A MACHINE LEARNING APPROACH TO PREDICTING STUDENT DROPOUT AND SUCCESS

EGBINE OGHENERO HENRIETTA

23015772

Department of Data Science

University of the West of England

Bristol, United kingdom.

Oghenero2.Egbine@live.uwe.ac.uk

**Introduction**

Higher education academic achievement is essential for employment, social justice, and economic expansion. The most significant challenge that higher education institutions need to solve to increase student success is dropout rates. The term "dropout" has no agreed-upon definition.

Higher education establishments keep a great deal of information about their students on file, which offers a large potential for information generation, knowledge creation, and monitoring. The lives of students and their families, higher education institutions, and society at large are all significantly impacted by school dropout and educational failure, which are barriers to economic growth, employment, competitiveness, and productivity in higher education. The ability to anticipate student outcomes has greatly improved in academic institutions as a result of the shift from traditional teaching approaches to data-driven decision-making. The purpose of this study is to improve intervention tactics and educational outcomes by identifying factors that contribute to academic achievement and student dropout using machine learning techniques.

# problem definition

The issue at hand pertains to guided learning. Since the goal is already known, it is supervised. The objective of this analysis is to address the problem of identifying potential dropouts, enrolled students, and graduates in an educational setting by developing a predictive model with machine learning techniques. Utilizing a variety of machine learning algorithms, the study entails examining an actual dataset in order to forecast student outcomes according to socioeconomic, academic, and demographic characteristics. In order to improve student retention rates and overall educational outcomes, educational institutions will be able to intervene early and provide targeted help to at-risk students by using a powerful predictive model that can reliably classify individuals into distinct groups.

## .

# DATASET OVERVIEW

Data from the first and second semesters, data at the time of student enrollment, demographic, socioeconomic, and macroeconomic statistics are all included in the dataset.

**Source and Composition:**

Source: The UCI Machine Learning Repository, which is renowned for its dependability and wide selection of instructional datasets, provided the data for this study.   
variables: The dataset contains a wide range of variables, including socioeconomic background data, academic performance indicators, student demographics, and engagement levels in academic activities.  
Volume: Several thousand student records make up the collection, each of which has their final academic status (Dropout, Enrolled, Graduate) noted. It contains 37 features and 4424 observations.

# METHODOLOGY

## Preprocessing of the Data

#### The following are a few of the data processing activities:

#### Managing Negative Values: The median was used to impute negative values because it is resistant to outliers.

#### Feature Encoding: To convert categorical data into a machine-readable format, label encoding was used to encode them. One example of this was the target column, which held enrolled graduates and dropouts.

#### Scaling of Features: To make sure that no variable predominates over others because of scale discrepancies, numerical features were normalized. This was achieved with standard scaler.

iv feature selection: I performed feature selection to determine the columns or features that best affect the target column . A list of them are below.

A graph with a bar chart

Description automatically generated with medium confidence

The above plot tries to give a pictorial representation

Of the most important features in the dataset.

## Exploratory Data Analysis (EDA) Distribution Analysis

The goal of the analysis was to ascertain whether the different classes are balanced or imbalanced by comprehending the target variable's distribution.

from the bar plot below we can see that we have an imbalanced dataset.A graph of a distribution of dropout

Description automatically generated

1. Features Analysis: A variety of characteristics were examined in relation to the graduation and dropout rates of students.

A graph of a dropout and graduate by gender

Description automatically generated

A graph of a dropout and graduate status

Description automatically generated

The dropout rate for the plots above appears to be somewhat similar across different genders, although a slight difference in proportions can be noted. But in general females graduate more. There is a noticeable difference in dropout rates between scholarship holders and non-scholarship holders, suggesting scholarships may have an impact on dropout rates. Meaning students with scholarship tend not to drop out.

1. Resampling: The class imbalance seen during EDA was another data processing task. To improve the model's capacity to generalize across less represented classes, I balanced the dataset using the sklearn modules so i imported resample method.

## Model Selection and Training

Two factors were considered in this analysis:

1. Feature Type: The kind of features in the dataset (such as text, numeric, or categorical) may influence the model selection. For instance, linear regression works well for regression problems, and naive bayes is especially useful for text categorization.
2. Interpretability: Knowing how the model generates decisions is important in a lot of applications. While neural networks and random forests are often regarded as "black boxes," models like logistic regression and random forests offer an easy way to analyse data.

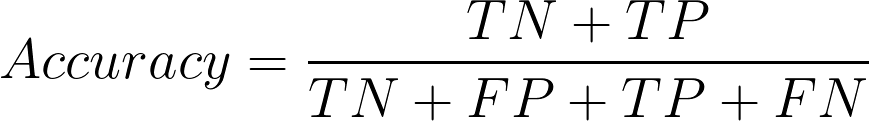
In this investigation, five machine learning algorithms were evaluated to find which one has a superior predictive capability:

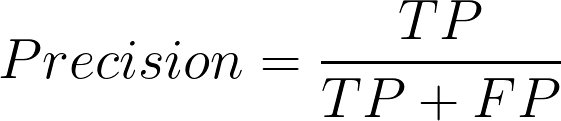
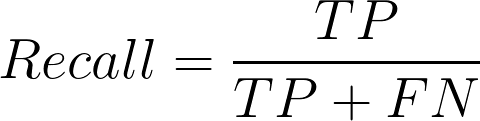
1. **K-NEAREST NEIGHBORS**: For classification and regression applications, K-Nearest Neighbors (KNN) is a straightforward and efficient supervised machine learning technique. To allocate the new data point to the class that appears the most frequently among its neighbors, KNN counts the number of data points in each class among the K nearest neighbors for classification tasks. We call this majority voting. KNN allocates the new data point to the class that appears most frequently among its neighbors after counting the number of data points in each class among the K nearest neighbors in classification tasks. It's referred to as majority voting here.
2. **GRADIENT BOOSTING METHOD:** Building on the concept of boosting, gradient boosting is a potent and adaptable machine learning technique that combines several weak learners to create a strong learner. When using gradient boosting in the context of decision trees, every new tree gradually outperforms the preceding ones by fixing the mistakes caused by the older trees. Gradient boosting is an adaptable technique that works with any differentiable loss function that is optimized while the system is learning. This loss function (e.g., squared error for regression, logarithmic loss for classification) is unique to the kind of problem being handled.
3. **SUPPORT VECTOR MACHINE:** Finding the hyperplane a line in two dimensions that best divides the classes in the feature space is the basic notion behind support vector machines (SVM). The goal of SVM is to maximize the difference between the class's nearest points and the hyperplane. The closest points are referred to as support vectors, and they are essential for figuring out where the hyperplane is.
4. **LOGISTIC REGRESSION:** The likelihood that a given input point belongs to a specific class is predicted via logistic regression. A logistic function, also known as the sigmoid function, is used to describe probability estimation and produces a number between 0 and 1. The likelihood that the dependent variable (target), given the predictors, is a success (often coded as 1) can be understood from this output. The output of the model can be subjected to a threshold to provide a binary forecast (0 or 1). Usually, a 0.5 threshold is applied. The outcome is anticipated as the positive class (1) if the predicted probability is greater than or equal to 0.5, and as the negative class (0) if it is less than 0.5. Maximum likelihood estimation is commonly used to estimate the logistic regression model's parameters, or coefficients (MLE).
5. **RANDOM FOREST CLASSIFIER:** A reliable and adaptable machine learning approach for classification (and regression) applications is the Random Forest classifier. To increase accuracy and reduce overfitting, it expands on the principles of decision trees by integrating several decision trees into a single model. The foundation of Random Forest is the idea of ensemble learning, which involves combining several weak learners in this example, decision trees to produce a strong learner.

## **Evaluation Methodology**

Although accuracy is one of the most logical performance measures for assessing classification models, it has a number of significant drawbacks, especially when datasets are unbalanced or when various errors have different prices attached to them. The f1 score, precision, recall, and confusion matrix were used in this investigation. The confusion matrix also gave a comprehensive picture of the model's performance across classes.

1. Accuracy: The percentage of accurately identified instances both true positives and true negatives out of all instances is known as accuracy. It offers a broad indicator of how accurate and successful the model is in making predictions. However, in cases when class distributions are unbalanced, accuracy might not be enough to evaluate model performance.



1. Precision: precision also known as sensitivity or true positive rate, measures the proportion of true positive predictions out of all actual positive instances, precision indicates the percentage of accurate positive predictions. The quality of positive predictions, that is, the proportion of projected positive cases that turn out to be true positives is the main emphasis of precision. In situations where the cost of false positives is significant, it is especially helpful.
2. Recall: The model's recall highlights its capacity to record every positive experience without omitting any. In applications where finding every positive example is essential, it is significant even if it results in some false positives.
3. F1 score: Precision and recall are harmonic means, and the F1 score strikes a balance between both. The F1 score is a valuable metric in situations where there is an unequal class distribution or if the costs of false positives and false negatives fluctuate. It takes into account both precision and recall. It offers a solitary figure that encapsulates the balance between recall and precision.

A black background with a black square

Description automatically generated with medium confidence

# MODEL COMPARISON

## KNN

The KNN model performed with about 72% accuracy on the test set. The precision, recall, and F1-scores for each class varied, indicating moderate performance across different classes.

## GRADIENT BOOSTING MODEL

On the test set, the accuracy of the GBM model was approximately 78%. Particularly for the enrolled and dropout groups, all the classes had relatively high precision, recall, and F1-scores.

A screenshot of a graph

Description automatically generated

## SUPPORT VECTOR MACHINE

The SVM model produced test set accuracy of roughly 70%. Precision, Recall, and F1-score: These metrics demonstrated modest performance across many classes and were similar to the KNN model.

## LOGISTIC REGRESSION

On the test set, the Logistic Regression model's accuracy was almost 70%. The precision, recall, and F1-scores were comparable to the SVM model, suggesting a moderate level of competence in various classes.

## RANDOM FOREST CLASSIFIER

On the test set, the Random Forest Classifier had a good accuracy with f1 score at 94%. For every class, the precision, recall, and F1-scores were high, suggesting better performance than the other models. Across all courses, it demonstrated the high accuracy along with consistently strong recall, precision, and F1-scores. Additionally, the Gradient Boosting Method functioned effectively, exhibiting excellent accuracy and consistent performance across several classes, but random forest outperformed all the other models.

A blue and white grid with numbers

Description automatically generated

MODELS SUMMARY

A screenshot of a graph

Description automatically generated

REJECTED ALGORITHMS AND REASONS.

1) Linear regression: This model is intended to solve regression issues rather than classification issues. So I rejected it because this is a classification problem Furthermore A linear relationship between the independent and dependent variables is the underlying assumption of linear regression. Poor prediction performance may result from linear regression's inability to correctly identify the underlying patterns in a nonlinear relationship.   
2) Decision trees: When the tree is deep and complicated, there is a chance that it will overfit the training set. Poor generalization performance on unknown data results from overfitting, which happens when the tree identifies noise in the training set instead of the underlying patterns. Small changes in the training data have a significant impact on decision trees' sensitivity.

**3)** K-NEAREST KNEIGHBOR: The Random Forest Classifier outperforms alternative learning algorithms like the K-Nearest Neighbors (KNN) algorithm, despite achieving a respectable level of accuracy in predicting the likelihood of student dropout. KNN is less appropriate for this predictive modelling task due to its interpretability issues and dependence on distance measurements, especially when dealing with high-dimensional datasets and intricate feature interactions.

**Future Research Directions:**

* **Deep Learning Models:** Discuss the potential of using deep learning models like recurrent neural networks (RNNs) to capture temporal patterns in student data (e.g., changes in attendance or performance over time).
* **Explainable AI (XAI):** Explore techniques for making the predictions of the machine learning model more interpretable. This can build trust in the model and provide educators with deeper insights into student behavior.
* **Longitudinal Studies:** Discuss the benefits of conducting longitudinal studies to track the long-term impact of interventions based on dropout risk predictions.

**CONCLUSION**

To sum up, predictive modelling presents a useful instrument for mitigating the risk of student dropout in academic settings. Institutions can enhance student retention rates and academic success by proactively identifying at-risk students and implementing tailored interventions through the utilization of machine learning algorithms and relevant dataset analysis. Predictive modelling adoption has enormous potential to improve student learning outcomes and provide a welcoming learning environment for all students.

REFERENCES

* 1. Suthaharan, S. and Suthaharan, S., 2016. Support vector machine. *Machine learning models and algorithms for big data classification: thinking with examples for effective learning*, pp.207-235.
  2. Kramer, O. and Kramer, O., 2013. K-nearest neighbors. *Dimensionality reduction with unsupervised nearest neighbors*, pp.13-23.
  3. Speiser, J.L., Miller, M.E., Tooze, J. and Ip, E., 2019. A comparison of random forest variable selection methods for classification prediction modeling. *Expert systems with applications*, *134*, pp.93-101.
  4. Streiner, D.L. and Norman, G.R., 2006. “Precision” and “accuracy”: two terms that are neither. *Journal of clinical epidemiology*, *59*(4), pp.327-330.
  5. Davis, J. and Goadrich, M., 2006, June. The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd international conference on Machine learning* (pp. 233-240).
  6. Lipton, Z.C., Elkan, C. and Narayanaswamy, B., 2014. Thresholding classifiers to maximize F1 score. *arXiv preprint arXiv:1402.1892*.
  7. Parmar, A., Katariya, R. and Patel, V., 2019. A review on random forest: An ensemble classifier. In *International conference on intelligent data communication technologies and internet of things*
  8. LaValley, M.P., 2008. Logistic regression. *Circulation*, *117*(18), pp.2395-2399.

**9)**Menze, B.H., Kelm, B.M., Masuch, R., Himmelreich, U., Bachert, P., Petrich, W. and Hamprecht, F.A., 2009. A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC bioinformatics*, *10*, pp.1-16.

10) Ding, J., Tarokh, V. and Yang, Y., 2018. Model selection techniques: An overview. *IEEE Signal Processing Magazine*, *35*(6), pp.16-34.

Identify applicable sponsor/s here. If no sponsors, delete this text box (*sponsors).*