Shodh Sari-An International Multidisciplinary Journal

2024, Vol. 03, Issue 04, 129-140
ISSN: 2959-1376
DOI: https://doi.org/10.59231/SARI7751

Classifying Student Performance: An In-Depth Analysis Using Machine Learning Algorithms

C, Suhasini[1]

[1]Research Scholar, Statistician, JSS AHER

B, Madhu[2]

[2]Professor & Deputy Dean (Research), JSS AHER

Abstract

In recent years, analyzing and predicting student performance has been a significant challenge for educational institutions. Early analysis of student data can reveal their strengths and weaknesses, aiding in the improvement of examination outcomes. Machine learning offers a powerful tool for predicting student performance by utilizing demographic and academic data. We can forecast the results and determine which classification algorithm has the highest accuracy from algorithms such as support vector machines, J48 decision trees, naive bayes, and simple logistic regression. As a result, the accuracy level is addressed in the conclusion of data statistics, featuring classification models. By analyzing the classification algorithms, the logistic regression approach provides a more efficient way for educators to identify patterns and trends in the performance of students. This study enables targeted processes for students who would normally perform poorly on examinations.

Keywords: machine learning, student performance, algorithm, examination

Introduction:

Assessment and evaluation methods need to be implemented in various ways to assess students' learning performance in educational institutions. To enhance educational outcomes, programs seek to provide an accurate overview of student success to promote data-driven decision-making. At significant moments in a course, assessments are crucial to providing evidence of student learning. For performance analysis and prediction, important attributes and the

@2024 International Council for Education Research and Training
ISSN: 2959-1376
2024, Vol. 03, Issue 04, 129-140
DOI: https://doi.org/10.59231/SARI7751

previous records of students are gathered. Subsequently, various data mining techniques and classification algorithms are applied to get deeper insights and predictions. The student's progress and adaptation of instruction are assessed by employing a variety of assessment tools, such as conversations, observations, and their academic results.

Due to a lack of prediction accuracy, improper attribute analysis, and insufficient datasets, the educational system is facing difficulties and challenges to effectively benefit from higher education. Hence, they focus on assessing students' achievement of overall expectations. By analyzing student performance data, educators can identify areas of strength and weakness, set achievable academic goals, and provide personalized support to students. To ensure the accuracy and effectiveness of assessment and evaluation strategies, it is crucial to align them with established learning standards and performance criteria. By providing students with clear expectations and rubrics, we empower them to take an active role in their learning and engage in self-assessment processes.

Student learning performance requires a comprehensive approach that integrates various assessment strategies, aligns with learning standards, involves students in self-assessment, and utilizes data analysis to drive continuous improvement and enhance educational outcomes. In recent years, various data mining techniques and classification algorithms have been used, such as Naïve Bayes, decision trees, neural networks, outlier's detections, and advanced statistical techniques. These techniques are applied to the student data to get information, to help in decision making support systems, pattern extraction, etc. Commonly, a student's academic performance is measured by their previous CGPA, but there are various other important attributes that affect the overall performance of their academic results.

The problem of accurate student performance prediction is still a challenging task due to various issues, and many other factors are involved in it. The main issues in the performance prediction methods are inefficiency and the use of improper attributes or variables. The objective of this survey is to conduct a comparative study and

@2024 International Council for Education Research and Training
ISSN: 2959-1376
2024, Vol. 03, Issue 04, 129-140
DOI: https://doi.org/10.59231/SARI7751

provide the best methods for performance prediction after analyzing the recent studies.

It represents a summary of previous literature from 2019 to 2023 that includes decision trees, KNN classifiers, the J48 algorithm, logistic regression, neural networks, clustering and classification, and NB tree classification algorithms.

Literature Review:

1. Chitra Jalota et al. (2019). In this paper, they discuss the use of data mining techniques in the educational field to predict student outcomes using classification algorithms like Navies Bayes, J48, and multilayer perception using the Weka tool for data analysis. The study is to compare the different classification algorithms and conclude that a multilayer perception classifier is the best to predict student performance. For further research, they preferred classifiers and clusters to enhance the accuracy and speed of data mining techniques in higher education.

2. Ali Salah Hashim et al. (2020). This study focuses on forecasting student success in higher education by identifying students at risk of falling before final examinations. They use several algorithms like decision trees, naïve bayes, and logistic regression to predict student performance, highlighting that logistic regression achieved the highest accuracy with 68.7%. For further study, they suggest that educational institutions develop policies to support students who are likely to fail and improve their overall academic performance.

3. Dhilipan et al. (2021). They explained in this paper the development of a machine learning model to predict student performance in higher education. They used various techniques like binomial, logistic regression, entropy, and KNN classifiers using student attributes like 10^{th} score, 12^{th} score, and previous semester results to predict their final grades. This study aims to assist students, educators, and academic institutions by providing insight and suggesting improvements in academic progress.

4. Sarah Alturki* et al. (2021). They discussed using educational data mining (EDM) to predict student performance. They include more distinctive attributes like gender, attendance, and student satisfaction feedback to improve accuracy and enhance learning outcomes. They are using classification algorithms. Navie Bayes and Random Forest were the most effective predictive methods to predict the student's academic performance for early intervention.

@2024 International Council for Education Research and Training
ISSN: 2959-1376
2024, Vol. 03, Issue 04, 129-140
DOI: https://doi.org/10.59231/SARI7751

5. Ismail O. Muraina et al. (2022). This paper studied the use of a decision tree to predict the student's academic performance in an advanced course using chi-squared automatic interaction detection to construct the decision-making process using various factors like attendance, practical assessment, and test scores to find the accuracy rate to predict the students' performance. They highlight the classification algorithm in machine learning techniques for enhancing educational outcomes.

6. Iddrisu Issah et al. (2023). They explained using machine learning to understand the relationship between students' data and their academic performance. In this study, they used binary classification techniques using the Knowledge Discovery Process (KDP) and Rapid Miner software to predict student performance. The random forest classification algorithm was most effective, with an accuracy 93.96%, indicating that a parent's education level significantly influenced the student's academic achievement.

Material and Methods:

Classification Algorithms:

Machine learning is a set of methods that detect relationships to predict the future or make decisions about the collected student data. Classification techniques are the most commonly used in machine learning to predict the availability of data into classes according to specific algorithmic rules. Classification is used to classify the data based on the training data set, and then it gives the patterns of the new data, which is known as the training set or supervised learning technique.

The classes are pretend before extracting the data using Navies Bayes, Decision Tree (J48), Support Vector Machine, and Logistic Regression classification algorithms as follows:

Decision Tree:

The decision tree method is one of the classification methods. A decision tree is a directional tool consisting of a root node with no input and internal nodes receiving one input. The basic structure of the decision tree consists of the following three sections: Nodes, branches, and leaves. Decision trees are essential in transforming data into information because they are simple and easily understandable. The classification model is easy to understand and performs faster than other classification methods. It is an aggregation of nodes that is meant to

@2024 International Council for Education Research and Training
ISSN: 2959-1376
2024, Vol. 03, Issue 04, 129-140
DOI: https://doi.org/10.59231/SARI7751

create a decision value for a class based on any estimated numerical values. Each node corresponds to a specific splitting attribute belonging to a different class, it separates them to reduce the error. The new node building is repeated until the ending criteria are satisfied. The prediction of a student's results is determined depending on the absolute majority of examples that reach this leaf during the generation.

Navies Bayes:

For classifying text documents, the Naive Bayes algorithm is a widely used classification technique in data mining. The excellent performance and excellent applicability are the reasons behind its broad preference. While being applied, the algorithm functions under specific assumptions. The element is classified according to the category with the highest probability, and then the algorithm determines the probability of each category for the given element. With a limited quantity of training data, this technique can nevertheless obtain high accuracy rates.

Support Vector Machine:

Support Vector Machines (SVM) is a well-known technique for binary classification. This method can be thought of as a more sophisticated perceptron that looks for a hyperplane that divides the data into two groups. Compared to traditional techniques like artificial neural networks, support vector machines (SVM) tackle overfitting problems and provide a sparse solution. These datasets have been prepared to categorize new text because each category contains labelled training data. They provide two significant benefits over more recent techniques, such as neural networks: quicker processing speeds and higher quality outcomes with fewer samples.

Logistic Regression:

This method uses as few variables as possible to get the best fit by creating a model that accurately shows the relationship between dependent and independent variables. When the dependent class variable contains more than two categories, logistic regression is applied to find the cause-and-effect relationship, which includes independent factors where variables are observed in binary, triple, and multiple categories. The output of the expected value according to the value of the probability is considered input in this method.

Proposed Work:

@2024 International Council for Education Research and Training 2024, Vol. 03, Issue 04, 129-140
ISSN: 2959-1376 DOI: https://doi.org/10.59231/SARI7751

In this proposed work, we collected secondary data using 1st year MBBS students' demographic information and their

academic results of 1st semester of UG students. To predict the student's performance using the Weka tool.

Data Set: The data set we used for better prediction is given in the Table-1

Table 1:

Attribute	**Description**
Gender	Male, Female
Neet Percentage-	40-60,60-80,80-100
Age	18-21,21-25
Nationality	India, Outside India (OCI)
Religion	Hindu, Muslim, Christean, Jain
Caste	GM, OBC
Father Qualification, Mother Qualification	UG, PG, PhD, Diploma
Parents Annual Income	0-20, 20-40,40-60,60-80
10th Score	50-70, 70-100
PUC Score	60-75, 75-90, 90 and above
Subjects 1, 2, and 3: Internal Assessment Theory	45-55,55-65,65-75,75-85,75-85,85-95
Subjects 1, 2, and 3: Practical	50-60,60-70,70-80,80-90,90-100
Subjects 1, 2, & 3: Attendance	75-80,80-85,85-90,90-95,95 and above
Semester Grade	Distinction, 1st Class, 2nd Class & Fail

Results and Discussion:

Table 2: The Comparative Study of Decision Tree (J48), Navies Bayes, Support Vector Machine, and Logistic Regression

Evaluation Criteria	**Classification Algorithms for Machine Learning**

@2024 International Council for Education Research and Training 2024, Vol. 03, Issue 04, 129-140
ISSN: 2959-1376 DOI: https://doi.org/10.59231/SARI7751

	Navies Bayes	Support Vector Machine	Decision Tree(J48)	Logistic Regression
Correctly Classified Instance	188	179	196	198
Incorrectly Classified Instance	11	20	3	1
Accuracy	94.472%	89.949%	98.492%	99.497%

The above **Table 1** shows that logistic regression has taken more than other classifiers, with an accuracy of 99.49%. The correctly classified instance is usually referred to as the accuracy of the model. Hence, logistic regression has more accuracy than other classification algorithms.

Table 3: Error Measurement for Classifiers: Decision Tree (J48), Navies Bayes, Support Vector Machine, and Logistic Regression

Evaluation Criteria	Classification Algorithms for Machine Learning			
	Navies Bayes	Support Vector Machine	Decision Tree(J48)	Logistic Regression
Kappa Statistics	0.9181	0.8456	0.9774	0.9925
Mean Absolute Error	0.0384	0.2584	0.0128	0.0285
Root Mean Squared Error	0.1394	0.325	0.0799	0.0778
Relative Absolute Error	11.45%	77.15%	13.80%	8.50%
Root Relative Absolute Error	34.11%	79.50%	19.53%	19.04%

In Table 3, It is explained that Logistic Regression is less that the other classification algorithm. Kapp statistics measured between non-random agreement observations and particular categorical variables. The root mean squared error and root relative absolute error is minimum compared to the other classifier. Therefore, logistic regression is

@2024 International Council for Education Research and Training
ISSN: 2959-1376
2024, Vol. 03, Issue 04, 129-140
DOI: https://doi.org/10.59231/SARI7751

efficient among the other classification technique.

Table 4: Class label accuracy for classification algorithms: Decision Tree (J48), Navies Bayes, Support Vector Machine, and Logistic Regression.

Classifier	TP	FP	Precision	Recall	F Measure	Class
Navies Bayes	0.915	0.019	0.977	0.915	0.945	1st Class
	0.951	0.013	0.951	0.951	0.951	2nd Class
	0.933	0.005	0.933	0.933	0.933	Fail
	1.00	0.040	0.891	1.00	0.942	Distinction
Support Vector Machine	0.968	0.162	0.843	0.968	0.901	1st Class
	0.756	0.019	0.912	0.756	0.827	2nd Class
	1.00	0.00	1.00	1.00	1.00	Fail
	0.857	0.00	1.00	0.857	0.923	Distinction
Decision Tree(J48)	0.989	0.019	0.979	0.989	0.984	1st Class
	0.976	0.00	1.00	0.976	0.988	2nd Class
	1.00	0.00	1.00	1.00	1.00	Fail
	0.980	0.007	0.980	0.980	0.980	Distinction
Logistic Regression	1.00	0.010	0.989	1.00	0.995	1st Class
	0.976	0.00	1.00	0.976	0.988	2nd Class
	1.00	0.00	1.00	1.00	1.00	Fail
	1.00	0.00	1.00	1.00	1.00	Distinction

Table 4 clearly shows that every classification algorithm based on true positive rate (TP), false positive rate (FP), precision, recall, and F measures is classified with accuracy. The logistic regression performed better than other classifiers within student academic results.

Table 5: Confusion Matrix

@2024 International Council for Education Research and Training 2024, Vol. 03, Issue 04, 129-140
ISSN: 2959-1376 DOI: https://doi.org/10.59231/SARI7751

Classifier	1st Class	2nd Class	Fail	Distinction	Class
Navies Bayes	86	1	1	0	1st Class
	2	39	0	0	2nd Class
	0	1	14	0	Fail
	6	0	0	49	Distinction
Support Vector Machine	91	10	0	7	1st Class
	3	31	0	0	2nd Class
	0	0	15	0	Fail
	0	0	0	42	Distinction
Decision Tree(J48)	93	1	0	1	1st Class
	0	40	0	0	2nd Class
	0	0	15	0	Fail
	0	0	0	48	Distinction
Logistic Regression	94	1	0	0	1st Class
	0	40	0	0	2nd Class
	0	0	15	0	Fail
	0	0	0	49	Distinction

The confusion matrix is very helpful to analyses the classifier.

Conclusion:

The work explores the classifiers in machine learning that influence decision-making using student's attributes. It has been discovered that logistic regression is best compared to another classifier. The study of experiments shows that data mining will be considered most useful in the educational field. Predicting a student's semester wise performance is a great concern for education institutes. Applying machine learning techniques using the Weka tool in the prediction of student performance is very helpful to find the abilities of students and their weaknesses. The student's demographic data is visible in the university examination in various subjects to apply the method of data mining using classification.

@2024 International Council for Education Research and Training
ISSN: 2959-1376
2024, Vol. 03, Issue 04, 129-140
DOI: https://doi.org/10.59231/SARI7751

Reference:

1. Abu-Dalbouh, H. M. (2021, October 15). Application of decision tree algorithm for predicting students' performance via online learning during coronavirus pandemic. *Journal of Theoretical and Applied Information Technology*, *99*(19), 4546–4556.

2. Ahmad, F., Ismail, N. H., & Aziz, A. A. (2015, April). The prediction of students' academic performance using classification data mining techniques. *Applied Mathematical Sciences*, *9*(129), 6415–6426. https://doi.org/10.12988/ams.2015.53289

3. Ali, R. H. (2022, September 30). Educational data mining for predicting academic student performance using active classification. *Iraqi Journal of Science*, 3954–3965. https://doi.org/10.24996/ijs.2022.63.9.27

4. Al-Luhaybi, M., Yousefi, L., Swift, S., Counsell, S., & Tucker, A. (2019). Predicting academic performance: A bootstrapping approach for learning dynamic bayesian networks. In. In *Artificial intelligence. Education. Proceedings, Part I*: 20th International Conference, AIED 2019, Chicago, IL, United States, June 25–29, 2019, *20* (pp. 26–36). Springer International Publishing.

5. Alturki, S., & Alturki, N. (2021). Using educational data mining to predict students' academic performance for applying early interventions. *Journal of Information Technology Education*, *20*, 121–137. https://doi.org/10.28945/4835

6. Binti Muhammad Zahruddin, N. A., Kamarudin, N. D., Mat Jusoh, R., Abdul Fataf, N. A., & Hidayat, R. (2023, December 31). Case study: Using data mining to predict student performance based on demographic attributes. *JOIV: International Journal on Informatics Visualization*, *7*(4), 2460–2468. https://doi.org/10.30630/joiv.7.4.02454

7. Dhilipan, J., Vijayalakshmi, N., Suriya, S., & Christopher, A. (2021, February 1) (Vol. 1055, No. 1, p. 012122). Prediction of students performance using machine learning. InIOP Conference Series. *IOP Conference Series: Materials Science and Engineering*. IOP Publishing, 1055(1). https://doi.org/10.1088/1757-899X/1055/1/012122

8. Dronyuk, I., Verhun, V., & Benova, E. (2019, January 1). Non-academic factors impacting analysis of the student's the qualifying test results. *Procedia Computer Science*, *155*, 593–598. https://doi.org/10.1016/j.procs.2019.08.083

@2024 International Council for Education Research and Training
ISSN: 2959-1376
2024, Vol. 03, Issue 04, 129-140
DOI: https://doi.org/10.59231/SARI7751

9. Hashim, A. S., Awadh, W. A., & Hamoud, A. K. (2020, November 1) (Vol. 928, No. 3, p. 032019). Student performance prediction model based on supervised machine learning algorithms. InIOP Conference Series. *Materials Science and Engineering*. IOP Publishing.

10. Hussain, S., Muhsion, Z. F., Salal, Y. K., Theodoru, P., Kurtoğlu, F., & Hazarika, G. C. (2019). Prediction model on student performance based on internal assessment using deep learning. *International Journal of Emerging Technologies in Learning*, *14*(8).https://doi.org/10.3991/ijet.v14i08.10001

11. Issah, I., Appiahene, P., Appiah, O., & Inusah, F. *Determining student demographic attributes influencing performance using binary classification in KDP model*.

12. Jalota, C., & Agrawal, R. (2019, February 14). Analysis of educational data mining using classification. In2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon) (pp. 243–247). IEEE Publications. https://doi.org/10.1109/COMITCon.2019.8862214

13. Muraina, I. O., Aiyegbusi, E., & Abam, S. (2022). Decision tree algorithm use in predicting students' academic performance in advanced programming course. *International Journal of Higher Education Pedagogies*, *3*(4), 13–23. https://doi.org/10.33422/ijhep.v3i4.274

14. Pamungkas, L., Dewi, N. A., & Putri, N. A. (2024, February 15). Classification of student grade data using the K-means clustering method. *Jurnal Sisfokom*, *13*(1), 86–91. https://doi.org/10.32736/sisfokom.v13i1.1983

15. Subarkah, A. F., Kusumawati, R., & Imamudin, M. (2023, November 28). Comparison of different classification techniques to predict student graduation. MATICS. *Jurnal Ilmu Komputer dan Teknologi Informasi* (Journal of Computer Science and Information Technology), *15*(2), 96–101.

16. Faniyi, A. O. (2023). Enhancing Student Academic Performance through Educational Testing and Measurement. *Edumania-An International Multidisciplinary Journal*, *01*(02), 162–171. https://doi.org/10.59231/edumania/8981

@2024 International Council for Education Research and Training
ISSN: 2959-1376
2024, Vol. 03, Issue 04, 129-140
DOI: https://doi.org/10.59231/SARI7751

17. Adeyanju, J. O., & Ajani, I. O. (2023). Educational Counseling Strategies for Curbing academic dishonesty among students in higher Institutions. *Edumania-An International Multidisciplinary Journal*, *01*(02), 210–221. https://doi.org/10.59231/edumania/8985

18. Naveen, & Bhatia, A. (2023). Need of Machine Learning to predict Happiness: A Systematic review. *Edumania-An International Multidisciplinary Journal*, *01*(02), 306–335. https://doi.org/10.59231/edumania/8991

19. Kumar, S., & Simran. (2024). Equity in K-12 STEAM education. *Eduphoria*, *02*(03), 49–55. https://doi.org/10.59231/eduphoria/230412

20. Sachin, S. (2024). SUSTAINABLE UNINTERRUPTED LEARNING – AN APPROACH TO BLENDED LEARNING. *Shodh Sari-An International Multidisciplinary Journal*, *03*(02), 86–101. https://doi.org/10.59231/sari7690

21. Bhagoji, M. D. (2024). Navigating Global Dynamics in Teacher Education: A Comprehensive Overview. *Shodh Sari-An International Multidisciplinary Journal*, *03*(01), 123–133. https://doi.org/10.59231/sari7660

Received on: June 05, 2024
Accepted on: Aug. 20,2024
Published on: Oct., 01.2024