Framework for Predicting Students' Performance – Review of Techniques

** Subhabaha Pal*

Abstract

With the advent of self-funding institutions in the education sector, retaining students along all the semesters of the academic program in order to maintain the desired cash-flow through re-registration has become a major challenge. A common tendency of students is discontinuing a course if they fail in any subject. Hence, predicting students' probable performance in semester-end examination and helping them raise performance level by providing necessary guidance have become the need of the day. Many research works have been directed towards this area. This paper attempted to provide a comprehensive review of all major researches that have been undertaken for prediction of student performance. It also attempted to suggest a common framework which can be used in studies related to student performance prediction. This framework can be used for predicting student performance on the basis of availability of relevant student data.

Keywords : Student performance prediction, Classification methods, Rule-based classification, Regression methods, Data-mining

I. INTRODUCTION

The efforts on the part of the government as well as private bodies to set up higher education institutes have been witnessed in both developed as well as developing countries. More self-funding educational institutions for higher education have come up in recent years with privatization of education. However, maintaining the quality of education by enhancing student performance everywhere has been a question in majority of the new institutions. Enhancing student performance should be the first priority of any educational institution as excellent records of academic achievements are perceived as one of the most important criteria for a high quality university [2]. In order to maintain quality of education and to enhance student academic performance, proper guidance to students should be given. Proper guidance calls for predicting student performance beforehand so that students can be counseled properly. Importance of predicting student performance lies in the fact that it is necessary for enhancing the students' academic performance [3, 4]. An economic aspect has also necessitated predicting students' performance. Retention is a major problem in higher education institutes (HEIs) today [6, 14, 15]. With advent of self-funding institutions in the educational sector, a major challenge for them is retaining students in the latter academic sessions. Discontinuing the course after not successfully completing the previous semester has become a regular affair. Li, Rusk, and Song [21] opined that 35% of the students cannot reach the 2ndor 3rd year and leave the course in between. Self-funding institutions are the most affected by this trend as these depend heavily on tuition fees and the loss of re-registration in the latter semester affects the cash in-flow to a large extent. Predicting academic performance of individual students may help in taking appropriate measures to increase the pass percentage of students ultimately leading to retaining the students for the latter academic sessions [1]. The scenario is more critical for distance learning courses where attrition percentage of un-successful students is quite high. Distance education has emerged in response to the need of providing access to those who would otherwise not be able to participate in face-to-face(F2F) courses [7]. Distance education encompasses those programs that allow the learner and instructor to be physically apart during the learning process and maintain communication in a variety of ways [8]. A major disadvantage in distance education is that the trainee lacks direct monitoring by the instructor which is possible in case of class-room training. Hence, predicting student performance in distance learning courses is more important so that the institute may properly guide the student in finishing semester

Manuscript received January 20, 2017; revised March 8, 2017; accepted March 9, 2017. Date of publication April 15, 2017;
* S. Pal is with T. A. Pai Management Institute, Karnataka, India-567104 (email: subhabaha@tapmi.edu.in)
Digital Object Identifier 10.17010/ijcs/2017/v2/i2/112040

examinations successfully, thus confirming the re-registration for the next semester. This will ultimately help in confirming the cash in-flow from re-registration in the latter semesters.

Several techniques have been proposed to predict student performance [11]. These techniques call for working with huge mass of data. Data mining is one of the most popular techniques for analyzing student performance. Educational data mining has widely been used since the last decade. Educational data mining is the process used to extract useful information and patterns from a huge educational database [12]. Romero and Ventura [13] introduced the topic 'educational data mining' and described different group of users, types of educational environments, and data they provide. They also listed the common tasks in the educational environment that have been resolved through data mining techniques. Naren, Elakia, Gayathri, and Aarthi [24] proposed that the current educational data-mining techniques including classification can be used to suggest career options for high school students and also to predict potentially violent behavior among students. Thai-Nghe proposed the use of recommender system for predicting student performance. Parack, Zahid, and Merchant [23] proposed the use of Apriori algorithm and K-Means clustering to profile students on the basis of various parameters like exam scores, term work grades, attendance, and practical exams. These useful information and patterns revealed through different data-mining techniques can also be used to predict and improve student performance [11, 17]. Student performance prediction is also useful for student placement services [50]. Angeline and James [50] discussed association rule-based classification techniques for student performance prediction at the time of placements.

A question that comes in the mind of each fresh researcher is how to proceed to predict student performance. In order to extract relevant data for analysis, it is required to know the data and attributes that should be considered for predicting student performance. Methodology also plays a huge role and it is needed to know which methodology should be used for prediction. It is necessary to have a specific framework which should be followed in order to make such a study. This paper was an attempt in the direction of proposing specific framework which can be used while conducting studies related to student performance management.

This paper reviewed all major researches that have happened in the area and tried to specify the attributes that the researchers have considered for predicting student performance. It also reviewed the methodologies used for the purpose. This paper provides a summary of the major works in this area. This paper also proposed a framework which one can follow in case one wishes to work on predicting student academic performance.

II. REVIEW OF LITERATURE

Many studies have been conducted in the area of student performance prediction. This section reviews the major works that have been done in this research area encompassing the prediction of student performance in both F2F and distance education. McKenzie and R. Schweitzer [30] is the earliest available work towards the direction of predicting student performance. The paper is about predicting the university performance of students based on their performance in the first semester. Li et al. [21] tried to identify the factors that serve as good indicators of whether a student will drop out or fail the program and it used the principal component analysis to establish the early warning indicators. Jishan, Rashu, Haque, and Rahman [22] proposed the use of discretization method called the Optimal Equal Width Binning and an over-sampling technique known as the Synthetic Minority Over-Sampling (SMOTE) to improve the accuracy of students' final grade prediction model for a particular course. Kumar [25] applied neural networks to predict whether a student will achieve his goal using his academic performance and attitude towards self-regulated learning. Several data-mining techniques have been proposed for accurately predicting student performance [11]. In [5, 52, 55], the authors reviewed the research works directed towards the application of data mining techniques in student performance prediction. Natek and Zwilling [19] focused on the small HEI datasets to predict student performance in terms of final grade using 2 data-mining tools. The paper opined that the data-mining tools can also be used for small set of data available in HEI.

Bin Mat, Buniyamin, Arsad, and Kassim [10] reviewed of the available academic analytics tools for few institutions and also proposed a theoretical model involving decision tree for making rules and neural network for prediction of academic performance. Some researches on educational data-mining (especially on student performance prediction) have also been directed in the field of distance learning which is gradually gaining its place in the education sector quite fast. Kotsiantis, Pierrakeas, and Pintelas [29] worked towards

the direction of predicting distance learning student performance using Hellenic Open University students' key demographic characteristics and their scores in few written assignments. One of the main component in distance education is e-learning portal which plays a major role in course delivery in this mode of education. Several researches have been directed towards predicting the performance of distance education students by analyzing their activities in the e-learning portal. The works of Tucker and Divinsky and Minaei-Bidgoli, Kashy, Kortemeyer, and Punch [28] were directed towards prediction of distance/online education student performance based on the activities data generated from e-learning portals. Shahiri, Husain, and Rashid [5], and Thakar, Mehta, and Manisha [52] opined that there are two main factors in predicting student performances; these are attributes and prediction method. Several factors/attributes related to students and the course of study are considered which are assimilated to develop a method/model in order to predict student performance. In [3, 12, 15, 16, 17, 18, 19] CGPA was used as the main attribute to predict student performance. According to Shahiri et al. [5], the main reason why most of the researchers are using CGPA is that the attribute CGPA has a tangible value for future educational and career mobility. In [12, 18, 23, 24, 25, 26, 28], the attributes like assignment marks, quizzes, lab-work, class test and attendance were grouped in a single attribute called internal assessment and it was used for student performance prediction. The studies [9, 10, 13, 16, 17, 19, 24, 32, 33] used attributes like gender, age, family background, and disability for student performance prediction. The reason behind utilization of demographic attribute like gender for performance prediction is that learning styles of male and female students are different [5, 10]. The results of Meit, Borges, Cubic, and Seibel [34] suggested that most of the female students have more positive learning styles and behaviors compared to male students. Simsek and Balaban [35] suggested that female students have effective learning strategies in their study. In [12, 18, 19, 24, 36], the authors considered extracurricular activities as attributes for predicting student performance. In [16, 17, 32, 33], the authors considered high-school background as an important attribute for prediction. In [16, 26, 37, 38], the authors opined that attributes like social interaction network contribute substantially in student academic performance. Psychometric factors which include student interest, study behavior, engagement time, and family support contributed to student performance. However, psychometric factors were rarely used as attributes because these focus more on valid qualitative data which is hard to get from respondents [5]. In [36, 39, 40, 41], the authors used psychometric factors as attributes for predicting student performance.

In academic performance prediction, choosing appropriate method plays an important role. In educational data mining method, predictive modeling is generally used in predicting student performance. For prediction, the different techniques used included classification, regression, and categorization [5]. There are different classification algorithms which are generally used to predict student performance – namely, decision tree, Artificial Neural Network, Naïve Bayes, K-Nearest Neighbour, and Support Vector machine. Garcia-Saiz and Zorrilla [31] used different classification techniques applied on distance learning-related educational datasets and proposed a meta-algorithm to pre-process the datasets to improve accuracy of the student performance prediction model.

The summary of the attributes and classification methods used for predicting student performance by different researchers in prominent works in the field is presented in table I with partial inputs from Shahiri et al. [5].

Though major researchers have used different classification techniques predominantly for predicting student performance, some researchers have undertaken work on the basis of regression methods. Table II presents review of the attributes and regression methods used by researchers for student performance-related prediction.

Few works have been done using regression method for predicting student performance. Researchers have pre-dominantly depended upon different classification techniques for performance prediction as evident from the previous review. Thai-Nghe's [11] is the only work using logistic regression for prediction purpose. However, it is mainly a forecasting related study where the performance of the student in a course is forecasted based on the past repeated performance in the same course. In real-life, both face to face (F2F) or distance education, the scenario where a student is appearing for the examination of the same course repeatedly is very rare. However, logistic regression has not been used for the cases where a student's performance needs to be predicted in a course at the end of the session on the basis of other attributes.

TABLE I.
REVIEW OF ATTRIBUTES AND CLASSIFICATION METHODS FOR PREDICTING STUDENT PERFORMANCE WITH ACCURACY/ MODEL OUTPUT

Methods	Authors	Attributes	Accuracy / Model Output
Decision Tree	Ahmad, Ismail, and Aziz [51]	Student demographic records, previous academic performance, family background information	67%
	Quadri and Kalyankar [15]	Gender, attendance, previous semester grade, parent education, parent income, scholarship, first child or not, working or not	75%
	Ruby and David [53]	Theory examination score, medium of study, previous course studied, UG percentage, stay, extra-curricular activities, family income	76%
	Cortez and Silva [48]	Past school grades (first and second periods), demographic, social and other school related data	93%
	Romero, Ventura, Espejo, and Hervás [27]	Internal assessments	76%
	Gray, McGuinness, and Owende [40]	Psychometric factors	65%
	Bunkar, Singh, Pandya, and Bunkar [42]	External assessments	85%
	Jishan et al. [22]	CGPA	91%
	Osmanbegovi? and Sulji? [16]	CGPA, student demographic, high school background, scholarship, social network interaction	73%
	Mayilvaganan and Kapalnadevi [18]	Internal assessments, CGPA, Extra-curricular activities	66%
	Ramesh, Parkavi, and Ramar [33]	Student demographic, high school background	65%
	Naren et al. [24]	Internal assessment, student demographic, extra-curricular activities	90%
	Natek and Zwilling [19]	External assessment, CGPA, student demographic, extra-curricular activities	90%
	Mishra, Kumar, and Gupta [36]	Psychometric factors, extra-curricular activities, soft skills	88%
	Ibrahim and Rusli [9]	Student demographic profile, information technology application knowledge, previous school whether boarding or non-boarding, programming knowledge, family financial status, CGPA	Square-root of Average Square Error (RASE) = 0.1769
Artificial	Wang and Mirovic [43]	Internal Assessments	81%
Neural	Gray et al. [40]	Psychometric factors	69%
Network	Arsad, Buniyamin, and Manan [44]	External assessments	97%
(ANN)	Jishan et al. [22]	CGPA	75%
	Osmanbegovi? and Sulji? [16]	CGPA, student demographic, high school background, scholarship, social network interaction	71%
	Ramesh et al. [33]	Student demographic, high school background	72%
	Oladokun, Adebanjo, and Charles-Owaba [32]	External assessment, student demographic, high school background	74%
	Kumar [25]	Internal assessment, external assessment	98%
	Ibrahim and Rusli [9]	Student demographic profile, information technology application knowledge, previous school whether boarding or non-boarding, programming knowledge, family financial status, CGPA	Square-root of Average Square Error (RASE) = 0.1714
Naïve Bayes	Ruby and David [53]	Theory examination score, medium of study, previous course studied, UG percentage, stay, extra-curricular activities, family income	80%
	Ahmad et al. [51]	Student demographic records, previous academic performance, family background information	67%
	Osmanbegović and Suljić [16]	cgpa, student demographic, high school background, scholarship, social network interaction	76%
	Ramesh et al. [33]	Student demographic, high school background	50%

	Jishan et al. [22]	CGPA	75%
	Mayilvaganan and Kapalnadevi [18]	Internal assessments, CGPA, Extra-curricular activities	73%
	Cortez and Silva [48]	past school grades (first and second periods), demographic, social and other school related data	91%
	Christian and Ayub [20]	Students' personal data, admission data, academic data	N.A.
K-Nearest	Gray et al. [40]	Psychometric factors	69%
Neighbour	Mayilvaganan and Kapalnadevi [18]	Internal assessments, CGPA, extra-curricular activities	83%
	Bidgoli et al. [28]	Internal assessments	82%
Support	Gray et al. [40]	Psychometric factors	78%
Vector Machine	Sembiring, Zarlis, Hartama, Ramliana, and Wani [39]	Psychometric factors	83%
	Mayilvaganan and Kapalnadevi [18]	Internal assessments, CGPA, extra-curricular activities	80%
	Hamalainen and Vinni [45]	Internal assessments, CGPA	80%
Rule Based Classification	Ahmad et al. [51]	Student demographic records, previous academic performance, family background information	71.3%
	Abuteir and El-Halees [17]	Matriculation GPA, gender, specialty, city, secondary school type, grade	N.A.

TABLE II.

REVIEW OF ATTRIBUTES AND REGRESSION METHODS FOR PREDICTING STUDENT PERFORMANCE WITH ACCURACY/ MODEL OUTPUT

Regression Method	Authors	Attributes	Accuracy / Model Output
Linear Regression	Cortez and Silva [48]	Past school grades (first and second periods), demographic, social and other school related data	85%
	Ibrahim and Rusli [9]	student demographic profile, information technology application knowledge, previous school whether boarding or non-boarding, programming knowledge, family financial status, cgpa	Square-root of Average Square Error (RASE) = 0.1848
Personal Linear Multi-regression	Elbadrawy, Polyzou, Ren, Sweeney, Karyois, and Rangwala [14]	E-Learning content access	Root Mean Square Error (RMSE) = 0.78
Exponential Smoothing	Thai-Nghe [11]	Performance in previous turns in same course	RMSE = 284
Logistic Regression	Thai-Nghe [11]	Performance in previous turns in same course	RMSE = 310

III. FRAMEWORK FOR PREDICTING STUDENT PERFORMANCE

In this section, a framework has been proposed which can be used when studies related to predicting student performance need to be undertaken.

As mentioned in [5], attributes play an important role in student performance prediction. Selecting proper attributes for student academic performance prediction makes the job half done.

A. Attributes Selection

1) Past Academic Performance: Past academic performance plays an important role in predicting student performance. It has been seen in different studies that students who have good performance in previous examinations have better chance of successfully completing the latter examinations. Suppose it is needed to predict whether a student will successfully complete a course / paper in a particular semester of an undergraduate academic program. In this case, it will be helpful if the following data on academic performance is considered –

a) Academic performance in 10^{th} standard (Total)

b) Academic performance in 12^{th} standard (Total)

c) Academic performance in all courses/papers in the previous semester of the academic program (if the course is not of the first semester)

In case of post graduate program course performance prediction, the additional data that can be considered is

the under-graduate program performance. The individual scores can be aggregated and an aggregate score can be calculated which may be used for predicting future academic performance. The studies [9, 15, 16, 17, 22, 48, 51, 53] consider the past academic performance as an important attribute for student academic performance prediction.

In the case of predicting the performance of students in quantitative subjects like mathematics, statistics, operation research, and analytics etc., the researcher may consider past performance in only mathematics-oriented subjects, if available. This reduces bias in the data.

2) Past Performance in the Course: Past performance by other students in the subject/paper also plays an important role. Suppose, past records show that out of the 100 students in the course, only 70 students can pass this course. This fact needs to be considered while framing a specific model for student performance prediction. Suppose, a post-graduate business management program subject is considered for predicting student performance. Now, in the business management program, the students from different undergraduate streams like engineering, arts, science, as well as commerce take admission. The past results of students from different streams are available in a particular subject. From that data, it will be known that out of total number of students, how many students from engineering stream have passed, how many from arts stream have passed etc. This will help in determining the probability of a student of a particular stream clearing a particular subject.

3) Student High-school/College Background: As discussed earlier, the studies [16, 17, 32, 33] consider high-school/previous college education background as important attributes for prediction. Suppose, a student is from arts program background, then his chance of passing a subject/paper in a quantitative subject in management will be less. This attribute is particularly true in case of management programs at the post-graduate level, where students from different under-graduate streams come and join the management program.

4) Demographic Data: Demography plays an important role in student performance. As mentioned earlier, the reason behind utilization of demographic attribute such as gender for performance prediction is that the learning styles of male and female students are different [5, 10]. Meit et al. and Simsek and Jalaban [34, 35] also considered gender as an important attribute in predicting student performance. In this kind of study, gender of the student should be considered as an important attribute and the model should be developed accordingly.

5) Academic Activities: Academic activities of the students give a glimpse of seriousness of the student in completing the subject successfully. Few studies like [9, 16, 27, 42, 43, 44, 53] consider activities like submitting internal assessments on time, accessing e-learning portal etc. as important attributes for predicting student performance. If students engross themselves in academic activities more, there are high chances that they will successfully complete the program. If the institute has any online learning portal, the login and access pattern of the student in the portal will be an important attribute in determining the future performance of the student in the course.

6) Extra-curricular Activities: As discussed earlier, studies such as [12, 18, 19, 24, 36] have considered extracurricular activities as an attribute for predicting student performance. If the students engross themselves in more activities other than academic activities, it may pose a negative impact on their performance. Socializing more through online portals like Facebook, Instagram, Twitter, and other websites may affect chance of students successfully completing a subject.

7) Family Background: Family background plays an important role in student academic performance. It is believed that students from less affluent economic background have more chances of being unsuccessful in a course due to their involvement in other activities to support family and a higher tendency to discontinue education [9, 51, 53]. Hence, factors like family background and family income should be considered important attributes in such studies.

8) Psychometric Factors: Psychometric factors like student behavior also play a role in student academic performance. In [36, 39, 40, 41], the authors included psychometric factors also as attributes for student performance prediction. However, as mentioned earlier, psychometric factors are rarely used.

9) Location: The current location of the student plays a role in student performance specifically, in case of distance education. If the student is from a metro city, there are better avenues for accessing internet which

helps in the study and it ultimately leads to better performance. Abuteir and El-Halees [17] suggested current location as an attribute in the study.

B. Method Selection

Selecting appropriate method for student performance prediction is the next major task in the study. The different techniques that can be used for student performance prediction include classification, regression, and categorization. Any of the above approaches can be taken for the study.

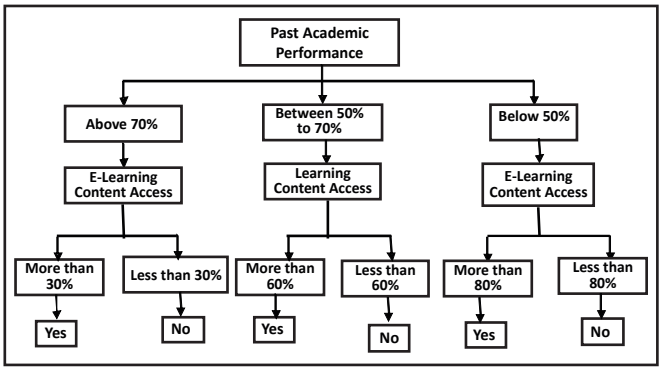
Classification is a popular approach. Different classification algorithms available for this kind of study are mentioned in the following sections:

1) Decision Tree: Decision tree is a popular algorithm which can be used for predicting student performance. This algorithm makes classification decision for a test sample with the help of tree like structure where –

i. Nodes in the tree are attribute names of the given data
ii. Branches in the tree are attribute values
iii. Leaf nodes are class labels

A simple example is given below on how the decision tree algorithm can be used for predicting student academic performance outcome. The classification is done based on two attributes – past academic performance and e-learning content access. Based on the data of these two attributes, it is predicted whether a student will pass (Yes) or fail (No) in a particular course/subject.

Fig. 1. Decision Tree Classification Algorithm for Predicting Student Performance Outcome



In [9, 15, 16, 18, 19, 22, 24, 27, 33, 36, 40, 42, 48, 51, 53], the authors used the decision tree classification algorithm for student performance prediction.

2) Artificial Neural Network: Neural networks is a field of Artificial Intelligence (AI) where the researchers, with inspiration from the human brain, find data structures and algorithms for learning and classification of data Nielsen 4i [57]. Many tasks that humans perform naturally fast such as recognizing a familiar face proves to be a very complicated task for a computer when conventional programming methodologies are used. By applying Neural Network techniques, a computer program can learn by examples, and can create an internal structure of rules to classify different inputs, such as recognizing images.

The same methodology can be used for student performance prediction also. Suppose data for 2000 students who have appeared for computer science examination are given with attributes such as past academic performance, demographic data, and e-learning access pattern. It is known whether the students have passed or failed. Now all the data inputs for the neural network classification model are training data and based on the data, the algorithm internally makes some rules which classifies students as successful and fail. When the rules are made, another 500 test students data can be input with those attributes but without the outcome (whether pass or fail). Then the neural network algorithm will classify these data as passed or failed based on the rules developed before (based on the 2000 students training data).

Rebym, Lek, Dimopoulos, Joachim, Lauga, and Aulagnier [56] proposed ANN classification method for use in behavioral sciences. The studies [9, 22, 25, 32, 33, 40, 43, 44] have used the ANN algorithm for student performance prediction.

3) Naïve Bayes: Naive Bayes algorithm is a classification technique based on Bayes' theorem with an assumption of independence among predictors [58]. A Naïve Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. Naïve Bayes model is easy to build and particularly useful for very large datasets. Along with simplicity, the Naïve Bayes model is known to outperform even highly sophisticated methods. Bayes theorem provides a way of calculating posterior probability P(B/A) from P(B), P(A) and P(A/B). The following is the Bayes probability theorem

$$P(B/A) = [P(A/B)*P(B)]/ P(A) \qquad (1)$$

Where,

a) P(B/A) is posterior probability of B given predictor A

TABLE III.
TRAINING DATA FOR NAÏVE BAYES CLASSIFICATION

Student	1	2	3	4	5	6	7	8	9	10
Previous Academic Performance	LT 50	GT 80	BT 50 to 80	BT 50 to 80	BT 50 to 80	LT 50	GT 80	GT 80	BT 50 to 80	LT 50
Performance in Mathematics	PASS	PASS	PASS	PASS	PASS	FAIL	PASS	PASS	FAIL	FAIL

b) P(B) is the prior probability of B
c) P(A/B) is the likelihood of A given B
d) P(A) is the prior probability of predictor A

The following simple example shows how the Naïve Bayes classifier can be used for student performance prediction problem. Suppose we have the data of 10 students on the attribute previous academic performance and whether they have passed in the subject mathematics in the current term. The previous performance of the students are classified as – Less than 50% (LT 50), Between 50% to 80% (BT 50 to 80) and Greater than 80% (GT 80).

The data is given in Table III.

As per the data, likelihood or probabilities are as follows-
P(LT 50) = 3/10 = 0.3
P(BT 50 to 80) = 0.4
P(GT 80) = 0.3
P(Pass) = 0.7
P(Fail) = 0.3

Now, we need to classify a student who has secured BT 50 to 80 and predict whether the student will pass or not, i.e., it is needed to find out the probabilities:
P(Pass/BT 50 to 80) and P(Fail/BT 50 to 80).
Now, P(Pass/BT 50 to 80) = P(BT 50 to 80/Pass).P(Pass)/P(BT 50 to 80)
and P(Fail/BT 50 to 80) = P(BT 50 to 80/Fail).P(Fail)/P(BT 50 to 80).

Now, P(BT 50 to 80/Pass) means likelihood that a student who has passed obtained between 50% to 80% in the previous examination.

P(BT 50 to 80/Pass) = 3/7
P(BT 50 to 80/Fail) = 1/3

Hence, P(Pass/BT 50 to 80) = P(BT 50 to 80/Pass)* P(Pass)/P(BT 50 to 80) = (3/7)*(0.7)/(0.4)=0.75
P(Fail/BT 50 to 80) = P(BT 50 to 80/Fail)*P(Fail)/P(BT 50 to 80) = (1/3)*(0.3)/0.4) = 0.25

Hence, a student who has secured between 50% to 80% in previous examination has more chance (75%) of passing in mathematics and only 25% chance of failing.

Hence, if the performance of the student who has secured between 50% to 80% in the previous examination is to be predicted through Naïve Bayes algorithm based on the training dataset mentioned above, it will be predicted that the student will pass the examination.

Naïve Bayes classifier should be used in student performance prediction in the manner mentioned above. [16, 18, 22, 33, 48, 51, 53] are some of the studies which have used the Naïve Bayes classifier for student performance prediction.

4) K-Nearest Neighbour: K-Nearest Neighbour (KNN) is another simple algorithm that stores all available cases and classifies new cases based on similarity measure (e.g. distance function) [59]. KNN is widely used in statistical estimation and pattern recognition since long and it is a widely recognized non-parametric technique.

In case of K-NN classification algorithm, a case is classified by a majority vote of its neighbor, with the case being assigned to the class most common among its K nearest neighbours measured by a distance function. If K = 1, then the case is simply assigned to the class of its nearest neighbor.

A simple example is given below to demonstrate how K-NN classification can be used for student performance prediction. In the below example, the e-learning portal access duration of 10 students are given and it is also known whether they have passed or failed in a certain subject for which they accessed the e-learning portal.

Now the performance of a student is there whose performance needs to be predicted. The student has accessed for 132 minutes. Now, if K is accepted as 1, then his nearest neighbor is student 4 (access time 133 minutes). As the student 4 has passed in the subject, his predicted performance will also be Pass.

TABLE IV.
TRAINING DATA FOR K-NN CLASSIFICATION

Student	1	2	3	4	5	6	7	8	9	10
E-Learning portal access time (in minutes)	110	125	130	133	140	134	85	200	40	45
Performance in mathematics	PASS	PASS	PASS	PASS	PASS	FAIL	PASS	PASS	FAIL	FAIL

However, if K is accepted as 3, i.e, K=3, then his nearest 3 neighbours will be Student 3 (130 Minutes), Student 4 (133 Minutes) and Student 6 (134 Minutes). Now, 2 students (Student 3 and Student 4) have passed and 1 student (Student 6) has failed. Hence, this predicted performance of the current student will be taken as 'Pass' as majority of his neighbours have passed.

Here, only one attribute, i.e., e-learning portal access time has been taken. Multiple attributes can also be considered. The distance will be Euclidean in case of two attributes. In case of attributes with categorical variable, the Hamming distance should be considered [59].

The studies [18, 28, 40] used K-NN algorithm for student performance prediction.

5) Support Vector Machine: Support Vector Machine (SVM) is a supervised machine learning algorithm which is used for classification [60]. In this algorithm, each data item is plotted as a point in n-dimensional space (where n is the number of attributes) with the value of each attribute being the value of a particular co-ordinate. Then the classification is performed by finding the hyper-plane that differentiates the classes very well. Given a set of training examples, each marked as belonging to one or the other two categories, a Support Vector Machine (SVM) training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible [61]. New examples are then mapped into the same space and predicted to belong to a category based on which side of the gap they fall.

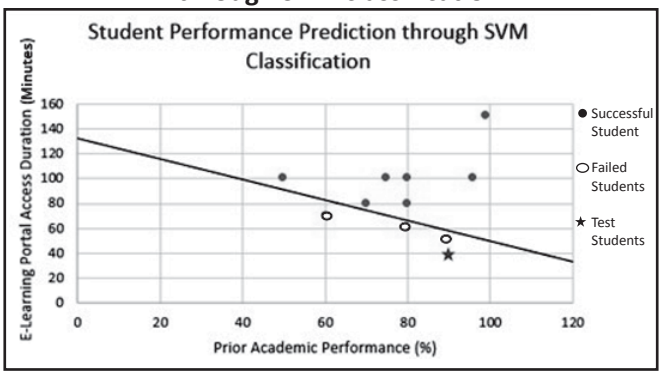
A simple example on use of SVM classification for student performance prediction is demonstrated based on example data mentioned in table V.

TABLE V.
TRAINING DATA FOR SUPPORT VECTOR MACHINE (SVM) CLASSIFICATION

Student	Prior Academic Performance	E-Learning Portal Access Duration (Mins)	Final Performance
1	99	150	PASS
2	96	100	PASS
3	75	100	PASS
4	80	60	FAIL
5	70	80	PASS
6	50	100	PASS
7	80	80	PASS
8	90	50	FAIL
9	60	70	FAIL
10	80	100	PASS

The data on prior academic performance and e-learning portal access for 10 students is given in table V. The final performances of the 10 students in mathematics are also given. The data is displayed in fig. 2. The black line separates the successful and failed students.

Fig. 2. Student performance prediction through SVM classification



Now, it is needed to classify a test student who accessed e-learning portal for 40 minutes. The data is marked as star. As the star falls on the side of the failed students, it is predicted that the student will fail in the mathematics examination.

The studies [18, 39, 45] have used the SVM classification techniques for student performance prediction. It is not as popular a classification method as the other student performance classification methods.

6) Rule-based Classification: Rule-based classification technique is somewhat similar to decision-tree classification. The user needs to define certain rules in terms of IF-THEN statement which is used to predict students; whether they will pass or not.

The studies [17, 51] use rule-based classification techniques for student performance prediction.

7) Logistic Regression: Regression is another approach which is used for student performance prediction. Logistic regression can be used for accurate prediction. The logistic model is as follows -

$$\hat{Y}_i = Logit(A_j * X_{ij}) \quad (2)$$

Where,

$\hat{Y}_i$ = Predicted Pass/Fail Status of the i^{th} student at the end of course (Values can be taken by $\hat{Y}_i$ are 0 = Fail, 1 = Pass)

A_j = Co-efficient of the j^{th} attribute in the equation
X_{ij} = Segment score of the i^{th} student for the j^{th} attribute.
How to calculate segment score?

Suppose the past performances of students is considered as an attribute. The students are grouped into several segments on the basis of their performance. A score is provided to each segment and students grouped in the same segment have the same score. The co-efficient for each attribute in the logistic regression model is calculated on the basis of training data. Then for the test students, the outcomes whether 'Pass' or 'Fail' can be predicted based on the logistic model.

The researchers can utilize attributes and models based on the data available and the classification/regression tool for performing the prediction. The attributes and models mentioned above can be used for any study in the direction of student performance prediction.

IV. SUMMARY

This paper focuses on the research area pertaining to prediction of student performance. The paper first presents a detailed review of the works that have happened on the topic prediction of student performance. Mainly classification based techniques have been used for the purpose of prediction. This paper discusses different classification techniques and how these techniques can be used for student performance prediction. It also suggests a common framework which can be useful in studies related to student performance prediction. It discusses in detail the attributes and methods which can be used in the framework for student performance prediction. This framework can be used for predicting student performance based on the availability of the relevant student data and tools for classification and regression.

REFERENCES

[1] A. Acharya and D. Sinha (2014), "Early prediction of students performance using mach. learning techn.," *Int. J. of Comput. Appl.*, vol.107, no. 1, pp. 37-43, 2014.

[2] Ministry of Edu., Malaysia, "Nat. higher Edu. strategic plan". [Online] Available: http://www.moe.gov.my/v/pelan-pembangunan-pendidikan-malaysia-2013-2025, 2015.

[3] F. Castro, A. Vellido, , À. Nebot, and F. Mugic, "Applying data mining techn. to e-learning problems," Book Chapter – 'Evolution of teaching and learning paradigms in intelligent environment', *Stud. in Computational Intell.,* vol. 62, pp. 183-221.

[4] C. Romero and S. Ventura, "Edu.al data mining: A survey from 1995 to 2005", *Expert Syst. with Appl.*, vol. 33, no. 1, pp. 135–146, 2007. doi: http://dx.doi.org/10.1016/j.eswa.2006.04.005

[5] A. M. Shahiri, W. Husain, and N. A. Rashid, "A Rev. on predicting student's performance using data mining techn.," *Procedia Comput. Sci.*, vol. 72, pp. 414-422, 2015. doi: http://dx.doi.org/10.1016/j.procs.2015.12.157

[6] A. Elbadrawy, A. Polyzou, Z. Ren, M. Sweeney, G. Karypis, and H. Rangwala, "Predicting student performance using personalized analytics," *Comput.*, vol. 49, no. 4, pp. 61-69, 2016. doi: 10.1109/MC.2016.119

[7] Y. Beldarrain, "Distance Edu. trends: Integrating new technologies to foster student interaction and collaboration", *Distance Edu.*, vol. 27, no. 2, pp. 139-153, 2007. DOI: 10.1080/01587910600789498

[8] D. Keegan, *Foundations of Distance Edu.* (2nd ed.). New York: Routledge, 1986.

[9] Z. Ibrahim and D. Rusli, "Predicting students' academic performance: Comparing artificial neural network, decision tree and linear regression", 21st Annu. SAS Malaysia Forum, 5th September, 2007, Kuala Lumpur.

[10] U. Bin Mat, N. Buniyamin, P. M. Arsad, and R. Kassim, "An overview of using academic analytics to predict and improve students' achievement : A proposed proactive intelligent intervention," in *Engineering Edu. (ICEED)*, *2013 IEEE 5th Conf. IEEE*, pp. 126-130. doi: 10.1109/ICEED.2013.6908316

[11] N. Thai-Nghe, "Personalized forecasting student performance", in *Proc. of the 11th IEEE Int. Conf. on Advanced Learning Technologies*, 2011, pp. 412–414, doi: 10.1109/ICALT.2011.130

[12] D. M. D. Angeline, "Assoc. rule generation for student performance analysis using Apriori algorithm, " *The SIJ Trans. on Comput. Sci. Engineering & its Appl.*, vol. 1, no. 1, pp. 12-16, 2013.

[13] C. Romero and S. Ventura, "Educational data mining: A Rev. of the state of the art," IEEE Trans. on Syst., Man, and Cybern., Part C (Appl. and Reviews), vol. 40, no. 6, 2010, pp. 601-618. doi : 10.1109/TSMCC.2010.2053532. URL http://dx.doi.org/10.1109/TSMCC.2010.2053532.

[14] A. Elbadrawy, A. Polyzou, Z. Ren, M. Sweeney, G. Karypis, and H. Rangwala, "Predicting student performance using personalized analytics," *Comput.*, vol. 49, no. 4, pp. 61-69, 2016. doi:

10.1109/MC.2016.119
[15] M. N. Quadri and N. V. Kalyankar, "Drop out feature of student data for academic performance using decision tree techn.," *Global J. of Comput. Sci. and Technol.*, vol. 10, no. 2, 2010.
[16] E. Osmanbegović and M. Suljić, "Data mining approach for predicting student performance," *Econ. Rev. – J. of Econ and Bus.*, vol. 10, no. 1, 2012.
[17] M. M. Abuteir and A. M. El-Halees, "Mining Edu.al data to improve students' performance: a case study," *Int. J. of Inform.,* vol. 2 , no. 2, 2012.
[18] M. Mayilvaganan and D. Kalpanadevi, "Comparison of classification techn. for predicting the performance of students' academic environment," in *Communication and Network Technologies (ICCNT), 2014 Int. Conf., IEEE*, 2014, pp. 113-118. doi: 10.1109/CNT.2014.7062736
[19] S. Natek and M. Zwilling, "Student data mining solution–knowledge Manage. system related to higher Edu. institutions," *Expert Syst. with Appl.*, vol. 41, no. 14, pp. 6400–6407, 2014. doi: http://dx.doi.org/10.1016/j.eswa.2014.04.024
[20] T. M. Christian and M. Ayub, "Exploration of classification using NBTree for predicting students' performance," in *Data and Software Engineering (ICODSE)*, *2014 Int. Conf., IEEE,* 2014, pp. 1-6. doi: 10.1109/ICODSE.2014.7062654
[21] K. F. Li, D. Rusk, and F. Song, "Predicting student academic performance," in *Complex, Intelligent, and Software Intensive Syst.* (CISIS)," 2013 Seventh Int. Conf., 2013, pp. 27-33. doi: 10.1109/CISIS.2013.15
[22] S. T. Jishan, R. I. Rashu, N. Haque, and R. M. Rahman, "Improving accuracy of students final grade prediction model using optimal equal width binning and synthetic minority over-sampling technique," *Decision Analytics*, vol. 2, no. 1, pp. 1–25, 2015. doi: 10.1186/s40165-014-0010-2
[23] S. Parack, Z. Zahid, and F. Merchant, "Application of data mining in Educational databases for predicting academic trends and patterns," in *Technol. Enhanced Edu. (ICTEE), 2012 IEEE Int. Conf.*, *IEEE*, 2012, pp. 1-4. doi: 10.1109/ICTEE.2012.6208617
[24] J. Naren, Elakia, Gayathri, and Aarthi, "Application of data mining in Educational database for predicting behavioural patterns of the students," *Int. J. of Comput. Sci. and Inform. Technologies*, vol. 5, no. 3, 2014, pp. 4649-4652.
[25] D. M. S. A. Kumar (2012), "Appraising the significance of self regulated learning in higher Edu. using neural networks", *Int. J. of Engineering Res. and Develop.,* vol, 1, no. 1, pp. 09-15, 2012.
[26] B. K. P. C. Tucker,and A. Divinsky, "Mining student-generated textual data in moocs and quantifying their effects on student performance and learning outcomes**,"** in *2014 ASEE Annu. Conf., Indianapolis, Indiana,* 2014. [Online] Available: https: //peer.asee.org /22840.
[27] C. Romero, S. Ventura, P. G. Espejo, and C. Hervás (2008), "Data mining algorithms to classify students," in *Edu.al Data Mining*, 2008.
[28] B. Minaei-Bidgoli, D. A. Kashy, G. Kortemeyer, and W. F. Punch, **"**Predicting student performance: An application of data mining methods with the Edu.al web-based system lon-capa," in *Proc. of ASEE/IEEE frontiers in Edu. Conf.*, 2003.
[29] S. Kotsiantis, C. Pierrakeas, and P. Pintelas, "Prediction of student's performance in distance learning using mach. learning techn.", *Appl. Artificial Intell.*, vol. 18, no. 5, pp. 411–426, 2004. doi: http://dx.doi.org/10.1080/08839510490442058
[30] K. McKenzie and R. Schweitzer, "Who succeeds at University? Factors predicting academic performance in first year Australian university students," *Higher Edu. Res. & Develop.*, vol. 20, no. 1, 2001, pp. 21-33.
[31] D. Garcia-Saiz and M. Zorrilla (2011), "Comparing classification methods for predicting distance students' performance," JMLR: Workshop and Conf. Proc. 17, 2nd Workshop on Appl. of Pattern Anal., 2011, pp. 26-32.
[32] V. Oladokun, A. T. Adebanjo, and O. E. Charles-Owaba (2008), "Predicting students academic performance using artificial neural network: A case study of an engineering course," *The Pacific J. of Sci. and Technol.*, vol. 9, no. 1, pp. 72–79, 2008.
[33] V. Ramesh, P. Parkavi, K. Ramar, "Predicting student performance: a statistical and data mining approach," *Int. J. of Comput. Appl.*, vol. 63, no. 8, pp. 35–39, 2013
[34] S. S. Meit, N. J. Borges, B. A. Cubic, and H.R. Seibel, "Personality differences in incoming male and female medical students**,"** Online Submission.
[35] A. Simsek, J. Balaban, "Learning strategies of successful and unsuccessful university students," C*ontemporary Educational Technol.*, vol. 1, no. 1, pp. 36–45, 2010.
[36] T. Mishra, D. Kumar, and S. Gupta, "Mining students' data for prediction performance," in *Proc. of the* 2014 *4th Int. Conf. on Advanced Computing & Communication Technologies*, ACCT '14, IEEE Comput. Soc., Washington, DC, USA, 2014, pp. 255-262. doi:10.1109/ACCT. 2014.105.

[37] C. Romero, M. López, J. Luna, and S. Ventura, "Predicting students' final performance from participation in on-line discussion forums," *Comput. & Edu.*, vol. 68, pp. 458–472, 2013.
[38] N. Thai-Nghe, T. Horváth, and L. Schmidt-Thieme, "Factorization models for forecasting student performance," in: *Educational Data Mining*, 2011.
[39] S. Sembiring, M. Zarlis, D. Hartama, S. Ramliana, and E. Wani, "Prediction of student academic performance by an application of data mining techn.," in: *Int. Conf. on Manage. and Artificial Intell. IPEDR*, vol. 6, 2011, pp. 110–114.
[40] G. Gray, C. McGuinness, and P. Owende, "An application of classification models to predict learner progression in tertiary Edu.," in *Advance Computing Conf. (IACC), 2014 IEEE Int., IEEE*, 2014, pp. 549-554. doi: 10.1109/IAdCC.2014.6779384
[41] I. Hidayah, A.E. Permanasari, and N. Ratwastuti, "Student classification for academic performance prediction using neuro fuzzy in a conventional classroom," in *Inform. Technol. and Elect. Engineering (ICITEE)*, 2013 Int. Conf., IEEE, 2013, pp. 221-225. doi: **:** 10.1109/ICITEED.2013.6676242.
[42] K. Bunkar, U. K. Singh, B. Pandya, and R. Bunkar, "Data mining: Prediction for performance improvement of graduate students using classification," in *Wireless and Opt. Commun. Networks* (WOCN), 2012 Ninth Int. Conf., IEEE, 2012, pp. 1-5. doi: 10.1109/WOCN.2012.6335530.
[43] T. Wang and A. Mitrovic, "Using neural networks to predict student's performance**,"** in *Comput. in Edu., 2002. Proc.. Int. Conf.*, IEEE, 2002, pp. 969-973.
[44] P. M. Arsad, N. Buniyamin, and J.-l. A. Manan (2013), "A neural network students' performance prediction model (NNSPPM)," in *Smart Instrumentation, Measurement and Appl. (ICSIMA)*, 2013 IEEE Int. Conf., IEEE, 2013, pp. 1-5. doi: 10.1109/ICSIMA.2013.6717966
[45] W. Hamalainen and M. Vinni, "Comparison of mach. learning methods for intelligent tutoring Syst.," in *Intelligent Tutoring Syst.*, Springer, 2006, pp. 525-534.
[46] N. Thai-Nghe, "Factorization techn. for predicting student performance," *Educational Recommender Syst. and Technologies: Practices and Challenges*, 2011, pp. 129–153.
[47] N. Thai-Nghe (2010), "Recommender system for predicting student performance," *Proc. of the 1st Workshop on Recommender Syst. for Technol. Enhanced Learning*, vol. 1, pp. 2811–2819.
[48] P. Cortez and A. Silva, "Using data mining to predict secondary school student performance," in *EUROSIS*, 2008, pp. 5 – 12.
[49] M. Wook, Y.H. Yahaya, N. Wahab, M.R. M. Isa, N.F. Awang, and H. Y. Seong, "Predicting NDUM student's academic performance using data mining techn.," in: *Comput. and Elect. Engineering*, *2009*. ICCEE'09. 2nd Int. Conf., vol. 2, IEEE, 2009, pp. 357-361. doi: 10.1109/ICCEE.2009.168
[50] D. M. D. Angeline and I. S. P. James, "Assoc. rule generation using apriori mend algorithm for student's placement," *Int. J. of Emerging Sciences*, vol. 2, no. 1, 2012, pp. 78–86.
[51] F. Ahmad, N. H. Ismail, and A. A. Aziz, "The prediction of students' academic performance using classification data mining techn.**,"** *Appl. Math. Sciences*, vol. 9, 2015, no. 129, pp. 6415-6426.
[52] P. Thakar, A. Mehta, and Manisha (2015), "Performance analysis and prediction in Educational data mining: A Res. travelogue," *Int. J. of Comput. Appl.*, vol. 110, no. 15, 2015.
[53] J. Ruby and K. David, "Predicting the performance of students in higher Edu. using data mining classification algorithms – A case study," *Int. J. for Res. in Appl. Sci. and Eng. Technol.*, vol. 2, no. 11, 2014.
[54] A. A. Aziz, N. H. Ismail, and F. Ahmad, "Mining students' academic performance**",** *J. of Theoretical and Appl. Inform. Technol.*, vol. 53, no. 3, 2013.
[55] P. Patil, "A study of student's academic performance using data mining techn.**,"** *Int. J. of Res. in Comput. Application and Robotics*, vol. 3, no. 9, pp. 59-63, 2015.
[56] D. Reby. S. Lek, I. Dimopoulos, J. Joachim, J. Lauga, and S. Aulagnier, "Artificial neural networks as a classification method in the behavioural Sciences," *Behavioural Processes*, 40, pp. 35-43, 1997. doi: http://dx.doi.org/10.1016/S0376-6357(96)00766-8
[57] F. Nielsen 4i, "Neural networks – Algorithms and Appl.," 2001. [Online] Available: http://www.glyn.dk/download/Synopsis.pdf
[58] S. Ray, "6 easy steps to learn naïve bayes algorithm (with code in Python)," 2015. [Online] Available: https://www.analyticsvidhya.com/blog/2015/09/naive-bayes-explained/
[59] Dr. S. Sayad, "K Nearest Neighbours – Classification". [Online] Available: http://www.saedsayad.com/k_nearest_neighbors.htm
[60] S. Ray, "Understanding support vector mach. algorithms from examples (along with code)," 2015. [Online] Available: https://www.analyticsvidhya.com/blog/2015/10/understaing-support-vector-mach.-example-code/

[61] "Support Vector Machines", Wikipedia, [Online] Available: https://en.wikipedia.org/wiki/Support_vector_mach.

About the Author

Dr. Subhabaha Pal was born in West Bengal, India on the August 21, 1980. He completed B.Sc. (Honors) in Statistics from the University of Calcutta, West Bengal in 2004, M.Sc. (Statistics) from the University of Calcutta, West Bengal, India in 2006, M.B.A. in Risk & Insurance from the ISBMA, Pune in 2009. He obtained Ph.D. degree from the University of Calcutta, West Bengal in 2015 for his research on the topic - 'A Study of the Underwriting Cycle and Growth Pattern in Non-life Insurance Sector in India'. He is a certified ABAP Workbench Consultant.

He started his career as Risk Management Quantification Specialist in Kuwait Petroleum Corporation, Kuwait in 2007. In 2008, he joined Sikkim Manipal University, Gangtok, India as Lecturer in Management and Commerce and also Manipal Global Education Services Pvt. Ltd. as Consultant. He acted as the first head of Sikkim Manipal University e-Learning portal EduNxt during his tenure with Sikkim Manipal University. He took up the responsibility of SAP Technical Lead in Manipal Global Education Services, Bangalore in 2011 and during this tenure, he had different roles viz. SAP Consultant and ABAP Lead. In June 2016, he joined T. A. Pai Management Institute, Manipal as Assistant Professor in Information Technology and Systems Management. Dr. Pal is a valued active member of the SAP Community Network (SCN) on SAP UI5/Mobility topics.

He has authored three books on Quantitative techniques, 29 research papers that have been published in reputed national and international peer-reviewed journals in the last 12 years. His major research areas include educational data mining, textual data mining, financial, and biological data modeling, e-learning, and m-learning.