Classification of Rice Varieties

Prepared for:

Dr. Samaher Alghamdi

CPIT- 440

Group number: 4

Section: CAR

Group leader:	**Group members:**	
Shuruq Hassan Baabdullah	Rahaf Mutaz Dawoud	Manar Mutlaq Altaiary
1906284	1911088	1906775

Contents:

Introduction:

Problem:

Rice, as one of the world's most frequently grown and consumed cereal crops. Rice is also one of Turkey's central sources of livelihood because of its cost-effectiveness and nutritional value. Rice goes through several industrial phases before reaching our tables. This process includes cleaning, color sorting, and classification.

To sum up, cleaning is the process of separating rice from foreign substances. Moreover, classification is the process of separating broken ones from sturdy rice. Furthermore, color extraction includes separating stained and striped rice from the white ones on the rice surface species. In this study, a computerized vision system was built in and developed to identify between two rice species. In addition, about 3810 rice grain pictures were obtained, analyzed, and feature inferences were formed.

Importance:

Physical appearance and cooking characteristics, aroma, taste, and smell are so important. From the perspective of the end consumer, it is the first feature of physical appearance that comes to mind from the criteria that stand out in the rice varieties that are sold packaged on market shelves. After production, it is seen that the need for technological methods increases because the calibration of rice, determination of its types, and separation of various quality elements are inefficient and time-consuming, especially in terms of those with high production volume. So that, when we look at the recent studies on cereal products using machine vision systems and image processing techniques, it is seen that the products are examined in terms of many physical properties such as color, texture, quality, and size.

Goal:

The ultimate goal is to distinguish between two proprietary rice species. Moreover, we can aim to reach this goal by analyzing the dataset. The result of the rice dataset will be expecting a prediction of the rice species kind of rice from the dataset (Cammeo or Osmania) which they are a Turkish brand of rice.

Figure 1 Osmania

Figure 2 CAMMEO

Data exploration:

To explore the date, we used some functions as describe below:

- we used the info () method to get a quick description of the data, in particular the total number of rows and columns, and each attribute's type and number of non-null values

```
RangeIndex: 3810 entries, 0 to 3809
Data columns (total 8 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   AREA          3810 non-null   int64
 1   PERIMETER     3810 non-null   float64
 2   MAJORAXIS     3810 non-null   float64
 3   MINORAXIS     3810 non-null   float64
 4   ECCENTRICITY  3810 non-null   float64
 5   CONVEX_AREA   3810 non-null   int64
 6   EXTENT        3810 non-null   float64
 7   CLASS         3810 non-null   object
dtypes: float64(5), int64(2), object(1)
memory usage: 238.2+ KB
```

Figure 3 info()

- we applied the describe () method to describe the whole dataset with its values statistically

	count	mean	std	min	25%	50%	75%	max
AREA	3810.0	12667.727559	1732.367706	7551.000000	11370.500000	12421.500000	13950.000000	18913.000000
PERIMETER	3810.0	454.239180	35.597081	359.100006	426.144752	448.852493	483.683746	548.445984
MAJORAXIS	3810.0	188.776222	17.448679	145.264465	174.353855	185.810059	203.550438	239.010498
MINORAXIS	3810.0	86.313750	5.729817	59.532406	82.731695	86.434647	90.143677	107.542450
ECCENTRICITY	3810.0	0.886871	0.020818	0.777233	0.872402	0.889050	0.902588	0.948007
CONVEX_AREA	3810.0	12952.496850	1776.972042	7723.000000	11626.250000	12706.500000	14284.000000	19099.000000
EXTENT	3810.0	0.661934	0.077239	0.497413	0.598862	0.645361	0.726562	0.861050

Figure 4 describe()

- All attributes are numerical, except the class field. Its type is object, specifically it is a text attribute. This attribute is categorical can find out what categories exist and how many rows (districts) belong to each category by using the value counts () method

```
Osmancik    2180
Cammeo      1630
Name: CLASS, dtype: int64
```

Figure 5 value_counts ()

Data visualization:

After we have explored our data, data visualization is needed. Here are some visualization representations to display the data in a graphical way.

1. Histogram: shows the number of instances (on the vertical axis) that have a given value range (on the horizontal axis). It is a representation of the distribution of data.

- We used the hist () to display the frequency of all numerical the attributes.
    - Area:

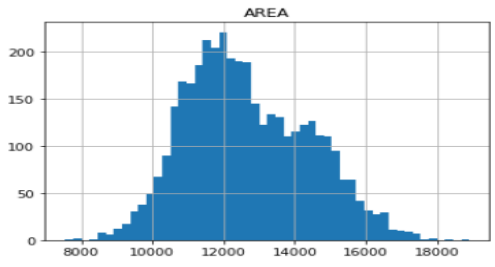

Figure 6 area

- Perimeter:

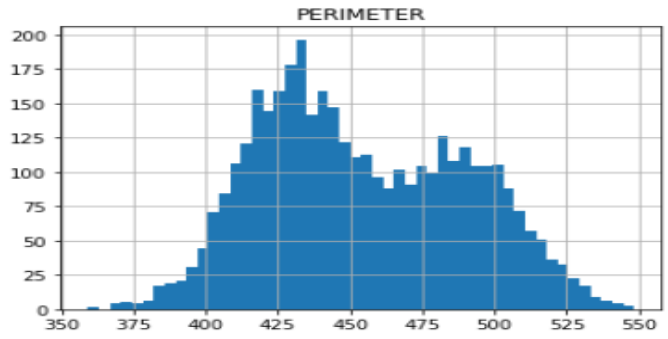

Figure 7 Perimeter

- Major Axis Length:

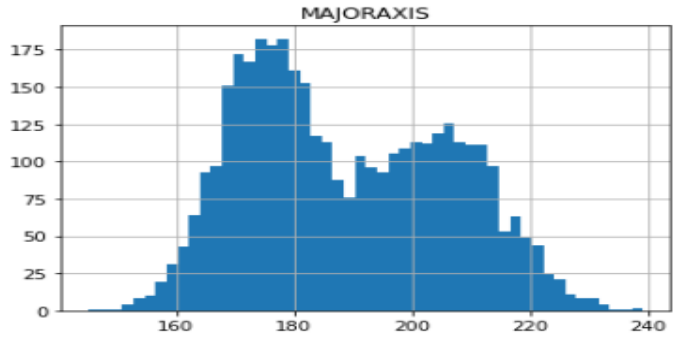

Figure 8 Major Axis Length

- Minor Axis Length:

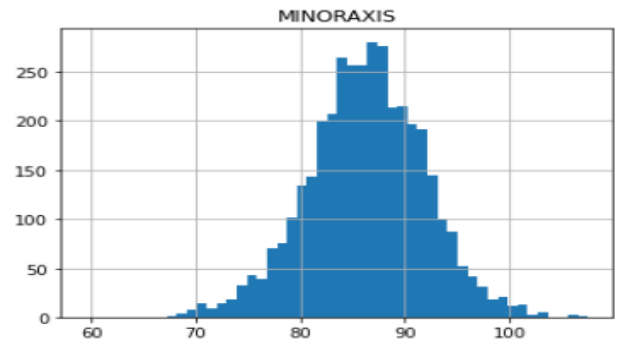

Figure 9 Minor Axis Length

- Eccentricity:

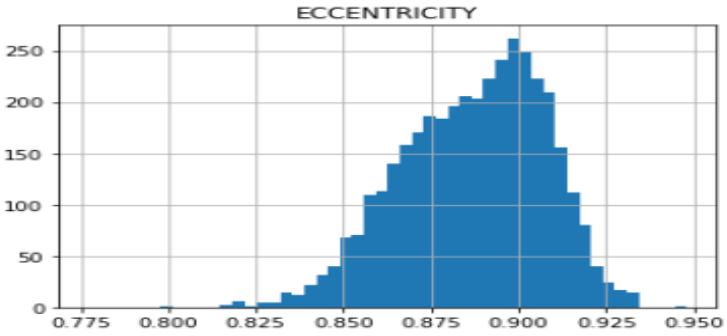

Figure 10 Eccentricity

- Convex Area:

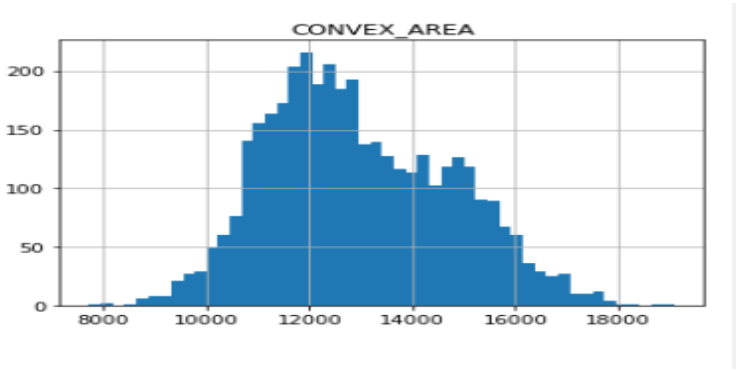

Figure 11 Concex Area

- Extent:

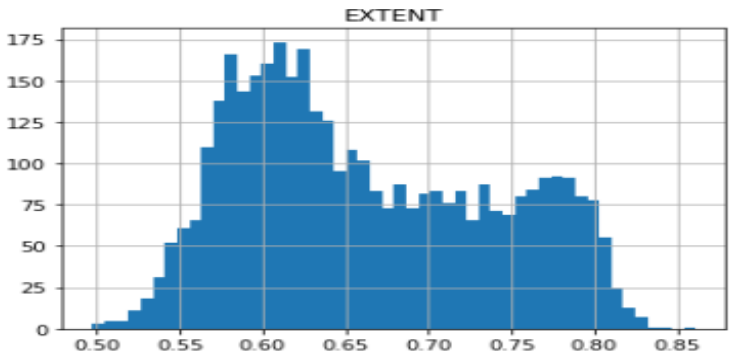

Figure 12 Extent

2. Bar: To plot histogram of the nominal attribute
    - Class:

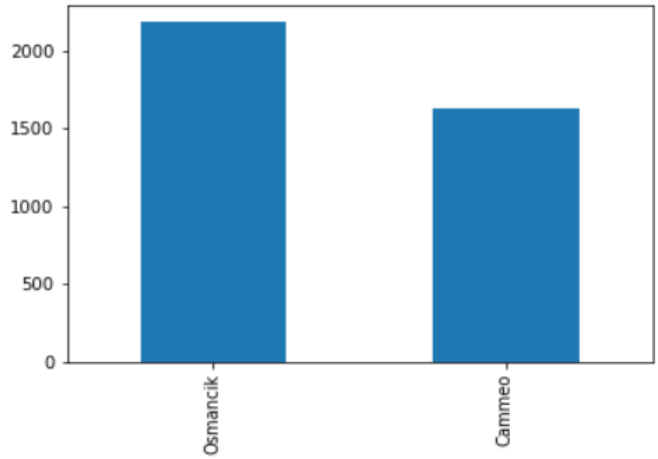

Figure 13 Bar: class

3. Scatter: to show the relationship between two attributes
    - MAJORAXIS and MINORAXIS:

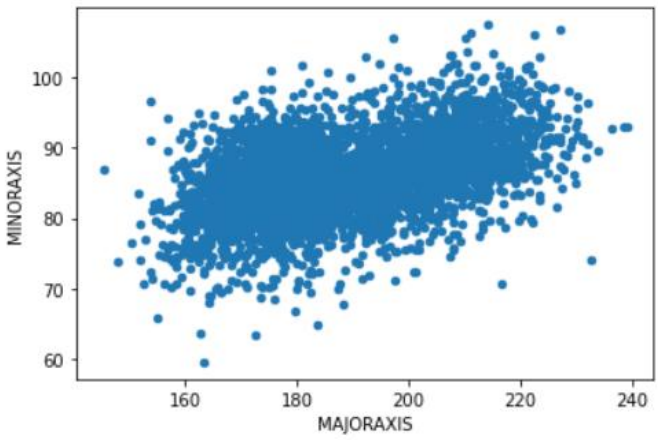

Figure 14 Scatter: majoraxis and minoraxis

4. Boxplot: it displays the 5 number summary (max, min, median, Q1, Q3), it helps to know the skewness of the data.

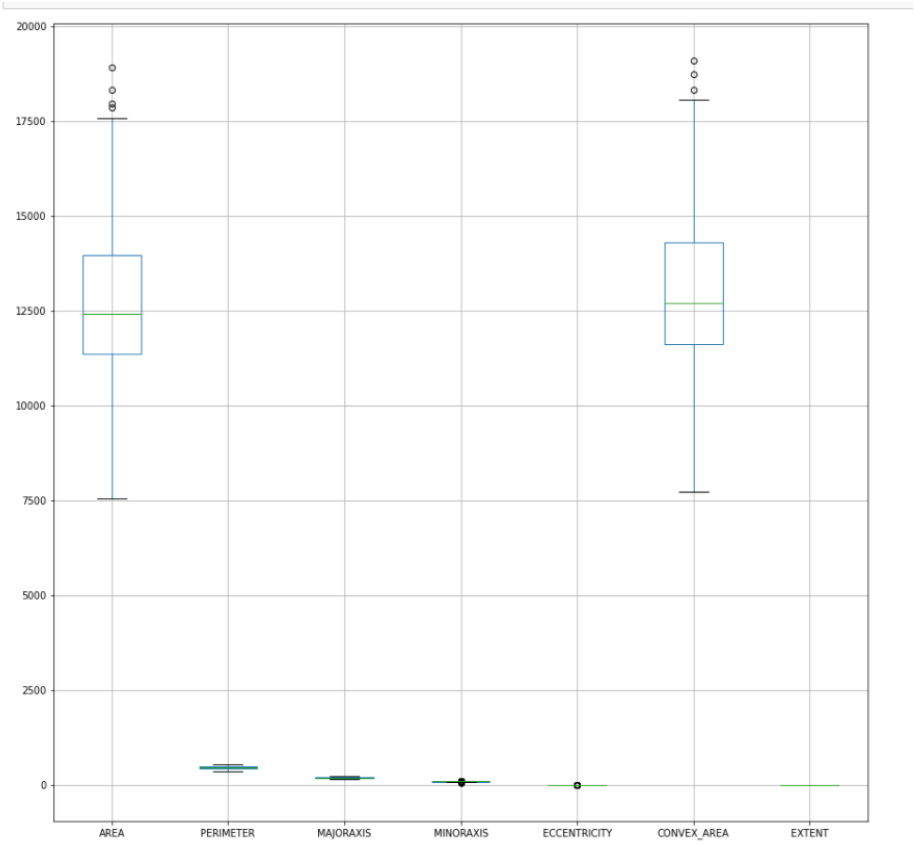

Figure 15 Boxplot: attribute

Data preprocessing:

Before applying algorithms of modeling, we should prepare and clean the data from any null values and outliers, duplication etc.

1. Duplication: In the beginning, we print the duplicated rows from the dataset by using this function duplicated () this function only checks for duplicates but does not remove them. if there are any duplicated rows than it should be removed by using the drop function Moreover, in our dataset we don't have any duplicate rows, so we do not have to implement the drop function

   Implement.

2. Sampling and splitting into train and test sets:

   Firstly, we split the dataset based on paper to 25% train 75% test.

   - Divide the dataset into Dependent & Independent variable:
     - Dependent: TargetClass
     - Independent: classWithoutTarget
   - Split the dataset into train and test using TargetClass and classWithoutTarget:
     - train_classWithoutTarget
     - test_classWithoutTarget
     - train_TargetClass
     - test_TargetClass

3. Missing Values:

   we applied a function to check if the data has missing values isnull ().and the output shows that there are no missing or null values. if there are, it will display as “Nan" we didn't find missing values.

4. Outliers:

   Outliers can skew statistical measures and data distributions which produce bad prediction and analysis of the data. First, we calculate the IQR of all the **numerical attributes**. After data visualization and by using boxplot, we found that we have 4 attributes that contain outlier. Then we found the indexes of the data frame by using index () method and dropped them.

5. Data transformation and scaling:

   Machine Learning algorithms don't perform well when the input numerical attributes have very different scales. There are two common ways to get all attributes to have the same scale: min-max scaling and standardization (z-score normalization). we use data transformation and scaling in Features selection.

6. Features selection: There are many methods to select the most important features. We used Random Forest Classifier, which is a classifier that contains several decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset First. Moreover, SGD Classifier and we apply the scaling and transformation on them, and the best was RFE because of the high accuracy:

   - RandomForestClassifie:0.94
   - SGDClassifier:0.51

Models Training

A training model is a dataset used to train a machine learning algorithm. It is made up of sample output data as well as the equivalent sets of input data that have an impact on the outcome. The training model is used to process the input data via the algorithm in order to compare the processed output to the sample output.

Algorithms:

- Decision Tree Classifier (CLF):

  Decision Trees are a type of supervised machine learning in which the data is continually split according to a parameter (you explain what the input is and what the related output is in the training data). Two entities, decision nodes and leaves, can be used to explain the tree. The decisions or final outcomes are represented by the leaves. And the data is separated at the decision nodes.

- Naïve Bayes (GNB FROM):

  The Bayes theorem is used in Nave Bayes algorithms, which is a classification strategy based on the strong assumption that all predictors are independent of one another. To put it another way, the assumption is that the presence of a feature in a class is unrelated to the presence of other features in the same class.

- Logistic Regression (LR):

  The supervised learning classification algorithm logistic regression is used to predict the likelihood of a target variable. Because the nature of the goal or dependent variable is dichotomous, there are only two classifications.

Cross validation:

We applied cross validation on Logistic Regression model after we fit the model. We predict training target set by using cross_val_predict. the result was:

Models' evaluation

Model evaluation is an important step in the creation of a model. It aids in the selection of the best model to represent our data and the prediction of how well the chosen model will perform in the future.

Performance metrics:

In model evaluation we used accuracy score method and the result were:

- Logistic Regression: 0.918
- Decision Tree Classifier: 0.926
- Naïve Bayes: 0.904

Comparison of models' performance using table or plots:

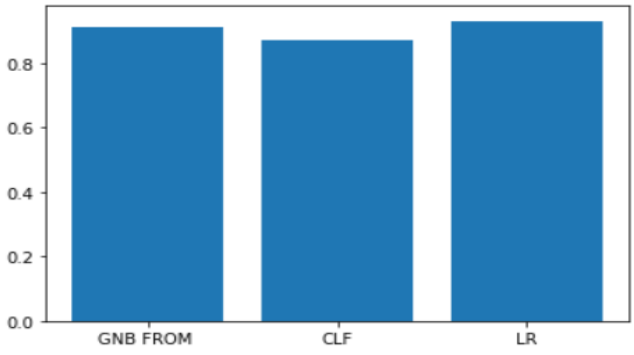

Figure 16 model performance

Selected model:

The best model based on performance was Decision Tree Classifier with accuracy 0.93.

Tools

We use a lot of libraries applying them in our dataset:

- Pandas to read the dataset
- From matplotlib library we use pyplot to draw information in a graphical way
- Seaborn
- %matplotlib inline
- from sklearn.model_selection we use train_test_split
- from sklearn.preprocessing we use StandardScaler to scale the dataset
- from sklearn.feature_selection we use RFE method to select features
- from sklearn.svm we use SVC
- from sklearn.ensemble we use RandomForestClassifier
- from sklearn.model_selection we use cross_val_score

Difficulties and challenges we have faced

The only difficulty we faced was choosing the right dataset for the project

Future work

We would try to apply new models and methods on the dataset to see if there's a better or more balanced output.

Work division

All the members of the group worked with the same effort and the work was divided between all of us fairly.

MANAR MUTLAQ ALTAIARY	SHURUQ HASSAN BAABDULLAH	RAHAF MUTAZ DAWOUD
33%	33%	33%

References

[1] M. K. Ilkay Cinar, "Classification of Rice Varieties Using Artificial Intelligence Methods," [Online]. Available: https://www.ijisae.org/IJISAE/article/view/1068.