DEC MINIPROJECT

Predicting Graduation and Dropout Rates

Siddhivinayak Kulkarni
Department of Computer Engineering and Technology
Dr. Vishwanath Karad MIT World Peace University
Pune 411038, Maharashtra, India

Group members

PG-27 Atharva Thorat
PRN:1032212077

PG-23 Aryan Goyal
PRN:1032212016

PG-24 Kunal Suryawanshi
PRN:1032212019

***Abstract*—Student dropout is a complicated and detrimental issue in the educational process, with costs to both students and institutions on a social and financial level. The tool presented in the paper uses machine learning techniques to forecast first-year undergraduate student dropouts. In order to calculate the risk of dropping an academic course, the tool takes into account personal information, secondary school academic records, and first-year course credits. In this paper, a broad study on the applications of the algorithms of machine learning is presented for predicting educational enrollment status. The research models in this situation. The used dataset includes a wide range of financial, academic, and demographic variables, offering a rich source of data for modelling. This study lays the groundwork for future developments in the field of predictive modelling for student enrollment and adds to the body of knowledge in that area. Researchers and practitioners looking to use similar predictive models in educational contexts can benefit greatly from the methodology and insights presented here.Focuses on comparing the efficacy of Support Vector Machine (SVM) and Gaussian Naive Bayes (GNB)**

Keywords—Machine Learning, Student Dropout, Classification, Naïve Bayes, SVM.

1. INTRODUCTION

In the realm of education, the success of high graduation prices and prevention of student dropouts are fundamental dreams for academic institutions and policymakers. The pursuit of better training represents not only a person's quest for knowledge and self-improvement but also a societal investment in human capital development. However, the persistent task lies in figuring out college students susceptible to dropout early enough to offer them vital guidance and interventions, ultimately guiding them toward successful graduation.

The consequences of high dropout rates are far-reaching and impacting on not only individual students but also their communities and societies. Dropouts regularly face diminished financial prospects, decreased earning ability, and restricted career possibilities that could perpetuate cycles of poverty and inequality. Moreover, the societal fees of misplaced capacity and productivity are giant, making the prediction and prevention of dropouts a count of good sized social and financial importance.

On the other hand, educational institutions benefit from high graduation rates, as they reflect the success of their programs and the fulfillment of their mission. Additionally, graduates contribute definitely to the group of workers, using financial increase and innovation. Thus, it is inside the high-quality interest of educational establishments to discover and guide students at the risk of not finishing their research.

This research paper delves into the crucial mission of predicting whether a pupil will in the end graduate or drop out. By leveraging the power of data analytics, gadget mastering, and statistical modeling, we aim to develop correct predictive fashions which can perceive at-hazard students early in their instructional journey. In doing so, we intend to make a contribution to the overarching aim of increasing graduation charges and reducing dropout rates in academic institutions.

To attain this, we will explore various factors that impact scholar results, along with demographic variables, academic performance metrics, attendance information, and socioeconomic indicators. By analyzing these elements and their interplay, we are searching to uncover styles and insights that could guide the development of effective predictive fashions. Once proven and refined, those fashions can function as valuable gear for educators, directors, and policymakers, enabling them to strategically allocate sources and interventions, thereby fostering a greater supportive and a successful educational environment.

In the following pages, we can delve into the literature surrounding scholar final results prediction, detail our method for statistics series and evaluation, gift our findings, and discuss the results of our research. Ultimately, we are hoping that this undertaking will make contributions to the wider challenge of nurturing the educational trips of all students, permitting them to attain their complete capacity and reap their instructional aspirations.

2. LITERATURE REVIEW

There hasn't been much quantitative research on the causes and solutions for student attrition in higher education . Feature engineering, which is frequently used in customer churn analysis, can be used to enhance student attrition prediction models.As alternatives to hand-engineered features in predicting student attrition, convolutional and recurrent neural networks can be investigated. Indicators like introductory STEM classes, remedial courses, and first-year interest groups are also taken into account by the study as potential predictors of student dropout.Although this information was missing from the dataset used for the study, financial standing and history are acknowledged as significant factors influencing students' decisions to stop their studies. The most effective dropout predictors are found using regularised linear regression and logistic regression.[1]

Various papers have explored the application of machine learning methods to foresee students' academic achievements. Higher-level education has been the subject of some studies that trained machine learning models with various datasets and features.In one study, features related to student provenance were included in a balanced dataset, whereas in the other, features related to family status and personal circumstances were used. The ID3 algorithm outperformed the other classification algorithms that were tested. In another study, webcam and eye-tracker data were used to predict student dropout using emotion analysis. Studies on high school education are less generalizable because different countries have different educational systems. However, the focus is different from the suggested solution. Numerous studies have brought forth the predictive models aimed at forecasting students' ultimate outcomes in blended or online education.[2]

3. METHODOLOGY

Each paragraph must show indentation meaning they should have both left and right-justified margins.

A. Data Preprocessing & Data Cleaning

- Firstly, we determined and dealt with any incomplete or missing data. It may be necessary to impute missing values using techniques like mean, median, and mode impute, as well as more sophisticated techniques.
- Outliers can be foretell as data points which significantly deviate from the remaining data; identify them and deal with them. Outliers may need to be eliminated, modified, or capped to achieve this.

- Normalised attributes to make them comparable in size. This is crucial for algorithms like gradient descent-based algorithms that are sensitive to the scale of features.

B. **Feature Selection**

Some features that we selected that are relevant and useful for prediction are:

- Marital Status: This feature might be relevant as it could potentially influence a student's decision to enroll. For instance, married individuals might have different enrollment patterns compared to single individuals.
- Application Mode: The mode through which the student applied (e.g., online application, in-person, etc.) could be indicative of their commitment to enrollment.
- Previous Qualification: This feature could be crucial in understanding a student's academic background, which may influence their decision to enroll in a new course.
- Admission Grade: The grade at which a student was admitted could serve as an indicator of their academic performance and commitment to the program.
- Scholarship Holder: This binary variable might play a significant role, as scholarship holders may have a higher likelihood of enrollment due to financial incentives.
- Age at Enrollment: Age could be a factor in enrollment decisions. Younger students might be more inclined to enroll compared to older individuals.
- Gender: Gender might have an impact on enrollment patterns, as there might be gender-specific considerations or preferences.

C. **Applied LabelEncoder**

Label encoding is a technique used in machine learning to handle categorical data, transforming non-numerical labels into numerical labels. It is particularly useful for algorithms that require numerical inputs. The label encoder assigns a unique integer to each unique category or label in the categorical feature.

Handling Class Imbalance :

A potential class imbalance was addressed using the SMOTE technique. For the minority class, this entails creating synthetic samples, ensuring a more balanced representation in the training data.

By using SMOTE, the model is prevented from being biased towards the majority class, which could result in incorrect predictions for the minority class. The dataset is balanced again through the use of synthetic instances, which improves the model's capacity to generalise to both classes.

D. **Prediction using Machine Learning Algorithms**

In this research paper, the predictive model consists of some machine learning algorithms are:

1. Support Vector Machines (SVM)
2. Gaussian Naive Bayes (NB)

➢ **Support Vector Machines :**

A strong and adaptable machine learning method called Support Vector Machines (SVM) is employed for both classification and regression tasks. It operates by identifying an ideal hyperplane in a high-dimensional feature space that effectively separates classes. SVM is especially useful when the data is unable to be segregated linearly since the chosen hyperplane optimizes the boundaries between the classes to the greater extent. It is accomplished by transforming the data into a space of higher-dimension, that enables a more distinct separation.

SVMs are effective supervised learning tools that may utilize both classification as well as regression applications. SVMs were first developed by Vapnik and Cortes in the 1990s, and they are now widely used in many fields, including machine learning, computer vision, biology, finance, and many more.

SVM's primary objective is for identifying the best hyperplane in a higher dimension feature space for separating classes. The margin between the classes is maximized by this hyperplane, creating a solid decision boundary. The data points which lie closer to the hyperplane are referred to as "support vectors" and they are very important in identifying the direction and location of the decision boundary. SVMs are especially useful in situations where the data is unable to be separated in a straight line. They use the "kernel trick" to plot the data into a space of high dimension where it might become linearly detached. The linear, polynomial, and radial basis function (RBF) kernels are frequently employed.

Support vectors play a crucial role in the position and orientation of the hyperplane and also lie closest to the decision boundary. The decision boundaries are resolved by the support vectors and can be used to make predictions for new data points.

Support Vector Machines stand as a formidable class of algorithms that continue to be widely used in diverse research and practical applications. Their way of handling linear and non-linear data, coupled with proper hyperparameter tuning, ensures their relevance and effectiveness in a wide array of real-world scenarios. Understanding the nuances of SVMs is imperative for researchers aiming to harness their full potential in their studies.

- **Gaussian Naive Bayes :**

Gaussian Naive Bayes (GNB) is a classification algorithm that operates by probabilistic algorithms based on the Bayes theorem and predicts feature independence. It is frequently employed in machine learning for text classification, disease diagnosis, and spam filtering.

GNB is rooted in Bayesian probability theory, specifically Bayes' theorem. It computes the conditional probability of a class given the feature values and makes predictions based on the maximum posterior probability.

GNB models the probability density function of each class as a Gaussian distribution. For each feature, it estimates the mean and variance of the feature's values within each class. This information is used to calculate the likelihood of a given feature value belonging to a particular class.

GNB is well-suited for continuous data due to its reliance on Gaussian distributions. It assumes that the feature values within each class follow a normal distribution, allowing it to effectively model continuous variables. While GNB is inherently designed for continuous data, it can be extended to handle categorical features through techniques like binning or encoding categorical variables as numerical values.

During the training phase, GNB calculates the mean and variance of each feature within each class. These statistics are essential for computing the likelihood of a feature value belonging to a class. To make predictions, GNB computes the likelihood of each class based on the feature values using Bayes' theorem. The class with the greatest likelihood is chosen as the forecasted class for a given set of features.

A fundamental and adaptable classification algorithm known for its ease of use and effectiveness is GNB. It has proven effective in various applications despite its underlying presumptions. When utilizing GNB for their particular research endeavors, researchers and practitioners should be aware of its limitations and make the most of its strengths.[3]

E. Model Evaluation Metrics :

In this project, which includes the application of Support Vector Machines (SVM) and Gaussian Naive Bayes (GNB) algorithms, several model evaluation metrics have been employed to evaluate the effectiveness of the predictive models.[4] These metrics play a crucial role in objectively quantifying the effective performance of a generalized model to new, unobserved and unseen data. Below are the detailed explanations of the evaluation metrics used:

1. Accuracy:

Definition: In classification tasks, accuracy is a key metric that quantifies the proportion of accurately predicted instances amidst all instances in a dataset.

2. Classification Report:

Definition: An in-depth analysis of the effectiveness of the model for every class in the dataset is provided in a classification report. The following metrics are included:

a) Precision: Precision, in other words, Positive Predictive Value, is the percentage of true positive predictions amidst every positive prediction.

b) Recall: Recall, in other words, Sensitivity or True Positive Rate, is the percentage of true positive predictions amidst every actual positives.

c) F1-Score: The F1-Score is the harmonic mean of precision and recall, it provides a fair assessment of the model's performance.

d) Support: Support represents the number of occurrences of each class in the true dataset.

These evaluation metrics give an all-encompassing picture of how well the predictive models perform in the context of student enrollment. They take into account the models' overall accuracy as well as how well they perform on particular classes, which is essential for identifying the models' strengths and weaknesses and selecting the best algorithm for a given problem.

4. CONCLUSION

Results of the models that we used for our prediction are:

1) Support Vector Machines -
   - Accuracy Score of training data: 78.185 %
   - Accuracy Score of testing data: 75.141 %
2) Gaussian Naive Bayes –

Table- 1. Naïve Bayes Performance

	Precision	Recall	F1-Score	Support
Dropout	0.81	0.69	0.75	316
Enrolled	0.37	0.26	0.21	151
Graduate	0.71	0.86	0.78	418
Accuracy			0.71	885
Macro Avg	0.63	0.60	0.61	885
Weighted avg	0.69	0.70	0.69	885

5. FUTURE SCOPE

This paper presents an enormous contribution to the field of educational enrollment prediction. The study lays the basis for numerous promising avenues of future studies:

1. Ensemble Methods and Advanced Algorithms:

- Future studies could discover the application of ensemble methods like Random Forests, Gradient Boosting, or advanced algorithms like Neural Networks. These fashions can also provide in addition enhancements in predictive accuracy and robustness.

2. Hyperparameter Optimization:

- Conducting a thorough hyperparameter tuning method for both SVM and Gaussian Naive Bayes models may want to doubtlessly yield even higher consequences. Techniques like grid search or randomized seek can be employed.

3. Feature Engineering and Selection:

- Further research into feature engineering strategies and more superior function selection techniques, including Recursive Feature Elimination or version-primarily based selection, may additionally cause a more optimized set of predictors.

4. Incorporating External Data Sources:

- Integrating additional external facts assets, along with socio-monetary indicators, local demographics, or instructional coverage changes, may additionally provide an extra comprehensive view and decorate prediction accuracy.

5. Longitudinal Studies:

- Conducting a longitudinal study to assess the lengthy-time period effectiveness and balance of the predictive models would provide treasured insights into their actual-international applicability through the years.

6. Generalization to Different Educational Institutions:

- Testing the fashions on datasets from different instructional institutions or contexts would assess their generalizability and flexibility to numerous settings.

7. Cost-Benefit Analysis:

- Evaluating the fee-effectiveness of implementing predictive fashions in instructional institutions, considering factors like aid allocation and intervention strategies, should provide sensible insights for choice-makers.

8. User-Friendly Deployment:

- Developing user-pleasant interfaces or equipment that allow academic institutions to effortlessly implement and interpret the predictive fashions in their choice-making tactics.

These future studies instructions goal to construct upon the inspiration laid with the aid of this observation and similarly advance the field of predictive modeling for instructional enrollment. Each avenue presents specific possibilities for innovation and contributes to the broader aim of improving academic planning and resource allocation.

References

[1] Aulck, L. (2016, June 20). *Predicting Student Dropout in Higher Education*. arXiv.org. https://arxiv.org/abs/1606.06364

[2] DelBonifro, F., Gabbrielli, M., Lisanti, G., & Zingaro, S. P. (2020, January 1). *Student Dropout Prediction*. Lecture Notes in Computer Science. https://doi.org/10.1007/978-3-030-52237-7_11

[3] M. Solis, T. Moreira, R. Gonzalez, T. Fernandez and M. Hernandez, "Perspectives to Predict Dropout in University Students with Machine Learning," 2018 IEEE International Work Conference on Bioinspired Intelligence (IWOBI), San Carlos, Costa Rica, 2018, pp. 1-6, doi: 10.1109/IWOBI.2018.8464191.

[4] STRECHT, P., CRUZ, L., SOARES, C., MENDES-MOREIRA, J. & ABREU, R. (2015) A comparative study of classification and regression algorithms for modelling students' academic performance. Proceedings of the 8th International Conference on Educational Data Mining, 392-95.

[5] Masci, C.; Johnes, G.; Agasisti, T. Student and school performance across countries: A machine learning approach. Eur. J. Oper. Res. 2018, 269, 1072–1085. [Google Scholar] [CrossRef][Green Version

[6] Chui, K. T., Fung, D. C. L., Lytras, M. D., & Lam, T. M. (2018). Predicting at-risk university students in a virtual learning environment via a machine learning algorithm. Computers in Human Behavior, 13(1), 63–75. https://doi.org/10.1177/1469787411429184 [Google Scholar]

[7] Kotsiantis, S. B., Pierrakeas, C. J., & Pintelas, P. E. (2003). Preventing student dropout in distance learning using machine learning techniques. International conference on knowledge-based and intelligent information and engineering systems. pp. 267–274. [Google Scholar]

[8] K. Chai, H. T. Hn and H. L. Cheiu, "Naive-Bayes Classification Algorithm", Bayesian Online Classif. Text Classif. Filter., pp. 97-104, 2002.

[9] V. Hegde and S. G. Kini, Multivariate and Multi-Behavioral Student Dropout Prediction Using Naïve Bayesian Algorithm.

[10] Mendez, G., Buskirk, T. D., Lohr, S., et al. Factors associated with persistence in science and engineering majors: An exploratory study using classification trees and random forests. Journal of Engineering Education - Washington-, 2008, 97(1): 57-70

[11] M. M. Tamada, J. F. de Magalhães Netto and D. P. R. de Lima, "Predicting and Reducing Dropout in Virtual Learning using Machine Learning Techniques: A Systematic Review," 2019 IEEE Frontiers in Education Conference (FIE), Covington, KY, USA, 2019, pp. 1-9, doi: 10.1109/FIE43999.2019.9028545.

[12] Yuda N. Mnyawami, Hellen H. Maziku & Joseph C. Mushi (2023) Enhanced Model for Predicting Student Dropouts in Developing Countries Using Automated Machine Learning Approach: A Case of Tanzanian’s Secondary Schools, Applied Artificial Intelligence, 36:1, DOI: 10.1080/08839514.2022.2071406