Classification Case

Overview

The aim of this work is to predict the gender class and evaluate the algorithms which are used. First the libraries we are going to use are installed to the environment.

```
library(ggplot2)
library(caret)
library(gridExtra)
library(randomForest)
```

The dataset is loaded and summary about the dataset is shown.

```
data <- read.csv("dataset.CSV", header = T)
str(data)
```

```
## 'data.frame':    10000 obs. of  3 variables:
##  $ Gender: Factor w/ 2 levels "Female","Male": 2 2 2 2 2 2 2 2 2 2 ...
##  $ Height: num  73.8 68.8 74.1 71.7 69.9 ...
##  $ Weight: num  242 162 213 220 206 ...
```

```
prop.table(table(data$Gender))
```

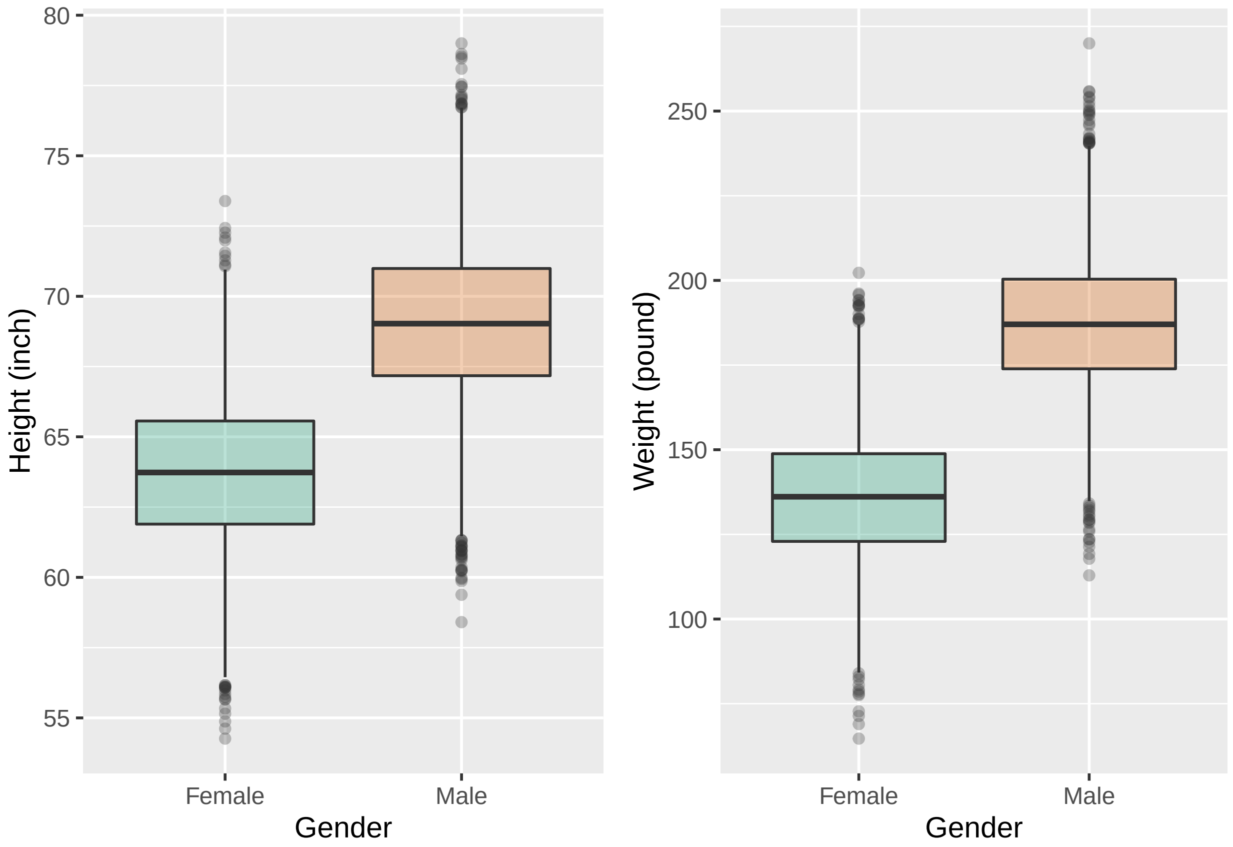
```
##
## Female   Male
##    0.5    0.5
```

Stastics of the dataset's features are shown on the boxplots below.

```
plot1 <- ggplot(data, aes(x=Gender, y=Height, fill=Gender)) +
  ylab("Height (inch)") +
  geom_boxplot(alpha=0.3) +
  theme(legend.position="none") +
  scale_fill_brewer(palette="Dark2")

plot2 <- ggplot(data, aes(x=Gender, y=Weight, fill=Gender)) +
  ylab("Weight (pound)") +
  geom_boxplot(alpha=0.3) +
  theme(legend.position="none") +
  scale_fill_brewer(palette="Dark2")

grid.arrange(plot1, plot2, ncol=2)
```



The dataset is split into train and test sets by half and half.

```
intrain <- createDataPartition(y=data$Gender,p=0.5,list=FALSE)
train <- data[intrain,]
test <- data[-intrain,]
```

Proportions of train and test sets' gender are found to be same.

```
prop.table(table(train$Gender)); prop.table(table(test$Gender))
```

```
##
## Female   Male
##    0.5    0.5
```

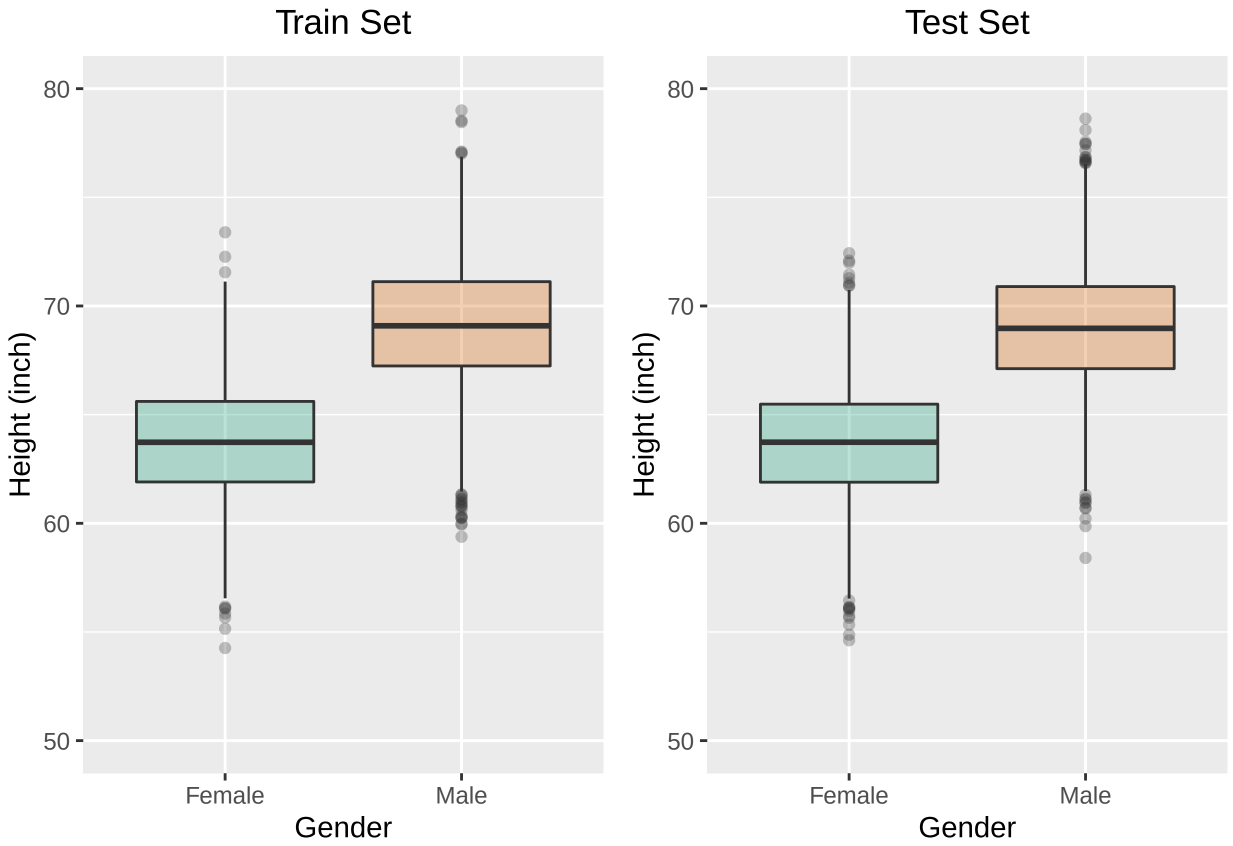
```
##
## Female   Male
##    0.5    0.5
```

Boxplots below also demonstrate that distribution of 'Height' is similar in both sets.

```
plot3 <- ggplot(train, aes(x=Gender, y=Height, fill=Gender)) +
  ggtitle("Train Set") +
  ylab("Height (inch)") +
  ylim(50,80) +
  geom_boxplot(alpha=0.3) +
  theme(legend.position="none", plot.title = element_text(hjust = 0.5)) +
  scale_fill_brewer(palette="Dark2")

plot4 <- ggplot(test, aes(x=Gender, y=Height, fill=Gender)) +
  ggtitle("Test Set") +
  ylab("Height (inch)") +
  ylim(50,80) +
  geom_boxplot(alpha=0.3) +
  theme(legend.position="none", plot.title = element_text(hjust = 0.5)) +
  scale_fill_brewer(palette="Dark2")

grid.arrange(plot3, plot4, ncol=2)
```

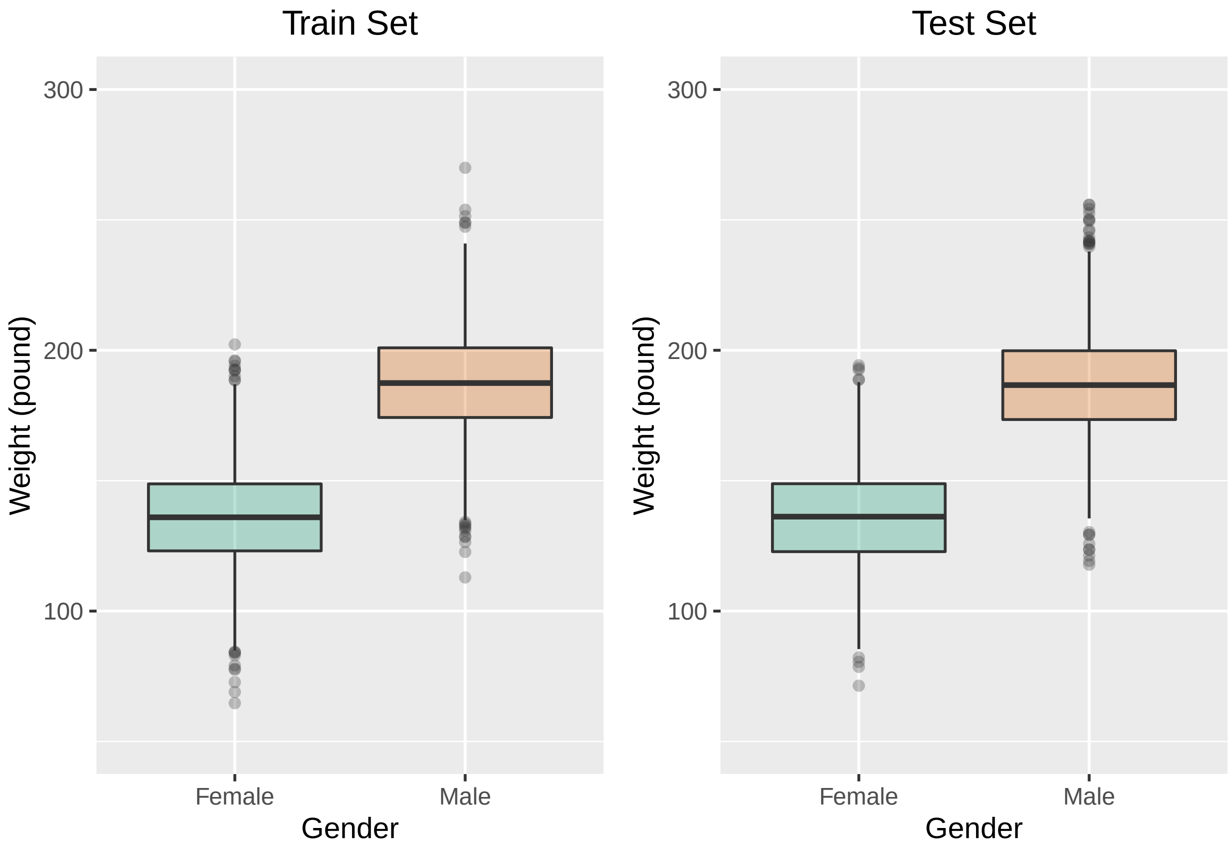


Boxplots below also demonstrate that distribution of ‘Weight’ is similar in both sets.

```
plot5 <- ggplot(train, aes(x=Gender, y=Weight, fill=Gender)) +
  ggtitle("Train Set") +
  ylab("Weight (pound)") +
  ylim(50,300) +
  geom_boxplot(alpha=0.3) +
  theme(legend.position="none", plot.title = element_text(hjust = 0.5)) +
  scale_fill_brewer(palette="Dark2")

plot6 <- ggplot(test, aes(x=Gender, y=Weight, fill=Gender)) +
  ggtitle("Test Set") +
  ylab("Weight (pound)") +
  ylim(50,300) +
  geom_boxplot(alpha=0.3) +
  theme(legend.position="none", plot.title = element_text(hjust = 0.5)) +
  scale_fill_brewer(palette="Dark2")

grid.arrange(plot5, plot6, ncol=2)
```


Train Set
Test Set
Weight (pound)
Female
Male
Gender