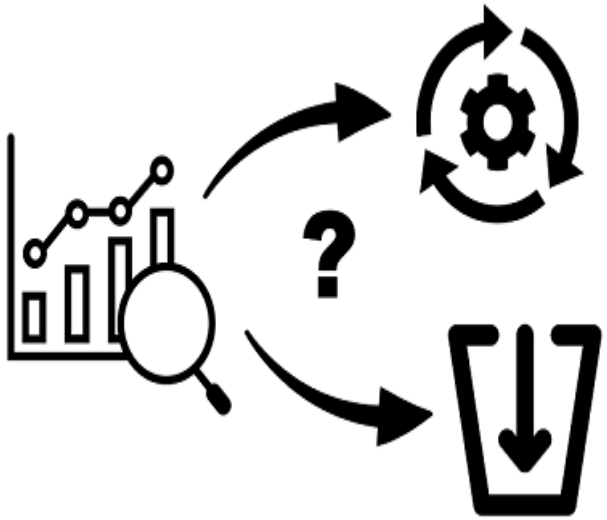
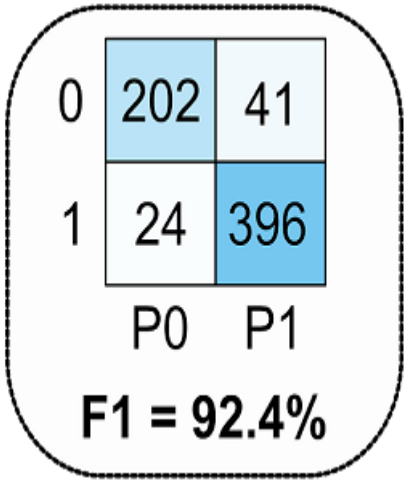




Fundamentos de Data Science

Unidad 4: Extracción de Conocimiento

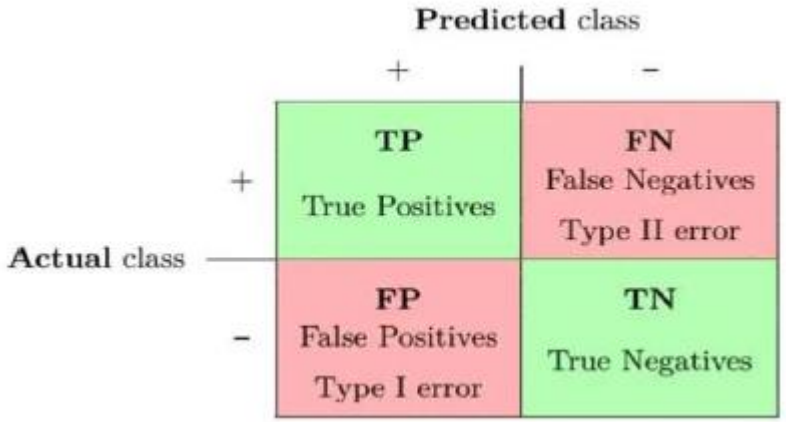
Semana 13 – La Modelización: Evaluación de Modelos Analíticos

Matriz de Confusión

Es una herramienta que permite la visualización del desempeño de un algoritmo que se emplea en el aprendizaje supervisado.

La Matriz de Confusión contiene información acerca de las predicciones realizadas por un **Método o Sistema de Clasificación**, comparando para el conjunto de individuos de la tabla de aprendizaje o de testing, la predicción dada versus la clase a la que estos realmente pertenecen.

La tabla muestra la matriz de confusión para un clasificador de dos clases:



Matriz de Confusión

La Accuracy es el número total de predicciones correctas dividido por el número total de predicciones.

La Precisión de una clase define cuan confiable es un modelo en responder si un punto pertenece a esa clase.

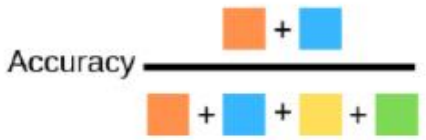
El Recall de una clase expresa cuan bien el modelo puede detectar a esa clase.

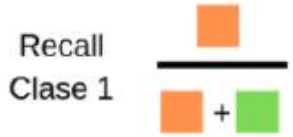
Falso Positivo es la proporción de casos negativos que fueron clasificados incorrectamente (Error Tipo I).

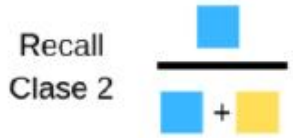
Falso Negativo es la proporción de casos positivos que fueron clasificados incorrectamente (Error Tipo II).

Matriz de Confusión

	Predicción Clase 1	Predicción Clase 2
Valor real Clase 1	Aciertos True Positive Clase 1	Fallos False Positive Clase 2
Valor real Clase 2	Fallos False Positive Clase 1	Aciertos True Positive Clase 2







Matriz de Confusión

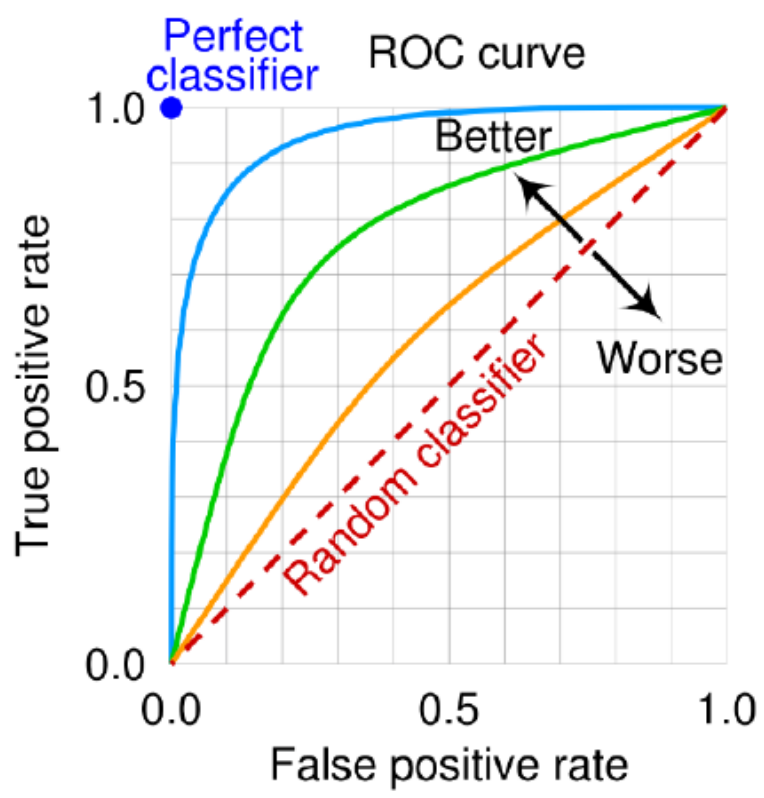
Tenemos cuatro casos posibles para cada clase:

- **Alta precision y alto recall**: el modelo maneja perfectamente esa clase
- **Alta precision y bajo recall**: el modelo no detecta la clase muy bien, pero cuando lo hace es altamente confiable.
- **Baja precisión y alto recall**: La clase detecta bien la clase pero también incluye muestras de otras clases.
- **Baja precisión y bajo recall:** El modelo no logra clasificar la clase correctamente.

Cuando tenemos un dataset con desequilibrio, suele ocurrir que obtenemos un **alto valor de precisión en la clase Mayoritaria y un bajo recall en la clase Minoritaria**

Curva de ROC

Es una representación gráfica que compara la tasa de falsos positivos con la de verdaderos positivos para distintos puntos de corte.



Un parámetro para evaluar la bondad de la prueba es el área bajo la curva que tomará valores entre 1 (prueba perfecta) y 0.5 (prueba inútil).