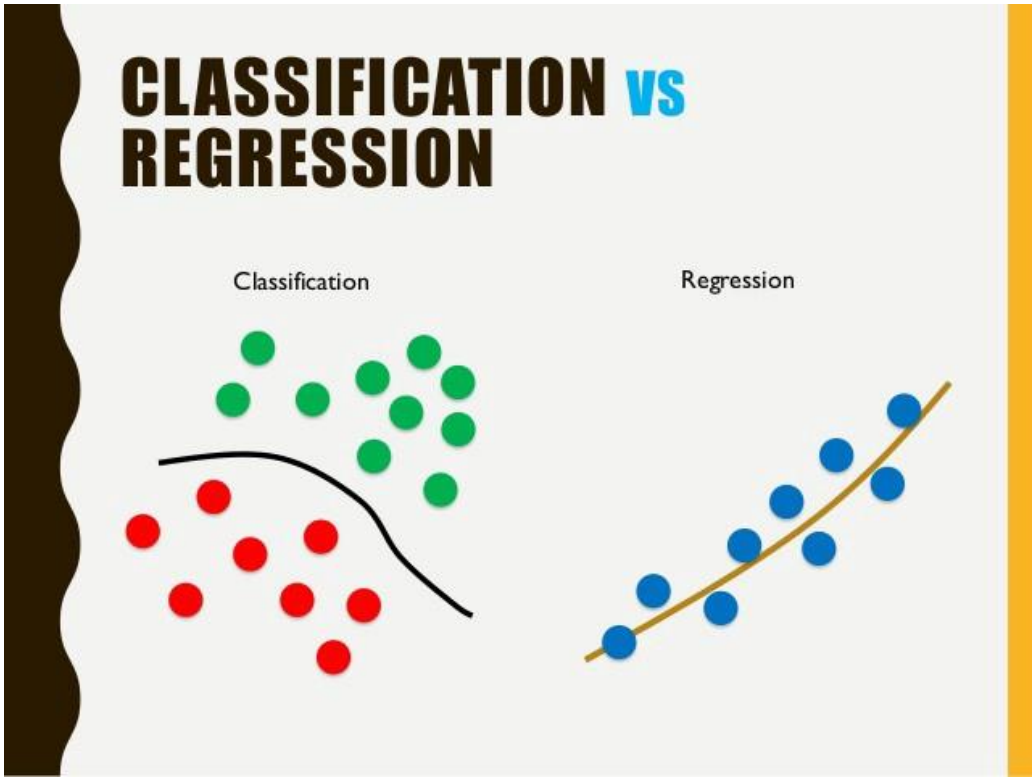
NED UNIVERISTY

OF ENGINEERING & TECHNOLOGY

Machine Learning Report



INSTRUCTOR:	Ms. Hameeza Ahmed

Group Members

SYED UMAIR HASAN	MUHAMMAD FAHAD ALAM
CS-18070	CS-18075
3rd Year, CIS	3rd Year, CIS

Rice Type Classification

A Description of Dataset

Problem of Interest

The data in the dataset was extracted from two kinds of rice, 'Gonen', 'Jasmine'. Our task is to train a model which will classify the type of rice, whether it is 'Gonen' or 'Jasmine'.

Data Source

The data was taken from the Kaggle website.

A Detailed explanation of the data attributes

Data issues and description

The dataset contains 18185 rows and 12 columns. All attributes are numeric variables and they are listed below:

- id
- Area
- MajorAxisLength
- MinorAxisLength
- Eccentricity
- ConvexArea
- EquivDiameter
- Extent
- Perimeter
- Roundness
- AspectRation
- Class

Fortunately, this dataset contains no missing values, there are no categorical features and the dataset is also well balanced. However, it contains some features which have outliers.

Data Preprocessing / Wrangling

Data Cleaning

As the dataset contains outliers in some of the features, therefore those outliers were removed by first calculating the interquartile range and then finding the upper and lower range. Any value which is out of that range is considered as an outlier and then removed.

Data Reduction

The 'Id' column was removed from the dataset as it was irrelevant. The 'Class' column was also dropped from the dataset as it was a target variable.

Data Editing

Many of the columns were not normalized in the dataset therefore , all the input columns were passed through a standard scaler to normalize the data.

Basic statistics of attributes

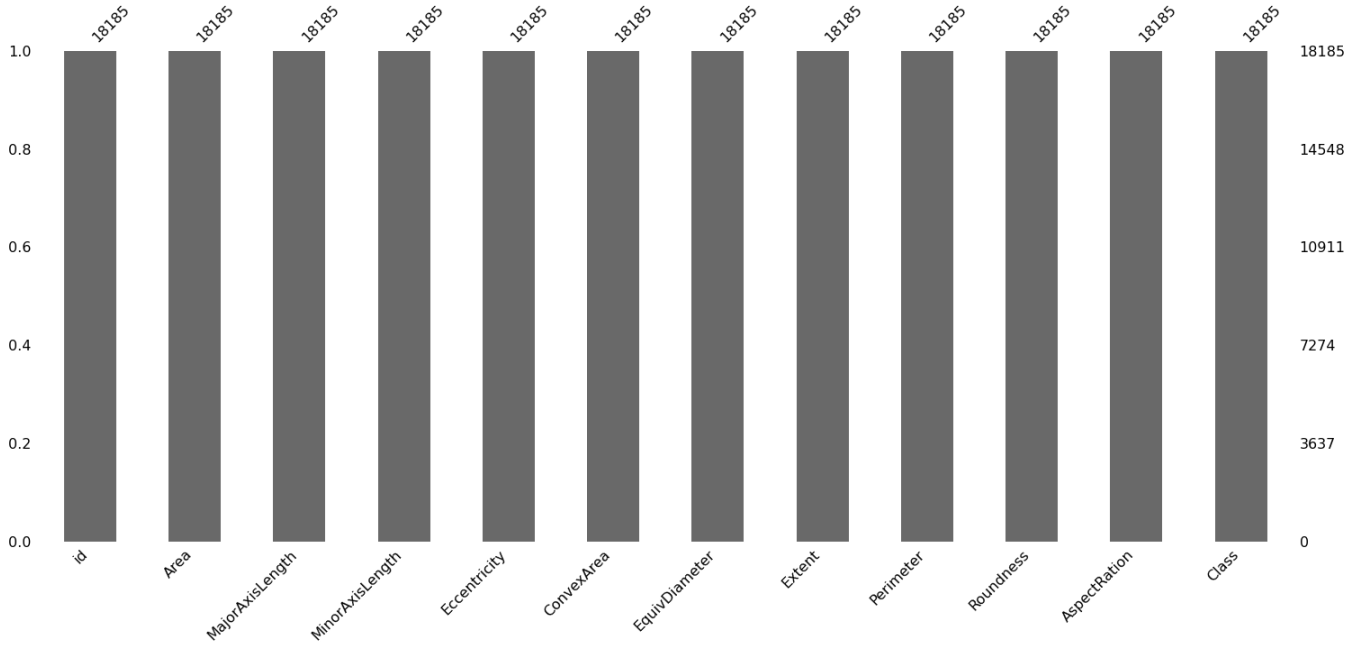
Description of Numerical Attributes

	Area	MajorAxisLength	MinorAxisLength	Eccentricity	ConvexArea	EquivDiameter	Extent	Perimeter	Roundness	AspectRation
count	17597.000000	17597.000000	17597.000000	17597.000000	17597.000000	17597.000000	17597.000000	17597.000000	17597.000000	17597.000000
mean	7111.463772	152.978496	59.961723	0.917226	7301.974086	94.689501	0.615732	353.964315	0.706197	2.615482
std	1412.424549	9.926544	10.004501	0.027213	1445.997988	9.407461	0.104490	25.733769	0.065566	0.424586
min	3698.000000	124.110890	36.177914	0.819338	3883.000000	68.618072	0.383239	280.060000	0.474142	1.744254
25%	6018.000000	146.610657	51.456124	0.892386	6180.000000	87.534882	0.536760	335.976000	0.650238	2.215959
50%	6709.000000	154.289429	55.917314	0.925461	6892.000000	92.423829	0.600259	354.276000	0.697445	2.639618
75%	8464.000000	160.280519	70.225361	0.941634	8687.000000	103.810883	0.694617	373.589000	0.768167	2.970556
max	10210.000000	180.332508	82.550762	0.966774	10659.000000	114.016559	0.886137	429.919000	0.854461	3.911845

Data visualization

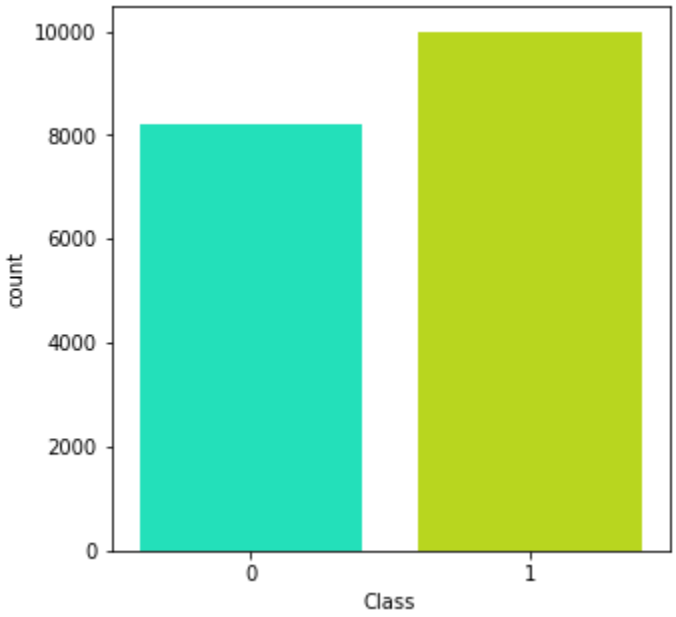
The distribution of the attributes

The bar graph shows that there are no missing values in dataset.

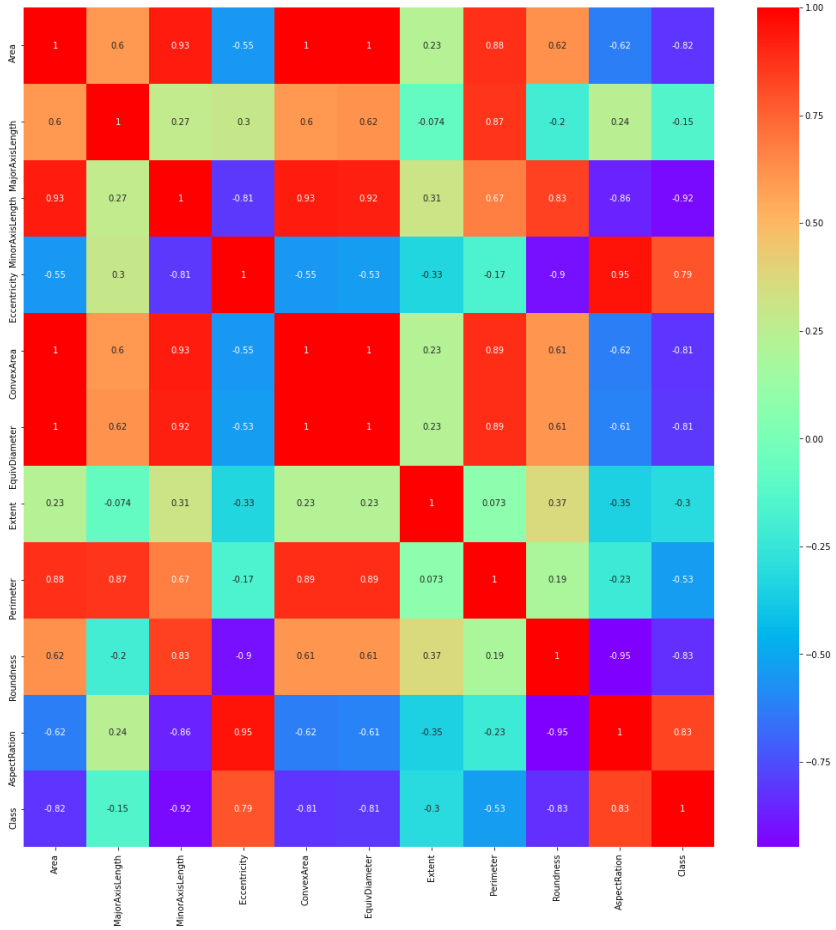


Checking whether dataset is balanced or not

The figure below shows that the dataset is balanced.

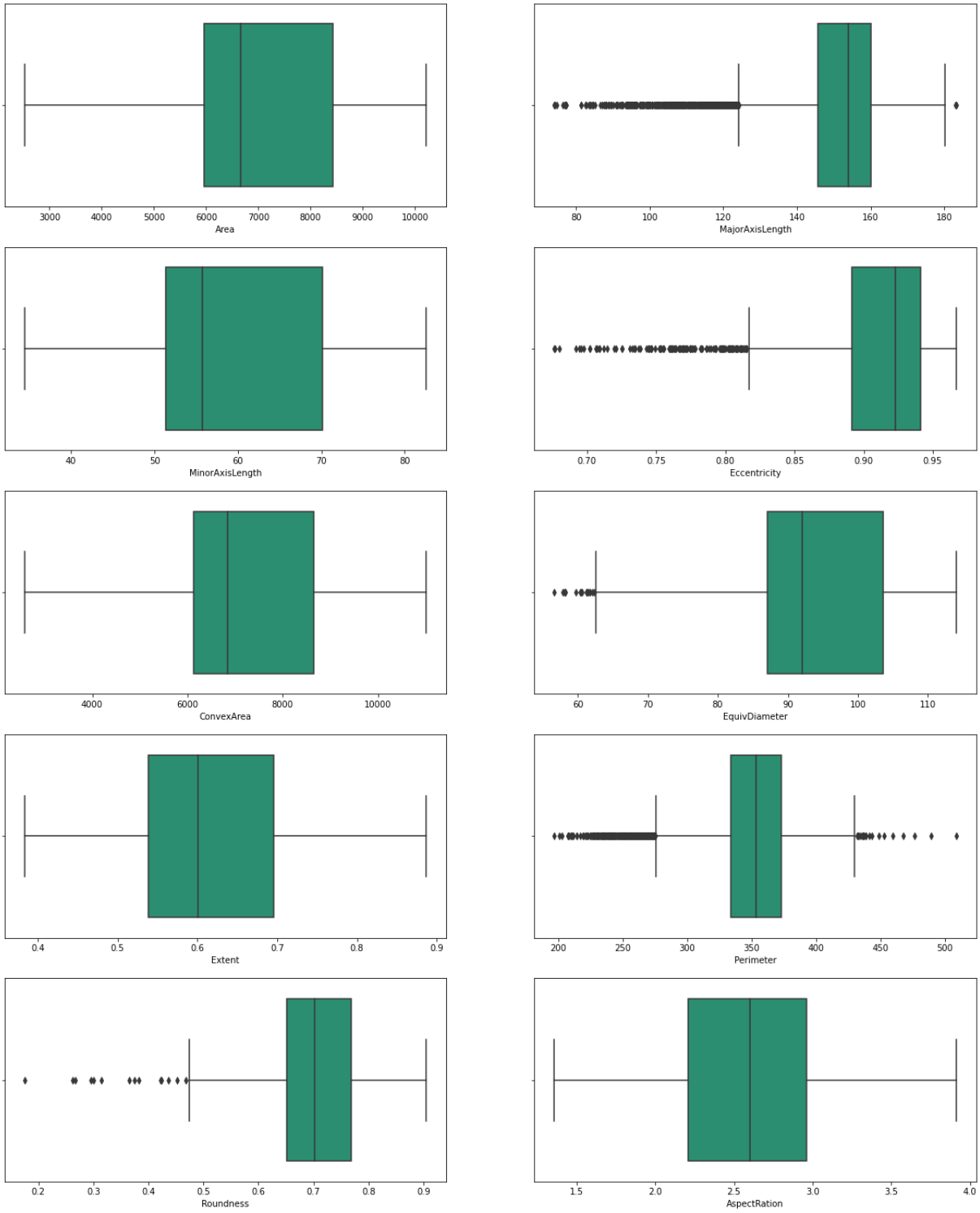


Correlation Matrix



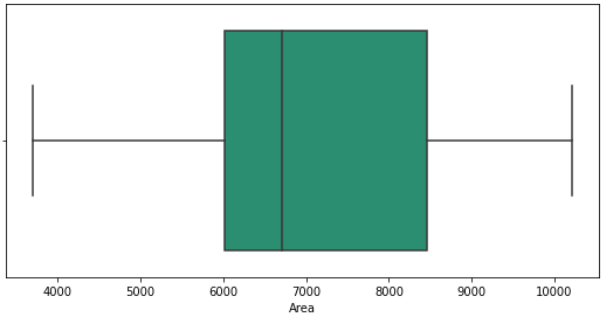
Box Plot before Outliers detection

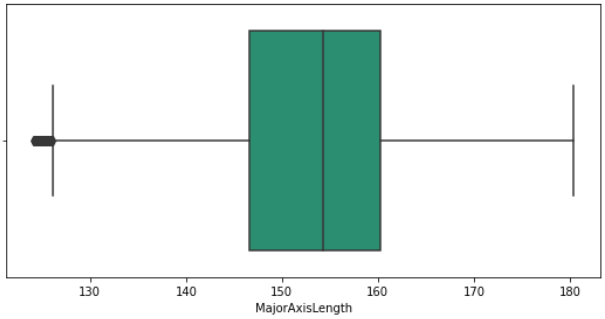
The figure below confirms that some features contain outliers.

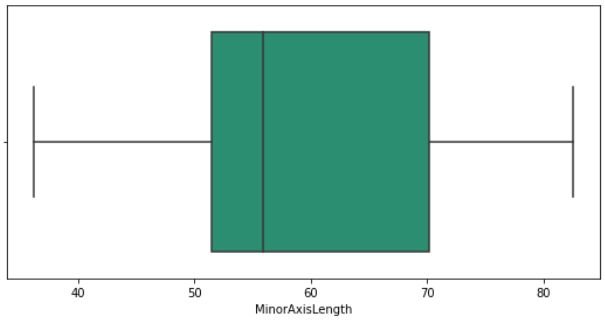


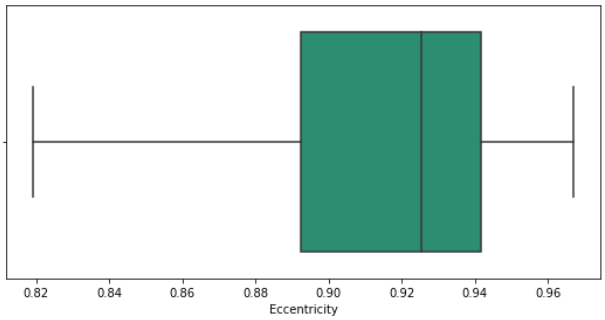
Box Plot after removing Outliers

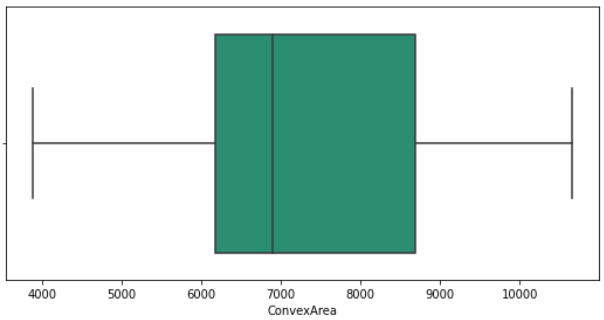
Now the outliers have been removed.

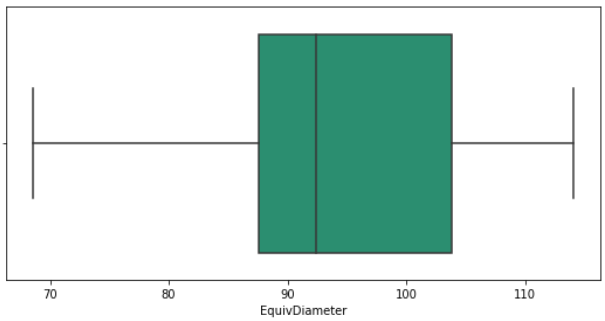


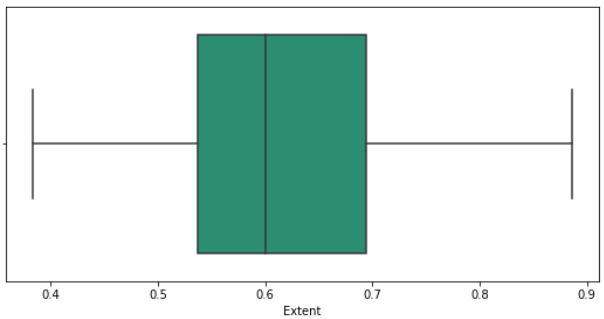


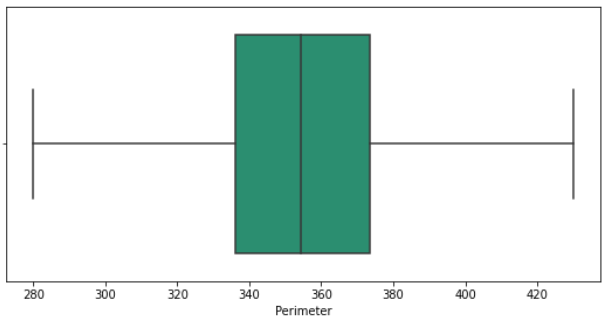


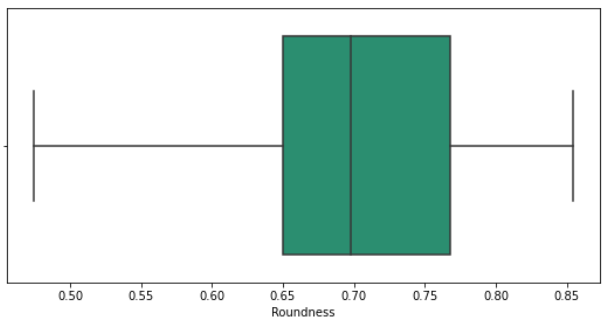


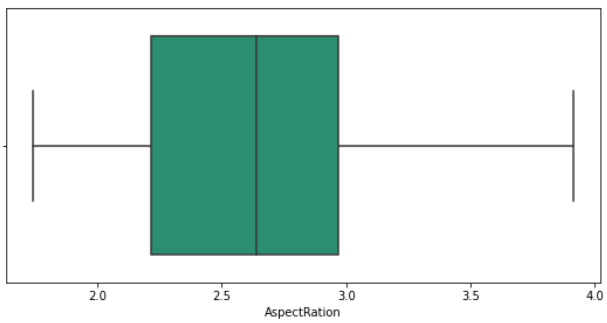




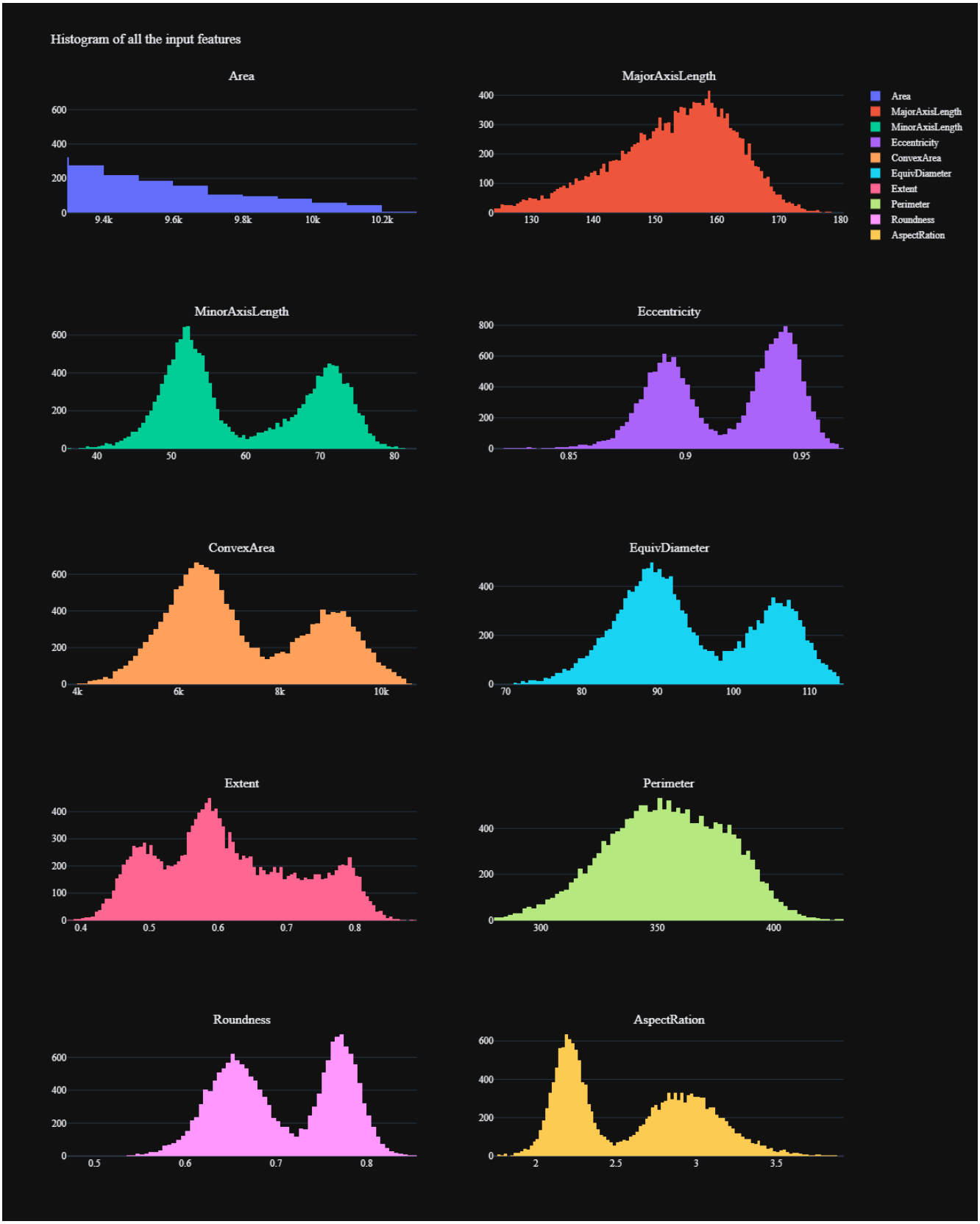








Histogram of all Features



Classification:

In machine learning, classification refers to a predictive modeling problem where a class label is predicted for a given example of input data. We have to select the input and output of the models.

Input:

As we have seen the correlation of the attributes of the dataset. Therefore, following attributes were taken as input:

- Area
- MajorAxisLength
- MinorAxisLength
- Eccentricity
- ConvexArea
- EquivDiameter
- Extent
- Perimeter
- Roundness
- AspectRation

We would use Degree Trunc, Distance,Section and Height as the input to the classification model.

Output:

'Class' is the target variable in our dataset.

- Jasmine - 1
- Gonen - 0

Model

We have split our data 30% and 70% for testing and training respectively. We applied Classification with Input Provided above and predicted Output.

We have used the following classification models:

1. Logistic Regression
   The logistic model is used to model the probability of a certain class or event existing such as pass/fail, win/lose, alive/dead or healthy/sick.

2. Naive Bayes Classifier
   A Naive Bayes classifier is a probabilistic machine learning model that's used for classification tasks. The crux of the classifier is based on the Bayes theorem.

3. KNeighborsClassifier
   KNeighborsClassifier is based on the k nearest neighbors of a sample, which has to be classified. The number 'k' is an integer value specified by the user. This is the most frequently used classifier of both algorithms.

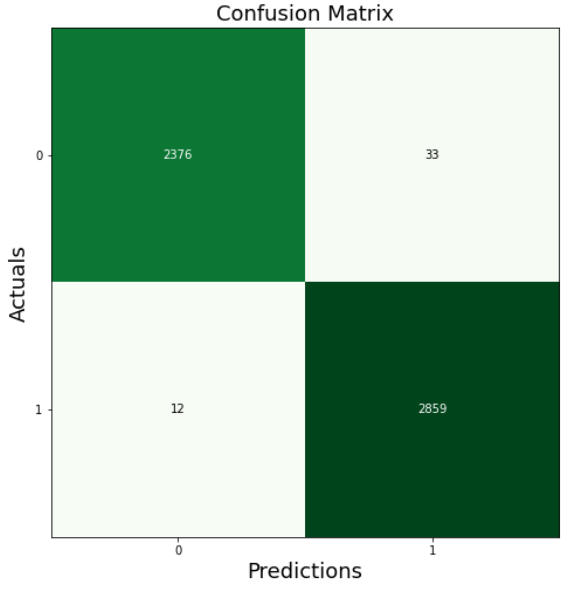
Evaluation of Models

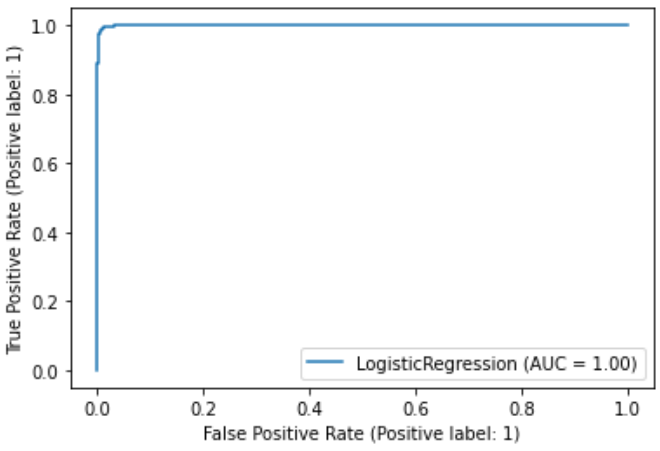
The following table shows the list of performance metrics of these models against the training data is as follows

	Logistic Regression	Naive Bayes Classifier	K-Neighbors Classifier
ACCURACY	99.14	98.90	98.86
RECALL RATE	Class 0 - 0.99 Class 1 - 1.00	Class 0 - 0.98 Class 1 - 1.00	Class 0 - 0.98 Class 1 - 0.99
PRECISION	Class 0 - 0.99 Class 1 - 0.99	Class 0 - 1.00 Class 1 - 0.98	Class 0 - 0.99 Class 1 - 0.99
FPR	Class 0 - 0.004 Class 1 - 0.013	Class 0 - 0.002 Class 1 - 0.021	Class 0 - 0.007 Class 1 - 0.015
AUC	1.00	1.00	0.99

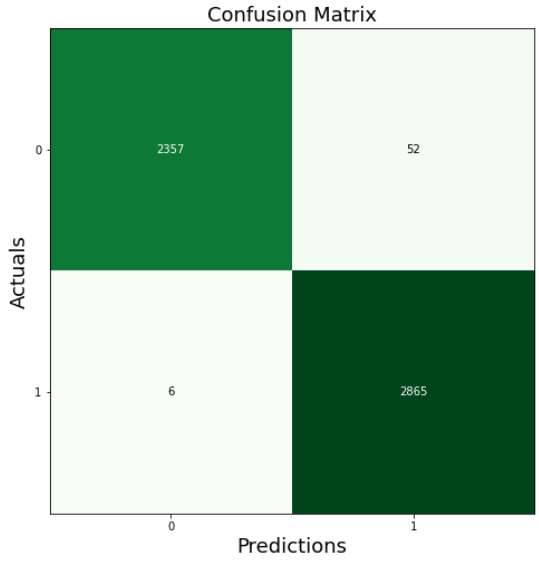
It is quite evident from the table that all models are correctly classifying the type of rice as accuracy is quite high along with high value of AUC. Logistic Regression is the most accurate.

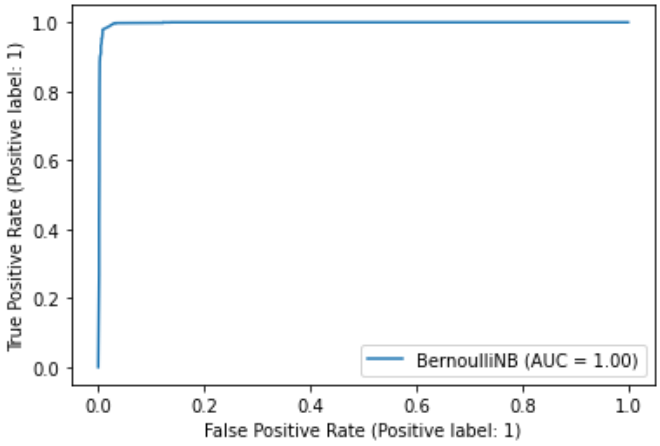
Logistic Regression



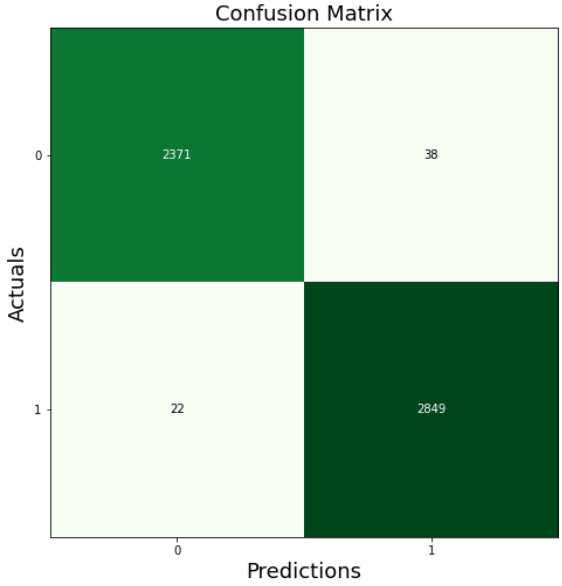


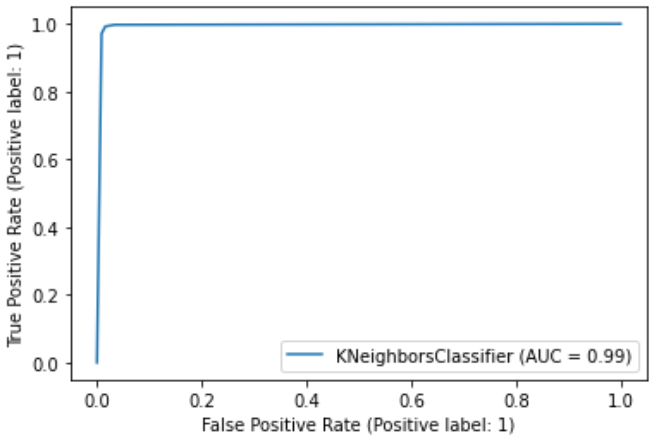
Naive Bayes Classifier





KNeighborsClassifier





Leave One Out Cross Validation

The accuracy score of LOOCV on all 3 models is 1.00. Which means all three models are generalized and working well on unseen data.

Conclusion

Conclusively, we get to know that the dataset was quite clean, and all of the features were relevant except for the 'id'. The models used were classifying the rice type accurately with high precision and low false positive rate.