MODEL TO PREDICT STUDENT DROPOUT AND ACADEMIC SUCCESS

A Report Submitted in
Partial Fulfillment of the Requirements
for the Degree of

BACHELOR OF ENGINEERING

in
Computer Science Engineering

Submitted by

YASH VERMA	**20CSE69 (2009005371074)**
SAIBA SAIFI	**20CSE47 (2009005371052)**
HARSHIT KUMAR	**20CSE22 (2009005371025)**
ANUSHKA	**20CSE09 (2009005371012)**

Under the supervision of
DR. RAJESH LAVANIA

INSTITUTE OF ENGINEERING AND TECHNOLOGY KHANDARI CAMPUS, AGRA

DR. BHIMRAO AMBEDKAR UNIVERSITY, AGRA

DEPARTMENT OF COMPUTER SCIENCE ENGINEERING
INSTITUTE OF ENGINEERING AND TECHNOLOGY KHANDARI CAMPUS, AGRA
DR. BHIMRAO AMBEDKAR UNIVERSITY, AGRA (282002)

DECLARATION

We the students of B.E. (Computer Science Engineering) of final year **I.E.T. Khandari Campus Agra,** declare that the project work and its presentation were carried out for the fulfillment of Bachelor Degree in Computer Science Engineering for session 2020-2024 under the guidance of **Dr. Rajesh Lavania**

We also declare that the total work for the same is original and nowhere it was used or submitted for the same.

DATE:-……

YASH VERMA	**20CSE69 (2009005371074)**
SAIBA SAIFI	**20CSE47 (2009005371052)**
HARSHIT KUMAR	**20CSE22 (2009005371025)**
ANUSHKA	**20CSE09 (2009005371012)**

CERTIFICATE

This is certify that Yash Verma(20cse69), Saiba Saifi(20cse47), Harshit Kumar(20cse22), Anusha (20cse09) has carried out the project work presented in this report entitled **"MODEL TO PREDICT STUDENTDROPOT AND ACADEMIC SUCCESS"** for the award of **Bachelor of Engineering at Institute of Engineering & Technology, Khandari Campus, Dr. Bhimrao Ambedkar University Agra (U.P.)** under my supervision and guidance. The project work and studies carried out by students themselves and the contents of the report do not form the basis for the award of any other degree to the candidate or to anybody else.

Dr. Rajesh Lavania
(Dept. of Computer Science Engineering)

Department In-charge
(Dept. of Computer Science Engineering)

ABSTRACT

As you know covid-19 is serious pandemic situation which the entire world is facing problem at present as a measure of prevention from deadly virus whose vaccine is not yet available at everywhere, WHO recommends application of alcohol-based sanitizers (60% alcohol content) to parts which are expose to the virus. People using hand sanitizers to wash hands frequently which have been proved effective till date. since sanitizers are effective in preventing covid-19, it would be a good idea to sanitize the whole body.

In this research, development of a short tunnel which sprays sanitizers when people pass through it is designed.

ACKNOWLEDGEMENT

We would like to thank "Department of Computer Science Engineering" Institute of Engineering & Technology, Khandari, Agra for providing us the knowledge, resources and opportunity as necessary component for completion of the project.

This project would not have been possible without the valuable assistance of many people to whom we are indebted. We would sincerely thank our respected project guide Dr. Rajesh Lavania, Department of Computer Science Engineering for motivating us, providing guidance, mighty support, valuable suggestion, timely co-operation and countless help till the successful completion of the project. We express our grateful thanks to……., Incharge Department of Computer Science Engineering for his valuable suggestion and fruitful encouragement throughout the duration of doing this project. Finally thank to all our staff member of Department of Computer Science Engineering who have directly or indirectly support us throughout the project.

We are also indebted to **Prof. Manu Pratap Singh, Director, Institute of Engineering & Technology, Khandari, Agra** for his cooperation, encouragement and providing us all the facilities during the project work.

Thank you all.

Introduction

Higher education institutions worldwide encounter the significant challenge of effectively addressing diverse student learning styles and academic performances to enhance both the students' learning experiences and the formative efficiency of the institution. The capacity to predict and proactively identify potential challenges faced by students is crucial for institutions seeking to develop strategies that support and guide those at risk of academic failure or dropout. Simultaneously, educational institutions amass substantial data annually, encompassing details about students' academic trajectories, demographics, and socio-economic backgrounds. The amalgamation of these factors creates an opportune environment for the application of machine learning approaches to forecast student performance.

Research Problem and Objective

The incidence of student attrition exhibits variability across studies, contingent upon the criteria used for dropout delineation, the data source employed, and the methodologies applied for computation. Frequently, scholarly examinations of dropout phenomena consider temporal distinctions, categorizing departures as either early or late. Owing to disparities in reporting practices, cross-institutional comparisons of dropout rates prove unfeasible. This investigation adopts a micro-level perspective, wherein transitions between academic fields and institutions are construed as instances of dropout, irrespective of the timing thereof. Notably, this approach yields substantially elevated dropout rates in contrast to the macro-level perspective, which exclusively accounts for students departing the higher education system without attainment of a degree.

As per an independent report commissioned by the European Commission, a significant proportion of students discontinue their higher education pursuits prematurely. Even in the most successful nation, Denmark, merely approximately 80% of students successfully conclude their academic endeavors, while in Italy, this figure dwindles to a mere 46%. The report underscores pivotal determinants of student attrition, identifying socioeconomic conditions as the predominant catalyst. This project aims to prognosticate these occurrences with precision.

Literature Review

Beaulac and Rosenthal conducted an analysis on an extensive dataset comprising 38,842 students enrolled at a prominent Canadian university. Their objective was to prognosticate academic success through the utilization of Random Forests (RF). The study employed the initial courses undertaken and corresponding grades of students to forecast program completion, delineating the primary program anticipated for completion in cases of success. The predictive accuracy for program completion yielded an overall rate of 79%, with a precision of 91% for students successfully completing their program and 53% for those who did not. The accuracy for predicting the significance of the outcomes was determined to be 47%.

Hoffait and Schyns, in a separate investigation, utilized a dataset comprising 6,845 students and employed standard classification methodologies, namely RF, Logistic Regression (LR), and Artificial Neural Networks (ANN). Their aim was to identify profiles of freshmen susceptible to encountering notable challenges in successfully completing their inaugural academic year. The achieved accuracy for the majority class was approximately 70%, while for the minority class, it was less than 60%, irrespective of the algorithm employed. Subsequently, RF was utilized to formulate a strategy aimed at enhancing prediction accuracy for specific classes of substantial interest. However, the developed approach did not consistently result in an augmented identification of students at risk.

A study involving 2,459 students from a European Engineering School was undertaken. Their focus was on predicting overall academic performance based on information available at the conclusion of the first year, encompassing demographic details, social factors, and academic metrics, including assessments from initial-year courses. Various predictive models, including Support Vector Machines, Naïve Bayes, Decision Trees (DT), RF, Bagging Decision Trees, and Adaptive Boosting Decision Trees, were employed. Notably, Random Forests and Adaptive Boosting Decision Trees yielded superior results, achieving an overall accuracy of 96%.

Dataset used

The dataset comprises approximately 50 entries, each delineating an individual student and encompassing 22 attributes. Its intended utilization involves benchmarking diverse algorithms to address analogous problem types, serving as a platform for assessing algorithmic performance. Furthermore, it serves as a valuable resource for machine learning training endeavours.

The dataset is categorized into two categories -> Categorical and Numerical

Numerical ->

* Curricular units 1st year(evaluations): The quantity of academic modules assessed by the student during the initial semester.
* Curricular units 1st year (approved): The quantity of academic modules successfully completed by the student during the initial semester.
* Curricular units 1st year(credited): The quantity of academic modules for which the student received credit during the initial semester.
* Curricular units 1st yesr (enrolled): The quantity of academic modules in which the student is registered during the inaugural semester.
* Age at enrollment: The age of the student at the point of enrollment.

Categorical ->

* Application mode: The approach employed by the student for the application process.
* Marital status: The current matrimonial status of the student.
* Gender: The gender of the student is requested.
* Daytime/evening attendance: Whether the student participates in daytime or evening sessions.
* Previous qualification: The educational credentials attained by the student prior to matriculating into tertiary education.
* Course: The course taken by the student.

* Scholarship holder: Determine if the student is a recipient of a scholarship..
* Debtor: Determine if the student has financial obligations.
* Tuition fees up to date: Please confirm the current status of the student's payment of tuition fees.
* Mother's qualification: The academic credentials of the student's mother.
* Father's qualification: The educational background of the student's father.
* Mother's occupation: The professional vocation of the student's mother.
* Father's occupation: The professional vocation of the student's father.
* Displaced: Determine if the student is an individual who has been displaced.
* Nationality: The student's country of origin.
* International: Determine whether the student is of international origin
* Educational special needs: Does the student possess any unique educational requirements?

Methodology

Steps:

1. Initial data preprocessing involved the removal of rows with missing values to uphold the integrity and completeness of the dataset.
2. Recognizing the predominant representation of students of Portuguese descent in the dataset and the limited variability in nationality, the irrelevant nationality feature was excluded from consideration for predicting student dropout.
3. The international feature, deemed to contribute insufficient predictive power, was also omitted to enhance the efficiency of the analysis.
4. Feature categorization was conducted to organize the data into distinct groups: Demographic data, Socioeconomic data, Macro-economic Enrollment Data, and Academic data. Subsequently, the correlation of these groups with the target variable was assessed.

5. Redundancy and potential overfitting concerns were addressed by identifying and removing features exhibiting similar correlations within the dataset.

6. The 'Enrolled' category, lacking relevance in predicting student outcomes, was removed during the feature selection process.

7. Label encoding was applied to represent the target variable, assigning the value '1' to 'graduated' and '0' to 'dropout.'

8. Features and target variables were segregated into X and Y, respectively, in preparation for model training.

9. The MinMaxScaler was applied to normalize the X data, mitigating potential biases arising from varying scales within the features.

10. The train-test split method was employed to partition the dataset, ensuring independent datasets for training and evaluation.

11. The construction and evaluation of five predictive models—Logistic Regression, Decision Tree Classifier, K-Nearest Neighbour Classifier, Random Forest Classifier, and eXtremeBoost Algorithm—were implemented, with a focus on calculating model accuracy.

Machine Learning

Machine learning (ML) is a subdomain of artificial intelligence (AI) that focuses on developing systems that learn—or improve performance—based on the data they ingest. Artificial intelligence is a broad word that refers to systems or machines that resemble human intelligence. Machine learning and AI are frequently discussed together, and the terms are occasionally used interchangeably, although they do not signify the same thing. A crucial distinction is that, while all machine learning is AI, not all AI is machine learning.

Machine Learning is the field of study that gives computers the capability to learn without being explicitly programmed. ML is one of the most exciting technologies that one would have ever come across. As it is evident from the name, it gives the computer that makes it more similar to humans: The ability to learn. Machine learning is actively being used today, perhaps in many more places than one would expect.

- Machine learning is data driven technology. Large amount of data generated by organizations on daily bases. So, by notable relationships in data, organizations makes better decisions.
- Machine can learn itself from past data and automatically improve.
- From the given dataset it detects various patterns on data.
- For the big organizations branding is important and it will become more easy to target relatable customer base.
- It is similar to data mining because it is also deals with the huge amount of data.

Logistic Regression in Machine Learning

Logistic regression is a **supervised machine learning algorithm** used for **classification tasks** where the goal is to predict the probability that an instance belongs to a given class or not. Logistic regression is a statistical algorithm which analyze the relationship between two data factors. The article explores the fundamentals of logistic regression, it's types and implementations.

Logistic regression is used for binary classification where we use sigmoid function, that takes input as independent variables and produces a probability value between 0 and 1.
For example, we have two classes Class 0 and Class 1 if the value of the logistic function for an input is greater than 0.5 (threshold value) then it belongs to Class 1 it belongs to Class 0. It's referred to as regression because it is the extension of linear regression but is mainly used for classification problems.

- Logistic regression predicts the output of a categorical dependent variable. Therefore, the outcome must be a categorical or discrete value.
- It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.
- In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1).

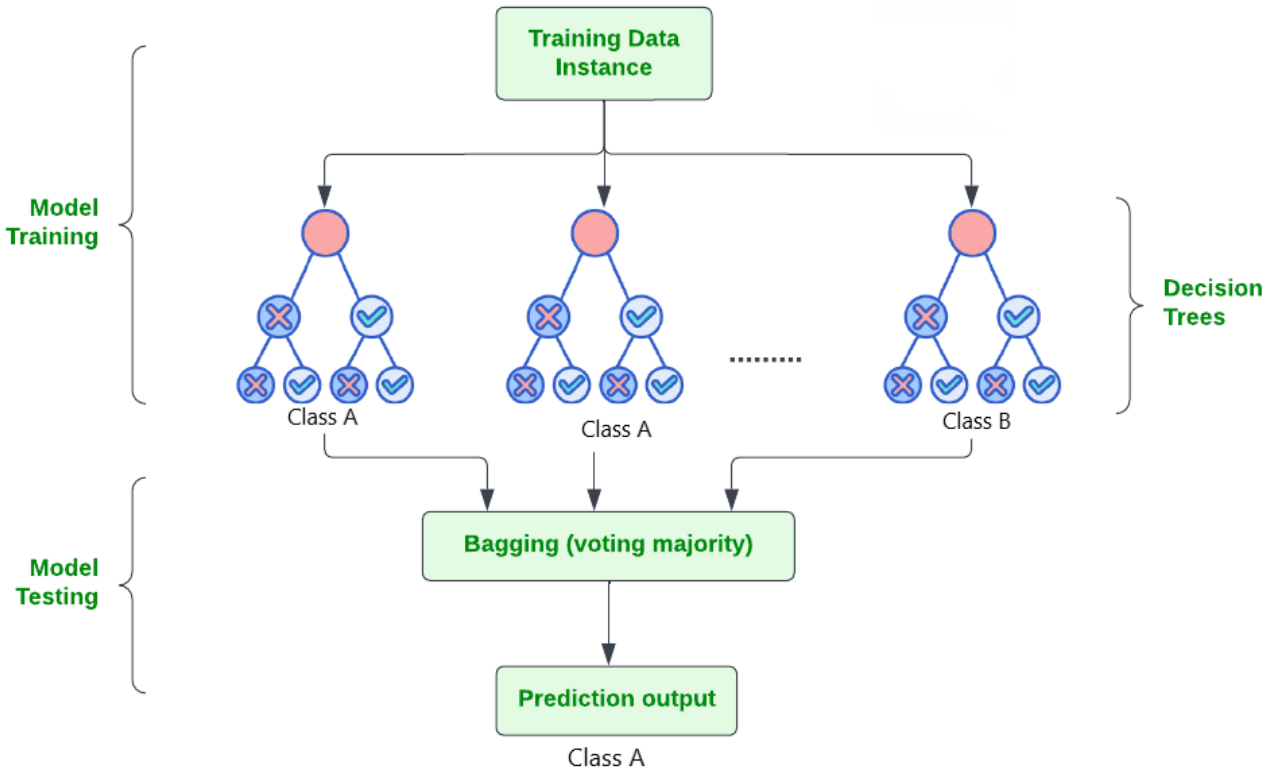
Decision Tree

A decision tree is one of the most powerful tools of supervised learning algorithms used for both classification and regression tasks. It builds a flowchart-like tree structure where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label. It is constructed by recursively splitting the training data into subsets based on the values of the attributes until a stopping criterion is met, such as the maximum depth of the tree or the minimum number of samples required to split a node.
During training, the Decision Tree algorithm selects the best attribute to split the data based on a metric such as entropy or Gini impurity, which measures the level of impurity or randomness in the subsets. The goal is to find the attribute that maximizes the information gain or the reduction in impurity after the split.

- **Root Node:** It is the topmost node in the tree, which represents the complete dataset. It is the starting point of the decision-making process.
- Decision/Internal Node: A node that symbolizes a choice regarding an input feature. Branching off of internal nodes connects them to leaf nodes or other internal nodes.
- **Leaf/Terminal Node:** A node without any child nodes that indicates a class label or a numerical value.
- **Splitting:** The process of splitting a node into two or more sub-nodes using a split criterion and a selected feature.
- **Branch/Sub-Tree:** A subsection of the decision tree starts at an internal node and ends at the leaf nodes.
- **Parent Node:** The node that divides into one or more child nodes.
- **Child Node:** The nodes that emerge when a parent node is split.
- **Impurity**: A measurement of the target variable's homogeneity in a subset of data. It refers to the degree of randomness or uncertainty in a set of examples. The **Gini index** and **entropy** are two commonly used impurity measurements in decision trees for classifications task

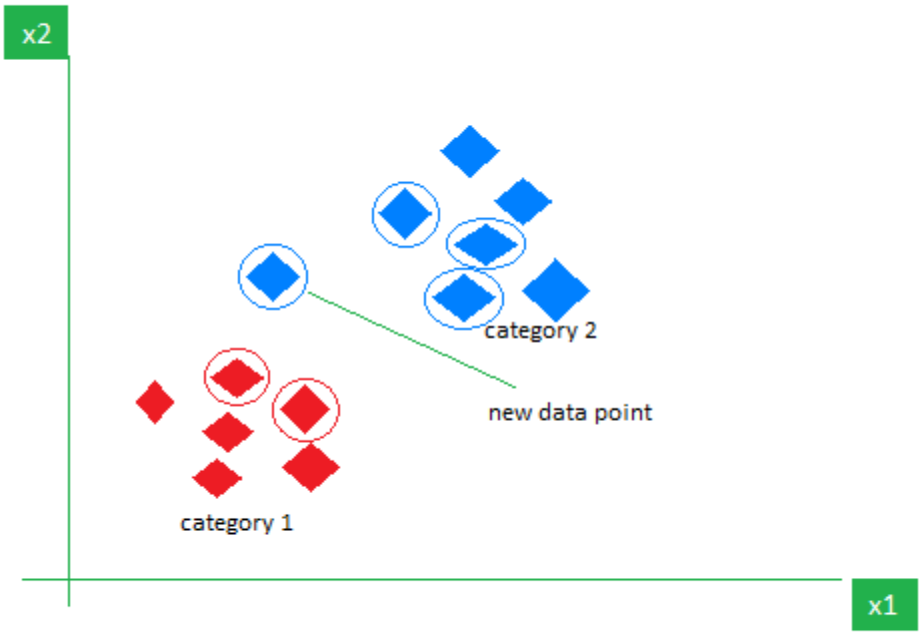
Random Forest

Random Forest algorithm is a powerful tree learning technique in Machine Learning. It works by creating a number of Decision Trees during the training phase. Each tree is constructed using a random subset of the data set to measure a random subset of features in each partition. This randomness introduces variability among individual trees, reducing the risk of overfitting and improving overall prediction performance. In prediction, the algorithm aggregates the results of all trees, either by voting (for classification tasks) or by averaging (for regression tasks) This collaborative decision-making process, supported by multiple trees with their insights, provides an example stable and precise results. Random forests are widely used for classification and regression functions, which are known for their ability to handle complex data, reduce overfitting, and provide reliable forecasts in different environments.



K-Nearest Neighbor(KNN)

The **K-Nearest Neighbors (KNN) algorithm** is a supervised machine learning method employed to tackle classification and regression problems. Evelyn Fix and Joseph Hodges developed this algorithm in 1951, which was subsequently expanded by Thomas Cover. The article explores the fundamentals, workings, and implementation of the KNN algorithm.
KNN is one of the most basic yet essential classification algorithms in machine learning. It belongs to the supervised learning domain and finds intense application in pattern recognition, data mining, and intrusion detection.
It is widely disposable in real-life scenarios since it is non-parametric, meaning it does not make any underlying assumptions about the distribution of data (as opposed to other algorithms such as GMM, which assume a Gaussian distribution of the given data). We are given some prior data (also called training data), which classifies coordinates into groups identified by an attribute.



XGBoost

XGBoost is an optimized distributed gradient boosting library designed for efficient and scalable training of machine learning models. It is an ensemble learning method that combines the predictions of multiple weak models to produce a stronger prediction. XGBoost stands for “Extreme Gradient Boosting” and it has become one of the most popular and widely used machine learning algorithms due to its ability to handle large datasets and its ability to achieve state-of-the-art performance in many machine learning tasks such as classification and regression.
One of the key features of XGBoost is its efficient handling of missing values, which allows it to handle real-world data with missing values without requiring significant pre-processing. Additionally, XGBoost has built-in support for parallel processing, making it possible to train models on large datasets in a reasonable amount of time.

XGBoost can be used in a variety of applications, including Kaggle competitions, recommendation systems, and click-through rate prediction, among others. It is also highly customizable and allows for fine-tuning of various model parameters to optimize performance.

XgBoost stands for Extreme Gradient Boosting, which was proposed by the researchers at the University of Washington. It is a library written in C++ which optimizes the training for Gradient Boosting.

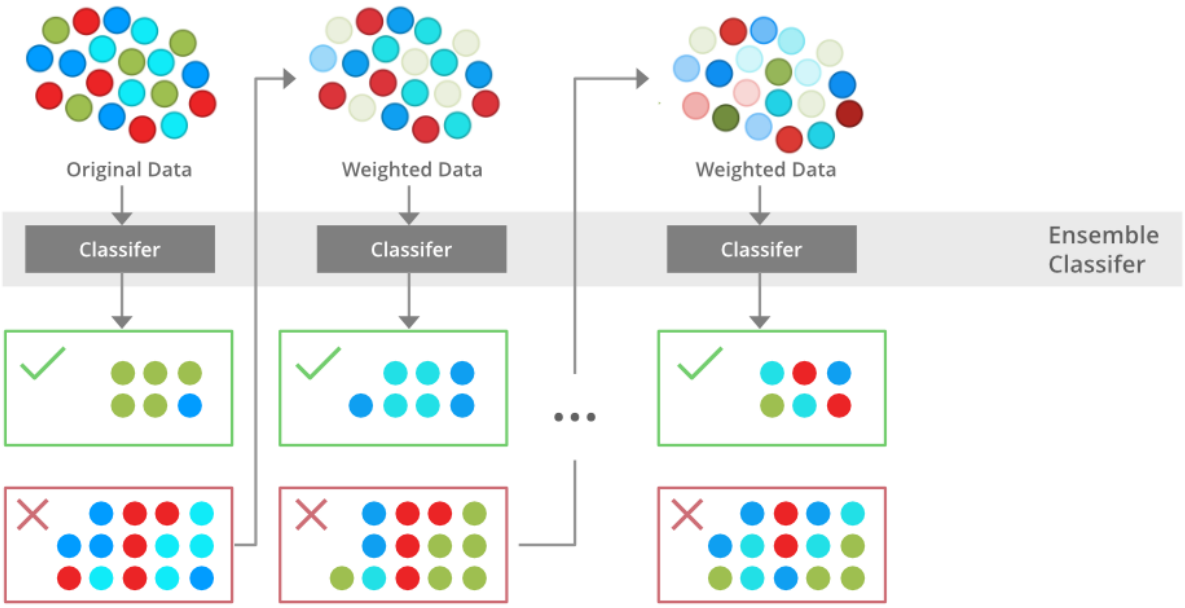
Bagging:

A Bagging classifier is an ensemble meta-estimator that fits base classifiers each on random subsets of the original dataset and then aggregate their individual predictions (either by voting or by averaging) to form a final prediction. Such a meta-estimator can typically be used as a way to reduce the variance of a black-box estimator (e.g., a decision tree), by introducing randomization into its construction procedure and then making an ensemble out of it.
Each base classifier is trained in parallel with a training set which is generated by randomly drawing, with replacement, N examples(or data) from the original training dataset, where N is the size of the original training set. The training set for each of the base classifiers is independent of each other. Many of the original data may be repeated in the resulting training set while others may be left out.

Bagging reduces overfitting (variance) by averaging or voting, however, this leads to an increase in bias, which is compensated by the reduction in variance though.

Boosting:

Boosting is an ensemble modelling, technique that attempts to build a strong classifier from the number of weak classifiers. It is done by building a model by using weak models in series. Firstly, a model is built from the training data. Then the second model is built which tries to correct the errors present in the first model. This procedure is continued and models are added until either the complete training data set is predicted correctly or the maximum number of models are added.



Advantages of XGBoost:

1. Performance: XGBoost has a strong track record of producing high-quality results in various machine learning tasks, especially in Kaggle competitions, where it has been a popular choice for winning solutions.
2. Scalability: XGBoost is designed for efficient and scalable training of machine learning models, making it suitable for large datasets.
3. Customizability: XGBoost has a wide range of hyperparameters that can be adjusted to optimize performance, making it highly customizable.
4. Handling of Missing Values: XGBoost has built-in support for handling missing values, making it easy to work with real-world data that often has missing values.

Disadvantages of XGBoost:

1. Computational Complexity: XGBoost can be computationally intensive, especially when training large models, making it less suitable for resource-constrained systems.
2. Overfitting: XGBoost can be prone to overfitting, especially when trained on small datasets or when too many trees are used in the model.
3. Hyperparameter Tuning: XGBoost has many hyperparameters that can be adjusted, making it important to properly tune the parameters to optimize performance. However, finding the optimal set of parameters can be time-consuming and requires expertise.
4. Memory Requirements: XGBoost can be memory-intensive, especially when working with large datasets, making it less suitable for systems with limited memory resources.

Results and Performance Metrics

The model's predictions, along with its accuracy, precision, and recall are analyzed.

	Model	Accuracy	Precision	Recall
0	Logistic Regression	0.898072	0.885010	0.959911
1	Decision Tree	0.852617	0.885135	0.875278
2	Random Forest	0.892562	0.888889	0.944321
3	KNN	0.818182	0.816367	0.910913
4	xgboost	0.902204	0.902128	0.944321

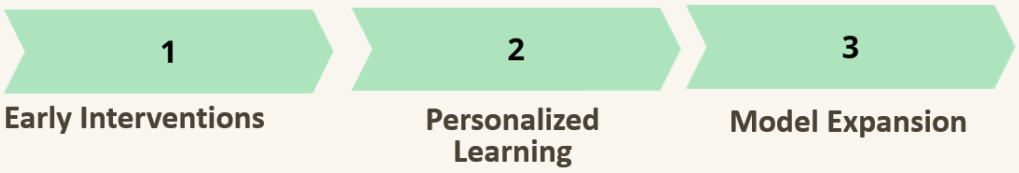
$$\text{Precision} = \frac{TP}{TP + FP} \qquad \text{Recall} = \frac{TP}{TP + FN}$$

Building the Machine Learning Model

1. **Logistic Regression**
2. **Decision Tree Classifier**
3. **Random Forest**
4. **K-NN**
5. **XG-Boost**

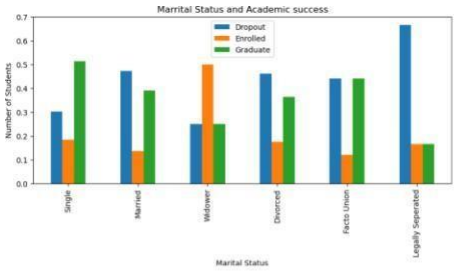
Potential Applications and Future Work

The model's insights enable the implementation of interventions, personalized learning, and targeted support to reduce dropout rates and improve student success. Future work involves expanding the model's scope and enhancing predictive accuracy.



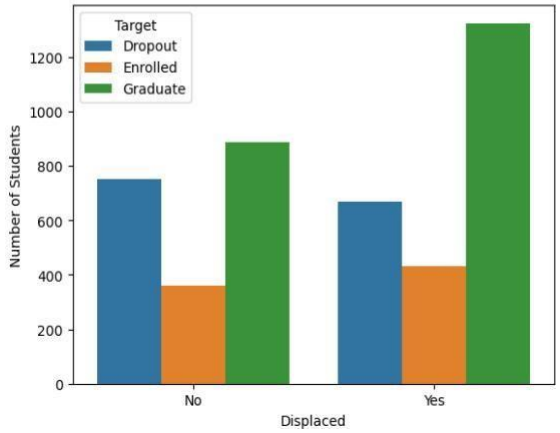
Result

Data-oriented result-



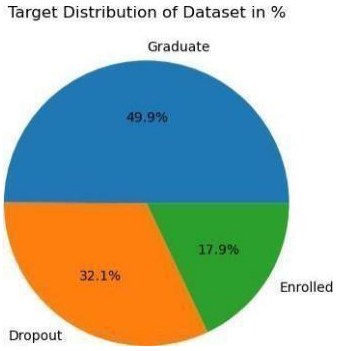
* Individuals who have undergone legal separation face an elevated risk of discontinuing their academic pursuits.

* Unmarried students exhibit a greater likelihood of successfully completing their education owing to their enhanced concentration and commitment.

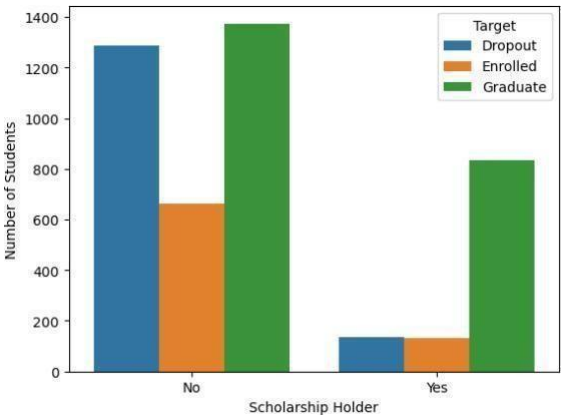


Individuals who have successfully completed their academic programs and obtained their

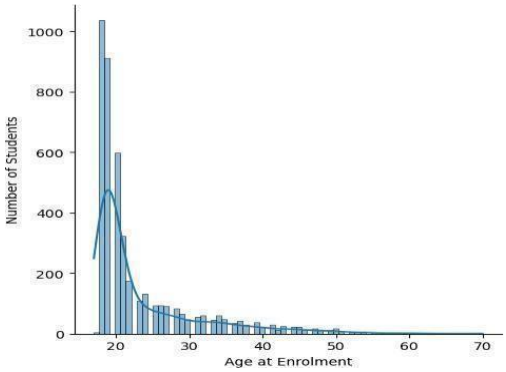
degrees are predominantly characterized as alumni or graduates.



About half of the student population within the dataset has successfully completed their academic programs.



Based on the depicted graph, it can be inferred that individuals who are recipients of scholarships exhibit a greater likelihood of successfully completing their academic programs as opposed to discontinuing their studies.



The depicted distribution plot indicates a positive skewness, suggesting that the enrollment age of students spans from 17 to 70 years, with the majority centered around the

mean age of 23 years.

Final Result:

Model	Accuracy	Precision	Recall
Logistic Regression	0.898072	0.885010	0.959911
Decision Tree	0.852617	0.885135	0.875278
Random Forest	0.892562	0.888889	0.944321
KNN	0.818182	0.816367	0.910913
Xgboost	0.902204	0.902128	0.944321

Precision and recall are two important metrics used to evaluate the performance of classification algorithms, particularly in the context of binary classification problems (where there are two classes: positive and negative). These metrics are often used in the field of machine learning and information retrieval.

1. Precision:

- Precision, also known as positive predictive value, measures the accuracy of the positive predictions made by a model. It answers the question: "Of all the instances predicted as positive, how many are actually positive?"

- Precision is calculated using the following formula:

$$\text{Precision} = \frac{}{}$$

where True Positives (TP) are the instances correctly predicted as positive, and False Positives (FP) are the instances incorrectly predicted as positive.

2. Recall:

- Recall, also known as sensitivity or true positive rate, measures the ability of a model to capture all the positive instances. It answers the question: "Of all the actual positive instances, how many were correctly predicted?"

- Recall is calculated using the following formula:

Recall = [redacted]

where True Negatives (TN) are the instances correctly predicted as negative, and False Negatives (FN) are the instances incorrectly predicted as negative.

In our project, emphasis is placed on prioritizing recall, whereby the significance lies in identifying students categorized as dropouts despite being erroneously marked as graduates. This approach considers students who have graduated but not marked as dropouts as comparatively less critical in the context of our objectives.

Managerial Implications

1. Early Identification and Intervention: Educational institutions can leverage dropout rate predictions to promptly identify students at risk early in the academic year, allowing for the implementation of targeted support and resources to mitigate the likelihood of dropout.
2. Strategic Resource Allocation: Schools and universities can enhance resource allocation efficiency by directing interventions and resources to areas with the highest predicted dropout rates, ensuring a more effective utilization of available resources.
3. Tailored Retention Strategies: Institutions have the opportunity to formulate personalized retention strategies aimed at enhancing student engagement and success, thereby diminishing dropout rates and fostering improved graduation outcomes.
4. Informed Policy Development: Governmental and educational authorities can utilize dropout rate predictions as a foundational element in shaping informed policy decisions related to education funding and program enhancements.
5. Counseling and Guidance Integration: School counselors can employ dropout rate predictions as a proactive tool to identify students in need of guidance and support, addressing both academic and personal challenges in a timely manner.
6. Facilitating Research Endeavors: Researchers can employ dropout rate predictions to investigate the multifaceted factors contributing to student attrition, facilitating the design of studies geared toward enhancing overall educational outcomes.

Limitations of the work and future improvements/extensions:

1. Geographic Extent: The present dataset is confined to a singular county, thereby limiting the broader applicability of predictive models. Future initiatives will entail the systematic aggregation of data from multiple counties, fostering a more diverse and representative dataset conducive to comprehensive analyses of student dropout and success rates.

2. Temporal Limitations: The current dataset exclusively captures academic outcomes up to the second semester, affording a constrained perspective on student performance. Prospective endeavors involve an extension of data collection efforts to encompass subsequent semesters, facilitating a longitudinal scrutiny of academic trajectories and engendering more precise prognostications over an extended timeframe.

4. Prolonged Machine Learning Training: Ongoing endeavors will focus on protracted machine learning model training, harnessing new data points and refining algorithms. This iterative methodology is designed to sustainably update and adapt the models, mirroring the dynamic landscape of student academic trajectories and ensuring the enduring relevance of the predictive tool.

3. Variable Inclusivity: Although the extant dataset comprises demographic, socioeconomic, macroeconomic, and academic parameters, future refinements will introduce supplementary variables. This augmentation aims to incorporate school and high school performance metrics, thereby providing a more holistic comprehension of the interconnected determinants influencing dropout rates and academic achievements.

Conclusion

In summation, this thesis project has successfully constructed predictive models for the anticipation of student attrition and academic achievement, utilizing an exhaustive

dataset obtained from an educational institution. The managerial implications underscore the potential for timely intervention, optimal resource allocation, personalized retention strategies, and well-informed policy formulation. Despite the notable accomplishments of the project, it is imperative to acknowledge certain limitations, including constraints pertaining to geography and temporality. Prospective enhancements are envisioned to amplify the representativeness of the dataset, extend temporal inclusivity, incorporate supplementary variables, and ensure a continual process of model refinement. These refinements are anticipated to significantly enhance predictive accuracy and broaden the applicability of the models, thereby addressing the intricate challenges associated with student attrition and academic success within higher education.

Bibliography

1. Realinho, V.; Machado, J.; Baptista, L.; Martins, M.V. Predicting Student Dropout and Academic Success. Data 2022, 7, 146. https://doi.org/10.3390/ data7110146

2. Aulck, L., Velagapudi, N., Blumenstock, J., & West, J. (2016 June). Predicting Student Dropout in Higher Education. ArXiv e-prints.

3. Github- https://github.com/SaibaSaifi/Student-Dropout-and-Academic-Success/blob/main/notebook%20(3).ipynb

4. Chen, R., & DesJardins, S. L. (2008). Exploring the effects of financial aid on the gap in student dropout risks by income Level. Research in Higher Education, 49(1), 1-18. https://doi.org/10.1007/s11162-007-9060-9.

5. Miguéis, V.L., Freitas, A., Garcia, P.J.V., Silva, A.: Early segmentation of students according to their academic performance: a predictive modelling approach. Decis. Support Syst. 115, 36–51 (2018).

6. Hoffait, A.S., Schyns, M.: Early detection of university Students with potential difficulties. Decis. Support Syst. 101, 1–11 (2017)

7. Beaulac, C., Rosenthal, J.S.: Predicting university Students' academic success and major using random forests. Res. High. Educ. 60, 1048–1064 (2019).