AMERICAN INTERNATIONAL UNIVERSITY-BANGLADESH

Faculty of Science and Technology

Assignment Cover Page

Assignment Title:	Data Warehousing and Data Mining Final Project			
Assignment No:	-		Date of Submission:	7 August 2022
Course Title:	Data Warehousing and Data Mining			
Course Code:	CSC4285		Section:	D
Semester:	Summer	2021-22	Course Teacher:	Akinul Islam Jony

Declaration and Statement of Authorship:

1. I/we hold a copy of this Assignment/Case-Study, which can be produced if the original is lost/damaged.
2. This Assignment/Case-Study is my/our original work and no part of it has been copied from any other student's work or from any other source except where due acknowledgement is made.
3. No part of this Assignment/Case-Study has been written for me/us by any other person except where such collaborationhas been authorized by the concerned teacher and is clearly acknowledged in the assignment.
4. I/we have not previously submitted or currently submitting this work for any other course/unit.
5. This work may be reproduced, communicated, compared and archived for the purpose of detecting plagiarism.
6. I/we give permission for a copy of my/our marked work to be retained by the Faculty for review and comparison, including review by external examiners.
7. I/we understand thatPlagiarism is the presentation of the work, idea or creation of another person as though it is your own. It is a formofcheatingandisaveryseriousacademicoffencethatmayleadtoexpulsionfromtheUniversity. Plagiarized material can be drawn from, and presented in, written, graphic and visual form, including electronic data, and oral presentations. Plagiarism occurs when the origin of them arterial used is not appropriately cited.
8. I/we also understand that enabling plagiarism is the act of assisting or allowing another person to plagiarize or to copy my/our work.

* *Student(s) must complete all details except the faculty use part.*
** Please submit all assignments to your course teacher or the office of the concerned teacher.

Group Name/No.: -

No	Name	ID	Program	Signature
1	ZISHAN AHMED	20-42085-1	BSc [CSE]	
2	SHAKIB SADAT SHANTO	20-43074-1	BSc [CSE]	
3	MOHAMMED ASHRAF	17-34496-2	BSc [CSE]	
4	REFAT HASAN	19-41109-2	BSc [CSE]	
5	-	-	-	
6	-	-	-	
7	-	-	-	
8	-	-	-	
9	-	-	-	
10	-	-	-	

Faculty use only		
FACULTYCOMMENTS	**Marks Obtained**	
	Total Marks	

Project Overview:

The practice of gathering useful data, trends, and other important information from a sizable number of data sets is known as data mining. Data mining is often referred to as knowledge discovery in data. Data mining is employed in a variety of fields, including business and research. Data mining is a cross-disciplinary field in computer science and statistics with the overall goal of extracting information from a data collection and structuring it for later use. Data mining uses a variety of classification techniques, including KNN, Naive Bayes, and Decision Tree.

K-nearest neighbors (KNN) is a type of supervised ML algorithm which can be used for both classification as well as regression predictive problems. It uses 'feature similarity' to predict the values of new data points which further means that the new data point will be assigned a value based on how closely it matches the points in the training set.

Naive Bayes is a machine learning model that is used for large volumes of data. It is a fast and uncomplicated classification algorithm. Even if you are working with data that has millions of data records the recommended approach is to start with Naïve Bayes.

A decision tree is a tree-like graph with edges representing answers and leaves representing the actual output or class label. Each node acts as a test case for some attribute, and each edge descending from that node corresponds to one of the possible answers. This process is repeated for every subtree rooted at the new nodes.

In this project we have implemented a classification-based data mining application on a real-world data set. Our goal is to apply various classification methods like Naïve Bayes, K-Nearest Neighbor, Decision Tree on the data set to compare their accuracy with Predictive accuracy, Confusion matrix. Classification refers to a predictive modeling problem where a class label is predicted for a given example of input data. Predictive accuracy describes whether the predicted values match the actual values of the target field within the incertitude due to statistical fluctuations and noise in the input data values. For our data set we have chosen a car purchase decision dataset. We have used Weka Tool for our project. Weka is a collection of machine learning algorithms for data mining tasks. We expect to have a good synopsis of our data set and find which classification method suits best for it.

Dataset Overview:

For our dataset, we have selected a real-world data set on car purchase decision. This dataset contains details of 1000 customers who intend to buy a car. The data set has 4 columns. They are Gender, Age, Annual Salary and Purchased.

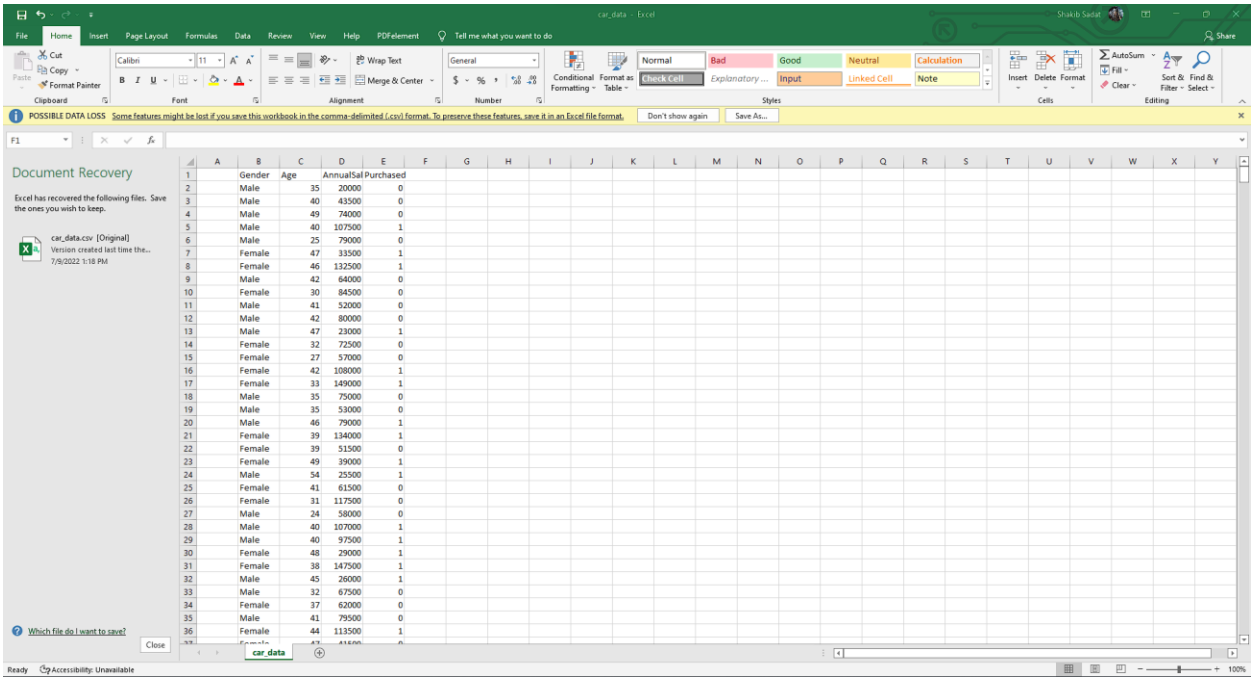

Fig 1: Car Purchase Dataset

The Purchased attribute is the decision or target attribute which has two values; 1 means Yes and 0 means No. The purchase decision will be based on the customers gender, age and annual salary.

Here's the url for the dataset:

https://www.kaggle.com/datasets/gabrielsantello/cars-purchase-decision-dataset

Model Development:

Naïve Bayes:

Naive Bayes methods are a set of supervised learning algorithms based on applying Bayes' theorem with the "naive" assumption of conditional independence between every pair of features given the value of the class variable. As we know, Naïve Bayes

Algorithm works with nominal attributes. So, in the data preprocessing step we had to make all attributes of our dataset from numeric to nominal. For doing that, we had to go to filter option in WEKA then select unsupervised then attributes then select numeric to nominal. For data cleaning step, we replaced missing values of the dataset with the means from the dataset. After completing data preprocessing steps, we applied the Naïve Bayes Algorithm to the dataset.

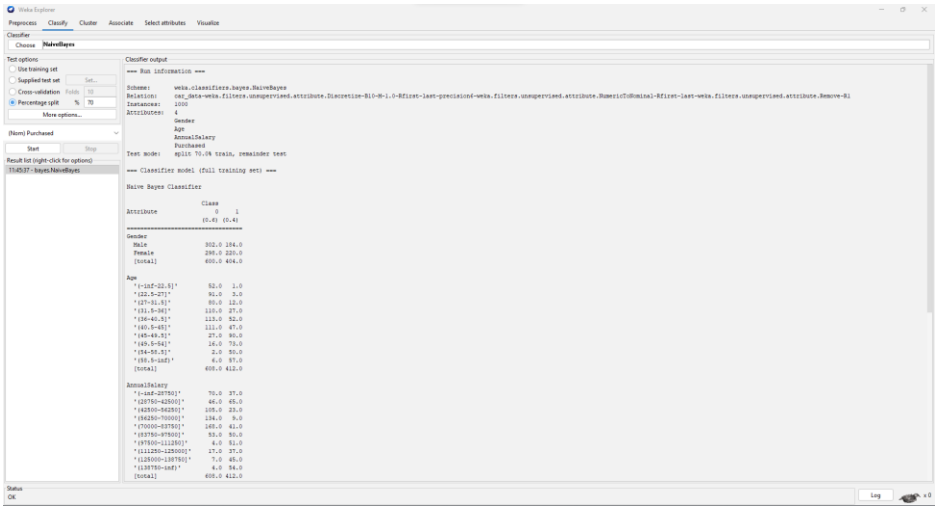

Fig 2: Naïve Bayes Algorithm (Summary)

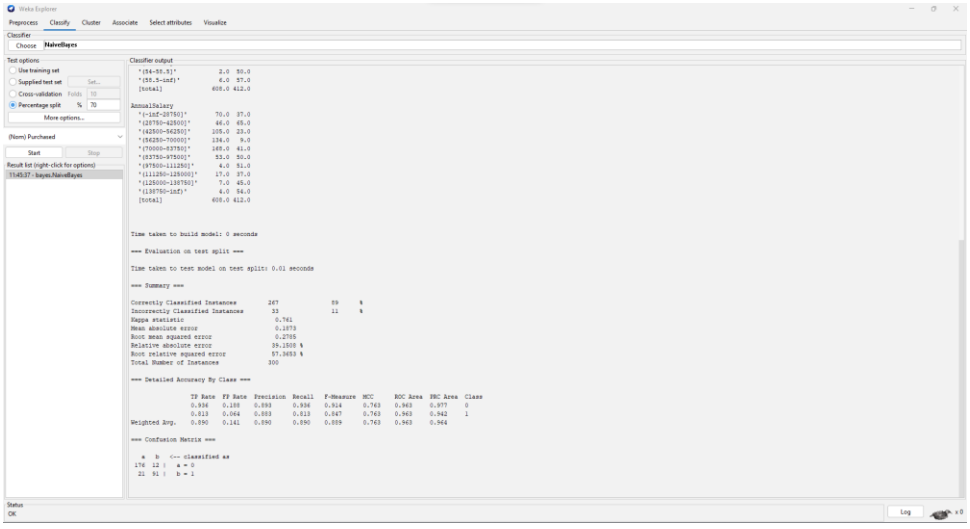

Fig 3: Naïve Bayes Algorithm (Summary)

Here we split the dataset to 70.0% training, remainder for testing.

Here,

Correctly Classified Instances	267	Accuracy = 89%
Incorrectly Classified Instances	33	Accuracy = 11%

Confusion Matrix:

n=300	Predicted: a=0	Predicted: b=1
Actual: a	176	12
Actual : b	21	91

Decision Tree:

A decision tree illustrates all possible outcomes for an input with a branching method. There is a root node, branches, and leaf nodes in this structure. The internal nodes represent tests on attributes, the branches represent test results, and the leaf nodes represent class labels. The topmost node in the tree is the root node.

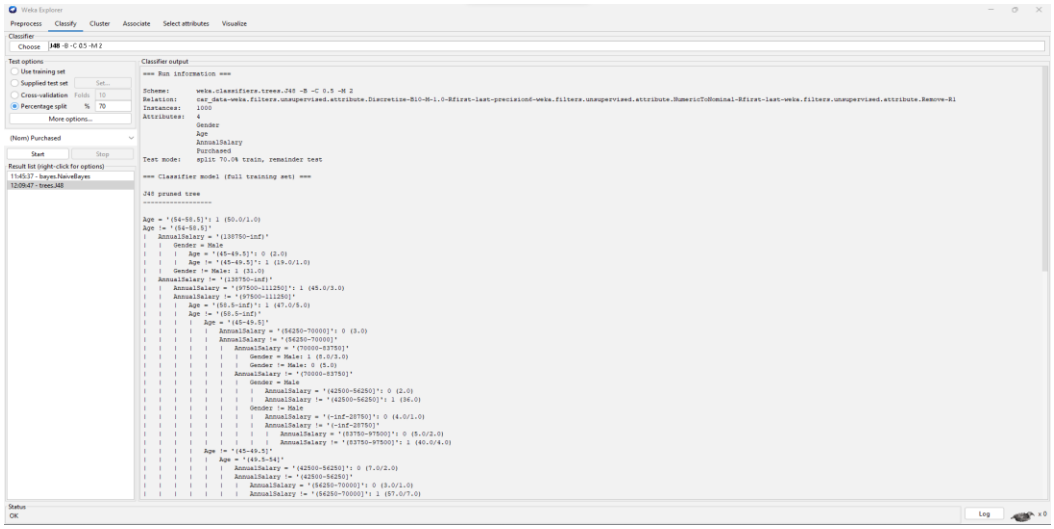

Fig 4: Decision Tree Algorithm (Summary)

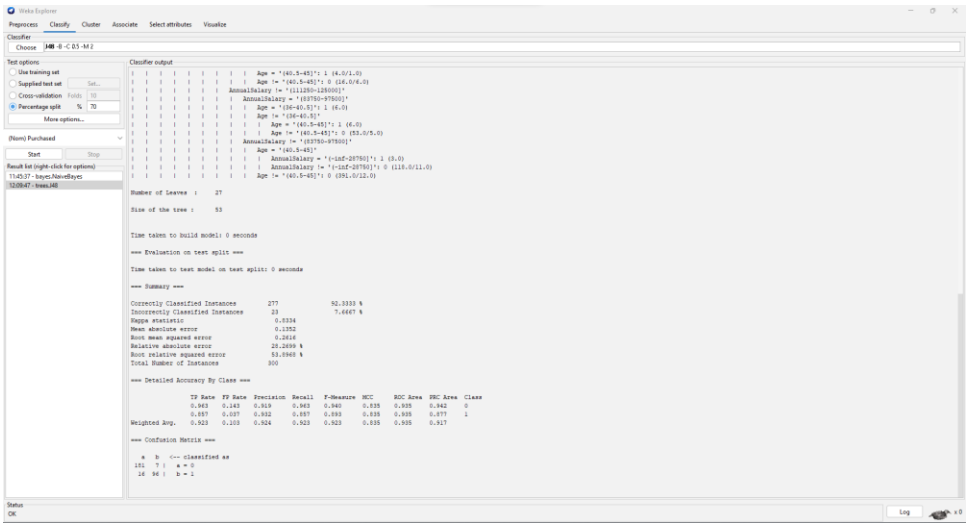

Fig 5: Decision Tree Algorithm (Summary)

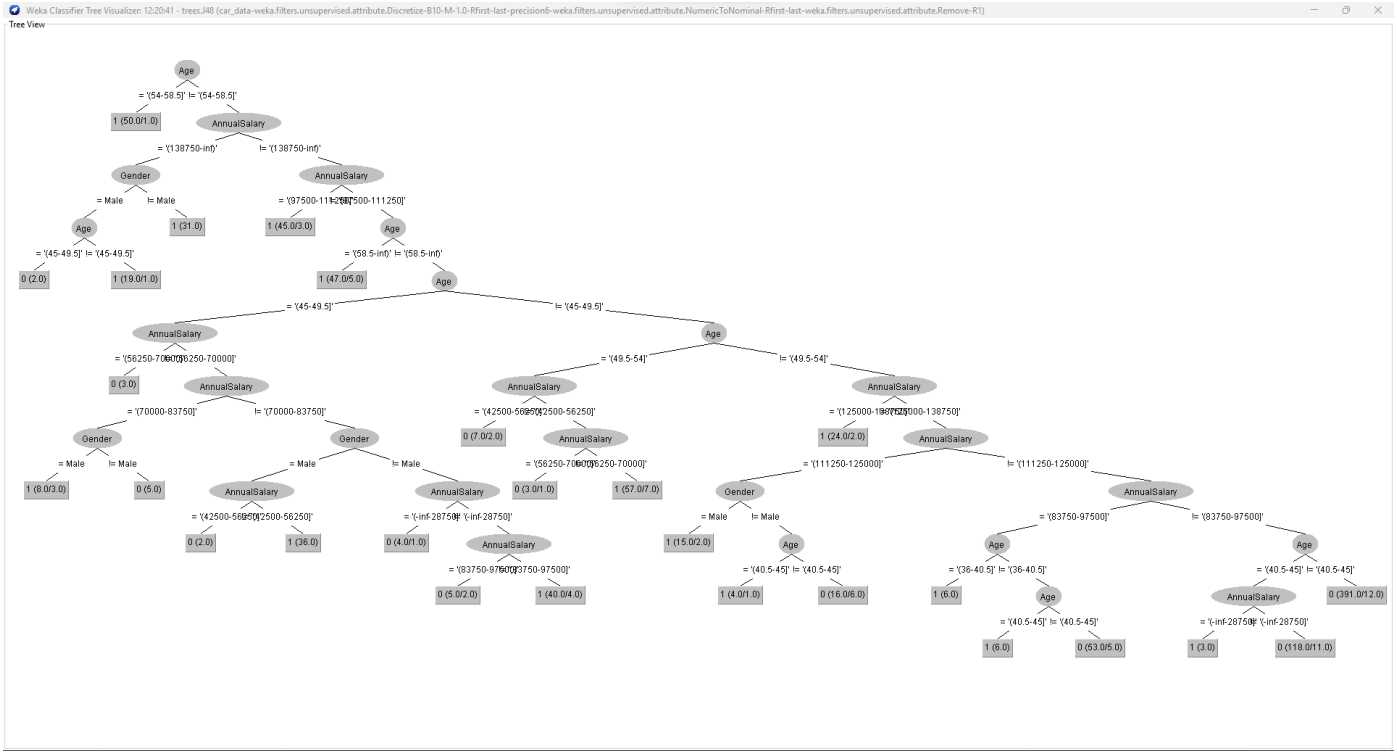

Fig 6: Decision Tree Visualization

Here,

Correctly Classified Instances	277	Accuracy = 92.33%
Incorrectly Classified Instances	23	Accuracy = 7.67%

Confusion Matrix:

n=300	Predicted: a=0	Predicted: b=1
Actual: a	181	7
Actual : b	16	96

K-Nearest Neighbor Classification:

One of the simplest machine learning algorithms, based on the supervised learning method, is K-Nearest Neighbor.

The K-NN method makes the assumption that the new case and the existing cases are comparable, and it places the new instance in the category that is most like the existing categories.

A new data point is classified using the K-NN algorithm based on similarity after all the existing data has been stored. This means that utilizing the K-NN method, fresh data may be quickly and accurately sorted into a suitable category.

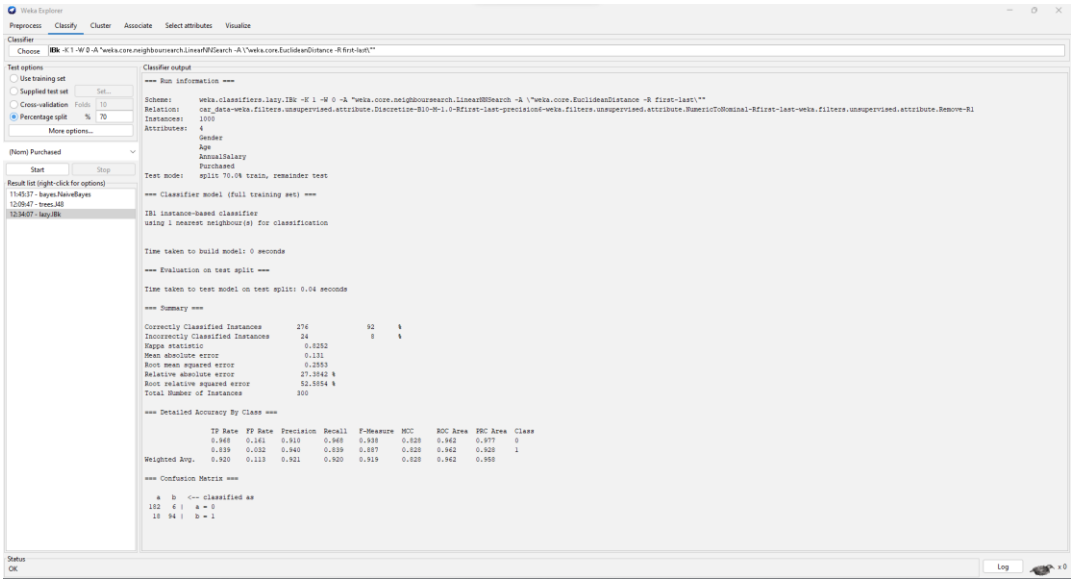

Fig 7:KNN Algorithm (Summary)

Here,

Correctly Classified Instances	276	Accuracy = 92%
Incorrectly Classified Instances	24	Accuracy = 8%

Confusion Matrix:

n=300	Predicted: a=0	Predicted: b=1
Actual: a	182	6
Actual : b	18	94

Data Visualization:

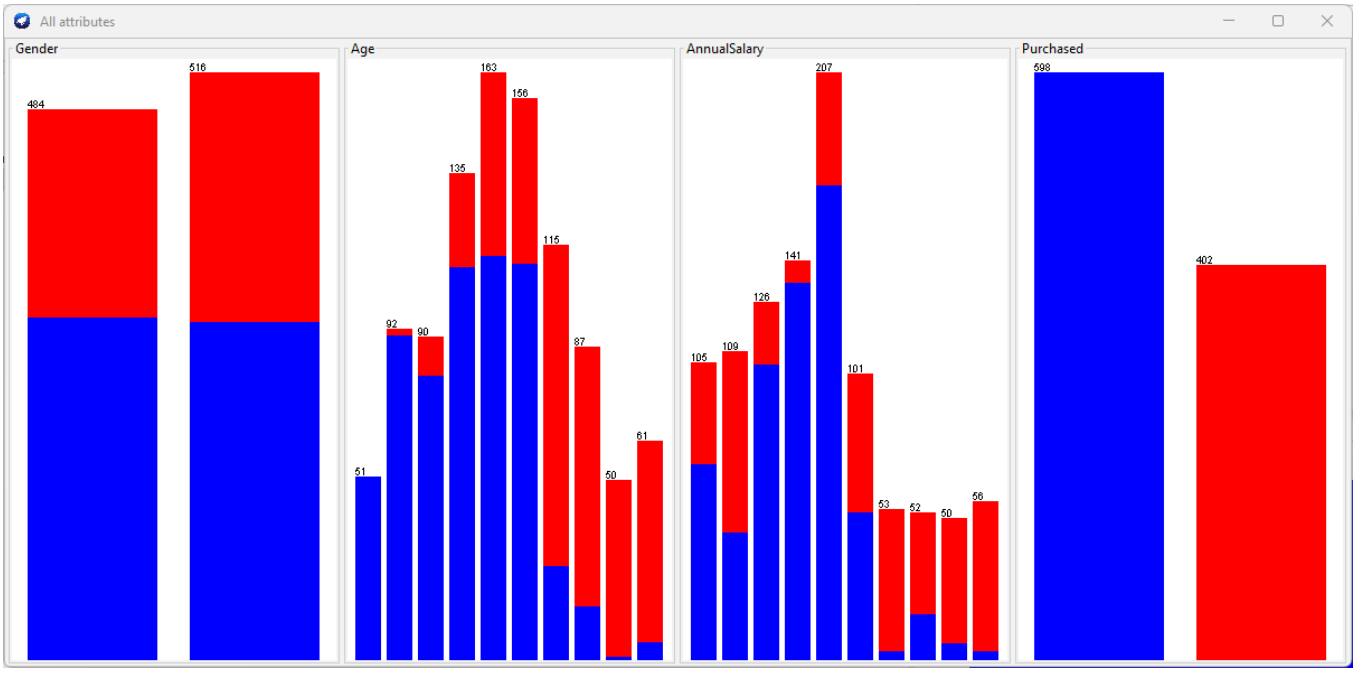

Fig 8: Visualization of All Attributes

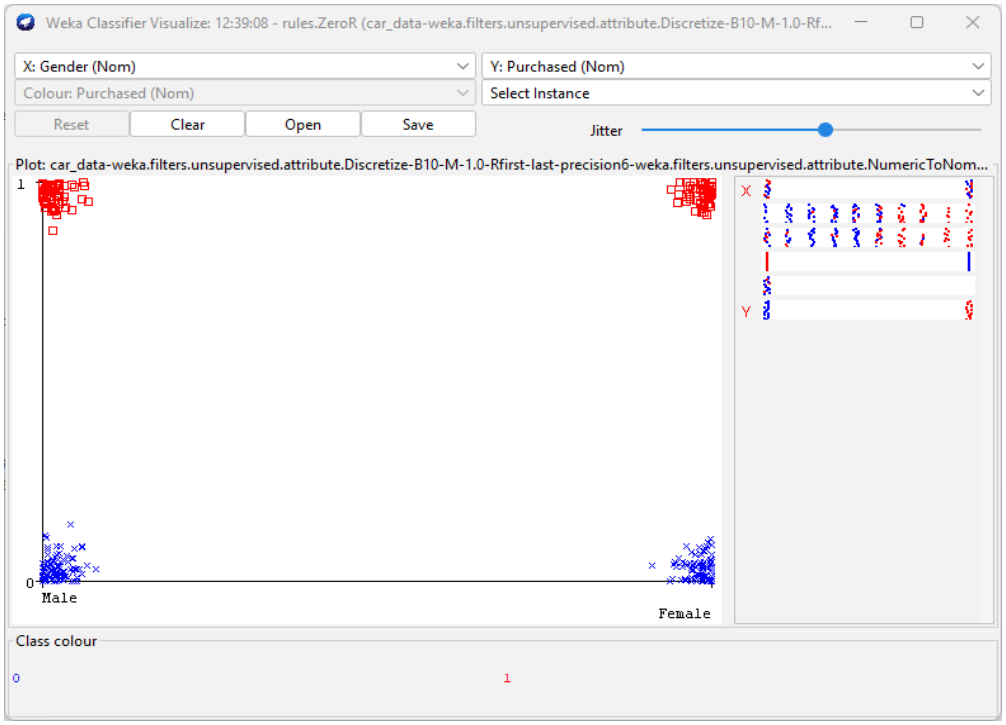

Fig 9: Visualization of Gender(X-axis) and Target variable Purchased(Y-axis)

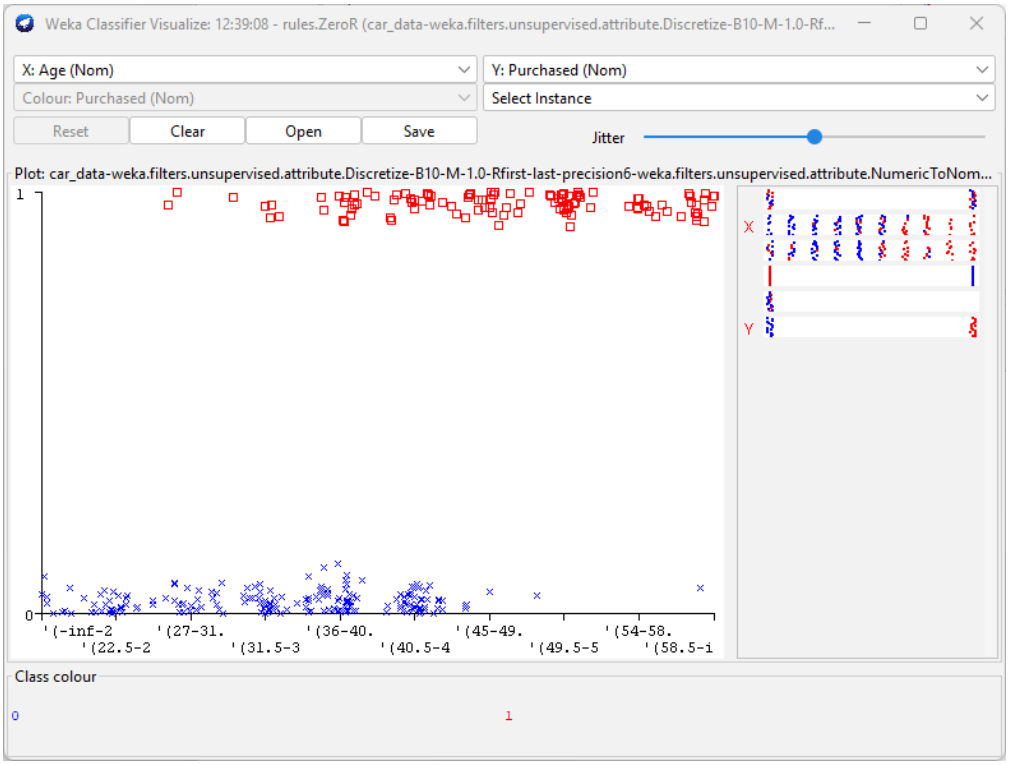

Fig 10: Visualization of Age(X-axis) and Target variable Purchased(Y-axis)

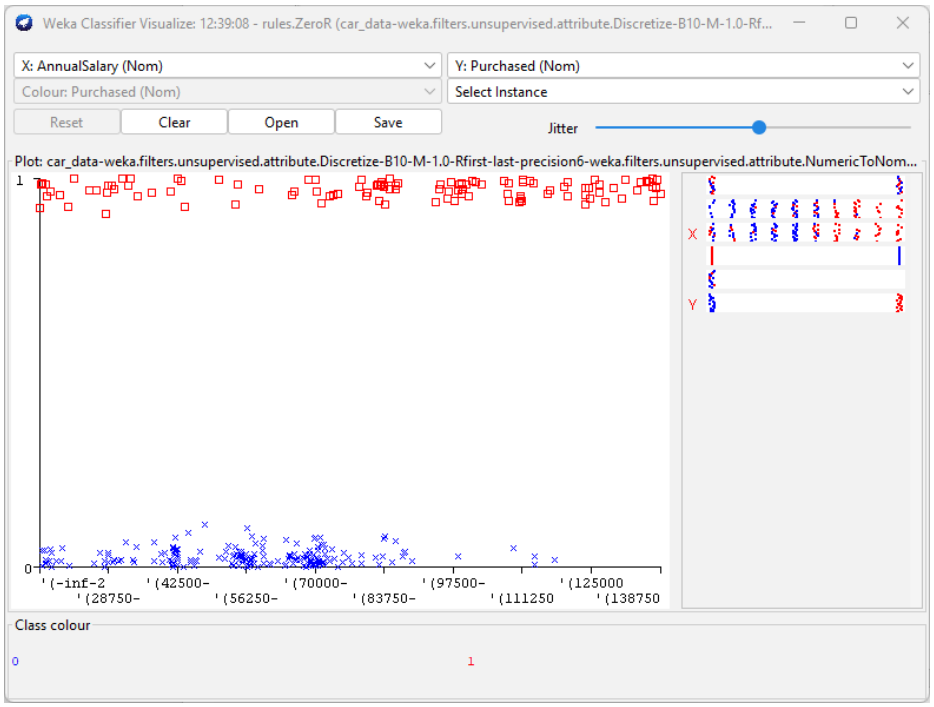

Fig 11: Visualization of AnnualSalary(X-axis) and Target variable Purchased(Y-axis)

Conclusion & Discussion:

Naïve Bayes:

There are 300 instances. From which 267(89%) are correctly classified and 33(11%) are incorrectly classified.

```
=== Summary ===

Correctly Classified Instances         267               89      %
Incorrectly Classified Instances        33               11      %
Kappa statistic                          0.761
Mean absolute error                      0.1873
Root mean squared error                  0.2785
Relative absolute error                 39.1508 %
Root relative squared error             57.3653 %
Total Number of Instances              300

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.936     0.188     0.893      0.936     0.914      0.763     0.963     0.977     0
               0.813     0.064     0.883      0.813     0.847      0.763     0.963     0.942     1
Weighted Avg.  0.890     0.141     0.890      0.890     0.889      0.763     0.963     0.964

=== Confusion Matrix ===

   a   b   <-- classified as
 176  12 |   a = 0
  21  91 |   b = 1
```

Fig 12: Naïve Bayes Model Accuracy with Confusion Matrix

Decision Tree:

There are 300 instances. From which 277(92.33%) are correctly classified and 23(7.67%) are incorrectly classified.

```
=== Evaluation on test split ===

Time taken to test model on test split: 0 seconds

=== Summary ===

Correctly Classified Instances         277               92.3333 %
Incorrectly Classified Instances        23                7.6667 %
Kappa statistic                          0.8334
Mean absolute error                      0.1352
Root mean squared error                  0.2616
Relative absolute error                 28.2699 %
Root relative squared error             53.8968 %
Total Number of Instances              300

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.963    0.143    0.919      0.963    0.940      0.835    0.935     0.942     0
               0.857    0.037    0.932      0.857    0.893      0.835    0.935     0.877     1
Weighted Avg.  0.923    0.103    0.924      0.923    0.923      0.835    0.935     0.917

=== Confusion Matrix ===

   a   b   <-- classified as
 181   7 |   a = 0
  16  96 |   b = 1
```

Fig 13: Decision Tree Model Accuracy with Confusion Matrix

KNN:

There are 300 instances. From which 276(92%) are correctly classified and 24(8%) are incorrectly classified.

```
=== Evaluation on test split ===

Time taken to test model on test split: 0.04 seconds

=== Summary ===

Correctly Classified Instances         276               92      %
Incorrectly Classified Instances        24                8      %
Kappa statistic                          0.8252
Mean absolute error                      0.131
Root mean squared error                  0.2553
Relative absolute error                 27.3842 %
Root relative squared error             52.5854 %
Total Number of Instances              300

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.968    0.161    0.910      0.968    0.938      0.828    0.962     0.977     0
               0.839    0.032    0.940      0.839    0.887      0.828    0.962     0.928     1
Weighted Avg.  0.920    0.113    0.921      0.920    0.919      0.828    0.962     0.958

=== Confusion Matrix ===

   a   b   <-- classified as
 182   6 |   a = 0
  18  94 |   b = 1
```

Fig 14: KNN Model Accuracy with Confusion Matrix

As we can see, the percentage of correctly classified instances of Nave Bayes is 89%, the percentage of correctly classified instances of KNN is 92%, and the percentage of correctly classified instances of Decision tree is 92.33%. Because the Decision Tree has a higher percentage of properly classified instances than Nave Bayes and KNN, we may conclude that it is better in this dataset.