Data Mining Project on Classification of Iris Flowers

Students Name

19BDS0016 SAI SREEKAR

20BDS0401 RUTHVIK REDDY

Institution Affiliation

Date

Data Mining Project on Classification of Iris Flowers

This project has been designed to provide a comprehensive analysis on the charecteristics of the iris flower. Iris flower is one of the most dominant floricultural products in the U.S. Just like any other flower, the irises are a wonderful attraction that is often too hard to resist. Most of these flowers are produced for commercial purpose which is later used to manufacture perfumes (Crampton, 2018). Additionally, rhizomes from iris flower can be used as a potpourri and natural toothpaste. Some parts of rhizomes can also be used to flavor beverages and food.

There are three common species of the iris flower. These are setota, Versicolor and virginica. These species are best classified according to how they are propagated. Some are propagated using bulbs while others are propagated using rhizomes (Pavlis, 2016). Alternatively, iris flowers can be classified using the type of petals they have. While most of them have three petals that are curved downwards, some have petals that either straight or fall away from the center of the flower.

The purpose of this project was to quantify the morphological variations of the three species of iris flower that were identified above. If given the raw data from the dataset one cannot correctly determine the species under which the flower fall. After analysis from this project, one will be able to determine the species under which the flower falls. It is also important to note that one species from the three is linearly separable from the other two.

In order to meet the objective of this project, I used r studio to analyze the data. R studio is the most common tool used for data classification and analysis. This is due to the fact that R studio provides an efficient platform whose output is eye appealing (Wickham & Grolemund, 2017). R studio supports various graphical applications and statistical computations. It also

offers one an opportunity to carry out some computational processes such as clustering, classification techniques, linear and nonlinear modeling. R programming language in itself is very adaptable and extensible when compared to other programming languages.

Data analysis

The dataset obtained from the website consisted of 50 samples representing the three species. All the three species of iris were classified using some attributes. The attributes that were measured included the length and the width of both the sepals and petals in centimeters. For purpose of easy analysis, five variables were used. Four of these variables were quantitative variables that were used to describe both the length and the width of parts of flowers in centimeters. The last variable known as Variable Species is a categorical splitting mechanism which classifies the species as setosa, Versicolor and virginica.

After obtaining the data from the source, it was necessary to convert them into a form that can be easily read by the R studio. In this case, I chose to convert the text file into a comma-separated value (CSV). After conversion, I imported the dataset into R studio. The next step was to plot the graph to shows the density distribution of the quantitative variable identified earlier on. The command used to obtain this is shown below.

```
plot(density(iris$Sepal.Length), col=iris$Species)
```

The resulting density graph is as shown below

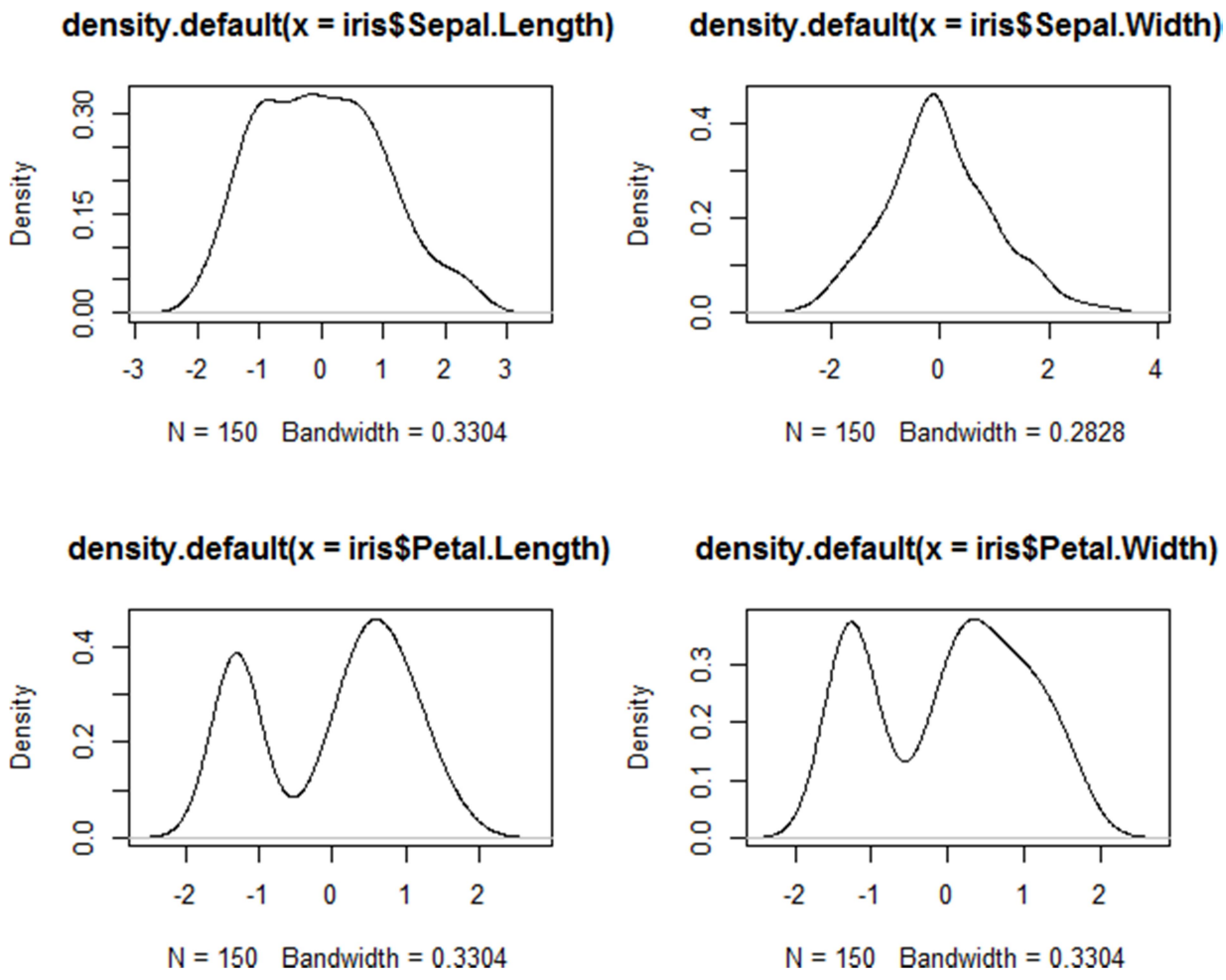


Figure 1. density graph

Alternatively, we can plot the histogram that shows the distribution curve for the variables identified above. The graph is meant to separately show the frequencies of the petals and the sepals against the length of the flower. The code that was used to obtain this histogram is given below. The code shows the length of the sepall. This code can be modified to represent the sepal width, petal lent and its respective width.

```
hist(iris$Sepal.Length, col="orange", breaks=18)
```

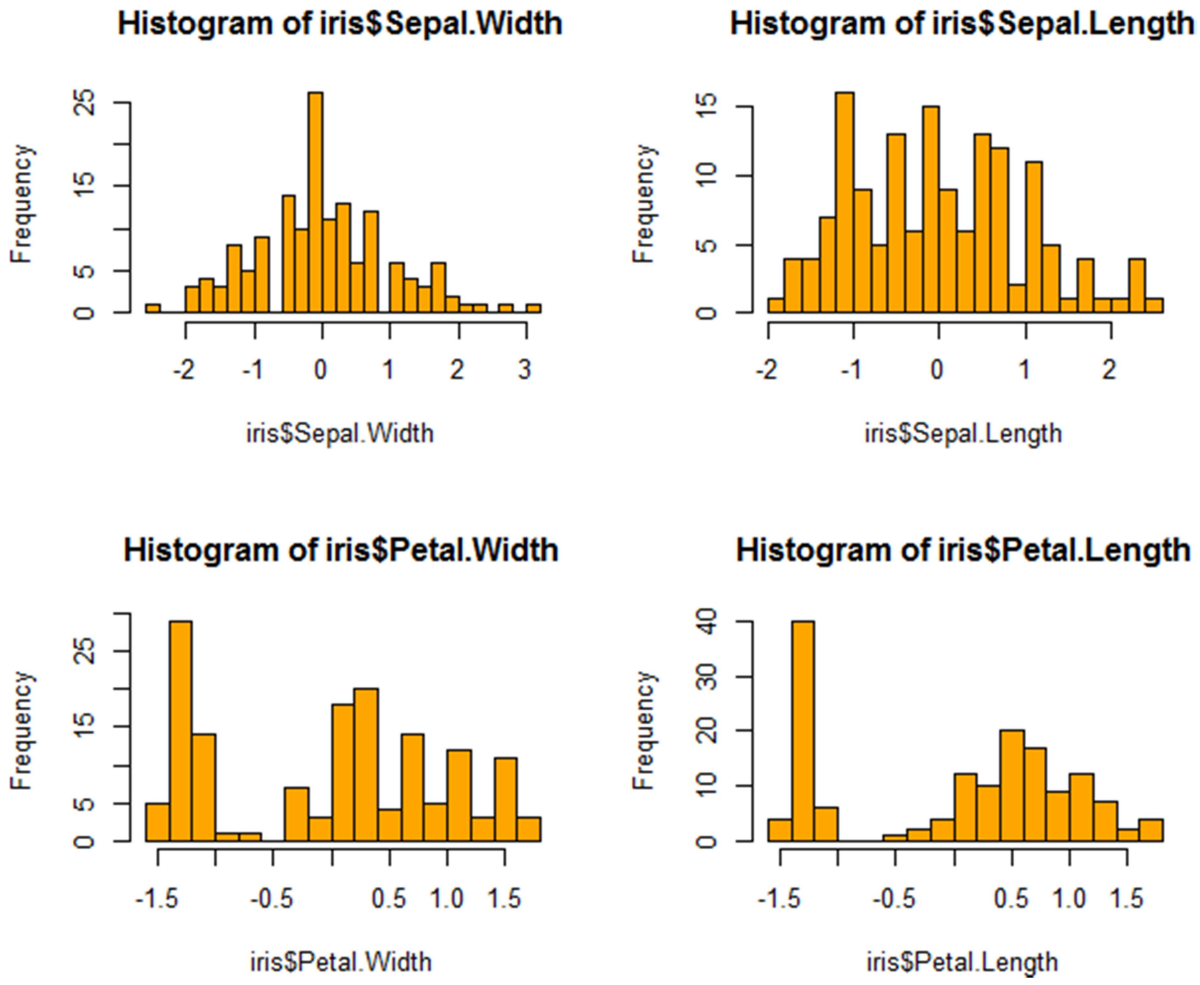


Figure 2. distribution histogram

Data relationship analysis

For purpose of the relationship, two graphs were used. The main goal of this as stated under the objectives was to find the relationship between the width and the length of both the sepal and the petals. One such graph obtained was the scattering box. This graph is mostly applicable when you want to scatter one variable on the axis which makes it easy to look for relationships between the variables. In other words, the tighter points will seem to 'hug the line'. In order to achieve this, I used the following command

ggplot(data = iris, aes(x = Sepal.Length, y = Sepal.Width, col = Species)) + geom_point() + geom_smooth(method="lm") + facet_grid(.~iris$Species)

This resulted in the graph shown below.

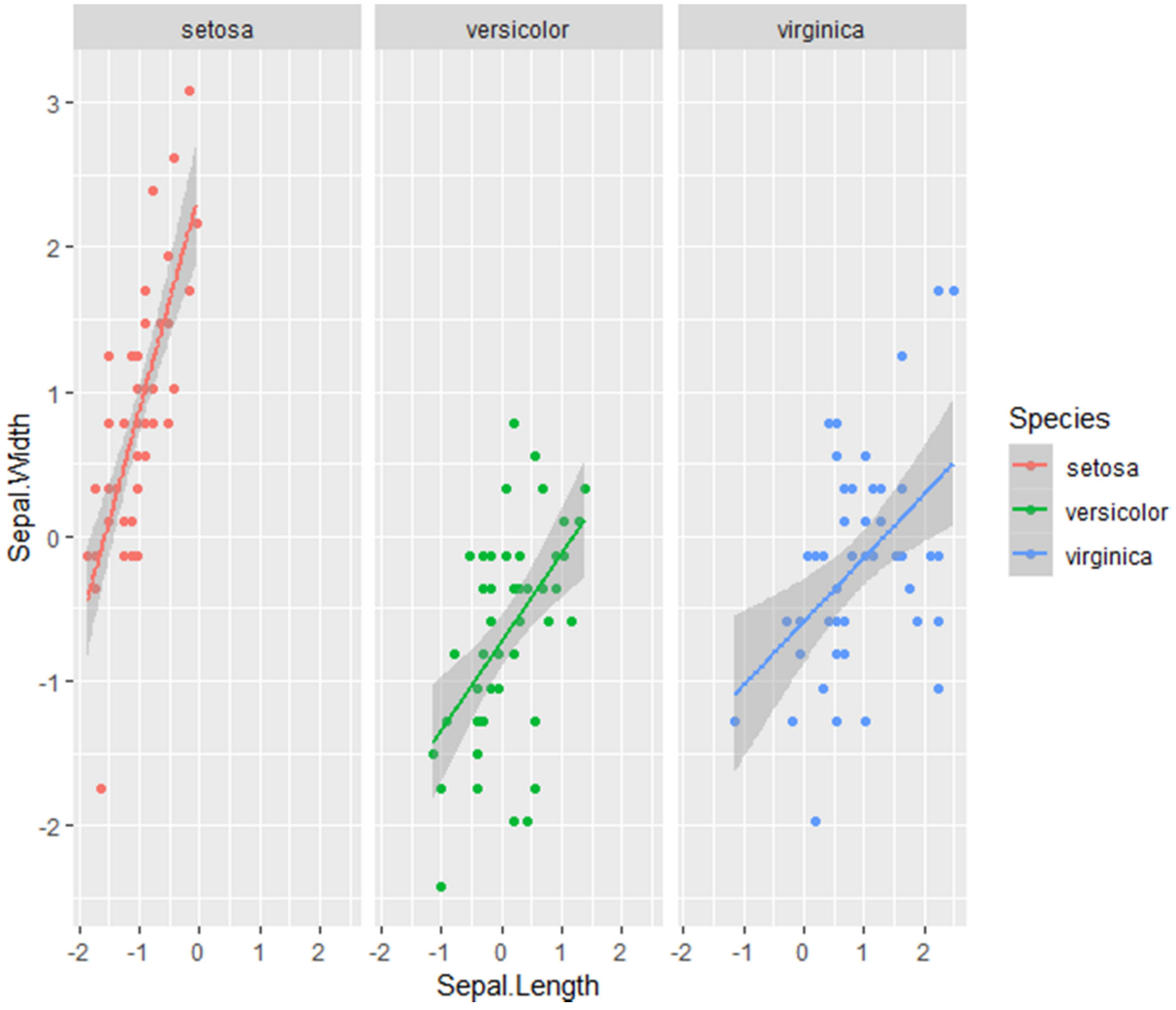
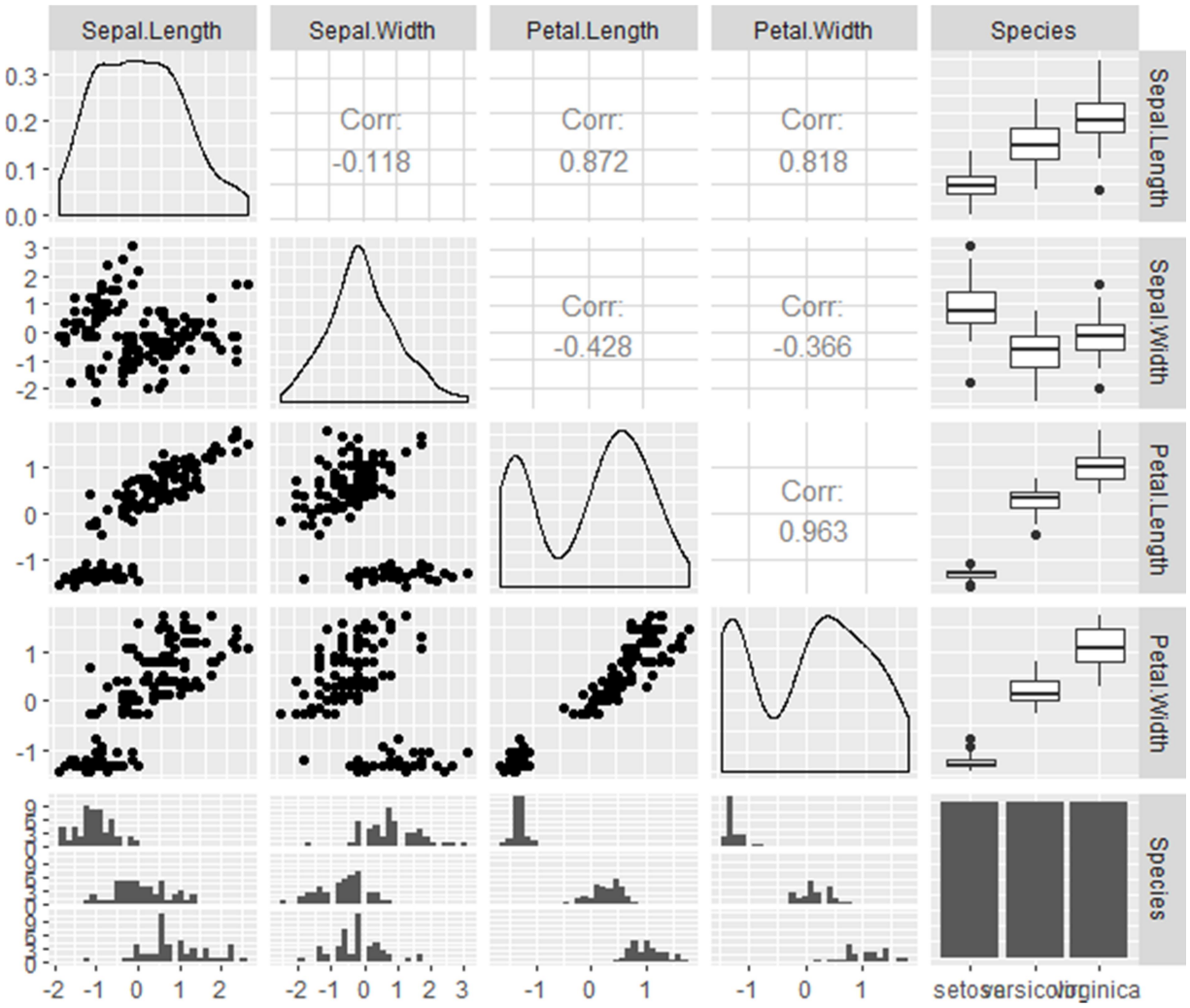


Figure 3. regression curve

Alternatively, a multiple regression line graph can be obtained from the above dataset. This is necessary data which requires a matrix of plots. When used, the graph obtained is useful

for quickly exploring the certain relationship between multiple columns and rows in a data frame (Fahrmeir, Kneib & Lang, 2013). The library function used for this process is the ggpairs which has both lower and upper arguments. With these arguments in place, one can choose the type of plot as either lower or upper diagonal of the matrix.



Classification of dataset

Having seen the different graphs, the next process is to classify the dataset into flower classes that can make it easy to predict. The process of data classification is a critical stage which requires extra keenness. This is because mistakes that arise in this stage may render the

whole analytical process useless. According to Brownlee (2016), a properly classified data helps one to build data that is easy to work with. Additionally, it enables one to figure out the possible results. During this stage, one may also discover that some attributes have a familiar distribution which may assist in scaling the project.

In the case of iris data set, I divided the dataset into training and test set to apply the KNN classification. From this dataset, it is simple to divide the data such that sixty percent is used as a training dataset while the rest is used as an a testing set. The code for this as shown below

```
set.seed(12366894)
setosa<- rbind(iris[iris$Species=="setosa",])
versicolor<- rbind(iris[iris$Species=="Versicolor",])
virginica<- rbind(iris[iris$Species=="Virginia",])
```

Validation of data

As seen from the graphs above one can make some inference concerning the general trends of some attributes portrayed by the graphs. Starting with the density plot graphs, it is evident that the graphs portray a curvilinear curve. The iris sepal width shows an almost Gaussian curve while petal length is a polynomial curve. From the graphs, it is also notable that there is a drastic decrease in the density as the bandwidth approaches two. Furthermore, it is evident from the histograms above that the curves do not observe a normal distribution curve.

The relationship graph has been presented in two way. The simplest way as seen above is the line correlation curve. As seen in figure 3, the graph represents how the petal lengths correlate with the three species of the iris flower. The general observation made is that the setosa

has a negative correlation while the Versicolor has a normal correlation. The petal size of virgica has a positive correlation with the petal length.

More detailed information can be obtained from the correlation matrix as seen in figure 4. From this figure, it is evident that there is a strong correlation between the length of the petal and its width. The correlation coefficient of 0.963 was registered. The length of the petal and the length of the sepals can also be said to be strongly correlated. The length of the petal and the width of the sepal are negatively correlated with a correlation coefficiency of 0.428 being registered. On the other hand, the lengths of the sepal and its width have a weak correlation. This can as well mean that two variables do not affect each other.

Conclusion

It is fair enough to say that the classification system above provided an accurate and efficient way of classifying the data. The overall structure used above reflects the knowledge that can uncover the hidden meanings within the dataset. The accuracy seen above can be attributed to the weights that were adjusted according to the association between the classes and various library rules available.

From the above case study, one can easily make inferences about the iris flowers. The general observation as seen above is that the range of petal length, petal width, sepal length and the width of the sepal are different for the different species of the iris flower. The range seen from the setosa flower is different while that of the Versicolor and virginicia are slightly closer to each other. A user can now correctly deduce on the type of the flower by just looking at the above-mentioned attributes.

References

Brownlee, J. (2016). Better Understand Your Data in R Using Descriptive Statistics (8 recipes you can use today). R Machine Learning. Retrieved from: https://machinelearningmastery.com/descriptive-statistics-examples-with-r/

Crampton, L. (2018). Forty Facts About Irises: Beautiful Flowers and Useful Plants; location retrieved from https://owlcation.com/stem/Facts-about-Irises-Beautiful-and-Interesting-Flowers

Fahrmeir, L., Kneib, T., & Lang, S. (2013). *Regression: Models, methods and applications*.

Pavlis, R. (2016). Iris Identification – Which Type of Iris Do I Have?; garden fundamentals. Retrieved from: http://www.gardenfundamentals.com/iris-identification-type/

Wickham, H., & Grolemund, G. (2017). *R for data science: Import, tidy, transform, visualize and model data*. Sebastopol: O'Reilly.