# DATA WAREHOSING & DATA MINING

**SEMESTER:** Fall 2021-2022

# FINAL-TERM ASSIGNMENT

## DATA MINING WITH WEKA


**SUBMITTED BY:**

**STUDENT NAME**: Sultana, Zafrin

**STUDENT ID:** 19-39345-1

**SECTION: D**

**DEPARTMENT:** CSE


**SUBMITTED TO:**

**COURSE TEACHER: TOHEDUL ISLAM**

# Table of contents Page

## INTRODUCTION

Data mining is the process of uncovering patterns and other valuable information from large data sets. It is also known as knowledge discovery in data (KDD). Data mining is used in many areas of research and business, including healthcare, education, sales and marketing, product development etc. It is a computer science and statistics multidisciplinary topic with the general purpose of extracting information from a data collection and transforming the information into an accessible structure for subsequent use. KNN, Naive Bayes, and Decision Tree are some of the classification algorithms used in data mining. I have chosen "Drug classification dataset" to classify the drug type by using three different classifier and find the best suited classifier for the dataset. Another part of this dataset's classifications job is to predict which drug might be appropriate for a future patient with same symptoms [1].

**Information about the dataset:** In this report, the used "Drug classification dataset", a CSV dataset file, collected from Kaggle.com which was used to predict the outcome of the drugs that might be accurate for the patient according to their health condition [2].

The targeted feature is:

- Drug type

The other feature sets are:

- Age
- Sex
- Blood Pressure Level (BP)
- Cholesterol Levels
- Na (Sodium) to K (Potassium) Ratio

**About the attribute:** The dataset contains 5 attribute and 1 class attribute which is the targeted feature to predict. This class attribute refers the type of drug to consume.

| Attribute | Representation in dataset | Data type |
|---|---|---|
| 1. Age (Age of patient) | Years | Numeric |
| 2. Sex (Sex of the patient) | M: Male; F: Female | Categorical (nominal) |
| 3. BP (Blood pressure of patient) | HIGH, NORMAL, LOW | Categorical (nominal) |
| 4. Cholesterol (Cholesterol levels) | HIGH, NORMAL | Categorical (nominal) |
| 5. Na_to_K (Sodium to Potassium Ratio in patients' blood) | Numeric value | Numeric (ratio-scaled) |
| 6. Drug (type of drug: Class attribute) | drugA, drugB, drugC, drugX, drugY | Class (nominal) |

There is total 200 instances of these 6 attributes and all these instances were used for classification. Here are the graphical details of the attributes:

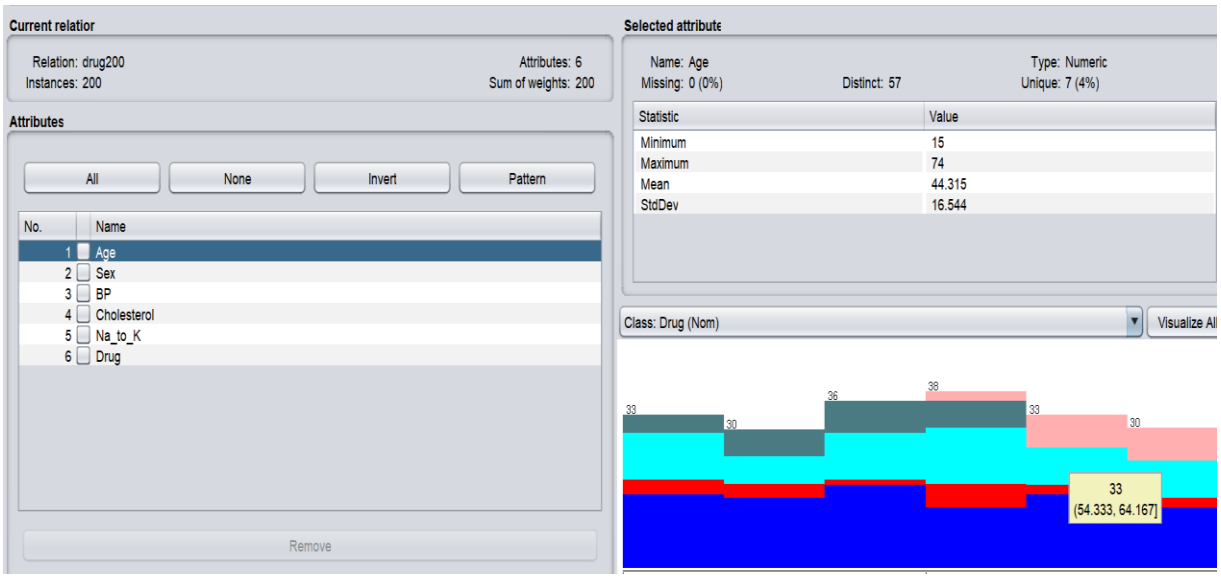


Figure 1: Selected dataset

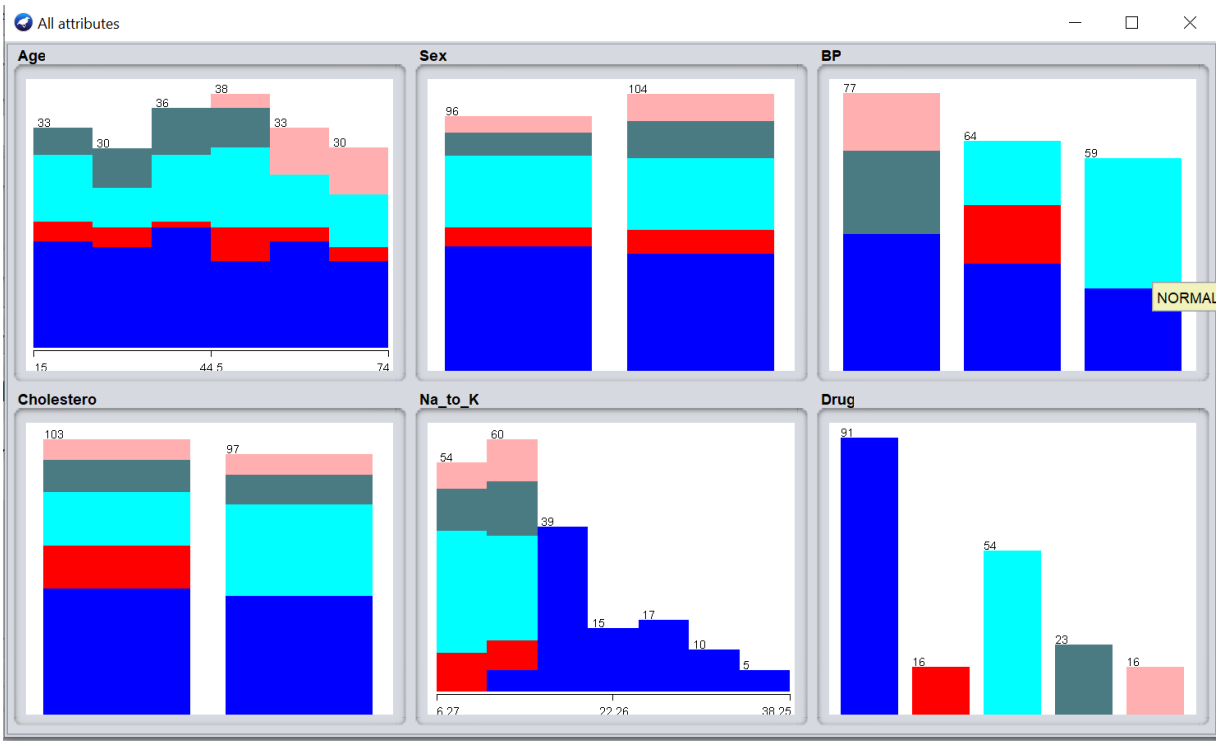


Figure 2: Details of all attribute

**Classifier:** A classifier is a machine learning model that is used to discriminate different objects based on certain features. Three kinds of classification have been used with same data to compare the result. In this process, Naïve Bayes, K-nearest Neighbour and Decision Tree classifier were used.

## RESULT OF THE CLASSIFIERS

Weka 3.8.5 version software was used to construct the classifier. Weka is a collection of machine learning algorithms for data mining tasks. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization.

**Applying naïve Bayes classifier:** Naïve Bayes algorithm is a supervised learning algorithm, which is based on **Bayes theorem** and used for solving classification problems. It is mainly used in *text classification* that includes a high-dimensional training dataset. It is a probabilistic classifier, which means it predicts on the basis of the probability of an object. While classifying the selected dataset, NaiveBayes format was selected from bayes folder.

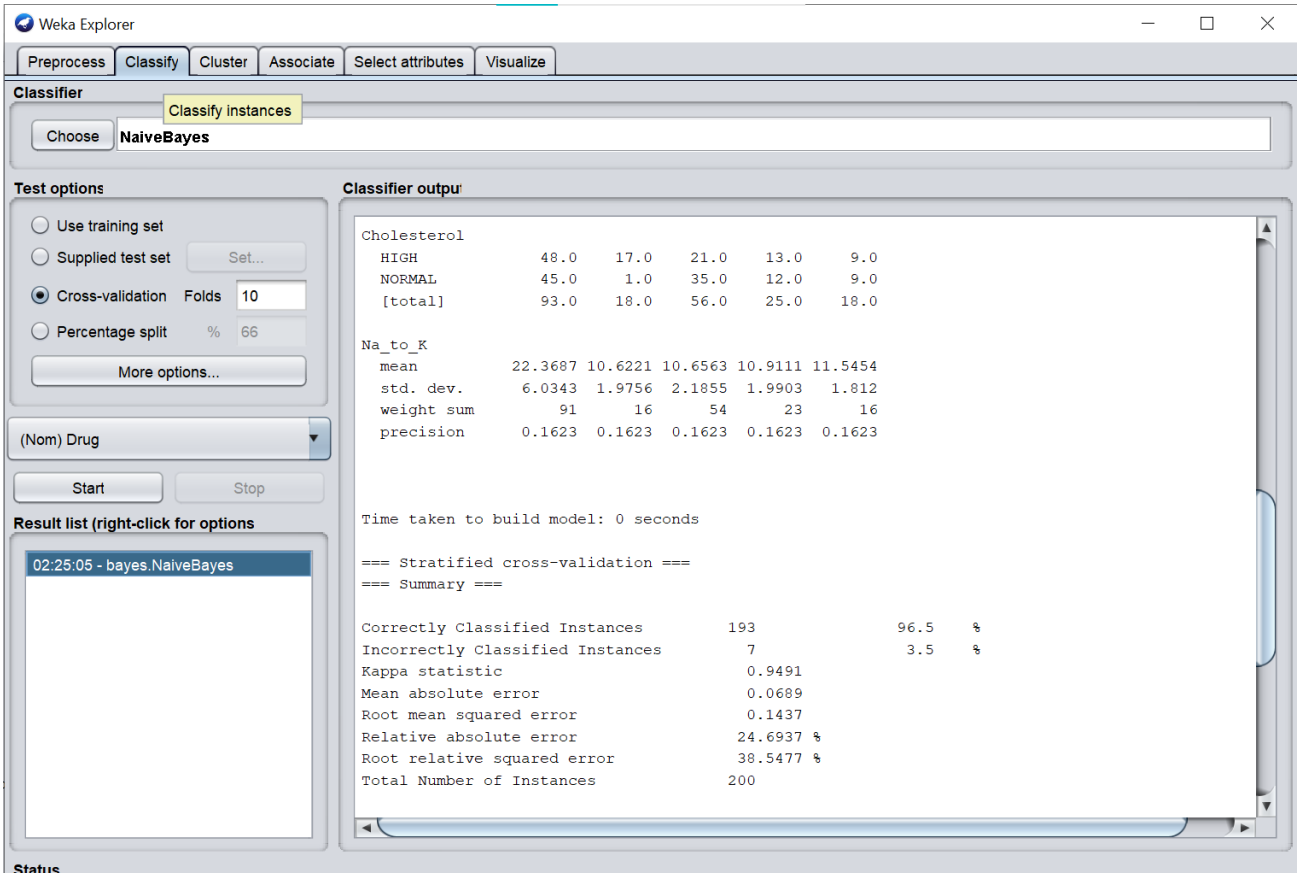


Figure 3: Naïve bayes classification

**Applying K-nearest Neighbour classifier:** K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique. KNN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. While classifying the selected dataset, IBk format was selected for KNN classification.

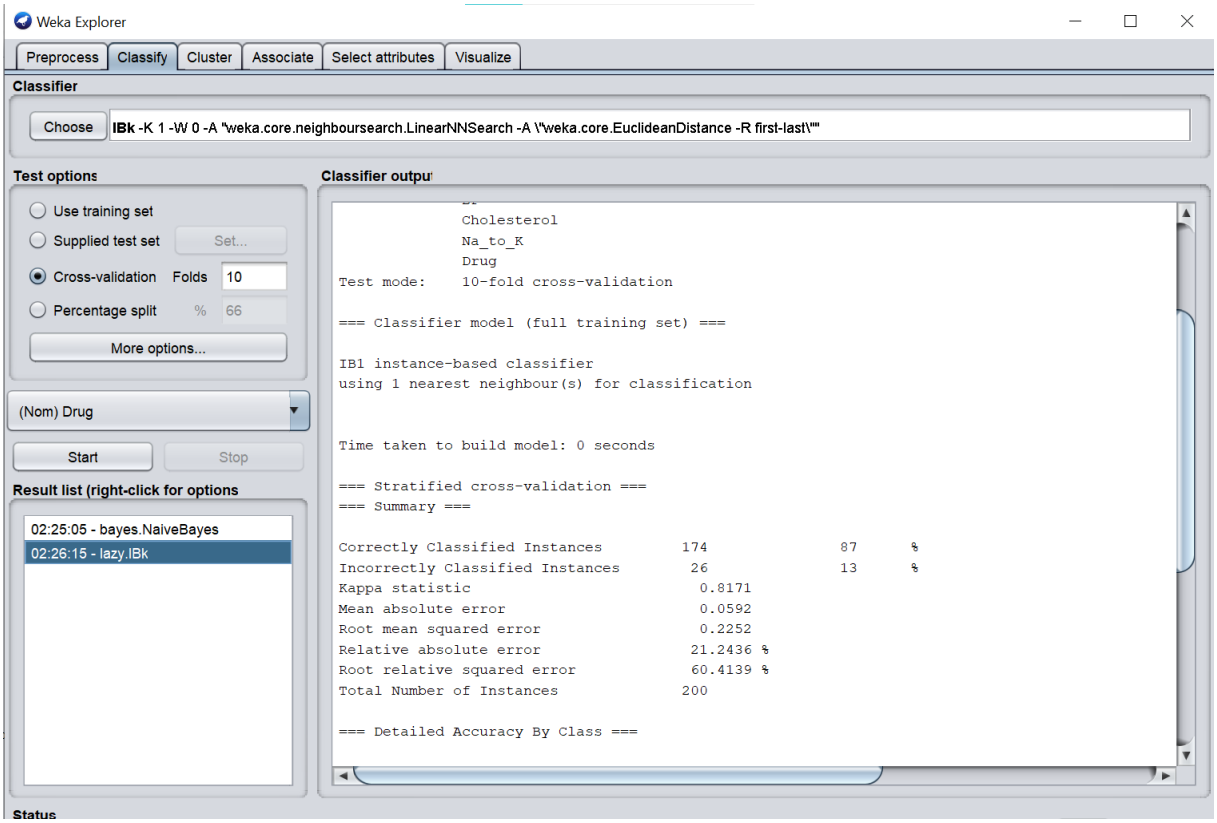


Figure 4: KNN classification

**Applying decision tree classifier:** Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. While classifying the selected dataset, J48 format was selected for Decision tree classification.

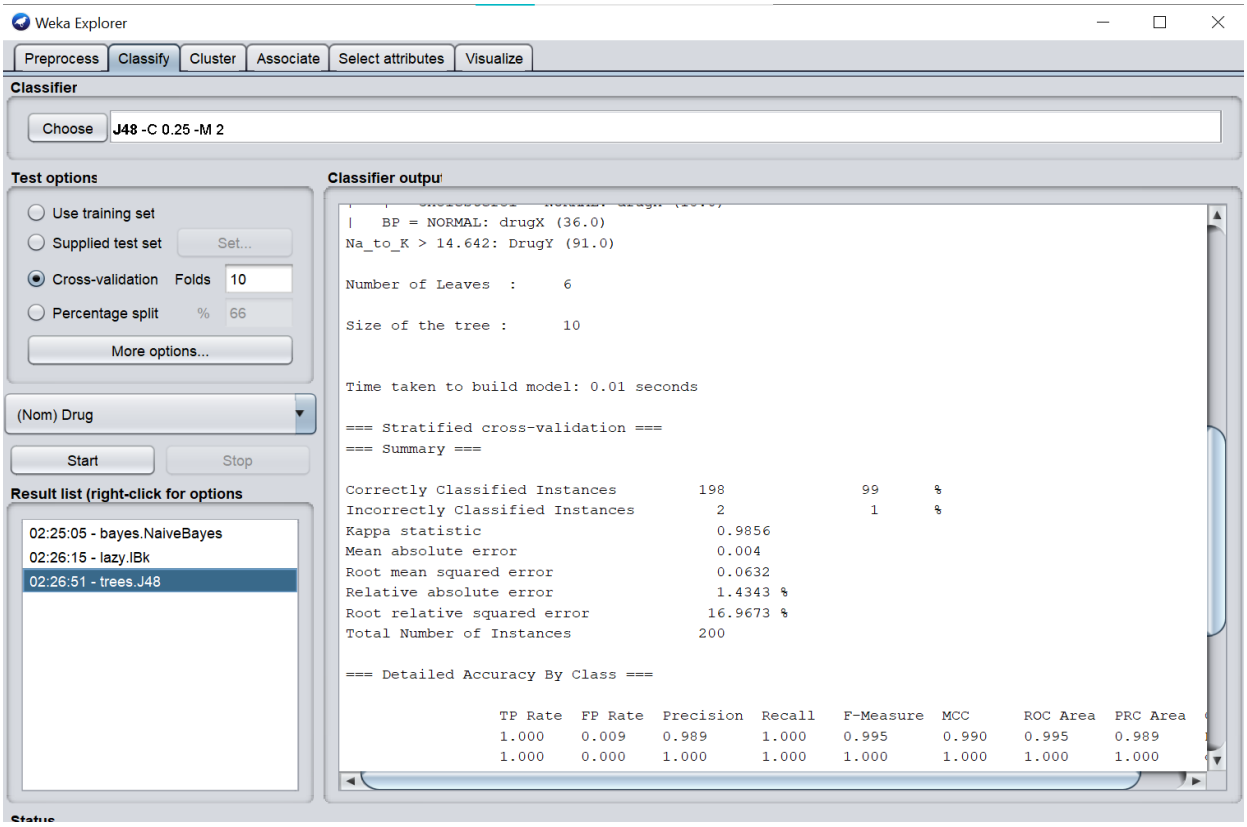

Figure 5: Decision tree classification

After applying three types of classifiers, the highest percentage of correctly classified instances is for Decision tree classifier with 99%. After that comes naïve bayes classifier with 96.5% and then KNN with 87% which is the lowest. The decision tree classifier is considered the best classifier for the dataset.

**Reason to choose Decision Tree classifier:** From the obtained result of the three classifiers, it is clearly seen that decision tree has the highest percentage of correctly classified instances which is 99%. As it has most accurate value so it would be more suitable classifier for the dataset. One of the advantages of decision trees is that their outputs are easy to read and comprehend without requiring statistical knowledge. Compared to other decision procedures, decision trees require less data preparation. Decision trees are more flexible and easier. Moreover, Decision tree is faster than KNN due to KNN's expensive real time execution. Another significant advantage of a decision tree is that it forces the consideration of all outcomes of a decision and traces each path to a conclusion. It creates a comprehensive analysis of the consequences along each branch and identifies decision nodes that need further analysis. This boost predictive models with accuracy, ease in interpretation, and stability.

Here is the summary of the Decision tree classifiers result:

- Correctly Classified Instances : 198 99 %
- Incorrectly Classified Instances : 2 1 %
- Kappa statistic : 0.9856

- Mean absolute error : 0.004
- Root mean squared error : 0.0632
- Relative absolute error : 1.4343 %
- Root relative squared error : 16.9673 %
- Total Number of Instances : 200

## PREPARING TEST-DATASET

One of the most important mechanisms in machine learning is to train your algorithm on a training set that is separate and distinct from the test set for which will be gauged its accuracy. To detect a machine learning behavior, a training dataset has been set which was extracted subset from the referred dataset. Then this model had been tested with a test dataset which is a subset to test the trained model. While preparing the test dataset, things that were made sure are that the dataset was large enough to yield statistically meaningful results. Also, it was representative of the data set as a whole. In other words, test set with unusual characteristics than the training set were not chosen. The suitable classifier is then used to predict the classification for the instances in the test set.

If the test set contains **N** instances of which **C** are correctly classified, **C** are correctly classified Predictive accuracy, **P = C/N**. There are 20 instances in this prepared test dataset.[3]

| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | Age | Sex | BP | Cholester | Na_to_K | Drug | |
| 2 | 18 | F | HIGH | HIGH | 37.188 | drugC | |
| 3 | 54 | M | NORMAL | HIGH | 24.658 | DrugY | |
| 4 | 41 | F | LOW | NORMAL | 18.739 | DrugY | |
| 5 | 23 | M | NORMAL | HIGH | 31.686 | DrugY | |
| 6 | 29 | F | HIGH | HIGH | 29.45 | drugB | |
| 7 | 36 | F | NORMAL | HIGH | 16.753 | DrugY | |
| 8 | 51 | M | HIGH | HIGH | 18.295 | DrugY | |
| 9 | 42 | F | LOW | NORMAL | 29.271 | DrugY | |
| 10 | 61 | F | LOW | HIGH | 18.043 | DrugY | |
| 11 | 47 | M | LOW | HIGH | 10.114 | drugC | |
| 12 | 56 | F | LOW | HIGH | 11.567 | drugC | |
| 13 | 49 | F | NORMAL | NORMAL | 9.381 | drugX | |
| 14 | 69 | F | NORMAL | HIGH | 10.065 | drugX | |
| 15 | 45 | M | LOW | NORMAL | 8.37 | drugX | |
| 16 | 45 | M | LOW | NORMAL | 10.017 | drugX | |
| 17 | 24 | F | NORMAL | HIGH | 10.605 | drugX | |
| 18 | 45 | F | HIGH | HIGH | 12.854 | drugA | |
| 19 | 31 | M | HIGH | NORMAL | 11.871 | drugA | |
| 20 | 60 | M | HIGH | NORMAL | 8.621 | drugB | |
| 21 | 53 | F | HIGH | NORMAL | 12.495 | drugB | |
| 22 | | | | | | | |

## PROCEDURE OF TESTING THE TEST DATASET

1. First, Weka 3.8.5 was opened and ‘Explorer’ option was chosen and such window named weka explorer was opened.

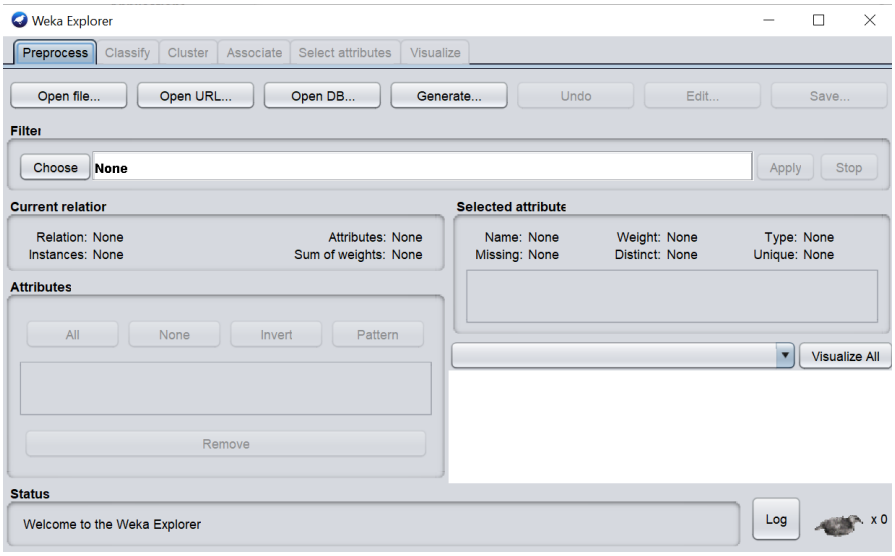


Figure 6: weka explorer

2. Then, the open file option was selected and the extracted CSV file , training dataset was selected from the device.

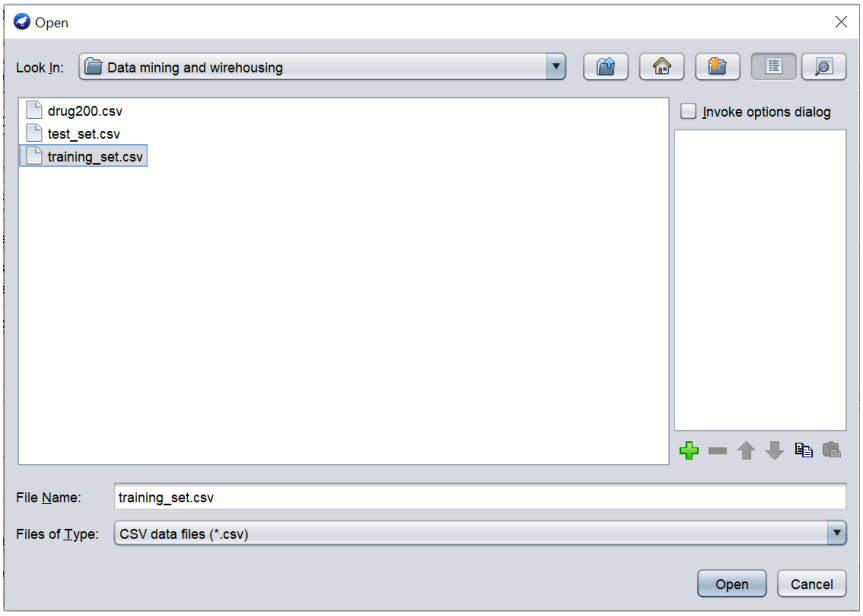


Figure 7: training data select

3. After the open option was clicked, the details of the dataset popped.

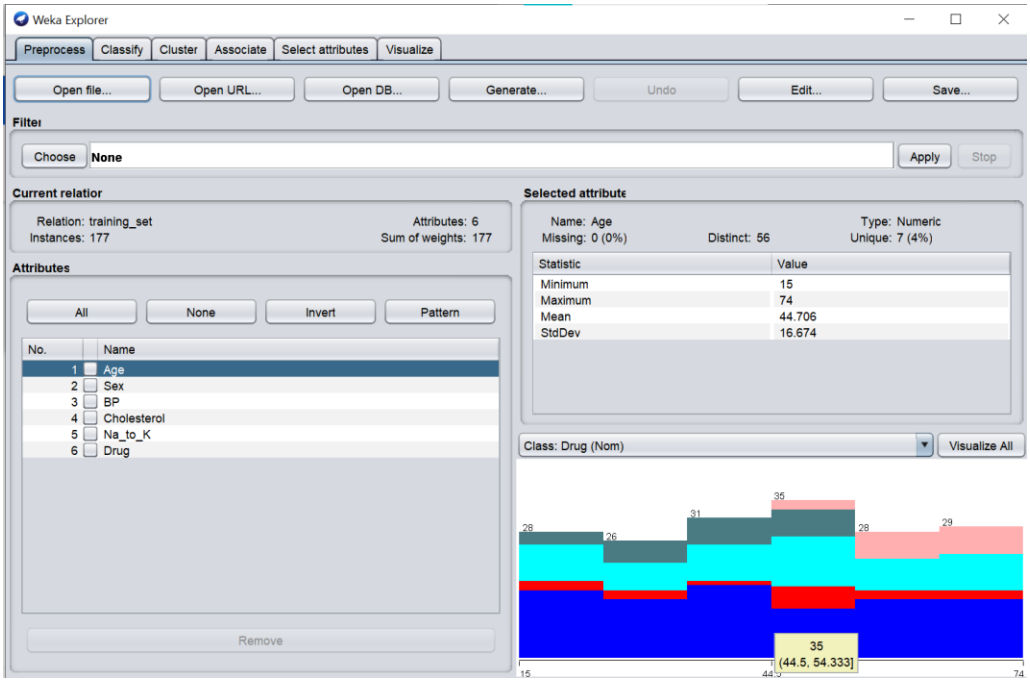


Figure 8: training dataset

4. Then the preferred classifier (Decision tree classifier) for the training dataset was selected. Then from the test options, use training set was chosen and the start option was selected.

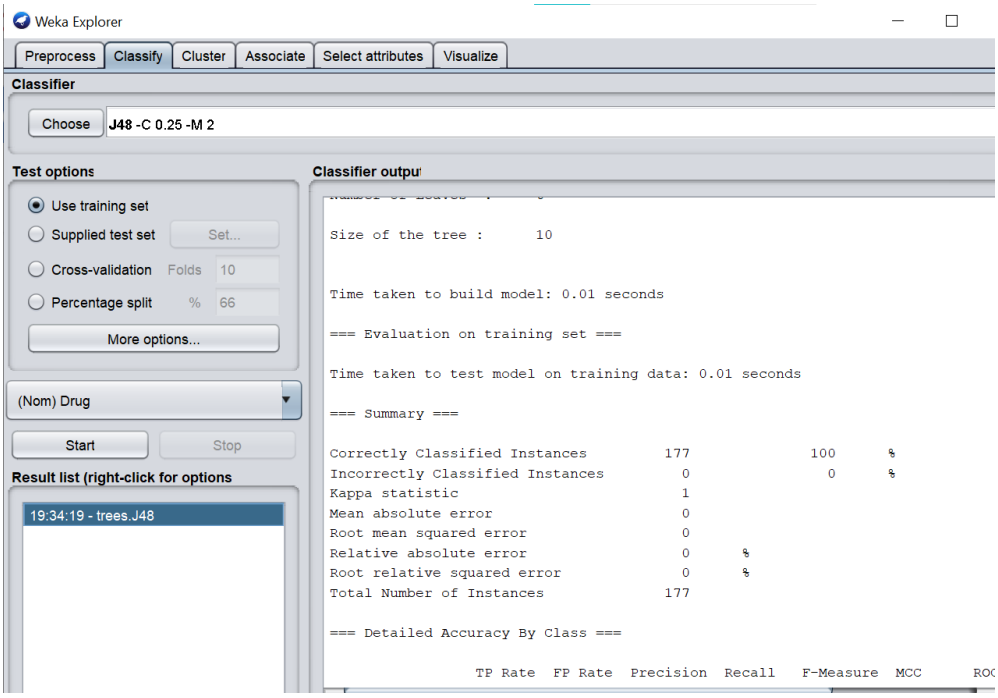


Figure 9: result of training set

5. To input the test set, from the test options, supplied test data was selected and then a window named test instances came.

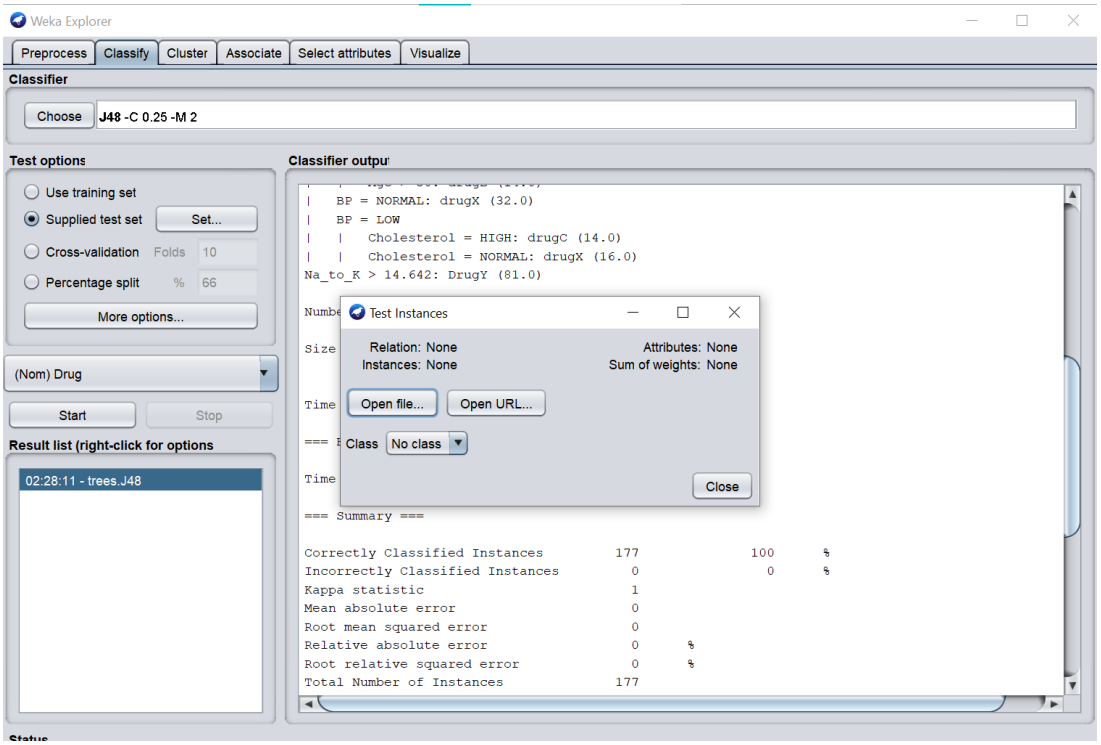


Figure 10: supplied test set selected

6. Then the open file option was chosen to insert the test data and then the dataset was opened.

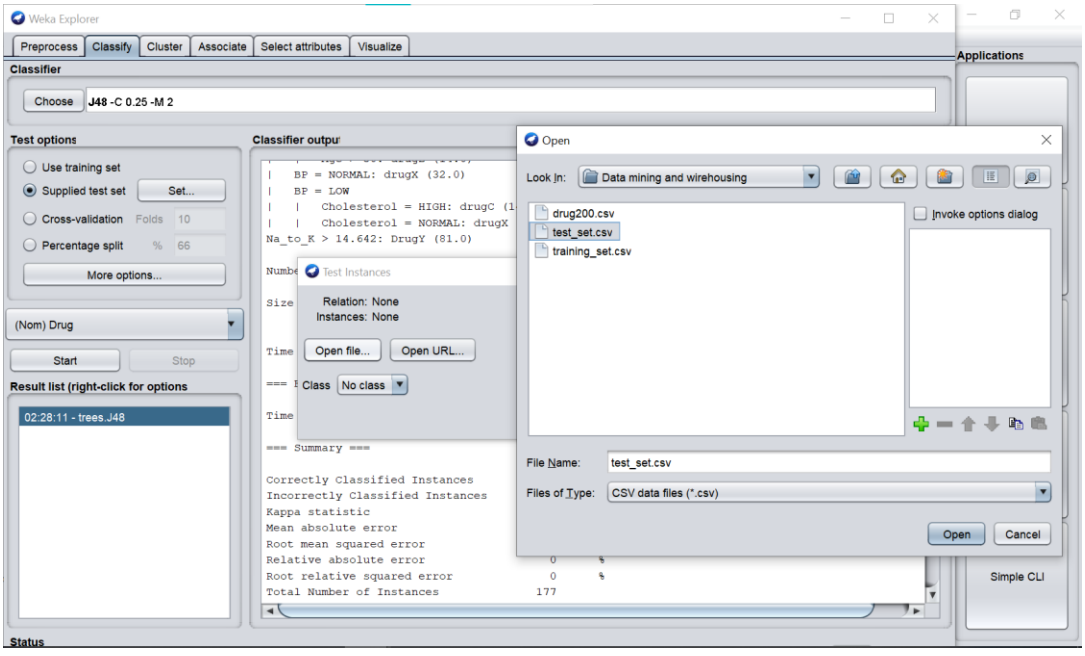


Figure 11: test set opened

7. To make sure the test set works properly, it was made sure that the output predictions was in PlainTest form.

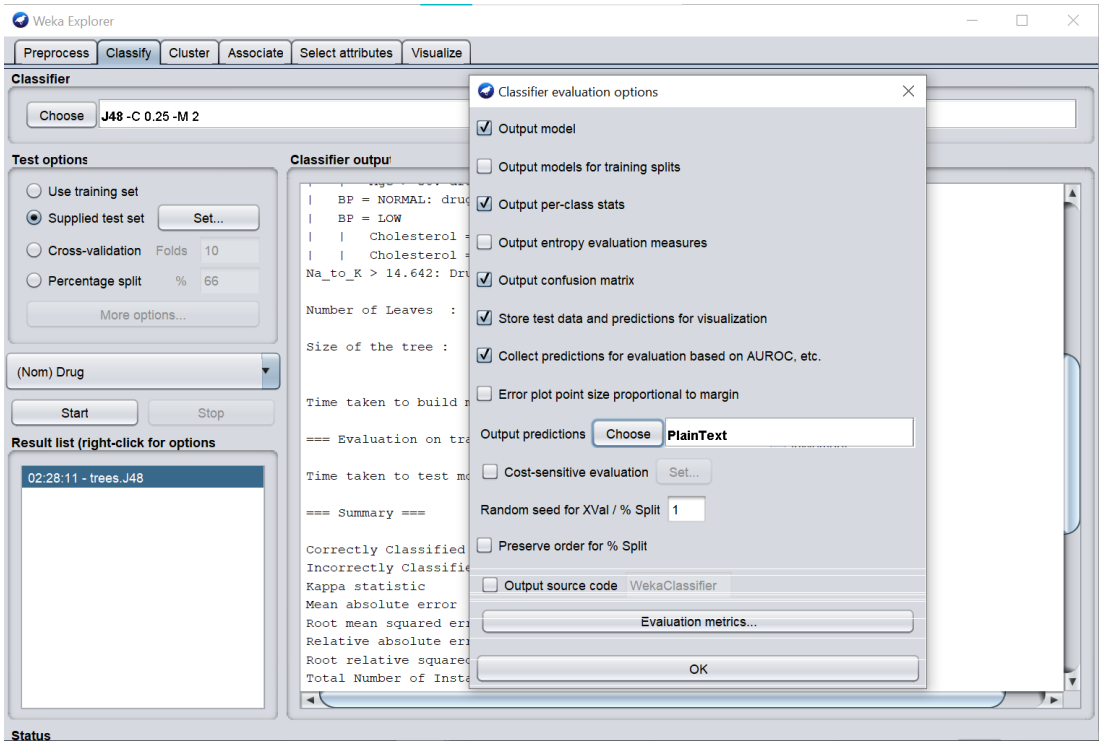


Figure 12: PlainText format chosen for output prediction

8. Then the start button was clicked.

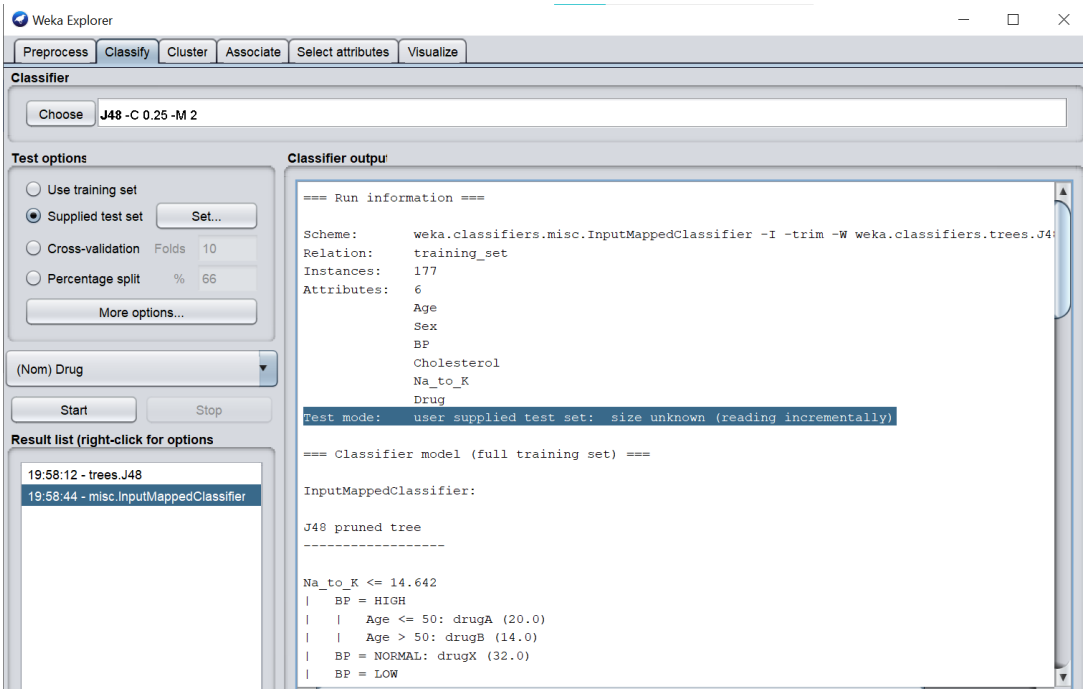


Figure 13: test set run

## RESULT OF TEST-DATASET MODEL

Once the start button was clicked, the output for the test dataset came. In the result, the test mode was ‘user supplied test set’ it means the classifier is evaluated on how well it predicts the class of a set of instances loaded from a file which was inputted by the user. The total time taken to build the model was 0.01 seconds and among the 20 instances, there were 2 instances where error was found. Which means in the test model, those two instances were not properly classified and the Machine learning model predicted it after finding the error.

In this model, Correctly classified instances % Value describes the amount of accuracy correctly classified instances provides by the algorithm. In this case, the percentage is 90% which is quite good.

Incorrectly classified instances % Value describes how much incorrect instances are given by the algorithm. In this case, the percentage is 10%.

Mean Absolute Error (MAE): It can define as statistical measure of how far an estimate from actual values i.e. the average of the absolute magnitude of the individual errors. It is usually similar in magnitude but slightly smaller than the root mean squared error. In this model the MAE is 0.04

Root Mean-Squared Error (RMSE): The Root Mean Square Error (RMSE) calculates the differences between values predicted by a model / an estimator and the values observed from the thing being modelled/ estimated. RMSE is used to measure the accuracy. It is ideal if it is small. In this case the RMSE is 0.2 which is ideal.

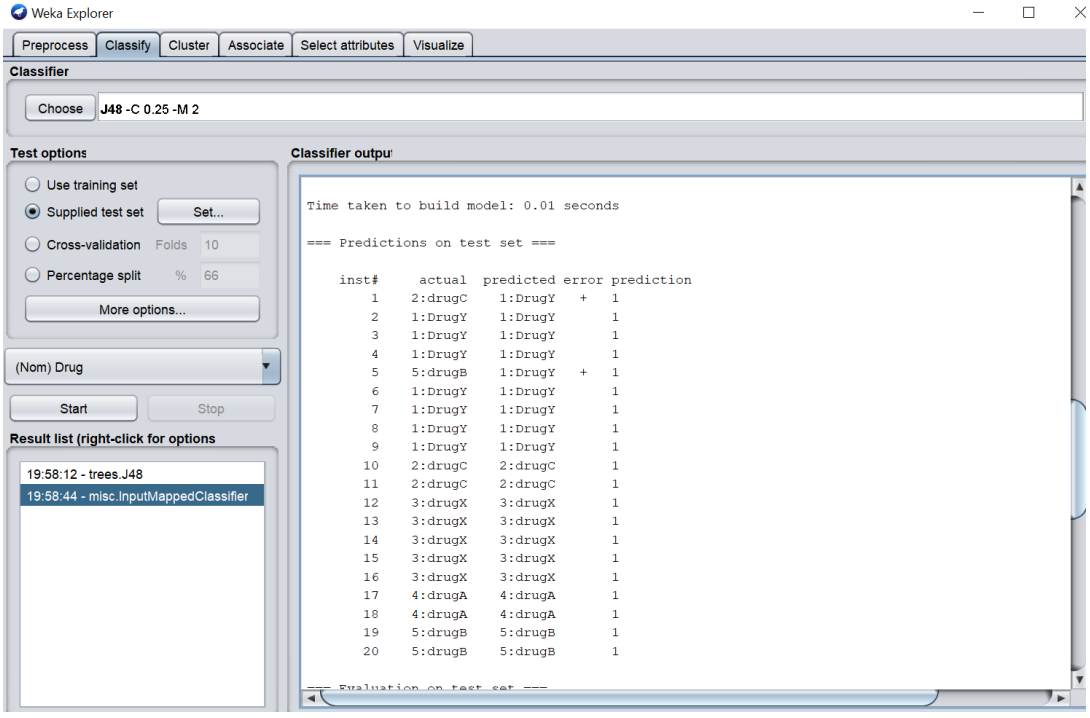


Figure 14: Prediction result of test set

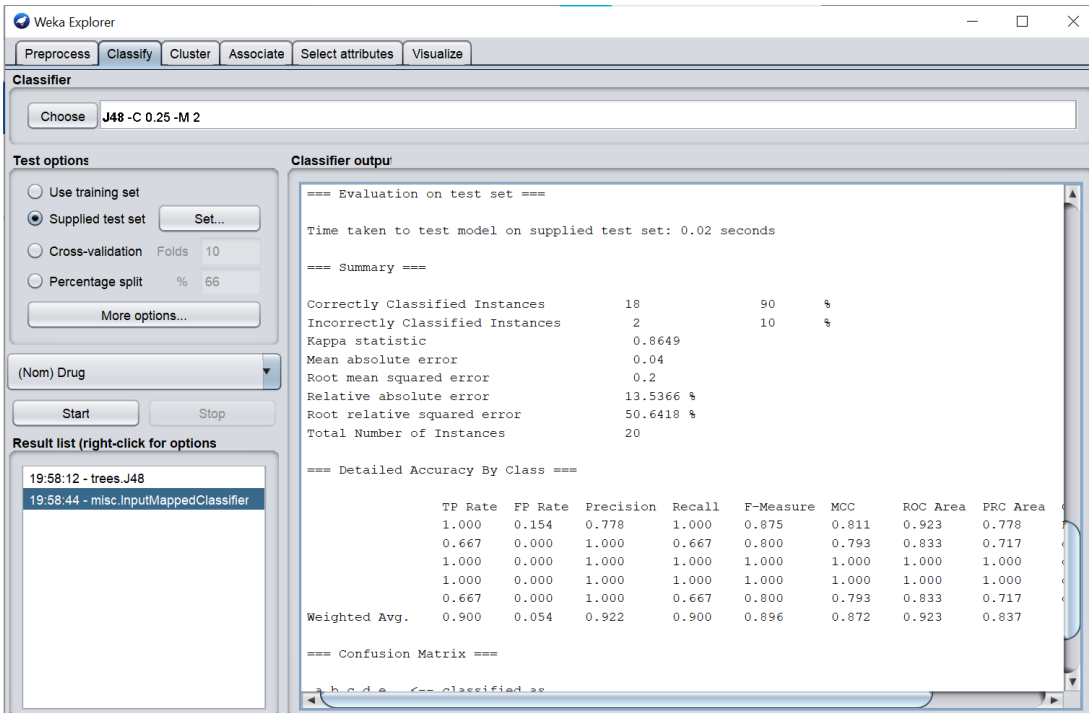


Figure 15: Result of accuracy

## DISCUSSION

The purpose of this report was to find a suitable classifier for the Drug classification dataset which will classify the drug as accurately as possible and will be able to predict the class from test dataset. After applying three different classifier which are KNN, naïve Bayes and decision tree, the best-chosen classifier for the dataset is decision tree classifier with 99% accuracy. Then a training set, extracted from original dataset was selected to prepare a Machine Learning Model. A prepared test dataset was used to test the model and finally the Models accuracy was 90% for the prepared test dataset. Creating training and testing dataset is an important concept in data science as it is used to improve generalization and minimize overfitting. This also helps to give an unbiased evaluation about the accuracy of the model itself.

## REFERENCES


1. About datamining, https://en.wikipedia.org/wiki/Data_mining
2. Drug classifier Dataset, https://www.kaggle.com/prathamtripathi/drug-classification
3. Training and test dataset, https://github.com/Zafrin-Jolly/Data-mining--Dataset