Club de Datos - Exactas UBA

Consigna para dataset Iris

August 31, 2021

Acerca del dataset

El conjunto de datos de Iris se utilizó en R.A. El artículo clásico de Fisher de 1936, The Use of Multiple Measurements in Taxonomic Problems, y también se puede encontrar en el Repositorio de aprendizaje automático de la UCI.

Incluye tres especies de iris con 50 muestras cada una, así como algunas propiedades de cada flor. Una especie de flor es linealmente separable de las otras dos, pero las otras dos no son linealmente separables entre sí.

Las columnas de este conjunto de datos son:

- Id
- SepalLengthCm
- SepalWidthCm
- PetalLengthCm
- PetalWidthCm
- Species

Un ejemplo de alguien que exploro este dataset y le ajusto un modelo de machine learning (k-nearest neighbors o knn) se encuentra en el siguiente link.

Objetivo Nivel Cero

El objetivo es hacer una exploración de los datos que logre informar, resumir o describir la información en el dataset.

Primero, vas a necesitar importar (cargar, abrir) los datos. El archivo es iris.csv.

Para entender el dataset, puede ser útil averiguar cuántas filas tiene, cuantas columnas, que valores toman esas columnas, cuales son los tipos de columna que hay (dtypes), cómo son las primeras filas.

Objetivo Exploración (Pandas - Seaborn)

Dar con algunas tablas y/o gráficos que nos puedan contar los puntos principales de esta historia, según la información en el dataset. Pueden probar la librería seaborn para armar fácilmente algunos gráficos interesantes. Sino, usar matplotlib, aunque puede ser más complicado para quien nunca la uso.

Objetivo Machine Learning (Scikit Learn)

El objetivo es dar con un modelo de clasificación (machine learning) que prediga la especie de iris a partir de sus medidas.

La idea es probar distintos modelos e ir evaluandolos por su capacidad de predecir correctamente.

Pueden evaluar los modelos usando medidas como precisión, recall, F1, matrices de confusión, etc, habiendo separado un training set de test set.

Se recomienda usar la libreria scikit-learn. La documentacion de Scikit Learn tiene ejemplos de modelos SVM, nearest centroid, nearest neighbors, gaussian process, decision tree, aplicados en el iris dataset.