Higher School of Agriculture of Mograne

Homework Project -2-

Group: 3^{rd} Engineer in agronomy

Prof. Bilel AMMOURI
Period: 2023 – 2024

Homework Project: The use of machine learning methods in classification of pumpkin seeds (Cucurbita pepo L.)[1]

Background:

Pumpkin seeds are widely consumed globally for their nutritional benefits. This study focuses on two important varieties, Ürgup Sivrisi.^{a}nd Çercevelik,"grown in Urgup and Karacaoren regions in Turkey. The goal is to develop classification models based on morphological measurements of pumpkin seeds, employing machine learning techniques.

Data

The dataset hosted on the homepage of Dr. Murat Koklu can be downloaded from there (muratkoklu).

The dataset encompasses the following variables: Area, Perimeter, Major Axis Length, Minor Axis Length, Convex Area, Equiv Diameter, Eccentricity, Solidity, Extent, Roundness, Aspect Ratio, Compactness. The target variable is Class (Urgup Sivrisi or Cercevelik).

Objectives:

1. **Data Acquisition**:
   - Obtain the dataset from (this link).
2. **Data Exploration and Preprocessing**:
   - Explore the dataset to understand its structure and characteristics..
   - Handle missing values, outliers, and any other data preprocessing steps deemed necessary.
   - Visualize the distribution of features and explore relationships.
3. **Feature Engineering**:
   - Extract relevant features or create new ones that might enhance model performance.
4. **Model Development**:
   - Split the dataset into training and testing sets.
   - Implement classification models (e.g., SVM, Random Forest, Logistic Regression, XGBoost, ...) using appropriate libraries (scikit-learn, TensorFlow, or PyTorch).
5. **Model Evaluation and Interpretation**:
   - Evaluate the models using appropriate metrics (accuracy, precision, recall, F1-score).
   - Interpret the results and discuss the performance of each model.
6. **Communication**:

[1] KOKLU, M., SARIGIL, S. and OZBEK, O. (2021). The use of machine learning methods in classification of pumpkin seeds (Cucurbita pepo L.). Genetic Resources and Crop Evolution, 68(7), 2713-2726. doi:
DOI: https://doi.org/10.1007/s10722-021-01226-0

- Summarize findings and insights in a clear and concise manner.
- Create visualizations to support your interpretations.

Deliverables:

1. A comprehensive report documenting the entire process, including data exploration, preprocessing, feature engineering, model development, evaluation, and interpretation.
2. A Jupyter notebook containing the code used for data analysis and model development.
3. Visualizations to aid in the understanding of the dataset and model performance.

Note:

Attached, you will find the paper related to this database.

Deadline:

Please submit your completed project by **January 15, 2024**.