

Academic Dashboard—Prediction of Institutional Student Dropout Numbers Using a Naïve Bayesian Algorithm

Aishwarya Suresh, H.S. Sushma Rao and Vinayak Hegde

Abstract Every year, many students enroll themselves on various courses offered by institutions. In that bundle of admissions, a few tend to fall out of their academic programs. Students drop out of their courses due to varied reasons. Analyzing these reasons in order to predict the dropout rate of an institution is of interest. In this research chapter, we are considering a few reasons such as student attendance, educational history, medical history, family background, disciplinary issues, attendance, etc. as factors to compute and predict future dropout rates of registered courses at institutions. To compute and predict dropout rate, a pre-survey and post-survey is conducted. By applying a Naïve Bayesian classifier we predict the probability of students dropping out. Early prediction of student dropout rates, will help to improve the performance of an organization, both professionally and economically.

Keywords Naive Bayesian · Dropout · Analysis · Prediction · Probability

1 Introduction

One of the mottos of every institution is to have a low number of dropouts. When a student joins an institution they have no idea whether they are likely to drop out of their academic programs during their educational journey. When a student drops out of a program or institution, it is normal for their institution to fail to record the reason why. Dropout numbers differ each year due to different reasons. Recording

A. Suresh (✉) · H.S. Sushma Rao · V. Hegde
Department of Computer Science, Amrita Vishwa Vidyapeetham Mysuru Campus,
Amrita University, Mysuru, Karnataka, India
e-mail: aishwarya3939@gmail.com

H.S. Sushma Rao
e-mail: sushmarao.1396@gmail.com

V. Hegde
e-mail: vinayakhegde92@gmail.com

these reasons helps a institution to know where they are lacking and also helps them to understand student mentality. The objective of this chapter is to analyze, and predict, the number of dropouts based on the parameters like attendance, educational history, medical history, family background, subject backlogs, etc. This may help an institution to take corrective measures to improve, and reduce, student dropout rates. Data for this chapter was gathered using survey forms. Two surveys were performed, i.e., a pre-survey and post-survey. The data was pre-processed by converting text into zeros and ones as well as finding and updating missing/null values. With use of the data a Naïve Bayesian classifier was used to predict the output.

The survey forms were based on the following research questions.

1. What makes a student drop out of an organization? Is there any student related character, academic factor, family background which could be linked to the dropping out?
2. How does dropping out affect an organization's growth?
3. Under which circumstances do students feel they need/must drop the course?
4. What impact does academic performance have on dropping out?

The Naïve Bayesian algorithm was used with the help of WEKA to discover and extract results. The rest of the chapter is organized as follows: Sect. 2 describes the literature survey; Sect. 3 describes the methodology used to carry out the research; Sect. 4 looks at research ethics; Sect. 5 describes experimental results; and Sect. 6 concludes the chapter.

2 Related Work

Sweta Rai, proposed a prototype machine learning tool based on classification which automatically recognizes whether a student will continue their study or drop out of their course based on certain factors commonly considered as being responsible for dropouts. The technique used here is based on a decision tree and extraction of hidden knowledge from a large dataset. Considering various factors, a discriminant analysis was used to extract frequent patterns and correlations from the dataset [1]. An association rule for mining was applied to the dataset. The association rule and the decision tree was carried out using the WEKA data-mining tool. The results gathered from these rules supported the fact that 0.68% of dropouts were down to personal problems. The main reason for dropouts was mostly linked to sickness (home sickness), adapting to new courses, and poor hostel facilities, along with simply not adjusting to the campus environment as well as low placement rates.

Abu-oda and El-halees (2015) made use of different data-mining approaches to predicting student dropouts from different courses. The data was collected historically from the first two years of study. To classify and predict the datasets, different

classifiers were used such as decision tree and Naïve Bayes. These methods were tested using 10-fold cross validation [2]. The result of the accuracy of the classifiers was 98.14% and 96.86%, respectively. From the result, it was also observed that one of the reasons for dropping out was relationships between students, which were not outwardly obvious. Yathongchai et al. (n.d.) considered three issues affecting student dropout rates. The factors were grouped thus: conditions related to students before admission, during study periods at university, and all the other factors [3]. Jadri (2010) concentrated on enhancing the efficiency of studying to understand the dropout problem. Statistical data processing was performed with data-mining methods. The first segment presents basic information on the structure of the student. The second segment presents an analysis using logistic regression, decision trees, and neural networks [4]. Models were identified according to SEMMA and were compared in order to select best at predicting. The attributes and parameters were listed and selected for on the basis of the analysis needed. The missing value replacement by most frequent value of neural network. Bayer et al. (n.d.) focused on school failure regarding dropouts due to social behavior and student records. This novel method for student failure prediction reduced the number of incorrectly classified dropouts. The classifier created, using only social behavior, had a data accuracy which did not exceed 69% [5]. By adding attributes describing social behavior this increased by 11%. The highest accuracy was obtained by PART, True Positive (TP). The best results were obtained using the decision tree learner, 82.53% and TP, 78.50%. This supports the hypothesis that four semesters represents a period model which can predict dropping out with high probability.

3 Methodology

The objective of this chapter is to find dropout rates of students from institutions. A predictive student dropout model may help institutions recognize/anticipate early dropouts. Two survey were conducted, i.e., one during student admission (pre-survey) and the other during the middle term of the course (post-survey), in order to collect data and predict which students are likely to drop out. Consideration was given to parameters like parental income, student medical history, student educational history, legal issues, and so on. The data gathered from the surveys were pre-processed by converting texts to zeros and ones. Using these sets of zeros and ones, an algorithm was used to compute and results. According to past results the Naïve Bayesian algorithm forms a better model than any other.

The Naïve Bayesian classifier provides a method for computing the probability using independent assumptions of predictors. The Bayesian classifier helps predict values from datasets.

3.1 Survey Form During Student Admission

At the time of student admission, a survey form was given to students. This form helped collect data, such as, student educational history, family background, parents' educational background, parents' economic background, student health, student disciplinary issues, previous failure in courses, etc. These were collected and saved in a database.

3.2 Survey Form for the Middle of Term

During the middle term of the course, another survey was issued, based on whether the student was adjusting to their course? Is the student settled their peers? Are they facing any difficulties from the institution or fellow students? Are they facing any difficulties in their hostel? Are they home sick? What is their attendance status? These and many more questions were posed on the survey form.

The data (categorical) from the survey was stored in the form of zeros and one in a database. Later, the date was retrieved and computed using the Naïve Bayesian algorithm in order to predict which student was likely drop out.

3.3 Work Methodology

See (Fig. 1).

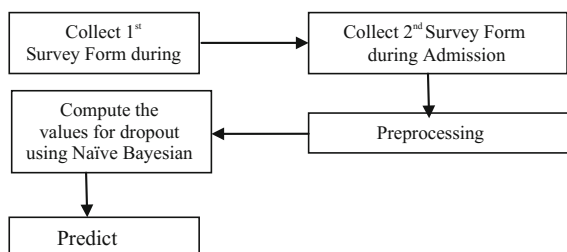


Fig. 1 Methodology undertaken during research. The data is collected via survey forms. Later, the data is pre-processed and a Naïve Bayesian classification is used to compute and predict the drop out values

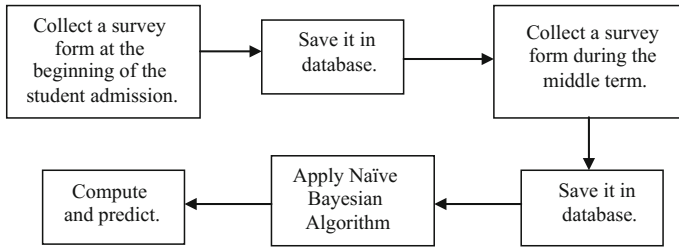


Fig. 2 A flow chart for the research

3.4 Predictions of Dropping Out

See (Fig. 2).

4 Research Ethics

Before conducting the survey, permission was requested from the head of the institution. Student drop out data was gathered through observation and interaction. Consent given by the institution was relayed to the student. The students understood they were participating in research, the confidentiality of research, what information was to be used, and the possible outcomes of the research. All relevant information was explained to the participants.

5 Experimental Results

5.1 Data Pre-processing

5.1.1 Converting Text to Zeros and Ones

The data gathered from the pre-survey and post-survey was in the form of text. This texts was converted into zeros and ones and stored in a database.

5.1.2 Finding Null Values and Separating Them

In the dataset collected, using this method, we separated the null values from the dataset.

5.1.3 Updating Missing Values

After separation of the datasets, we updated the null values using values relevant to other corresponding values.

5.2 Classifier Used

5.2.1 The Algorithm

The Naive Bayesian classifier uses the concept of independent assumptions existing between predictors, based on the Bayes' Theorem. A Naive Bayesian classifier uses the fact that the value of a predictor (x) on a given class (c) is independent of the values of other predictors. This assumption is called class conditional independence.

$$P(c/x) = P(x/c)P(c)/P(x) \quad (1)$$

$P(c|x)$ is the posterior probability of a class (target) given predictor (attribute)

$P(c)$ is the prior probability of the class

$P(x|c)$ is the likelihood of the probability of a predictor given class

$P(x)$ is the prior probability of the predictor

5.3 Dataset

The datasets are collected using surveys. Two survey forms were designed, i.e., for post-survey and pre-survey. Pre-survey was completed during the admission of the student. Post-survey was completed during the middle term of the course. The questions were framed in a way to obtain a YES or NO answer, which will then be pre-processed into zeros and ones. The frequency count of these is considered and, based on the survey's questions, categorization is undertaken, which helps understand which category is the main reason for a student dropping out. The summarized data is categorized thus: lack of attendance, number of failing subjects, satisfactory rating of hostel facilities, involvement in disciplinary issues, financial status, confidence in English speaking and writing.

5.4 Evaluation and Measurements

See (Tables 1, 2).

Table 1 Dropout predictions and probability distribution

— Prediction on test split —				
Institution	Actual	Predicted	Error	Probability distribution
1	1:No	1:No	*0.834	0.166
2	2:Yes	2:Yes	0.028	*0.972
3	1:No	1:No	0.99	0.01
4	2:Yes	2:Yes	0.129	*0.871
5	1:No	2:Yes	0.338	*0.662
6	1:No	1:No	*0.608	0.392
7	1:No	1:No	*0.864	0.136
8	2:Yes	1:No	*0.781	0.219
9	2:Yes	1:No	*0.563	0.437
10	2:Yes	2:Yes	0.223	*0.777

Table 2 Detailed accuracy of the dataset

== Detailed accuracy analysis by class ==							
TP rate		FP rate	Precision	Recall	F-measure	ROC area	Class
	0.902	0.128	0.881	0.902	0.892	0.964	No
	0.872	0.098	0.895	0.872	0.883	0.964	Yes
Weighted average	0.888	0.113	0.888	0.888	0.887	0.964	

==== Confusion matrix ====

a b < - classified as

37 4 | a = No

5 34 | b = Yes

==== Evaluation of test set ====

==== Summary ==

Correctly classified instances	71	88.75%
Incorrectly classified instances	9	11.25%
Kappa statistic	0.7747; agreement of prediction with true class	
Mean absolute error	0.19; not squared before averaging	
Root mean squared error	0.288; squared before averaging, so large errors have more influence	
Relative absolute error	38.0192%; relative values are ratios, and have no units	
Root relative squared error	57.6279%; total number of instances equals 80	

Table 3 Summary of student responses to survey questions

No	Label	Count
1	No	39
2	Yes	37

5.5 Results and Discussions

See (Table 3).

The total data collected was 76 of which 39 counts were NO and 37 counts were YES. We came to the conclusion that students who said No represented a negative indication of satisfaction towards academia. We observed through graphical representation that all students who had low attendance, a higher number of course failures, and had no chance of attending an exam because of a lack of attendance would go on to drop their course between terms.

5.6 Visualization

See (Figs. 3, 4, 5).

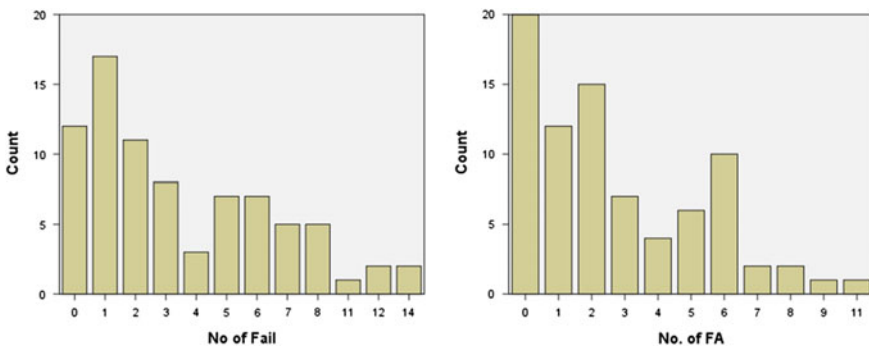


Fig. 3 Graphical representation of both academically failed students and lack of student attendance

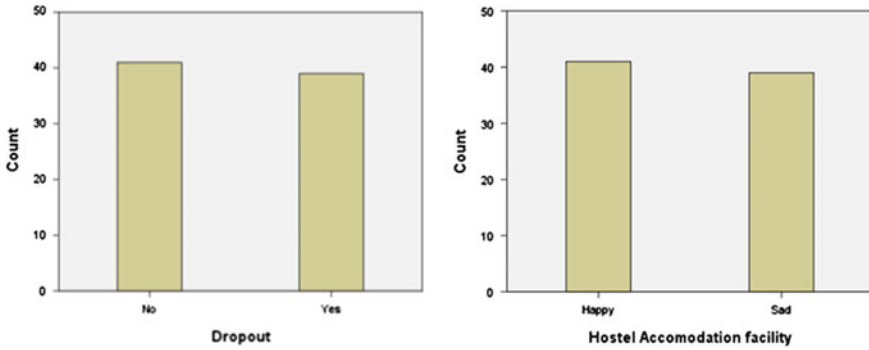


Fig. 4 Representation of willingness to take a decision to drop a course and/or continue a course

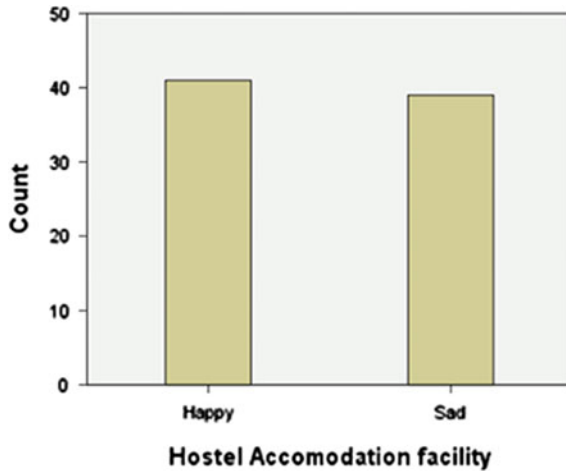


Fig. 5 Representation of individual opinions about residence in student hostels

6 Conclusion

The research was undertaken not only to identify potential dropouts, but also to help institutions understand the reasons for students dropping out. It helps institutions put in place corrective measures to reduce the dropouts. This chapter serves as a platform for predicting dropouts using a Naïve Bayesian classifier. Institutions can therefore have a clear idea of students who are dropping out, who are likely to get dropped, along with the reasons for this. In the future, prediction of the rates of students dropping out can be made.

References

1. Rai S (n.d.) Student's dropout risk assessment in undergraduate course at residential university, pp 1–69
2. Abu-oda GS, El-halees AM (2015) Data mining in higher education: university student dropout case study, vol 5(1), pp 15–27
3. Yathongchai W, Yathongchai C, Kerdprasop K (n.d.) Factor analysis with data mining technique in higher educational student drop out, pp 111–116
4. Jadri M (2010) Student dropout analysis with application of data mining methods, pp 31–46
5. Bayer J, Byd' H, Jan G (n.d.) Predicting drop-out from social behavior of students, (Dm)
6. Abu-Oda GS, El-Halees AM (2015) Data mining in production management and manufacturing. *Int J Data Mining Knowl Manage Process (IJDKP)* 5(1):97–106
7. Muzenda A (2014) Students perceptions on determinants of dropouts from colleges, vol 5, no 1, pp 114–118