Student Academic Performance Prediction using Machine Learning

Simran Singhani[1], Shruti Desai[2], Radhika Bailurkar[3], Rashmi Mantri[4]

[1] Professor [2]Student

[1]simran.s@xavierengg.com

[2]shrutidesai108@gmail.com

[3]rbailurkar@gmail.com

[4]rashmi.m.97@gmail.com

Abstract*— Predicting student academic performance has been an important and necessary research topic in many academic institutions. Student performance prediction is an area of concern for the Educational institutions. At the University level learning system where semester wise grading exists, the method or rule adopted to identify the candidates who pass or fail differs depending on various factors such as the course, the department of study and so on. In this study the result of a student is predicted to identify their progress in the next semester using various machine learning techniques. However measuring academic performance of students is challenging since students academic performance hinges on diverse factors. This work focuses to find a way to predict a student's academic performance using the machine learning approach. This can be done by using the previous records of the student rather than applying course dependent formulae to predict the student's final grade. We train model on a relatively large real world students dataset, and the experimental results show the effectiveness of the proposed method which can be applied into academic pre-warning mechanism for students. In this work, various algorithms of machine learning are used such as Logistic Regression, Multiple Linear Regression, Decision trees, Random Forest, Support Vector Machine (SVM) to find out the best model to predict the result of the students.

Keywords— **Machine learning, SVM, Multiclass Logistic regression, Random forest, Naïve Bayes, Decision tree, KNN**

I. INTRODUCTION

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that makes day to day tasks easier. The process of learning begins with observations or data, direct experience, or instruction, in order to look for patterns in data and make better decisions in the future based on the examples that we provide. The primary and the most important aim is to allow the computers learn automatically without human intervention or assistance and adjust actions accordingly.

Machine learning enables us to analyze of massive quantities of data. While it generally delivers faster, more accurate results in order to identify profitable opportunities or dangerous risks, it may also require additional time and resources to train it properly. Often Machine learning is combined with AI to give better and improved results.

II. MACHINE LEARNING TECHNIQUE

The major machine learning techniques are given as follows

- Supervised technique
- Unsupervised technique
- Semi-supervised technique
- Reinforcement technique

Supervised machine learning algorithms can apply what has been learned in the past to new data using labeled examples to predict future events. It tries to find pattern in the given dataset. Starting from the analysis of a known training dataset, the learning algorithm produces an inferred function to make predictions about the output values. The system is able to provide targets for any new input after sufficient training. The learning algorithm can also compare its output with the correct, intended output and find errors in order to modify the model accordingly.

In contrast, unsupervised machine learning algorithms are used when the information used to train is neither classified nor labeled. Unsupervised learning studies how systems can infer a function to describe a hidden structure from unlabeled data. The system doesn't figure out the right output, but it explores the data and can draw inferences from datasets to describe hidden structures from unlabeled data.

Semi-supervised machine learning algorithms fall somewhere in between supervised and unsupervised learning, since they use both labeled and unlabeled data for training – typically a small amount of labeled data and a large amount of unlabeled data.

Usually, semi-supervised learning is chosen when the acquired labeled data requires skilled and relevant resources in order to train it / learn from it. Otherwise, acquiring unlabeled data generally doesn't require additional resources.

Reinforcement machine learning algorithms is a learning method that interacts with its environment by producing actions and discovers errors or rewards. Trial and error search and delayed reward are the most relevant characteristics of reinforcement learning. This method allows machines and software agents to automatically determine the ideal behavior within a specific context in order to maximize its performance. Simple reward feedback is required for the agent to learn which action is best; this is known as the reinforcement signal.

III. Data Collection

Initially the data was collected in raw format in the form of excel sheet. It consisted of students cgpa, grade, extracurricular activities, attendance, feedback and roll no. This data consisted of all the records of students from computer engineering, IT engineering and Electronics engineering branch.

IV. Data Pre-processing

In this we first rounded off the CGPA value to get a single digit CGPA. Then converted the feedback of professors for students in a rating format from 1 to 5 as well as for attendance by specifying the range and assigning numbers in that range.

Table 1

Rating	***Attendance***
1	>80
2	80-60
3	60-40
4	40-20
5	<20

Table 2

Rating	***Feedback***
1	Poor
2	Average
3	Good
4	Excellent
5	Outstanding

Table 3

Grade	***CGPA***
O	>=9
A	8-8.9
B	7-7.9
C	6-6.9
D	<6
F	0(Fail)

Table 4

Rating	***Extra Curricular***
1	Sports
2	Technical
3	Cultural
4	Certificates
5	Other

The excel files were then converted into csv format. This csv files are then loaded into Anaconda Navigator application. Then we divide our data into fragments. Each fragment depicting each semester result of the student.

Then we calculated the mean and median of attendance score , feedback score and extracurricular score.Then we standardize the data to transform it into appropriate range of values. Then we divided the data for training and testing. The 70% of data is trained against 30% of data which is used for testing. After this we have applied different learning models on the data.

In the first step, we randomly divide our available data into two subsets: a training and a test set. Setting test data aside is our work-around for dealing with the imperfections of a non-ideal world, such as limited data and resources, and the inability to collect more data from the generating distribution. Here, the test set consists of

New, unseen data to our algorithm; it's important that we only touch the test set once to make sure we don't introduce any bias when we estimate the generalization accuracy. Typically, we assign 2/3 to the training set, and 1/3 of the data to the test set. Common training/test splits are 60/40, 70/30, 80/20, or even 90/10.

After we set our test samples aside, finally we pick a learning algorithm that we think could be appropriate for the given problem. A quick definition of hyperparameters are the parameters of our learning algorithm, or meta-parameters if you will. And we have to specify these hyperparameter values manually – the learning algorithm doesn't learn them from the training data in contrast to the actual model parameters. Since hyperparameters are not learned during model fitting, we need some sort of "extra procedure" or "external loop" to optimize them separately – this holdout approach is ill-suited for the task. So, for now, we have to go with some fixed hyperparameter values – we could use our intuition or the default parameters of an off-the-shelf algorithm if we are using an existing machine learning library.

Our learning algorithm fit a model in the previous step. The next question is: How "good" is the model that it came up with? The answer of the question goes like, since our learning algorithm hasn't "seen" this test set before, it should give us a pretty unbiased estimate of its performance on new, unseen data! So by taking this test set and use the model to predict the class labels. Then, we take the predicted class labels and compare it to the "ground truth," the correct class labels to estimate its generalization accuracy.

Finally, we have an estimate of how well our model performs on unseen data. So, there is no reason for with-holding it from the algorithm any longer. As a rule of thumb, the model will have a better generalization performance if the algorithms uses more informative data – given that it hasn't reached its capacity, yet.

A. F1Score

F1 Score is used to measure a test's accuracy

F1 Score is the Harmonic Mean between precision and recall. The range for F1 Score is [0, 1]. It tells you how precise your classifier is (how many instances it classifies correctly), as well as how robust it is (it does not miss a significant number of instances).

High precision but lower recall, gives you an extremely accurate, but it then misses a large number of instances that are difficult to classify. The greater the F1 Score, the better is the performance of our model. Mathematically, it can be expressed as :

$$F1 = 2 * \frac{1}{\frac{1}{precision} + \frac{1}{recall}}$$

B. Precision

It is the number of correct positive results divided by the number of positive results predicted by the classifier.

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

C. Recall

It is the number of correct positive results divided by the number of ***all*** relevant samples (all samples that should have been identified as positive).

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

V. Literature Survey

A. *Machine Learning Approach for Tracking and Predicting Student Performance in Degree Programs - Jie Xu, Kyeong Ho Moon and Mihaela van der Schaar*

The paper highlights on Students Academic performance throughout the degree program to give us an insight about it. Many students have different backgrounds at home and different attitudes towards college degrees. Students evolving progress needs to be incorporated into the predction. The objective is to predict the final cumulative GPA of the student at a certain course at the end of the term. For this they have used Ensemble Based Progressive Prediction. The result is given in the form of an histogram indicating Students vs Selected Courses. A latent factor model-based course clustering method was developed to discover relevant courses for constructing base predictors. An ensemble-based progressive prediction architecture was developed to incorporate students' evolving performance into the prediction.

B. *Predicting Students Performance in Educational Data Mining - Bo Guo, Rui Zhang, Guang Xu, Chuangming Shi and Li Yang*

By predicting the academic performance the student using machine learning algorithms students can evaluate their performance and chose their career effectively. The attributes considered for this study are educational attributes such as Background and Demographic data consisting of gender, age family, etc, Past study data, School Assessment data, Study data and personal data like personality, attention,etc. This real data set was collected fro 100 junior high schools in Hubei Province. After preprocessing the dataset approriate algorithm is applied. The algorithm used in this study are Naïve Bayes, SVM, MLP, SVM, SPPN. It is observed by calculating the accuracies of above mentioned algorithm that SPPN has the highest accuraacy in all.

C. *SVM Kernel based Predictive Analytics on Faculty Performance Evaluation - E Deepak, G Sai Pooja, R N S Jyothi, S V Phani Kumar, K V Kishore*

This paper gives us the performance of the faculty members which is evaluated on the basis of different parameters are taken for assessment and predicted by building models using data mining techniques. Techniques used in this paper are Machine Learning, Statistics and Artificial Intelligence. The prime algorithm used in this paper is SVM. A suitable kernel function provides the good learning capacity to SVM and is also used to find out the inner product of two data vectors which are already transformed into a feature space. The parameters used in this algorithm are Faculty Profile, Quality of Teaching , Maintaing relationships, Learing Assessments, Counselling and Mentoring, Administrating Functions, R & D and Organizational Qualities. Performance of various kernels is evaluated with the data and models with SVM-PUKF yields better accuracy by 97.84% when compared with other three standard kernels.

D. *Mining Educational Data to Predict Student's academic Performance using Ensemble Methods- Elaf Abu Amrieh , Thair Hamtini and Ibrahim Aljarah*

The paper focuses on the emerging filed of Educational Data Mining(EDM) in research field. Data Mining methods used in this paper are Decision Tree, Naïve Bayes and Artificial Neural Networks. The students features considered where Demographical features like Nationality, Gender Place of Birth,etc, Academic Background like school levels, grades, semester, total no of absent days,etc, Parents Participations on learning processes like parent answering survey, parent school satisfaction,etc, Behavioral features like Discussion groups, visited resources, raised hand in class, viewing announcements,etc. This paper introduces ensemble methods for students performance. The predictions of ensemble methods is usually more accurate than the traditional method which makes use of only one method. The Ensemble methods used are bagging, boosting and random forest. The results are visualized in the form of bar graphs to make the results more clearer. The accuracy of students predictive model using behavioral features achieved upto 22% improvement comparing to the results when removing such features and it achieved upto 26% accuracy improvement using ensemble methods. After completing the training process, the predictive model is tested using unlabeled newcomer students, that achieved accuracy moere than 80%. This results proves how realistic predictive model is. Lastly, this model can help educators to understand learners, identify weak learners, to improve learning process and trimming down academic failure rates.

E. Performance Prediction of Engineering Students using Decision Trees – R. R. Kabra, R. S. Bichkar

The main focus of this paper is decision tree classifier using supervised learning methods. This paper describes the model that predicts the academic performance of the engineering students in contact education system. If we know in advance which students are likely to fail, the colleges or the teachers can take the necessary actions (like increasing tuition hours per week) to improve the results. Institute's success highly depends upon students' success in that institute. Knowing the reasons of failure of student can help the teachers and administrators to take necessary actions so that the success percentage can be improved. The data is collected from S. G. R. Education Foundation's College of Engineering and Management.
The institute has been started in the year 2008 and is affiliated to University of Pune in Maharashtra, India. This study shows that students past academic performance can be used to create the model using decision tree algorithm that can be used for prediction of student's performance in First Year of engineering exam. From the confusion matrix it is clear that the true positive rate of the model for the FAIL class is 0.907, that means model is successfully identifying the students who are likely to fail. These students can be considered for proper counseling so as to improve their result.

F. Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models - Shaobo Huang, Ning Fang

The paper makes use of four types of mathematical models to predict student academic performance in engineering dynamics - multiple linear regression model, the multilayer perception network model, the radial basis function network model, and the support vector machine model. Data were collected from a total of 323 undergraduate students who took dynamics in four semesters: 128 students in Semester #1, 58 students in Semester #2, 53 students in Semester #3, and 84 students in Semester #4. For each of the 323 students, nine data points (Y, X1, X2, X3, ., X8) were collected, where Y is the score on the dynamics final comprehensive exam

- X1: Cumulative GPA
- X2: Static Grade
- X3 & X4: Calculus 1& 2 grade
- X5: Physics Grade
- X6: Dynamics score in mid term exam 1
- X7: Dynamics score in mid term exam 2
- X8: Dynamics score in mid term exam 3

From the research conducted following is observed.

1. The type of mathematical model has only a slight effect on the average prediction accuracy and the percentage of accurate predictions.
2. The combination of predictor variables has only a slight effect on the average prediction accuracy but a profound effect on the percentage of accurate predictions.

Some of the limitations observed were:

3. the predictive models developed in the present study only take into account eight cognitive (X1- X8)factors. A significant amount of research has suggested that learning is an extremely complex process involving many psychological factors such as learning styles, self-efficacy, achievement goals, motivation, interest, and teaching and learning environment. These psychological factors will be considered in our future modeling work to develop a more accurate predictive model.
4. the grades that a student earned in pre-requisite courses (i.e., X2–X5) might not truly reflect the student's knowledge of those topics.

VI. Algorithms Used

Classification is a form of data analysis that extracts models describing important data classes. Such models also called as classifiers help predict class labels. It is subdivided into two parts: supervised learning and unsupervised learning. Data Classification is a two step process, consisting of a learning step or training phase, where classification model is constructed and a classification step, where the model is used to predict class labels for the given data. In the first step, we try to build a classifier on a predetermined set of data classes. This process is also called as learning step, where a classification algorithm tries to build the classifier by learning from a given data set. The next step, the built classifier is used for classification. Then calculate the accuracy of it.

A. K Nearest Neighbor

The purpose of K Nearest neighour algorithm is to classify new objects based on the attributes and the training data. To classify objects based on the training data that has the closest distance to a new object is based on Euclidean equation formula.

1. Load the data

2. Initialize K = number of nearest neighbors

3. For each example in the data. Calculate the distance between the query example and the current example from the data. Add the distance and the index of the example to an ordered collection
4. Sort the ordered collection of distances and indices from smallest to largest (in ascending order) by the distances
5. Pick the first K entries from the sorted collection

6. Get the labels of the selected K entries

7. If regression, return the mean of the K labels

8. If classification, return the mode of the K labels

B. Naive Bayes

Naïve Bayes algorithm works on Bayesian Classifiers and on Bayes Theorem. It assumes all the attributes are conditionally independent for evaluating class conditional probability. The most important naïve bayes assumption is that each feature makes an independent and equal contribution to the outcome.

P(H/X) = P(X/H)P(H) / P(X)
P(H) = Probability of event before evidence is seen or priori of H.
P(X) = Probability of event before evidence is seen or priori of X.
P(H|X) = is posteriori probability of X. Probability of event after evidence is seen.
P(X|H) = is posteriori probability of H. Probability of event after evidence is seen.

Conditional probability serves as the base for the naïve bayes algorithm.

C. Multiclass Logistic Regression

Instead of y=0,1 we will expand our definition so that y=0,1...n. Basically we re-run binary classification multiple times, once for each class.

algorithm
1. Divide the problem into n+1 binary classification problems (+1 because the index starts at 0?).

2. For each class…

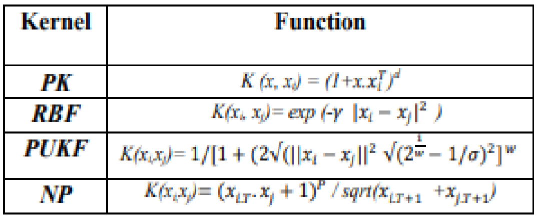
3. Predict the probability the observations are in that single class.

4. prediction = <math>max(probability of the classes)

For each sub-problem, we select one class (YES) and lump all the others into a second class (NO). Then we take the class with the highest predicted value.

D. Support Vector Machine

Vladimir N. Vapnik introduced SVM in 1995. This becomes the most popular supervised learning algorithm for classification and regression. The classification performed by SVM is perfected by constructing an N-dimensional Hyper plane. Hyper plane separation, Maximum-Margin Hyper plane, Soft margin and Kernel Methods are the four basic concepts of SVM. The separation of classes with maximum margin, which is created by hyper planes is the main idea in SVM. Due to its interesting ability, it uses various kernel functions to handle high dimensional data.**.** Kernel function is important in SVM for good and effective classification. A suitable kernel function provides the good learning capacity to SVM and is also used to find out the inner product of two data vectors which are already transformed into a feature space. The use of this kernel function makes the operations feasible and can reduce the computational effort. There are four types of kernel functions are proposed in this paper and used for experimentation. Related to this Study, Pearson VII function is also compared with Radial Basis and Polynomial, which are commonly used kernels and they are shown in the Table.



Kernel	Function
PK	$K(x, x_i) = (1+x.x_i^T)^d$
RBF	$K(x_i, x_j) = exp(-\gamma \ \lvert x_i - x_j \rvert^2 \ )$
PUKF	$K(x_i,x_j) = 1/[1 + (2\sqrt{(\lVert x_i - x_j \rVert^2} \sqrt{(2^{\frac{1}{w}} - 1/\sigma)^2}]^w$
NP	$K(x_i,x_j) = (x_{i,T} . x_j + 1)^P / sqrt(x_{i,T+1} \ + x_{j,T+1})$

Fig 1Support Vector Machine

E. Decision Tree

1. Assign all training instances to the root of the tree. Set curent node to root node.
2. For each attribute
    a. Partition all data instances at the node by the value of the attribute.
    b. Compute the information gain ratio from the partitioning.
3. Identify feature that results in the greatest information gain ratio. Set this feature to be the splitting criterion at the current node.
    a. If the best information gain ratio is 0, tag the current node as a leaf and return.
4. Partition all instances according to attribute value of the best feature.
5. Denote each partition as a child node of the current node.
6. For each child node:
    a. If the child node is "pure" (has instances from only one class) tag it as a leaf and return.
    b. If not set the child node as the current node and recurse to step 2.

7. Random Forest

As the name suggests, Random Forest is a collection of decision trees. The individual decision trees are generated using a random selection of attributes at each node. Each tree depends on the value of random vector sampled independently and with the same distribution for all trees of the forest.

Using the principle of bagging, Random Forests can be generated. One big advantage of random forest is, that it can be used for both classification and regression problems.

1. Randomly select "K" features from total "m" features where k << m
2. Among the "K" features, calculate the node "d" using the best split point
3. Split the node into daughter nodes using the best split
4. Repeat the a to c steps until "l" number of nodes has been reached
5. Build forest by repeating steps a to d for "n" number times to create "n" number of trees

VII. Conclusions

Predicting the student's performance is the most effective way to learners in upgrading their teaching and learning processes. Better results could be drawn with the Random forest algorithm resulting in better prediction of a student cgpa in the next semester. Further analysis is necessary to better understand and improve these results. This method will aid the educational institutions to monitor the performance of students in an effective and systematic way. Lastly, this model can help educators understand learners, identify weak learners, to improve learning processes and bring down academic failure rates. It also can help the administrators to improve the learning system outcomes.

To increase the pergentage of accurate prediction, psychological factors such as learning styles, self-efficacy, achievement goals, motivation, interest, and teaching and learning environment can be considered in future modelling work. The result may prove to be a more realistic predictive model. The accuracy of Logistic Regression is 42% , Linear SVM is 42%, SVC with kernel is 53%, Decision tree is 53% and random forest is 57%. Thus we get the highest accuracy using Random forest model which is 57%, that will be used for further prediction of results of the students.

References

[1] Bo Guo, Rui Zhang, Guang Xu, Chuangming Shi and Li Yang, "Predicting Students Performance in Educational Data Mining",2015 IEEE

[2] E Deepak, G Sai Pooja, R N S Jyothi, S V Phani Kumar, K V Kishore,"SVM Kernel based Predictive Analytics on Faculty Performance Evaluation", Vol 2016

[3] Elaf Abu Amrieh , Thair Hamtini and Ibrahim Aljarah," Mining Educational Data to Predict Student's academic Performance using Ensemble Methods", International Journal of Database Theory and Application, Vol 9,10 September 2016

[4] R. R. Kabra, R. S. Bichkar," Performance Prediction of Engineering Students using Decision Trees", International Journal of Computer Applications, Vol 36, 11 December 2011

[5] Shaobo Huang, Ning Fang," Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models", Computers & Education, Vol 2012 (2002) The IEEE website. [Online]. Available: http://www.ieee.org/

[6] Jaiwen Han,Micheline Kamber," Data Mining Concepts and Techniques"

[7] Jie Xu, Kyeong Ho Moon and Mihaela van der Schaar," A Machine Learning Approach for Tracking and Predicting Student Performance in Degree Programs", IEEE, Vol 2016

[8] Qasem A. Al-Radaideh, Eman Al Nagi,"Using Data Mining techniques to Build a Classification Model for Predicting Employees Performance", International Journal of Advanced Computer Science and Applications, Vol 3 No 2, 2012

[9] C. Romero, S. Ventura," Educational Data Mining: A Review of the State of the Art", IEEE transactions, Vol 40 No 6, 2017