*Student Name*

*Student id*

*Table of Contents*

*Data Choice5,6*

*Background Information7*

*Data Description7,11*

*Data Pre-Processing11,16*

*Data Mining16,23*

*Discussion of Findings24*

*References25*

*Appendices25,26*

*Table of Figures*

*Figure 15*

*Figure 25*

*Figure 36*

*Figure 46*

*Figure 57*

*Figure 68*

*Figure 79*

*Figure 810*

*Figure 910*

*Figure 1011*

*Figure 1112*

*Figure 1212*

*Figure 1313*

*Figure 1413*

*Figure 1514*

*Figure 1615*

*Figure 1715*

*Figure 1816*

*Figure 1917*

*Figure 2017*

*Figure 2118*

*Figure 2219*

*Figure 2319*

*Figure 2420*

*Figure 2520*

*Figure 2621*

*Figure 2722*

*Figure 2822*

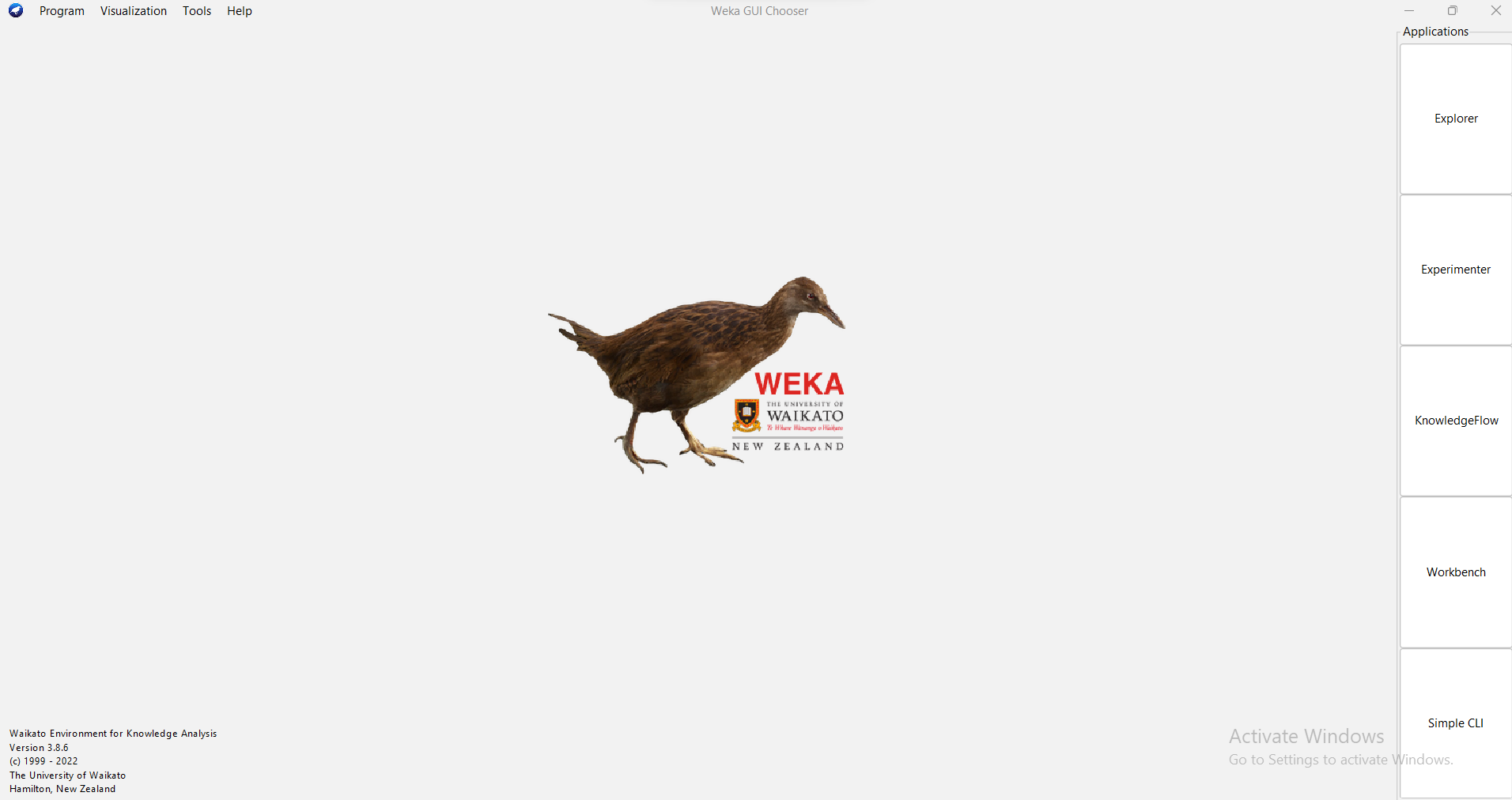
*Figure 2922*

*Figure 3023*

***Data Choice:***

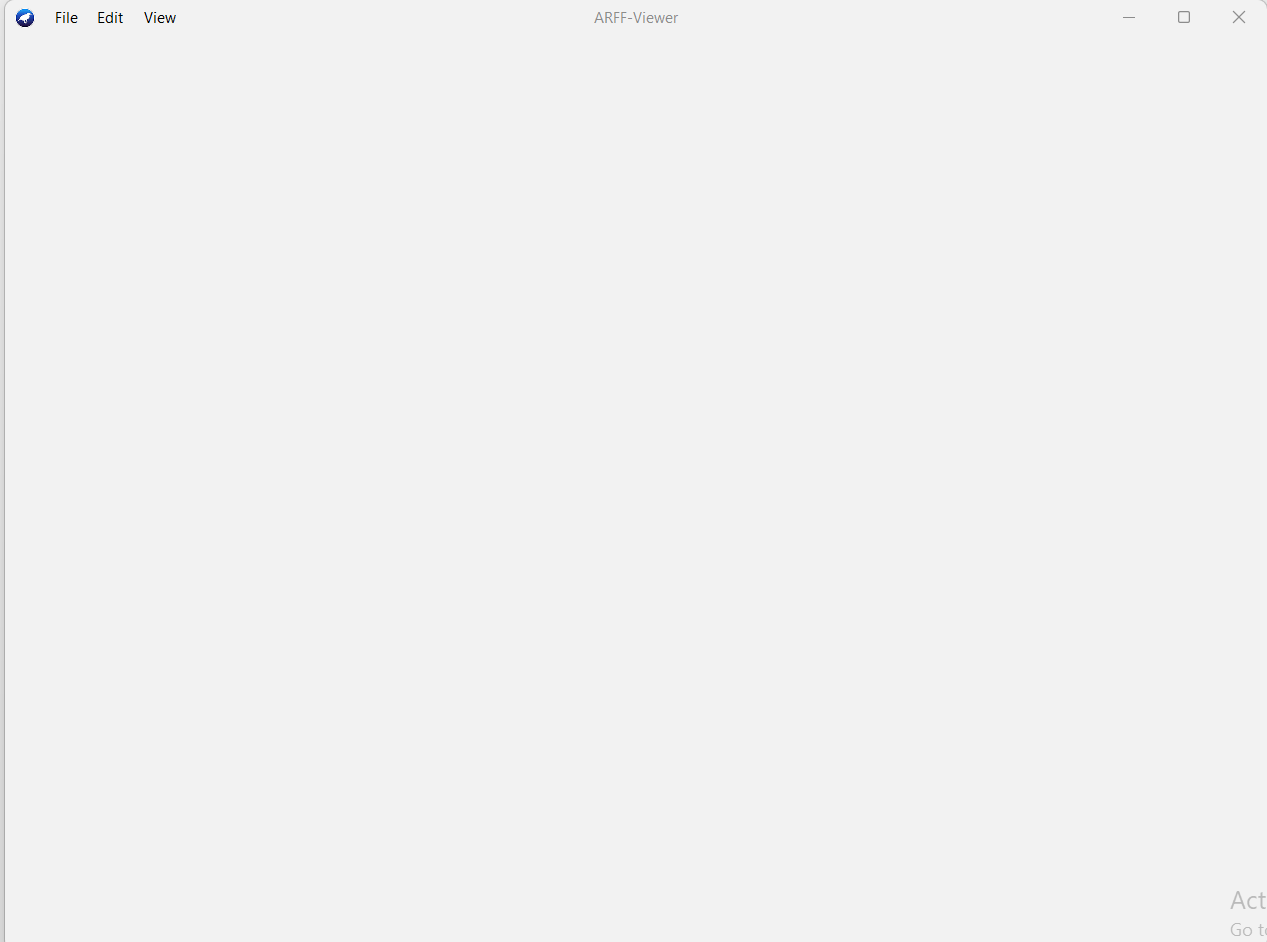
*The data which I have chosen is Diabetes Prediction. It has the attributes greater than 5 and its classification data to predict the diabetes in person on specific values and tell 0 or 1 which is yes or no.*

***Start Screen of Weka:***

**

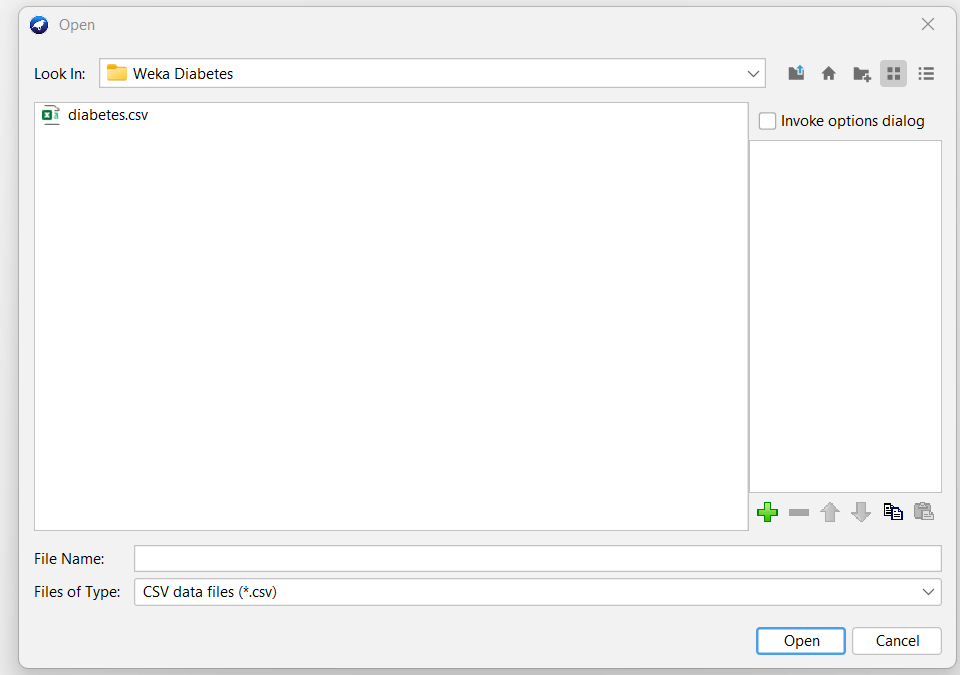
**Figure 1**

***Arrf viewer by Clicking the tools option on the top:***

**

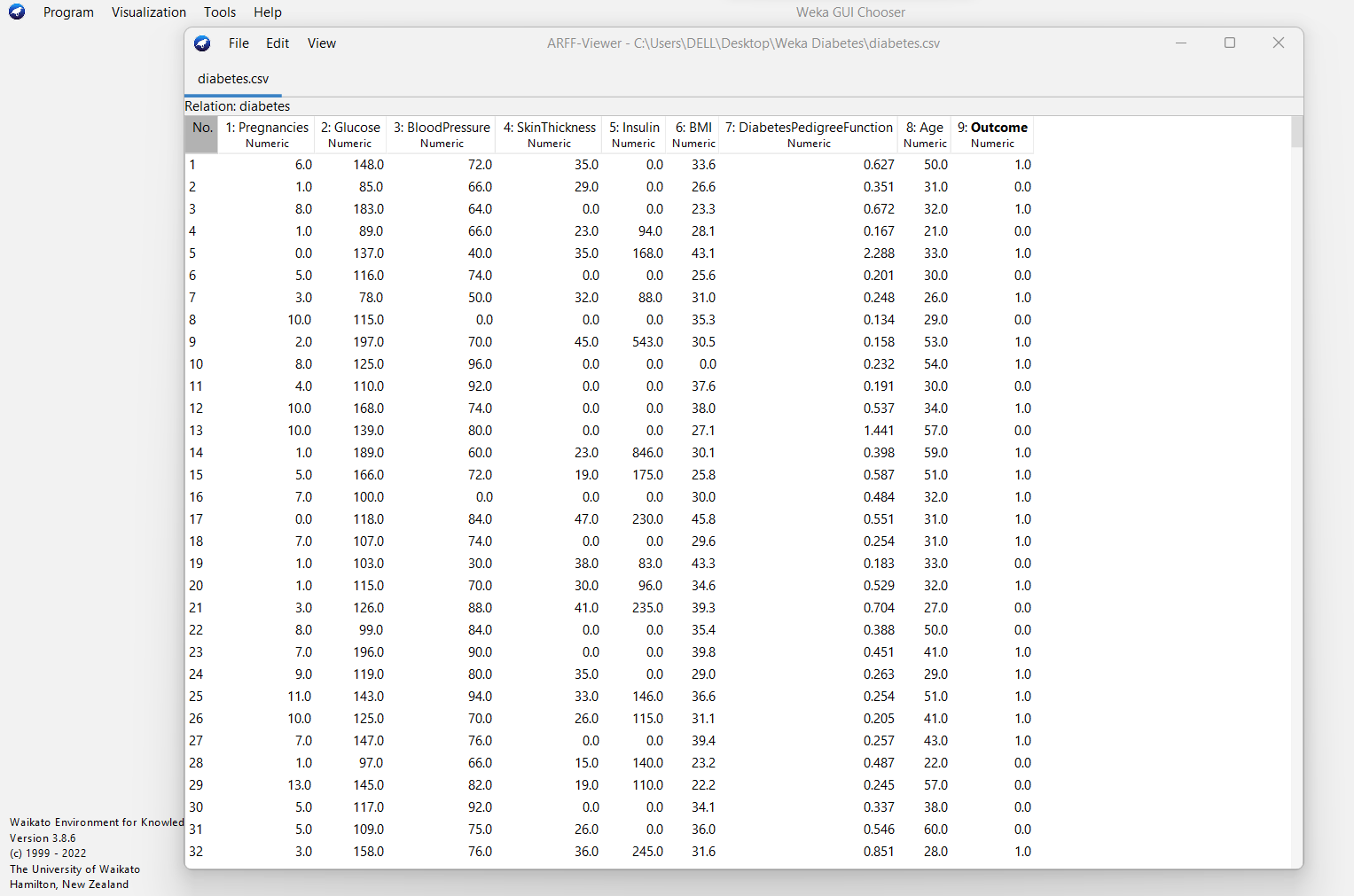
**Figure 2**

***By clicking file choose the csv file:***

******

**Figure 3**

***Data Conversion from csv to arff:***

**

**Figure 4**

***Background Information:***

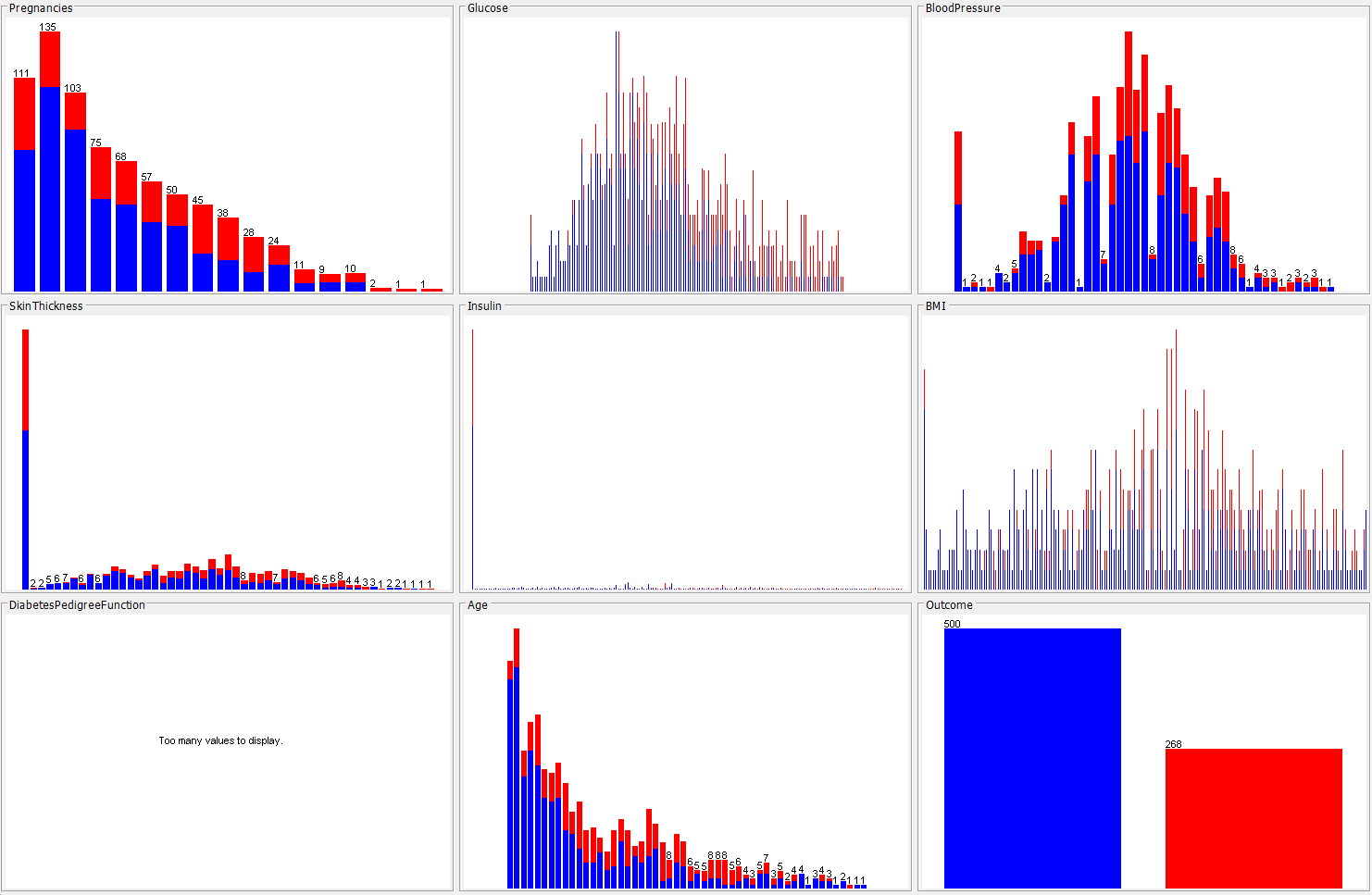
*Data is gathered from the Kaggle repository. This data basically belongs to the National Institute of diabetes, digestive and kidney diseases. The purpose of this data is to predict whether the patient has diabetes or not based on the given parameters. All the female patients from Pima Indian heritage are here in the dataset. The purpose is to predict whether or not the patient has diabetes. The data mining helps the organization to get the reliable information that can benefit the organization or institute. As data mining also uses the new and legacy systems for the organization as our data also used the legacy and new data because the assumptions are made by seeing the old records and results. And, with the new record the data should be updated.* (CHAUHAN, 2022)

***Data Description:***

*There is total 768 instances in the data and total of 9 attributes that includes:*

* *Pregnancies*
* *Glucose*
* *Blood Pressure*
* *Skin thickness*
* *Insulin*
* *BMI*
* *Diabetes Pedigree function*
* *Age*
* *Outcome (CLASS)*

***All Attributes Visualization:***

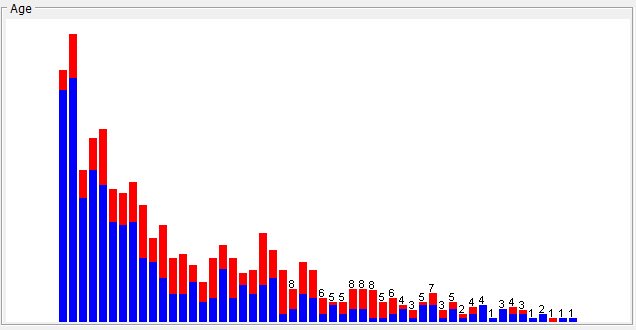
**

**Figure 5**

*The Data has no missing values inside the data, as well as it’s the numeric data all the columns are of numeric types. Each column has its own value in terms of finding and detecting the diabetes disease because diabetes includes all the column used in this data the value and count may differ but it matters a lot. The 5 columns are:*

***Age:***

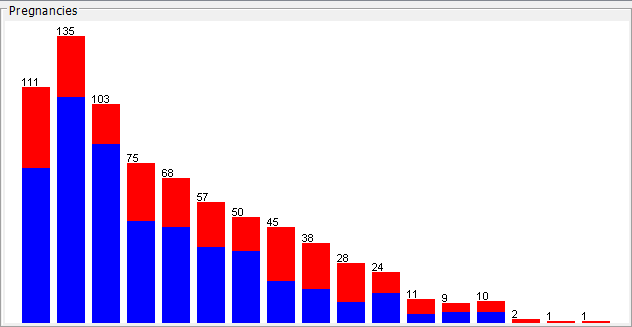
*It has no missing value and 5 are the unique values in this column. Minimum age in the column is 21 and maximum age in the column is 81. The visualization is:*

**

Figure

***Pregnancies:***

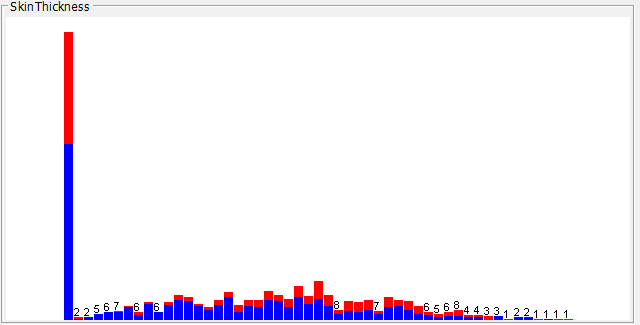
*This column also has no missing value. It is of numeric type. The minimum value in this column is 0 and the maximum value is 17. The visualization is:*

**

Figure

***Skin thickness:***

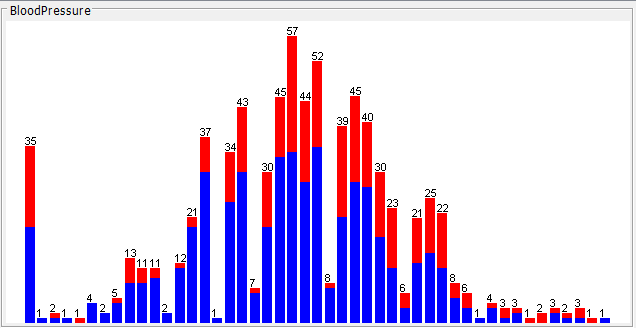
*This column also has no missing value. It is of numeric type. This column has 5 unique values. The minimum value is 0 and maximum value is 99. The visualization is:*

**

Figure

***Blood Pressure:***

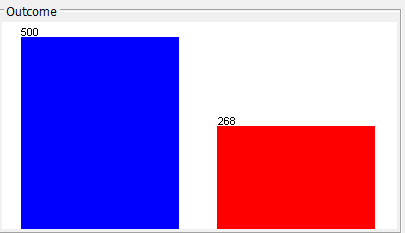
*This column is having no missing value. This column is also of numeric type. The minimum value is 0 and maximum value is 122. The graph is given below:*

**

**Figure 9**

***Outcome:***

*This column is the class label that has two values 0 and 1. This is of numeric type and has no missing values. The graph is:*

**

**Figure 10**

***Data Pre-Processing:***

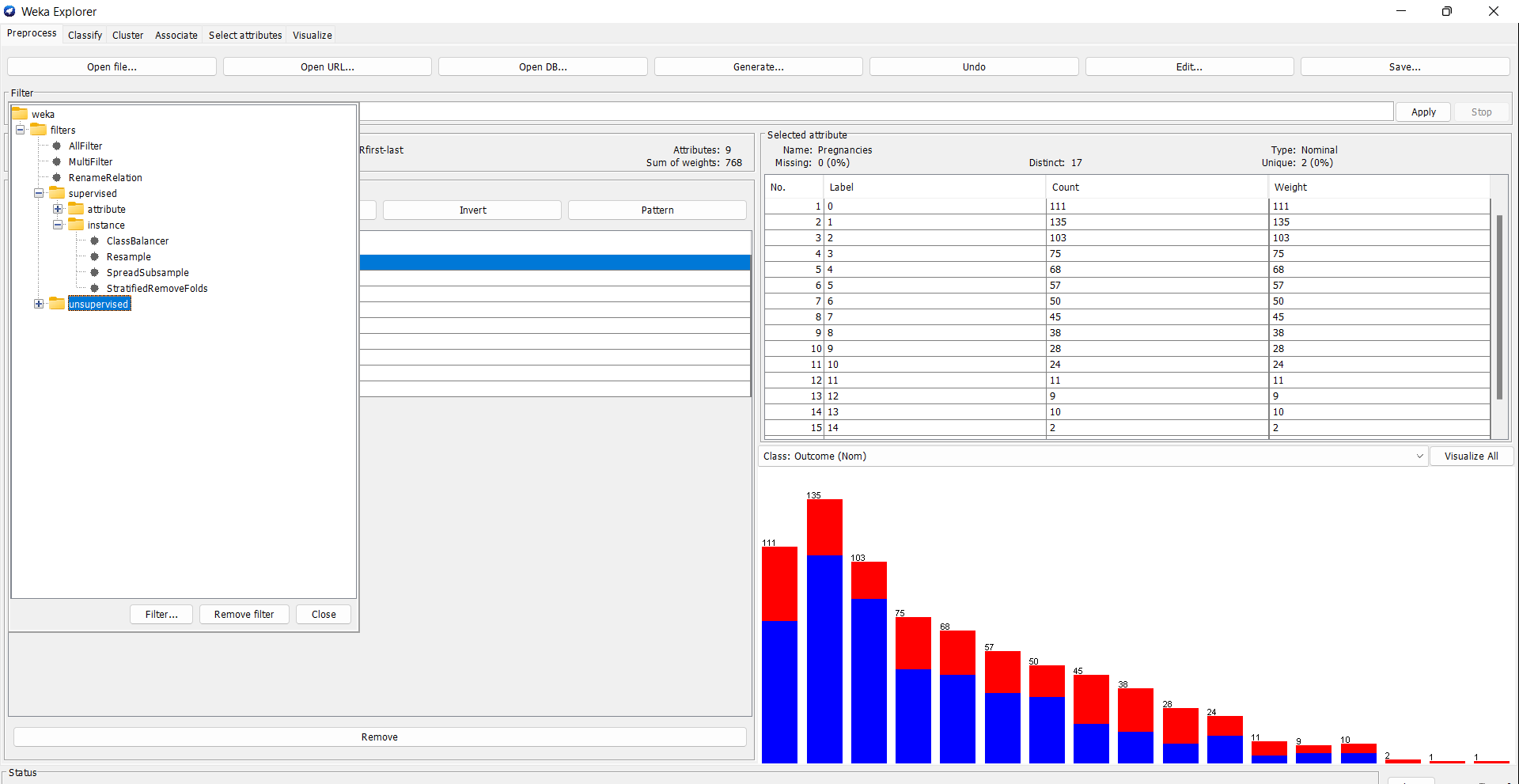
*The preprocessing techniques includes dealing with missing values, using the various filters in the preprocess tab in WEKA and dimension reduction. This preprocessing is the main goal before data analysis to make sure that the data is ready to use for any kind of analysis, it includes:*

* *Removing missing values*
* *Data Cleaning*
* *Sampling Data*
* *Data Transformation*
* *Imbalanced Data*
* *Feature Engineering*
* *Dimension Reduction*

***Under Sampling:*** *It is the process which includes the deletion of majority class and balancing the values of class. This process helps in reducing the risk of the analysis and as well the results of machine learning algorithms. In our diabetes data we had 500 data was of 1 class out of 700 so that was biased we used the filter and divided the class equally. The main advantage includes reduction of ADC cost.*

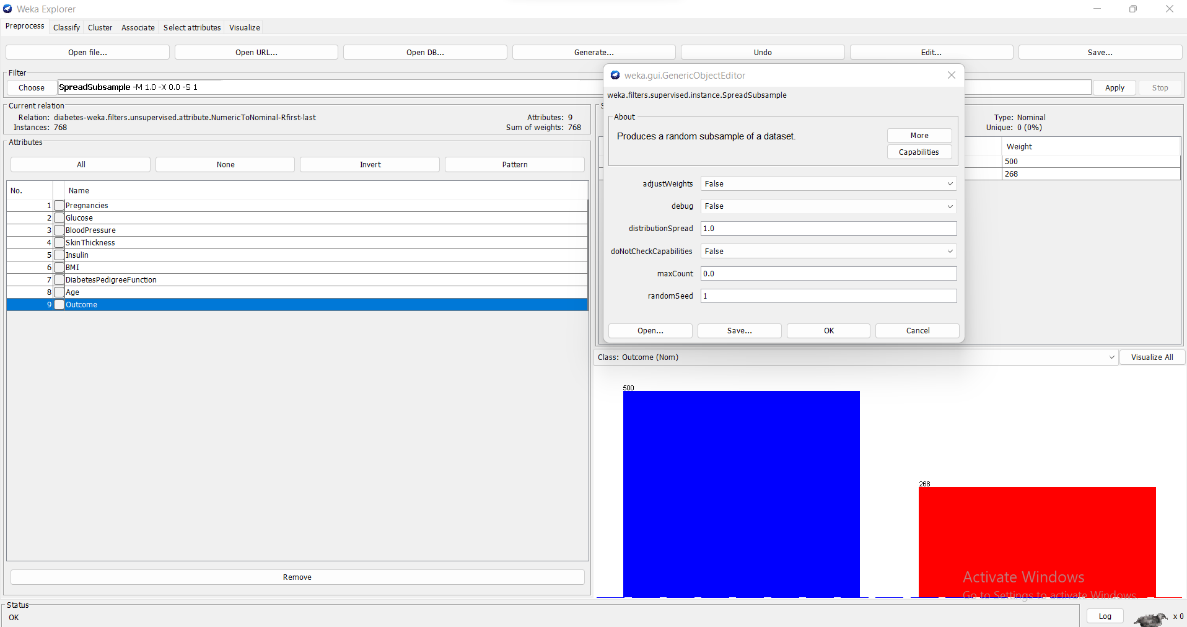
***Follow the following steps to apply the under sampling:***

***Click the choose button under the filter:***



**Figure 11**

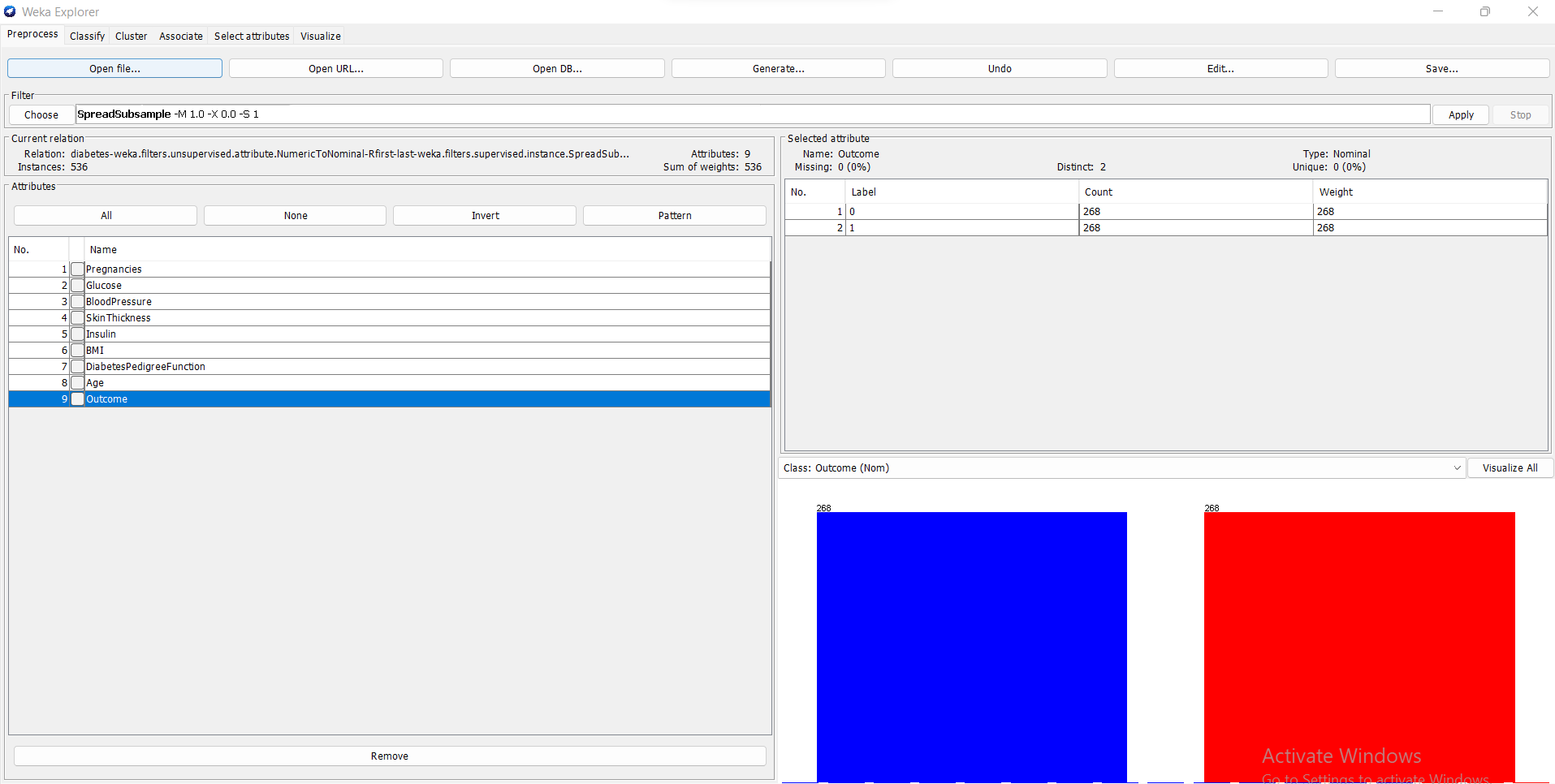
***The values will be set as default for my diabetes data you can edit and change if you want to change depending on the size and type of data you are applying this filter to when you are done editing the values of parameters then click apply button to apply the filter:***



**Figure 12**

***After the under-sampling class:***

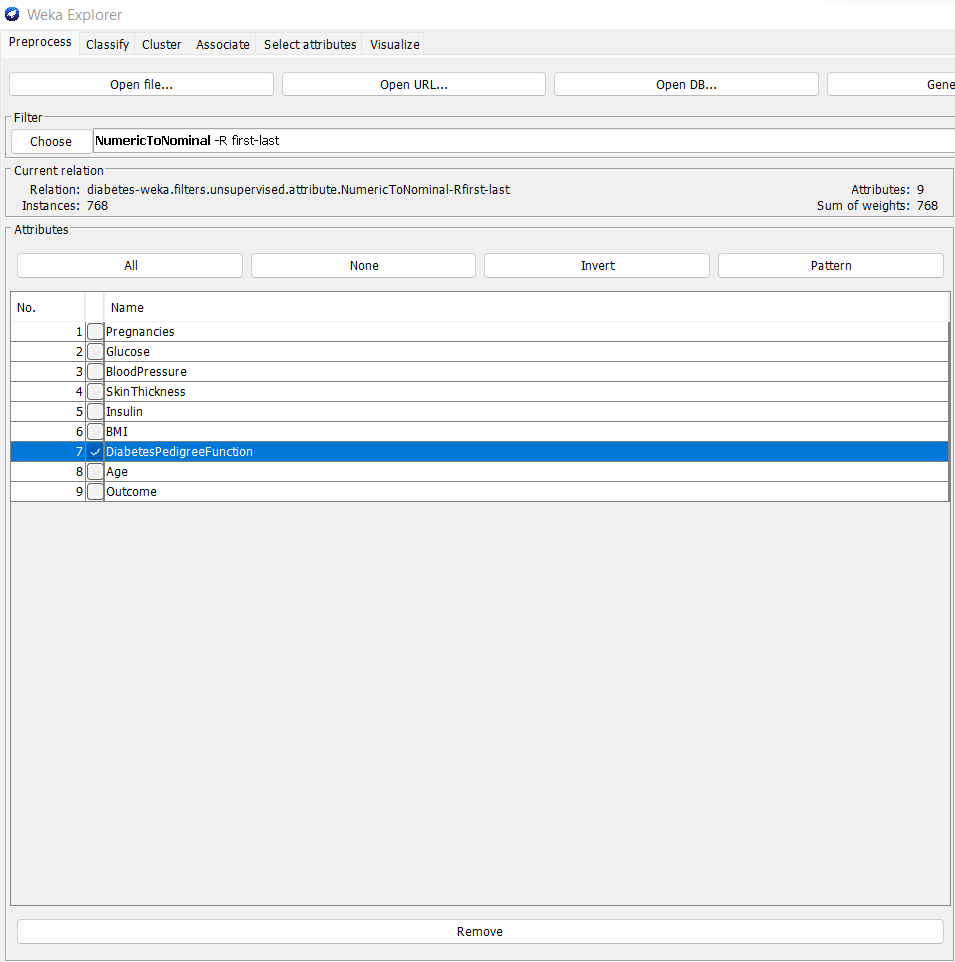
***When the filter is applied the class values were imbalanced now you will see that the class values will be equal.***



**Figure 13**

***Dimensionality Reduction:*** *Dimension reduction is removing the columns that are of not use or may be is the not the requirement. For example, as our data is of women patients so it includes the column called pregnancies what if we want to use the data for men patients then we will remove pregnancies column. So, here we have removed one column named diabetes pedigree function as shown below:*

***Simply select the column as below which you want to remove from the data and click the remove button to remove that column.***

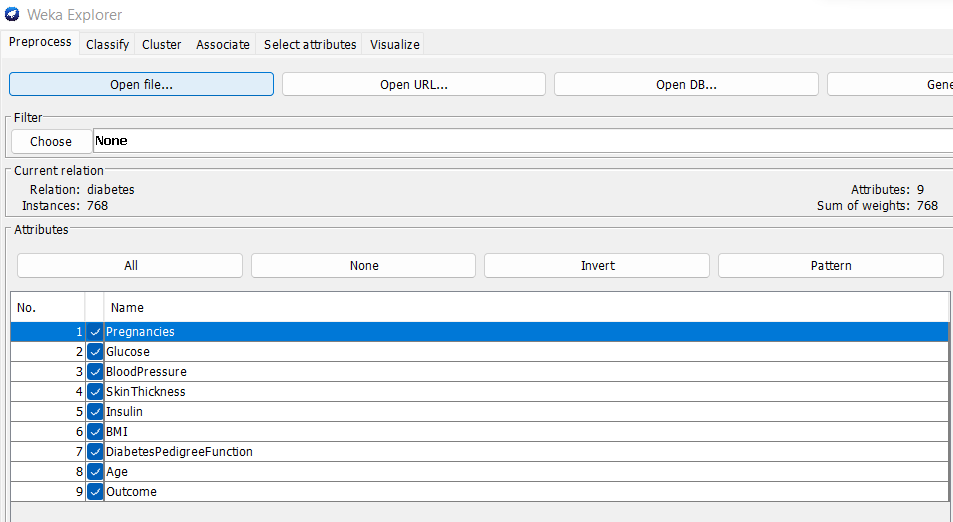
**

**Figure 14**

***Feature Engineering:***

*In this process I have used an unsupervised filter and converted the numeric values to the nominal as the for the Machine Learning algorithms we need to convert these columns from numeric to nominal. Below is the process we used while using the filter:*

***Select all the attributes by clicking the ALL button in the tab:***



**Figure 15**

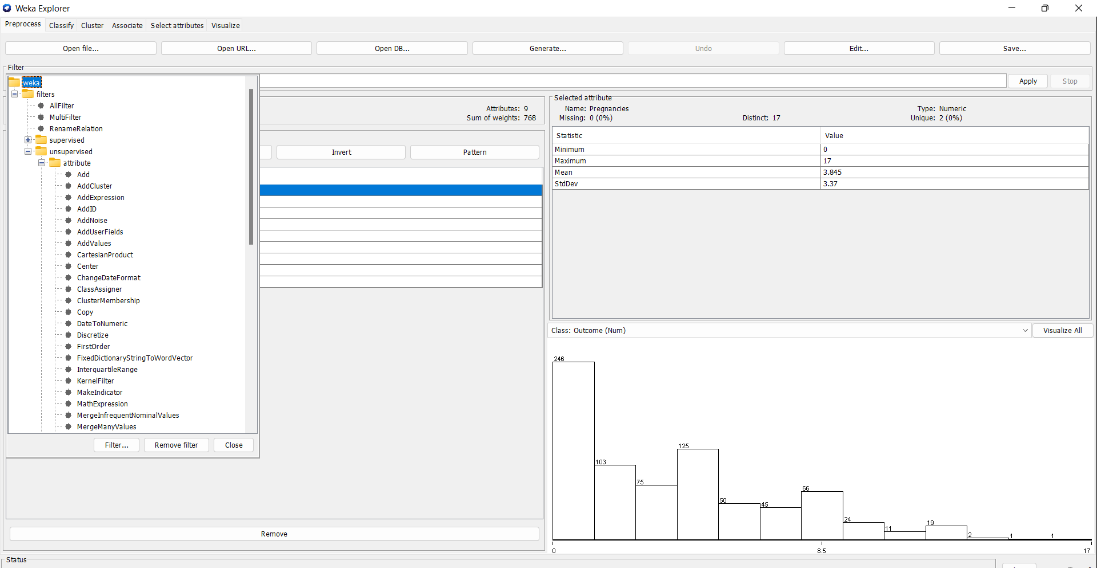
***Click the choose button below the Filter button:***

*And follow the following steps:*

*Filter >>*

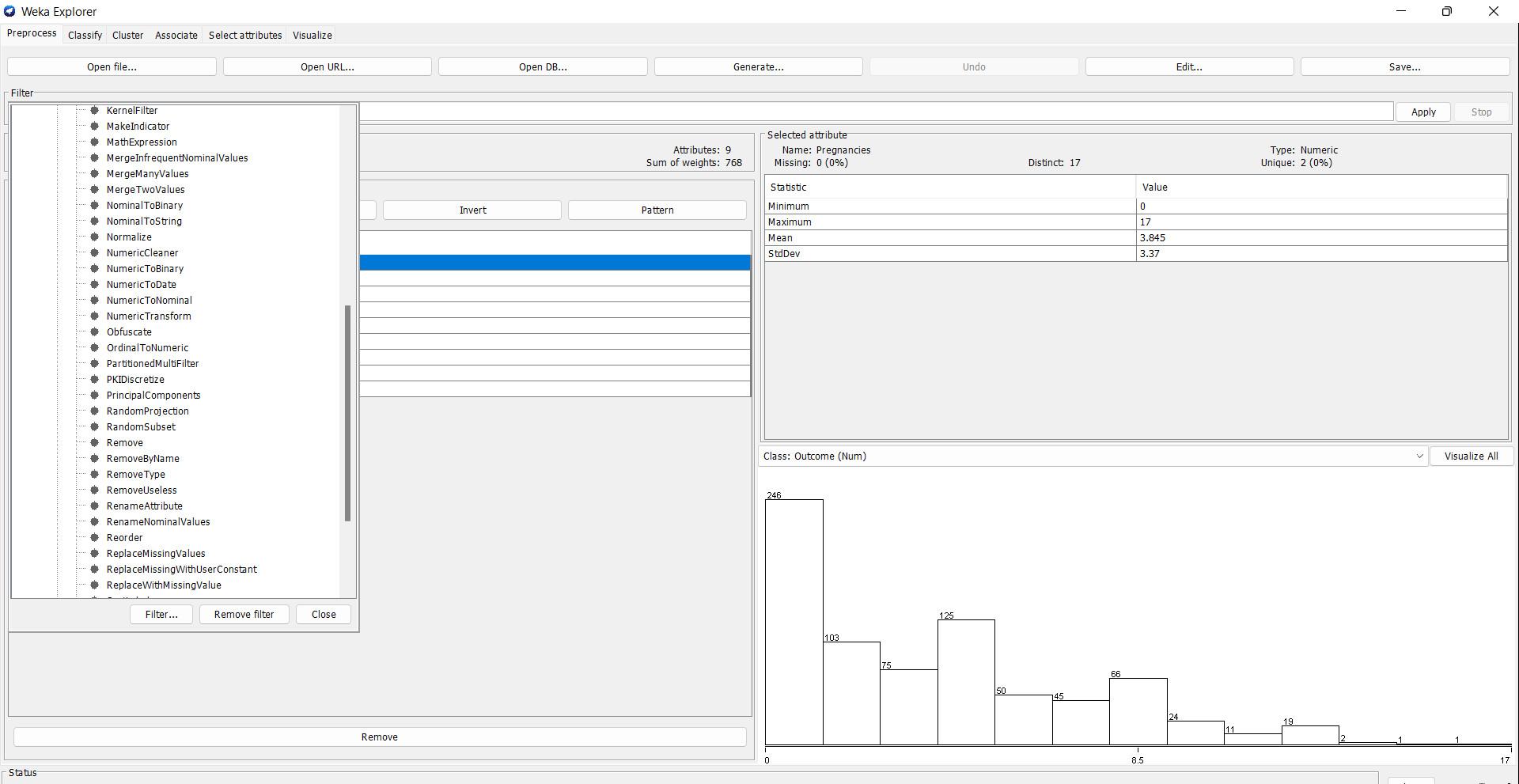
*Unsupervised >>*

*attributes >>*



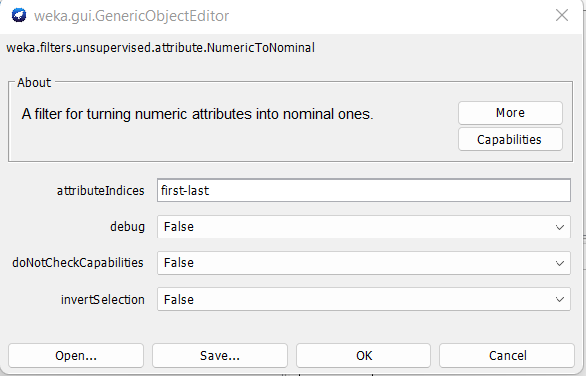
**Figure 16**

***Choose the Numeric to Nominal Option as given below:***



**Figure 17**

***Click the chosen filter and edit the filter values according to the needs of your data for mine data it will be set as default:***



**Figure 18**

***Data Mining:***

*Algorithms in data mining includes the set of calculation that create model for the data. To create the model, it checks the trend of the data you provide it.*

*I have used three data mining algorithms; the algorithms and result are:*

* *J48 (Classification)*
* *K Means (Clustering)*
* *Association Rule Mining (Association)*

***J48 Tree:***

*J48 is the classification is algorithm and used in WEKA. It is the same as the decision tree. It uses the top-down approach a which can also be recursive divide and conquer strategy. By splitting the root node, you find out each possible attributes value and split instances of each value.*

*The algorithm is used based on train/test split and the results includes the accuracy of the algorithm, Classification Report and Confusion Matrix.*

*The training size is 66% and remaining is the testing size for the data mining model for the given data.*

***Following are the steps followed to apply J48:***

***Click on the classify tab for applying data mining classification algorithm:***

******

**Figure 19**

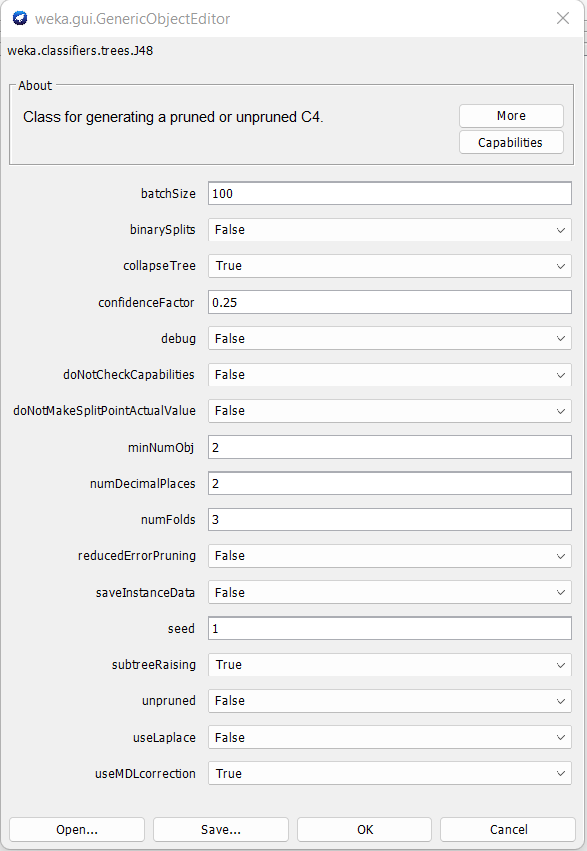
***Do the following steps:***

***Click choose button and click classifiers >> trees >> J48***



**Figure 20**

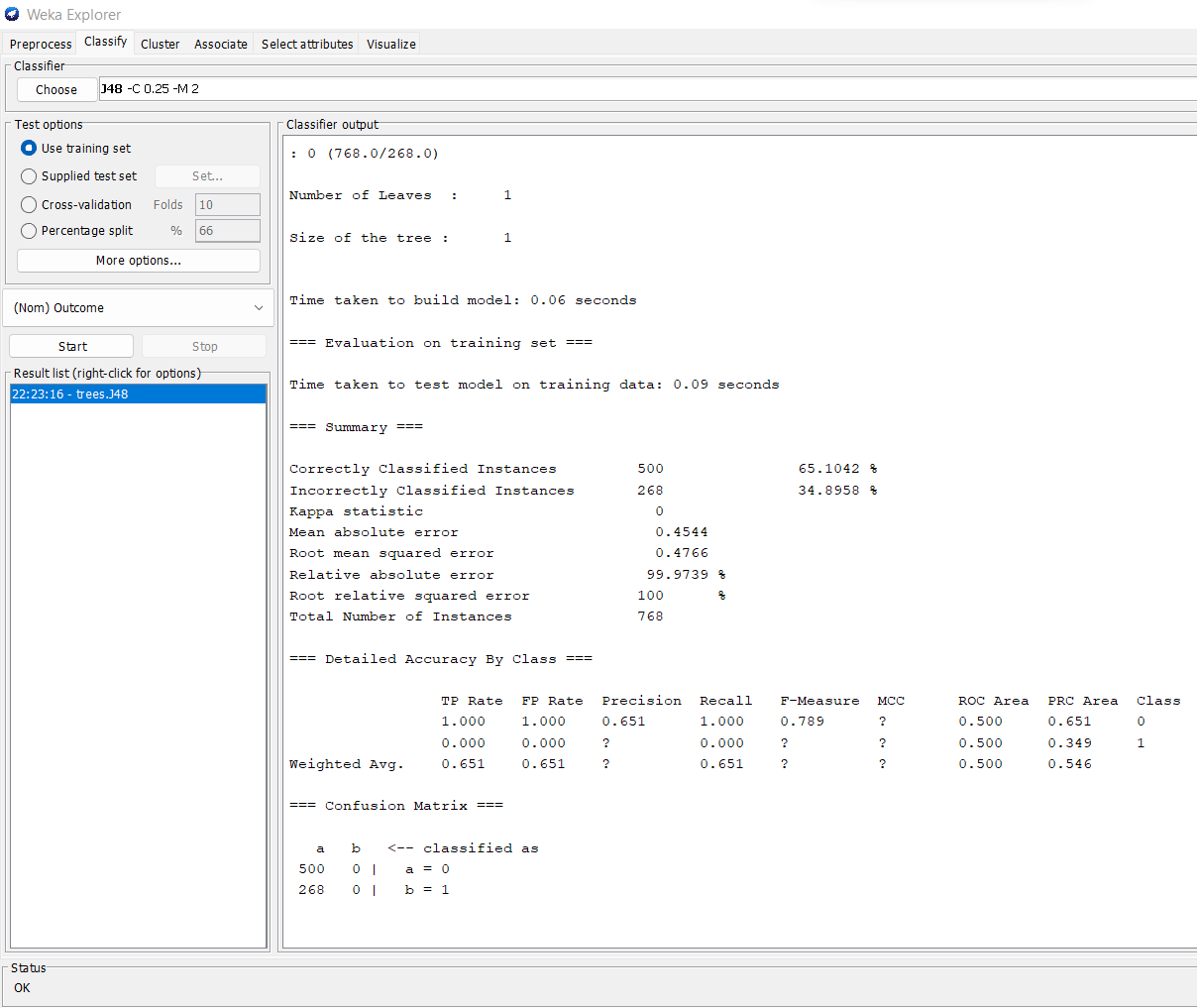
***Click the J48 tab in the algorithms and set the parameters according to the data and classes you have for my diabetes data I have set the values to the default:***

******

**Figure 21**

*Then choose the test option we used the training set and testing set where the training size is 66% and rest is the testing size click the start button and on the other side of the screen all the results will be shown.*

***The results are given below:***



**Figure 22**

***K Means:***

*It is iterative algorithms the partition the dataset into the subgroups or clusters unless they belong form the single group. It is the clustering algorithm which chooses two random clusters, and the distances are measured between all the clusters, so the closest distances clusters are placed together and other in the other cluster. So, like all we have two clusters.*

***Following are the steps followed:***

*Click the cluster tab on the top to apply cluster data mining algorithms.*

******

**Figure 23**

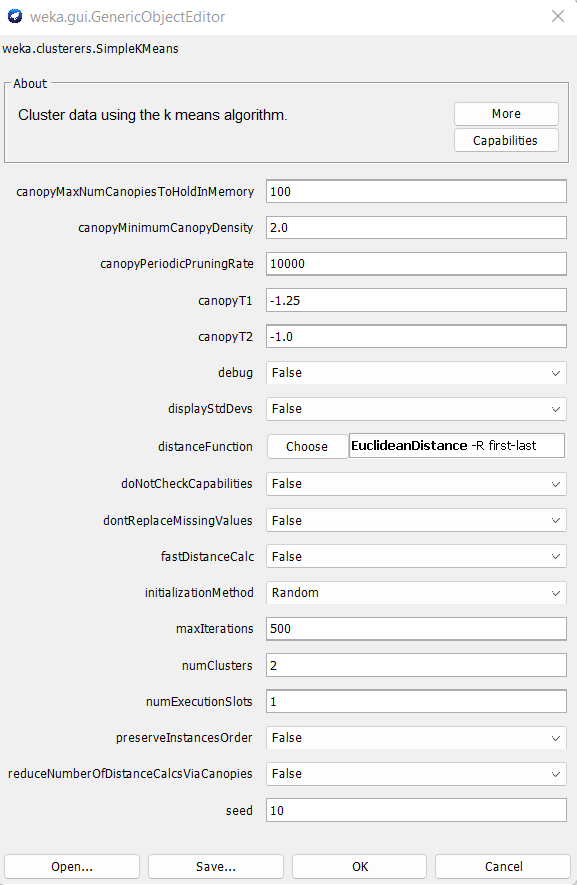
***Click the choose between to select the cluster algorithm:***

Graphical user interface, application

Description automatically generated

**Figure 24**

***Choosing the parameters for the clustering algorithm as in classification algorithms I have set all the parameters set to default so does here in clustering algorithms I have set it to default as well.***

******

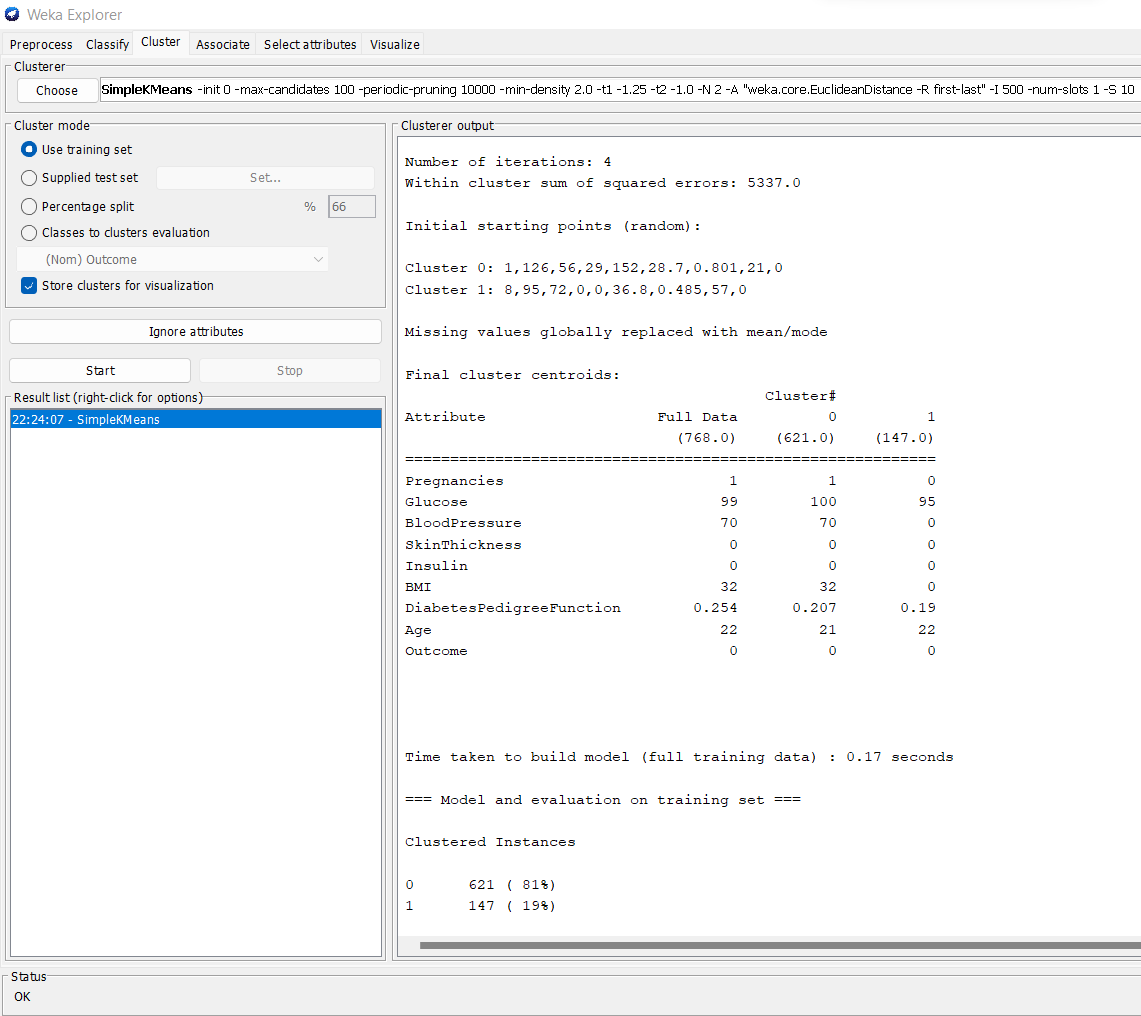
**Figure 25**

***I have used the same training and testing size of the clustering algorithm as I used in classification algorithm. 66% of training and rest is the testing size. Click start to run the algorithm and on the other side of the screen all the results will be shown.***

***Here are the results:***

***Total of two clusters:***

* *0 has 621 values (Cluster 1)*
* *1 has 147 values (Cluster 2)*



**Figure 26**

***Apriori:***

*Association rule mining is the data mining technique in which the rules are generated against the values. Rules are created as the most commonly occurring values in the data. Like the which element or number in the data is occurring mostly or right after any specific value inside the data. So those commonly occurred are made the rules that if A comes than there is chance that B will come right after that.*

***Following steps are followed:***

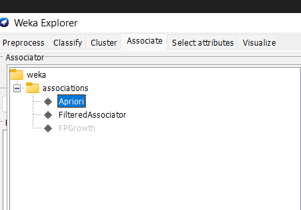
***Choose the association tab on the top to apply association algorithms:***

***Graphical user interface, text, application

Description automatically generated***

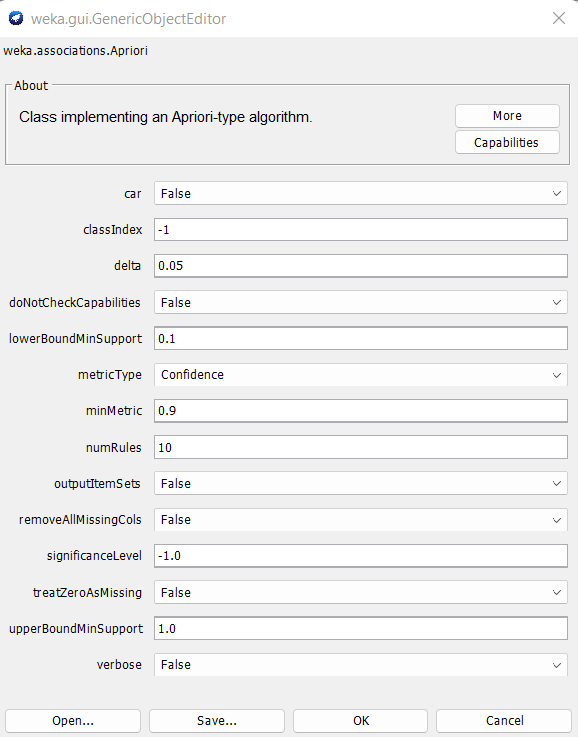
**Figure 27**

***Click to choose the algorithm:***

******

**Figure 28**

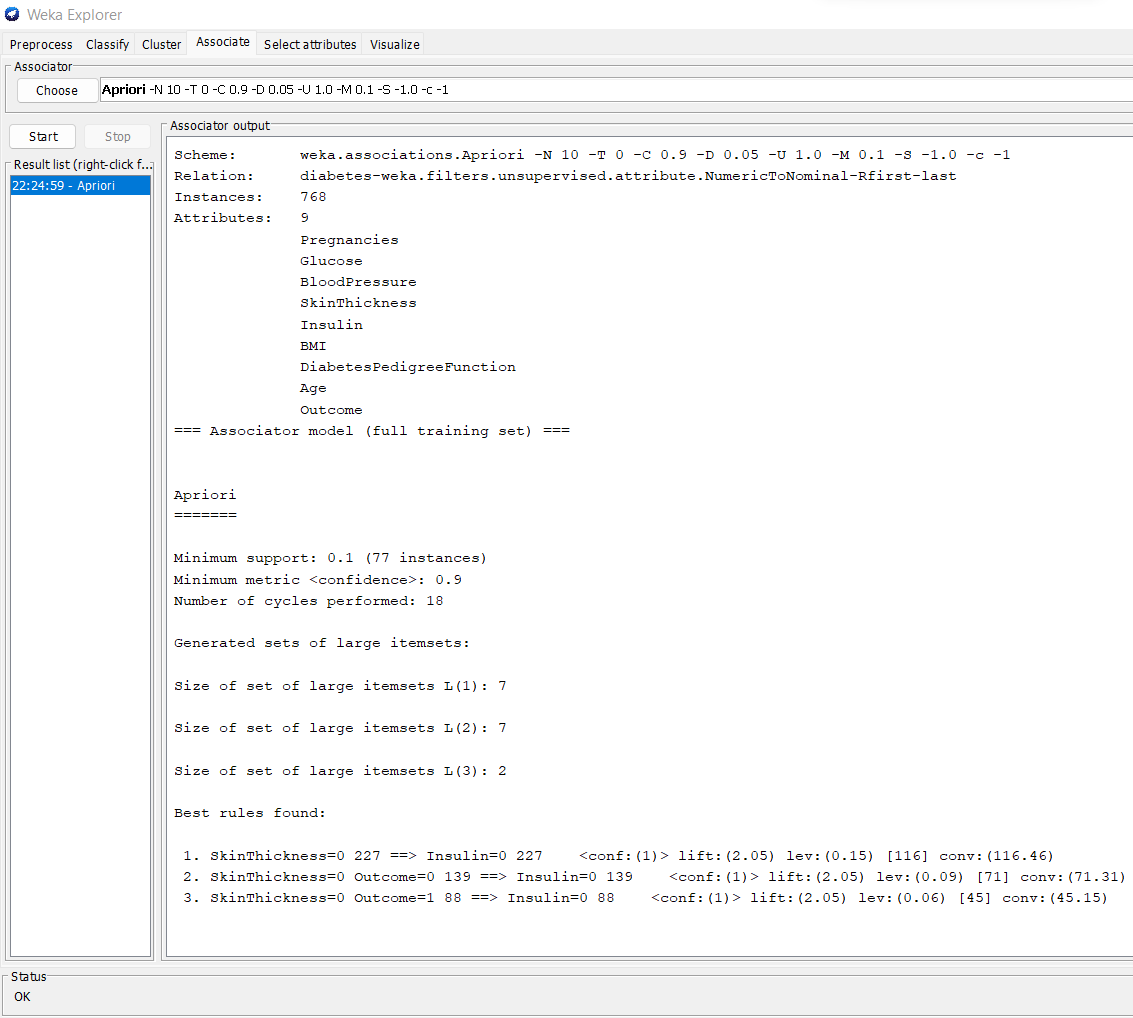
***The parameters are set as default as in all algorithms above:***

******

**Figure 29**

*Click the start button and on the other side of the screen all the attributes and the rules generated against those attributes.*

*There are total of 3 best rules generated for this diabetes data:*



**Figure 30**

(Ian H. Witten, 2016)*.*

***Discussion of Findings:***

*Data Analysis has quite an impact in modern era for the successful growth of any kind of business. This helps in identifying the different trends and patterns from the data. Data analysis finds and helps which model suits for the business what to sale what not to sale, when to sale and when not to sale, how much to sale all the things are not suggested by the data analysis by seeing the new and legacy systems that’s the property of data mining as well.* (Maheshwari, 2017)*. The data is taken from the Kaggle and downloaded as csv file as for the first requirement of WEKA was to convert from csv to arrf that is the file which is readable by WEKA. The data usually must go through the preprocessing methods missing values as in our data we had no missing values, dimension reduction was to remove the unnecessary columns if there are any, feature engineering includes the conversion from numeric to nominal for better usage of algorithms. Under sampling is done in case the class has the bias values and the 0 and 1 has the huge difference that can cause the results to be biased as well for that we used the filter and make the values equal for both 0 and 1. The data deals with missing values by converting the missing values into mean, mode or average of that column or row. The data did well in supervised and unsupervised models. As the data is classification data it produced better results on decision tree.*

*The data mining algorithms including classification, closeting and association were fit for the data produced most accurate results. All the results including classification reports, confusion matrix, accuracy of model, the confidence of the rules generated in apriori are discussed above defined the data very well. All these algorithms are used for the betterment of data, these data mining algorithms gives:*

* *Meaning and information of the data for the organization*
* *It helps saving the cost.*
* *It helps the organization to make profitable decision (Association)*

*The main findings are that it is observed through the graph and algorithms that the diabetes disease mostly fond in the older age’s patients. Another factor that affects the result is the glucose level. the most the glucose level of patient the higher the chance of patient to be the diabetic patient.*

# ***References***

*CHAUHAN, A. (2022, September 1). Predict Diabetes. Retrieved from Kaggle: https://www.kaggle.com/datasets/whenamancodes/predict-diabities?resource=download*

*Ian H. Witten, E. F. (2016). Data Mining: Practical Machine Learning Tools and Techniques. Elsevier Science.*

***Appendices:***

***Appendix A:***

*Data Choice*

***Appendix B:***

*Background information*

***Appendix C:***

*Data Description*

*C.1 Age*

*C.2 Pregnancy*

*C.3 Skin Thickness*

*C.4 Blood Pressure*

*C.5 Outcome*

***Appendix D:***

*Data Preprocessing*

*D.1 Removing Missing Values*

*D.2 Dimension reduction*

*D.3 Filter*

***Appendix F:***

*Data Mining*

*F.1 J48*

*F.2 K Means*

*F.3 Apriroi*

***Appendix G:***

*Discussion of finding*