****

**AMERICAN INTERNATIONAL UNIVERSITY-BANGLADESH**

**Data Warehouse and Datamining**

**Section: A**

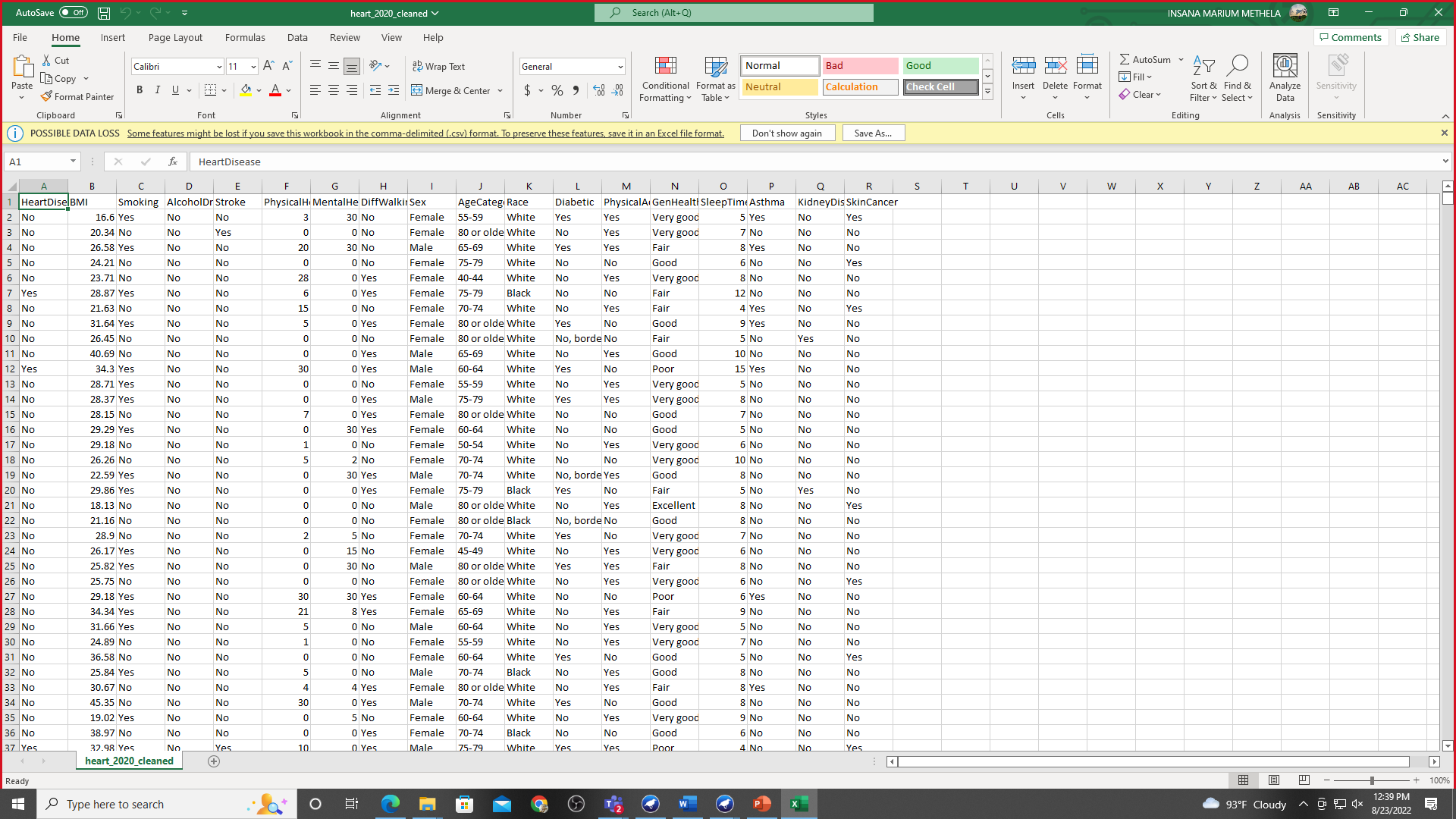
**Dutta, Anindya(18-37634-1)**

**Introduction:**

This project's objective is to predict the causes of heart patients die. Heart disease data is a dataset that we have used in our project to apply classification. The dataset covers information about the factors that have the greatest impact on heart disease. For supervised learning we use heart disease data & unsupervised we use date fruit dataset. The dataset is collected from [www.kaggle.com](http://www.kaggle.com).

Dataset:

The dataset’s first 37 instances (out of 398) are shown as a sample in the figure. For the full dataset please go to <https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease> .



**Figure: Personal Key Indicators of Heart Disease**

Attributes:

Heart Disease= (yes/no)

BMI= (16.6-58.54)

Smoking= (yes/no)

Alcohol Drinking= (yes/no)

Stroke= (yes/no)

Physical Health= (0-30)

Mental Health = (0-30)

DeffWalking= (yes/no)

Sex= (Male/Female)

Age Category=ALL

Race= (White, Black, Asian, American Indian /Alaskan Native)

Diabetic = (yes/no, borderline diabetes/yes (during pregnancy))

Physical activity = (yes/no)

GenHealth= (Very good, Good, Fair, Excellent, Poor)

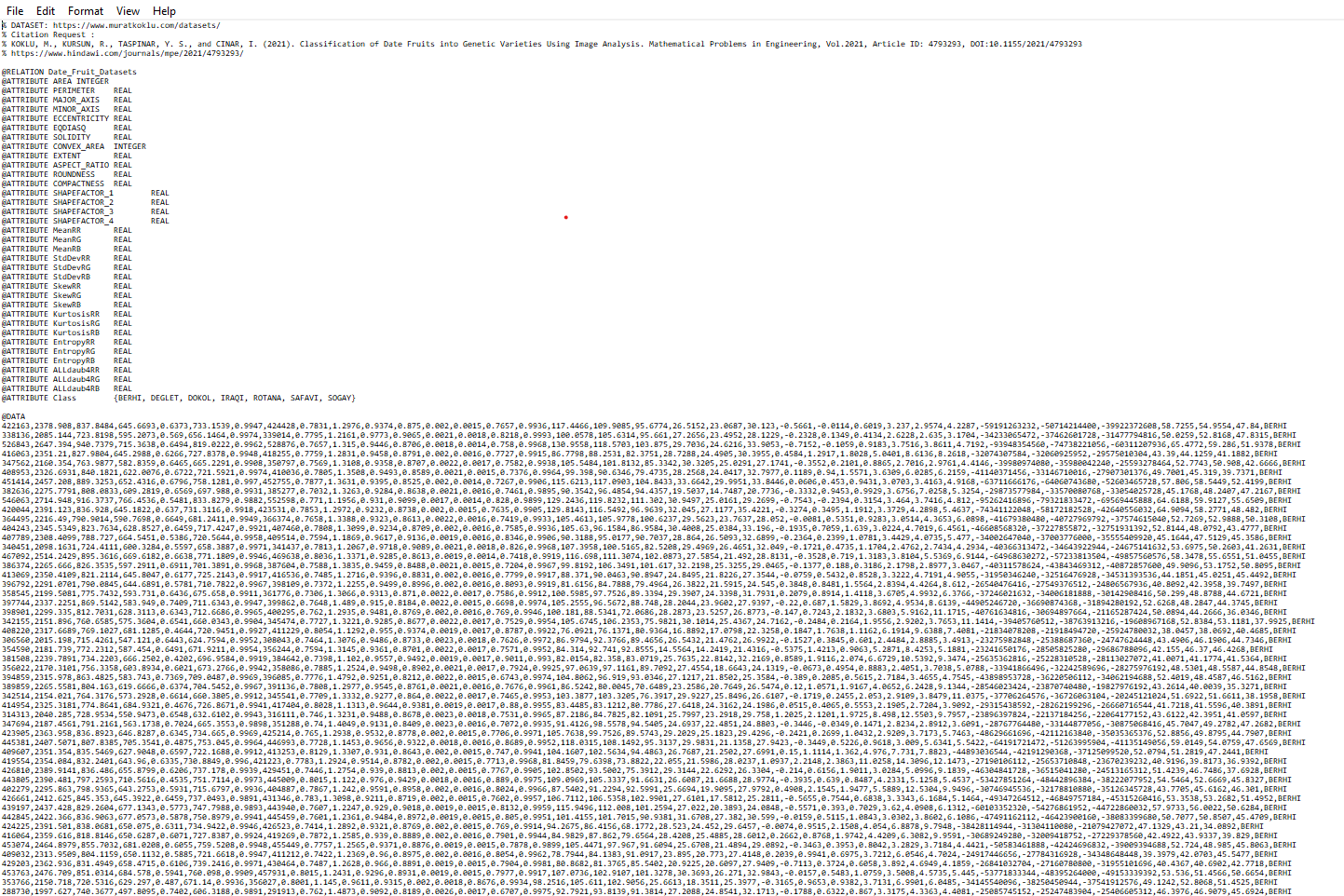
Sleep Time= (2-15)

Asthma= (yes/no)

Kidney Disease= (yes/no)

Skin Cancer= (yes/no)

We also collect another unsupervised dataset for task three. We collected this dataset from <https://www.kaggle.com/datasets/muratkokludataset/date-fruit-datasets>



Procedure:

The following step will be to select categorize from Weka. Weka is a software package that includes visualization tools and algorithms for data analysis and predictive modeling. We will utilize Weka to analyze data for this task.

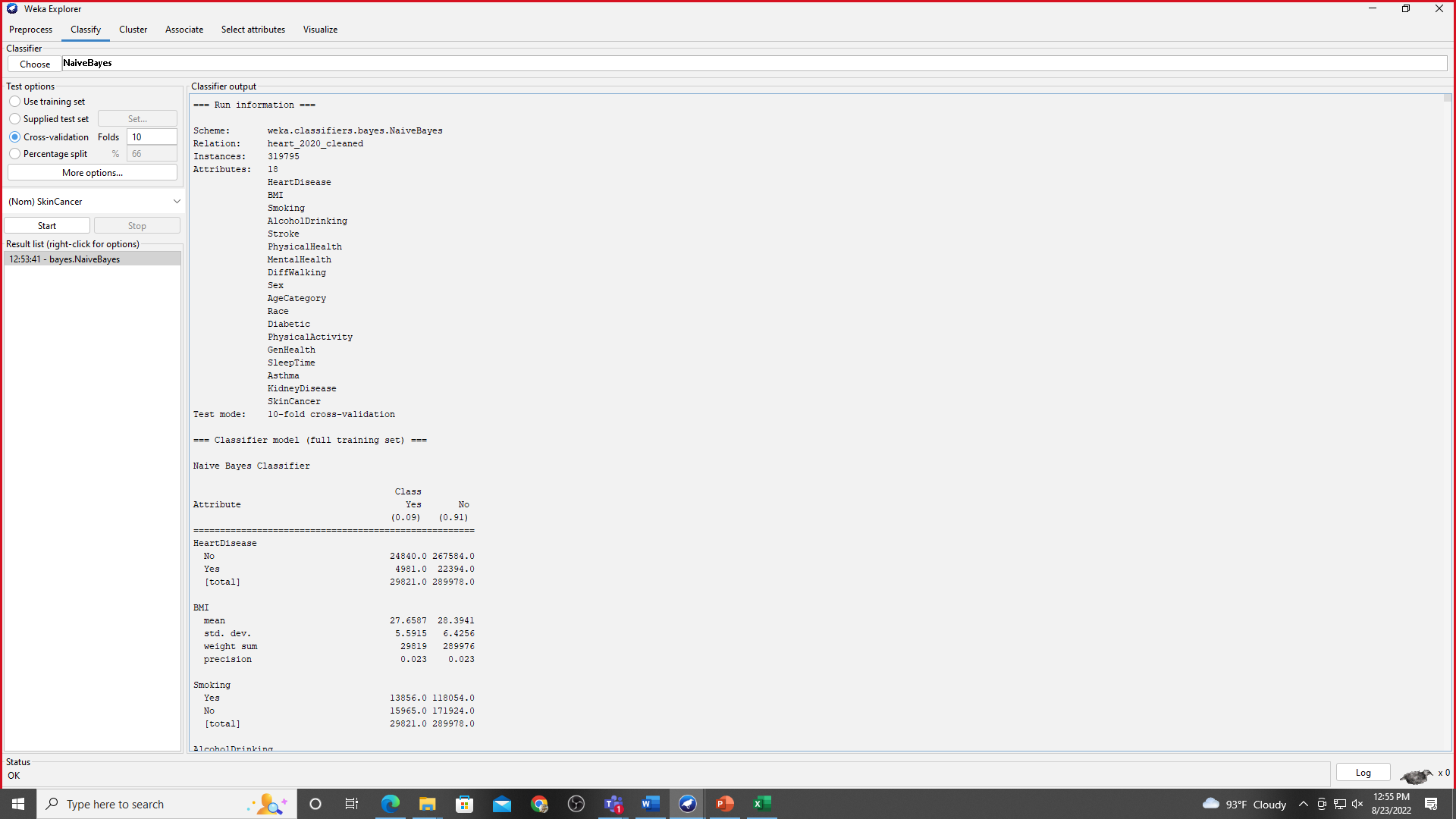
We have chosen 2 classifications from the WEKA classifiers list. Those are NaiveBayes and J48.

**Task1:**

**NaiveBayes:**

It is a classification technique based on Bayes’ Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. NaiveBayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods. We must follow the following steps given below to apply NaiveBayes classification method.

**Step-1 Select the Naïve Bayes Classifier**



**Figure: Appling Naïve bayes**

Text

Description automatically generated with medium confidence

**Figure: Appling Naïve bayes**

**2.Decision Tree-(J48)**

Decision Tree is the classification technique that consists of three components root node, branch (edge or link), and leaf node. Root represents the test condition for different attributes, the branch represents all possible outcomes that can be there in the test, and leaf nodes contain the label of the class to which it belongs. The root node is at the starting of the tree which is also called the top of the tree. J48 is an algorithm to generate a decision tree that is generated by C4.5 (an extension of ID3). It is also known as a statistical classifier. Now we have to do the following steps to apply the j48 decision tree.

**Step-1: Select J48 Classifier**

Graphical user interface, text, application, Word

Description automatically generated

**Figure: Applying j48 Classifiers**

Graphical user interface, text

Description automatically generated with medium confidence

**Figure: Applying j48 classifiers**

**Graphical user interface

Description automatically generated**

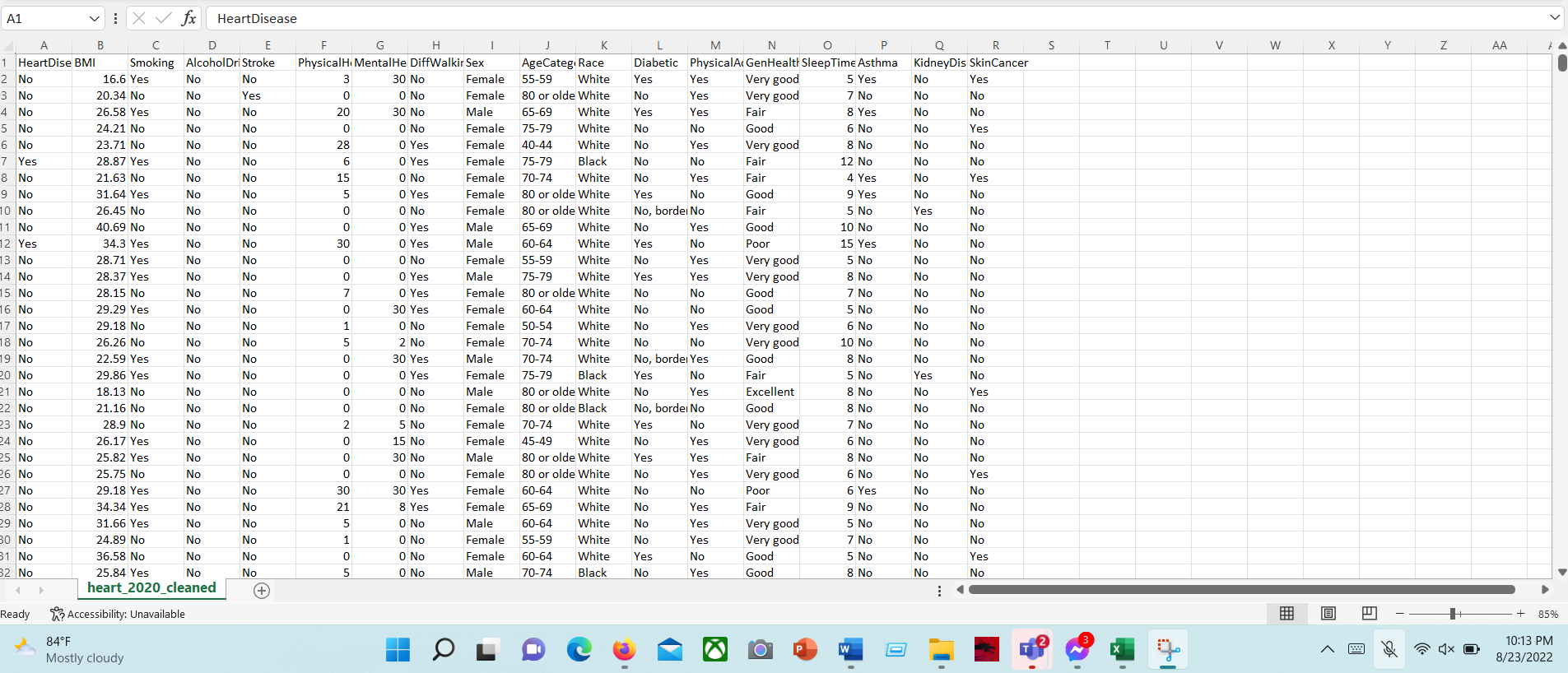
**Figure: Graph**

**Graphical user interface, chart, line chart

Description automatically generated**

**Figure: Graph**

**Task 2:**



**Figure: Tested dataset**

Table

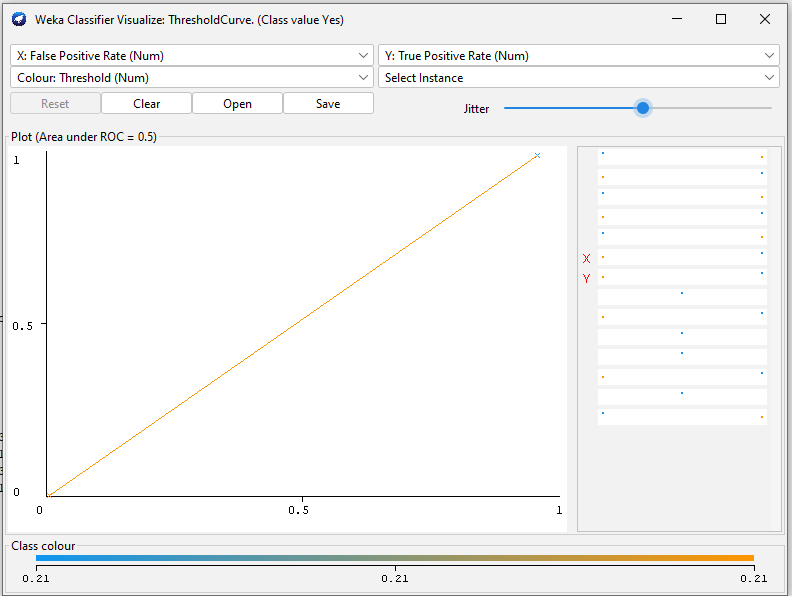
Description automatically generated

**Figure: Tested data to j48 Classifiers**

Graphical user interface

Description automatically generated

**Figure: j48 classifies**



**Figure: Graph**

**Task 3:**

https://www.kaggle.com/datasets/muratkokludataset/date-fruit-datasets

Text

Description automatically generated

**Figure: Appling k means clustering Algorithm**

**Chart, scatter chart

Description automatically generated**

**Figure: Graph**

**Result:**

**For task 1:**

When the Naïve Bayes classifier is used in our dataset, 304 instances are correctly identified and a rate of 76.3819% instances were found as correct. The other 94 instances which were erroneously categorized, the percentage of those incorrectly classified instances were 23.6181%. In J48 trees, 315 cases were successfully identified, with a 79.1457 % success rate and other 83 instances which were erroneously categorized, the percentage of those incorrectly classified instances were 20.8543%.  As a result, the accuracy of the J48 classifiers is higher. The number of right cases is significantly higher than the other two. So J48 is the best classifier. The best classifier is J48 with 2.7638% higher correctness than Naïve Bayes.

**For task 2:**

For test data set we select supervised learning data. For low percentage of error we select j48 classifier, In J48 trees, total 40 instances cases were successfully identified, with a 77.5 % success rate and other 9 instances which were erroneously categorized, the percentage of those incorrectly classified instances were 22.5. Here we can see in instance 1,10,14,17,26,28,38,39,30 actual value yes when we test this we get predicted value no.

**For task 3:**

Here two clusters are visible. They are 582 (65%) & 316(35%). Incorrectly clustered instances is 500 (55.67%).

**Discussion:**

In this paper here we use two classifier (Naïve Bayes & Decision tree j48) & k means clustering algorithm. We collected this data from Kaggle. For task 1 we use Personal Key Indicators of Heart Disease dataset. Task 1 we use naïve bayes and j48 and finding which one is most efficient. In naïve bayes classifiers we get 76.38% accuracy and decision tree j48 we get 79.14% accuracy. . So we can say decision tree j48 is most efficient .Task 2 we use same dataset for test then we apply j48 classifier and get 77.5% accuracy. For task 2 we use Date Fruit Datasets. For Clustering we apply k means Clustering algorithm.