Performance Analysis of Students using weka tool

Abstract— In recent years the amount of data stored in educational database is growing rapidly. The stored database contains hidden information which if used aids improvement of student's performance and behaviour. In this paper predictive modelling approach is used for extracting this hidden information. Data is collected, a predictive model is formulated with concepts of clustering, classification, association rule generation, predictions are made, and the model is validated as additional data becomes available. The predictive models will help the instructor to understand how well or how poorly the students in his/her class will perform, and hence the instructor can choose proper pedagogical and instructional interventions to enhance student learning outcomes.

1.INTRODUCTION

An educational institution needs to have an approximate prior knowledge of enrolled students to predict their performance in future academics. This helps them to identify promising students and also provides them an opportunity to pay attention to and improve those who would probably get lower grades. There is no absolute scale for measuring knowledge but examination score is one scale which shows the performance indicator of students. So the performance influencing factors are taken into account and rules are mined in order to predict the performance of the students.

There are various factors that are to be taken into account in order to predict the student's performance such as first generation cadet, medical history, etc.. So this tool is helpful in predicting the performance of the students' considering all the influential factors. A rule based methodology is provided by using appropriate algorithm, which provides insight to students to enhance their results and for tutors to provide a better coaching to students. The main objective of the project is to identify academically at risk student in order to give them special coaching and to predict the student's result.

Data mining is the process by which new relationship and facts are discovered among data which will be useful in knowledge mining. It is used to find out the hidden relationship among data sets and attributes. In data mining, clustering is the process by which a data set is separated into groups called clusters, it is done in such a way that all the data points that are having similar type of values lies in same cluster, based on the distribution cluster centres are created. In the field of data mining clustering plays very important role.

The phase of clustering is followed by the labelling of the clusters. The range of the marks are assigned to different cluster based on the numerical value of the cluster centre. After this labelling the marks that are in nominal values are changed to ordinal values. Then this data set is fed into weka tool for the purpose of association rule mining. Apriori algorithm is used. Apriori is an algorithm for frequent item set mining and association rule learning. Apriori algorithm uses bottom up approach where frequent subsets are extended one at a time. This step is known as candidate generation. Apriori uses breadth-first search and a Hash tree structure to count candidate item sets efficiently. Association rules are mined with the help of this algorithm. The rules that are generated with confidence level more than 0.9 are taken into consideration or deriving knowledge. After this step, the dataset with numerical values are again fed into the weka tool which predicts the performance of the third test of the students. The algorithm used here is Linear Regression algorithm to identify relationship between dependent and independent variables in training dataset. Once the relationships between dependent and independent variables are found, then a linear regression model is created. Accuracy is based on consideration of various factors and attributes.

2. **LITERATURE REVIEW**

Ishwank Singh, Sai Sabitha, Abhay Bansal are research scholars of Amity University Uttar Pradesh, Noida, proposed a system for analysing student performance using K-Means Clustering. In this project the data for a section of Computer science student of the class of

2012–2016 was considered. Data pre-processing was done by applying suitable weights to relevant attributes and removing the missing, irrelevant data .They applied K-means algorithm on the pre-processed data set for obtaining the cluster. Rapid Miner studio was used for doing analysis. Using this analysis the clusters were labelled and classified into different categories based on performance of the students in the test. Academic Performance, Co-curricular activities and Overall performance of the students are analysed. Comparative plots and analysis are done.

Bishal Dey Sarkar,Sonali Shankar proposed a system for Performance analysis of student learning metric using K-mean clustering approach. They collected the data set released by The Harvard University regarding their students who registered for the online course for year AY2013 on May 14, 2014. The paper aims to analyse the performance of Harvard students registered for online course based on average grades and other attributes of students belonging to different countries. The raw set of data was preprocessed and it is fed into the data mining tool TANAGRA. K-Means clustering is used. The K-value for random cluster generation is deduced with the help of Silhouette Index (SI). the preprocessed data set is fed to Cluster engine to obtain different values of WSS. Within Cluster sum of Square (WSS) denotes the compactness between the points of the different cluster. Clusters with minimum WSS value is taken for analysis and thus The average performance of the students belonging to different countries is analysed.

Kartika Maharani, Teguh Bharata Adji and Noor Akhmad Setiawan published a paper in Yogyakarta, Indonesia. In order to improve the academic performance of Indonesian students a promising step to implement the student's academic performance predictive factor applying Educational Data Mining (EDM) (as it is able to transform unnoticed data to useful one)was done. Predictive factors affecting academic performance of students can be implemented with classification technique. Classification creates a model concluding new class label from input combination of other aspects. Statistical classification classifies groups of individual aspects based on characteristic information in a subject. Students' characters are taken from demographic factor, social, family support, psychological factor etc.. They can be considered

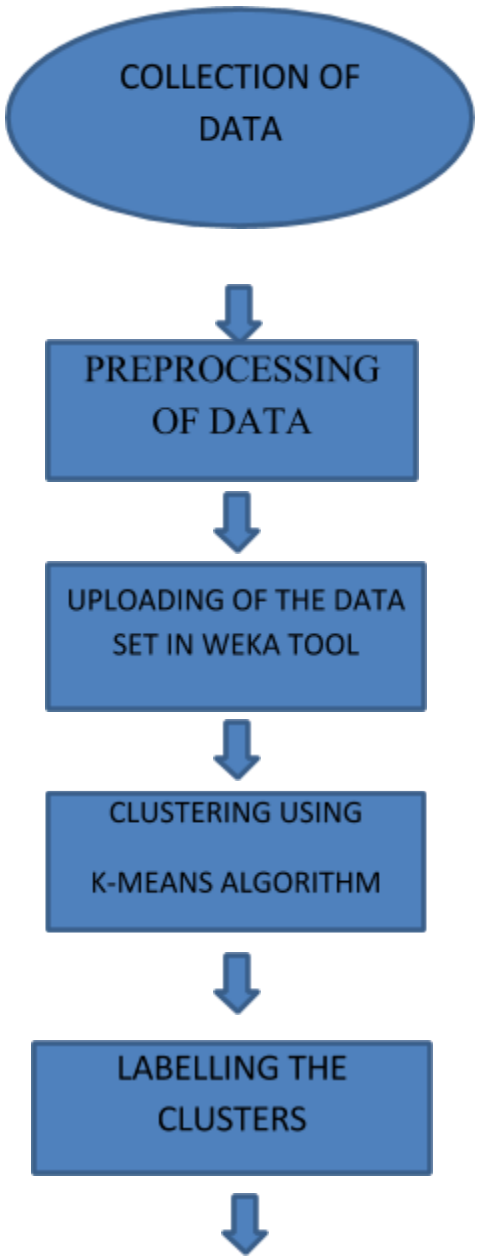
as the influential determinants of student performance. The raw data was preprocessed using data cleansing and data transformation techniques. Chawla proposed the combination of oversampling and under-sampling which can achieve better classifier accuracy named SMOTE (Synthetic Minority Over-sampling Technique). Artificial generated dataset is obtained in order to improve accuracy. From this new dataset, some variations of feature selections will be accomplished.Gain Ratio, Principal Component Analysis, Classifier Subset Evaluator were used for feature selection. Each selected features are then tested by classifiers (Naive Bayes) and being validated (Cross Validation). Three feature selection algorithms were applied on this expanded dataset in order to predict significant factors influencing student performance.

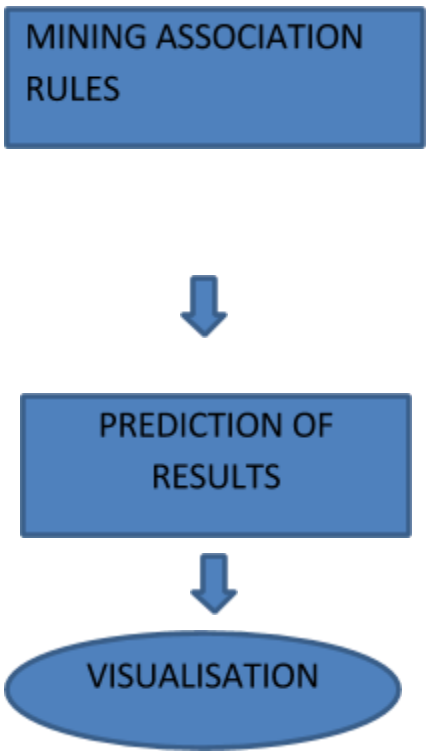
3. METHODOLOGY

1. The data set of students studying from various schools are collected. Collected data set are analysed for external attributes apart from marks such as medical fitness, first generation cadet, sports performance, attendance.
2. The collected data set is pre processed in order to fill the empty cells, filtering test absentees data, scaling of marks.
3. After pre-processing the dataset is loaded into weka tool for clustering. K-Means Clustering algorithm is used to cluster the dataset based on the marks that they have scored. K-values are changed and the corresponding clusters that are obtained are analysed. Optimum value of k is found out in order to fine tune the accuracy of the system.
4. The optimum value for our dataset that we have uploaded was found to be 4, so corresponding labelling is done in order to group the students into different categories based on their performance.
5. The numerical values of the attributes are changed to ordinal values. Then the modified dataset is again fed into weka tool in order to generate association rules.
6. Apriori algorithm is used to generate the association rules. The rules that are with confidence value more than 90% are taken into consideration and useful insight and knowledge are derived.

7. The sample dataset with numerical values that consist of some 20 students data, is again fed into weka tool in order for predicting their third test marks. With help of linear regression algorithm a formula is formulated using the insight of various attributes that are taken into account such as first generation, medical fitness, sports performance and extra curricular activities. Here these attributes are given values from 0 to 1 based on the cadets that are mentioned. The formula obtained is then fed to rest of the dataset to predict the students' third test marks.
8. Plot against different attributes, performance plot, influencing factors that affect students' test performance, cluster outliers are all visualised as charts of any sort using visualisation tab that's available in weka. Thus useful insight can be drawn from the visualisation.

The architecture diagram of the proposed system is as follows:





4.DATASET AND EXPERIMENTAL SETUP

The data set was available in online, scaling has been done, and extra data has been added in order to fine tune the accuracy. The proposed system requires following Specification: The data set should contain extra attributes like first generation, medical fitness, sports apart from the normal mark attributes. Pre processing is done before uploading the data set to the weka tool. WEKA version no:3.6.9 is used .

Pre processing of the data set includes the following steps:

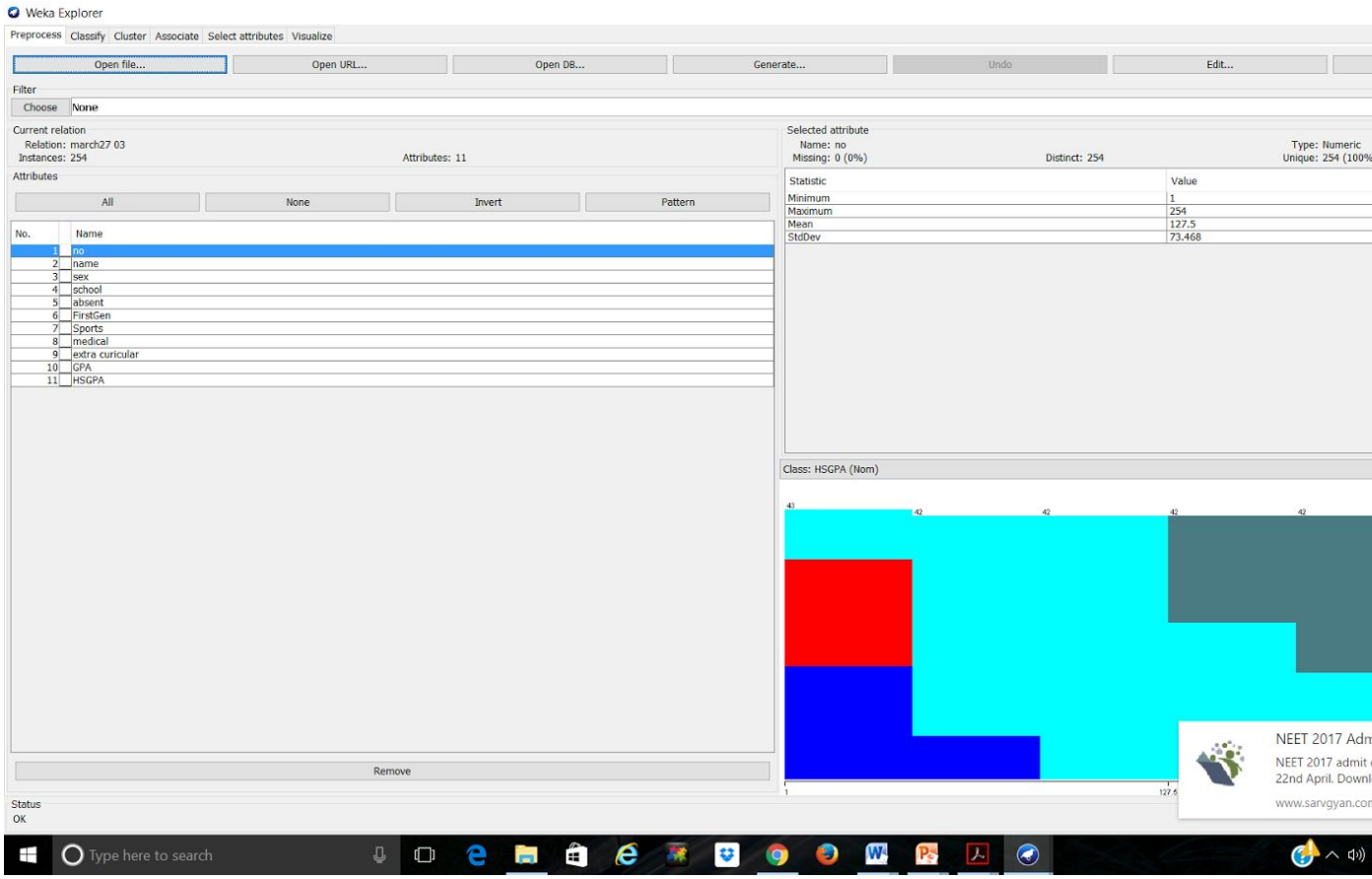
- Data Cleaning- It includes the removal of irrelevant data, noise reduction. In our project absentees students data are removed.
- Data Integration- It implies the collection of data from various sections and

schools.

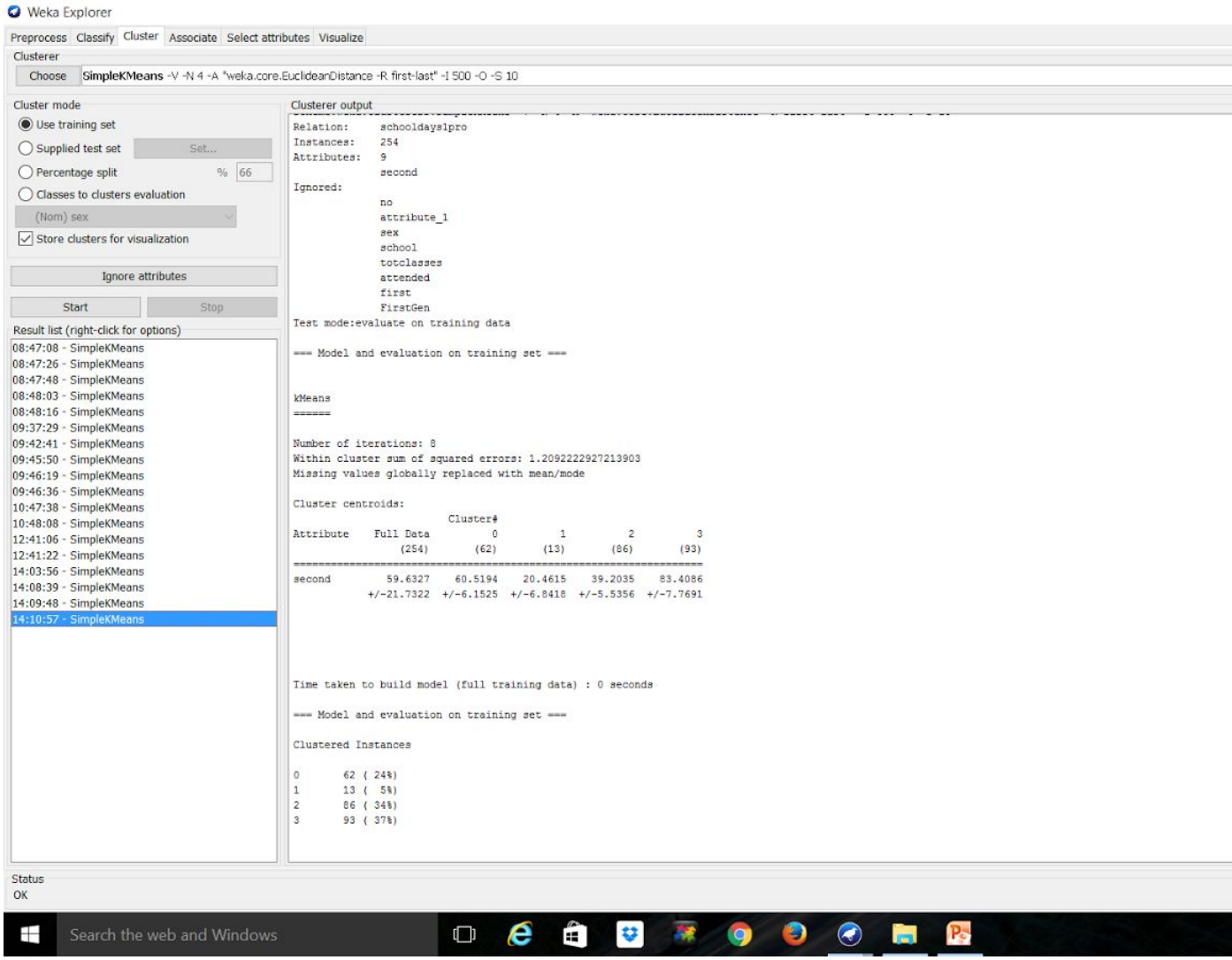
- Data Transformation- Transforming the data into a suitable format that can be used as input for a particular Learning Analytics method.

5. RESULTS

After opening the weka tool, the dataset is fed into the tool and pre processing is done.

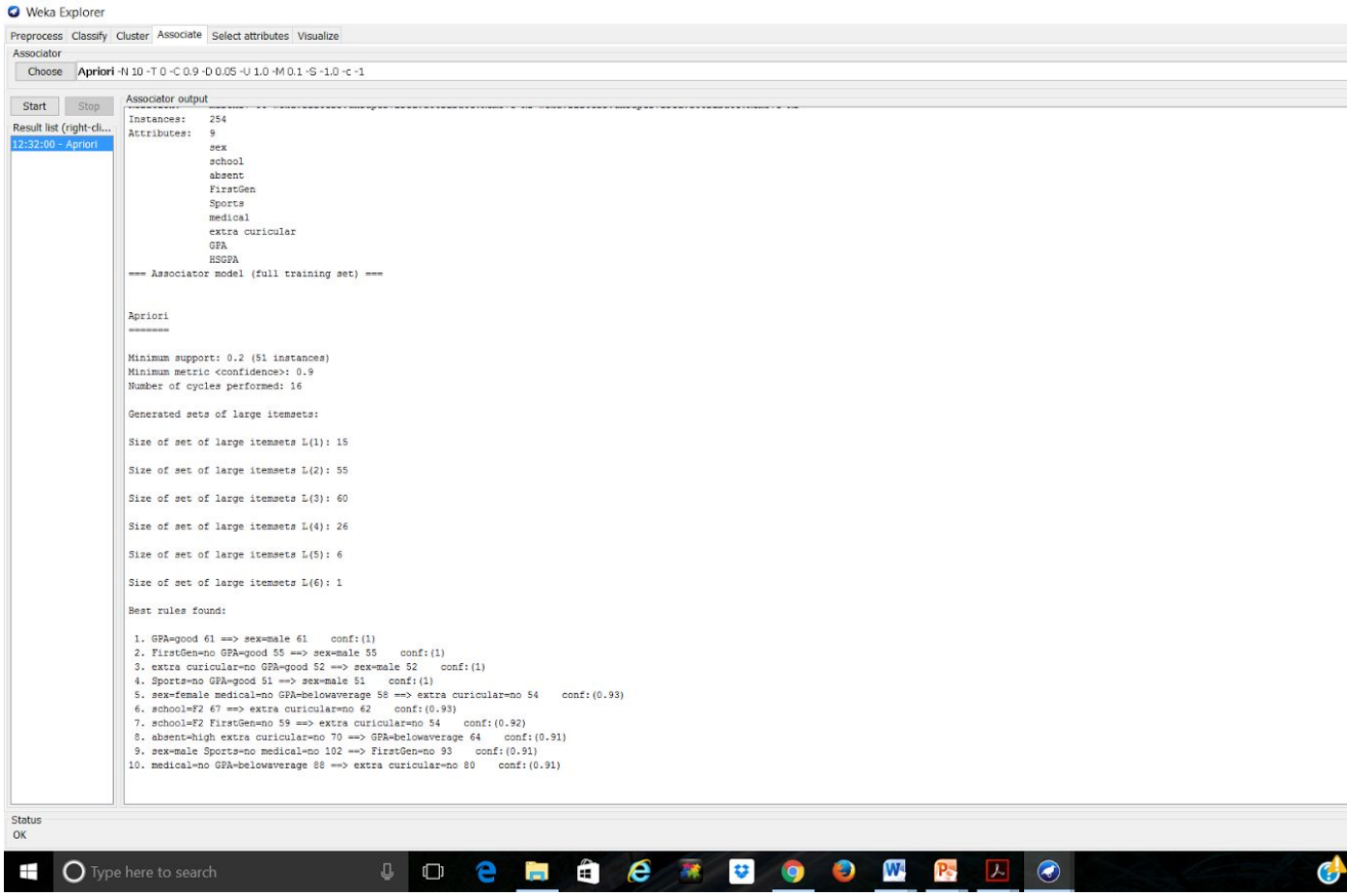


The pre processed data is then clustered using Simple K-Means algorithm for both the test marks separately and with the former combined. After series of analysis the optimum value of k was found to be 4.



The clusters are labelled with the range specified and they are categorised into groups.

The nominal attributes in the data set are converted to ordinal attributes and then the modified dataset is loaded in the weka tool for generating association rules using Apriori algorithm.



Some of the useful rules that are obtained are as follows:

1.sex=female absent=high medical=no 28 ==> GPA=belowaverage 28 conf:(1)

2.sex=male absent=high medical=no 29 ==> HSGPA=aboveaverage 29 conf:(1)

3. sex=female HSGPA=belowaverage 40 ==> GPA=belowaverage 40 conf:(1)

4. medical=no HSGPA=belowaverage 27 ==> GPA=belowaverage 27 conf:(1)

Gender determination:

6.school=F1 extra curicular=no GPA=good 28 ==> sex=male 28 conf:(1)

7. absent=low GPA=good HSGPA=aboveaverage 28 ==> sex=male 28 conf:(1)

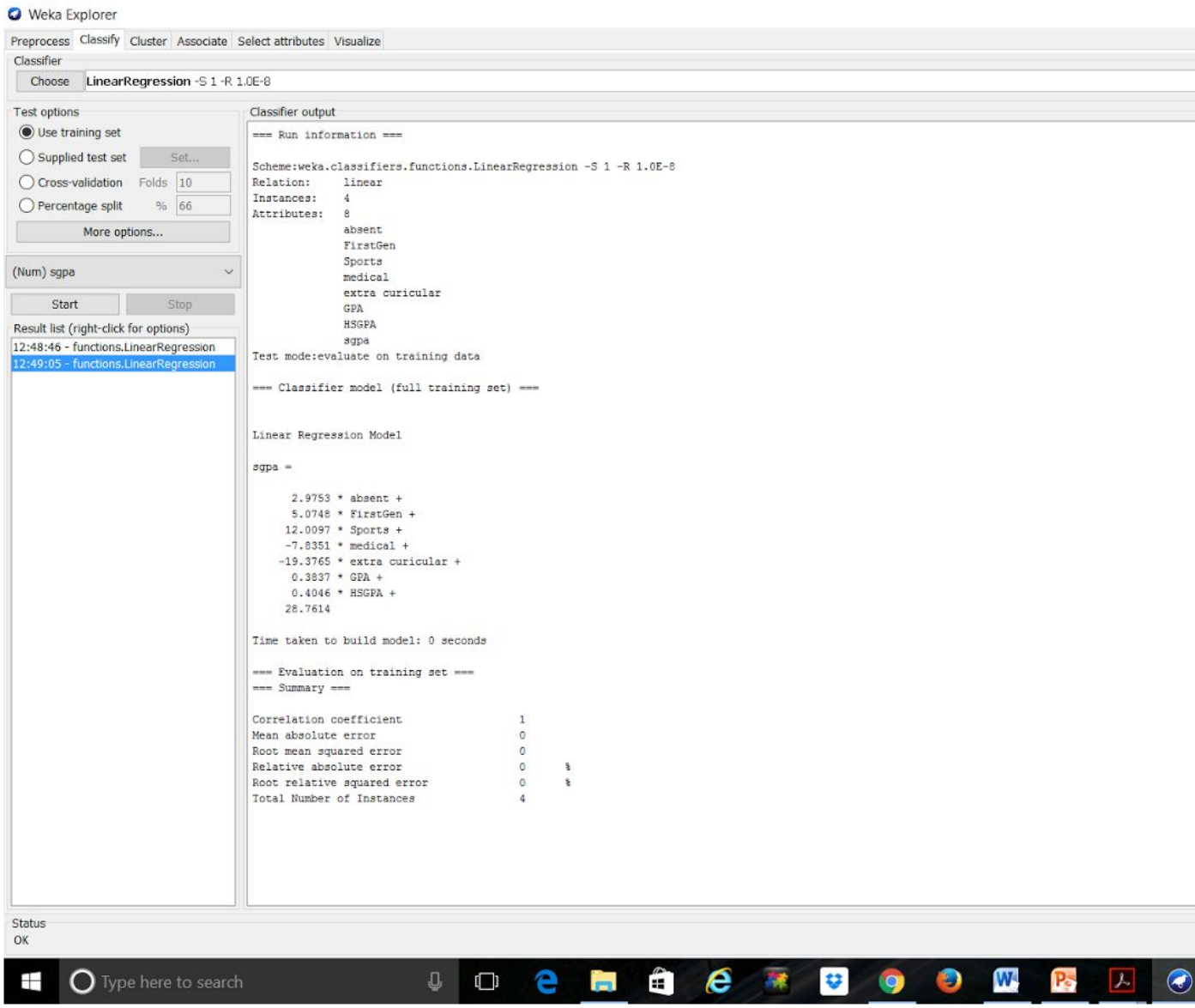
8.absent=low Sports=no GPA=good 32 ==> sex=male 32 conf:(1)

9.sex=female absent=high extra curicular=no 35 ==> GPA=belowaverage 35 conf:(1)

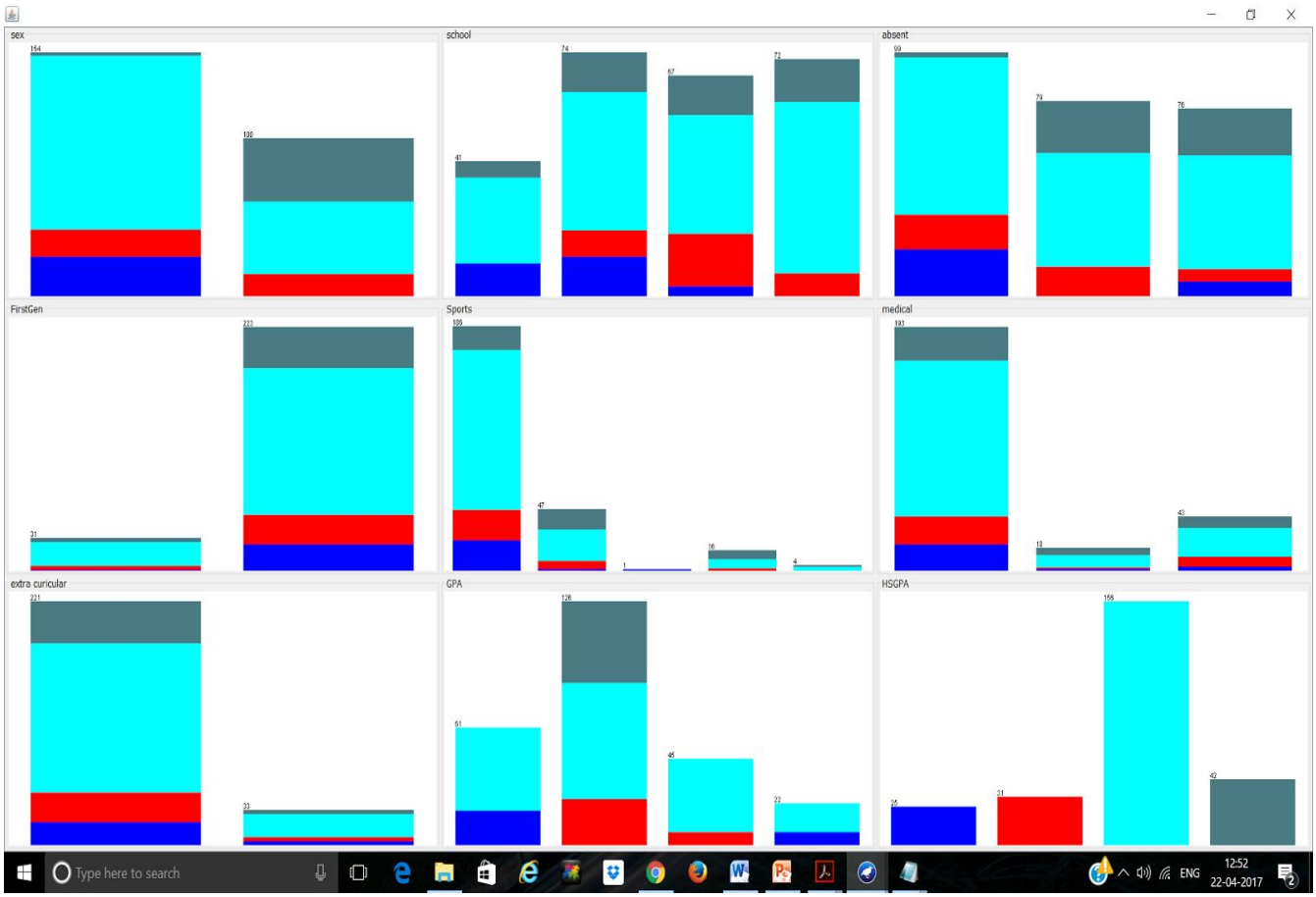
10. Sports=no medical=no extra curicular=no GPA=good 35 ==> sex=male 35 conf:(1)

11. school=F1 GPA=belowaverage 27 ==> sex=female 26 conf:(0.96)

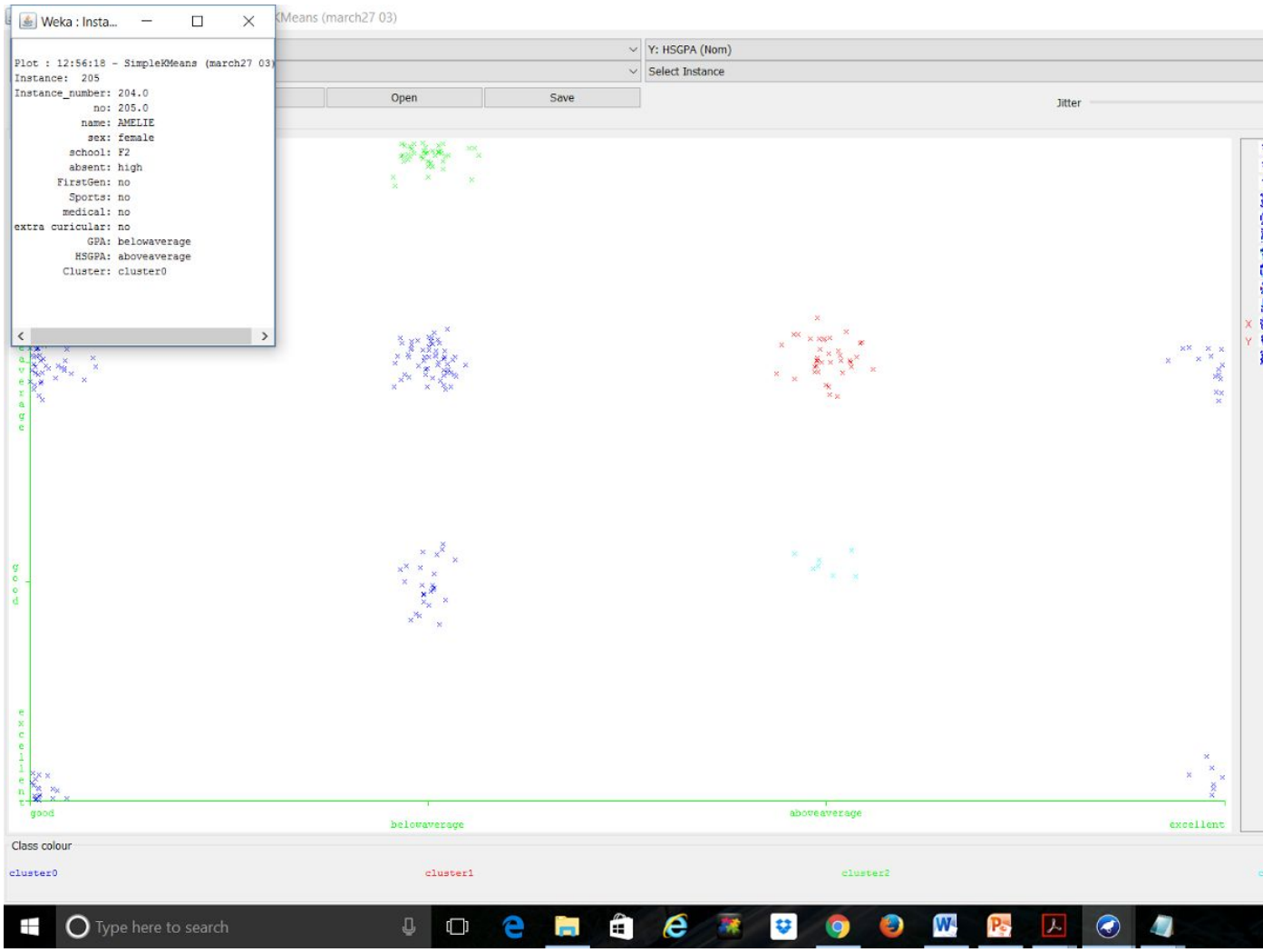
The sample data set that contains information of 10 students in numeric format is fed into weka tool in order to generate a formula that will be useful in predicting the third test marks.



The visualisation part is shown as follows:



The cluster visualisation is shown below:



This picture shows the outlier information also. By clicking the cross mark the information regarding the student is displayed in a separate window.

6. CONCLUSION

The project can be further developed by adding large number of attributes, considering past analysis of test marks in order to create a consistency in the student's performance. Rank card and report card can be generated. Messages can be sent to the students who need to work on for their progress and specific area of interest can also be notified which helps them to choose their field of study. Various algorithms can be applied to the dataset in order to compare the consistency of the results. Alternative algorithms are

available in weka.

Implementing an application of this scale would require a very strong database which will be able to store a lot of values from many schools. Scaling factors has to be taken into consideration. The accuracy of the results that are generated using different algorithms should be taken into account in order to improve the consistency of the result. Notification sender requires the concept of networking that has to be taken into account for further extension of the project.

7.ACKNOWLEDGEMENT

1.Singh, A. S. Sabitha and A. Bansal, "Student performance analysis using clustering algorithm," 2016 6th International Conference - Cloud System and Big Data Engineering (Confluence), Noida, 2016, pp. 294-299.

doi: 10.1109/CONFLUENCE.2016.7508131

2. S. Shankar, B. D. Sarkar, S. Sabitha and D. Mehrotra, "Performance analysis of student learning metric using K-mean clustering approach K-mean cluster," 2016 6th International Conference - Cloud System and Big Data Engineering (Confluence), Noida, 2016, pp. 341-345.

doi: 10.1109/CONFLUENCE.2016.7508140

3. K. Maharani, T. B. Adji, N. A. Setiawan and I. Hidayah, "Comparison analysis of data mining methodology and student performance improvement influence factors in small data set," 2015 International Conference on Science in Information Technology (ICSITech), Yogyakarta, 2015, pp. 169-174.

doi: 10.1109/ICSITech.2015.7407798