# Preprocessing and Classification of Data Analysis in Institutional System using Weka

Reena Thakur
G.N.I.E.M. Nagpur

A.R. Mahajan
P.I.E.T. Nagpur

## ABSTRACT
In today's world, an organization generates more information in a week than most people can read in a lifetime. It is humanly impossible to study, decipher, and interpret all that data to find useful information. By applying data mining techniques people can work on the extraction of hidden, historical and previously unknown large databases. In this paper we have used powerful data mining technology weka tool for the preprocessing, classification and analysis of institutional result of Computer science & engineering UG students. Here efficient information have been mined from the university result. Results show the analysis of marks, pass or fail, percentage of attendance etc.

## Keywords
Classification, clustering, weka, data mining.

## 1. INTRODUCTION
Data Mining is a process of recognizing unique, potentially required, preprocessed valid and ultimately understandable patterns in massive amounts of data [2]. Data mining techniques can be classified into two unsupervised and supervised learning techniques. Data mining, "a major way of creating knowledge", is a useful way of studying medicine, genetics, bioinformatics, education [2].

At present huge amounts of data is being accumulated. Traditional way of mining data is manual but in case of large quantities this task becomes tedious. To overcome this condition Data mining tools have benn used. In this paper we are using WEKA Tool for the analysis of Institutional data. By using Data mining techniques, knowledge could be mined from from the data large in size as well as in dimensionality.

This paper uses Institutional Data Mining Techniques (IDM) to provide more accurate result analysis. Data mining is a dynamic technology to deal and extract the hidden potential data which is to be converted to useful information. It discovers information within the data that queries and reports can't effectively reveal. After gathering data from the university result, data mining technique need to be applied to determine pass, atkt and failed student.

## 2. LITERATURE SURVEY
In [9], the author used an approach to classify students to achieve the result of their grades based on extraction of features from web based educational system. The author used various classifiers for classification and compared their performance on the dataset. Further, we proceeded with the combination of classifiers to improve the classification performance. Again to improve the prediction accuracy they have used Genetic Algorithm. This method is of considerable usefulness in identifying students at risk early, especially in very large classes, and to allow the instructor to provide appropriate advising in a timely manner.

In paper [10], the author analyzed about association rules and its uses in Educational data mining to work on learned data.

In paper [11], the author focused on how DM is useful in improving the performance of the higher educational students. The author applied association rule, classification rule and cluster analysis on the dataset based on the subject database system and the considered Moodle e-learning of student.

In paper [11], the author studied the relationship between the students result appeared in the University entrance examination & their success rate using cluster analysis & K-means algorithm techniques. They were grouped the university student according to their characteristic, forming cluster & clustering process carried out using the K-means clustering.

The author in [12] surveyed on the application of DM using various techniques and shown how they are useful in traditional educational system.

## 3. INSTITUTIONAL DATA MINING
Educational Data Mining is a vital issue concerned with developing methods for researching the information from large data that come from educational settings, and using these techniques to better understand students result, and the settings which they learn in [7]. Data mining is an infusion of interesting (non-trivial, inexplicit, previously unknown and latent important) patterns or digging knowledge from vast amount of data. As vast amount of student database is stored in Institutional databases , so in order to find the users required data and to find the predictive relationships, various data mining techniques such as classification, clustering, association etc are to be used. We can use the data mining in educational system as: predicting drop-out student, relationship between the student university entrance examination results & their success, predicting student's academic performance, discovery of strongly related subjects in the undergraduate syllabi, knowledge discovery on academic achievement, classification of students' performance in computer programming course according to learning style.

## 4. METHODOLOGY
We are using Weka Tool for the preprocessing and analysis of result. Fig shows the GUI of this tool. Weka (Waikato Environment for Knowledge Analysis) is a popular suite of machine learning software written in Java, developed at the University of Waikato, New Zealand. WekaTool is free software available under the GNU (General Public License). The Weka workbench contains preprocess, a collection of visualization tools , algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access to this functionality [7]. Weka is a collection of machine learning algorithms for solving real-world data mining problems. It is written in Java and runs on almost any platform. The algorithms can either be applied directly to a dataset or called from your own Java code [8].

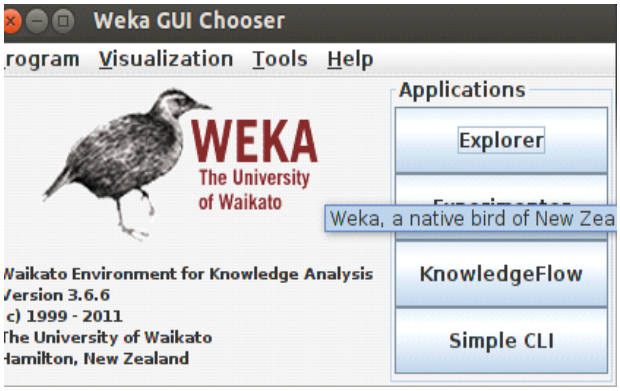



**Fig (a) Weka GUI Chooser**

As shown in fig(a) the Weka GUI Chooser consists of four applications:

Explorer: This is an environment where data can be explored with WEKA.

Experimenter: This is an environment for performing experiments and conducting statistical tests between learning schemes.

Knowledge Flow: This is an environment which provides the drag and drop interface and supports incremental learning.

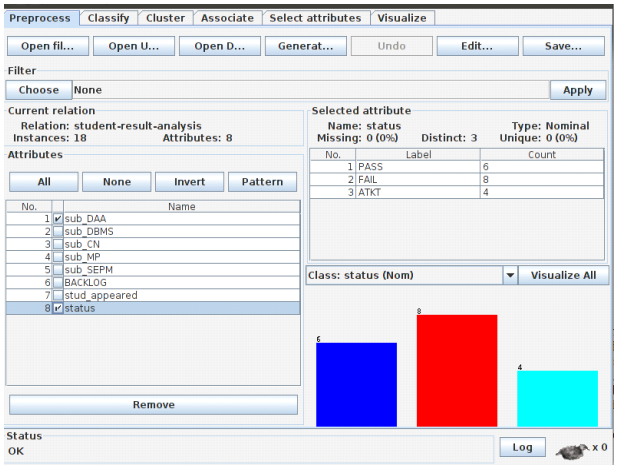
Simple CLI: Provides a simple command-line interface that allows direct execution of WEKA commands for operating systems that do not provide their own command line interface.

## 5. EXPERIMENTAL WORK AND RESULTS

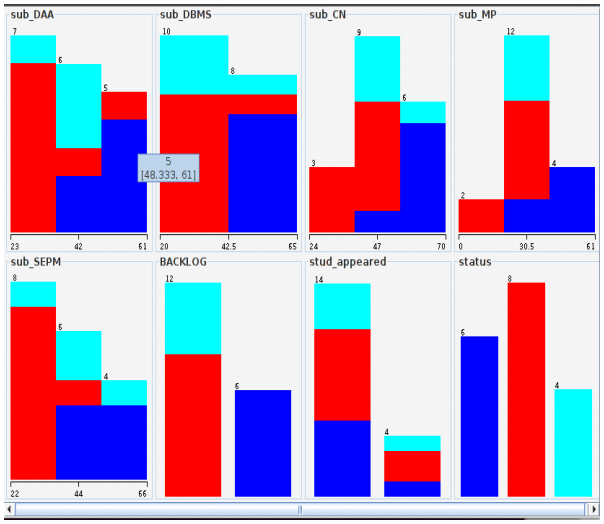
The primary data required is taken from Guru Nanak Institution, Department of Computer Scince & Engineering, Nagpur, Maharashtra. This database is designed in Notepad which is arranged according to the compatible format. Further, the data is saved with extension ARFF (Attribute Relation File Format) format to process in WEKA. An ARFF file is an ASCII text file that describes a list of instances sharing a set of attributes. ARFF files were developed by the Machine Learning Project at the Department of Computer Science of The University of Waikato for use with the Weka machine learning software.

Then start with the Weka tool use the explorer application and select the preprocess button followed by this open the result analysis data set. After this choose flter allfilter, mutifilter, supervised and unsupervised filter algorithm which can be used to transform the data from one format to other e.g. numeric attributes into discrete ones. It is also possible to delete instances and attributes according to specific criteria on the preprocess screen. You can visualize a graph for particular attribute also.

After processing the ARFF file in WEKA the list of all attributes, statistics and other parameters can be utilized as shown in Fig(b).

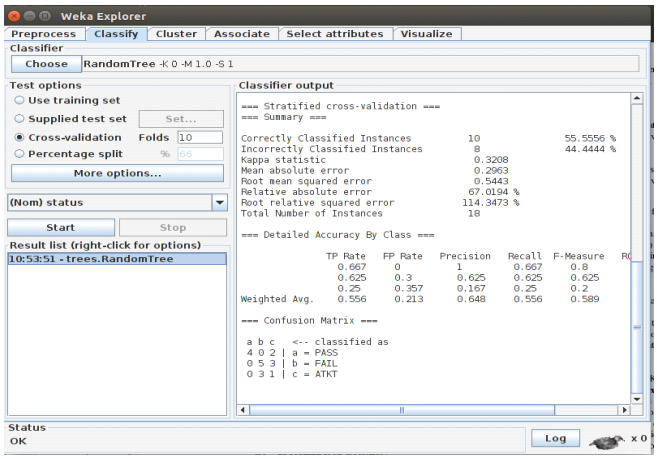



**Fig(b):Weka3.6.10 Explorer window open**



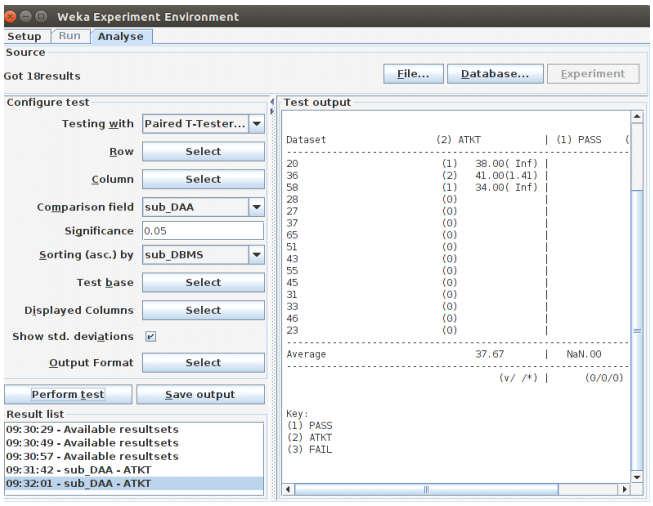

**Fig(c) Graph of subject with Student Result dataset**

We have considered the result of third year student of UG of our institute for result analysis. There are five subjects DAA(Design Analysis & Algorithm), DBMS(Database Management System), MP(Microprocessors), CN(Computer Network) and SEPM(Software Engineering & Project Management). The graph for each subject, backlog, status(Pass/Fail/ATKT) is shown in the fig(c).

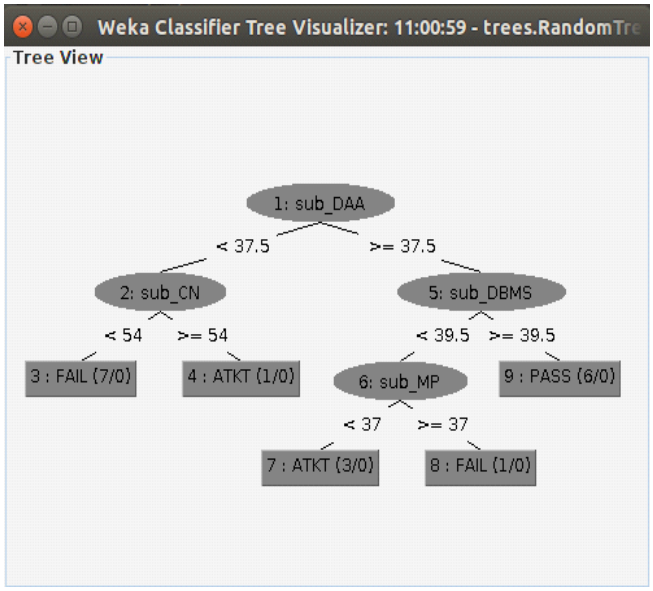



**Fig(d) Classification using classifier**

During classify, ZeroR classifier is considered which predicts the majority of class in training data. It predicts the mean for numeric value & mode for nominal class. The fig(d) shows the confusion matrix which displays the accuracy of solution to the classification problem where four students are passed, eight students are failed and six are having ATKT. The Fig(e) shows the analysis of the selected subject by using Weka Experiment Environment where column 1 shows pass, column 2 shows fail and column 3 shows ATKT students.




**Fig(f) Classification using decision tree**




**Fig(g) Analysis of DAA subject using weka experiment environment**

## 6. CONCLUSION AND FUTURE WORK

This paper includes the study of Data Mining tool applied to Institutional systems. By using Weka tool you can preprocess the data, classify the data for various subjects and can analyse the result data. Here we have used the third year engineering student university result. On University result, we have discovered the knowledge by using different data mining techniques.

For future work, more data mining techniques such as neural nets, genetic algorithm, k-nearest neighbor, Naive Bayes, etc can be applied on complex systems. Comparison of the result using various data mining tools can be the future work. This will be also useful in e-educational system.

## 7. REFERENCES

[1] Jiawei Han and Micheline Kamber, Data Mining Concepts and Techniques, 2nd ed.,Morgan Kaufmann publishers, SanFrancisco, 2006.

[2] Fayyad, U., & Stolorz, P. (1997). Data mining and KDD: promise and challenges. Future generation computer systems, 13(2), 99-115.

[3] Guerra L, McGarry M, Robles V, Bielza C, Larrañaga P, Yuste R. (2011). Comparison between supervised and unsupervised classifications of neuronal cell types: A case study. Developmental neurobiology, 71(1): 71-82.

[4] Yoo I, Alafaireet P, Marinov M, Pena-Hernandez K, Gopidi R, Chang JF, Hua L. (2012). Data mining in healthcare and biomedicine: a survey of the literature. Journal of medical systems, 36(4): 2431-2448.

[5] An Introduction to weka data mining tool.

[6] Romiro C. and Ventura S., "Educational data mining- A survey from 1995-2005" Expert systems with applications(33) 135-146. 2007.

[7] International Educational Data Society www.educationaldatamining.org

[8] Kifaya (2009) Mining student evaluation using associative classification and clustering communication of the IBIMA vol 11 IISN 1943-7765.

[9] Ritu Chauhan, Harleen Kaur, M.Afshar Alam, ―Data Clustering Method for Discovering Clusters in Spatial Cancer Databaset, International Journal of Computer Applications (0975 – 8887) Volume 10– No.6, November 2010

[10] J.R Quinlan. C4.5: Programs for Machine Learning. Morgan Kaufman, 1993.

[11] Behrouz.et.al., (2003) Predicting Student Performance: An Application of Data Mining Methods with The Educational Web-Based System Lon-CAPA © 2003 IEEE, Boulder, CO

[12] Sheikh, L Tanveer B. and Hamdani, S., "Interesting Measures for Mining Association Rules". IEEE- INMIC Conference December. 2004

[13] Alaa el-Halees (2009) Mining Students Data to Analyze e-Learning Behavior: A Case Study.

[14] Erdogan and Timor (2005) A data mining application in a student database. Journal of Aeronautic and Space Technologies July 2005 Volume 2 Number 2 (53-57)