STATS 4M03/6M03: Multivariate Analysis

Final Project

Analysis of Kernels of Three Different Wheat Varieties

November 24, 2023

1 Introduction

1.1 Abstract

This report focuses on the seeds dataset, specifically on distinguishing the differentiation of three seed varieties through multiple multivariate analysis, such as K-Nearest Neighbors (KNN), Classification Trees, and Random Forests. The analysis begins with an Exploratory Data Analysis (EDA) to gain a comprehensive overview of the dataset, followed by the in-depth analytical methods.

1.2 The Data

This dataset is obtained from UC Irvine Machine Learning Repository. Also, the dataset "comprised kernels belonging to three different varieties of wheat: Kama, Rosa and Canadian" (UCI Data), including 210 samples (each wheat class was represented by 70 samples) and 7 variables of wheat kernels: area, perimeter, compactness, length of kernel, width of kernels, asymmetry coefficient, and length of kernel groove.

1.2.1 Exploratory Data Analysis (EDA)

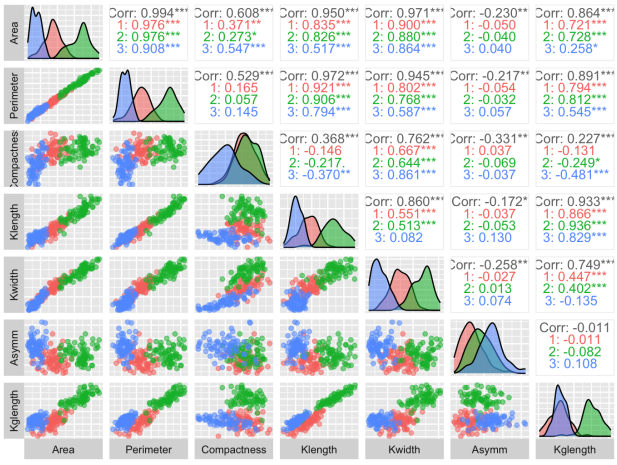


Figure 1: Pair Plot

All 7 variables follow a linear distribution, which suggests linearity. Area, Perimeter and Kwidth has clear separation between each group of seeds.

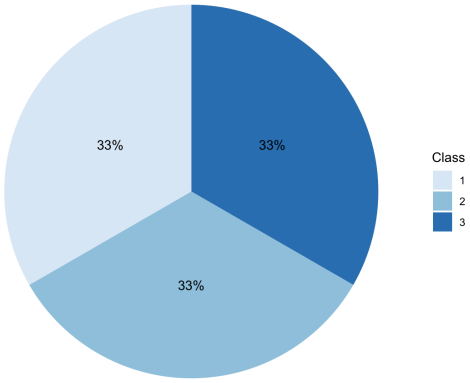


Figure 2: Pie chart

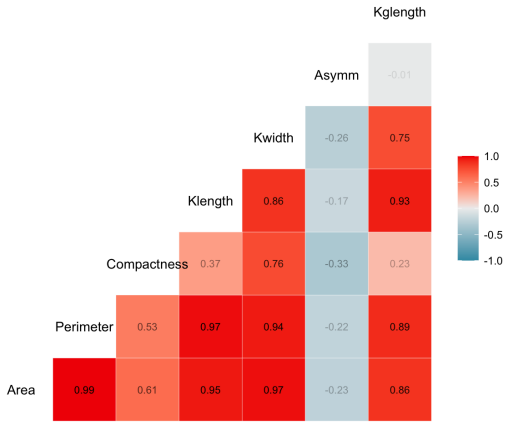


Figure 3: Detailed Correlation

The left graph plots compares compares the relative size of each class in dataset, and every class constitutes equal number of seeds' sample. And the right graph illustrates the correlation between each variable. Each variable has a positive correlation with each other except for Asymm, which depicting a negative correlation with the rest. While the correlations are strong and positive, Compactness shows a relatively weaker positive correlation. Perimeter and Area has a strongest positive correlation, at 0.99 and compactness and Asymm has a strongest negative correlation, at -0.33.

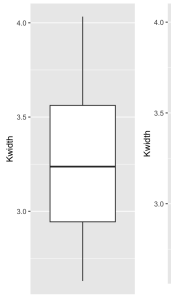
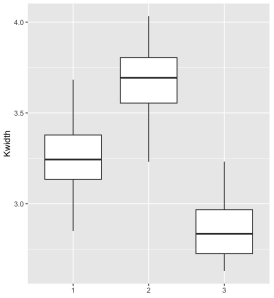


Figure 4: Kwidth Grid



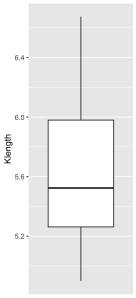
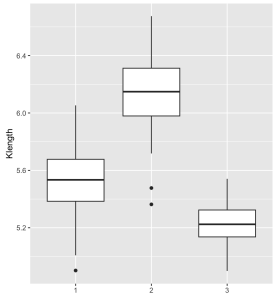
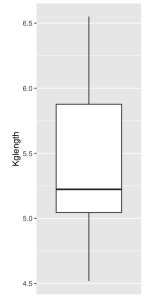


Figure 5: Klength Grid





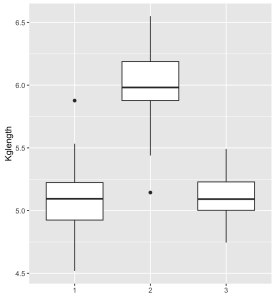


Figure 6: Kglength Grid

These plots compare the distribution for variable Klength, Kwidth and Kglength for both whole classes and each class individually. Kwidth shows a great skewness, while both Klength and Kglength has outliers.

1.2.2 Data Preparation

First of all, In order to make different codes get consistent results, we set the seed to a specific value(1) in R to ensure the repeatability of the random process.

The data split is also necessary. The dataset is split into an 80% training set and a 20% test set. By

the 80-20 split, there is a large portion of the data available to effectively train the model, while still retaining a large amount of data for evaluation.
Besides,

2 Methodology

2.1 K-Nearest Neighbourhood (KNN)

K-nearest neighbor (KNN) is a simple and efficient algorithm that makes predictions or classifications of an unknown class based on the labels of the k closest observations with a known class. In this project, KNN is applied for classification, predicting the class of each kernel based on seven features. For this algorithm, the parameter is the value of k.

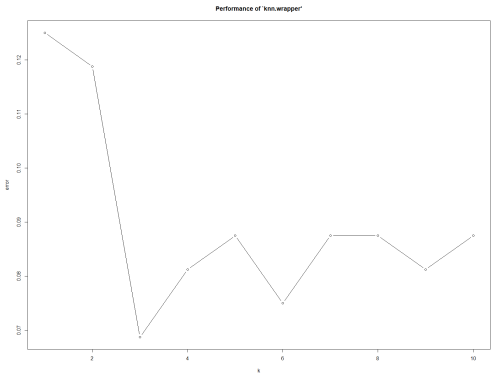


Figure 7

From the figures above, the optimal value of k was determined by 10-fold cross-validation, which resulted in k = 3 as the best choice, the model has the smallest error.

2.2 Classification Trees

Classification tree is a type of predictive modeling that built through a process known as binary recursive partitioning. This is an iterative process of splitting the data into partitions and then splitting it up further on each of the branches. Initially, all objects are considered as a single group. the group is split into two subgroups using a criteria, say high values of a variable for one group

and low values for the other. The two subgroups are then split using the values of a second variable. The splitting process continues until a suitable stopping point is reached.

This is the Classification tree plotted based on the data set.

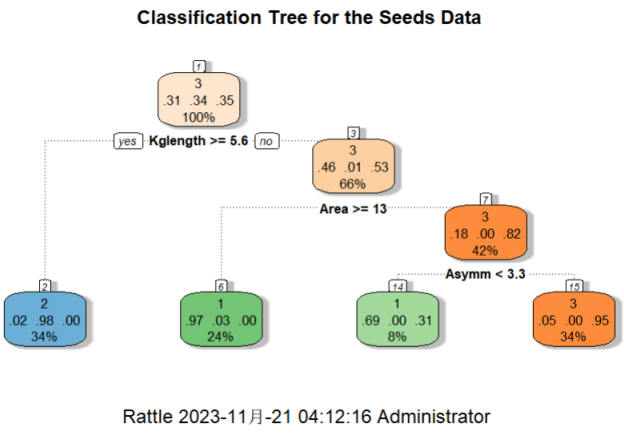


Figure 8: Tree

From the graph it can be concluded that almost all observation with Kglength $\geq$ 5.6 belongs to class 2. From those data that Kglength $<$ 5.6, 97 % of the observation with Area $\geq$ 13 belong to class 1. And from rest of the data, which, Kglength $<$ 5.6 and also Area $<$ 13, 95 % of the observation that Asymm $\geq$ 3.3 belong to class 3, 5 % belong to class 1, and for those obervation that Asymm $<$ 3.3, 69 % of those belong to class 1, other belong to class 3

2.3 Random Forests

Random Forest is an ensemble learning method used for both classification and regression tasks. It is based on generating a large number of decision trees at training time, and then using the training set to recognize, classify, and make predictions for the class. In the seeds dataset, out of 210 data, 160 are used as the training set, while 50 are used as the test set. Random Forests are applied to determine parameters, such as the number of decision trees to build and the minimum number of random samples required for each individual tree.

In the above figure, by constructing 500 decision trees, using 1 to 7 random variables in each tree, the optimal values can be determined by 5-fold cross validation with ntree=100 and mtry=5.

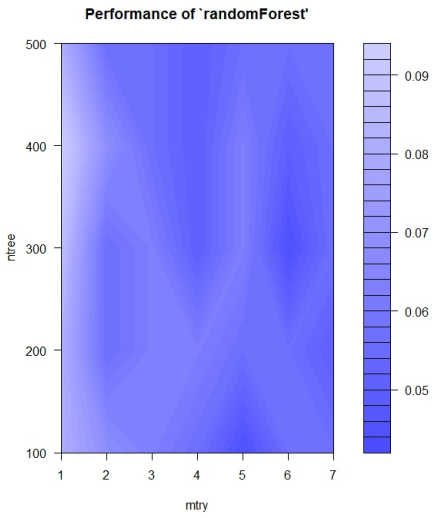


Figure 9

Subsequently, by building 100 decision trees, each employing 5 random variables, the calculations of Out-of-Bag Error Estimation: $MSE_{oob} = \frac{1}{M}\sum_{m=1}^{M}(\widehat{f}^{m}(x) - \widehat{f}_{bag}(x))^2$ indicated that three treatments in the dataset had 92.5% accuracy in categorizing the data (OOB error rate=7.5%).

Based on the case of ntree=100 and mtry=5, construct the variance importance plot:

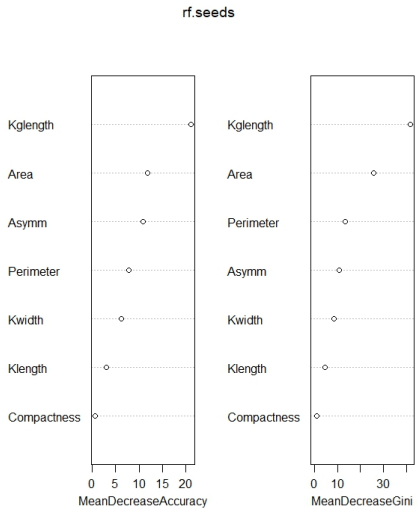


Figure 10: varImpPlot

In the variance importance plot, the higher the value of Mean Decrease Accuracy or Mean Decrease Gini Score, the higher the importance of the variables in the model. From Figure 10, it can

be inferred that, for the kernels of wheat varieties, Kglength is the most important variable.

3 Discussion

Based on the Methodology part above, we can use the code to compute that, for the Classification Tree method, we finally get the table as:

KNN Table

	Kama	**Rosa**	**Canadian**
Kama	19	2	1
Rosa	1	14	0
Canadian	0	0	13

Classification Tree Table

	Kama	**Rosa**	**Canadian**
Kama	19	1	3
Rosa	0	15	0
Canadian	1	0	11

Random Forests Table

	Kama	**Rosa**	**Canadian**
Kama	19	0	1
Rosa	1	15	0
Canadian	1	0	13

Based on the table on the left, which is the table of classification tree, the ARI and the misclassification rate can be calculated, ARI is around 0.7132, and the misclassification rate is 10%. This means that the model is accurate for the majority of observation, but may have room for improvement. Based on the table in the middle, which is the table of Knn, The table shows that most test samples are correctly classified, The misclassification error is 8%, and the ARI is 0.75, which indicates a moderate agreement between the predicted and true labels. Based on the table in the right, which is the table of random forests, the ARI and the misclassfication rate can be calculated, the table shows that the ARI of the model is 0.817, and the misclassification rate is 6%, which means the strong performance in classification accuracy. Most of the data have been correctly classified.

	pram	**misclassification rate**	**ARI**
KNN	$k = 3$	8%	0.75
classification tree	$Klength, Area, Asymm$	10%	0.71
random forests	$ntree = 100, mtry = 5$	6%	0.82

Compare these three methods, we can conclude that the ARI of the Tree model is 0.71, the misclassification rate is 10%, and the ARI of the Knn model is 0.75, its misclassification rate is 8%, and the ARI of the Random Forests is 0.82, its misclassification rate is 6%, since we need to choose the method that is most accurate, we want our ARI to be larger, also we need out our misclassification

rate to be smaller cause the smaller the misclassification rate, the better the observation can be classified by our model. So comparing these three methods, we find that the Random Forests have the highest ARI and lowest misclassification rate, so we recognize it is the best model out of these three, and the Tree model has the lowest ARI and highest misclassification rate, so we think it is the least accurate model.

4 Conclusion

In summary, the three classification methods of KNN, decision tree and random forests are applied to distinguish the differentiation of three wheat class varieties (Kama, Rosa, Canadian). For KNN, the best parameter is choosing 3-nearest neighbourhood, which indicates that there are three closest points to the sample point. For decision tree, the 3 parameters length of kernel groove, area and asymmetry coefficient are used for classification. This explains that as long as the kglength $<$ 5.6, the area $<$ 13 and the asymm $\geq$ 3.3 it will be categorized as a Canadian wheat grain. While the other variables were classified into Kama and Rosa respectively(Figure 8). For the random forests, the optimal parameter is ntree=100, mtry=5, which suggests constructing 100 decision trees, each with 5 variables for classification, yields the highest accuracy. However, the variation between groups makes it impossible to confidently predict the wheat variety for each kernel in the test dataset. These errors are primarily associated with the Kama variety, this is because the ranges of the Kama variables in the dataset all overlap with the ranges of the Rosa and Canadian varieties. Conversely, since there is no overlap between Canada seed and Rosa seed, they will not be misclassified. Moreover, this dataset isn't perfect. There are some outliers within the data for the kernels of the three different wheat varieties. For more accurate results and lower misclassification rates, data processing of outliers is one of the most effective methods in further research.

5 Bibliography

1. Charytanowicz,Magorzata, Niewczas,Jerzy, Kulczycki,Piotr, Kowalski,Piotr, and Lukasik,Szymon. (2012). seeds. UCI Machine Learning Repository. https://doi.org/10.24432/C5H30K.

2. RStudio Team. (2023). RStudio: Integrated Development Environment for R (Version 2023.06.2+561)

[Computer software]. RStudio, PBC. Retrieved from https://www.rstudio.com/

3. Breiman L., Friedman J.H., Olshen R.A. and Stone, C.J. (1984). Classification and Regression Trees. Chapman & Hall/CRC.

4. Zhou, Z.-H. (2012), Ensemble Methods: Foundations and Algorithms, Boca Raton: Chapman & Hall/CRC Press.