HIGHER EDUCATION STUDENT'S PERFORMANCE EVALUATION USING MACHINE LEARNING TECHNIQUES

BY

MD.Kamrul Hasan Efty
ID: 201-15-3561

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

MD.Assaduzzaman

Lecturer

Department of CSE

Daffodil International University

Co -Supervised By
Amir Sohel

Lecturer

Department of CSE

Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

JANUARY 202

APPROVAL

This Research paper titled” **HIGHER EDUCATION STUDENT’S PERFORMANCE EVALUATION USING MACHINE LEARNING TECHNIQUES**”, submitted by MD.Kamrul Hasan Efty, ID No: 201-15-3561 to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 19 January 2023.

BOARD OF EXAMINERS

Dr. Touhid Bhuiyan
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Chairman

Naznin Sultana
Associate Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Abdus Sattar
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

19.1.23

Dr. Md. Sazzadur Rahman
Associate Professor
Institute of Information Technology
Jahangirnagar University

External Examiner

DECLARATION

We hereby declare that, this project has been done by us under the supervision of **MD.Assaduzzaman ,Lecturer, Department of CSE** Daffodil International University. We also declare that neither this project nor any part of this paper has been submitted elsewhere for award of any degree or diploma.

Supervised by:

18.01.2023

MD.Assaduzzaman

Lecturer

Department of CSE

Daffodil International University

Co -Supervised by :

Amir 18/01/23

Amir Sohel
Lecturer

Department of CSE

Daffodil International University

Submitted by:

Kamrul Hasan

MD.Kamrul Hasan Efty

ID: -201-15-3561

Department of CSE

Daffodil International University

ACKNOWLEDGEMENT

First we express our heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the final year project/internship successfully.

We really grateful and wish our profound our indebtedness to MD.Assduzzaman, Lecturer, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of "machine learning" to carry out this project. His endless patience ,scholarly guidance ,continual encouragement , constant and energetic supervision, constructive criticism , valuable advice ,reading many inferior draft and correcting them at all stage have made it possible to complete this project.

We would like to express our heartiest gratitude to Professor Dr. Touhid Bhuiyan,Head, Department of CSE, for his kind help to finish our project and also to other faculty member and the staff of CSE department of Daffodil International University.

We would like to thank our entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, we must acknowledge with due respect the constant support and patients of our parents.

ABSTRACT

Failure and success in the classroom have real-world implications for achieving economic success in the knowledge-based economy. Using early detection markers (such as age, reading frequency, and CGPA), this research aims to forecast the likelihood of students' academic performance in order to provide prompt and effective remediation. On the basis of secondary data acquired from students' information systems, a machine learning approach was employed to create a model. In this paper, our main aim is to predict student performance for 3 specific factors student scientific book reading frequency, extra work conditions, and weekly study time. So we are using five machine learning algorithms KNN, Random forest, Decision tree, Linear regression, and GBC, and also use almost 1200 student attribute datasets. For students with extra work conditions random forest algorithms given the highest 99 % accuracy. For student scientific book reading frequency random forest and decision tree are given the highest 98 % accuracy. For students weekly study hours random forest and KNN given highest 97 % accuracy.

TABLE OF CONTENTS

CONTENTS **PAGE NO**

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: INTRODUCTION

1.1 Introduction

In numerous disciplines, including economics, medicine, etc. machine learning algorithms have been employed. One of the fields concerned with investigating data trends in a learning context is education machine learning. Predicting student performance to enhance the current educational environment is one of the most crucial uses. This research attempts to predict student performance using a machine learning algorithm. With the help of early detection factors (such as work, CGPA) this article aims to forecast the likelihood of student academic performance . One way to predict student's behavior is to use machine learning algorithm to analyze past student's behavior. Machine learning is a technique where computer how to learn and comprehend to given parameters. It can be used to find out information about the past that can be used to predict the future. One common method is called correlation. That is, it is used to predict whether an outcome will be in one eye view (such as graph). It is used to predict a value on a continuous scale. Machine learning algorithm can be used to predict student's behavior in many different ways. For example, machine learning algorithm can be used to predict:

- Whether a student reading frequency during exam time.
- Whether a student does work along with study.
- Student scientific temper

In this paper, we use almost 1200 undergraduate student data including their real life data as like as reading frequency , expected CGPA, expected salary, family condition. We use MLA to define student performance during exam and after completing graduation. In order to manage resources in higher education institutions and reduce failure, weak students are detected so that remediation can also be organized for them. Teachers can also identify students who are at risk, allowing them to support at-risk boys and successfully lead weaker students [1]. The majority of students spent four years at university unable to acquire the essential information and understanding. In addition to having little technical systems, many students leave universities with low test results. They have a very difficult time finding employment because students lack both theoretical and practical experience. [2]. There are two reasons for this: first, it is important to identify the students who will conduct well on the college course exam so that scholarships can be given to them, and second, it is even more crucial to determine the students who may struggle so that some sort of remediation can be provided to them. Students' academic success is affected by a variety of variables, including their prior academic performance, economic situation, family background, how they perform on final exams, etc[3].

1.2 Motivation

Academic achievement and failure serve as important turning points for achieving economic success in the knowledge-based economy. The purpose of this work is to predict the frequency of students' academic achievement using early identification markers to support in timely and effective remediation (such as work and CGPA).

1.3 Objective

In order to better manage resources in tertiary institutions and prevent failure, it is important to identify the "weak" pupils so that some sort of remediation can be provided for them.

1.4 Expected Outcome

1. Professionalism
2. Scientific temper
3. Reading frequency

CHAPTER 2: LITERATURE REVIEW

According to R. Kabra and R. Bichkar [1] work, decision tree algorithm can be utilized to create a model that can be used to predict students' success on engineering first-year exam. It is evident from the confusion matrix that the true the model's true positive rate for the FAIL class is 0.9, which suggests the students most likely to fail are successfully being identified by the model. In this case, we use random sampling of the training data is used for each of the many single trees that make up a random forest. Usually, they have better accuracy than simple decision trees. The next illustration demonstrates how adding more trees increases the decision boundary's accuracy and stability. These students can be taken into account for suitable therapy in order to enhance their output. If we take into account additional examples and include the attributes that show the present performance (such as attendance, test scores, etc.), the model's accuracy is likely to increase.

Zulfiker [2] the outputs of the basis algorithms were collected using a weighted voting approach, and 7 foundation algorithms were used in this work to estimate the student's final grades. Additionally, it is obvious from the study that employing the weighted voting method to aggregate the base classifiers has improved accuracy

The study's drawback is that no attempt was made to compare the effectiveness of the technique to that of other approaches, as shown by other studies.

Acharya A [3] This research aims to predict student performance using MLAs. From among the five MLA classes, an appropriate representative was selected, trained, and then tested. They were shown to predict better outcomes. Thus, educators are in a position to recognize the exceptional and, more importantly students who might not perform well at the end of the term examination. The data set and properties are derived by doing a survey of a group of computer science students in a few Kolkata undergraduate colleges. It should be noted, a broad nature to the research methods developed. It might be used for both full-time and distant learning courses, including some that use web-based instruction used. While the testing set has 104 instances, the training set has 309 instances. Decision tree are the most practical approach to generate the set, according to a thorough analysis. Consequently, C4.5 was employed to create the choice tree Kappa Measurements and F-Measure were measure the prediction algorithm's effectiveness. Average The training dataset's F-Measure value was found to be 0.79. While it was discovered that for the testing dataset was 0.66. as of later value appears to be relatively low, possibly as a result of the three are only 104 testing examples total similar outcomes are shown by Kappa Statistic as well.

On a number of fronts, the study approaches this report suggests can be improved. First off, students who receive less than 40% of the grade have been given a "F." As a result, there is no differentiation made between the students who received 38% and those who received 10%. Second, a number of students enroll in a course, take the midterm examinations, but for one reason or another are unable to take the final exams. Due to a missing attribute, these students are not taken into account for prediction. Finally, combining multiple classifiers may improve prediction accuracy (CMC). It is also possible to use genetic algorithms for this.

Yakubu [4] In order to predict whether or not students at a private institution in Nigeria will attain a passing CGPA, this research will use various methods to identify significant factors. The methodology applied was machine learning and the set of data was divided into test data (i.e., 30%) and training data (i.e., 70%). 84.7% accuracy was achieved in predicting the students' academic success or failure using a logistic regression model to analyze the data. Using the test data set, the model's accuracy was then evaluated, and it was found to be 83.5% accurate.

Enughwure [5] Everyone benefits hugely from the skill of predicting student achievement, but especially educators and students. To make sure their students receive the most effective learning plan from them, the administrators can assess their teaching methods. Students can assess their learning styles to determine the activities that will best suit their needs. The authors of this work reviewed earlier research on MLAs for predicting student performance. Since the majority of the authors of the articles under review are from lowliterate nations, this will aid academics in those nations in tracking student progress and improving their literacy rates. The researchers' go-to analytical tool for predicting student success is Weka. The supervised learning approach was primarily used in this research area when it comes to machine learning techniques. The majority of researchers made an effort to predict students' performance using their academic, social-economic, educator ability data. This approach maintains that a student's academic success is not just dependent on his academic contribution. The performance of pupils in the majority of the reviewed work is predicted using the total academic curriculum of the student.

Ahmed [6] One of the important areas is performance prediction for students (EDM). In this study, a dataset was produced to predict students' final exam results in preparation. The Models were created utilizing 20 attributes and 450 university students from a proposed dataset using custom machine learning methods. The GBDT algorithm leads the other techniques, with an accuracy of 89.1%. Using techniques of machine learning to improve the efficiency of educational institutions is a potential area for EDM. Institutions of higher learning, such schools, can use it in this area. This study showed that, employing a number of carefully chosen algorithms on the dataset, student performance may be predicted with a reasonable accuracy level.

Ghorbani [7] In order to address the problem of unlabeled data when projecting student performance, this study compares various resampling techniques, including Borderline SMOTE, Random Over Sampler, SMOTE, SMOTE-ENN, SVM-SMOTE, and SMOTE-Tomek. This paper uses a variety of machine learning classifiers, including RN, KNN, Artificial Neural Network, XGB, SVM(Radial Basis Function), DT, Logistic Regression, and Nave Bayes, to be able to better evaluate well how resampling methods perform in solving the unequal problem.

Altabrawee [8] In this study, four machine learning techniques were used to create a classifier that can estimate student performance in a computer science course offered by Al-Muthanna University's (MU) College Of Humanities. This research basically examines the relationship between students' usage of social media as a learning aid and the amount of time they spend on these websites. The ANN (fully linked feed forward layered ANN) model already had best performance (0.807), as well as the best classification accuracy (77.04%). The decision tree model also identified five other variables as crucial in influencing pupils' academic performance. These effects are the outcome of using performance measure features.

Masood [9] On two public student databases, they applied 11 highly machine learning models to make predictions about the future using the data from the research. The outputs of multiple machine learning models have been compared in order to identify which one is the best among them based on accuracy and F-measure. After to training the models with the databases, feature extraction approaches were utilized to achieve the goals. The importance of educational data mining (EDM) has increased in the modern era as a result of the fact because student concerns are growing along with technological innovation.

The best machine learning models in the case of these two public databases that we looked at in our research are "DT" and "Random Forest," as accuracy is determined by the quality of the database, the number of data, and the machine learning model. Tables 4 and 5's results show that "DT" and "Random Forest" are the best machine learning models because they are both almost 100% accurate for the first database and 97% accurate for the second.

Aggarwal [10] According to the comparison, models made using the Inter Perceptron and Random Forest classification methods have the highest classification accuracy, with an accuracy of 92.3%. When using Multi-Layer Classification model (MLP), which results in a relative absolute error of 22.4%, the error is as small as it can get.

Marwaha [11] This study uses ML models to try to determine the effects of various variables on the prediction of students' performance. Academic, demographic, social, and mental factors are taken into consideration when establishing a feature space. In order to choose a subset of features with greater predictive potential, the features selection is crucial. The primary goal was to raise the standard of higher education institutions by developing individuals who needed special attention so that the appropriate authorities could participate and correct the problem.

Singh [12] Using Python code and supporting packages like Scikit-learn, Pandas, and Numpy, among others, the research is carried out on a dataset provided by the student. The student dataset is split into a training set and a test set, each comprising 80% of the dataset. The middle 50% of the facts are represented by the plot's 25th and 75th percentage points surrounding the data. To represent the overall distribution of observations, draw a line at the median (50th percentile), with whiskers below and above the box.

Rivas [13] The goal is to use those models to assess data from a digital environment and build performance models for predicting how a student will succeed or fail during the academic year. Since the advent of virtual classrooms, teachers have been able to follow their students' use of digital resources and analyze some of the factors that may have helped to their student achievement lack thereof—by identifying their learning habits. The elements that most significantly affect a student's performance are then discovered, and improvements to those factors are made in order to improve the pass rate for students.

Karthikeyan [14] Data about students' performance from academic and other classroom activities in the university during course time are included in the dataset use for research purposes. One of the reasons students do badly in academic activities and even drop out of classes is the extraordinary development of information systems, such as social media, which may distract them from their actual course. The first method used an improved K-Means algorithm with SVM, which was improved to reduce the number of training examples used during classification. The second method used a 2-step method that combined the benefits of various classifiers to improve efficiency. This work proposed methods to increase the student achievement prediction method.

Shruthi [15] The purpose of this research is to identify the students who need extra assistance in order to reduce the failure rate and to take the appropriate steps in preparation for the exam that will be presented the upcoming semester. In order to project a student's performance in the upcoming semester using data from previous students, an administers is used to a student database in this paper. To predict the performance at the end of the semester, data on the student's previous database was collected, comprising attendance, seminar, and assignment marks. The Nave Bayes algorithm is applied in this context since there are numerous methods for classifying data.

Ramesh [16] The goal of this work is to determine the variables that affect students' performance on final examination and develop a suitable data mining method to project students' grades in order to issue timely and appropriate notifications to students who are at risk. According to the results of the theory testing, parents' professions have a major impact on predicting grades, whereas the type of school attended has no bearing on students' performance. To determine the connection between various attributes and the grade a student received, researcher employed the Chi-square test. One of the most popular and basic parametric and non-parametric tests used in statistical work is the chi-square test (x2).

Khan [17] The goal of this study is to create a prediction system that can warn a student to his or her potential results at the start of the semester. We used WEKA to apply 11 classification algorithms to a data source in order to do this. We came to the realization that the algorithms in the Decision Tree family achieve high accuracy, with J48 being the most acceptable at 88%. The outcomes of this research will improve both the students' and the teachers' performance. In addition to lowering the fail rate and taking appropriate action for the next semester's exam, this study will help identify the students who needed additional help.

Alloghani [18] The study's primary goal was to evaluate and predict student academic performance using educational data mining methods. In terms of predicting grade levels, the analysis's findings indicate that neural networks performed better than the other methods. While undergraduate learning behaviors like attendance in class and resource use obviously have an impact on their performance, their performance varies from country to country. Nave Bayes, Neural Networks, and decision trees (C4.5 and CART) are some of the classification methods employed in this paper.

Shahiri [19] This paper's main goal is to give a complete review of the data mining methods that have been applied to predicting student performance. The researchers' two most popular techniques for predicting student performance under the classifications techniques are the neural network and decision tree. Educators and students can both profit from enhancing their teaching and learning processes by using functions of the application of their pupils. The performance of students has been predicted in earlier research using a range of analytical techniques. In summary, the conceptual on predicting student performance has motivated us to conduct research to be used in our environment.

Suguna [20] An ideal model for multiple linear regression is produced using backward elimination, which lowers the unnecessary independent variable and improves the model's efficiency. Evaluating performance measures like R-Squared numbers and adjusted R-Squared values allows for the evaluation of the experimental findings and the evaluation of the model's efficacy.

CHAPTER:3 RESEARCH METHODOLOGY

3.1 Proposed Model

In this chapter, I go over the suggested model for predicting the risk students for student performance in good details. This section gives a description of the system's architecture. In Section 3.2, the dataset containing student performance data is described. The methods of machine learning that were studied and implemented for the prediction are discussed in Section 3.3. Our research methods is displayed in figure 1 below.

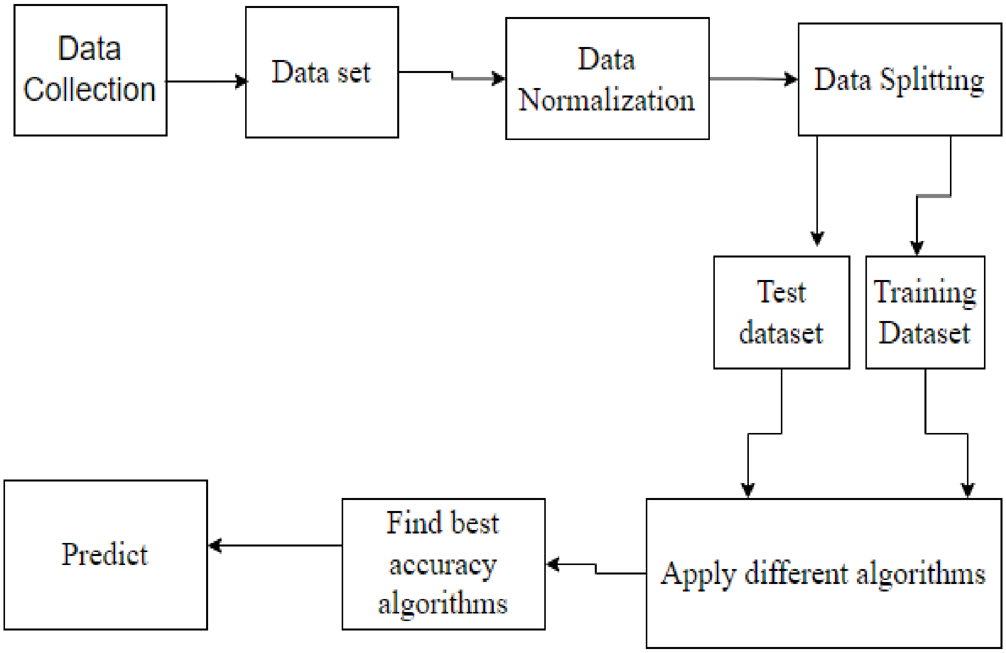


Figure 3.1 : Proposed model

3.2 Collection of Data

We are collecting this dataset from our local various universities where we collect student academic performance like scientific book reading frequency and weekly study hours. In this dataset, we use different class levels for predicting student performance evaluation like (yes or no).

3.2.1 Dataset

The data set's row number was reduced to a level that was appropriate for using different machine learning algorithms by condensing it into a single CSV file. Machine learning algorithms needs a lot of data to predict anything.

3.2.2 Attribute Information

TABLE 3.2.2 ATTRIBUTE INFORMATION FOR DATASET

Attribute Name	Data type	Category
Student ID	Numerical	Range 0-100
Age	Numerical	18-26 or above
Gender	Nominal	Female, male
Higher education type	Nominal	Private, State, Other
Scholarship	Nominal	25 %-100 %
Work	Nominal	Yes, No
Activity	Nominal	Yes, No
Partner	Nominal	Yes, No
Salary	Numerical	135 (USD) – 410 (USD)

Transport	Nominal	Bus, Private, car/taxi , bicycle, Other
Living	Nominal	Rental, Dormitory, With family, Other
Mother education	Nominal	Primary school, Secondary school, High school, University, MSc., Ph.D.
Father education	Nominal	Primary school, Secondary school, High school, University, MSc., Ph.D.
Siblings	Nominal	1 – 5 or above
Kids	Nominal	Parental status: Married, Divorced, Died - one of them or both
Mother job	Nominal	Retired, Housewife, government officer, Private sector employee, Self-employment, Other
Father job	Nominal	retired, government officer, private sector employee, self-employment, other
Study hours	Nominal	None , <5 hours, 6-10 hours, 11-20 hours, more than 20 hours
Reading frequency	Nominal	None, Sometimes, Often
Reading frequency scientific	Nominal	None, Sometimes, Often
Attendance dept. seminar	Nominal	Yes, No
Impact project	Nominal	Positive, Negative, Neutral

Attendance	Nominal	Always, Sometimes, Often
Preparation study		Alone, With friends, Not applicable
Preparation exam	Nominal	Closest date to the exam, regularly during the semester, never
Take class note	Nominal	Never, sometimes, always
Listen	Nominal	Never, sometimes, always
Classroom	Nominal	Not useful, useful, not applicable
Cumulative CGPA	Numerical	<2.00, 2.00-2.49, 2.50-2.99, 3.00-3.49, above 3.49
Expected GPA	Numerical	<2.00, 2.00-2.49, 2.50-2.99, 3.00-3.49, above 3.49
Like discuss	Nominal	Never, sometimes, always

3.2.3 Pre-processing of Data

Preparing the data for analysis is the initial stage in creating a prediction model. The model becomes more effective when the data is transformed in the right way. Searching for any missing values was the initial stage in processing this dataset. Average values were used to fill in any missing data, and so on. The dataset was resolved in this manner before any algorithmic techniques were used.

3.2.4 Splitting Dataset

The dataset must be divided into two sections: one is for training the model and the other for model testing before using any machine learning technique. Data partitioning is the process of achieving this. This must be done before to using any methods of machine learning.

3.3 Learning Models

The most popular data mining techniques for pattern and classification recognition are listed in the section that follows. This article was written using five machine learning approaches. The algorithms are known by the names logistic regression, KNN, SVC, Random Forest, Gradient Boosting and Decision Tree. Using this data, six different models were created and evaluated. The algorithms were carefully put into practice, and the generated data were carefully reviewed.

3.3.1 Decision Tree

The machine learning technique known as a decision tree uses branches to represent all potential decisions' results in relation to specific criteria. Each tree branch correlates to one or more results from the initial dataset. The rule sets that make up the tree structure are actually hierarchical structures that are organized from leaf nodes to root attributes. [23, 24]. The top node of the tree, the root, is the only node that does not have any incoming branches, and all of the outgoing branches represent each row according to the dataset. Use the internal node in the tree, which has both incoming and outgoing branches, to test the attribute. The terminal node or leaf is the descending node that has the single upward branch. This node is the final node of the tree, even though there may be other tree structures that show the most recent calculations [25].

3.3.2 Logistic regression

Logistic regression is frequently used to examine and highlight the relationship between a binary variable (such as "pass" or "failed") and a number of predicted variables [26]. There will be a search for the most effective model to describe the relationship between the independent and dependent variable sets. Despite being developed at the same time as linear regression, logistic regression has a different response for binary and continuous data [27].

3.3.3 KNN

KNN is a fundamental machine learning technique that assigns grades to objects based on the decision of their neighbors. The object is allocated into the class that its closest neighbors share the most. K is a real integer that is typically modest. The object is put in the class of its nearest neighbor if k is equal to 1. In binary (two class) classification issues, choosing k as an odd integer helps to break ties. The choice of parameter value in this algorithm could be important. [27, 29, 30].

3.3.4 Random forest

A Random Forest classifier develops a range of decision trees that are trained on multiple parts of the same training set in order to improve classification rates and address the fitting difficulties problem [31]. Random Forest generates K numbers of trees each time with different properties if editing is not used. Decision Tree only examines one tree, allocating the output that is most frequently produced to that instance, as compared to Random Forest, which checks the test data against all the trees that were produced [32].

3.3.5 Gradient boosting

Gradient boosting classifiers are a subclass of machine-learning methods that integrate a lot of unsuccessful learning models into a powerful predicting model. Gradient boosting is usually carried via using decision trees. The Support Vector Algorithm is a machine learning method that is based on an effective cost function optimization technique. It can be employed for both data classification and prediction[33].

CHAPTER 4: RESULTS & DISCUSSION OF EXPERIMENTS

4.1 Experimental Results

The accuracy and scores of each algorithm were then utilized to evaluate which algorithm was best at predicting student progress in terms of scientific interest, reading frequency, and CGPA after the successful execution of the Machine Learning Model Creation. The study results contain an analytical portion where each possible score for each algorithmic application and approach may be evaluated.

4.2 Find best algorithm

Each approach must be used multiple times with varied performance levels in order to determine the optimum algorithm. The methods must be examined using the dataset. It is challenging to find a suitable algorithm because there are so many requirements that must be met.

4.3 Confusion matrix

A unique table structure used in training set, a subfield of machine learning that focuses on statistical categorization, is a confusion matrix, also known as an error matrix. The examples in a predict class are represented by each column of the matrix, whereas the instances in an actual class are represented by each row (or vice versa) [34]. Second, depending on the situation, it's crucial to check these mean values for inter categorization using either a micro mean or a macro mean. Before diving deeper, it's crucial to comprehend the four basic components that are used to build the various evaluation measures. The four categories of false positives and false negatives are false positives (FP), false negatives (FN), and true positives (TP) (FN).

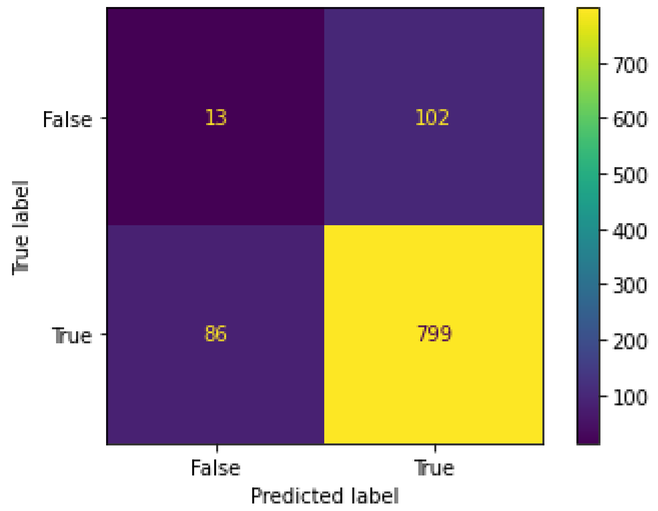


Figure 4.4.1 : Confusion matrix of random forest (Student work)

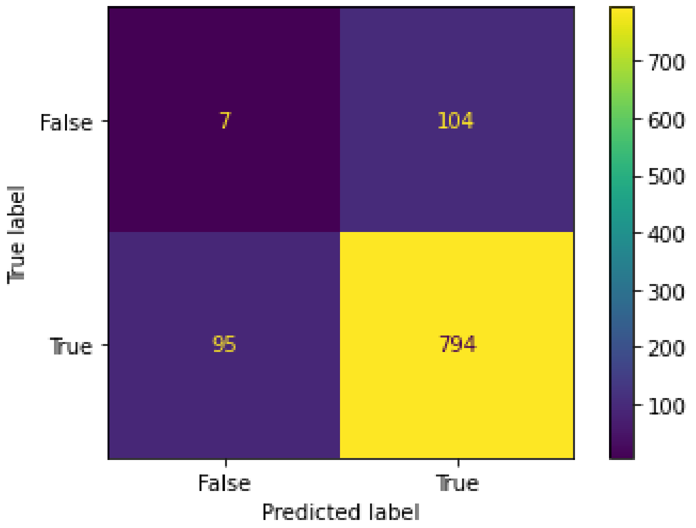


Figure 4.4.2 : Confusion matrix of random forest (Scientific reading frequency)

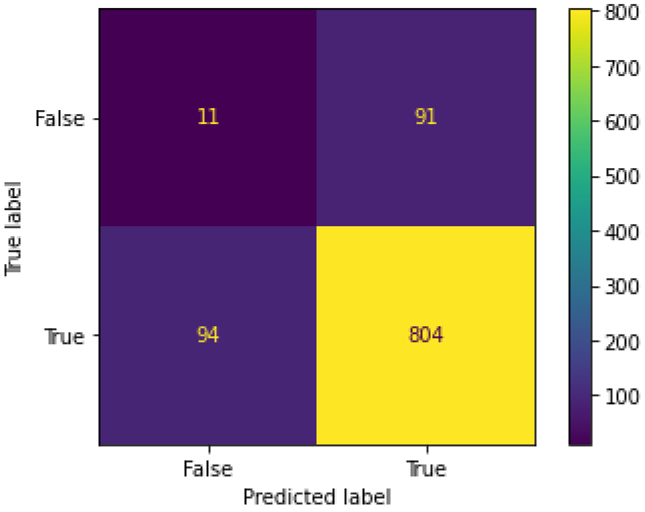


Figure 4.4.3 : Confusion matrix of random forest(weekly study hour)

A. True Positive (TP)

Positive sets indicate that the classifier correctly identified them. The letter TP stands in for it in the name.

B. True Negative (TN)

Positive tuples that were misclassified by the classifier are referred to as negative tuples. These tasks may be represented with the letter TN.

C. False Positive (FP)

Now, our interest has been peaked by the incorrect classification of these tuples with negative labels as positive. The use of FP may be used to indicate this kind of relationship.

D. False Negative (FN)

These Today, our focus is on the classification of these tuples with negative labels as positives. The use of FP can demonstrate this type of link. The classifier misclassified these positive tuples as negative. It is referred to as FN.

Precision

Precision is the capacity of a classifier to prevent classifying something as positive that is actually negative. The percentage of true positives to the sum of true positives and false positives is how it is defined for each class.

$$\text{Precision} = \frac{TP}{TP+FP}\ldots\ldots\ldots\ldots\ldots\ldots\ldots..(1)$$

Recall

Recall refers to a classifier's ability to recognize each successful instance. According to one definition, it is the ratio of true positives to all true positives and false negatives for each class.

$$\text{Recall} = \frac{TP}{TP+FN}\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots..(2)$$

F1 Score

The F1 score, which ranges from 0.0 to 1.0 based on recall and precision is a weighted conjunction. F1 scores are less accurate than accuracy measures because precision and recall are taken into account when computing them. It is usually suggested that, when comparing classifier models, the weighted average of F1 be used instead of overall accuracy.

$$\text{F1 Score} = \frac{2*(Recall*Precision)}{Recall+Precision}\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots.(3)$$

Accuracy

That how a model performs across all classes is a measure of its accuracy. This is beneficial when each class is equally important. By dividing the complete number of forecasts by the guess that was successfully predicted, it is determined.

$$\text{Accuracy} = \frac{(TP+TN)}{(TP+FN+FP+TN)} \ldots\ldots\ldots\ldots\ldots\ldots\ldots.(4)$$

4.4 Classification Report

A classification report is a figure used in machine learning to assess the system's overall effectiveness. The F1 Score, support, recall, and accuracy of a trained classification model are displayed in conjunction with such a training dataset. Performance information for a machine learning model based on classification is represented by this number. The support, accuracy, recall, and F1 score for the model are shown in this table. It offers a clearer view of the trained model's general efficacy. In order to assess the categorization outcome produced by machine learning models, it is critical to be familiar with all of the metrics provided in the study. These data were used to create and assess five different models.

TABLE 4.5.1 CLASSIFICATION REPORT FOR READING FREQUENCY(SCIENTIFIC)

Classifiers Name	Accuracy	Class level	Precision	Recall	F1 score
KNN	74 %	None	0.82	0.67	0.74
		Sometimes	0.75	0.86	0.80
		Often	0.61	0.58	0 .60
Random forest	98%	None	0.94	1.00	0.97
		Sometimes	1.00	1.00	1.00
		Often	1.00	0.90	0.98
GBC	91%	None	0.98	0.94	0.96
		Sometimes	0.94	1.00	0.97
		Often	1.00	0.92	0.96
LR	71 %	None	0.69	0.71	0.70
		Sometimes	0.72	0.87	0.79
		Often	0.71	0.34	0.46
DT	98 %	None	0.93	1.00	0.97
		Sometimes	1.00	0.95	0.97
		Often	1.00	1.00	1.00

TABLE 4.5.2 : CLASSIFICATION REPORT FOR STUDENT WORK

Classifiers Name	Accuracy	Class level	Precision	Recall	F1 score
KNN	79%	Yes	0.71	0.57	0.67
		No	0.82	0.89	0.85
Random forest	99%	Yes	1.00	0.96	0.98
		No	0.98	1.00	0.99
GBC	93%	Yes	1.00	0.78	0.98
		No	0.91	1.00	0.95
LR	76%	Yes	0.74	0.39	0.51
		No	0.77	0.94	0.84
DT	97%	Yes	0.96	0.96	0.96
		No	0.98	0.98	0.98

TABLE 4.5.3 CALSSIFICATION REPORT FOR STUDENT READING FREQUENCY

Classifiers Name	Accuracy	Class level	Precision	Recall	F1 score
GBC	93%	None	0.82	0.92	0.87
		< 5 hours	0.95	0.96	0.95
		6-10 hours	1.00	0.87	0.93
		11-20 hours	1.00	1.00	1.00
		More than 20 hours	1.00	1.00	1.00
Random forest	97 %	None	0.95	1.00	0.97
		< 5 hours	0.96	0.98	0.97
		6-10 hours	1.00	0.90	0.95
		11-20 hours	1.00	1.00	1.00
		More than 20 hours	1.00	1.00	1.00

DT	97%	None	1.00	1.00	1.00
		< 5 hours	0.98	0.98	0.98
		6-10 hours	0.94	0.94	0.94
		11-20 hours	1.00	1.00	1.00
		More than 20 hours	1.00	1.00	1.00
KNN	71 %	None	0.78	0.58	0.67
		< 5 hours	0.70	0.88	0.78
		6-10 hours	0.75	0.43	0.55
		11-20 hours	0.75	0.75	0.75
		More than 20 hours	0.20	0.25	0.22
	60 %	None	0.38	0.26	0.31

LR		< 5 hours	0.64	0.86	0.73
		6-10 hours	0.42	0.20	0.27
		11-20 hours	0.86	0.75	0.80
		More than 20 hours	1.00	0.25	0.40

4.5 Result Analysis

We want this paper to predict student performance as reading frequency, scientific temper, and student additional working conditions. So we use 5 machine-learning algorithms random forest, KNN, GBC, logistic regression, and decision tree classifiers. According to our data set, we use 3 attributes additional work, reading frequency weekly, and reading scientific books. For students with extra work conditions random forest algorithms given the highest 99 % accuracy. For student scientific book reading frequency random forest and decision tree are given the highest 98 % accuracy. For students weekly study hours random forest and KNN given highest 97 % accuracy.

CHAPTER 5: FUTURE SCOPE & CONCLUSION

5.1 Future scope

The dataset can be increased in the future to perform better and be more accurate by collecting data from different private and public universities in Bangladesh. In the future, it will be possible to contrast the create a plan in this study to the approaches offered in other studies. Chances of success can be obtained by preprocessing data using normalization techniques and oversampling techniques like Synthetic Minority Over-sampling Technique, according to several studies (SMOTE). These techniques can be used to preprocess the collected dataset, which will enhance performance.

5.2 Conclusion

Student performance prediction is a very important factor in higher education. An early and perfect methodology can predict student performance. Our proposed methodology can predict student performance effectively. Previously much research showed some gaps which did not predict student performance like as professionalism, scientific temper, and weekly reading frequency. In this paper, we predict professionalism, scientific temper, and weekly reading frequency by using machine learning algorithms. Professionalism, scientific temper, and weekly reading frequency random forest and DT algorithms give the highest accuracy and it helps to improve student weakness properly.

REFERENCES

1. R. Kabra and R. Bichkar, “Performance prediction of engineering students using decision trees,” International Journal of computer applications, vol. 36, no. 11, pp. 8–12, 2011.

2. Zulfiker, M.S., Kabir, N., Biswas, A.A., Chakraborty, P. and Rahman, M.M., 2020. Predicting students' performance of the private universities of Bangladesh using machine learning approaches. International Journal of Advanced Computer Science and Applications, 11(3).

3. Acharya A, Sinha D. Early prediction of students performance using machine learning techniques. International Journal of Computer Applications. 2014 Jan 1;107(1).

4. Yakubu, M.N. and Abubakar, A.M., 2021. Applying machine learning approach to predict students' performance in higher educational institutions. Kybernetes.

5. Enughwure, A.A. and Ogbise, M.E., 2020. Application of machine learning methods to predict student performance: a systematic literature review. Int. Res. J. Eng. Technol.(2020). Retrieved April, 6, p.2021.

6. Ahmed, D.M., Abdulazeez, A.M., Zeebaree, D.Q. and Ahmed, F.Y., 2021, June. Predicting university's students performance based on machine learning techniques. In 2021 IEEE International Conference on Automatic Control & Intelligent Systems (I2CACIS) (pp. 276-281). IEEE.

7. Ghorbani, R. and Ghousi, R., 2020. Comparing different resampling methods in predicting students' performance using machine learning techniques. *IEEE Access*, *8*, pp.67899-67911.

8. Altabrawee, H., Ali, O.A.J. and Ajmi, S.Q., 2019. Predicting students' performance using machine learning techniques. *JOURNAL OF UNIVERSITY OF BABYLON for pure and applied sciences*, *27*(1), pp.194-205.

9. Masood, M.F., Khan, A., Hussain, F., Shaukat, A., Zeb, B. and Ullah, R.M.K., 2019, November. Towards the selection of best machine learning model for student performance analysis and prediction. In *2019 6th International Conference on Soft Computing & Machine Intelligence (ISCMI)* (pp. 12-17). IEEE.

10. Aggarwal, D., Mittal, S. and Bali, V., 2019. Prediction model for classifying students based on performance using machine learning techniques. *International Journal of Recent Technology and Engineering*, *8*(2S7), pp.496-503.

11. Marwaha, A. and Singla, A., 2020. A study of factors to predict at-risk students based on machine learning techniques. In *Intelligent Communication, Control and Devices* (pp. 133-141). Springer, Singapore.

12. Singh, R. and Pal, S., 2020. Application of machine learning algorithms to predict students performance. *Int. J. Adv. Comput. Res*, *29*, pp.7249-7261.

13. Rivas, A., Fraile, J.M., Chamoso, P., González-Briones, A., Rodríguez, S. and Corchado, J.M., 2019, April. Students performance analysis based on machine learning techniques. In *International Workshop on Learning Technology for Education in Cloud* (pp. 428-438). Springer, Cham.

14. Karthikeyan, K. and Kavipriya, P., 2017. On Improving student performance prediction in education systems using enhanced data mining techniques. *International Journal of Advanced Research in Computer Science and Software Engineering*, *7*(5).

15. Shruthi, P. and Chaitra, B.P., 2016. Student performance prediction in education sector using data mining.

16. Ramesh, V.A.M.A.N.A.N., Parkavi, P. and Ramar, K., 2013. Predicting student performance: a statistical and data mining approach. *International journal of computer applications*, *63*(8).

17. Khan, I., Al Sadiri, A., Ahmad, A.R. and Jabeur, N., 2019, January. Tracking student performance in introductory programming by means of machine learning. In *2019 4th mec international conference on big data and smart city (icbdsc)* (pp. 1-6). IEEE.

18. Alloghani, M., Al-Jumeily, D., Hussain, A., Aljaaf, A.J., Mustafina, J. and Petrov, E., 2018, September. Application of machine learning on student data for the appraisal of academic performance. In *2018 11th International conference on developments in ESystems engineering (DeSE)* (pp. 157-162). IEEE.

19. Shahiri, A.M. and Husain, W., 2015. A review on predicting student's performance using data mining techniques. *Procedia Computer Science*, *72*, pp.414-422.

20. Suguna, R., Shyamala Devi, M., Bagate, R.A. and Joshi, A.S., 2019. Assessment of feature selection for student academic performance through machine learning classification. *Journal of Statistics and Management Systems*, *22*(4), pp.729-739.

21. Kumar, A.S. and Joshna, K., 2021. Student’s Performance Analysis with EDA and Machine Learning Models.

22. Guleria, P., Thakur, N. and Sood, M., 2014, December. Predicting student performance using decision tree classifiers and information gain. In *2014 International conference on parallel, distributed and grid computing* (pp. 126-129). IEEE

23. Hamoud, A., 2017. Applying association rules and decision tree algorithms with tumor diagnosis data. *International Research Journal of Engineering and Technology*, *3*(8), pp.27-31.

24. Basu, K., Basu, T., Buckmire, R. and Lal, N., 2019. Predictive models of student college commitment decisions using machine learning. *Data*, *4*(2), p.65.

25. Millar, R.B., 2011. *Maximum likelihood estimation and inference: with examples in R, SAS and ADMB*. John Wiley & Sons.

26. G. Fitzmaurice and N. Laird, "Multivariate Analysis: Discrete Variables (Logistic Regression)," 2001.

27. Basu, K., Basu, T., Buckmire, R. and Lal, N., 2019. Predictive models of student college commitment decisions using machine learning. *Data*, *4*(2), p.65.

28. Hastie, T., Tibshirani, R. and Friedman, J., 2009. Additive models, trees, and related methods. In *The Elements of Statistical Learning* (pp. 295-336). Springer, New York, NY.

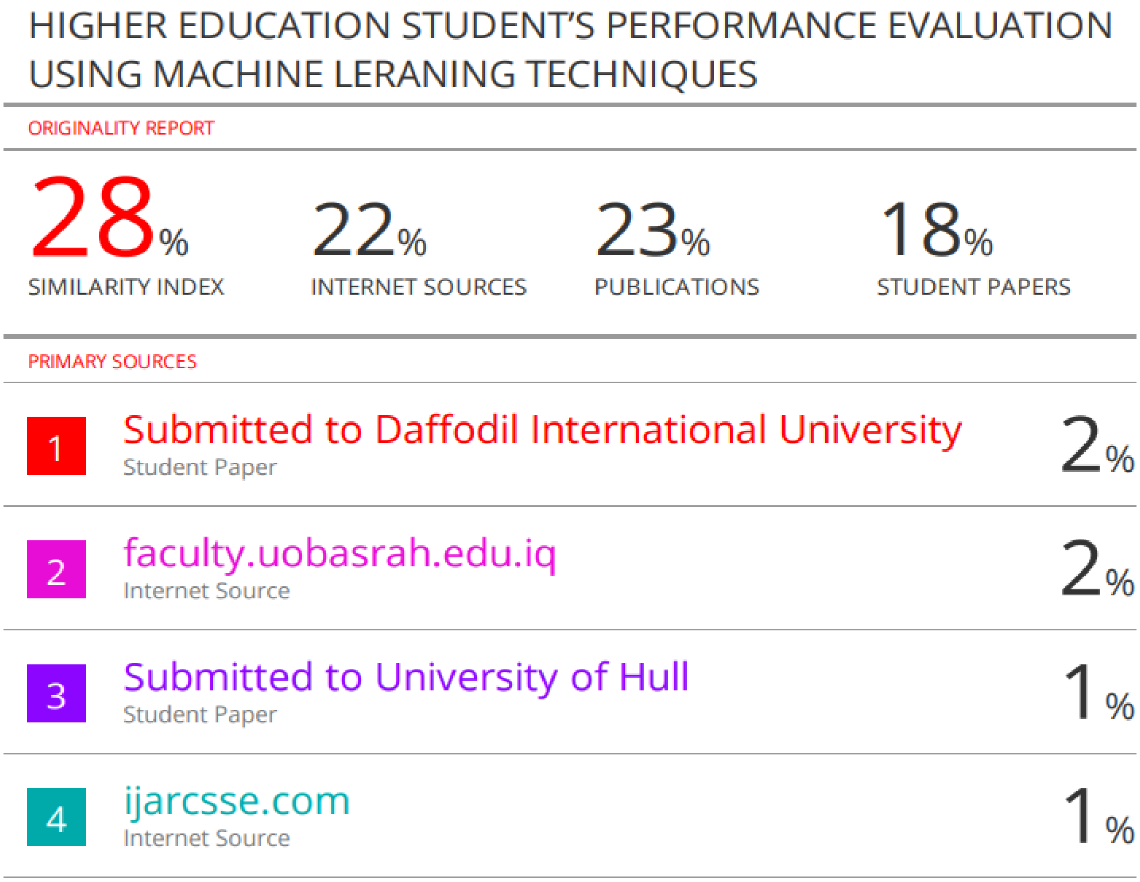
29. Chaurasia, V. and Pal, S., 2017. A novel approach for breast cancer detection using data mining techniques. *International journal of innovative research in computer and communication engineering (An ISO 3297: 2007 Certified Organization) Vol*, *2*.

30. Mahboob, T., Irfan, S. and Karamat, A., 2016, December. A machine learning approach for student assessment in E-learning using Quinlan's C4. 5, Naive Bayes and Random Forest algorithms. In *2016 19th International Multi-Topic Conference (INMIC)* (pp. 1-8). IEEE.

31. Mishra, T., Kumar, D. and Gupta, S., 2014, February. Mining students' data for prediction performance. In *2014 Fourth International Conference on Advanced Computing & Communication Technologies* (pp. 255-262). IEEE.

32. Sekeroglu, B., Dimililer, K. and Tuncal, K., 2019, March. Student performance prediction and classification using machine learning algorithms. In *Proceedings of the 2019 8th International Conference on Educational and Information Technology* (pp. 7-11).

33. Haghighi, S., Jasemi, M., Hessabi, S. and Zolanvari, A., 2018. PyCM: Multiclass confusion matrix library in Python. *Journal of Open Source Software*, *3*(25), p.729.



HIGHER EDUCATION STUDENT'S PERFORMANCE EVALUATION USING MACHINE LERANING TECHNIQUES

ORIGINALITY REPORT

28%	22%	23%	18%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	2%
2	faculty.uobasrah.edu.iq Internet Source	2%
3	Submitted to University of Hull Student Paper	1%
4	ijarcsse.com Internet Source	1%