Jaidan Dovala
CPSC 4310
June 3, 23
Milestone III

Introduction

The purpose of this study is to identify the best classifier that accurately recognizes the type of dry beans by using a variety of classification techniques. The following methods will be employed: LR, DT, and SVM. The code for this project can be found here.

Preparing The Dataset

In order to ensure that there was no interference that would have affected the accuracy of the relevant models, we had to prepare the dataset before we could begin making predictions and training and testing it.

Steps:

I. First, we had to gather our dataset. The choice of data depends on the problem you are trying to solve. The dataset I used in this project was 'Dry_Beans_Dataset' which was collected from **UCI Machine Learning Repository** **(UC Irvine)**

II. In this phase, I searched the provided dataset for any characteristics that weren't int or float, as well as any missing data, null values, duplicates, and other features. The 'Class' feature was the sole feature in this dataset that didn't belong to one of those two kinds. Because it was categorical, I had to change it to an int.

III. Next splitting the data into training and test subsets using 90/10 technique. As it achieves slightly higher accuracy score than a 80/10 split.

IV. Lastly was feature selection. Using this technique, I was able to see what the feature importance was. Doing this is what makes the dataset unique.

Classifiers Used

Logistic Regression Classifier (LR)

Building a categorial dependent variable or a categorical outcome variable uses the probabilistic statistical classification technique known as the logistic regression model. To forecast binary or multiple answers to a categorical dependent variable, it uses more than one independent variable.

The Logistic Regression model was trained using the top 10 selected features obtained through the SelectKBest feature selection method. It achieved a high training score of 91.26% and a respectable testing score of 92.14%. This indicates that the model performed well in predicting the type of dry beans based on the selected features.

Accuracy:
Logistic Regression Model Train Score is: 0.9125642909625276
Logistic Regression Model Test Score is: 0.92143906020558

Decision Tree Classifier (DT)

The easiest and most straightforward classifiers are decision trees. The pattern is given a class number via a decision tree that filters it through a test, with an internal node representing tests and leaf nodes representing categories.

The Decision Tree Classifier was trained with a maximum depth of 12 and the Gini criterion. While it achieved a high training score of 96.07%, the testing score was slightly lower at 89.21%. This suggests that the model might have overfit the training data, resulting in reduced generalization performance on unseen data.

Accuracy:
Decision Tree Model Train Score is: 0.960731488284758
Decision Tree Model Test Score is: 0.8920704845814978

Support Vector Machine (SVM)

The goal of a support vector machine is to identify an ideal decision boundary that, using a hyperplane, divides the n-dimensional feature vectors into two classes. To train the Support Vector Machine model and move the feature vectors into a higher dimensional space, a kernel is used. After that, the machine learning problem is treated as a convex optimization problem.

The dataset was split into training and testing sets, and the SVM model was trained on the training set. A linear kernel was chosen for the SVM classifier. The model achieved a training score of 91.60% and a testing score of 91.80%.

Accuracy:
Support Vector Machine Model Train Score is: 0.9160462382445141
Support Vector Machine Model Test Score is: 0.9180135174845724

Results

Based on the dataset for dry beans' analysis. The Logistic Regression model performed well in predicting the type of dry beans, earning training and testing scores of 0.913 and 0.921, respectively. A potential overfitting on the training data is indicated by the Decision Tree classifier's slightly lower testing score of 0.892 but higher training score of 0.961. Last but not least, the Support Vector Machine model produced a training score of 0.916 and a testing score of 0.918, showing consistent performance on both the training and testing datasets. These results imply that the task is well suited for Logistic Regression and Support Vector Machine, whereas the Decision Tree model should be used with caution due to the danger of overfitting.