JOURNAL OF ADVANCEMENT IN
ENGINEERING AND TECHNOLOGY

Journal homepage: http://scienceq.org/Journals/JAET.php

Research Article Open Access

A Framework for Student Academic Performance Using Naive Bayes Classification Technique

[1]Y Divyabharathi, [2]P Someswari

[1,2] Department of Computer Science Engineering,, GMR Institute of Technology, Andhra Pradesh, INDIA.

*Corresponding author: Y Divyabharathi, E-mail: ydivyabharathi@gmail.com

Received: April 25, 2018, Accepted: June 05, 2018, Published: June 05, 2018.

ABSTRACT

The real fact in the education institute is the significant growth of the educational data. Data mining techniques are used to extract the useful information and to predict the student academic performance. The main aim of this paper is to construct predictive model for student academic performance. As there are many classification techniques are available, in this paper naive bayes classification technique is used. This paper presents and analyses the experience of applying certain data mining methods and techniques on student data in order to prevent academic risk and desertion.

Keywords: *Data Mining, Predictive modeling, Academic risk prevention, Academic Performance, Educational data mining.*

1. INTRODUCTION

Educational Data Mining has interesting research area and it has become a vital need for the academic institutions to improve the quality of education. One important reason for educating individuals is to create an enabling environment. For this reason, the quality of students and their academic achievement has become critical and drawn much attention from the research community as it plays a significant role in determining the worth of graduates who will be responsible for economic and social growth of the country. Student low academic performance [1] in the engineering dynamics course has been a long-standing problem. Before designing and implementing any instructional interventions to improve student learning in engineering dynamics, it is important to develop an effective model to predict student academic performance in this course so the instructor can know how well or how poorly the students in the class will perform.

To predict student academic performance using data mining techniques, data mining is the process of sorting through large data sets to identify patters and establish relationships to solve problems through data analysis. Data mining tools allow enterprises to predict future trends. This study focused on developing and validating mathematical models that can be employed to predict student academic performance in institutions. Student dropouts and failures as a phenomenon has been extensively studied and modelled. To identify the possible causes in order to seek strategies to prevent it.

In which predicting student academic performance using naïve Bayes classification algorithm, Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data. For example, a classification model [2] could be used to identify loan applicants as slow, medium, or high credit risks. The main reason for constructing a predictive model for student academic performance using naive Bayes is student dropouts and failures in an educational institute.

This project should allow the institutions to make timely decisions and design better strategies to prevent academic risk as a phenomenon and its consequences, decreasing the likelihood of dropping out. In this we are implemented Naive Bayes technique in order to exploit

the information that the institution collects from students. This practice is known as Educational Data Mining.

2. EXISTING SYSTEM

The real fact in the education institute is the significant growth of the educational data. Data mining techniques are used to extract the essential information and to explore the relationships between variables stored in the data warehouse. Data mining models [3] have been developed to predict student academic performance. In order to identify the possible causes for student desertion by constructing the predictive model using decision tree, artificial neural networks and other classification techniques. By using data mining techniques, we can monitor the student academic performance, the system should alert according to the grades obtained by students in each period. These alerts classify students into three levels of risk low, medium, high risk. So that act on the causes of risk before the risk occurs.

3. PROPOSED SYSTEM

Student low academic performance and drop outs as the major standing problem so that we need to prevent the academic performance of a student before the risk occurs. IN ORDER to predict the student academic performance Naive Bayes classification technique [4] is used. By using this model, we can take timely decisions in order academic risk of a student. Predictive modelling is the process that uses the data mining and probability to estimate outcomes. The objective of this task is to predict the risk of a particular student based on the other attributes like marks and attendance. So that the instructor can know how well or how poorly the students in the class will perform. This study focused on developing and validating mathematical models that can be employed to predict student academic performance in educational institutions.

4. PROPOSED METHODOLOGY

The main aim of the proposed system is to build the classification model that classifies a students' performance. The classifiers, has been built by combining the Standard for Data Mining that includes student performance and finally application of data mining techniques which is classification in present study. In other words, using this Naive Bayes algorithm, we wanted to be able to guide student towards achievement of good score that we felt they would enjoy doing. Naive Bayes methods classify instances in this we take the student data set and the data set contains the id, name, mid averages of five subjects, attendance and risk. In which we take the class label as risk and that should be predicted. In this We Generalized Naive Bayes algorithm for predicting the student's performance as pass or fail. Once the student is found at the risk of failure he/she can be provided guidance for performance improvement.

For building a predictive model of academic performance of students based on data mining is necessary to develop four major phases. First they must be extracted and prepared feed the data mining process[5]. Secondly, the process must be implemented data mining consists of the data pre-processing, algorithm execution, analysis results - rules. Thirdly, the predictive model must be formulated analysing, selecting and defining the set of rules that allow a proper prediction of academic performance in relation to the institutional context and objectives for the model. Finally, the predictive model must be validated by applying it to different data sets with characteristics similar to that was used during the mining process. This paper presents the development of the first two stages, analysing the data mining process [6] and its applicability in building predictive models in educational settings. Fig. 1 illustrates the method used.

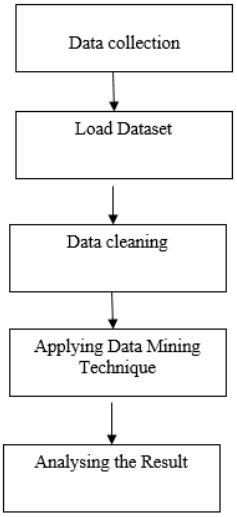


Fig 1: Project work flow diagram

4.1 DATA COLLECTION

Data collection is the systematic approach to gathering and measuring information from a variety of sources to get a complete and accurate picture of an area of interest. Data collection enables a person or organization to answer relevant questions, evaluate outcomes and make predictions about future probabilities and trends. Data mining requires a significant amount of data to provide meaningful results. Data collection is a component of research in all fields of study including physical and social sciences, humanities, and business. In which collected the educational data of students. Collected the records of 500 students from educational institutes [7]. The data should contain the name, id, mid averages of five subjects, attendance, and risk factor.

	id	name	C	JAVA	DE	PYTHON	DBMS	Attendan	Risk
2	14341a0101	THALATHOTI HARSHITHA	16	17	16.5	16	16	75	no
3	14341a0102	BEHARA VENKATA RAMA RATNAM	16	16	16	16	16	77	no
4	14341a0103	BURALLI NARENDRA KUMAR	16	16	16	16	16	88	no
5	14341a0104	CHOWDAVADA VARA PRASAD	16	17	16.5	16	16	99	no
6	14341a0105	JUJJUVARAPU SUMANTH	13	15	14	13	13	77	no
7	14341a0106	PETTA BHARGAV	12	19	15.5	12	12	74	yes
8	14341a0107	VARADI KUMAR RAJU	17	19	18	17	17	75	no
9	14341a0108	ABHISHEK KANKIPAATI	4	10	7	4	4	76	yes
10	14341a0109	ADABALA SAI MOHAN	12	14	13	12	12	45	yes
11	14341a0110	ADAPA INDRAJA	13	13	13	13	13	46	yes
12	14341a0111	ADAPAKA SANGAMITRA	15	9	12	15	15	42	yes
13	14341a0112	ADHIKARLA MADHURI	15	12	13.5	15	15	63	yes
14	14341a0113	AJAY KUMAR REDDY YEKKANTI	16	14	15	16	16	63	yes
15	14341a0114	AKHIL JAVVADI	14	8	11	14	14	63	yes
16	14341a0115	ANALA KIRANMAYI	13	9	11	13	13	85	yes
17	14341a0116	ANDHAVARAPU MEGHANA	18	18	18	18	18	85	no
18	14341a0117	ANGINA VASUDHA	17	14	15.5	17	17	95	no
19	14341a0118	ANUSHA PEDDINA	16	15	15.5	16	16	96	no
20	14341a0119	ATYAM DIVYA	0	10	5	0	0	60	yes
21	14341a0120	BADE DINESH	18	19	18.5	18	18	82	no
22	14341a0121	BAKKA ROHINI	19	12	15.5	19	19	83	no
23	14341a0122	BALAGAM AVANI	17	19	18	17	17	81	no
24	14341a0123	BALIVADA SAI TEJA	16	19	17.5	16	16	78	no
25	14341a0124	BANDARU SRI DATTA	18	12	15	18	18	79	no

Fig 2: Trained Data set

4.2 LOAD THE DATA

The main objective is to generate a predictive model for the student academic performance for this purpose student's data is selected and extracted from the database information[8,9] corresponding to students. Data mining requires a significant amount of data to provide meaningful results. In this regard dataset consists of anonymous records of 500 students from educational institutional
database, which were delivered in four separate spreadsheets were obtained.

4.3 DATA PREPROCESSING

Data pre-processing is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Data pre-processing is a proven method of resolving such issues. Data pre-processing prepares raw data for further processing. Data pre-processing is used database-driven applications such as customer relationship management and rule-based applications.
Data goes through a series of steps during pre-processing:

- Data Cleaning: Data is cleansed through processes such as filling in missing values, smoothing the noisy data, or resolving the inconsistencies in the data.
- Data Integration: Data with different representations are put together and conflicts within the data are resolved.
- Data Transformation: Data is normalized, aggregated and generalized.
- Data Reduction: This step aims to present a reduced representation of the data in a data warehouse.
- Data Discretization: Involves the reduction of a number of values of a continuous attribute

by dividing the range of attribute intervals.

During this phase, some pre-processing techniques are applied for the collected data to prepare it for the mining techniques. At first, some irrelevant attributes, e.g. student name, nationality, and campus are eliminated. And all the data related to the general and elective courses are also eliminated. In this mainly focusing on the program mandatory courses and re-arranged the Table so that each student has the following attributes: Student ID, final GPA, and the course grades student took during a four-year study program. In the final step, the numerical attributes are converted into categorical ones. For

example, the students final GPA is categorized into five groups: excellent, very good, good, average and poor. In the same way, the students' grade is assigned for each course into: A+, A, B+, B, C+, C, D, D+ and F. The following table demonstrates a sample of the data

Table 1: Sample of the dataset

Student ID	Final GPA	JAVA	Operating System
1	Excellent	A+	A
2	Good	C	B+

4.3.1 DATA CLEANING

Data cleansing is the process of altering data in a given storage resource to make sure that it is accurate and correct. There are many ways to pursue data cleaning in various software and data storage architectures; most of them center on the careful review of data sets and the protocols associated with any particular data storage technology.

Data cleansing is sometimes compared to data purging, where old or useless data will be deleted from a data set. Although data cleansing can involve deleting old, incomplete or duplicated data, data cleansing is different from data purging in that data purging usually focuses on clearing space for new data, whereas data cleansing focuses on maximizing the accuracy of data in a system. A data cleansing method may use parsing or other methods to get rid of syntax errors, typographical errors or fragments of records. Careful analysis of a data set can show how merging multiple sets led to duplication, in which case data cleansing may be used to fix the problem.

Many issues involving data cleansing are similar to problems that archivists, database admin staff and others face around processes like data maintenance, targeted data mining and the extract, transform, load methodology, where old data is reloaded into a new data set. These issues often regard the syntax and specific use of command to effect related tasks in database and server technologies like SQL or Oracle. Database administration is a highly important role in many businesses and organizations that rely on large data sets and accurate records for commerce or any other initiative.

4.4 NAIVE BAYES CLASSIFICATION ALGORITHM

The Bayesian Classification represents a supervised learning method as well as a statistical method for classification. Assumes an underlying probabilistic model and it allows us to capture uncertainty about the model in a principled way by determining probabilities of the outcomes. It can solve predictive problems. This Classification is named after Thomas Bayes, who proposed the Bayes Theorem. Bayesian classification provides practical learning algorithms and prior knowledge and observed data can be combined. Bayesian Classification provides a useful perspective for understanding and evaluating many learning algorithms. It calculates explicit probabilities for hypothesis and it is robust to noise in input data.

Naive Bayes is a classification technique based on Bayes Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. For example, a fruit may be considered to be an apple if it is red, round, and about 3 inches in diameter. Even if these features depend on each other or upon the existence of the other features, all of these properties independently contribute to the probability that this fruit is an apple and that is why it is known as 'Naive'.

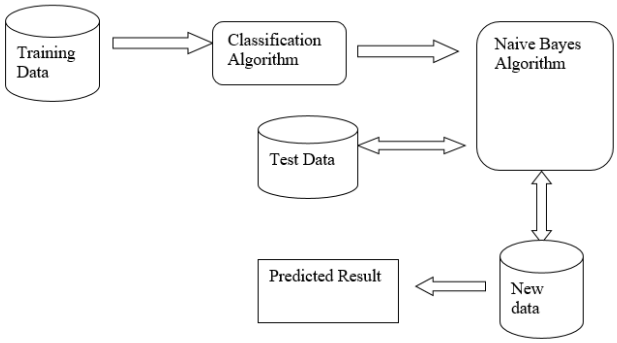
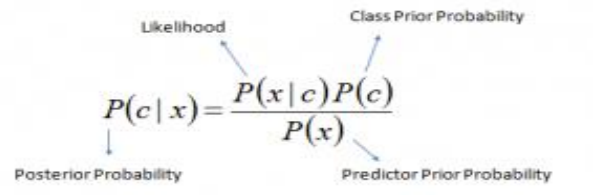


Fig 3: Prediction using classification technique

This is one of the important techniques for tracking or monitoring academic risk, performed in the information system and academic management. In this system alerts according to the grades obtained by students in each period and attendance and its relationship to other historical records of performance are generated. These alerts classify students into two levels of risk: low and high. The latter triggers a process of support and intervention to the student, led by Student Support Program. By using naive Bayes classification technique identifying the causes of academic risk and act on these causes before risk occurs. So it allow the institution to make timely decisions and design better strategies to prevent academic risk as a phenomenon and its consequences, decreasing the likelihood of dropping out. In this, in order to exploit the information that the institution collects from students Naive Bayes technique is implemented. This practice is known as Educational Data Mining.

Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods. Bayes theorem provides a way of calculating posterior probability P(c|x) from P(c), P(x) and P(x|c). Look at the equation below:



$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Above,

- $P(c|x)$ is the posterior probability of *class* (c, *target*) given *predictor* (x, *attributes*).
- $P(c)$ is the prior probability of *class*.
- $P(x|c)$ is the likelihood which is the probability of *predictor* given *class*.
- $P(x)$ is the prior probability of *predictor*.

5. RESULTS AND DISCUSSIONS

Data mining techniques can predict the student academic performance based on the historical data and find out the retention rate of the students.

The prediction results of the classification methods are presented in Table 4.1. In the confusion matrixes the rows represent the actual and the columns represent the predictions. A confusion matrix is a table that is often used to describe the performance of

a classification model (or "classifier") on a set of test data for which the true values are known.

Table 2: Confusion Matrix

N=150	Predicted-no	Predicted-yes	Total
Actual-no	50(tn)	6(fp)	56
Actual-yes	4(fn)	91(tp)	95
Total	54	97	

Below shows the prediction accuracy and misclassification for the output variable values. As the results indicate, the classification method performed reasonably well in predicting the student academic performance. Based on the result,Naive bayes algorithm produced the best prediction results with 94% overall accuracy.

Accuracy

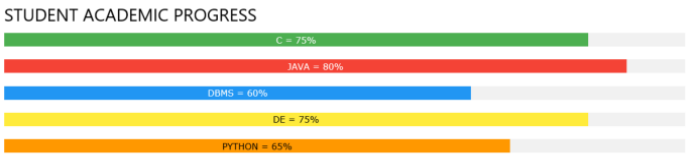
(Tp+tn)/150=((91+50)/150)*100)

=0.94=94%

Misclassification

(Fp+fn)/150=((6+4)/150)*100)

=0.06=6%



6. CONCLUSION AND FUTURE WORK

In which the student academic performance is successfully predicted using Naive Bayes classification technique. It helps the management take timely action to improve the student performance through extra coaching and counselling. In this paper the main focus is on the student academic performance in a specific subject based on their performance of test result components during the performance by applying the Naive Bayes Classification algorithm.

Future work include applying data mining techniques on an expanded data sets which consider extracurricular activities and other vocational courses completed by students, which may have a significant impact on the overall performance of the students.

REFERENCES

1. Liga Paura, Irina Arhipova “Student Dropout Rate in Engineering Education Study Program”, Jelgava, 25-27.05.2016.
2. Abeer Badr EI Ahmed, Ibrahim sayed Elaraby “Data Mining: A prediction for Student's Performance Using Classification Method” World Journal of Computer Application and Technology, 2014.
3. Mr.Navin, Prof. Vijay Anand “Analysis of Student performance in Education System in using Methodologies of Data mining” International Journal of Advanced Innovative Technology, in Engineering (IJAITE), Vol. 1, Issue 3, May-2016.
4. BahaSen, EmineUcar “Evaluating the achievements of computer engineering department of distance education students with data mining methods”, Elsevier, 2012.
5. MonikaGoyal1 and Rajan Vohra2 “Applications of Data Mining in Higher Education”, IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 2, No 1, March 2012.
6. I. Davidson and G. Tayi, “Data quality data preparation using matrices for classification mining” European Journal of Operational Research 764-772. 2009.
7. MS Bhullar and A. Kaur "Use of data mining in education sector" Lecture Notes in Computer Science and Engineering 2200, pp. 513-516, 2012.
8. Nikhil Rajadhyax, Rudresh Shirwaikar, Data Mining on Educational Domain, 2012.
9. J.K. JothiKalpana, K. Venkatalakshmi(2014)” Intellectual Performance Analysis of Students by Using Data Mining Techniques” International Journal of Innovative Research in Science, Engineering and Technology Volume 3, Special Issue 3, March 2014 ISSN (Online) : 2319 – 753 ISSN (Print) : 2347 – 6710.

Citation: Y Divyabharathi *et al.* (2018). A Framework for Student Academic Performance Using Naive Bayes Classification Technique, J. of Advancement in Engineering and Technology, V6I3.08. DOI: 10.5281/zenodo.1277183.