# Multiclass classification and price prediction of dry beans using Different Algorithms

# A white paper on statistical Analysis on *Dry Beans* Data Set

## Introduction

Dry bean cultivation is practiced in Turkey and Asian countries usually in the form of populations containing mixed species of seeds. As different populations containing different genotypes are cultivated, the final products contain different species of seeds. Finally, when the dry bean seeds obtained from population cultivation are released to the market without being separated by species, the market value decreases immensely. Because of the wide range of genetic diversity of dry bean, seed classification is essential for marketing and production. With an aim for establishing a sustainable agricultural system, Koklu et al. provided methods for obtaining uniform seed varieties from crop production. Additionally, Koklu et al. created distinct multiclass classification algorithms for seven different registered varieties of dry beans with comparable characteristics to achieve consistent seed categorization which also addressed the problems of market value.

Our aim of this project to explore the potential of the various learning methods and give a recommendation on which type of algorithm we should further develop for the task of sorting white beans. We are expected to develop an automated method that predicts the value of a harvest from a 'population cultivation' from a single farm that has been presented at market. Since each of the beans has a different value at market the cost of an error depends on the actual type of white bean and what it is predicted class. Additionally, we will provide some measure of our predictive systems.

## Problem Statement

The given data contains *3000 observations* with *9 variables* describing the structural morphology of dry beans. Among these observations, variable '*CLASS*' will be used as a target variable. We have two main objectives for our *Dry Beans* Data. The objective of this study is to build multiclass classification algorithms and calculate the accuracy of the built models. Furthermore, we will predict price for different models and choose the best one.

## Methodology

### Data Analysis

The data set is taken from the given 'labeled.csv' in R and has total of 3000 observations in 9 variables. For our convenience, we have ignored variable 'X' indicating index in our dataset. Using summary statistics, we have gathered an idea about statistical distribution of features of dry bean varieties (in Pixels). There are large differences in the range of variables, the variables with larger ranges can dominate over those with small ranges which may lead to biased results. We have also made a class covariance among six classes and had an idea about the class variances which measures variability from the average or mean. The variance of each variable by class shows evidence of non-constant variance.

To start off the *Exploratory Data Analysis*, we have plotted *density plots*, *histograms*, and *boxplot* to see the status of distribution, variability, and presence of any outliers. Density plot represents distribution of a numeric variable using kernel density estimate to show the probability density function of the variable. We have seen bimodal distribution in data for variable '*Perimeter*' and class '*BOMBAY*' is off from other classes. The histograms from the labeled data reveal that the variables exhibit multimodal behavior.  This indicates that at least one of the bean classes differs greatly from the others. Additionally, boxplot (variables 'Beans Area' and 'Perimeter' with respect to each class) displayed the distribution of data based on a five number summary ("minimum", first quartile (Q1), median, third quartile (Q3), and "maximum"). Median value for the class 'Bombay' was the highest and 'Dermason' was lowest. We can also infer variability within and in between classes from the boxplot. The class 'Bombay' has the highest variability. To our next step, we have used pairs plot and correlation matrix for feature selection process. But we saw that most of the variables except for Eccentricity and Extent are highly correlated with each other. To validate our inference, we have further done two-way-anova for all our numeric variables (*Numeric variables~Class*). For the Two-way ANOVA, we assumed the observations within each class are normally distributed and have equal variances. All the p-values are less than the significant level $\alpha$=0.05 meaning they have significant differences in the mean values for all of 6 classes. Based on p-values, we have decided to keep all the variables in our model and then developed algorithms accordingly.

## Statistical Model Development

We have divided our data set into 60% (1800) training and 40% (1200) test data. Using Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Mclust, Mclust 'EDDA' and KNN algotihms, we created confusion matrix for actual and predicted model and calculated accuracy. From the actual and predicted class, we tried to compare the price of a particular farm having different types of beans. LDA model is a dimensionality reduction technique used to solve classification problems. Using the Linear combinations of predictors, LDA tries to predict the class of the given observations. QDA Model is the general form of Bayesian discrimination. Like LDA, the QDA classifier results from assuming that the observations from each class are drawn from a Gaussian distribution and plugging estimates for the parameters into Bayes' theorem to perform prediction. Mclust and Mclust EDDA are powerful and popular packages which allows modelling of data as a Gaussian finite mixture with different covariance structures and different numbers of mixture components.

*LDA Model*: Basically, it helps to find the linear combination of original variables that provide the best possible separation between the groups. Based on the training dataset, each has 16.67% data. The "proportion of trace" that is printed when I type "linear" (the variable returned by the lda() function) is the percentage separation achieved by each discriminant function. For example, for the Beans Data we get the same values as just calculated (84.22%, 33.79%, 12.53%, 6.26% and 1.51%). We have further created confusion matrix based on training and test data. The accuracy of our LDA model is 84.88% (Training Data) and 87.41% (Test Data). The results obtained from train and test sample almost equal. Furthermore, we have predicted price based on LDA model. For our convenience, we have converted each class of beans as factor:  Bombay=1,Cali = 2 Dermanson = 3, Horoz=4, Seker =5, Sira = 6 Actual Price on test data is $8.62 and $8.643.

*QDA Model*: Like LDA, the QDA classifier results from assuming that the observations from each class are drawn from a Gaussian distribution and plugging estimates for the parameters into Bayes' theorem to perform prediction. However, unlike LDA, QDA assumes that each class has its own covariance matrix. We have created confusion matrix based on training and test data. The accuracy of our QDA model is 90.22% (Training Data) and 91.41% (Test Data). The results obtained from train and test sample almost equal. Furthermore, we have predicted price based on QDA model. Actual Price on test data is $8.62 and $8.63.

*MclustDA and Mclust EDDA Model*: Finite mixture models are being used increasingly to model a wide variety of random phenomena for clustering, classification, and density estimation. mclust is a powerful and popular package which allows modelling of data as a Gaussian finite mixture with different covariance structures and different numbers of mixture components, for a variety of purposes of analysis.

The accuracy of our MclustDA model is 91.41% (Training Data) and 92.5% (Test Data). The results obtained from train and test sample almost equal. Furthermore, we have predicted price based on QDA model. Actual Price on test data is $8.62 and $8.66.

On the contrary, we got less accuracy for the EDDA model with 89.58% (Training Data) and 89.41% (Test Data). The results obtained from train and test sample almost equal. Furthermore, we have predicted price based on QDA model. Actual Price on test data is $8.62 and $8.66.

## Conclusion

The 3000 observations containing 9 variables were analyzed using multiclass classification algorithms. Four different models with different variables were produced and examined to predict classes for the test data set. Based on accuracy, we can say the sequence would be*: McIsutDA > QDA > MclustEDDA > LDA* . Again, if we look at the BIC of MclustDA and MclustEDDA. MclustDA has the greatest BIC. So our conclusion is further validified with the BIC value from the two models. After price prediction, we have seen that MclustDA and MclustEDDA would perform better.

So, combining accuracy and price prediction, I would select ‘MclustDA’ for building future model.