Data Mining: An analysis to predict academic performance of students based on varying factors

Shruti Bijawat *, Deepak Moud **

*(Department of Computer Science, Poornima Institute of Engineering & Technology, Jaipur)
** (Department of Computer Science, , Poornima Institute of Engineering & Technology, Jaipur)

Abstract

Data Mining is a study to extract the required essential information from the given large amount of databases. It is a combination of machine learning, statistics and visualization techniques to discover and extract knowledge. Today the amount of data stored in educational database is increasing with a great speed. These databases contain hidden information for improvement of students' performance. Educational data mining is used to study the data available in the educational field and bring out the hidden knowledge from it. This study is conducted in accordance to build a model based on multiple linear regression analysis between a students' performance and the various surrounding factors affecting him, like the his previous studying medium, his physics-chemistry-maths marks in 12th std., attendance in Engineering during 4 years, gender, and finally on the factor that he avails the hostel faculty or not. Since the academic performance is influenced by many factors, therefore it is essential to develop predictive data mining model for students' performance so as to identify the difference between high learners and slow learners
student.

Keywords- Data Mining, Academic Performance, Regression analysis ,

I. INTRODUCTION

Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations. Educational data mining is a new emerging technique of data mining that can be applied on the data related to the field of education. There are increasing research interests in using data mining in education. This new emerging field, called Educational Data Mining, concerns with developing methods that discover knowledge from data originating from educational environments.

The ability to predict a student's performance is very important in educational environments. A very promising tool to attain this objective is the use of Data Mining. Data mining techniques are used to operate on large amount of data to discover hidden patterns and relationships helpful in decision making.

II. KDD AND DATA MINING FUNCTIONALITIES

Data Mining, also popularly known as *Knowledge Discovery in Databases* (KDD), refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. While data mining and knowledge discovery in databases (or KDD) are frequently treated as synonyms, data mining is actually part of the knowledge discovery process. The following figure (Figure 1.1) shows data mining as a step in an iterative knowledge discovery process.

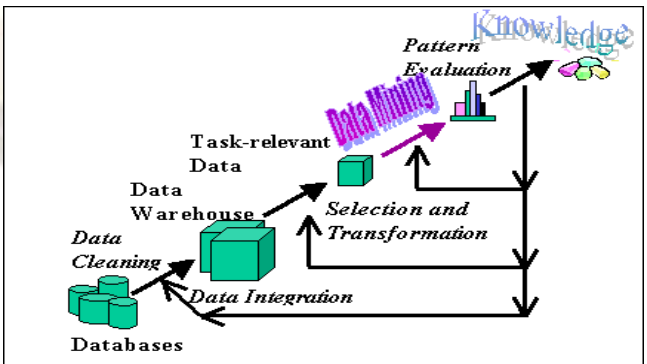


Figure 1.1: Data Mining is the core of Knowledge Discovery process

The Knowledge Discovery in Databases process comprises of a few steps leading from raw data collections to some form of new knowledge. The iterative process consists of the following steps:

- **Data cleaning**: also known as data cleansing, it is a phase in which noise data and irrelevant data are removed from the collection.
- **Data integration**: at this stage, multiple data sources, often heterogeneous, may be combined in a common source.
- **Data selection**: at this step, the data relevant to the analysis is decided on and retrieved from the data collection.
- Data transformation: also known as data consolidation, it is a phase in which the selected data is transformed into forms appropriate for the mining procedure.
- Data mining: it is the crucial step in which clever techniques are applied to extract patterns potentially useful.
- Pattern evaluation:in this step, strictly interesting patterns representing knowledge are identified based on given measures.
- Knowledge representation: is the final phase in which the discovered knowledge is visually represented to the user. This essential step uses visualization techniques to help users understand and interpret the data mining results.

The KDD is an iterative process. Once the discovered knowledge is presented to the user, the evaluation measures can be enhanced, the mining can be further refined, new data can be selected or further transformed, or new data sources can be integrated, in order to get different, more appropriate results.

Data mining derives its name from the similarities between searching for valuable information in a large database and mining rocks for a vein of valuable ore. Both imply either sifting through a large amount of material or ingeniously probing the material to exactly pinpoint where the values reside. It is, however, a misnomer, since mining for gold in rocks is usually called "gold mining" and not "rock mining", thus by analogy, data mining should have been called "knowledge mining" instead. Nevertheless, data mining became the accepted customary term, and very rapidly a trend that even overshadowed more general terms such as knowledge discovery in databases (KDD) that describe a more complete process. Other similar terms referring to data mining are: data dredging, knowledge extraction and pattern discovery.

III. STUDENTS DATA SET AND ITS PREPROCESSING

The data set used in this paper contains the information of the Engineering students ,collected from the Poornima Institue of Enginneing & Technology, Jaipur for a period of four years from 2008 to 2012. This data set consists of 91 records and 18 attribute. Table 1 presents the attributes and their description that exist in the data set as taken from the source database.

Table 1: The Engineering Students Data Set Description

Attributes	Description	Selected Fields
student_id	Registration no. of the student	
student_name	Name of the student	
gender	{M,F}	√
Dob	Date of Birth	
medium	{eng=1 ,hindi=2}	√
stream	{science}	
marks in 12	Physics, Chemistry and Mathematics marks are incuded	√
City	City of the student	
State	Its State	
Address	Student's address	
B.tech Stream	{Computer Science}	
Attendance in B.Tech	Attendance of the students	√
Hostel facility	{Y=1 N=2}	√
B.tech I year marks	{A – 100% - 80%, B – 79% - 70%, C – 69% - 60%, D – 59% - 40%, E - < 40%}	
B.tech II year marks	{A – 100% - 80%, B – 79% - 70%, C – 69% - 60%, D – 59% - 40%, E - < 40%}	
B.tech III year marks	{A – 100% - 80%, B – 79% - 70%, C – 69% - 60%, D – 59% - 40%, E - < 40%}	

B.tech IV year marks	{A – 100% - 80%, B – 79% - 70%, C – 69% - 60%, D – 59% - 40%, E - < 40%}	
Final Aggregate B.tech Marks	Aggregate marks of B. Tech	√

Poornima Group of Colleges endeavours to create a dynamic environment that facilitates interaction and dialogue. To augment the learning resources and promote interdisciplinary research, Poornima Group of Colleges has developed different academic departments in each Institution to generate the spirit of competition.Each department strives for excellence in a defined area of academic pursuit. This is done with an aim to develop it as a Centre of Excellence in the chosen field. Poornima Institue of Enginneing & Technology, Jaipur grants their students an Engineering degree in seven technical specialty, including branches like Computer Science, Information Technology, Cvil , Mechanical, Electronics, Electronic Instrumentation & Control and Electrical Engineering.

As part of the data preparation and preprocessing of the data set and to get better input data for data mining techniques, we did some preprocessing for the collected data before loading the data set to the data mining software, irrelevant attributes were removed. The attributes marked as selected as seen in Table 1 are processed to apply the data mining methods on them. The attributes such as the Student_Name or State, etc. are not selected to be part of the mining process; this is because they do not provide any knowledge for the data set processing and they present personal information of the students, also they have very large variances or duplicates information which make them irrelevant for data mining.

The following steps are performed as part of the preparation and preprocessing of the data set:

- The data set contains 12 missing values in various attributes from 91 record, the records with missing values are ignored from the data set since it doesn't consider a large amount of data. The numbers of records are reduced to 188 record.

,

IV. APPLYING DATA MINING TECHNIQUES ON THE DATA SET

The Figure 1.2 shown below shows the working methodology adopted for data mining techniques methodology used in this paper. The methodology starts from the problem definition, then preprocessing which are discussed in the introduction and the data set and preprocessing sections, then we come to the data mining methods which are association, classification ,clustering, and outlier detection, followed by the evaluation of results and patterns, finally the knowledge representation process.

In this section; we describe the results of applying the data mining techniques to the data in our data set, for each of the four data mining tasks; Association, classification, clustering and outlier detection, and how we can benefited from the discovered knowledge.

We have used SPSS (Statistical Package for the Social Sciences)software for analysing our data set. IBM SPSS is a data management and analysis product produced by IBM SPSS, Inc. in Chicago, Illinois. Among its features are modules for statistical data analysis, including descriptive statistics such as plots, frequencies, charts, and lists, as well as sophisticated inferential and multivariate statistical procedures like analysis of variance (ANOVA), factor analysis, cluster analysis, and categorical data analysis. SPSS is particularly well-suited to survey research, though by no means is it limited to just this topic of exploration.

SPSS is a very powerful and user friendly program for statistical analysis. Descriptive statistics is been used to summarize and present data in a meaningful manner so that the underlying information can be easily understood. SPSS is a comprehensive and flexible statistical analysis and data management solution. SPSS can take data from almost any type of file and use them to generate tabulated reports, charts, and plots of distributions and trends, descriptive statistics, and conduct complex statistical analyses. SPSS is available from several platforms; Windows, Macintosh, and the UNIX systems.

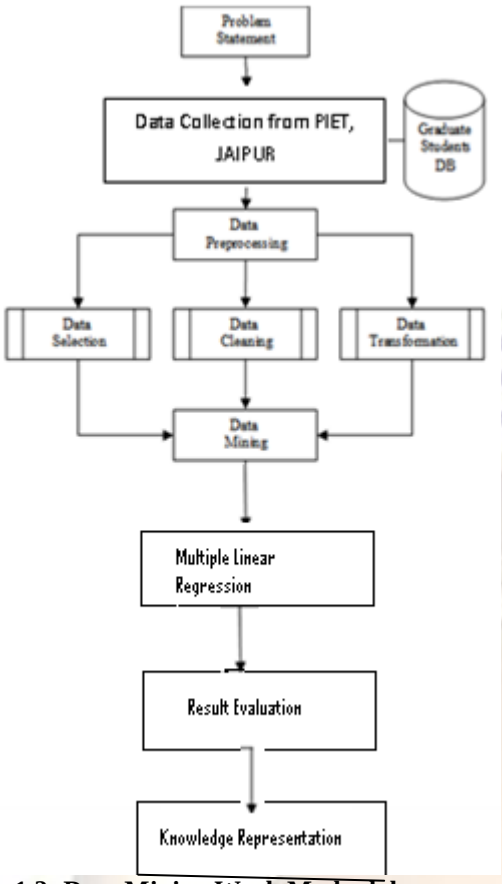


Figure 1.2: Data Mining Work Methodology.

We have used Multiple Linear Regression analysis to assess the relationship between one dependent and several independent variables.Regression co-efficient is a measure of how strong each Independent vaiable predicts the Dependent variable.

Regression analysis is used when we want to predict the value of a variable based on the value of another variable. In this case, the variable we are using to predict the other variable's value is called the independent variable or sometimes the predictor variable. The variable we are wishing to predict is called the dependent variable or sometimes the outcome variable.

V. RESULT ANALYSIS :

SPSS generates quite a few tables in its results section for a linear regression. In this session, we are going to look at the important tables. The first table of interest is the **Model Summary** table. This table provides the R and R^2 value. The R value is 0.723, which represents the simple correlation and, therefore, indicates a high degree of correlation. The R^2 value indicates how much of the dependent variable, MPercentage(which is final percentage of the Engineering students) can be explained by the independent variable, such as Attendence, gender, hostel , Medium and PCM Marks. In this case, 52.3% can be explained, which is very large.

Table 2

Variables Entered/Removed

Model	Variables Entered	Variables Removed	Method
1	Attendance, PCMmarks, Hostel, Gender, SMedium[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: MPercentage

Table 3

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.723[a]	.523	.490	3.864

a. Predictors: (Constant), Attendance, PCMmarks, Hostel, Gender, SMedium

The next table4 is the **ANOVA** table. This table indicates that the regression model predicts the outcome variable significantly well. We will have a look at the "Regression" row and go to the **Sig.** column. This indicates the statistical significance of the regression model that was applied. Here, $P < 0.0005$ which is less than 0.05 and indicates that, overall, the model applied is significantly good enough in predicting the outcome variable.

Table 4

ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1193.934	5	238.787	15.996	.000[a]
	Residual	1089.731	73	14.928		
	Total	2283.665	78			

a. Predictors: (Constant), Attendance, PCMmarks, Hostel, Gender, SMedium

b. Dependent Variable: MPercentage

The table5, **Coefficients**, provides us with information on each predictor variable. This provides us with the information necessary to predict MPercenntage from Attendence, PCMmarks, Hostel, Gender, SMedium. We can see that all the

Independent variables (Gender, Hostel, SMedium, PCMmarks, Attendance) along with the constant contribute significantly to the model (by looking at the **Sig.** column). By looking at the **B** column under the **Unstandardized Coefficients** column we can present the regression equation as:

MPercentage = 29.586 + 5.274(Gender) + (-3.373)(SMedium) + (-1.462)(Hostel) + 0.042(PCMmarks) + .375(Attendance)

Table 5

Coefficients[a]

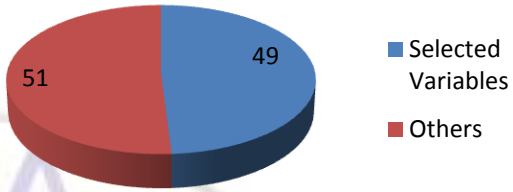
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	29.586	8.788		3.366	.001
	Gender	5.274	.969	.482	5.445	.000
	SMedium	-3.373	1.226	-.312	-2.752	.007
	Hostel	-1.462	1.050	-.116	-1.393	.168
	PCMmarks	.042	.009	.530	4.841	.000
	Attendance	.375	.112	.282	3.354	.001

a. Dependent Variable: MPercentage

VI. CONCLUSION

After the analysis of the result generated from SPSS using Multiple Linear Regression methods we have concluded that a all our Independent variables contribute to 49% in the performance or the final Marks (Dependent variable) of the Engineering students. Using the Significance value calculated we can state that our model is good enough to predict the results form the provided input.

Performance Evaluation Factors



VII. FUTURE SCOPE

This paper is a research work to predict the academic performance of Engineering students based on varying factors like Attendence, gender, hostel , Medium and PCM Marks . Our next work would be to include more factors such as Marks of each semester of Engineering, Qualification of faculty members and the behavioral factors of students for performance evaluation.

REFERENCES

- Baker, R., & Yacef, K. (2009). The State of Educational Data mining in 2009: A Review Future Visions. *Journal of Educational Data Mining, 1*
- Berson, A., Smith, S., & Thearling, K. (2011). An Overview of Data Mining Techniques
- Research in Higher Education Journal Educational data-mining research
- Blikstein, P. (2011). *Using learning analytics to assess students' behavior in open programming tasks*.
- Dunham, M. (2003). *Data Mining: Introductory and Advanced Topics*
- Guan, J., Nunez, W., & Welsh, J. (2002). Institutional strategy and information support: the role of data warehousing in higher education.174.
- Luan, J. (2002). *Data Mining and Knowledge Management in Higher Education Applications*. Paper presented at the Annual Forum for the Association for Institutional
- *A survey of Education data mining research*, Richard A. Huebner,Norwich University