MULTICLASS CLASSIFICATION AND PRICE PREDICTION OF DRY BEANS USING DIFFERENT ALGORITHMS

STAT 602 Mid Term Project
Spring 2022

Prepared By
Md Mominul Islam (101009250)

PRJECT OUTLINE

- Background Study
- Project Objective
- Data Preview
- Developed Algorithm
- Exploratory Data Analysis
- Statistical Model Development
- Statistical Model Accuracy Summary
- Project Conclusion
- References

BACKGROUND STUDY

- 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. [2]
- Because of the wide range of genetic diversity of dry bean, seed classification is essential for marketing and production.
- To establish a sustainable agricultural system, Koklu et al. [1] provided method for obtaining uniform seed varieties from crop production.
- To address the problem of market value, Koklu et al. [1] created distinct classification algorithms for seven different registered varieties of dry beans with comparable characteristics in order to achieve consistent seed categorization.

[1] Computers and Electronics in Agriculture 174 (2020): 105507.
[2] Journal of Agriculture and Food Sciences 26.1 (2012): 15-26.

PROJECT OBJECTIVE

- For the model classification, koklu et al. have taken 13,611 grain images of seven different registered dry beans and applied different classification algorithms [1].
- Multilayer perceptron (MLP), Support Vector Machine (SVM), k-Nearest Neighbors (kNN), Decision Tree (DT) classification models were created with 10-fold cross validation and performance metrics were compared [1].
- Overall correct classification rates have been determined as 91.73%, 93.13%, 87.92% and 92.52% for MLP, SVM, kNN and DT, respectively [1]
- For our project, we will do a brief exploratory data analysis with the beans data and apply different classification algorithms to justify the published result.
- Next, with a certain price for every class of beans, we will predict the price for our data set.
- Finally, based on our algorithms and predicted price, we will recommend on which type of algorithm should be further developed for the task of sorting white beans

[1] Computers and Electronics in Agriculture 174 (2020): 105507.

DATA PREVIEW

- We were given a data set of beans comprised of 3000 records and 9 variables describing the structural morphology of dry beans.
- There are 8 independent and one dependent variables (with 6 different classes)

Independent Variables	Target/Dependent Variable
Area, Perimeter, Major Axis Length, Minor Axis Length, Eccentricity, Convex Area and Extent	Class (Dermason, Cali, Seker, Sira, Horoz and Bombay)

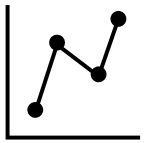
DATA REFORMATTING

- For our convenience, we have ignored variable ‘X’ indicating index in our dataset.
- We looked for missing values and typos in our dataset but found none.
- In our categorical target variables, there are 500 records of each class

Bombay	Cali	Dermason	Horoz	Seker	Sira
500	500	500	500	500	500

DEVELOPED ALGORITHM

- Initially, we performed different EDA analysis. Summary Statistics by Class, Density plot to see the distribution od data, histogram, Correlation matrix to find correlation among different features and pairs plot.
- With our EDA, we have successfully analyzed different features of our data set based on Beans Class.
- We have done ANOVA to find any significant features in our data set.



- Next, we developed a 5 different classification algorithms using ‘CLASS’ as target variable.
- We have split our data into 60 % training data and 40% test data and train our model.
- Furthermore, we predicted and checked our results using test data.
- With all of our built model, we predicted price for whole set of data of a definite farm and compared our results.
- Lastly, we compared the results obtained from models and came to a conclusion.

STATISTICAL DISTRIBUTION OF FEATURES OF DRY BEAN VARIETIES (IN PIXELS).

The minimum, maximum, mean and standard deviation data of the features obtained for all dry bean samples are given in Table 1

Parameter	Mean	Standard Dev.	Median	Minimum	Maximum	Range
Area	69875	49578.52	48714.5	20645	251320	230675
Perimeter	1012.24	347.75	941.9	384.17	2164.1	1779.93
Major Axis Length	362.05	124.52	332.9	161.52	740.97	579.45
Minor Axis Length	225.19	73.35	202.73	106	473.39	367.39
Eccentricity	0.76	0.1	0.77	0.3	0.94	0.64
ConvexArea	70944.1	50382.27	50807.5	8912	259965	251053
Extent	0.75	0.05	0.77	0.57	0.85	0.28

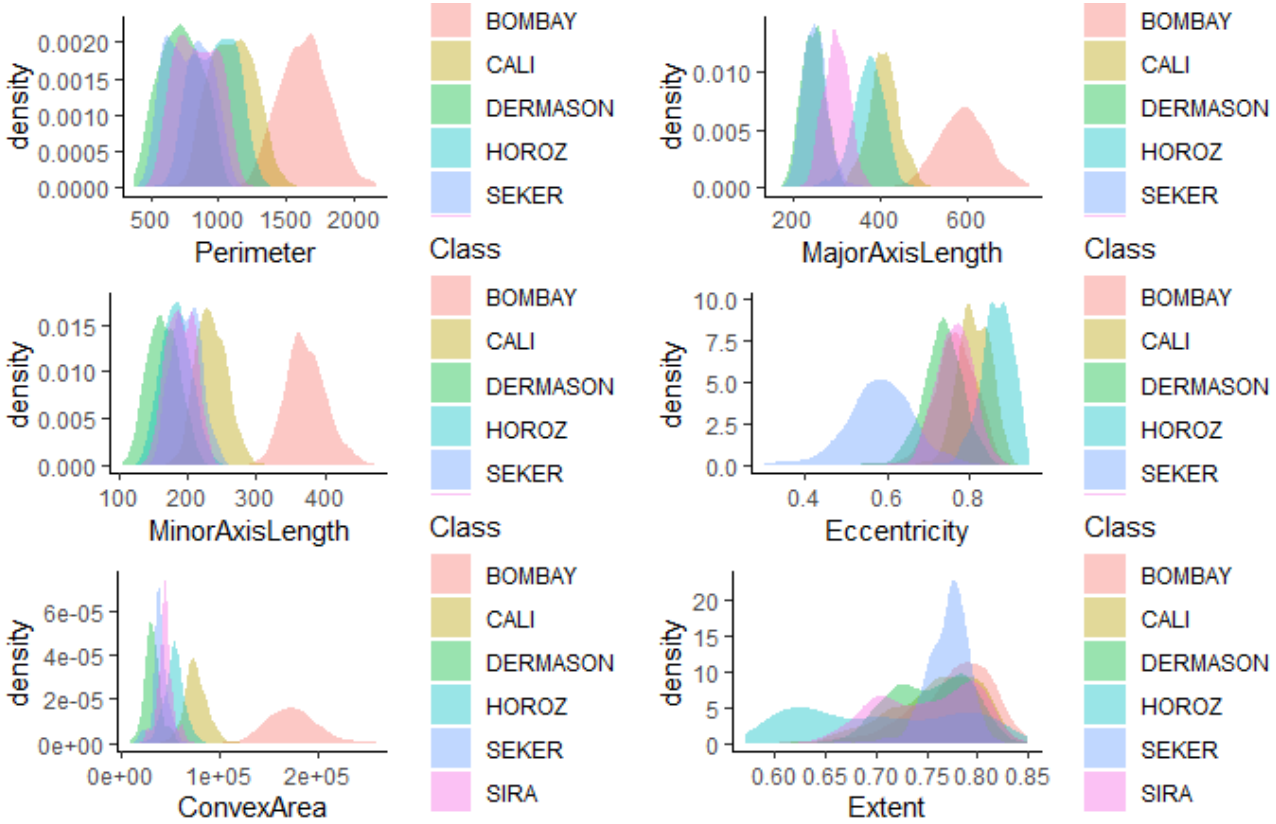
- The variables**, Area and Convex Area**, had the largest range for all four datasets.
- There are large differences in the range of variables, the variables with larger ranges can dominate over those with small ranges which may lead to **biased results.**
- In these results, the mean area for the dry beans is 69,875, and the median area is 48,714.5. The data appear to be skewed to the right, which explains why the mean is greater than the median.
- While **Extent and Eccentricity** appears to be skewed to the left as mean is less than the median.

SUMMARY STATISTICS (CLASS COVARIANCE)

Class	Area	Perimeter	Maj.Axis.	Min.Axis.	Eccentricity	Convex Area	Extent
BOMBAY	5.53E+08	32015.8	3177.899	826.684	0.547263	259965	0.850243
CALI	91528410	26272.79	1188.178	491.4581	0.618366	117510	0.842753
DERMASON	24651963	22913.81	696.7121	498.7684	0.549495	56174	0.847196
HOROZ	56765885	24960.58	1252.489	456.5644	0.722737	82462	0.842089
SEKER	22567179	24170.41	736.8507	419.4165	0.300635	65674	0.81831
SIRA	22641401	23977.06	782.4715	401.8085	0.609884	73945	0.841802

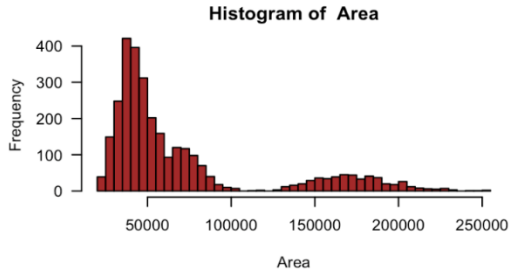
- From above table, we can have an idea about the class variances which measures variability from the average or mean.
- The variance of each variable by class shows evidence of non-constant variance.

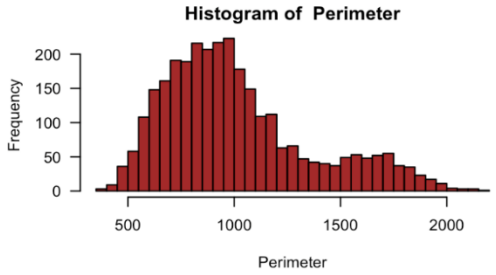
EXPLORATORY DATA ANALYSIS (DENSITY PLOT)

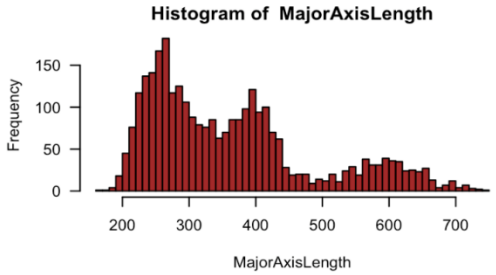


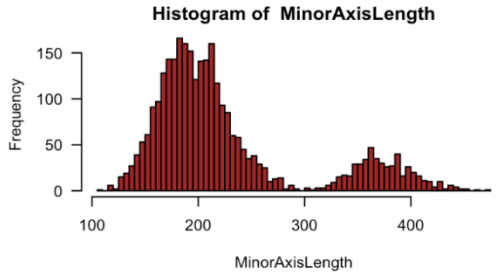
- Density plot represents distribution of a numeric variable using kernel density estimate to show the probability density function of the variable
- For variable ‘Perimeter’ we can see bimodal distribution in data and class ‘BOMBAY’ is off from others.
- For Eccentricity, class ‘SEKER’ is left skewed.
- In the ‘Extent’ parameter, we can see bimodal distribution and class ‘HOROZ’ is right skewed.
- For variables related to area, length and perimeter, class ‘BOMBAY’ has bell shaped distribution but not similar to other classes.

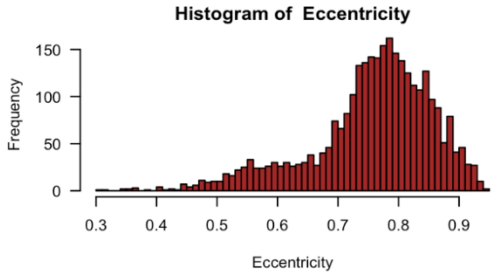
EXPLORATORY DATA ANALYSIS (HISTOGRAM)

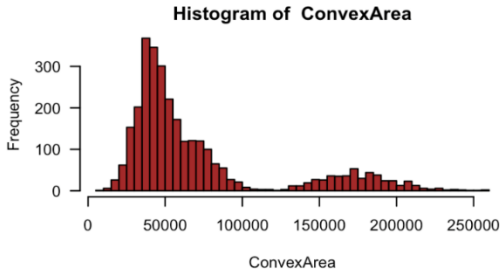


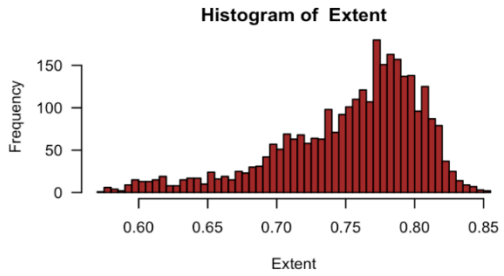






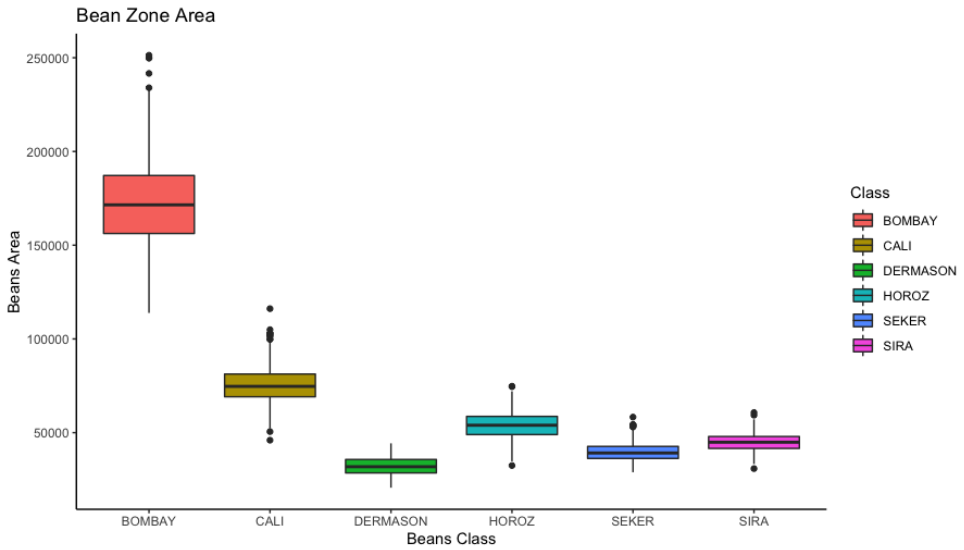


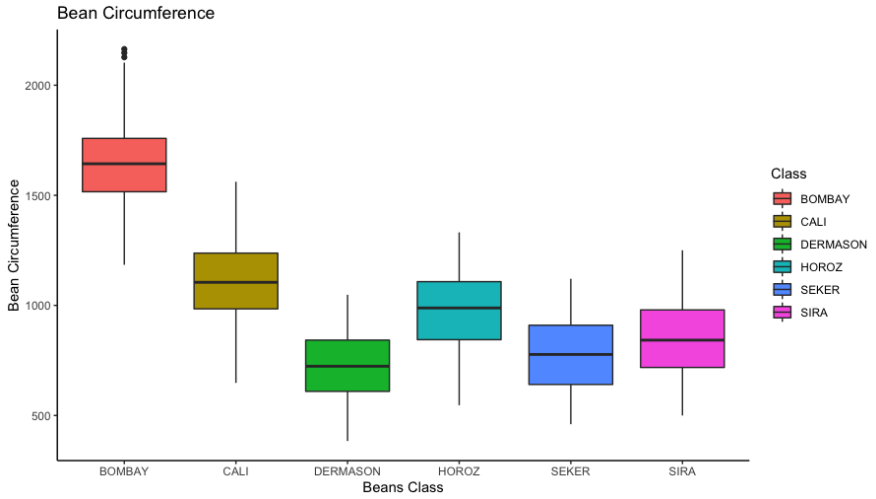




- The histograms from the labeled data reveal that the variables exhibit multimodal behavior.
- This indicates that at least one of the bean classes differs greatly from the others.
- Further investigation revealed that the kind of BOMBAY beans is to blame for the multimodality.

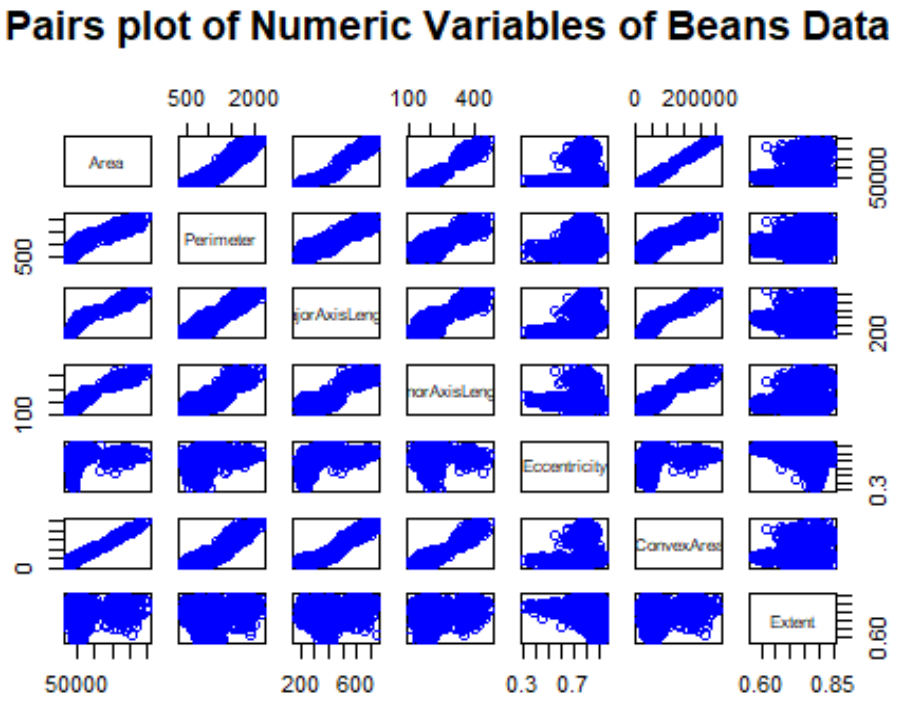
EXPLORATORY DATA ANALYSIS (BOXPLOT)

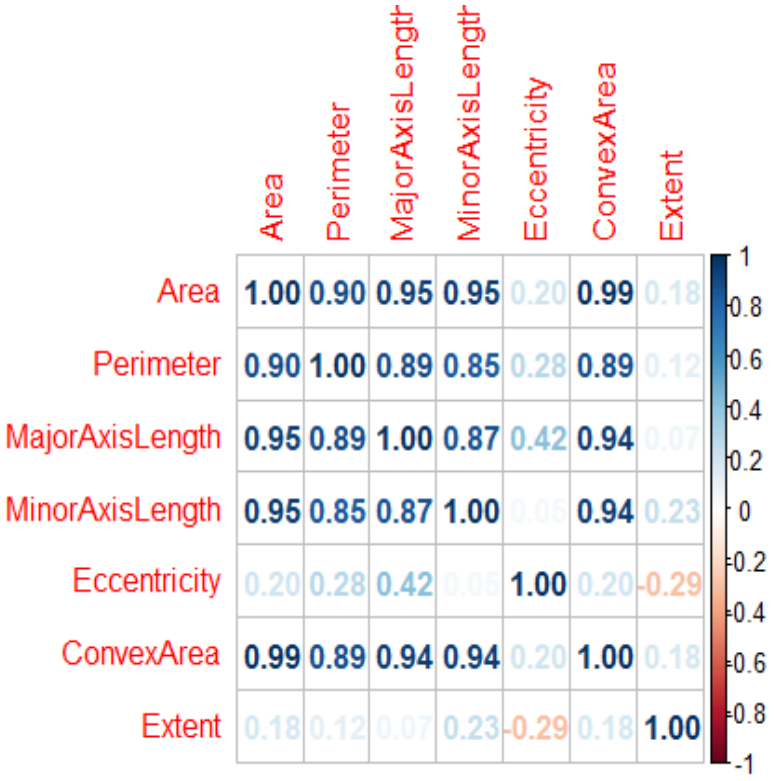




- A boxplot is a standardized way of displaying the distribution of data based on a five number summary ("minimum", first quartile (Q1), median, third quartile (Q3), and "maximum")
- Plotted variables 'Beans Area' and 'Perimeter' with respect to each class.
- For class 'BOMBAY', we can say that their range is highest.
- All of the classes have similar variability except the 'BOMBAY'.
- These inference is justified when we built our model.
- Median value of the class 'DERMASON' is lowest for both of the two parameters.
- No skewness is observed except some outliers in the class 'BOMBAY'

FEATURE SELECTION





- Most of the variables except for Eccentricity and Extent are highly correlated with each other.
- With using pairs plot, we can see Eccentricity and Extent are not having any correlation which is further validated with the correlation matrix on the right.

ANALYSIS OF VARIANCE

Parameter	Area	Perimeter	Major Axis Length	Minor Axis Length	Eccentrcity	ConvexArea	Extent
Sum Square	6.87E+12	285666122	42590497	14591140	22.933	7.10E+12	1.707
Mean Square	1.40E+12	57133224	8518099	2918228	4.587	1.42E+12	0.3414
F Statistic	10877	2221	6523	5658	1715	8442	157.1
p-Value	2.00E-16	2.00E-16	2.00E-16	2.00E-16	2.00E-16	2.00E-16	2.00E-16

- From the pairs plot and correlation matrix, we have seen the variables eccentricity and extent has little correlation compared to other variables.
- To justify the assumption, we have 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% training and 40% test data.
- Using Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Mclust , and Mclust 'EDDA', 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***: 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. [4]
- ***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 in order to perform prediction. [5]
- ***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 [3]

[3] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5096736/pdf/nihms793803.pdf
[4] https://www.geeksforgeeks.org/linear-discriminant-analysis-in-r-programming
[5] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. (2013). An introduction to statistical learning : with applications in R. New York :Springer,

LINEAR DISCRIMINANT ANALYSIS (LDA) MODEL

Confusion Matrix and Accuracy – Training Data

Actual	Predict					
	BOMBAY	CALI	DERMASON	HOROZ	SEKER	SIRA
BOMBAY	300	0	0	0	0	0
CALI	0	277	0	24	1	1
DERMASON	0	0	233	3	22	47
HOROZ	0	12	0	244	0	22
SEKER	0	1	15	0	248	4
SIRA	0	10	52	29	29	226

LDA Model Accuracy on Training Data: **84.88%**

- We have 1800 records in our training data set with 300 in each classes of beans. Our LDA model has an accuracy of 84.88% in terms of predicting beans classes.

Confusion Matrix and Accuracy – Test Data

Actual	Predict					
	BOMBAY	CALI	DERMASON	HOROZ	SEKER	SIRA
BOMBAY	199	0	0	0	0	0
CALI	1	184	0	9	1	3
DERMASON	0	0	159	3	6	27
HOROZ	0	8	1	171	0	14
SEKER	0	0	4	0	181	1
SIRA	0	8	36	17	12	155

LDA Model Accuracy on Test Data: **87.41%**

PRICE PREDICTION USING LDA MODEL

	Train Data		Test Data	
Class	Actual no of Beans	Predicted Beans	Actual Beans	Predicted Beans
Bombay	300	300	200	199
Cali	303	303	200	198
Dermason	305	305	200	195
Horoz	278	278	200	194
Seker	268	268	200	186
Sira	346	346	200	228

- From our training and test data, we have calculated actual and predicted number of beans per class using LDA Model.
- These number are used further to calculate total actual and predicted price of beans.
- 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 Weight (gm)	Actual Weight (lbs)	Actual Price ($)	Predicted Weight(gm)	Predicted Weight(lbs)	Predicted Price ($)
Train Data	1260	2.77	12.93	1253.59	2.76	13.013
Test Data	840	1.85	8.62	836.12	1.84	8.6431

- From Our train Data, the actual price was $12.93, and the predicted price is $13.013
- From our test data, actual price is $1.85, and predicted price is $8.643

QUADRATIC DISCRIMINANT ANALYSIS (QDA) MODEL

Confusion Matrix and Accuracy – Training Data

Actual	Predict					
	BOMBAY	CALI	DERMASON	HOROZ	SEKER	SIRA
BOMBAY	300	0	0	0	0	0
CALI	0	290	0	12	0	2
DERMASON	0	0	257	3	14	21
HOROZ	0	9	0	268	0	20
SEKER	0	1	14	0	266	14
SIRA	0	0	29	17	20	243

QDA Model Accuracy on Training Data: **90.22%**

- We have 1800 records in our training data set with 300 in each classes of beans. Our QDA model has an accuracy of 90.22% in terms of predicting beans classes.

Confusion Matrix and Accuracy – Test Data

Actual	Predict					
	BOMBAY	CALI	DERMASON	HOROZ	SEKER	SIRA
BOMBAY	200	1	0	0	0	0
CALI	0	189	0	5	1	2
DERMASON	0	0	173	3	0	20
HOROZ	0	7	1	185	0	12
SEKER	0	1	2	1	188	4
SIRA	0	2	24	6	11	162

QDA Model Accuracy on Test Data: **91.41%**

PRICE PREDICTION USING QDA MODEL

	Train Data		Test Data	
Class	Actual no of Beans	Predicted Beans	Actual Beans	Predicted Beans
Bombay	300	300	200	201
Cali	303	304	200	197
Dermason	305	295	200	196
Horoz	278	297	200	205
Seker	268	295	200	196
Sira	346	309	200	205

- From our training and test data, we have calculated actual and predicted number of beans per class using QDA Model.
- These number are used further to calculate total actual and predicted price of beans.
- 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 Weight (gm)	Actual Weight (lbs)	Actual Price ($)	Predicted Weight(gm)	Predicted Weight(lbs)	Predicted Price ($)
Train Data	1260	2.777	12.9304	1260.45	2.778	12.974
Test Data	840	1.85	8.62	841.51	1.855	8.639

- From Our train Data, the actual price was $12.93, and the predicted price is $12.97
- From our test data, actual price is $1.85, and predicted price is $8.639

MCLUSTDA MODEL

Confusion Matrix and Accuracy – Training Data

Actual	Predict					
	BOMBAY	CALI	DERMASON	HOROZ	SEKER	SIRA
BOMBAY	300	0	0	0	0	0
CALI	0	290	0	9	1	0
DERMASON	0	0	257	0	14	29
HOROZ	0	12	3	268	0	17
SEKER	0	0	14	0	266	20
SIRA	0	2	21	20	14	243

MclustDA Model Accuracy on Training Data: **91.41%**

- We have 1800 records in our training data set with 300 in each classes of beans. Our 'MclustDA' model has an accuracy of 91.41% in terms of predicting beans classes.

Confusion Matrix and Accuracy – Test Data

Actual	Predict					
	BOMBAY	CALI	DERMASON	HOROZ	SEKER	SIRA
BOMBAY	200	0	0	0	0	0
CALI	0	193	0	4	1	2
DERMASON	0	0	172	1	3	24
HOROZ	0	7	3	184	0	6
SEKER	0	1	1	0	191	7
SIRA	0	3	16	7	4	170

MclustDA Model Accuracy on Test Data: **92.5%**

PRICE PREDICTION USING MCLUSTDA MODEL

	Train Data		Test Data	
Class	Actual no of Beans	Predicted Beans	Actual Beans	Predicted Beans
Bombay	300	300	200	200
Cali	303	304	200	204
Dermason	305	295	200	192
Horoz	278	297	200	196
Seker	268	295	200	199
Sira	346	309	200	209

- From our training and test data, we have calculated actual and predicted number of beans per class using MclustDA Model.
- These number are used further to calculate total actual and predicted price of beans.
- 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 Weight (gm)	Actual Weight (lbs)	Actual Price ($)	Predicted Weight(gm)	Predicted Weight(lbs)	Predicted Price ($)
Train Data	1260	2.777	12.9304	1260.45	2.778	12.974
Test Data	840	1.85	8.62	841.05	1.854	8.66

- From Our train Data, the actual price was $12.93, and the predicted price is $12.97
- From our test data, actual price is $1.85, and predicted price is $8.66

MCLUST 'EDDA' MODEL

Confusion Matrix and Accuracy – Training Data

Actual	Predict					
	BOMBAY	CALI	DERMASON	HOROZ	SEKER	SIRA
BOMBAY	300	0	0	0	0	0
CALI	0	286	0	11	1	2
DERMASON	0	0	249	0	15	36
HOROZ	0	18	3	257	0	22
SEKER	0	1	15	0	263	21
SIRA	0	3	34	21	13	229

MclustEDDA Model Accuracy on Training Data: **89.58%**

- We have 1800 records in our training data set with 300 in each classes of beans. Our 'MclustEDDA' model has an accuracy of 89.58% in terms of predicting beans classes.

Confusion Matrix and Accuracy – Test Data

Actual	Predict					
	BOMBAY	CALI	DERMASON	HOROZ	SEKER	SIRA
BOMBAY	200	0	0	0	0	0
CALI	0	188	0	7	1	4
DERMASON	0	0	163	1	5	31
HOROZ	0	8	3	177	0	12
SEKER	0	1	1	0	188	10
SIRA	0	4	25	9	5	157

MclustEDDA Model Accuracy on Test Data: **89.41%**

PRICE PREDICTION USING MCLUSTEDDA MODEL

	Train Data		Test Data	
Class	Actual no of Beans	Predicted Beans	Actual Beans	Predicted Beans
Bombay	300	300	200	200
Cali	303	304	200	204
Dermason	305	295	200	192
Horoz	278	297	200	196
Seker	268	295	200	199
Sira	346	309	200	209

- From our training and test data, we have calculated actual and predicted number of beans per class using MclustDA Model.
- These number are used further to calculate total actual and predicted price of beans.
- 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 Weight (gm)	Actual Weight (lbs)	Actual Price ($)	Predicted Weight(gm)	Predicted Weight(lbs)	Predicted Price ($)
Train Data	1260	2.777	12.9304	1260.45	2.778	12.974
Test Data	840	1.85	8.62	841.05	1.854	8.66

- From Our train Data, the actual price was $12.93, and the predicted price is $12.97
- From our test data, actual price is $1.85, and predicted price is $8.66

MODEL COMPARISONS

			Training Data		Test Data	
Model	**Train Accuracy (%)**	**Test Accuracy (%)**	**Actual Price ($)**	**Predicted Price ($)**	**Actual Price ($)**	**Predicted price ($)**
LDA	**84.88**	**87.41**	**12.93**	**13.013**	**8.62**	**8.643**
QDA	**90.22**	**91.41**	**12.93**	**12.974**	**8.62**	**8.639**
Mclust DA	**91.41**	**92.5**	**12.93**	**12.97**	**8.62**	**8.66**
Mclust EDDA	**89.58**	**89.41**	**12.93**	**12.97**	**8.62**	**8.66**

- We have developed four kinds of classification algorithms and predicted the price of beans.
- Compared to all models, Mclust DA performed better with a test accuracy of 92.5% and the actual and predicted price on test data are very close.
- We have also obtained similar results using QDA approach.

CONCLUSION

Initially, we have done EDA to see the distribution of data using Histograms and boxplots, our descriptive statistics helped to understand the variability and ranges.

Using two-way-ANOVA, we have done the feature selection criterion.

Based on our data, we have built four different classification algorithms (60% Train and 40% Test Data)and compared the accuracy of each models.

Accuracy:
MclustDA >QDA> MclustEDDA>LDA

Finally, we have compared the predicted price for all the four models and found that QDA can most accurately predict price base on test data set.

REFERENCES

1. Mazhar, K. A. R. A., et al. "Seed size and shape analysis of registered common bean (Phaseolus vulgaris L.) cultivars in Turkey using digital photography." *Journal of Agricultural Sciences* 19.3 (2013): 219-234.
2. Koklu, Murat, and Ilker Ali Ozkan. "Multiclass classification of dry beans using computer vision and machine learning techniques." *Computers and Electronics in Agriculture* 174 (2020): 105507.
3. James, Gareth, et al. *An introduction to statistical learning*. Vol. 112. New York: springer, 2013.
4. Lecture Notes and Materials From Stat 601 and Stat 602

Internet Resources

1. http://www.sthda.com/english/wiki/one-way-anova-test-in-r
2. https://www.guru99.com/r-anova-tutorial.html
3. https://gexijin.github.io/learnR/importing-data-and-managing-files.html
4. https://agroninfotech.blogspot.com/2021/11/rapid-publication-ready-anova-table-in-r.html
5. https://gexijin.github.io/learnR/index.html
6. https://stats.oarc.ucla.edu/r/dae/multinomial-logistic-regression/
7. https://rpubs.com/lmorgan95/ISLR_CH4_Solutions
8. https://rpubs.com/zlzlzl2/754880
9. https://rpubs.com/Richie222/853114
10. https://pages.cms.hu-berlin.de/EOL/gcg_quantitative-methods/Lab11_LDA_Model-assessment.html#Linear_Discriminant_Analysis
11. MCLust : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5096736/pdf/nihms793803.pdf
12. LDA: http://rstudio-pubs-static.s3.amazonaws.com/35817_2552e05f1d4e4db8ba87b334101a43da.html
13. https://stats.oarc.ucla.edu/r/dae/multinomial-logistic-regression/
14. https://sites.stat.washington.edu/mclust/
15. https://bradleyboehmke.github.io/HOML/model-clustering.html
16. Textbook: An Introduction to Statistical Learning With Appliaction in R by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani
17. STAT 602 and 601 Lecture Slides and Videos
18. https://alessiopassalacqua.github.io/getstrongeR-Rstats/readings.html
19. https://beta.rstudioconnect.com/content/2025/dplyr.nb.html
20. https://alessiopassalacqua.github.io/getstrongeR-Rstats/readings.html
21. https://www.webpages.uidaho.edu/~stevel/519/literatures/MCLUST%20for%20R.pdf
22. https://bradleyboehmke.github.io/HOML/model-clustering.html