Student Academic Performance Prediction Analysis Based on Lifestyle and Previous Academics Using Data Science and Machine Learning

Fathima Anwar
Assistant Professor
Computer Science and Engineering
MEA Engineering College
Perinthalmanna, India

Muhammed Suhair K.
Assistant Professor
Computer Science and Engineering
MEA Engineering College
Perinthalmanna, India

Razeen Abdul Gafoor
Computer Science and Engineering
MEA Engineering College
Perinthalmanna, India

Sabah Sayed K.
Computer Science and Engineering
MEA Engineering College
Perinthalmanna, India

Shahabas M. P.
Computer Science and Engineering
MEA Engineering College
Perinthalmanna, India

Shahul P.
Computer Science and Engineering
MEA Engineering College
Perinthalmanna, India

Abstract—Students academic performance has the potential to be revolutionised by machine learning and data science in the realm of education which can understand and predict the performance of student. By analyzing historical data, educators can gain insights into the factors that influence student success and use this information to develop tailored learning strategies. The objective of predicting undergraduate student performance using machine learning algorithms was looked into in this study. The classification methods of support vector machines, decision trees, random forests, and logistic regression were all tested. The key factors that were found to influence student performance included study hours, internet usage, and hours of sleep. These results imply that data science and machine learning can play a significant role in enhancing student learning and ensuring their academic success. Further investigation is required to completely comprehend the potential utilization and generalizability of these methods in various educational environments.

Index Terms—Logistic Regression, SVM, Decision Tree, Ran-dom Forest, Voting Classifier

I. Introduction

Predicting student academic performance is a vital aspect of education that has the potential to greatly benefit students, educators, and the overall learning experience. These results simply that data science and machine learning can play a significant role in enhancing student learning and ensuring their academic success, as they are capable of analyzing vast amounts of data and identifying patterns that might not be noticeable to humans. With the capacity to predict student performance, teachers can spot students who could be in risk of falling behind and give them the individual assistance they need to succeed. This can greatly improve the overall learning experience for students and the achievement gap between high- and low-performing students be closed. By Intervening early, educators can provide additional support to struggling students, leading to a better learning experience for all students.

Machine learning algorithms are also highly effective in analyzing large amounts of data and identifying patterns that may not be visible to humans. This results in more precise projections of student performance and a great understanding of the unique needs of each student. Machine learning algorithms that predict student academic success have significant potential to enhance overall learning and close the achievement gap. Utilizing machine learning methods allow educators to intervene early and provide targeted support to students who may be struggling, leading to a better learning experience for all students. The ability to analyze large amounts of data and identify patterns that may not be noticeable to humans leads to more accurate predictions and a better understanding of individual student needs.

II. Related Works

It has been discovered in the study that student-centric learning has been greatly beneficial in enhancing the learning experience for students. The support of data mining applications, through the early detection of low-performing students, has aided instructors in optimizing their educational strategies and keeping the students motivated [1]. The historic results of a course were utilized in the evaluation of accuracy in student success. Several classification algorithms were tested and linear discrimination analysis was found to be the most accurate [1].

The use of TMS data in predicting academic performance was investigated. Multiple machine learning models were applied and the best model was found to be the linear SVM model [2]. The importance of different TMS items in making predictions was also explored and two items were identified as the most important. Further research is suggested to include more data and explore the potential use of other features and classification models to improve prediction accuracy [2].

Decision Tree and Random Forest are compared as classification techniques were employed to predict academic performance of students. The findings show that Decision Tree performed better than Random Forest in terms of effectiveness. [3]. A variety of factors affecting students' learning behavior were revealed through the experiments, such as the influence of spending time with friends or family. The findings of this study can be utilized for students' learning behavior evaluation, as a reference for lecturers and parents, and as a basis for further studies that involve expanded datasets and more classification techniques [3].

Based on students academic success and first semester grades, a machine learning model was created to predict the success of students at the end of their first year in university. [4]. The implementation of the model was able to accurately predict whether a student would pass or fail their exam with 87% accuracy. Further exploration and improvement of the model could be conducted to increase its accuracy and usefulness [4].

The focus was on predicting the future success of students in distance learning systems through analysis of data generated by learning management systems. Support vector machine and logistic regression are two classifiers, were used with features selected through the gain ratio feature selection method [5]. The findings demonstrated that the primary influences on academic grades were student satisfaction, system interaction, and punctuality in class, and that the support vector machine technique utilising sequential minimal optimization was better suited to predicting future performance [5].

This study highlights the significance of data mining in educational institutions and its potential to improve students' performance through effective decision making [6]. The results from this experiment, using five classifiers in WEKA, indi-cated that among the classifiers tested, Multilayer Perceptron performed the best. Future work could include an expansion of the dataset and the exploration of other, strategies for data mining like association and clustering [6].

The data classification is a useful data mining technique for knowledge discovery in predicting students' performance in exams [7]. The experiments conducted on popular decision tree classifiers revealed the best classifier for predicting students' performance in the First Year of engineering exams. The outcomes demonstrated that the FAIL class had a high true positive rate of 0.786 for the C4.5 decision tree algorithm, indicating that it can effectively identify students who are likely to fail. This can be beneficial for proper counseling to improve their results. Machine learning algorithms like the C4.5 decision tree can be applied to student data to produce short and accurate predictions for incoming students [7].

It has been demonstrated through experiments that the proposed GBDT algorithm, which is capable of incrementally updating the classification model built upon GBDT, is able to achieve improved model building/updating time while retaining the same classification accuracy compared to the straightforward GBDT method. Further testing in real-world applications and adaptations to the Spark distributed computing platform for efficiency improvement, as well as the suitability for transfer learning, are planned as future work [8].

Students pursuing bachelor's and master's degrees had their academic performance predicted using both the fuzzy genetic algorithm and the decision tree algorithm. The decision tree algorithm identified a higher number of students at risk, leading to increased attention from lecturers for better results in final exams [9]. The fuzzy genetic algorithm, on the other hand, resulted in more students being classified as passed and provided a mental satisfaction for these students. This created a friendly environment between professors and students, which led to respected businesses hiring knowledgeable students, leading to a secure future for students and prestige for the institution [9].

To predict students academic performance prior to admission or promotion to higher courses in an academic, an intelligent decision support system (IDSS) was created [10]. A decision tree-based technique called Logistic Model Trees (LMT) was used to understand the association between attributes and grades after key performance-influencing factors were determined through a subjective process. A real-world dataset was used to test the suggested model with a predictive accuracy, providing guidance for parents, education institutions, and students to make informed decisions about continuing or quitting a program [10].

III. Methodology

A. Data Collection

Recently conducted a study in which the collected data from students of different departments using Google Forms and Google Sheets. Our goal was to predict the performance of individual students and determine whether they would pass or fail their courses. To gather the data, a survey was created using Google Forms and distributed it to students from a variety of departments. The survey included questions about the students' study habits, course workload, and other factors that believed to impact their academic performance. Once the data was collected, it was entered into Google Sheets and began analyzing it. Each student's performance was accurately predicted after carefully examining the data.

In many cases, our predictions matched up with the actual outcomes, which suggests that our methodology was effective. In conclusion, our study shown that it is feasible to estimate each student's performance using information gathered via Google Forms and Google Sheets. Although there are undoubt-edly additional factors that might affect a student's academic success, our results indicate that these tools can be helpful in identifying students who may be at risk of suffering in their courses.

B. Data Description

TABLE I
DATA DESCRIPTION

Feature	Data Type	Description
Gender	Binary	Students gender: ”0” - female or ”1” - male
Age	Numeric	Student’s age (numeric: from 18 to 23)
SSLC	Numeric	Percentage (numeric: 0-9)
Plus Two	Numeric	Percentage (numeric: 0-9)
Relationship	Binary	Binary: ”0” - committed or ”1” - single
Annual Income	Binary	Binary: 1 - ”above one lakh”, 0 - ”below one lakh”
Frequent Sickness	Binary	Binary: yes ”1” or no ”0”
Assignment Submission	Numeric	Assignment submission status numeric: 1”after the deadline” , 2”deadline”, 3”before the deadline”
Self Learner	Binary	Binary: yes ”1” or no ”0”
Study Hour	Numeric	Study time (numeric: 0 to 30 minute, 30 minute to 1:00 hr, 1:00 to 2:00 hr, 2:00 to 3:00 hr, 3:00 to 4:00 hr, 4:00 to 5:00 hr,5:00 to 6:00 hr)
PCM	Numeric	Percentage (numeric: 0 -9)
Sleep Hour	Numeric	Sleeping hour (numeric: 0 to 1:00 hr , 1:00 to 2:00 hr, 2:00 to 3:00 hr, 3:00 to 4:00 hr, 4:00 to 5:00 hr, 5:00 to 6:00 hr, 7:00 to 8:00 hr,)
Interest in B. Tech Studies	Binary	Binary: yes ”1” or no ”0”
Insta Reels Spend	Numeric	Reels spending time (numeric: 0 to 30 minute, 30 minute to 1:00 hr, 1:00 to 2:00 hr, 2:00 to 3:00 hr, 3:00 to 4:00 hr, 4:00 to 5:00 hr>)
Activities	Binary	Extra-curricular activities (binary: yes or no)
Exam Fear	Binary	Binary: yes ”1” or no ”0”
Study Preparation	Numeric	Numeric(1 -”day” , 2- ”week”,3 ”month”)
Programming	Binary	Programming interest (binary: yes ”1” or no ”0”)

C. Classification Algorithms

1) Logistic Regression: One of the machine learning techniques for predicting student academic achievement is logistic regression. It is a statistical method used to predict a binary outcome (yes/no, pass/fail, etc.) based on one or more independent variables. In the case of student performance analysis, the binary outcome would be whether a student passes or fails a course, and the independent variables could be factors such as study hours, internet usage, and hours of sleep.

In Logistic Regression, the relationship between the dependent variable (student performance) and independent variables (study hours, internet usage, and hours of sleep) is modeled using a logistic function. In order to predict the student’s performance, one can use the logistic function’s output of a probability between 0 and 1. To enhance the model’s fit to the data, the logistic regression algorithm modifies the independent variable coefficients.

The use of Logistic Regression in student performance analysis provides valuable insights into the factors that impact student performance. By examining the coefficients of the independent variables, educators can determine which factors have the greatest impact on student performance and use that information to make targeted interventions. The results of Logistic Regression can also be used to identify students who are at risk of failing and provide them with targeted support to improve their performance.

2) Support Vector Machine: Support Vector Machines (SVM), a type of machine learning technique, is used for classification and regression analysis. SVM can be used to evaluate a huge amount of student data and spot trends in performance when predicting student academic success. Once future performance is predicted using this data, instructors can identify students who may be at danger of falling behind.

The support vectors, or data points nearest to the border, are used by SVM to create a boundary that divides the data into multiple classes. SVM then uses this boundary to classify new data into the correct class. In the context of student performance prediction, this means that SVM can analyze a student’s previous performance data and classify them as a high-performing or low-performing student, allowing educa-tors to provide targeted support accordingly.

The use of SVM in student performance prediction holds great promise for improving the overall learning experience and closing the achievement gap. SVM’s ability to analyze large amounts of data and identify patterns that may not be noticeable to humans leads to more accurate predictions and a better understanding of individual student needs.

3) Decision Tree: A strategy for supervised learning for classification and regression analysis is the decision tree algorithm. In student performance prediction, the decision tree is trained on a set of input features and the corresponding target variables (student performance). The algorithm then creates a model of choices and results that resembles a tree and can be used to forecast student performance.

Each internal node of the tree represents a decision based on one of the input features. The branches from the node represent the possible consequences of the decision, and the leaf nodes represent the predicted target variable (student performance). By following the decisions and consequences, the decision tree can predict student performance for new data. The decision tree algorithm is particularly useful in student performance prediction because it can handle non-linear relationships be-tween input features and the target variable, and it is easy to interpret and understand. Additionally, it can handle both categorical and numerical data, making it versatile for a wide range of student performance data.

Another advantage of the decision tree algorithm for student performance prediction is its ability to handle missing values in the input features. In real-world scenarios, student performance data may not always be complete, and some features may have missing values. Decision trees can handle such missing values by using statistical measures such as mean or median imputation, which replace missing values with the average or median value of the available data for that feature. This enables the algorithm to continue training and making predictions even with incomplete data, making it a valuable tool for student performance prediction in practical applications.

4) Random Forest: Using a machine learning process called Random Forest, numerous decision trees are built, and the results are used to predict student academic success. Each subset of characteristics is chosen at random, and a decision tree model is trained on each subset. The outputs of all the decision trees are then combined to make a final prediction.

Random Forest reduces the overfitting that often occurs in decision trees by combining the results of multiple trees, which reduces the variance in the predictions. This leads to more robust and accurate predictions of student performance, which can be used to provide targeted support to struggling students and improve the overall learning experience.

Another advantage of the Random Forest algorithm for student performance prediction is its ability to identify the most important features for predicting student performance. Random Forest calculates a measure of the importance of each feature in the model, which can be used to prioritize interventions and support for students based on the factors that have the greatest impact on their academic success. This can help educators and policymakers make more informed decisions about resource allocation and intervention strategies, leading to better outcomes for students. Overall, Random Forest is a powerful machine learning technique that can be used to improve student performance prediction and support the success of all learners.

D. System Framework

The Fig. 1 outlines the process of building and deploying a machine learning model using different algorithms. It all starts with the initial dataset that will be used to train the model. A subset of the dataset, known as the train dataset, is selected to train the machine learning model. The accuracy and effectiveness of the trained model are tested on another separate subset of the dataset known as the test dataset.

Various machine learning algorithms such as Support Vector Machine, Logistic Regression, Decision Tree, Random Forest, and Voting Classifier can be used to train the model. Support Vector Machine is a commonly used algorithm for classifi-cation and regression analysis, while Logistic Regression is often used for classification analysis. Decision Tree works by creating a tree like model of decisions and their possible consequences, while Random Forest is a collection of decision trees used for classification, regression, and other tasks. Voting Classifier is an ensemble learning algorithm that combines the predictions from multiple algorithms to improve the accuracy of the final result.

The next step is building the model by training it using the selected algorithm and the train dataset. Once the model has

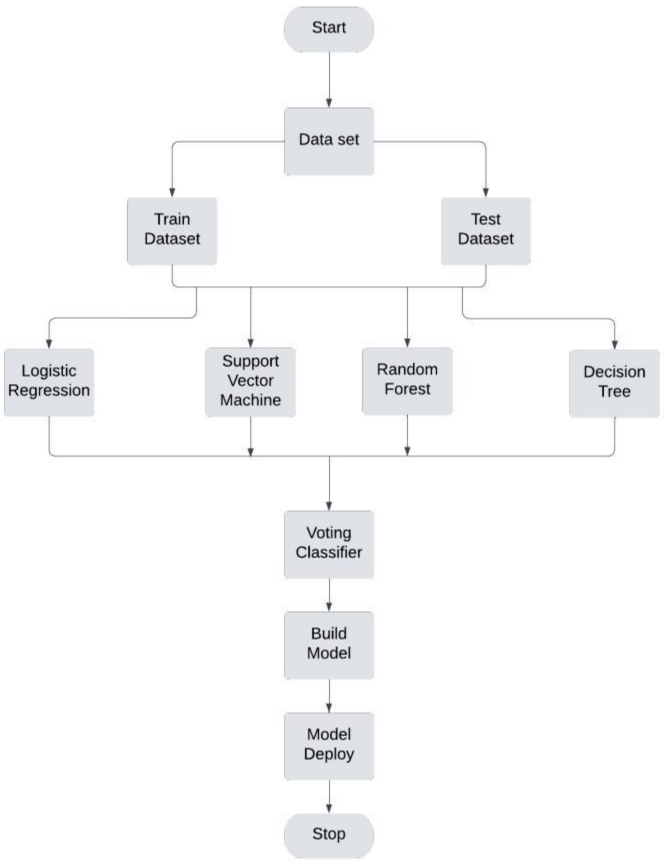

Fig. 1. System Framework Flow Chart

been built and tested, it can be deployed for use in prediction and analysis. However, it is important to note that the accuracy and effectiveness of the machine learning model depend on the algorithm and dataset used. The process stops when the model is built and deployed for use.

Data Collection : The process of creating a machine learning model requires the collecting of data as a crucial stage. It involves gathering and acquiring the data that will be used to train and test the model. This data can come from a variety of sources, such as databases, web scraping, or user input. Once the data has been collected, it is important to perform data preprocessing. This involves cleaning and preparing the data for use in the model. Data preprocessing can include tasks such as filling in missing values, normalizing numerical data, and encoding categorical data.

Training and Testing: The data must then be divided into a training dataset and a testing set as the next stage. The test set is used to assess the model’s performance after it has been trained on the train set. It is important to have a separate test set to make sure the model is not overfitted to the training data. Once the data has been collected, preprocessed, and split into a train set and test set, it is time to build the model. There are several types of models that can be used, such as support vector machines (SVMs), logistic regression, decision trees, and random forests. Each type of model has its own strengths and weaknesses, and the appropriate model will depend on the specific problem at hand. Once the model has been built and trained on the train set, it is time to deploy it. This involves making the model available for use in real-world applications, such as predicting outcomes or classifying data. Deployment can involve integrating the model into an existing software system or building a standalone application that utilizes the model.

In summary, building a machine learning model involves several steps, including data collection, data preprocessing, splitting the information into a test set and a train set, building the model, and deploying the model. Each of these steps is crucial to creating a successful machine learning system.

System with Voting Classifier: Use a voting classifier in the process of building a machine learning model, the outcome would depend on the specific classifiers that are using and how well they perform on the training and test data. If the individual classifiers are able to make accurate predictions on their own and their predictions are reasonably consistent with each other, then the voting classifier may be able to make more accurate predictions than any of the individual classifiers. However, if the individual classifiers have poor performance or their predictions are highly inconsistent with each other, then the voting classifier may not provide any improvement over the individual classifiers.

E. Model Improvement

1) Voting Classifier: A Voting Classifier is an ensemble learning method that combines multiple base models to form a prognosis with more accuracy. In the context of student performance analysis, a Voting Classifier can be used by combining multiple base models that use different algorithms to predict student performance. By adding together the results of the base models' forecasts and casting a majority vote, the final prediction is determined.

This helps to address the limitations of individual algorithms and reduce the risk of overfitting. The use of a Voting Classifier in student performance prediction can lead to a more robust and accurate prediction, improving the overall learning experience and closing the achievement gap between high- and low-performing students.

IV. Implementation

Data collection, preprocessing, and modelling are steps in the machine learning process used to evaluate student academic performance. The data collected contains various features of students such as attendance, marks, age, gender, etc. These features are converted into numerical values using python libraries such as numpy, pandas, and seaborn for data preprocessing and visualization.

Machine learning techniques are used to examine the student performance once the data has been preprocessed.

Logistic regression, SVM, decision trees, and random forests are some of the techniques employed in this investigation. These models are trained and evaluated using K-fold cross-validation, and the effectiveness of the models is assessed using the analysis of the findings. The voting classifier is also used to combine the predictions of multiple models to increase accuracy.

Model Deployment: The final step in the student academic performance analysis process is deployment. The models are deployed using streamlit and joblib libraries to create an interactive web application for accessing the models. The joblib library is used to store the trained models, and streamlit is used to create an interface for accessing the models. This allows users to easily access the models and make predictions about student performance.

Using machine learning to evaluate student academic perfor-mance is a challenging procedure that necessitates a thorough understanding of data analysis and machine learning tech-niques. The use of python libraries and techniques such as k-fold cross-validation, voting classifier, and model deployment allows for accurate predictions about student performance. The final result is a user-friendly web application that provides insights into student performance and helps predict future performance.

V. Result And Discussion

A. Perfomance Matrices

TABLE II
Perfomance Matrices

Model Name	Precision	Recall	F1-Score
Logistic Regression	0.88	0.88	0.88
Support Vector Machine	0.81	1	0.9
Decision Tree	0.75	0.66	0.75
Random Forest	1	0.88	0.94
voting classifier	0.91	0.91	0.91

The precision, recall, and F1 score for the logistic regression model were all 0.88. The SVM model showed 0.8 precision, 1 recall, and 0.9 F1 score. The decision tree model's accuracy, recall, and F1 score were all 0.75. Another SVM model achieved an F1 score of 0.94, a precision of 1, and a recall of 0.88. The precision, recall, and F1 score for the voting classifier model were all 0.91.

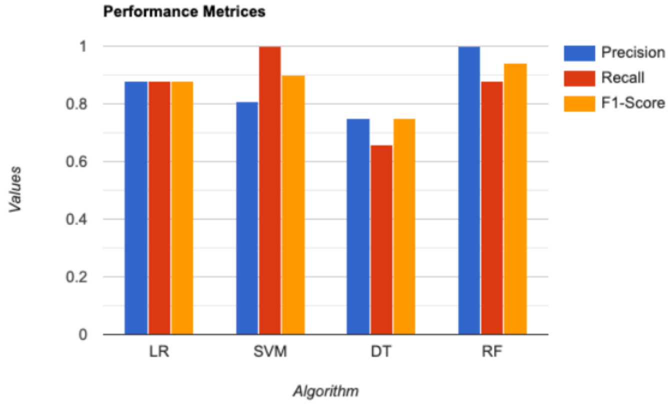

Fig. 2. Classification Perfomance Matrices

B. Confusion Matrix

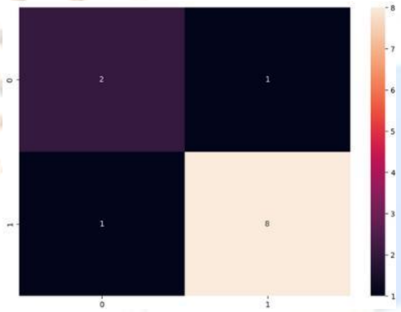

Fig. 3. Logistic regression

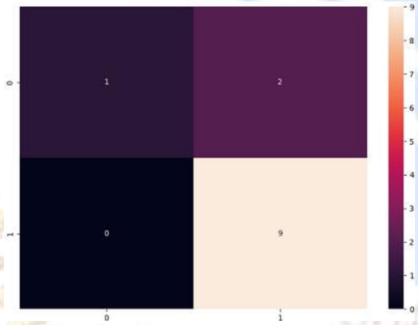

Fig. 4. Support Vector Machine

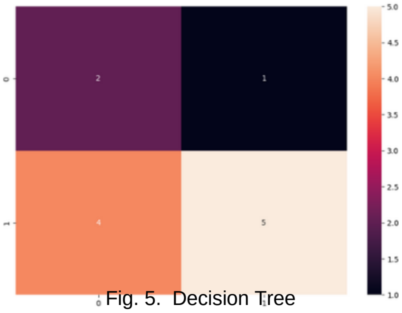

Fig. 5. Decision Tree

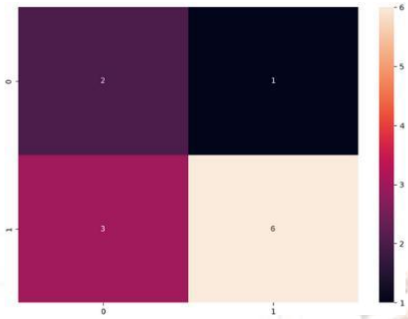

Fig. 6. Random Forest

The findings from the study suggest that data science and machine learning techniques can provide valuable insights into student academic performance and help to identify areas for improvement. Key factors found to be important in the prediction were study hours, internet usage and hours of sleep. However, the limitations of the study, including the small sample size and specific educational context, should be considered in interpreting the results. Further research is needed to explore the generalizability of these findings and the potential applications of these techniques in other educational settings.

The use of data science and machine learning techniques in education has raised concerns about student privacy and data security. To address these concerns, it is crucial to implement robust data governance policies that safeguard student data and ensure that it is used only for its intended purposes. It is also essential to ensure that students are informed about the use of their data and have the right to opt-out if they choose to do so. By adopting a transparent and ethical approach to data science and machine learning in education, can harness the power of these technologies to enhance student learning and promote academic success.

VI. Conclusion And Future Work

There are many advantages for the educational system in using machine learning algorithms to predict student academic performance. By accurately predicting a student's performance, teachers can identify and help those who might be at risk of falling behind, enhancing the entire learning process and narrowing the achievement gap between high- and low-performing students. Utilizing technology and study tools, such as machine learning, can also assist students in learning about computers and enhance their educational experience. With a clear understanding of the factors that impact student performance, and the use of a large, diverse dataset, the development of predictive models can enable educators to make accurate predictions about student success, identify and correct performance faults, and ultimately help students achieve their academic goals.

Future work in the area of utilising machine learning techniques to predict student performance may examine new and reducing methods to increase prediction accuracy. An-other area of focus could be expanding the scope of data analyzed to include a wider range of factors that impact stu-dent performance, such as socio-economic status and learning environment. Additionally, exploring ways to incorporate the predictions generated by these algorithms into educational practices and decision making can also be an important aspect of future work.

Acknowledgment

We would like to thank everyone who took part in the study because it wouldn’t have been possible without them. Also like to thank the institution for providing the necessary resources and support to conduct the study. According to the study’s findings, data science and Approaches to machine learning can be used to forecast students academic success. It is hoped that these findings will advance the area of education and enhance students’ learning opportunities.

References

[1] H. Gull, M. Saqib, S. Z. Iqbal and S. Saeed, ”Improving Learning Experience of Students by Early Prediction of Student Performance using Machine Learning,” 2020 IEEE International Conference for Innovation in Technology (INOCON), 2020.

[2] M. R. Rimadana, S. S. Kusumawardani, P. I. Santosa and M. S. F. Erwianda, ”Predicting Student Academic Performance using Machine Learning and Time Management Skill Data,” 2019 International Sem-inar on Research of Information Technology and Intelligent Systems (ISRITI), 2019.

[3] F. J. Kaunang and R. Rotikan, ”Students’ Academic Performance Prediction using Data Mining,” 2018 Third International Conference on Informatics and Computing (ICIC), 2018.

[4] K. Al Mayahi and M. Al-Bahri, ”Machine Learning Based Predicting Student Academic Success,” 2020 12th International Congress on Ul-tra Modern Telecommunications and Control Systems and Workshops (ICUMT), 2020.

[5] E. S. Bhutto, I. F. Siddiqui, Q. A. Arain and M. Anwar, ”Predict-ing Students’ Academic Performance Through Supervised Machine Learning,” 2020 International Conference on Information Science and Communication Technology (ICISCT), 2020.

[6] C. Jalota and R. Agrawal, ”Analysis of Educational Data Mining using Classification,” 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), 2019.

[7] Kumar, Surjeet Pal, Saurabh, ”Data Mining: A Prediction for Per-formance Improvement of Engineering Students using Classification”, World of Computer Science and Information Technology Journal, 2012.

[8] Zhang, Chongsheng Zhang, Yuan Shi, Xianjin Almpanidis, ”Data Min-ing A Prediction for Performance Improvement of Engineering Students using Classification”, World of Computer Science and Information Technology Journal, 2012.

[9] Hashmia Hamsa, Simi Indiradevi, Jubilant J. Kizhakkethottam, ”Student Academic Performance Prediction Model Using Decision Tree and Fuzzy Genetic Algorithm”, Procedia Technology, Volume 25, 2016.

[10] F. Aman, A. Rauf, R. Ali, F. Iqbal and A. M. Khattak, ”A Predictive Model for Predicting Students Academic Performance,” 2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA), 2019.