**ASSIGNMENT 1**

**Problem 1: Differentiate between the following terms:**

1. Classification and Clustering

Classification is a supervised data mining technique. In classification we have a set of predefined classes and the class label of the data objects are known.

Clustering allows a user to make groups of data to determine patterns from the data, and find whether there is some relationship between the objects without class label.

1. Classification and Prediction

The decision tree applied to existing data is called as classification. When the decision tree is applied to new data, for which the class is unknown that is termed as prediction. Also predicting class labels is classification, and predicting values is prediction

1. Data Warehouse and Database

Data warehouse is a repository of data in which data from various sources is saved. It is used for reporting and data analysis.

Database is an organized collection of data. In this data can be modified/updated.

1. Data Mining and OLAP

Data Mining deals with extracting interesting patterns, ratios from large sets of data. It combines many methods from artificial intelligence, statistics and database management.

OLAP (Online Analytical Processing) tool provides multidimensional data analysis and they provide summaries of the data.

1. Machine Learning and Statistics

Machine learning is focused about scalability and using the predictions to make decisions.

Statistics is focused on analyzing existing data, and drawing valid conclusions.

**Problem 2: Discuss whether or not each of the following activities is a data mining task.**

1. Monitoring the heart rate of a patient for abnormalities.

This is a data mining task. Build a model of the normal behavior of heart rate and an unusual heart behavior. This area of data mining known as anomaly detection. This can be considered as a classification problem, if it contains both normal and abnormal heart behavior.

1. Computing the total sales of a company.

This is not a data mining task, as there are no patterns to extract from the data.

1. Sorting a student database based on student identification numbers.

No, this is a simple database query.

1. Predicting the outcomes of tossing a (fair) pair of dice.

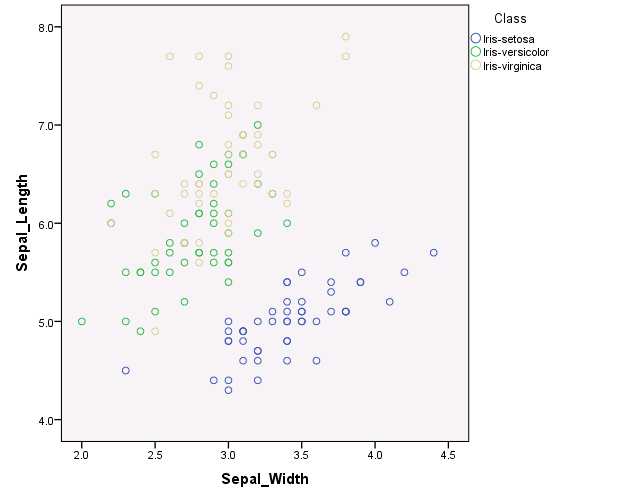
This is not a data mining task. This a problem related to probability calculations.

1. Monitoring seismic waves for earthquake activities.

Yes, this is a data mining task. In this case, build a model of various kinds of seismic waves associated with earthquake activities. This area of data mining is called as classification.

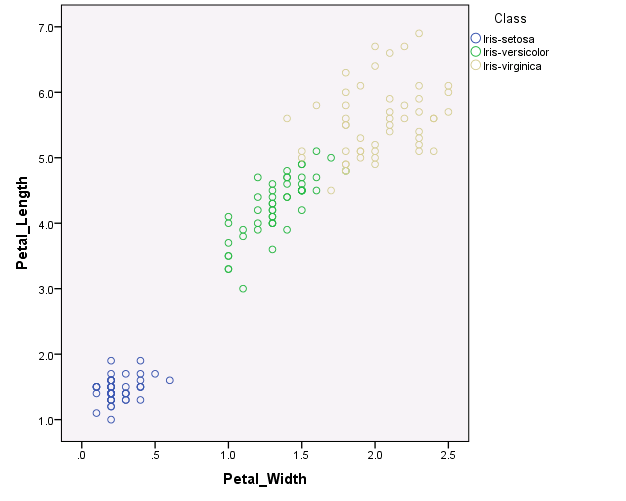
**Problem 3:**

1. Visualize and interpret the relationship between the two sepal variables, sepal length and sepal width. Provide the scatterplot that you created to visualize the data along with your interpretation. When you plot the data, you may want to use different colors/signs for representing the data points belonging to the different three class species. Do you think that a classification algorithm will be successful in classifying the data with respect to these two variables? Justify your answer.



The scatterplot for sepal variables, sepal length and sepal width are plotted in the graph. It is clearly evident from the above graph, the Iris-setosa is separated from the rest of the two class species. However Iris-versicolor and Iris-virginica have smaller region overlapping with each other. Hence it can stated that, Iris-setosa might be failed on the other two.

1. Repeat part a. for the petal variables.

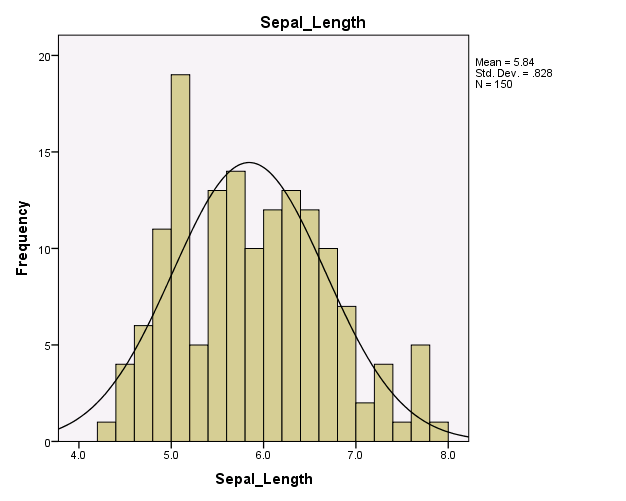
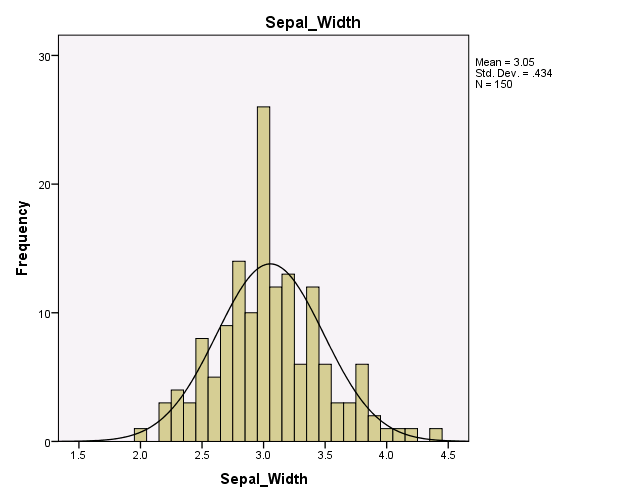


The scatterplot for petal variables are plotted. From the above graph we can see that, the Iris-Setosa is very well separated from the rest of the class species. Whereas the Iris-Versicolor and Iris-Virginica are overlapped for a very smaller area.

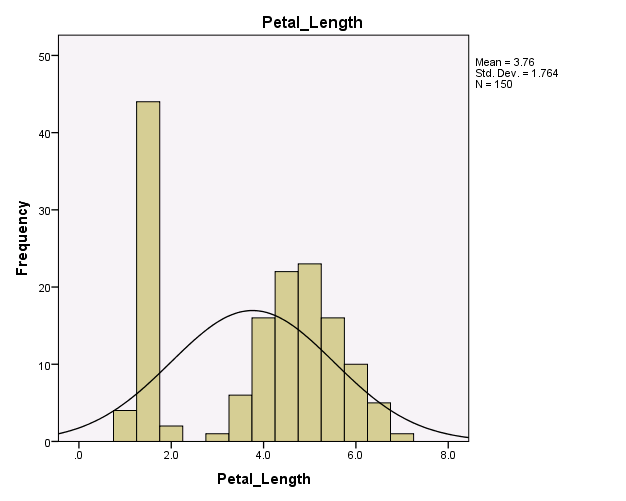
Also it is very clear from the graph the Iris-Setosa has the shortest petal length and width, Iris-Versicolor with average and Iris-Virginica has the longest length and width. Hence we can easily classify Iris species based on the petal variables.

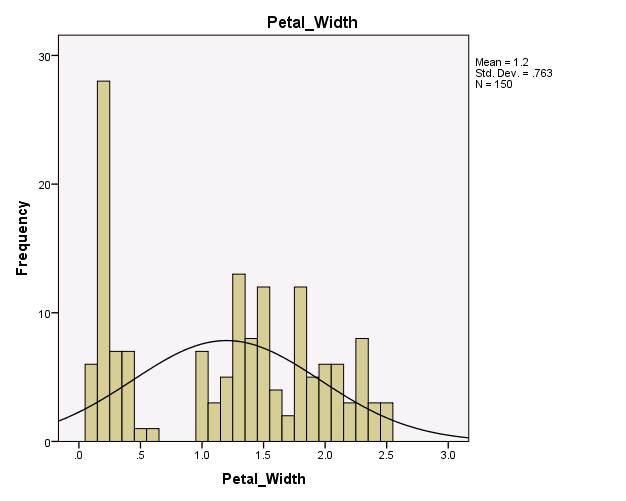
1. Draw the histograms of the four variables and interpret the distributions of each one of the four variables.

The distribution of the sepal variables are almost normal.

Whereas the distribution of the petal variables are skewed to the right.





1. Determine if there are any outliers in the data with respect to the sepal length.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Descriptive Statistics** | | | | | | | |
|  | | | | | | | |
|  | N | Range | Minimum | Maximum | Mean | Std. Deviation | Variance |
| Sepal\_Length | 150 | 3.6 | 4.3 | 7.9 | 5.843 | .8281 | .686 |
| Valid N (listwise) | 150 |  |  |  |  |  |  |

According to \*IQR Rule, values lesser than Q1 – (Q3- Q1) or greater than Q3+ (Q3-Q1) can be considered s outliers. In above case Since Q1-1.5\*IQR is smaller than minimum and Q+1.5\*IQR is greater than maximum, there are no outliers detected

1. Repeat d. for the petal length.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Descriptive Statistics** | | | | | | | |
|  | N | Range | Minimum | Maximum | Mean | Std. Deviation | Variance |
| Petal\_Length | 150 | 5.9 | 1.0 | 6.9 | 3.759 | 1.7644 | 3.113 |
| Valid N (listwise) | 150 |  |  |  |  |  |  |

In above case Since Q1-1.5\*IQR is smaller than minimum and Q+1.5\*IQR is greater than maximum, there are no outliers detected