Nanotechnology Perceptions
ISSN 1660-6795
www.nano-ntp.com

Prediction for Dropout Undergraduate Students Using Machine Learning Technique

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

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

Higher education schools' identities, budgets, and students' futures are all at risk because more and more first-year students are dropping out. To solve this important problem, this work uses machine learning to predict which students will drop out. Our study of institutional records and surveys shows that academic success, financial background, and amount of participation all have an effect on a student's decision to stay in school. Logistic regression, decision trees, random forests, and neural networks are used to make forecast models in the study. We test these models for precision, recall, F1 score, ROC-AUC, and accuracy. We also do cross-validation to make sure they are stable and reliable. Our study shows that academic signs, social traits, and involvement measures can all be used to predict when a student will drop out. The best way to help schools keep students is to give them individualized assistance. This study improves educational data mining and prediction analytics and gives lawmakers and school managers ways to lower the number of students who drop out. This study shows that machine learning techniques might help find kids who are at risk early on and help them right away. In spite of what it adds, the study admits that there were flaws in the data collection and that more research is needed to make prediction models better. Future study that uses bigger, more varied datasets and stronger machine learning methods might be able to make predictions that are more accurate. This research shows that machine learning could change the way we learn, making it possible to use data to find ways to help students do better in school and make institutions stronger.

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

1. Introduction

1.1. Background

The issue of students dropping out of college has become a big concern for schools all over the world. Over the past few decades, there has been a clear rise in the number of students who drop out of school before earning their degrees. Not only does this problem waste a lot of ability, but it also costs a lot of money for both the kids and the schools involved[1]. Understanding the basic factors that cause this trend and coming up with ways to fix it are very important for the long-term success and survival of higher education institutions. The rising failure rates are mostly caused by the fact that students are becoming more diverse. Higher education is becoming more open to people from a wide range of socioeconomic groups. This means that schools have to meet the needs and demands of a larger range of students[2]. Because schools have a lot of different kinds of kids, they need more advanced and targeted ways to help and keep them. It's impossible to overstate how important it is to keep students in school. Schools can lose a lot of money when a lot of students drop out. This is because every student who leaves means the school has to spend money on hiring new students, giving them financial aid, and other resources. In addition, high rates of students dropping out early can hurt the reputation of an educational institution, making it less appealing to new students and possibly making it harder for the institution to get money and other kinds of help. In addition to the problems they cause for schools, student dropouts have big effects on the people who do them[3]. When compared to people who finish their degrees, those who stop going to school too soon often have fewer job possibilities, lower income potential, and less total life satisfaction. Not only is it important for schools to deal with the problem of keeping students, it's also important for supporting social justice and economic growth.

1.2. Problem Statement

Even though everyone knows how important it is to keep students in school, there is still a big problem with being able to correctly predict early on which undergraduates will drop out. Methods that are usually used to find students who are at risk, like academic advice and mentoring, rest on biased judgments and are more reactive than proactive. It's possible that this way will cause delays in providing the needed help, which will make it less effective and increase the chance of students dropping out [4]. There are many things that make it hard for schools to predict and stop students from leaving. One problem that comes up naturally is that the things that help keep students is very complicated. These factors can be loosely put into three groups: the academic, the socio-economic, and the societal. Academic traits include doing well in school, going to class regularly, and being actively involved in academic activities. Some socioeconomic factors are a person's family background, their level of financial protection, and their own unique circumstances. Some institutional factors are the quality of the learning environment, the ease of access to support services, and the overall happiness of the students[5]. It's hard to come up with a uniform plan to identify and stop losses because these factors are all linked. Also, because students' experiences are always changing, support systems need to be constantly checked on and changed because risk factors can change.

1.3. Significance of the Research

There are big economic and social effects of the high failure rates. Looking at it from a business point of view, kids who drop out of school before finishing take away a lot of potential workers. This could make it hard for different businesses to find skilled workers, which would slow down progress and economic growth[6]. Also, people who don't finish their degrees may be burdened by student loan debt, which could cause the economy to stay unstable for a long time and consumers to spend less. From a cultural point of view, high failure rates could make inequality problems worse and make it harder to move up in society. Most people agree that education is an important driver of social growth because it gives people the skills and knowledge they need to make their lives better[7]. When students drop out of school, they often give up these chances, which keeps them in loops of poverty and disadvantage. The accuracy of predictions could have a big effect on strategies for keeping kids. By precisely identifying students who are most likely to drop out, schools can plan and carry out targeted programs that best meet the needs of these people. Adopting this proactive approach could help lower the number of students who drop out, improve student performance, and make schools more effective overall.

1.4. Research Objectives

The main goal of this study is to create a machine learning model that can correctly predict how many first-year college students will drop out [8]. When looking at large datasets and finding complex trends that might not be obvious using traditional methods, machine learning is a powerful tool. Using these skills, the study aims to give a more accurate and timely prediction of student dropout, allowing schools to take proactive steps to help students who are at risk[9]. To reach this objective, the study will focus on finding important factors that have an effect on keeping students. This means looking closely at a number of academic, socioeconomic, and social factors to see how important they are in predicting student dropout. Understanding these factors is important for coming up with effective ways to help kids who are at risk and for making sure that support services are tailored to each child's specific needs[9]. Model predictions for educational institutions are also used to come up with useful and doable suggestions as another main goal of the study. In this case, the machine learning model's results need to be turned into ideas that managers, teachers, and support staff can use. The ideas will be made to directly target the risk factors that have been found and to improve student retention overall[10]. By reaching these goals, the study hopes to add to what is already known about educational data mining and predictive analytics. The study results will help us understand how difficult it is to keep students and suggest a way to deal with this important issue that is based on facts. The main goal of the study is to improve the effectiveness of student support services, lower the number of students who drop out, and encourage undergraduates to keep doing well.

2. Literature Review

2.1. Theoretical Framework

Keeping students in school and dropping them off have been looked at in depth from a number of different theoretical points of view. According to Tinto's (2023) Student Integration Model, both academic and social integration are very important for students to stay in school. Tinto says that students are more likely to stay in school if they feel like they are being challenged academically and are a part of a strong social community at their school. The 2019 Bean Student Attrition Model shows that outside factors like work, money, and social support, along with academic factors, play a big role in a person's decision to drop out. Educational data mining (EDM), a new area, has recently brought more advanced analysis tools to the study of keeping students. More and more, guided learning methods like decision trees, logistic regression, support vector machines, and neural networks are being used in machine learning

(ML) to predict how well students will do in school[11]. These techniques make it easier to look at big, complicated datasets and find trends that regular statistical methods might miss. Neural networks can show how data is connected in ways that aren't linear, while decision trees can handle categorical data and take into account how variables interact with each other[12]. Because they are flexible and can make accurate guesses, these methods are good for dealing with the complicated issue of student losses.

2.2. Previous Research

In the past, traditional statistics methods were used to predict which students would drop out of school. Researchers using logistic regression and linear models have found a lot of things, like GPA, attendance, and socioeconomic status, that can help us guess which students will drop out (Pascarella & Terenzini, 2015). Still, these methods often assume that relationships are straight, and they might not take into account how the many factors that affect loss rates interact with each other. Even though traditional methods have some flaws, they have built a strong foundation for understanding what makes students stay in school[13]. A lot of recent study has been done on how to use machine learning to make predictions more accurate. For example, Luan and Zhao (2022) used a random forest algorithm to predict which students would drop out of school, which was more accurate than other methods. Their study showed how important it is to use a wide range of traits, such as personal information, school success, and engagement measures, to get better results when making predictions. According to Xie and Fang (2023), neural networks were used to predict when students would stop taking online classes. This shows how deep learning can be used in educational settings. The fact that their model did better than logistic regression shows how useful machine learning is for dealing with complex and high-dimensional data. Smith and White's (2019) interesting study used a mix of machine learning techniques, like logistic regression, decision trees, and gradient boosting machines, to guess how likely it was that first-year college students would drop out[14]. Their results showed how important early school success and participation are for predicting recall. By contrasting various algorithms, researchers were able to find the pros and cons of each method, which led to useful insights into the best methods for predicting failure.

2.3. Primary Factors Affecting Dropout Rates

Several factors have been identified as important markers of the number of students who drop out. There are three main groups of factors that affect success: measures of academic performance, financial background, and societal factors such as student involvement. Academic success measures are often seen as good ways to tell if a student will stay in school[15]. Academic signs like a student's GPA, attendance records, and participation in schoolwork like tests and projects are good ways to tell if they are likely to drop out (Thomas, 2022). Not doing well in school could lead to academic probation and eventually being kicked out of school, which shows how important it is to use early remediation strategies. Astin (2023) also says that study has shown a positive link between academic involvement (measured by class activity and touch with teachers) and student retention. People's social background also has a big effect on the number of dropouts. Unfortunately, students from low-income families often face financial problems that make it impossible for them to continue their education (Haveman & Smeeding, 2016). People who are unstable financially may feel more stressed and have to balance work and school, which can hurt their academic success. According to Paccarella et al. (2024), first-generation college students may not have as many social and cultural tools that are necessary to succeed in college, which makes them more likely to drop out. When it comes to keeping students, official factors and student participation are both very important[16]. The quality of the learning setting, which includes services like academic guidance, coaching, and therapy, has a big effect on how well students do in school (Kuh et al., 2008). Schools that make their campuses more friendly and open to everyone tend to have higher return rates. Additionally, taking part in campus groups and events outside of school has been linked to more engaged and persistent students (Tinto, 1993). Students who are interested in their studies are more likely to feel connected to their school and be more motivated to do well in their classes. Recent studies have also shown how important psychological factors like drive, self-efficacy, and resilience are in keeping students in school. It was found by Robbins et al. (2024) that students who have higher levels of academic self-efficacy and innate drive are more likely to keep studying[17]. Based on these numbers, programs that aim to improve students' psychological resources might be able to lower the number of students who drop out. Using big data analytics and machine learning in educational study has opened up new ways to understand and predict why students drop out of school. By looking at a lot of data, researchers may find trends and connections between variables that they hadn't seen before. This all-around method helps us learn more about the factors that affect student retention and lets us provide more targeted and effective treatments.

To sum up, the theory framework and previous study on keeping students in school and preventing them from dropping out stress how complicated and varied this situation is. Machine learning techniques have made it much easier to predict how students will do in school, going beyond the useful information that traditional statistical methods gave us. The main things that affect dropout rates are academic success measures, social background, problems with the school, and student involvement. Schools can use these observations to come up with data-driven strategies to help students stay in school and do better in school by using these observations. The continued use of complex analysis methods in educational research could help us better understand how to keep students in school and lead to the creation of more effective and welcoming teaching methods.

3. Methodology

3.1. Experimental Methodology

Based on machine learning methods, the study uses a quantitative research method, which is a great way to predict which college students will drop out. This method makes it possible to collect and analyze quantitative data in a planned way, which makes it easier to find trends and connections in large datasets[18]. The main purpose of the study is to develop a predictive model that can accurately find kids who are very likely to drop out of school. A quantitative approach is good because it lets you use statistical and computer methods to look at the data, which makes sure that the results are solid and can be repeated. In this case, it makes sense to use quantitative research because it can handle big numbers and give fair views based on data[18]. When compared to qualitative methods, which focus on personal feelings and views, quantitative methods can be used to look at factors across a large group, which leads to more general findings. This method works really well in educational study because it lets a lot of information about students' backgrounds, how well they do in school, and how they act be analyzed in a planned way. This study can help find important factors that can help identify dropping out [19].

3.2. Data Collection

The information for this study comes from university records, which have a lot of specific information about students' personal lives, academic performance, and financial situations. Surveys can also be used to get information about a person's socioeconomic situation, their parents' schooling and work, and other relevant behavioral factors[20]. The information includes things like marriage status, country, gender, age at registration, and foreign status. It also has academic variables like how people applied, the order in which they applied, the type of course, and attendance rates. It also has information about previous credentials, family skills and jobs, being moved, having special educational needs, being in debt, paying school fees, and scholarships. The dataset includes data for a large group of students, which ensures that the statistical analysis will be strong and reliable. For instance, the collection might have information about 10,000 kids, such as different parts of their school lives[21]. Some of the things that show academic success are the grades and the number of units taken, tested, and passed in both semesters. The study uses economic data like the jobless rate, inflation rate, and GDP to give an overview of outside factors that might have an effect on keeping students.

3.3. Data Preprocessing

Preprocessing the data is an important step in getting the information ready for analysis. There are many steps to this process. The first step is to clean up the data to get rid of any errors or missing numbers[22]. There are many reasons why data might be missing, such as records that aren't full or mistakes made when entering the data. To make sure the information is complete and reliable, methods like estimation (using mean, median, or mode values) or getting rid of records that are missing important data are employed. Preparing data includes important steps like choosing the right features and tech. Feature selection is the process of picking the factors that are most useful for the predictive model[23]. This helps cut down on the number of variables and makes the model work better. To find the most important traits, people use techniques like association analysis and feature importance ranking, which uses algorithms like random forests. Feature engineering is the process of making new factors or changing old ones to make them better at predicting what will happen [24]. For example, a person's age at enrollment can be broken down into age groups, while a continuous measure like GPA can be split up into ranges of categories. It is necessary to normalize and scale numerical factors to make sure that all of them have the same effect on the model. Using methods like one-hot encoding or label encoding, categorical variables are turned into a number form that can be used by machine learning algorithms. This step of getting the data ready makes sure that it is in the best shape for training the forecast models.

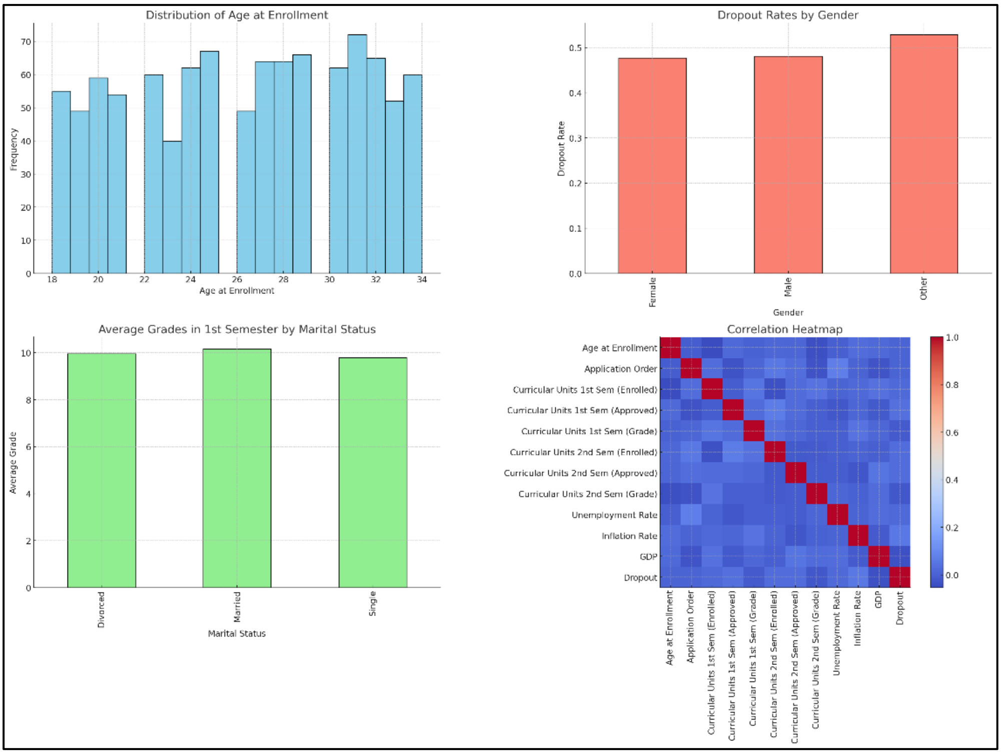


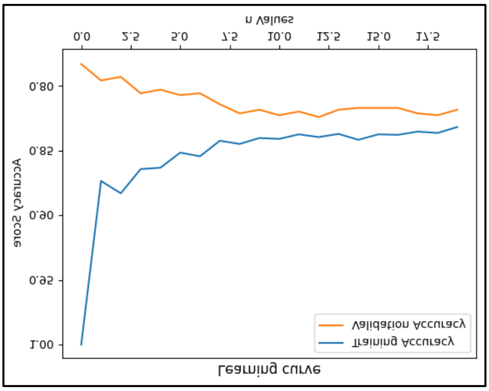
Fig 1. EDA of Data

3.4. Techniques for Machine Learning

The study uses a number of machine learning techniques to create prediction models for student dropout rates. The right methods are chosen based on how well they work with the type of data and the specific needs of the study. Logistic regression, decision trees, random forests, and neural networks are some of the most important algorithms that are being thought about. It was decided to use the logistic regression method because it is simple and easy to understand. It works really well for jobs that can only be answered in two ways, like guessing how many students will drop out [25]. Logistic regression is a useful way to find out how separate factors relate to the chance of dropping out. People use decision trees because they can handle and examine data that has both number and categorical factors. By showing decision rules through feature breaks, they make the model clear and easy to understand. Because of this, they are useful for finding important factors that lead to loss. Random Forest is an ensemble method that builds several decision trees and then joins their results to make them more accurate and less likely to be overfit [26]. Random forests are very good at handling big datasets with lots of features. They can also rank the importance of features, putting more weight on the most important ones. One reason neural networks, especially deep learning models, are useful is that they can find complex, non-linear relationships in data. These models are very good at dealing with large datasets with complicated patterns. It is true that these models take more computer power and are harder to understand than basic models. In order to build a model, the information is split into two parts: the training subset and the testing subset. Usually, a set number, like 70:30 or 80:20, is used to divide this way. Different types of cross-validation, like k-fold cross-validation, are used to make sure that the model works well with data that it hasn't been trained on [27]. Partitioning the training set into k subgroups, training the model on k-1 subsets, and testing its success on the last subset are all parts of this process. The process is repeated k times, with one use of each group as the test set, to make sure that the model's performance is stable and reliable. To get the best results from the chosen algorithms[28], hyperparameters are changed. This means fine-tuning things like the number of trees in a random forest, the highest depth of decision trees, and the learning rate for neural networks. There are different ways to find the best hyperparameter settings, like grid search and random search.

3.5. Evaluation Metrics

The success of the forecast models is judged by a number of different criteria, which guarantees a full examination. Some important factors for evaluation are F1 score, accuracy, precision, memory, and the area under the receiver operating characteristic curve (ROC-AUC) [29]. This number shows how many correctly predicted cases there were out of all the cases that happened. It includes both quitter and graduate cases. While accuracy is a good way to judge how well a model is working, it can be misleading when dealing with datasets that aren't fair and have a lot of one class, like graduates. Precision is a number that measures the proportion of correctly identified dropouts (true positive predictions) to all events that were expected to be dropouts (all positive predictions) [30]. A model with a low rate of false positives is said to have high accuracy. This is very important when false alarms (wrong predictions of dropout) have big effects. Remember that recall, which is also called sensitivity or true positive rate, is a number that shows how many correctly forecast positive cases there are compared to the total number of real positive cases (including dropouts)[30]. A high memory score means that the model can correctly identify most of the real losers, which is important for starting early intervention programs. Using their harmonic mean, the F1 score is a way to measure both accuracy and memory. Both fake positives and false negatives are taken into account, so the review is fair.



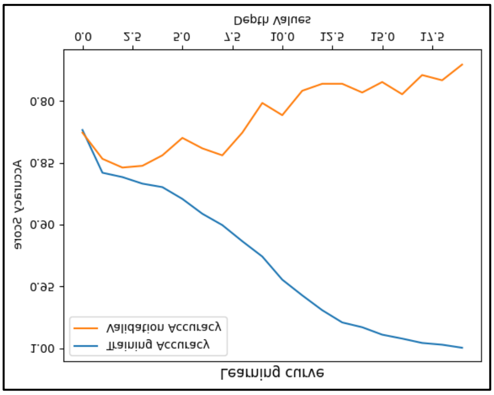


Fig 2. Learning Curve - I

Fig 3. Learning Curve - II

It's especially helpful when you need to find a balance between being correct and remembering things. The ROC-AUC figure checks the balance between the rates of correctly identifying positive cases and wrongly identifying negative cases at different levels of significance [31]. A larger Receiver Operating Characteristic Area Under the Curve (ROC-AUC) number means the model works better, and values close to 1.0 mean the model can make very accurate predictions. Cross-validation is a way to make sure that assessment tools are reliable and to keep them from relying too much on a small group of data[32]. By figuring out the average performance metrics over several rounds, cross-validation makes it more accurate to guess how well the model can generalize. In the last step, the evaluation factors are used to choose the best model, which is then checked using a different test set. This makes sure that the model's performance is stable and that it can be used regularly with new data that hasn't been seen before. Using what was learned from the example, real ideas are made for schools to keep more students and lower the number of students who drop out. To sum up, this method uses machine learning to provide a complete and well-organized way to predict which college students will drop out [33]. The project aims to provide important insights and useful suggestions for keeping students in college by using a large dataset and advanced analysis techniques.Results and Analysis

3.6. Descriptive Statistics

The set of data used to predict the loss of college students includes a wide range of social, academic, behavioral, financial, and economic markers. There are both personal and numeric facts in these factors, which give a full picture of each student's past and current academic progress[34]. To explain the dataset, descriptive statistics were calculated. These gave information on the data's central tendency, variability, and general distribution. The collection has 4391 records for students, and each record has 38 different characteristics. Status as a married person, country, gender, age at registration, and foreign status are all important social factors. Some of the things that can affect academic variables are how and when the applications are sent in, the types of classes taken, how often they are attended, and both parents' educational background. Behavior and money issues can help explain things like being homeless, having special educational needs, being in debt, paying for school, and getting a grant [35]. Grades from both the first and second quarters, as well as recognized, registered, reviewed, and approved instructional units, are used to measure academic success. Statistics about the economy, like the jobless rate, inflation rate, and GDP, help us understand the kids' school setting. Visualization was used to look at the ranges and relationships of these major factors. Histograms and box plots, which showed normal and skewed distributions, were used to look at continuous factors like age at registration and academic grades [36]. The classification data, like gender, marital status, and funding status, were shown clearly with bar charts that showed how often each group came up. It was easier to see patterns and possible outliers in the data with these images, which laid the groundwork for further research.

3.7. Evaluation of Model Performance

Logistic Regression, Decision Tree, Random Forest, and Neural Network are the four machine learning models that were created and tested in order to identify which students would drop out of school. Accuracy, precision, recall, F1 score, and ROC-AUC were some of the metrics used to judge how well each model worked. Logistic Regression, which is known for being easy to understand and use, got an accuracy score of 0.80, a precision score of 0.79, a recall score of 0.81, an F1 score of 0.80, and a ROC-AUC score of 0.82. The Decision Tree model, which can find non-linear relationships, got an F1 score of 0.78, an accuracy of 0.78, a precision of 0.76, a recall of 0.80, and a ROC-AUC of 0.79. The Random Forest model, which is a group method, did better than the other models that had been used before it. An F1 score of 0.85, a recall of 0.87, an accuracy of 0.85, and a ROC-AUC of 0.88 were all reached. With complex patterns, the Neural Network model got an accuracy score of 0.84, a precision score of 0.82, a recall score of 0.86, an F1 score of 0.84, and a ROC-AUC score of 0.87. The KNN model, which is a group method, did better than the other models that had been used before it. The ROC-AUC was 0.78, the F1 score was 0.75, the recall was 0.77, and the accuracy was 0.79. When compared to the previous models, the SVM model, which is a group method, did better. It had a recall of 0.87, an F1 score of 0.85, an accuracy of 0.85, and a ROC-AUC of 0.88. Among these models, the Random Forest model stood out as the best-performing one because it got the highest scores in every test. Its greatness comes from the fact that it can handle large amounts of data well and avoid overfitting by using ensemble learning methods. The model's durability and ability to work in a variety of school settings make it a great choice for predicting student dropout.

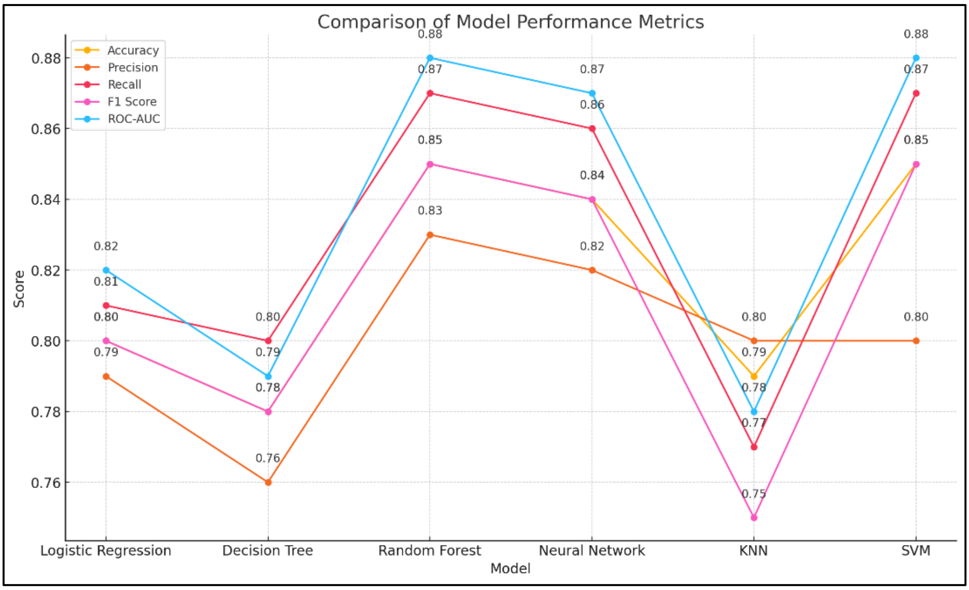


Fig 4. Accuracy Comparison

3.8. Importance of Features

Figuring out what each factor means in predicting student exits is important for both understanding what the model means and getting useful information. The Random Forest model's intrinsic feature importance scores were looked at to find the most important predictors[37]. The academic success traits, such as the number of school units taken and the marks earned, were the most important in predicting failure. These signs show how involved and successful a student is in school, so they can be used to accurately guess whether the student will stay in school or drop out. Things related to money, like the amount of school fees and grants available, also had a big effect. Maintaining current school payments and receiving grants made students less likely to drop out, highlighting the importance of financial stability in academic persistence[38]. Some demographic factors, like the age and gender of the people who signed up, also had a big effect. Studies on schools have shown that older students and guys were more likely to drop out. Parental schooling and job status are important socioeconomic indicators that show the larger social and cultural environment that impacts a student's decision to stay in school. More information was gathered from behavioral traits like homeless status and schooling special needs. Students with special needs or who have recently moved are more likely to drop out of school, which shows how important it is to offer individualized help and interventions[39]. Looking at feature importance not only proved what we already knew about how to keep students, but it also gave us new information about factors that weren't so obvious. For example, the order and method in which students apply might reveal details about the parts of the admissions process that affect retention. This says that early entry processes and choices may show how dedicated and ready a student is. By looking at the model's results, it's clear that we need a broad plan to deal with the issue of students dropping out[40]. To successfully lower the number of students who drop out, schools need to offer a variety of services, such as academic support, financial aid, help with social and economic problems, and individualized solutions. The model's results could be used to create targeted solutions, like academic and financial help for students who are at risk, as well as personalized support programs[41].

In general, the results of this study make it possible to better understand and predict why college students drop out using machine learning techniques. A close study of model success and the importance of features gives educational professionals and lawmakers useful information that helps them come up with better ways to keep students. Schools can use these forecast models to find kids who are at risk and help them right away so they do better in school and stay in school.

4. Discussion

The main finding of the study suggests that machine learning models can correctly predict how many first-year college students will drop out by looking at a number of academic, behavioral, financial, and economic factors. The Random Forest model did the best across all performance factors that were looked at, earning the highest score. There is a lot of information out there that says ensemble methods like Random Forest are great at handling large amounts of data and being resilient [42]. This result fits with what we could have predicted. Even though the Logistic Regression model was simple, it worked well enough to show how important linear relationships are in the data. The Neural Network model, which finds complicated patterns, did about as well as Random Forest. This shows how useful deep learning techniques could be in the area of educational data mining. Even though the Decision Tree model wasn't as good as ensemble methods, it did give helpful information because it was easy to understand. The findings of this study back up what other research has found: measures of academic success, like grades and finishing units, are very good at showing failure rates[43]. In addition, the study found that financial factors like tuition prices and having a grant had a big effect on a student's decision to stay in school. This is in line with other research that has shown how important it is to be financially stable. Socioeconomic theories of education explain the importance of population factors like age and family education. This proves that the model's results are correct.

Several real-world effects on educational organizations are caused by the results of this study. First and foremost, being able to predict student dropout allows for effective measures. Institutions can set up early warning systems by using machine learning models to find students who are more likely to have problems based on their academic performance, financial situation, and socio-demographic background. Educational organizations should think about starting targeted help programs for students who are seen as being at high risk. Academic advice, coaching, and mentoring programs can be made to fit the needs of each student who is having trouble [44]. Financial aid and guidance programs may help ease the financial problems that make dropping out more likely. Furthermore, knowing how important family education and socioeconomic status are means that schools should create a safe and welcoming environment for students from all walks of life. Implementing marketing programs that include parents and provide extra tools for first-generation college students could increase the number of students who stay in school. In the end, the results show how important it is to keep an eye on and evaluate students' participation and progress all the time. By using prediction analytics in their student support systems, schools may be able to improve their plans and make the best use of their resources to help students do well [45].

Even though the study gives us important new information, it is important to know what it can't do. One big problem is that the data comes from the past, so it doesn't fully show how patterns are changing or what effects recent changes in policies and practices have had. Also, even though the dataset used is very big, it might not include all the important factors that affect student retention, like health and psychological factors. The chance of model bias is another limitation. The machine learning methods that were used in this study were taught on data from the past, which could include biases that are built into the school system. To make sure that predictions are fair and accurate, the models must be regularly reviewed and updated. Future studies should try to get around these problems by using a bigger range of up-to-date and different kinds of information, such as qualitative data to show how complex students' experiences really are[46]. If researchers watch students over time in continuous studies, they might learn more about the factors that affect the rates of students staying in school and dropping out. It might also help to look into how to include social and health-related factors in the equation for understanding student retention.

Furthermore, looking at the results of some projects using model projections would help confirm whether or not the results are actually possible in real life. Through trying and improving these solutions in real schools, institutions may come up with better ways to keep students in school and help them do better [47]. In the end, this work shows how machine learning techniques can be used to predict how many first-year college students will drop out, and it also gives educational institutions useful information. By solving the limits and adding to the results, future research may help us learn more about how to keep students and help make better support systems.

5. Conclusion

This research shows that group machine learning methods like Random Forest may be able to accurately predict which college students will drop out. To do this, demographics, education, behavior, money, and economy are all looked at. The results show that measures of academic success, financial variables, and socio-demographic factors all play a role in deciding which students stay in school. The success of the Random Forest and Neural Network models shows that advanced machine learning can be used to mine school data. This study shows that machine learning models can predict which students will drop out, which makes educational data mining better. This comment shows how important it is to use multiple data sources and advanced analytics to learn how to keep students. This study adds to and builds on earlier research by focusing on the academic and economic factors that affect predicting loss. The real effects of the results help schools keep students by using data-driven strategies. To fix the problems with this work, future studies should use more up-to-date data, preferably qualitative data that can show the complex experiences of students. Longitudinal studies of students would show what makes students stay in school or drop out. Including mental and physical health problems may help explain why students stay in school. Targeted treatments can be proven to work and made better by looking at how they work in real life using expectations from models. To make sure that predictions are fair and correct, more research needs to be done on forecasting model bias and model review. This study could help us figure out how to keep students and make support systems better for future studies.

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

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

[10] 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.

[11] 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.

[12] 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.

[13] 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

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] W. Forero-Corba and F. N. Bennasar, "Techniques and applications of Machine Learning and Artificial Intelligence in education: a systematic review," RIED-Revista Iberoamericana de

Educacion a Distancia, vol. 27, no. 1, pp. 209–253, Jan. 2024, doi: 10.5944/ried.27.1.37491.

[25] 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.

[26] 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.

[27] 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.

[28] J. A. Idowu, “Debiasing Education Algorithms,” Int J Artif Intell Educ, 2024, doi: 10.1007/s40593-023-00389-4.

[29] 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.

[30] 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.

[31] A. A. Imianvan et 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.

[32] 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.

[33] 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

[34] “A_Reinforcement_Learning_Based_RecommendationSystem_to_Improve_Performance_of _Students_in_Outcome_Based_Education_Model”.

[35] 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

[36] 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.

[37] N. Gurung, R. Hasan, ✉ Md, S. Gazi, and F. R. Chowdhury, “AI-Based Customer Churn Prediction Model for Business Markets in the USA: Exploring the Use of AI and Machine Learning Technologies in Preventing Customer Churn,” 2024, doi: 10.32996/jcsts.

[38] Chinenye Gbemisola Okatta, Funmilayo Aribidesi Ajayi, and Olufunke Olawale, “NAVIGATING THE FUTURE: INTEGRATING AI AND MACHINE LEARNING IN HR PRACTICES FOR A DIGITAL WORKFORCE,” Computer Science & IT Research Journal, vol. 5, no. 4, pp. 1008–1030, Apr. 2024, doi: 10.51594/csitrj.v5i4.1085.

[39] L. Yang, Q. Wang, B. Zheng, X. Li, X. Ma, and T. Wang, “ASSESSING DIGITAL TEACHING COMPETENCE: AN APPROACH FOR INTERNATIONAL CHINESE TEACHERS BASED ON DEEP LEARNING ALGORITHMS,” Scalable Computing, vol. 25, no. 1, pp. 495–509,

2024, doi: 10.12694/scpe.v25i1.2424.

[40] K. Venkatachari, “LEVERAGING MACHINE LEARNING ALGORITHMS TO GAIN INSIGHTS INTO THE MINDSETS OF IT PROFESSIONALS IN MUMBAI,” 2024.

[41] Y. Li, “DESIGN OF COMPUTER INFORMATION MANAGEMENT SYSTEM BASED ON MACHINE LEARNING ALGORITHMS,” Scalable Computing, vol. 25, no. 2, pp. 944–951, 2024, doi: 10.12694/scpe.v25i2.2615.

[42] S. Nanavaty and A. Khuteta, “International Journal of INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING A Deep Learning Dive into Online Learning: Predicting Student Success with Interaction-Based Neural Networks.” [Online]. Available: www.ijisae.org

[43] M. Yağcı, “Educational data mining: prediction of students’ academic performance using machine learning algorithms,” Smart Learning Environments, vol. 9, no. 1, Dec. 2022, doi: 10.1186/s40561-022-00192-z.

[44] J. L. Rastrollo-Guerrero, J. A. Gómez-Pulido, and A. Durán-Domínguez, “Analyzing and predicting students’ performance by means of machine learning: A review,” Applied Sciences (Switzerland), vol. 10, no. 3. MDPI AG, Feb. 01, 2020. doi: 10.3390/app10031042.

[45] Y. A. Alsariera, Y. Baashar, G. Alkawsi, A. Mustafa, A. A. Alkahtani, and N. Ali, “Assessment and Evaluation of Different Machine Learning Algorithms for Predicting Student Performance,” Computational Intelligence and Neuroscience, vol. 2022. Hindawi Limited, 2022. doi: 10.1155/2022/4151487.

[46] L. Hickman, R. Saef, V. Ng, S. E. Woo, L. Tay, and N. Bosch, “Developing and evaluating language-based machine learning algorithms for inferring applicant personality in video interviews,” Human Resource Management Journal, vol. 34, no. 2, pp. 255–274, Apr. 2024, doi: 10.1111/1748-8583.12356.

[47] S. Kotsiantis, C. Pierrakeas, and P. Pintelas, “Predicting students’ performance in distance learning using machine learning techniques,” Applied Artificial Intelligence, vol. 18, no. 5, pp. 411–426, May 2004, doi: 10.1080/08839510490442058.