**A close up of a sign

Description automatically generated**

**AMERICAN INTERNATIONAL**

**UNIVERSITY-BANGLADESH**

FINAL TERM PROJECT

Course Title: DATA WAREHOUSING AND DATA MINING Section: [C]

Semester: Spring 2021-2022

Course Teacher: TOHEDUL ISLAM

Date of Submission: 21-04-2022

Submitted by:

|  |  |  |
| --- | --- | --- |
| NO. | Name | ID |
| 1 | FAHAD MOLLAH | 18-36702-1 |
| 2 | MD. ARIFUL ISLAM | 18-36842-1 |
| 3 | AHSUN MAHMUD RIAD | 18-36814-1 |

**Contents**

[Introduction 2](#_Toc101402788)

[Result 3](#_Toc101402789)

[Naïve Bayes: 3](#_Toc101402790)

Classifiers Error  [4](#_Toc101402792)

[K- Nearest Neighbor Classification 4](#_Toc101402791)

[Test Data 5](#_Toc101402793)

[Naïve bayes (After Applying Test data) 6](#_Toc101402794)

[Discussion 8](#_Toc101402795)

# **Introduction**

“Salary Prediction” dataset is used as a supervised learning dataset for this project. We used the information in the dataset to predict whether a person makes over 50K a year. In this dataset, there are total 32561 number of instances. And number of attributes is 15 which are age, workclass, fnlwgt, education, education-num, marital-status, occupation, relation, race, sex, capital-gain, capital-loss, hours-per-week, native-country, salary. In these attributes, there are 6 numeric data and 8 categorical data. We used the Naïve Bayes, K-nearest neighbor to classify the data.

“Diabetes” Data set is used as unsupervised learning data set to apply K-means algorithm. There are 768 number of instances.

Whatever If required to find-out the data set then visit the given following links:

|  |  |
| --- | --- |
| Datasets | |
| Supervised | <https://www.kaggle.com/datasets/ayessa/salary-prediction-classification> |
| Unsupervised | <https://www.kaggle.com/code/yusatll/unsupervised-learning-with-diabetes-dataset> |

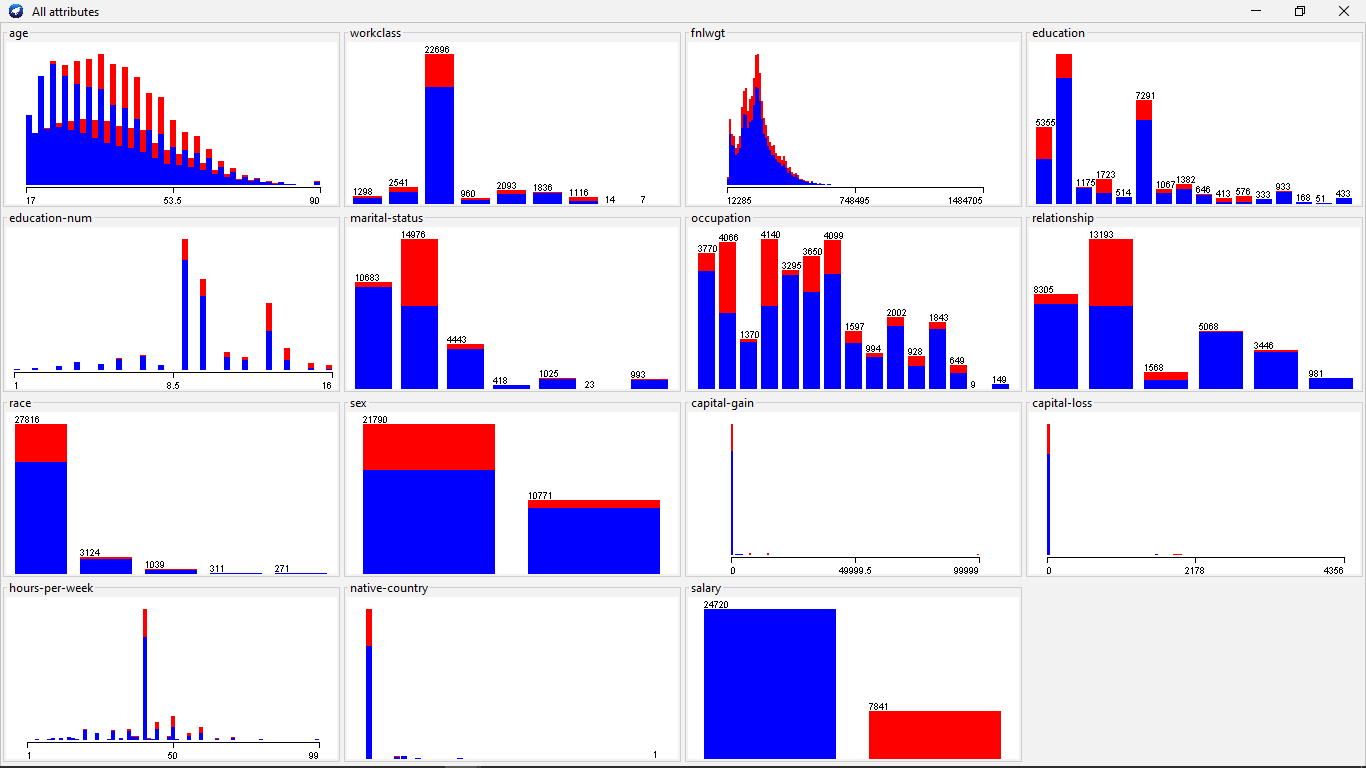


Figure: Dataset Visualization (Salary Prediction)

# **Result**

We applied the Percentage split approach. We divided the dataset into two parts: 60% and 40%. The first 60% is used to train the model, while the remaining 40% is used to test the model.

## **Naïve Bayes:**

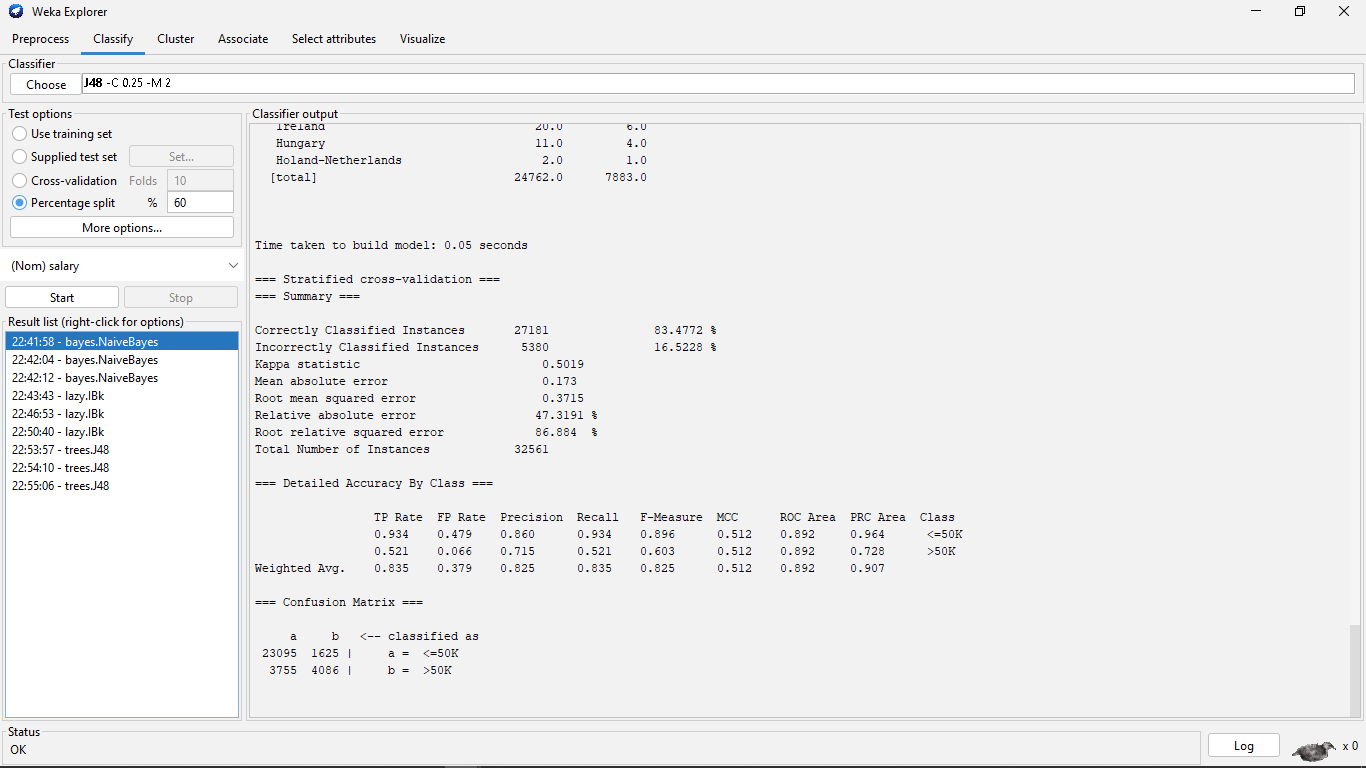
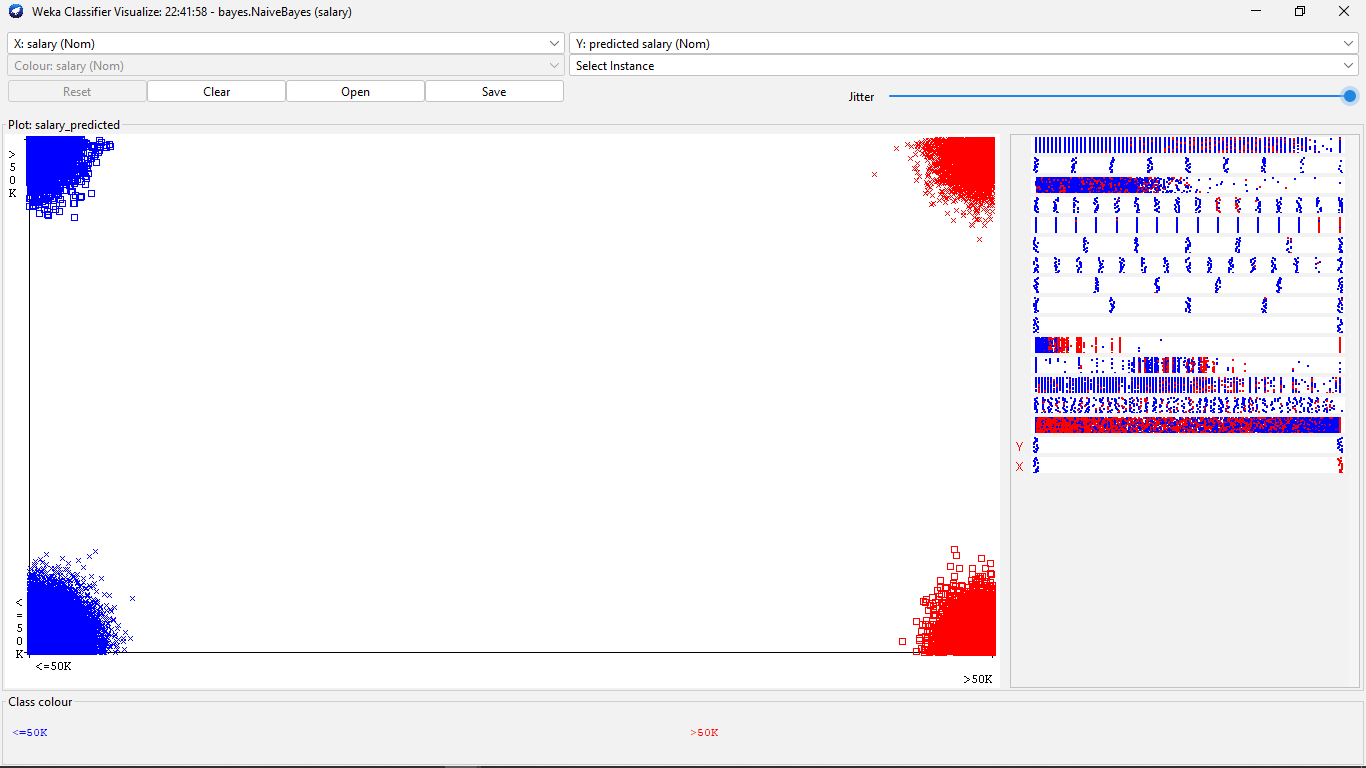


Figure: Naïve Bayes Classification

**Classifiers Error:**



The above figure shows the error rate of the models we applied. The blue color denotes the salary <-50,

and the red indicates salary >50k

## **K- Nearest Neighbor Classification**

## 

Figure: K- Nearest Neighbor Classification

We are applied two classifiers (Naïve bayes and KNN) from Weka on the selected data set. For naïve Bayes classifier, the value of the correctly classified instances is 27181 which was 83.47% of the dataset and incorrectly classified instances rate was 16.52 %.

On the other hand, when we applied KNN classification correctly classified instances 27090 whether incorrectly classified instances were 5471 which was 83.19% & 16.80% of the dataset. At the end, I decided to choose naïve Bayes for the best classifier. Because the rate of the highest correct classifier was 83.47%. That’s why we choose naïve bayes to complete the next step.

## **Test Data**

We have picked randomly 25% instances from the dataset as test data which were 8000 instances.

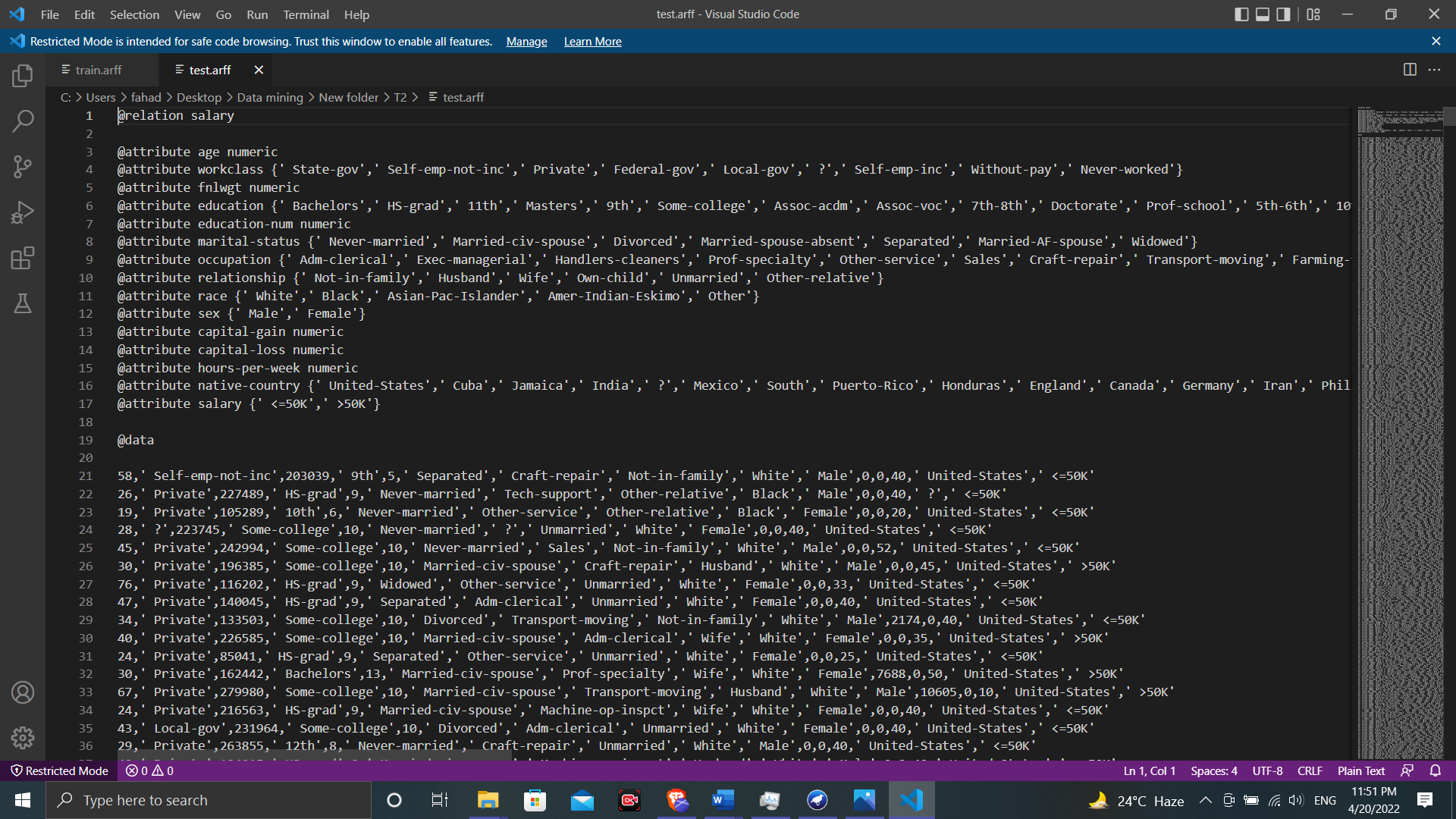


Figure: Test Data set

## **Naïve bayes (After Applying Test data)**

We have applied Naïve bayes Classification using test data where total number of instances are 8000. After applying naïve bayes correctly classified instances were 6696 which is about 83.7% and incorrectly classified instances were 16.3%.

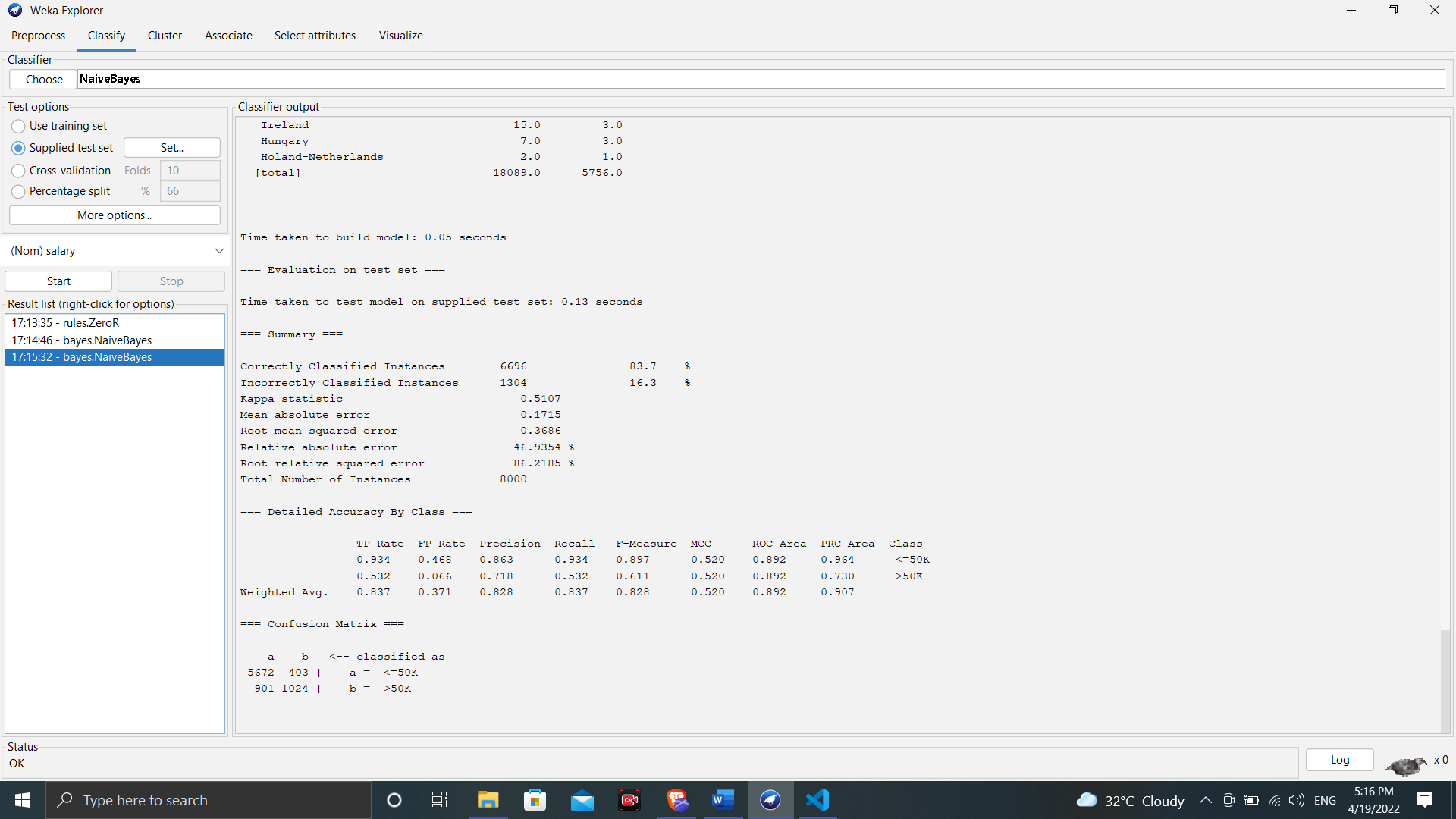


Figure: Naïve Bayes (After Applying Test data)

**K means clustering:**

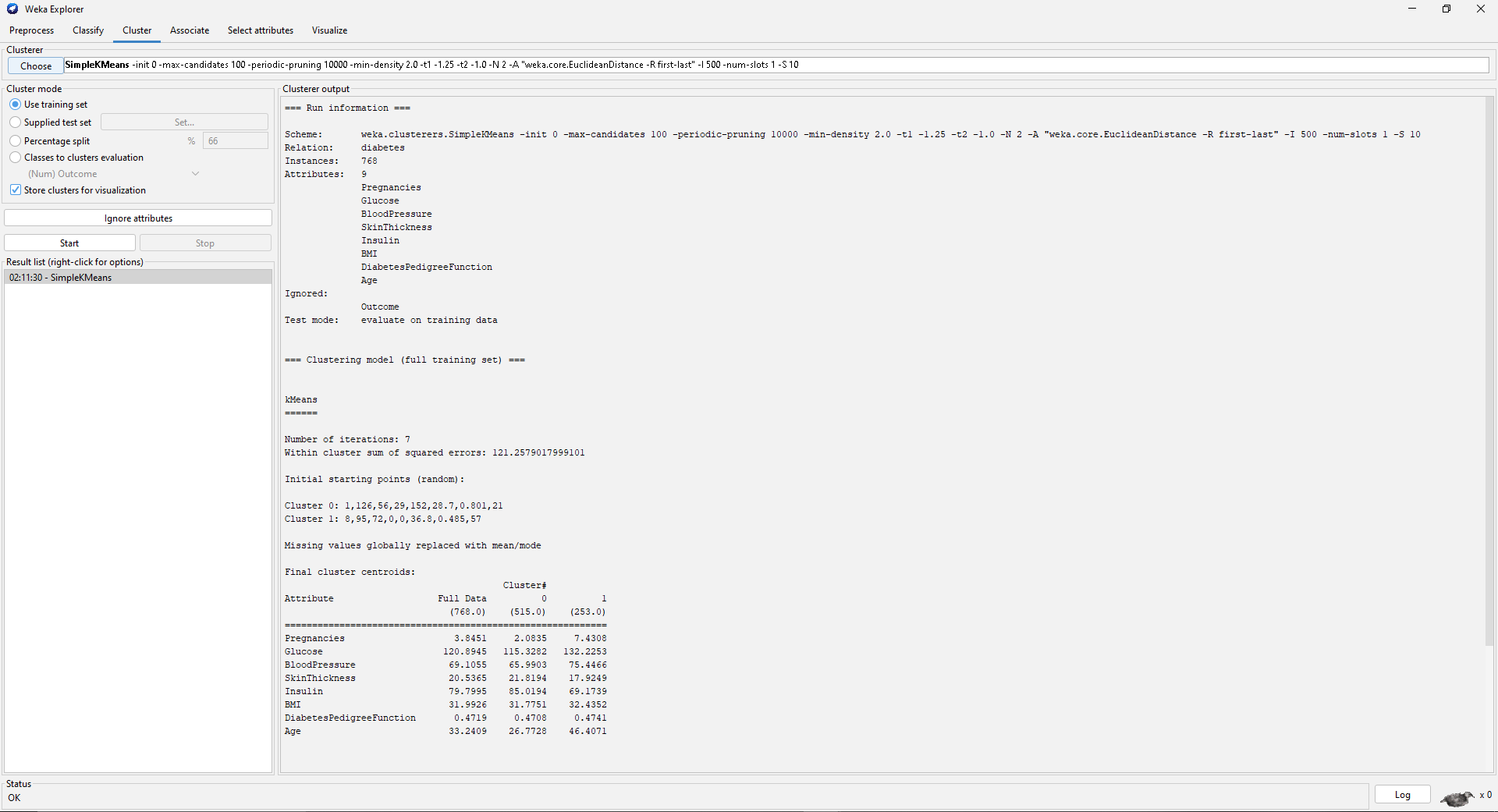
****

Figure: K-Means Clustering (Unsupervised Learning)

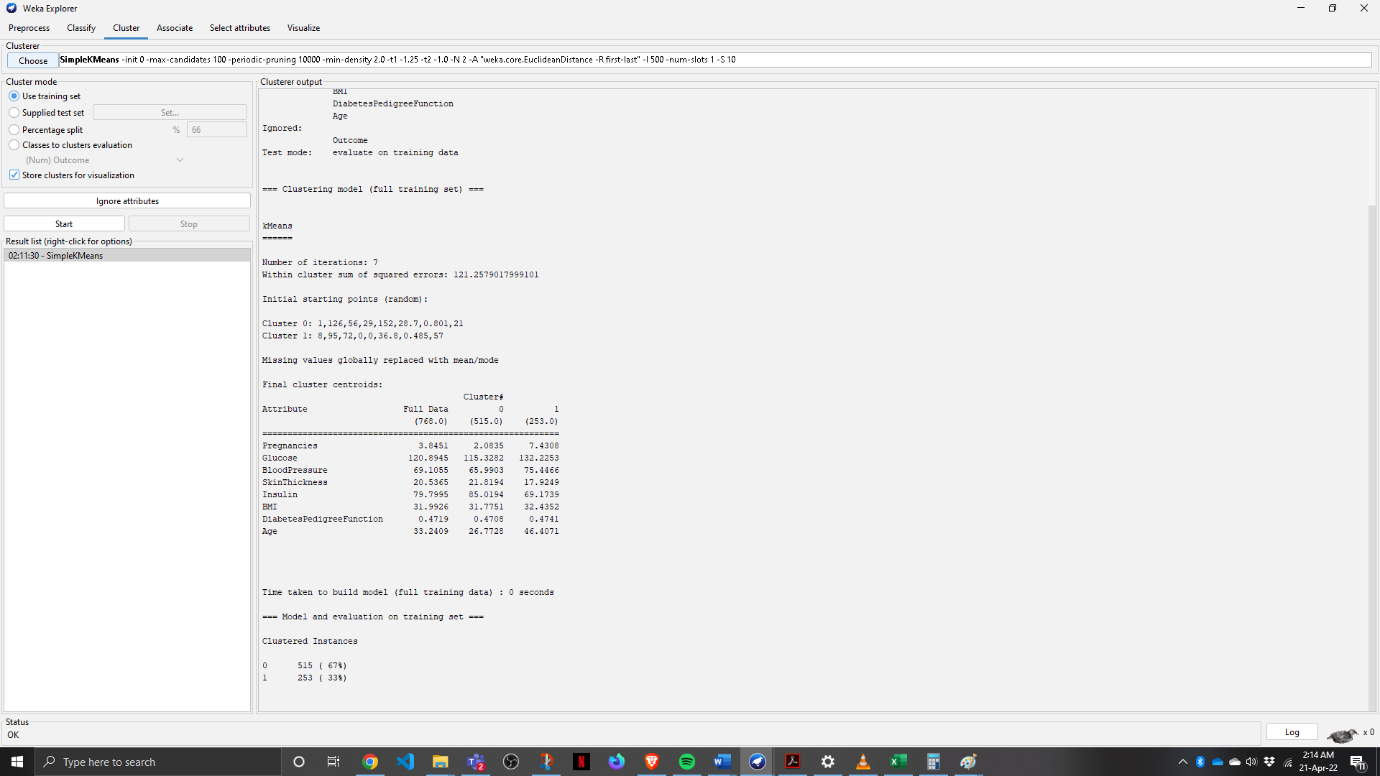
****

Figure: K-Means Clustering (Unsupervised Learning)

# **Discussion**

In this experiment we become familiar with Weka Software. We understand the uses of various types of algorithms and find out the dataset. And this achievement will provide a useful resource in the future.