PREDICTING STUDENTS PERFORMANCE USING PERSONALIZED ANALYTICS

1G.Sujatha
SRMIST, Chennai,Tamil Nadu,India
sujatha.g@ktr.srmuniv.ac.in

2Sindhu.S
SRMIST, Chennai,Tamil Nadu,India
sindusaravan@gmail.com

3Savaridassan.P
SRMIST, Chennai,Tamil Nadu,India
savaridassan.p@ktr.srmuniv.ac.in

ABSTRACT:
This paper focuses on predicting student performance using personalized analytics based on educational databases. Predicting student performance becomes more challenging due to the huge amount of databases. We present two different approaches to improving student achievements. Both approaches are validated on one course which was offered to students of the school of computing between the years of 2012 and 2015. The first approach is based on regression algorithms to predict student performance. Regression is a data mining function that predicts a number. The main goal is to find how well a student can perform in the programming language by predicting grades based on their school background and performance in mathematics and English. In the model build (training) process, a regression algorithm estimates the value of the dependent variable as a function of the predictors in the build data based on the independent variables. These relationships between predictors and target are summarized in a model, which can then be applied to a different dataset in which the target values are unknown. The Second approach is to find the error rate of regression algorithms by using root mean square error. The obtained results reveal that the school background also plays a major role in predicting grades. Finally, we can identify the students who are at risk and provide better additional training for the weak students.

KEYWORDS: Regression ,Prediction, Traindata, Testdata, Analytics.

INTRODUCTION:
Measuring academic performance of students is more challenging since
student’s academic performance depends on diverse factors like personal, socio-economic, psychological and other environmental variables [3,4]. The scope of this paper is to predict the student grades using the best algorithms with high accuracy[16,10]. There is a critical need to develop innovative approaches that ensure students become graduates in a timely fashion and are well trained and workforce ready in their field of study. From the paper [12, 13], The higher secondary education is important in a student’s life because it is one of the main factors that are going to decide the future of the student based on their mark in the higher secondary examination, they are going to get the college education and the field of information technology based on their knowledge in programming. We present methods that draw on techniques from recommender systems to accurately predict students’ programming course grades. Data mining [15] provides many tasks that are used to predict the student’s performance. In this paper, the regression task is used to evaluate the performance of a student and as there are many approaches that are used for data regression, the multi-linear regression and SVM, random forest and regression tree methods was used here. For this study, recent real world data has collected.

Information[refer table1] like the medium of study, syllabus, intermediate background etc., were collected. mathematics marks are also collected to know the logical ability of the student and English marks to know the communication and understanding level of the student. This study is more useful for identifying weak students in the programming at the beginning of the semester and the identified students can be assisted by the educators so that their performance is better in future. This study investigates the accuracy of some regression techniques for predicting performance of a student.

The Main Objectives of this Study:

- Predict the student's success or failure
- Find the best regression algorithm on student data.
- Predict the final grade

The problem of the student final grade prediction in a particular course has recently been addressed using data mining techniques. Researchers usually examine study-related
records, e.g. the age, gender and the field of study [12,13] because of their easy availability in university information systems. The most typical way how to obtain such data is to conduct questionnaires but it tends to have a lower response rate. Therefore, only the data originated from the college are considered for our experiments.

**LITERATURE REVIEW:**

*Prediction in single course :-*

**Title :** Predicting student academic performance in an engineering dynamics course:A comparison of four types of predictive mathematical models  
**Author :** S. Huang, N. Fang  
**Year :** 2013  
**Description:**  
The work presented in this paper does not predict performance at degree level but at a course level. However, it is interesting as it suggests a kind of upper bound for the accuracy that can be achieved when predicting performance at the end of a degree. They employed four mathematical models namely multiple linear regression, multilayer perception networks, radial basis functions and support vector machines to predict student's academic performance in an engineering dynamics course. It has worked on the data of 323 undergraduate students who took dynamics course at Utah State University in four semesters. Their predictor variables were the students’ cumulative GPA; grades earned in four pre-requisite courses i.e. statistics, calculus I, calculus II and physics; and scores on three dynamics midterm examinations. The paper used six combinations of predictor variables to develop a total of 24 predictive mathematical models. For all the four models, they achieved an average prediction accuracy of 81%–91%. This work shows that previous marks can predict the grade of a course with high accuracy.

*Prediction by Using Data Mining Classification :-*

**Title :** Student Performance Prediction by Using Data Mining Classification Algorithms, International Journal of Computer Science and Management Research.  
**Author :** DorinaKabakchieva  
**Year :** 4 November 2012  
**Description:**  
The research is focused on the development of data mining models for predicting student performance, based on their personal, higher secondary and university
performance by using One Rule Learner, Decision Tree, Neural Network and K-Nearest Neighbour classification methods and found that Neural Network model predicts with high accuracy than other three models. The present study differs from other works in various aspects. First, by using the predicting values that effect the programming skills most are only considered, no socio-economic data is considered. It also gives some limit to what can be achieved when predicting graduation performance. Indeed the predictors include midterm examinations that can be expected to correlate well with the final exam of the course, more than marks of single courses with the graduation mark.

**Factors Affecting Student's Performance**

**Title:** Factors Affecting Student Performance : A Case of Private Colleges Bangladesh e-Journal of Sociology

**Author:** Hijazi, S. T., and Naqvi, R.S.M.M

**Year:** 2006.

**Description:**

A study on the student's performance by selecting a sample of 300 students (225 males, 75 females) from a group of colleges under Punjab University in Pakistan. The conclusion that was stated as Student's attitude towards attendance in class, hours spent in study on daily basis after college, students' father's income, mother's age and mother's educational qualification are significantly related to student performance by means of simple linear regression analysis, it was found that the factors like mother's educational qualification and student's father's income where highly interconnected with the student academic performance. The present study differs from other works in various aspects. First, by using the predicting values that effect the programming skills most are only considered, no socio-economic data is considered.

** Analyzing University Data for Determining Student Profiles**

**Title:** Analyzing University Data for Determining Student Profiles and Predicting Performance, Proceedings of the 4th International Conference on Educational Data Mining Scheduling Independent Tasks.

**Author:** D. Kabakchieva , K. Stefanova, V. Kisimov

**Year:** July 6-8, 2011.

**Description:**

The study predicts student's university performance by using student's
personal and pre-university characteristics. They experimented on the data of 10330 students of a Bulgarian educational society, each student being described by 20 attributes. (e.g., gender, date of birth, living area, and country, place and total score from previous classes, current semester scores, total university grades, etc.). In the papers they have applied different data mining algorithms such as the decision tree C4.5, Naive Bayes, Bayesian networks, K-nearest neighbors (KNN) performance at the end of the degree. It is easier to gather marks of students than their socio-economic data. Therefore if a reasonable prediction can be reached without socio-economic data, it makes the implementation of a performance support system in a university easier. If courses can be identified with a major impact on graduation performance, then measures can be taken at the level of those courses, making also the implementation of a performance support system easier. In this study, the performance of a student at the end of the degree will be a class A, B, C, D or E, which represents the interval in which her/his final mark lies. This allows to differentiate between strong and weak students.

**EVALUATION:**

We used a dataset collected from ‘Qubacollege of engineering’. The dataset contains attributes that effect most in the programming marks of a student dataset and it contains one-course instances of 3000 students in a span of 3 years of data from computer science students. The data has been divided as 70% as train data and remaining as test data.

<table>
<thead>
<tr>
<th>Grade</th>
<th>Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 - 49</td>
<td>Fail</td>
</tr>
<tr>
<td>50 - 64</td>
<td>Weak programmer</td>
</tr>
<tr>
<td>65 - 79</td>
<td>Good programmer</td>
</tr>
<tr>
<td>80-100</td>
<td>Excellent programmer</td>
</tr>
</tbody>
</table>

*Table I represents the percentage evaluation of student’s marks*

<table>
<thead>
<tr>
<th>ATTRIBUTE</th>
<th>DESCRIPTION</th>
<th>POSSIBLE VALUES</th>
</tr>
</thead>
<tbody>
<tr>
<td>c</td>
<td>Marks in c programming</td>
<td>Percentage range (0-100)</td>
</tr>
<tr>
<td>M</td>
<td>Marks in maths</td>
<td>Percentage range (0-100)</td>
</tr>
<tr>
<td>E</td>
<td>Marks in English</td>
<td>Percentage range (0-100)</td>
</tr>
<tr>
<td>S</td>
<td>Higher secondary school board of study</td>
<td>(state, ICSE , CBSE)</td>
</tr>
<tr>
<td>l</td>
<td>Medium of study</td>
<td>(local,english)</td>
</tr>
<tr>
<td>bck</td>
<td>Group of study</td>
<td>(biology,maths)</td>
</tr>
<tr>
<td>Cs</td>
<td>He/she has computer programming experience or studied</td>
<td>(yes/no)</td>
</tr>
<tr>
<td>int</td>
<td>Internal marks of c programming</td>
<td>Percentage range (0-50)</td>
</tr>
</tbody>
</table>

*Table II represents the variables after preprocessing the dataset*
Methodology:

Student Characteristics
Regression is the most often used technique for student performance prediction. Researchers usually examined study-related (SR) data [6,7,10]. Our study-related data contained attributes such as higher secondary school background, the medium of study, syllabus, mathematics and English marks. We built a multi linear regression for programming course based on the training set and evaluated the results using various models(SVM, RF, Decision tree). The method [15,16] that achieved best results was subsequently validated on the test set.

Grade prediction
The regression [11,17] is a commonly used technique for student grade prediction. And also SVM Reg., Random Forest, decision Tree are used to validate the models each other. The baseline model predicts the programming marks of the training set of a given instances of attributes. In addition to accurately predicting student’s performance, [7,9,13] the multi-regression model can be used to analyze how the different features contribute to the predicted grades and thus gain some insights about the student’s performance. For a proper analysis of the estimated model parameters, it is more convenient that all the attributes have non-negative values which will make all the model’s components to contribute additively to the predicted grades.

IMPLEMENTATION:
Multi-linear regression model:

\[ Y = 0.0267(m) + 0.00444(e) + (-0.25824(sicse)/0.59825(sstate)) - 0.49231(llocal) + 0.56192(bckmaths) - 0.59242(csyes) + 1.79781(int) \]

\[ \text{lm(formula} = c \sim m + e + s + l + bck + cs + \text{int, data} = \text{trai}) \]

COMPARING APPROACHES:
In comparison with the method using all grades, both approaches had positive effects on the number of calculations. RMSE gives the standard deviation of the model prediction error. A smaller value indicates better model performance.

In this case, residuals of models are considered and RMSE is calculated for both multi linear regression (4.8) and SVM (4.3) has almost same error
rate so any of the models can be used to test the data and predict the values.

Fig 1.1 represents the RMSE rate for algorithms

ROOT MEAN SQUARE ERROR

Root mean square is calculated by identifying residuals and squaring residuals and finding mean of the squared residuals and finally calculating the root of the mean squared gives RMSE.

By using RMSE we can identify that which algorithm gives high accuracy results for the given datasets. In this case we observe [from fig 1.1] that SVM, linear model has least RMSE.

Experimental results

Fig 1.2 represents the graphical output for four algorithms in a test data

Applying linear model

Linear model coefficients

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>Local</th>
<th>Icecream</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.79761</td>
</tr>
</tbody>
</table>

SVM model coefficients

Number of support vectors: 414
Conclusion:
Predicting student’s performance would boost the results of student’s grades and gives teachers a better approach for teaching the students who are at risk of failure. Regression models, tree based models and created to make the best predictions with high accuracy. The basic idea is to increase the efficiency of the prediction results using various algorithms. Thus by finding the RMSE, we observed that SVM, linear model gives optimum results (from figure 1.2).

References:


