ST494 Statistical Learning Project
Identifying and Clustering Iris Species Using Statistical Learning

Prepared for

Dr Sunny Wang

Shihui Zhou(181473200)
Zixuan Deng(160691110)

April 2nd 2020

Contents

1. Introduction

The Fisher's or Anderson's iris data gives the measurements in centimeters of the variables sepal length and width and petal length and width, respectively, for 50 flowers from each of 3 species of iris. The species are Iris setosa, versicolor, and virginica. Figure 1 is the picture for all three species.

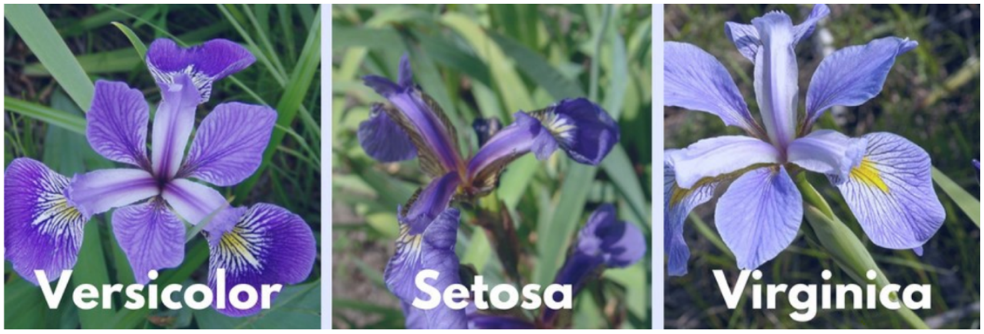


Figure 1: Picture of different Iris Flowers. [1]

1.1 Objective

In this project, we would like to apply both supervised learning and unsupervised learning methods to analyze the iris data. For the supervised learning part, we decide to apply classification algorithms to classify the iris species according to the corresponding length and width of petals and sepals as the species of iris are labelled in the data set. On the other hand, we can also apply clustering techniques to cluster our observations into 3 groups, which is typically an unsupervised learning technique.

[1] Av, S. (2019, October 13). Classification of Iris dataset. Retrieved from https://medium.com/@sa2253/classification-of-iris-dataset-b4310ddf0482

2. Data

2.1 Data Descriptions

In the iris data set, there are 150 observations of iris recorded and each observation contains 5 variables - Sepal.Length, Sepal.Width, Patel.Length, Patel.Width, and Species. So, there are 750 values in total contained in iris data set. And there is no missing value out of the 750 values. Table 1 illustrates more specific information for each of the 5 variables.

We select Species, which is the only categorical variable, to be the response variable for our classification models, and the results of statistical summary for the response variable are displayed in Table 2. And we can visualize Species by the bar chart displayed in Figure 2.1.1. As each of the 3 classes of Species contains the equal number of observations (50 out of 150 observations), the data set is balanced so the resampling strategies for imbalanced data is no need to be applied. In addition, we also take a look of all of the numerical explanatory variables. And Table 3 summarizes the statistical information for each of the independent variable.

2.2 Exploratory Visualizations

We first illustrate the data without group by species. Figure 2.2.1 is the histogram of four explanations. It seems the frequency of sepal width does form a normal distribution. The global maximal is located around 3cm. By looking at the rest of histograms, we cannot find any evidence of distribution among the count of explanatory variables. Some bins are empty. From the histogram of petal length and histogram of petal width, we find an interesting thing that the histograms have been split into two parts. We see that most petal lengths are located around 1cm.

When we look at the histogram with colored different species, we find that versicolor and virginica follow similar patterns in the histograms of sepal length, sepal width. Setosa usually has the highest number in the bin that has a global maximum. By looking at figure 2.2.2 (iris density plot), we also get the similar conclusion as the histogram. The density plot of setosa has the thinnest tail among three species, and it is usually located in the left or right corner of the whole density plot.

2.3 Data Preprocessing

There is no missing value in the iris data set. Thus, we do not need to consider the data cleaning techniques for this case further. Moreover, we notice that the unit used for each variable is the same, so we do not scale the dataset until some algorithms are required.

Based on figure 2.3.1, the box plot, we find there are only 4 outliers in 150 observations. As the outlier is in the amount of 2.67% in the whole dataset, we do not consider doing anything to handle these outliers.

As there are only 150 observations in the dataset, which is a pretty small dataset, we need to have enough observations in the test dataset. We first randomly shuffle the dataset and then use 70% of observations to build the training dataset, and remaining observations belong to the test dataset. Table 4 shows the number of three species in the train and test dataset. We see that the number of observations in three observations does not have a huge difference.

We used best subset selection to select the feature. There are four combinations of features. Looking at figure 2.3.2, the plot of Cp, Adjusted R Square suggests no.3 selection, while BIC suggests no.2 selection. Since the dataset only has four explanatory variables, we choose no.3

selection, which includes three variables. From figure 2.3.3, we choose variables “ Sepal.Width”, “Petal.Length”, “Petal.Width” to build models.

3. Algorithms and Techniques

3.1 Supervised Learning

Supervised Learning is the traditional part in statistical learning. Within the label of the given predictive variable, we can do classification of iris species. We have built models using multinomial logistic regression, KNN, LDA, QDA, Random Forest, Decision Tree, and SVM. Since we only use two datasets to build the model, k-fold cross validation is necessary to tune the parameter in each model. Caret package in R provides an effective function to train, predict and tune parameters for models. In our case, we use 10-fold cross validation to set the trainControl.

Figure 3.1.1 shows the summary of training accuracy and Kappa in each model. Overall, LDA performs better than other methods. Training accuracy gives us an idea how well the model trained, this does not mean LDA has the best performance in the Iris dataset. Sometimes, the model performs well in the training dataset and it has a high misclassification rate in the testing dataset, which is called overfitting. Table 5 shows the accuracy and Kappa of optimal model, and we notice that QDA has the highest accuracy and kappa, which is 0.9778 and 0.9659 respectively. However, since the Iris dataset does not have class imbalance problems, we mainly look at the accuracy of the model as the criteria to evaluate model performance. Logistic regression, LDA, Decision Tree and Random Forest do not perform well as the other algorithms. This makes sense because some variables in the dataset have strong relationships, which makes the tree based and logistic regression algorithm not as accurate as other models because they all require independent observations.

3.1.1 QDA

So, we will analyze the best performance model - QDA. QDA classifier assumes each class of Y are drawn from Gaussian distribution. But, QDA is not like LDA, QDA do not assume the predictor variables have common variance. From Figure 3.1.1.1, we see the output only contains the group means. But it does not contain the coefficients of the linear discriminants as QDA classifier involved a quadratic, rather than a linear function of the predictors. Figure 3.1.1.2 is the performance of QDA. The confusion matrix illustrates the result, as we could see that there is only one observation misclassified.

3.1.2 Hyperparameter

By using caret package, we tune the hyperparameters for each model. For LDA and QDA, there is no hyperparameter. For logistic regression, we use 0.5 as the threshold. We use k =9 for KNN model. In Support vector machine, sigma parameter in the RBF kernel determines the reach of a single training instance. If the value of sigma is low, then every training instance will have a far reach. Conversely, high values of sigma mean that training instances will have a close reach. The cost parameter decides how much an SVM should be allowed to “bend” with the data. For a low cost, it is a smooth decision surface and for a higher cost we classify more points correctly. Our SVM model has sigma 1.231 and cost parameter 0.5. For complexity parameter in decision tree is the minimum improvement in the model needed at each node. In decision tree model, the optimal value for Cp is 0. Mtry defines the number of variables randomly sampled as candidates at each split. Optimal Mtry for randorm forest is 2, and the number of tree we use default value 500.

3.2 Neural Network

Constructing a neural network to predict the Species by the selected variables (Sepal.Width, Patel.Length, and Patel.Width) is also applied in this project. Firstly, we treat Sepal.Width, Patel.Length, and Patel.Width as the three inputs for the neuron of the neural network, and the output layer also contains three parts, which are the three classes for Species (setosa, versicolor, virginica). Then a hidden layer with size 10 is constructed. Figure 3.2.1 is the visualization of the neural network.

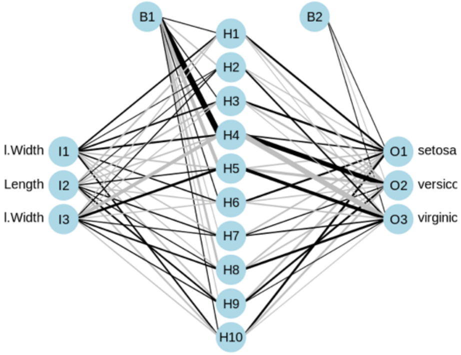


Figure 3.2.1: Visualization of the Neural Network

There are 73 weights assigned in our 3-10-3 neural network, and we found that the decay for the final model is 0.01. The overall accuracy of our neural network for iris data is 95.56%, and the 95% confidence interval is (0.8485,0.9946). We also obtained a confusion matrix to summarize our predicted results.

As shown in Table 6, the neural network we built can predict observations whose type are setosa and virginica completely perfect. In addition, there are 2 false predictions existed as they predict these two observations to be virginica but the true prediction for them should be versicolor. Overall, the neural network is able to predict the majority observations (43 out of the total 45 observations) correctly, which results in the 95.56% accuracy.

3.3 Clustering

In this project, we also tried some unsupervised algorithms to explore the structure of the iris data by grouping them into 3 cases, as there are 3 classes of Species in total. For this part, we have considered two popular algorithms for cluster analysis:

1. Hierarchical Cluster Analysis
2. The K-Means Algorithm

For the Hierarchical cluster analysis, we considered to use the Euclidean distance to determine the similarity or dissimilarity. First of all, we measure how close our clusters are by plotting dendrograms for the three methods - single linkage, complete linkage, and average linkage. From Figure 3.3.1, we can see that the dendrogram for single linkage is quite like two strings of clusters, and the right part contains more strings of clusters. On the other hand, the other dendrograms look more similar, as they both look compact and spherical. Then, we plot the data to see how the clusters look like by Patel.Length vs Patel.Width. As we can see from Figure 3.3.4, the observations tend to have close numbers of observations for virginica and versicolor, as these two clusters are closer with each other compared to the cluster for setosa.

On the other hand, we also tried the K-means algorithm in this project. Consequently, we have made K-means with 3 clusters and which have corresponding sizes 38, 50, and 62. And the Within cluster sum of squares by cluster for the corresponding three clusters are 23.87947, 15.15100, 39.82097. Furthermore, we have set cluster 1 to be virginica, cluster 2 to be versicolor, and cluster 3 to be setosa. Figure 3.3.5 has shown the plotting results of how the three clusters distributed.

From Figure 3.3.5, we observed that cluster 2 is quite pure, but there is one observation belonging in cluster 1 that was grouped in cluster 2 by mistake. Similarly, there is also one observation belonging to cluster 2 that was grouped in cluster 2, and cluster 1 and cluster 2 are sort of mixed at the top right corner.

We also plotted centres of clusters. In Figure 3.3.6, we set the Sepal length to be the x-axis and Sepal Width to be the y-axis, while we set the Petal length as x-axis and Petal Width as y-axis in Figure 3.3.7s. For both of the plots, we discovered that observations in virginica and versicolor groups tend to have more similar values as these two clusters are close to each other and their centres are also close. In contrast, the setosa group seems to have significantly different lengths and widths for patel as the centre of cluster 2 is far away from the rest clusters in Figure 3.3.3.

4. Conclusion

To Summary, within seven classification models we tried, QDA has the best accuracy. All tree-based algorithms perform as not good as other models. However, all models have accuracy above 90%. And this might because the dataset is simple because we do not need to do any data clean steps. The dataset quality is pretty good, with no missing value, no imbalanced class. Even there is limited outliers, but this does not influence the model accuracy.

We also tried two clustering algorithms to see the structure of iris data. As these two algorithms are not supervised learning algorithms, so we cannot measure the performance of them. By using three different criteria, we have done Hierarchical Cluster Analysis. We see that each dendrogram for different linkage has a lot difference. We have also made K-means with 3 clusters and which have corresponding sizes 38, 50, and 62. We found Virginica and versicolor groups are very similar.

Appendix

Variable names	Description	Data type
Sepal.Length	The length of sepal in cm	numerical
Sepal.Width	The width of sepal in cm	numerical
Patel.Length	The length of patel in cm	numerical
Patel.Width	The width of patel in cm	numerical
Species	The species of iris	categorical

Table 1: Data Description and Data Types

Class number	Variable Name	Number of observations
1	setosa	50
2	versicolor	50
3	virginica	50

Table 2: Data Summary for Species

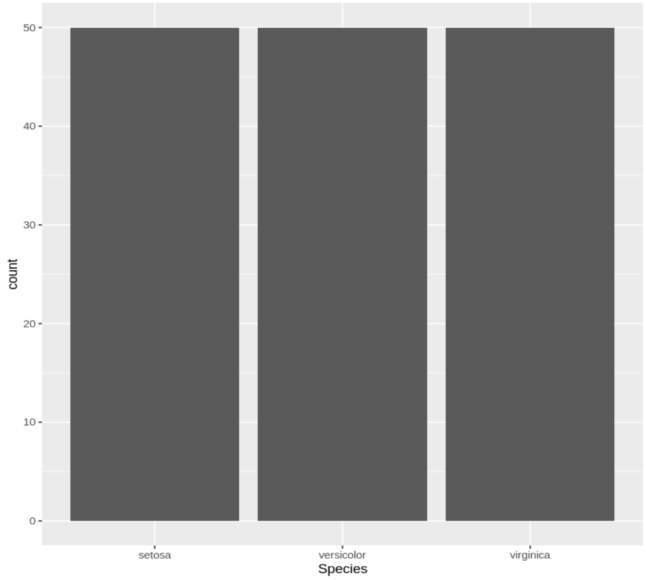


Figure 1.2.1: Number of Observations for Each of Species

Variable name	Min	1st quartile	Median	Mean	3rd quartile	Max
Sepal.Length	4.300	5.100	5.800	5.843	6.400	7.900
Sepal.Width	2.000	2.800	3.000	3.057	3.300	4.400
Patel.Length	1.000	1.600	4.350	3.758	5.100	6.900
Patel.Width	0.100	0.300	1.300	1.199	1.800	2.500

Table 3: Data Summary For Explanatory Variables

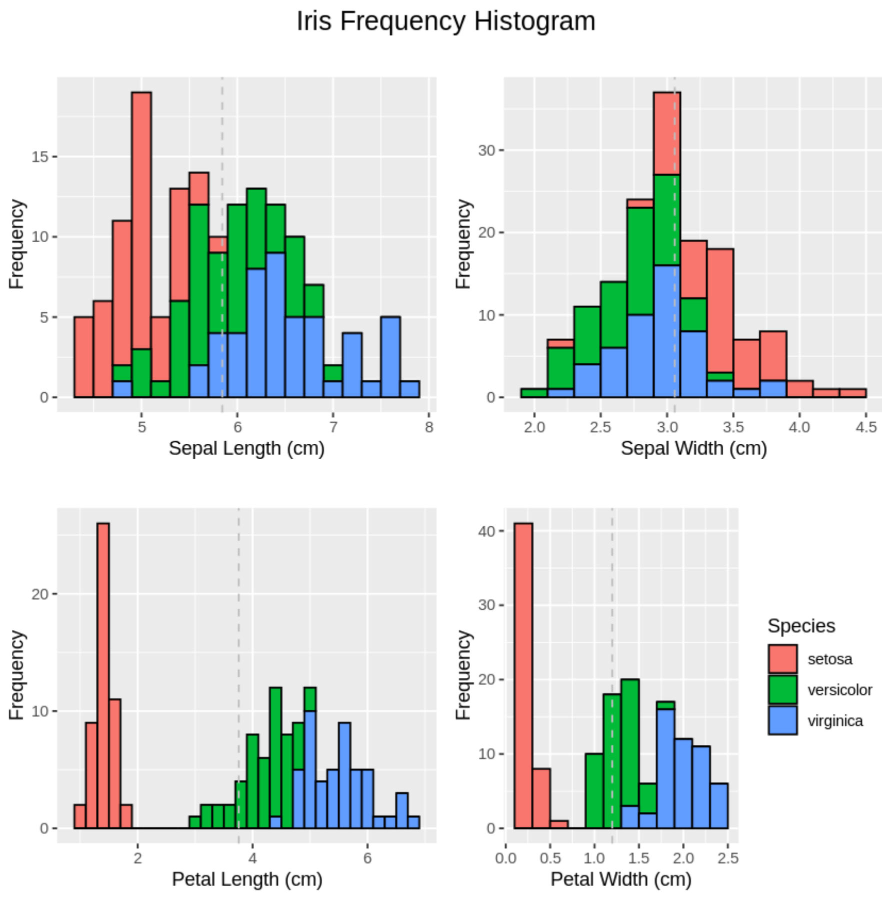


Figure 2.2.1: Iris Frequency Histogram

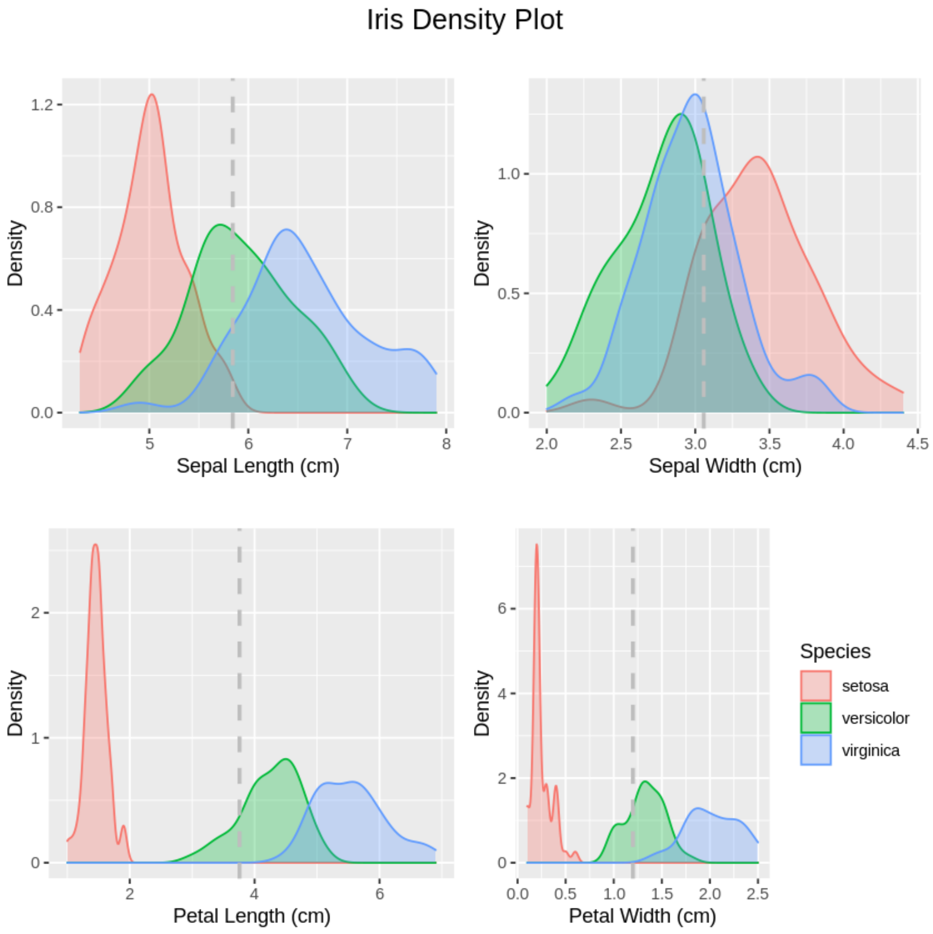


Figure 2.2.2: Iris Density Plot

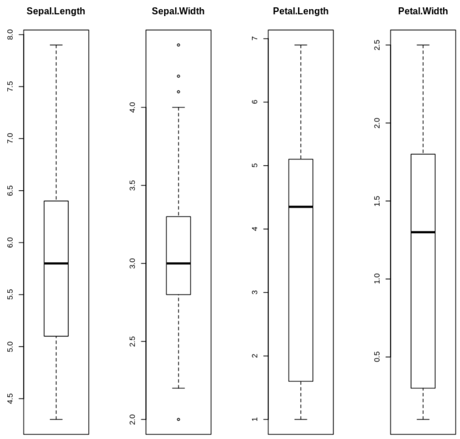


Figure 2.3.1: Box Plot

	Setosa	Versicolor	Virginica
Train Dataset	38	31	36
Test Dataset	12	19	14

Table 4: Species in Train & Test Dataset

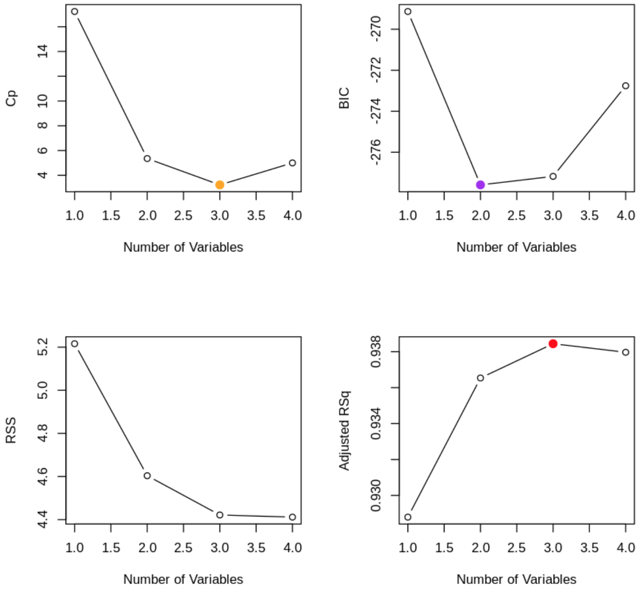
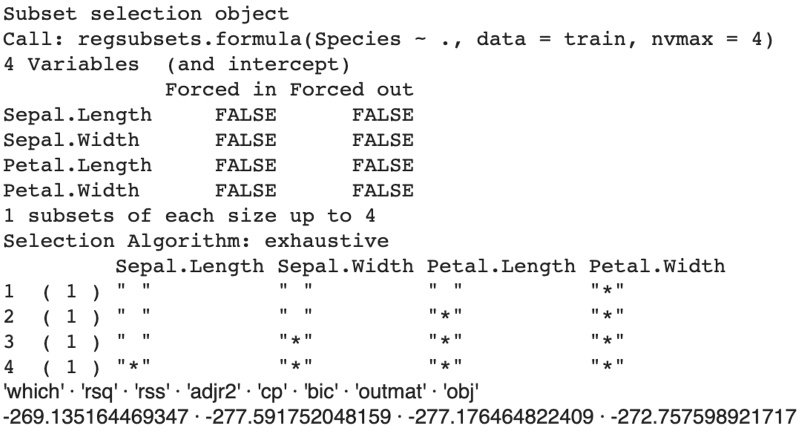
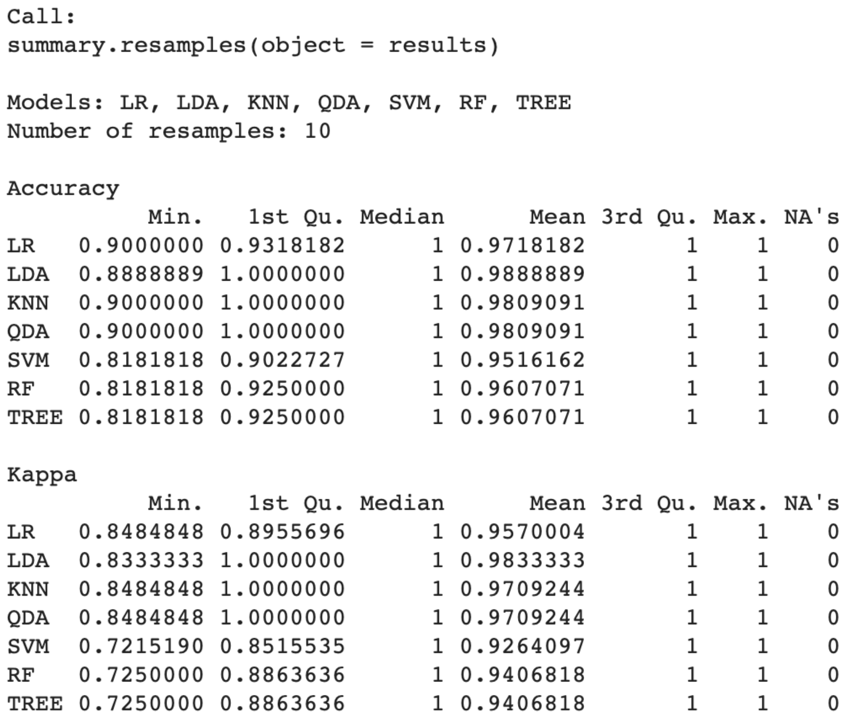


Figure 2.3.2: Plot of Cp, BIC, RSS, Adj Rsq



```
Subset selection object
Call: regsubsets.formula(Species ~ ., data = train, nvmax = 4)
4 Variables  (and intercept)
             Forced in Forced out
Sepal.Length     FALSE      FALSE
Sepal.Width      FALSE      FALSE
Petal.Length     FALSE      FALSE
Petal.Width      FALSE      FALSE
1 subsets of each size up to 4
Selection Algorithm: exhaustive
         Sepal.Length Sepal.Width Petal.Length Petal.Width
1  ( 1 ) " "          " "         " "          "*"
2  ( 1 ) " "          " "         "*"          "*"
3  ( 1 ) " "          "*"         "*"          "*"
4  ( 1 ) "*"          "*"         "*"          "*"
'which' · 'rsq' · 'rss' · 'adjr2' · 'cp' · 'bic' · 'outmat' · 'obj'
-269.135164469347 · -277.591752048159 · -277.176464822409 · -272.757598921717
```

Figure 2.3.3: Subset Selection Object



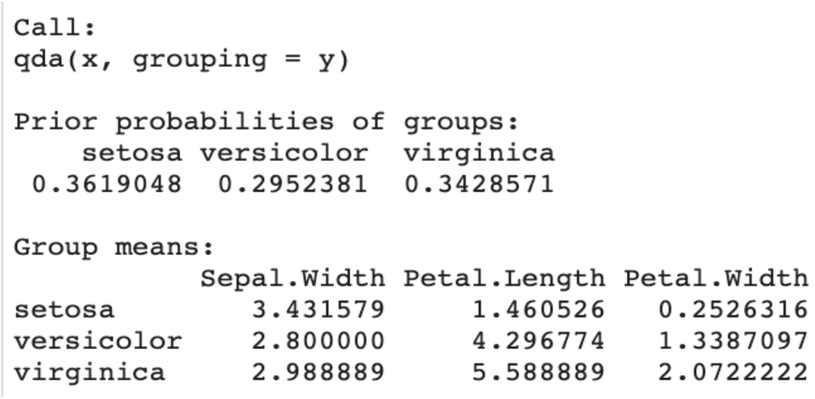
```
Call:
summary.resamples(object = results)

Models: LR, LDA, KNN, QDA, SVM, RF, TREE
Number of resamples: 10

Accuracy
          Min.   1st Qu. Median      Mean 3rd Qu. Max. NA's
LR   0.9000000 0.9318182      1 0.9718182       1    1    0
LDA  0.8888889 1.0000000      1 0.9888889       1    1    0
KNN  0.9000000 1.0000000      1 0.9809091       1    1    0
QDA  0.9000000 1.0000000      1 0.9809091       1    1    0
SVM  0.8181818 0.9022727      1 0.9516162       1    1    0
RF   0.8181818 0.9250000      1 0.9607071       1    1    0
TREE 0.8181818 0.9250000      1 0.9607071       1    1    0

Kappa
          Min.   1st Qu. Median      Mean 3rd Qu. Max. NA's
LR   0.8484848 0.8955696      1 0.9570004       1    1    0
LDA  0.8333333 1.0000000      1 0.9833333       1    1    0
KNN  0.8484848 1.0000000      1 0.9709244       1    1    0
QDA  0.8484848 1.0000000      1 0.9709244       1    1    0
SVM  0.7215190 0.8515535      1 0.9264097       1    1    0
RF   0.7250000 0.8863636      1 0.9406818       1    1    0
TREE 0.7250000 0.8863636      1 0.9406818       1    1    0
```

Figure 3.1.1: Summary of model Accuracy

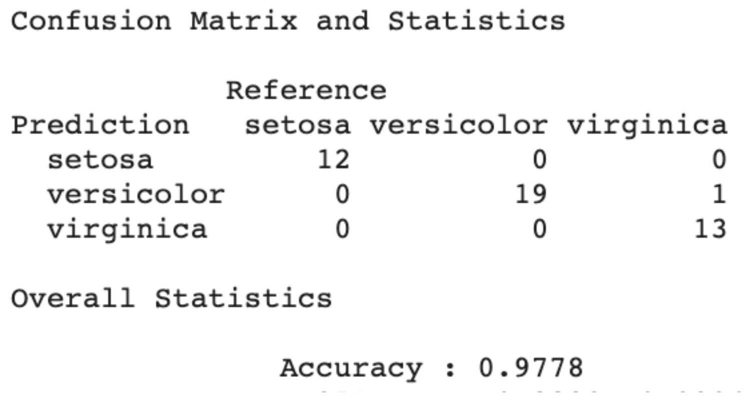


```
Call:
qda(x, grouping = y)

Prior probabilities of groups:
    setosa versicolor  virginica 
 0.3619048  0.2952381  0.3428571 

Group means:
           Sepal.Width Petal.Length Petal.Width
setosa        3.431579     1.460526   0.2526316
versicolor    2.800000     4.296774   1.3387097
virginica     2.988889     5.588889   2.0722222
```

Figure 3.1.1: Optimal model of QDA



```
Confusion Matrix and Statistics

            Reference
Prediction   setosa versicolor virginica
  setosa         12          0         0
  versicolor      0         19         1
  virginica       0          0        13

Overall Statistics

               Accuracy : 0.9778
```

Figure 3.1.1.2: Output of QDA

	Logistic Regression	LDA	QDA	KNN	SVM	Tree	Random Forest
Accuracy	0.9333	0.9333	0.9778	0.9556	0.9556	0.9111	0.9333
Kappa	0.8976	0.8969	0.9659	0.9315	0.932	0.863	0.8976

Table 5: Accuracy & Kappa of Final optimal model

Prediction	setosa	versicolor	virginica
setosa	12	0	0
versicolor	0	19	2
virginica	0	0	12

Table 6: Confusion Matrix of Neural Network

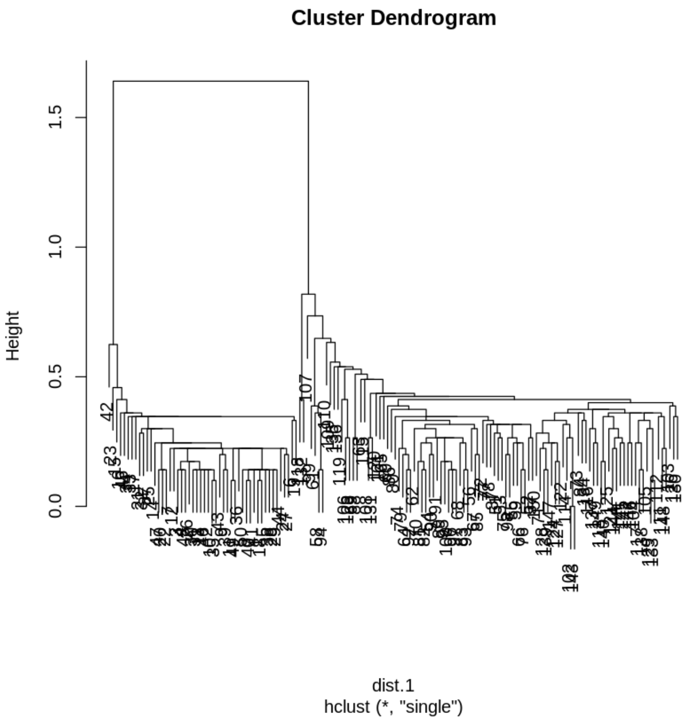


Figure 3.3.1: Cluster Dendrogram with Single Linkage

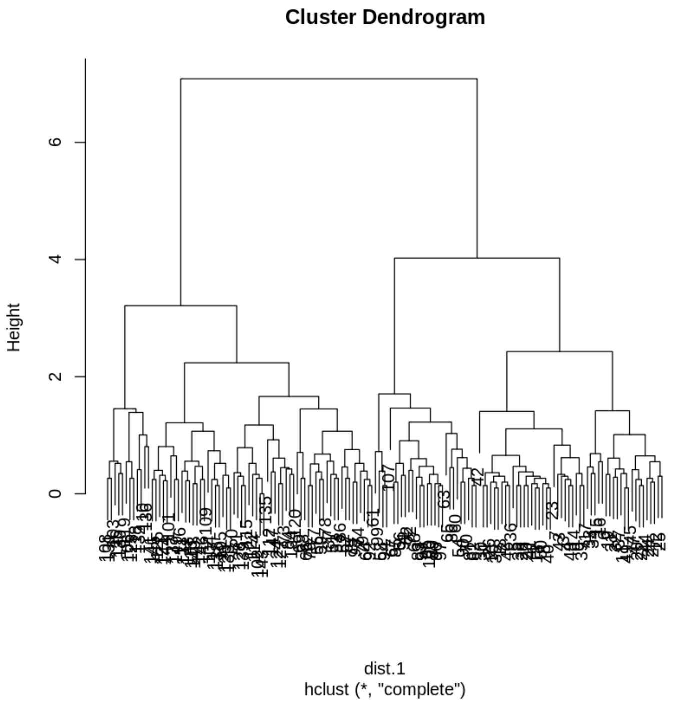


Figure 3.3.2:: Cluster Dendrogram with Complete Linkage

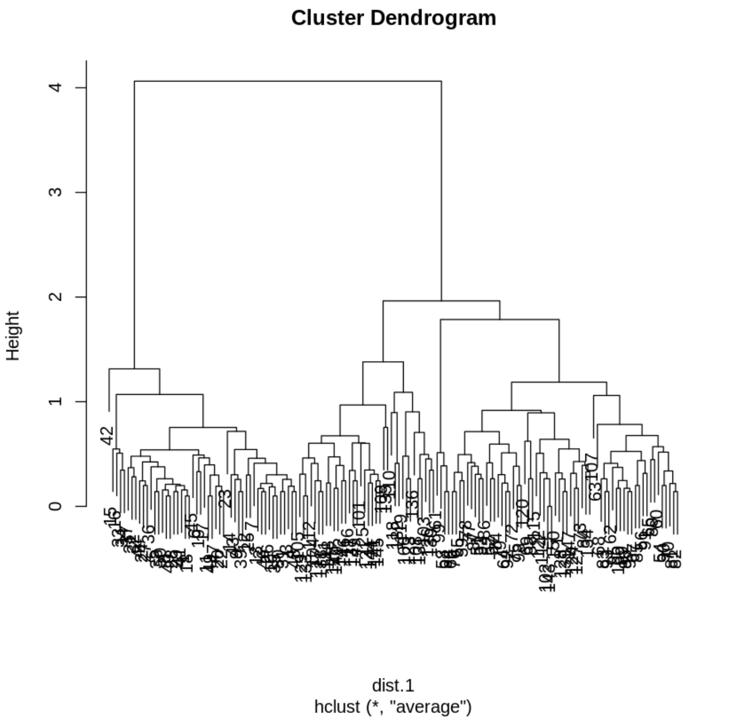


Figure 3.3.3:: Cluster Dendrogram with Average Linkage

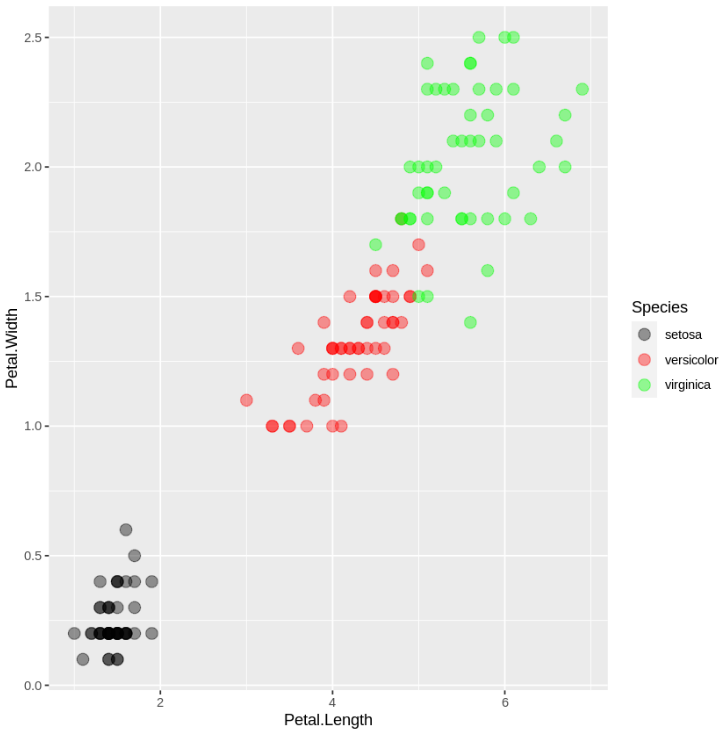


Figure 3.3.4: Hierarchical Cluster Analysis Plot (Patel.Length vs Patel.Width)

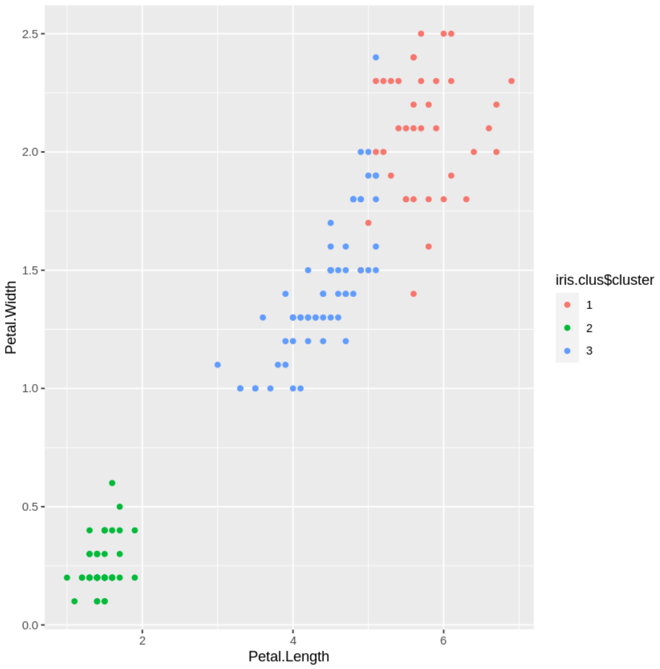


Figure 3.3.5: the K-mean Plot of Iris Data

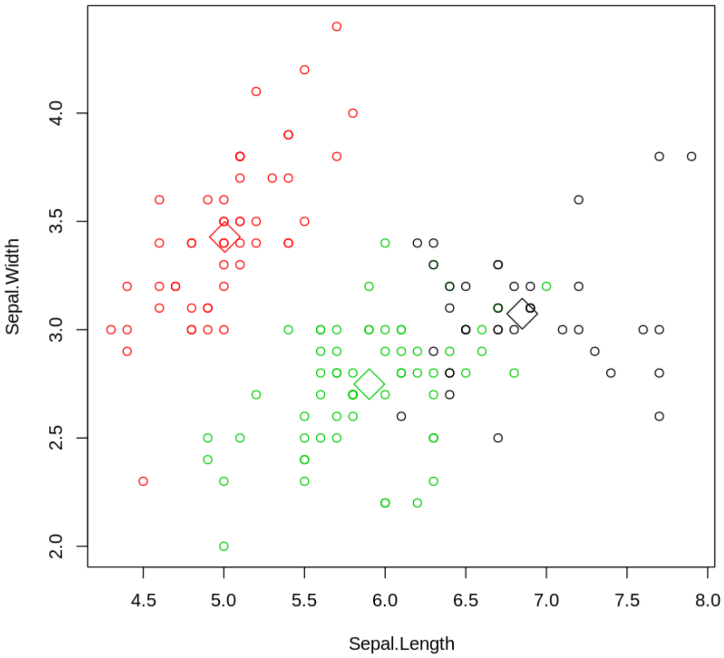


Figure 3.3.6: K-Means Centre Sepal.Length vs Sepal.Width

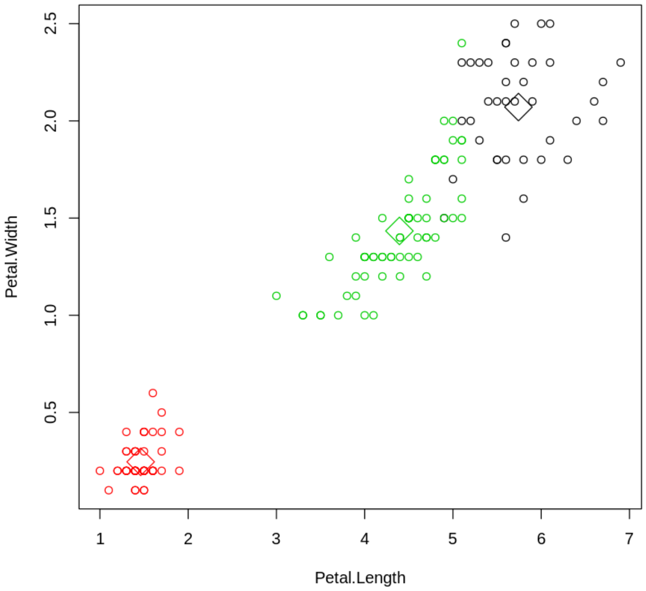


Figure 3.3.7: K-Means Centre Petal.Length vs Petal.Width