# Iris Dataset Analysis Report

*Manuel Alexis Mena Nieves*

*9/19/2019*

## 1. Introduction

This is a report based on the Iris Dataset which is a multivariate data set introduced by the british statistician and biologist Ronal Fisher in his 1936 paper "The use of multiple measurements in taxonomic problems". The dataset consists of 50 samples from each of three species of Iris (Iris setosa, Iris versicolor and Iris virginica). Four features were measured from each sample: the length and the width of the sepals and petals, in centimeters.

## 2. Understanting the data

The first step in our analysis is to understand the data. In this case, we want to know how many columns and rows has the dataset.

```
dim(iris)
```

```
## [1] 150   5
```

Then we want to see the structure of the data. We see that we have four numeric features, in this case the lengths and widths of Sepal and Petal, and one factor which comprises the three species value.

```
str(iris)
```

```
## 'data.frame':    150 obs. of  5 variables:
##  $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
##  $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
##  $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
##  $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
##  $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...
```

The range of the features is between 0.1 and 7.9 centimeters. The mean of the features is between 1.199 and 5.843 centimeters. We don't see NA values, son we can consider the data is tidy. In our case, we want to understand the relations between the features and the species.

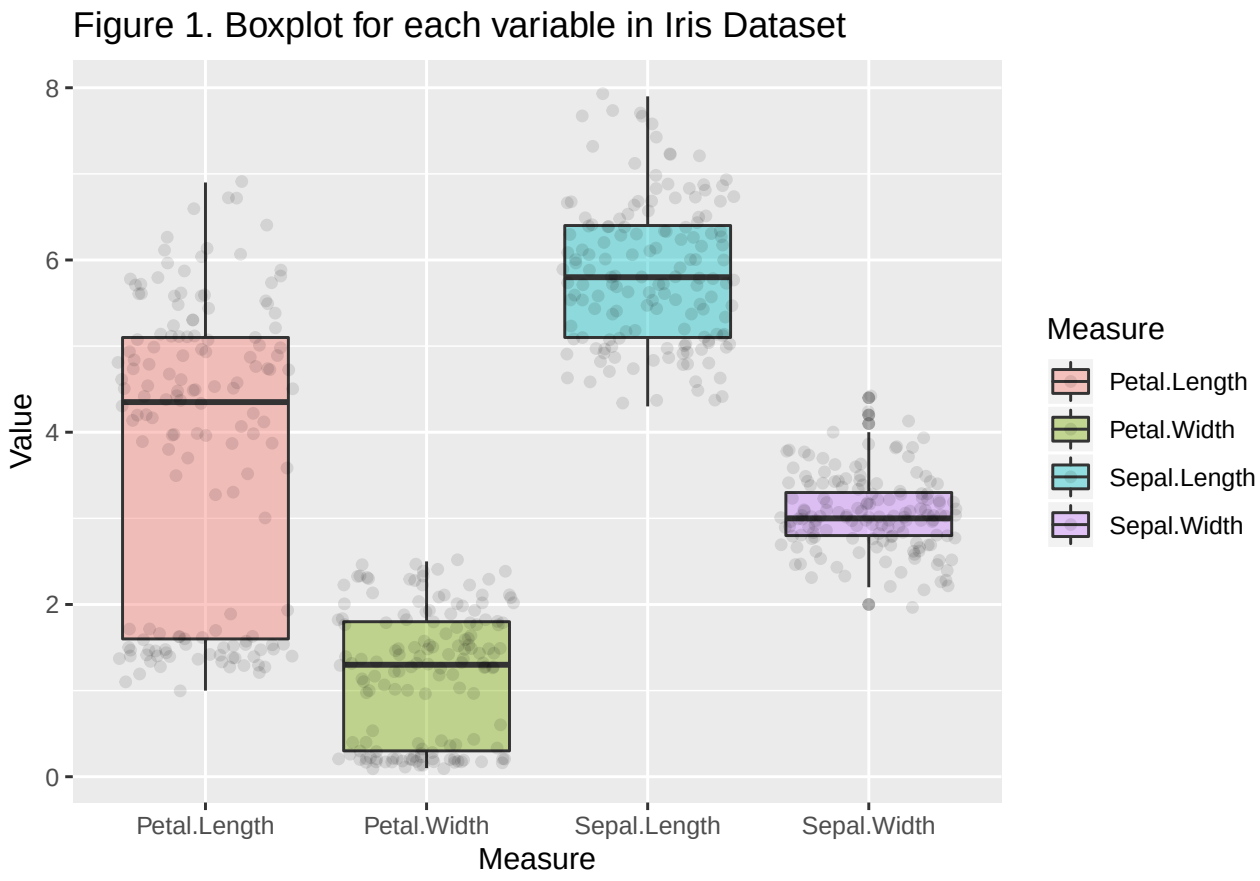
```
summary(iris)
```

```
##   Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
##  Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
##  1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
##  Median :5.800   Median :3.000   Median :4.350   Median :1.300
##  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
##        Species
##  setosa    :50
##  versicolor:50
##  virginica :50
##
##
##
```

## 3. Exploratory Analysis

To begin with the exploratory analysis, we make a boxplot for each of the four measures. We can observe that sepal length and width distributions are almost centered, but petal length and width distributions are not. The sepal lengths have the highest median value and the petal widths the lowest median value.

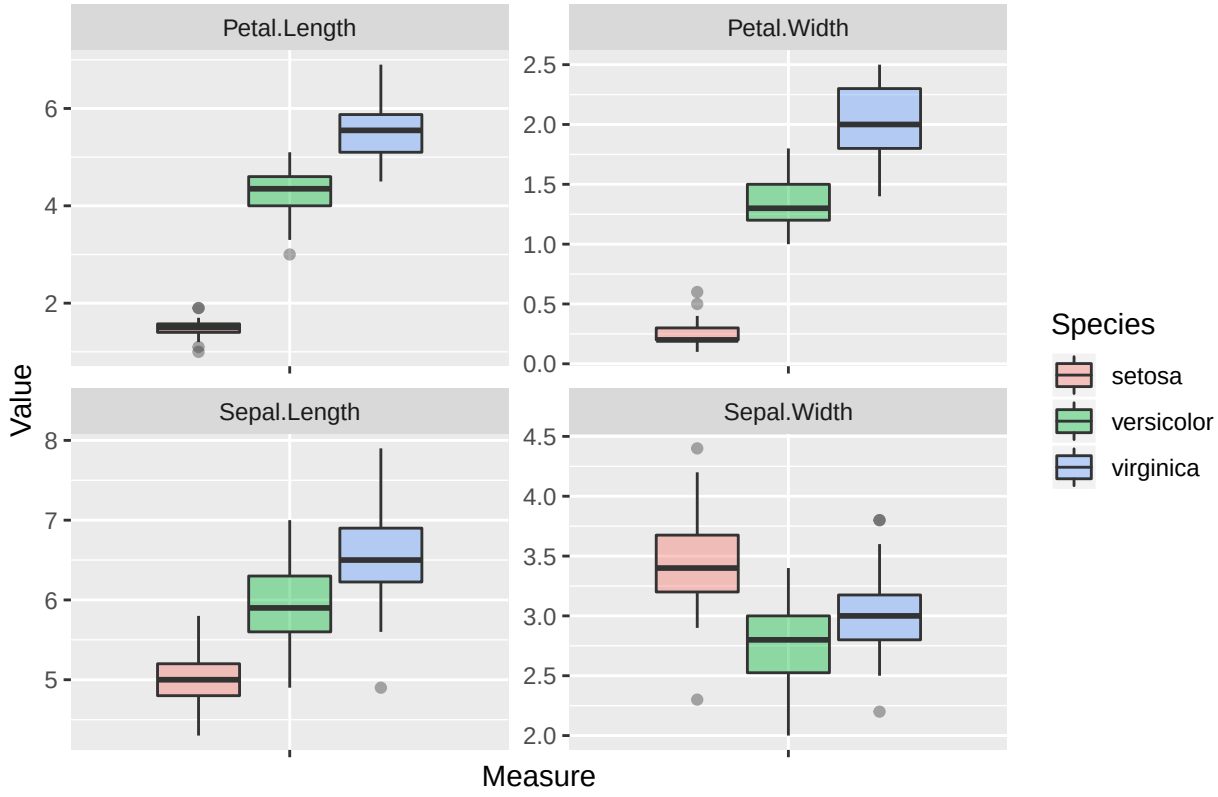
```
iris %>% gather(Measure,Value, -Species) %>%
  ggplot(aes(x = Measure, y = Value, fill = Measure)) +
  geom_boxplot(alpha = 0.4) + geom_jitter(alpha = 0.1) +
  labs(title = "Figure 1. Boxplot for each variable in Iris Dataset")
```




We can look at the distribution of each measure stratified by species and see that petal length separates setosa from versicolor and virginica. Petal width also separates setosa from versicolor and virginica. This implies that the measures related to petal have some importance in the species.

```
iris %>% gather(Measure, Value, -Species) %>%
  ggplot(aes(x = Measure, y = Value, fill = Species)) +
  geom_boxplot(alpha = 0.4) + facet_wrap(~ Measure, scales = "free") +
  theme(axis.text.x = element_blank()) +
  labs(title = "Figure 2. Stratified boxplot in Iris Dataset")
```

Figure 2. Stratified boxplot in Iris Dataset




By plotting petal length versus petal width we can see that the specie Setosa is clearly separated, and versicolor can be separated from virginica but they both have a little space in common. Even in that situation it can be used to define an algorithm with a good accuracy.

```
iris %>% ggplot(aes(x = Petal.Length, y = Petal.Width, colour = Species)) +
  geom_point(aes(shape = Species), cex = 2) +
  labs(title="Figure 3. Petal length vs. petal width plot")
```

Figure 3. Petal length vs. petal width plot


## 4. Modelling the data

We define two set for the data modelling, in this case one for training which represents the 70% if the data, and the other set for testing our models. The modelling uses the caret package, and we set a cross validation control for the algorithms we will use.

```
library(caret)
set.seed(1979)
index <- createDataPartition(iris$Species, times = 1, p = 0.3, list = FALSE )
train_set <- iris[-index,]
test_set <- iris[index,]
fitControl <- trainControl(method = "repeatedcv", number = 10, repeats = 5)
```

The first algorithm we try is the decision tree, since the exploratory analysis suggest we can obtain a decision rule from petal measurements. We see a

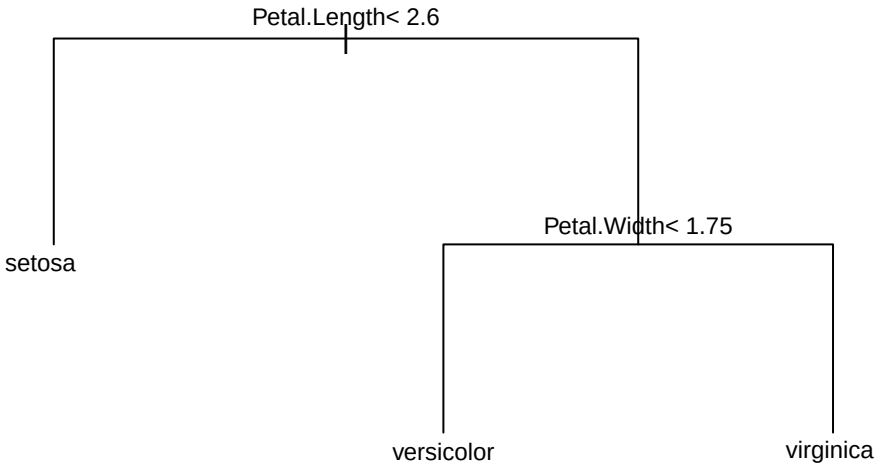
```
train_rpart <- train(Species ~ ., data = train_set, method = "rpart", trControl = fitControl)
y_hat_tree <- predict(train_rpart,test_set)
confusionMatrix(y_hat_tree, test_set$Species)
```

```
## Confusion Matrix and Statistics
##
##              Reference
## Prediction   setosa versicolor virginica
##   setosa         15          0         0
##   versicolor      0         14         2
##   virginica       0          1        13
##
```

```
## Overall Statistics
##
##                Accuracy : 0.9333
##                  95% CI : (0.8173, 0.986)
##     No Information Rate : 0.3333
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.9
##
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: setosa Class: versicolor Class: virginica
## Sensitivity                 1.0000            0.9333           0.8667
## Specificity                 1.0000            0.9333           0.9667
## Pos Pred Value              1.0000            0.8750           0.9286
## Neg Pred Value              1.0000            0.9655           0.9355
## Prevalence                  0.3333            0.3333           0.3333
## Detection Rate              0.3333            0.3111           0.2889
## Detection Prevalence        0.3333            0.3556           0.3111
## Balanced Accuracy           1.0000            0.9333           0.9167
```

```
plot(train_rpart$finalModel, margin = 0.1, main = "Figure 4. Decision tree for Iris Dataset")
text(train_rpart$finalModel, cex = 0.75)
```

**Figure 4. Decision tree for Iris Dataset**

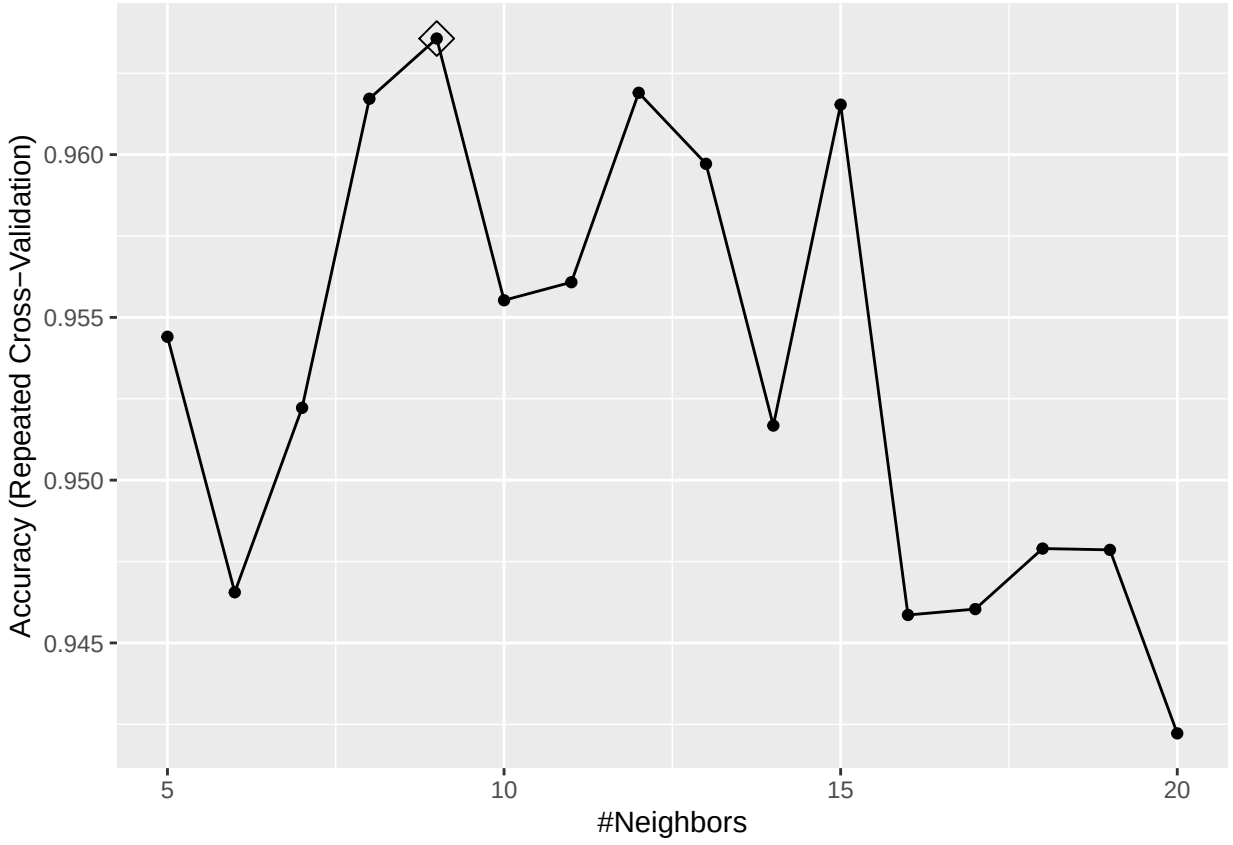



```
train_knn <- train(Species ~., data = train_set, method = "knn", trControl = fitControl,
                   tuneGrid = data.frame(k = seq(5,20,1)))
y_hat_knn <- predict(train_knn, test_set)
confusionMatrix(y_hat_knn, test_set$Species)
```

```
## 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
```

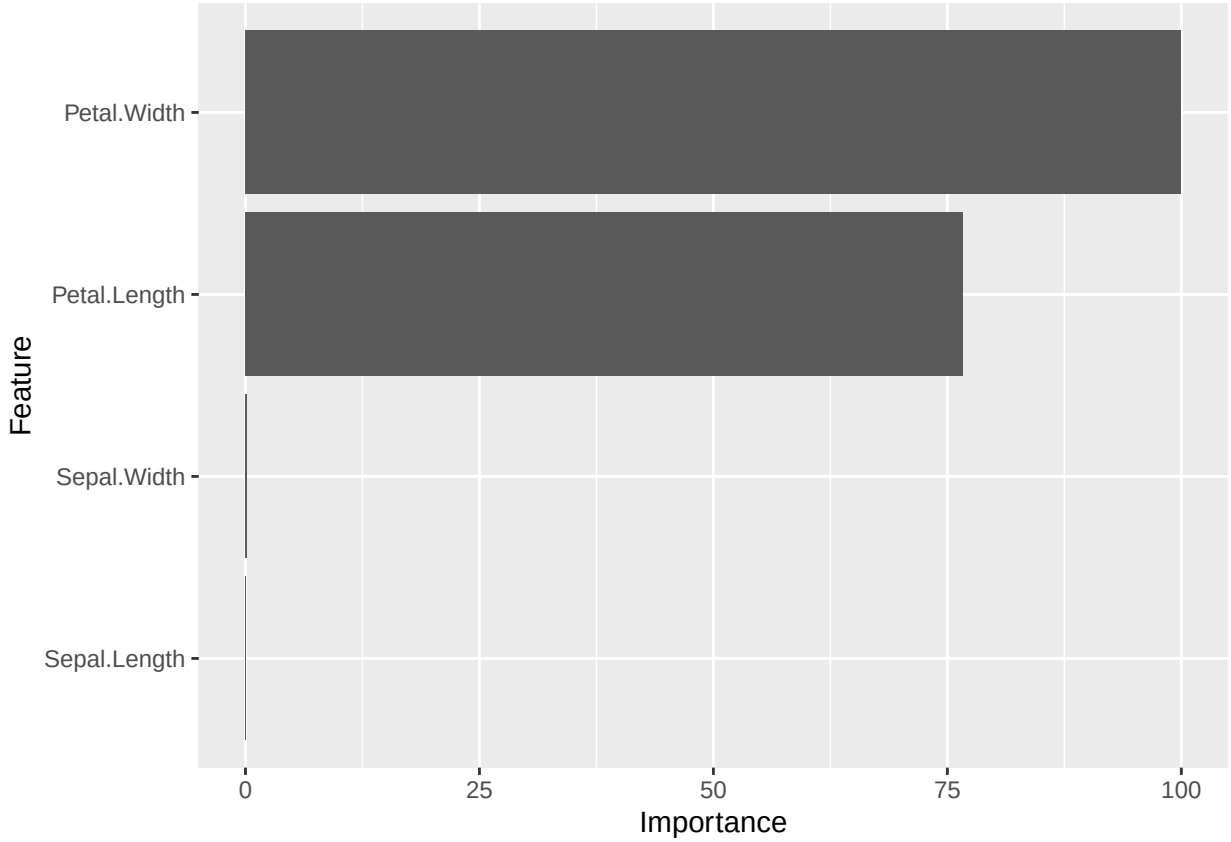
```
ggplot(train_knn,highlight = TRUE)
```



```r
train_rf <- train(Species ~., data = train_set, method ="rf", trControl = fitControl)
y_hat_rf <- predict(train_rf, test_set)
confusionMatrix(y_hat_rf, test_set$Species)
```

```
## Confusion Matrix and Statistics
##
##             Reference
## Prediction   setosa versicolor virginica
##   setosa         15          0         0
##   versicolor      0         14         1
##   virginica       0          1        14
##
## Overall Statistics
##
##                Accuracy : 0.9556
##                  95% CI : (0.8485, 0.9946)
##     No Information Rate : 0.3333
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.9333
##
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: setosa Class: versicolor Class: virginica
## Sensitivity                 1.0000            0.9333           0.9333
## Specificity                 1.0000            0.9667           0.9667
## Pos Pred Value              1.0000            0.9333           0.9333
## Neg Pred Value              1.0000            0.9667           0.9667
## Prevalence                  0.3333            0.3333           0.3333
## Detection Rate              0.3333            0.3111           0.3111
## Detection Prevalence        0.3333            0.3333           0.3333
## Balanced Accuracy           1.0000            0.9500           0.9500
```

```r
ggplot(varImp(train_rf))
```




## 5. Conclusions

```r
# a quite detailed info
```

```r
# a brief summary
```