DSCI 403
Srijita Ghoshal
March 19, 2024

Underlying Factors that Contribute to University Dropout Rates

Abstract

As universities continue to grow and improve their diversity, they must provide methods to support students from different backgrounds. Although many universities already provide resources for students (like the Colorado School of Mines), it is an iterative process where research is still being conducted. Certain characteristics such as a student's economic situation, whether they are a first-generation student, and their attendance are all key factors that have been proven to affect academic success. However, it should also be important to explore demographic factors like gender, age, nationality, and disabilities to see whether they affect graduation rates. Situational factors like debt and timely tuition payments could also influence student success. Researching and analyzing these trends can allow institutions to provide better support methods for individuals who fit into these demographics. This project explores machine learning models to discover which models best predict a sample's graduation rate. Additionally, these models can enlighten which demographic or situational factors are influencing these rates. To analyze the sample data, machine learning models such as k-nearest neighbors, logistic regression, random forest, and neural networks are being utilized. These models will be used to discover which model has the most accurate predictions. The goal is that exploratory research in this field can lead to higher graduation rates and overall satisfaction from the student bodies that make up these universities. Furthermore, the most successful models can be utilized by universities to determine which students may require additional support for graduation success.

Overview

To explore the elements that are contributing to university dropout rates, a large dataset of many features is needed. Oftentimes, it is challenging to find a huge dataset that is clean and ready to utilize for machine learning. However, this task was simplified by searching through Kaggle, an online website where many preprocessed datasets can be found. From there, a dataset with 36 features was chosen. Overall, these features included information about a student's marital status, previous qualifications, gender, ethnicity, debt, grades, special needs, etc [1]. From a general view, these features seem to cover demographic factors, situational factors, and economic factors as well such as GDP, unemployment rate, and inflation rate. Within this dataset, the feature that states whether a student graduated, enrolled, or dropped out is called 'Target'.

Although universities need to expand resources for students, the analysis of educational success factors is crucial for societal development. If machine learning models can better predict students who may be at risk given certain factors, then universities around the globe can work to prevent these students from dropping out. They can also set up aid programs to guide these students. This benefits society from a larger view because more students will feel supported by their universities, and will be more likely to graduate. The barriers to educational success will be lowered and higher education will be more obtainable despite diverse backgrounds. In short, this leads to more college graduates and a better-educated society. This project will assist in solving this problem.

The motivation for utilizing this dataset is to determine which features may be hindering student success. Since the dataset covers many attributes for each student, it can provide a lot of insight into specific factors that are hurting graduation rates. Similarly, this dataset can uncover factors that are contributing to successful graduation. Therefore, this makes it easier for universities to encourage the positive factors and assist in signs of the negative factors.

Another goal is to find which models work best in predicting student graduation rates. Therefore, this dataset is a good size in which to test different models given that it has a significant set of features and has around 4,500 rows. The data is not too large to be inefficient during the runtime of these models, but it is not too small to cause overfitting.

One of the questions that this project seeks to answer is: what demographic or situational factors are the causes behind student dropout rates? Another question that will be explored is: what machine learning model is the best in predicting student graduation success?

Related Work

Although research in this field is not novel, the methods and the conclusions from this data can be unique. Conclusions that are drawn from different datasets can also be different from prior research that already exists. However, it is important to examine and acknowledge the existing research in this domain. The first study is conducted by a Canadian research team. Their experiment details a machine learning model called RG-DMML in which they use an ensemble algorithm to predict student retention and graduation [2]. This RG-DMML model utilizes a sophisticated version of the k-nearest neighbors model coupled with an industry process of data mining [2]. The work conducted by this research team is indeed similar to the project defined in this paper; however, the project in this paper focuses on multiple different machine learning models to determine the best-performing model. This is unique from the existing research. The study done by the RG-DMML team focuses only on this one model and whether it is successful in predicting student graduation and retention rates [2].

Another related study was conducted by a Turkish researcher, Mustafa Yağcı. His research utilizes different machine-learning models to predict the academic performance of students [3]. This study utilizes models like random forest, support vector machine, linear regression, neural networks, and k-nearest neighbors; however, the focus of these models is to predict the final exam grades of a language class at the University of Turkey [3]. Although many of the models used in this study are the same as the ones used in this paper, the domain is very different between the two. This paper is unique because it hopes to find the best machine-learning method for predicting university graduation rates, whereas the existing study focuses on grade prediction.

The final related study was executed by the University of Nevada, Las Vegas (UNLV). This study utilized machine learning models such as logistic regression, decision trees, support vector machines, and neural networks to identify whether a student will graduate [4]. Although the purpose of this study is very similar to the purpose defined in this paper, the datasets are quite different and contain different features. The dataset that is examined in the study has demographic features such as race, age, income, gender, and Nevada residency [4]. It also has academic features such as GPA, loans, grants, and high school grades [4]. However, it lacks any economic features included in this paper's dataset such as GDP, unemployment rate, and inflation rate. These economic factors could influence student graduation rates. Furthermore, the dataset in the UNLV study contains only students from UNLV and nowhere else, whereas the dataset in this paper comes from multiple different universities in South America [5]. The data in this paper is broad because it can allow the study to be more generalized, instead of focusing on one single university and its specific demographics.

Data Acquisition

As mentioned previously, this dataset comes from Kaggle. The dataset contains a total of 36 features. Each row in the data represents a university student from South America and their information. These features fall into these categories: age, nationality, gender, application information, grades, prior college qualifications, mother/father's qualifications, debt, course load, and economic factors. Most of the data is categorical and has been cleaned by the authors. Therefore, almost all the columns represent the categories using numeric definitions; however, the only column that has not been separated in this way is the 'Target' column, and it will be used as the intended target of the study. This column defines whether a student graduated or not. The string definitions include 'Graduated', 'Enrolled', and 'Dropout.' The table below explains some of the features of this dataset.

Table 1 - Feature Definitions

Data Column	Explanation
Marital Status	Whether or not the student is married
Course	The course that the student is in
Previous Qualifications	The level of education that the student had before attendance
Nacionality	The nationality of the student
Debtor	Whether a student has taken on debt
Education and Special Needs	If a student has special needs
Tuition Fees up to Date	Whether a student has paid all of their tuition
Scholarship Holder	If a student has scholarships
International	If a student is international
Unemployment Rate	The current unemployment rate of the country at the time
Curricular units 1st sem (grade)	The grade received in the first semester

It is important to note that these are not the only features of the study in this dataset, but they represent the gist of the information that is recorded. In short, all 36 features seem very useful in understanding the multiple factors that can contribute to graduation rates. Therefore, they will all be used in the machine learning models. However, certain specific features such as 'Scholarship Holder', 'Tuition fees up to date', 'Curricular units 1st sem (grade)', 'Nacionality', and 'Educational and special needs' will all be examined independently in machine models to see if they affect dropout rates.

Although the dataset was posted to Kaggle, the original owners of this data are UCI (University of California, Irvine). The data comes from their machine learning repository and was published in 2021 [5]. There is no direct license that is defined by UCI, but the authors of the dataset mention that if the data is used in a scientific publication, then to 'kindly' cite it [5]. Therefore, the use of this data seems to be open to creative research and is publicly hosted on a machine learning repository.

Preprocessing

The cleansing for this dataset was very minimal. Considering that the data originally comes from a machine learning repository, it is already well structured and organized. All categorical variables have been defined with numerical values to represent each group. The target variable was the only outlier. However, it was easy to apply the get_dummies() function in

Pandas to define ‘Target_Graduate’, ‘Target_Dropout’, and ‘Target_Enrolled’ as numerical values.

Since the data was clean, it was simple to visualize it as well and to perform some exploratory statistics. The first visualization that was performed was a correlation matrix which made it clear how each feature was correlated to each other. The following figure is a heatmap matrix of all the features showing the different correlations between them. The graph is very dense; however, the color bar on the right side can provide some insight into how closely correlated or not correlated the features are.

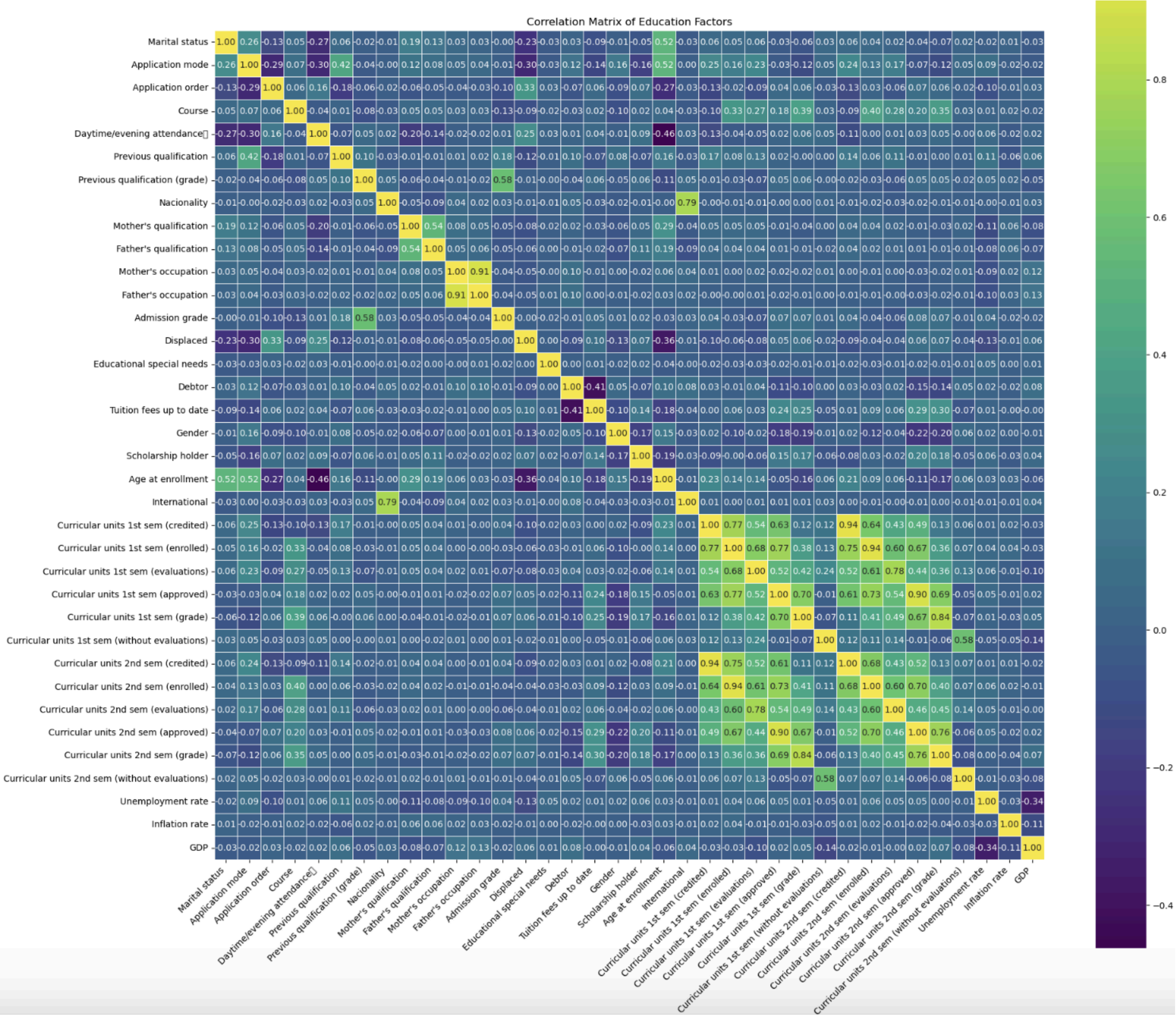

Figure 1 - Correlation Matrix of Feature Set

Now that a broad overview of the data has been established, subparts can be explored even further. The next visualization is an age distribution of all the students at enrollment. In the figure below, it is evident that this distribution is skewed toward most students being in their early 20s. The right tail shows that very few students are over this age and that there are rarely any students who are above 60 years of age.

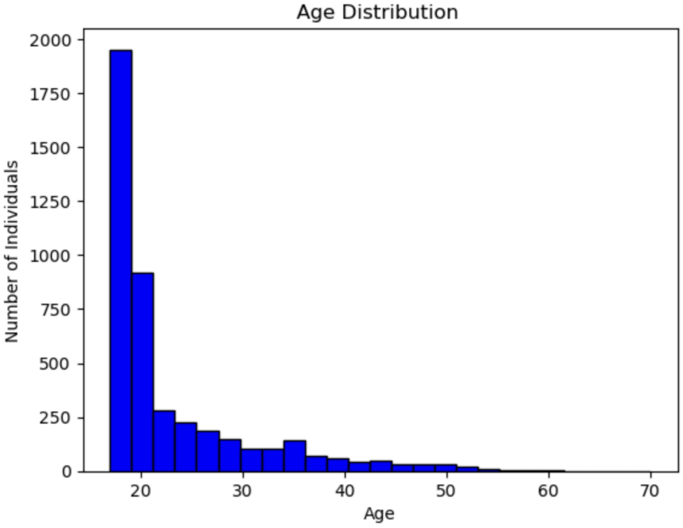

Figure 2 - Age Distribution at Enrollment

The next feature of the dataset that was examined was the percentage of scholarship holders. This was graphed using a pie chart. In the figure below, it can be examined that 24.8% of students in this data were scholarship holders. This leaves about ¾ of the population to not hold any scholarships.

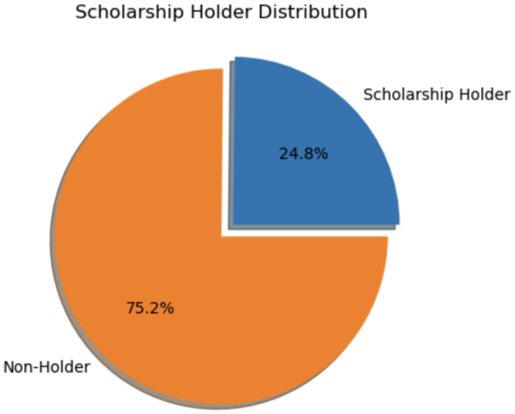

Figure 3 - Scholarship Holders

Another item that was visualized was the makeup of the students who had graduated, enrolled, and dropped out. This visualization is important because it provides insight into the distribution of the target variable. The figure below is a pie chart showing the makeup. About 50% have graduated, 17.9% enrolled, and 32% dropped out.

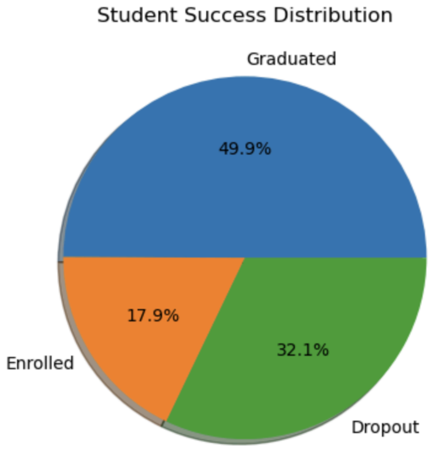

Figure 4 - Student Graduation, Enrollment, and Dropout Rates

The last item that was visualized was the academic scores before student enrollment. This was denoted as the ‘Admission Grade’ category in the dataset. The scores have been standardized on a scale of 0 to 200. The graph below shows the histogram of the grades and draws a bell curve showing the approximately normal distribution.

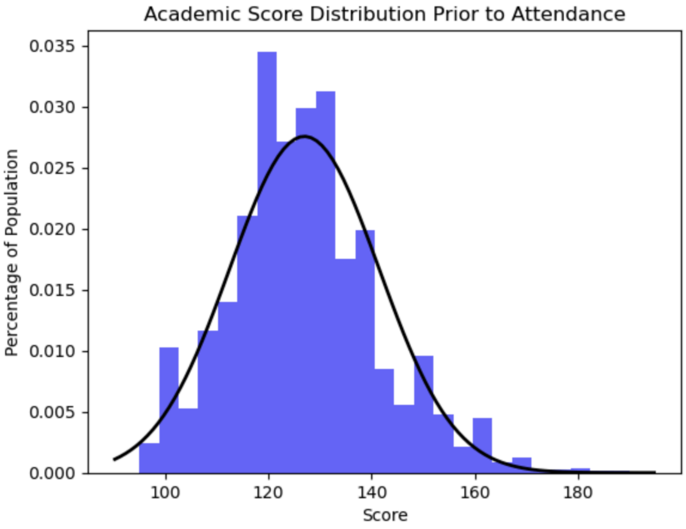

Figure 5 - Academic Grades Before University Distribution

Overall, the dataset was not too large. It contains about 4,500 rows and 36 features, so the dimensionality was not concerning. To verify this, the data was run through some test models to see if the learning was slow or presented any alarm signs. It was also run through testing dimensionality reduction. Both methods had good performance, so the final experiment did not use dimensionality reduction. Instead, the goal was to focus on as many different features, so there could be more explainable causes of university dropout rates.

Model Selection

Now that the data has been explored, it is appropriate to pick the most viable machine learning models. A multitude of models were selected to compare which ones would perform the best. The biggest factor in considering model selection was the fact that the target variable was a categorical classification. This meant that the models selected would need to be compatible with categorical data. After careful consideration and thought, the models that were selected were random forest, k-nearest neighbors, logistic regression, and neural networks.

Random forest was chosen as one of the models because of its ability to perform either classification or regression [6]. In the case of this dataset, the model is required to perform classification. Furthermore, a random forest is a collection of many decision trees, a machine-learning model that uses branches and leaves to show the decision-making process. Since a random forest consists of many decision trees, there is a lot more accuracy in its predictions. Decision trees alone can lead to overfitting [7]. It is notable that since random forests are a collection of many decision trees, the model is computationally expensive [7]. However, a random forest can output the feature importance: how largely a feature affects the predictions of the model [8]. This is quite helpful in understanding which factors may be contributing to dropout rates in universities.

Moving on to K-nearest neighbors, which is a model used for classification. This model classifies new data points given the distance from neighboring points [6]. The advantages of this model include the fact that it is easy to implement and time efficient because there is no training period [9]. This model was chosen to be in this study because of its simplicity in comparison to the other models. However, some drawbacks of this model include the fact that it is inefficient for large datasets, struggles with high dimensionality, and is sensitive to noise in data [9]. Despite these issues, it is important to compare the performance of this model against the others.

As for logistic regression, this model is great for determining the probability of a target given a set of features, or even a single feature [10]. This model was chosen because it was a reliable tool to explore the relationship between certain features of the dataset and the likelihood of dropping out of university. Relationships between single and multiple features can be examined. However, this model's largest drawback is its struggle to interpret complex relationships between variables. Instead, a neural network is much more capable of this [11].

This leads to the final model in this study, neural networks. This model works best with complex or multi-faceted data because neural networks are modeled after the deep-thinking process of human brains. This is achieved by assigning weights to the inputs and moving them through an activation function that outputs the predictions. These units that do the computation are called perceptrons [12]. Since this model works well with dimensional data it is a strong contender in analyzing student success factors. Predictions in student success could include many different factors that neural networks would likely excel at comprehending. It can be a reliable tool in predicting a student's likelihood to drop out given certain factors. Like random forests, neural networks work with both classification and regression problems, but they will only be used for classification in this study [7]. However, one of the biggest issues with a neural network is that it is computationally and memory-heavy [13]. This can be a problem with large datasets, but the dataset in this paper is not large enough to pose a problem. In short, the intricate nature of this model should perform well with nuanced data like student dropout factors, and it will be very interesting to compare its performance against the less complex models in this study.

Results and Evaluation

Given that the models were established with purpose, the final step was to evaluate their performance individually and against each other. The evaluation began with the random forest model. The goal of this model was to understand the feature importance when it came to student dropout rates. Therefore, this model was given all the features of the dataset just to generalize the data to see which attributes would prevail. The target of this model was set to be the students who dropped out of university. The model was given 1,000 estimators (decision trees) with a max depth of 10. Once the learning was complete. The feature importance graph was outputted below:

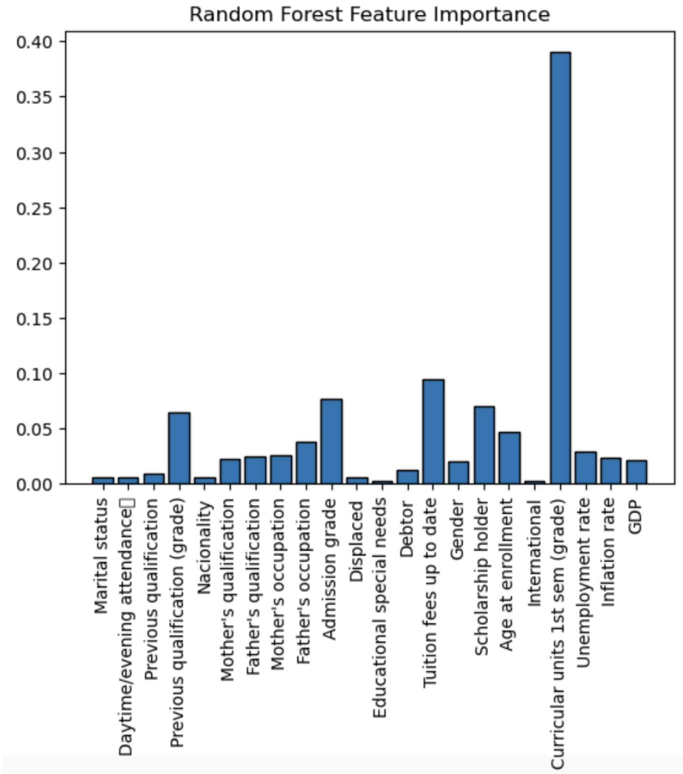

Figure 6 - Random Forest Feature Importance

Interestingly, the most pivotal feature in this classification is the 1st-semester grades of students, and the model has marked this with great importance. This is not a surprising fact since students in general are more likely to succeed if they maintain good grades within their first few years of college. Often, good grades set a precedence for success and motivate against dropping out. However, the random forest model marked that tuition fees being up to date, admission grades, and previous qualifications were the next most significant factors. The tuition fees being up to date can be a notable influence on student success because students cannot continue if they are unable to pay off tuition. The admission grades and previous qualifications can also be a factor since these attributes can determine the previous work ethic of a student. If they have high achieving grades from their high school resume, then they are likely to be high achievers due to their experience.

It is also notable that the random forest classifier did not mark demographic factors such as marital status, nationality, special needs, international, and displaced to be influential in the decisions of the model. These factors contributed least to importance. The only slightly significant demographic factor seems to be the age at enrollment. Other economic factors such as GDP, unemployment rate, and inflation rate also seem to be of minimal significance which indicates the model did not find them to be very important in decision making.

In comparison to the results of all the models researched in this study, the random forest had the worst performance. In fact, its R-squared value came out to be 0.456 which indicates that the model could not confidently explain the variance in the model. Therefore, it is important to take the feature importance of the model lightly considering its performance was weak, but there is no need to count it off entirely. The low R-squared value is often associated with data that tries to explain human behavior [14]. Data in this domain has a lot of variance by nature, and it can be hard to statistically explain this variance, but the feature importance can still be insightful [14]. In all, the random forest model still provided some understanding of student dropout factors.

The next model that was evaluated was the k-nearest neighbors model. All the features of the dataset minus the target were fed into the model. To evaluate the best performance of this model two techniques were utilized: parameter search and k-fold cross-validation. The model was tuned to see which k would output the highest accuracy scores: from 1 to 20 neighboring data points. Then, to ensure that the scores were as accurate as possible, each k had a k-fold cross-validation performed on it. The data was split into 5 folds for each validation. Lastly, the mean score was taken from the performance and compared against the regular accuracy without k-fold validation. This outputted the graph below:

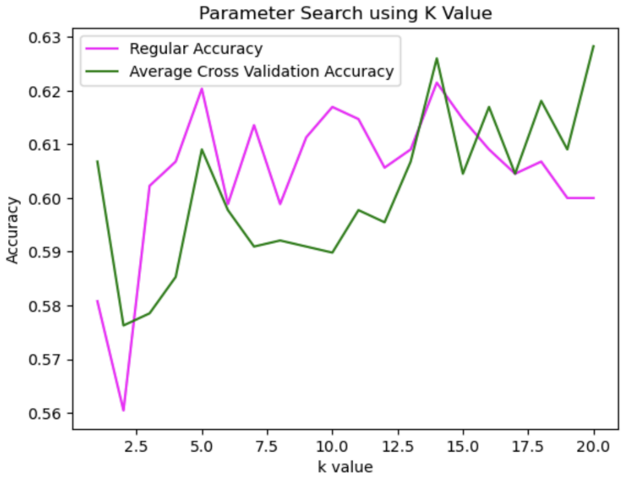

Figure 7 - KNN with K-Fold Cross Validation and Parameter Search

The results of the model show that the KNN model performed best with k = 14 with both the regular accuracy and average cross-validation accuracy being around 0.620 to 0.625. This means that the model was decent in classifying students likely to drop out given the many features.

The next model that was utilized was logistic regression to predict if a student will drop out given different sets of features. This model performed best when given the whole set of features and the target of students who dropped out. In fact, this scored the second highest out of all of the different models in this study with an accuracy score of 0.831. However, there was more learning to be done to test how certain factors impact student dropout rates. Therefore, the table below outlines the tested features and the outcomes of the machine learning. The target was kept as the dropout column. The 'Situational Factors' shown in the table below describe the set of all factors minus any demographic attributes.

Table 2 - Logistic Regression Scores with Different Feature Subsets

Factors	**Accuracy Score**
Scholarship Holder	0.685
Tuition Fees Up to Date	0.767
Curricular Units (1st Sem) Grade	0.759
Nationality	0.684
Educational Special Needs	0.667
Purely Demographic Factors	0.723
Purely Situational Factors	0.833

In short, logistic regression performed best when it came to explaining situational factors such as debt, grades, economic factors, prior qualifications, etc. The model performed worse when predicting using scholarship holders, nationality, and educational special needs. However, overall the logistic regression model excelled in comparison to KNN and random forest models.

The last model left to evaluate was the neural network. This model was written simply with only two hidden layers. Each layer utilized the ReLU activation function with 100 hidden neurons. To ensure the best accuracy of the model, it was trained and tested through 250 epochs. The output of the epochs is shown in the graphs.

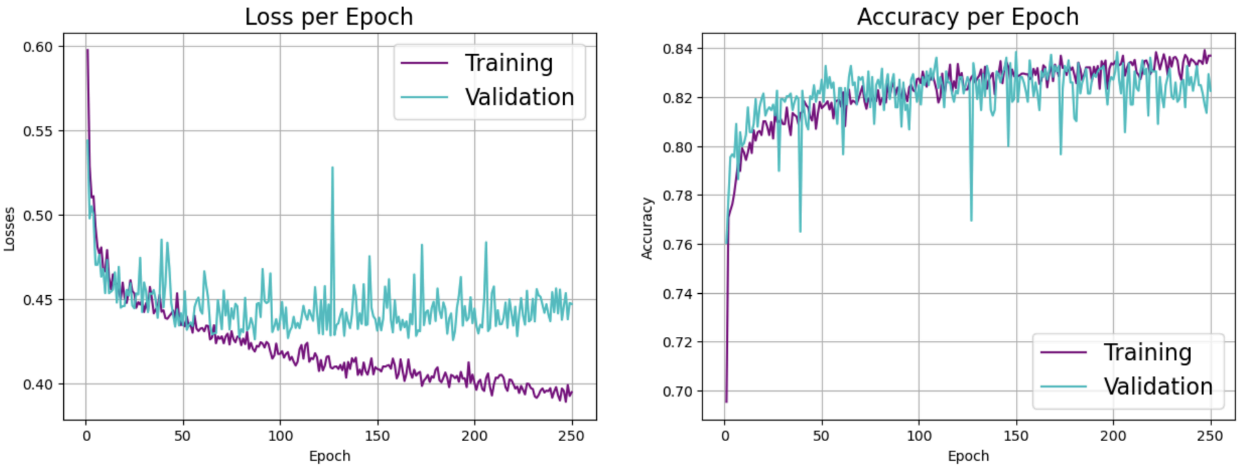

Figure 8 - Loss and Accuracy Per Epoch

As seen above, the loss begins to increase in the validation data as around 50 epochs are reached. The deviation between the validation data begins to increase from here; however, it does not drift dramatically. Therefore, the model can predict the values without significant error [15]. This is validated by the accuracy seen for each epoch, in which the training dataset reaches a maximum of 0.839. This is the highest out of all models in the study; however, it is only marginally more accurate than the logistic regression model. It is also notable that the training and validation scores are quite close together with each epoch meaning that the model is generalizing well to the data [16]. Therefore, its performance is quite good and given its accuracy scores, it takes the lead as the best predictor of student dropout rates. However, it should be noted that this model is the slowest and resource-intensive out of all that was compared. When it comes to the least heavyweight model, logistic regression performed the best.

Ethics

Working with real data always has ethical implications. Even more, working with student data can have large concerns regarding privacy. One of the biggest issues that must be adhered to is FERPA compliance. This is governed by U.S. law and requires that educational institutions must limit third-party access to student records [17]. Therefore, when performing machine learning on student data it is important that this data is kept extremely confidential and that students cannot be individually identifiable by this data. Furthermore, this data must be obtained in an ethical manner in which students consent to their data being utilized for further study. Some students may not want their information available and being tested.

Another ethical implication that needs to be avoided is wrongfully flagging students as more likely to drop out. Given that machine learning models can identify certain factors that may influence student success, it would be unethical to flag a student as not capable of graduating based on these factors alone. These models cannot explain real data with 100% accuracy. It may mark someone as less likely to graduate, but it cannot guarantee the outcome. Instead, the model should be used as a tool for monitoring students who already show concern rather than a preemptive measure.

Furthermore, institutions should not utilize this tool to dismiss applicants given that the applicant's factors show concern. This would deny them a fair chance at education. There should be admission of a diverse body of students with many different backgrounds and situational factors. This diversity is what fosters a stronger education given the many different viewpoints. To exclude individuals from entering a school based on their background is discrimination and does not comply with the ethical standards of society.

Future Work

This study could be expanded upon by exploring the feature importance of dropout rates in more depth. Although the random forest model was able to explore this importance to some degree, it was lacking in accuracy. The feature importance of a more accurate model should be evaluated, so that more reliable conclusions can be drawn about these factors. Perhaps, the feature importance produced by the neural network could be examined given that it performed the best out of all models in the study.

Furthermore, the neural network used in this study was very simple with only two hidden layers. To see better results in complex data, a more robust deep learning model should be used. In the future, a model with more hidden layers and neurons should be utilized. Additionally, the neural network used in this study only utilized one type of activation function (ReLU) and perhaps different activation functions would yield different results. This should be explored to determine if the model can be tuned to provide better performance.

Notably, the neural network used in this study did not assign any weights to the model. This leads to the weights being set randomly [18]. This can cause issues such as overfitting, vanishing gradient problems, or exploding gradient problems [19]. Therefore, it is recommended that the weights be set using some form of weight initialization function. Then, the model can be run and optimized to figure out which weight function works best. This creates models with high accuracies [19]. These types of distinctions are what distinguish a regular neural network from one that is more complex and deep in learning.

For a project that determines a student's likelihood to drop out, it is of utmost importance that the model used for machine learning is accurate. Although the neural network used in this project was the most successful out of all models, it has quite a few aspects that could be improved upon. Making these improvements would ensure a more reliable and accurate model.

The last item that could be explored is the question of does student life at university play into dropout rates. Items such as quality of dining, social event attendance, club involvement, etc. could be examined. This would require the collection of new data preferably around different universities around the United States. The data used in this study only came from South America and the demographics in the U.S. could be very different. Therefore, the U.S. may have different factors that influence dropout rates. Overall, these aspects need to be explored further to generalize the study to the United States.

Sources

[1] Vasu_Avasthi, “Student’s dropout and Academic Success Dataset,” Kaggle, https://www.kaggle.com/datasets/missionjee/students-dropout-and-academic-success-dataset (accessed Mar. 19, 2024).

[2] 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, https://www.sciencedirect.com/science/article/pii/S2666920X24000067 (accessed Mar. 19, 2024).

[3] M. Yagci, “Educational Data Mining: Prediction of Students’ academic performance using machine learning algorithms - smart learning environments,” SpringerOpen, https://slejournal.springeropen.com/articles/10.1186/s40561-022-00192-z (accessed Mar. 19, 2024).

[4] E. C. Ploutz, Machine Learning Applications in Graduation Prediction at the University of Nevada, Las Vegas , https://digitalscholarship.unlv.edu/cgi/viewcontent.cgi?article=4312&context=thesesdissertations (accessed Mar. 19, 2024).

[5] Realinho,Valentim, Vieira Martins,Mónica, Machado,Jorge, and Baptista,Luís. (2021). Predict Students' Dropout and Academic Success. UCI Machine Learning Repository. https://doi.org/10.24432/C5MC89.

[6] “Choosing the Best Machine Learning Classification model and avoiding overfitting,” Choosing the Best Machine Learning Classification Model and Avoiding Overfitting - MATLAB & Simulink, https://www.mathworks.com/campaigns/offers/next/choosing-the-best-machine-learning-classification-model-and-avoiding-overfitting.html (accessed Mar. 19, 2024).

[7] G. L. Team, “Random Forest algorithm in Machine Learning: An overview,” Great Learning Blog: Free Resources what Matters to shape your Career!, https://www.mygreatlearning.com/blog/random-forest-algorithm/ (accessed Mar. 19, 2024).

[8] T. Shin, “Understanding feature importance in machine learning,” Built In, https://builtin.com/data-science/feature-importance (accessed Mar. 19, 2024).

[9] A. Soni, “Advantages and disadvantages of KNN,” Medium, https://medium.com/@anuuz.soni/advantages-and-disadvantages-of-knn-ee06599b9336 (accessed Mar. 19, 2024).

[10] “What is logistic regression?,” IBM, https://www.ibm.com/topics/logistic-regression (accessed Mar. 19, 2024).

[11] “Advantages and disadvantages of logistic regression,” GeeksforGeeks, https://www.geeksforgeeks.org/advantages-and-disadvantages-of-logistic-regression/ (accessed Mar. 19, 2024).

[12] M. Banoula, “What is Perceptron? A beginners guide for 2023: Simplilearn,” Simplilearn.com, https://www.simplilearn.com/tutorials/deep-learning-tutorial/perceptron (accessed Mar. 19, 2024).

[13] N. Donges, “4 disadvantages of Neural Networks,” Built In, https://builtin.com/data-science/disadvantages-neural-networks (accessed Mar. 19, 2024).

[14] “Are low R-squared values always a problem?,” Quora, https://www.quora.com/Are-low-R-squared-values-always-a-problem (accessed Mar. 19, 2024).

[15] Shankar297, “Understanding loss function in deep learning,” Analytics Vidhya, https://www.analyticsvidhya.com/blog/2022/06/understanding-loss-function-in-deep-learning/ (accessed Mar. 19, 2024).

[16] Aesir, “Validation accuracy is always close to training accuracy,” Data Science Stack Exchange, https://datascience.stackexchange.com/questions/42606/validation-accuracy-is-always-close-to-training-accuracy (accessed Mar. 19, 2024).

[17] L. M. University, “Ferpa - rights and privacy act,” FERPA - Rights and Privacy Act - Loyola Marymount University, https://registrar.lmu.edu/ferpa-rightsandprivacyact/#:~:text=The%20Family%20Educational%20Rights%20and,the%20parent%20to%20the%20student. (accessed Mar. 19, 2024).

[18] V. Kurama, “An introduction to Deep Learning and tensorflow 2.0,” Built In, https://builtin.com/machine-learning/introduction-deep-learning-tensorflow-20 (accessed Mar. 19, 2024).

[19] “Weight initialization techniques for deep neural networks,” GeeksforGeeks, https://www.geeksforgeeks.org/weight-initialization-techniques-for-deep-neural-networks/ (accessed Mar. 19, 2024).