STUDENT PERFORMANCE PREDICTION USING CLASSIFICATION DATA MINING TECHNIQUES

[1]Vinaya Patil, [2]Shiwani Suryawanshi, [3]Mayur Saner, [4]Viplav Patil, [5]Bhushan Sarode

Students
Department of Computer Engineering
SSBT COET Bambhori, Jalgaon

Abstract: **Students opting engineering as their disciple is increasing rapidly. But due to various factors and inappropriate primary education in India dropout rates are high. Students are unable to excel in core engineering subjects which are complex and mathematical, hence mostly get drop / keep term (kt) in that subject. With the help of data mining techniques we can predict the performance of students in terms of grades and dropout for a subject can be predicted. In the proposed system, Naïve Bayes algorithm is used. Based on the rules obtained from the developed technique, the system can derive the key factors influencing student performance.**

Keywords: **dropout, prediction, classification, data mining, education.**

- **Introduction**

Data mining has been used in the areas of Science and Engineering, such as Education ,Genetics ,Medicine , Bioinformatics and electrical power engineering.Data mining techniques and tools are used to extract meaning from large set of data generated to peoples learning activities. It has been widely used in the areas of Business to analyze the Customer Relation Management, Human Resource management, marketing etc., Data Mining has high impact in the Business sector, Education is also tapping into the power of Data Mining.

Data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information. Information that can be used to increase revenue, cuts Costs, or both .It can be classified as Supervised and Unsupervised learning. In the supervised learning classification requires the training data has to specify what we are trying to learn (the classes) and where as in unsupervised learning the training data doesn't specify what we are trying to learn (the clusters). Supervised learning is analogous to human learning from past experiences to gain new knowledge in order to improve our ability to real world tasks . Various algorithms are used to perform supervised learning and few among them are Symbolic Machine Learning algorithm, Semi symbolic machine learning algorithm, Nearest Neighbor Algorithm, Naive Bayes algorithm.

The Naive Bayes algorithm is a simple probabilistic classifier which is based on Bayes theorem with strong and naive Independence assumptions. It is one of the most basic classification techniques with various applications in email spam detection, personal email sorting, document categorization, sexually explicit content detection, language detection and sentiment detection. Despite the naive design and oversimplified assumptions that this technique uses, Naive Bayes performs well in many complex real-world problems. Naive Bayes algorithm is highly scalable and requires a number of parameters linear in the number of variables. A Naive Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem (from Bayesian statistics) with strong (naive) independence assumptions.

- **Related work**

M. Wook, Y. Hani Yamaya, N. Wahab, M. Rizal Mohd Isa, N. Fatimah Awang and H. Yann Seong compared two data mining techniques which are: Artificial Neural Network and the combination of clustering and decision tree classification techniques for predicting and classifying student's academic performance. As a result, the technique that provides accurate prediction and classification was chosen as the best model.

Using this model, the pattern that influences the student's academic performance was identified.S. Kumar Yadav, B. Bharadwaj and S. Pal obtained the university students data such as attendance, class test, seminar and assignment marks from the students' database, to predict the performance at the end of the semester using three algorithms ID3, C4.5 and CART and shoes that CART is the best algorithm for classification of data.

Thai Nghe, P. Janecek and P. Haddawy compared the accuracy of decision tree and Bayesian network algorithms for predicting the academic performance of undergraduate and postgraduate students at two very different academic institutes. These predictions are most useful for identifying and assisting failing students, and better determine scholarships. As a result, the decision treeclassifier provides better accuracy in comparison with the Bayesian network classifier.

M. Alam and S. A. Alam have presented a novel algorithm implementing decision trees to maximize the profit based objective function under resource constraints. More specifically, they take any decision tree as input, and mine the best actions to be chosen in order to maximize the expected net profit of all the customers.NBTree -The Naive Bayesian tree learner, NBTree (Kohavi 1996), combined Naive Bayesian classification and decision tree learning.

- **Methodology**

In this research proposed data mining technique is for predicting student's academic performance by analyzing student's feedback using Naïve Bayes algorithm. The research process includes the following process.

A. Data Selection
B. Data Transformation
C. Implementation of Naive Bayes algorithm
D. Classification

- Data Selection:

The experiment was carried out using the standard data sets of students using only 8 high impact attributes (ssc ,hsc ,cet ,traveltime ,attendance .assignment ,unit testpercent ,university resultpercent). The 9th attribute in represents the class that is to be predicted by the algorithm.

Attribute	Description	Possible ESTIMATED Values
ssc	MARKS OBTAINED IN SECONDARY	ssc>=45 && ssc<=60 = low ssc>60 && ssc<75=MEDIUM ssc>=75 && ssc<=100=high .
hsc	MARKS OBTAINED IN HIGHER SECONDARY	hsc>=45 && hsc<=60 = low hsc>60 && hsc<75=MEDIUM hsc>=75 && hsc<=100=high .
cet	MARKS OBTAINED IN ENGINEERING PRE ENTRANCE	ssc>=45 && ssc<=60 = low ssc>60 && ssc<75=MEDIUM ssc>=75 && ssc<=100=high .
traveltime	TRAVEL TIME REQUIRED FOR STUDENT TO ATTEND COLLEGE FROM HIS ACCOMODATION PLACE.	traveltime>=0 &&traveltime<=1=low traveltime>=1 && traveltime<=3=medium traveltime>=3 && traveltime<=4 =high
attendance	ATTENDANCE IN ENGINEERING CLASSES IN PERCENTAGE.	attendance>=15 && attendance<=40 = low attendance>40 && attendance<75=MEDIUM attendance>=75 && attendance<=100=high
Assignment	ASSIGNMENT WORK MARKS ACHIEVED IN PERCENTAGE.	ASSIGNMENTPERCENT>=45 && ASSIGNMENTPERCENT<=60 = low ASSIGNMENTPERCENT>60 && ASSIGNMENTPERCENT<75=MEDIUM ASSIGNMENTPERCENT>=75 && ASSIGNMENTPERCENT<=100=high
Unit Testpercent	INTERNAL UNIT COLLEGE LEVEL ACADEMIC TEAST PERFORMANCE PERCENTAGE .	UNIT-TESTPERCENT>=45 && UNIT-TESTPERCENT<=60 = low UNIT-TESTPERCENT>60 && UNIT-TESTPERCENT<75=MEDIUM UNIT-TESTPERCENT>=75 && UNIT-TESTPERCENT<=100=high
University Resultpercent	END SEMISTER EXAM RESULT PERCENTAGE	UNIVERSITY_RESULTPERCENT>=45 && UNIVERSITY_RESULTPERCENT<=60 = low UNIVERSITY_RESULTPERCENT>60 && UNIVERSITY_RESULTPERCENT<75=MEDIUM UNIVERSITY_RESULTPERCENT>=75 && UNIVERSITY_RESULTPERCENT<=100=high

- Implementation of Naive Bayes algorithm :

The Naive Bayesian algorithm is based on Bayes theorem with independence assumptions between predictors. A Naive Bayesian model is easy to build, with no complicated iterative parameter estimation which makes it particularly useful for very large datasets. Despite its simplicity, the naive Bayesian classifier often does surprisingly well and is widely used because it often outperforms more sophisticated classification methods.Bayes theorem provides a way of calculating the posterior probability(c|x), from P(c),P(x), and P(x|c). Naive Bayes classifier assumes that the effect of 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.

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

where,

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

- Classification :

Classification rule is generated based on the classification process based on users request or research needs. This can be derived specially for the needs on better understanding for each class of data in a database.

MAP: Maximum A Posterior rule generation for feedback prediction.

Assign x to c* if

P(C=c* | X=x)>P(C=c|X=x) c≠c* ,c=c1,..,cl (2)

Generative classification with the MAP rule
Apply Bayesian rule to convert them into posterior probabilities from (1) and (2).

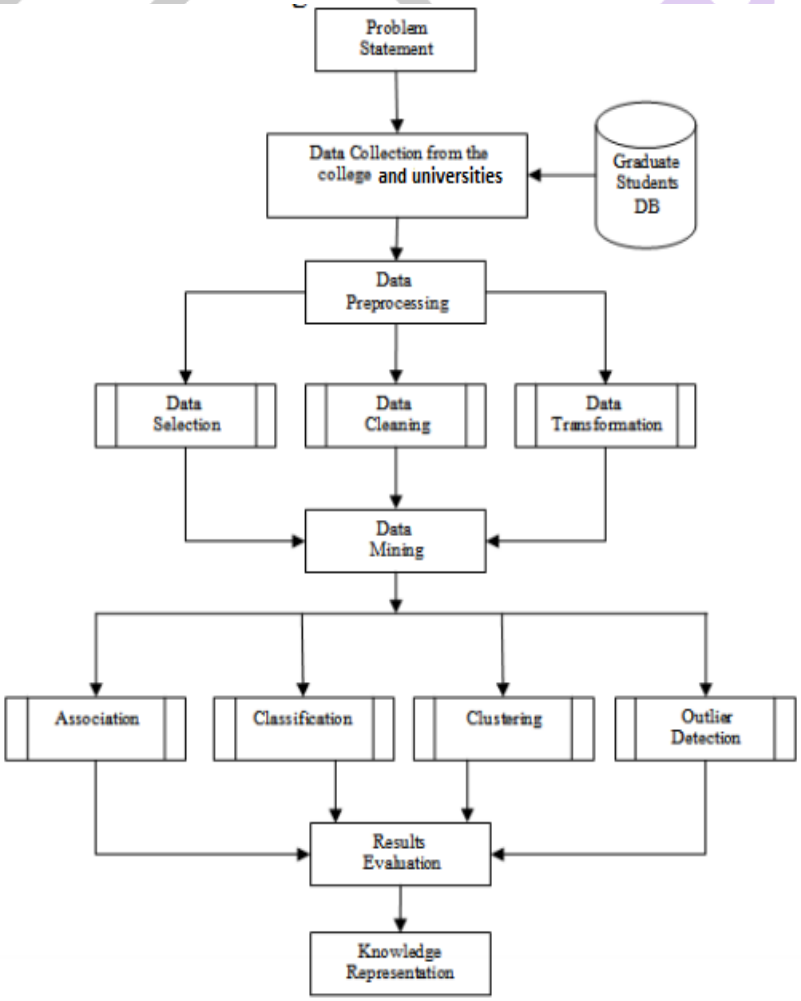
$$P(C = c_i \mid \mathbf{X} = \mathbf{x}) = \frac{P(\mathbf{X} = \mathbf{x} \mid C = c_i)P(C = c_i)}{P(\mathbf{X} = \mathbf{x})}$$
$$\propto P(\mathbf{X} = \mathbf{x} \mid C = c_i)P(C = c_i)$$
$$\text{for } i = 1,2,\cdots,L$$

The implementation work is based on the collected data which possess various data mining aspects. The Student data is taken into account for the performance prediction. The proposed research work is categorized into two modules. First the feedback results are analyzed and same is compared with the internal test performance.

- **System architecture**

Before applying the data mining techniques on the data set, there should be a methodology that governs our work. Figure depicts the work methodology used which is based on the framework proposed in. The methodology starts from the problem definition, then pre-processing which are discussed in the introduction and the data set and pre-processing sections, then 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 of our case study, for each of the four data mining tasks; Association, classification, clustering and outlier detection, and how benefited from the discovered knowledge can.

- **Implementation**



For Implementation we had taken standard datasets from Gulabrao Deokar College of Engineering Jalgaon which contains almost 300 tuples. The dataset is taken of students of 2009 batch. We calculated travel time attribute constraint from Google map distance calculator from the address of the specified candidate. The inclusion of travel time constraint affected a lot on Student performance.

We had taken MYSQL SERVER as an server for linking our datasets values to our code. On the basis of YES/NO class we predicted whether student will successfully complete his/her engineering or not. After calculation several values of specified cases

From datasets we calculated mean values.

Algorithm	TP	FP	PRECISON	RECALL	FMEASURE
NAÏVE BAYES	0.947	0.474	0.922	0.773	0.761

- **Result and analysis**

There are some parameters on the basis of which we evaluated the performance of the classifiers such as ssc-marks, hsc-marks, travel time, age, Entrance Marks. The Accuracy of a classifier on a given test set is the percentage of test set tuples that are correctly classified by the classifier. The Error Rate or Declassification rate of a classifier M, which is 1 Acc (M), where Acc (M) is the accuracy of M. The Confusion Matrix is a useful tool for analysing how well your classifier can recognize tuples of different classes. The sensitivity and specificity measures can be used to calculate accuracy of classifiers. Sensitivity is also referred to as the true positive rate (the proportion of positive tuples that are correctly identified), while Specificity is the true negative rate (that is, the proportion of negative tuples that are correctly identified). These measures are defined as follows:

$$TPrate = \frac{T_p}{T_p + F_n}$$

- True Positive Rate: It is the proportion of actual positives which are predicted as positive. The formula is defines as, Where, Tp stands for true positive and Fn stands for false negative.

- False Positive Rate: It is the rate of negatives tuples that are incorrectly labelled. The formula is given as follows,

From a total number of standard student's record data set, we had chosen sample 500 students record for our analysis. It

$$FP \ \ rate \ \ of \ \ Class \ \ "YES" = \frac{F_n}{T_p + F_n}$$

$$FP \ \ rate \ \ of \ \ Class \ \ "NO" = \frac{F_p}{T_p + F_p}$$

Algorithm	T_p	F_p	Precision
Naive Bayes	0.947	0.474	0.922

demonstrates number of Eligible and Non eligible candidates for engineering faculty. Numbers of Eligible candidates are 166 and Number of Non eligible candidates are 34.

- F-measure: This is a measure that combines precision and recall, a harmonic mean of precision and recall, is known as the traditional F-measures.

$$Fmeasure = 2 * \frac{(Precission * Recall)}{(Precession) + (Recall)}$$

- Recall: Recall in information retrieval is the fraction of the documents that are relevant to the query and that are successfully retrieved. The formula for recall is given as below,

$$Recall = \frac{(T-Pos)}{(T-Pos) + (F-neg)}$$

- **Conclusion and Future work**

Data mining techniques allow a high level extraction of knowledge from raw data, offering interesting possibilities for the education domain. In this study a model was developed based on some selected input variables collected through questionnaire method. After testing some hypothesis, some of most incensing factors were identified and taken to predict the grades. Data mining techniques are applied to predict the performance of the students and found that Naive Bayes algorithm is best suited to predict the grades. We designed a tool using JAVA framework to predict whether the student will be able to complete engineering successfully or not. Our Code achieved an accuracy of 49.5 percent which shows the half potential efficiency of Naive Bayes algorithm.

The obtained results from hypothesis testing reveals that type of College is not influence student performance and on the other hand travel time; Academic Marks plays a major role in predicting Possibility. As a result, having the information generated through our experiment, institution would be able to identify students provide better additional training. Therefore, it seems to us that data mining has a lot of potential for education. Furthermore, we intent to enlarge the experiments to collect additional features like psychological factors which disturb the students, motivational efforts taken by the teachers and e-learning materials available to the students.

References

[1] M. Al-Razgan, A. S. Al-Khalifa, and H. S. Al-Khalifa, "Educational data mining: A systematic review of the published literature 20062013," in Proc. the 1st International Conference on Advanced Data and Information Engineering, 2013, pp. 711-719.

[2] Weka 3: http://www.cs.waikato.ac.nz/ml/weka/

[3] Sonali Agarwal, G. N. Pandey, and M. D. Tiwari, "Data Mining in Education: Data Classification and Decision Tree Approach", in the Journal of e-Education, e-Business, e-Management and e-Learning, Vol. 2, No. 2, April 2012, pp. 140-144.

[4] Md.Sarwar kamal, Linkon Chowdhury, Sonia Farhana Nimmy, "New Dropout Prediction for Intelligent System", in the International Journal of Computer Applications (0975 – 8887) Volume 42– No.16, March 2012, pp. 26-31.

[5] Carlos Márquez-Vera, Cristóbal Romero Morales, and Sebastián Ventura Soto, "Predicting School Failure and Dropout by Using Data Mining Techniques" in the IEEE Journal Of Latin-American Learning Technologies, Vol. 8, No. 1, February 2013, pp. 7-14.