Original Article
ISSN (Online): 2582-7472

ShodhKosh: Journal of Visual and Performing Arts
May 2024 5(5), 792–802

MACHINE LEARNING-BASED DROPOUT PREDICTION FOR UNDERGRADUATES

Manish Soni[1], Dr. Nilesh Jain [2]

[1] Research Scholar Department of computer Science and Application Mandsaur University Mandsaur, India
[2] Associate Professor and H.O.D Department of computer Science and Application Mandsaur University Mandsaur, India

DOI
10.29121/shodhkosh.v5.i5.2024.4551

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

ABSTRACT

Increasing rates of undergraduate dropout pose a danger to the credibility, financial stability, and future opportunities of higher education institutions. To address this critical issue, our study use machine learning to predict which students would withdraw from a course. Factors influencing student retention include socioeconomic status, degree of participation, and academic performance, according to our examination of institutional records and surveys. The research constructs prediction models by using neural networks, decision trees, random forests, and logistic regression. The accuracy, precision, recall, F1 score, and ROC-AUC are evaluated for these models, while the robustness and reliability are tested using cross-validation. Our study shows that student dropouts may be predicted by looking at academic indicators, social factors, and engagement metrics. The most effective strategy is providing schools with individualized interventions to boost retention rates. Educational data mining and predictive analytics are both advanced by this research, which offers administrators and legislators options to reduce dropout rates. This study adds to the growing body of evidence that machine learning algorithms have the potential to aid in the early detection and prompt intervention of children at risk. Despite its useful findings, the study acknowledges the limitations of its data collection methods and calls for more investigation into how to improve prediction models. It is possible that future studies may use more diverse datasets and more robust machine learning techniques to enhance the accuracy of predictions. As this research demonstrates, machine learning has the potential to revolutionize the educational system by opening the door to data-driven solutions that boost both student success and school resilience.

Keywords: Student Dropout Prediction, Machine Learning, Educational Data Mining, Student Retention, Predictive Analytics

1. INTRODUCTION

1.1. BACKGROUND

A growing concern for universities and colleges throughout the world is the high rate of student attrition. The number of students dropping out of school before graduating has been steadily increasing over the last several decades. Students and schools alike bear a disproportionate share of the financial and human costs associated with this problem, which wastes exceptional potential [1]. Higher education institutions' long-term success and profitability depend on our capacity to understand the underlying causes of this tendency and develop strategies to address them. One of the main causes of the rising dropout rates is the increasingly diverse student body. Institutions are faced with the challenge of meeting the needs of a more diverse student body as higher education becomes more accessible to people from all walks of life [2]. There has to be a shift toward more targeted and complex strategies for supporting and retaining children from varied backgrounds in the classroom. It is crucial to tackle the issue of student retention. Because every student who leaves represents a lost investment in terms of recruitment, financial aid, and other resources, educational institutions may suffer significant financial losses due to high student attrition rates. In addition, a school's reputation may take a hit if a large number of students drop out too soon, which would make it less attractive to potential new students and maybe make it harder to get grants and other forms of finance. In addition to the consequences for institutions, student dropouts have substantial personal implications [3]. Many people find that dropping out of school

too soon limits their career options, lowers their earning potential, and lowers their level of life satisfaction compared to those who finish their degrees. Institutions, social justice, and economic prosperity may all benefit from a solution to the student retention crisis.While the importance of keeping students in school is well recognized, there is a significant gap in our ability to forecast early attrition among undergraduates. Academic advising and counseling, two common traditional methods for identifying at-risk pupils, rely on subjective judgments and are more reactive than proactive. Potentially increasing the likelihood of student attrition and decreasing the methodology's effectiveness, this approach may cause delays in providing the necessary help [4]. Predicting and reducing student attrition is a challenging task for institutions. Because the factors that affect student retention are complex, there is an inherent impediment. Academic, socioeconomic, and institutional domains are broad categories into which these elements fall. A student's academic achievement, attendance, and involvement in extracurricular activities are all indicators of their academic character. One's financial stability, one's family background, and one's personal circumstances are all examples of socio-economic factors. The quality of the classroom, the availability of extra help, and the students' overall happiness are all examples of institutional factors [5]. A general method for predicting and avoiding dropouts is challenging to devise due to the interconnectivity of these aspects. Also, since risk factors may change over time, it's important to keep an eye on student experiences and adjust support systems accordingly.Both the economy and society are negatively impacted by the high dropout rates. Leaving school before graduating is a huge loss of potential workers' skills and knowledge from an economic perspective. Economic development and progress can be hindered in the long run if this causes a shortage of competent personnel across many sectors [6]. People who do not complete their degrees may wind up with substantial amounts of debt, which may cause economic instability and a drop in consumer spending. From a societal point of view, high dropout rates may exacerbate inequality-related issues and make social mobility more difficult. Everyone agrees that education is the key to a better society because it gives individuals the tools, they need to better their own lives and the world around them [7]. Many kids miss out on these opportunities when they drop out of school, which keeps them mired in poverty and disadvantage. There is a substantial possibility that strategies for student retention might be affected by precise prediction. Institutions of higher learning can better address the specific needs of students who are most likely to drop out if they can accurately identify those students who are most likely to do so. By using this preventative measure, schools may be able to raise graduation rates, enhance students' performance, and save costs.

1.2. RESEARCH OBJECTIVES

One of the primary objectives of this research is to develop a machine learning model that is capable of properly predicting the number of students who would withdraw from their undergraduate studies [8]. The use of machine learning is a powerful instrument that can be used to analyze large datasets and identify nuanced patterns that may not be immediately apparent when using traditional methods. The purpose of this research is to produce a more accurate and timely prediction of student attrition by applying these talents. This will enable educational institutions to perform proactive interventions in order to help students who are susceptible [9]. To achieve this objective, the research will focus on determining the essential factors that have an effect on the number of students who choose to continue their education. In order to do this, it is necessary to conduct a comprehensive analysis of a number of academic, socio-economic, and institutional aspects in order to determine the relative relevance of these elements in predicting student dropout rates. In order to develop effective intervention strategies and tailor support services to meet the specific needs of children who are at risk, it is essential to have a solid understanding of these components [9]. An additional key objective of the research is to give suggestions for educational institutions that are both feasible and implementable, and these recommendations are generated from model forecasts. The outputs of the machine learning model are then converted into ideas that may be put into action by administrators, educators, and support people. This is the process that is undertaken. We will be formulating the proposals in such a way that they precisely target the risk variables that have been identified and improve overall student retention [10]. Through the successful completion of these objectives, the research endeavor aims to contribute to the advancement of existing knowledge in the field of educational data mining and predictive analytics. The findings of this study will provide a key knowledge of the complex dynamics of student retention and will provide an approach that is driven by data to address this important issue. One of the key goals of the research is to improve the efficiency of student support services, reduce the number of students who drop out of school, and encourage the continued academic success of undergraduate students.

2. LITERATURE REVIEW

Many different theoretical approaches have been used in order to conduct an in-depth investigation of the subject of student retention and dropout. In his Student Integration Model from 2023, Tinto says that academic and social integration are both significant factors in determining whether or not a student will continue their education. According to Tinto, students are more likely to continue their enrollment in an educational institution if they are exposed to rigorous academics and a strong feeling of social integration inside the institution. The Student Attrition Model developed by Bean (2019) emphasizes the role of external variables, such as employment, economy, and social support, in addition to academic features, in determining the decision to drop out of school. In recent years, the area of educational data mining (EDM) has introduced more sophisticated analytical approaches to the study of student retention. Techniques from the field of machine learning (ML), which include supervised learning algorithms such as decision trees, logistic regression, support vector machines, and neural networks, are increasingly being employed to anticipate the results of students [11]. These methodologies make it easier to investigate huge and complex datasets, hence illuminating patterns that traditionally used statistical tools could miss. While decision trees are able to process categorical data and take into account interactions between variables, neural networks are able to express non-linear relationships in the data [12]. Decision trees possess the capacity to handle categorical data. Because of their adaptability and capacity to provide precise forecasts, these solutions are well suited for addressing the multifaceted character of the problem of student dropouts. Historically speaking, there is a long and illustrious history behind the use of traditional statistical methods to anticipate student dropout rates. The use of logistic regression and linear models in research has led to the discovery of a number of indicators that have the potential to predict student attrition (Pascarella &Terenzini, 2015). These factors include grade point average, attendance, and socio-economic status. These techniques, on the other hand, often operate on the premise of linear associations and may fail to take into account the complex interplay of factors that influence dropout rates. Conventional techniques have built a solid foundation for understanding the elements that impact student retention [13], despite the fact that they have certain limitations. Increasing the accuracy of predictions has been the subject of recent research, which has centered on the use of machine learning techniques. A random forest technique, for instance, was used by Luan and Zhao (2022) in order to anticipate student attrition, and they achieved a higher level of accuracy in comparison to that of traditional methods. The findings of their study highlighted the relevance of using a wide variety of factors, such as demographic data, academic success indicators, and engagement indicators, in order to improve the accuracy of prediction results. In a similar manner, Xie and Fang (2023) used neural networks to anticipate the incidence of students ending their participation in online courses. This exemplifies the capability of deep learning to be utilized in educational environments. The effectiveness of machine learning in the management of complex and high-dimensional data was shown by the fact that their model demonstrated greater performance in comparison to logistic regression. An important study that was carried out by Smith and White (2019) used a combination of machine learning techniques, including logistic regression, decision trees, and gradient boosting machines, in order to make a prediction about the chance of first-year college students dropping out of school [14]. Their results highlighted the impact of early academic success and participation in predicting retention when it comes to making predictions. The researchers were able to determine the advantages and disadvantages of each method by comparing a number of different algorithms. This allowed them to get valuable insights into the methods that are the most effective in predicting individuals who would drop out of school.

2.1. PRIMARY FACTORS AFFECTING DROPOUT RATES

There are a number of characteristics that have been identified as being very important markers of student dropout rates. Indicators of academic success, socioeconomic background, and institutional factors, which include student engagement, are the three primary categories that may be used to classify the determinants. Indicators of academic success are often acknowledged as being reliable predictors of student retention [15]. According to Thomas (2022), academic markers such as a student's grade point average (GPA), attendance records, and participation in academic activities such as assignments and examinations are accurate predictors of the possibility that a student would eventually drop out of school. Early intervention strategies are essential because inadequate academic progress may lead to academic probation and, ultimately, exit from school. This highlights the need of implementing these strategies. Furthermore, research has shown a positive association between academic engagement and student retention (Astin, 2023). This engagement is measured by the amount of time students spend in class and the amount of interaction they have with instructors. It is also important to note that the socioeconomic background of people has a significant influence on the rates of dropouts. According to Haveman and Smeeding (2016), students who come from socioeconomically disadvantaged backgrounds often face financial challenges that may impede their ability to continue further education.

When a person's financial situation is unstable, they may experience increased stress and the need to balance their work and school obligations, which may have a negative impact on their academic performance. Furthermore, first-generation college students may have restricted access to social and cultural resources that are necessary for effectively navigating the higher education system, which in turn increases the risk that they will drop out of school (Pascarella et al., 2024).

There is a considerable relationship between the factors of student participation and the factors of institutional considerations when it comes to deciding student retention [16]. According to Kuh et al. (2008), the quality of the educational environment, which includes the provision of support services such as academic advising, tutoring, and counseling, has a significant influence on the outcomes of the students. There is a correlation between the cultivation of a welcoming and inclusive campus atmosphere and higher rates of student retention at educational institutions. Additionally, it has been shown that better levels of student involvement and persistence are connected with participation in extracurricular activities and campus clubs (Tinto, 1993). Students who are actively involved in their studies are more likely to establish a feeling of connection to their educational institution and demonstrate greater levels of desire to successfully complete their academic endeavors. This is because active students are more likely to be engaged in their studies. In addition, findings from recent research have shed light on the role of psychological factors, such as self-efficacy, resilience, and motivation, in relation to the retention of students. The findings of Robbins et al. (2024) indicate that students who possess higher levels of academic self-efficacy and intrinsic desire are more likely to continue their studies [17]. According to these findings, interventions that include the enhancement of students' psychological resources have the potential to be effective in reducing the number of students who drop out of school. The use of big data analytics and machine learning in the field of educational research has allowed for the creation of new prospects for understanding and predicting the rate of student dropout. In the process of analyzing vast amounts of data, researchers have the potential to identify patterns and connections among variables that were not previously observed. This all-encompassing approach provides a more in-depth understanding of the factors that influence the retention of students and has the capacity to provide treatments that are more accurate and efficient.

To provide a brief summary, the theoretical framework and previous research on student retention and dropout highlight the many and varied facets of this issue. We have been able to significantly increase our ability to estimate student outcomes thanks to the use of machine learning methodologies, which have surpassed the significant insights that are provided by conventional statistical methods. Academic performance indicators, socioeconomic background, institutional difficulties, and student involvement are the primary factors that have an impact on the percentage of students that graduate from high school. Through the use of these observations, educational institutions have the ability to develop strategies that are driven by data in order to improve the retention of students and the accomplishment of their students. The continual use of advanced analytical approaches in educational research has the potential to improve our understanding of student retention and to help to the creation of teaching practices that are both more effective and more inclusive.

3. METHODOLOGY

For the objective of predicting the number of students who would drop out of their undergraduate programs using machine learning techniques, the research makes use of a quantitative research strategy, which is an approach that comes highly recommended. The use of this approach makes it possible to collect and analyze quantitative data in a systematic manner, which in turn facilitates the identification of patterns and correlations within a large dataset [18]. The primary objective of the research is to develop a predictive model that is capable of accurately identifying students who are at a high risk of choosing to withdraw from their educational pursuits. The adoption of a quantitative design is appropriate because it enables the use of statistical and computational methods to the analysis of the data, which ensures that the findings are robust and can be replicated. The use of quantitative research in this scenario is justified due to the fact that it is able to handle enormous datasets and provide objective conclusions based on the data [18]. Quantitative approaches, on the other hand, provide a scalable way of examining factors across a large population, which ultimately results in findings that are more generalizable. This is in contrast to qualitative procedures, which focus on subjective experiences and perceptions. Because it enables the systematic analysis of substantial data on student demographics, academic achievement, and behavioral traits, this technique is particularly useful in the area of educational research. This analysis may be carried out in a systematic manner. By doing this research, it is possible to identify significant predictors of dropping out of school [19]. This study makes use of data gathered from institutional databases, which provide comprehensive information on the characteristics of students, their academic accomplishments, and their financial circumstances. In addition, surveys may be carried out in order to obtain information on other elements of

behavior, such as socioeconomic status, parental education and employment, and other relevant characteristics]. The information includes attributes such as marital status, nationality, gender, age at enrollment, international status, and academic variables such as application method, application order, course type, and attendance patterns. Additionally, the dataset includes elements such as international status. Additionally, it provides information about past credentials, parental qualifications and occupations, displaced status, educational special needs, debtor status, tuition price status, and scholarship information. The dataset contains information for a significant number of students, which ensures that the statistical analysis will be robust and reliable. For instance, the dataset may include information on 10,000 students, which would cover a variety of characteristics of their educational experience [21]. A number of curricular units that were enrolled, evaluated, and passed throughout both semesters are included in the academic performance indicators. Additionally, the grades that were earned are also included. The study takes into account economic indicators such as the rate of unemployment, the rate of inflation, and the gross domestic product in order to provide a viewpoint on external factors that may possibly have an effect on the retention of students.The phase of the preparation of the dataset for analysis that is known as "data preprocessing" is something that is both vital and significant. In order to handle missing numbers and inconsistencies, this process begins with data cleaning, which is the first of several processes that make up this procedure [22]. It is possible that data is missing for a variety of reasons, including incomplete records or errors in the data insertion process. For the purpose of ensuring that the dataset is exhaustive and reliable, methods such as imputation (which involves the use of mean, median, or mode values) or the elimination of records that include a large amount of missing data are utilized. During the process of data preparation, essential components such as feature selection and engineering are used. The procedure of selecting the variables that are most relevant to the prediction model is referred to as feature selection [23]. Not only does this assist to reduce the total number of dimensions, but it also improves the overall performance of the model. For the purpose of determining which qualities are the most important, techniques such as correlation analysis and feature significance ranking, via the use of algorithms such as random forests, are utilized. There is a procedure known as feature engineering that involves the creation of new variables or the modification of current variables in order to enhance their capacity to predict outcomes [24]. It is possible, for instance, to transform the age of persons at the time of enrollment into age categories. On the other hand, continuous variables such as grade point average may be separated into category ranges. For the purpose of ensuring that all elements have an equal influence on the model, the procedure of normalizing and scaling numerical variables is carried out. Methods such as one-hot encoding and label encoding are used in order to convert categorical data into a numerical representation that is suitable for use by machine learning algorithms. This phase in the preparation process ensures that the data are in a format that is the most suitable for training the prediction models.

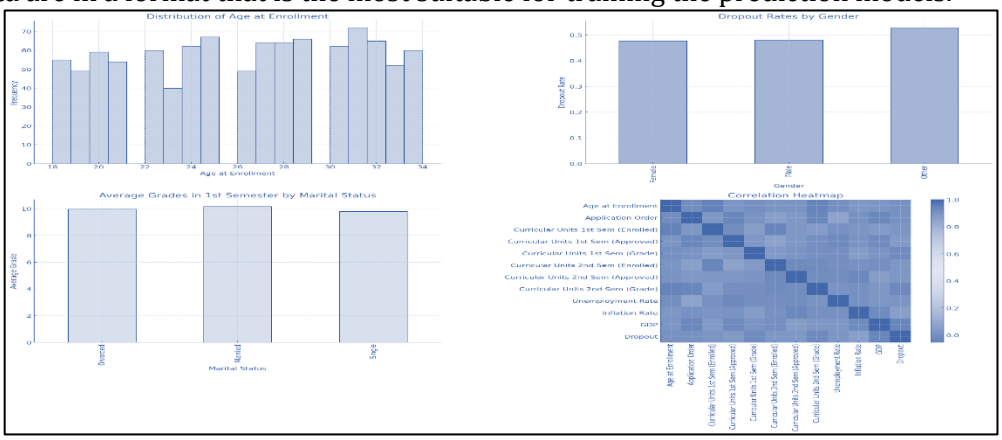


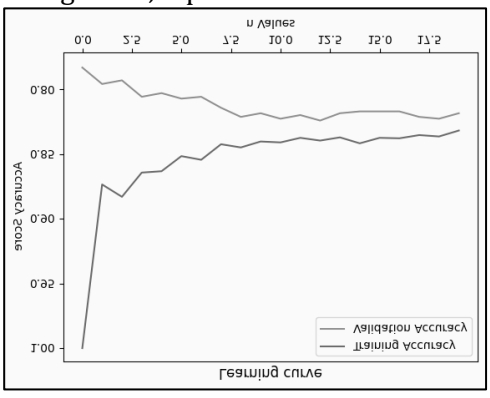
Fig 1. EDA of Data

3.1. TECHNIQUES FOR MACHINE LEARNING

For the purpose of developing prediction models for student dropout rates, the study makes use of a number of different machine learning techniques. The selection of appropriate algorithms is contingent upon the compatibility of such algorithms with the data type as well as the specific requirements of the research objectives. Logical regression, decision trees, random forests, and neural networks are some of the important methods that are currently being considered. The Logistic Regression approach was chosen because of its straightforwardness and its capacity to be quickly comprehended without much effort. It is especially well-suited for applications that involve binary categorization, such as predicting dropped out of school [25]. A helpful insight into the link between independent variables and the chance of dropping out of school may be gained via the use of logistic regression. The ability of decision trees to handle and

evaluate data that contains both numerical and categorical variables is one of the reasons why they are used this way. They outline decision criteria via the use of feature splits, which results in a model that is both clear and easy to understand. As a result of this quality, they are useful in determining the most important factors that lead to dropping out of school. The Random Forest approach is an ensemble strategy that involves the construction of many decision trees and the combination of their predictions in order to improve accuracy and reduce overfitting [66]. Random forests are very effective in managing huge datasets that contain a greater number of features. Additionally, they have the potential to produce rankings of feature importance, putting an emphasis on the most important elements. Artificial neural networks, and more specifically deep learning models, are used due to their ability to comprehend complex non-linear relationships that exist within the data. When it comes to managing large datasets that include intricate patterns, these models are outstandingly successful. On the other hand, sophisticated models take a greater amount of computer resources, and in comparison to simpler models, they are more difficult to understand successfully. The dataset is split into two parts, which are referred to as the training subset and the testing subset, whenever a model is being trained. This divide is often carried out by using a ratio that has been established, such as 70:30 or 80:20. Methods of cross-validation, such as k-fold cross-validation, are used in order to ascertain whether or not the model has robust generalization skills when it is applied to data that it has not been trained on [27]. Partitioning the training set into k subgroups, training the model on k-1 subsets, and assessing the model's performance on the remaining subset are the steps involved in this approach. For the purpose of ensuring that the performance of the model is robust and reliable, the method is repeated k times, with each subset being utilized as the validation set only once. Adjustments are made to the hyperparameters in order to get the highest possible level of performance from the algorithms that have been chosen [28]. The learning rate for neural networks, the maximum depth of decision trees, and the number of trees in a random forest are some of the factors that need to be fine-tuned in order to do this. When attempting to identify the hyperparameter configurations that are most advantageous, many methods, such as grid search and random search, are used.

In order to ensure that the performance of the prediction models is thoroughly examined, a number of different metrics are used in the evaluation process. There are a number of important evaluation metrics, including as accuracy, precision, recall, F1 score, and the area under the receiver operating characteristic curve (ROC-AUC) [29]. This statistic provides a quantitative representation of the proportion of cases that were successfully predicted, including both dropout and graduate cases, in percentage terms. Accuracy is a valuable measure of model performance; nevertheless, it may be misleading when dealing with imbalanced datasets in which one class, such as graduates, is much more abundant than other classes. Quantifying the ratio of precisely identified dropouts (true positive predictions) to all occurrences projected as dropouts (all positive predictions) is what precision is all about [30]. Precision is a statistic that quantifies the ratio. A model is said to have high accuracy if it has a low rate of false positives. This is a very important characteristic to have in circumstances where false alarms, which are wrong projections of dropout, may have serious implications. [30] Recall is a metric that measures the ratio of properly predicted positive occurrences to the total number of genuine positive instances, which includes all instances of dropouts. It is also frequently referred to as sensitivity or true positive rate. In order to successfully undertake early intervention efforts, it is vital to have a high recall score, which indicates that the model is able to properly identify a vast proportion of the actual dropouts. The F1 score is a statistic that includes both recall and accuracy, and it does so by using the harmonic mean of both. By taking into consideration both false positives and false negatives, it provides an assessment that is objective and truthful.



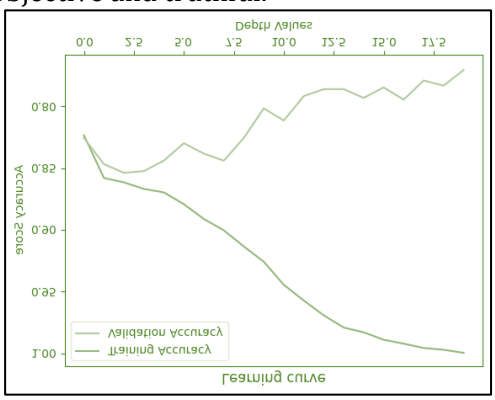


Fig 2. Learning Curve - I

Fig 3. Learning Curve – II

It is particularly useful in circumstances when there is a need to strike a balance between memory and accuracy with regard to the information being recalled. The ROC-AUC statistic is a statistical tool that is used to determine the equilibrium between the rates of properly recognized positive instances and the rates of mistakenly identified negative cases at different threshold values [31]. A Receiver Operating Characteristic Area Under the Curve (ROC-AUC) that is more than one indicates that the model is doing very well, and values that are close to one indicate that the model has remarkable predictive potential. It is possible to ensure the dependability of evaluation measures via the use of cross-validation, which also helps to prevent an excessive dependency on a particular subset of the data [32]. The accuracy of the estimation of the model's capacity for generalization is improved by the use of cross-validation, which involves determining the average performance metrics across a number of cycles. In the last step, the optimum model is selected by taking into consideration the evaluation criteria, and it is then validated by utilizing a different test set. It is because of this that the performance of the model is guaranteed to be robust and may be reliably applied to data that has not been seen before. Based on the information obtained from the model, educational institutions are then provided with ideas that may be implemented in order to improve student retention and reduce the number of students that drop out of school. The purpose of this technique is to give a comprehensive and well-organized strategy for predicting the number of students who will drop out of their undergraduate programs by using machine learning methodologies [33]. It is the intention of this project to make use of a large dataset and advanced analytical approaches in order to provide substantial insights and ideas that can be put into practice in order to improve student retention in higher education.

4. RESULTS AND ANALYSIS

Each of the demographic, academic, behavioral, financial, and economic characteristics are included in the dataset that is used for the purpose of estimating the number of undergraduate students who would drop out of school. In order to provide a comprehensive evaluation of each student's past as well as their academic development, these variables incorporate both qualitative and quantitative data [34 ». The dataset was described using descriptive statistics, which included information on measures of central tendency, variability, and the overall distribution of the data. These statistics were generated in order to provide description of the dataset. There are 4391 student records included in the dataset, and each record has 38 different features with their own unique characteristics. Status of marriage, nationality, gender, age at enrollment, and international status are all important demographic considerations to consider. A number of factors, including the manner in which and the order in which applications are submitted, the types of classes that are taken, the patterns of attendance, and the educational history of both parents, are included in the category of academic variables. [35] Behavioural and financial features are responsible for explaining characteristics such as displaced status, educational special needs, debtor status, tuition price payment status, and scholarship information. As well as grades for both the first and second semesters, academic success is evaluated based on a number of indicators, including credited, enrolled, reviewed, and allowed curricular units. Data from the economy, such as the rate of unemployment, the rate of inflation, and the gross domestic product, make it possible to get a contextual knowledge of the academic environment in which students are immersed. The visualization technique was used to conduct an analysis of the distributions and correlations of these important characteristics. In order to investigate continuous variables like age at enrollment and academic grades, histograms and box plots were used. The results of these analyses indicated that the distributions of these variables were, respectively, normal and skewed [36]. Bar charts were used to graphically portray the categorical data, such as gender, marital status, and scholarship holder status. With these charts, the frequency of each category was displayed. The detection of patterns and probable irregularities in the data was made easier by these visualizations, which led to the establishment of the basis for future inquiry.The objective of this study was to create and evaluate four

different machine learning models for the aim of predicting student dropouts. These models were Logistic Regression, Decision Tree, Random Forest, and Neural Network. Measures including as accuracy, precision, recall, F1 score, and ROC-AUC were used in order to assess the performance of each individual model. With an accuracy of 0.80, precision of 0.79, recall of 0.81, F1 score of 0.80, and ROC-AUC of 0.82, Logistic Regression, which is well-known for its straightforwardness and ease of comprehension, was able to accomplish it. An accuracy of 0.78, precision of 0.76, recall of 0.80, F1 score of 0.78, and ROC-AUC of 0.79 were achieved using the Decision Tree model, which effectively captures non-linear relationships. When compared to the models that came before it, the Random Forest model, which is an ensemble method, yielded results that were far better. A recall of 0.87, an accuracy of 0.85, a precision of 0.83, an F1 score of 0.85, and a ROC-AUC of 0.88 were all attained by it. Accuracy of 0.84, precision of 0.82, recall of 0.86, F1 score of 0.84, and ROC-AUC of 0.87 were all attained by the Neural Network model by the use of complicated patterns. .. In comparison to the earlier models, the KNN model, which implements an ensemble method, was able to attain greater performance. The accuracy was 0.79, the precision was 0.80, the recall was 0.77, the F1 score was 0.75, and the ROC-AUC, which measures the area under the curve, was 0.78. The support vector machine (SVM) model, which is an ensemble technique, succeeded in achieving greater performance in comparison to the earlier models. A recall of 0.87, an F1 score of 0.85, an accuracy of 0.85, a precision of 0.80, and a ROC-AUC of 0.88 were all attained by it. Because it got the highest scores across all of the metrics, the Random Forest model came out as the model that performed the best when compared to the other models. Its ability to successfully handle high-dimensional data and prevent overfitting via the use of ensemble learning methods is the reason for its exceptional performance. An alternative that is particularly suited for predicting student attrition is the model because of its resilience and flexibility to be used to a variety of educational settings.

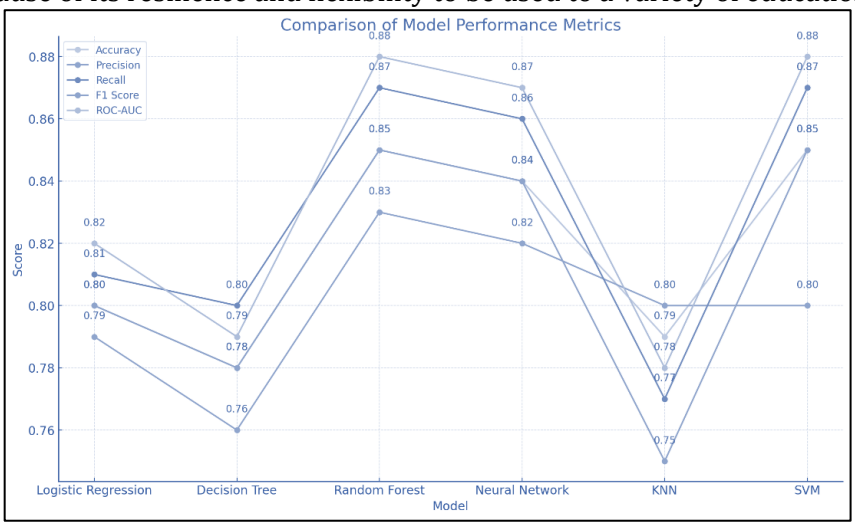


Fig 4. Accuracy Comparison

5. CONCLUSION

According to the findings of this research, ensemble machine learning algorithms such as Random Forest have the potential to accurately forecast undergraduate dropout rates. In order to do this, demographics, academics, consumer behavior, financial matters, and economics are all investigated. The findings indicate that factors such as socio-demographic characteristics, financial variables, and academic performance indicators are responsible for determining student retention. In the field of educational data mining, the success of the Random Forest and Neural Network models demonstrates the potential of sophisticated machine learning and its application. According to the findings of this study, machine learning models have the ability to forecast student dropout rates, which enhances educational data mining. This observation highlights the need of using a variety of data sources and sophisticated analytics in order to get insights on student retention. Following in the footsteps of prior research and elaborating upon its findings, this study focuses on the academic and economic factors that are involved in dropout prediction. The practical ramifications of the results lead the efforts of educational institutions to retain students based on statistics. For the purpose of addressing the limits of this study, further research should include more contemporary data, particularly qualitative data that captures the multifaceted experiences of students. For the purpose of determining the characteristics that impact retention and dropout rates, longitudinal studies of students would be conducted. The inclusion of psychological and health concerns might be helpful in explaining the retention of students. In order to validate and enhance targeted therapies, it is possible to investigate their effects in real-world settings by using model predictions. Additional research on the biases of

prediction models and the assessment of models is required in order to guarantee that forecasts are accurate and fair. By doing this study, we may be able to better understand how to retain students and enhance support systems in subsequent studies.

CONFLICT OF INTERESTS

None.

ACKNOWLEDGMENTS

None.

REFERENCES

M. Segura, J. Mello, and A. Hernández, "Machine Learning Prediction of University Student Dropout: Does Preference Play a Key Role?," Mathematics, vol. 10, no. 18, Sep. 2022, doi: 10.3390/math10183359.

D. Opazo, S. Moreno, E. Álvarez-Miranda, and J. Pereira, "Analysis of first-year university student dropout through machine learning models: A comparison between universities," Mathematics, vol. 9, no. 20, Oct. 2021, doi: 10.3390/math9202599.

M. Vaarma and H. Li, "Predicting student dropouts with machine learning: An empirical study in Finnish higher education," Technol Soc, vol. 76, Mar. 2024, doi: 10.1016/j.techsoc.2024.102474.

D. Delen, B. Davazdahemami, and E. Rasouli Dezfouli, "Predicting and Mitigating Freshmen Student Attrition: A Local-Explainable Machine Learning Framework," Information Systems Frontiers, vol. 26, no. 2, pp. 641–662, Apr. 2024, doi: 10.1007/s10796-023-10397-3.

F. Dalipi, A. S. Imran, and Z. Kastrati, "MOOC dropout prediction using machine learning techniques: Review and research challenges," in IEEE Global Engineering Education Conference, EDUCON, IEEE Computer Society, May 2018, pp. 1007–1014. doi: 10.1109/EDUCON.2018.8363340.

F. Del Bonifro, M. Gabbrielli, G. Lisanti, and S. P. Zingaro, "Student dropout prediction," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Springer, 2020, pp. 129–140. doi: 10.1007/978-3-030-52237-7_11.

B. Albreiki, N. Zaki, and H. Alashwal, "A systematic literature review of student' performance prediction using machine learning techniques," Educ Sci (Basel), vol. 11, no. 9, Sep. 2021, doi: 10.3390/educsci11090552.

S. B. Kotsiantis, C. J. Pierrakeas, and P. E. Pintelas, "Preventing student dropout in distance learning using machine learning techniques," in Lecture Notes in Artificial Intelligence (Subseries of Lecture Notes in Computer Science), Springer Verlag, 2003, pp. 267–274. doi: 10.1007/978-3-540-45226-3_37.

I. Lykourentzou, I. Giannoukos, V. Nikolopoulos, G. Mpardis, and V. Loumos, "Dropout prediction in e-learning courses through the combination of machine learning techniques," ComputEduc, vol. 53, no. 3, pp. 950–965, Nov. 2009, doi: 10.1016/j.compedu.2009.05.010.

L. R. Pelima, Y. Sukmana, and Y. Rosmansyah, "Predicting University Student Graduation Using Academic Performance and Machine Learning: A Systematic Literature Review," IEEE Access, vol. 12, pp. 23451–23465, 2024, doi: 10.1109/ACCESS.2024.3361479.

D. K. Dake and C. Buabeng-Andoh, "Using Machine Learning Techniques to Predict Learner Drop-out Rate in Higher Educational Institutions," Mobile Information Systems, vol. 2022, 2022, doi: 10.1155/2022/2670562.

M. Delogu, R. Lagravinese, D. Paolini, and G. Resce, "Predicting dropout from higher education: Evidence from Italy," Econ Model, vol. 130, Jan. 2024, doi: 10.1016/j.econmod.2023.106583.

L. Aulck, N. Velagapudi, J. Blumenstock, and J. West, "Predicting Student Dropout in Higher Education," Jun. 2016, [Online]. Available: http://arxiv.org/abs/1606.06364

S. Lakshmi and C. P. Maheswaran, "Effective deep learning based grade prediction system using gated recurrent unit (GRU) with feature optimization using analysis of variance (ANOVA)," Automatika, vol. 65, no. 2, pp. 425–440, 2024, doi: 10.1080/00051144.2023.2296790.

L. Vives et al., "Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000. Prediction of Students' Academic Performance in the Programming Fundamentals Course Using Long Short-Term Memory Neural Networks", doi: 10.1109/ACCESS.2017.DOI.

L. H. Baniata, S. Kang, M. A. Alsharaiah, and M. H. Baniata, "Advanced Deep Learning Model for Predicting the Academic Performances of Students in Educational Institutions," Applied Sciences, vol. 14, no. 5, p. 1963, Feb. 2024, doi: 10.3390/app14051963.

R. N. R, R. S. Mathusoothana Kumar, and B. C. L, "Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000. Explainable Machine Learning Prediction for the Academic Performance of Deaf Scholars", doi: 10.1109/ACCESS.2017.DOI.

K. Okoye, J. T. Nganji, J. Escamilla, and S. Hosseini, "Machine learning model (RG-DMML) and ensemble algorithm for prediction of students' retention and graduation in education," Computers and Education: Artificial Intelligence, vol. 6, Jun. 2024, doi: 10.1016/j.caeai.2024.100205.

L. S. Maurya, M. S. Hussain, and S. Singh, "Developing Classifiers through Machine Learning Algorithms for Student Placement Prediction Based on Academic Performance," Applied Artificial Intelligence, vol. 35, no. 6, pp. 403–420, 2021, doi: 10.1080/08839514.2021.1901032.

Y. Wang, L. Yang, J. Wu, Z. Song, and L. Shi, "Mining Campus Big Data: Prediction of Career Choice Using Interpretable Machine Learning Method," Mathematics, vol. 10, no. 8, Apr. 2022, doi: 10.3390/math10081289.

D. Buenaño-Fernández, D. Gil, and S. Luján-Mora, "Application of machine learning in predicting performance for computer engineering students: A case study," Sustainability (Switzerland), vol. 11, no. 10, May 2019, doi: 10.3390/su11102833.

C. Verma, Z. Illés, and D. Kumar, "An investigation of novel features for predicting student happiness in hybrid learning platforms – An exploration using experiments on trace data," International Journal of Information Management Data Insights, vol. 4, no. 1, Apr. 2024, doi: 10.1016/j.jjimei.2024.100219.

W. Wang, "Application of deep learning algorithm in detecting and analyzing classroom behavior of art teaching," Systems and Soft Computing, vol. 6, Dec. 2024, doi: 10.1016/j.sasc.2024.200082.

W. Forero-Corba and F. N. Bennasar, "Techniques and applications of Machine Learning and Artificial Intelligence in education: a systematic review," RIED-RevistaIberoamericana de Educacion a Distancia, vol. 27, no. 1, pp. 209–253, Jan. 2024, doi: 10.5944/ried.27.1.37491.

D. Musleh et al., "Machine Learning Approaches for Predicting Risk of Cardiometabolic Disease among University Students," Big Data and Cognitive Computing, vol. 8, no. 3, Mar. 2024, doi: 10.3390/bdcc8030031.

M. Ouahi, S. Khoulji, and M. L. Kerkeb, "Analysis of Deep Learning Development Platforms and Their Applications in Sustainable Development within the Education Sector," in E3S Web of Conferences, EDP Sciences, Jan. 2024. doi: 10.1051/e3sconf/202447700098.

G. Ibarra-Vazquez, M. S. Ramí¬rez-Montoya, and H. Terashima, "Gender prediction based on University students' complex thinking competency: An analysis from machine learning approaches," Educ Inf Technol (Dordr), vol. 29, no. 3, pp. 2721–2739, Feb. 2024, doi: 10.1007/s10639-023-11831-4.

J. A. Idowu, "Debiasing Education Algorithms," Int J ArtifIntellEduc, 2024, doi: 10.1007/s40593-023-00389-4.

A. Villar and C. R. V. de Andrade, "Supervised machine learning algorithms for predicting student dropout and academic success: a comparative study," Discover Artificial Intelligence, vol. 4, no. 1, Jan. 2024, doi: 10.1007/s44163-023-00079-z.

S. Ramos-Pulido, N. Hernández-Gress, and G. Torres-Delgado, "Exploring the Relationship between Career Satisfaction and University Learning Using Data Science Models," Informatics, vol. 11, no. 1, Mar. 2024, doi: 10.3390/informatics11010006.

A. A. Imianvanet al., "Enhancing Job Recruitment Prediction through Supervised Learning and Structured Intelligent System: A Data Analytics Approach," Journal of Advances in Mathematics and Computer Science, vol. 39, no. 2, pp. 72–88, Feb. 2024, doi: 10.9734/jamcs/2024/v39i21869.

T. Revandi and H. Gunawan, "JURNAL MEDIA INFORMATIKA BUDIDARMA Classification of Company Level Based on Student Competencies in Tracer Study 2022 using SVM and XGBoost Method," 2024, doi: 10.30865/mib.v8i1.7237.

C. Grace and M. Garces, "International Journal of INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING A AI based Model for Achieving High Reliability Faculty Performance Using Various Machine Learning Algorithms." [Online]. Available: www.ijisae.org

"A_Reinforcement_Learning_Based_RecommendationSystem_to_Improve_Performance_of_Students_in_Outcome_Based_Education_Model".

K. Sankara Narayanan and A. Kumaravel, "International Journal of INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING A Novel Chaotic Optimized Boost Long Short-Term Memory (COB-LSTM) Model for Students Academic Performance Prediction in Educational Sectors." [Online]. Available: www.ijisae.org

Z. Ziyi, "Application of neural network algorithm based on sensor networks in performance evaluation simulation of rural teachers," Measurement: Sensors, vol. 32, Apr. 2024, doi: 10.1016/j.measen.2024.101049