoid*]
- **Gamma:** [*scale, auto*, 0.001, 0.01, 0.1, 1]
- **Degree:** [2, 3, 4, 5, 6, 7, 8, 9, 10]

**Mejor Configuración:**

- $C = 1$ , kernel = rbf, gamma = scale

### 4.3.9 Naive Bayes Gaussiano

- **Precisión (Acc):** 0.822
- **F1 Score:** 0.83

**Grid de Hiperparámetros:**

- **Suavizado:** Cualquier suavizado

**Mejor Configuración:**

- Suavizado: Cualquiera

### 4.3.10 Random Forest

- **Precisión (Acc):** 0.943
- **F1 Score:** 0.94

**Grid de Hiperparámetros:**

- **Criterion:** [*gini, entropy*]
- **Max Depth:** [*None, 3, 5, 7, 9*]
- **Min Samples Split:** [2, 5, 10]
- **Min Samples Leaf:** [1, 2, 4]
- **Max Features:** [*None, sqrt, log2*]

**Mejor Configuración:**

- Criterion: [*entropy, gini*], Maxdepth = *None*, min samples split = [2, 5], min samples leaf 1, max features = [*sqrt, log2*].

## 4.4 Importancia de las Características

La importancia de las características de las mejores configuraciones.

#### 4.4.1 Random Forest

Característica	Importancia
ASTV	0.139807
ALTV	0.109941
MSTV	0.104823
Mean	0.091579
AC	0.063645
Mode	0.061986
Median	0.060633
DP	0.047945
LB	0.045324
MLTV	0.045132
Variance	0.040531
UC	0.039166
Width	0.030551
Min	0.030109
Max	0.027147
FM	0.020801
Nmax	0.018407
DL	0.011128
Tendency	0.007652
Nzeros	0.003405
DS	0.000287

Atributos que mejoran accuracy: [][ASTV, ALTV, MSTV, Mean, AC, Mode, Median, DP, LB, Variance]

#### 4.4.2 Regresión Logística

Característica	Importancia (Permutación)
Mean	0.098487
AC	0.084113
ASTV	0.057069
Median	0.031631
DP	0.029740
LB	0.023404
Variance	0.022270
UC	0.022080
ALTV	0.019243
Max	0.018109
Nmax	0.014374
Mode	0.011348
Min	0.005910
MSTV	0.004208
FM	0.003830
Tendency	0.003546
MLTV	0.002979
Nzeros	0.002837

DL	0.001655
Width	0.000189
DS	0.000000

#### 4.4.3 SVM

Característica	Importancia (Permutación)
ASTV	0.050355
ALTV	0.037069
UC	0.030638
AC	0.026903
DP	0.018345
Mean	0.015887
Mode	0.014988
Median	0.014043
Nmax	0.011915
MSTV	0.009125
DL	0.005059
Variance	0.004775
Nzeros	0.004586
Min	0.004444
Max	0.004350
MLTV	0.003357
Tendency	0.003026
FM	0.002459
Width	0.001418
DS	0.000000
LB	-0.000804

#### 4.4.4 Naive Bayes Gaussiano

Característica	Importancia (Permutación)
AC	0.057163
DP	0.018676
ALTV	0.015461
ASTV	0.005106
DS	0.002695
UC	0.002364
FM	0.001371
Variance	0.001087
Nzeros	0.001040
Nmax	-0.000993
Tendency	-0.001040
Mode	-0.001324
Max	-0.001371
Min	-0.001986
LB	-0.001986

MLTV	-0.002222
Width	-0.002459
Median	-0.003310
MSTV	-0.004965
Mean	-0.005768
DL	-0.006809

#### 4.4.5 Importancia de las Características (Coeficientes Absolutos de Regresión Logística)

Característica	Importancia (Coeficientes)
AC	3.619754
Mean	2.520677
LB	1.134264
ASTV	0.867591
Variance	0.547877
Nmax	0.522037
UC	0.514121
Max	0.468159
DP	0.309519
Min	0.245054
MSTV	0.233617
Mode	0.216870
Median	0.143711
MLTV	0.121284
Nzeros	0.103791
DL	0.076451
Tendency	0.071444
Width	0.029727
ALTV	0.025537
FM	0.024773
DS	0.000035

## **Chapter 5**

# **Conclusiones**

Explicar que aprendimos con la realización de este trabajo. Qué nos muestran los resultados.



# Bibliography