Tarea 4

Introducción

La empresa irlandesa *Dole Food Company* busca desarrollar una máquina capaz de distinguir entre plátanos de buena calidad (*Good Quality*) y aquellos de mala calidad (*Bad Quality*). Se ha encargado a un equipo del departamento de robótica de la empresa la creación de esta máquina, y han contratado sus servicios para desarrollar un modelo de minería de datos que permita llevar a cabo esta distinción.

Luego de entregar el primer informe (Tarea 3), donde realizo la clasificación con un algoritmo KNN, usted cree que quizás se pueden hacer una mejor clasificación usando otros algoritmos supervisados. Para comprobar su hipótesis, ahora compara los valores de accuracy, precisión, recall y f1-score de tres algoritmos de clasificación: KNN, Árbol de Decisión y Bayes.

Datos

La base de datos la pueden encontrar en WebCursos y contiene los siguientes atributos:

- **Size:** tamaño de la fruta
- **Weight**: peso de la fruta
- **Sweetness**: dulzura de la fruta
- **Softness**: suavidad de la fruta
- **HarvestTime**: cantidad de tiempo transcurrido desde la cosecha de la fruta
- **Ripeness**: madurez de la fruta
- **Acidity**: acidez de la fruta
- **Quality**: calidad de la fruta (Buena o Mala)

Entregables

- En Webcursos, usted debe subir el código Python utilizado, en formato Jupyter Notebook (ipynb).
- El nombre del Archivo debe ser **Tarea3_iniciales1_iniciales2.ipynb**
- Ejemplo, el Notebook de Alfonso Tobar y Sofía Álvarez se debe entregar como **Tarea3_AT_SA.ipynb**.

Código

El código Python escrito en Jupyter Notebook debe:

1. Abrir el archivo (.csv).
2. Separar en los datos en conjuntos de **entrenamiento** (80%) y **prueba** (20%) usando la función train_test_split(). Recuerde usar una semilla para reproducir sus resultados
3. Utilizando el set de Train, entrene un modelo Naive Bayes, Decision Tree y Logistic Regression. Para ello encuentre los mejores Hiperparámetros de cada modelo utilizando un 5-Fold Cross Validation:
    a. Cada modelo debe venir pre-procesado utilizando MinMaxScaler() dentro de un Pipeline.
    b. Para cada modelo varíe los siguientes hiperparámetros según corresponda:
        i. DecisionTree: MaxDepth: [1,10]
        ii. Naive Bayes: alpha : [0.1,1] variando en pasos de 0.1.
        iii. LogisticRegression: C: [0.001,100] usando una escala logarítmica.
    c. Escoja bajo qué métrica se validarán los modelos de K-Fold y reporte el valor promedio en gráficos apropiados.
4. Indicar qué algoritmo entrega el mejor modelo de clasificación reportando los resultados de Accuracy, Recall y Precision en el **test set**.
5. Compare los resultados obtenidos con el mejor modelo obtenido en la Tarea 3. ¿Hay alguna mejora?

El Jupyter Notebook debe explicar:

- Todos los pasos del código **en detalle**.
    - Si se entrega un código sin su explicación, entonces será evaluado con nota 1.
    - Si se entrega código con código explicado parcialmente, solo las partes explicadas serán evaluadas.