APRENDIZAJE AUTOMÁTICO CON PYTHON

TERCERA SESIÓN

TOULOUSE LAUTREC

TOULOUSE LAUTREC

EDUCACIÓN CONTINUA

ALEXANDER VALDEZ PORTOCARRERO

CONTENIDO

Introducción al Machine Learning

SESIÓN - 1

Introducción al Machine Learning

- Fundamentos de Machine Learning
- Tipos de Aprendizaje
- Librerías de Python para Machine Learning
- Algoritmos de Regresión
- Evaluación de Modelos de Regresión

Introducción al Machine Learning

SESIÓN - 2

Clasificación

- Algoritmos de Clasificación
- Balanceo de Datos
- Regresión Logística
- Evaluación de Modelos de Clasificación
- Árboles de Decisión

CONTENIDO

Introducción al Machine Learning

SESIÓN - 3

Clustering

- Algoritmos de Agrupación
- Clustering No Jerárquico: Kmeans, PAM, CLARA
- Clustering Jerárquico: AGNES, DIANA
- Clustering Basado en Densidad: DBSCAN

Introducción al Machine Learning

SESIÓN - 4

Reducción de Dimensionalidad

- Análisis de Componentes Principales
- Análisis Factorial
- Selección de Variables

Normas de clase online:

- Habrá un **break de 10 min** después de la parte teórica y antes de la parte práctica en Google Collaboratory.

- La evaluación se realizará **durante la sesión 4 como trabajo final** y obtendrás puntos adicionales en función de las tareas resueltas y enviadas al correo del profesor:

TAREAS PRESENTADAS	PUNTOS ADICIONALES
Ninguna	0
1	+2
2	+4
3	+6

ALGORITMOS DE CLASIFICACION

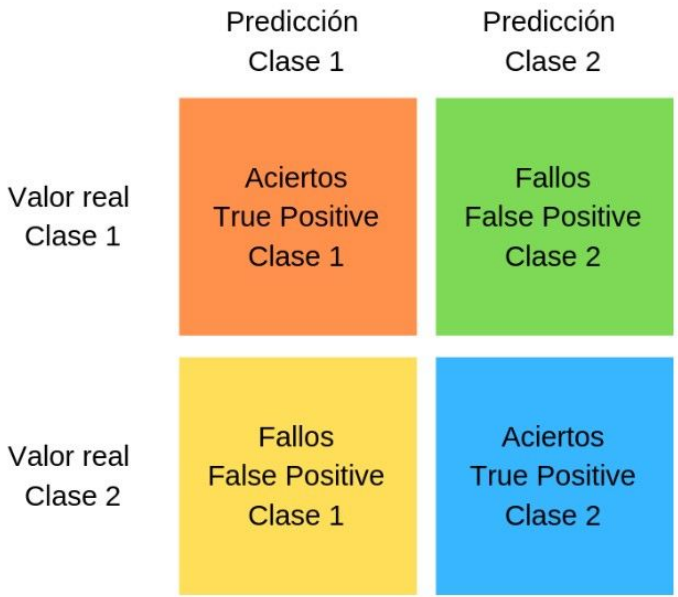
1. **Regresión Logística**
2. **K-vecino más cercano**
3. Clasificador Naive Bayes
4. Perceptron
5. Máquinas de Vectores de Soporte (SVM)
6. Árboles de Decisiones

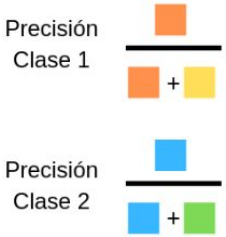
BALANCEO DE DATOS

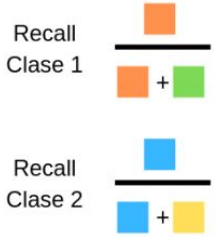
- ¿Qué son los problemas de clasificación de Clases desequilibradas?
  - Requerimiento de balanceo de datos.

- ¿Cómo nos afectan los datos desbalanceados?

EVALUACIÓN DE MÉTRICAS Y MATRIZ DE CONFUSIÓN







EJEMPLO: MÉTRICAS Y MATRIZ DE CONFUSIÓN

$$\text{Accuracy} = \frac{990 + 0}{990 + 0 + 10 + 0}$$

$$\text{Precisión Clase 1} = \frac{990}{990 + 10}$$

$$\text{Precisión Clase 2} = \frac{0}{0 + 0}$$

$$\text{Recall Clase 1} = \frac{990}{990 + 0}$$

$$\text{Recall Clase 2} = \frac{0}{0 + 10}$$

CASO PRÁCTICO:

- Utilizaremos el dataset Credit Card Fraut Detection de la web de Kaggle.
- El dataset consta de 285.000 filas con 31 columnas (features).
- La información es privada, no sabemos realmente qué significan los features y están nombradas como V1, V2, V3, etc,excepto por las columnas Time y Amount
- Y nuestras clases son 0 y 1 correspondiendo con “transacción Normal” ó “Hubo Fraude”. Como podrán imaginar, el **set de datos está muy desequi**

ESTRATEGIAS PARA EL MANEJO DE DATOS BALANCEADOS

1. Ajuste de Parámetros del modelo.
    a. Penalización para compensar
2. Modificar el Dataset.
    a. Subsampling en la clase mayoritaria
3. Muestras artificiales.
    a. Oversampling de la clase minoritaria

Estrategia: Combinamos resampling con Smote-Tomek

4. Balanced Ensemble Methods.
    a. Ensamble de Modelos con Balanceo

ANÁLISIS SUPERVISADO

NOTEBOOK DE BALANCEO DE DATOS

ANÁLISIS SUPERVISADO

NOTEBOOK DE REGRESIÓN LINEAL

ANÁLISIS SUPERVISADO

NOTEBOOK DE BAYES

ANÁLISIS SUPERVISADO

NOTEBOOK DE ARBOL DE DECISION

ANÁLISIS SUPERVISADO

NOTEBOOK DE RANDOM FOREST

ANÁLISIS SUPERVISADO

NOTEBOOK DE k-NEAREST NEIGHBOR

APRENDIZAJE AUTOMÁTICO CON PYTHON

Gracias por su atencion.