

Comparación de Técnicas de Aprendizaje Automático Supervisado Aplicados a Datos Cardiólogos

Autor: Nicolás Seivane

Tutora: Andrea Rey

Fecha

Universidad Nacional de Hurlingham

Índice General

1	Introducción	11
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 Multiclase: Cardiotocografía Predicción	13
2	Métricas de Rendimiento Utilizadas	17
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	Área Bajo la Curva ROC (<i>ROC AUC</i>)	19
2.3	Métricas para caso Multiclase	19
2.3.1	Matriz de Confusión (Multiclase)	20
2.3.2	<i>Precision</i>	20
2.3.3	<i>Recall</i>	20
2.3.4	<i>F-measure</i>	20
2.3.5	Área Bajo la Curva ROC (<i>ROC AUC</i>)	21
3	Descripción de los Métodos Utilizados	23
3.1	Regresión Logística	23
3.1.1	Función de Probabilidad	23
3.1.2	Función Logit	23
3.1.3	Estimación por Máxima Verosimilitud	24
3.1.4	Hiperparámetros	24
3.2	Árboles de Decisión	25
3.2.1	Conceptos Fundamentales	25
3.2.2	Bosques Aleatorios (Random Forest)	26
3.2.3	Hiperparámetros	26
3.3	Clasificador Naïve Bayes	26
3.3.1	Caso Continuo (Naïve Bayes Gaussiano)	27
3.4	Máquinas de Soporte Vectorial (SVM)	27
3.4.1	Margen Rígido (Hard Margin)	27
3.4.2	Margen Suave (Soft Margin)	27

3.4.3	Formulación Dual y Kernel Trick	27
3.4.4	Hiperparámetros	28
4	Resultados	29
4.1	Introducción	29
4.1.1	Dataset Binario	29
4.1.2	Regresión Logística	29
4.1.3	Máquinas de Soporte Vectorial (SVM)	29
4.1.4	Naive Bayes Gaussiano	30
4.1.5	Random Forest	30
4.2	Importancia de las Características	30
4.2.1	Random Forest	31
4.2.2	Regresión Logística	31
4.2.3	SVM	31
4.2.4	Naive Bayes Gaussiano	31
4.2.5	Importancia de las Características (Coeficientes Absolutos de Regresión Logística)	32
4.2.6	Dataset Multiclase	32
4.2.7	Regresión Logística	32
4.2.8	Máquinas de Soporte Vectorial (SVM)	33
4.2.9	Naive Bayes Gaussiano	33
4.2.10	Random Forest	33
4.3	Importancia de las Características	33
4.3.1	Random Forest	34
4.3.2	Regresión Logística	34
4.3.3	SVM	35
4.3.4	Naive Bayes Gaussiano	35
4.3.5	Importancia de las Características (Coeficientes Absolutos de Regresión Logística)	36
5	Conclusiones	37

Índice de Figuras

Índice de Tablas

1.1	Tipo de atributo del conjunto Binario.	12
1.2	Tipo de atributo del conjunto Multiclase.	14
2.1	Matriz de Confusion	18
2.2	Matriz Confusion Multiclase	20

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

Las enfermedades cardiovasculares son la causa número uno de muerte globalmente, con un estimado de 17.9 millones de vidas cada año, aproximadamente 31% de todas las muertes globales. La idea central de este trabajo es encontrar la técnica de aprendizaje automático mas óptima para poder realizar predicciones de si un paciente tiene altas probabilidades de tener insuficiencia cardíaca.

Este dataset [1], llamado *HeartFailure* fue creado mediante la combinación de cinco datasets independientes en 11 atributos comunes, logrando el dataset más grande de información de enfermedades cardiovasculares utilizado para investigación. Los cinco datasets utilizados son:

- Cleveland: 303 observaciones
- Hungarian: 294 observaciones
- Switzerland: 123 observaciones
- Long Beach VA: 200 observaciones
- Stalog (Heart) Data Set: 270 observaciones

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	-

Tabla 1.1: Tipo de atributo del conjunto Binario.

La cantidad de registros y los tipos de atributos utilizados para este trabajo fueron los siguientes, considerando que estos registros ya fueron previamente procesados para poder ser utilizados en las técnicas de aprendizaje automático. Se informara la distribución de los valores si el atributo es categórico, en caso de ser numérico, se informara la media de los valores, junto al valor máximo y mínimo que poseen. En la Tabla 1.1 se muestran que tipo de datos son los atributos del dataset que serán utilizados en este trabajo.

Cantidad de registros: 918

Cantidad de atributos: 11

Atributos Categóricos: 5

Atributos Numéricos: 6

Descripción atributos:

A continuación se realizara una pequeña descripción de cada atributo anteriormente señalado. Si el atributo es categórico, se informara la distribución de los valores que posee dicho atributo y en caso de resultar ser un atributo numérico, se informara la media de los valores de dicho atributo, junto al valor máximo y mínimo que posee. Se realiza lo anterior para contextualizar los atributos y ver los rangos de valores con que se trabajara.

Age: Este atributo refiera a la 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: Refiere al Sexo de los pacientes; hay una distribución 78.98% M (masculinos) y hay 21.02% F (femeninos)

ChestPainType: Tipo del dolor en el pecho, del cual hay varias clasificaciones; Tiene una distribución 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: Se esta describiendo la Presión sanguínea en reposo, donde hay una distribución de 51.09% de mujeres, codificadas en 1 y 48.91% de hombres, codificados en 0.

Cholesterol: Este atributo es el 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: Es la 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: Son los 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: Este atributo es el Máximo ritmo cardíaco registrado, tiene media: 136.79, valor máximo: 202.00 y valor mínimo: 60.00

ExerciseAngina: Es la 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 máximo: 6.20 y valor mínimo: -0.10

ST_Slope: The slope of the peak exercise ST segment; hay 43.08% Up, 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 Multiclase: Cardiotocografía Predicción

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.

En el dataset utilizado 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).

La cantidad de registros y los tipos de atributos utilizados para este trabajo fueron los siguientes, considerando que estos registros ya fueron previamente procesados para poder ser utilizados en las técnicas de aprendizaje automático. En la Tabla 1.2 se muestran los tipos de las variables utilizadas en este trabajo.

Cantidad de registros: 2115

Cantidad de atributos: 21

Atributos Categóricos: 0

Atributos Numéricos: 21

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	Catégorico (string)

Tabla 1.2: Tipo de atributo del conjunto Multiclae.

Descripción atributos:

A continuación se realizara una pequeña descripción de cada atributo anteriormente señalado. Si el atributo es catégorico, se informara la distribución de los valores que posee dicho atributo y en caso de resultar ser un atributo numérico, se informara la media de los valores de dicho atributo, junto al valor máximo y mínimo que posee. Se realiza lo anterior para contextualizar los atributos y ver los rangos de valores con que se trabajará.

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 Aprendizaje Automático. Para alcanzar este objetivo, se utilizarán **métricas de rendimiento** que permiten cuantificar la capacidad del algoritmo de clasificar, ergo son herramienta que utilizamos para saber que tan bien predice un modelo una clase o que tan bien puede clasificar entre varias clases.

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). Esto se debe a que es perfectamente normal y sumamente esperable que el modelo funcione correctamente con el conjunto de datos que se utiliza para el entrenamiento del modelo, la idea fundamental es poder tener el mismo rendimiento o incluso uno mejor que con el conjunto de 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. Es importante señalar que un grupo k_i es utilizado solo una vez para entrenar, el resto de veces será utilizado como parte del conjunto de prueba. El valor final estimado de la métrica, denotado por \widehat{M} , es el promedio de los valores obtenidos de cada grupo, es decir,

$$\widehat{M} = \frac{1}{k} \sum_{i=1}^k M_i, \quad (2.1)$$

donde M_i es el valor de la métrica de evaluación obtenido en el i -ésimo grupo utilizado como conjunto de prueba, para $i = 1, 2, \dots, k$.

Dentro de este trabajo no sólo se evaluarán distintos modelos, sino que se utilizarán distintos *datasets* para lograrlos. A continuación se señalarán las métricas que se utilizarán en cada caso y se podrán apreciar algunas diferencias, leves, pero diferencias en sí.

2.2 Métricas para caso Binario

2.2.1 Matriz de Confusión

Una matriz de confusión, que se puede observar en la Tabla 2.1, es una forma simple de saber de que forma está clasificando el algoritmo, donde una clase es considerada **positiva** P y la otra **negativa** N . La matriz de confusión clasifica las predicciones en:

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

Tabla 2.1: Matriz de Confusion

- **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 negativos.

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, ergo se puede obtener un *Accuracy* alto si predice una clase muy bien, que tiene una distribución mucho mayor que la otra, mientras que la de menor distribución casi no la predice. En términos de la Matriz de Confusión la formula seria la siguiente:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} = \frac{TP + TN}{\text{Total}}, \quad (2.2)$$

y en términos del conjunto de predicciones y valores verdaderos, se tiene que n_{samples} : representa la cantidad total de ejemplos en la muestra, mientras que \hat{y}_i es el valor predicho del i -ésimo ejemplo, e y_i es el valor verdadero correspondiente:

$$\text{accuracy}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} 1(\hat{y}_i = y_i), \quad (2.3)$$

por lo tanto, se puede simplificar como la siguiente formula:

$$\text{Accuracy} = \frac{\text{Número de predicciones correctas}}{\text{Número total de muestras}} \quad (2.4)$$

2.2.3 Precision

El *Precision* mide la probabilidad de que la predicción positiva del clasificador sea correcta, en otras palabras, mide que tan bien predice las clases positivas el modelo. En términos de la Matriz de Confusión, se puede expresar lo anterior de la siguiente manera;

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2.5)$$

2.2.4 Recall

El *Recall* o también conocido como Sensibilidad o Tasa de Verdaderos Positivos (TPR). Mide la probabilidad de que el clasificador detecte un caso positivo cuando en verdad lo es. En términos de la Matriz de Confusión se puede entender a la *Recall* de la siguiente manera:

$$\text{Recall} = TPR = \frac{TP}{TP + FN} = \frac{TP}{P} \quad (2.6)$$

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 β es igual a 1. Un clasificador perfecto tiene un valor $F1 = 1$. Fórmula General (F_β):

$$F_\beta = \frac{(1 + \beta^2)\text{precision} \times \text{recall}}{\beta^2\text{precision} + \text{recall}}, \quad (2.7)$$

donde la fórmula del F1-score ($\beta = 1$) en términos de Precision y Recall se puede notar de la siguiente manera:

$$F1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}, \quad (2.8)$$

o en términos de la Matriz de Confusión, de forma más simplificada:

$$F1 = \frac{2TP}{2TP + FP + FN} \quad (2.9)$$

2.2.6 *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, una métrica muy útil para comparar el desempeño entre modelos distintos o entre un mismo modelo con hipermetamorfosis distintos. **La Curva ROC** es un gráfico que ilustra el rendimiento de un clasificador binario a media 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. El **AUC** mide justamente el área debajo de la Curva ROC.

Ejes utilizados para el gráfico:

- Eje Y: TPR
- Eje X: FPR

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 Multiclase

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

La métrica ponderada por la presencia de la clase, M_{weighted} , se calcula como el promedio de la métrica por clase M_l , donde cada contribución es ponderada por el tamaño de la clase $|y_l|$, siendo L es el conjunto de etiquetas o clases. Donde $\hat{M}_{\text{weighted}}$ es el valor estimado de la métrica promedio ponderada.

$$\hat{M}_{\text{weighted}} = \frac{1}{\sum_{l \in L} |y_l|} \sum_{l \in L} |y_l| \cdot M_l \quad (2.10)$$

		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

Tabla 2.2: Matriz Confusion Multiclase

2.3.1 Matriz de Confusión (Multiclase)

La matriz de confusión multiclase, que se puede ver en la Tabla 2.2, 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 que antes habíamos utilizados para la matriz de confusión del caso binario:

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

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 más simple, utilizando la Matriz de Confusión:

$$\text{Precision}_l = \frac{TP_l}{TP_l + FP_l} \quad (2.11)$$

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 más simple, utilizando la Matriz de Confusión:

$$\text{Recall}_l = \frac{TP_l}{TP_l + FN_l} \quad (2.12)$$

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 β es igual a 1. Un clasificador perfecto tiene un valor $F1 = 1$.

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

$$F_{1,\text{weighted}} = \sum_{l \in L} w_l F_{1,l}, \quad \text{con } w_l = \frac{n_l}{\sum_{i \in L} n_i} \quad (2.13)$$

2.3.5 Área Bajo la Curva ROC (ROC AUC)

Para extender la métrica ROC AUC a clasificación multiclase se emplea el enfoque **One-vs-Rest (OVR)**. Para cada clase l , se considera la clase l como positiva y el resto como negativas, luego 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:

$$AUC_{\text{OVR, weighted}} = \sum_{l \in L} w_l AUC_l, \quad (2.14)$$

Chapter 3

Descripción de los Métodos Utilizados

3.1 Regresión Logística

El modelo de **Regresión Logística** (LR, por su equivalente en inglés *Logistic Regression*) es una técnica del análisis de datos utilizada para establecer relaciones entre las variables predictoras y la clase a la cual pertenece cada registro. Posteriormente, el modelo permite predecir la probabilidad de que un nuevo registro pertenezca a una clase determinada. Este tipo de modelo de regresión es justamente utilizado para los problemas no linealmente separables, como la mayoría de problemas.

A diferencia de la regresión lineal múltiple, la regresión logística predice una probabilidad (valor entre 0 y 1). Ambos modelos son lineales en sus parámetros, pero difieren en la naturaleza de la variable dependiente. El objetivo es estimar los coeficientes de regresión que maximizan la verosimilitud de los datos observados, o encontrar los coeficientes que mejor funcionan para los datos de entrenamiento.

El modelo busca modelar la probabilidad condicional de que una observación pertenezca a la clase objetivo y_i , siendo

$$P(y_i = 1 \mid \mathbf{x}_i), \quad (3.1)$$

donde $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ik})$ es el vector de características de la observación i , el cual también llamamos registro.

3.1.1 Función de Probabilidad

La función logística define la probabilidad de pertenencia de x_i a la clase 1 como:

$$p(\mathbf{x}_i) = P(y_i = 1 \mid \mathbf{x}_i) = \frac{\exp(\beta_0 + \sum_{j=1}^k \beta_j x_{ij})}{1 + \exp(\beta_0 + \sum_{j=1}^k \beta_j x_{ij})}, \quad (3.2)$$

donde β_0 es el intercepto y β_j los coeficientes asociados a cada predictor.

3.1.2 Función Logit

La función inversa de la función logística, denominada *logit*, relaciona el logaritmo de las *odds* con un modelo lineal, siendo *odds* lo que se utiliza para analizar si la probabilidad de ocurrencia de un evento -caso/no caso- difiere o no en distintos grupos,

$$\text{logit}(p(\mathbf{x}_i)) = \ln \left(\frac{p(\mathbf{x}_i)}{1 - p(\mathbf{x}_i)} \right) = \beta_0 + \sum_{j=1}^k \beta_j x_{ij} \quad (3.3)$$

Esta transformación asegura que las probabilidades estén acotadas entre 0 y 1, mientras que la combinación lineal de predictores puede tomar cualquier valor real, siendo este ultimo punto un factor que realiza el calculo de coeficientes muy difícil.

3.1.3 Estimación por Máxima Verosimilitud

Los coeficientes de regresión se estiman mediante el método de **Máxima Verosimilitud** (MLE), donde la función de log-verosimilitud a maximizar

$$\ell(\beta_0, \beta_1, \dots, \beta_k) = \sum_{i=1}^n [y_i \ln(p(\mathbf{x}_i)) + (1 - y_i) \ln(1 - p(\mathbf{x}_i))], \quad (3.4)$$

en donde la solución analítica no existe, por lo que se utilizan métodos numéricos iterativos para obtener los parámetros óptimos.

3.1.4 Hiperparámetros

- **Parámetro de Regularización (C):** Controla la complejidad del modelo. Valores pequeños de C implican mayor regularización (menor sobreajuste), mientras que valores grandes permiten mayor flexibilidad del modelo.
- **Penalización (*penalty*):** Es un término regulador (Ω) que se suma a la función de coste original (J) para formar una función ajustada \tilde{J} . Controla la capacidad del modelo y reduce el error de generalización.

- Regularización L1 (Lasso).

$$\tilde{J}_{L1}(\beta) = J(\beta) + \alpha \sum_{j=1}^k |\beta_j| \quad (3.5)$$

- Regularización L2 (Ridge).

$$\tilde{J}_{L2}(\beta) = J(\beta) + \frac{1}{2} \alpha \sum_{j=1}^k \beta_j^2 \quad (3.6)$$

- Regularización Elastic Net

$$\tilde{J}_{EN}(\beta) = J(\beta) + \alpha \left[\rho \sum_{j=1}^k |\beta_j| + \frac{1-\rho}{2} \sum_{j=1}^k \beta_j^2 \right] \quad (3.7)$$

- **Algoritmo de Optimización (*solver*):** El *solver* es el algoritmo numérico encargado de minimizar la función de coste regularizada $\tilde{J}(\beta)$.

- Método Newton-CG (basado en segunda derivada).

$$\beta^{(t+1)} \leftarrow \beta^{(t)} - \mathbf{H}^{-1} \nabla_{\beta} \tilde{J}(\beta^{(t)}) \quad (3.8)$$

- Método BFGS (quasi-Newton).

$$\beta^{(t+1)} \leftarrow \beta^{(t)} - \mathbf{B}^{-1} \nabla_{\beta} \tilde{J}(\beta^{(t)}) \quad (3.9)$$

- **Estrategia Multiclase (*multi_class*):** La regresión logística está diseñada originalmente para clasificación binaria. Para extenderla a múltiples clases se emplean estrategias como:

- *one-vs-rest* (OvR): Entrena un clasificador por clase.
- *multinomial*: Optimiza una única función de verosimilitud multinomial conjunta.

3.2 Árboles de Decisión

El aprendizaje mediante **Árboles de Decisión** (RF, por su equivalente en inglés *Random Forest*) es un método no paramétrico que utiliza divisiones jerárquicas sobre los atributos de los datos, construyendo reglas de decisión del tipo *if-else* para predecir el valor de una variable objetivo. El objetivo principal es encontrar las divisiones (particiones) que maximicen la pureza de los nodos hijos, es decir, que minimicen la impureza del nodo resultante. Es un método mucho más sencillo de realizar, donde se crea un camino de decisión, por lo cual según los valores de los atributos de un registro podemos predecir que a que clase pertenecerían, donde el cómputo pesado se encuentra en la creación del propio camino.

3.2.1 Conceptos Fundamentales

La probabilidad de que un ejemplo en el nodo t pertenezca a la clase C_k se define como

$$p(k|t) = \frac{N_k(t)}{N(t)}, \quad (3.10)$$

donde $N(t)$ es la cantidad total de ejemplos en el nodo t , y $N_k(t)$ la cantidad de ejemplos de la clase C_k . Es importante este fenómeno porque es un costo computacional muy barato el calcular esta probabilidad y se asemejan a las probabilidades equiprobables, donde la probabilidad de una clase en un nodo esta dada por la cantidad de clases que tienen en el mismo.

Impureza del Nodo

La impureza de un nodo t se mide mediante una función ϕ que depende de las probabilidades de clase en dicho nodo, donde la función de impureza i es un hiperparametro en sí mismo.

$$i(t) = \phi(p(1|t), p(2|t), \dots, p(K|t)). \quad (3.11)$$

La impureza es máxima cuando las clases están perfectamente mezcladas y mínima (cero) cuando el nodo contiene solo una clase, en donde la impureza máxima seria que la probabilidad de cada clase sea aleatoria, ya que están perfectamente mezcladas.

Entropía de Shannon de un conjunto de datos D

$$H(D) = - \sum_{k=1}^K \frac{N_k(D)}{N(D)} \log_2 \left(\frac{N_k(D)}{N(D)} \right) \quad (3.12)$$

Índice de Gini de un nodo t

$$\text{Gini}(t) = 1 - \sum_{k=1}^K [p(k|t)]^2 = 1 - \sum_{k=1}^K \left(\frac{N_k(t)}{N(t)} \right)^2 \quad (3.13)$$

Disminución de Impureza

La reducción de impureza generada al dividir el nodo t en dos nodos hijos t_1 y t_2 mediante una partición s se calcula como

$$\Delta i(s, t) = i(t) - q_1 i(t_1) - q_2 i(t_2), \quad (3.14)$$

donde $q_j = \frac{N(t_j)}{N(t)}$ para $j = 1, 2$.

3.2.2 Bosques Aleatorios (Random Forest)

El algoritmo de **Random Forest** (RF) combina múltiples árboles de decisión independientes contruidos sobre subconjuntos aleatorios de los datos (muestreo con reemplazo o *bootstrap*). Cada árbol se entrena sobre un subconjunto de atributos aleatorios en cada división, lo que introduce diversidad y reduce la varianza.

La predicción final para clasificación se obtiene mediante el voto mayoritario de los árboles

$$\hat{y} = \operatorname{argmax}_{c \in \mathcal{C}} \sum_{m=1}^M \mathbb{I}(h_m(\mathbf{x}) = c), \quad (3.15)$$

donde $h_m(\mathbf{x})$ es la predicción del árbol m .

3.2.3 Hiperparámetros

- **Criterio de Partición:** Función de impureza utilizada (e.g., Índice de Gini o Entropía de Shannon).
- **Algoritmo de Construcción:** *ID3* emplea la ganancia de información (entropía), mientras que *CART* utiliza el índice de Gini y genera árboles binarios.
- **Número de Atributos Muestreados (RF):** En Random Forest, típicamente se seleccionan \sqrt{a} o $\ln(a)$ atributos por partición, donde a es la cantidad total de atributos.
- **Número de Árboles (RF):** Cantidad de árboles a construir en el bosque.

3.3 Clasificador Naïve Bayes

El **Clasificador Naïve Bayes** (CNB) es un método supervisado probabilístico basado en el *Teorema de Bayes*, que asume independencia condicional entre los atributos dado la clase, lo cual sabemos que no es posible que estas mismas variables sean independientes entre sí. La regla de clasificación, o la probabilidad de que se de una clase dado un registro,

$$P(Y = C_k | X_1 = x_1, \dots, X_d = x_d) = \frac{P(X_1 = x_1, \dots, X_d = x_d | Y = C_k) P(Y = C_k)}{P(X_1 = x_1, \dots, X_d = x_d)} \quad (3.16)$$

Probabilidad Condicional: Bajo el supuesto de independencia condicional, la probabilidad condicional puede expresarse como

$$P(X_1 = x_1, \dots, X_d = x_d | Y = C_k) = \prod_{j=1}^d P(X_j = x_j | Y = C_k), \quad (3.17)$$

en donde puede ser expresado para una clase k cuando se presenta un registro j

$$\hat{\theta}_{jmk} = P(X_j = x_{jm} | Y = C_k) = \frac{\#\{X_j = x_{jm} \wedge Y = C_k\}}{\#\{Y = C_k\}} \quad (3.18)$$

Probabilidad a Priori: Otra estimación fundamental a calcular y expresar es la a priori, en la cual se asemeja a la probabilidad de que un nodo pertenezca a una clase k en *Random Forest*

$$\hat{\pi}_k = P(Y = C_k) = \frac{\#\{Y = C_k\}}{N} \quad (3.19)$$

3.3.1 Caso Continuo (Naïve Bayes Gaussiano)

Cuando los atributos son continuos y se asume distribución normal, las verosimilitudes se estiman con la función de densidad Gaussiana:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp \left[-\frac{1}{2} \left(\frac{x - \mu}{\sigma} \right)^2 \right], \quad (3.20)$$

por lo cual la decisión final se obtiene como

$$\hat{Y} = \underset{C_k}{\operatorname{argmax}} \left[\log P(Y = C_k) + \sum_{j=1}^d \log f(x_j \mid \mu_{jk}, \sigma_{jk}^2) \right] \quad (3.21)$$

3.4 Máquinas de Soporte Vectorial (SVM)

Las **Máquinas de Soporte Vectorial** (SVM) buscan encontrar el hiperplano que mejor separa las clases, maximizando el margen M (la distancia mínima entre el hiperplano y las observaciones más cercanas, llamadas vectores de soporte). En donde estos vectores que llamamos soportes, se buscan que sean de distantes clases y tengan la distancia más cercanas posibles.

3.4.1 Margen Rígido (Hard Margin)

$$\text{Minimizar } \frac{1}{2} \|\mathbf{w}\|^2 \quad \text{sujeto a } y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + \beta) \geq 1 \quad (3.22)$$

3.4.2 Margen Suave (Soft Margin)

$$\text{Minimizar } \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^n \xi_i \quad \text{sujeto a } \begin{cases} y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + \beta) \geq 1 - \xi_i \\ \xi_i \geq 0 \end{cases} \quad (3.23)$$

3.4.3 Formulación Dual y Kernel Trick

$$\text{Maximizar } -\frac{1}{2} \sum_{i,\ell=1}^n \alpha_i \alpha_\ell y_i y_\ell K(\mathbf{x}_i, \mathbf{x}_\ell) + \sum_{i=1}^n \alpha_i \quad (3.24)$$

$$\text{sujeto a } 0 \leq \alpha_i \leq C, \quad \sum_{i=1}^n \alpha_i y_i = 0 \quad (3.25)$$

Funciones Kernel Comunes

Kernel Lineal:

$$K(\mathbf{a}, \mathbf{b}) = \langle \mathbf{a}, \mathbf{b} \rangle \quad (3.26)$$

Kernel Radial (RBF o Gaussiano):

$$K(\mathbf{a}, \mathbf{b}) = \exp(-\gamma \|\mathbf{a} - \mathbf{b}\|^2) \quad (3.27)$$

Kernel Polinómico:

$$K(\mathbf{a}, \mathbf{b}) = (\langle \mathbf{a}, \mathbf{b} \rangle + r)^d \quad (3.28)$$

3.4.4 Hiperparámetros

- **Parámetro de Regularización (C):** Controla el equilibrio entre la maximización del margen y la penalización de errores de clasificación.
- **Tipo de Kernel:** Lineal, Polinómico, Radial (RBF) o Sigmoido.
- **Parámetros del Kernel:** Por ejemplo, γ o σ^2 en RBF; grado d y coeficiente r en el kernel polinómico.

Chapter 4

Resultados

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

4.1 Introducción

A continuación se mostraran los resultados obtenidos a través de los métodos previamente señalados, se detallaron los hiperparametros con mejor resultado en cuanto a lo relacionado a las métricas utilizadas.

4.1.1 Dataset Binario

4.1.2 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*]

4.1.3 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

4.1.4 Naive Bayes Gaussiano

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

Grid de Hiperparámetros:

- **Suavizado:** Cualquier suavizado

Mejor Configuración:

- Suavizado: Cualquiera

4.1.5 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.

4.2 Importancia de las Características

La importancia de las características de las mejores configuraciones.

4.2.1 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

4.2.2 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

4.2.3 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

4.2.4 Naive Bayes Gaussian

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

4.2.5 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

4.2.6 Dataset Multiclase

4.2.7 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]

4.2.8 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

4.2.9 Naive Bayes Gaussiano

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

Grid de Hiperparámetros:

- **Suavizado:** Cualquier suavizado

Mejor Configuración:

- Suavizado: Cualquiera

4.2.10 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*].

4.3 Importancia de las Características

La importancia de las características de las mejores configuraciones.

4.3.1 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: $\llbracket [ASTV, ALTV, MSTV, Mean, AC, Mode, Median, DP, LB, Variance]$

4.3.2 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

4.3.3 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

4.3.4 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

4.3.5 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

- [1] fedesoriano. Heart failure prediction dataset. <https://www.kaggle.com/fedesoriano/heart-failure-prediction><https://www.kaggle.com/fedesoriano/heart-failure-prediction>, September 2021.