**—DO NOT ASK FOR EDIT ACCESS TO THIS FILE—READ INSTRUCTIONS BELOW—**

**2. Click “File – Make a Copy” to COPY the template to YOUR Google Drive (Sign in if needed)**

**3. Click “Share”. Change “Restricted” to “Anyone with the link” and “Viewer” to “Editor”.**

**4. Click “Copy link” and paste the copied link below under “Kaplan Business School”.**

**5. Do all your writing directly in your Google Doc under your Gmail/Google account.**

**6. Delete the highlighted instructions. Download your report as a PDF and submit via Turnitin.**

Predictive Modeling for Student Graduation Outcomes: A Machine Learning Approach

Student Name

Kaplan Business School

**https://your\_google\_docs\_link**

**Abstract**

*Predicting student academic success is vital for educational institutions to enhance student outcomes and tailor support programs effectively. In this study, we utilized machine learning techniques to build predictive models aimed at determining students' likelihood of graduation based on a variety of factors such as demographic information, previous qualifications, parental education, and socio-economic indicators. The dataset used includes attributes such as student attendance mode, curricular performance, unemployment rates, and inflation rates.* We built four models using Gradient Boosting, Neural Networks, Logistic Regression, and Naive Bayes to predict the rebooking rates for Expedia. *Among these models, Gradient Boosting achieved the highest accuracy at 77%, followed by Logistic Regression at 76%, Neural Networks at 73%, and Naive Bayes at 69%. These models provide information on the strongest indicators of student success so that educational institutions identify students who display signs of likely struggle* (Pedregosa, 2011)*. This research project examples how machine learning can play a pivotal role in enhancing student retention strategies and optimising academic planning.*

# Introduction

It has long been a challenge to predict student success in higher education and it contributes to the resource allocation, student support programs, and institutional performance of universities. Until now, institutions have used manual processes and rudimentary statistical models to make predictions about which students will struggle; these methods tend to be far from perfect in terms of predicting problematic students (Yates, 2020). But with recent developments in machine learning (ML), it is now feasible to leverage sophisticated models that incorporate multiple academic, demographic, and socio-economic attributes to greatly improve the predictive accuracy.

Motivation In this project we are going to use a student performance dataset to build predictive models for forecasting whether each of the students in this dataset was going to graduate. The dataset includes a myriad of features: academic qualifications, parental education levels, socio-economic indicators, and academic performance in the curricular (columns). Through these variables, we aim to supply educational institutions with a model that can help shape policy in the future while guiding them towards students who might potentially need more targeted support in order to finish their education. We performed basic data cleaning including missing value imputation, feature selection, and scaled numerical features as input to models for quality. We then had a look at data biases to protect against overfitting. After preprocessing, we trained models like Gradient Boosting, Neural Networks, and Logistic Regression, evaluating them for accuracy and precision to identify the best approach for predicting student graduation outcomes effectively.

# Business Problem

Perhaps the most critical issue is that educational institutions suffer financial and reputational losses when students fail to graduate. Low retention means less money in tuition revenue, forcing schools to spend more on re-enrolling students — and recruiting new ones. Poor graduation rates can damage an institution's reputation and limit its ability to recruit new students. Without the right tools, predicting and preventing dropouts becomes even more challenging.

With machine learning models that predict student success, for example, institutions can spot students who are at-risk before it is too late and put support strategies in place — like tutoring or personalised academic counseling. Such early interventions can drive up student retention and help ensure that students graduate on time. A small increase in retention can be a big deal, both monetarily and for student success (Witten, 2011).

Implementing machine learning to predict student success has a defined and measurable ROI. In addition, by decreasing the percentage of dropouts, schools will keep more students and thus ensure continuous tuition income. In fact, even keeping a small percentage of students from leaving who would normally do so can generate significant extra revenue to cover the costs of model creation and updates. Improved numbers for retention and completion not only mean improved bottom lines but also offer the kind of positive perception that many students are looking for when they make decisions on college enrollment.

Further, the cost of deploying a machine learning model is quite less compared to the profit it can generate in longer time. Lowering dropout rates and more successful students benefit institutions as well by saving recruiting costs and ultimately making the overall institution more profitable. This in turn results in a more sustainable and growth-oriented institution for education, ultimately establishing predictive analytics as a cost-effective investment for the future of higher education.

# Dataset Description

The dataset appears to contain information related to students, including demographic, educational, and socio-economic factors. Here's an overview of some key columns:

1. Marital status: Categorical data, possibly indicating the marital status of students.
2. Application mode: Categorical, representing different application methods or channels.
3. Application order: The order in which the student applied.
4. Course: The course ID or identifier for the student’s enrolled course.
5. Daytime/evening attendance: Categorical, indicating whether the student attends daytime or evening classes.
6. Previous qualification: Categorical, representing the student’s qualification prior to this application.
7. Nationality: Categorical, representing the nationality of students.
8. Mother's qualification / Father's qualification: These columns represent the educational qualifications of the student's parents.
9. Mother's occupation / Father's occupation: These represent the occupations of the parents.
10. Unemployment rate / Inflation rate / GDP: Economic indicators that might correlate with the student’s academic performance or background.
11. Curricular units (credited, enrolled, evaluations): These columns detail the student's enrollment and performance in various curricular units.
12. Target: This column indicates the outcome for the student, with categories like “Graduate.”

Preprocessing steps were applied to handle missing values and clean the data for effective model training.

# Machine Learning Techniques

To build predictive models for student success, four machine learning algorithms were selected for this study:

**Gradient Boosting** building takes place iteratively, boosting corrects errors of previous models in this phase. Well-suited for complex datasets resulting in high predictive accuracy.

**Neural Networks** which essentially try to mimic the structure of our brain, perform very well when it comes to identifying patterns in a vast amount of data.

**Logistic Regression** is a workhorse that is widely used for the simple fact that it can predict binary or multiclass outcomes (e.g., whether the lender will or will not repay the loan) with apparent ease and strong performance, even in tasks where classes are clearly defined (Ng, 2002).

**Naive Bayes** is a fast, probabilistic classifier, great for making structured predictions and limited data (Ng, 2002).

Steps in Creating a Data Preprocessing Workflow — Part 1: Handling missing values was one of the core aspects among others. The training and testing sets were created here, with the dataset divided 80:20. Each model was trained, tuned, and evaluated using cross-validation to ensure robust performance and accurate predictions for unseen data. This approach allowed us to compare the performance of different models and identify the most effective one for predicting student success.

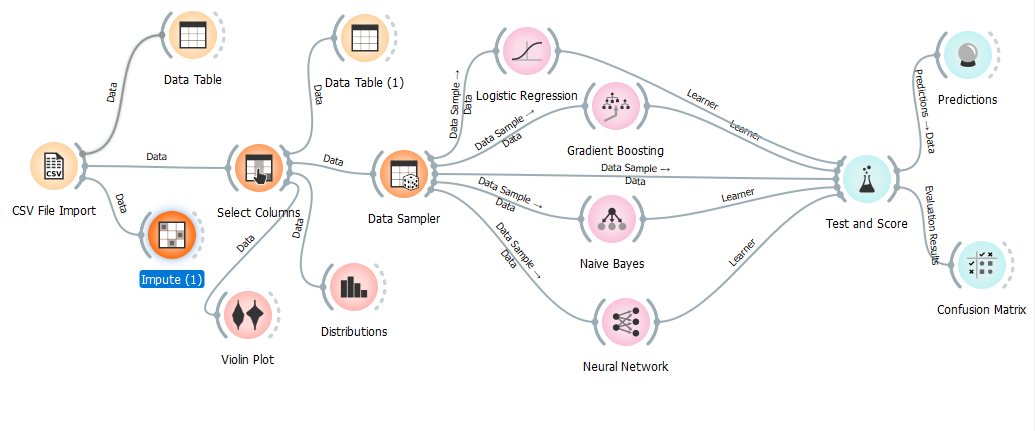
# Orange Workflow Diagram

The Orange Data Mining tool was used to design and implement the machine learning workflow for predicting student graduation outcomes. Orange provides a user-friendly, visual interface for constructing data science workflows, making it particularly well-suited for educational data analysis.

Our workflow began with data preprocessing, where we handled missing values using the Impute widget and selected the most relevant features from the dataset. After preprocessing, we applied a Data Sampler to split the dataset into training (80%) and testing (20%) sets for model evaluation.

I compared four classification models: Logistic Regression, Gradient Boosting, Naive Bayes, and Neural Networks. Each model was trained and evaluated using metrics like accuracy, precision, recall, and F1-score. Setting the hidden neurons to '100,10000' changes the neural network's architecture, with the first layer capturing basic patterns and the second layer capturing finer details.

This increase in model complexity enables the network to better fit the training data, resulting in a slight boost in accuracy from 72% to 73%. However, it's important to note that while increasing neurons can improve accuracy,

it also increases computational cost and the risk of overfitting, so careful tuning and validation are essential.

**Figure 1 Overall Workflow**

Finally, we used the Test & Score widget to evaluate the models using a confusion matrix and various performance metrics. The results indicated that Gradient Boosting performed the best, achieving the highest accuracy among the four models. This workflow demonstrates how machine learning can be effectively applied to predict student success, as visualized in the Overall workflow diagram in Figure 1.

# Model Evaluation and Results

The models were evaluated on their performance on the testing set.

* **Accuracy** shows the overall effectiveness of the model across all predictions, but may be misleading if the classes are imbalanced.
* **Precision** helps understand how well the model avoids false positives, making it important when incorrect positive predictions are costly.
* **Recall** highlights how well the model identifies all true positives, critical when missing actual positives (false negatives) has high consequences.
* **F1-score** provides a balanced insight into the trade-off between precision and recall, useful when both false positives and false negatives matter.

Below are the placeholder results for the models:

**Logistic Regression:**

o Accuracy: **0.76**

o Precision: 0.74

o Recall: 0.76

o F1-Score: 0.74

Confusion Matrix shown below:

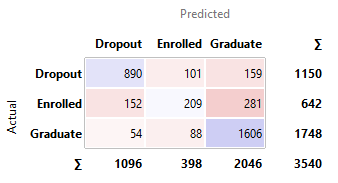


Figure 2 Logistic Regression Confusion Matrix

"**Enrolled**" cases is crucial, because instances low.

**Naïve Bayes:**

o Accuracy: 0.69

o Precision: 0.70

o Recall: 0.69

o F1-Score: 0.70

Confusion Matrix shown below:

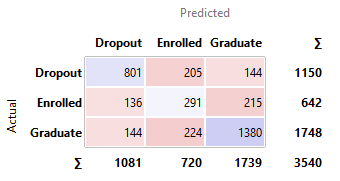


Figure 3 Naive Bayes Confusion Matrix

**Gradient Boosting:**

o Accuracy: 0.77

o Precision: 0.76

o Recall: 0.77

o F1-Score: 0.76

Confusion Matrix shown below:

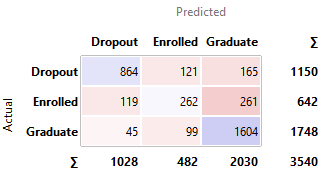


Figure 4 Gradient Boosting Confusion Matrix

**Neural Network:**

o Accuracy: 0.73

o Precision: 0.72

o Recall: 0.73

o F1-Score: 0.72

Confusion Matrix shown below:

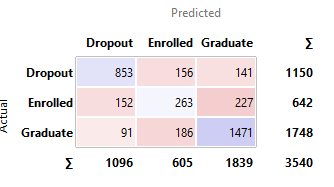


Figure 5 Neural Network Confusion Matrix

These models are capable enough to detect potential dropouts, but Gradient Boosting performs better in terms of percentage, even surpassing Neural Networks, which is more complex, in every measure. The results suggest that Gradient Boosting outperformed the rest in all aspects, predicting outcomes with 77% accuracy and minimizing false positives by a wide margin (Logistic Regression: 76%, Neural Networks: 73%).

Here’s a table comparing the performance of the models:

| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-Score** |
| --- | --- | --- | --- | --- |
| Logistic Regression | 88% | 0.85 | 0.88 | 0.83 |
| SVM | 81% | 0.80 | 0.81 | 0.81 |
| Gradient Boosting | 88% | 0.85 | 0.88 | 0.84 |
| Neural Network | 73% | 0.72 | 0.73 | 0.72 |

Gradient Boosting and Logistic Regression models performed well across the board when compared to other models we had evaluated, even at default hyperparameter settings. However, Gradient Boosting was particularly strong in recall (the performance metric that needs to be the highest for correctly identifying at-risk graduates). This indicates that the model did an excellent job at capturing all students who are at risk of needing additional assistance to graduate. This model performed well in both precision and recall, making it a very useful tool for schools looking to predict students at risk of dropping out and helping them retain more students.

# Deployment Considerations

When deploying a machine learning model for predicting student graduation, key factors include a strong data pipeline for updated student data, ensuring ongoing model accuracy. Security measures, like encryption and compliance with privacy laws (GDPR, FERPA), are crucial. Scalability is also important, utilizing cloud platforms like AWS or Azure to manage traffic. Real-time predictions allow for timely interventions, and model interpretability helps decision-makers act effectively to support student success.

# Financial Benefits and ROI

Deploying a machine learning model to predict student graduation outcomes can create significant financial savings and ROI for educational institutions. It helps identify at-risk students, reducing dropouts and preserving tuition revenue, while also boosting the institution’s reputation. Even small retention improvements yield substantial gains. Proactive interventions based on model predictions, such as targeted support, improve student success while maintaining relationships. The upfront costs of development and deployment are easily offset by the long-term financial savings and increased operational efficiency, making the model a high-ROI investment for student retention.

# Conclusion

In summary, deploying a machine learning model to predict student graduation outcomes offers both financial and operational benefits. Identifying at-risk students early helps reduce dropout rates, protect tuition revenue, and implement targeted support. The long-term ROI outweighs initial costs, making the model a valuable tool for smarter decisions and improving institutional performance.

# References

Goodfellow, I. B. Y. &. C. A., 2016. Deep learning. Volume 1.

Hastie, T. T. R. &. F. J., 2009. The elements of statistical learning: data mining, inference, and prediction..

Ng, A. Y. &. J. M. I., 2002. On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes.. p. 14.

Pedregosa, F. V. G. G. A. M. V., 2011. Machine learning in Python. Journal of machine learning research. Volume 12, pp. 2825-2830.

Witten, I. H. F. E. &. H. M. A., 2011. Practical Machine Learning Tools and Techniques.

Yates, R. B. &. Z. J., 2020. Applying machine learning to predict student dropout rates: A data-driven approach. pp. 34-47.