Assignment 1 - DSC 441

Problem 1 (5points): Differentiate between the following terms:

a. classification and clustering

Classification means that there are defined classes or groups. With classification, each data object belongs to a class. It is used in supervised learning where items are assigned predefined labels.

Clustering means that there are no predefined classes, but data objects are grouped together because there is a relationship between them.
It is used in unsupervised learning where similar items are grouped together based on their features and properties.

b. classification and prediction

Classification is a machine learning algorithm to predict the class of given data points in a data set. Classes are called as labels or categories.

Prediction derives the relationship between a thing you know and a thing you need to predict for future reference. It is not necessarily related to future events, but the used variables are unknown.

c. feature selection and feature extraction

Feature selection is the process of selecting relevant attributes for use in the construction of a model. It selects a subset of the existing features

Feature extraction on the other hand is the process of combining attributes to create new non-redundant attributes. It transforms the existing features into a lower dimension.

d. data mining and SQL

Data mining is the process of discovering interesting patterns and knowledge from large amounts of data. The information is implicit in this case, meaning intelligent methods need to be applied to the data in order to discover patterns.

SQL: structured query Language is used to communicate with a database, and it is responsible for querying and editing information stored in a certain database management system.

e. data warehouses and data marts

A **Data Warehousing** (DW) is process for collecting and managing data from varied sources to provide meaningful business insights. A Data warehouse is typically used to connect and analyze business data from heterogeneous sources. The data warehouse is the core of the BI system which is built for data analysis and reporting.

A **data mart** is a simple section of the data warehouse that delivers a single functional data set. In a human resources database, we could create data marts for Employees, Benefits, or Payroll to name a few. It holds data about a specific subject related to the task.

Problem 2 (5 points):
Discuss whether or not each of the following activities is a data mining task.

(a) Monitoring the heart rate of a patient for abnormalities.
Yes, this would be an example of a data mining task Data as it uses preexisting data to draw inference about the abnormalities.

(b) Computing the total number of courses offering by a university.
No, this would not be considered an example of data mining task as we can directly count the number of courses from given data.

(c) Sorting a student database based on student identification numbers.
No, this is not an example of data mining task as it can also be accomplished with a single SQL query without going through pattern evaluation

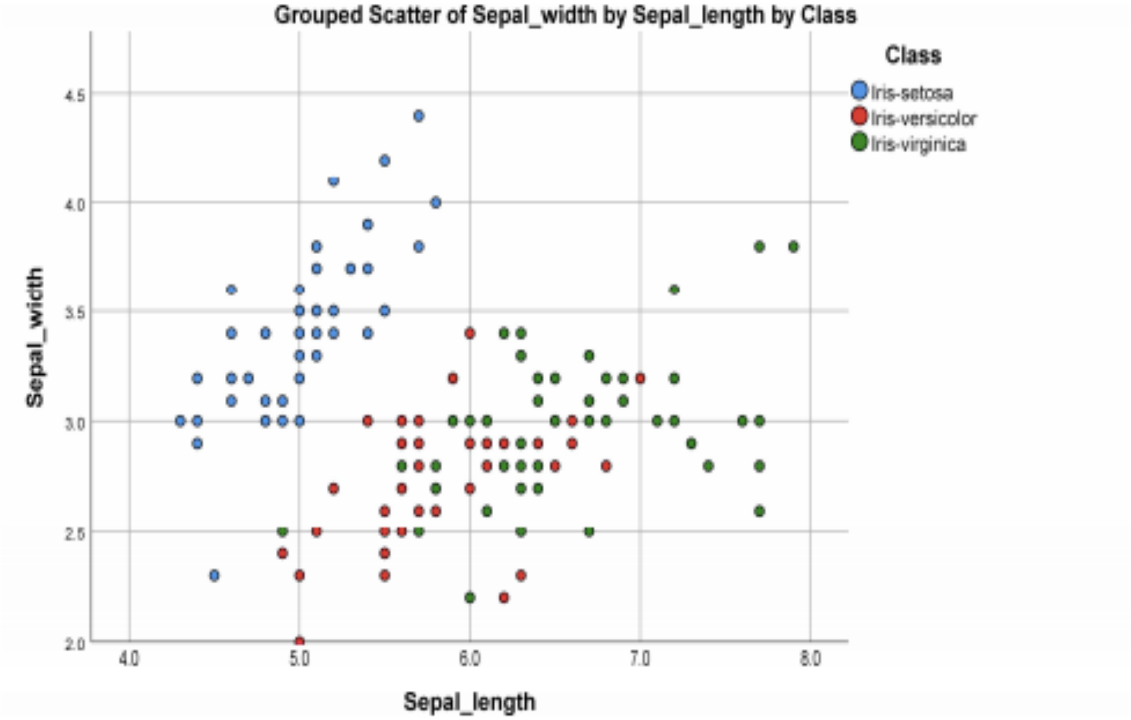
(d) Predicting the outcomes of tossing a (fair) pair of dice.
No, it can be solved using probability and not using data mining.

(e) Monitoring seismic waves for earthquake activities.
Yes, Data mining task as it uses previously stored data from the dataset to predict future earthquake activities.

Problem 3 (15 points): Fisher's iris data (download the IRIS dataset from http://archive.ics.uci.edu/ml/datasets/Iris) consists of measurements on the sepal length, sepal width, petal length, and petal width of 150 iris specimens. There are 50 specimens from each of three species.
Use SPSS software to answer the following questions:

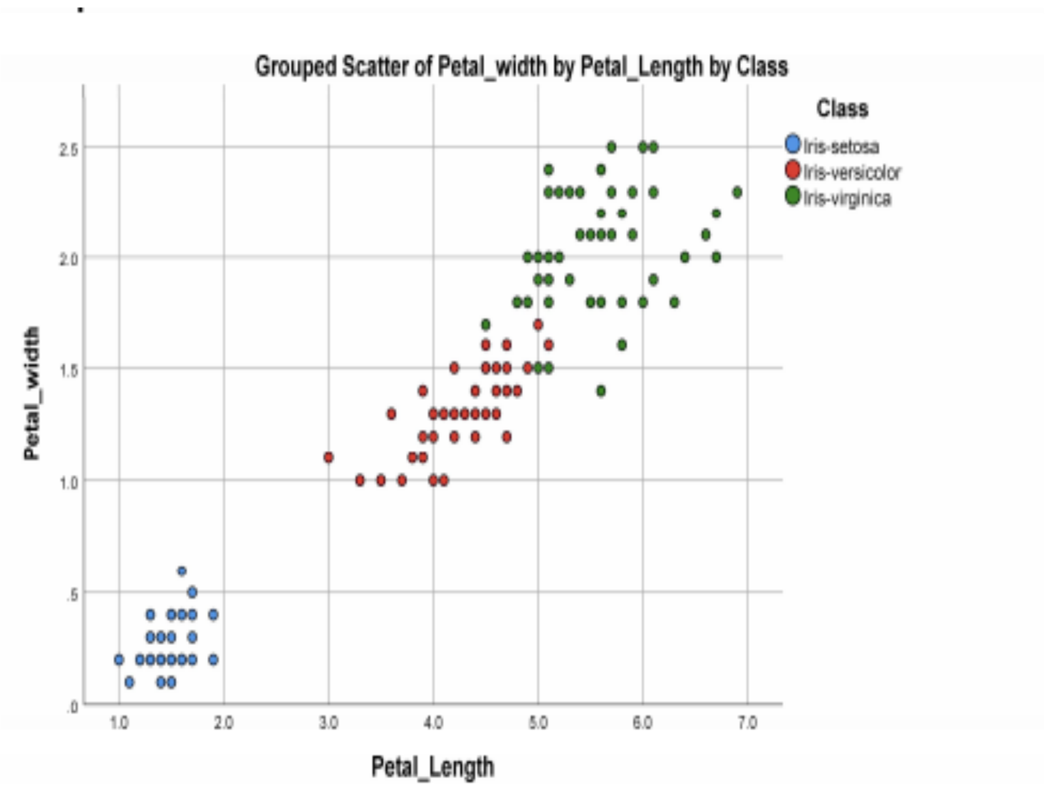
A. Visualize and interpret the relationship between the two sepal variables, sepal length and sepal width. Provide the scatterplot that you created to visualize the data along with your interpretation. When you plot the data, you may want to use different colors/signs for representing the data points 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.

```
[DataSet1]
```



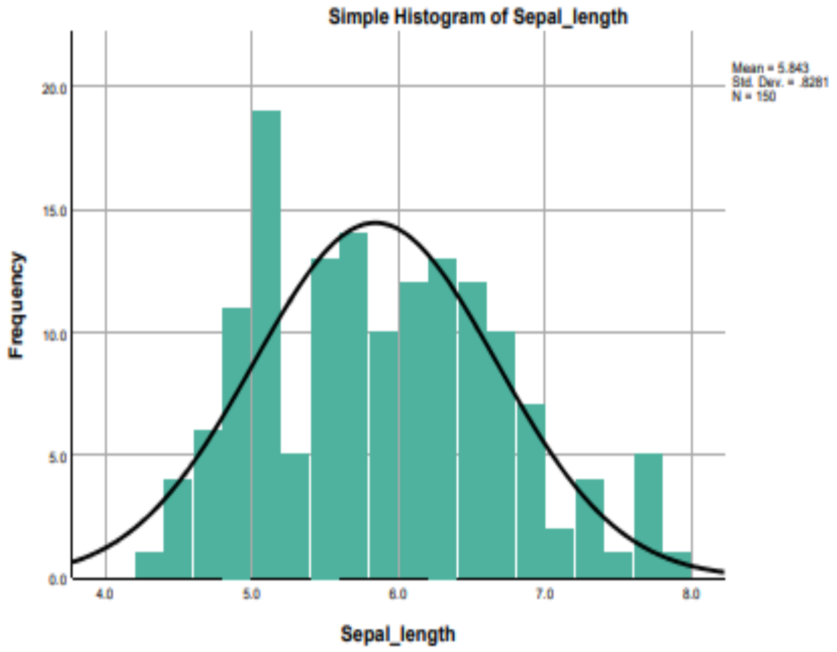
. I don’t think that classifying the data would work when analyzing the relationship between sepal length and sepal width. When we look at the scatterplot, we can see the two distinct clusters of data in Iris-setosa and the other two classes Iris-versicolor and Iris-virginica. However, there is a lot of overlap between Iris-versicolor and Iris-virginica. Hence, the classification based on these two variables will be unsuccessful.

b. Repeat part a. for the petal variables.

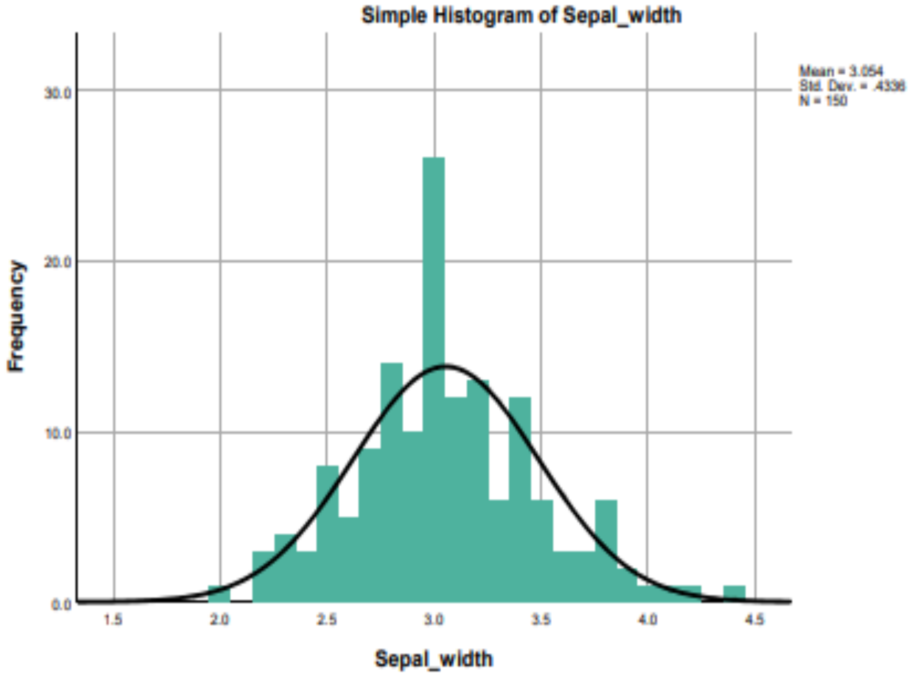


we can observe that the three classes, Iris-setosa, Iris-virginica and Iris-versicolor can easily be classified in 3 different classes with a few overlaps between Iris-versicolor and Iris-virginica. Hence the classification algorithm based on the two petal variables will be successful.

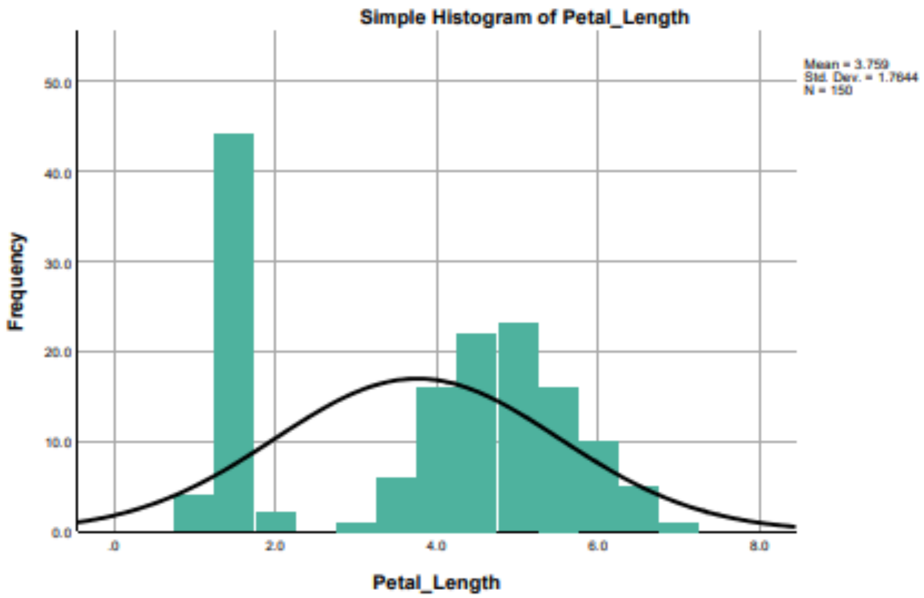
c. Draw the histograms of the four variables and interpret the distributions of each one of the four variables.



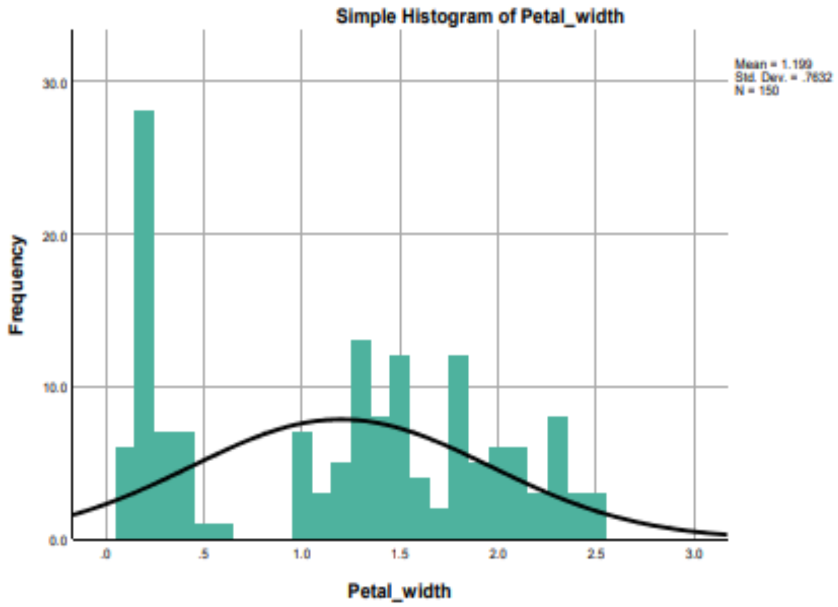
The mean of sepal length is 5.843 and the standard deviation is 0.8281.The data is skewed



The mean of sepal width is 3.054 and the standard deviation is 0.4336. It is symmetric and unimodal.

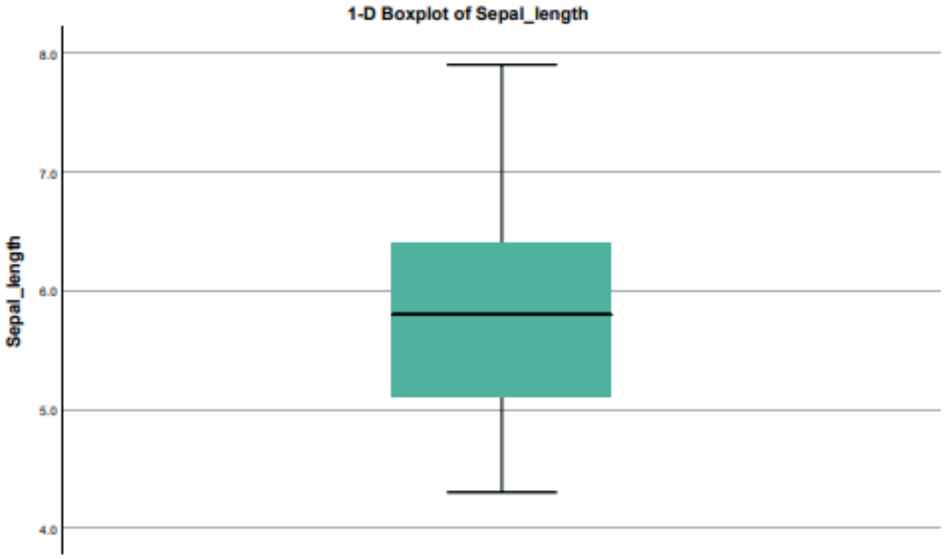


The mean of petal length is 3.759 and the standard deviation is 1.764. It is a Bimodal Distribution.



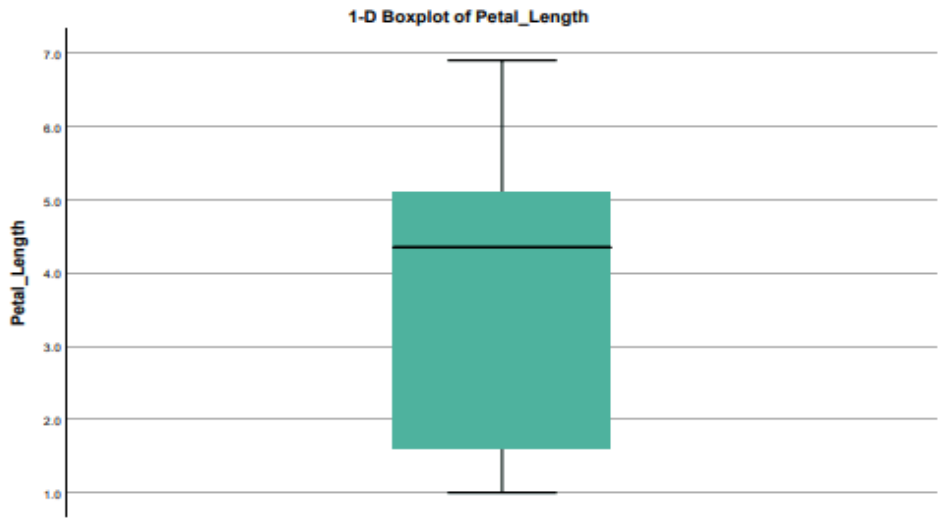
The mean of petal width is 1.199, and the standard deviation is 0. 7632.The distribution is random and has many peaks.

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



There are no outliers.

e. Repeat d. for the petal length.



There are no outliers

Problem 4 (5 points): The following paper presented at the ACM KDD 2017 Workshop on Machine Learning Meets Fashionshowcases an interesting application of data science to fashion and social media: "Identifying Fashion Accounts in Social Networks" by Doris Jung-Lin Lee, Jinda Han, Dana Chambourova, and Ranjitha Kumar:
https://kddfashion2017.mybluemix.net/final_submissions/ML4Fashion_paper_21.pdf

Read the paper and briefly answer the following questions:

1. What was the data used for the study? Include descriptions on the type of data and the size of the data.

   A dataset used for this study consists of 10,230 twitter accounts. Crowdsourcing technique is used to label each account. The accounts are grouped into fashion and not fashion groups as labels. Out of 10,230 accounts, 2734 accounts were labeled as "fashion" and 5,500 "non-fashion"

2. Was the data preprocessed or cleaned before applying any modeling techniques?

   No, the data is not preprocessed or cleaned before applying any modeling techniques

3. Did the authors solve a classification, a prediction, or a clustering problem as part of the pattern discovery stage? Justify your answer.
   The author solved a classification problem as part of pattern discovery stage as the study is trying to classify whether a twitter account is fashion related or not with the labels fashion related and not fashion related. The algorithms used in this paper are support-vector machines and Naïve Bayes for classification problems.

4. For the problem identified, which algorithm(s) the authors use to solve that problem?

   The author uses support-vector machine (SVM) and Naïve Bayes algorithm to solve the classification problem.