World Journal of Engineering Research and Technology

WJERT

www.wjert.org

Impact Factor Value: 5.924

DROPOUT STUDENT PREDICTION USING NAIVE BAYS CLASSIFIER

[*1] Dr.P.S.S.Akilashri, [2]P. Sundari and [3]M. Vijayalakshmi

[1] HOD, PG & Research Department of Computer Science, National College (Autonomous), Tamilnadu, India.

[2]Assistant Professor, PG & Research Department of Computer Science, National College (Autonomous), Tamilnadu, India.

[3]PG Scholar, PG & Research Department of Computer Science, National College (Autonomous), Tamilnadu, India.

Article Received on 24/08/2020 **Article Revised on 14/09/2020** **Article Accepted on 04/10/2020**

***Corresponding Author**
Dr.P.S.S.Akilashri
HOD, PG & Research Department of Computer Science, National College (Autonomous), Tamilnadu, India.

ABSTRACT

The objectives of this research work is to identify relevant attribute from socio-demographic, academic and institutional data of first year students from undergraduate at the University and design a prototype machine learning tool which can routinely distinguish whether the student persist their revise or drop their learning using classification technique based on decision tree. For powerful decision making tool different parameter are need to be considered such as socio-demographic data, parental attitude and institutional factors. The generated knowledge will be quite useful for tutor and management of university to develop policies and strategies related to increase the enrolment rate in University and to take precautionary and consultative procedures and thereby diminish student dropout. It can also use to find the reasons and relevant factors that affect the dropout students.

KEYWORDS: Predicting dropout student, machine learning algorithm, classification, decision tree, Prediction, Classification, Data Mining, and Education.

INTRODUCTION

Machine learning provides tools for analysis of large quantities of data automatically such as feature selection. Feature selection is to choose a subset of input data most useful for analysis and future prediction by eliminating features, which are irrelevant or of no predictive information. Feature selection is use for increasing the predictive accuracy and reducing complexity of learner results. One of the biggest dispute that superior education faces is to progress student dropout rate.[1] Student dropout is a demanding task in higher education and it is reported that about one fourth of students dropped college after their first year. Student dropout has developed into an indication of educational performance and enrolment management. Recent study consequences illustrate that intervention programs can have considerable effects on dropout, particularly for the first year. To successfully exploit the imperfect support resources for the intervention programs, it is pleasing to recognize in progress students who be inclined to necessitate the support most.[2]

The earlier prediction of dropout student is challenging task in the higher education. Data analysis is one technique to level down the rate of dropout students and augment the enrollment velocity of students in the university. It is fact that student dropout quite often in the first year of graduation. Dropout in residential university is caused by academic, family and personal reasons, campus environment and infrastructure of university and varies depending on the educational system adopted by the university. Thus, this learning is moderately functional for enhanced planning and implementation of education program and infrastructure to augment the enrollment rate of students in particular courses provided by the university.[3] The main aim of this paper is to design a classification model using decision tree induction algorithm and classifier rules to predict whether student will graduate or not using the historic data. In this paper, ID3 decision tree algorithm is used to design a conceptual model. Information like age, parent's qualification, parent's occupation, academic record, attitude towards university was accumulated from the student's residing in university campus, to predict list of students who need special attention.[4]

Data mining combines machine learning, statistics and visualization techniques to realize and extract knowledge. Educational Data Mining carries out tasks such as prediction, clustering, association mining, refinement of statistics for human judgment, and discovery with models. Moreover, EDM can resolve many problems based on educational domain. Data mining is non-trivial extraction of implicit, previously unknown and potentially useful information

from huge amounts of data. It is used to predict the prospect trends from the knowledge pattern.[5] The main objective of this dissertation is to exploit data mining methodologies to find students which are expected to drop out their first year of manufacturing. In this examine, the categorization assignment is used to estimate preceding year's student dropout data and as there are lots of approaches that are used for data classification, the Bayesian classification practice is used here. Information similar to marks in High School, marks in Senior Secondary, students family position etc. were composed from the student's management system, to predict list of students who necessitate particular attention.[6]

Problem Definition

In commonly of preceding learning the variables considered to be predictive of dropout intentions are associated to the student instructive surroundings, his/her actual performance at the university and socioeconomic factors. Wide assortments of approaches are used to categorize and authenticate the consequence of such variables for dropout objective prediction. The author of was the first to centre on the dropout problem and encourage the research on this issue.[7] In a statistical descriptive analysis is used. This study concludes that previous academic performance, first year academic performance, class attendance and enrolment date are variables that are directly linked to dropout. In the authors study dropout intention and learning outcomes simultaneously and create a conceptual model that directly relates the two concepts.[8] The conclusion strained from this research is that the level of academic satisfaction is important to predict dropout intention. An additional approach is approve in where the authors complete dropout intention prediction using logistic regression with categorical variables such as stage of study of the parents, parent's profession, sex and first year educational performance. Our contribution uses also machine learning techniques, but evaluate five diverse classification models to predict dropout intention.[9]

Proposed System

The 'Prediction System' Approaches all about institutional practices and processes that are taken into consideration, the student's concerns of the level of the knowledge they receive. The prediction system is a management information system for education establishments to manage student data. An Online prediction scheme is a routine feedback generation arrangement that provides the appropriate feedback to the teachers as per the categories like always, poor, usually, very often, occasionally.[10]

METHODOLOGY

The approach using Naïve byes classifier cognitive was optional for assessment of student's presentation. The elements measured for wide-ranging evaluation of students was educational, attendance and extra curriculum activities.[11] The information was obscured between academic data set that was extractable through data mining methods. The classification assignment was used to assessment performance of student and as there are lots of approaches that are used for data classification, naïve byes classifier was used as included information was extracted that defined performance of student in conclusion semester examination.[12] It was facilitate in recognized drop-out and students who compulsory additional care, authorized lecturer to provide appropriate counsel. In academic organization, performance of student was staunch by inner assessment and previous semester examination.[13] It was carried out using lecturer based upon performance of student in educational actions such as quiz's, discussion, assignments, attendance and lab work. Final semester examination was consequence by student which was obtained minimum marks to pass semester in concluding semester examination.[14] The scheme started from problem definition, then describes data set and pre-processing method executed, and experiments of grades, knowledge representation development.[15]

Architecture Diagram

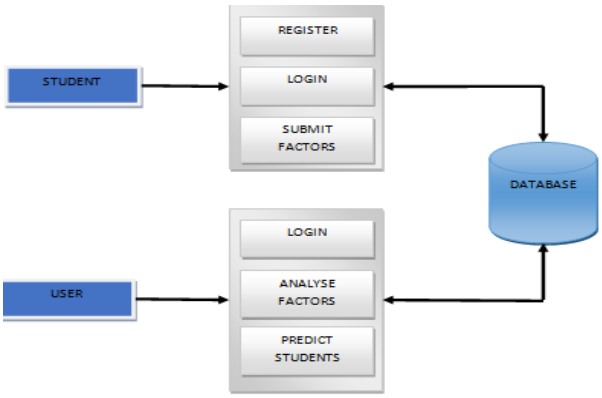


Figure 1: Architecture diagram.

Modules

- Student
- Admin

Student

The student needs to obtain registered with the application by incoming all the details in the registration contact. A student can get logged in by inflowing college name, username, and password. After logging in, each student needs to submit their feedback of previous courses.

Admin

Admin needs to get logged in with valid username and password. Admin need to record student feedback details. The analysis is done based on statistical data for the period 2000 to 2009. The data series were processed in EViews involving the statistical graphics elaboration, the application of ANOVA technique to determine whether there are differences between different populations of students in relative to premature school separation, the correlation matrix production based on statistics from the four regions and the elaboration of box - plot chart distributions to determine whether the data series principles from progress regions be different significantly involving them.

Data Flow Diagram

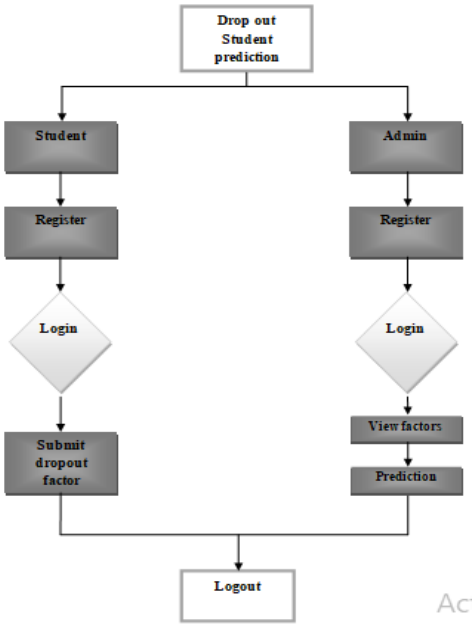


Figure 2: Data Flow diagram.

Data Selection and Transformation

The fields were nominated that was required for data mining which was selected variable, although some of data for variables was mined from data set. The student's performances were conscious that was worn to calculate current students' performance in their approaching semester.

Data clustering

It was arithmetical and unsupervised data investigation technique which was classified duplicate information into homogenous cluster. It was used to activate large data set to determine association and concealed pattern assist to construct decision efficacy and with rapidly. The cluster analysis was student to subdivision huge data set into subsets referred for instance clusters. Every cluster was group of information items that was related to each other was placed surrounded by equivalent cluster but was unconnected to items in other clusters

Classification

It was typically applied technique in academic data mining that predicts cluster exists in data set. Classification was usually applied technique in academic data mining that predicts collection exists in data set. It was worn by scholars in academic stadium to greater understand behaviours of student, to enhanced training capability, and to allocate exchange resolution for inconvenience arises in Department of CS and SE. Classification was exposed representation for predicting academic performance of student to recognize students at danger. The semester effect in order to predict CS and SE student's concluding results.

Data Mining Step

The aim of this step is to extract a hidden pattern from huge amount of data by applying intelligent task such as classification, clustering and association etc. For discovering student dropout pattern, classification using decision tree technique was used. Classification is a supervised learning. It is two-step process. In first step, model is built using historical data. In second step, the model is used to classify future test data for which the class levels are unknown.

Decision tree one of the most popular model, because it is simple and easy to understand. It is flow-charting approximating tree arrangement, where each internal node denotes investigation on an attribute, every branch represents a conclusion of the test, and leaf node represents class. For building a model that would categorize the students into the two classes, depending on the historical data.

Feature Selection

In this manuscript, employ feature selection for supervised machine learning tasks on the basis of correlation between features. It contains a good feature subset that is highly correlated with class otherwise it is irrelevant. Correlation measures the strength of linear

organization involving two variables. The range of correlation is -1.0 and +1.0. If the correlation is positive, the relation is positive. If it is negative, the relation is negative. Feature selection has two approaches forward selection and backward selection. It selects a subset of input variables by eliminating irrelevant features. In this paper, Correlation-Based Feature Selection (CFS) is use to find the feature subsets that are highly correlated with the class but minimal correlation between features combined with search strategy best-first search (BFS). CFS measures correlations between nominal features, so numeric features are first discretized. BFS starts with empty set of features and generate all possible single feature expansions. The separation with uppermost assessment is preferred and prolonged in the identical approach by adding single features.

ID3

ID3 (Iterative Dichotomizer 3) algorithm is used for building the decision tree using information theory invented. It builds the decision tree from top down, with no backtracking. Information Gain is used to decide on the greatest characteristic for classification.

Entropy

It is a compute of uncertainty concerning a source of message. It ranges from 0 to 1. When entropy is 1 means dataset is homogenous. Entropy is calculated by prescription:

$$Entropy(s) = \sum_{i=1}^{c} -P_i Log_2 P_i$$

ID3 Algorithm

Step 1: compute classification entropy.

Step 2: for each attribute, calculate information gain using classification attribute.

Step 3: select attribute with highest information gain.

Step 4: remove node attribute, for future calculation.

Step 5: repeat steps 2-4 until all attribute have been used. Function ID3 (Input attribute, Output attribute, Training data)

{

If (Training data is empty)

{

Return a single node with Failure;

}

If (all records in training data have positive value)

{

Return a single node with level positive.

}

If (all records in training data have negative value)

{

the single node with level negative;

}

If (input attribute is empty)

{

Return a single node with the value of the most frequent value;

}

Otherwise

{

Compute information gain for each attribute; Split the attribute with highest information gain value; Return a tree with root node X and arcs X1, X2…, Xm; Recursively call the ID3 function until all attribute have been used.

Bayesian Classification

The Naïve Bayes Classifier practice is principally suitable when the dimensionality of the inputs is high. Despite its simplicity, Naive Bayes can repeatedly outperform further sophisticated classification methods. Naïve Bayes representation identifies the characteristics of dropout students. It shows the probability of equally input attribute for the predictable state.

A Naive Bayesian classifier is a simple probabilistic classifier based on communicate Bayesian theorem (from Bayesian statistics) with strong (naive) independence assumptions. By the employ of Bayesian theorem we preserve write

$$P(Ci \mid X) = \frac{P(X \mid Ci)P(Ci)}{P(X)}$$

Naïve Bayes Algorithm

The student performance was predicted expend data mining technique named classification regulations. The NB classification algorithm was used by administrator to predict student performance in opportunity semester based on previous semester consequence and behaviour.

A Naïve Bayes classifier was easy probabilistic classifier found on recounting Bayes theorem by naive impartiality assumptions. Naïve Bayes classifiers were trained tremendously expeditiously in supervised education position. It was effortless to appreciate, necessary training statistics to parameters approximation Unresponsive to unconnected features, handled real and distinct records well.

Naive Bayesian Classification Algorithm

Let D be a training deposit of tuples and their related class labels. As customary, each tuple is symbolize by an n-dimensional attribute vector, X=(x1, x2,…, xn), depicting n measurements absolute on the tuple from n attributes, equally, A1, A2,….., An.

Assume that there are m classes, C1, C2,,…,Cm. Given a tuple, X, the classifier will forecast that X belongs to the class having the highest posterior probability, trained on X. That is, the naïve Bayesian classifier predicts that tuple x belongs to the class Ci if and only if.

P (Ci|X) > P (Cj|X) for $1 \leq j \leq m, j \neq i$

Thus we capitalize on P(Ci|X). The class Ci for which P(Ci|X) is maximized is called the maximum posteriori hypothesis. By Bayes' theorem.

$$P(Ci \mid X) = \frac{P(X \mid Ci)P(Ci)}{P(X)}$$

RESULT AND DISCUSSION

The massive data stored in academic dataset that restricted valuable information for predict performance of students. The classification was used to predict end arrangement of students and as there are frequent approaches as data classification, a naïve bye in linear classifier technique was used here. Using measured and experimental every cluster; it was form stand influential characteristics of every Cluster and assessment concerning every clusters as displayed in Table 1 which was naïve byes classifier predicted in percentage of cluster as C1 96.8%, C2 93.9% and C3 98.8% that was most excellent cluster.

Table: Frequency distribution of students in dropout.

	Frequency	Percent	Cumulative Percent
No	183	83.2	83.2
Yes	37	16.8	100.0
Total	220	100.0	

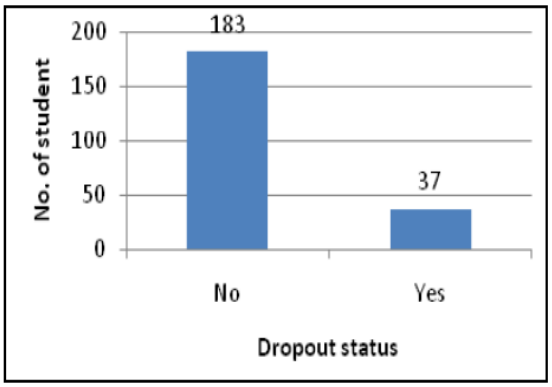
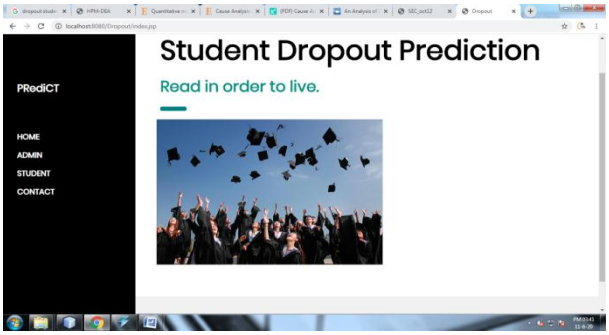
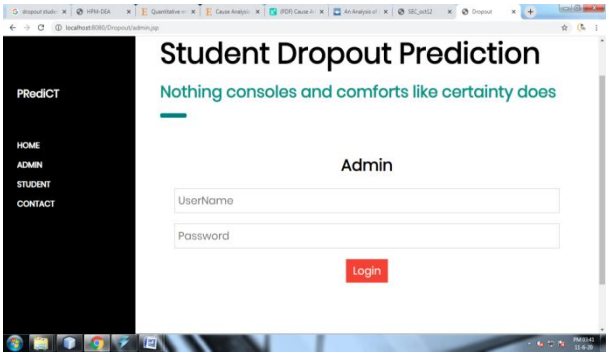


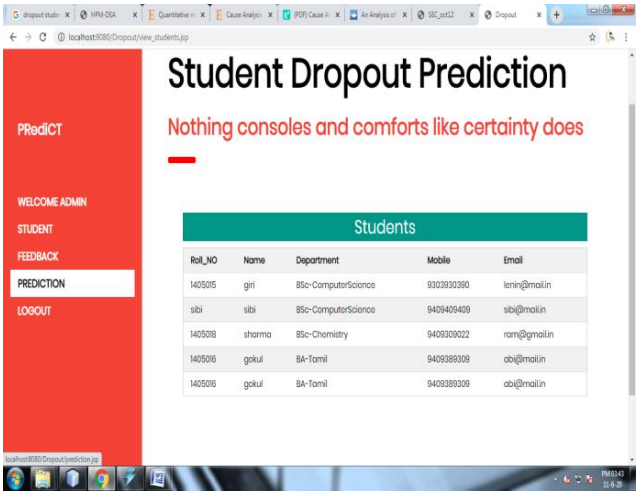
Figure 3: Frequency Distribution of students in dropout.

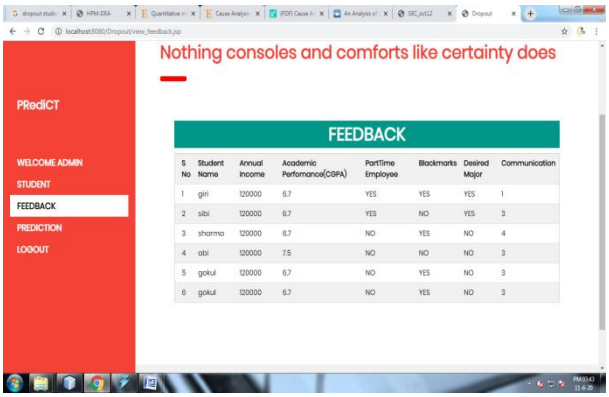
The data collected from 220 students was analyzed to study the frequency distribution against each factor of those students who have completely decided to drop out during the course of study programme. The dropout variable has two possible values such as Yes (students who have completely decided to dropout), and No (students not interested to dropout) and based on these two groups.

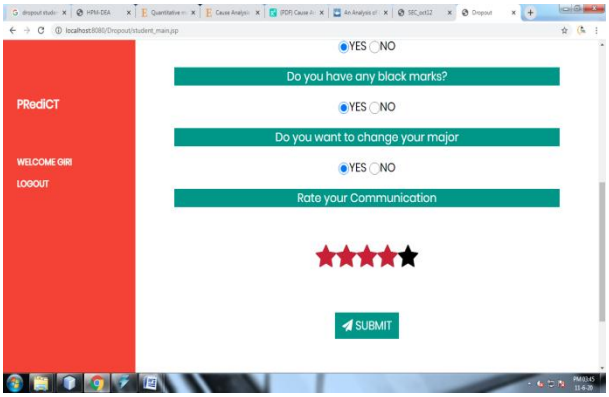
OUTPUT RESULT




Student Dropout Prediction
Nothing consoles and comforts like certainty does
PRediCT
HOME
ADMIN
STUDENT
CONTACT
Admin
UserName
Password
Login


Student Dropout Prediction
Nothing consoles and comforts like certainty does
PRediCT
WELCOME ADMIN
STUDENT
FEEDBACK
PREDICTION
LOGOUT
Students
Roll_NO Name Department Mobile Email
1405015 giri BSc-ComputerScience 9303930390 lenin@mail.in
sibi sibi BSc-ComputerScience 9409409409 sibi@mail.in
1405018 sharma BSc-Chemistry 9409309022 ram@gmail.in
1405016 gokul BA-Tamil 9409389309 abi@mail.in
1405016 gokul BA-Tamil 9409389309 abi@mail.in


Nothing consoles and comforts like certainty does
PRediCT
WELCOME ADMIN
STUDENT
FEEDBACK
PREDICTION
LOGOUT
FEEDBACK
S No Student Name Annual Income Academic Performance(CGPA) PartTime Employee Blackmarks Desired Major Communication
1 giri 120000 6.7 YES YES YES 1
2 sibi 120000 6.7 YES NO YES 3
3 sharma 120000 6.7 NO YES NO 4
4 abi 120000 7.5 NO NO NO 3
5 gokul 120000 6.7 NO YES NO 3
6 gokul 120000 6.7 NO YES NO 3


YES NO
Do you have any black marks?
YES NO
Do you want to change your major
YES NO
Rate your Communication
PRediCT
WELCOME GIRI
LOGOUT
SUBMIT

CONCLUSION

This proposed system has explored the application of the naive bays methods in higher education, where they are not usually applied. It is found out that this area abounds in unused data that, unfortunately, are not stored in an appropriate way. The investigation has detached the origin of students' dropout (e.g. family income, CGPA, etc). It has also resolute the characteristic outline of the student disposed to drop out at the Faculty of Economics in Split. The obtained data should, in the earliest stage, be used to raise awareness on the possibilities and need to use the data mining models and methods at the institution in which this research has been carried out. The planned structure of data warehouse will permit maintain in deliberate decisions and observe of the dropout trend. All this will also support the process, which also aims at enhancing the efficiency of studying.

REFERENCE

1. Al-Hawaj, A. Y., Elali, W., Twizell, E. H. (Ed.) Higher Education in the Twenty-First Century: Issues and Challenges, Taylor & Francis Group, London, 2008.
2. Deem, R., Hillyard, S., Reed, M. Knowledge, Higher Education, and the New Managerialism: The Changing Management of UK Universities, Oxford University Press Inc., New York, 2007.
3. K Akilashri, GS Kumar Efficient Routing Solutions For Avoiding Network Traffic Using Fine Grained Tream Based Measurement Scheme, 2001.
4. Gamberger, D., Šmuc, T. Poslužitelj za analizu podataka [http://dms.irb.hr]. Zagreb, Hrvatska: Institut Rudjer Bošković, Laboratorij za informacijske sustave, 2001.
5. Garača, Ž. ERP sustavi, Ekonomski fakultet, Sveučilište u Splitu, 2009.
6. GFME The Global Management Education Landscape: Shaping the future of business schools, Global Foundation for Management Education, 2008.
7. Hair, J., Anderson, R., Babin, B. Multivariate Data Analysis, Prentice Hall, 2009.
8. Halmi A. Multivarijantna analiza u društvenim znanostima, Alinea, Zagreb, 2003.
9. Klepac, G., Mršić, L. Poslovna inteligencija kroz poslovne slučajeve, Lider press, 2006.
10. Knust, M., Hanft, A. (Ed.) Continuing Higher Education and Lifelong Learning: An International Comparative Study on Structures, Organisation and Provisions, Springer Science & Business Media, Heidelberg, 2009.
11. Matignon, R. Data Mining Using SAS Enterprise Miner TM, John Wiley & Sons, Inc., Hoboken, New Jersey, 2007.

12. McKelvey, M., Holmén, M. (Ed.) Learning to Compete in European Universities: From Social Institution to Knowledge Business, Edward Elgar Publishing, Inc., Massachusetts, 2009.
13. P.S.S Akilashri, E Kirubakaran Analysis of Automatic Crack Detection in Metal, International Journal of Recent Development in Engineering and Technology, www.ijrdet.com (ISSN 2347 - 6435 (Online), 2014; 2(1).
14. P.S.S. Akilashri: Analysis of Security Algorithms used to secure Cloud Environment, Dec-2018; 06(11).
15. P.S.S. Akilashri: Association Rule Mining Classification using J48 & Navie Bayes, Dec-2018; 06(11): 216-220.