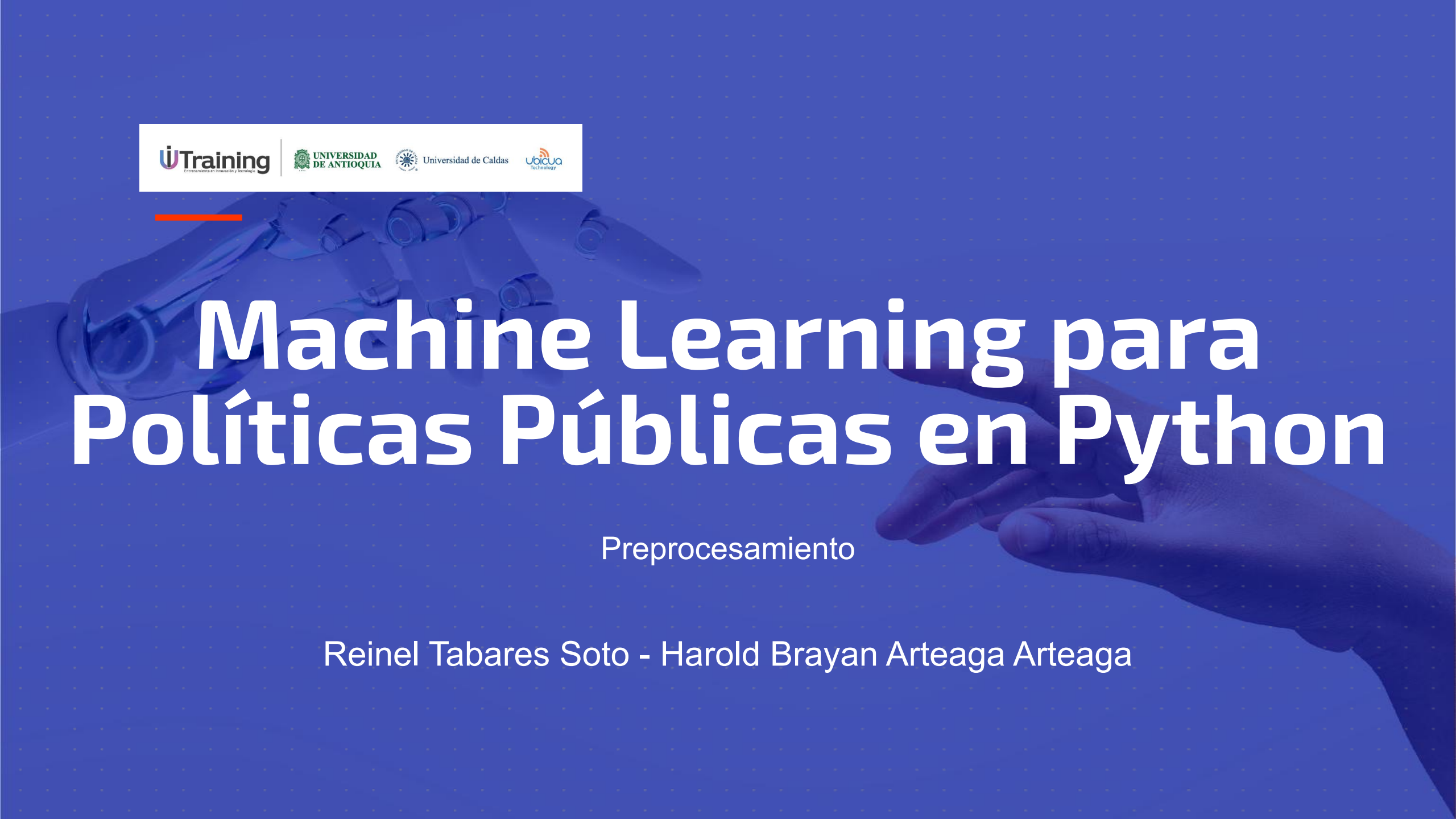


Machine Learning para Políticas Públicas en Python

Preprocesamiento

Reinel Tabares Soto - Harold Brayan Arteaga Arteaga

Agenda

Contenido del módulo 3

- Análisis exploratorio de datos - EDA.
- K-Nearest Neighbors (modelo base para sección práctica).
- Métricas de medición de rendimiento.
- Transformación de características.
- Balance de clases.
- Reducción de dimensionalidad.

Análisis exploratorio de datos - EDA.

Algunas recomendaciones para realizar un EDA

Análisis exploratorio de datos - EDA.

- Verificar la cantidad de clases
- Verificar datos faltantes y outliers
- Verificar el tipo de datos que componen el dataset
- Analizar histogramas de las features
- Analizar correlaciones entre las features con mapas de calor
- Analizar los gráficos entre features con pairplot
- Analizar los gráficos como boxplot, catplot (ejemplo: bar y violin), entre los labels y features seleccionadas

Conjuntos de entrenamiento y prueba

Análisis exploratorio de datos - EDA.

- Conocidos en inglés como conjuntos (sets) de train y test
- No podemos entrenar y probar el modelo con el mismo conjunto de datos
- Se suele separar un conjunto aleatorio de datos para probar (evaluar) los resultados del modelo
- El resto de los datos son usados para entrenar
- Una separación común es 30% para el conjunto de prueba y 70% para el conjunto de entrenamiento

Dataset sobre el diagnóstico de cáncer

Análisis exploratorio de datos - EDA.

- **Descripción dataset :**
  https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)

- **Descargar dataset:**
  https://github.com/stedy/Machine-Learning-with-R-datasets/blob/master/wisc_bc_data.csv

Carpeta con la Data requerida por los Notebooks link

A programar

Link

Análisis exploratorio de datos - EDA.

- **Librerias**
- **Lectura del dataset**
- **Eliminar columnas innecesarias del dataset**
- **Análisis exploratorio de datos (Exploratory data analysis - EDA)**
- **División en datos de entrenamiento y testing**
- K-Nearest Neighbors
- Métricas
- Métricas bonitas
- KNN con preprocesamiento de las features
- Balance de clases
- Reducción de dimensionalidad

Transformación de características.

Normalización de datos

Transformación de características.

Normalización: ajustar los valores medidos en diferentes escalas respecto a una escala común

- Normalización min-max $$z_i = \frac{x_i - min(x)}{max(x) - min(x)}$$
- Normalización z-score $$z_i = \frac{x_i - \overline{x}}{\sigma}$$

Carpeta con la Data requerida por los Notebooks link

A programar

Link

Transformación de características.

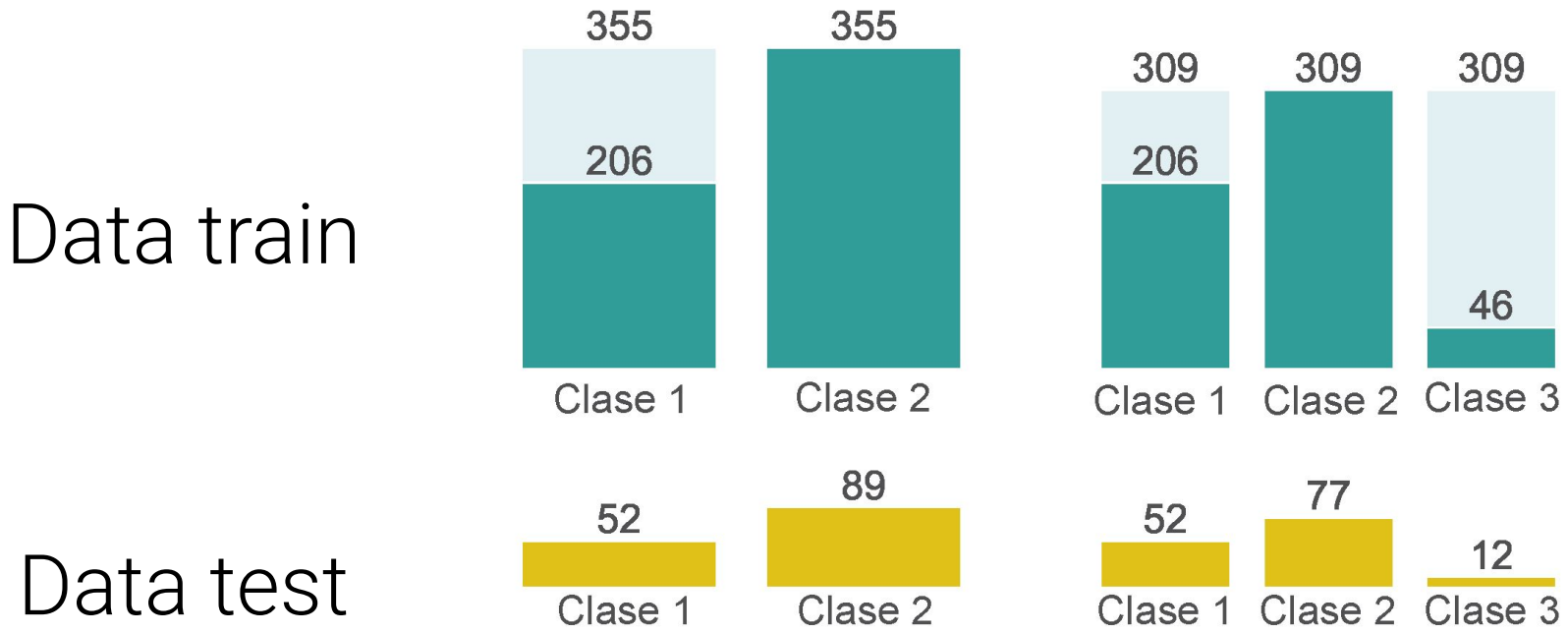
- Librerias
- Lectura del dataset
- Eliminar columnas innecesarias del dataset
- Análisis exploratorio de datos (Exploratory data analysis - EDA)
- División en datos de entrenamiento y testing
- **K-Nearest Neighbors**
- **Métricas**
- **Métricas bonitas**
- **KNN con preprocesamiento de las features**
- Balance de clases
- Reducción de dimensionalidad

Balance de clases.

¿Cuándo balancear?

Balance de clases.

Se deben explorar los datos, si estos se encuentran desbalanceados, se puede realizar el balance de clases.



Técnicas para balancear clases

Balance de clases.

- **A la clase mayor**
  - Random
  - SMOTE
    - BorderlineSMOTE
    - KMeansSMOTE
    - SVMSMOTE
    - SMOTEN
  - ADASYN

Técnicas para balancear clases

Balance de clases.

- **A la clase menor**
  - Random
  - ClusterCentroids
  - NearMiss

Reducción de dimensionalidad.

Análisis de Componentes Principales (PCA)

Reducción de dimensionalidad.

- Técnica utilizada para describir un conjunto de datos en términos de **nuevas variables** («componentes») **no correlacionadas**. Los componentes se ordenan por la cantidad de varianza original que describen, por lo que la técnica es útil para reducir la dimensionalidad de un conjunto de datos.
- Técnicamente, el PCA busca la proyección según la cual los datos queden mejor representados en términos de mínimos cuadrados. Esta convierte un conjunto de observaciones de variables posiblemente correlacionadas en un conjunto de valores de variables sin correlación lineal llamadas **componentes principales**.
- Se puede seleccionar las primeras componentes para representar los datos.

Carpeta con la Data requerida por los Notebooks link

A programar

Link

Transformación de características.

- Librerias
- Lectura del dataset
- Eliminar columnas innecesarias del dataset
- Análisis exploratorio de datos (Exploratory data analysis - EDA)
- División en datos de entrenamiento y testing
- K-Nearest Neighbors
- Métricas
- Métricas bonitas
- KNN con preprocesamiento de las features
- **Balance de clases**
- **Reducción de dimensionalidad**