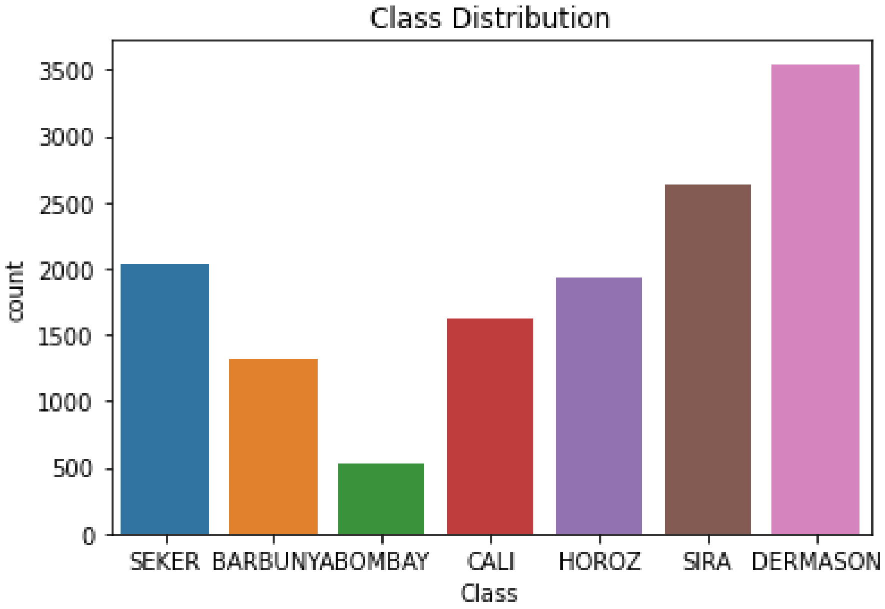
About the Dataset

Dry Bean Dataset was donated to UC Irvine Machine Learning Repository in 2020 from Selcuk University's Faculty of Technology in Turkey. A high-resolution camera was used to capture 13,611 images of 7 different types of dry beans [1]. Which is equivalent to 13, 611 instances and 7 classes. This dataset used a computer vision system to distinguish between the 7 different types of dry beans and contains no missing data. It has 16 attributes split into 12 dimensions and 4 shape form measurements [1]. The attributes are measured as integer or as continuous values. The attributes recorded in integer are area and convex area. The attributes recorded as continuous values are Perimeter, Major axis length, minor axis length, aspect ratio, eccentricity, equivalent diameter, extent, solidity, roundness, compactness, ShapeFactor1, ShapeFactor2, ShapeFactor3, and ShapeFactor4 [1]. The classes of the dry beans are Seker, Barbunya, Bombay, Cali, Dermosan, Horoz and Sira [1].

Approaches & Results

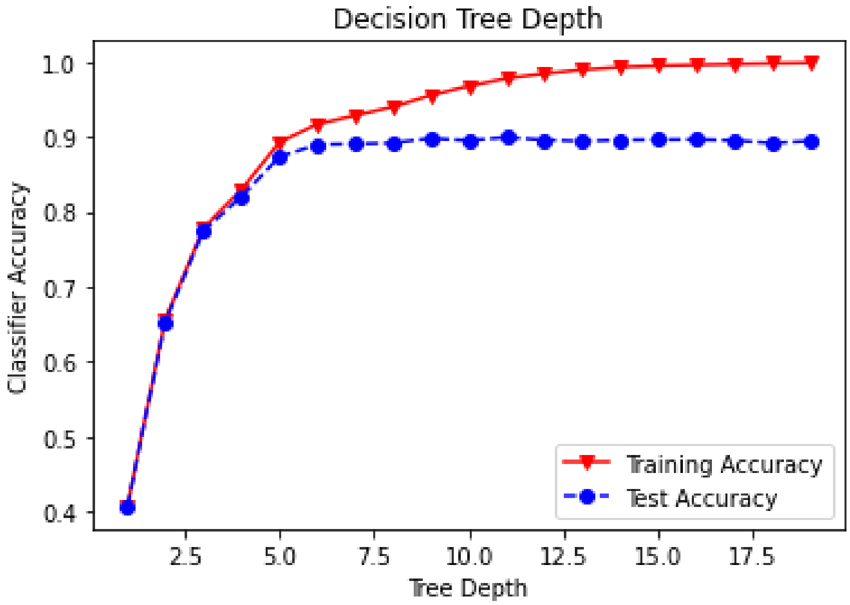
Decision Tree, K Nearest Neighbours, Naïve bayes, Support Vector Machine, and Logistic Regression classification methods were used to compare which classifier produced the highest accuracy score. Firstly, the required packages were imported, and dataset fetched from the repository. Next, the data was split into 80% training and 20% testing sets with parameter “stratify=y” to ensure that there is an equal portion of the classes. In addition, the “random_state” was set to produce comparable results across different classifiers. Lastly, the data had to be scaled because there was an imbalance between the classes.

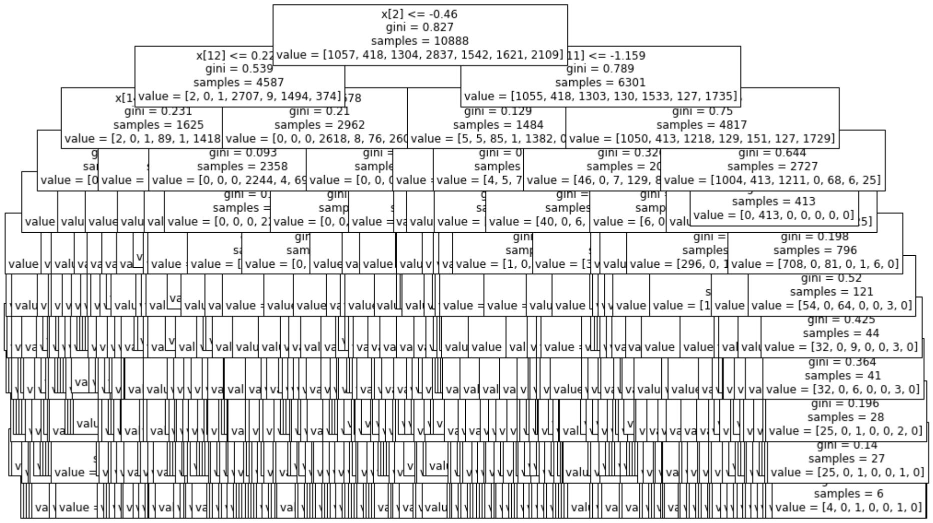
The disparity between the number of instances belonging to each of the 7 classes of dry beans is shown below. Note, Dermason has 3546 instances, whereas Bombay only has 522 instances.



Decision Tree

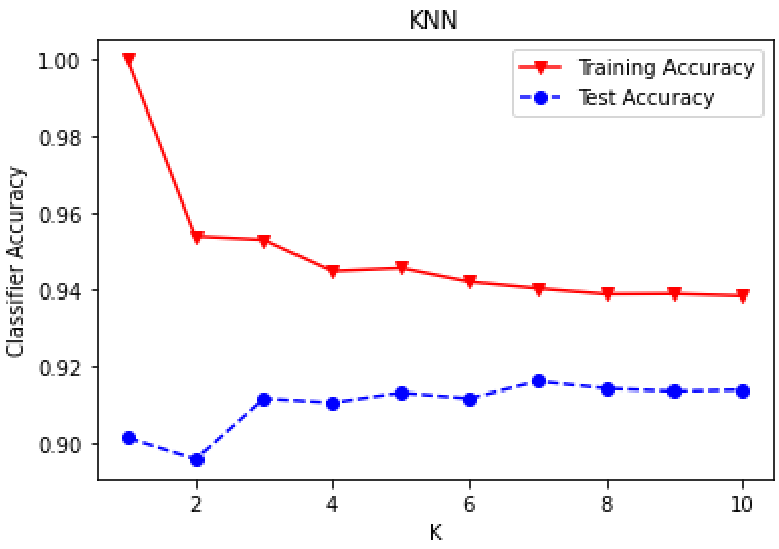
The most important attribute to building the ideal decision tree was the depth of the tree. The accuracy score varied depending on the depth. A loop was used to iterate through the tree depths from 1 to 19 to find which depth produced the highest accuracy score. When tree depth was 11, the highest accuracy score for decision tree classifier was found to be approximately 89.9%.




x[2] <= -0.46
gini = 0.827
samples = 10888
value = [1057, 418, 1304, 2837, 1542, 1621, 2109]
x[12] <= 0.22
gini = 0.539
samples = 4587
value = [2, 0, 1, 2707, 9, 1494, 374]
11] <= -1.159
gini = 0.789
samples = 6301
value = [1055, 418, 1303, 130, 1533, 127, 1735]
gini = 0.231
samples = 1625
value = [2, 0, 1, 89, 1, 1418
gini = 0.21
samples = 2962
value = [0, 0, 0, 2618, 8, 76, 260
gini = 0.129
samples = 1484
value = [5, 5, 85, 1, 1382, 0
gini = 0.75
samples = 4817
value = [1050, 413, 1218, 129, 151, 127, 1729]
gini = 0.093
samples = 2358
value = [0, 0, 0, 2244, 4, 69
gini = 0.644
samples = 2727
value = [1004, 413, 1211, 0, 68, 6, 25]
samples = 413
value = [0, 413, 0, 0, 0, 0, 0]
gini = 0.198
samples = 796
value = [708, 0, 81, 0, 1, 6, 0]
gini = 0.52
samples = 121
value = [54, 0, 64, 0, 0, 3, 0]
gini = 0.425
samples = 44
value = [32, 0, 9, 0, 0, 3, 0]
gini = 0.364
samples = 41
value = [32, 0, 6, 0, 0, 3, 0]
gini = 0.196
samples = 28
value = [25, 0, 1, 0, 0, 2, 0]
gini = 0.14
samples = 27
value = [25, 0, 1, 0, 0, 1, 0]
samples = 6
value = [4, 0, 1, 0, 0, 1, 0]

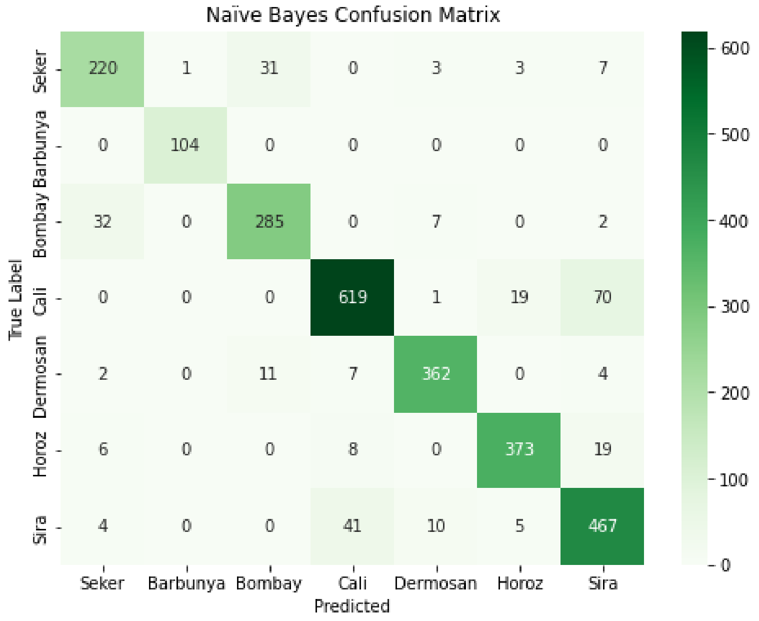
K Nearest Neighbour

The most important attribute for the Nearest Neighbour classifier is the k value. A loop was used to iterate through the k values from 1-10 to find the k value that produced the highest accuracy score. The highest testing accuracy for KNN classifier is 91.6 % when K is 7.



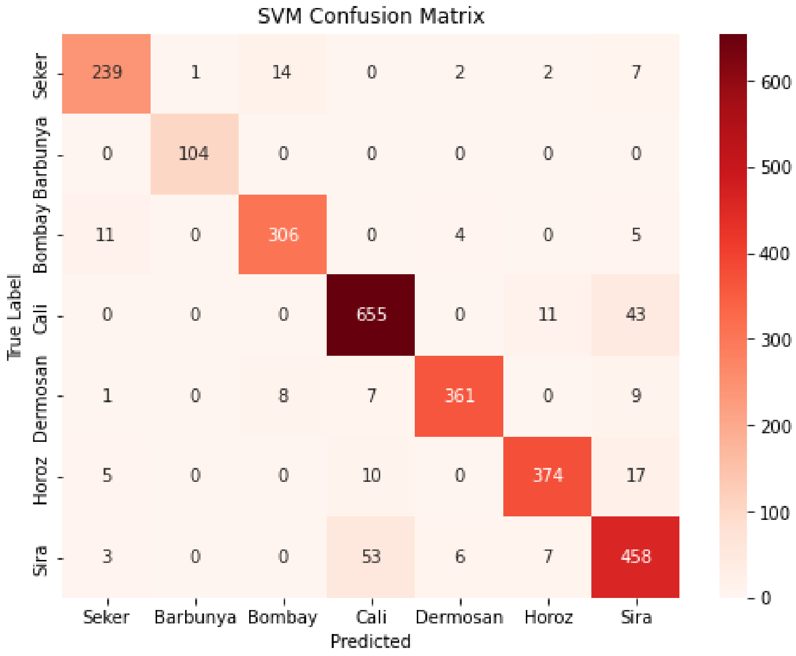
Naïve Bayes

The most important feature of Naïve Bayes is the assumption that the attributes are independent of each other. GaussianNB was created to fit and predict the test set. The accuracy for Naïve Bayes classifier is 89.2%.



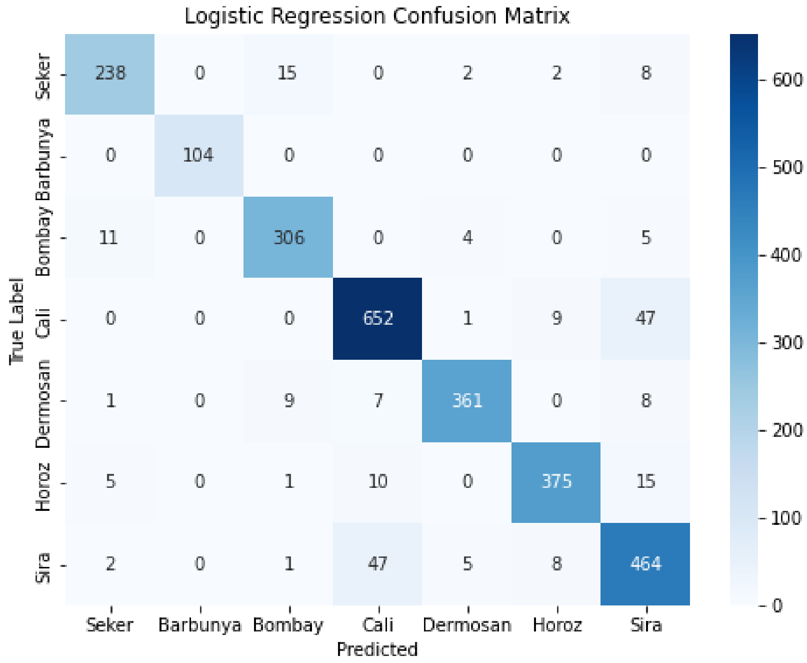
SVM

For SVM, the C value hyperparameter depends on what is prioritized. A small C value will produce larger margins and higher training errors, but it prioritizes simplicity. In contrast, a large c value will produce smaller margins and minimize the amount of training errors. For this model, the default C value of 1 was kept as it did not require adjustment to prioritize either small or large margins. As a result, the accuracy for SVM classifier is 91.7%.



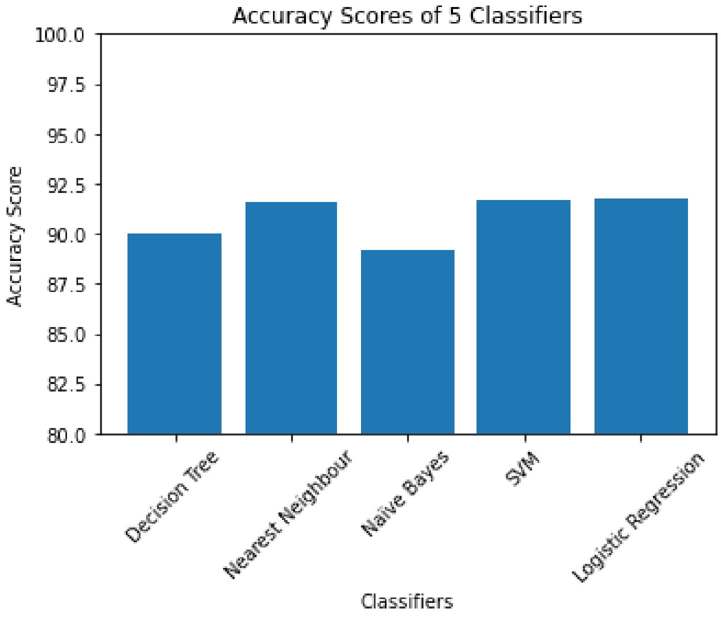
Logistic Regression

One of most important attributes in Logistic Regression is the magnitude of the feature coefficients. Typically, larger coefficients would have more influence on the classifier. Therefore, the data was standardized in order to minimize the imbalance of the coefficients of the different features. A multinomial logistic regression approach was used since the Dry Bean dataset has 7 classes. In addition, max iteration was set to 1000 for optimization and to minimize computational time. This Logistic Regression model produced an accuracy of 91.8%.



Conclusion

After comparing 5 different models for accuracy, all produced similar results. The best performing classifier was Logistic Regression at 91.8%. However, the difference between the best and worst performing classifiers was within 3%.



References

[1] Dry Bean Dataset. (2020). UCI Machine Learning Repository. https://doi.org/10.24432/C50S4B.