MODERN APPPLIED STATITICS - FINAL PROJECT

Group 11 - Kenneth Annan; Iftekhar Chowdhury; Jie Hu; Siyi Lui; Richard Acquah-Sarpong

Contents

Introduction

Dry bean- Phaseolus vulgaris L. is a major cultivated grain species in the genus Phaseolus that is widely consumed worldwide for its edible legume and pea pods (Heuze et al., 2015). Nevertheless, selecting the best seed species is one of the main concerns for both bean producers and the market. Since different genotypes are cultivated worldwide, it is important to separate the best seed variety from the mixed dry bean population, otherwise the market value of these mixed species of beans could drop enormously (Varankaya & Ceyhan, 2012). The aim of our project is to develop an automated method to multiclass classification of dry beans that could predict the net worth of a given bean species harvested from a 'population cultivation' from a single farm when presented in the market.

Data Exploration

For this project, we used two datasets namely 'labeled' and 'unlabeled' sets. The labeled (training) dataset contains 3000 observations and 8 variables. The dependent variable has 6 levels (Classes): BOMBAY, CALI, DERMASON, HOROZ, SEKER, and SIRA. Each class has 500 observations. The unlabeled dataset is drawn from the three samples namely Sample A, B, and C. The total observations for sample A, B, and C are 777, 1373, and 982 respectively. Roundness, which is the measure of how closely the shape of beans approaches a perfect circle, was calculated and added as an additional predictor variable (Koklu & Ozkan, 2020). Tables 1 through 4 show the summary statistics of the variables in the labeled data, Sample A, B, and C, respectively . 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 will dominate over those with small ranges which may lead to biased results, therefore it is necessary to transform/scale these variables before fitting our distance-based models (i.e., KNN and SVM). Table 7 shows that Bombay has the highest grams and price per seed. The price per seed is the product of the price per pound and price per seed divided by the total weight of 453.592 grams. The histograms from the labeled data (Figure 1) show evidence of multimodality behavior in the variables. This means that at least one of the classes of beans is very distinct from the others. The multimodality behavior is also shown in the histograms from Sample A (Figure 2), but not from Sample B or C (Figures 3 & 4). The boxplots from the labeled data (Figure 5) show that BOMBAY and CALI beans are very distinct from the other beans. It can be seen from the boxplots that Roundness and Extent seems to be a strong predictor for the SEKER. Eccentricity seems to be a good predictor to HOROZ. The violin plots for each class (Figure 5) shows that most of the class distributions are approximately normal except for the distributions for Roundness and Extent. From these distributions, we expect BOMBAY and CALI to be easily predicted by our models. We expect to see very low predictions of BOMBAY for Sample B and C, because there is no multimodality behavior in their histograms. The variance of each variable by class shows evidence of non-constant variance (Tables 5 & 6). Based on the normality distribution and non-constant variance, we expect the QDA model to perform well. Most of the variables except for Eccentricity, Extent, and Roundness, are highly correlated (Figure 6) in each dataset. This behavior is also seen in the correlation of the variables by classes (Figure 7). The principal component analysis (Figure 8) indicates that the first 3 principal components, which are new variables that are constructed as linear combinations or mixtures of the initial variables, explained more than 90% of all variance in the dataset.

Methodology

This is a multiclass classification problem. Therefore, we tried five supervised classifiers including Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), K-Nearest Neighbors (KNN), Random Forest (RF), and Support Vector Machines (SVM). In this project, we first use the labeled dataset to fit these models with different feature selections, and compare their leave-one-out cross validation (LOOCV) performance, then choose the model with the best performance and do prediction on sample A, B, and C datasets. The last step is to use bootstrap technique to do a simulation and measure the prediction accuracy.

Feature Selection: We have tried four different feature combinations. The first one uses all the 8 features. The second one uses only four variables: Area, Eccentricity, Extent, and Roundness in order to get rid of the

effect of high correlation among other features. The third one uses only three variables: Area, Eccentricity, and Extend in order to see the effect of the newly added Roundness variable. The last one uses the first three principal components to check whether PCA helps denoise the dataset.

Model description: Both LDA and QDA classifiers are based on Bayes' theorem, with the assumption that every class is normally distributed. However, LDA has constant variance assumption among all classes, while QDA relaxes the assumption of constant variance among all classes. The LDA produces a linear decision boundary, while QDA produces a non-linear decision boundary. The QDA also requires more training data due to its non-constant variance assumption compared to LDA. The Random Forest classifier contains a large number of individual decision trees, where each individual tree in the random forest produces a class prediction and the class with the most votes becomes our model's prediction. LOOCV is used to select the optimal number of features, 'mtry' and optimal number of trees, 'ntree'. The KNN classifier predicts the observation class by finding the majority of the classes of the k-nearest training data points. Where, 'nearest' implies minimum Euclidean distance. LOOCV is used to find the optimal k (Figure 9). The SVM classifier identifies the best hyperplane that acts as a decision boundary among the different classes. We use a radial kernel for the SVM model in this project.

Results & Discussion

Our results (Figure 10) indicate that there is not much of a difference in performance for each of the five models while using all variables, four selected variables, and three selected variables. All these models underperform when three principal components are used. For each type of feature combination, QDA, Random Forest, and SVM consistently outperform other models. We selected QDA as our final model with all variables for predictions on sample A, B, and C. From Table 13, we see that all samples have a small number of predictions for BOMBAY. The classes with the highest prediction are CALI and SEKER for sample A, DERMASON for sample B, and HOROZ for sample C. We also visualize this comparison in Figure 12. Then we calculated the predicted price for each sample, the result is shown in the 'Predicted.Net.Worth' column of Table 17. Sample A is predicted to have higher price than Sample B and C. Finally, we construct the probability of count of each class given that the predicted class is one of the six classes (Table 16), and use it to do a bootstrap simulation and get the prediction interval for each sample (Table 17). Sample A has a narrower 2.5% to 97.5% price prediction interval, compared to Sample B and C.

Conclusion & Recommendation

QDA, Random Forest, and SVM did a good job of predicting the beans classes, their LOOCV accuracy rates are all 90%. Extent and Eccentricity are good predictors, Area, Perimeter, MajorAxisLength, MinorAxisLength, and ConvexArea are highly correlated, either one of them can be a good predictor. The newly added variable Roundness does not add much prediction power to our models. With the final model we used, QDA, it has a better prediction accuracy for Sample A compared to Sample B and C. We recommend using any one of these three models (QDA, Random Forest, and SVM), and including Extent, Eccentricity and at least one of the five highly correlated features as predictors to automate the classification of dry beans.

Appendix

Data exploration

Statistic summary

Table 1: Statistical distribution of features of dry beans varieties (in pixels) - Label

	n	mean	sd	median	min	max	range	se
Area	3000	69874.978	49578.516	48714.500	20645.000	251320.000	230675.000	905.176
Perimeter	3000	1012.238	347.749	941.897	384.169	2164.100	1779.931	6.349
MajorAxisLength	3000	362.048	124.520	332.901	161.517	740.969	579.452	2.273
MinorAxisLength	3000	225.193	73.350	202.735	106.003	473.395	367.391	1.339
Eccentricity	3000	0.756	0.102	0.773	0.301	0.945	0.644	0.002
ConvexArea	3000	70944.115	50382.269	50807.500	8912.000	259965.000	251053.000	919.850
Extent	3000	0.753	0.052	0.766	0.571	0.850	0.279	0.001
Roundness	3000	0.840	0.294	0.771	0.391	2.056	1.664	0.005

Table 2: Statistical distribution of features of dry beans varieties (in pixels) - Sample A

	n	mean	sd	median	min	max	range
Area	776	60857.161	27800.948	57064.500	25438.000	236248.000	210810.000
Perimeter	776	967.796	245.790	948.700	448.388	1952.837	1504.449
MajorAxisLength	776	338.230	96.130	340.992	187.939	728.672	540.733
MinorAxisLength	776	223.890	36.682	220.362	158.560	425.634	267.074
Eccentricity	776	0.700	0.133	0.740	0.222	0.896	0.674
ConvexArea	776	61431.496	28731.925	56920.000	19481.000	240540.000	221059.000
Extent	776	0.764	0.034	0.771	0.624	0.839	0.215
Roundness	776	0.846	0.292	0.776	0.442	1.994	1.552

Table 3: Statistical distribution of features of dry beans varieties (in pixels) - Sample B

	n	mean	sd	median	min	max	range
Area	1373	35831.824	7170.688	35431.000	19037.000	59464.000	40427.000
Perimeter	1373	747.579	163.305	745.685	361.623	1180.854	819.231
MajorAxisLength	1373	256.920	34.212	254.191	171.036	374.017	202.981
MinorAxisLength	1373	175.298	25.281	174.232	108.597	255.549	146.952
Eccentricity	1373	0.718	0.078	0.731	0.370	0.879	0.509
ConvexArea	1373	36149.414	9272.275	35657.000	9344.000	67922.000	58578.000
Extent	1373	0.756	0.038	0.761	0.625	0.840	0.215
Roundness	1373	0.904	0.385	0.801	0.379	2.231	1.852

Table 4: Statistical distribution of features of dry beans varieties (in pixels) - Sample C

	n	mean	sd	median	min	max	range
Area	982	50634.567	13126.088	50965.000	20077.000	114858.000	94781.000
Perimeter	982	916.364	191.883	908.451	392.745	1410.768	1018.023
MajorAxisLength	982	343.353	62.700	358.140	175.513	515.070	339.557
MinorAxisLength	982	185.840	27.556	183.423	114.305	296.549	182.244
Eccentricity	982	0.826	0.068	0.839	0.621	0.949	0.328
ConvexArea	982	51256.353	14490.685	51111.000	9637.000	116575.000	106938.000
Extent	982	0.725	0.067	0.734	0.551	0.847	0.296
Roundness	982	0.814	0.292	0.742	0.399	1.882	1.483

labeled data: variance check for each classes

Table 5: Variance of distribution

Class	Var.Area	Var.Perimeter	var.Maj.Axis.	var.Min.Axis.	var.Eccentricity
BOMBAY	23509.529	32015.80	3177.8992	826.6840	0.5472634
CALI	9567.048	26272.79	1188.1780	491.4581	0.6183656
DERMASON	4965.074	22913.81	696.7121	498.7684	0.5494947
HOROZ	7534.314	24960.58	1252.4894	456.5644	0.7227374
SEKER	4750.492	24170.41	736.8507	419.4165	0.3006355
SIRA	4758.298	23977.06	782.4715	401.8085	0.6098838

Table 6: Variance of distribution (continuted)

Class	var.ConvexArea	var.Extent	var.Roundness
BOMBAY	259965	0.8502428	1.262811
CALI	117510	0.8427527	1.512446
DERMASON	56174	0.8471957	2.055745
HOROZ	82462	0.8420894	1.620064
SEKER	65674	0.8183099	1.990205
SIRA	73945	0.8418021	1.718602

Price per seed

Table 7: distribution of types of dry beans and prices per seed

	price per pound	grams per seed	price per seed
BOMBAY	$5.56	1.92	0.023535
CALI	$6.02	0.61	0.008096
DERMASON	$1.98	0.28	0.001222
HOROZ	$2.43	0.52	0.002786
SEKER	$2.72	0.49	0.002938
SIRA	$5.40	0.38	0.004524

Histogram of each feature

Histogram of each feature - Labeled Data

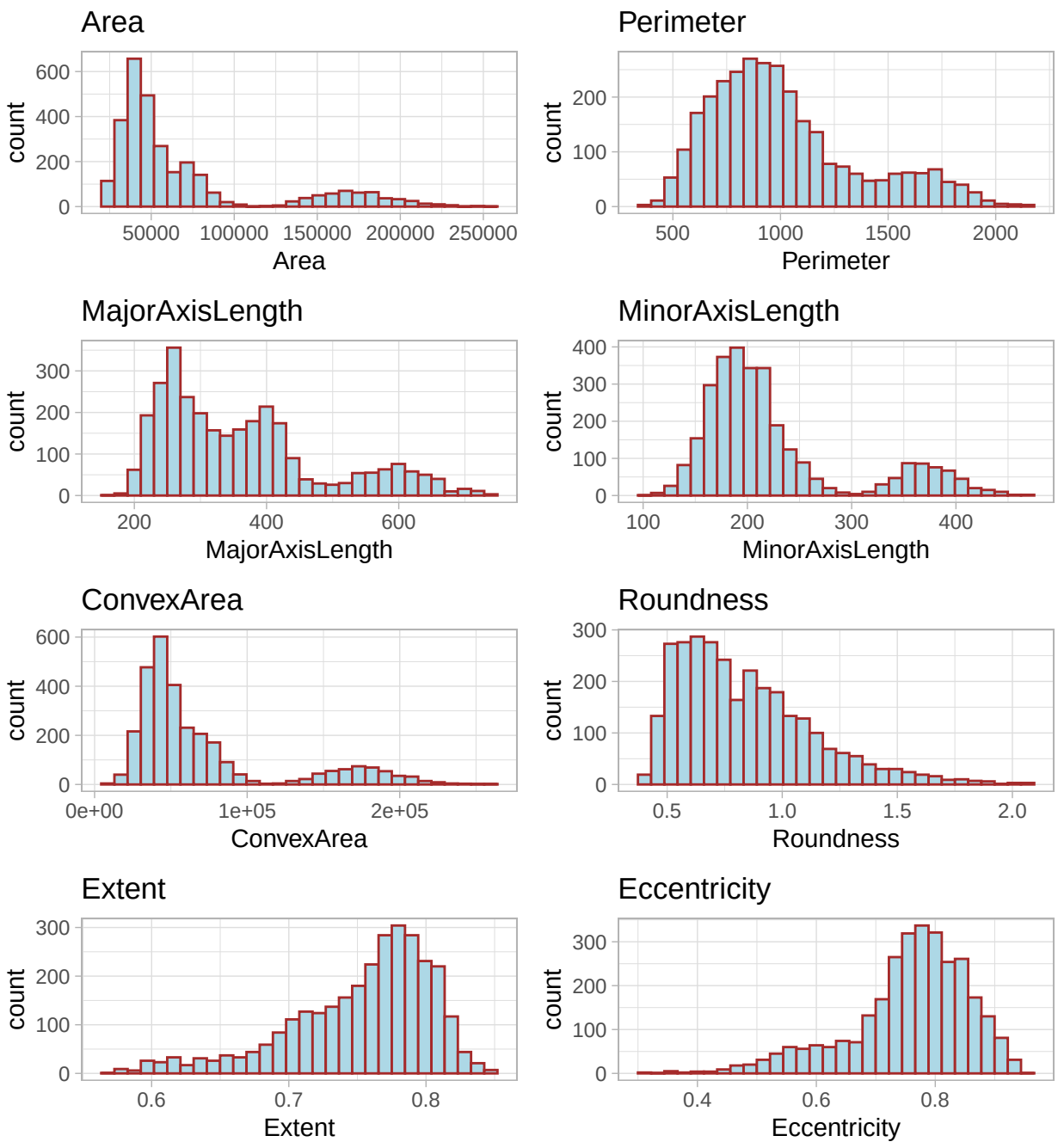

Figure 1: Histograms of Variables - Labeled Data

Histogram of each feature - Sample A

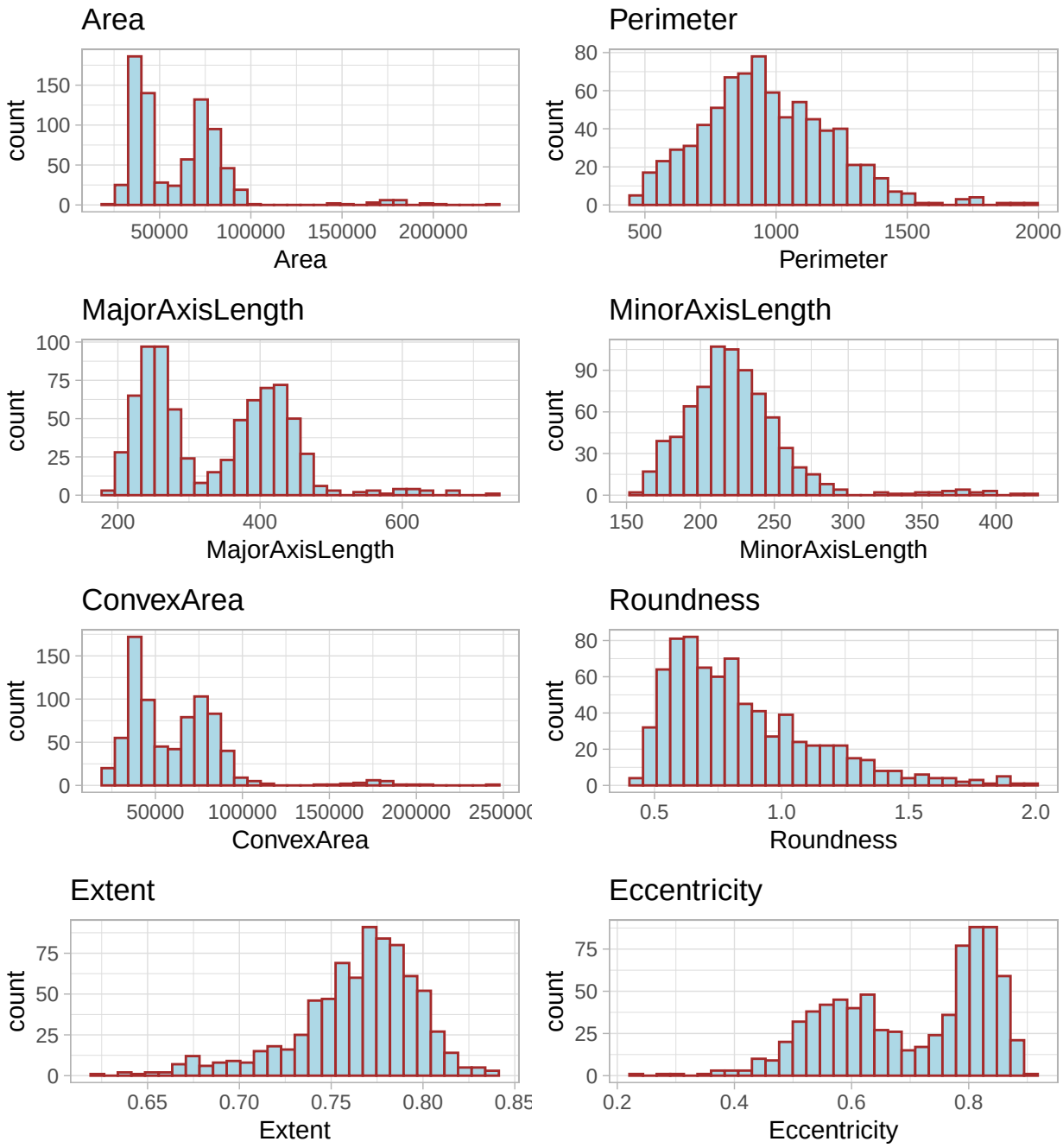

Figure 2: Histograms of Variables - Sample A

Histogram of each feature - Sample B

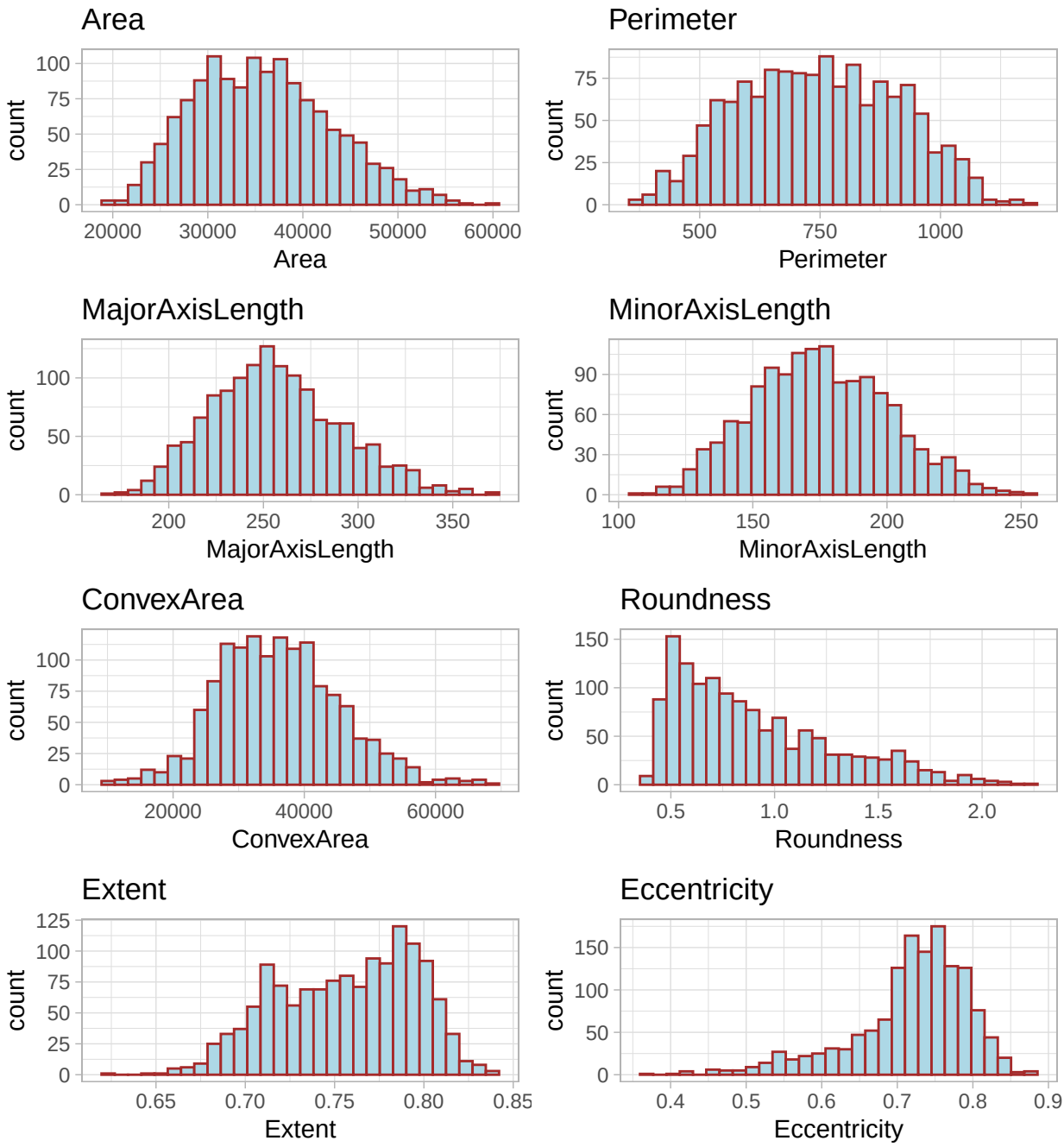

Figure 3: Histograms of Variables - Sample B

Histogram of each feature - Sample C

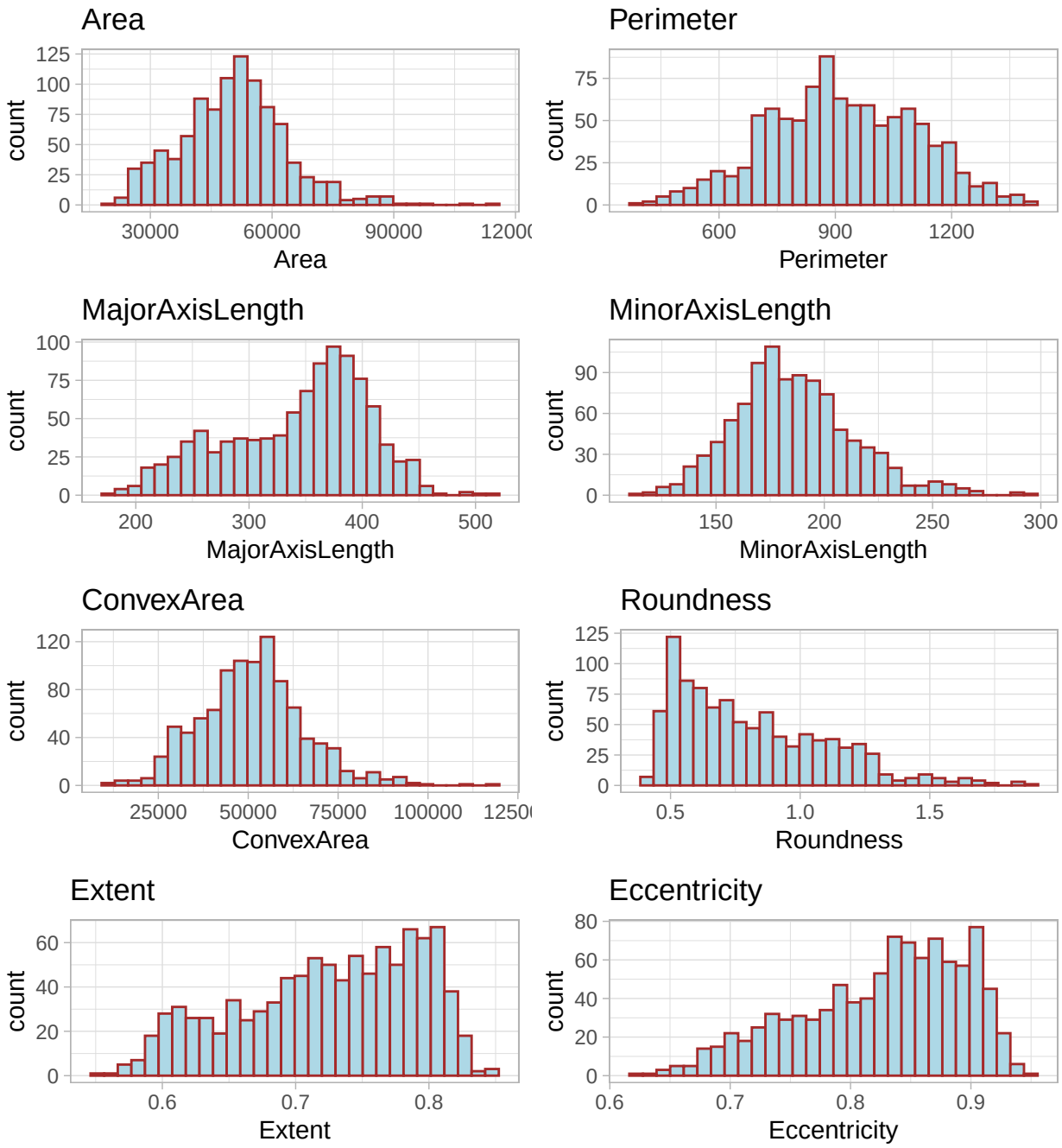

Figure 4: Histograms of Variables - Sample C

Boxplot and Violin plots for each class

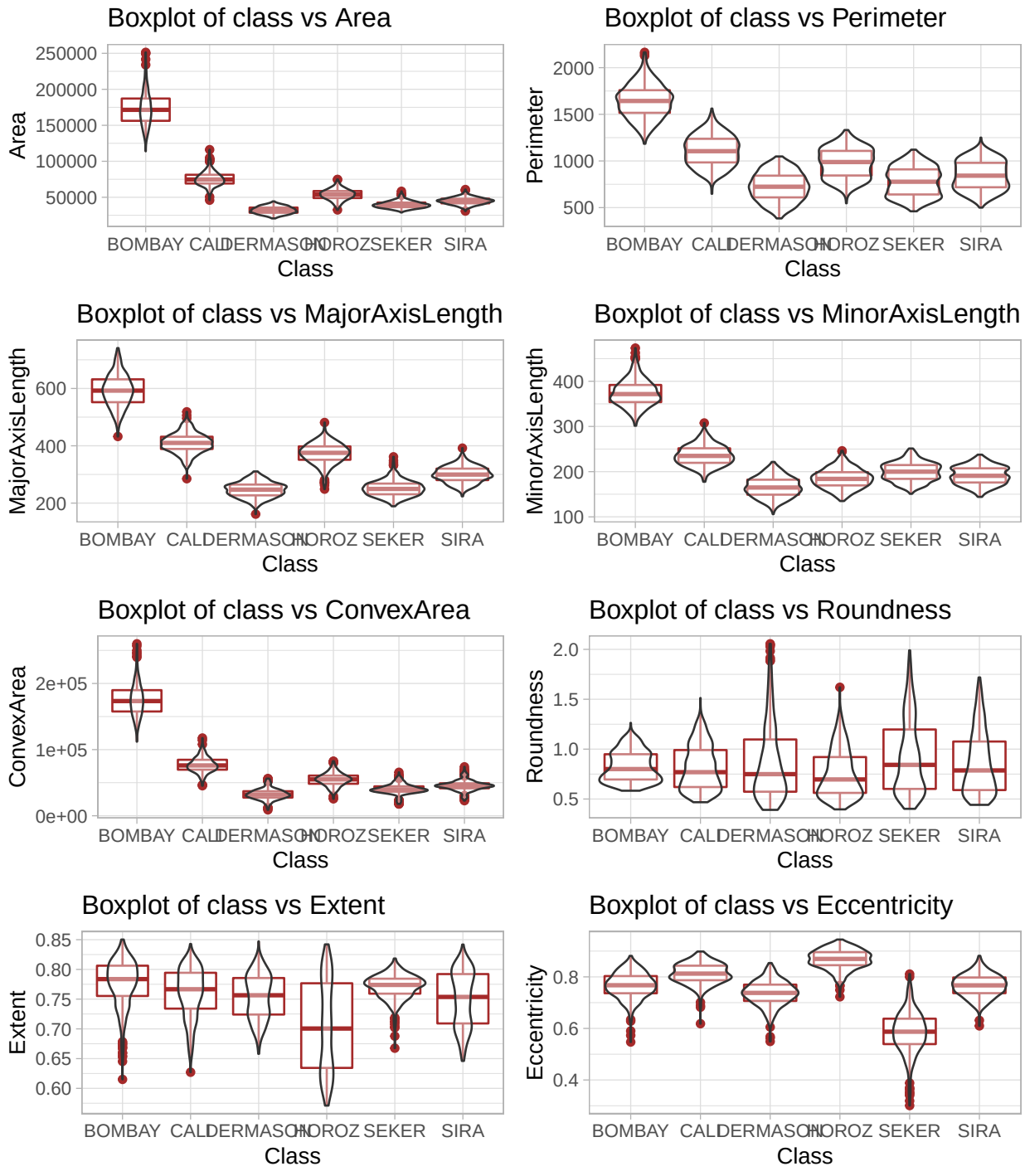

Figure 5: Boxplots and Voilin Polts of Variables by Classes

Correlation Plot

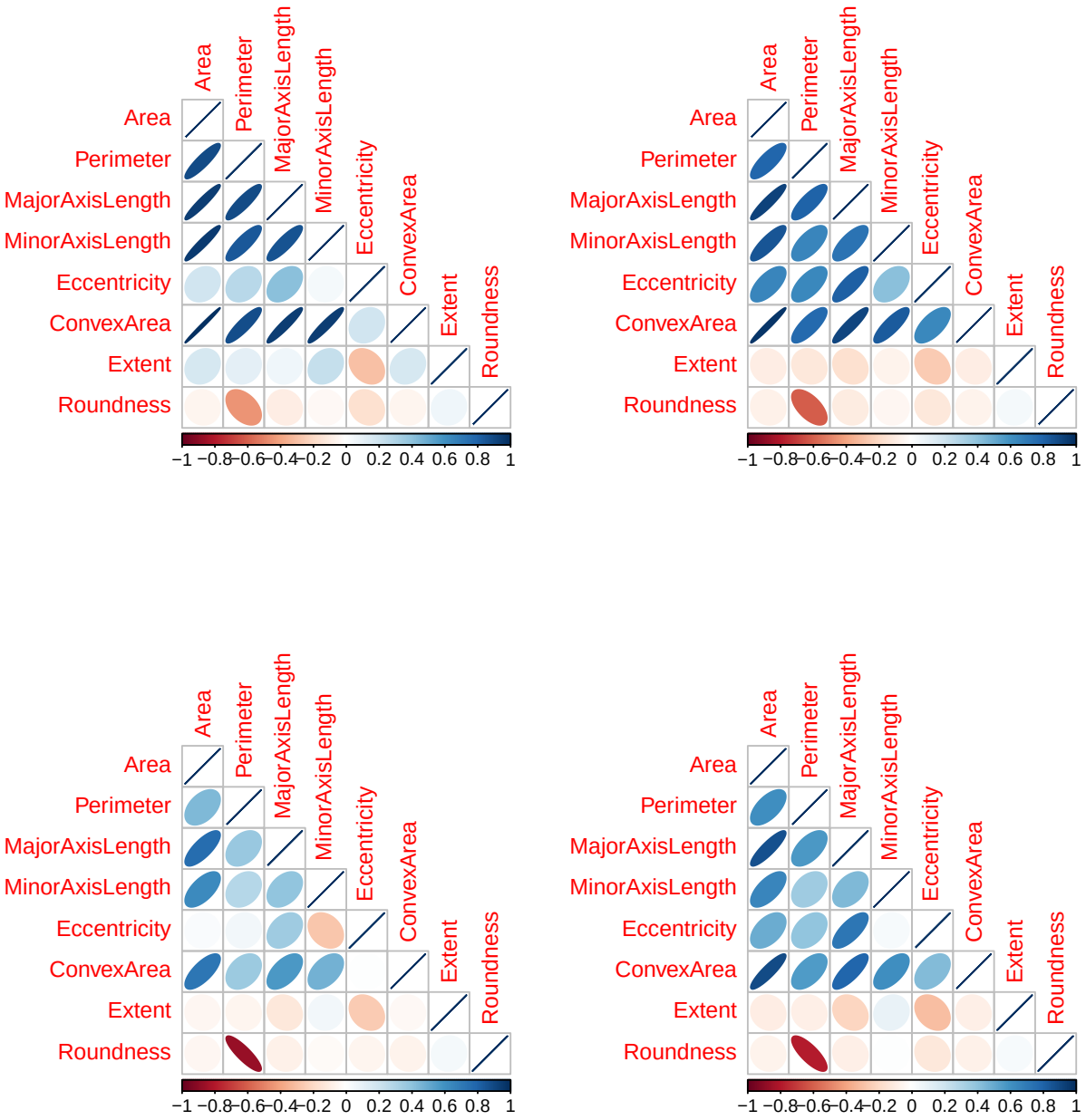

Figure 6: Correlation plot

Correlation plot by class for labeled dataset

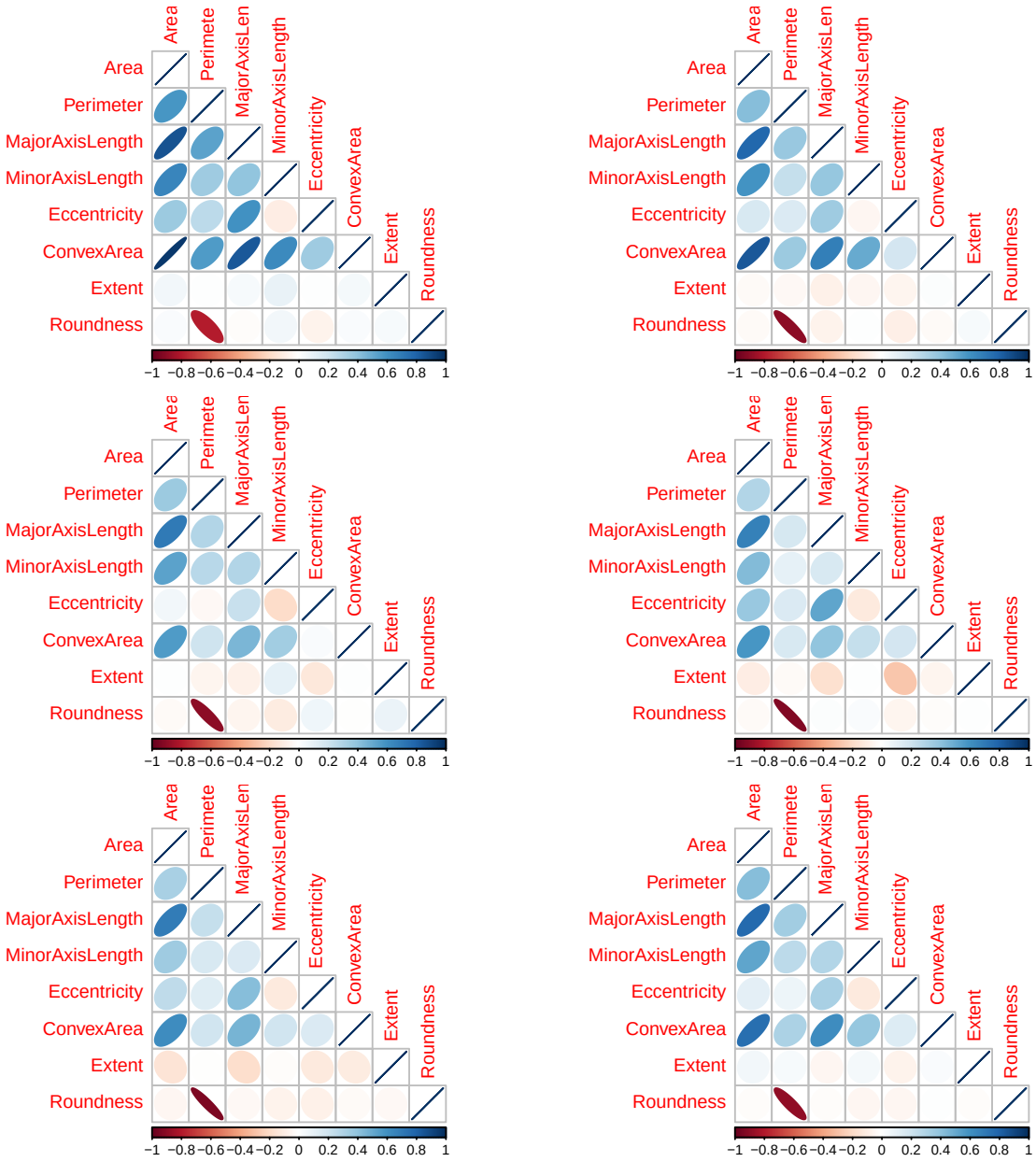

Figure 7: correlation plot by class for labeled dataset

Principle components analysis

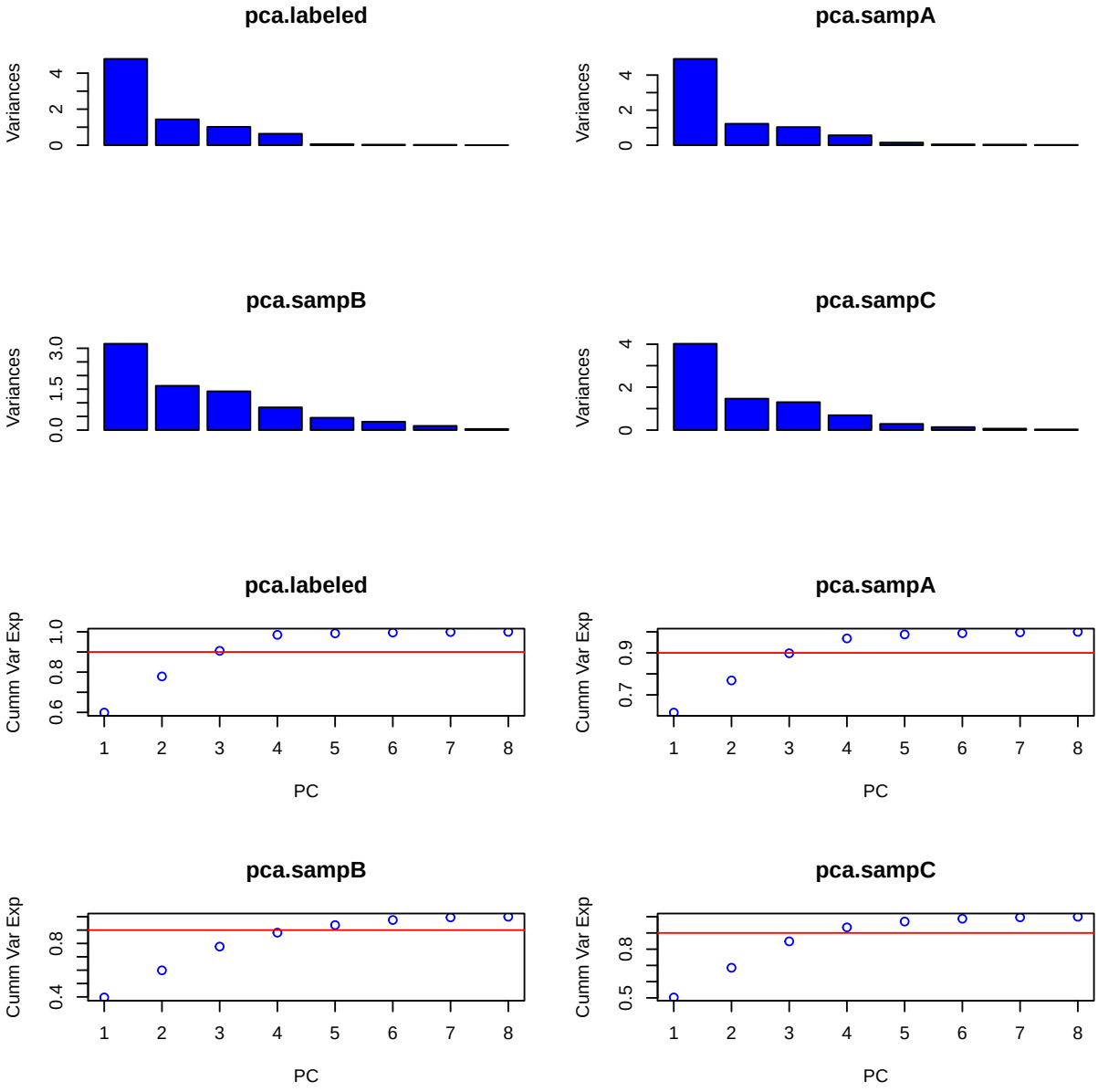

Figure 8: Variance explained by each components

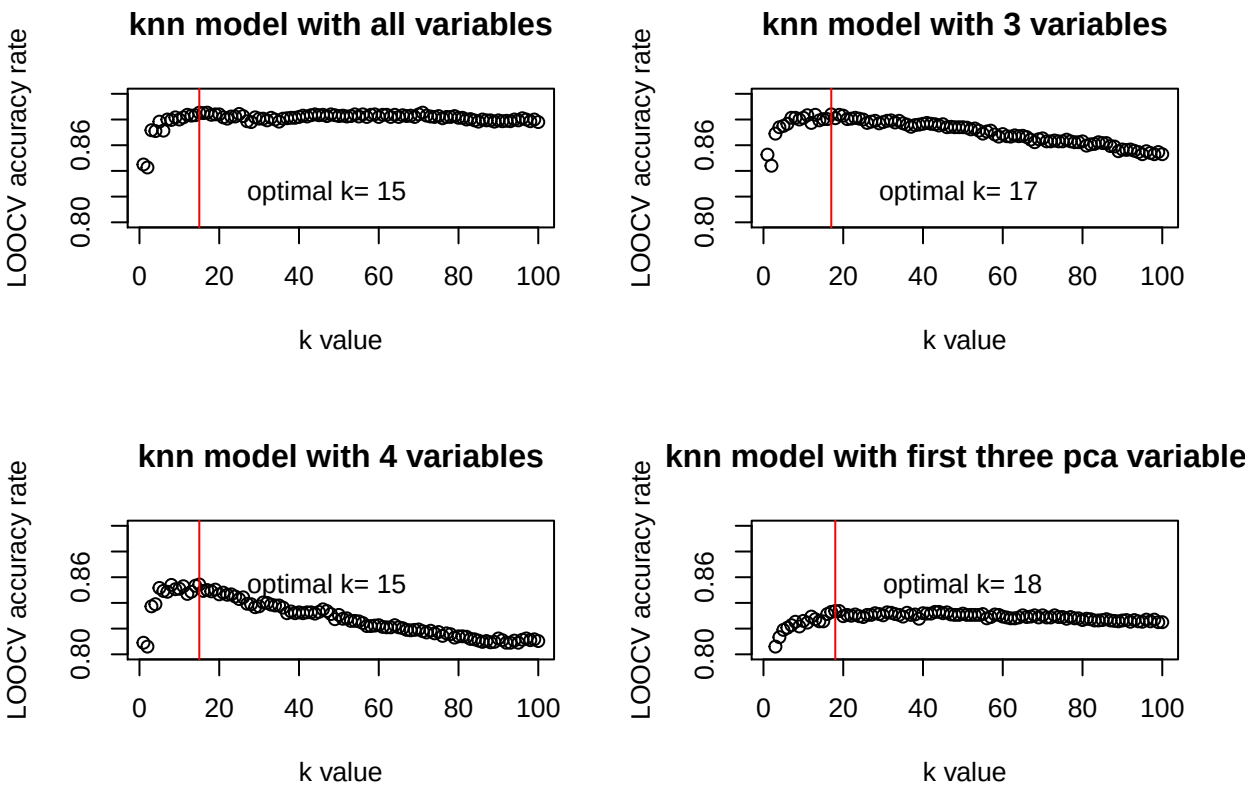

Figure 9: optimal k value choices plots for knn model

Table of Performance Measures

Table 8: Average LOOCV Accuracy Rate across Classes

	lda	qda	RandomForest	knn.sc	svm.sc
all.var	0.86	0.90	0.90	0.88	0.90
3var	0.87	0.90	0.90	0.88	0.90
4var	0.87	0.90	0.90	0.85	0.89
3pca	0.81	0.83	0.82	0.83	0.83

Table 9: Average LOOCV Precision across Classes

	lda	qda	RandomForest	knn.sc	svm.sc
all.var	0.87	0.90	0.90	0.89	0.90
3var	0.88	0.90	0.90	0.89	0.90
4var	0.87	0.90	0.90	0.86	0.89
3pca	0.82	0.83	0.82	0.84	0.90

Table 10: Average LOOCV Recall across Classes

	lda	qda	RandomForest	knn.sc	svm.sc
all.var	0.86	0.90	0.90	0.88	0.90
3var	0.87	0.90	0.90	0.88	0.90
4var	0.87	0.90	0.90	0.85	0.89
3pca	0.81	0.83	0.82	0.83	0.83

Table 11: Average LOOCV Specificity across Classes

	lda	qda	RandomForest	knn.sc	svm.sc
all.var	0.97	0.98	0.98	0.98	0.98
3var	0.97	0.98	0.98	0.98	0.98
4var	0.97	0.98	0.98	0.97	0.98
3pca	0.96	0.97	0.96	0.97	0.97

Table 12: Average LOOCV F1.score across Classes

	lda	qda	RandomForest	knn.sc	svm.sc
all.var	0.86	0.90	0.90	0.89	0.90
3var	0.88	0.90	0.90	0.88	0.90
4var	0.87	0.90	0.90	0.85	0.89
3pca	0.81	0.83	0.82	0.84	0.84

Graph of Performance measures

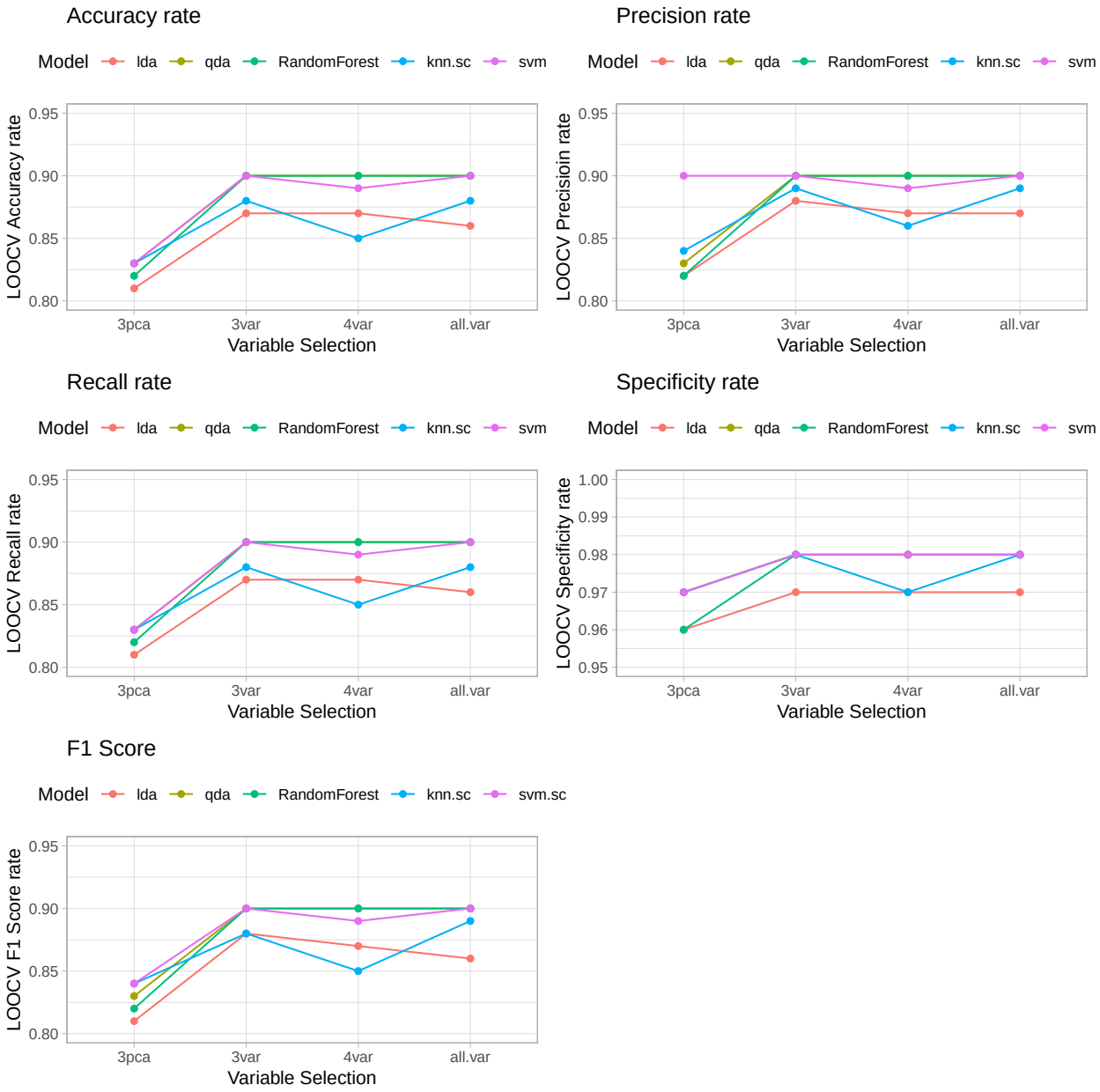

Figure 10: Model Performance

Visualize best-selected model:qda with all variables

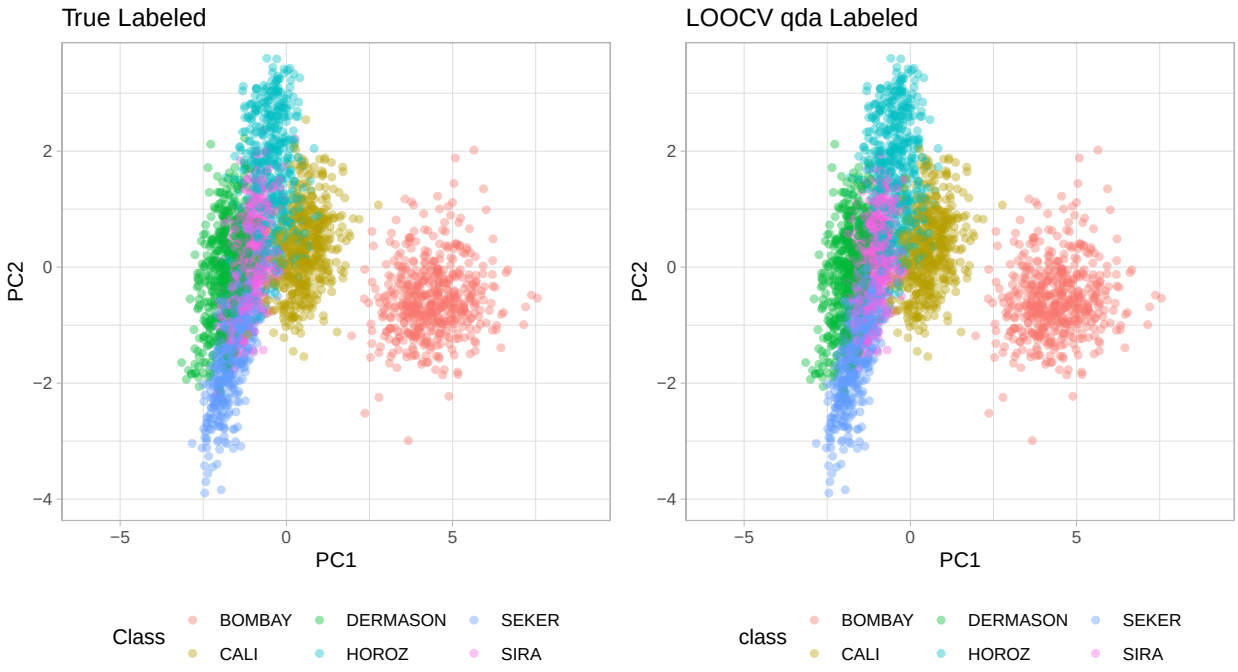

Figure 11: Final selected model (QDA)

Classes prediction result

Table 13: Prediction result for each sample

	BOMBAY	CALI	DERMASON	HOROZ	SEKER	SIRA	Num.obs.
sampleA	22	359	12	12	345	26	776
sampleB	0	1	779	15	238	340	1373
sampleC	1	102	161	540	8	170	982

Visualize classes prediction

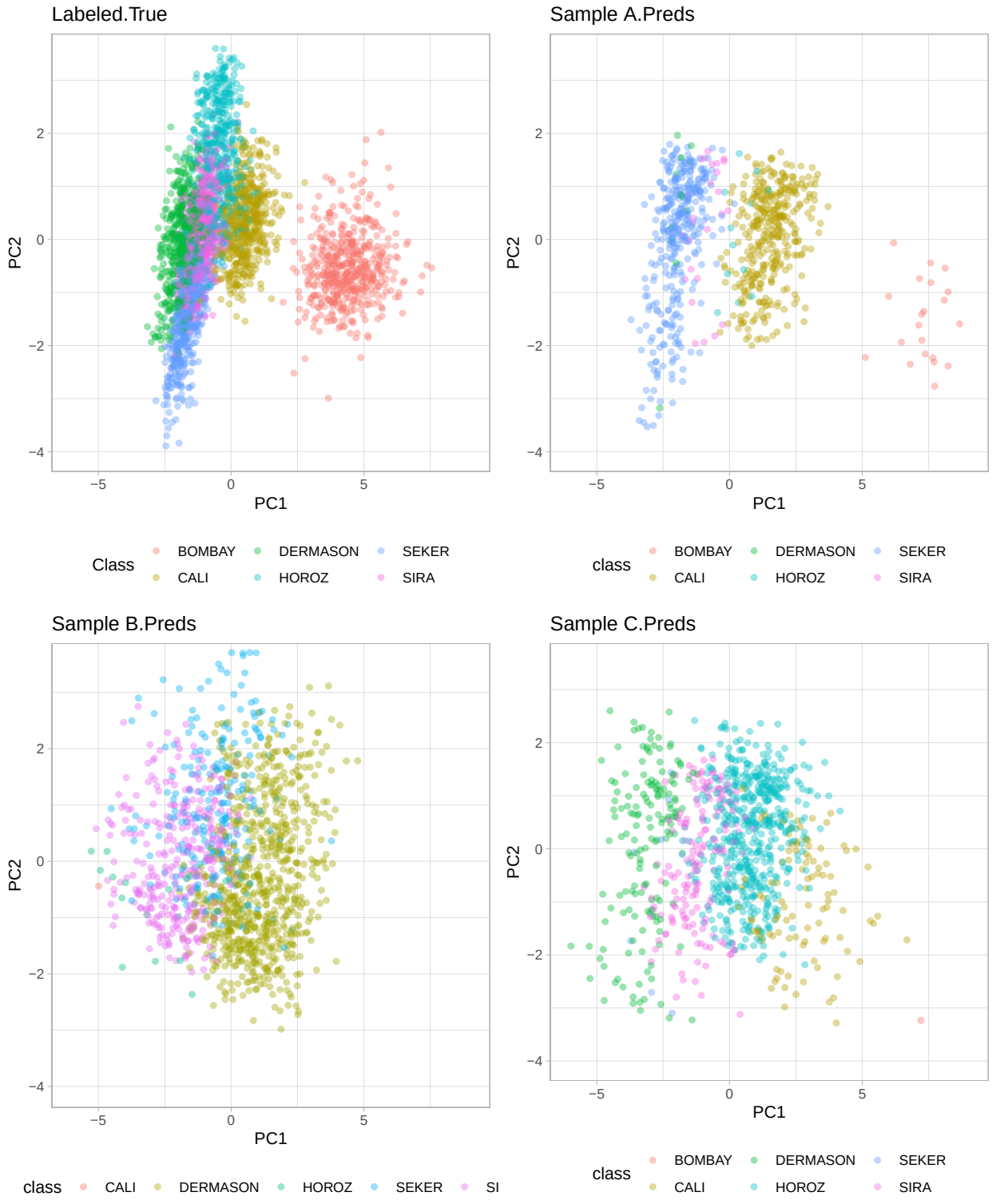

Figure 12: Prediction Visualization

Confusion matrix of label data with LOOCV QDA

Table 14: Confusion matrix of label data with LOOCV QDA (all variables)

	True.BOMBAY	True.CALI	True.DERMASON	True.HOROZ	True.SEKER	True.SIRA
Pred.BOMBAY	500	0	0	0	0	0
Pred.CALI	0	479	0	19	1	2
Pred.DERMASON	0	0	416	6	13	38
Pred.HOROZ	0	16	3	449	0	36
Pred.SEKER	0	2	16	0	454	24
Pred.SIRA	0	3	65	26	32	400

Table 15: rowsums of confusion matrix

	Num.Preds
Pred.BOMBAY	500
Pred.CALI	501
Pred.DERMASON	473
Pred.HOROZ	504
Pred.SEKER	496
Pred.SIRA	526

Table 16: multinominal distribution estimation

	True.BOMBAY	True.CALI	True.DERMASON	True.HOROZ	True.SEKER	True.SIRA
Pred.BOMBAY	1	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
Pred.CALI	0	0.9560878	0.0000000	0.0379242	0.0019960	0.0039920
Pred.DERMASON	0	0.0000000	0.8794926	0.0126850	0.0274841	0.0803383
Pred.HOROZ	0	0.0317460	0.0059524	0.8908730	0.0000000	0.0714286
Pred.SEKER	0	0.0040323	0.0322581	0.0000000	0.9153226	0.0483871
Pred.SIRA	0	0.0057034	0.1235741	0.0494297	0.0608365	0.7604563

Prediction result and accuracy

Table 17: prediction result and accuracy (in dollars)

	0%	2.5%	97.5%	100%	Predicted.Net.Worth	Range
samp.A	4.45	4.48	4.57	4.59	4.60	0.09
samp.B	3.22	3.25	3.39	3.44	3.24	0.14
samp.C	3.35	3.37	3.49	3.55	3.34	0.13

References

1. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York: springer.

2. Heuzé V., Tran G., Nozière P., & Lebas F. (2015). Common Bean (Phaseolus vulgaris), Feedipedia.org – Animal Feed Resources Information System – A programme by INRA, CIRAD, AFZ and FAO, http://www.feedipedia.org/node/266 (accessed on 29 April 2021).

3. Koklu, M., & Ozkan, I. A. (2020). Multiclass classification of dry beans using computer vision and ma-chine learning techniques. Computers and Electronics in Agriculture, 174, 105507. doi:10.1016/j.compag.2020.105507

4. Varankaya, S., & Ceyhan, E. (2012). Problems Encountered in Bean Farming in the Central Anatolia Region and Solution Suggestions. Selçuk Tarım Bilim. Journal. 26, 15–26.

5. https://en.m.wikipedia.org/wiki/Sensitivity_and_specificity

6. https://www.geeksforgeeks.org/loocvleave-one-out-cross-validation-in-r-programming/

7. https://towardsdatascience.com/what-is-out-of-bag-oob-score-in-random-forest-a7fa23d710

8. https://alekhyo.medium.com/interview-questions-on-svm-bf13e5fbcca8://alekhyo.medium.com/interview-questions-on-svm-bf13e5fbcca8