Tracking and Predicting Student Performance in Degree Programs Using Machine Learning Approach

[1]Kailash R. Yadav,[2]Omansh K. Singh,[3]Ajay U. Choudhary, [4]Prof. Manisha Singh
[123]Student BE Computer, [4]Assistant Professor,
[1]Computer Engineering Department,
[1]Dhole Patil College of Engineering, Pune, India

Abstract : Machine Learning research fields are now fully emerged as interesting areas for research and development, which are now exposing useful knowledge from large datasets for many purposes like predicting student's performance in case of Educational Data Mining and Learning Analytics. It can be beneficial for taking rapid actions in today's educational systems. Existing techniques have used attributes which are mostly related to academic performance, family background, emotional and social influences; while attributes regarding family expenses and student's personal information are usually ignored. In this paper, Machine Learning techniques, scikit-learn library and its algorithms are applied to predict that a student will be able to perform good in his graduation or not. Predicting performance of a student accurately based on their ongoing academic records is very important for effectively carrying out necessary interventions to ensure students on-time and satisfactory graduation.

Predicting student's performance in completing degrees (e.g. college programs) is much less studied and faces new challenges: (1) Students coming from different backgrounds and have selected different courses;(2) Since the courses are not equal it is not possible to make accurate predictions; (3) The evolving progress of student needs to be incorporated into the prediction. In this paper, we develop a new machine learning technique for predicting student performance in degree programs which is able to address the above key challenges. The proposed methodology has two major features. First, a bilayer structure comprising of multiple base predictor attributes and a cascade of predictors which is developed for making predictions based on students evolving performance states. On the other hand, a data-driven approach based on latent factor models and the probabilistic matrix factorization which aims to discover the course relevance and which is also important for constructing the efficient base predictors.

***IndexTerms* – Prediction, Performance, Data Mining, Learning Analytics.**

I. Introduction

Since it's apparent that higher education has a major and direct role to play in the nation's economy and its progress. So, there are lots of institutions of higher education are available across the country. Basically, the quality of education is judged by the factor that how much the student is capable to do the task in reliable and efficient manner.

The student's performance which predicted by the system will be helpful in identifying the students who are at risk of failure and thus management can give timely help and will be helpful in taking essential steps to coach the students to improve his/her performance. Classification techniques have been applied to predict the academic performance of the students based on their socio-economic condition and previous academic performances. This paper explores the link between emotional skills of the students along with socio economic and previous academic performance parameters in order to predict academic performance using data mining techniques. The emotional skills like assertion, leadership, stress management, etc. are obtained, using standard Emotional Skill assessment process ESAP. Data mining tasks can be either descriptive or predictive. Descriptive data mining uses technique of association rule mining, clustering etc. to find patterns hidden in large data set which helps in intelligent decision making. Predictive data mining constructs models using rule set, decision tree, neural nets, and support vectors etc. to predict the class of a new data set. The objective of this paper is to predict the third semester performance of graduation students. The reason behind considering only the third semester for prediction is the observation that most of the students drop out of the course after first year and also students normally take a year to get adapted and accustomed in an academic environment. The decision tree CART (Classification and Regression Tree) have been used to build the model and the main contribution of this paper is the model comparison along with finding the impact of various attributes on student's performance.

II. Literature Survey

Using 24 different predictor variables which includes demography, Turkish, scores in maths, religion and ethics, science and technology, level determination exams, etc. for predicting Turkish secondary education result. A paper by Sen, Ucar and Delen [1]. Applications of Artificial NN (Neural Network), SVM (Support Vector Machine), Multiple Regression and Decision showed that the most important predictor variables are determination exam, scholarship, number and success level. Bharadwaj and Pal [2] focuses on Previous Semester marks, class test grade, seminar performance, Assignments done, attendance and practical to predict end semester marks of students. Data set of 50 students from year 2007 to 2010 graduation of Purvanchal University was considered. This paper helps to calculate Split info and gain ratio of each predictor and products prediction rules.

Bidgoli, Koshy, Kortemeyer and Punch [3] have used tree classifiers and non-tree classifiers to predict the grades of students which are enrolled with online education, Latest Learning Online Network with Computer Assisted Personalized Approach (LON – CA PA) which was developed at Michigan State University and they found that with the combination of multiple classifiers we can enhance the accuracy of prediction.

Ramaswami and Bhaskaram have used Chi-squared Automatic Interaction Detector (CHAID) in [5] to classify 12th class students of selected Tamil Nadu schools. With demographic details, student's health, coaching and hours spent on study ath home etc. have been studied. Prediction Accuracy obtained was 44.69.

Kabakchieva in [4] used few data mining classification Techniques on 10330 students, with 14 attributes which includes personal details, secondary educational scores, entrance exam scores and admission year etc. The students are classified into five categories viz. excellent, very good, good, average and bad. The 10 fold cross-validation and percentage splitting was used for all the classifiers like J48, K-nearest and Bayesian.

III. Related Work

The purpose of this technique is to help us classify the graduation student performance. This classifier is built by combining process of Dataset creation, data understanding, data segregation, data preparation, modeling and applying of machine learning algorithms to predict future performance of student.

A. Data Generation:

A data collected of around 250 students from the university with a sample of 250 students was collected having 22 attributes including academics, social and emotional as shown in Table I.

B. Data processing:

The collected data is saved in CSV format file. The data should be clean so by deleting missing values, removing redundant data and having consistency we can achieve a good dataset with accurate values.

C. Data segregation and Modeling:

By using 80-20 strategy we have to divide data as 80% training data and 20% test data.
For modeling we are using scikit-learn an open source library for python. This library comes with NumPy, SciPy, pandas, sklearn and matplotlib which can be used for various machine learning applications.

D. Classification:

For classification we are using Decision tree algorithms like C4.5 and CART (Classification and Regression Trees). The C4.5 algorithm is basically a decision tree algorithm and build decision tree. C4.5 helps converting the trained trees into sets of if-then rules CART supports numerical target variables(regression). As per testing these algorithms gives us better results with smaller data sets and higher number of attributes.

E. Performance Analysis:

For predicting performance the model we have built based on the attributes which are dependent on each other. For Example, if a student's family income is not so good and also he has taken an education loan also previous year's results are not so good then there are more chances of student's performance will be poor. Based on predicted values we display a class of performance grade.

Table I: Attributes Description

Attributes Name	Values	Description
Gender	Male, Female	Gender
FE	Midschool, Inter, Grad, Postgrad	Father's Education
FO	Govtjob, Pvtjob, Business	Father's Occupation
MO	Govtjob, Pvtjob, Business, Housewife	Mothers Occupation
FI	MIG, HIG, LIG, VHIG	Annual Family Income
LOAN	Yes, No	Educational loan taken
EARLYLIFE	Metro, City, Village	15 years of life spent
MEDIUM	English, Other	Medium
TENTH	BLAVG, AVG, ABAVG, EXCL	%marks in 10th
TWELVTH	BLAVG, AVG, ABAVG, EXCL	% marks in 12th
GRAD	BLAVG, AVG, ABAVG, EXCL	% marks in Graduation
GRADDEGTYPE	Regular, Distance	Type of Graduation Degree
GRADDEGSTREAM	CS, NCS	Graduation Degree Stream
GAPYEAR	Yes, No	Gap year in education
ACADEMICHRS	INSUF, SUF, OPTIMAL	Hours spent on academics

ASSERTION	D, S, E	Assertiveness of the student
EMPATHY	D, S, E	Empathy of the student
DECISIONMAKING	D, S, E	Decision making ability of the student
LEADERSHIP	D, S, E	Leadership ability of the student
RIVE	D, S, E	Drive of the student
STRESSMGMT	D, S, E	Stress management skill of the student
HOURS SPENT	Integer Value	Hours spent per day on academics

IV. Proposed methodology and Algorithm

A decision tree is a decision support algorithm that uses a tree-like graph or model of decisions and their possible consequences, including all the chances of event outcomes, resource costs, utility and attributes. It is one way to display an algorithm that only contains conditional control statements and based on decisions predict a particular output class.

4.1Population and Training data

We have selected around 250 student's data in csv format for training our algorithm. Using the split function, we split data for 80% training and 20% test data randomly, the data is then preprocessed for encoding. The data encoding is required for our algorithm to study because the data is string class format.

4.2 Encoding technique

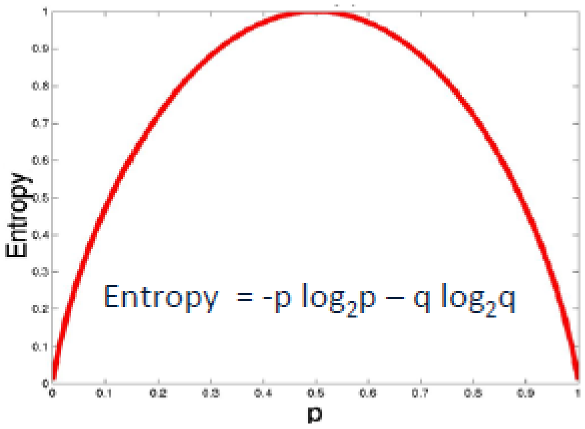
Here we use Label Encoding for converting the labels into numeric form so as to convert it into the machine-readable form. Machine learning algorithms can then decide in a better way on how those labels must be operated. It is an important pre-processing step for the structured dataset in supervised learning. Here we have predefined the encoded labels for decoding when the output is predicted according to class respectively.

4.3 Mathematical Model

The core algorithm for building decision trees called ID3 by J. R. Quinlan which employs a top-down, greedy search through the space of possible branches with no backtracking. ID3 uses Entropy and Information Gain to construct a decision tree. In ZeroR model there is no predictor, in OneR model we try to find the single best predictor but decision tree includes all predictors with the dependence assumptions between predictors. In Decision tree CART algorithm the Gini Index is used to calculate, split the subset and find the prediction.

A. Entropy:

A decision tree is built top-down from a root node and involves partitioning the data into subsets that will contain instances with similar values. In order to define information gain precisely, we begin by defining a measure commonly used in information theory, called entropy that consists of impurity value of an arbitrary collection of examples.



$$H(S) = \sum_{c \in C}^{n} -\mathrm{p(c)} \log_2 \mathrm{p(c)}$$

Where,

S - The Current (data) set for the entropy is being calculated.

C - Set of classes in S . c = {yes|No}

P(c) – The proportion of the number of elements in class c to the number of elements in set S.

B. Information gain:

The information gain is based on the decrease in entropy after a data-set is split on an attribute. Constructing a decision tree is all about finding attribute that returns the highest information gain (i.e., the most homogeneous branches). Information gain is calculated as follows,

$$\text{Entropy (attributes)} = \text{entropy } (f1, f2)$$

- Calculate the entropy of every target.
- The result is the Information Gain, or decrease in entropy.
- Choose attribute with the largest information gain as the decision node, divide the dataset by its branches and repeat the same process on every branch.

C. Gini Index:

The CART algorithm can be used for building both Classification and Regression Decision Trees. The impurity (or purity) measure used in building decision tree in CART is Gini Index. The decision tree built by CART algorithm is always a binary decision tree (each node will have only two child nodes).

Gini index says, if we select two items from a population at random then they must be of same class and probability for this is 1 if population is pure.

$$Gini = \sum_{i\neq j}^{n} \mathrm{p(i)p(j)}$$

Where, i and j are the levels of target variables.

Steps to calculate Gini index:

1. It works with categorical target variable “Success” or “Failure”.
2. It performs only Binary splits.
3. Higher the value of Gini higher the homogeneity.
4. CART (Classification and Regression Tree) uses Gini method to create binary splits.

Steps to Calculate Gini for a split:

1. Calculate Gini for sub-nodes, using formula sum of square of probability for success and failure (p^2+q^2).
2. Calculate Gini for split using weighted Gini score of each node of that split.

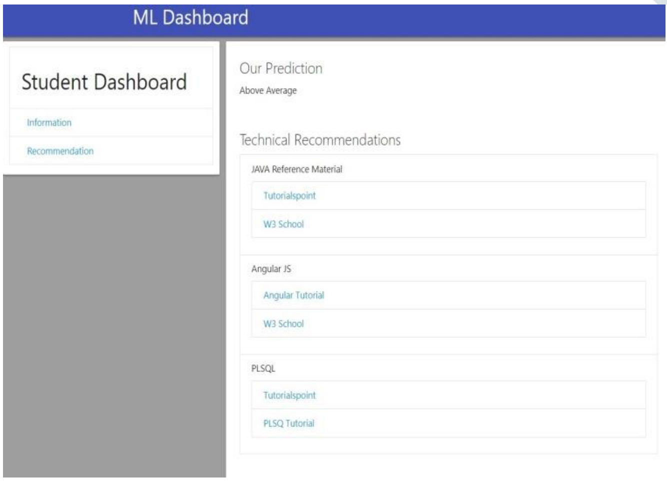
V. Results

Our Project lets the user know his/her performance based on the details and values provided to the system and will give the result in following values→

1. Good
2. Average
3. Above Average.
4. Excellent etc.

In case if the user has more than 65 percent then he will get the result as above average and incase if the user has less than 60 percent then the user will get result as average, similarly the user will get result for all other values.

The user will be able to get his/her predictions based on the academic records like his/her marks and some more attributes like his father occupation or how much time the particular student has put in his/her studies.



Our Prediction

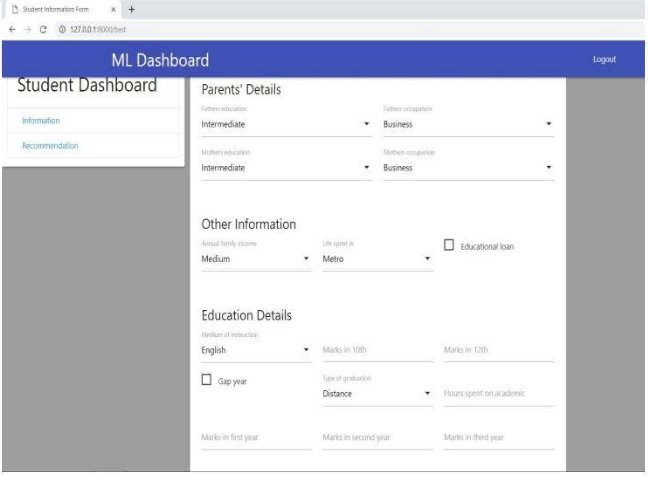
Above Average

Technical Recommendations

JAVA Reference Material
- Tutorialspoint
- W3 School

Angular JS
- Angular Tutorial
- W3 School

PLSQL
- Tutorialspoint
- PLSQ Tutorial



Parents' Details

Fathers education: Intermediate
Fathers occupation: Business
Mothers education: Intermediate
Mothers occupation: Business

Other Information

Annual family income: Medium
Life spent in: Metro
□ Educational loan

Education Details

Medium of instruction: English
Marks in 10th
Marks in 12th
□ Gap year
Type of graduation: Distance
Hours spent on academic
Marks in first year
Marks in second year
Marks in third year

VI. CONCLUSION

In the world of so much rush where everyone is running behind technology, students are not so much focused towards their studies and achieving their career goals, so our projects aims and letting them know where they are lying in terms of their performance based on some academic records and values provided by them.

Effect of social and emotion parameters on placement of students is established. In future work we will be focusing on all other branches of Engineering and also Post graduation student's performance analysis will be added.

REFERENCES

[1] B. Sen, E. Uçar and D. Delen, "Predicting and Analyzing Secondary Education Placement-Test Scores: A Data Mining Approach", Expert System with Application, Volume 39, Issue 10, 2012.

[2] B.K.Bhardwaj and S.Paul , "Mining Educational Data to Analyze Students Performance", International Journal Advanced Computer Science and application Vol. 2 No. 6 , 2011 .

[3] B. M. Bidgoli, D.Koshy, G.Kortemeyer, W.F.Punch, "Predicting Student Performance: An Applicant of Data Mining methods with an educational web based system" , 33rd ASEE/ IEEE .frontiers in Education Conference 20004.

[4] D.Kabakchieva, "Predicting Student Performance by using Data Mining methods for classification." , Cybernetics and Information Technologies, Volume 13, 2013.

[5] M. Ramaswami, and R. Bhaskaran, "A CHAID Based Performance Prediction Model in Educational Data Mining", International Journal of Computer Science, Vol. 7, Issue 1, No. 1.of 2010.

[6] R.S.J.D Baker and K.Yacef, "The State of Educational Data Mining in 2009: A Review and Future Visions" , Journal of Educational Data Mining, 1, Vol 1, No 1, 2009.

[7] M. Wook, Y.H.Yahaya, N. Wahab, M. R.M. Isa, N. F. Awang a International Conference nd H.Y. Seong, "Predicting NDUM Student'sAcademic Performance Using Data Mining Techniques, Paper presented at International Conference of Computer and Electrical Engineering, ICCEE. December 28-30. 2009.

[8] N. S. Shah, "Predicting Factors that Affect Students ' Academic Performance By Using Data Mining," Pakistan Business Review, January 2012.

[9] DeLong,C., P.Radclie, L. Gorny. Recruiting for Retention: Using Data Mining andMachine Learning to Leverage the Admissions Process for Improved Freshman Retention. –In: Proc. of the Nat. Symposium on Student Retention, 2007.

[10] Cortez, P., A. Silva. Using Data Mining to Predict Secondary School Student Performance.EUROSIS. A. Brito and J. Teixeira, Eds. 2008, 5-12.