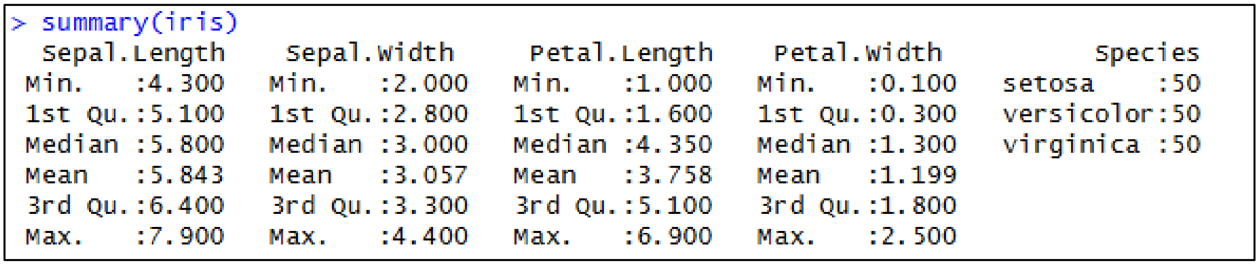
K-NEAREST NEIGHBOURS

Q1: Data Exploration

The data set is the built-in iris data set in R, with 150 observations of iris flowers from three species: setosa, versicolor, and virginica. There are four numerical features in each observation: Sepal.Length, Sepal.Width, Petal.Length, and Petal.Width.

Summary Statistics:

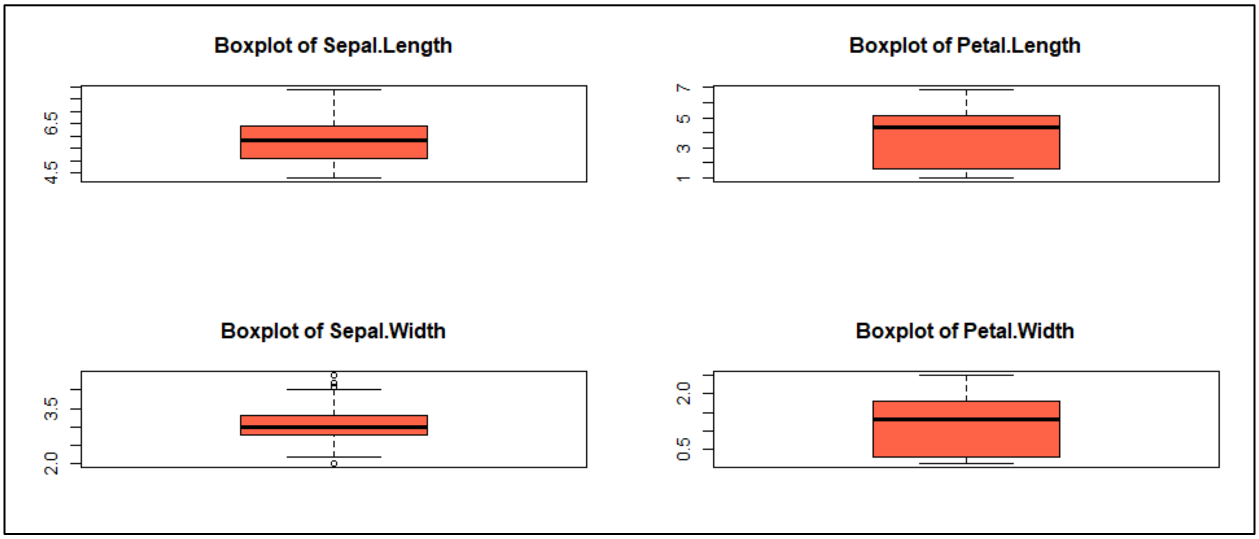
- No missing values were found (colSums(is.na(iris)) resulted in all zeros).
- The summary statistics show Sepal.Length ranges from 4.3 to 7.9, and Petal.Length from 1.0 to 6.9, quite a large range in the flower sizes.
- Balanced species: 50 observations in each class.



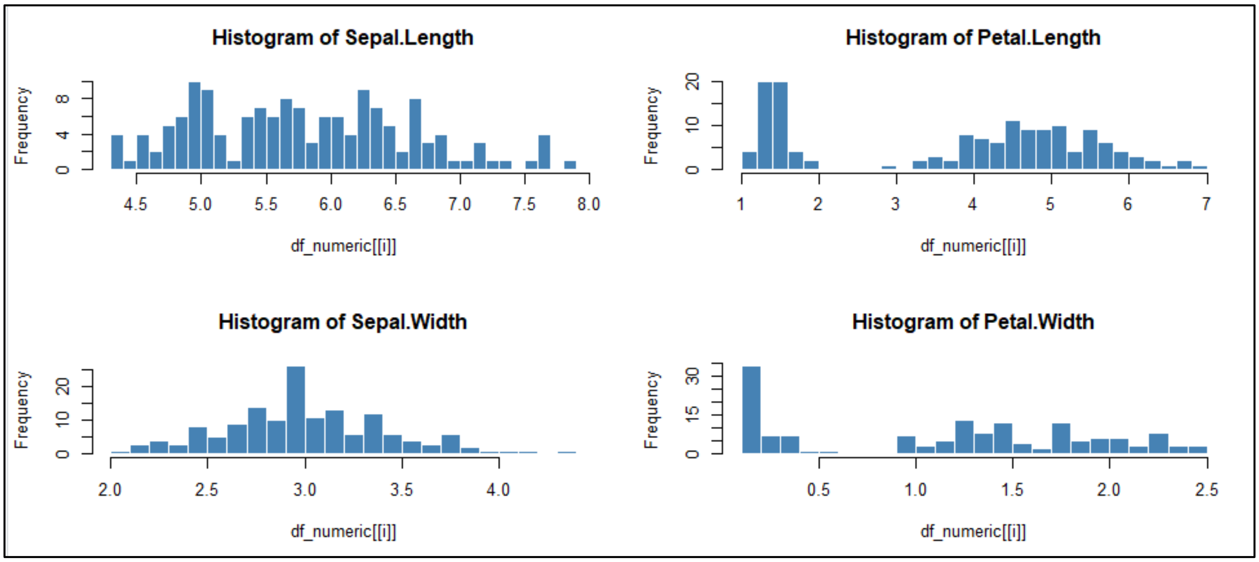
```
> summary(iris)
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width          Species  
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   setosa    :50  
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   versicolor:50  
 Median :5.800   Median :3.000   Median :4.350   Median :1.300   virginica :50  
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199                  
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800                  
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500                  
```

Distribution Insights:

- Boxplots suggest potential outliers in Sepal.Width and Petal.Width, with a greater spread in Sepal.Width.



- Histograms suggest that Petal.Length and Petal.Width are right-skewed, while Sepal.Length and Sepal.Width are approximately symmetric.

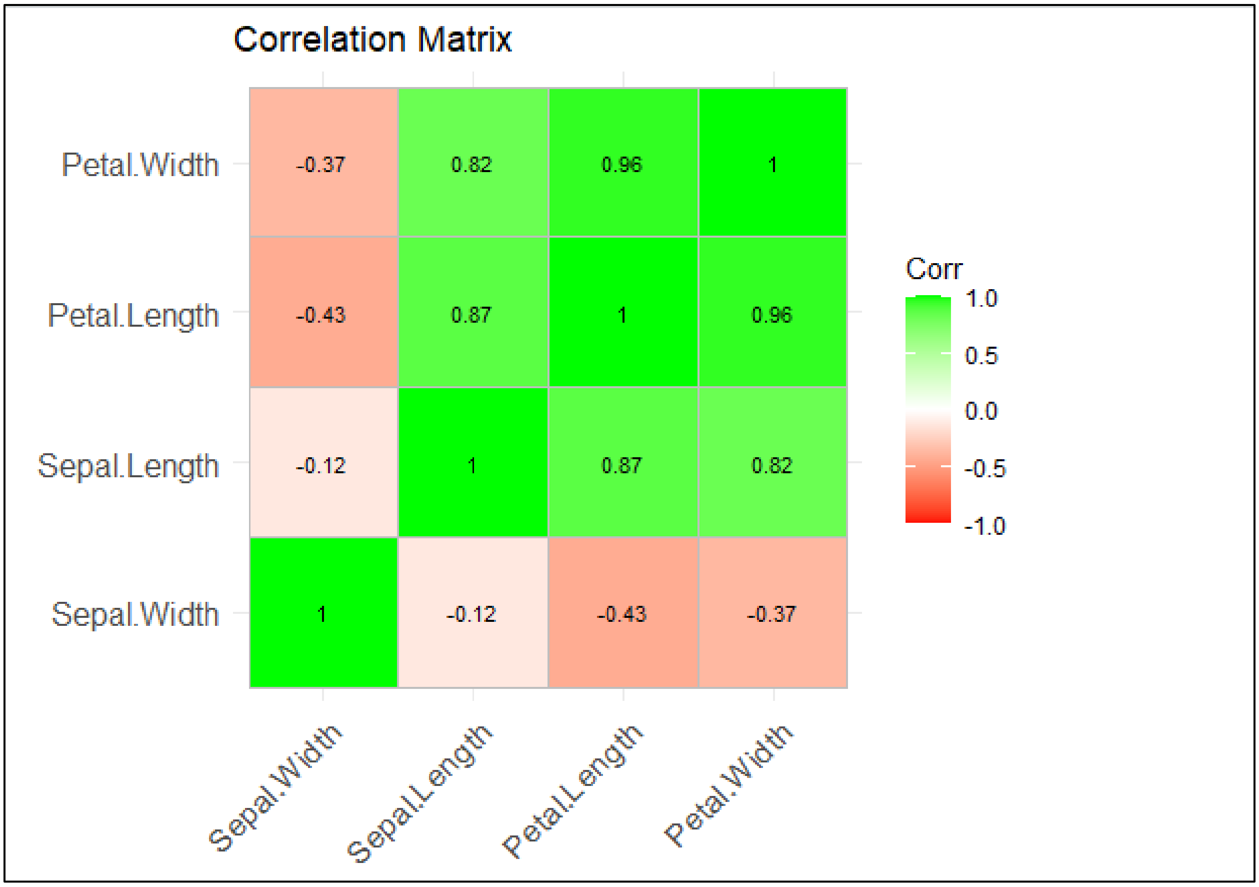


Correlation Analysis:

- Petal.Length and Petal.Width have a very strong positive correlation (0.96).

- Sepal.Length also correlates highly with Petal.Length (0.87).

- Sepal.Width has weak or negative correlations with the other features, especially Petal.Length (-0.43) and Petal.Width (-0.37).



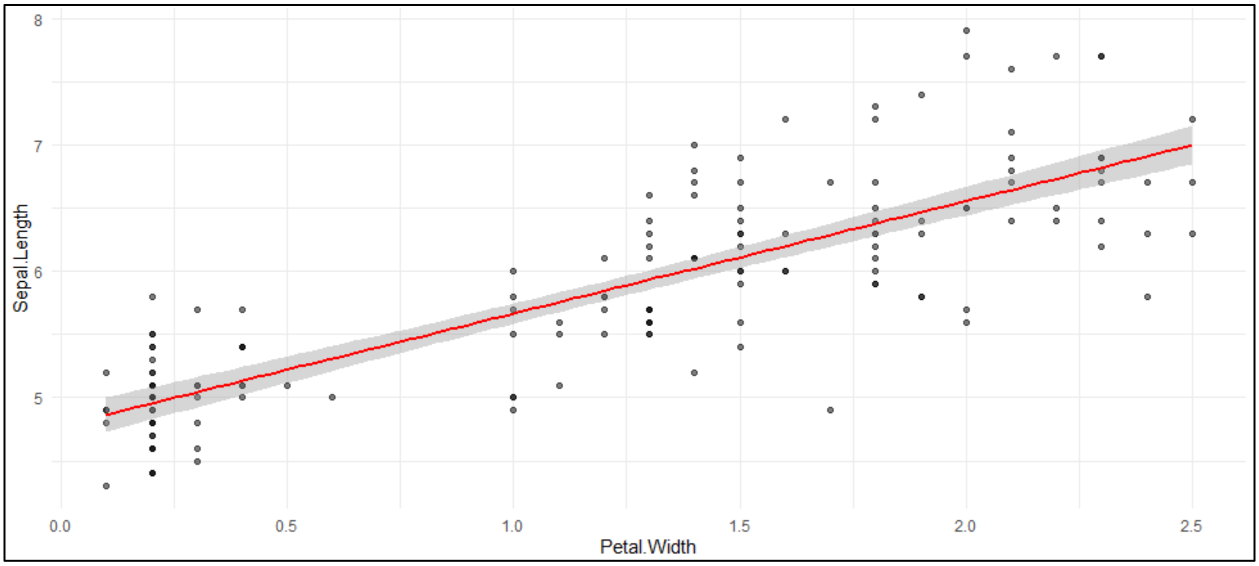
Scatterplots:

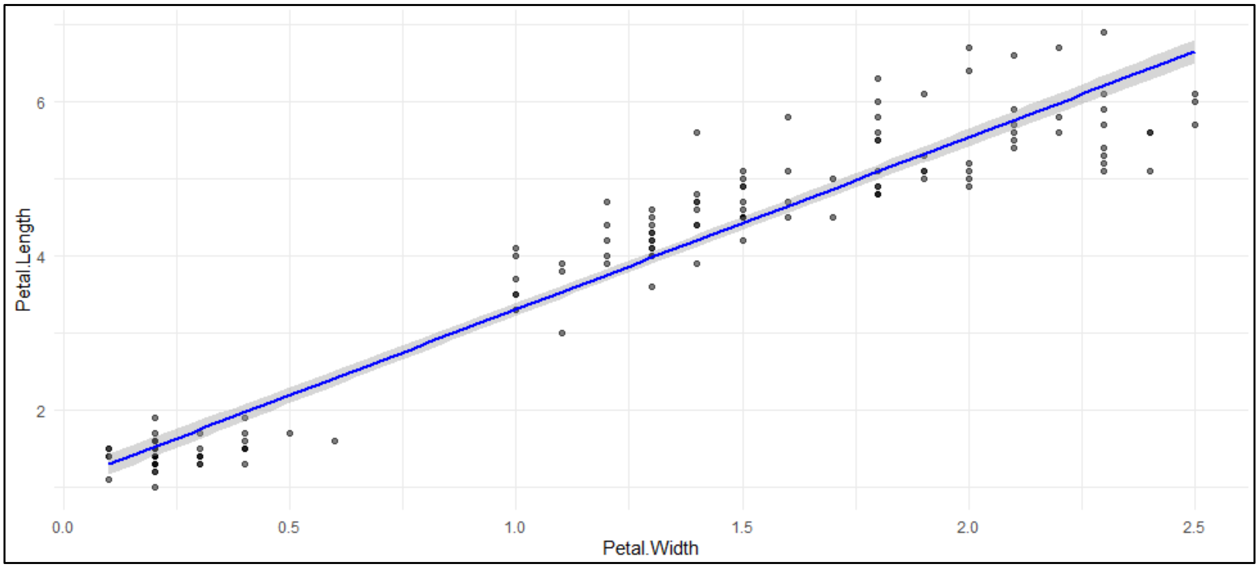
- Petal.Width vs Sepal.Length shows a positive linear trend.

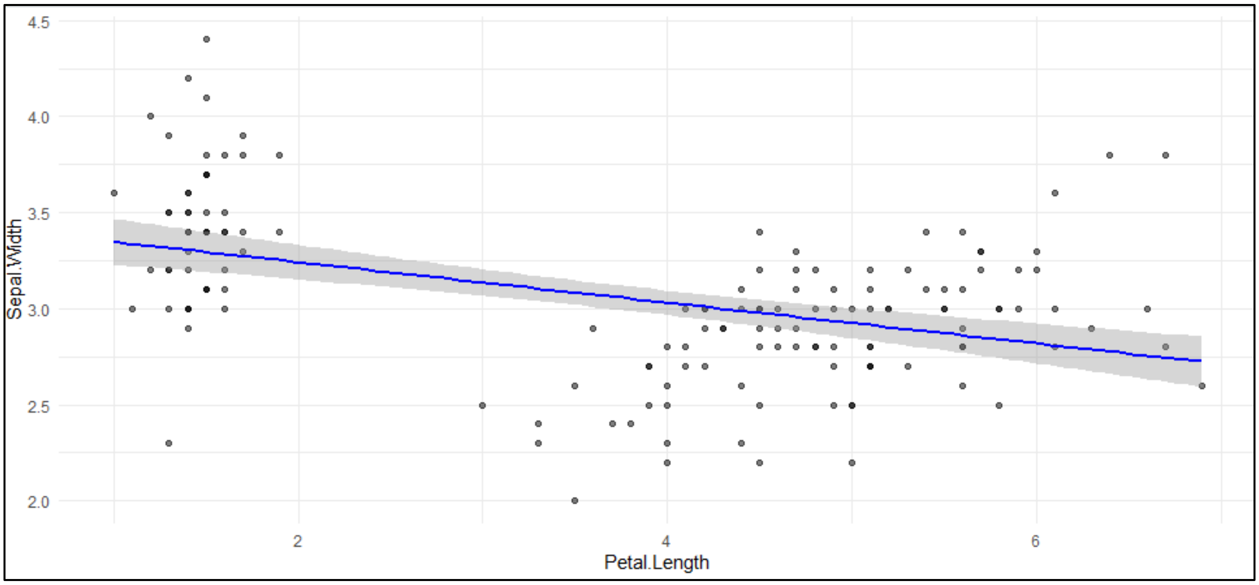
- Petal.Width vs Petal.Length has a very strong positive relationship.

- Petal.Length vs Sepal.Width shows a negative trend.

These relationships suggest Petal features are more informative to distinguish species, and Sepal.Width may have less importance in separation.


8
7
6
5
Sepal.Length
0.0
0.5
1.0
1.5
2.0
2.5
Petal.Width


6
4
2
Petal.Length
0.0
0.5
1.0
1.5
2.0
2.5
Petal.Width



Q2: Building the Model

The data was divided using set.seed(16) to make it reproducible. A 70/30 split was then established using createDataPartition(), with 105 observations for training and 45 for testing.

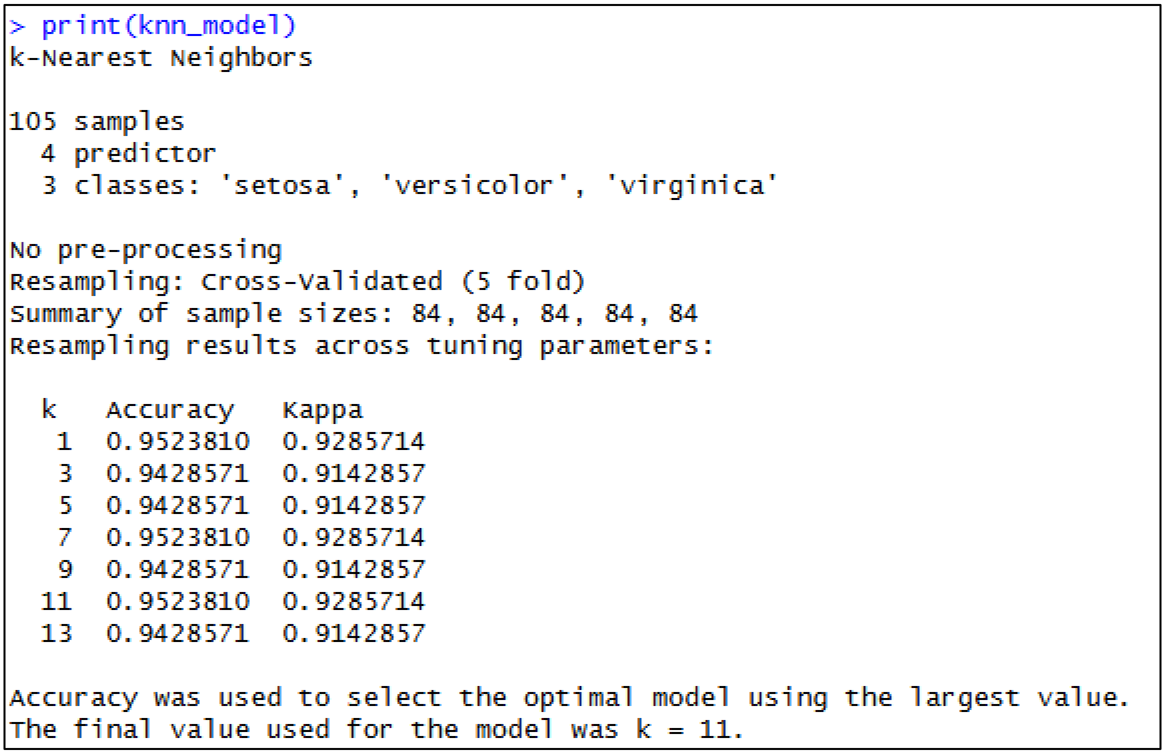
The four numeric variables were then scaled using preProcess() with center and scale parameters, which is required for KNN since it is based on Euclidean distances.

Q3: Finding the Best K

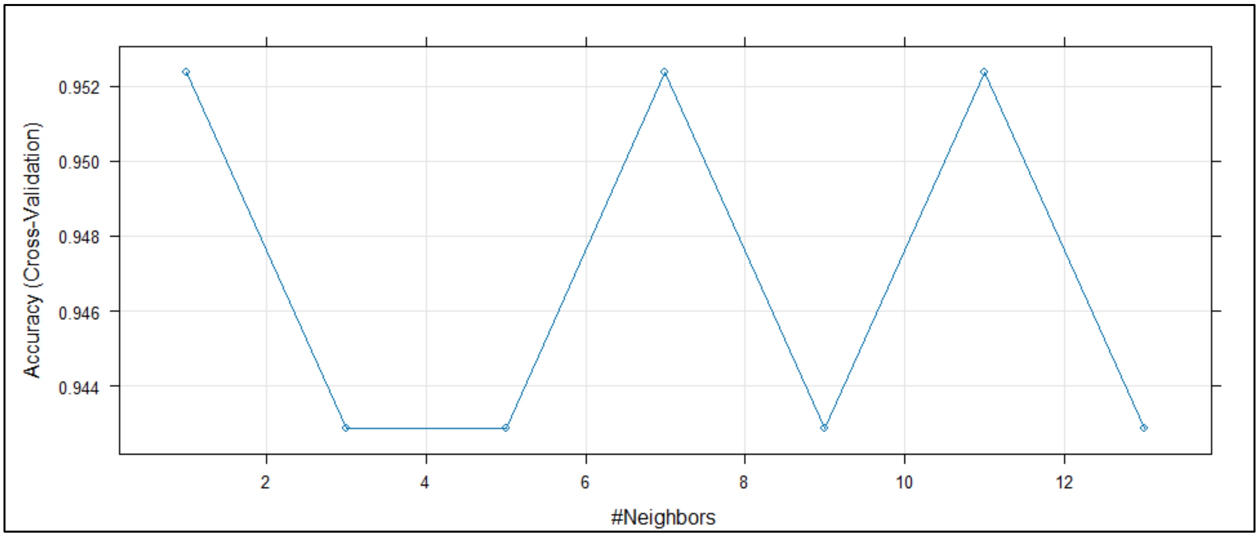
A KNN model was trained with caret::train() for a grid of K values: 1, 3, 5, 7, 9, 11, and 13. Cross-validation (5-fold) was used to measure performance.

Accuracy vs K Plot:

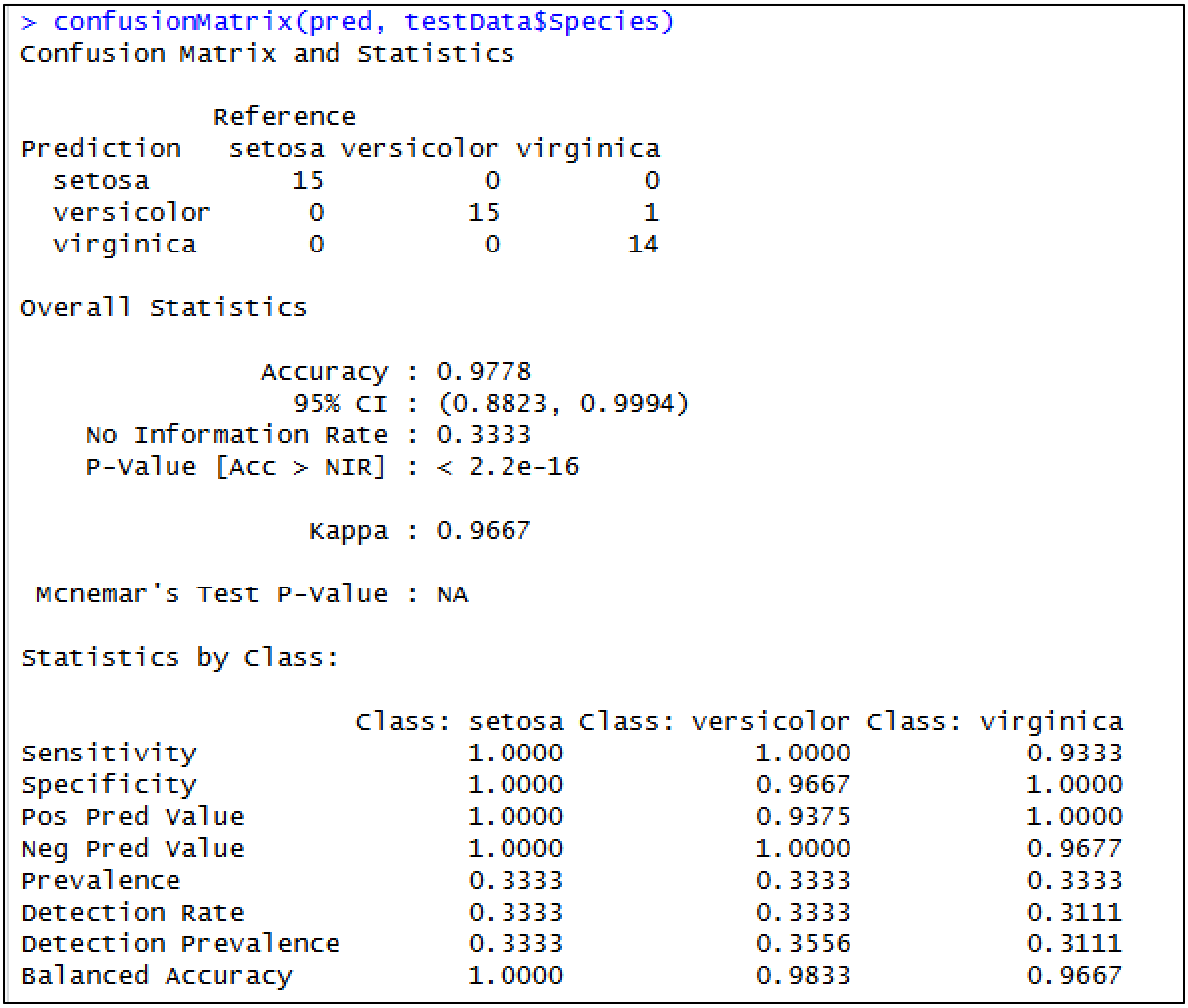
The plot shows that accuracy is the highest with K = 1, 7, and 11, all of 95.24%. K = 11 was selected as the best model based on highest accuracy.



```
> print(knn_model)
k-Nearest Neighbors

105 samples
  4 predictor
  3 classes: 'setosa', 'versicolor', 'virginica'

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 84, 84, 84, 84, 84
Resampling results across tuning parameters:

  k   Accuracy   Kappa
   1  0.9523810  0.9285714
   3  0.9428571  0.9142857
   5  0.9428571  0.9142857
   7  0.9523810  0.9285714
   9  0.9428571  0.9142857
  11  0.9523810  0.9285714
  13  0.9428571  0.9142857

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was k = 11.
```



Q4: Model Interpretation



```
> confusionMatrix(pred, testData$Species)
Confusion Matrix and Statistics

            Reference
Prediction   setosa versicolor virginica
  setosa         15          0         0
  versicolor      0         15         1
  virginica       0          0        14

Overall Statistics

               Accuracy : 0.9778
                 95% CI : (0.8823, 0.9994)
    No Information Rate : 0.3333
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9667

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: setosa Class: versicolor Class: virginica
Sensitivity                 1.0000            1.0000           0.9333
Specificity                 1.0000            0.9667           1.0000
Pos Pred Value              1.0000            0.9375           1.0000
Neg Pred Value              1.0000            1.0000           0.9677
Prevalence                  0.3333            0.3333           0.3333
Detection Rate              0.3333            0.3333           0.3111
Detection Prevalence        0.3333            0.3556           0.3111
Balanced Accuracy           1.0000            0.9833           0.9667
```

With K = 11, predictions were performed on the test set. The confusion matrix is:

Interpretation:

- Accuracy: 97.78% — extremely high, i.e., very good predictive performance.

- No Information Rate (NIR): 33.33% — the accuracy is significantly higher than this baseline.

- P-value [Acc > NIR]: < 2.2e-16 — statistically significant better than random guessing.

- Kappa: 0.9667 — perfect above-chance agreement.

- McNemar's Test: Not applicable here because of perfect classification in certain classes.

Class-wise Metrics:

- Sensitivity: 100% for setosa and versicolor, 93.33% for virginica.
- Specificity: 100% for setosa and virginica, 96.67% for versicolor.
- Positive Predictive Value: All above 93%, setosa and virginica at 100%.
- Balanced Accuracy: All above 96%, setosa at 100%.

In general, the model is outstanding in performance for all classes with minimal misclassification. The lone error was a virginica classified as versicolor, as might be expected with their identical feature space.

Q5: Will produce a Scatterplot of Sepal Features

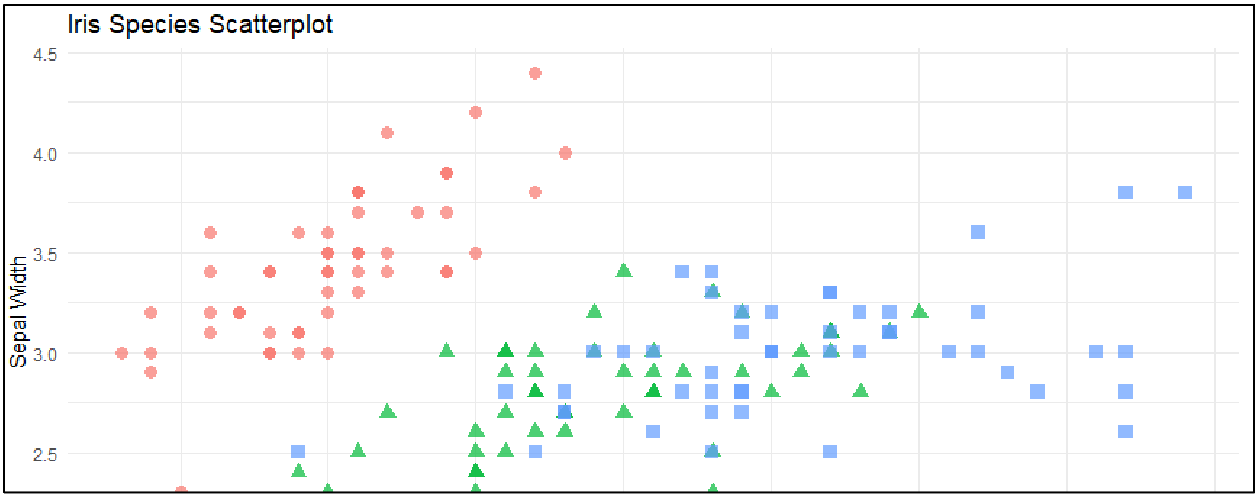
The Sepal.Length vs Sepal.Width scatterplot shows clear clustering:

•Setosa is tightly clustered with shorter Sepal.Length and wider Sepal.Width.

•Virginica has longer Sepal.Length and narrower Sepal.Width.

•Versicolor is mid-range, with moderate values.

This plot shows that using only Sepal features gives some separability, but not as strong as Petal features.



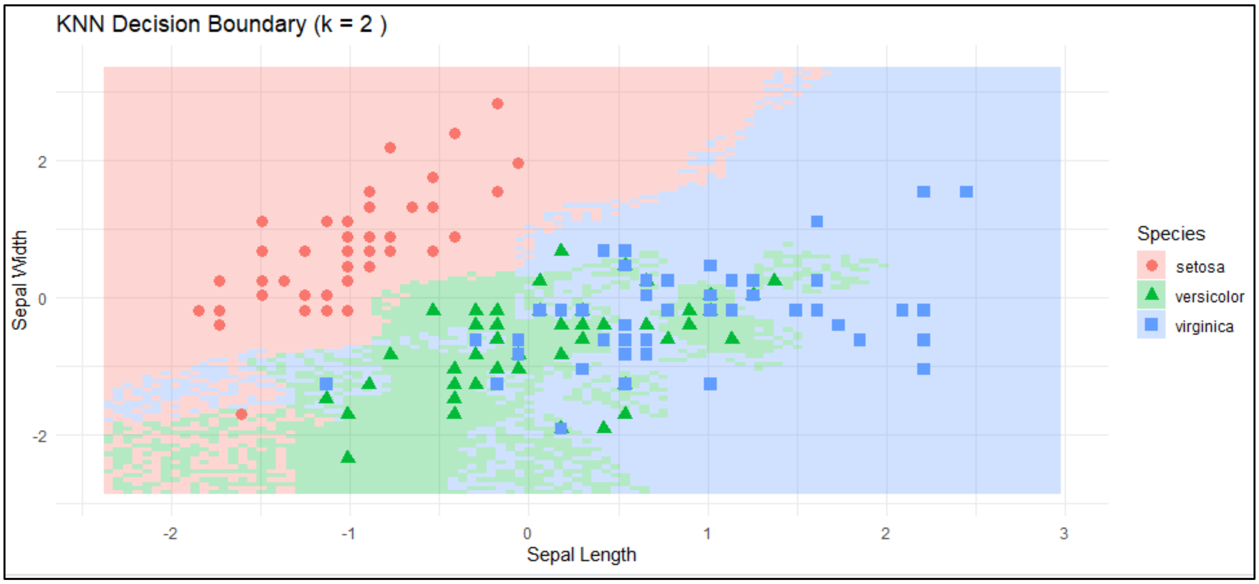
Q6: Decision Boundary Analysis

What is a decision boundary?

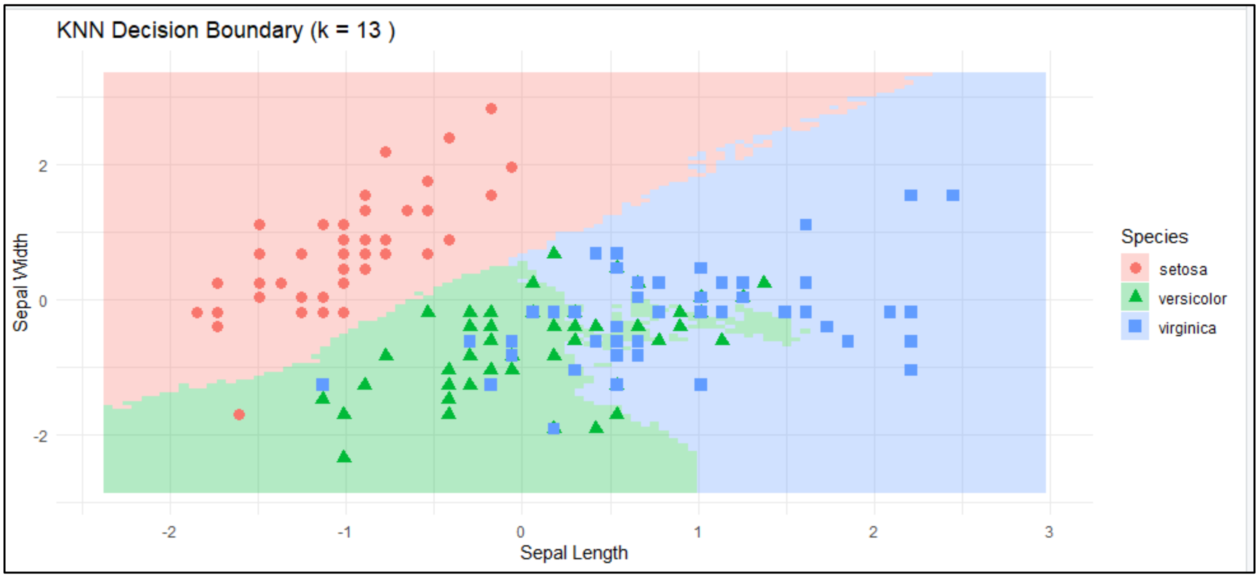
In KNN, a decision boundary is a region in feature space where the algorithm switches from predicting one class to predicting another. It's characterized by proximity of training points and the value of K.

Plots for K = 2 and K = 13:

- K = 2: The decision boundaries are discontinuous and extremely sensitive to local oscillations. This is characteristic of overfitting — the model's sensitive to close points.



- K = 13: Generalized and smoother boundaries. The model considers a broader neighborhood, hence not as sensitive to noise.



Why are KNN boundaries hard to interpret?

Because the fact that KNN is non-parametric and instance-based means the boundaries are entirely defined by the distribution of the training set. They are non-linear and irregualr and become particularly so with small values of K, so harder to explain than equation-based models.