Dissertation on

Student's Performance Prediction in Online Education System Using Machine Learning Techniques: A Survey

Thesis submitted towards partial fulfillment
of the requirements for the degree of

Master in Multimedia Development

Submitted by

Soumita Das Guha Roy

EXAMINATION ROLL NO.: M4MMD24004B

UNIVERSITY REGISTRATION NO.: 163785 of 2022-2023

Under the guidance of

Prof. Dr. Matangini Chattopadhyay

School of Education Technology

Jadavpur University

Course affiliated to

Faculty of Engineering and Technology

Jadavpur University

Kolkata-700032

India

2024

Master in Multimedia Development
Course affiliated to
Faculty of Engineering and Technology
Jadavpur University
Kolkata, India

CERTIFICATE OF RECOMMENDATION

This is to certify that the thesis entitled **"Student's Performance Prediction in Online Education System Using Machine Learning Techniques: A Survey**" is a bonafide work carried out by **SOUMITA DAS GUHA ROY** under our supervision and guidance for partial fulfillment of the requirements for the degree of Master in Multimedia Development in School of Education Technology, during the academic session 2022-2023.

SUPERVISOR
School of Education Technology
Jadavpur University,
Kolkata-700 032

DIRECTOR
School of Education Technology
Jadavpur University,
Kolkata-700 032

DEAN - FISLM
Jadavpur University,
Kolkata-700 032

Master in Multimedia Development
Course affiliated to
Faculty of Engineering and Technology
Jadavpur University
Kolkata, India

CERTIFICATE OF APPROVAL **

This foregoing thesis is hereby approved as a credible study of an engineering subject carried out and presented in a manner satisfactory to warranty its acceptance as a prerequisite to the degree for which it has been submitted. It is understood that by this approval the undersigned do not endorse or approve any statement made or opinion expressed or conclusion drawn therein but approve the thesis only for purpose for which it has been submitted.

Committee of final examination
For evaluation of Thesis

** Only in case the thesis is approved.

DECLARATION OF ORIGINALITY AND COMPLIANCE OF ACADEMIC ETHICS

I hereby declare that this thesis contains literature survey and original research work by the undersigned candidate, as part of her **Master in Multimedia Development** studies.

All information in this document has been obtained and presented in accordance with academic rules and ethical conduct.

I also declare that, as required by this rule and conduct, I have fully cited and referenced all materials and results that are not original to this work.

NAME: SOUMITA DAS GUHA ROY

EXAMINATION ROLLS NUMBER: M4MMD24004B

THESIS TITLE: STUDENT'S PERFORMANCE PREDICTION IN ONLINE EDUCATION SYSTEM USING MACHINE LEARNING TECHNIQUES: A SURVEY

SIGNATURE: DATE:

Acknowledgement

I am indebted to several individuals in academic circles who have contributed to this thesis completion. The acknowledgement transcends the reality of formality when I would like to express deep gratitude and respect to all those people behind the screen who guided, inspired and helped me with the completion of my thesis work.

I would like to acknowledge my sincere appreciation to my project guide as well as the Director of the School of Education Technology, Prof. Dr. Matangini Chattopadhyay mam for her immense guidance, valuable advice and constructive suggestions throughout this thesis work. She always guided me by answering my doubts and forcing me to read research and work on this topic. I was fully allowed to have the necessary freedom to exercise a thoughtful and scientific approach to the problem.

Also, I would like to thank my family for their encouragement and for providing moral support. There have been a lot of people from whom I've learned a lot of things in the course of my life. Lastly I would like to thank everyone from whom I have gathered knowledge and enriched myself.

Soumita Das Guha Roy
Master in Multimedia Development
School of Education Technology
Jadavpur University
Kolkata: 700 032

Date:

Contents **Pages**

List of Figures

List of Table

List of Abbreviations

The following abbreviations are used in this manuscript:

RF = Random Forest
LG = Logistic Regression
NN = Neural Network
SVM = Support Vector Machine
DT = Decision Tree
NB = Naïve Bayes
KNN = K-nearest neighbor

Executive Summary

Predicting student's performance is one of the most important topics for learning contexts such as schools, colleges, universities. It helps to design effective mechanisms that improve academic results, avoid drop out and find out students at risk. These are benefitted by studying student's activity, behavior, communication with instructor, results. Prediction is processed by handling huge volume of data collected from software tools for technology-enhanced learning. Thus, analyzing and processing these data can give us important information about student's knowledge and relationship between them. This information is the source that feeds promising algorithms and methods able to predict student's performance.

In this study, almost 40 papers are analyzed to show different modern techniques that are widely applied for predicting student's performance, dropout percentage and number of students who are at risk. These techniques and methods are based on various machine learning algorithms.

Chapter 1

1.0 Introduction

1.1 Overview

The global educational system is poorly affected during COVID-19 pandemic. In this period traditional classroom is swiftly transited into online classroom. This shift was too much challenging for remote education. UNESCO reported that by April 2020, school closures affected around 1.6 billion learners worldwide, disrupted establishment of educational practices and adopted online educational environment rapidly [1].

Now across the globe, educational institutes are enhancing their e-learning instructional mechanism in accordance with the aspirations of present day students that are widely tech-savvy using numerous technology tools- computers, tablets, mobiles and internet for educational purpose. Therefore, the pandemic has fast-tracked the digital evolution in educational system, posing vital questions about future teaching and learning methods and it is also crucial to analyze the experience and outcomes of this method to guide post-pandemic educational practices.

There is often a great need to be able to predict future students' performance and behavior in order to improve curriculum design and plan for academic support and guidance on the curriculum offered to the students. This development in the education sector have been truly inspired and effected by using Educational Data Mining (EDM) [2]. EDM is an evolving research domain that consist the concept of 'Data Mining in Education'. It is emerging discipline aimed at using statistical and machine learning methods to analyze large data for a better understanding of students' behavior patterns and their learning environment [3]. Several EDM studies have been discovered different ML techniques to trace variables that significantly influence student's adaptability, performance, dropout, engagement and interaction during online class.

Among those studies, some of them target analyzing variables that generated based on student's online activities [4][5] whereas some studies are also used for predicting students' early dropout based on their interaction logs in online environment, student's academic performance based on their emotional wellbeing and interaction on various e-learning platform, students' adaptability level in

online education system etc. Although data mining leads to knowledge discovery, machine learning algorithms help us to provide the needed tools for these purposes. In literature, the majority of studies try to collect variables and predicting student's performance at the end of the course. The results obtained from those studies are useful to identify the significant variables that influence the student's ability the most.

1.2 Problem Statement

A survey on student's performance prediction in online education system using machine learning techniques.

1.3 Objectives

The purpose of this survey paper is to provide better understanding on

- Students' performance
- Prediction of students' early dropout based on their interaction in online environment
- Predicting the Performance of Students at Risk Using ML
- Prediction of students' adaptability level in online education system
- Optimizing student engagement in edge-based online learning with advanced analytics
- Comparing different resampling methods in predicting students' performance using Machine Learning (ML)
- Predicting at-risk students at different percentages of course length for early intervention using Machine Learning (ML)

Chapter 2

2.0 Machine Learning Techniques

Machine learning is a branch of computer science that focuses to enable AI using given data and algorithms to imitate the way that humans learn and day by day it improves its accuracy. Machine learning uses algorithm to analyze large number of data, identify patterns and then make decisions. The more amounts of data you can provide o system, the better it can learn and improve its accuracy in making decision. The process of building machine learning model involves some steps.

- Defining the problem and success criteria
- Identifying required data
- Determining the model's features and training it.
- Evaluating model's performances based on data
- Deploying the model and monitoring its performances
- Continuously refining the using machine learning model.

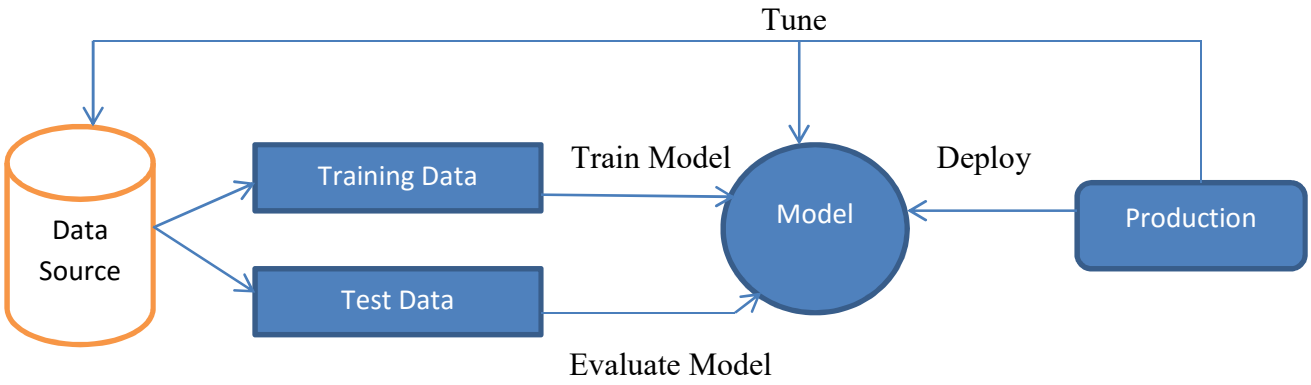

Figure 1 Basic Machine Learning Workflow [8]

Machine learning algorithm such as K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Decision Tree, Random Forest (RF) , Logistic Regression(LR) etc. are usually trained using daily, weekly, or monthly students activity data to find student's adaptability and performance. Deep learning algorithm (DL) is also useful to create predictive model because they can process raw unprocessed data. [6] Described in their journal that Recurrent Neural Network (RNN) algorithm trained on raw data of student's record to predict student's learning performance at

the end of the session and the result of this algorithm showed that RNN gave superior performance compared to other methods. Here some basic concepts of Machine Learning algorithms are described-

KNN

It is the most basic and straightforward classification techniques. This method is used when less or no information about the data. KNN is suitable when reliable parameters to estimate probability is unknown or hard to find out. Parameter K determines how many neighbors will be selected for doing estimation. The result is dependent upon the choice of K. If the value of K is small, the estimation value will tend to poor. Greater value of K cause over-smoothing and miss out on important pattern. The aim is to choose a suitable value of K is to balance out over fitting or under fitting data. The KNN is commonly based on Euclidean distance between a test sample and the trained sample [7]. By default, the KNN () function use Euclidean distance with the equation-

$$D(x, y) = \sqrt{((x2-x1)^2 + (y2-y1)^2)}$$

Where D is the Euclidean distance and x and y are characteristics.

SVM

The SVM is based on plotting data in n dimensional space with n number of features and making a hyper plane to distinguish the classes which are generally used for regression. The dimension of the hyper plane depends on the number of features. For instance, if there are two input features, the hyper plane is simply line and if there are three input features, the hyper plane becomes a 2D plane. As the number of features increases beyond three, the complexity of visualizing the hyper plane also increases.

Decision Tree (DT)

A DT has a tree-shaped structure where each node represent an attributes, each link show a decision making rules and each leaf represent outcomes. It is used for both continuous and discrete data sets. First, DT begins with a node which is called root and then from this node, user split every nodes recursively according to decision tree algorithm based on if-the questions. In the result of decision tree each branch represents a possible scenario and its probable outcomes.

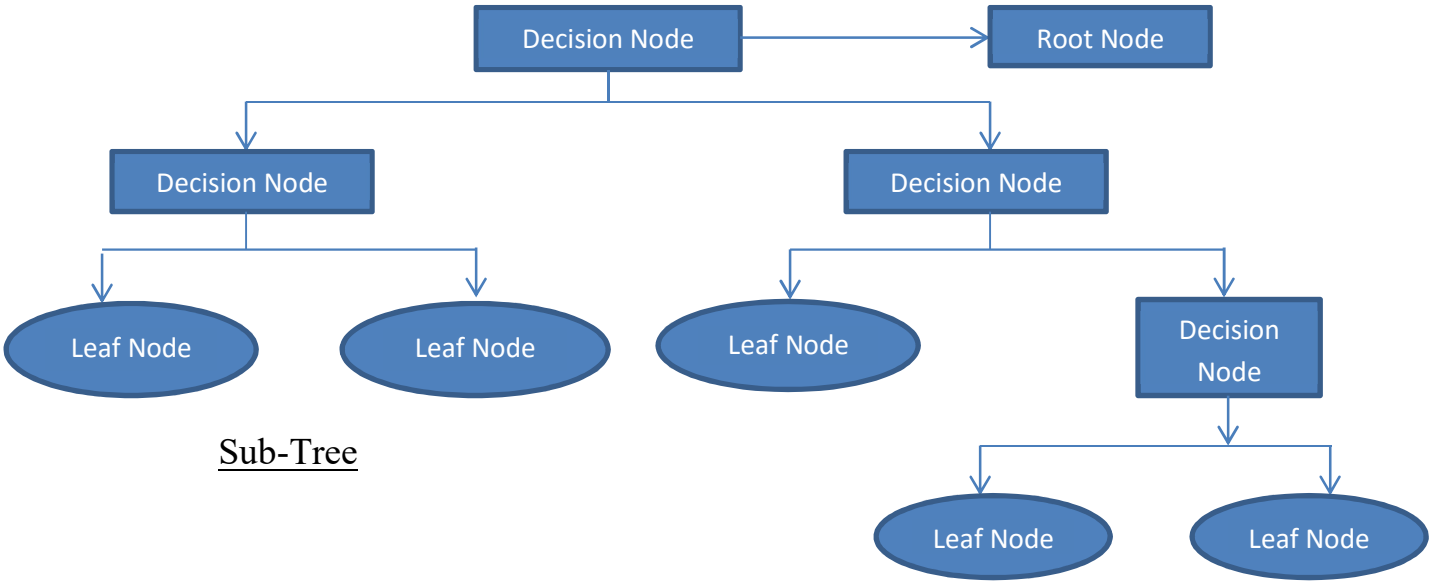

Figure 2 Structure of decision tree algorithm [8]

Random Forest (RF)

As compared to other algorithm, Random Forest is an expert solution for the major problems and falls under the ensemble classifier for weaker model combining to create a powerful one. Ensemble methods are most promising area for research. It consists a set of classifier where prediction are brought together to build new instances. Ensemble learning classifier is an efficient technique to improve prediction accurately and make complex problems into sub-problems. Briefly we can say that numerous decision trees are combined in random forest. To identify an

object having attributes, each tree gives a classification which is called as a vote and the forest has ability to choose the classification with maximum vote.

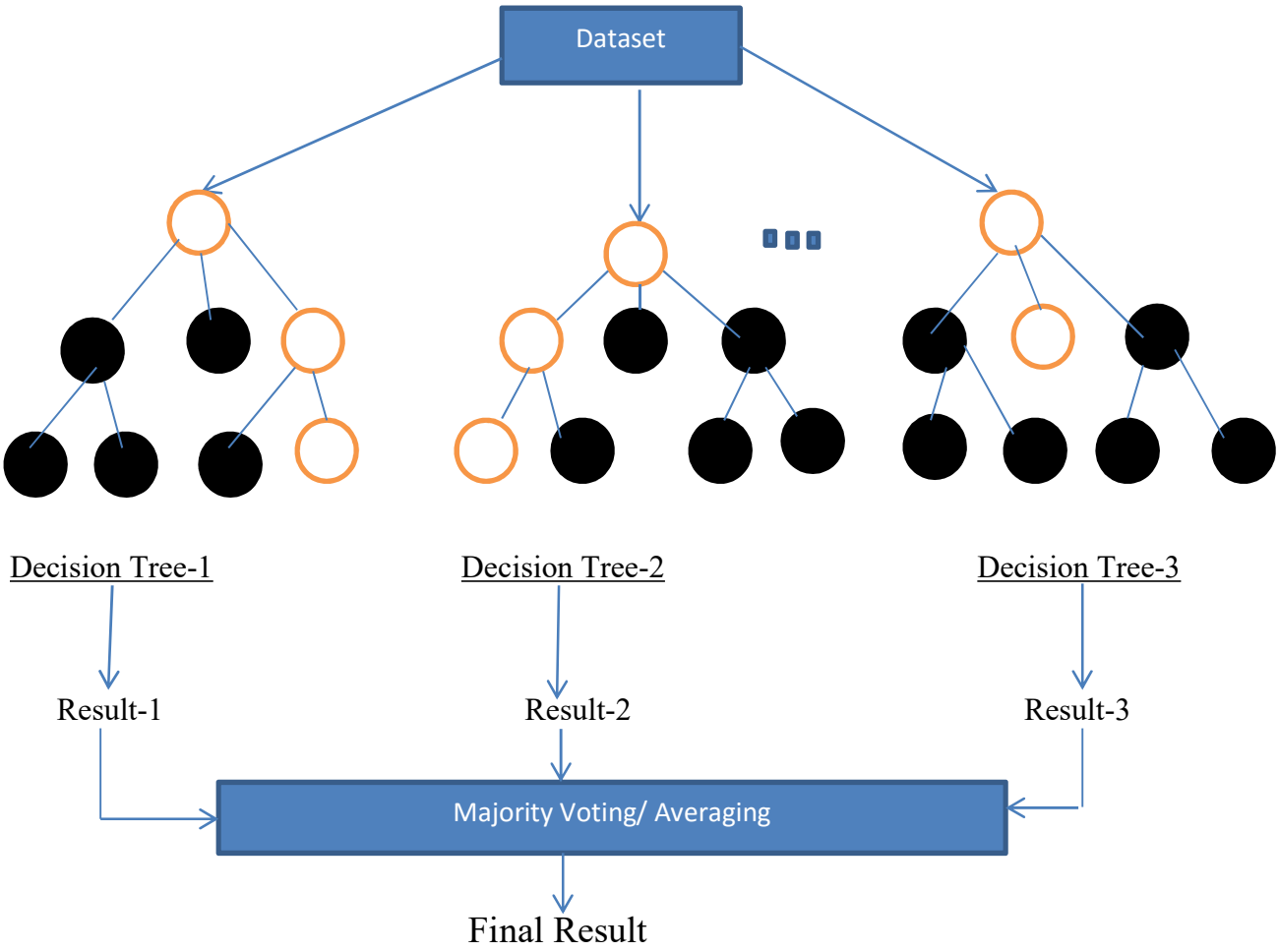

Figure 3 An example of a Random Forest structure considering multiple.[7]

Logistic Regression

Basically linear regression determines the relationship between two or more variables affecting each other and it helps us to make prediction by doing analysis on the variation [9]. Model requiring minimum two independent variables is known as multiple linear models. How independent variables affect the dependent one, the equation describe below [10].

$$Y= \beta0 + \beta1X1 + \beta2X2 + \ldots\ldots + \beta nXn \quad [7]$$

Here X is an independent variable and Y is a dependent variable. β1, β2….βn are the unknown constant and β0 are that kind of constant which can create a line of best fit. Linear regression is not a convenient algorithm for categorical output. Logistic regression was developed to solve the classification problem. The purpose of Logistic Regression is to identify a function from the dataset to calculate the probability where a new data entry belongs to one of the known classes.

Naïve Bayes

It makes use of a simple probabilistic function for classification. It computes a set of probabilities by calculating the frequency and combination of values in a dataset. It allows all attributes to contribute to the final decision equally. Naïve Bayes works well with high-dimensional data and is insensitive to irrelevant data or noises. Its simplicity and low execution time makes it an ideal choice for predictive analysis.

Deep Learning

Deep learning is a sub part of Artificial Neural Network (ANN). ANN is a brain neural system inspired algorithm that consists of layers with connected nodes and is include in ML. It has input and output layers as well as hidden layers. It is used for image recognition, speech recognition, machine translation, medical diagnosis. The idea of hidden layers is important to combine the values in the preceding layer to solve more complicated function of the input. [11] states a research paper where ANN is used to provide customized learning in cyber security professional. This

method is very efficient and effective to solve the problem of ‘one-size-fits-all’ learning. It is challenging to understand raw data for computer. In those cases deep learning decomposes challenging problems into a series of compacted concepts where different layer of predictive model describes each part [12]. Implementation of common machine learning algorithm is usually repetitive for a huge number of trial error methods. Selecting various kind of algorithm will produce different results which is acceptable in some scenario. When other algorithm has some limitations, in that case deep learning is a technique to explain high or low level of abstraction in a given dataset where typical machine learning will fail.

Chapter 3

3.0 Literature Review

E-learning is the delivery of education and all related activities using various electronic mediums like internet. It has provided many benefits such that learner's flexibility and increasing interaction in the form of digital activities by using online learning system. In current situation educational system is going to be digitization. For those changes, student has to take challenges for adapting online education system.

In this work, we have used data and information of research papers published by Journal of Physics, Education Science, Educational Technology and Society, IEEE, Applied Science, ELSEVIER, Research Gate etc. Some keywords are employed to ensure the relevance of the gathered literature, some terms such as online learning, COVID-19, virtual classroom, distance learning, lockdown, e-learning, teaching strategy, student engagement, remote education, educational innovation, post pandemic education. The search period is limited to articles between 1st January, 2019 to 1st march, 2024 to ensure that the collected information and data included in this survey are current.

Yet, some classic references are also added which are published in earlier. These fundamental sources of the data have been included to provide historical context and a comprehensive understanding of the educational evolution. This approach allows us to address the development in online and distance learning, educational technology. This work is structured to explore impact of online learning on students and predicting its' consequences methodically.

Accurate prediction of students' performance and identification of students at risk on online education platform described three approaches;
(i) prediction of educational performance
(ii) identification of students at risk
(iii) prediction of students' early dropout
For above those approaches, most research works show that prediction of students' academic performances is a vital area of interest, with 8 studies undertaken between 2018 and 2023. Identifying students' at risk was also a major research works with 7 research studies undertaken the same period. But the most important studies that were prediction of students' early dropout contained 12 research

studies. Each research is unique and individual in the methodology used and selected attributes are also different from others to determine relevant algorithm applied during classification.

In the research papers, [40] [41] have studied improvement of the online education system. They sowed a remarkable difference between student's performances, satisfaction and benefit of online education system. Those papers showed that 85% of students replied that they learnt more in online education. Mainly researchers tried to improve assessment system for student and teachers and also self and peer assessment for student and teacher.

In [42] [43] researchers has studied that COVID-19 is a problem in worldwide education system. For Corona disease, more than 100 countries closed their schools, colleges and universities. Their study shows that effect of COVID -19 is very horrible and they found several barriers which can hamper interaction between students as well as instructor during lockdown. It affected most in rural area where not have any digital skill, technological background and poor electricity, have network issues. For that reason student's performance got affected and no of drop out increased. So in our work, we tried to find out which machine learning mechanism will work better in the given scenario.

In [14] there is a record 82.26% with Naïve Bayes theorem on a small dataset of 215 students without any balancing technique. But online learning features was not included as an attributes. [15] indicated to remove all missing data from initial 142110 students for predicting performance in online education system and after that with the remaining 72010, accuracy of up to 70% were recorded with Random forest.

[18][19] state that it was possible to predict final student's performance with behavioral supplemented data. The system obtained a weekly ranking of each student's probability of belonging to one of these three levels; high, moderate, low performance. The research concluded that prediction mechanism can improve by exploiting cognitive and non-cognitive characteristic of student.

In this matter[23] [24] stated most important factors for predicting school dropout risk which effected on student's commitment and consistency of their studying process with the help of online resources. Researcher found that higher education

institutions in United States faced a major problem of student consistency, attention, intelligence in the area of science, technology, engineering and mathematics. More than 60% of the dropout happened in first two years.

[25][26] applied a machine learning based technology to analyze students' academic performance to help out instructor as well as learner and educational body that are curious to extract method that can improve individual's educational performance. Their assessment performed the analysis of the past result with individual attribute such that age, demographic distribution, behavior, attitude towards review, family income and family background by implemented various machine learning algorithm. They concluded three significant model i.e. linear regression, supervised learning and deep learning.

[29][30][31] stated an experiment that conducted among the final year university students. They used some unique features like lecture attendance, learning environment, intermediate assessment marks, quiz mark to find some factors as a potential feature for analyzing individual's performance during courses. They implemented DT, KNN, and RF on the self-generated data and claimed 75% accuracy in the prediction of performance with simple small easily accessible data.

Another research paper [32] worked on the same topic and they applied standard analysis tools like IBM-SPSS on the data set collected from the collaboration of institute in between 2012-2019.

[34] examined the factors affecting the students' performance. They classified the students' merit into three classes, fast learner, average learner, slow learner. The data was collected from affiliated colleges to universities where 45 features were extracted from the dataset.

Chapter 4

4.0 Evaluation Methods and Analysis

We have gathered many different research papers based on the topic of student's performances using different machine learning algorithm and we have divided this into some wide groups i.e. student's performance, student's drop out, predicting the student's performance at risk, optimizing student's engagement in edge-based online learning, comparing different resampling methods in predicting students' performance, students' knowledge. In this case student's performance takes the majority of prediction efforts and followed by student dropout, students' performance at risk and other objectives.

4.1 Student's Performance

Evaluating the predictive model is an essential part to determine the accuracy of student's performance. Some important performance metrics to find out the right machine learning techniques are as follows:

Accuracy- It is the ratio of correct predictions to total number of input. It is mostly used to assess the quality of classifier's solution.

$$\text{Accuracy} = \frac{true\ positive + true\ negative}{total\ number\ of\ sample}$$

Precision- It is the number of correct positive results divided by all samples labeled as positive by the algorithm.

$$\text{Precision} = \frac{true\ positive}{true\ positive + fa \quad positive}$$

Recall (Sensitivity) = It is the number of correct positive results divided by all total number of positive instances dataset. Positive instances include true positive and false negative.

$$\text{Recall} = \frac{true\ positive}{true\ positive + false\ negative}$$

F-measure = It refers to a harmonic mean of precision and recall. It has a high recall value with low precision.

$$\text{F-measure} = \frac{2 * precision * recall}{precision + r}$$

Using above metrics researchers has gone through different methodology to evaluate results. [13] Did research that capability of seven algorithms such that SVM, KNN, deep learning and other traditional algorithms. Though having recorded 91 % accuracy with random forest, the attributes didn't work for several online learning tools. A large no. of dataset of 35593, which were taken for predicting student's performance are imbalanced records.

According to [14] there is a record 82.26% with Naïve Bayes theorem on a small dataset of 215 students without any balancing technique. But online learning features was not included as an attributes. [15] indicated to remove all missing data from initial 142110 students for predicting performance in online education system and after that with the remaining 72010, accuracy of up to 70% were recorded with Random forest.

[16] highlights on the importance of emotional and engagement as two very crucial aspects in learning engagement. The researcher discussed about how these two factors are interlinked and how they jointly influence academic platform. Authors also discussed the relationship between instructor-learner discussion, cognitive presence or social interaction between learner and instructor via online medium.

In [17][18] the performance was predicted considering particular first semester courses. Their goal was to represent knowledge as a set of grades from their passed courses and to find similarity among students to predict their performances. In small courses with few students, the research was carried out with large matrices which represented students, grades and assignments. The result was not as much accurate as expected because more information about student was needed. Accuracy is important since it can be very useful to plan intervention in educational system to improve the result of the teaching learning process, saving government resources and educator's time and effort.

[19] state that it was possible to predict final student's performance with behavioral supplemented data. The system obtained a weekly ranking of each student's probability of belonging to one of these three levels; high, moderate, low performance. The research concluded that prediction mechanism can improve by exploiting cognitive and non-cognitive characteristic of student.

Data obtained from previous research was important, even better than applying course dependent formulas for predicting performance [20]. Students' performance and activities on the learning platform provide much feedback needed for performance prediction. The commonly applied algorithms in early prediction using static and dynamic data were; KNN, NB, SVM, DT, RF.

4.2 Prediction of students' early dropout based on their interaction logs in online environment

Various studies focused on the dropout rate to predict the performance of students rather than predicting the likelihood of students' dropout. The author of [21] found that following factors are highly informative to predict school dropout: family history, socioeconomics, status of family, exam result, high school grade etc. It that unbalanced data of classroom was a major problem for prediction. To solve this problem, the authors compared between different data balancing techniques to improve the percentage of accuracy. Among all the techniques, to improve accuracy of prediction, SVM technique achieved the best performance. Nowadays, higher educational institute like college, university are trying to use data collected from database of universities to identify students at risk dropping out [22]. The data which valid the Moodle Engagement Analytics Plugin Learning analysis tool is used here. High dropout rates are very serious matter for online learning process.

The author [23] proposes a technique which is considered as a combination of different classifiers to predict a set of attributes of student's performance over time. They selected student's personal characteristics and academic involvement as an input attributes. They implemented some prediction models using DT, ANN (Artificial neural network). In this matter [24] stated most important factors for predicting school dropout risk which effected on student's commitment and consistency of their studying process with the help of online resources. Researcher found that higher education institutions in United States faced a major problem of student consistency, attention, intelligence in the area of science, technology, engineering and mathematics. More than 60% of the dropout happened in first two years.

In this research paper [25] they tried to detect models in order to predict early dropout through their activity with the course content and the impact on their performance. They surveyed on a dataset taken from an open university at UK which help them to extract important features from the behavior of students in a weekly basis. They tracked on the basis of two components: a.) students' interaction with activities of the course according to the predict probability to set calibrated near 0.50, b.) students' interaction with the activities of the course according to assessment and final scores. In this scenario nonlinear SVM performed better than logistic regression. Due to logistic regression is a linear model, SVM's nonlinearity served dropout prediction task better.

[26] applied a machine learning based technology to analyze students' academic performance to help out instructor as well as learner and educational body that are curious to extract method that can improve individual's educational performance. Their assessment performed the analysis of the past result with individual attribute such that age, demographic distribution, behavior, attitude towards review, family income and family background by implemented various machine learning algorithm. They concluded three significant model i.e. linear regression, supervised learning and deep learning.

[27] compared numerous resampling techniques for predicting students' dropout using two datasets. Firstly they tried to handle data unbalancing problems while computing adequate solution for prediction. They applied many algorithms on balanced data like RF, KNN, ANN, SVM, DT, LG and NB. They researched that combination of RF with balancing techniques provided 77.7% accuracy as the best result.

[28] implemented machine learning algorithms to calculate and enhance the undergraduate students' dropout. They collected the data from the university and measured various factors for assessment like enrolment type, gender, age admission mark, birth city, marital status, nationality, k-12 subjects etc. from these data they found that admission marks was significantly important and single student performed better than the married one and age was also mattered.

[29] applied machine learning based methodology to predict the performance of student. They collected circular and non-circular activity data from institutes and

suggested fuzzy neural network which is based on gradient based error correction and was limited efficient in overall. They compared various methodologies with several NB, KNN, RF, ANN, SVM etc. They did multiple experiments on the mentioned methodologies and claimed 96.04% accuracy in the prediction of students' performance.

[30] stated an experiment that conducted among the final year university students. They used some unique features like lecture attendance, learning environment, intermediate assessment marks, quiz mark to find some factors as a potential feature for analyzing individual's performance during courses. They implemented DT, KNN, and RF on the self-generated data and claimed 75% accuracy in the prediction of performance with simple small easily accessible data.

[31] described a classification of the machine learning algorithms for prediction the students' performance that is NB, RF, DT, ZeroR. They have done a huge no of experiments by using Weka tools and claimed 80% accuracy in their self-generated dataset.

Another research paper [32] worked on the same topic and they applied standard analysis tools like IBM-SPSS on the data set collected from the collaboration of institute in between 2012-2019. Lastly they conclude that if they collected sufficient data, it could be a easier to apply advanced algorithm and can achieved more than 98% accuracy using modern programming tools and developing languages.

Prediction of possible students' dropout is critical to determine necessary remedial. The most used approaches are identifying dropout features, curriculum, retention rate, dropout factors, syllabus, interactivity etc. Students' characteristics were commonly used attributes for research works. The commonly applied algorithms to predict dropout were; DT, SVM, KNN and NB.

4.3 Predicting the Performance of Students at Risk Using ML

Predicting students' performance provide a positive benefit for increasing students' retention rate, effective enrollment management, improving targeted marketing and overall educational effectiveness. To decrease drop out students, intervention

programs in school curriculum is important for those who are at risk of failing in graduation. Timely identification and prioritization of the students at risk are only solution to success of this program. Here some summarized research works are given.

[33] carried out a review that compared the performance of supervised learning classifier, SVM and KNN using the data of universities. The total dataset was converted into numeric forms before analyzing. The dataset was comprised in students' languages and mathematics courses. The proposed classification is evaluated using precision, recall and F-measure. Researchers recorded remarkable improvement in accuracy for both languages which was 82.82% and mathematics which was 82.27% in the student dataset. Comparing with RF, NB, LR, SVM, KNN and DT algorithms, the proposed approach attained accuracy, precision, recall and F-measure scores. In the result, SVM achieved high accuracy among them.

[34] examined the factors affecting the students' performance. They classified the students' merit into three classes, fast learner, average learner, slow learner. The data was collected from affiliated colleges to universities where 45 features were extracted from the dataset. Twelve top features were marked as important features for predicting students' performances.

[35] compared academic features and discussed about nonacademic features like demographic information by applying eight different ML algorithms. They utilized a dataset which was based on Indian which consisted of 6807 students with academic and non-academic features. They applied some oversampling methods to reduce skewness of given dataset they examined that 93.2% F1 score with Decision Tree, 90.3% with Logistic, 91.5% with Multi-Layer perception, 92.4% with Support vector machine,93.8% with Random Forest and 92.355 with voting. They also explained that the academic performance is not only dependent on the academic features but it was a high influential on demographic information as well. They also used nonacademic features for predicting students' performance.

[36] utilized the concept of auto machine learning to enhance accuracy by exploiting features to start a new academic program. They achieved 75.9% accuracy with auto ML with lower false prediction rate. They also employed pre-

admission data and start intervention and consulting session before starting academic improvement, so who need immediate help may survive in the society.

[37] evaluated usage of neural network in the field of EDM with aspect of selection of classification. They utilized several neural networks in various student databases to check their performance. According to them neural network dominated other several algorithms such as Naïve Bayes, Support Vector Machine, Random Forest and Artificial Neural Network to evaluate student performance successfully.

[38] proposed usage of machine learning methods for final grade prediction of educational learners by using historical data. In this method matrix factorization, LR and user item filtering performance were compared with past dataset. Dataset contained with a log file of students obtained when students interacted first time with computer aided system. The result recommended based on matrix factorization low average root mean squared error.

[39] stated that many researchers had utilized the machine learning in the advanced level to predict students' performance effectively. However, they were not able to provide any competent lead for underperforming student. They targeted to beat the limitation and worked for marking the explainable human characteristics that helped to determine the student who will have poor performance academically. They applied SVM, RF, and DT. They got more than 75% accuracy to identify the factors which are enough to spot which student will not be able to pass this term.

4.4 Comparative analysis

The commonly applied algorithms to predict performance and identify risk factor were DT, LR, SVM, NB, and SVM. To identify retention rate, dropout factor, and early prediction were commonly used attributes to find out dropout features.

Based on data collected to predict student's performances and behavior, the most widely used technique was supervised learning as it provides accurate and reliable result. In particular, SVM was the most used by the authors and provided most accurate predictions. In addition to SVM, DT, NB, RF have also been well studied algorithmic proposal that generated good result.

Table 1: Prediction of student performance, student's early dropout and student's performance at risk

Approach	Attributes	Algorithms	Count	References
Performance prediction	Socio demographic, Teaching performance, Student's activity	NB,SVM,DT ANN	6	[12,14,15,17,18 20]
Identification of students at risk	At-risk of failing to graduate, Early prediction, Final GPA results	SVM,RF,NB, KNN, Deep learning	7	[1,5,6,11,13, 16,19]
Predict the difficulties of learning platform	Difficulties on e-learning system	ANN,SVM,DT	4	[7,8,9,10]
Features for dropout prediction	Student's personal characteristics and academic performance	DT, SVM, RF, NB	10	[21,22,23,24 25,26,29,30, 31,32]
Dropout factors	Evaluation of useful models	ANN,SVM	2	[27,28]
Retention Rate	Freshman student	DT, ANN	2	[26,30]

Chapter 5

5.0 Conclusions and Future Scopes

Following the mentioned literature review published between 2018 to 2023 the following conclusion are made-

- Most studies used less data than required one to train the machine learning methods. That's why it is a fact that to get accurate results we need massive data.
- Only few studies have focused on data balancing which is considered important for obtaining better performances.
- The temporal nature of features which is used for at risk and drop out of students' prediction has not been studied to its potential.
- It was also seen that prediction of student's at risk and drop out for on campus students utilized the data set with a minimum no of instances. Machine learning algorithms trained on small data set never give satisfactory results.
- Less attention has been paid to feature tasks such that students' demography, academic and e-learning interaction session logs where types of feature can influence prediction of performances.
- Most of the studies used SVM, DT, KNN, NB and only few have investigated with the help of deep learning.

Future research will focus more on developing efficient method to practically deploy ML based performance prediction methodology and provide automatic needed remedial actions to help the students as early as possible.

References

1. Simone Nomie Sato, Emilia Conde Moreno, Alejandro Rubio-Zarapuz, Athanasios A. Dalamitros , " Navigating the New Normal: Adapting Online and Distance Learning in the Post-Pandemic Era", Educ. Sci. 2024, 14, 19. https://doi.org/10.3390/edusci14010019, pp. 1-25 , 2024.
2. Muhammad Adnan, Asad Habib, Jawad Ashraf, Shafaq Mussadiq, Arsalan Ali Raza, Muhammad Abid, Maryam Bashir and Sana Ullah khan, " Predicting at risk students at different percentages of course length for early intervention using Machine Learning Models", Institute of Computing, Kohat University of Science and Technology, Pakistan, IEEE ACCESS, , Volume 9, pp.7519-7539, January 5, 2021.
3. Kudratdeep Aulakh, Rajendra Kumar Roul, Manisha Kaushal, " E-Learning enhancement through educational data mining with Covid -19 outbreak period in backdrop: a review", International Journal of Educational Development(101), ELESVIER, 19 May, 2023.
4. S.M.Jayaprakash, E.W.Moody, E.J.M.Lauria, J.R.Regan and J.D.Baron, "Early alert of academically at risk students: An open source analytics initiative,"J. Learn.Analytics, vol. 1, no.1, pp. 6-47, May, 2014.
5. S.Palmer, "Modeling engineering student academic performing using academic analytics, Int.J.Eng.Edu" vol.29, no.1, pp. 132-138, 2013.
6. G.Korosi and R.Farkas," Mooc performance prediction by deep learning from raw clickstream ,pp. 474-485, data," in Proc. Int. Conf. Adv. Compute. Data Sci. Valletta, Malta: Springer, 2020.
7. Yudish Teshal Badal, Roopesh Kevin Sungkur, " Predictive modeling and analytics od student's grades using machine learning algorithms", Education and Information Technology(2022) 28:3027-3057 part of Springer nature, https://doi.org/10.1007/s10639-022-11299-8, pp. 3028-3057, 8th September, 2022.
8. Hafeez, M.A, Rashid, Tariq.H.Abideen, Alotaibi, S.S and Sinky, M.H, "Performance improvement of decision tree: a robust classifier using Tabu search algorithm. Applied Science, 11(15), 6728, 2021.
9. Uyanik, G.K., & Guler.N, "A study on Multiple Liner Regression Analysis", Procedia-Social and Behavioral Sciences,[online] 106, Available

at: https://www.sciencedirect.com/science/article/pii/S1877042813046429[Accessed: 20 November 2021].

10. Petrovski, A., Petruseva, S.,Zileska,P.V., " Multiple liner regression model for predicting bidding price.", Technics Technologies Education Management, 10(1), 386-393, 2015.
11. Sungkur, R.K., Maharaj, " A review of intelligent techniques for implementing smart learning environment.", Proceeding of the 3rd Inernational Conference on Communication, Devices and Computing.Lecture Notes in Electrical Engineering, vol 851. Springer, Singapore. https://doi.org/10.1007/978-981-16-9154-669, 2022.
12. Di Fransco, G & Santurro, M, " Machine learning, artificial neural networks and social research", Quality & Quantity, 55(3), pp. 1007-1025, 2020.
13. Adnan, M., Habib, A., Ashraf, Mussadiq, Raza, bashir, M.,& Khan, S.U, "Predicting at Risk Students at Different Percentages of Course Length for Early Intervention using Machine Learning Models, IEEE Access, [online] pp. 7519-7539, 9, 10th December,2021.
14. Ko, C.Y & Leu, F.Y, "Examining successful attributes for undergraduate students by applying machine learning techniques", IEEE Transaction on Education, 64(1), pp. 50-57, 2021.
15. Tarik, a., Aissa, H& Yousef, F, "Artificial intelligence and machine learning to predict student performance during COVID-19", Procedia Computer Science, 184, pp. 835-840, 2021.
16. Liu, S., Liu, Z, Peng, X., &Yang,Z., " Automated detection of emotional and cognitive engagement in MOOC discussions to predict learning achievement", Computers & Education, Vol-181, ISSN 104461, pp. 0360-1315, 2022.
17. Bydzovska, H., "Student performance prediction using collaborating filtering methods." Lect.Notes Comput.Sci, 9112, pp. 550-553, 2019.
18. Bydzovska, H., "Are collaborative filtering method suitable for student performance prediction?" Lect.Notes Comput.Sci, 9273, pp. 425-430, 2019.
19. Villagrá-Arnedo, C.; Gallego-Duran, F.; Compan-Rosique, P.; Llorens-Largo, F.; Molina-Carmona, R., "Predicting academic performance from behavioral and learning data", Int. J. Des. Nat. Ecodyn, pp. 239-249, 2020.

20. Shanthini, A.; Vinodhini, G.; Chandrasekaran, R., "Predicting students' academic performance in the University using Meta decision tree classifiers." J. Comput. Sci, pp. 654–662., 2019.
21. Nandeshwar, A., Menzies, T,Nelson,A. " Learning patterns of university student retention.", Expert Syst.Appl.38, pp. 14984-14996, 2020.
22. Liu,D., Richards, D, Froissard, C, Atif, " A validating the effectiveness of the moodle engagement analytics plugin to predict student academic performance", In proceeding of the 21st Americas Conference on Information Systems(AMCIS 2018) , Fajardo, Puerto Rico, 13th August, 2021.
23. Dewan,M, Lin,F.;Wen,D,Kingshuk, " Predicting Dropout-prone student in e-learning education system", in proceeding of the 2017 IEEE 12th Intl Conference on Ubiquitous Intelligent and Computing and 2017 IEEE 12th Intl Conference on Autonomic and Tusted Computing and 2017 IEEE 15th Intl Conference on Scalable Computing and Communication and Its Associated Workshop (UIC-ATC-ScalCom), Beijing, China, pp. 1735-1740, 14th August, 2019.
24. Saqr,M.; Fors, U, Tedre,M, " How learning analytics can early predict under-achieving students in a blended medical education course.", Med tech, pp. 757-767, 2022.
25. Ahmed A. Mubarak, Han Cao & Weizhen Zhang, " Prediction of students early dropout based on their interaction logs in online learning environment", interactive Learning Environment, vol. 30, No. 8, pp. 1414-1433, 2022.
26. Oydeji, AO, Salami, A.M, Folorunsho, O, Abolade, "Analysis and prediction of student academic performance using machine learning", JITCE (J. Inf. Technol. Comput. Eng), pp. 10-15, 2020.
27. Ghorbani, R, Ghousi, R, "Comparing different resampling methods in predicting students' performance using machine learning technique." IEEE Access, pp. 67899-67911, 2020.
28. Ahusban, Shatnawi,Yasin,M.B, Hmeidi, " Measuring and enhancing the performance of undergraduate student using machine learning tools", In proceeding of the 2020 11th International Conference on Information and Communication Systems (ICICS), Copenhagen, Denmark, 24-26 August, pp. 261-265, 2020.

29. Hussain, Talpur, N, Afeb, M.U., "A novel met heuristic approach to optimization of Neuro Fuzzy system for students' performance prediction." J. Soft comput. Data Min, pp. 1-9, 2020.
30. Wakelam, E, Jeffries, A.Davey, Sun Y, "The potential for students' performance prediction in small cohorts with minimal available attribute", Br. J. Educ. Technol, pp. 347-370, 2020.
31. Walia, N.; Kumar, M.; Nayar, N.; Mehta, G. Student's Academic "Performance Prediction in Academic using Data Mining Techniques." In Proceedings of the International Conference on Innovative Computing & Communications (ICICC); Springer: Berlin/Heidelberg, Germany, 2020.
32. Gafarov, F.; Rudneva, Y.B.; Sharifov, U.Y.; Trofimova, A.; Bormotov, P., " Analysis of Students' Academic Performance by Using Machine Learning Tools.", In proceedings of the International Scientific Conference "Digitalization of Education: History, Trends and Prospects" (DETP 2020), Yekaterinburg, Russia,; Atlantis Press: Paris, France, pp. 574–579, 23–24 April 2020.
33. Hussain, M.; Zhu, W.; Zhang, W.; Abidi, S.; Ali, S., "Using machine learning to predict student difficulties from learning session data." Artif. Intell. Rev., pp. 381–407, 2020.
34. Kaviyarasi, R.; Balasubramanian, T, "Exploring the high potential factors that affect students. Acad. Perform.", Int. J. Educ. Manag. Eng, 2019.
35. Aggarwal, D.; Mittal, S.; Bali, V. "Significance of Non-Academic Parameters for Predicting Student Performance Using Ensemble Learning Techniques." Int. J. Syst. Dyn. Appl. (IJSDA), pp. 38–49, 2021.
36. Zeineddine, H.; Braendle, U.; Farah, A., "Enhancing prediction of student success: Automated machine learning approach", Comput. Electr. Eng., 2021.
37. OuahiMariame, S.K. Feature Engineering, "Mining for Predicting Student Success based on Interaction with the Virtual Learning Environment using Artificial Neural Network." ,Ann. Rom. Soc. Cell Biol., pp. 12734–12746, 2021.
38. Buenaño-Fernández, D.; Gil, D.; Luján-Mora, S., "Application of machine learning in predicting performance for computer engineering students:" A case study on Sustainability, In Proceedings of the International Conference on Innovative Computing & Communications (ICICC); Springer, 2019.

39. Reddy, P.; Reddy, R. “Student Performance Analyser Using Supervised Learning Algorithms. 2021”, Available online: https: //easychair.org/publications/preprint/QhZK, 2021.
40. Md. Mahmudul Hasan Suzan, Nishat Ahmed Samrin, Ai Amin Biswas, Md. Aktaruzzaman Pramanik, “Students’ adaptability level prediction in online educational system using machine learning algorithms”, 12th International Conference on Computing Communication and Networking Technology, IEEE, 2021.
41. R.Afrouz and B.R.Crisp, “ Online education in social work, effectiveness, benefits and challenges :A scoping review ”, Australian Social Work, vol. 74, no.1,pp. 55-67, 2021 .
42. E.M.Onyma, N.C.Eucheria,A. Sharma and A.O.Alsayed, “ Impact Of Coronavirus Pandemic on Education”, Journal of Education and practice, vol. 11, no. 13, pp. 108-121, 2020.
43. R.E. Baticulon, J.J.Sy , J.C.B Reyes, “Barriers to online learning in the time of covid-19 : a national survey”, Medical Science educator, pp. 1-12, 2021.