Prediction Of Student Performance Using Weka Tool

Gurmeet Kaur[1], Williamjit Singh[2]
1 Student of M.tech (CE), Punjabi university, Patiala
2 (Asst. Professor) Department of CE, Punjabi University, Patiala
E-mail: [1]grmtkaur76@gmail.com, [2]williamjeetsingh@gmail.com

Abstract. Data mining is widely used in educational field to find the problems arise in this field. Student performance is of great concern in the educational institutes where several factors may affect the performance. For prediction the three required components are: Parameters which affect the student performance, Data mining methods and third one is data mining tool. These Parameters may be psychological, personal, environmental. We conduct this study to maintain the education quality of institute by minimizing the diverse affect of these factors on student's performance. In this Paper, Prediction of student Performance is done by applying Naïve bayes and J48 decision tree classification techniques WEKA tool. By applying data mining techniques on student data we can obtain knowledge which describes the student performance. This knowledge will help to improve the education quality, student's performance and to decrease failure rate. All these will help to improve the quality of institute.

Keywords: Prediction, naïve bayes, j48, Weka Tool

1. Introduction

Data Mining (DM), or Knowledge Discovery in Databases (KDD), is an approach to discover useful information from large amount of data [3]. DM techniques apply various methods in order to discover and extract patterns from stored data. The pattern found will be used to solve a number of problems occurred in many fields such as education, economic, business, statistics, medicine, and sport. The large volume of data stored in those areas demands for DM approach because the resulting analysis is much more precise and accurate.

Data mining techniques are used to operate on large volumes of data to discover hidden patterns and relationships helpful in decision making. While data mining and knowledge discovery in database are frequently treated as synonyms, data mining is actually part of the knowledge discovery process. The sequences of steps identified in extracting knowledge from data are shown in Fig 1.

In recent years, there has been increasing interest in the use of DM to investigate educational field. Educational Data Mining (EDM) is concerned with developing methods and analyzing educational content to enable better understanding of students' performance [2]. It is also important to enhance teaching and learning process. The data can be collected form historical and operational data reside in the databases of educational institutes. The student data can be personal or academic. Also it can be collected from e-learning systems which have a large amount of information used by most institutes. Educational Data Mininguses many techniques such as Decision Trees, Neural Networks, Naïve Bayes, K-Nearest neighbor, and many others.

Prediction models that include all personal, social, psychological and other environmental variables are necessitated for the effective prediction of the performance of the students [15]. The prediction of student performance with high accuracy is beneficial for identify the students with low academic achievements initially. It is required that the identified students can be assisted more by the teacher so that their performance is improved in future.

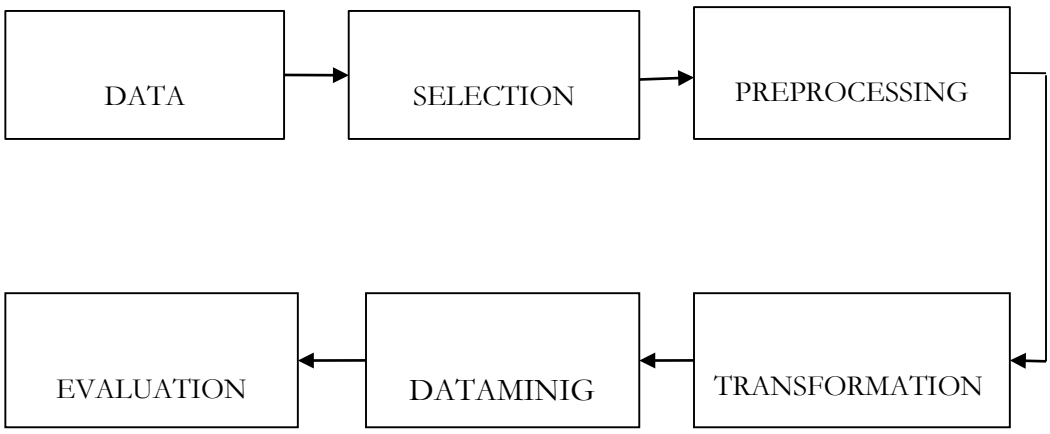

Fig 1: steps of extracting knowledge

2. Literature Survey

Alaa M. El-Halees et al. [1], proposed a case study that Educational data mining concerns with developing methods for discovering knowledge from data that come from educational domain. They used educational data mining to improve graduate students' performance, and overcome the problem of low grades of graduate students. In their case study they tried to extract useful knowledge from graduate students data collected from the college of Science and Technology – Khanyounis. The data include fifteen years period [1993-2007]. After preprocessing the data, they applied data mining techniques to discover association, classification, clustering and outlier detection rules. In each of these four tasks, they presented the extracted knowledge and describe its importance in educational domain.

Sajadin Sembiring et al. [13], applied data mining techniques analyze the relationships between student's behavioral and their success and to develop the model of student performance predictors. They used Smooth Support Vector Machine (SSVM) classification and kernel k-means clustering techniques. The results of this study reported a model of student academic performance predictors by employing psychometric factors as variables predictors. This study expressed the strong correlation between mental condition of student and their final academic performance.

Different methods and techniques of data mining were compared during the prediction of students' success, applying the data collected from the surveys conducted during the summer semester at the University of Tuzla, the Faculty of Economics, academic year 2010-2011, among first year students and the data taken during the enrollment by Edin Osmanbegović et al[6]. The success was evaluated with the passing grade at the exam. The impact of students' socio-demographic variables, achieved results from high school and from the entrance exam, and attitudes towards studying which can have an affect on success, were all investigated. They have shown that Naïve Bayes classifier outperforms in prediction decision tree and neural network methods.

Bayesian classification method is used on student database to predict the students division on the basis of previous year database by Saurabh Pal et al. [14].This study shown that academic performances of the students

are not always depending on their own effort. Our investigation shows that other factors have got significant influence over students' performance. The study proposed by them is helped to which needed special attention to reduce failing ration and taking appropriate action at right time.

Table 1: Sample of data mining techniques used for prediction of student performance

Author	**year**	**DM technique**	**Accuracy**
Saurabh Pal et. al	2011	Naïve bayes	Not assigned
Anwar M. A	2012	Apriori algorithm	Not assigned
Vaibhav P. Vasani et. al	2014	Naïve Bayes	86.4%
		J48	95.9%
Sembiring et al.	2011	Naïve bayes	82.4%
		K-Means	93.7%
		DecisionTree	80.2%
Edin Osmanbegovet.al	2012	Naïve Bayes	76.48%
		Multilayer Perception	71.2%
		J48	73.98%

Classification of the data collected from students of polytechnic institute has been discussed by Vaibhav P. Vasani et al [16]. This data was pre-processed to remove unwanted and less meaningful attributes. These students were then classified into different categories like brilliant, average, weak using decision tree and naïve Bayesian algorithms. The processing was done using WEKA data mining tool. This paper also compared results of classification with respect to different performance parameters. It is also observed that Decision tree (J48) gives better result than Naïve Bayesian algorithm in terms of accuracy in classifying the data.

3. Data Set and Attribute Selection

I used dummy data to obtain the results of classification, clustering and association rules. This data set consist 52 instances and each instance consists of 9 attributes in Table 2. Initially data is collected in excel sheet.

The values for some of parameters are defined for the present investigation as follows:

Gender – Gender parameter have two values M for male students and F for female students.

Hometown – The environment of student's hometown. Its value is divided into two values: Urban and Rural.

PrevSemGrad – Previous Semester Marks/Grade obtained by student. It is split into four class values: First – >60%, Second – >45% and <60%, Third – >36% and < 45%, Fail < 40%.

Sem – Seminar Performance obtained. Seminar performance is evaluated into three classes: Poor – Presentation and communication skill is low, Average – Either presentation is fine or Communication skill is fine, Good – Both presentation and Communication skill is fine.

Atte – Attendance of Student. Minimum 70% attendance is compulsory to participate in Final Examination. Attendance is divided into three classes: Poor - <60%, Average - > 60% and <80%, Good - >80%.

Sports - Sports is divided into two classes: Yes –student participated in Sports, No – Student not participated in Sports.

SenSecGrade - It is split into four class values: Excellent – >90%, Above_average – <85% and <70%, Below_average <70 – <50% and < 45%, Average - 70%.

FamInc - It is split into four class values Below_20000, 20000_30000, 30000_40000, 40000_50000.

Med – it is divided into three values based on the teaching language. These are English, Punjabi, Hindi.

Table 2: Parameter description and possible values.

Parameter	**Description**	**Value**
Gender	Gender	{M,F}
Hometown	Location of house	{rural, urban}
FamInc	Family Income	{>20000,20000_30000,30000_40000,40000_50000}
PrevSemGrade	Previous semester	{First>60%,Second>45&<60%,Third>36& <45%,Fail<36%}
Atten	Attendance	{Poor, Average, Good}
Med	Medium	{English, Punjabi, Hindi}
SenSecGrade	+2 grade	{>Average, Average, <Average, Excellent}

Sem	Seminar performance	{Poor ,Average, Good}
Sports	Participation in sports	{Yes, No}

4. Implementation Using Weka Tool

Steps to be performed for implementing the algorithms in WEKA tool, are as follows:

Step-1: Preprocessor:

The preprocessor import the dataset into the tool and preprocess it. Output of preprocessor shown in Fig 2.

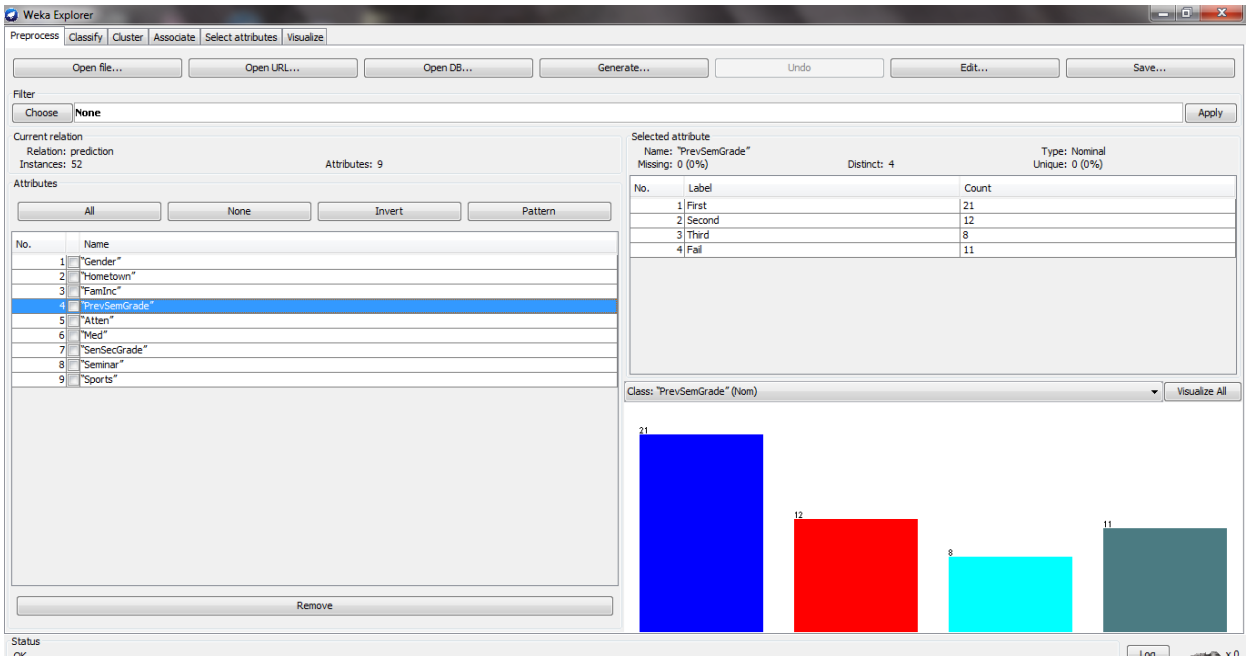

Fig 2: Output of Preprocessing

Left hand side of the above screen shows detail of relation name, number of attributes and number of records. Right hand side gives details of attribute values, type, and number of distinct values. Specification of every attribute is displayed in the right bottom of the screen.

Step 2: Classification

The Classify panel allows the user to apply classification and regression algorithms to the dataset estimate the accuracy of the resulting model. In this paper we used two algorithms for classification thatare Naïve Bays

algorithm and J48 tree on the basis of previous semester grades. The outputs of naïve bayes is shown in Fig 4. A decision tree is obtained from algorithm J48 as shown in Fig 5.

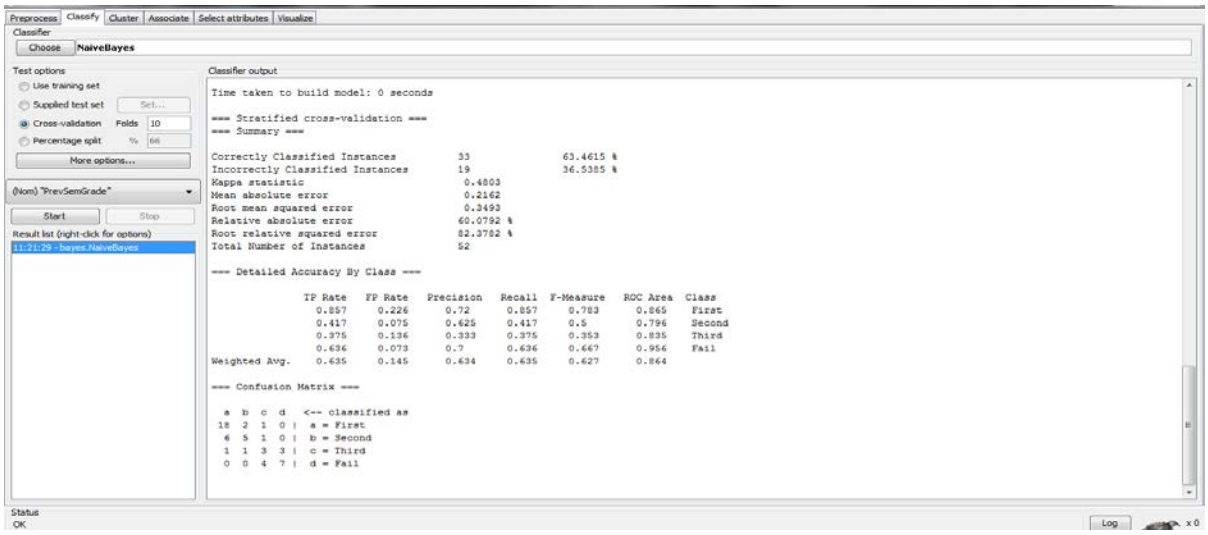

Fig 4: Output of naïve bayes classification

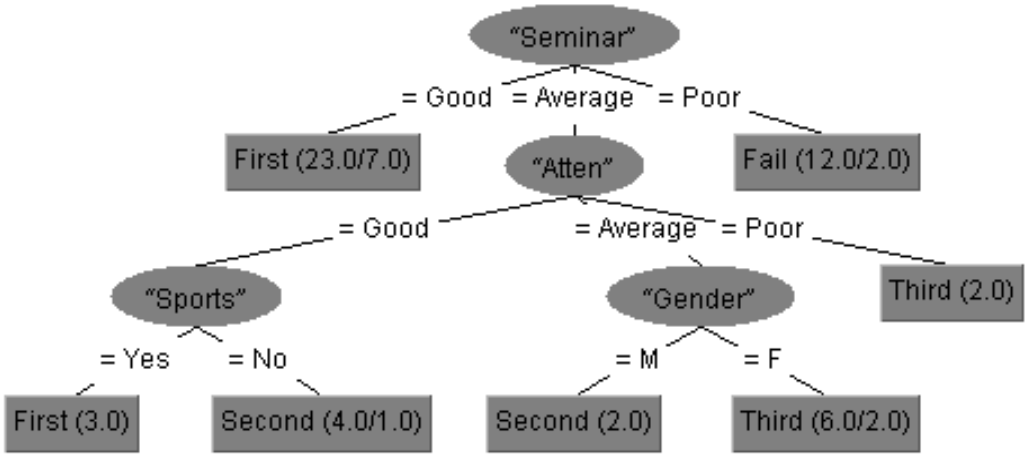

Fig 5: Decision tree obtained from J48

5. Findings and Results

The Proposed application has evaluated on dummy data of 52 students. While the output of all the algorithms in previous section. Now in this section we will discuss all the findings and results obtained from them. First of all results of classification using naïve bayes and j48 algorithms. The performance of algorithms is evaluated on the basis of recall and precision. Precision is defined as number of correct positive prediction over total number of positive prediction and recall is defined as number of correct positive prediction over total number of positive cases. A high precision indicates that algorithm returns more relevant results than irrelevant and high recall means that most of the results retuned by the algorithms are relevant.

Classification matrix of J48 and naïve bayes shown in Table3and Table4 respectively:

Table 3: Classification matrix of J48

PrevSemGrade		Prediction			
		First	**Second**	**Third**	**Fail**
Actual	**First**	18	2	1	0
	Second	6	5	1	0
	Third	1	1	3	3
	Fail	0	0	4	7

Table 4: Classification matrix of Naïve bayes

PrevSemGrade		Prediction			
		First	**Second**	**Third**	**Fail**
Actual	**First**	16	4	1	0
	Second	9	2	1	0
	Third	1	1	2	4
	Fail	0	0	2	9

The performance comparison of J48 and naïve bayes is shown in Table 5.

Table 5: Performance Comparison of J48 and Naïve bayes

	J48		**Naïve bayes**	
	Precision	**Recall**	**Precision**	**Recall**
First	0.63	0.66	0.72	0.85
Second	0.44	0.33	0.62	0.41
Third	0.50	0.50	0.33	0.37
Fail	0.76	0.90	0.70	0.63
Weighted average	0.59	0.61	0.63	0.63
Correctly Classified Instances	61.53%		63.59%	
Incorrectly ClassifiedInstances	38.46%		36.41%	

6. Conclusion

This above study shows that data mining will be considered most useful in educational field. Predicting student's academic performance is of great concern to the education institutes. By applying data mining techniques and tools in prediction of student performance is helpful to identify the abilities of students, their interests and weaknesses and also helpful to cluster the students according to their performance. Naïve bayes and j48 decision tree are used for classification. Naïve bayes provide 63.59 % accuracy and j48 Provide 61.53% accuracy.

References

[1] Alaa M. El-Halees, Mohammed M. Abu Tair, "Mining Educational Data to Improve Students' Performance: A Case Study", International Journal of Information and Communication Technology Research, 2012.

[2] Azwa Abdul Aziz, Nur Hafieza Ismail, Fadhilah Ahmad, "Mining Students' Academic Performance" Journal of Theoretical and Applied Information Technology, 31st July 2013.

[3] Brijesh Kumar Baradwaj, " Mining Educational Data to Analyze Students Performance", International Journal of Advanced Computer Science and Applications, 2011

[4] D.I. Reea, D.J. Brewer b , L.M. Argys, " How Should We Measure The Effect Of Ability Grouping On student Performance?" Economics of Education Review 19, 2000.

[5] Edin Osmanbegović, Mirza Suljić, "Data Mining Approach For Predicting Student Performance", Economic Review – Journal of Economics and Business, May 2012.

[6] Eric Eide , Mark H. Showalter, "The Effect Of School Quality On Student Performance: A Quantile Regression Approach" Economics Letters 58 (1998),1998.

[7] Hanumathappa. M, Margaret Mary.T, "Educational Data Mining And Prediction Of Learning Disabilities" , Asia Pacific journal of research,2013.

[8] K. Rajeswari, Suchita Borkar, Saul "Predicting Students Academic Performance Using Education Data Mining", International Journal of Computer Science and Mobile Computing, IJCSMC, Vol. 2, Issue. 7, July 2013.

[9] Kartik N. Shah , Srinivasulu Kothuru, and S. Vairamuthu, "Clustering Students' Based On Previous Academic Performance", International Journal of Engineering Research and Applications (IJERA), May-Jun 2013.

[10] Mr. LOBO L . M . R . J, Sunita B A her, "Data Mining in Educational System using WEKA", nternational Conference on Emerging Technology Trends (ICETT) 2011.

[11] Naseer Ahmed, "Information Mining in Assessment Data of Students' Performance", International Journal of Engineering Science, 2012.

[12] Prashant M. Dolia, Ph.D, Nikhil P. Shah, Jaimin N. Undavia, "Prediction of Graduate Students for Master Degree based on Their Past Performance using Decision Tree in Weka Environment", International Journal of Computer Applications (0975 – 8887), Volume 74– No.11, July 2013.

[13] Sembiring,M. Zarlis , Dedy Hartama ,Ramliana S , Elvi Wani, "Prediction Of Student Academic Performance By An Application Of Data Mining Techniques", International Conference on Management and Artificial Intelligence,2011.

[14] Saurabh Pal, "Data Mining: A Prediction For Performance Improvement Using Classification", International Journal of Computer Science and Information Security, April 2011.

[15] Tripti Mishra, Dr. Dharminder Kumar, Dr. Sangeeta Gupta, "Mining Students' Data for Performance Prediction", Fourth International Conference on Advanced Computing & Communication Technologies,2014