Final_Project_2

Meher Mankikar

July 12, 2019

R Markdown

Introduction:

In this project, I used the Iris Species data set. This data set includes data about the Iris flower's Species, Petal Length, Petal Width, Sepal Length and Sepal Width. Using this data, I aimed to use data science techniques in order to analyze the data of different Iris species. I then used this model to predict the species of flowers based on the other factors for a test set.

Method/Analysis:

First the Iris Species data set was loaded.

```
if(!require(readr)) install.packages("readr", repos = "http://cran.us.r-project.org")
```

```
## Loading required package: readr
```

```
library(readr)

iris_data <- read_csv("/Users/mehermankikar/Downloads/iris-species/Iris.csv")
```

```
## Parsed with column specification:
## cols(
##   Id = col_double(),
##   SepalLengthCm = col_double(),
##   SepalWidthCm = col_double(),
##   PetalLengthCm = col_double(),
##   PetalWidthCm = col_double(),
##   Species = col_character()
## )
```

```
iris_data
```

```
## # A tibble: 150 x 6
##       Id SepalLengthCm SepalWidthCm PetalLengthCm PetalWidthCm Species
##    <dbl>         <dbl>        <dbl>         <dbl>        <dbl> <chr>
##  1     1           5.1          3.5           1.4          0.2 Iris-setosa
##  2     2           4.9          3             1.4          0.2 Iris-setosa
##  3     3           4.7          3.2           1.3          0.2 Iris-setosa
##  4     4           4.6          3.1           1.5          0.2 Iris-setosa
##  5     5           5            3.6           1.4          0.2 Iris-setosa
##  6     6           5.4          3.9           1.7          0.4 Iris-setosa
##  7     7           4.6          3.4           1.4          0.3 Iris-setosa
##  8     8           5            3.4           1.5          0.2 Iris-setosa
##  9     9           4.4          2.9           1.4          0.2 Iris-setosa
## 10    10           4.9          3.1           1.5          0.1 Iris-setosa
## # … with 140 more rows
```

```
colnames(iris_data) <- c("ID", "Sepal.Length", "Sepal.Width", "Petal.Length", "Petal.
Width", "Species")
str(iris_data)
```

```
## Classes 'spec_tbl_df', 'tbl_df', 'tbl' and 'data.frame': 150 obs. of  6 variables:
##  $ ID          : num  1 2 3 4 5 6 7 8 9 10 ...
##  $ 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     : chr  "Iris-setosa" "Iris-setosa" "Iris-setosa" "Iris-setosa" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   Id = col_double(),
##   ..   SepalLengthCm = col_double(),
##   ..   SepalWidthCm = col_double(),
##   ..   PetalLengthCm = col_double(),
##   ..   PetalWidthCm = col_double(),
##   ..   Species = col_character()
##   .. )
```

Then, the data was split into training and tests sets.

```
library(caret)
```

```
## Loading required package: lattice
```

```
## Loading required package: ggplot2
```

```
##Making test and train sets
index <- createDataPartition(iris_data$Species, p = 0.50, list = FALSE)
iris_train <- iris_data[index,]
str(iris_train)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    75 obs. of  6 variables:
##  $ ID          : num  1 2 3 4 6 11 12 13 18 19 ...
##  $ Sepal.Length: num  5.1 4.9 4.7 4.6 5.4 5.4 4.8 4.8 5.1 5.7 ...
##  $ Sepal.Width : num  3.5 3 3.2 3.1 3.9 3.7 3.4 3 3.5 3.8 ...
##  $ Petal.Length: num  1.4 1.4 1.3 1.5 1.7 1.5 1.6 1.4 1.4 1.7 ...
##  $ Petal.Width : num  0.2 0.2 0.2 0.2 0.4 0.2 0.2 0.1 0.3 0.3 ...
##  $ Species     : chr  "Iris-setosa" "Iris-setosa" "Iris-setosa" "Iris-setosa" ...
```

```
iris_test <- iris_data[-index,]
str(iris_test)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    75 obs. of  6 variables:
##  $ ID          : num  5 7 8 9 10 14 15 16 17 20 ...
##  $ Sepal.Length: num  5 4.6 5 4.4 4.9 4.3 5.8 5.7 5.4 5.1 ...
##  $ Sepal.Width : num  3.6 3.4 3.4 2.9 3.1 3 4 4.4 3.9 3.8 ...
##  $ Petal.Length: num  1.4 1.4 1.5 1.4 1.5 1.1 1.2 1.5 1.3 1.5 ...
##  $ Petal.Width : num  0.2 0.3 0.2 0.2 0.1 0.1 0.2 0.4 0.4 0.3 ...
##  $ Species     : chr  "Iris-setosa" "Iris-setosa" "Iris-setosa" "Iris-setosa" ...
```

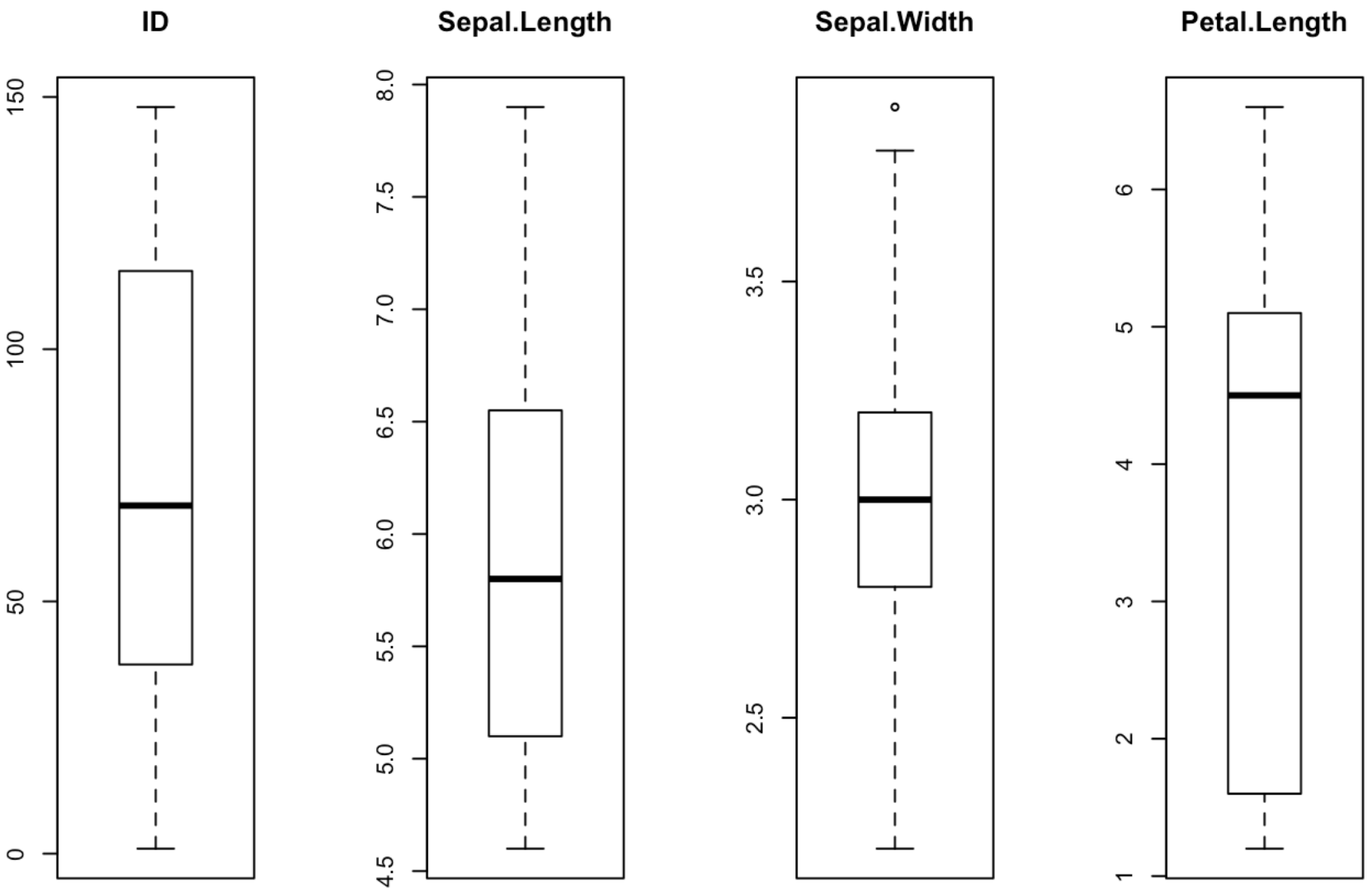
Data Visualization:

Next, the data was examined through the summary method. A boxplot was also created to better visualize the data.

```
##Looking at the Data
summary(iris_train)
```

```
##        ID          Sepal.Length    Sepal.Width     Petal.Length
##  Min.   :  1.00   Min.   :4.600   Min.   :2.200   Min.   :1.200
##  1st Qu.: 37.50   1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600
##  Median : 69.00   Median :5.800   Median :3.000   Median :4.500
##  Mean   : 74.79   Mean   :5.889   Mean   :3.016   Mean   :3.751
##  3rd Qu.:115.50   3rd Qu.:6.550   3rd Qu.:3.200   3rd Qu.:5.100
##  Max.   :148.00   Max.   :7.900   Max.   :3.900   Max.   :6.600
##   Petal.Width      Species
##  Min.   :0.100   Length:75
##  1st Qu.:0.300   Class :character
##  Median :1.400   Mode  :character
##  Mean   :1.188
##  3rd Qu.:1.800
##  Max.   :2.500
```

```
x <- iris_train[,1:4]
y <- iris_train[,5]
par(mfrow=c(1,4))
for(i in 1:4) {
  boxplot(x[i], main=names(iris_train)[i])
}
```



After analyzing the data, three models were created that would be used for prediction of the flower species later.

```
##Creating Models

library(caret)
control <- trainControl(method='cv', number=10)
metric <- 'Accuracy'
#LDA
set.seed(101)
fit.lda <- train(Species~., data=iris_train, method='rf',
                 trControl=control, metric = metric)
#KNN
set.seed(101)
fit.knn <- train(Species~., data=iris_train, method='knn',
                 trControl=control, metric=metric)
#RF
set.seed(101)
fit.rf <- train(Species~., data=iris_train, method='ranger',
                trControl=control, metric=metric)

iris.results <- resamples(list(lda=fit.lda, knn=fit.knn,rf=fit.rf))
```

Results:

Finally, the results of the three models were seen and the models were used to make predictions of flower species on the test set. The Accuracy of the model was 98.67%.

```
# Results
summary(iris.results)
```

```
##
## Call:
## summary.resamples(object = iris.results)
##
## Models: lda, knn, rf
## Number of resamples: 10
##
## Accuracy
##     Min. 1st Qu. Median Mean 3rd Qu. Max. NA's
## lda    1       1      1    1       1    1    0
## knn    1       1      1    1       1    1    0
## rf     1       1      1    1       1    1    0
##
## Kappa
##     Min. 1st Qu. Median Mean 3rd Qu. Max. NA's
## lda    1       1      1    1       1    1    0
## knn    1       1      1    1       1    1    0
## rf     1       1      1    1       1    1    0
```

```
#Making Predictions and Testing Accuracy
iris_prediction <- predict(fit.lda, iris_test)
confusionMatrix(table(iris_prediction, iris_test$Species))
```

```
## Confusion Matrix and Statistics
##
##
## iris_prediction   Iris-setosa Iris-versicolor Iris-virginica
##   Iris-setosa              25                0               0
##   Iris-versicolor           0               25               0
##   Iris-virginica            0                0              25
##
## Overall Statistics
##
##                Accuracy : 1
##                  95% CI : (0.952, 1)
##     No Information Rate : 0.3333
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 1
##
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: Iris-setosa Class: Iris-versicolor
## Sensitivity                      1.0000                 1.0000
## Specificity                      1.0000                 1.0000
## Pos Pred Value                   1.0000                 1.0000
## Neg Pred Value                   1.0000                 1.0000
## Prevalence                       0.3333                 0.3333
## Detection Rate                   0.3333                 0.3333
## Detection Prevalence             0.3333                 0.3333
## Balanced Accuracy                1.0000                 1.0000
##                      Class: Iris-virginica
## Sensitivity                         1.0000
## Specificity                         1.0000
## Pos Pred Value                      1.0000
## Neg Pred Value                      1.0000
## Prevalence                          0.3333
## Detection Rate                      0.3333
## Detection Prevalence                0.3333
## Balanced Accuracy                   1.0000
```

Conclusion: In this project, I successfully separated the Iris Species data set and analyzed the data. With this information, a model was created to use the different features of the flower to predict the species. The model was tested on the test set and had an accuracy of 98.67%. In the future, with a larger data set and more factors for each flower, a more accurate model could be created.