Amal Jyothi College of Engineering
Kanjirappally, Kerala

STUDENT ALCOHOL CONSUMPTION USING WEKA

INTEGRATED MCA SEMINAR REPORT

Submitted in the partial fulfillment of the requirements for the Award of the Degree in

Integrated Master of Computer Applications

By

TINCYMOL M T

Reg No: AJC16MCA-I57

Under the Guidance Of

Ms. GRACE JOSEPH

January 2021

DEPARTMENT OF COMPUTER APPLICATIONS

AMAL JYOTHI COLLEGE OF ENGINEERING

KANJIRAPPALLY

CERTIFICATE

This is to certify that the seminar report, "**STUDENT ALCOHOL CONSUMPTION Using WEKA**" is the bonafide work of **TINCYMOL M T (Reg.No: AJC16MCA-I57)** in partial fulfillment of the requirements for the award of the Degree of Integrated Master of Computer Applications under APJ Abdul Kalam Technological University during the year *2020-21.*

Ms GRACE JOSEPH
Internal Guide

Ms. SONA MARIA SEBASTIAN
Coordinator

Fr. Rubin Thottupuram
Head of the Department

ACKNOWLEDGEMENT

First and foremost, I thank God almighty for his eternal love and protection throughout the seminar. I take this opportunity to express my gratitude to all who helped me in completing this seminar successfully. It has been said that gratitude is the memory of the heart. I wish to express my sincere gratitude to our manager **Rev. Fr. Dr. Mathew Paikatt** and Principal **Dr. Z V Lakaparampil** for providing good faculty for guidance.

I owe a great depth of gratitude towards our Head of the Department Fr. Rubin Thottupuram for helping us. I extend my whole hearted thanks to the seminar coordinators Fr. Rubin Thottupuram and Ms. Sona maria Sebastian for their valuable suggestions and for overwhelming concern and guidance from the beginning to the end of the seminar. I would also like to express sincere gratitude to my guide, Ms. Grace Joseph for her inspiration and helping hand.

I thank our beloved teachers for their cooperation and suggestions that helped me throughout the seminar. I express my thanks to all my friends and classmates for their interest, dedication, and encouragement shown towards the seminar. I convey my hearty thanks to my family for the moral support, suggestions, and encouragement to make this venture a success.

ABSTRACT

Educational data mining is the process of applying data mining tools and techniques to analyze data for educational purpose. This paper carries out educational data mining to study the student alcohol consumption through a public dataset which includes student attributes and their grades. The decision tree algorithm and the random forest algorithm are applied to perform classification and to analyze the variable importance. The regression model is then employed to illustrate the relationship between alcohol consumption level and the students' final grades. Our analysis provides knowledge on the relationship between student characteristics and alcohol consumption. The study also compares performance of the decision tree algorithm and the random forest algorithm.

In our society, there are number of problems arises due to the students consuming alcohol during its teen age. Retrieving exact and accurate students which consumes alcohol is the main task. It is a real-world problem in our society. The major challenge for finding alcohol addicted students with respect to given data is to find accurate and efficient method which takes less time to generate results. There is large amount of data available, but getting the right information accessible when needed is very important. The availability of educational data has been growing rapidly, and there is a need to analyse hedge amount of data generated from this educational ecosystem. Educational data mining (EDM) has been emerged as a process of applying data mining tools and techniques to analyse the data at educational institutions. This area of research is gaining

popularity due to potential benefits to the educational field. Educational institutions use educational data mining (EDM) to gain deep and through knowledge to enhance its assessment, evaluation, planning, and decision making in its educational programs. EDM helps academic programs to identify and discover hidden patterns in the data. These extracted patterns can be used for finding students who are consuming alcohol and its affect on their academic performance. In our proposed system we will use some educational institutes student's data and generate prediction weather student is alcohol addicted or not, we will do this by using clustering, classification, and filtering methods of data mining.

CONTENTS

1. INTRODUCTION

1.1 Waikato Environment for Knowledge Analysis (WEKA)

The foundation of any Machine Learning application is data - not just a little data but a huge data which is termed as Big Data. To train the machine to analyze big data, our big data needs lots of preprocessing. Once the data is ready, one can apply various Machine Learning algorithms such as classification, regression, clustering and so on to solve the problem.

The type of algorithms that we apply is based largely on our domain knowledge. We can test the different algorithms under the same class to build an efficient machine learning model. One can prefer visualization of the processed data and thus requires visualization tools. Weka is a software that accomplishes all the above with ease and lets you work with big data comfortably.

Waikato Environment for Knowledge Analysis, developed at the University of Waikato, New Zealand, is free software licensed under the GNU General Public License. Weka provides tools for data preprocessing, implementation of several Machine Learning algorithms, and visualization tools so that you can develop machine learning techniques and apply them to real-world data mining problems.

Weka supports several standard data mining tasks, more specifically, data preprocessing, clustering, classification, regression, visualization, and feature selection. All of Weka's techniques are based on the assumption that the data is available as one flat file or relation, where each data point is described by a fixed number of attributes (numeric or nominal attributes,

some other supported attribute types). Weka provides access to SQL databases using Java Database Connectivity and can process the result returned by a database query. Weka provides access to deep learning with Deeplearning4j. It is not capable of multi-relational data mining, but there is separate software for converting a collection of linked database tables into a single table that is suitable for processing using Weka.

Features and other information of WEKA includes,

- It is an open-source tool with Graphical User Interface in the form of "Explorer", "Experimenter" and "Knowledge Flow".
- Portability, since it is fully implemented in the Java programming language and thus runs on almost any modern computing platform.
- A comprehensive collection of data preprocessing and modelling techniques.
- It contains 49 data preprocessing tools, 76 classification and regressions algorithms, 8 clustering algorithms, 3 algorithms for finding association rule 15 attribute selection algorithms and 10 feature selection algorithms are present in WEKA.

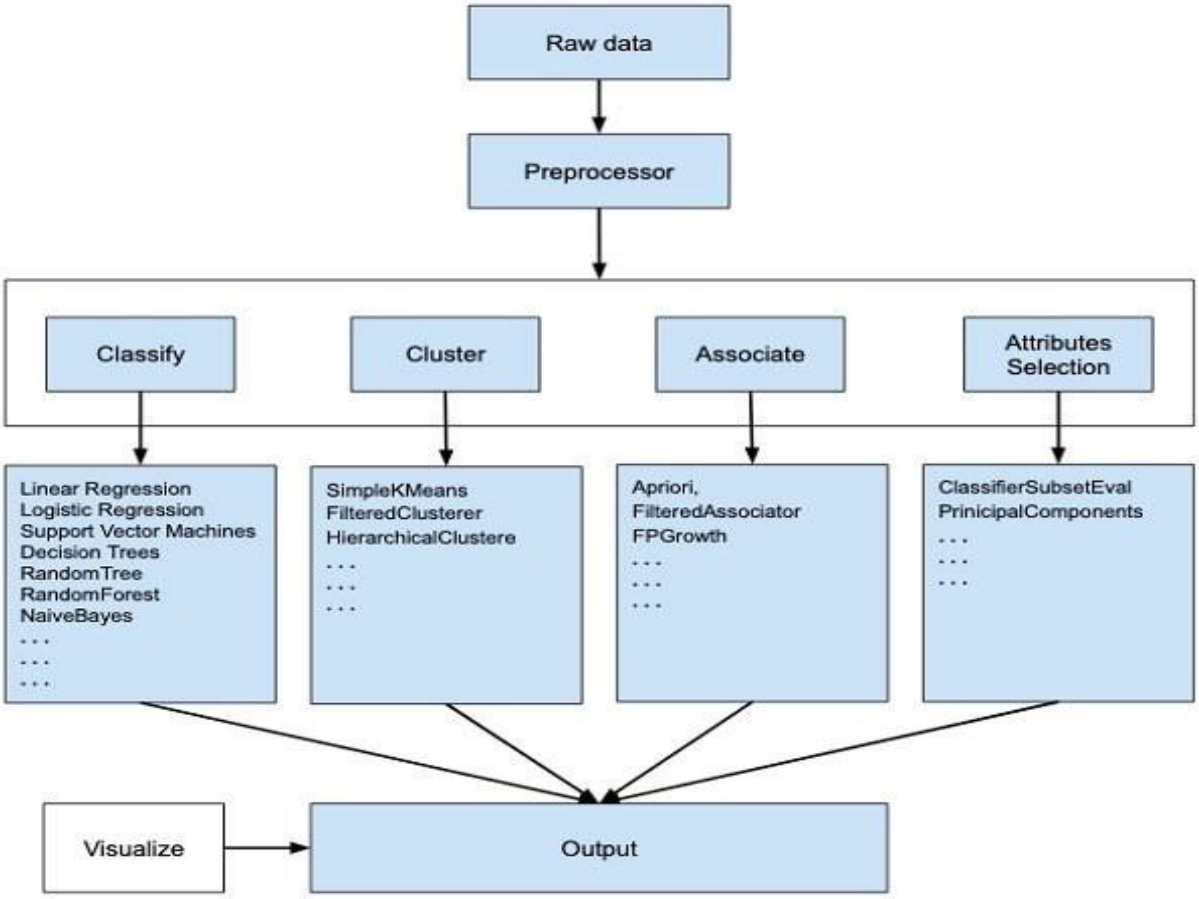
Using WEKA, users can develop custom code for machine learning.

1.2 History of Weka

- In 1993, the University of Waikato in New Zealand began development of the original version of Weka, which became a mix of Tcl/Tk, C, and Makefiles.
- In 1997, the decision was made to redevelop Weka from scratch in Java, including implementations of modeling algorithms.
- In 2005, Weka received the SIGKDD Data Mining and Knowledge Discovery Service Award.
- In 2006, Pentaho Corporation acquired an exclusive licence to use Weka for business intelligence. It forms the data mining and predictive analytics component of the Pentaho business intelligence suite. Pentaho has since been acquired by Hitachi Vantara, and Weka now underpins the PMI (Plugin for Machine Intelligence) open-source component.

1.3 Working of Weka Tool

What WEKA offers is summarized in the following diagram –



If we observe we understand that there are many stages in dealing with Big Data to make it suitable for machine learning –

1. You will start with the raw data collected from the field. We use the data preprocessing tools provided in WEKA to cleanse the (null/irrelevant) data. The preprocessed data its then stored in your local storage for applying ML algorithms.

2. Depending on the kind of ML model that we are trying to develop we would select one of the options - Classify, Cluster, or Associate. The Attributes Selection allows the automatic selection of features to create a reduced dataset.

3. Under each category, WEKA provides the implementation of several algorithms. We would select an algorithm of your choice, set the desired parameters and run it on the dataset.

4. We can also use WEKA to obtain the statistical output of the model processing. It provides you a visualization tool to inspect the data.

   The various models can be applied on the same dataset. You can then compare the outputs of different models and select the best that meets your purpose. Thus, the use of WEKA results in a quicker development of machine learning models on the whole.

2. STUDENT ALCOHOL CONSUMPTION USING j48 ALGORITHM WITH THE HELP OF WEKA.

It would be very difficult to manually go through the huge set of academic records to identify the student trends and behaviour and the pattern in which they learn. Educational data mining (EDM) has been emerged as a process of applying data mining tools and techniques to analyze the data at educational institutions. This area of research is gaining popularity due to potential benefits to the educational field. Educational institutions use educational data mining (EDM) to gain deep and through knowledge to enhance its assessment, evaluation, planning, and decision making in its

educational programs. EDM helps academic programs to identify and discover hidden patterns in the data. These extracted patterns can be used for finding students who are consuming alcohol and its affect on their academic performance

2.1 INTRODUCTION TO DECISION TREE ALGORITHM

(J48 ALGORITHM)

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. Another more advanced decision tree algorithm that we use here is the C4.5 algorithm, called J48 in Weka. J48

classifier 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. Decision tree J48 is the implementation of algorithm ID3 (Iterative Dichotomiser 3) developed by the WEKA project team. You can review a visualization of a decision tree prepared on the entire training data set by right clicking on the "Result list" and clicking "Visualize Tree".

Decision trees can support classification and regression problems. Decision trees are more recently referred to as Classification and Regression Trees (CART). They work by creating a tree to evaluate an instance of data, start at the root of the tree and moving town to the leaves (roots) until a prediction can be made. The process of creating a decision tree works by greedily selecting the best split point in order to make predictions and repeating the process until the tree is a fixed depth. After the tree is constructed, it is pruned in order to improve the model's ability to generalize to new data. Decision tree can be constructed moderately quick compare to other methods of classification. Decision tree classifiers obtain like or better accuracy when compare with other classification methods.

- The topmost node in the Decision tree is called the Root node
- The bottom-most node is called the Leaf node
- A node divided into sub-nodes is called a Parent node. The sub-

nodes are called Child nodes.

2.2 STEPS

Step 1: Download Weka and Install if it is not pre-installed.

Visit the Weka Download page and locate a version of Weka suitable for your computer (Windows, Mac, or Linux).

Step 2: Start Weka. This may involve finding it in program launcher or double clicking on the weka.jar file. This will start the Weka GUI Chooser. The Weka GUI Chooser lets you choose one of the Explorer, Experimenter, KnowledgeFlow and the Simple CLI (command line interface).

Step 3: Open the data/mushroom.arff Dataset. Click the **"Open file..."** button to open a data set and double click on the **"data"** directory. Weka provides a number of small common machine learning datasets that you can use to practice on. Select the "**student-mat.csv**" file to load the Iris dataset.

Step 4: Select and Run an Algorithm. Now that you have loaded a dataset, it's time to choose a machine learning algorithm to model the problem and make predictions. Click the "**Classify**" tab. This is the area for running algorithms against a loaded dataset in Weka. You will note that the "**ZeroR**" algorithm is selected by default. Here we use "**J48**" algorithm, so select the "**J48**" algorithm from tree option. Click the "Start" button to run this algorithm.

Step 5: Review Results. After running the "**J48**" algorithm, you can note the results in the "Classifier output" section and the presented result is a summary of those predictions.

2.3 Features

- It handles classification with the missing values in the data.
- It can be applied to both discrete and continuous variables.
- It also performs the pruning of the tree.
- It can handle high dimensional data.
- It replaces internal node with a leaf node and thus reduces the error rate.

2.4 Limitations

- A small change in the data can cause a large change in the structure of the decision tree causing instability.
- For a Decision tree sometimes, calculation can go far more complex compared to other algorithms.
- Decision tree often involves higher time to train the model.
- Decision tree training is relatively expensive as the complexity and time has taken are more.
- The Decision Tree algorithm is inadequate for applying regression and predicting continuous values.

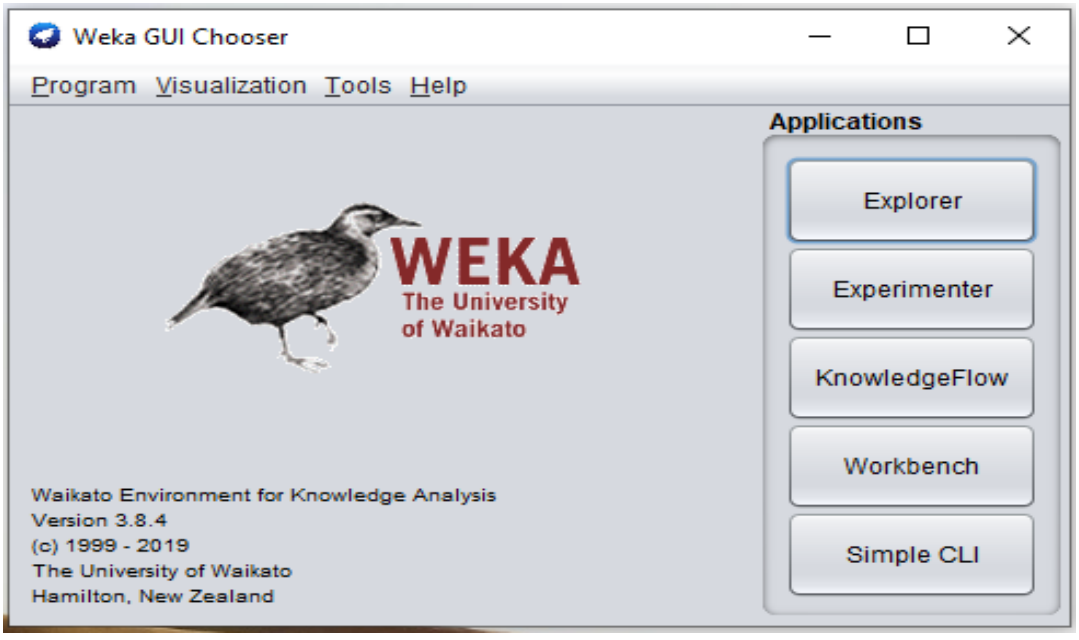
3. DEMONSTRATION OF STUDENT ALCOHOL CONSUMPTION USING WEKA GROUPING USING j48 CLASSIFY ALGORITHMS WITH THE HELP OF WEKA.

i. Download the installation file from WEKA's official website.

ii. Double click on the downloaded file to run installation.

iii. Click on the Weka 3.8.4 icon to start Weka.

The WEKA GUI Chooser application (allows to run five different types of applications - Explorer, Experimenter, KnowledgeFlow, Workbench, Simple CLI) will start and you would see the following screen –



a) **Explorer:** The Weka Knowledge Explorer is an easy-to-use graphical user interface that harnesses the power of the weka software.

b) **Experimenter:** The Weka Experiment Environment enables the user to

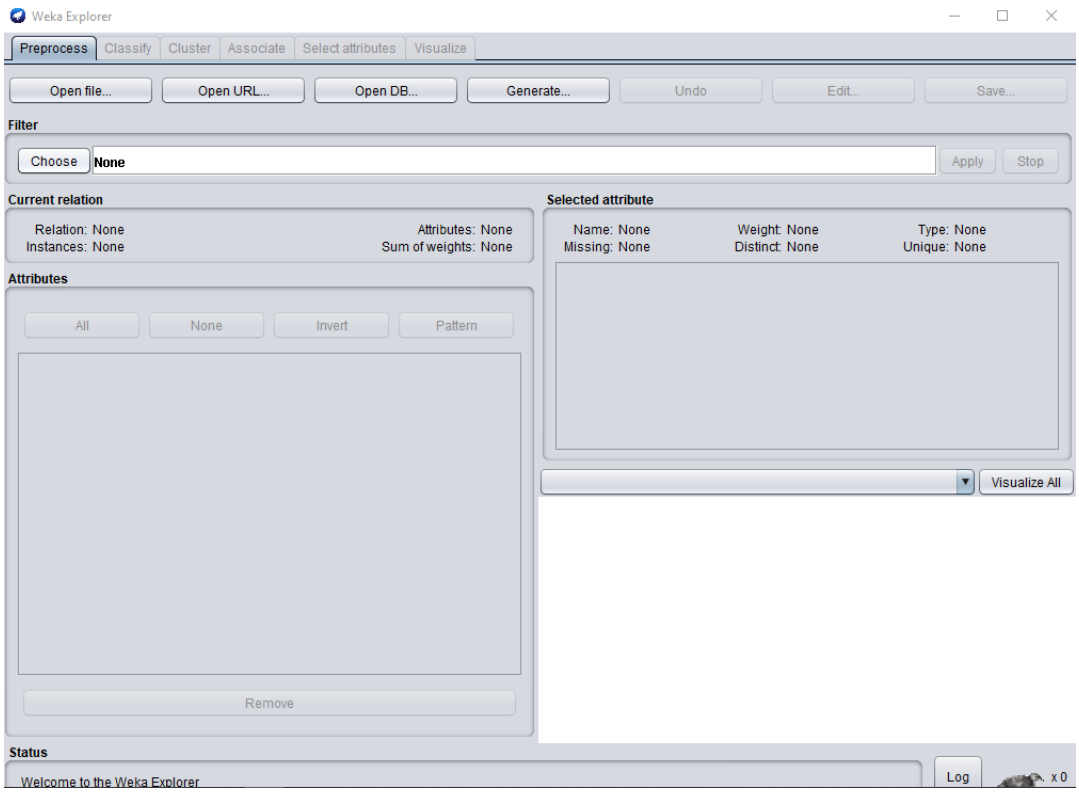
create, run, modify, and analyse experiments in a more convenient manner than is possible when processing the schemes individually.

c) **Knowledge Flow**: The Knowledge Flow Interface is an alternative to the Explorer, and it lets you lay out filters, classifiers, and evaluators interactively on a 2D canvas.

d) **Workbench:** The Weka Workbench is an environment that combines all of the GUI interfaces into a single interface. It is useful if you find yourself jumping a lot between two or more different interfaces, such as between the Explorer and the Experiment Environment.

e) **Simple CLI:** Simple CLI is a simple command line interface provided to run Weka functions directly.

IV. Here I have used explorer. When you click on the Explorer button in the Applications selector, it opens the following screen –



When you open the explorer, on the top you will see several tabs listed as below.

- **Preprocess:** Data preprocessing is a data mining technique which is used to transform the raw data in a useful and efficient format.
- **Classify:** The classifier panel allows you to configure and execute any of the weka classifiers on the current dataset. You can choose to perform a cross validation or test on a separate dataset. Classification errors can be visualized in a pop-up data visualization tool. If the classifier produces a decision tree it can be displayed graphically in a pop-up tree visualizer.
- **Cluster:** From the cluster panel you can configure and execute any of the weka clusters on the current dataset. Clusters can be visualized in a pop-up data visualization tool.
- **Associate:** From the associate panel you can mine the current dataset for association rules using the weka associators.
- **Select Attributes:** This panel allows you to configure and apply any combination of weka attribute evaluator and search method to select the most pertinent attributes in the dataset. If an attribute selection scheme transforms the data then the transformed data can be visualized in a pop-up data visualization tool.
- **Visualize Panel:** This panel displays a scatter plot matrix for the current dataset. The size of the individual cells and the size of the points they display can be adjusted using the slider controls at the bottom of the panel**.**

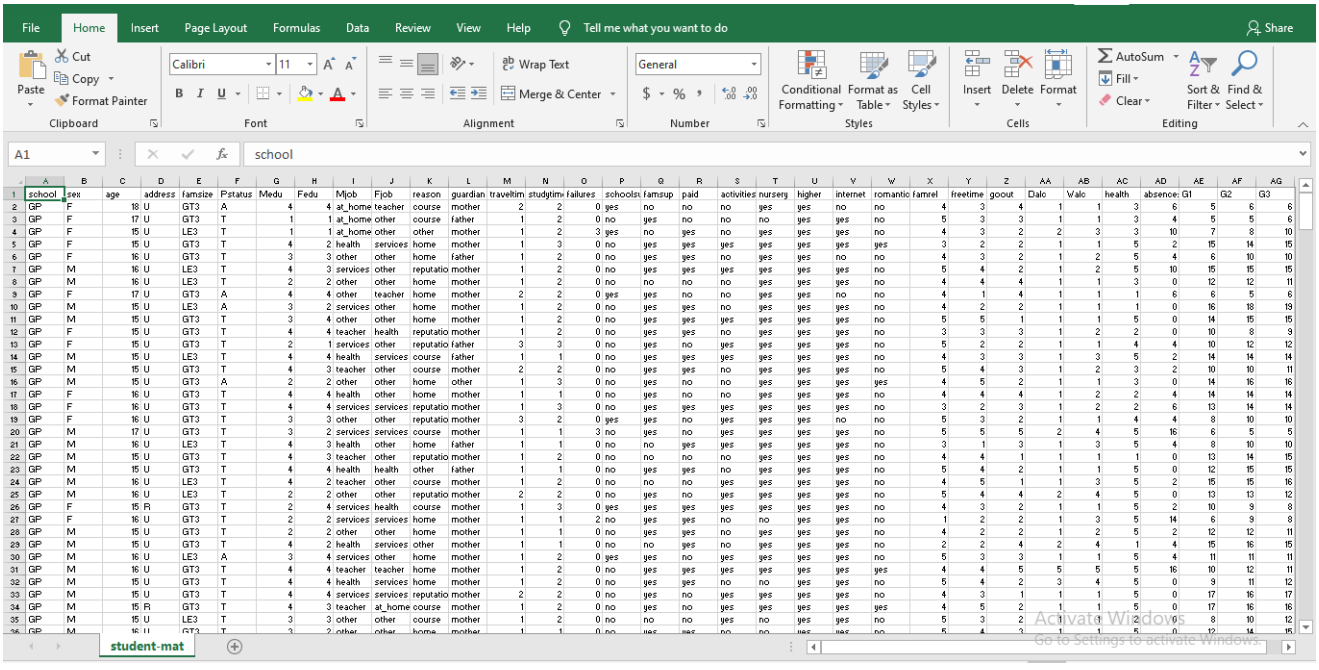
The WEKA Explorer windows show different tabs starting with preprocess (which is active) as the data set is first preprocessed before applying algorithms to it.

Each of this tab contain several pre implemented machine learning algorithms. Now we want to open the data set.

There are three ways to implement this.

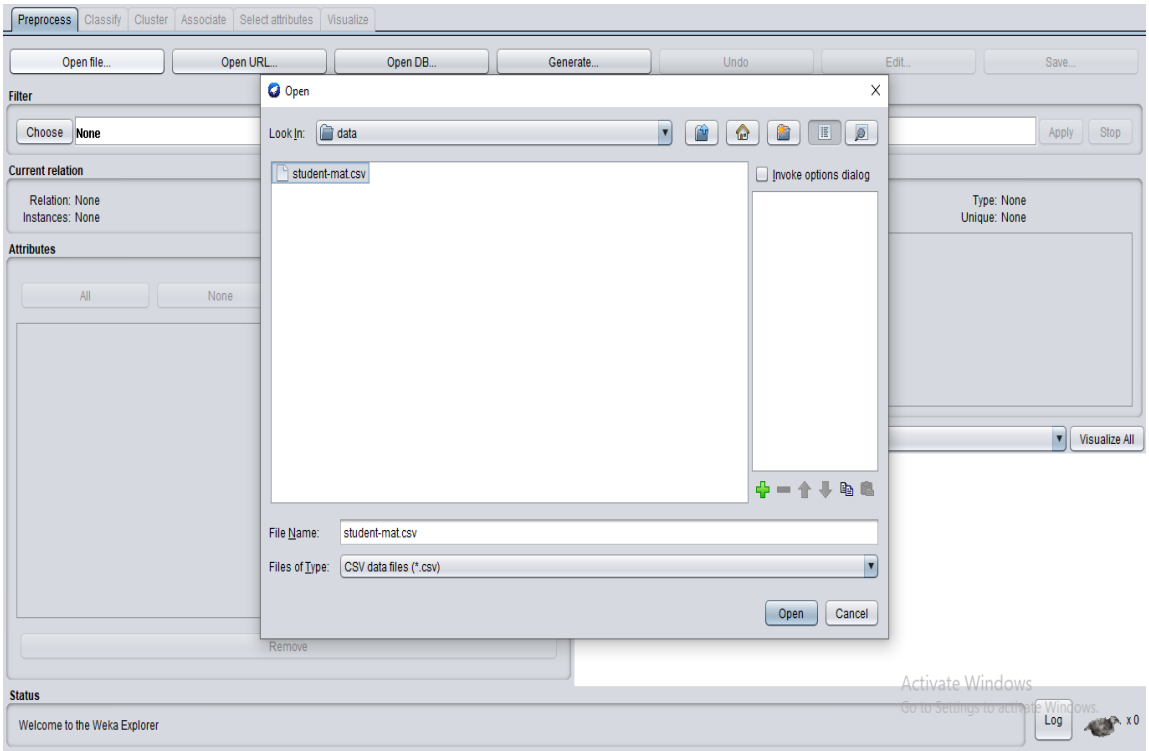
- Open file
- Open URL
- Open DB

V. To load the data from the local file system, click on the Open file button under the Preprocess tab. A directory navigator window opens as shown in the following screen –

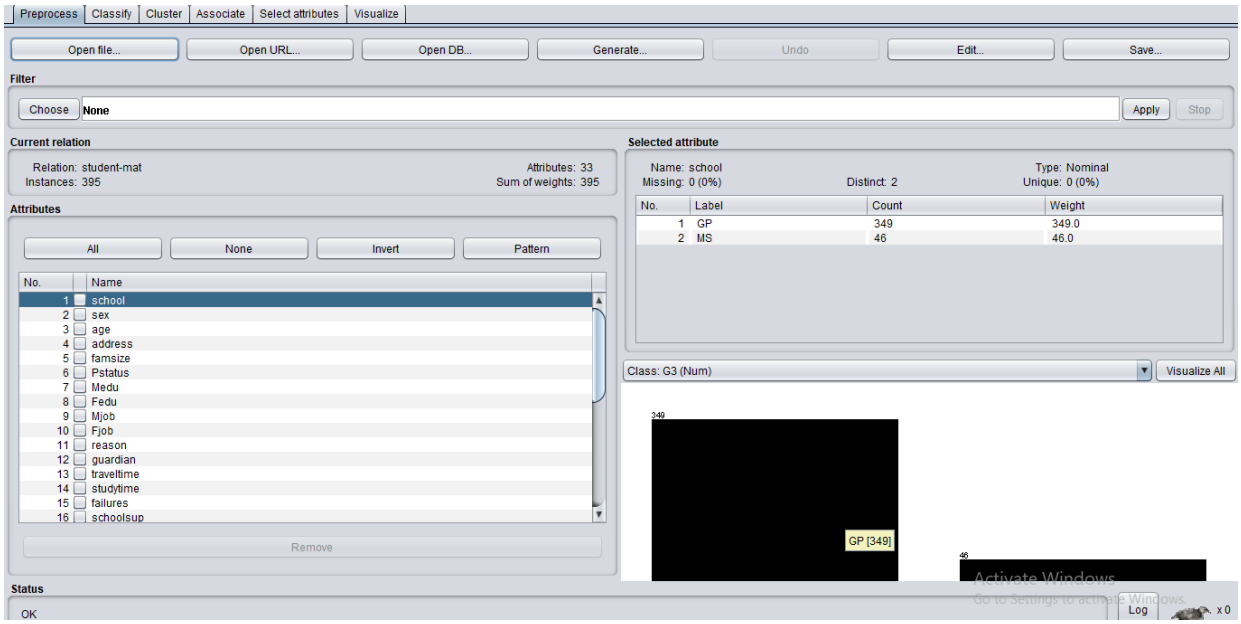


dataset

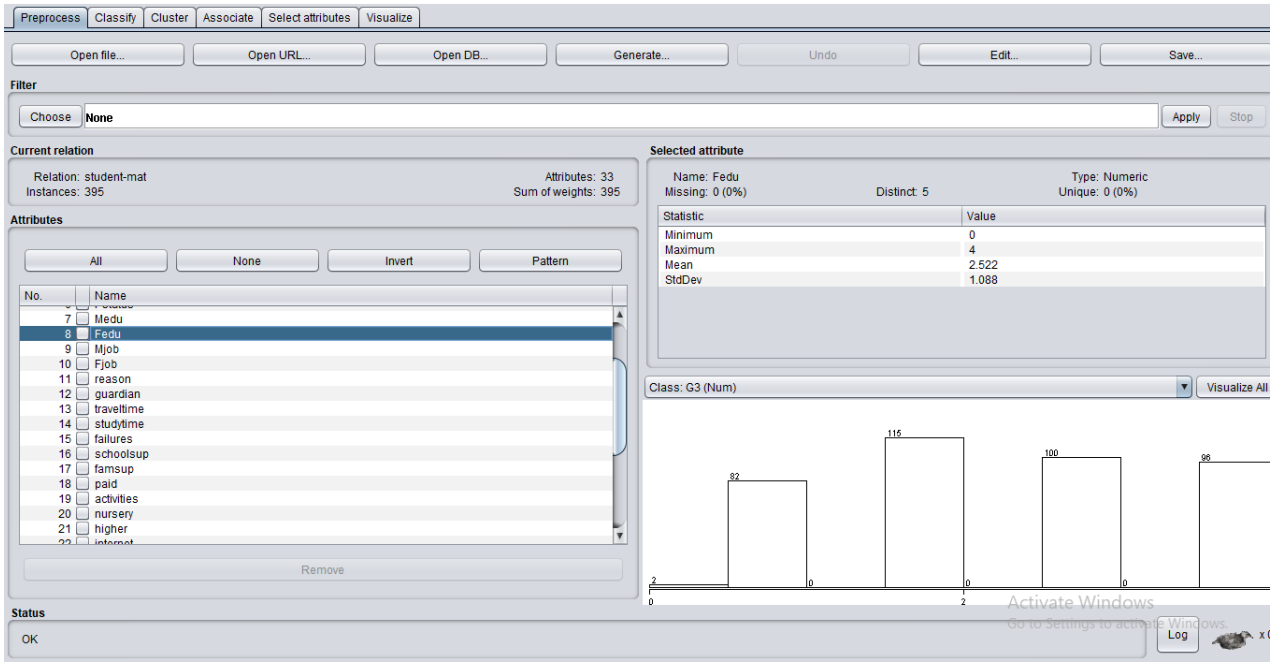
Attribute	Description (Domain)
sex	student's sex (binary: female or male)
age	student's age (numeric: from 15 to 22)
school	student's school (binary: *Gabriel Pereira* or *Mousinho da Silveira*)
address	student's home address type (binary: urban or rural)
Pstatus	parent's cohabitation status (binary: living together or apart)
Medu	mother's education (numeric: from 0 to 4[a])
Mjob	mother's job (nominal[b])
Fedu	father's education (numeric: from 0 to 4[a])
Fjob	father's job (nominal[b])
guardian	student's guardian (nominal: mother, father or other)
famsize	family size (binary: ≤ 3 or > 3)
famrel	quality of family relationships (numeric: from 1 – very bad to 5 – excellent)
reason	reason to choose this school (nominal: close to home, school reputation, course preference or other)
traveltime	home to school travel time (numeric: 1 – $<$ 15 min., 2 – 15 to 30 min., 3 – 30 min. to 1 hour or 4 – $>$ 1 hour).
studytime	weekly study time (numeric: 1 – $<$ 2 hours, 2 – 2 to 5 hours, 3 – 5 to 10 hours or 4 – $>$ 10 hours)
failures	number of past class failures (numeric: n if $1 \leq n < 3$, else 4)
schoolsup	extra educational school support (binary: yes or no)
famsup	family educational support (binary: yes or no)
activities	extra-curricular activities (binary: yes or no)
paidclass	extra paid classes (binary: yes or no)
internet	Internet access at home (binary: yes or no)
nursery	attended nursery school (binary: yes or no)
higher	wants to take higher education (binary: yes or no)
romantic	with a romantic relationship (binary: yes or no)
freetime	free time after school (numeric: from 1 – very low to 5 – very high)
goout	going out with friends (numeric: from 1 – very low to 5 – very high)
Walc	weekend alcohol consumption (numeric: from 1 – very low to 5 – very high)
Dalc	workday alcohol consumption (numeric: from 1 – very low to 5 – very high)
health	current health status (numeric: from 1 – very bad to 5 – very good)
absences	number of school absences (numeric: from 0 to 93)
G1	first period grade (numeric: from 0 to 20)
G2	second period grade (numeric: from 0 to 20)
G3	final grade (numeric: from 0 to 20)



Navigate to folder where the files are stored. When you open the file, your screen looks like as shown here –

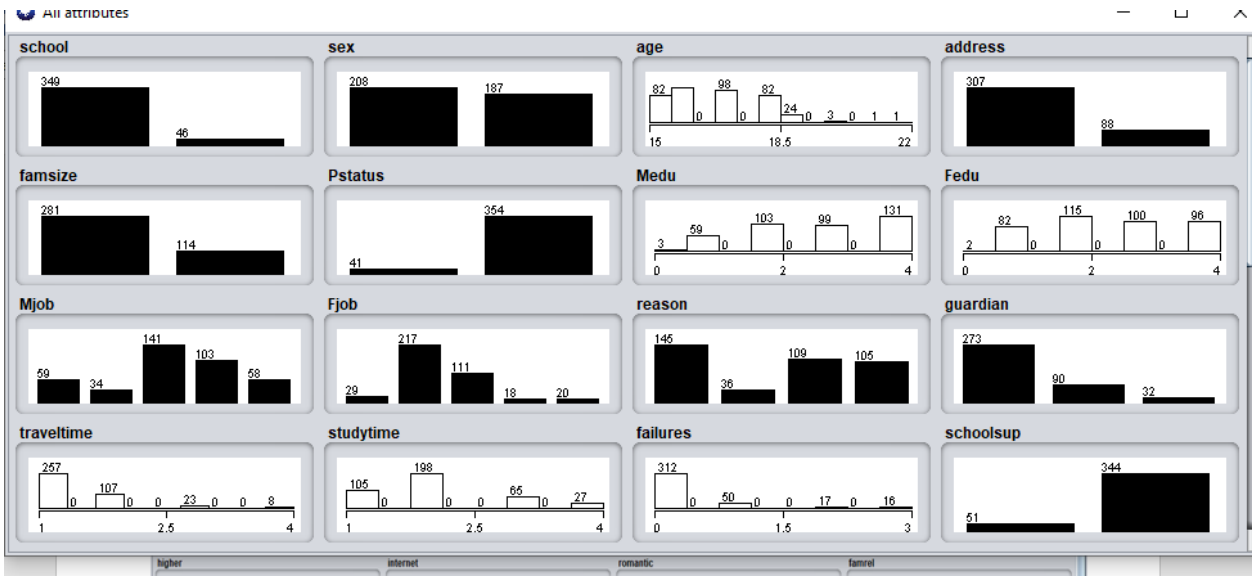


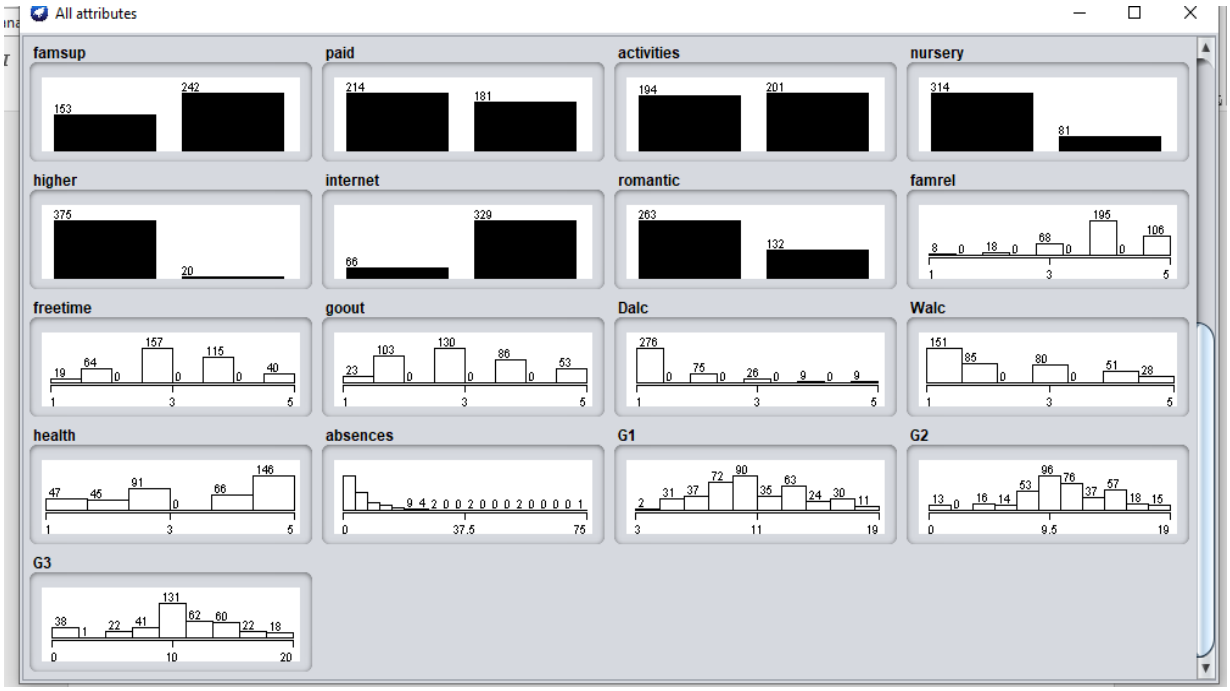
VI. The student database contains 33 field. We can select an attribute from the list, while clicking on its further details of the attributes are shown on the right-hand side in the window.



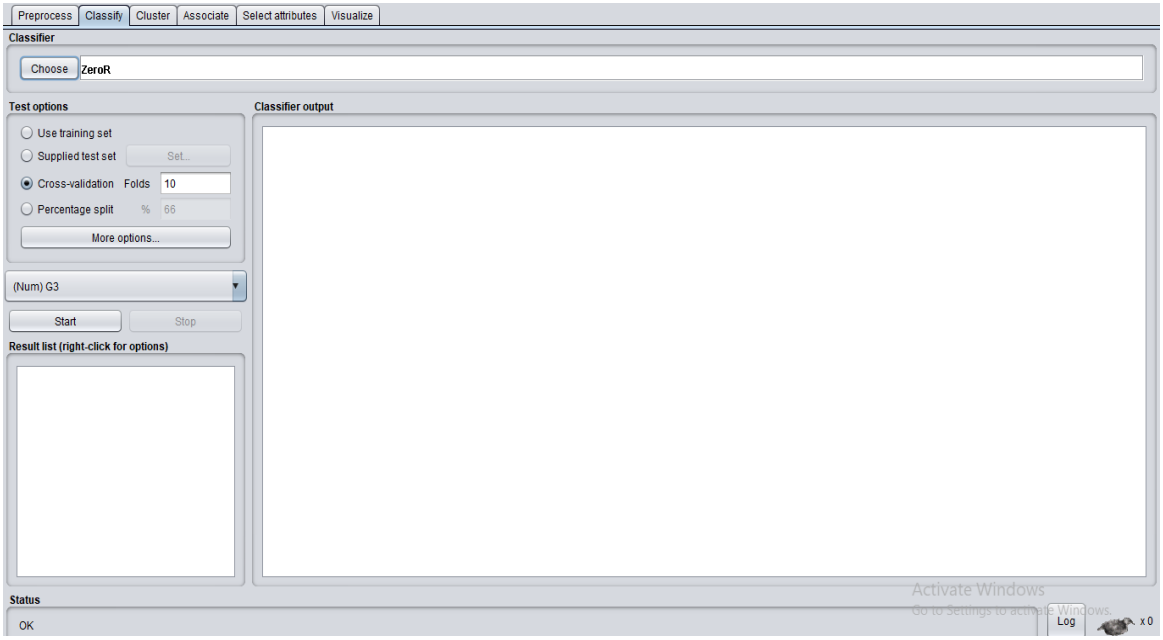
In the right-hand side of the window, we can observe several thinks about the dataset. Such as name and type of the attribute, the number of distinct value and also choose the count and weight in the terms of percentage.

VII. In the right-hand side of the window there is an option to visualize of the graph. Click on visualize all button.it is as shown below.





a. click on the classify tab and click on tools button.



There are several text options available. They are listed below.

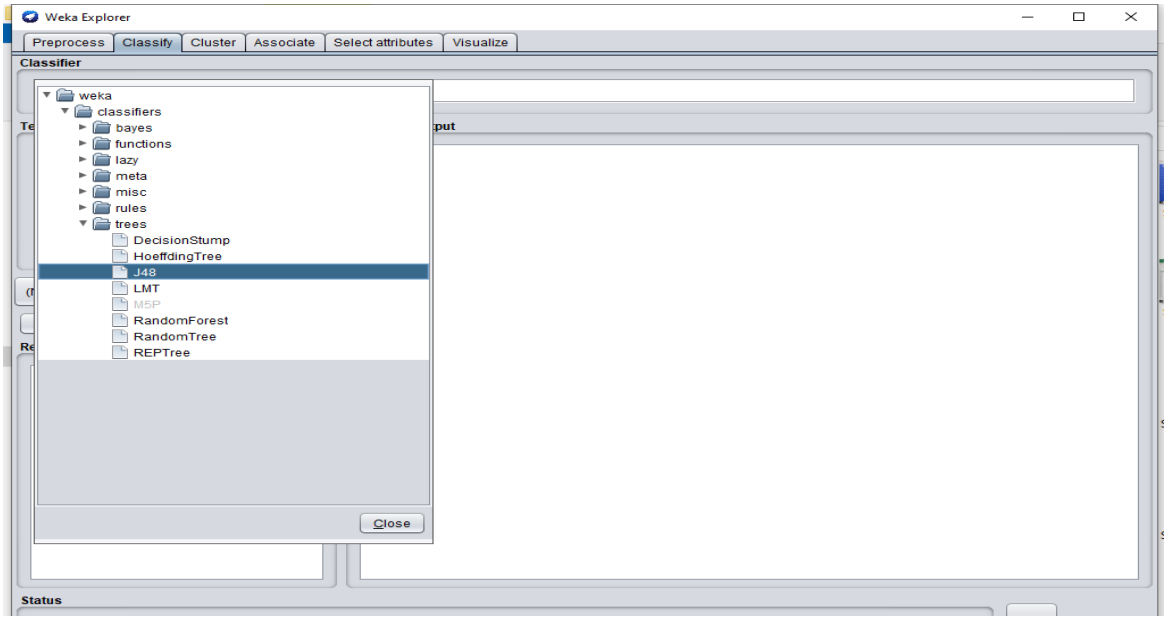
- **Use training set:** Classifies your model based on the dataset which you originally trained your model with.
- **Supplied test set**: Controls how your model is classified based on the dataset you supply from externally. Select a dataset file by clicking the Set button.
- **Cross-validation:** The cross-validation option is a widely used one, especially if you have limited number of datasets. The number you enter in the *Fold* section are used to divide your dataset into Fold numbers (let's say it is 10). The original dataset is randomly partitioned into 10 subsets. After that, Weka uses **set 1** for testing and **9 sets** for training for the first training, then uses **set 2** for testing and the **other 9 sets** for training, and repeat that 10 times in total by incrementing the set number each time. In the end, the average success rate is reported to the user.

- **Percentage split:** Divide your dataset into train and test according to the number you enter. By default, the percentage value is 66%, it means 66% of your dataset will be used as training set and the other 33% will be your test set.

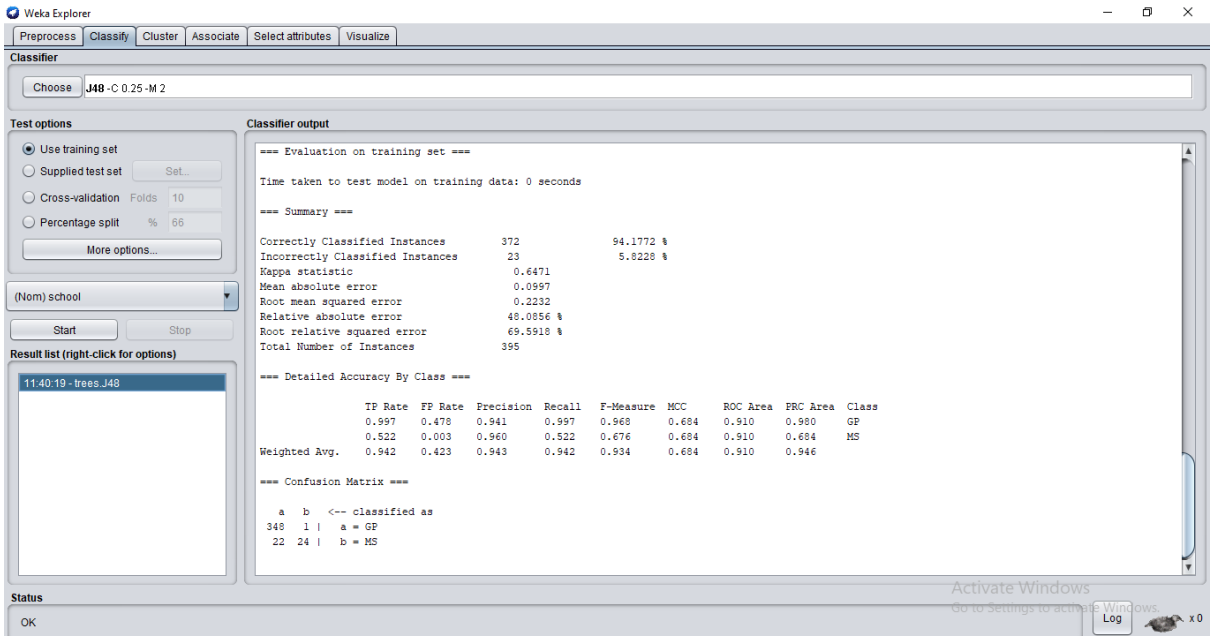
Next,you have to select the classifier for that click on the choose button and select the following classifier.

Weka->classify->tree->J48

Here we choose J48 (decision tree algorithm) for training our data set. This is shown the screenshot below.



Click on the start button to start the classification process. After some time, the classification result is shown on the screen as below.



Here Run *Information* will give detailed results. gives detailed information about the dataset and the model used. J48 is a decision tree. The *Classifier Model* part illustrates the model as a tree and gives some information about the tree, like number of leaves, size of the tree, etc. Next is the s*tratified cross-validation* part and it shows the error rates. By checking this part, you can see how successful your model is. You can see a *Confusion Matrix* and detailed *Accuracy Table* at the bottom of the report. F-Measure and ROC Area rates are important for the models and they are developed according to a confusion matrix. A confusion matrix represents the True Positive, True Negative, False Positive and False Negative rates. The algorithm was run with 10-fold cross-validation: this means it was given an opportunity to make a prediction for each instance of the dataset (with different training folds) and the presented result is a summary of those predictions.

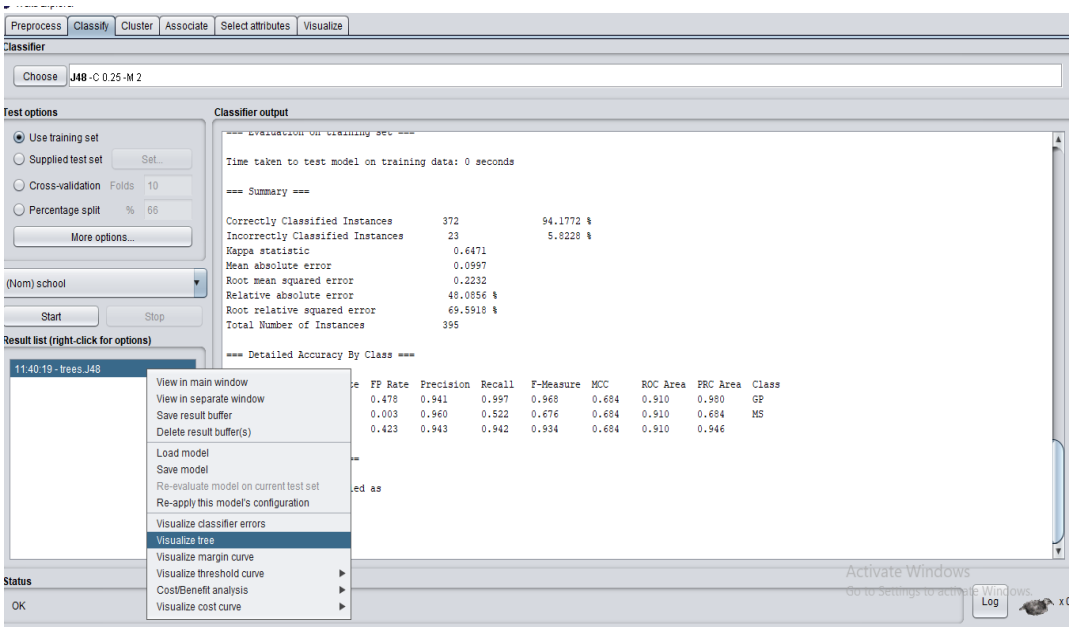
The total number of instances are 395. the output shows 372 Correctly Classified Instances and 23 Incorrectly Classified Instances. The Kappa statistic (The kappa statistic is frequently used to test interrater reliability.) is 0.6471 and the Mean

absolute error (the mean absolute error is an average of the absolute errors, where is the prediction and the true value.) is 0.0997.the Root mean squared error (It represents the sample standard deviation of the differences between predicted values and observed values) is 0.2232. Relative absolute error and Root relative squared error are 48.0856 %, 69.5918 %.

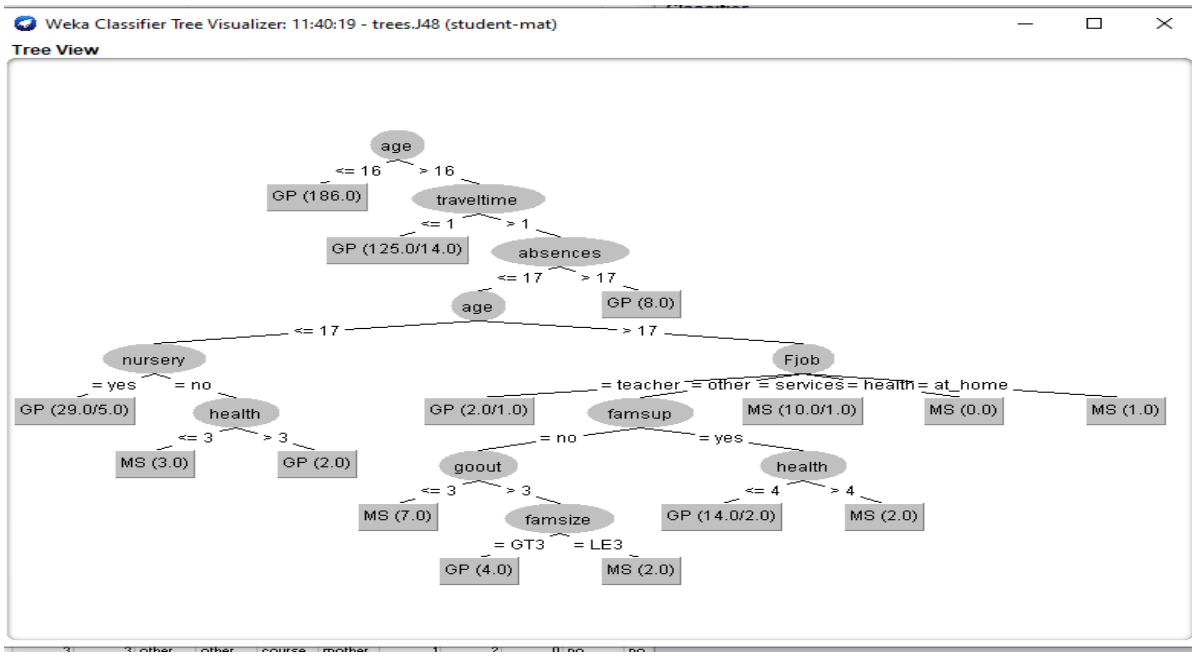
Confusion Matrix

True Positive means that the actual class of classification is equal to the class your model set. The dataset which consists of student alcohol consumption, some of them have consuming alcohol some of them don't. You're creating a model to student alcohol consumption automatically in terms of their academic's results. So, you give a student information to your model and your model's result shows the student academics performance. In this case, our positive result is being healthy and negative result is finding student alcohol consumption affect their academic performance.

To see the visual representation of the results, right click on the result in the Result list box. Several options would pop up on the screen as shown here –



Select **Visualize tree** to get a visual representation of the traversal tree as seen in the screenshot below:



4.NEW PROBLEM CONCEPT

5. CONCLUSION

Through this study we are acknowledged with the significant factors that contribute in the indulgence of teenagers in alcoholic activities and affect their academic performance in secondary schools.

WEKA is a powerful tool for developing machine learning models. It provides implementation of several most widely used ML algorithms. Before these algorithms are applied to your dataset, it also allows you to preprocess the data. The types of algorithms that are supported are classified under Classify, Cluster, Associate, and Select attributes. The result at various stages of processing can be visualized with a beautiful and powerful visual representation. This makes it easier for a Data Scientist to quickly apply the various machine learning techniques on his dataset, compare the results and create the best model for the final use.

Thus, Weka is a comprehensive software that lets you to preprocess the big data, apply different machine learning algorithms on big data and compare various outputs. This software makes it easy to work with big data and train a machine using machine learning algorithms.

6. REFERENCES

- https://www.tutorialspoint.com/weka/weka_quick_guide.htm
- https://www.softwaretestinghelp.com/weka-tutorial/
- https://youtu.be/Z4VZsF96QfU
- https://youtu.be/HCA0Z9kL7Hg
- https://www.kaggle.com/uciml/student-alcohol-consumption
- "Understand why children drink alcohol", Drinkaware.co.uk, https://www.drinkaware.co.uk/advice/underagedrinking/understand-why-children-drink-alcohol/