Evaluation of Student's Performance Using Decision Tree Model

DR. SANTOSH SINGH[1], SHERILYN KEVIN[2], KUNIKA JANGID[3], ABHISHEK SINGH[4]

[1] *Head of Department / Department of IT, Thakur College of Science and Commerce*
[2] *Assistant Professor /Department of IT, Thakur College of Science and Commerce*
[3, 4] *PG Student, Department of IT, Thakur College of Science and Commerce*

Abstract— Predicting and evaluating student academic performance is a crucial job for educational institutions since it enables teachers to modify their methods and help problematic pupils on time. In recent years, deep learning models have demonstrated impressive skills in a range of disciplines, including education. This study presents an in-depth analysis and assessment of the application of machine learning techniques for predicting and evaluating student academic achievement. The study starts with a discussion of the significance of accurate performance evaluation and its influence on students' learning experiences. Traditional evaluation systems usually fail to capture the intricacies of individual learning processes and appropriately predict future performance. Machine learning algorithms, which use complex patterns and correlations in educational data, provide a viable alternative. Machine learning models are a broad class of computing methods that play an important role in artificial intelligence. These models employ a variety of methodologies, ranging from supervised learning, in which they learn to map inputs to labelled outputs, to unsupervised learning, in which hidden patterns in unlabeled data are discovered. Semi-supervised learning bridges the gap by using both labelled and unlabeled data. Furthermore, reinforcement learning models teach agents how to make decisions and conduct actions by interacting with their surroundings, which is a basic notion in AI research.

Indexed Terms: Student Assessment, Neural Networks, Classification Models, Regression Models, Predictive Modeling.

I. INTRODUCTION

Evaluating student academic performance is a critical component of education systems around the world. The application of machine learning models to improve student academic evaluation has become increasingly popular in today's educational environment. With the help of these models, large volumes of student data may be effectively analyzed and interpreted, yielding insightful information on academic performance, learning trends, and possible areas for development. Educators and institutions may customize personalized learning experiences and support systems for students by leveraging the power of machine learning to make data-driven decisions.

This study is to investigate the use of several machine learning approaches in assessing students' academic performance, emphasizing the importance, difficulties, and possible ramifications for educational establishments. In order to support academic achievement, educators may anticipate student performance, identify at-risk children, and provide focused interventions by utilizing predictive analytics, classification, and clustering algorithms. Furthermore, the use of machine learning models into academic evaluation can aid in the creation of adaptive learning systems, promoting an active and flexible learning environment that accommodates a range of learning styles and demands.

This study aims to provide a complete knowledge of the function of machine learning techniques in altering student academic evaluation through an in-depth analysis of the different machine learning approaches. Furthermore, the purpose of this research is to give educators and policymakers with beneficial advice on how to efficiently implement machine learning models

into present educational frameworks, therefore creating an academic atmosphere that values data-driven decision-making and continual progress. Furthermore, the incorporation of machine learning models into academic evaluation can contribute in the development of adaptive learning systems, fostering an active and adaptable learning environment that caters to a variety of learning styles and needs.

II. LITERATURE REVIEW

Samah Fakhri Aziz in this research has taken a custom dataset with 1040 instances and 17 attributes to which the student performance evaluation has been done. The data is pre-processed and the lost values are checked. The algorithm taken is random forest and gradient boosting decision tree, the applied data are divided into 80% and 20 % respectively. The results of each algorithm are analyzed and compared in terms of accuracy, prediction and error rate [1]. In this paper the "plot importance" function in Scikit-learn library help in plotting the important features that affect student final grade to get an accurate evaluation of our model, the dataset containing 1044 students is split into train and test dataset in 70% to 30% ratio using the 'train_test_split' function in sci-kit learn. This study used several ML classification models to predict student academic performance. Results showed a range of accuracy from 33% to 98% and a Range of cross-validation from30% to 37%. The XGBoost Model is the most suitable ML model by achieving 97.12%accuracy and 35.67% cross-validation [2]. In this research paper decision tree, K Nearest Neighbor, Entropy, binomial regression, linear regression are used to implement the system. The proposed system handles the student individual marks that include 10th, 12th mark and their semester mark. This method can be used by instructors to evaluate and schedule better methods for improving their academics. The accuracy of binomial logistic regression is 97.05%, Decision Tree 88.23%, Entropy 91.19% and K-NN 93.71% [3] . In this project, the data were taken from the Student Information System (SIS), where all student records are stored at a State University in Turkey. RF, NN, LR, SVM, NB and KNN were employed to predict students' academic performance. According to this result, it can be said that students' midterm exam grades are an important predictor to be used in predicting their final exam grades. The proposed model predicted the final exam grades of students with 73% accuracy. According to this result, it can be said that academic achievement can be predicted with this model in the future [4]. In this study, ML classification models were designed and assessed using a dataset obtained from an e-learning environment. This study employed the knearest neighbour (k-NN) approach to classify the user knowledge classes. There were 403 instances, or records, in the dataset, each with values for the different attributes. This paper suggested a classification approach for student's knowledge according to their level of expertise, depicted. The dataset used in this work serves as the source data for a novel model based on ML methods that forecasts student performance on the knowledge assessment [5]. There are students that drop out of university programs in the initial first year. According to a study in 2018, Sweden had a dropout percentage of 29% of full-time bachelor's students. This study used a quantitative method to approach and answer the research questions. Upon completion of the project, several aspects were looked at. The first aspect was reviewing how well defined were the goals and research. The technical purpose of this study was to evaluate two or more ML algorithms for performance in terms of accuracy, precision, recall, f1 score, and prediction [6]. The KDD methodology is employed in this study to conduct the experiments. This method is widely used by scientific researchers in the data mining field. The dataset used in this study was provided by the Deanship of Admission and Registration databases at SQU. the sample consisted of 33 features and 37,599 students, and the data represented student cohorts from fall 2009 to summer 2019. The optimal individual algorithm is with an accuracy of 82.4% [7]. In this paper the education data is used from the different colleges and the student data is obtained. The dataset has 649 samples (students) and 33 variables for evaluation purpose 70 and 30 percent data is used as training and testing .In this the variable is defined and it is referred as pass and fail parameter. The results in this case were also satisfactory, with the best algorithm based on the AUC metric being Random Forest (as compared to SVM and kNN). It mostly concerns classifying a sample in the immediately neighboring classes, with almost 81.3% of data points getting classified in the correct class [8]. In this paper research was conducted by the Kingdom of Saudi Arabia to

ascertain the pre- and post-COVID-19 conditions of online learning. More than 50% of students with "unsatisfactory" marks have seen their performance improve, which is the most notable change. The "Maharat" platform applies a survey of quality assurance of online training program standards in consideration. The predictions of academic achievement using SVM and three non-ensemble approaches (Naive Bayes, k-Nearest Neighbor, and Decision Tree) are evaluated [9]. In this research the three data modules are defined, according to the high low discrimination performance is analyzed in the term of high low data points. The features blocked in dashed boxes (including LyE, HurstE, DFA, and LSTMbased features) are proposed in this study, Outcome-based education (OBE) has emerged as a new school of thought in education and has recently enjoyed wide acceptability and adoption. The knowledge obtained in this study can also potentially contribute to related research among K-12 students [10]. The methods presented in have two drawbacks for students' answer script evaluation and introduced a new approach for evaluating students' answer scripts using fuzzy numbers associated with degrees of confidence of the evaluator. It has used two methods that are bootstrapper and convex normalized fuzzy numbers. A typical execution of fuzzy rules for the first-level node Category 11 in hierarchical FRBS [11]. The present study aimed to check the relationships between SET and academic achievement, starting from the knowledge offered by previous studies. This study is carried out in a different context to most previous works. The sample included 1,538 students of the National Polytechnic School from Ecuador, enrolled in eight different faculties and schools and studying 28 different degrees. Of these students, 68.6% were male and 31.4% were female. The total percentage of variance explained from the final academic achievement at the level of the students was 25%, while at the level of the sections, it was 33% [12].

.

III. METHODOLOGY

3.1 Algorithm

A decision tree is a versatile and intuitive machine learning model that is widely used for both classification and regression tasks. It's a graphical representation of a decision-making process, which resembles an inverted tree with branches and leaves. At its core, a decision tree partitions data based on a series of binary decisions. It starts with a "root" node representing the entire dataset and then recursively splits the data into subsets by asking questions about feature attributes. These questions, also known as "decisions," are designed to maximize the separation of data points with similar outcomes. As the tree grows, it forms internal "nodes" representing decision points and "branches" connecting nodes, showing the outcome of each decision. Finally, at the end of each branch, we find "leaf nodes" that represent the model's predictions or class labels.

An overview of the Decision Tree methodology that used in research article is provided in Fig.1
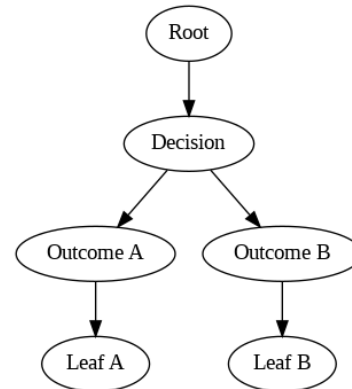

Figure 1: Decision Tree

In this representation, "Root" represents the starting point of the decision tree. The "[Decision]" node represents a binary decision based on a feature attribute. The tree branches into two possible outcomes, "[Outcome A]" and "[Outcome B]." Each of these outcomes leads to a "Leaf" node, which represents the model's prediction or class label. This is a basic representation. A Decision trees can become much more complex with multiple nodes, decisions, and leaves, depending on the data and the problem being addressed

The key advantage of decision trees lies in their transparency and interpretability. It allows users to understand the decision-making process and how specific features influence the model's predictions. This interpretability makes decision trees an invaluable tool in various domains, including finance,

healthcare, and natural language processing. However, decision trees can be prone to over fitting, which means they may create highly complex and detailed trees that perform poorly on unseen data. Techniques like pruning and ensemble methods, such as Random Forests, are employed to mitigate this issue. So basically a decision tree is a fundamental machine learning algorithm that offers a visual and straightforward approach to decision-making. Its simplicity and interpretability make it an excellent choice for tasks where transparency and insight into the model's operations are critical. While it has its limitations, it remains a valuable tool in the broader field of machine learning and data science.

The working of the Decision Tree model is shown in figure 2.
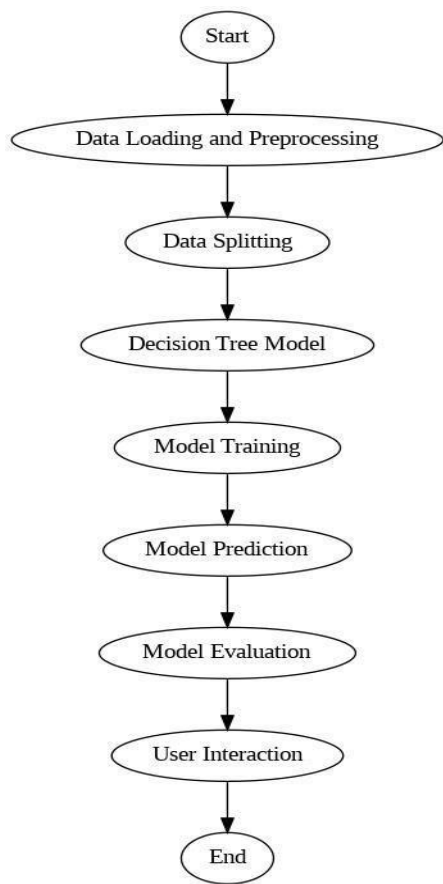
Figure 2: Working of the model

3.2 DATASET

This is a collection of educational data collected by a learning management system (LMS) called Kalboard360. Kalboard360 is a multi-agent LMS designed to facilitate learning using cutting-edge technology. Such a system gives users synchronous access to learning resources from any Internet device. A student performance tracking tool called the Experience API (XAPI) is used to collect data. The XAPI is part of the Training and Learning Architecture (TLA), which enables tracking of learning progress and learner actions, such as reading an article or watching a training video. The Experience API helps providers of learning activities define objects that describe the learner, the activity, and the learning. The dataset consists of 1032 student records and 16 characteristics. Characteristics fall into three main categories: (1) Demographic characteristics such as gender and ethnicity. (2) Academic background characteristics such as educational level, grade level and tenure. (3) Behavioural characteristics such as raising your hand in class, accessing resources, answering parent questionnaires, and being satisfied with school. The data consists of 600 men and 432 women. The students are from different origins, such as 279 students are from Kuwait, 272 students are from Jordan, 48 students are from Palestine, 42 students are from Iraq, 170 students are from Lebanon, 120 students are from Tunisia, 11 students are from Saudi - Arabic. Arabia, 40 students are from Egypt, 70 students from Syria, 69 students from the United States, Iran and Libya, 49 students from Morocco and one student from Venezuela.

Data are collected during two study periods: 500 student records are collected in the first semester and 532 student records are collected in the second semester. The data also includes an attendance indicator, so students are classified into two categories according to days of absence: 587 students are absent for more than 7 days and 445 students are absent for less than 7 days.

This dataset also includes a new main class; this quality is the birth of parents in learning. Parent participation feature have two sub features: Parent Answering Survey and Parent School Satisfaction. There are 270 of the parents answered survey and 610

are not, 422 of the parents are satisfied from the school and 188 are not.

Attributes:
- Gender - students gender (nominal: Male and Female')
- Nationality- students nationality (nominal:' Kuwait',' Lebanon',' Egypt',' Saudi Arabia',' USA',' Jordan',' Venezuela',' Iran',' Tunis',' Morocco',' Syria',' Palestine',' Iraq',' Lybia')
- Place of birth- student and; s Place of birth (nominal:' Kuwait',' Lebanon',' Egypt',' SaudiArabia',' USA',' Jordan',' Venezuela',' Iran',' Tunis',' Morocco',' Syria',' Palestine',' Iraq',' Lybia')
- Educational Stages-educational level student belongs (nominal:
- 'lowerlevel','MiddleSchool','HighSchool')
- Grade Levels- grade student belongs (nominal: 'G-01', 'G-02', 'G-03', 'G-04', 'G-05', 'G-06', 'G-07', 'G-08', 'G-09', 'G-10', 'G-11', 'G-12 ')
- Section ID- classroom student belongs (nominal:'A','B','C')
- Topic- course topic (nominal:' English',' Spanish', 'French',' Arabic',' IT',' Math',' Chemistry', 'Biology', 'Science',' History',' Quran',' Geology')
- Semester- school year semester (nominal:' First',' Second')
- Parent Responsible for student
- (nominal:'mom','father')
- Raised hand- how many times the student raises his/her hand on classroom (numeric:0-100)
- Visited resources- how many times the student visits a course content(numeric:0-100)
- Viewing announcements-how many times the student checks the new announcements (numeric:0-100)
- Discussion groups- how many times the student participate on discussion groups (numeric:0-100)
- Parent Answering Survey- parent answered the surveys which are provided from school or not (nominal:'Yes','No')
- Parent School Satisfaction- the Degree of parent satisfaction from school (nominal: 'Yes', 'No')
- Student Absence Days-the number of absence days for each student (nominal: above-7, under-7)

The students are classified into three numerical intervals based on their total grade/mark:

Low-Level: interval includes values from 0 to 50,
Middle-Level: interval includes values from 51 to 75,
High-Level: interval includes values from 76-100.

	gender	NationaliT	PlaceofBir	StageID	GradeID	SectionID	Topic	Semester	Relation	raisedhan	VisiTedRe	Announce	Discussion	ParentAnsw	Parentschoi	StudentAbsence
2	M	KW	KuwaiT	lowerleve	G-04	A	IT	F	Father	15	16	2	20	Yes	Good	Under-7
3	M	KW	KuwaiT	lowerleve	G-04	A	IT	F	Father	20	20	3	25	Yes	Good	Under-7
4	M	KW	KuwaiT	lowerleve	G-04	A	IT	F	Father	10	7	0	30	No	Bad	Above-7
5	M	KW	KuwaiT	lowerleve	G-04	A	IT	F	Father	30	25	5	35	No	Bad	Above-7
6	M	KW	KuwaiT	lowerleve	G-04	A	IT	F	Father	40	50	12	50	No	Bad	Above-7
7	F	KW	KuwaiT	lowerleve	G-04	A	IT	F	Father	42	30	13	70	Yes	Bad	Above-7
8	M	KW	KuwaiT	MiddleScl	G-07	A	Math	F	Father	35	12	0	17	No	Bad	Above-7
9	M	KW	KuwaiT	MiddleScl	G-07	A	Math	F	Father	50	10	15	22	Yes	Good	Under-7
10	F	KW	KuwaiT	MiddleScl	G-07	A	Math	F	Father	12	21	16	50	Yes	Good	Under-7

Figure 3: Student Dataset

IV. RESULTS

In this work, 1032 data were used to train the model. In the following project the main objective is to monitor the student data and analyse and give the result whether the student is performing well or not. The result can be presented in the various forms such as letter grades, numerical scores, or percentage depending upon the grading system used by the educational institute. The project's interactive feature, allowing users to input student attributes and obtain predictions using different models, enhances its usability and practicality. This feature facilitates the exploration of various scenarios and the assessment of the model effectiveness in real-time situations. The dataset undergoes pre-processing, which includes encoding categorical variables, shuffling the data, and splitting it into training and testing sets. This preparation step is crucial for ensuring the robustness and reliability of the models' predictions.

The dataset splits into a training set (70%) and a testing set (30%). A Decision Tree Classifier is instantiated and trained on the training data, and predictions are made on the test data. The code calculates and reports the accuracy of the Decision Tree model using scikit-learn's classification report and accuracy calculation. The model's precision, recall and f1-score for classifies the performance respectively as shown in Fig.4. The Confusion matrix shown in Fig.5 helps in understanding the classification of student's performance dataset.

Figure 4: Precision, Recall, Support and F1-Score

Further, the script offers an interactive option for users to input their personal attributes, such as gender, nationality, grade, and more. Based on these attributes, the trained Decision Tree model predicts the students' performance, categorizing it as "Excellent," "Good," or "Bad". In summary, this code provides a comprehensive machine learning solution for educational data analysis, where a Decision Tree classifier is used to predict students' performance levels. The model gives an accuracy of 99.07% for detecting the performance is well or not.
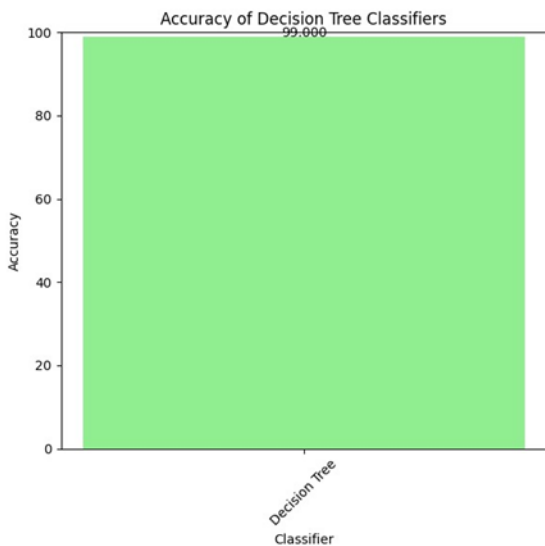
Figure 5: Accuracy vs Decision Tree

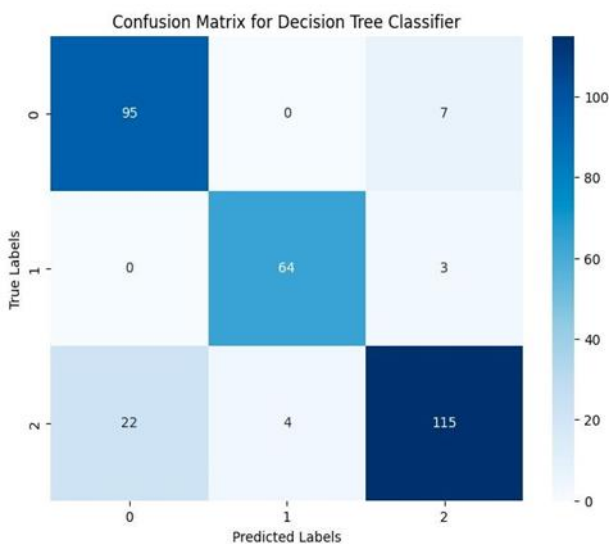Figure 7: Student's Performance
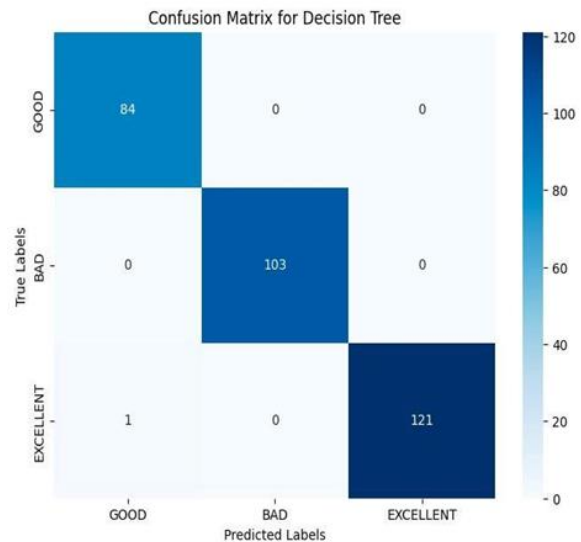

Figure 6: Confusion Matrix of Decision Tree

Figure 8: Confusion Matrix of Student's Performance

CONCLUSION

This study has conducted extensive research to explore how machine learning can empower educators with informed decision-making. Predicting student academic performance is a critical aspect of addressing the ongoing academic performance challenges among students. The ,study employed several machine learning classification models to forecast student academic performance, resulting in a range of accuracy rates, with the highest achieving 99%, and a range of cross-validation results, which reached up to 37%. Notably, the Decision Tree model emerged as the most suitable machine learning model, boasting an accuracy rate of 97.12% and a cross-validation rate of 35.67%.Furthermore, the findings of the study underscored the significance of various factors that influence student academic performance, including the number of school absences, current health status, social activities such as spending time with friends, the availability of free time after school, and the quality of family relationships.

This machine learning solution has the potential to serve as a valuable tool for educational institutions, teachers, and students alike. It offers insights into the key determinants of academic success, providing a systematic and data-driven approach to comprehend and predict student outcomes. These insights can play a pivotal role in facilitating well-informed decisions and interventions aimed at supporting and enhancing academic achievement.

COMPLIANCE AND ETHICAL STANDARDS

REFRENCES

[1] Samah Fakhri Aziz "Students' Performance Evaluation Using Machine Learning Algorithms"

[2] Opeyemi Ojajuni1, Foluso Ayeni2(B), Olagunju Akodu3, Femi Ekanoye4, Samson Adewole4, TimothyAyo4, Sanjay Misra5, and Victor Mbarika6 "Predicting Student Academic Performance Using Machine Learning".

[3] J.Dhilipan1, N.Vijayalakshmi2, S.Suriya3, Arockiya Christopher4 "Prediction of Students Performance using Machine learning".

[4] Mustafa Yağcı "Educational data mining: prediction of students' academic performance using machine learning algorithms".

[5] Nuha Alruwais 1 andMohammed Zakariah 2 "Evaluating Student Knowledge Assessment Using Machine Learning Techniques".

[6] Alabbas Hayder "Predicting Student Performance Using Machine Learning: A Comparative Study Between Classification Algorithms".

[7] Lamees Al-Alawi, Jamil Al Shaqsi, Ali Tarhini & Adil S. Al-Busaidi "Using machine learning to predict factors affecting academic performance: the case of college students on academic probation".

[8] Charalampos Dervenis,Vasileios Kyriatzis,Spyros Stoufis, Panos Fitsilis - "Predicting Students' Performance Using Machine Learning Algorithms".

[9] Mohammed Nasser Alsubaie "Predicting student performance using machine learning to enhance the quality assurance of online training via Maharat platform".

[10] Ashmina Khan, Prof. K. N. Hande "A survey on Prediction and Analysis of Students Academic Performance Using Machine Learning Technique" .

[11] Satish S. Salunkhe EMAIL logo, Ashok Deshpande and Yashwant Joshi "Degree of Certainty in Students' Academic Performance Evaluation Using a New Fuzzy Inference System "

[12] Tarquino Sánchez Raquel Gilar-Corbi, Juan-Luis Castejón Jack Vidal and Jaime León

[13] "Students' Evaluation of Teaching and Their Academic Achievement in a Higher Education Institution of Ecuador ".