Multiclass Classification and Price Prediction of Dry Beans

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.

The purpose of this project is 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. In doing so, we will 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 [Figure 2] the cost of an error depends on the actual type of white bean, and it’s predicted class.

The algorithm selection will be based on the algorithm that produces the highest accuracy in classifying the bean type. The algorithm with the highest classification accuracy will serve to minimize the market value cost of classification errors between dry bean varieties.

Problem Statement

There are two datasets used in this project: The *labeled* and *unlabeled* dataset. The labeled dataset, which contains the classes of dry beans, will be used to train the various machine learning models, hence also referred to as the training dataset. The training dataset contains 3000 observations and 8 variables. The dependent variable has 6 levels (Classes): BOMBAY, CALI, DERMASON, HOROZ, SEKER, and SIRA. These classes represent the different varieties of dry beans. Each class has 500 observations in the training set.

The unlabeled dataset is made up of the combination of three separate samples, namely, Sample A, B, and C. There are 7 features which does not include Class, thus making the data unlabeled. The objective of this study is to build multiclass classification algorithms and select the best model based on the highest accuracy achieved in predicting bean type in the labeled data. Using this best selected model, we will predict the market value of a one-pound sample of beans from each cultivation in the unlabeled data set. Finally, we will give a measurement of the accuracy of the market value we predicted by providing a 95% confidence interval of the price that should be paid for each one-pound sample.

Data Exploration

The data set is taken from the provided ‘labeled.csv’ dataset and has total of 3000 observations with 9 features [Figure 1]. 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 features. The features with larger ranges may dominate over those with small ranges which may lead to biased results. Features with larger ranges may benefit through feature engineering.

To start off the *Exploratory Data Analysis*, we have plotted *density plots* [Figure 3], and *boxplots* [Figure 4 & 5] 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. Boxplot (variables *Log_Area* and *Eccentricity* with respect to each class) display the distribution of data based on a five number summary (“minimum”, first quartile (Q1), median, third quartile (Q3), and “maximum”). The range is the highest for the *BOMBAY* class for *Log_Area* and *SEKER* for *Eccentricity*. We can also infer variability within and in between classes from the boxplot. The class ‘Bombay’ has the highest variability. For our next step, we have used a covariance matrix that explicitly provides the correlations between the features [Figure 6]. This plot shows that many features are highly correlated and that, due to this, only a subset of the features should be needed for use in the development of the classification algorithm.

Feature Selection

To select which features to use in the development of the algorithms, feature selection is applied using best subset selection [Figure 7] as described in sections 6.1 and 6.5 of James, Witten, Hastie & Tibshirani. In this selection method, a least squares regression is applied for each possible subset of features in a standard linear model. The best subset of features is then selected based on the best adjusted R squared measurement of the associated model. Using this method, the model with the highest adjusted R-squared value used three features, Area, Eccentricity and Extent.

Statistical Model Development

In developing the algorithms, the *Holdout Method* was used in which 2000 of the 3000 samples were randomly selected to be used as the training data and remaining 1000 samples were used for the test data. Using *Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), KNN, Random Forest and GAM* algorithms, we created confusion matrices for actual vs predicted bean class and have provided accuracy, precision and recall for each model [Figure 8]. To deal with classification issues, LDA is a dimensionality reduction method. For example, LDA uses Linear combinations of predictors to forecast an observation's class. Discrimination with QDA is the most common kind of Bayesian discrimination. To make predictions, the QDA classifier uses estimates of the parameters from the Bayes theorem to assume that the data for each class are selected from a Gaussian distribution. An unsupervised learning algorithm, KNN is utilized for classification and regression. Random forest is a regression technique that employs the ensemble learning approach for regression. It is possible to expand a typical linear model by including non-linear functions for each of the variables, while keeping the model's additivity. GAMs may be used with both quantitative and qualitative answers, much as linear models.

From the performance measures table, we can see that GAM performs the best with 90.90% accuracy, 91.05% precision, and 90.91% Recall. QDA performs second best followed by Random Forest, LDA and KNN.

After selecting the best model based on accuracy, we have used our *Generalized Additive Model (GAM)* in predicting number of beans in the unlabeled data set (Sample A, B and C) along with the bean value (in USD) per class. Note that the GAM model uses feature engineering where a natural spline with four degrees of freedom was used on the log transform of the Area feature. Bean Counts were calculated by predicting bean class for each cultivation using the GAM model [Figure 10]. Bean Values were calculated by taking the grams per seed for the predicted bean class, divided by grams per pound and multiplied by dollar per pound for the predicted bean class [Figure 11]. The bootstrap is a widely applicable and extremely powerful statistical tool bootstrap that can be used to quantify the uncertainty associated with a given estimator or statistical learning method. We have illustrated the bootstrap on Sample A, B and C with 95% confidence interval in which we wish to determine the best investment allocation under a simple model [Figure 12].

Conclusion

From the above analysis, the GAM algorithm is recommended as the best model to use in classifying between white bean varieties. After selecting the GAM model, we have predicted beans count and values in dollar for the unlabeled data set [Figure 10 & 11]. The exact dollar amount calculated per bean classification error has been calculated by subtracting the incorrectly classified bean market price from the actual bean market price. Using the bootstrap approach, a 95% confidence interval was produced which serves as an estimate of the accuracy of the prediction of the market value of a one-pound bean sample from each cultivation [Figure 12].

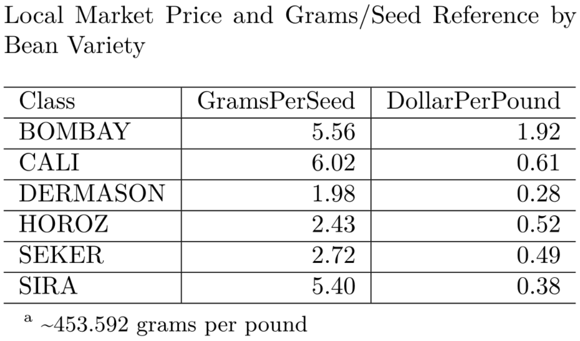
Appendix

Figure 1

Data Feature Descriptions (Koklu & Ozkan, 2020)

- Area (A): The area of a bean zone and the number of pixels within its boundaries.
- Perimeter (P): Bean circumference is defined as the length of its border.
- Major axis length (L): The distance between the ends of the longest line that can be drawn from a bean.
- Minor axis length (l): The longest line that can be drawn from the bean while standing perpendicular to the main axis.
- Eccentricity (Ec): Eccentricity of the ellipse having the same moments as the region.
- Convex Area (C): Number of pixels in the smallest convex polygon that can contain the area of a bean seed.
- Extent (Ex): The ratio of the pixels in the bounding box to the bean area.
- Class: One of the six bean types/varieties (BOMBAY, CALI, DERMASON, HOROZ, SEKER, SIRA)

Figure 2



Local Market Price and Grams/Seed Reference by Bean Variety

Class	GramsPerSeed	DollarPerPound
BOMBAY	5.56	1.92
CALI	6.02	0.61
DERMASON	1.98	0.28
HOROZ	2.43	0.52
SEKER	2.72	0.49
SIRA	5.40	0.38

[a] ~453.592 grams per pound

Figure 3

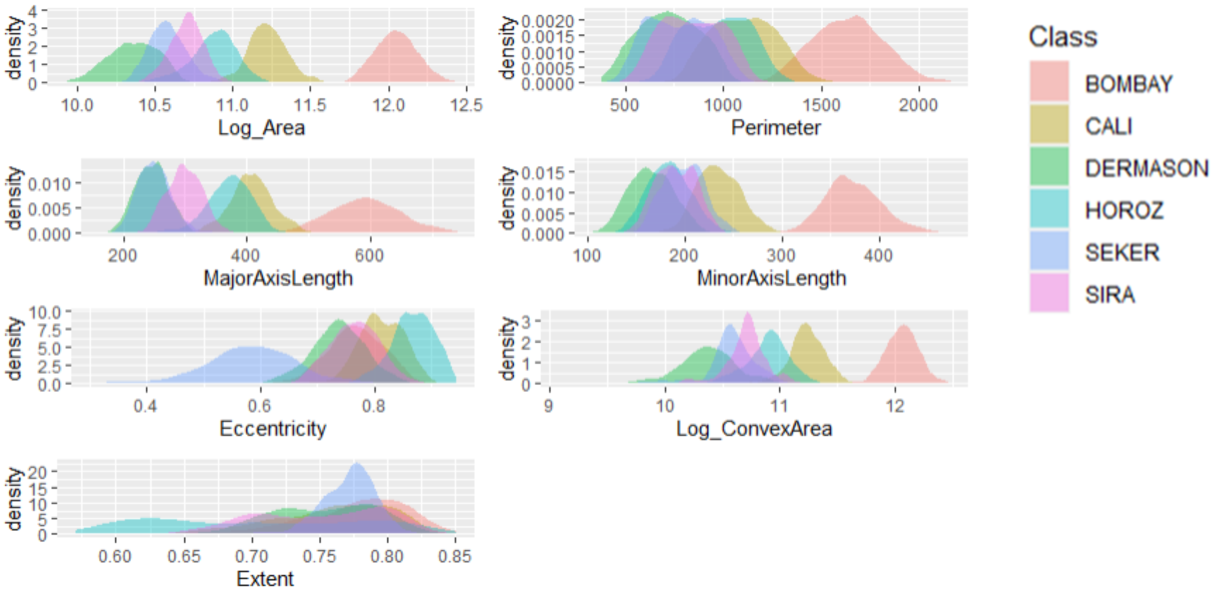


Figure 4

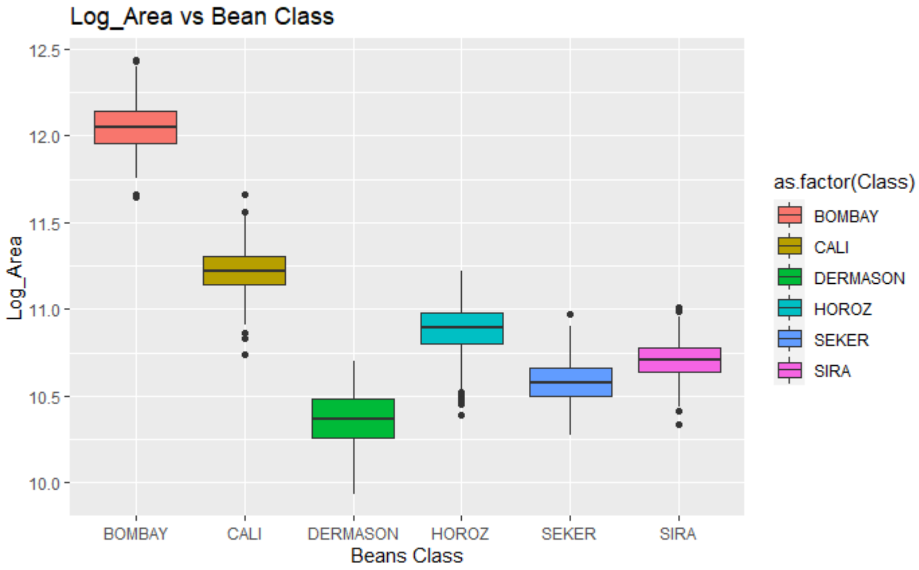


Figure 5

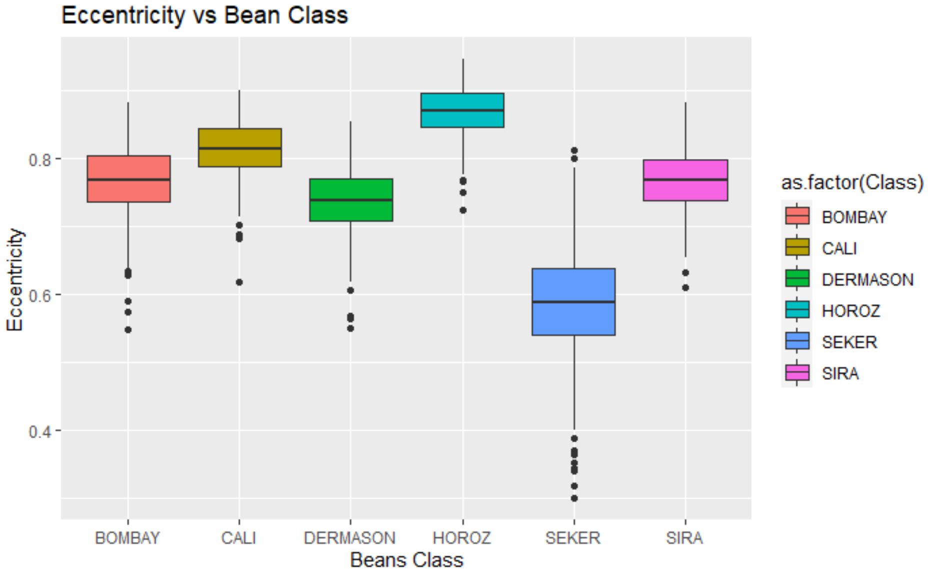
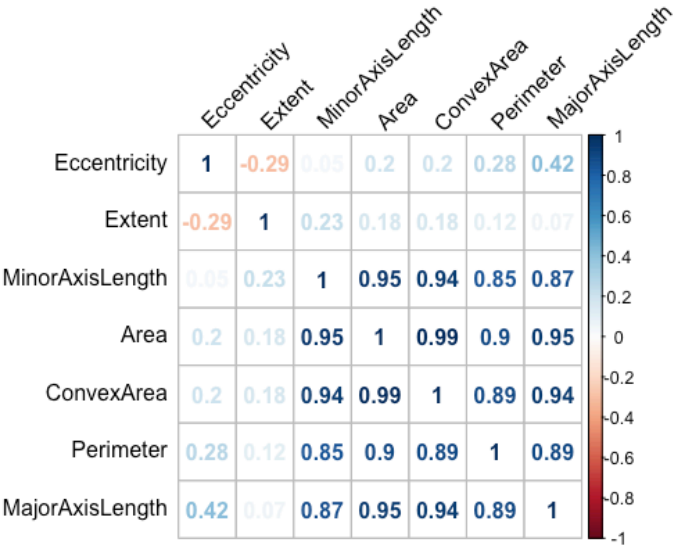


Figure 6

Feature Covariance Matrix



	Eccentricity	Extent	MinorAxisLength	Area	ConvexArea	Perimeter	MajorAxisLength
Eccentricity	1	-0.29	0.05	0.2	0.2	0.28	0.42
Extent	-0.29	1	0.23	0.18	0.18	0.12	0.07
MinorAxisLength	0.05	0.23	1	0.95	0.94	0.85	0.87
Area	0.2	0.18	0.95	1	0.99	0.9	0.95
ConvexArea	0.2	0.18	0.94	0.99	1	0.89	0.94
Perimeter	0.28	0.12	0.85	0.9	0.89	1	0.89
MajorAxisLength	0.42	0.07	0.87	0.95	0.94	0.89	1

Figure 7

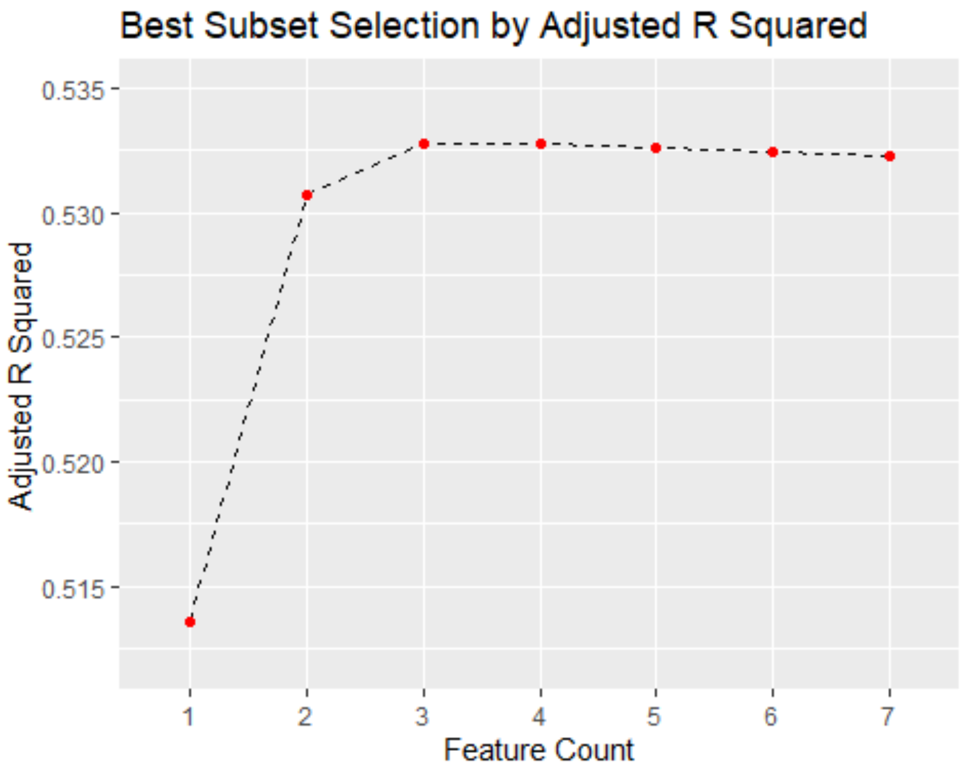
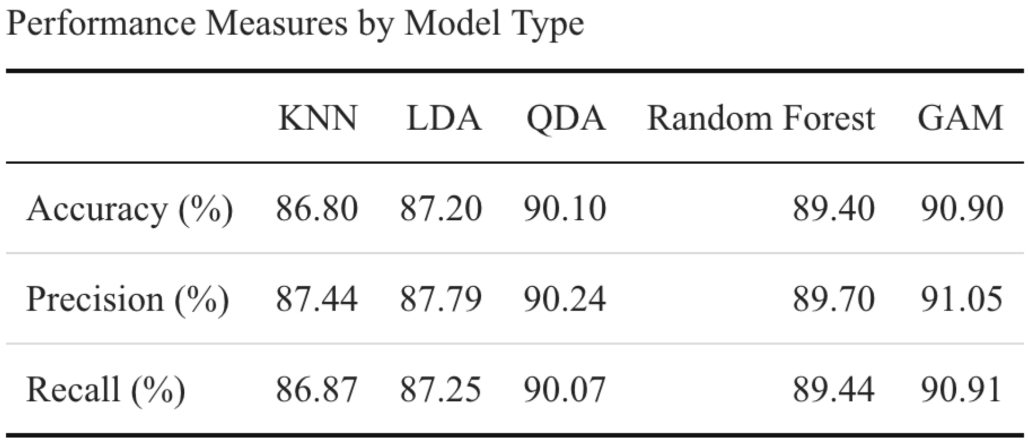


Figure 8



Performance Measures by Model Type

	KNN	LDA	QDA	Random Forest	GAM
Accuracy (%)	86.80	87.20	90.10	89.40	90.90
Precision (%)	87.44	87.79	90.24	89.70	91.05
Recall (%)	86.87	87.25	90.07	89.44	90.91

Figure 9

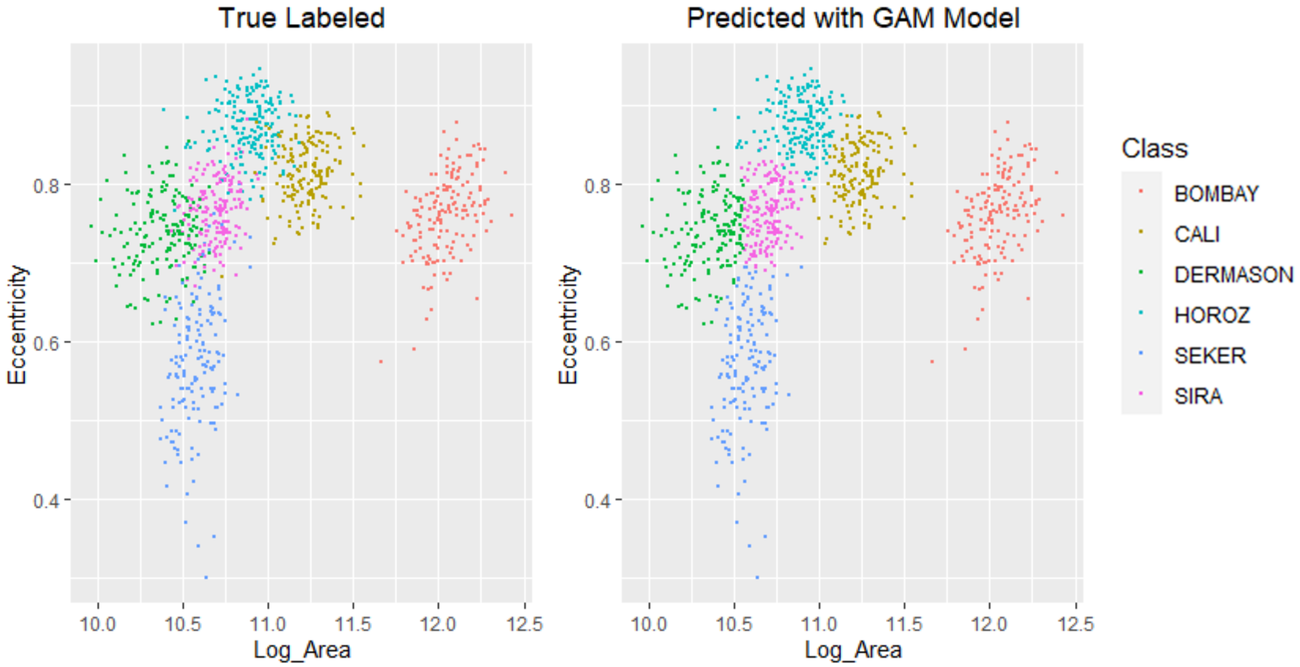


Figure 10

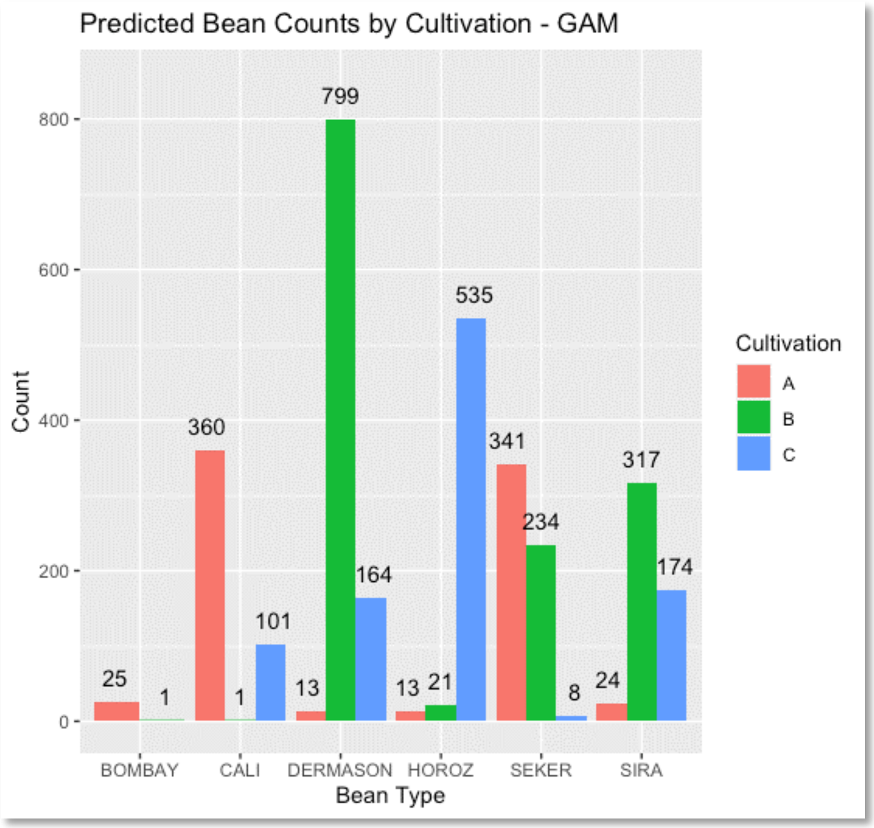


Figure 11

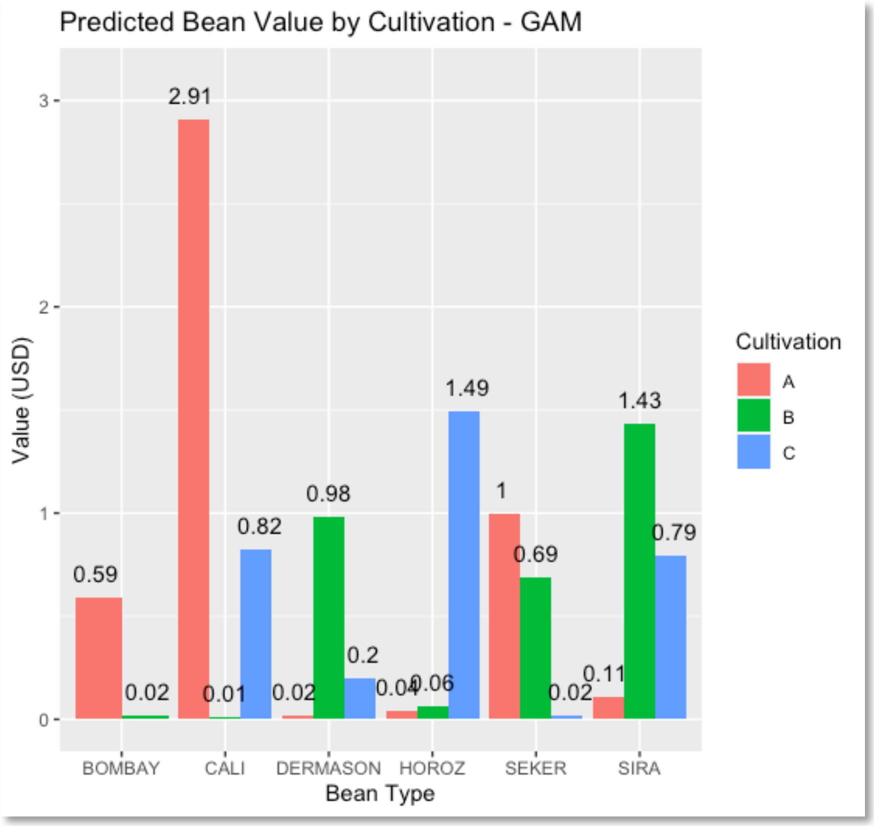


Figure 12

Value (USD) of One Pound Bean Cultivation (Calculated from Bootstrap of 1,000 One Pound Bean Samples)

	Mean	Standard Error	Lower Confidence Interval	Upper Confidence Interval
A	4.6680	0.1122	4.4511	4.8792
B	3.1894	0.0566	3.0788	3.3010
C	3.3189	0.0581	3.2022	3.4342

References

Koklu, M., Ozkan, I. A. (2020). Multiclass classification of dry beans using computer vision and machine learning techniques. Computers and Electronics in Agriculture, 174.

James, G., Witten, D., Hastie, T., Tibshirani, R. (n.d.). 6. Linear Model Selection and Regularization. In An Introduction to Statistical Learning with Applications in R.

Hossin, Mohammad & M.N, Sulaiman. (2015). A Review on Evaluation Metrics for Data Classification Evaluations.

International Journal of Data Mining & Knowledge Management Process. 5. 01-11. 10.5121/ijdkp.2015.5201.