Predicting Student Performance using Advanced Learning Analytics

Ali Daud a,d
Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia
adfmohamad@kau.edu.sa

Naif Radi Aljohani a
Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia
nrajohani@kau.edu.sa

Rabeeh Ayaz Abbasi a,b
Department of Computer Science, Quaid-i-Azam University, Islamabad, Pakistan
rabbasi@kau.edu.sa

Miltiadis D. Lytras c
Computer Information Systems Department, The American College of Greece, Greece
mlytras@acg.edu

Farhat Abbas d
Department of Computer Science and Software Engineering, International Islamic University, Islamabad, Pakistan
farhatabbas421@gmail.com

Jalal S. Alowibdi e
Faculty of Computing and Information Technology, University of Jeddah, Saudi Arabia
jalowibdi@uj.edu.sa

ABSTRACT
Educational Data Mining (EDM) and Learning Analytics (LA) research have emerged as interesting areas of research, which are unfolding useful knowledge from educational databases for many purposes such as predicting students’ success. The ability to predict a student’s performance can be beneficial for actions in modern educational systems. Existing methods have used features which are mostly related to academic performance, family income and family assets; while features belonging to family expenditures and students’ personal information are usually ignored. In this paper, an effort is made to investigate aforementioned feature sets by collecting the scholarship holding students’ data from different universities of Pakistan. Learning analytics, discriminative and generative classification models are applied to predict whether a student will be able to complete his degree or not. Experimental results show that proposed method significantly outperforms existing methods due to exploitation of family expenditures and students’ personal information feature sets. Outcomes of this EDM/LA research can serve as policy improvement method in higher education.

CCS CONCEPTS
• Computing methodologies–Supervised learning by classification • Applied computing–Education

KEYWORDS
Learning Analytics (LA); Educational Data Mining (EDM); Student Performance Prediction; Family Expenditures; Students Personal Information

1. INTRODUCTION
Students are the main stakeholders of institutions/universities and their performance plays a significant role in a country’s social and economic growth by producing creative graduates, innovators and entrepreneurs [26]. Educational Data mining has emerged as very important area of research to reveal presentable and applicable knowledge from large educational data repositories. Data mining algorithms are used to obtain the hidden information and desired benefits from these large data repositories [17]. There is a critical demand for academic institutions to maintain and to integrate large datasets of learners for multipurpose decision making. The use of web technology has also become an integral part of the current era of education in many universities, increasing the actual amount of data about students, teachers and their interactions with learning and educational systems [7,15]. Higher education plays an important role in the development of a society. It is a field which provides a large amount of data about participants such as students, teachers, facilities and curricula [21]. The performance of students is a main concern of various stakeholders including educators, administrators and corporations. For recruiting fresh graduates, academic achievement is the main factor considered by the recruiting agencies. Therefore, graduates have to work hard for excellent grades, so that they may rise up to the expectations of recruiting agencies [26].

The sources of educational data may be broadly divided into two categories. The first category comprises centralized educational systems such as LMS; ‘centralized’ here means that the educational data in the analytics come from one source [4]. The second comprises the de-centralized educational data retrieved from different systems’ resources, such as WWW data, massive open online courses that make use of many systems to deliver the learning materials [5,18]. Recently, analysis of educational data, for instance learning analytics, academic analytics, educational data mining, predictive analytics and learners’ analytics has emerged as an innovative area of research [22]. The commonality between all of these terms is the use of educational data for multiple purposes. Recently, a new term has been introduced: ‘educational data science’ that clarifies how different disciplines and researchers with different research interests and backgrounds can work in this area [15].

Educators, parents and institutions would like to know the answer of the question “Is it possible to predict the performance of a student that is enrolled in an educational institution?” For example, will he complete his degree or not? However, currently the process of learning has been declared as an individual’s effort. Therefore, developing models for evaluating learning efforts of a student is not an easy task [16]. Recently, data mining techniques have been used to provide new insights for this problem. There are several diversified influential factors to evaluate the students’ performance. These can be identified by using data mining approaches in educational sector [17]. Data mining for an educational system is an iterative process for hypotheses development and testing. Fig. 1 shows the applications of data mining in educational systems. Student performance evaluation system can help in decision making for awarding scholarships or in other words targeting the right students.

The main contributions of this paper are as follows:

1. The impact of these features is determined as per their efficiency and most influential features are shortlisted.
2. The performance of standard models (both discriminative and generative) is analyzed by comprehensive experiments with baseline and proposed features.
3. The overall finding is that Learning Analytics based on personalized features can improve prediction of students’ performance. For sure the generalization of these findings requires additional research for the incorporation of additional features like talent, skills and personal competencies from different online web sources.

Rest of the paper is arranged as follows: Section 2 describes the related work. The problem definition is presented in Section 3. Section 4 provides applied models, performance measures, data collection and construction of feature space. Section 5 presents experiments conducted by using five classifiers with baseline and proposed feature sets. Finally, section 6 concludes this work.

2. RELATED WORK

The research problem of students’ performance prediction can be analyzed through diverse angles. In the current literature, a number of complimentary approaches provide a baseline for such an analysis. In an ideal scenario, a rich dataset with student identity along with numerous characteristics could be the basis for advanced learning analytics. The problem is that in most of the cases, not all the data are available for the dynamic construction of the student identity, further limited by lack of access to various sources. In Fig. 2, we briefly present some of the most representative methods of Applied Educational Data Mining and Learning Analytics based on a comprehensive literature review.

Student performance prediction has got a lot of attention from the educational data mining researchers. Typical data mining methods have been employed to deal with different tasks related to the students. A survey of data mining techniques for traditional educational systems such as adaptive web-based and content management systems is presented in [16]. An association rule based mining method is applied for selection of weak students in a school and is found effective [8]. Genetic Algorithm is used to assign the weights for the modeling of students’ grade for three levels (binary, 3-level and 9-level) [9]. It shows that the combination of multiple classifiers leads to a significant improvement in classification. A model is proposed for predicting student performance using six machine learning techniques for distance learning education, which is quite different from the traditional educational system [6]. The experimental results show that demographic and performance features are better predictors for predicting student performance.

A regression model is applied to predict the test score of subject for school students [14]. It concludes that mixed-effect models present best performance as compared to Bayesian network. A prediction model (CHAID) is developed to predict the performance of higher secondary school students, which is critical before getting admission into universities [16].

The grades of graduate students are predicted using Naïve Bayesian and Rule Induction classifiers [25]. Clusters are made from students’ data and the outliers are successfully identified. A model is presented to estimate the abilities of students and competence of teachers in order to predict the future student outcomes [19]. It shows that demographic profiles and personality traits features are correlated and have high impact on student performance. Similarly student performance evaluation and engineering students’ abilities are analyzed for improved recruitment process by using data mining methods [12].
3. PROBLEM DEFINITION

The formal definition of student performance prediction problem is described as:

Given \( n \) training samples \((X_i, z_i), (X_2, z_2)\ldots (X_m, z_m)\), where \( X_i \) is the feature vector for student \( a_i \) and \( A \) is the set of \( n \) students where \( A = \{a_1, a_2, a_3 \ldots a_n\} \). The \( X_i \in \mathbb{R}^m \) and \( m \) is the total number of features and \( z_i \) is the student performance status (degree completed or dropped) where \( z_i \in \{-1, +1\} \). To predict the performance of a student, the following prediction function is proposed:

\[
z = F(A/X)
\]

Where,

\[
F(A/X) = \begin{cases} 
\geq 0 & \text{if } z = +1, \text{ completed} \\
< 0 & \text{if } z = -1, \text{ dropped}
\end{cases}
\]

**Learning Task:** Goal is to learn a predictive function \( \hat{F}(.) \) or alternatively to predict whether a student will complete his/her degree or not. It is written as:

\[
\hat{z} = \hat{F}(A/X)
\]

4 METHODOLOGY

4.1 Models

Two types of classification models (discriminative and generative) are used to learning the desired predictive function \( \hat{F}(.) \). Two generative and three discriminative models are used for experimental analysis. They are selected on the basis of their frequent usage in the existing literature. The list of methods are as follows:

1. Support Vector Machine (SVM) [discriminative]
2. C4.5 [discriminative]
3. Classification and Regression Tree (CART) [discriminative]
4. Bayes Network (BN) [generative]
5. Naive Bayes (NB) [generative]

4.2 Performance Evaluation

For performance evaluation, three standard evaluation metrics (precision, recall and F1-score) are used. 5-fold cross validation is used for comparison with baseline methods. These performance evaluation parameters are defined as:

\[
\text{Precision} = \frac{TP}{TP + FP}
\]

\[
\text{Recall} = \text{sensitivity} = \frac{TP}{(TP + FN)}
\]

\[
\text{F1 score} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}
\]

4.3 Data Collection

For experimental purpose, the data of graduate and undergraduate students has been collected from different universities of Pakistan during the period (2004 to 2011). Initially, about 3000 student records were collected. Pre-processing is applied to obtain the most relevant characteristics of students. After removing inconsistencies and duplications in the dataset, we considered 776 student instances for experiments. The main goal of this research is to predict the student’s performance i.e. “will he/she complete his/her degree or will he/she drop”. The dataset consists of 690 instances of students, who have completed their academic degrees.
(true values) and 86 instances that are dropped in the midway or at the end (false values). In the first step, 20 students’ instances have been used (10 completed, 10 dropped), and then 40, 60, 80 and 100 instances of dataset are used. However, after 100 records/tuples, the performance remains unchanged as with the increase in instances in the data set. So, 100 instances of students (50 completed, 50 dropped) for experimental setup are selected. The distribution of the dataset is presented pictorially in Fig. 3.

Figure 3 Characteristics of dataset.

4.4 Construction of Feature Space
Feature set is constructed by considering four categories of characteristics related to student and his family. Initially, a pool of 33 features is constructed by combining some existing (baseline) and proposed features and then feature subset selection process is applied to remove/reduce the number of redundant features. Information Gain and Gain ratio are used to select the best feature subset.

Overall, four categories of features are collected (some from literature and some are proposed in this research work). Out of twenty-three, thirteen are our proposed features. Table 1 presents the description of each feature, its category and status (proposed or existing/old). Then feature subset selection process is adapted in the following manner: First of all, for comprehensive analysis of features’ comparison and best features selection, two measures, information gain and gain ratio are used. A threshold of 0.01 is used to identify the best feature subset. Finally, we get 23 features in which 13 are new (proposed) and 10 are old as shown in Table 1. We found larger information gain values of selected features as compared to gain ratio values. Because our dataset does not contain equal number of samples for both classes and information gain is biased towards maximum attribute values.

5. EXPERIMENTAL RESULTS
In this section, comprehensive experiments are presented using data set that is designed based on student’s information acquired from different universities of Pakistan as described in Section 4.3. The dataset consists of 100 student records (tuples) and 23 features. Therefore, we get a $100 \times 23$ feature matrix. Default parameters are used for all classifiers using Weka 3.7. Five-fold cross validation method is used to evaluate the accuracy of all the classifiers. Experiments are conducted in two ways: Firstly, influence of individual feature for the student’s performance is analyzed. Secondly results of classifier on baseline methods and proposed feature sets are evaluated.

5.1 Individual Feature Analysis
This section evaluates the impact of each feature for the prediction of student’s performance. Twenty-three features (selected by the feature extraction process) are selected for experiments. In experiments five classifiers are used (BN, NB, SVM, C4.5 and CART) to analyze the influence of each feature for predicting the performance of students as shown in Fig. 4. We find the “natural gas” expenditure is the best predictor for the desired student’s performance using C4.5 classification method as shown in Fig. 4. BN and NB methods show second and third highest F1-scores using same features. Other family expenditure features also play important roles. The “Stock Value” feature has the lowest performance for prediction and all classifiers present same F1-score (0.333). By analyzing performance of best and worst features, that conclude the proposed proposition based on the family expenditure feature set improves classification accuracy.

“Self Employed” is found to be the second-best feature that also belongs to proposed feature set of student’s personal information and all classifiers show 0.77 F1-score which represents better performance by using proposed feature. The third best feature is “Location” which belongs to the baseline feature set. If we critically analyze the impacts of other proposed features in comparison with old features, better accuracy is obtained by using our proposed feature space as compared to the old feature space. Hence, it can be concluded that students’ “natural gas” expenditure, “electricity” expenditure, “self-employed” and “location” characteristics are most influential for prediction of his/her performance in academics.

5.2 Comparisons
Performance of classifiers is analyzed using four baseline methods and our proposed feature sets based method and results are critically analyzed. The performance of experiments is evaluated by F1 score. The purpose of this experiment is to analyze the influence of proposed and existing features based methods for the student performance prediction task.

The feature sets proposed by [16,12,25,23] are considered as baseline for comparison as shown in Fig 5. Proposed method significantly outperforms baseline methods as shown in Fig. 5. SVM performs best for our proposed feature sets with F1 score of 0.867, which is 13% more as compared to second best method for SVM model. BN and NB classifiers overall perform better in case of most methods as compared to C4.5 and CART. For C4.5 model the performance of most methods is very low and unstable.

5.3 Discussion
This research work presents the student academic prediction methods that use four different types of features namely: family expenditure, family income, student personal information and family assets. It also adapts the process of feature subset selection in order to identify the most effective determinants for student academic performance prediction. It is evident from the comparative analysis that our proposed features are important predictors and achieved F1-score of 86% (Fig. 5) on real life undergraduate students’ data.
Table 1 Features Distribution.

<table>
<thead>
<tr>
<th>Category</th>
<th>Name</th>
<th>Description</th>
<th>Status</th>
<th>Info. Gain</th>
<th>Gain Ratio</th>
<th>Features Used</th>
</tr>
</thead>
<tbody>
<tr>
<td>Family Expenditure</td>
<td>Electricity Bill</td>
<td>Average of Electricity bills for last six months</td>
<td>New</td>
<td>0.38</td>
<td>0.05</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Natural Gas Bill</td>
<td>Average of Gas bills for last six months</td>
<td>New</td>
<td>0.26</td>
<td>0.06</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Telephone Bill</td>
<td>Average of Telephone bills for last six months</td>
<td>New</td>
<td>0.10</td>
<td>0.04</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Water Bill</td>
<td>Average of Water bills for last six months</td>
<td>New</td>
<td>0.10</td>
<td>0.06</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Food Expenses</td>
<td>Average of food expenses for last six months</td>
<td>New</td>
<td>0.09</td>
<td>0.03</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Miscellaneous Expenditure</td>
<td>Average of Miscellaneous Expenditures for last six months</td>
<td>New</td>
<td>0.11</td>
<td>0.02</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Medical</td>
<td>Average of Medical Expenditures for last six months</td>
<td>New</td>
<td>0.06</td>
<td>0.01</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Family Expenditure on Education</td>
<td>Average of Family Expenditure on education for last six months</td>
<td>New</td>
<td>0.35</td>
<td>0.04</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Accommodation Expenses</td>
<td>Average of Accommodation Expenses for last six months</td>
<td>New</td>
<td>0.27</td>
<td>0.25</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Studying Family Members</td>
<td>Total number of studying family members of student</td>
<td>Old</td>
<td>0.008</td>
<td>0.003</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Dependent Family Member</td>
<td>Total number of dependent family members of student</td>
<td>Old</td>
<td>0.02</td>
<td>0.007</td>
<td></td>
</tr>
<tr>
<td>Family Income</td>
<td>Father Income</td>
<td>Per month income of father/guardian of student</td>
<td>Old</td>
<td>0.29</td>
<td>0.04</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Mother Income</td>
<td>Per month income of mother of student</td>
<td>Old</td>
<td>0.02</td>
<td>0.03</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Land Income</td>
<td>Per month income from land of family of student</td>
<td>Old</td>
<td>0.02</td>
<td>0.05</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Miscellaneous Income</td>
<td>Per month miscellaneous income of family of student</td>
<td>Old</td>
<td>0.08</td>
<td>0.03</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Earning Hands</td>
<td>Total number of Earning hands of student’s family</td>
<td>Old</td>
<td>0.007</td>
<td>0.005</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Father Status</td>
<td>Status of father of student: alive or deceased</td>
<td>New</td>
<td>.0008</td>
<td>.0001</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Father Retired</td>
<td>Father retired or in service</td>
<td>New</td>
<td>0.002</td>
<td>0.003</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Guardian Alive</td>
<td>Is student’s guardian alive</td>
<td>New</td>
<td>0.003</td>
<td>0.004</td>
<td></td>
</tr>
<tr>
<td>Student Personal Information</td>
<td>Gender</td>
<td>The gender of the student (male or female)</td>
<td>Old</td>
<td>0.004</td>
<td>0.005</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Marital Status</td>
<td>Marital status of student (married or unmarried)</td>
<td>New</td>
<td>0.003</td>
<td>0.01</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>House Owner Ship</td>
<td>Student have his/her own house</td>
<td>New</td>
<td>0.08</td>
<td>0.10</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Previous Program Scholarship</td>
<td>Scholarship received or not in previous academic program</td>
<td>New</td>
<td>.0002</td>
<td>.0003</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Previous Institution Type</td>
<td>Type of student previous institution</td>
<td>Old</td>
<td>0.001</td>
<td>0.002</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Self Employed</td>
<td>Is student is self employed</td>
<td>New</td>
<td>0.06</td>
<td>0.04</td>
<td>✓</td>
</tr>
<tr>
<td>Family Assets</td>
<td>Land Value</td>
<td>Current value of lands belongs to student’s family</td>
<td>Old</td>
<td>0.04</td>
<td>0.02</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Bank Balance</td>
<td>Bank balance of student’s family</td>
<td>Old</td>
<td>0.05</td>
<td>0.07</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Stock Value</td>
<td>Value of Shares/Bonds belong to student’s family</td>
<td>Old</td>
<td>0.01</td>
<td>0.08</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>House Value</td>
<td>Value of house belong to student’s family</td>
<td>Old</td>
<td>0.14</td>
<td>0.03</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>House Condition</td>
<td>Structure of house belong to student’s family</td>
<td>New</td>
<td>0.06</td>
<td>0.04</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Miscellaneous Asset Value</td>
<td>Any other assets related to student</td>
<td>Old</td>
<td>0.04</td>
<td>0.02</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Location</td>
<td>Type of Location where student resides; urban or rural</td>
<td>Old</td>
<td>0.03</td>
<td>0.04</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>No of Vehicles at home</td>
<td>How many vehicles belong to family of a student</td>
<td>Old</td>
<td>0.005</td>
<td>0.008</td>
<td></td>
</tr>
</tbody>
</table>

Figure 4 Impact of selected individual features on classification accuracy.
because having a comfortable living accommodation enables the students, both these factors are very helpful for students in achieving better performance in their studies. Last but not the least, the house condition of a student is also an influential factor due to intuitive reasons provided in discussions. In future, WWW research on Digital Learning and Learning Analytics should be focused on the following directions:

- Which kinds of methods and flexible applications permit the construction of critical learners’ data from the WWW, e.g., mining of social media content can be a basis for personal expenditure?
- Which are the possibilities to proceed to dynamic profiling of personal characteristics of students from the Deep Web?
- Which are the standards for codifying critical students’ information in the WWW, and how this can envision future WWW based learning services?

Learning analytics of mobile and ubiquitous learning environments from the perspective of human computer interaction [1,2] also require detailed exploration in addition to aforementioned traditional and web based features.

**REFERENCES**


