Multiclass classification of dry bean using Machine Learning

1. Introduction

Dry bean (Phaseolus vulgaris L) is the most important & most produced pulse in the world (Koklu et al., 2020). The United States, as a global leader in dry bean production produces between 1.5 to 1.7 million acres of edible dry beans a year. Seed classification, whether manual or automatic, is essential for both marketing & production of dry beans (Koklu et al. 2020). Manual classifications are difficult & require a lot of time, which essentially makes it an inefficient process. ML classifications like the one developed in this analysis can help automate these process & allow farmers to bring to market well sorted dry beans. This will also increase the value of the seeds.

2. Objective

The primary objective of this analysis is to develop a classification model that can predict the value of a harvest from a population of cultivation from a single farm that has been presented at market. The net worth of a single pound of dry beans from harvest will be predicted using statistical learning methods that were taught in Modern Applied Statistics I & II at SDSU.

3. Methodology

The initial analysis process will begin by performing exploratory data analysis (EDA). This phase is a critical step to understand the distribution of the data & uncover patterns or trends that may not be immediately apparent.

Principal component analysis (PCA) will be used for feature decomposition order to reduce the dimension of the data. This will allow us to understand & visualize in a 2-dimensional plane the intricate relationships between the covariates while maintaining as much information as possible from the original dataset.

Furthermore, boruta algorithm which is a wrapper built around random forest classification algorithm will be utilized to isolate important features that will be used to construct various competing classification models.

Model evaluation is an integral part of statistical modeling aimed at gaging a models' ability to explain the functional form of the data. In predictive analytics, this is typically done by evaluating performance metric such as: 1) accuracy; 2) sensitivity; 3) specificity; 4) CV error rate; 5) ROC curve. Finally, linear regression will be utilized in order to predict the value of 1-pound sample of dry beans.

4. Exploratory Data Analysis

The aim of this section is to properly address the following about the data: 1. statistical properties of the explanatory variables by means of summary statistics. 2. distributions of the independent variables to ensure normality assumption is not violated.

Our labeled dataset contains 3,000 rows & 8 columns. Each row represents a sample observation or an individual dry bean that belongs to a particular class of white beans. 7 out of the 8 columns are morphometric measurements recorded in each observation in the form of pixel count. The remaining column is 'class', which is the categorical target variable containing 6 classes of beans. Table 1 in the analysis paper contains the summary statistics of the dataset.

Since many of the commonly used classification algorithms implicitly assume input data is normally distributed, we check the distribution of all explanatory variables by class. For example, the QQ-plots show normal distribution in almost all the variables with the exception of few variables, namely 'Area', 'Perimeter', & 'Extent'. The distribution of 'Area' for the Bombay bean seems to have observations that occur outside of the confidence band indicating potential presence of outliers. 'Perimeter' is even more uncooperative in following the normality line & 'Extent' is pushing the boundary of breaking the normality assumption.

The presence of outliers is further supported by the box plots that demonstrate left skewness in 'Area' & its associate, e.g., 'MajorAxisLength'; in contrast, 'Extent' seems to be right skewed. The boxplots suggest that these variables are related, meaning if 'Area' contains outliers it is highly likely those related variables will contain outier as well. It is also concerning to notice that 'Extent' is directly calculated from 'Area'. These relationship & potentially other hidden relationships can cast ambiguity in the interpretation of coefficient estimates of models constructed with these highly correlated variables. That is, the unique contribution of the individual explanatory variable to the model might become difficult to isolate.

It is evident in the density plots that some classes of beans have unique distributions discernible from other classes of beans. For example, Bombay's area distribution is significantly different from other classes of beans, rarely overlapping with any of them. This is also true for variables like 'MajorAxisLength', 'MinorAxisLength' & 'ConvexArea' as they are related to 'Area'. Hence, a model will likely find Area or any of its related variables to be significant.

Moreover, 'Eccentricity' which measures how different a shape is from true circle indicates Seker bean distribution having significantly different distribution than other classes. Notice 'Extent' wildly overlapping distribution for all 6 classes of beans. This might a be a precursor for redundant predictor.

5. Principal Component Analysis

we have established that many of the explanatory variables exhibit high level of positive correlations. This multicollinearity will complicate model interpretability even if the overall accuracy of the model is not downgraded. Hence, performing a principal

component analysis (PCA) is highly appropriate to help us choose a smaller number of representative principal components that collectively explain most of the variability in the original dataset.

To Perform PCA on the original labeled dataset, it is important to examine the variance & the mean of each individual explanatory variable. Interestingly enough, we see that 'Area', 'perimeter' & 'ConvexArea' seem to have extremely high mean & variance than the other explanatory variables. To account for this discrepancy in the data, we standardized the data so the mean is 0 & the standard deviation is 1.

According to the PCA literature, there is no unified interpretation of the principal components. However, one can interpret the first principal component as positively correlated with 'Area', 'Perimeter', 'MajorAxisLength', 'MinorAxisLength', & 'ConvexArea'. Increasing any of these variables will positively increase the other positively correlated variables by about 42-45%. This suggests that beans that have larger area will also have larger perimeter, major axis length, minor axis length, & convex area; the opposite is also true.

To summarize, the first principal components explain about 68% of the total variance in the data, followed by the second principal which captures approximately 20% of the variability in the data. The third principal component explains about 9% of the variance. Together the first, second & third principal components account for 96% of the variance in the dataset. Therefore, we will use the first three principal components to fit multiple classification models.

6. Model Fit

K-Nearest Neighbor

Initially, a KNN model using all the features of the dry bean was built. That resulted in an optimal k of 26 neighbors using 10-fold cross-validation. We then used the trained model to validate it on the partitioned 20% of the labeled dry bean. In this validation stage we obtained a model accuracy of 87%. This KNN model used Bombay as a reference point & predicted all of the 159 observations that were Bombay. The KNN model gave us 100% classification rate for Bombay, 95% for Cali, 79% for Dermason, 82% for Horoz, 89% for Seker, & 76% for Sira. The confusion matrix as well as model performance metrics are summarized in tables in the annotated analysis report attached to this submission.

Moreover, we built a KNN model using the principal components that we obtained from the principal component analysis of the dataset. Three principal components were selected that explained 96% of total in the data. Using only those three principal components we obtained a model accuracy of 86%.

Linear Discriminant Analysis

We fitted an LDA model with all the feature & then we fitted another LDA model with variables extracted from Boruta package. These variables were: Area, Eccentricity & Extent. The output of the initial model that contains all features from the original data set shows the prior probabilities, which has equal proportions of classes in the training data, group means, coefficients of linear discriminant, & Proportion of trace. The value for each LD is scaled which means it will have a mean of 0 and variance of 1. The group mean shows the mean values for different predictors when they fall into a predicted class. for instance, there is a clear difference between the proportion of Areas & Perimeter between different white bean classes. The variation in mean groups does not exist as much for variables Eccentricity & Extent. The full LDA model had an accuracy of 84.22% while the reduced LDA model through Boruta method an accuracy of 85.89%.

Quadratic Discriminant Analysis

Similarly, we fitted a QDA model with all the features & the reduced features. The output shows the prior probabilities, which shows equal proportions of classes in the training data, & group means. The QDA model output shows the prior probabilities, which shows equal proportions of classes in the training data, & group means. As a result of performing QDA, we obtained an overall accuracy of 89.55%. The other important model metrics are summarized in tables in the analysis paper.

7. Model Comparison

In order to decide which model generalized the data, the model accuracy as well as the confusion matrix table along with the ROC curves will be used as basis for deciding which model performed best. The class-specific sensitivity & specificity will also be used as well & are all summarized in tables in the analysis report attached to this submission. The model accuracies are as follows:

8. Dry Bean Value

Furthermore, linear regression model was created in order to predict the price of 1 lb of dry beans for 3 different samples. Moreover, the data was bootstrapped in order to provide some value of uncertainty and confidence in the value prediction. Sample A was predicted to have a value of $4.78 with a 95% CI of ($4.68, $4.88); sample B has value of $3.15 with a 95% CI of ($3.00, $3.30); sample C has a value of $3.77 with a 95% CI of ($3.64, $3.89).

Multiclass classification of dry bean using Machine Learning

9. Conclusion & Recommendation

Accurate prediction of dry bean is of most importance due to the dire need to make the process of segregating different types of dry beans more efficient. We have trained various models to see which statistical algorithm would suit such application. Some of the variables have nonlinear relationships and the LDA models did not perform as well as QDA. The two best competing models for dry bean classification application are KNN & QDA and both had relatively good accuracies. These two models were able efficiently segregate the 6 classes of dry beans with relative ease. The most efficient model was the QDA and had an accuracy of almost 90%.

Recommendation

1. The use of methods that can adequately deal with nonlinearity is particularly useful in prediction of dry beans.

10. Reference

1. *Nabi, J. (2018, December 23). Machine Learning — Multiclass Classification with Imbalanced Dataset. Medium. https://towardsdatascience.com/machine-learning-multiclass-classification-with-imbalanced-data-set-29f6a177c1a*
2. *9.2.3 - Optimal Classification | STAT 508. (2021). PennState: Statistics Online Courses. https://online.stat.psu.edu/stat508/lesson/9/9.2/9.2.3*
3. *Mutuvi, S. (2019, April 9). Introduction to Machine Learning Model Evaluation - Heartbeat. Medium. https://heartbeat.fritz.ai/introduction-to-machine-learning-model-evaluation-fa859e1b2d7f*
4. Kassambara, Charles, Visitor, Kassambara, & SFd. (2018, March 11). Discriminant analysis essentials in R. Retrieved May 07, 2021, from http://www.sthda.com/english/articles/36-classification-methods-essentials/146-discriminant-analysis-essentials-in-r/#:~:text=Discriminant%20analysis%20is%20used%20to,one%20or%20multiple%20predictor%20variables.&text=Linear%20discriminant%20analysis%20(LDA)%3A,class%20of%20a%20given%20observation
5. Linear vs. Quadratic Discriminant analysis - comparison of algorithms. (2021, January 15). Retrieved May 07, 2021, from https://thatdatatho.com/linear-vs-quadratic-discriminant-analysis/
6. Shekhar, P. (2018, January 04). How to perform logistic Regression, LDA, & QDA in R. Retrieved May 07, 2021, from https://datascienceplus.com/how-to-perform-logistic-regression-lda-qda-in-r/#:~:text=QDA%20Assumption%3A,%2Dspecific%20covariance%20(%CF%83k2).
7. Hastie, T., Friedman, J., & Tisbshirani, R. (2017). *The Elements of statistical learning: Data mining, inference, and prediction*. New York: Springer.
   WandeRum. (n.d.). WandeRum/multiROC. Retrieved May 07, 2021, from https://github.com/WandeRum/multiROC
8. *Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. An Introduction to Statistical Learning: with Applications in R. New York :Springer, 2013.*