**Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)**
ISSN：1671-5497
E-Publication Online Open Access
Vol: 41 Issue: 12-2022
DOI 10.17605/OSF.IO/Z7FQA

# STUDENT PERFORMANCE PREDICTION IN E-LEARNING ENVIRONMENT USING MACHINE LEARNING

## USAMA JABBAR
Department of Computer Science, The Superior University Lahore, Pakistan.
Email: Phcs-f21-015@superior.edu.pk

## SHAHID MEHMOOD
Department of Computer Science, University of Central Punhab Lahore, Pakistan.
Email: shahid.mehmood01@ucp.edu.pk

## SABA RAMZAN
Sharif College of Engineering and Technology Lahore, Pakistan. Email: sabaramzan01@sharif.edu.pk

## MUHAMMAD WASEEM IQBAL
PhD, Associate Professor, Department of Software Engineering, Superior University Lahore, 54000, Pakistan. Email: waseem.iqbal@superior.edu.pk

## SABAH ARIF
Lecturer, Department of Computer Science, Superior University Lahore, Pakistan.
Email: sabah.arif@superior.edu.pk

## MUHAMMAD ZUBAIR*
Department of Computer Science, Thal University, Bakhar, Pakistan.
Corresponding author Email: Sheikhzubair333@gmail.com

## MUHAMMAD JAVAID IQBAL
Department of Computer Sciecne and Information Technology, the Superior University Lahore, Pakistan.
Email: javaid.iqbal@superior.edu.pk

**Abstract:**

Automatic prediction of students in e-learning environments is challenging due to a large amount of education data. This issue is addressed by educational data mining. Education data mining is essential in analyzing and discovering a meaningful pattern in education data. These entire meaningful patterns are helpful in student performance improvement and in making a decision. We proposed a student performance prediction model using data mining techniques in this study. We used an enhanced educational dataset with student behavioural features. These features show the student learner's interactivity with the e-learning system. We applied various classification algorithms to our data set: decision tree, random forest, support vector, and Logistic Regression. We first trained the model using this classification algorithm, then obtained results of the models were compared. The model is evaluated based on performance assessment metrics, Accuracy, precision, and f measure and recall values. The experiment results show that random forest performs better than other tested models. Moreover, we demonstrate different visualization, which is helpful in the interpretation of the result. Furthermore, this study reveals that student behavior features strongly correlate with student performance.

**Keywords:** Machine Learning; E-Learning; Prediction; Student Performance.

**Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)**
ISSN：1671-5497
E-Publication Online Open Access
Vol: 41 Issue: 12-2022
DOI 10.17605/OSF.IO/Z7FQA

## 1. INTRODUCTION

Teams Now day's, massive innovations in information and communication technologies have changed how teachers teach, and students learn. E-learning has various advantages, such as delivering online course content and not requiring students to sit physically in the classroom. E-learning is not as much of expensive as compared to the conventional mode of learning. E-learning is becoming a widely used method of teaching and an essential part of developing web-based education systems. Due to the brunt of the COVID-19 pandemic, e-learning has been extensively used internationally due to its high temporal and spatial flexibility, the low threshold of learning new things and using excellent learning resources. Yet, This method makes it difficult for teachers to understand their students' learning status and questions were raised about the quality of e-learning. Studying the prediction of learning performance serves as a foundation to modify their teaching strategies for pupils who might have issues, in studies by forecast student performance basis on the future Exams, eliminate the risk of student failure [1].

There is various kind of e-learning platforms, such as massive open online courses(MOOC), digital electronic education system (DEED), which provide course material, assignment task and conduct examination. In these online learning environments, students can enroll themselves and learn the courses at any location. Moreover, they provide a platform for controlling administration activities, student management and education services. In the learning management system, the teacher can manage and provide learning material and examine student performance in education tasks like every teacher. Student activity is stored in the system [2, 3].

Conversely, the data stored in LMS are raw and give us no meaningful information or exit measurement approach [1, 2, 3]. Through depth analysis of log data, the teacher can evaluate the whole learning performance of students and what kind of learning issues they might have. In the literature review, many studies have found that information about student learning activity stored in LMS can use to predict future student performance [4]. Recognizing the students with learning problems and taking essential measures for them is one of the significant steps to improve their future Performance. However, besides the advantages of online learning, most institutions face the challenge of analyzing student log data to predict student learning progress. They used a traditional statistical method to analyze the student log data and achieved less accuracy. Data mining has excellent potential to explore and get important information in extensive data. Today data mining is widely used in different industrial applications like health, finance, the education sector, bioinformatics etc. Data mining techniques are commonly used in education to analyze and classify student learning difficulties. In education, the data mining technique is getting more attention to assess large datasets and extract valuable knowledge that supports the prediction model.

**Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)**
ISSN：1671-5497
E-Publication Online Open Access
Vol: 41 Issue: 12-2022
DOI 10.17605/OSF.IO/Z7FQA

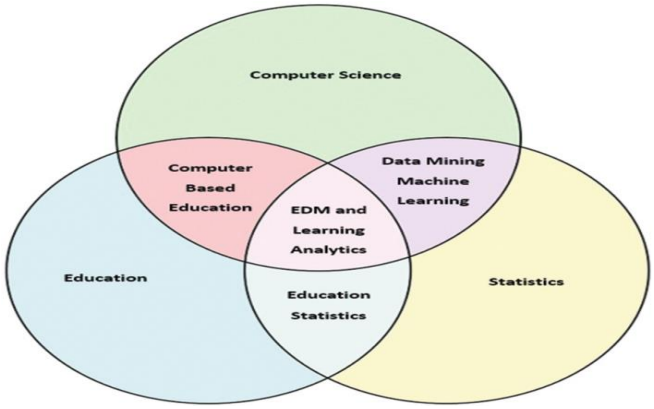


**Figure 1: Different interrelated field**

Educational data mining (EDM) is a promising domain that concentrates on mining extensive educational data and understanding student behaviour that supports classifying students who have learning problems in early phases. Educational data mining (EDM) provides an efficient prediction technique to analyze massive academic databases and extract useful information [2]. In many past, researchers used data mining to employ educational records to predict students' current Performance in the registered course and future grades, pass or fail. The literature reveals that most researchers used student past history data and demographic information to predict student performance.

In many situations, these predictive models fail because they include historical data or non-academic factors such as (economic status, gender, and age) not beneficial to the teacher or student [1]. Student progress prediction in sessions and classes such as academic activities, assignments, quizzes and exams provides in-depth course information. To achieve this objective, researchers employ machine learning techniques on data gather from conventional universities' education and online learning systems. In the traditional university education system, researchers used conventional statistical methods such as survival analysis and linear mixed model (learning activity, weather condition, heart rate) as input features to predict student performance [5]. In literature, many researchers compare different machine learning classifiers and find out which algorithm achieves the best result in terms of accuracy. Abu saa investigates a study that discovers the most effective classifier to predict student academic performance. In his research, they used personal (gender, age, learning activity) and social as input features [6].

Huang et al. investigate a study to forecast student academic performance in engineering courses using machine learning methods. In his research, the author used an SVM classifier that used course grades over all semesters as input and final scores as output. This study reveals that a multi-linear regression classifier to predict student performance achieves the best result in terms of accuracy [7]. Kaur et al. analysed the student's difficulties in a mathematics course. In this study, the author used an ad boost

**Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)**
ISSN：1671-5497
E-Publication Online Open Access
Vol: 41 Issue: 12-2022
DOI 10.17605/OSF.IO/Z7FQA

classifier that used test grades as input feature and then compared the result between existing models and the proposed solution [8]. Most of the study focuses on predicting student performance in upcoming sessions. Dataset used collected from traditional universities. The researcher used input features such as previous history data, grade point average (GPA), and course grade and non-academic input features such as (gender and age). However, these input features did not imitate a positive effect on student performance.

Many studies have found from the literature that forecast student grade pass or fail. But these studies negatively affect student performance if they already know neither his future Performance nor help the teacher improve his teaching method. In this study, we addressed these problems and proposed an education decision support system. In this system, we used the learning management system (LMS) log data to forecast students' progress in the registered course. The log data set consists (of assignment marks, quiz marks, and attendance) as an input feature.

Our primary objective is as follows:

- How can we predict student performance using LMS log data.
- To investigate which machine learning classification algorithm show better accuracy to classify student performance.

The outline of this paper is as follows: section 2 conduct literature review in the context of educational data mining and the current research gap. Section 3 proposes prediction model, section 4 performs an experiment, and section 5 and 6 concludes and describes future direction.

## 2. LITERATURE REVIEW

Education quality act a vital part in the progress of any country. However, the quantity of educational data rises daily due to the advancement of the admission system, learning management system and e-learning. Therefore, the most challenging tasks is to analyze and predict student progress in large amounts of educational data. Data mining study is an innovative area of study put into practice in different areas of education. It acquires significant importance in the industry because it helps analysts analyze the data to identify important facts, patterns, and meaningful information that helps make future decisions. Education data mining is essential in exploring and discovering meaningful ways to education data. The purpose of our study is to incorporate data mining techniques in education. It is helpful to forecast future student progress in terms of grades, examination results and study engagement.

Vasile Paul Bresfelean stated that Understanding student-related issues and student performance is a complex task. It collects students' past and present data from institutions and surveys. After collecting the student educational data, they apply a data mining clustering technique to the dataset. The author illustrates their experiments based on data clustering and classification technique to predict the student profile

**Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)**
ISSN：1671-5497
E-Publication Online Open Access
Vol: 41 Issue: 12-2022
DOI 10.17605/OSF.IO/Z7FQA

success or failure [9]. In this research, the author suggested student performance predictive model using supervised learning techniques. In his proposed model, the ensemble-based method is used in the classifier to minimize classification and prediction problems. In this study, experiment data sets of the Alentejo region of Portugal were taken from the machine learning repository. This study uses three different supervised machine learning algorithms multilayer perceptron, j48 nonnested generalization in the Experiment dataset. The experiment findings demonstrate that the ensemble prediction model attain high accuracy compared to the existing classifier [10].

Pauziah Mohd Arsad stated that early predicting student performance is significant for both academic society and students' future decisions. They used the data of bachelor and diploma-level students. There are two variables used in the data one input variable, which is student results in fundamental subjects, and cumulative grade point average CGPA is the output variable. They employed artificial neural network techniques on collected data to predict the students' Performance. He evaluated the attaintment of the proposed model using mean square error and coefficient correlation. Experiment results show that the first three semesters strongly influence the final semester CGPA. They ignore the learner's behaviour towards learning and cognitive difficulty while studying learning material [11].

According to Raza Abidi et al., sustainable learning and teaching is the main issue in creating arenewable education system. They evaluate the Performance of confused students who failed to perform their educational task given by their teacher on an intelligent tutoring system. They used the data from Assignment, which is an online tutoring platform. They have used seven different machine learning classifiers to employ on students' data to predict the vulnerable student group who failed to complete their Assignment. Evaluating and measuring the Performance of machine learning models based on precision, accuracy and f-measure [12].

Sadiq Hussain analysis the student performance and predict the overall future academic grades. Measuring academic Performance is critical because variables are correlated and none correlated. He developed a regression model using deep learning based on student datasets. He collected the dataset from three different colleges in India and applied the proposed model to it. Evaluating and measuring the Performance of deep learning models based on mean absolutes error (MAE) coefficient correlation and mean square error (MSE) [13]. Yahia Bashar and Hitham Alhussian are exploring the various machine learning classifers used to forecast students' academic Performance. However, it is critical to achieving better and more reliable predictions in massive amounts of educational data. The use of machine learning and deep learning model significantly impact data analysis and provide an accurate forecast. In this study, the author reviews various machine learning classes, comparesres them and identifies which classifier is best in terms of precise student performance predictions. His findings reveal that neural network is the best Performance for achieving the highest prediction accuracy [14].

**Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)**
ISSN：1671-5497
E-Publication Online Open Access
Vol: 41 Issue: 12-2022
DOI 10.17605/OSF.IO/Z7FQA

Ouafae El Aissaoui1 and Yasser El Alami El Madani stated that predicting student performance is helpful in academic institutions for making student educational policies. However, identifying the factor that affects student performance is the most challenging task. In contrast, accurate student performance prediction will benefit teacher improve their learning and teaching method. Machine learning and data mining are two techniques which are frequently used to forecast student performance in educational data mining. He proposed a methodology to build a student prediction model using multiple linear Regression (MLA), which is a supervised machine learning technique. There are three stages in his proposed methodology first stage is preprocessing and analysing the data attribute. Then the second phase consists of selecting the more significant point in the experiment. In the third step, we applied multiple linear Regression on details and compared their Performance using k fold validation technique [15].Pavlo D. Antonenko analyzes the web server log data to forecast student learning behaviour using cluster data mining approaches in an online learning environment. Clustering is the method of making a group based on similar data objects. They employed hierarchical clustering, and non-hierarchical k means clustering on web server log data to analyze the learner behaviour. At the same time, connect in attempting problem-solving activities in an online learning environment [16].

In this study, the author proposed a methodology using data mining techniques based on pre-admission tests to predict candidate performance that assist universities in admission decision-making. In this study, the author used the data set of a public sector university in Saudi Arabia. This dataset consistsof2039 students enrolled in computer science programs from 2016 to 2019 and were used for experimental purposes that validate his proposed methodology. The experiment result show candidate's early university test score and general aptitude test score most appropriately predict the candidate's future Performance. These two test scores gain more weight in the admission system. In this study, the author used an artificial neural network in his proposed predicted classifier. He compares the accuracy comparison between the existing classification technique and the proposed predicted classifier based on precision, accuracy and f-measure. His findings reveal that neural network is the best Performance for achieving the highest prediction accuracy [18].

In this study, the author proposed a recommender system using three different machine learning classifiers to predict the action required based on their academic outcomes. This study addresses the problem related to multi-class labels, classification and reliable accuracy. A case study of Saudi Arabia college was formulated to demonstrate the proposed predictive classification model. The experiment findings demonstrate that the ensemble prediction model accomplish high accuracy compared to the existing classifier [19].

Luis A. Mendoza-Chate explores machine-learning techniques to predict students' academic Performance. However, it is critical to achieving better and more reliable predictions in massive amounts of educational data. The use of machine learning and deep learning model significantly impact data analysis and provide an accurate forecast.

**Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)**
ISSN：1671-5497
E-Publication Online Open Access
Vol: 41 Issue: 12-2022
DOI 10.17605/OSF.IO/Z7FQA

In this study, the author reviews various machine learning classifiers, compares them and identifies which classifier is best in terms of precise student performance prediction. His findings reveal that neural network is the best Performance for achieving the highest prediction accuracy [20].

**Table 1: Survey of Different Existing Approaches**

| Author | Research problem | Technique | Approach |
|---|---|---|---|
| [1] | Student Performance | Clustering, Classification | J48,k-mean, Decision tree |
| [2] | academic Performance | Classification | J48,multilayer perceptron |
| [3] | Students Performance prediction | Classification | artificial neural network, Regression |
| [4] | Students Assignment task Performance prediction | Classification | SVM, Naïve Bayes,Decision tree, neural network classifier, J48 |
| [5] | Students' Performance and future grade prediction | Regression, Classification | Multiple regression Model, Random forest model |
| [6] | Student academic prediction | Classification | Artificial neural network |
| [7] | Students Performance prediction | Regression, Classification | Multiple regression Model |
| [8] | Predict the student learning behaviour | Clustering, Classification | hierarchical clustering and non-hierarchical k mean |
| [9] | Students' Performance prediction in an e learning environment | Classification and Regression | Regression, Decision tree. |
| [10] | Predict the learner's needs | Classification and information retrieval technique | content-based filtering and collaborative filtering |
| [11] | Predict the Selection of the Right Undergraduate Major | Classification | Random forest, SVM, Decision tree |
| [12] | Students' Performance prediction in university admission | Classification | Artificial neural network |
| [13] | Student course improvement prediction | Multi-class classification | K-nearest neighbour, |
| [14] | Identification of usage of E-learning | Quantitative | Empirically Evaluation |
| [15] | Predicting the Academic Performance | Classification | SVM, Naïve Bayes,Decision tree, neural network classifier |

## 3. METHODOLOGY

In this section, we present the detail of our proposed approach. We used a variety of data mining approaches to forecast student progress in online learning environment. The primary step of our proposed method is shown in figure 1, and details are given below.

### Machine learning techniques

Several types of data mining approaches have been used to develop prediction models. Four kinds of DM classifiers are tested, as this study's prediction model is illustrated below.

**Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)**
ISSN：1671-5497
E-Publication Online Open Access
Vol: 41 Issue: 12-2022
DOI 10.17605/OSF.IO/Z7FQA

## Support Vector Machine

A Support Vector Machine (SVM) is a kind of supervised ML classifier and is extensively used in regression and classification problems. The SVM can be used for both numeric and Categorical data. It is a linear classifier that can be used to predict the class of an observation using a training set. It gives highly accurate results and a lesser amount of overfitting. The ideal hyperplane improves the border between the two classes where the support vector is located. It converts the original input data into high-dimension feature space. It then discovers the hyperplane in a new dimension by separating the two classes. The hyperplane finds by an algorithm using a support vector.

$$W.X + B = 0 \quad (1)$$

Where W donates a weight vector W= {w1, w2, w3… wn} B is frequently referred to as bias, and N is several features, if there are two attributes, A1 and A2, in the training set, then x1 and x2 are the values of the features, hence the point that is separating hyperplane class A1 can be written as:

$$W.X + B > 0$$

Hence the point that separating hyperplane class A2

$$W.X + B < 0$$

## Decision tree

A decision tree (DT) classifier is extensively used in classification problems. It works in parallel and serial form according to massive data. In DT, construction does not need any factor or domain knowledge. The DT is widely useful in the decision-making process. It has a tree-like structure in which ovals represent leaves and a rectangle represents internal nodes. Each internal node consists of two children. Each node produces child nodes until subgroup is not divided further and generates meaningful information. The internal node and leaf represent the dataset features and the values of attributes. The terminal node is the down most nodes representing the target output value.

## Random forest Classifier

Random forest is a kind of supervised ML classifier and is extensively used in regression and classification problems. It works equally on categorical and numeric variables. One of the essential aspects of random forest classifiers is to generate several individual decision trees based on the given data and display the final target output based on the majority voting. The creation of decision trees does ‘not examine all attributes. It randomly selects the root node in the decision tree structure [25].

**Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)**
ISSN：1671-5497
E-Publication Online Open Access
Vol: 41 Issue: 12-2022
DOI 10.17605/OSF.IO/Z7FQA

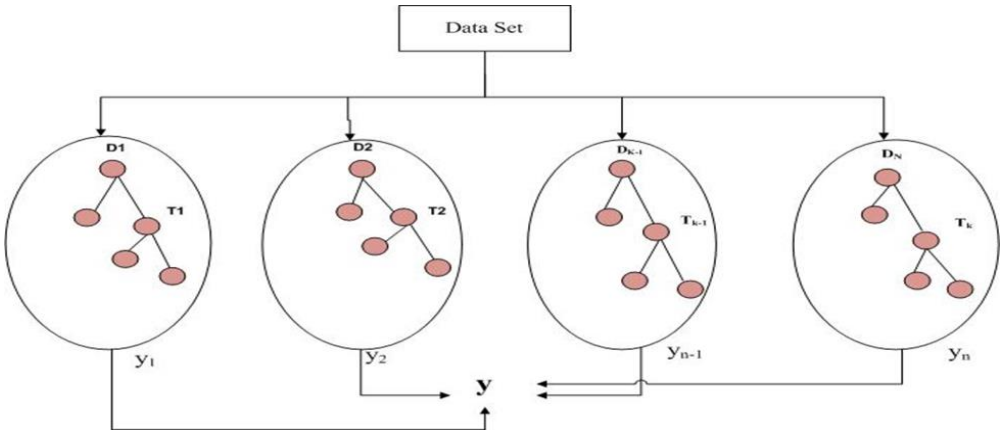


**Figure 2: Random forest model**

## Logistic Regression

Logistic Regression is a supervised machine-learning technique widely used to solve Regression-based problems. This technique uses a collection of independent variables to predict the category dependent variable. Logistic Regression can get from linear regression expression [21].

The equation for a straight line may be expressed as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + .....\beta_n X_n \quad (2)$$

Linear Regression is helpful in that case when the output value of the prediction is continuous. In the above equation, y is the output value, β_0 is the intercept line, β_1 is the independent variables, and X_1 is the features.

$$\log\left[\frac{y}{1-y}\right] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + .....\beta_n X_n \quad (3)$$

In equation 2, log(y) is a predictor variable, and x is the independent variable. Beta is a coefficient parameter in a logistic regression model that is anticipated via the highest likelihood function. In this method, we examine multiple iterations of beta values for the best fit of the log. Logistic Regression tries to find the best parameters estimation for this purpose by capitalizing on the likelihood function produced in the entire iteration.

**Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)**
ISSN：1671-5497
E-Publication Online Open Access
Vol: 41 Issue: 12-2022
DOI 10.17605/OSF.IO/Z7FQA

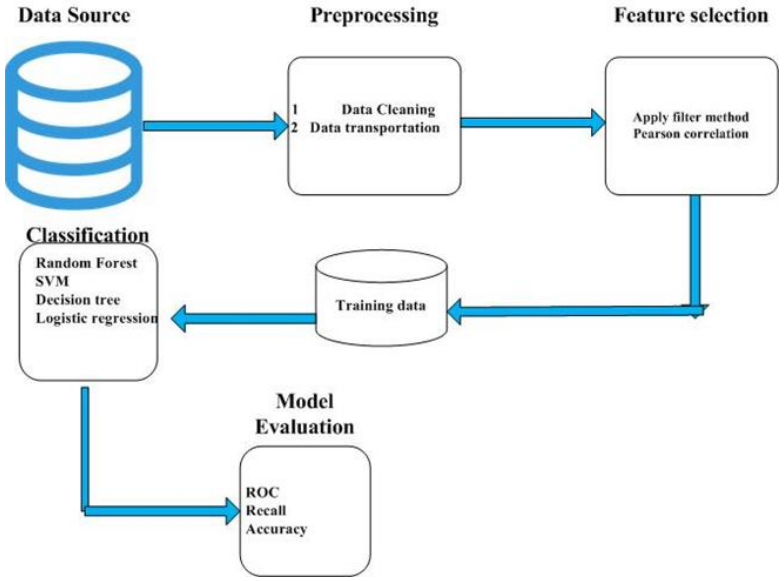


**Figure 3: Student performance Prediction Model**

## Data Set Collection

The data set of student academic performance used in the study has been collected from kal board 360 learning management system. The data set used in the study has been published on the Kaggle website. It consistsof480 student records, including 175 females and 305 males. It includes 16 features which include academic and non-academic features.

## Data Preprocessing

Data preparation is a crucial stage of the data mining process. We collect data from multiple sources, such as log files, databases, and excel files. The extracted data is in raw form, which is not in good shape for preprocessing. The date might be missing, noisy and uneven. The accomplishment better results, we must clean the data.

**Table: 2. Structure of dataset**

| Name of attributes | Description | Data types |
|---|---|---|
| Gender | Student gender detail | object |
| Nationality | Student belong to which country | object |
| Birthplace | Birth place of the student | object |
| Grade_id | The detail of student obtaining grade | object |
| Section | The detail of student belong to which section | object |
| Topic | Student learning course detail | object |
| Semester | The detail of student enroll semester | object |

**Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)**
ISSN：1671-5497
E-Publication Online Open Access
Vol: 41 Issue: 12-2022
DOI 10.17605/OSF.IO/Z7FQA

| | | |
|---|---|---|
| Relation | Relationship with the student | object |
| raised hands | The aspect of How many times student raise their hands in learning session | int |
| Visited Resources | The detail of How many times student visited LMS to view course content | int |
| Announcements View | How many times student visited LMS to view educational news | int |
| Parent Answering Survey | Student guardian satisfaction about student learning progress | object |
| Parent school Satisfaction | Student guardian satisfaction about school | object |
| Student Absence Days | Student attendance detail | object |
| Class | Student classification detail | object |

## Data integration and transformation

We gather data from multiple sources, which are in raw form. When we integrate various files into one file, redundancy arises in the dataset. The data set consists of two files taken from the Kaggle website. These files consist of student engagement and academic performance data extracted from the learning management system kalboard 360. In this step, we integrate various files into one file. Several classifiers do not apply to continuous data. There we use the discretization method to transform the desired numerical data into the nominal form.

## Experiments and results

This section predicts students' performance in the online learning environment. To answer our research objective, we conduct several experiments on our dataset. We employed ML classifiers on the dataset using the Jupiter notebook tool to build a predictive model. In this study, we used three ML classifiers, namely Random forest, Support vector, and decision tree, which were experimented on our dataset. These classifiers are evaluated using 10 fold cross-validation method. Each data set in this technique is divided into 10 subsets of equal size. In which 9 subsets are used for training and one for testing

**Table 3: Result of Implemented Classifiers**

| Classification Techniques | Accuracy | Recall | F Measure |
|---|---|---|---|
| Random Forest | 0.89 | 0.89 | 0.88 |
| Support vector | 0.83 | 0.82 | 0.82 |
| Logistic Regression | 0.81 | 0.81 | 0.81 |
| Decision tree | 0.82 | 0.80 | 0.80 |

**Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)**
ISSN：1671-5497
E-Publication Online Open Access
Vol: 41 Issue: 12-2022
DOI 10.17605/OSF.IO/Z7FQA

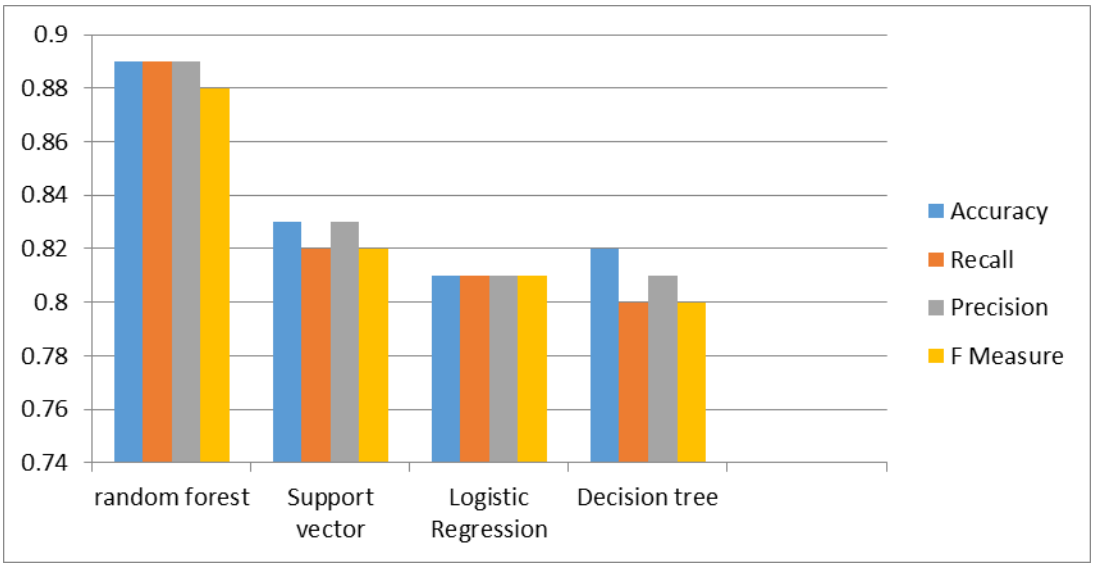


**Figure 4: comparison of different classifiers**

A comparison of the four machine learning classifiers' results is visualised in figure 4 above. Our Experiments show promising results accuracy achieved between 0.81% and 0.88%. The implemented classifiers, namely random forest, support vector, Logistic Regression, and decision tree, have an overall accuracy of 0.81%, 0.82%, 0.83%, and 0.88%, respectively.

### Model Evaluation

In this study, we first train the classifier on a given dataset and thenassess the Performance of our proposed predictive model. After obtaining the prediction result from the model on test data, we identified the number of true negatives, true positives and false negatives, and incorrect positive metrics used to assess the predictive model's performance.

### Performance Metric

### Accuracy

The first evaluation metric was accuracy, which is used to correctly predict the number of students with high and low Performance in the course [30].

$$\text{Accuracy} = \text{TP} + \text{TN}/\text{TP} + \text{TN} + \text{FP} + \text{FN} \quad (4)$$

### Precision recall

In this step, we calculate the recall that points out a fraction of the student dataset. The student with low Performance and classifier correctly identifies as having a low-performance student. An ML classifier is considered to have satisfactory Performance if they have good recall. And precision corresponds to the percentage of correctly classified accurate instances.

**Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)**
ISSN：1671-5497
E-Publication Online Open Access
Vol: 41 Issue: 12-2022
DOI 10.17605/OSF.IO/Z7FQA

Precision = TP/TP+ FP (5)

Recall = TP/TP+FN (6)

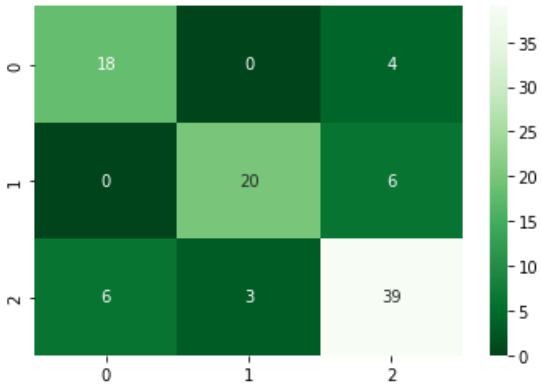


**Figure 5: Decision tree**

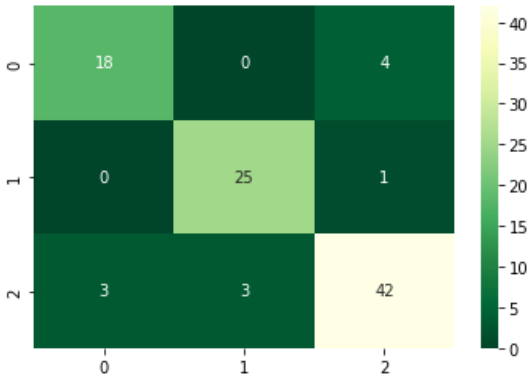


**Figure6: Random Forest**

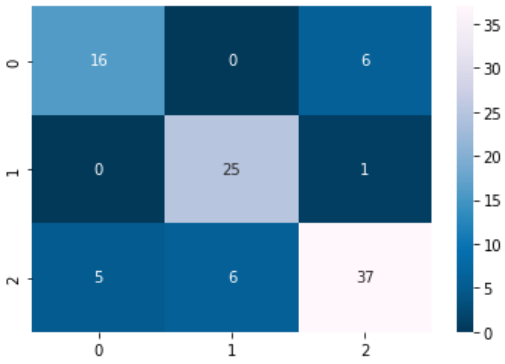


**Figure 7: Logistic Regression**

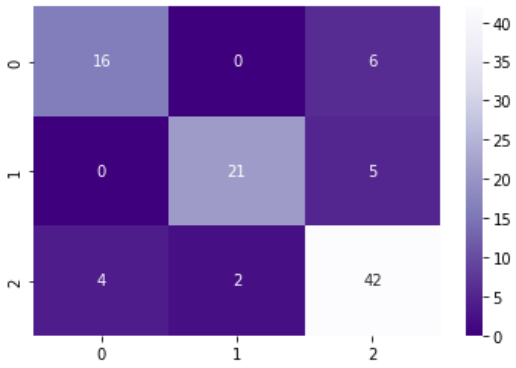


**Figure8: Support Vector Machine**

The confusion matrix is deployed to check the Performance of our proposed predictive model. It is in N*N matrix form to estimate the Performance of Machine learning classifiers where N belongs to the total number of target classes. In the above figures, we visualize the confusion metrics of implemented classifiers. The diagonal values show the correctly classify the values

## 4. CONCLUSION

The student's academic progress is a critical quality measurement indicator in e-learning. The prediction of academic achievement offers a foundation for teachers to adapt their teaching approach for students who may have study difficulties. This study used machine learning algorithms and statistical techniques to analyse the student data. The dataset used in the study is extracted from the Kaggle website, which is publicly

ISSN：1671-5497
E-Publication Online Open Access
Vol: 41 Issue: 12-2022
DOI 10.17605/OSF.IO/Z7FQA

available. After collecting data, we converted it into a suitable form before passing it into the machine learning model. We select the feature based on the information gain filter method from the dataset. Then we applied machine learning classifiers to the data set: support vector, Logistic Regression, random forest, and decision tree. We evaluated them by divide the data into 80% for training and 20% for testing. The experiment result demonstrates random forest achieves 89% higher accuracy among other tested classifiers. In Table 3 and figure 4, confusion metrics indicate that random forest is most appropriate for classifying student performance into high, medium and low categories.

## Future work

In future, we will enhance our proposed predictive model to classify disabled students' academic performance. Disabled students have learning difficulties such as hearing, blindness and memory. Classification disables students' academic performance will help them to enhance their learning.

### References

1. Bresfelean, Vasile Paul, Mihaela Bresfelean, Nicolae Ghisoiu, and Calin-Adrian Comes. “Determining Student’s Academic Failure Profile Founded on Data Mining Methods.” In ITI 2008 - 30th International Conference on Information Technology Interfaces. pp. 317–22. Cavtat/Dubrovnik, Croatia: IEEE, 2008
2. Imran, Muhammad, Shahzad Latif, Danish Mehmood, and Muhammad Saqlain Shah. "Student Academic Performance Prediction Using Supervised Learning Techniques." International Journal of Emerging Technologies in Learning (IJET)., vol. 14, pp 92. (July 24, 2019)
3. Arsad, Pauziah Mohd, Norlida Buniyamin, and Jamalul-lail Ab Manan. "A Neural Network Students' Performance Prediction Model (NNSPPM)." In 2013 IEEE International Conference on Smart Instrumentation, Measurement and Applications (ICSIMA), pp 1–5. Kuala Lumpur, Malaysia: IEEE, 2013. https://doi.org/10.1109/ICSIMA.2013.6717966
4. Abidi, Syed, Mushtaq Hussain, Yonglin Xu, and Wu Zhang. "Prediction of Confusion Attempting Algebra Homework in an Intelligent Tutoring System through Machine Learning Techniques for Educational Sustainable Development." Sustainability., vol. 11, no. 1,pp 105.(December 25, 2018):105. https://doi.org/10.3390/su11010105
5. Hussain, Sadiq, Silvia Gaftandzhieva, Md. Maniruzzaman, Rositsa Doneva, and Zahraa Fadhil Mushin. "Regression Analysis of Student Academic Performance Using Deep Learning." Education and Information Technologies., vol. 26, no. 1,pp 783–98 . (January 2021) https://doi.org/10.1007/s10639-020-10241-0
6. Baashar, Yahia, Gamal Alkawsi, Nor'ashikin Ali, Hitham Alhussian, and Hussein T Bahbouh. "Predicting Student's Performance Using Machine Learning Methods: A Systematic Literature Review." In 2021 International Conference on Computer & Information Sciences (ICCOINS), pp 357–62. Kuching, Malaysia: IEEE, 2021. https://doi.org/10.1109/ICCOINS49721.2021.9497185
7. El Aissaoui, Ouafae, Yasser El Alami El Madani, Lahcen Oughdir, Ahmed Dakkak, and Youssouf El Allioui. "A Multiple Linear Regression-Based Approach to Predict Student Performance." In Advanced Intelligent Systems for Sustainable Development (AI2SD'2019), edited by Mostafa Ezziyyani, pp 9–23. Advances in Intelligent Systems and Computing. Cham: Springer International Publishing, 2020. https://doi.org/10.1007/978-3-030-36653-7_2

ISSN：1671-5497
E-Publication Online Open Access
Vol: 41 Issue: 12-2022
DOI 10.17605/OSF.IO/Z7FQA

8. M. E. Bulger, R. E. Mayer, K. C. Almeroth, and S. D. Blau, "Measuring learner engagement in computer-equipped college classrooms," Journal of Educational Multimedia and Hypermedia, vol. 17, no. 2, pp. 129–143, 2008
9. D. T. Tempelaar, B. Rienties, and B. Giesbers, "In search for the most informative data for feedback generation: learning analytics in a data-rich context," Computers in Human Behaviour, vol. 47, pp. 157–167, 2015
10. Antonenko, Pavlo D., Serkan Toy, and Dale S. Niederhauser. “Using Cluster Analysis for Data Mining in Educational Technology Research.” Educational Technology Research and Development., vol. 60, no. 3, pp 383–98. (June 2012). https://doi.org/10.1007/s11423-012-9235-8
11. Hu, Ya-Han, Chia-Lun Lo, and Sheng-Pao Shih. "Developing Early Warning Systems to Predict Students' Online Learning Performance." Computers in Human Behavior., vol.36, no.3, pp 469-78. (July2014).
12. Aissaoui, Ouafae EL, and Lahcen Oughdir. "A Learning Style-Based Ontology Matching to Enhance Learning Resources Recommendation." In 2020 1st International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET), pp 1–7. Meknes, Morocco: IEEE, 2020. https://doi.org/10.1109/IRASET48871.2020.9092142
13. Alsayed, Alhuseen Omar, Mohd Shafry Mohd Rahim, Ibrahim AlBidewi, Mushtaq Hussain, Syeda Huma Jabeen, Nashwan Alromema, Sadiq Hussain, and Muhammad Lawan Jibril. "Selection of the Right Undergraduate Major by Students Using Supervised Learning Techniques." Applied Sciences 11, vol .no. 22, pp 106-39 (November 11, 2021): https://doi.org/10.3390/app112210639
14. Mengash, Hanan Abdullah. "Using Data Mining Techniques to Predict Student Performance to Support Decision Making in University Admission Systems." IEEE Access.vol.8, pp55462-55470 (2020).
15. Yanes, Nacim, Ayman Mohamed Mostafa, Mohamed Ezz, and Saleh Naif Almuayqil. "A Machine Learning-Based Recommender System for Improving Students Learning Experiences." IEEE Access, vol .no.8, pp 201218–35 (2020). https://doi.org/10.1109/ACCESS.2020.3036336
16. Baashar, Yahia, Gamal Alkawsi, Nor'ashikin Ali, Hitham Alhussian, and Hussein T Bahbouh. "Predicting Student's Performance Using Machine Learning Methods: A Systematic Literature Review." In 2021 International Conference on Computer & Information Sciences (ICCOINS), pp 357–62. Kuching, Malaysia: IEEE, 2021. https://doi.org/10.1109/ICCOINS49721.2021.9497185
17. S. M. Jayaprakash, E. W. Moody, E. J. M. Laur´ıa, J. R. Regan, and J. D. Baron, "Early alert of academically at-risk students: an open source analytics initiative," Journal of Learning Analytics, vol. 1, no. 1, pp. 6–47, 2014
18. P. Leitner, M. Khalil, and M. Ebner, "Learning analytics in higher education—a literature review," in Learning Analytics: Fundaments, Applications, and Trends: A View of the Current State of the Art to Enhance E-Learning, A. Peña-Ayala, Ed., pp. 1–23, Springer, Berlin, Germany, 2017
19. O. Corrigan, A. F. Smeaton, M. Glynn, and S. Smyth, "Using educational analytics to improve test performance," in Proceedings of EC-TEL 2015 10th European Conference on Technology Enhanced Learning, pp. 15–18, Toledo, Spain, September 2015
20. Sperandei, S. (2013). Understanding logistic regression analysis. Biochemiamedica 24 (1), 12–18. doi:10.11613/BM.2014.003
21. Liao, S. N., Zingaro, D., Alvarado, C., Griswold, W. G., & Porter, L. (2019). Exploring the value of different data sources for predicting student performance in multiple cs courses. In Proceedings of the 50th ACM Technical Symposium on Computer Science Education (pp. 112–118)

ISSN：1671-5497
E-Publication Online Open Access
Vol: 41 Issue: 12-2022
DOI 10.17605/OSF.IO/Z7FQA

22. Amazona, M. V., & Hernandez, A. A. (2019). Modelling student performance using data mining techniques: Inputs for academic program development. In Proceedings of the 2019 5th International Conference on Computing and Data Engineering (pp. 36–40)

23. Wei, D., & Burrows, A. N. (2016). Tracking students'Performance to assess correlations among computer science programming series courses. Journal of Computing Sciences in Colleges, 32(1), 9–16

24. Shahani, Niaz Muhammad, Xigui Zheng, Xiaowei Guo, and Xin Wei. 2022. "Machine Learning-Based Intelligent Prediction of Elastic Modulus of Rocks at Thar Coalfield" Sustainability 14, no. 6: 3689. https://doi.org/10.3390/su14063689

25. Elbadrawy A, Studham RS, Karypis G (2015) Collaborative multi-regression models for predicting students'Performance in course activities. In: 5th International conference on learning analytics and knowledge (LAK' 15), pp 103–107. https://doi.org/10.1145/2723576.2723590

26. Shahani NM, Zheng X, Liu C, Hassan FU and Li P (2021) Developing an XGBoost Regression Model for Predicting Young's Modulus of Intact Sedimentary Rocks for the Stability of Surface and Subsurface Structures. Front. Earth Sci. 9:761990. doi: 10.3389/feart.2021.761990

27. B. Bahati, F. Uno, and M. Tedre, "Can student engagement in online courses predict performance on online knowledge surveys?" International Journal of Learning,Teaching and Educational Research, vol. 16, no. 3, pp. 73–87,2017

28. M. Wells, A. Wollenschlaeger, D. Lefevre, G. D. Magoulas, and A. Poulovassilis, "Analyzing engagement in an online management programme and implications for course design," in Proceedings of Sixth International Conference on Learning Analytics & Knowledge, pp. 236–240, Edinburgh, UK, April

29. 2016 Iqbal, M. J., Bajwa, U. I., Gilanie, G., Iftikhar, M. A., & Anwar, M. W. (2022). Automatic brain tumor segmentation from magnetic resonance images using superpixel-based approach. Multimedia Tools and Applications, 1-19.

30. Iqbal, M. J., Bajwa, U. I., Gilanie, G., Iftikhar, M. A., & Anwar, M. W. (2022). Automatic brain tumor segmentation from magnetic resonance images using superpixel-based approach. Multimedia Tools and Applications, 1-19.