Sultan Qaboos University

Department of Computer Science

COMP3602: Data Analysis and Visualization with Python

Lab Test - Spring 2024

April 25th, 2024 @ 8:00 AM – 9:30 AM

Time: 90 Minutes

Total Marks: 30

Important Notes for Jupyter Notebook File:

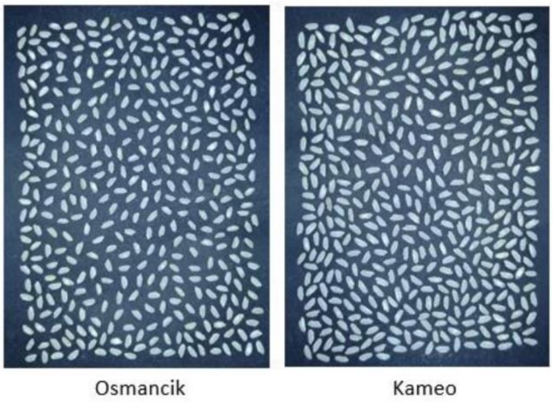
- You should download the jupyter notebook from Elearn and rename it as "ID.ipynb" where "ID" represents your student ID
- Do not modify the content or order of any cells with questions.
- Use the given empty cells to write code and feel free to add extra cells if needed.
- When you need to write an answer, make sure the cell is Markdown

Cheating/Copying Rule

- Any attempt of cheating or copying from any online or offline source will lead directly to an F grade
- Automatic software can be used to detect the above-mentioned cases
- Internet is not allowed.

Dataset

Among the certified rice grown in TURKEY, the Osmancik species, which has a large planting area since 1997 and the Cammeo species grown since 2014 have been selected for the study. When looking at the general characteristics of Osmancik species, they have a wide, long, glassy and dull appearance. When looking at the general characteristics of the Cammeo species, they have wide and long, glassy and dull in appearance.



The rice grain's images were taken for the two species, processed and feature inferences were made. 7 morphological features were obtained for each grain of rice.

Attribute Information:

1. Area: Returns the number of pixels within the boundaries of the rice grain.
2. Perimeter: Calculates the circumference by calculating the distance between pixels around the boundaries of the rice grain.
3. Major Axis Length: The longest line that can be drawn on the rice grain, i.e. the main axis distance, gives.
4. Minor Axis Length: The shortest line that can be drawn on the rice grain, i.e. the small axis distance, gives.
5. Eccentricity: It measures how round the ellipse, which has the same moments as the rice grain, is.
6. Convex Area: Returns the pixel count of the smallest convex shell of the region formed by the rice grain.
7. Extent: Returns the ratio of the region formed by the rice grain to the bounding box pixels
8. Class: Commeo and Osmancik.

Address the following questions using the Jupyter Notebook file.

1. [**1 Point**] Download the dataset from elearn and then load it to your notebook.
2. [**1 Point**] Utilize appropriate functions from suitable libraries to determine if the dataset is balanced.
3. [**1 Point**] Show the information about the dataset including the total number of observations, the total number of attributes and the type of each attribute in this dataset
4. **[1 Point]** For feature selection drop any unusable feature from the dataframe
5. [**1 Point**] a) Display a count of missing observations in each attribute. b) Remove the observations with missing values.
6. **[1 Point]** Create a list that holds the numeric attributes columns' names.
7. **[2 Point]** Utilize the appropriate function to Show the five-number summary for all numeric attributes
8. [**6 Points**] a) Show the correlation table. b) Use appropriate plots to show the relationship between every two numeric columns in the dataset, with each data point coloured according to the class attribute. (*Make sure plots have a title, xlabel and ylabel*).
9. [**2 Points**] a) Show the distribution of the 'Area' for the 'Cammeo' class and the 'Osmancik' class separately. b) Point out the differences between the two histograms.
10. [**6 Points**] a) Use appropriate visualization to identify all attribute that has outliers and all the attributes with no outliers. b) Use an appropriate way to scale the data and redo the visualization
11. [**2 Points**] Perform appropriate preprocessing techniques on the categorical attribute(s) in the dataset.
12. [**6 Points**] Perform multivariate data analysis for classification using the following steps for the k nearest neighbor.
    a. Create feature (X) of the scaled data and target (y)
    b. Split the dataset into a training set (70%) and testing set (30%), set random_state to 42. Each set (training and test) should contain approximately the same percentage of samples of each target class as the complete set.
    c. Run the Knn classifier with **one** neighbor

    d. Print a classification report.
    e. Write a conclusion (in 2 sentences) summarizing the most important finding of this task.

13. [**BONUS 10 Points**] Grid Search is a method used to find optimal hyperparameters that result in the highest prediction accuracy for a model.
    a. Apply Logistic regression classification on the sets you have prepared for the previous question.
    b. Apply Grid search to find the best hyperparameter for the logistic regression model.
    c. Print a classification report of the best LR model
    d. Compare the results of LR model and Knn model (in short)