Engineering Universe for scientific Research and management
ISSN (Online): 2319-3069
Vol. XIV Issue XI
November 2022

A Literature Review on Student Performance Prediction Using Machine Learning Approach

Mahalya Chopra[1], Mr. Gajendra Singh[2]

M. Tech Scholar, Computer Science and Engineering, SSSUTMS, Sehore, (MP), India[1]
Assistant Professor, Computer Science and Engineering, SSSUTMS, Sehore, (M.P.), India[2]
Email: mahalyachopra@gmail.com[1]

Abstract

The ability to predict student performance is crucial in understanding the student succession rate. Education is the Power, and by forecasting educational success using the right metrics, we will be able to address student weakness at the appropriate moment by utilising accurate pedagogies and techniques. Various machine learning technologies, including supervised, unsupervised, and reinforcement learning, have been created to predict student performance. Using historical observations, machine learning enables us to learn and generate accurate predictions. In this study, we give a literature review on the topic of predicting student achievement using machine learning techniques, along with the benefits and drawbacks of various machine learning approaches.

Keywords: *Student success, Prediction, Machine learning, Supervised learning, Reinforcement learning.*

1. Introduction

Student performance is an essential part in higher learning institutions. This is because one of the criteria for a high quality university is based on its excellent record of academic achievements [1]. There are a lot of definitions on students performance based on the previous literature. Usamah et al. (2013) stated that students performance can be obtained by measuring the learning assessment and co-curriculum [2]. However, most of the studies mentioned about graduation being the measure of student's success. Students are very important part of an educational institute and also for the country. In a crowded class a teacher can't monitor every students. So it becomes very difficult for a teacher to give attention to every student in the class equally. A classroom is filled with a lot of introvert and extrovert students. When we started giving tuition to students we feel the importance of monitoring every student. A teacher should know before a student falls behind. So we decided to research on it how we can predict a student's condition using Artificial Intelligence and a student's previous academic career. [3] Student's academic data is the most important thing for this research, because it indicates most of the things about a student. Like how much he/she studies, what type of subjects he likes and subjects he doesn't like. An IQ test and a physiological test can also help a lot for this research. If we could know how much time he spends for studies and how much he spends for a hobby then we could understand what type of motivation he needs from his teacher. Teachers are the mentors or coaches for a student. We feel that a teacher should know about his/her student's future success probabilities on respected fields or courses. If a teacher is contained with the knowledge of a student's result before the semester ends according to student's previous data. The teacher can provide help and can take the necessary and immediate steps to improve a student's condition. Also, if a student can know the result prediction he or she can also take necessary steps to improve himself or herself. This research is not only about improving the academic result. The main goal of this research is about knowing if the student is learning. A subject's result is depends on many things. Usually every subject have class test marks, attendance mark, assignment marks, presentation Marks, mid-term examination marks, final examination marks. The summation of every test is equal to the result of the subject. It also depends on some other things like the physiological data of the student, IQ score, how much time he or she spend on studying etc. We have used the curriculum of Daffodil International University (DIU). Where every course carry 100 marks and the marks are distributed in class test, attendance, presentation, assignment, mid-term examination and final examination. So, it is a clear indication that, the final examinations performance mostly depends on the other marking attributes like the class test, midterm examination, and mostly on attendance. If a student is regular in the class and can carry a good remark in class test and mid-term, then he/she can perform well in the final examination. But, what if when a student carry good mark in mid-term and has poor performance in attendance or class test! Alternatively, what would happen in a case when he/or she has an excellent presentation skill but cannot perform in the main examination! It is said that 'No one is perfect

in this earth' and it is also applicable for a student. However, we believe that we can boost our perfectness to a maximum level according to our personal capacity. So primarily, we would predict the performance of the final examination in this research according to students past event's report. In Machine Learning, K-Nearest Neighbors, SVC, Decision Tree Classifier, Random Forest Classifier, Gradient Boosting Classifier, Linear Discriminant Analysis algorithm can be applied to predict the future result from some existing attributes of students. In this research, we present the literature survey for the prediction of student performance using machine-learning approach. Also, present the advantages and disadvantages of these techniques.

2. Review of Literature

This section of the research work describes the earlier work done in the field of student performance prediction by the various researchers using different approaches and algorithm for the efficient and accurate analysis of student record.

Ahamed et al. (2017) Students need to have an effective education to take advantage of all the latest tools available. Even with a proper education, they are failing to reap its benefits; reasons involve social, economic and psychological factors a student faces during their adolescence. Our research is directed towards this particular problem of educational effectiveness. We have surveyed a large number of students across different districts in Bangladesh. Pre-processing was done thoroughly; the use of data balancing, dimensionality reduction, discretization and normalization in combinations has allowed us to derive the best model that could predict the academic performance based on different factors during the adolescence [4]. ***Acharya and Sinha (2014)*** presented a set of attributes are first defined for a group of students majoring in Computer Science in some undergraduate colleges in Kolkata. Since the numbers of attributes are reasonably high, feature selection algorithms are applied on the data set to reduce the number of features. Five classes of Machine Learning Algorithm (MLA) are then applied on this data set and it was found that the best results were obtained with the decision tree class of algorithms. It was also found that the prediction results obtained with this model are comparable with other previously developed models [5]. ***Almasri et al. (2019)*** presented in three folds that include the following: (i) providing a thorough analysis about the selected features and their effects on the performance value using statistical analysis techniques, (ii) building and studying the performance of several classifiers from different families of machine learning (ML) techniques, (iii) proposing an ensemble meta-based tree model (EMT) classifier technique for predicting the student performance.(e experimental results show that the EMT as the ensemble technique gained a high accuracy performance reaching 98.5% (or 0.985). In addition, the proposed EMT technique obtains a high performance, which is a superior result compared to the other techniques [6]. ***Lubna Mahmoud Abu Zohair (2019)*** presented to prove the possibility of training and modeling a small dataset size and the feasibility of creating a prediction model with credible accuracy rate. This research explores as well the possibility of identifying the key indicators in the small dataset, which will be utilized in creating the prediction model, using visualization and clustering algorithms. Best indicators were fed into multiple machine learning algorithms to evaluate them for the most accurate model. Among the selected algorithms, the results proved the ability of clustering algorithm in identifying key indicators in small datasets. The main outcomes of this study have proved the efficiency of support vector machine and learning discriminant analysis algorithms in training small dataset size and in producing an acceptable classification's accuracy and reliability test rates [7]. ***Imran et al. (2019)*** proposed a student performance prediction model based on supervised learning decision tree classifier. In addition, an ensemble method is applied to improve the performance of the classifier. Ensemble methods approach is designed to solve classification, prediction problems. This study proves the importance of data preprocessing and algorithms fine tuning tasks to resolve the data quality issues. The experimental dataset used in this work belongs to Alentejo region of Portugal which is obtained from UCI Machine Learning Repository. Three supervised learning algorithms (J48, NNge and MLP) are employed in this study for experimental purposes. The results showed that J48 achieved highest accuracy 95.78% among others [8]. ***E. T. Lau, L. Sun, Q. Yang (2019)*** presented an approach with both conventional statistical analysis and neural network modelling/prediction of students' performance. Conventional statistical evaluations are used to identify the factors that likely affect the students' performance. The neural network is modelled with 11 input variables, two layers of hidden neurons, and one output layer. Levenberg–Marquardt algorithm is employed as the backpropagation training rule. The performance of neural network model is evaluated through the error performance, regression, error histogram, confusion matrix and area under the receiver operating characteristics curve. Overall, the neural network model has achieved a good prediction accuracy of 84.8%, along with limitations [9]. ***Micheline Apolinar –Gotardo (2019)*** proposed decision tree Model which showed that finals had the highest instance and in predicting student performance in the Data Structures and Algorithms subject. It also showed that Finals has the highest factor to receive either of the following remarks: Pass, Failed or Conditional. The model was also able to identify 85.31% accuracy for the attribute Pass, 79.41% accuracy for the attribute Conditional and 91.67% accuracy for the attribute Failed. Further, the Decision Tree Model

likewise revealed that for the student to pass the Data Structures and Algorithms subject they should have a grade higher than 66.12% in Midterms and a grade higher that 72.30% in Finals. The use of the data driven system can be used by institutions to track student performance. Data analysis is a key component to further strengthen their policies and do intervention programs where it is highly needed. Further, for more improvement of this study additional data mining techniques can be applied [10]. ***Afeni et al.(2019)*** proposed a system which can predict the performance of students from their previous academic record using concepts of data mining techniques under Classification. The dataset contains information about students, such as gender, age, SSCE grade, UTME score, post UTME score and grade in students first year. ID3 (Iterative Dichotomiser 3) and C4.5 classification algorithms was applied on the data to predict the academic performance of students in future examinations [11]. ***Duzhin and Gustafsson (2018)*** suggested a machine-learning algorithm that accounts for students' prior knowledge. Their algorithm is based on symbolic regression that uses non-experimental data on previous scores collected by the university as input. Results of their algorithm shows that clickers were a more effective teaching strategy compared to traditional handwritten homework; however, online homework with immediate feedback was found to be even more effective than clickers. [12] ***Utkarsh Verma et al. (2022)*** uses various machine learning (ML) techniques to predict the student's academic performance using the real data collected (comprising the academic history and personal habits of the students). Furthermore, a comparison of ML techniques on different evaluation metrics has been presented. It will assist the students to keep a track of their academic performance and accordingly, manage their study pattern to help them perform well in future.[22] ***J. Dhilipan (2021)*** proposed a prediction system using their 10th, 12th and previous semester marks. The study is evaluated using Binomial logical regression, Decision tree, and Entropy and KNN classifier. In order to attain their higher score, this framework would assist the student to recognize their final grade and improve their academic conduct.[23].

3. Machine Learning Techniques

Machine learning is a paradigm that may refer to learning from past experience (which in this case is previous data) to improve future performance. The sole focus of this field is automatic learning methods. [13] Learning refers to modification or improvement of algorithm based on past "experiences" automatically without any external assistance from human. It is classified into three categories namely supervised, unsupervised and reinforcement learning. The flow diagram of machine learning is shown below:

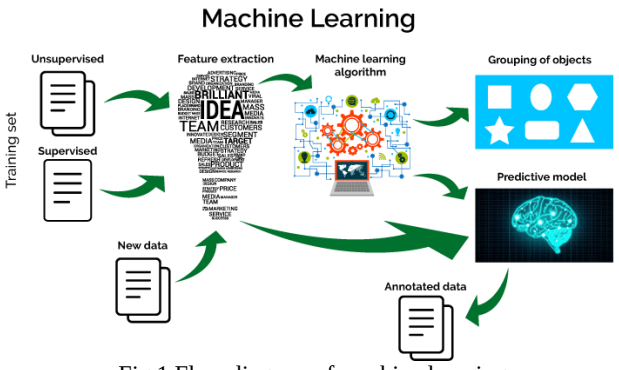

Fig.1 Flow diagram of machine learning

Supervised learning

In Supervised learning we study an objective function that can be used to forecast the values of a separate class feature as accepted or not- accepted. Machine learning algorithm makes predictions on agreed set of example whereas supervised learning algorithms searches for model within the charge labels assigned to data points. This algorithm consists of an ending changeable which is to be predicted starting, a specified set of predictor's i.e. sovereign variables. Using these set of variables, we produce a purpose that map input to wanted outputs. The training procedure continues awaiting the model achieves stage of accurateness on the training data. This complete procedure helps in decrease of spending on physical review for significance and coding. Examples of supervised learning: Neural Networks, Regression, Decision tree, KNN, Logistic Regression, SVM, Naive Bayes etc.

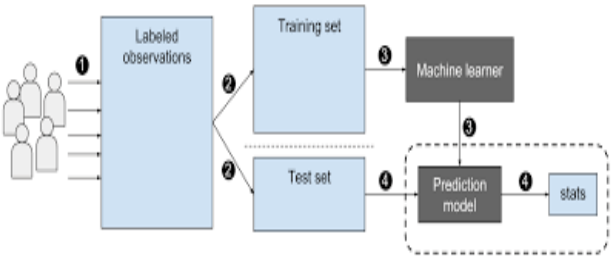

Fig. 2 Supervised Learning

Support Vector Machine

SVM (support vector machine) are supervised learning models with connected learning algorithm that analyze data after which they are used for categorization. Categorization refers to which images are interrelated to which class or data sector set of categories. Learning Classification is considered an instance of supervised learning in machine which refers to task of inferring a meaning from branded training data. Training data in image repossession development can be accurately acknowledged images that are put in a scrupulous class .Where each class belong to dissimilar category of images. In the SVM training algorithms model is build in which the new examples are assigned to one grouping class or other. In this model depiction of examples in

categories are done with clear gaps that are as vast as promising.[14, 17] Classification of data is a widespread commission in machine learning. Machine learning explores the learning and building of algorithms that can be trained from and make predictions on data. Let there are a quantity of descriptions every belong to one of two classes, and our main purpose is to make your mind up to which group a new likeness will be assigned to. Different images are put in (p-dimensional vector) and we necessitate to know whether we can take apart such points with hyper-plane (p-1).There are various hyper-planes which may catalogue the data. But we have to decide the best as per maximum margin of separation. The two main numerical operations: Nonlinear mapping of an contribution patter to superior dimensional feature space. Construction of a most favourable hyper plane for separating the patterns in the advanced dimensional space obtained from first process.

Input: Set of training samples i.e. x1; x2; x3:::;xn and the productivity result is y. In this we can obtain a lot of features as necessary.

Output: Set of weights w, one for each feature, whose linear grouping predicts the value of y. at this time we use the optimization of maximizing the margin to diminish the amount of weights that are non zero to presently a few that keep in touch to the imperative features that matter in deciding the hyper plane .These non zero weights communicate to the support vector.

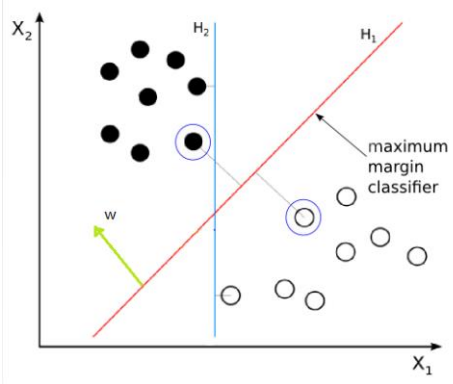

Fig. 3 H1 does not separate the class, H2 does excluding with diminutive margin, H3 separate with maximum margin

Advantages

SVM offers best categorization presentation on the training data. SVM supply more good organization for pure arrangement of the future data. It doesn't make any strong supposition on data. It doesn't greater than fit the data.[20]

Disadvantages

More than one SVM class may recognize or all SVM'S may refuse the data points. In such case data points cannot be classified.

Applications

SVM is frequently used for stock advertise forecasting by different financial institutions. As for comparing qualified concert of the stocks of unlike companies of matching sector. So this relative judgment of stocks helps in supervision speculation based decisions.[21]

Unsupervised Leaning

Learning valuable formation with-out characterized classes, optimization condition, feedback signal, or any former information further than the raw data is referred as unverified learning. In this algorithm, we don't have any objective unpredictable to approximation means here we don't have several label linked with data points or we can speak class label of education data are indefinite [18-19]. These algorithms are used for organizing the data into the group of bunches to explain its arrangement i.e. cluster the data to disclose significant partitions and hierarchies. It creates data look easy and prepared for analysis. Examples: K-means, Fuzzy clustering, Hierarchical clustering. Input data is not labelled and doesn't have a identified result. A model is equipped by deducing construction current in the input data. This may be to remove broad rules. It may during a mathematical procedure to methodically reduce dismissal.

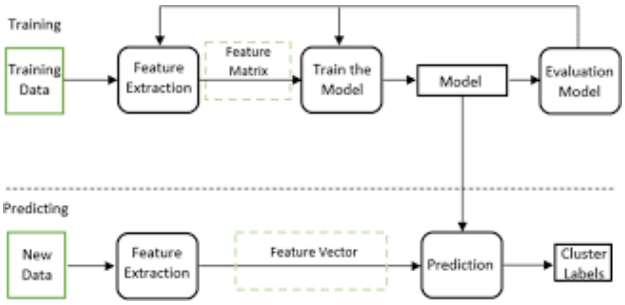

Fig. 4 Unsupervised Learning

K-Mean Algorithm

K-mean is a partitioned - clustering algorithm. It aims to divider the agreed n explanation into K clusters. The mean of every group is establish and the image is positioned in a cluster, whose mean has the least Euclidean detachment with the image attribute vector. Due to the multifaceted distribution of the image data, the k-mean clustering frequently cannot disconnect images with dissimilar concepts well sufficient. Clustering like weakening describes the class of difficulty and the group of methods [13, 16]. Clustering methods are characteristically prepared into two modelling approaches as Centroid-based and Hierarchical. The most accepted amid all is K-mean which essentially comes underneath the grouping of clustering in unconfirmed learning. K-mean is a type of unconfirmed algorithm which solves the clustering difficulty. Its practice follows a uncomplicated and simple way to categorize a specified data set throughout a convinced number of clusters (take as K clusters). Data points inside a cluster are homogeneous and heterogeneous to peer groups. Let the set data points be x1; x2; :::xn where xi1; xi2; ::xir is a vector in a re-valued space X Rr and here r is the number of attributes in the data. This algorithm partitions the

participation data into clusters. Every cluster with its centroid. Here k is specified by user.

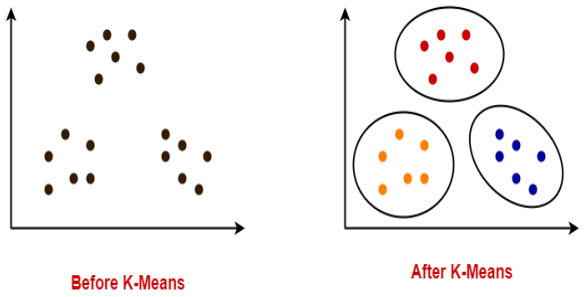

Fig. 5 K-mean clustering [14]

Advantages of K-Mean

- Easy to understand and realize.
- Efficient: Time complexity: (tkn) where n =number of data points, k =number of clusters and t = number of iterations.
- If both k and t are small, it is considered as linear algorithm

Disadvantages of K-Mean

- This algorithm is merely applicable if the signify is defined. The user needs to identify k .
- This algorithm is sensitive to outliers (data points that are very far away from other data points).
- Not apposite for discovering clusters that are not hyper-spheres.

Reinforcement learning

Using this algorithm, the machine is qualified to make explicit decisions. These algorithms prefer an exploit, based on every data point and later learn how superior the decision was in this the machine is showing to an environment where it trains.

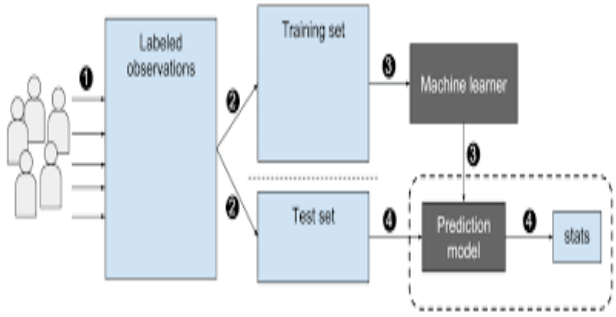

Fig.6 Reinforcement Learning [15]

Two kinds of reinforcement learning methods are:

Positive Reinforcement

It is defined as an event, which occurs because of specific behavior. It increases the strength and the frequency of the behavior and impacts positively on the action taken by the agent. This type of Reinforcement helps you to maximize performance and sustain change for a more extended period. However, too much Reinforcement may lead to over-optimization of state, which can affect the results.

Negative Reinforcement

Negative Reinforcement is defined as strengthening of behavior that occurs because of a negative condition which should have stopped or avoided. It helps you to define the minimum stand of performance. However, the drawback of this method is that it provides enough to meet up the minimum behavior.

Advantages

- This learning model is very similar to the learning of human beings. Hence, it is close to achieving perfection.
- The model can correct the errors occurred during the training process.
- Once the model corrects an error, the chances of occurring the same error are very less.
- It can create the perfect model to solve a particular problem.
- Robots can implement reinforcement-learning algorithms to learn how to walk.

Disadvantages

- Reinforcement learning as a framework is wrong in many different ways, but it is precisely this quality that makes it useful.
- Too much reinforcement learning can lead to an overload of states which can diminish the results.
- Reinforcement learning is not preferable to use for solving simple problems.

The curse of dimensionality limits reinforcement learning heavily for real physical systems.

4. Summary

In order to provide students with a high-quality education and help them enhance their academic performance, early performance prediction of students is crucial. Predicting student achievement normally aids teachers and students in advancing their respective teaching and learning processes. The majority of the research papers assessed in this study employ internal assessment and cumulative grade point average as their datasets. However, they employ supervised learning strategies for prediction, and in this strategy, decision trees and support vector machines are frequently employed. Following the study, it was determined that an ensemble method will be necessary in the future for forecasting and improving student performance.

References

[1] M. of Education Malaysia, National higher education strategic plan (2015). URL http://www.moe.gov.my/v/pelan-pembangunan-pendidikan-malaysia-2013-2025

[2] U. bin Mat, N. Buniyamin, P. M. Arsad, R. Kassim, An overview of using academic analytics to predict and improve students' achievement: A proposed proactive intelligent intervention, in: Engineering Education (ICEED), 2013 IEEE 5th Conference on, IEEE, 2013, pp. 126–130.

[3] Hasan et al. "Machine Learning Algorithm for Student's Performance Prediction", 10th ICCCNT 2019, In proceeding of IEEExplore , pp 1-7.

[4] Ahamed et al., "An intelligent system to predict academic performance based on different factors during adolescence", Journal of Information and Telecommunication, 2017 Vol. 1, No. 2, 155–175.

[5] Anal Acharya, Devadatta Sinha, "Early Prediction of Students Performance using Machine Learning Techniques", International Journal of Computer Applications (0975 – 8887) Volume 107 – No. 1, December 2014.

[6] Almasri et al., "EMT: Ensemble Meta-Based Tree Model for Predicting Student Performance", Hindawi Scientific Programming Volume 2019, Article ID 3610248, 13 pages.

[7] Lubna Mahmoud Abu Zohair, "Prediction of Student's performance by modelling small dataset size", International Journal of Educational Technology in Higher Education (2019) 16:27.

[8] Imran et al., "Student Academic Performance Prediction using Supervised Learning Techniques", iJET – Vol. 14, No. 14, 2019.

[9] E. T. Lau · L. Sun, · Q. Yang "Modelling, prediction and classification of student academic performance using artificial neural networks", SN Applied Sciences (2019) 1:982 | https://doi.org/10.1007/s42452-019-0884-7.

[10] Micheline Apolinar –Gotardo," Using Decision Tree Algorithm to Predict Student Performance", Indian Journal of Science and Technology, Vol. 12(5), DOI: 10.17485/ijst/2019/v12i5/140987, February 2019.

[11] Afeni et al., "Students' Performance Prediction Using Classification Algorithms", Journal of Advances in Mathematics and Computer Science 30(2): 1-9, 2019; Article no.JAMCS.45438 ISSN: 2456-9968.

[12] Duzhin, F. and Gustafsson, A., " Machine learning based app for self-evaluation of teacher-specific instructional style and tools", Education Sciences Journal in MDPI 2018, 8, 7, 1-15.

[13] A. Smola and S. Vishwanathan, Introduction to Machine Learning. United Kingdom at the University Press, Cambridge, October 1, 2010.

[14] Sunpreet Kaur, Sonika Jindal "A Survey on Machine Learning Algorithms", International Journal of Innovative Research in Advanced Engineering (IJIRAE)2016,Issue 11, Volume 3.

[15] [Online]. Available: www.analyticsvidhya.com.

[16] Soumi Ghosh et al , " Comparative Analysis of K-Means and Fuzzy C-Means Algorithms", International Journal of Advanced Computer Science and Applications, Vol. 4, No.4, 2013.

[17] J. Han, M. Kamber and J. Pei, "Data Mining: Concepts and Techniques", 3rd Edition, MK Series, 2012.

[18] Sheeraz Ahmad Peerzada, Jitendra Seethalani, "Machine Learning and Its Implications on Educational Data Base (U-DISE)", Smart Intelligent Computing and Applications, 2020, Springer.

[19] Vaseem Naiyer, Jitendra Sheetlani, Harsh Pratap Singh, "Software Quality Prediction Using Machine Learning Application", Smart Intelligent Computing and Applications, 2020, Springer

[20] V Aruna, et al. "Implementation Of Sequential Pattern Algorithms In Web Usage Mining", Turkish Journal of Physiotherapy and Rehabilitation, 2021 Vol 32, Issue 2.

[21] SI Pasha et al., "A Novel Model Proposal Using Association Rule Based Data Mining Techniques for Indian Stock Market Analysis", Annals of the Romanian Society for Cell Biology, 2021.

[22] Utkarsh Verma; Chetna Garg; Megha Bhushan; Piyush Samant; Ashok Kumar; Arun Negi, "Prediction of students' academic performance using Machine Learning Techniques", International Mobile and Embedded Technology Conference (MECON), 2022, In proceeding of IEEExplore.

[23] J. Dhilipan, N.Vijayalakshmi, S.Suriya, Arockiya Christopher, "Prediction of Students Performance using Machine learning", IOP Conf. Series: Materials Science and Engineering 1055 (2021) 012122.