IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

STUDENT PERFORMANCE PREDICTION SYSTEM USING MACHINE LEARNING

Aniket Kalmani[1], Akshay Rakhunde[2], Prachi Badekar[3], Sonakshi Kokate[4], S.L. Yedage[5]

[1-4] Department of Computer Engineering, KJ College of Engineering and Management Research, Pune Maharashtra, India

[5] Professor, Department of Computer Engineering, KJ College of Engineering and Management Research, Maharashtra, India

Abstract - *COVID-19 pandemic has affected various sectors of the global economy including the unexpected closure of schools and colleges. Because of this sudden closure teaching and learning process have gone online which has affected student performance. Student's academic performance needs to be predicted to help an instructor identify struggling students more easily and giving teachers a proactive chance to come up with supplementary resources to learners to improve their chances of increasing their grades. Early indications regarding students' progress help academics to optimize their learning strategies and focus on diverse educational practices to make the learning experience successfully. In this work we created a machinebased learning model to predict a student's educational performance. The developed model relied on the student's previous data and performance in the last stage of the college.*

Key Words*: *Early Grade Prediction , Logistic Regression,
Linear Discriminant Analysis, K-nearest neighbours, Educational Data Mining (EDM) , Fuzzy System , Data Processing , Data Mining, Learning Management System (LMS)

1.INTRODUCTION

The performance of students in their academics is a big turning point in their career. Due to the covid 19 pandemic teaching and learning has been introduce in an online mode, which has affected on some student's performance. Instructor should know the performance of student in order to makes teaching and learning process more effective. To provide a better quality of education. Providing better study materials. Students would about their results in the future and get notified on how to avoid bad results. In the Student Prediction System, we Engineered Machine Learning to adapt the changes in the information and to Find out the different predictions on the given data. By using Machine Learning, we implemented the system which can identify the any student depend upon their qualities and attributes and also on their academic performance. This may get lot of help to the different companies to find out the perfect matched candidates for doing any specific job. Companies can use our algorithm to Find out the right and correct candidate. Teachers can be tension free from identifying which student performance is good or not. Machine Learning is Plays a most important role in our project which helps to perform different operation on the inputted data which we have given to the model for training. Data Cleaning and make the data linear is the very important task in our project. We used the different machine learning libraries like ski-kit learn, xgboost modifier, numpy and visualization libraries like pandas, matplotlib. In this Project, we implement Using different machine learning models and compare their accuracies on given datasets and choose proper model depends upon different parameter and the Outcomes.

1.1 LITERATURE SURVEY

Our research idea is kind a different from others similar researches. H. M. Rafi hasan and AKM Shahariarazad rabbi said in their research that , parents profession, leaving environment, sycological issue are also important behind student performance. But collecting this kind of data is to difficult especially physiological data.

Brigesh Kumar Bardwaj , saurabh pal they research on how to gather data for making the system more intelligent by using data mining process and by using clustering how data can be preprocessed and how results are evalulated are along with dependancies is calculated.

Pratik Nanavati, Abhishek masurkar reaserched in educational data mining (EDM) which uses machine learing

and data mining techniques to explore data from educational settings. In that they describe about how can we identify patterens from the raw data.

Erick Ferenando, Toffic Darvis they studied that in the education system the distance learning is the most vital factor and they used different pricised machine learining model. In that they serves as a metrics of overall system performance.

Abdul Aziz , asaf uddowla golap there is fuzzy logic system introduce by them in that they takes input then carry out calculations and finally provides and output values which ellustrate the process of converting crisp input values into fuzzy values using fuzzy membership functions. In that they given parameter for outcome like high, low, very high, very low.

Michael Kuehn , jared Estad they made a research simple approach to solving the problem more efficiently. They used different data structures like Graphs and Binary Trees to Refinement of the outcome and the correlation between prospective subject for future sense. In that they develop different dynamic notes for classifying data into different categories. They made it in the simplistic and statanderize format.

Gita R.B , S. G Totad they provide a simple interface for maintaince of student information system. They gathred all the information related to stakeholders , faculty and management to deploy that into a web based information management system.

Parnit kaur , Manprit Singh, Gurpreet Singh Josan in this research they work on identifying slow learners among students and displaying it by a predictive data mining model using classification based algorithms. This paper mainly shows that the importance of the prediction and classification based data mining algorithms in the field of eduction and also presents some promising future lines. Mabel Christina In this research paper they tells about different aspects of data mining and tools Like WEKATOOL that actualize a substanscial accumulation of machine learning calculations and is generally utilize in information mining applications.

2. PROPOSED METHODOLOGY

Our proposed methodology started with gathering dataset as similar as researchers made earlier. So we try to collect students informations like School name, sex, age, family size, mother education, father education, study time, family support, mother job, father job, higher education, is it alchohloic or not , health is good or not such parameters we used in our system. This is a classification problem.

2.1 Dataset

We used a new dataset for the proposed model for training we have 1023 sudents of data of previously mentioned attributes. But the main problem is finding real data because academic data is so confidential for both students and the institutes. Using dummy data for this system model is not a wise decision because when we first used dummy data we faced overfitting problem and also the accuracy rate of existing system was very low.

Table -1: Data description table

NO.	NAME	DESCRIPTION
1	G1	Grade given to students.
2	Absenses	No. of Absenses (Numeric value 0-93)
3	Activites	Extra curricular activites (low to hight)
4	Freetime	Freetime In a Day
5	Health	Overall Health Status of students.
6	Age	Students Age
7	Fedu	Father Education
8	Travel Time	How much time taken by students for travel
9	Medu	Mother Education
10	Schools up	Extra educational support

2.2 Correlation

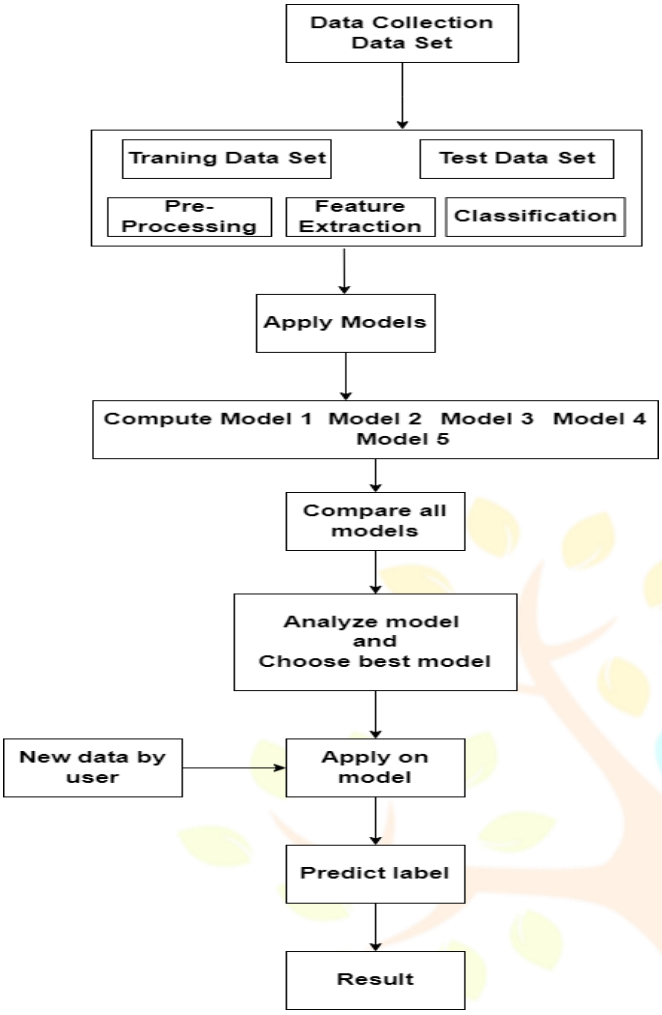
If the value of y increases with the value of x, then we can say that the variables have a positive correlation. If the value of y decreases with the value of x then we can say that the variables have a negative correlation.

Chart -1: Correlation

3. PROPOSED SYSTEM

In our student performance prediction system, the first stage is data collection and preparation. In this phase we gathered different attributes of information which is needed for system. After that we clean the data and by handling if there are any null values are available. After completing preparation, we choose some data from that dataset for testing and training. As per norms we used 80 % for Training dataset and remaining 20 % for testing dataset. After that we Try to Identify some common features along with the given input attributes by using feature extraction technique. To

classify that input attributes in the different categories, we decided to move further with machine learning classification models.



We used 7 type of machine learning models for generating proper results. We used Random Forest Model, Logistics Regression Model, SVM Model, Decision Tree Model, ADA Boost Model, XGBoost Model, K Cross Validation Model are used for implementing model accurately. According to any changes in dataset we try to perform different operations on that dataset and we test that dataset using each and every model as mentioned above. And we compare all the model's performance and accuracy and choose the best appropriate model for further prediction system. For our dataset we executed all the models and compare their performance and accuracy of all the 7 models as we mentioned above. And we found out that there is a best accuracy for XGBOOST Model of 95 %. So, we selected a XG Boost model for the deployment of whole system.

After deciding the best model for prediction system, we deployed that system on the web by using flask. In which the system can take parameters form the users and compare it with the dataset and predict the outcomes in which category the users lie in i.e. Satisfactory, Good, very Good, Poor, Excellent, Failure.

4. ARCHITECTURE OF THE MODEL

Our proposed model is about predicting the student performance. We used Random forest Model, Logistics Regression Model, SVM Model, Decision Tree Model, ADA Boost Model, Boost Model, K Cross Validation Model are used for implementing model accurately

.

4.1 Random Forest

Random forest is a popular machine learning algorithm related to supervised learning methods. It can be used for both classification problems and regression in machine learning. It is based on the concept of team learning, which is the process of combining multiple classifiers to solve a complex problem and improve model performance. Because a random forest combines multiple trees to predict the dataset class, it is possible that some decision trees may be able to predict the correct output and others may not. But together all the trees predict the correct result. Therefore, below are two assumptions about a better Random Forest classifier. There should be some actual values in a dataset function variable for the classifier to predict exact results, not a guess. The predictions from each tree must have a very low correlation Logistics Regression Logistic regression is one of the most popular machine learning algorithms that is subject to the Supervised Learning technique. Predict categorical dependent variable using a given set of independent variables. Logistic regression predicts the outcome of the categorical dependent variable. Therefore, the result must be a categorical or a discrete value. It can be Yes or No, 0 or 1, True or False, and so on, but instead of specifying an exact value as 0 and 1, it gives probabilistic values that lie between 0 and 1.

Logistic regression is very similar to linear regression, except as to how they are used. Linear regression is used to solve regression problems, while logistic regression is used to solve classification problems.

In logistic regression, instead of fitting the regression line, we fit an "S" shaped logistic function that predicts two maximum values (0 or 1).

4.2 SVM Model

Support Vector Machine or SVM is one of the most popular supervised learning algorithms that is used to solve classification and regression problems. However, it is primarily used for classification problems in machine learning.

The goal of the SVM algorithm is to create the best decision line or boundary that can divide n-dimensional space into classes, so that we can easily put a new data point in the correct category in the future. This best decision boundary is called a hyperplane.

SVM selects extreme points / vectors that help in creating the hyperplane. These extreme cases are called support vectors, hence the algorithm is referred to as the support vector machine.

4.3 Decision Tree

A decision tree is a supervised learning technique that can be used for both classification and regression problems, but is most often the preferred method for solving classification problems. It is a tree-structured classifier where the interior nodes represent the characteristics of the dataset, the branches represent the decision rules, and each leaf node represents the result.

There are two nodes in a decision tree which are a decision node and a leaf node. Decision nodes are used to make arbitrary decisions and have many branches, while Leaf nodes are the result of these decisions and do not contain any further branches.

The decisions or the test are performed on the basis of features of the given dataset.

It is a graphical representation that allows you to obtain all possible solutions to a problem / decision based on the given conditions.

It is called a decision tree because, like a tree, it starts with a root node that expands into successive branches and builds up a tree-like structure.

4.4 ADA Boost Model

First of all, AdaBoost is short for Adaptive Boosting. Basically, Ada Boosting was the first truly successful boost algorithm developed for binary classification. It is also the best starting point for understanding amplification. Moreover, modern methods of strengthening rely on AdaBoost, primarily stochastic gradient enhancing machines.

In general, AdaBoost is used with short decision trees. In addition, the first tree is created, and the performance of the tree on each training instance is used. We also use it to weigh how much attention is paid to the next tree. Thus, attention should be paid to each training instance when creating. Consequently, training data that is unpredictable takes more weight. Although, while easy to predict cases have less weight.

4.5XG Boost Model written in C ++. It is a kind of software library that was designed essentially to improve the speed and performance of the model. Recently, it has dominated applied machine learning. XGBoost models dominate many Kaggle competitions.

In this algorithm, decision trees are created in sequential form. Weights play an important role in XGBoost. All independent variables are assigned weights which are then entered into a decision tree that forecasts the results. The weight of the variables falsely predicted by the tree is incremented and these variables are then fed to a second decision tree. These individual classifiers / predictors are then combined to produce a stronger and more precise model. It can work on problems with regression, classification, ranking and user-defined prediction.

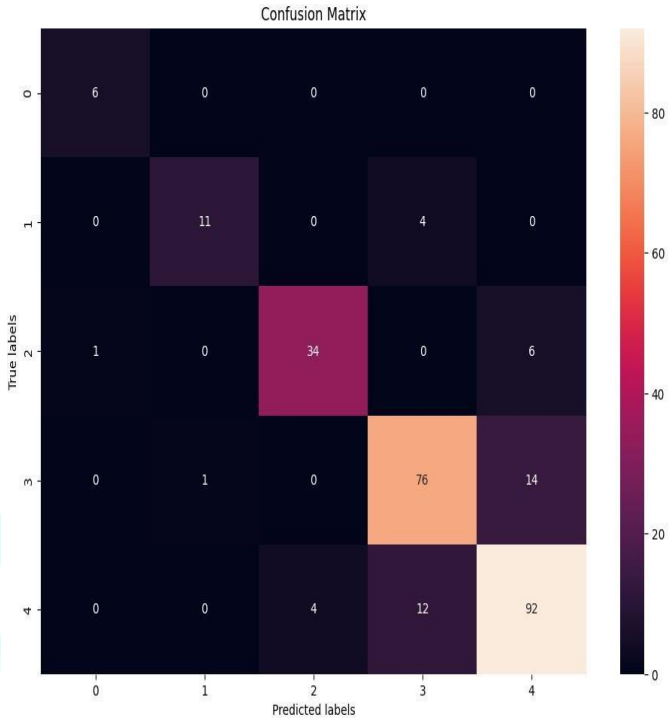
4.6 K Cross Validation

Cross-validation is a technique of validating the performance of a model by training it on a subset of the inputs and testing it on a previously invisible subset of the inputs. It can also be said that it is a technique that checks how a statistical model generates itself into an independent data set.

In machine learning, there is always a need to test the stability of a model. This means that it is only based on the training dataset; we cannot fit our model to the training dataset. For this purpose, we reserve a specific sample of the dataset that was not part of the training set. Then we test our model on this sample before implementation, and the entire process is cross-validated. This is different from the general breakdown of the train test.

5. MATH: - 5.1

Confusion Matrix:



The confusion matrix is the N x N matrix used to evaluate the performance of the classification model, where N is the number of target classes. The matrix compares the actual target values with those predicted by the machine learning model. This gives us a holistic view of how well our classification model works and what kinds of errors it is making.

True Positive (TP) The predicted value matches the actual value. The actual value was positive and the model predicted a positive value. We calculate TP Rate using,

TP = np. diag(cm)

TPR= TP/ (TP + FN) **True Negative (TN)**

The predicted value matches the actual value. The actual value was negative and the model predicted a negative value. We calculate TN Rate using,

TN = cm.sum () - (FP + FN + TP)

TNR= TN/(TN+FP)

False Positive (FP) – Type 1 error

The predicted value was falsely predicted The actual value was negative but the model predicted a positive value

Also known as the **Type 1 error**

We calculate FP Rate using, FP = cm.sum(axis=0) - np. diag(cm) FPR = FP/(FP+TN) False Negative (FN) – Type 2 error

The predicted value was falsely predicted The actual value was positive but the model predicted a negative value Also known as the **Type 2 error**

We calculate FN Rate using, FN = cm.sum(axis=1) - np. diag(cm) FNR = FN/(TP+FN)

Overall Accuracy: -

We calculate overall accuracy using,

ACC= (TP+TN) / (TP+FP+FN+TN)

6. MODEL PERFORMANCE AND RESULT

Our model got 92.64 % using Random Forest Classifier, 92.18 % using SVM Model , 93.40% using Decision Tree,92.34 % of K Cross Validation, 90.59 % using Logistics Regression.d 91.72 % using ADA Boost, 95.56 % using XG Boost Model. Using XGBoost Accuracy on predicting student performance we found our model works so good on predicting performance of the students.

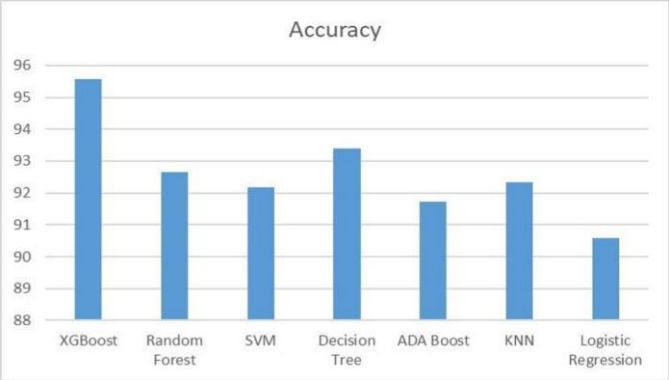


Chart -2: Accuracy

7. CONCLUSIONS

Companies and educational institutions use learning management systems to create and manage lessons, courses, quizzes and other training materials. Student's success needs to be predicted to help an instructor identify academic performance and helps with identifying struggling students more easily and giving teachers a proactive chance to come up with supplementary resources to learners to improve their chances of increasing their grades. It may be difficult for students to learn virtually than in a traditional class hence the student's performance varies due to difference methods of delivering the course materials. Various machine learning models were used to predict student success using the learning management system. We aim to extend the study by collecting more additional features such as encouraging and motivational strategies taken by facilitators and teachers and considering more materials available for students in an Elearning platforms. We also intent to use more interesting and detailed data set to predict student academic performance in our future studies.

REFERENCES

[1] Hina Gull, Madeeha Saqib, Sardar Zafar Iqbal, Saqib Saeed, Improving Learning Experience of Students by Early Prediction of Student Performance using Machine Learning Department of Computer Information Systems, College of Computer Science and Information Technology, Imam Abdulrahman Bin Faisal University, P.O. Box No. 1982, Dammam, SaudiArabia1hgull@iau.edu.sa,2mssaeed@iau.edu.sa,

[2] Pratik Nanavati, Abhishek Masurkar, Chaitanya Shinde, Aaryaman Singh, Prof. Sumitra Sadhukhan5 Student Information System and Performance Prediction in Educational Data Mining Student, Department of Computer Engineering, MCT"s Rajiv Gandhi Institute of Technology, Mumbai, India 1Asst. Professor

Department of Computer Engineering, MCT"s Rajiv Applications (0975 – 888 Gandhi Institute of Technology, Mumbai, India5

[3] Abdul Aziz, Md. Asaf-uddowla Golap and M. M. Hashem Student's Academic Performance Evaluation Method Using Fuzzy Logic System Department of Computer Science and Engineering Khulna Univeresity of Engineering & Technology, Khulna-9203, Bangladesh abdulaziz@cse.kuet.ac.bd, asaf.golap@kuet.ac.bd, hashem@cse.kuet.ac.bd

[4] Dina Fitria Murad,Bambang Dwi Wijanarko ,Erick Fernando,Willy Johan Widjaja Saputra,TaufikDarwis,Lena Prediction Learning Achievement Indicators in Distance Learning Students Information Systems DepartmentBINUS Online Learning Bina Nusantara University Jakarta, 11480, Indonesia dmurad@binus.edu

[5] Machine Learning Algorithm for Student's Performance Prediction H.M. Rafi Hasan ,KM Shahariar Azad Rabby *Dept. of Computer Science and Engineering Daffodil nternational University* Dhaka, Bangladesh azad15-5424@diu.edu.bd

[6] Parneet Kaura,Manpreet Singhb,Gurpreet Singh Classification and prediction based data mining algorithms to predict slow learners in education sector Josanc*aScholar, Department of CSE, Punjab Technical University,Jalandhar 144603,IndiabAssistant Professor, Department CSE&IT, GNDEC, Ludhiana, Punjab, IndiacAssistant Professor, Department of CSE & IT, Punjabi University, Patiala, Punjab, India.*

[7] ELVIRA POPESCU 1, (Member, IEEE), AND FLORIN LEON2 Predicting Academic Performance Based on Learner Traces in a Social Learning

Environment1Computers and Information Technology Department, University of Craiova, 200585 Craiova, Romania2Department of Computer Science and Engineering, Gheorghe Asachi Technical University of Iasi, 700050 Iasi, Romania

[8] R. Sumitha, E.S. Vinothkumar "Prediction of Students Outcome Using Data Mining Techniques"International Journal of Scientific Engineering and Applied Science (IJSEAS) – Volume-2, Issue-6,June 2016

[9] Renza Campagni, Donatella Merlini, Renzo Sprugnoli, Maria Cecilia Verri, "DataMining Models for Student Careers", Science Direct - Expert Systems with Applications 42 (2015) 5508–5521.

[10]S.R.Bharamagoudar, Geeta R.B., S.G.Totad, "Web Based Student Information Management System",International Journal of Advanced Research in Computer and Communication EngineeringVol.2,Issue6,June2013

[11]V.Ramesh Assistant Professor Department of CSA, SCSVMV University Kanchipuram India "Predicting Student Performance A Statistical and Data Mining Approach", International Journal of Computer