ISSN No. 0976-5697

Volume 7, No. 3, May-June 2016

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Mining Educational Data from Student's Management System

Oshin Mundada
B.E., Computer Science and Engineering
Maharashtra, India

Abstract: Database management systems have grown exponentially since their inception. This has led to a huge amount of data, but not knowledge. The same is the case for the educational sector, where data is plenty, but the advantages that can be inferred from this data are lean. Thus, educational data mining is a manifesting branch of knowledge concerned with evolving methods for discovering knowledge from data which comes from the educational domain. This paper is the outcome of a research carried out on students of a coaching class. The data includes five years of student data to which several mining techniques have been applied. We have used the classification rules to predict the performance of the student in competitive exams. This helps the teachers in early identification of the weaker students or students who need more attention and allow them to act appropriately, eventually increasing the result of the class.

Keywords: Educational Data Mining (EDM), Prediction, Classification, Naïve Bayesian model

I. INTRODUCTION

Data Mining, alternatively known as Knowledge Discovery in Databases (KDD), is a discipline used to extract potentially useful information from extensive data. Data mining is bestdescribed as the union of historical and recent developments instatistics, AI, machine learning and Database technologies. These techniques are then used together to study data and find previously-hidden trends or patterns within [1].

Data Mining has been used in multifold fields of varying diversities, such as finance, telecommunication, retail, biological, etc. One such field in that of education and this area of inquiry is specifically termed as Educational Data Mining. It is of pronounced significance to achieve better quality and results in the learning methodologies of students and teaching methodologies of educators. The repository of data from multitudinous students with analogous experience has given us anchorage for predicting the diverse factors like personal, social and psychological affecting students' performance.

In this paper, we have used data of a coaching class for students of XI and XII, preparing for the entrance examinations of Indian Institute of Technology. Mining the data available from their student's management system about the student's background, results of weekly tests and previous trends, we analyze the probability of the student's good or bad performance in the final examination. Based on this, we show the predicted rank, college and course achievable for each student.

Mining in the educational domain allows the teachers to assist students with focus on student's weaknesses, early detection of dropout students and consequently provision of appropriate counseling. It helps the students to track their strengths and shortcomings and subsequently work and improve on the same.

This paper uses various techniques of Educational Data Mining like Classification, Prediction, decision tree and Naïve Bayesian to achieve the desired results.

II. DEFINITION AND TECHNIQUES / MAIN APPROACHES:

Educational Data Mining makes use of multiple techniques like decision tree, rule induction, neural networks, k-nearest neighbor, naïve Bayesian and many others. From this common listing, clustering, prediction and relationship mining are considered universal method across all types of data mining.

Data mining is often reckoned as a step in the KDD process. The steps followed for extracting knowledge from data in this paper is depicted in the diagram below:

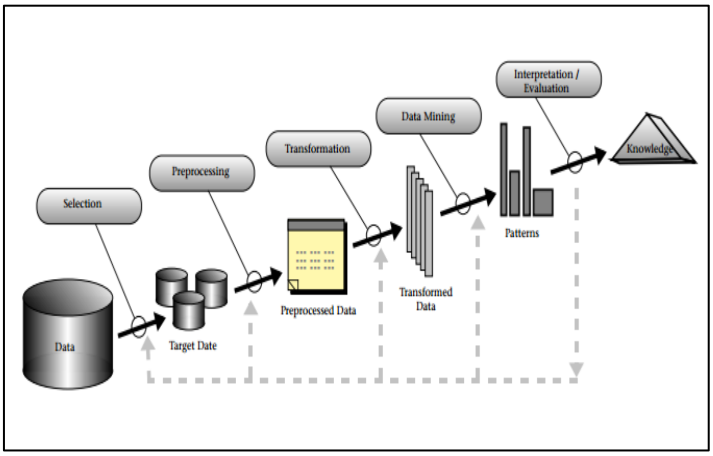

Fig.1: Diagrammatical depiction of steps of KDD

Here is a brief introduction to the methods and techniques used in this paper.

A. Classification

© 2015-19, IJARCS All Rights Reserved

Classification is the most commonly applied data mining technique, which employs a set of pre-classified examples to develop a model that can classify the population of records at large. This approach frequently employs decision tree or neural network-based classification algorithms. The data classification process involves learning and classification. In Learning the training data are analyzed by classification algorithm. In classification test data are used to estimate the accuracy of the classification rules. If the accuracy is acceptable the rules can be applied to the new data tuples. The classifier-training algorithm uses these pre-classified examples to determine the set of parameters required for proper discrimination. The algorithm then encodes these parameters into a model called a classifier. [2]

B. Decision Trees

A decision tree is a tree structure which classifies an input sample into one of its possible classes. Decision trees are used to extract knowledge by making decision rules from the large amount of available information. A decision tree classifier has a simple form which can be compactly stored and that efficiently classifies new data. [3]

C. Statistics

It is a technique to identify outlier fields, record using mean, mode, etc. and hypothetical testing. This technique is used to improve the student's management system and student response system.[4]

D. Prediction

It is a technique which predicts a future state rather than a current state. This technique is useful to predict success rate, drop out, and retention management of students.[5]

E. Naïve Bayes

Naïve Bayesian classifiers assume that the effect of an attribute value on a given class is independent of the values of the other attributes. This assumption is called class-conditional independence. It is made to simplify the computations involved and, in this sense, considered "naïve". [6]

III. DATA MINING PROCESS

In accord with the current scenario, the student has to appear for two examinations before getting admitted to the Indian Institute of Technology. His/her performance in these examinations is based on several factors, including his XII board results. This study weighs the different factors, and predicts the final outcome of whether the student will get admittance from the present 17 IITs in India.

The Data Mining process used in this paper is shown in the diagram below:

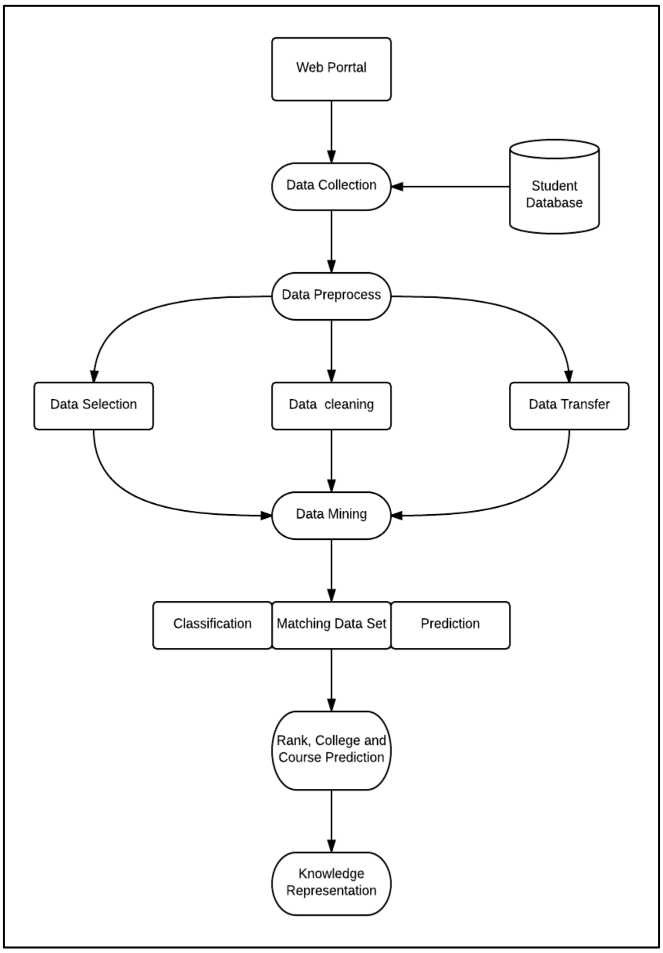

Fig.2: Process of Data Mining

A. Data Collection

The data used in this study comes from the student database of the web portal of the institute. The students are required to fill certain information during admission. Also, the performance of the students in the weekly tests is recorded through the online portal. In this step, the data stored in different tables is joined. Errors are removed, if any.

B. Data Preprocessing

The next step is that of data processing, which includes data selection, data cleaning and data transformation. Only those fields are selected which are required. Some derived values are calculated. The variables considered during the study are depicted in Table 1.

Table 1: VARIABLES UNDER CONSIDERATION

Variable	Description	Possible Values
SEX	Sex	{Male, Female}
ATT	Attendance	{Good, Average, Poor}
CAT	Category	{General, SC, ST, OBC}
BRD	X Board	{SSC, CBSE, ICSE}
BRDP	X Board Percentage	{ A > 90%, B > 80% and <90%, C > 70% and <80%, D > 55% and <70%, E > 35% and <55%,

		F < 35% }
FC	Foundation Course	{Yes, No}
TP	Test Performance	{Good, Average, Poor}
ER	End Result	{First, Second, Third, None}

The domain values for these variables in the present study are defined as follows:

- **SEX** -Sex of student. Gender of the student is considered. It can be either *Male* or *Female*
- **ATT –** Attendance of Student.This variable defines the attendance of the student in class. It is divided into 3 categories: *Good* being >75%, *Average* being between 50% to 75%, and *Poor* being <50%.
- **CAT –** Category/Caste. As the IITs provide reservation seats for different categories, three categories are considered: *General, SC, ST, and OBC.*
- **BRD –** Board from which student passed X examination. The current three boards in India are: State Board/*SSC*, *CBSE* and *ICSE*.
- **BRDP –** X Board Percentage. The performance of the student in the X Board Examination. We have divided it into six grades - A > 90%, B > 80% and <90%, C > 70% and <80%, D > 55% and <70%, E > 35% and <55%, F < 35%.
- **FD –** Foundation Course. This is divided into two classes: Yes – Student has attended foundation course, No – Student has not attended foundation course.
- **TP –** Test Performance. This parameter depicts the performance of the students in the weekly conducted tests. This is divided into three categories: Good, Average and Poor.
- **ER –** End Result. End result is the performance of the student in the final IIT examination with four values: First – The student gets into the first 7 ranked IITs, Second - The student gets into the next 5 ranked IITs, Third – The student gets into IITs ranked from 13 to 17, None – The student does not get into any IIT. The category division is shown in the table below:

Table 2: Classification of IIT colleges

First	IIT Kharagpur, IIT Bombay, IIT Kanpur, IIT Madras, IIT Delhi, IIT Guwahati, IIT Roorkee
Second	IIT Bhubaneshwar, IIT Gandhinagar, IIT Hyderabad, IIT Jodhpur, IIT Patna
Third	IIT Ropar, IIT Varanasi, IIT Indore, IIT Mandi, IIT Dhanbad

C. Implementation of Mining Algorithm

Various algorithms and techniques like Classification, Clustering, Regression, Artificial Intelligence, Neural Networks, Association Rules, Decision Trees, Genetic Algorithm, Nearest Neighbor method etc., are used for knowledge discovery from databases. Classification is one of the most frequently studied problems by data mining and machine learning (ML) researchers. It consists of predicting the value of a (categorical) attribute (the class) based on the values of other attributes (the predicting attributes). There are different classification methods. In the present study we use the Bayesian Classification algorithm.

Bayes classification has been proposed that is based on Bayes rule of conditional probability. Bayes rule is a technique to estimate the likelihood of a property given the set of data as evidence or input Bayes rule or Bayes theorem is-

$$P(h_i|x_i) = \frac{P(x_i|h_i)P(h_i)}{P(x_i|h_i) + \ P(x_i|h_2)P(h_2)}$$

The approach is called "naïve" because it assumes the independence between the various attribute values. Naïve Bayes classification can be viewed as both a descriptive and a predictive type of algorithm. The probabilities are descriptive and are then used to predict the class membership for a target tuple. The naïve Bayes approach has several advantages: it is easy to use; unlike other classification approaches only one scan of the training data is required; easily handle mining value by simply omitting that probability. An advantage of the naive Bayes classifier is that it requires a small amount of training data to estimate the parameters (means and variances of the variables) necessary for classification. Because independent variables are assumed, only the variances of the variables for each class need to be determined and not the entire covariance matrix. In spite of their naive design and apparently over-simplified assumptions, naive Bayes classifiers have worked quite well in many complex real-world situations.

The data set of 50 students used in this study is obtained from the students of the institute. It is shown in Table 3. The data set consists of 33 males and 17 females.

Table 3: DATA SET

SEX	ATT	CAT	BRD	BRDP	FD	TP	ER
Male	Good	General	CBSE	A	Yes	Good	First
Female	Average	SC	SSC	A	No	Average	Third
Male	Poor	ST	SSC	C	No	Poor	No
Female	Good	General	ICSE	B	Yes	Good	First
Male	Average	OBC	CBSE	D	Yes	Average	Third
Female	Good	General	SSC	D	No	Poor	No
Male	Good	SC	SSC	B	No	Good	Second
Male	Average	General	CBSE	A	No	Average	Third
Male	Poor	OBC	ICSE	C	Yes	Average	No
Female	Poor	ST	SSC	B	No	Poor	No
Male	Good	General	SSC	C	No	Average	Third
Male	Good	General	SSC	A	Yes	Good	Second
Female	Good	ST	CBSE	B	Yes	Average	Third
Male	Average	SC	SSC	B	Yes	Good	First
Female	Average	OBC	CBSE	A	No	Average	Second
Male	Average	General	ICSE	C	Yes	Good	Third
Male	Poor	General	CBSE	B	No	Good	Third
Female	Good	General	ICSE	B	Yes	Good	Second
Male	Good	General	CBSE	B	Yes	Good	Second
Male	Good	General	SSC	A	Yes	Good	Second
Female	Average	ST	CBSE	E	Yes	Poor	No
Female	Poor	SC	SSC	C	No	Poor	No
Male	Average	ST	CBSE	A	No	Average	Second
Male	Average	ST	CBSE	A	Yes	Good	First
Female	Poor	SC	SSC	A	Yes	Average	Second
Male	Good	General	SSC	E	Yes	Poor	No
Male	Good	General	SSC	C	Yes	Average	Third
Male	Good	SC	ICSE	A	No	Good	First

© 2015-19, IJARCS All Rights Reserved

Male	Average	ST	SSC	B	Yes	Average	Third
Male	Poor	OBC	CBSE	C	No	Poor	No
Female	Good	General	CBSE	A	Yes	Good	First
Male	Good	SC	SSC	B	No	Good	First
Male	Average	General	SSC	A	No	Good	Second
Male	Poor	SC	CBSE	B	Yes	Average	Third
Female	Good	OBC	ICSE	C	No	Average	Third
Male	Average	ST	SSC	B	No	Good	First
Male	Good	SC	SSC	E	Yes	Poor	No
Female	Average	General	CBSE	A	Yes	Good	First
Female	Average	OBC	SSC	C	No	Average	Third
Male	Good	General	ICSE	A	No	Good	Second
Male	Good	SC	SSC	B	Yes	Poor	No
Male	Average	ST	ICSE	D	Yes	Average	Third
Female	Good	General	SSC	C	Yes	Good	Third
Male	Poor	General	CBSE	A	No	Good	Third
Male	Average	OBC	CBSE	B	Yes	Average	Second
Male	Good	General	ICSE	A	Yes	Average	Second
Female	Average	SC	CBSE	B	No	Average	Second
Female	Poor	OBC	ICSE	D	Yes	Poor	No
Male	Good	General	CBSE	B	Yes	Good	Second
Male	Average	ST	SSC	A	No	Good	Second

IV. RESULTS AND DISCUSSION

The features that were used for prediction model construction are shown above. For both variable selection and prediction model construction, we have used WEKA tool.

In the study it is found that the students' performance is highly dependent on their performance in the weekly tests. This is shown in Figure 3.

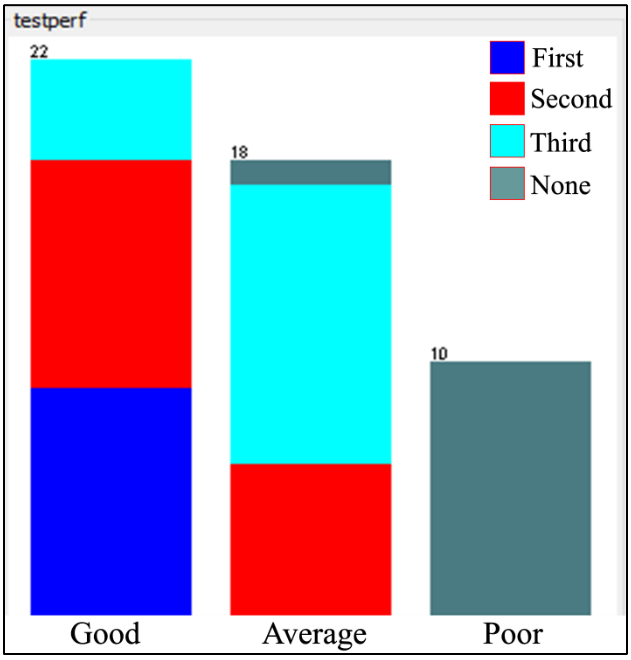

Fig.3: Relationship between TP and ER

It is also found that the second high potential variable for students' performance is their category. The relationship between students' caste and their result in end examination is shown in Figure 4.

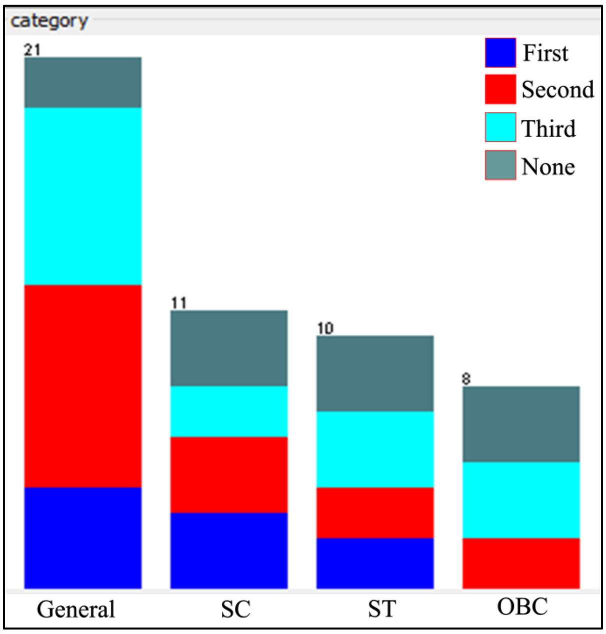

Fig.4: Relationship between CAT and ER

It is found that the third high potential variable for students' performance is performance in board examination. IIT claims the examination to be based on the CBSE syllabus. The relationship between students' board performance and their end result in IIT examination is shown in Fig 5.

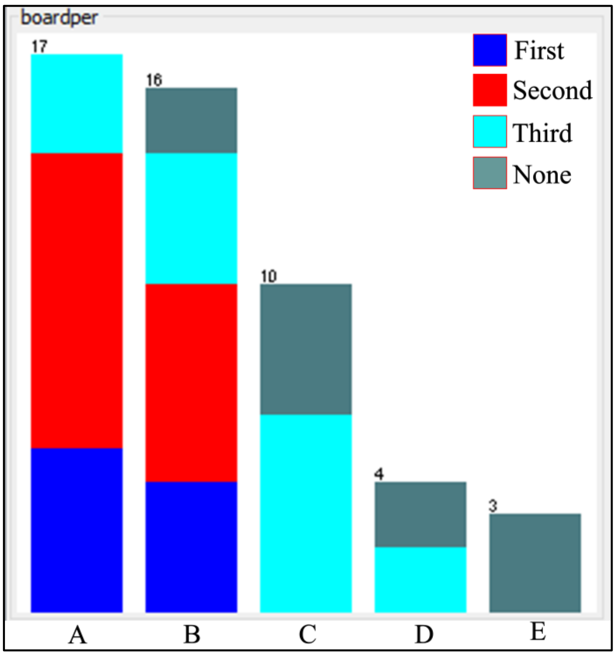

Fig.5: Relationship between BRDP and ER

In this paper, Bayesian classification method is used on student database to predict the student will get into which class of IIT. This study will help the students and the teachers to improve the performance of the student. This study will also work to identify those students which need special attention to reduce failing ratio and taking appropriate action at right time.

Present study shows 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. This proposal will improve the insights over existing methods.

V. CONCLUSION

Owing to the above results and discussions, we see that there are certain internal and external factors that impact a student's performance. The number of correctly classified instances comes to 72%, with an accuracy of 0.833. Only the Naïve Bayes classification method achieves this level of accuracy.

Now, with this data in hand, whenever a new student enrolls, he/she can be classified correctly so as to achieve maximum result. The students who need special attention, constant motivation, more focus on foundation, etc. can be figured to help them accordingly.

VI. ACKNOWLEDGMENT

This research is a consequence of thesupport from Velocity Classes, a coaching institute for aspirants of IIT. I thank them and Ms. ParminderKaur for her comments and insights, that helped improve the manuscript.

VII. REFERENCES

[1] DharminderKaur, Deepak Bhradwaj, "Rise of Data Mining: Current and Future Application Areas" International Journal of Computer Science Issues, Vol. 8, Issue 5, No 1, September 2011

[2] Brijest Kumar Baradwaj, Saurabh Pal, "Mining Educational Data to Analyze Students' Performance", International Journal of Advanced Computer Science and Applications, Vol.2, No. 6, 2011

[3]ShahrukhTeli, PrashastiKanikar, "A Survey on Decision Tree Based Approaches in Data Mining", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5, Issue 4, 2015

[4] Campbell, J.P., and Oblinger, D.G. (2007,Oct.). Academic Analytics. EDUCAUSE, Washington, D.C. [Online],Available:http://net.educause.edu/ir/library/pdf/pub6101.pdf,2007.

[6] Jiawei Han, MichelineKamber, Jian Pei, "Data Mining Concepts and Techniques", Morgan Kaufmann

[7] Rajni Jindal and Malaya Dutta Borah, "A Survey On Educational Data Mining And Research Trends", International Journal of Database Management Systems (IJDMS) Vol.5, No.3, June 2013

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

[9] Umamaheswari. K, S. Niraimathi, "A Study on Student Data Analysis Using Data Mining Techniques",International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 8, August 2013

[10] Mohammed M. Abu Tair, Alaa M. El-Halees, "Mining Educational Data to Improve Students' Performance: A Case Study", International Journal of Information and Communication Technology Research, Volume 2 No. 2, February 2012

© 2015-19, IJARCS All Rights Reserved