Predicting Bean Type Based on Metamorphic Measurements

Analysis by Gavin Gunawardena

Utilizing a dataset from Seed Size and Shape Analysis of Registered Common Bean (Phaseolus vulgaris L.) Cultivars in Turkey Using Digital Photography from the Journal of Agricultural Science

Intro and Objective

- Objective: Test various predictive analytics algorithms for multiclass classification to sort beans based on their morphometric measurements
- Dataset
  - 3000 observations
  - 500 observations per class
  - 6 classes
  - No additional assumptions
- Measures
  - Accuracy
  - Weight differential between predictions and actual values
  - Cost differential between predictions and actual values

Intro and Objective

- Dependent Variable Classes:

ClassID <dbl>	Class <chr>	dollars_per_lb <dbl>	grams_per_seed <dbl>	approx_lbs_per_seed <dbl>	approx_dollars_per_seed <dbl>	approx_dollars_per_gram <dbl>
1	BOMBAY	5.56	1.92	0.0042	0.0234	0.0123
2	CALI	6.02	0.61	0.0013	0.0078	0.0133
3	DERMASON	1.98	0.28	0.0006	0.0012	0.0044
4	HOROZ	2.43	0.52	0.0011	0.0027	0.0054
5	SEKER	2.72	0.49	0.0011	0.0030	0.0060
6	SIRA	5.40	0.38	0.0008	0.0043	0.0119

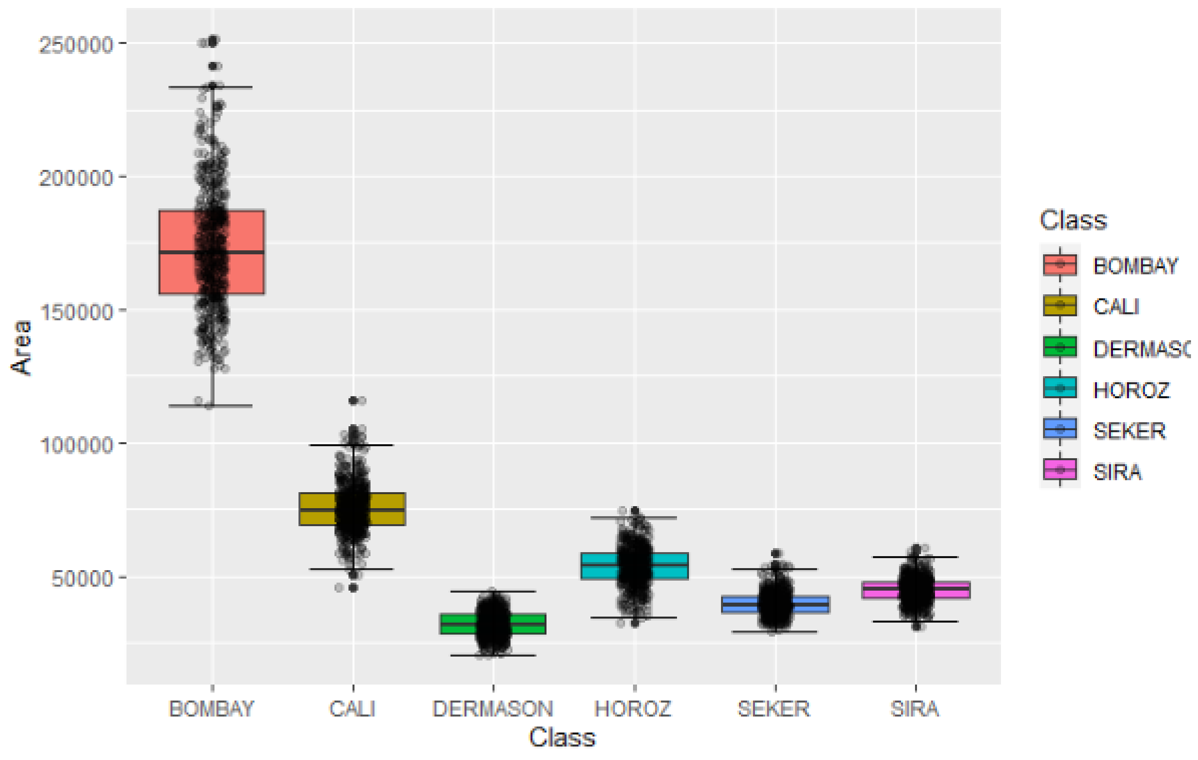
- Independent Variables:

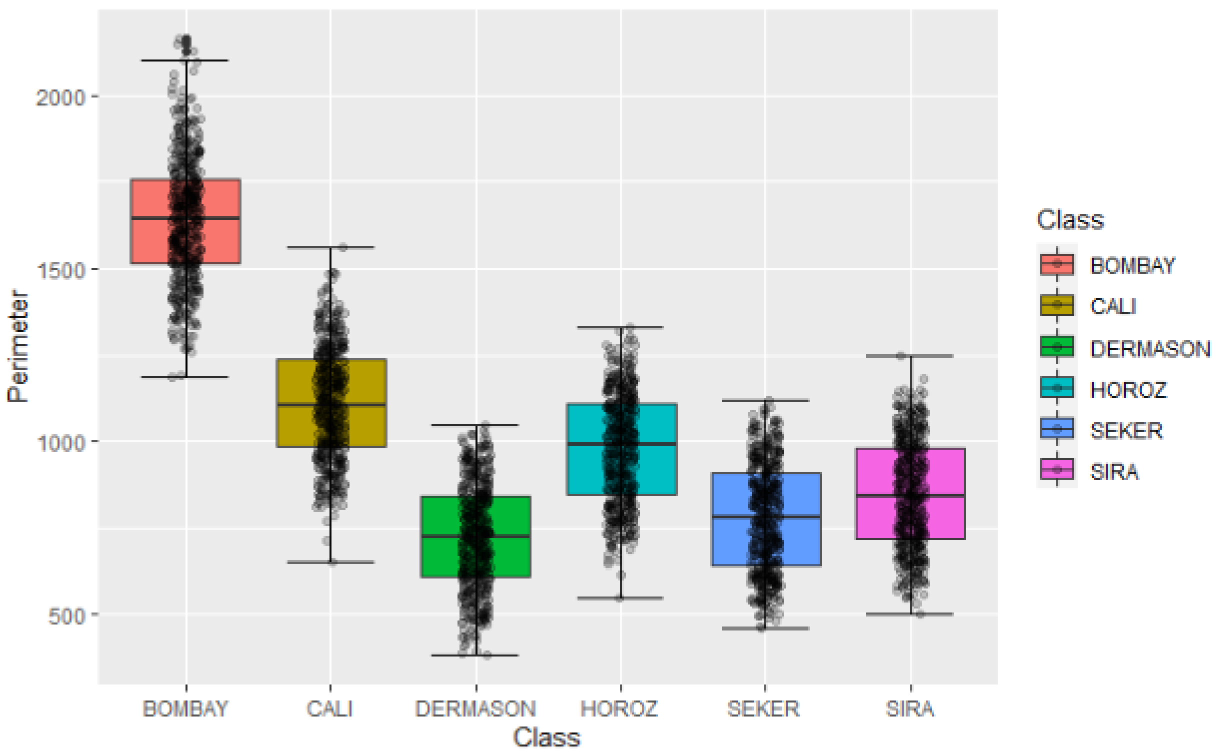
ID <dbl>	name <chr>	description <chr>
1	Area	The area of a bean zone and the number of pixels within its boundaries.
2	Perimeter	Bean circumference is defined as the length of its border.
3	MajorAxisLength	The distance between the ends of the longest line that can be drawn from a bean.
4	MinorAxisLength	The longest line that can be drawn from the bean while standing perpendicular to the main axis.
5	Eccentricity	Eccentricity of the ellipse having the same moments as the region.
6	ConvexArea	Number of pixels in the smallest convex polygon that can contain the area of a bean seed.
7	Extent	The ratio of the pixels in the bounding box to the bean area.

Methods

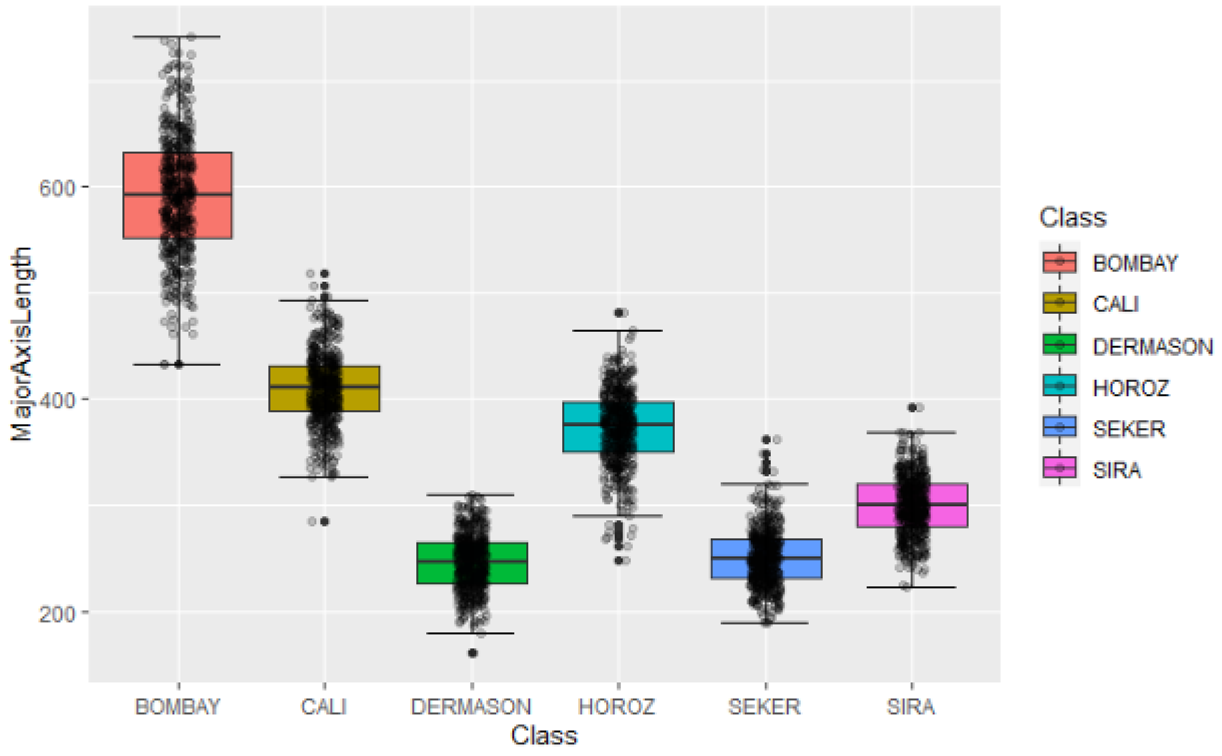
- Validation Technique
  - Leave One Out Cross-Validation
  - 80/20 data split with 80% of the data being used for training and validation while the last 20 is used to test the final model(s) and obtain a final accuracy rate
- Normalization/Standardization Technique
  - Z-Score
- Feature Selection Method
  - Best Subset Selection with residual sum of squares and then adjusted R^2 as the criteria for feature selection
- Statistical Models
  - Multinomial Logistic Regression
  - Linear Discriminant Analysis
  - K-Nearest Neighbors Classifier
  - Quadratic Discriminant Analysis
  - Naïve Bayes

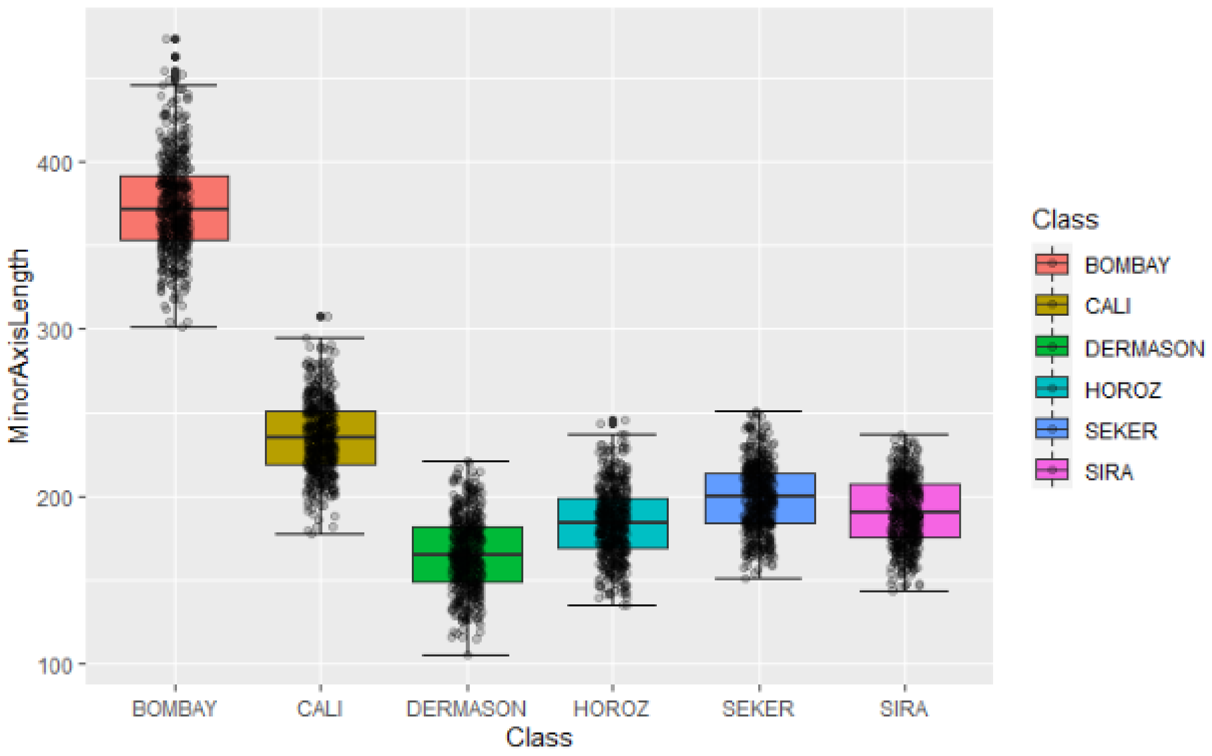
Exploratory Analysis



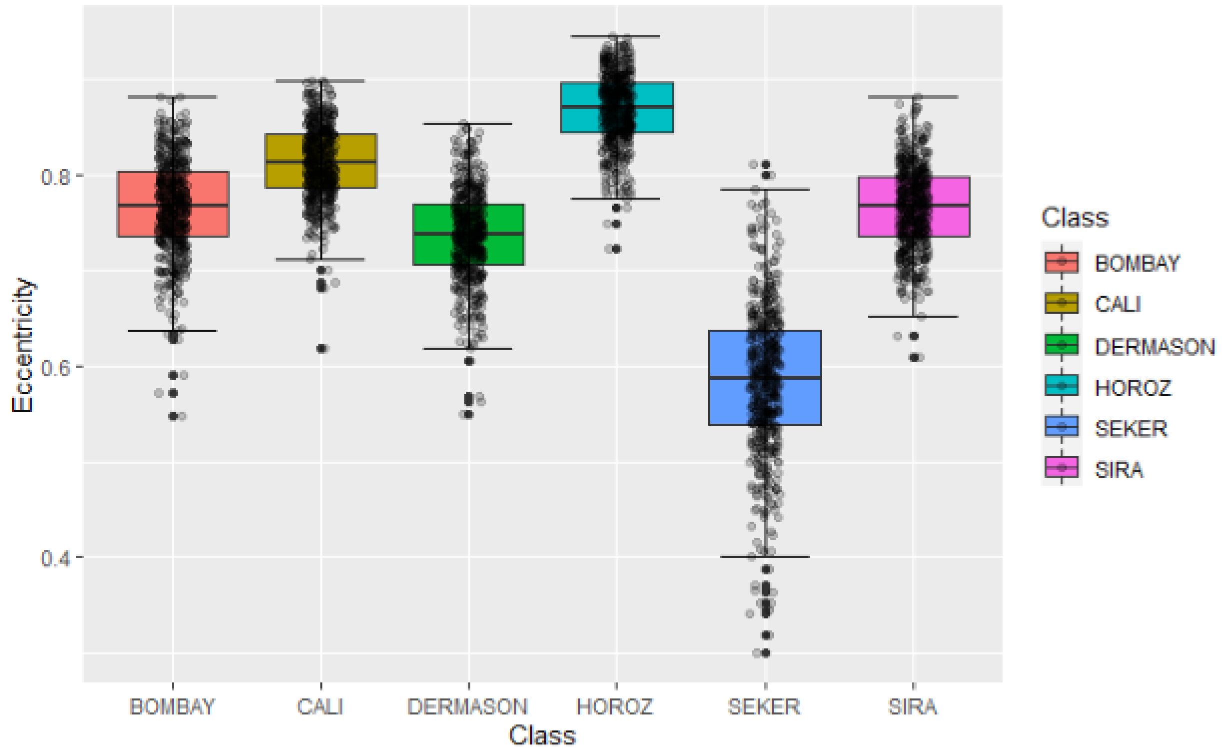


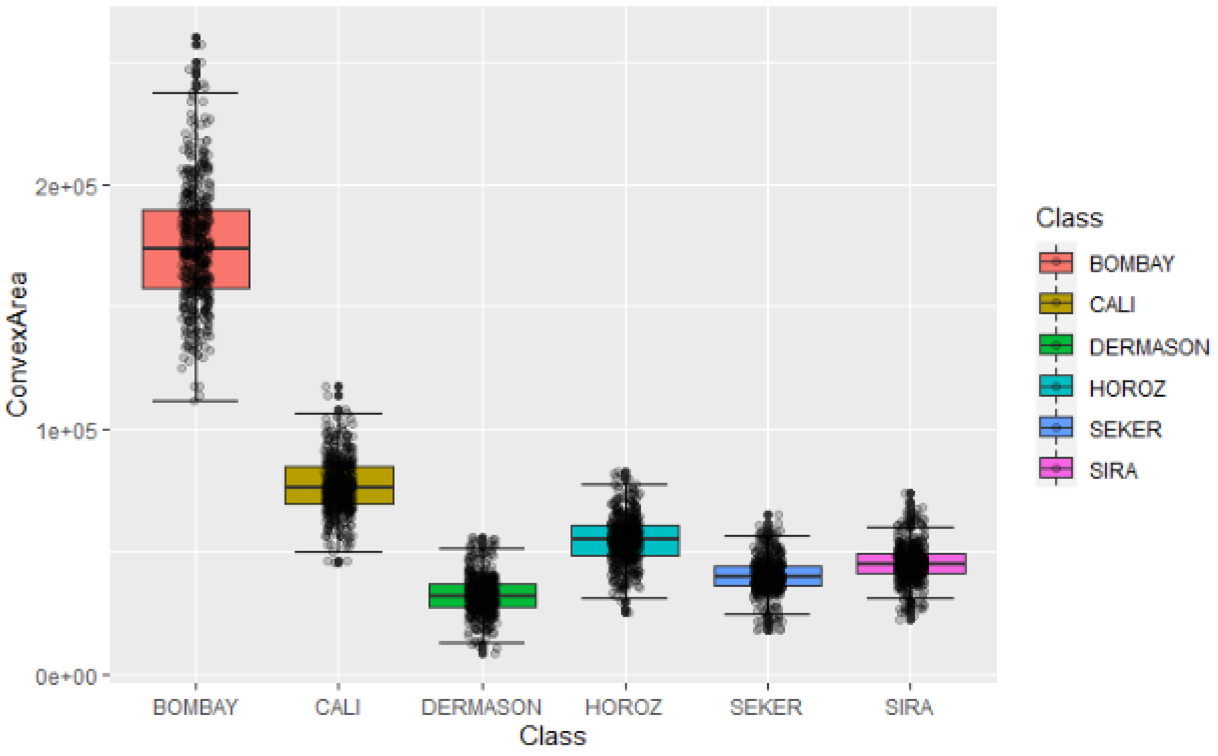
Exploratory Analysis



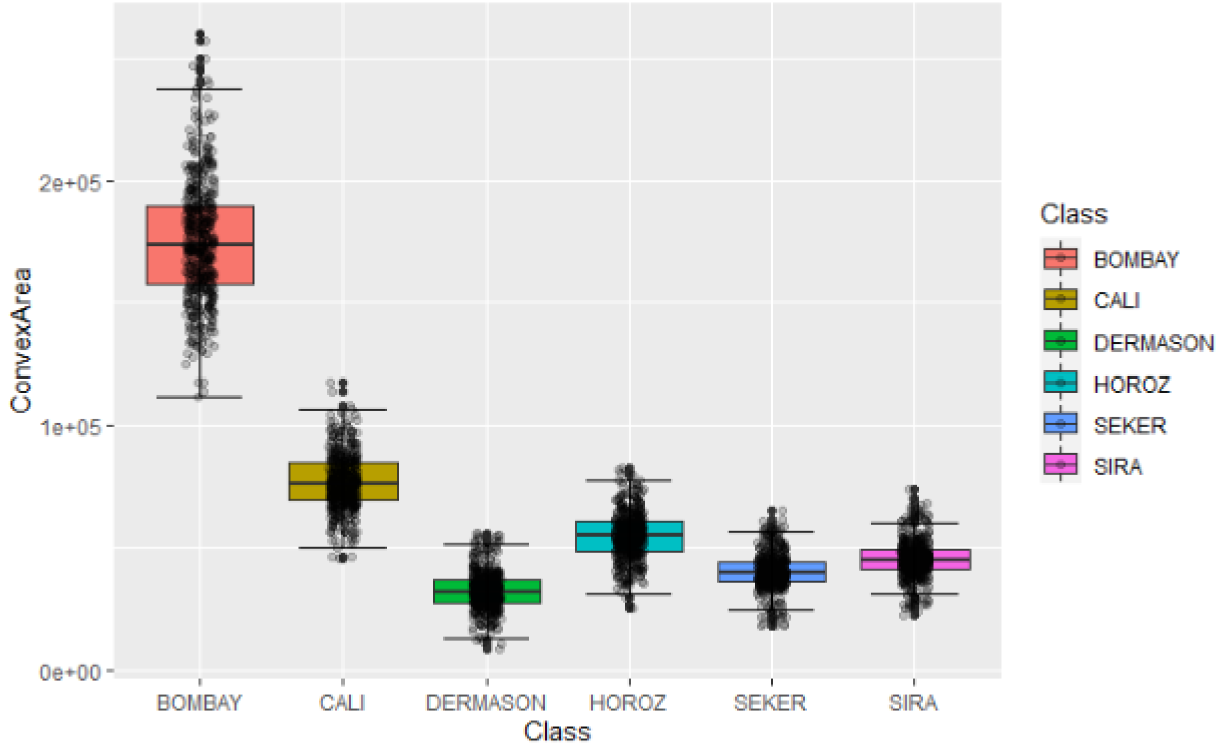


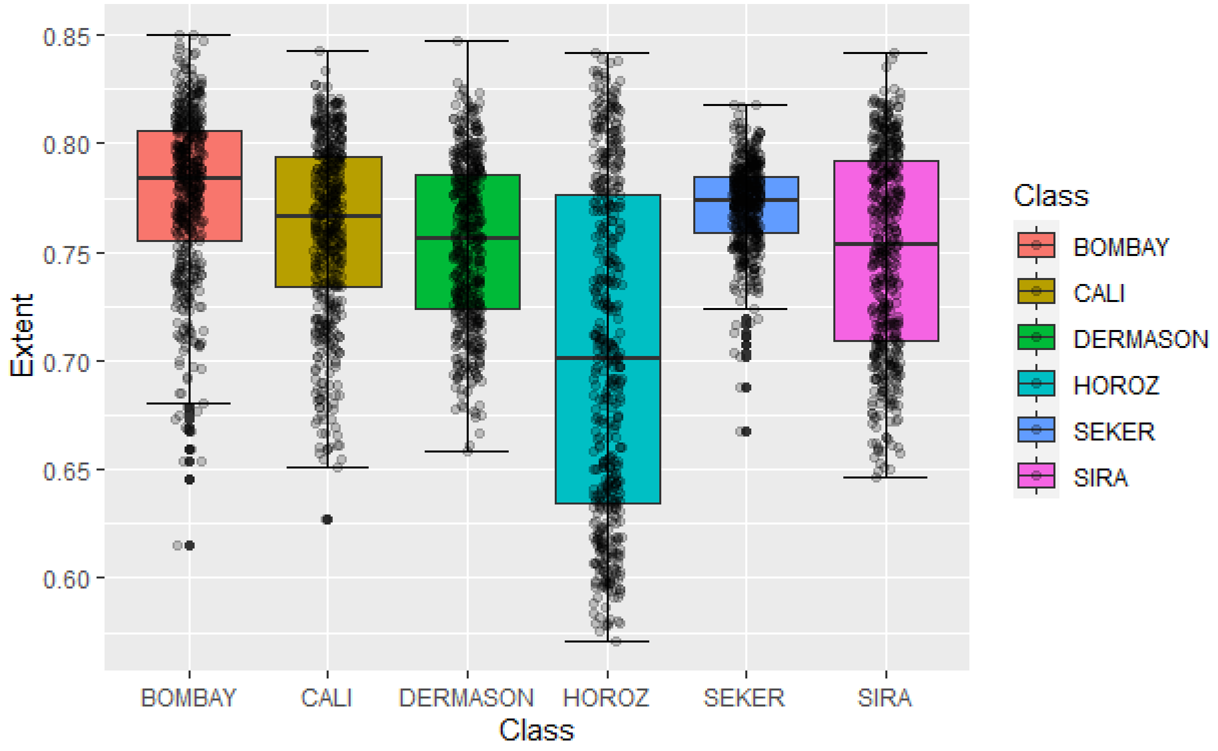
Exploratory Analysis



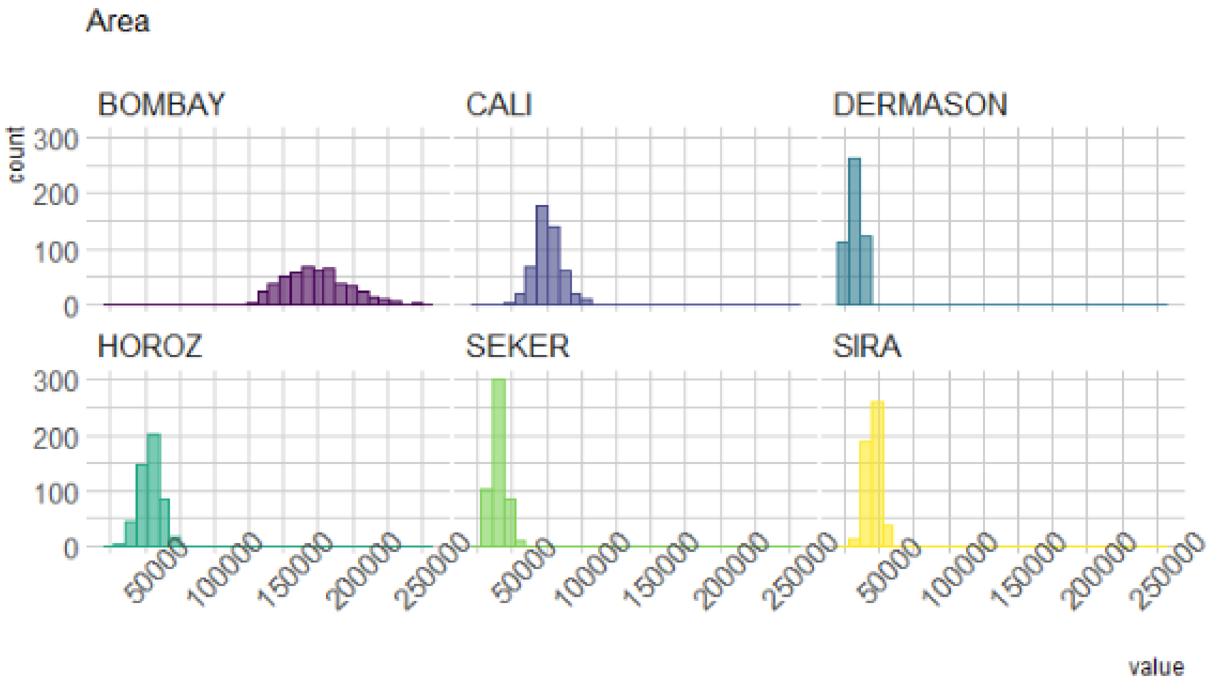


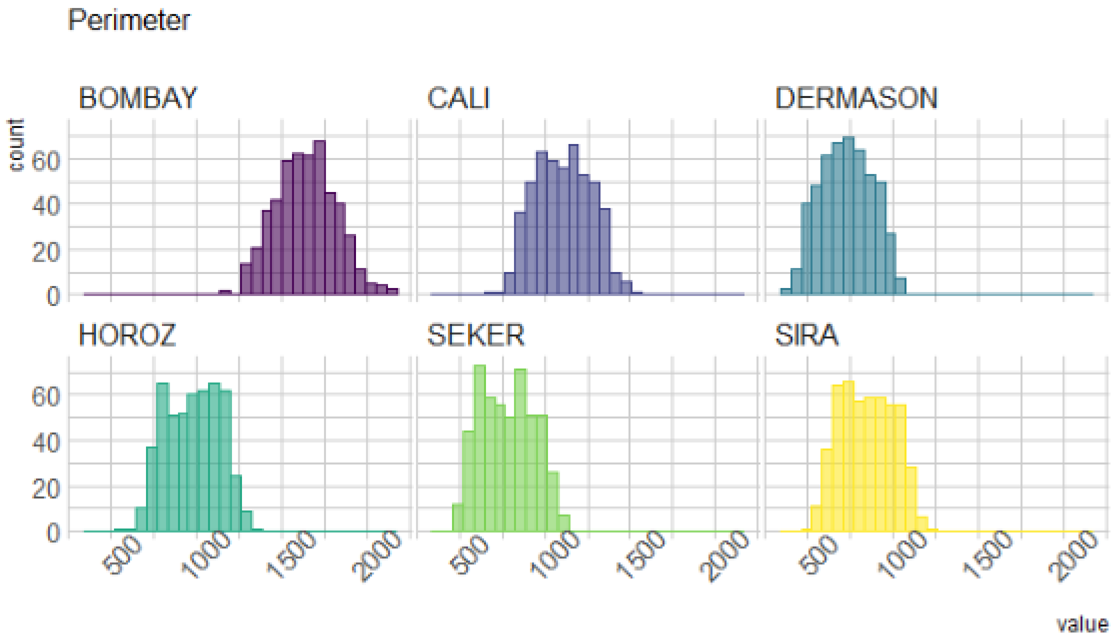
Exploratory Analysis



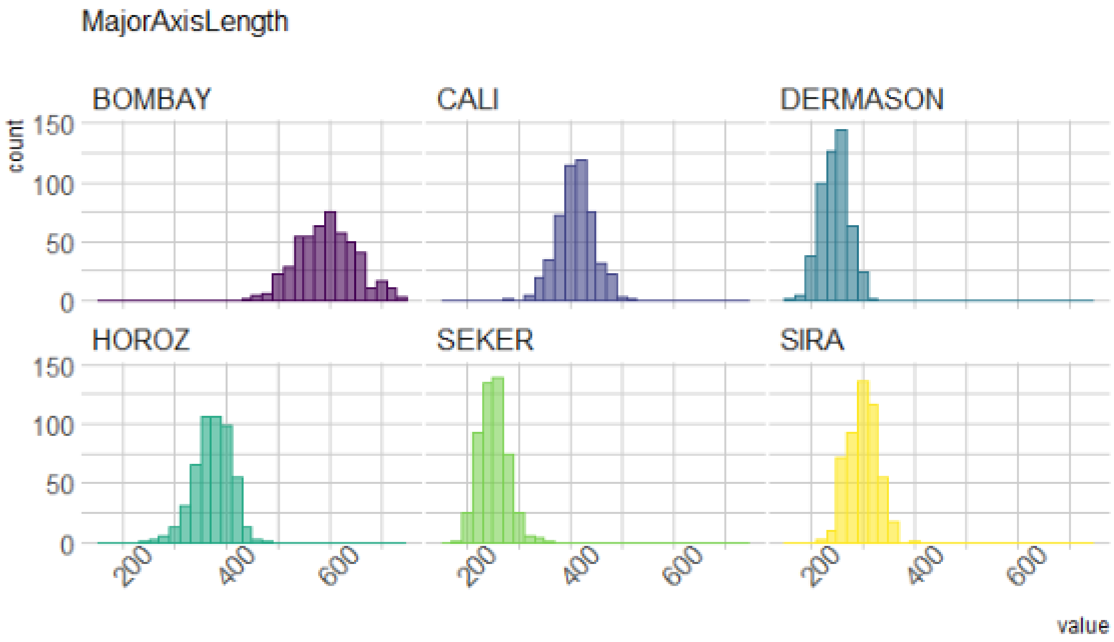


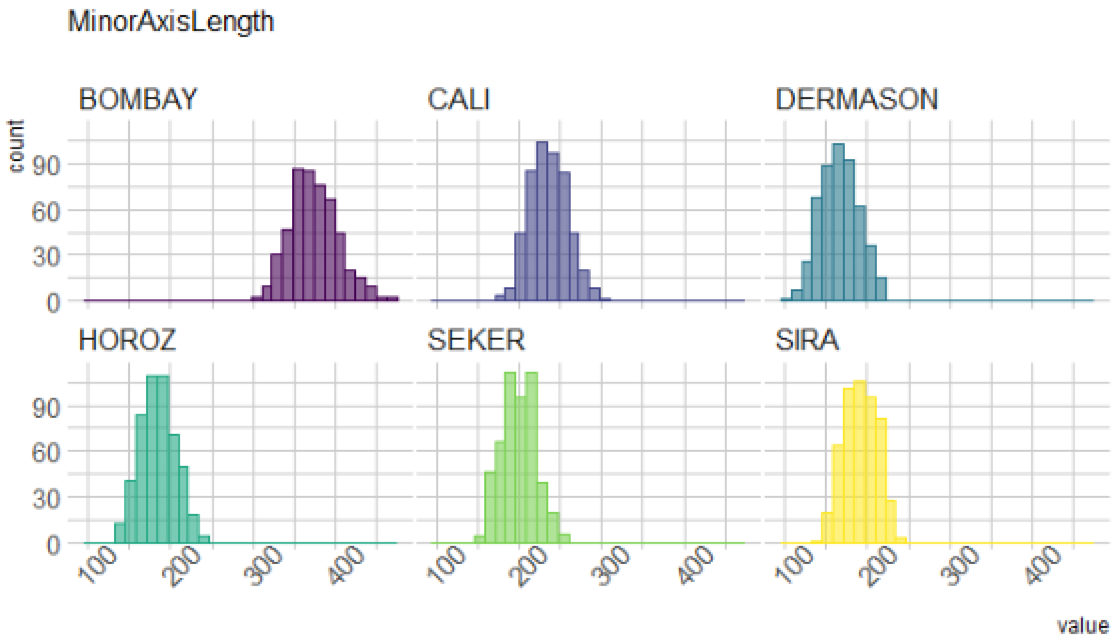
Exploratory Analysis



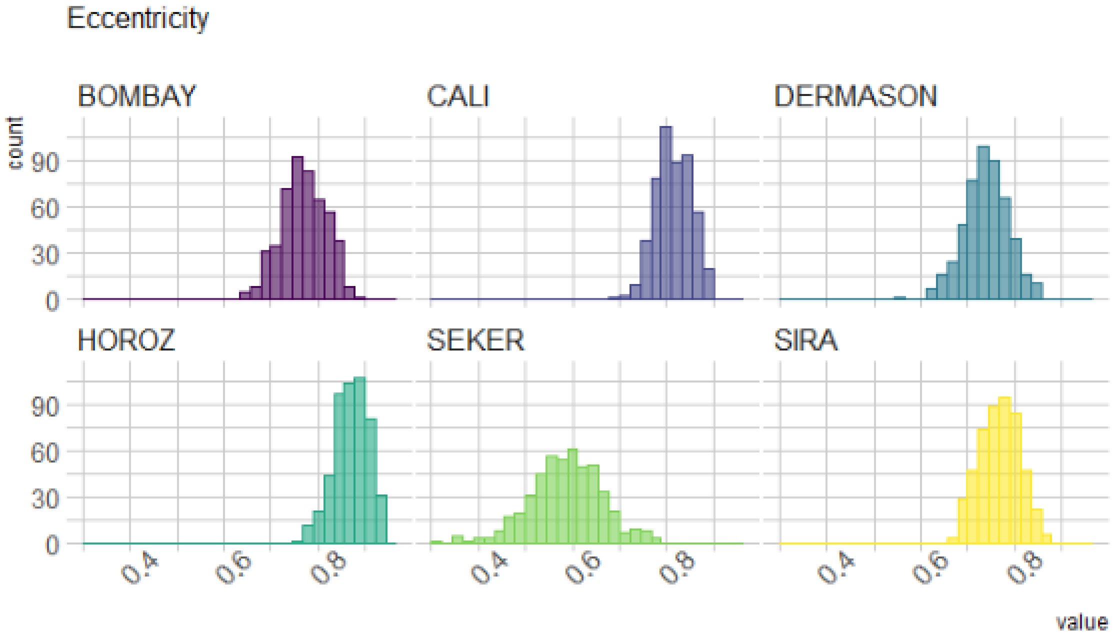


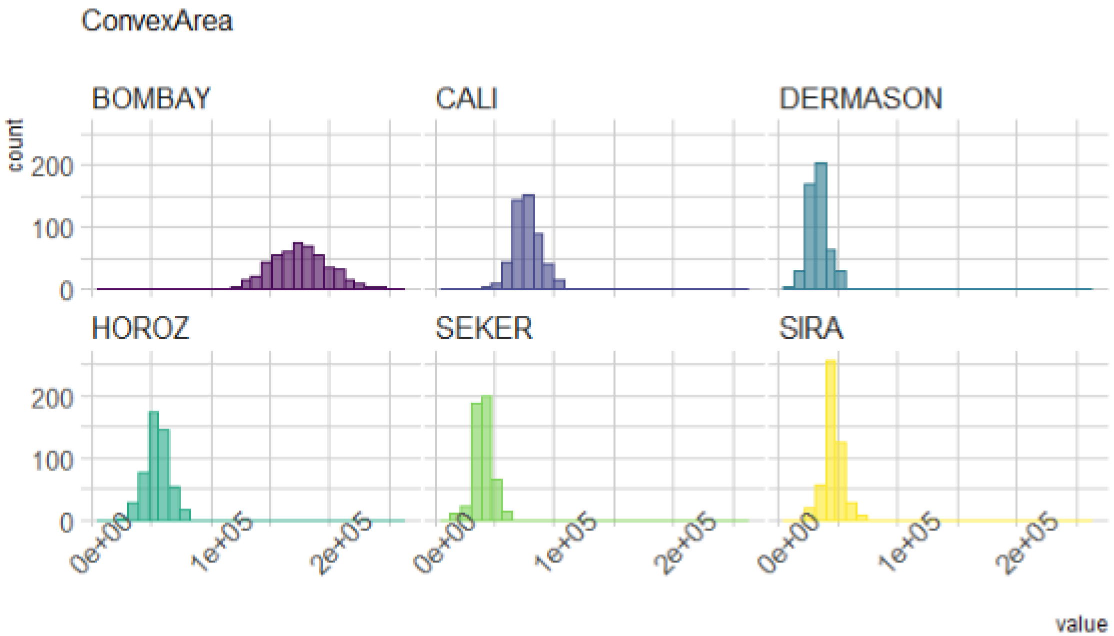
Exploratory Analysis



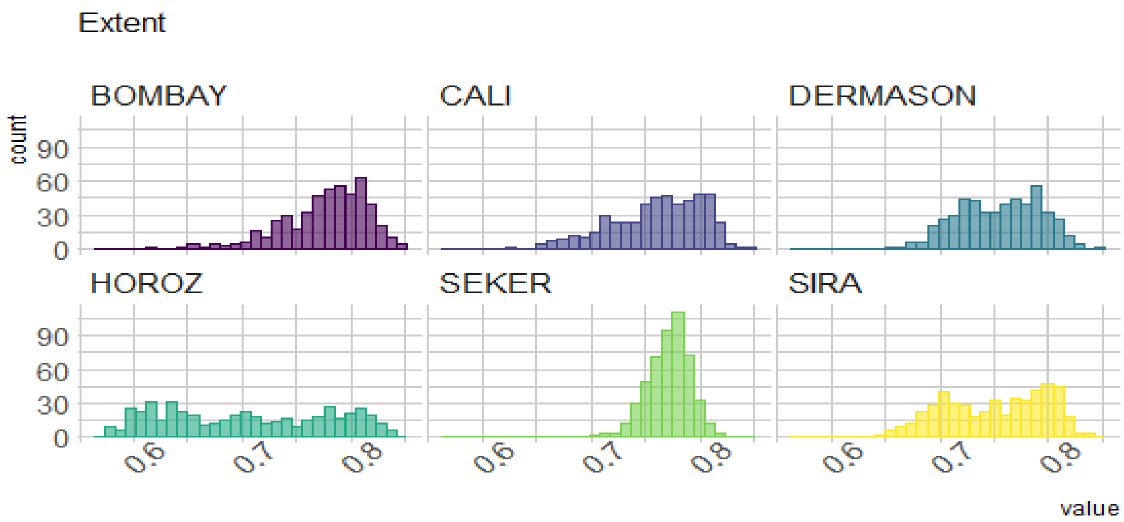


Exploratory Analysis





Exploratory Analysis



Class <chr>	Freq <int>
BOMBAY	500
CALI	500
DERMASON	500
HOROZ	500
SEKER	500
SIRA	500

- Results:
  - Most of the data is normally distributed with the least normally distributed variable being Extent
  - The dataset has many outliers as revealed by the boxplots. These will be removed in order optimize the predictive models.

Exploratory Analysis

- Results After Outlier Removal
  - Dataset was reduced from 3000 observations to 2731

Class <chr>	Freq <int>
BOMBAY	460
CALI	468
DERMASON	470
HOROZ	474
SEKER	422
SIRA	437

Predictive Analysis

- Best Subset Selection results:
  - The below sets of features have the highest RSS for their variable count
  - The set with 4 features was chosen as it had the highest adjusted R^2 score

	ID	(Intercept)	Area	Perimeter	MajorAxisLength	MinorAxisLength	Eccentricity	ConvexArea	Extent	RSquared
1	1	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	0.516706108713031
2	2	TRUE	TRUE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	0.533035489974861
3	3	TRUE	TRUE	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE	0.536511679697907
4	4	TRUE	TRUE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	0.537073700856129
5	5	TRUE	TRUE	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	0.53690714422355
6	6	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	0.536726908670107
7	7	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	0.536514755893378

Results Analysis

- Leave-One-Out Cross Validation Results and chosen models:

	ModelID	Model_Name	Accuracy	Weight_Differential	Cost_Differential
1	1	Multinomial Logistic Regression	0.9849	0.15	0.0393
2	2	Linear Discriminant Analysis	0.978	6.95	0.0275
3	3	K-Nearest Neighbors	0.9876	5.01	0.012
4	4	Quadratic Discriminant Analysis	0.9679	1.05	0.0951
5	5	Naive Bayes Analysis	0.9611	1	0.0383

- 2 models were chosen as they were both optimal but based on different criteria

Results Analysis

- Test dataset results:

	ModelID	Model_Name	Accuracy	Weight_Differential	Cost_Differential
1	1	Multinomial Logistic Regression	0.9835	0.22	0.0055
2	2	K-Nearest Neighbors	0.9817	1.21	0.0232

- The results on the test dataset indicated strong results on all three measures and by both models.

Conclusions

- Multinomial Logistic Regression and K-Nearest Neighbors were the most accurate when making predictions on bean types based on the morphometric measurement features
  - Both had accuracy rates of over 98% on the validation and test datasets
  - Both also had the lowest weight differential and cost differential respectively compared to the other models when tested on the validation dataset

References

- Bobbitt, Z. (2019, April 20). *How to Normalize Data in R*. Retrieved from Statology: https://www.statology.org/how-to-normalize-data-in-r/
- Clark Science Center. (2016). *Lab 8 - R*. Retrieved from Clark Science Center @ Smith College: https://www.science.smith.edu/~jcrouser/SDS293/labs/lab8-r.html
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning with Applications in R Second Edition.* Springer.
- KARA, M., SAYINCI, B., ELKOCA, E., ÖZTÜRK, İ., & ÖZMEN, T. B. (2013). Seed Size and Shape Analysis of Registered Common Bean (Phaseolusvulgaris L.) Cultivars in Turkey Using Digital Photography. *Journal of Agricultural Sciences*, 220-221.
- UCLA: Statistical Consulting Group. (n.d.). *Multinomial Logistic Regression | R Data Analysis Examples*. Retrieved from UCLA Advanced Research Computing: https://stats.oarc.ucla.edu/r/dae/multinomial-logistic-regression/
- Xiaozhou, Y. (2020, May 9). *Linear Discriminant Analysis, Explained*. Retrieved from Towards Data Science: https://towardsdatascience.com/linear-discriminant-analysis-explained-f88be6c1e00b