Student Performance Analysis for Australian Schools

Technical Report

Prepared By: Srijana Bhusal

Table of content

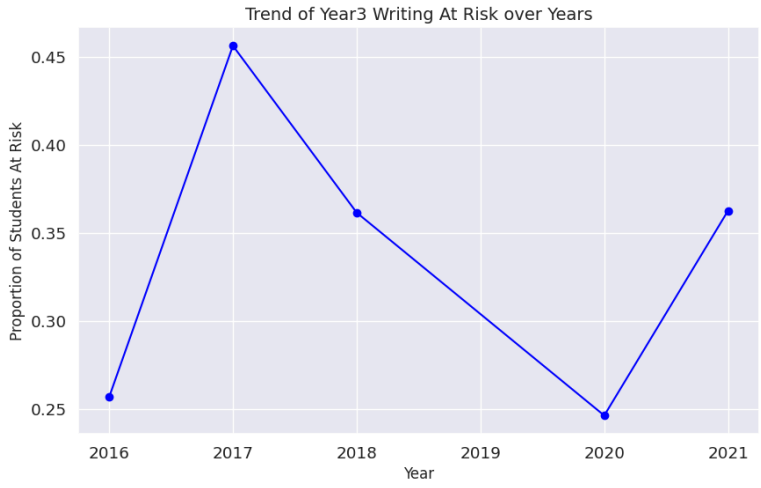
Executive summary

The objective of this project is to develop a predictive model that identifies students at risk of underperforming in Year 3 writing, a critical focus for enhancing educational outcomes. The significance of this project lies in its potential to improve intervention strategies for students, thereby supporting their academic success. Data was collected from 40 Australian schools, providing a robust foundation for analysis. Our methodology employs both supervised and unsupervised machine learning models; specifically, we use logistic regression and decision trees to predict whether students are at risk, while K-means clustering identifies underlying patterns among student characteristics. Key findings reveal that logistic regression outperforms decision trees, achieving an accuracy of 0.74 compared to 0.72, with superior precision and recall metrics. As a result, logistic regression is recommended for guiding targeted interventions. In conclusion, stakeholders are encouraged to implement the logistic regression model for predicting at-risk students while considering further data collection and standardization to enhance model accuracy and adhere to data protection policies. This approach will enable the direct application of the model to identify students at risk in the future, ultimately benefiting the educational community.

Introduction

The project aims to support a consortium of forty Australian primary schools in collaboration with Data2Intel to pre-identify underperforming students and help them with the necessary support. We plan on providing solutions based on our machine learning models to predict the students at risk by identifying the factors that are affecting student performance ratings. The major intention of this project is to identify areas where early interventions can be applied in the foundational schooling years for long-term academic success.

For this, we are provided with information on 2000 students from 40 different primary schools with 33 features associated with each student from 2016 to 2020. The risk indicators are associated with each student which are obtained from the NAPLAN result of Year 3.



Objective

The objective of this project is to develop a predictive model that can help identify students at risk of underperforming in Year 3 writing.

Approach

In the project, we use supervised and unsupervised models. We have labeled data explaining student's past performance which can be used to predict future performance making supervised algorithms suitable for prediction(*Google Scholar*, 2024). We will use two predictors: logistic regression and decision tree to predict the target variable. However, there might be data patterns that might not be explained by the supervised models (Naeem et al., 2023), hence we use unsupervised models to uncover these patterns which can help us in student profiling and targeting customized interventions.

For supervised learning, we use logistic regression as the problem deals with the target which follows binary classification i.e. student is at risk or not. This model provided

probabilistic weightage of features through equation making it very easy to understand the contribution of each feature in predicting the target(Jessen & Menard, 1996).
Our target is to identify whether the students are at risk or not, i.e. a categorical variable. We have 3 categorical variables and the remaining numerical variables, referred to as features or predictors to predict the target.

We use a decision tree since it can properly handle nonlinear relationships between the variables(Mendonça et al., 2023). Furthermore, a decision tree helps the nontechnical stakeholders easily understand how decision-making is being done. K mean clustering is chosen for understanding the hidden patterns among the features without taking into account the target variable. It will help cluster students with similar characteristics together making it easy to tailor the interventions per segment.

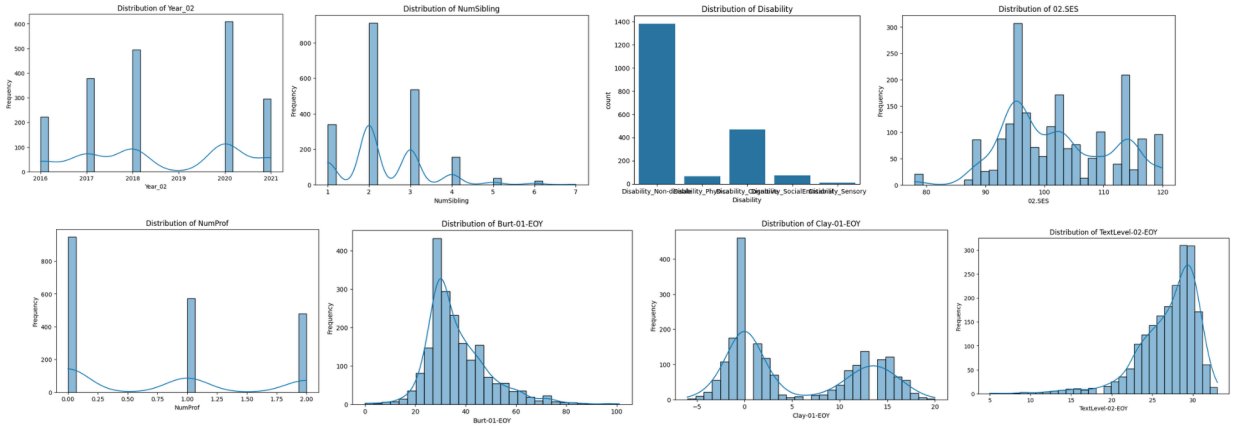
Hence, the major target of developing machine learning models is to predict students who are at risk of underperforming year 3 tests and grouping students with similar characteristics together to apply customized interventions.

Data preparation and Exploratory Data Analysis (EDA)

The dataset has been collected from 40 Australian schools which has information of 2000 students about their demography, school characteristics, family background, and assessment scores for literacy and numeracy along with their results in the Year 3 NAPLAN test. There is information about 2000 students explained by 34 different attributes.

We have performed univariate, bivariate, and multivariate analyses of all the attributes, and below are the findings on some important features:

Univariate analysis:

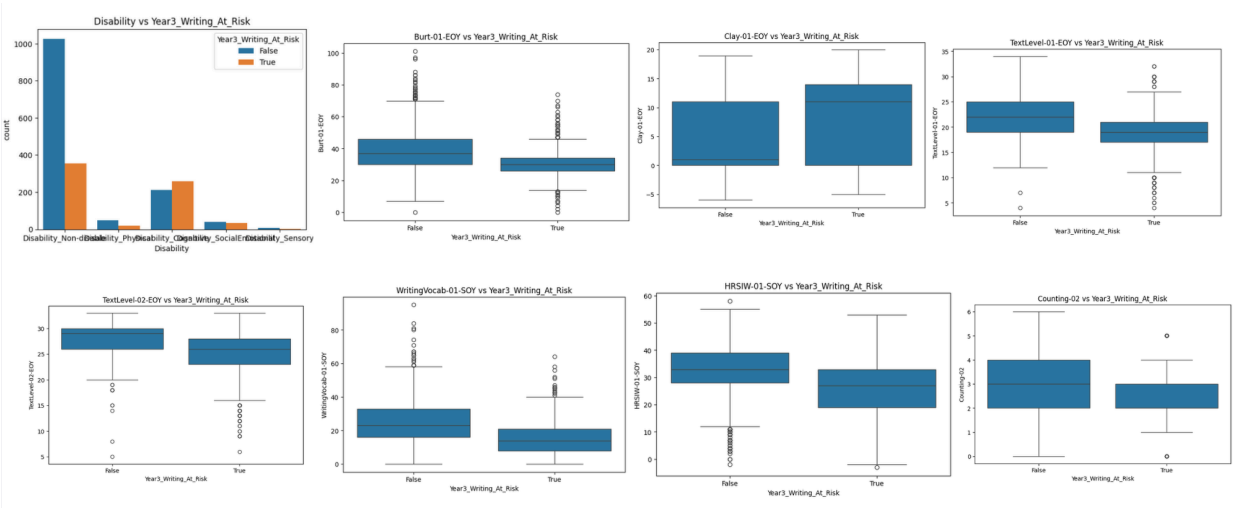


The univariate plot of the year shows a missing value from year 9, most probably due to the COVID-19 outbreak, hence the number of students undertaking the NAPLAN test is surging during the year 2020 as the tests were postponed due to the pandemic. Most of the students have

2 siblings with the maximum number being 7. Out of 2000 students, 1381 students are not suffering from any disabilities. Among the ones suffering from a disability, cognitive disability is the most frequent one followed by socioemotional, physical, and sensory respectively. The average SES score is 102 with the highest being 120 and the lowest being 78 which shows a notable disparity among the socioeconomic status of schools. In most cases, no parents of the school children have professional jobs. The average score of the Burt test at the end of the year is 36 with a minimum being 0. More than 50% of students have a Burt score above 29. The distribution is highly skewed with a few outliers. The clay test shows an inconsistent distribution with the average score being 5, the maximum being 20 and the minimum being -6. However, all the tests are subjected to maintain a range from 0 and above. Hence, we will be reconsidering the negative scores as 0 for further analysis. For the TextLevel test, the average score is 21 with the highest being 34 and lowest being 4, showing a skewed distribution.

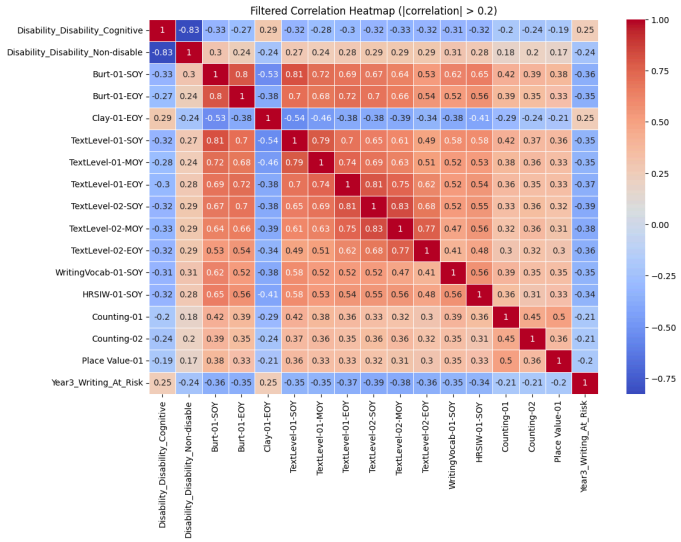
Bivariate analysis:

Bivariate analysis has been done by pairing the feature target variables with the target variable.



Students with disability are at high risk of underperforming. The cognitive ability has a notable impact on student performance. Students performing low on the Burt test also contribute more to the risk. The Clay test has a high impact on students' performance. Students performing poorly at TextLevel, writing vocabulary, HRSW and counting are more at the risk of underperforming.

Multivariate analysis



The above heatmap contains only the feature with corr values >=0.2. Features like cognitive disability, Burt test, Clay test, TextLevel test, writing vocabulary, hearing and recording ability, and counting are highly correlated to the target variable in comparison to the other variables. Since the features in numeracy and literacy at the start and end of the year have very high correlations, we will be using their average for further processing.

Data Preprocessing:

No null data were found in the dataset. However, there were unexpectedly many negative data. The scores in the test are supposed to be 0 to a positive range but since few of the tests are done locally, e.g. Clay test, we can observe few negative values. Hence, to maintain the data sanity, the negative numbers have been converted into 0. Similarly, the student ID column has been dropped as it doesn't provide any significant contribution to the prediction. Similarly, based on the NSPLAN method of calculating the relative growth of students, we have used a similar technique to capture information from multiple columns and devised new columns under "relative growth". We have used an average score for the test to make our analysis clear and easier and reduce multicollinearity.

Converting categorical into numerical data:

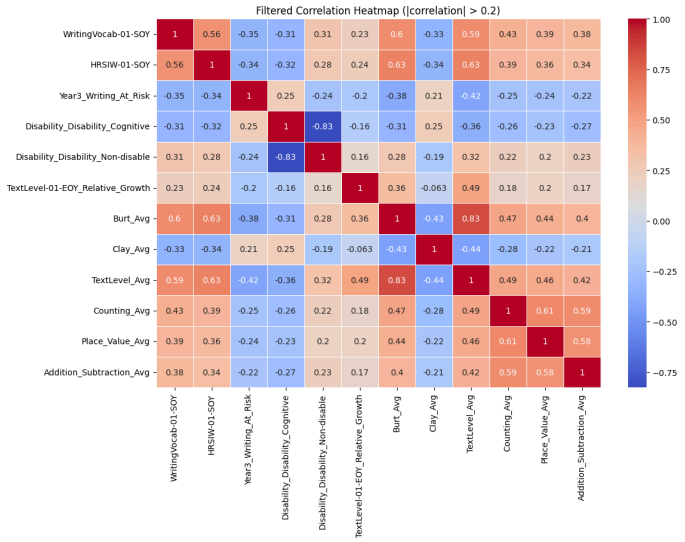
Two categorical data: Gender and Disability have been converted from categorical to numerical data using dummy variables through a one-hot encoding technique. Conversion is essential for machine learning models to process information from these features. Since there is no relationship between the values in Gender i.e. male and female or the values among disability types, no ordinality or natural ranking can be assumed for the values. Hence, we have selected a one-hot encoding technique for converting categorical data types into numerical ones to prevent the model from assuming any ordinal relationship among the categories (Poslavskaya & Korolev, 2023).

Feature scaling:

Different columns had different data ranges. Since we are planning on using logistic regression, the weightage of features would be affected by the difference in data range. Higher values would be assigned higher weightage making erroneous predictions. Hence, we have scaled all the values in the range of 0 to 1 using min-max scaling.

Feature selection:

A new heatmap has been generated using the new values.



All the values with corr values >=0.2 are selected for further prediction. Since the tests have been independently conducted, we have not omitted the tests with high correlations among each other.

Model development and evaluation

Supervised Machine Learning

For both the supervised models, we have divided the dataset into an 80:20 ratio, with 80% dataset allocated for training and 20% for testing purposes.

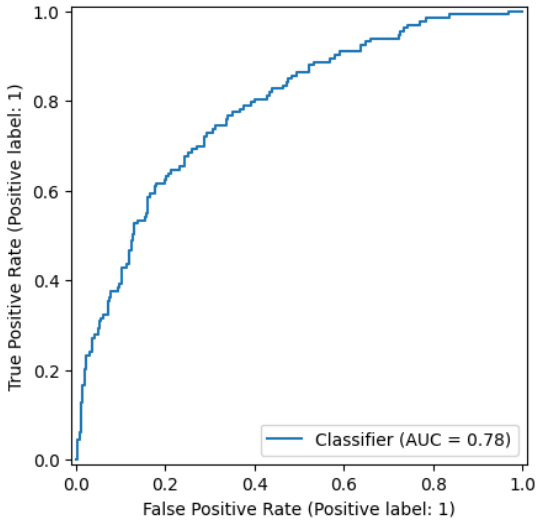
Logistic Regression

Confusion Matrix

	Predicted 0	Predicted 1
Actual 0	235	32
Actual 1	71	62

Classification Report

	Precision	Recall	F1-Score	Support
0.0	0.77	0.88	0.82	267
1.0	0.66	0.47	0.55	133
Accuracy			0.74	400
Macro Avg	0.71	0.67	0.68	400
Weighted Avg	0.73	0.74	0.73	400



Based on the confusion matrix, logistic regression has been able to correctly detect 62 students at risk and 235 students correctly as not at risk. However, 32 students, who were not at risk have been classified at risk giving rise to false alarms. Similarly, 71 students who were actually at risk were categorized as not at risk which reduced the model accuracy. 66% of the time, the model can predict the students at risk correctly. However, the recall of 47% only suggests that the model is still missing out on many students who were supposed to be detected at risk of underperformance.

Pros and cons of using Logistic Regression:

Because of its ease of use and interpretability, logistic regression is a popular statistical model for tasks involving binary categorization. Its ability to give stakeholders with clear probabilistic interpretations of the results is one of its main advantages. This enables them to comprehend the various elements that influence a student's likelihood of underperforming. Furthermore, logistic regression is less likely to overfit, particularly when dealing with a high number of observations, and performs well with data that can be divided linearly. Its efficacy, however, may be constrained when handling intricate, nonlinear interactions between variables, which could cause

the model to become overly simplistic. It also assumes that there is a linear relationship between the independent factors and the dependent variable's log odds, which may not always be the case.

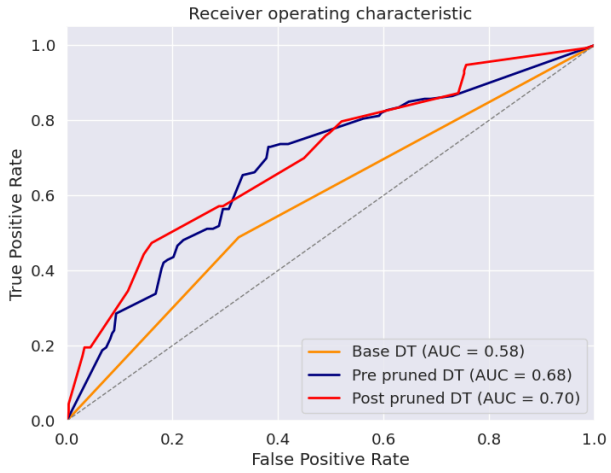
Decision Tree

Confusion Matrix

	Predicted 0	Predicted 1
Actual 0	224	43
Actual 1	70	63

Classification Report

	Precision	Recall	F1-Score	Support
0.0	0.76	0.84	0.80	267
1.0	0.59	0.47	0.53	133
Accuracy			0.72	400
Macro Avg	0.68	0.66	0.66	400
Weighted Avg	0.71	0.72	0.71	400



The decision tree was able to correctly identify 63 students at risk and 224 students who were not at risk. However, 43 students who were not at risk were classified as students at risk which might give rise to false alarms. Also, 70 students who were actually at risk were identified to be at no risk leading to lower accuracy of the model. The model can accurately predict students at risk 59% of the time. The low recall of 47% suggests that the model has missed many students who were to be identified as risk.

Pros and cons of using a decision tree:

Decision trees provide both technical and non-technical stakeholders with a simple and adaptable method for classifying activities. Their main benefit is that they can simulate interactions and nonlinear correlations between characteristics, which makes it possible to conduct more intricate

decision-making procedures. Additionally, decision trees can be graphically interpreted, which facilitates comprehension of the decision-making process and makes them appropriate for sharing insights with different stakeholders. However they can be prone to overfitting, which can result in poor generalization on data that hasn't been seen, particularly with tiny datasets or when the tree is allowed to develop deeply. Decision trees may also be susceptible to subtle variations in the data, which could lead to varying tree architectures and interpretations and, ultimately, unstable predictions from the model.

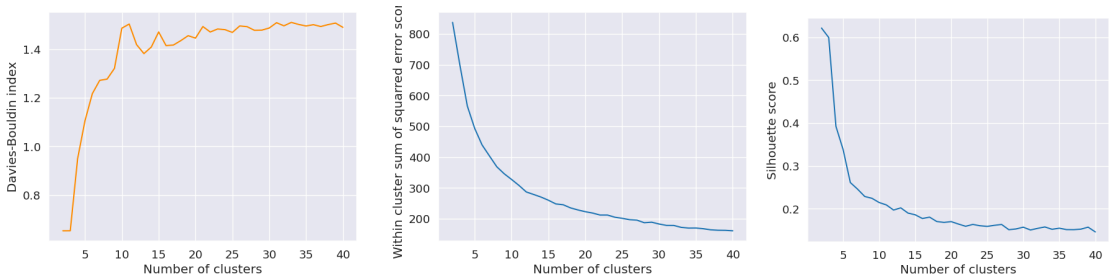
Comparison between Logistic Regression and Decision Tree

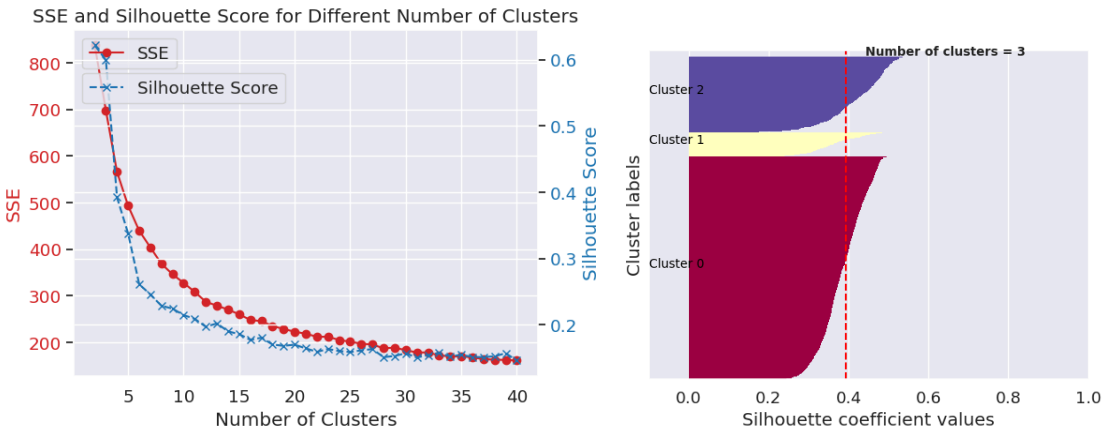
Logistic regression performed better than the decision tree both in terms of accuracy and precision. Compared to the Decision Tree, which has an accuracy of 0.72, Logistic Regression has a higher overall accuracy of 0.74, demonstrating its superior capacity to categorize instances correctly across both classes. This benefit is further highlighted by the accuracy measurements, where the Decision Tree achieves a precision of 0.76 for class 0.0 and 0.59 for class 1.0, whereas Logistic Regression achieves a precision of 0.77 for class 0.0 and 0.66 for class 1.0. While both models show difficulties in finding true positives for class 1.0, Logistic Regression once more wins in recall. In terms of recall, Logistic Regression leads once more with a recall of 0.88 for class 0.0, indicating a good ability to identify students at risk of underperforming in writing, even though both models show difficulties in recognizing true positives for class 1.0. Logistic regression's balanced performance is further demonstrated by its F1-scores, which are 0.55 for class 1.0 and 0.82 for class 0.0, respectively. This is notably true for the more common class 0.0. Additionally, Logistic Regression regularly performs better overall and is robust across classes, as seen by the macro and weighted averages of the performance indicators. Logistic Regression is a more effective model in this educational analytics setting due to its greater precision, recall, and F1-scores, as well as its clearer explanatory power regarding feature impacts, despite Decision Trees' advantages in interpretability and visualization.

Although both models struggle to identify class 1.0 cases, Logistic Regression is the best option since it provides stakeholders with the performance and interpretability that they need to successfully serve children who are in need. Therefore, it is advised that the Logistic Regression model be used to guide focused interventions for elementary school students who are at risk of writing poorly, while also taking into account the possibility of improving the model further by adjusting hyperparameters or investigating alternative algorithms.

K mean clustering:

Highly correlated features i.e. WritingVocab-01-SOY, HRSIW-01-SOY, Burt_Avg, TextLevel_Avg, Counting_Avg, Place_Value_Avg have been used to develop clusters. While a random k value of 4 was selected in the beginning, we selected the optimal k value as 3 after looking at the WCSS score, DBI, Silhouette scores, and the elbow curve.





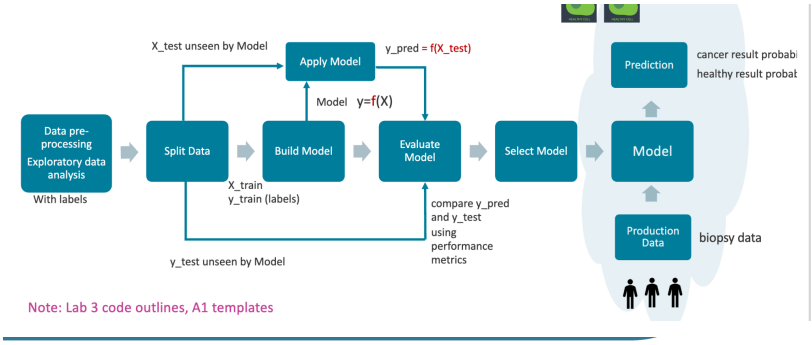
The elbow curve suggests a tradeoff between the SSE score and the Silhouette score where a lower SSE score and lower Solhouette score are considered ideal for clustering. The balance has been assumed to be at k=3 from the curve above.

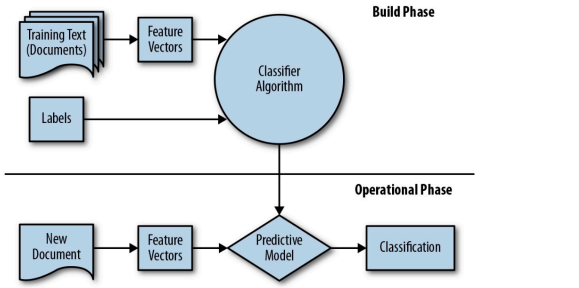
Solution recommendation:

Based on the model, the client is suggested to use logistic regression due to its high accuracy and simplicity. It is also suggested that more data could be gathered to improve the model's accuracy. Also, the dataset had some erroneous values, hence the data collection step could be standardized so that the ranges remain within the boundary. While collecting the dataset, it is recommended to follow the standard data protection policies and anonymize the data so that the personal information remains safe.

Technical recommendations:

The model has been developed using Python 3 in Google Colab IDE. Libraries and packages like Numpy, Pandas, Matplotlib, ScikitLearn, Seaborn, etc. have been used. The model has been developed using the following process as shown in the diagram.





Bengfort et al., 2018

Several tactics must be put into practice in order to keep the predictive model relevant and accurate. To make sure the model satisfies defined benchmarks, it should undergo regular evaluations using a separate validation dataset to track important performance metrics like accuracy and recall. Student performance trends will be updated with new demographic and educational data regularly, bringing the training dataset up to date. Setting up a feedback mechanism will enable stakeholders and educators to offer suggestions for model modifications, and regular retraining will aid in the identification of novel patterns. Reevaluating feature relevancy is a good idea, and working with educational specialists can provide new information about how educational demands are changing. Stakeholder trust can be fostered by maintaining model transparency, which will also enable effective tracking of changes through version control and thorough documentation.

Reference

Google Scholar. (2024). Google.com. https://scholar.google.com/scholar?q=Supervised%20Machine%20Learning:%20A%20Review%20of%20Classification%20Techniques

Jessen, H. C., & Menard, S. (1996). Applied Logistic Regression Analysis. *The Statistician*, *45*(4), 534. https://doi.org/10.2307/2988559

Mendonça, Y. V. S., Naranjo, P. G. V., & Pinto, D. C. (2023). The Role of Technology in the Learning Process: A Decision Tree-Based Model Using Machine Learning. *Emerging Science Journal*, *6*, 280–295. https://doi.org/10.28991/esj-2023-sied-020

Naeem, S., Ali, A., Anam, S., & Ahmed, M. M. (2023). An Unsupervised Machine Learning Algorithms: Comprehensive Review. *International Journal of Computing and Digital Systems*, *13*(1), 911–921. https://doi.org/10.12785/ijcds/130172

Poslavskaya, E., & Korolev, A. (2023, December 28). *Encoding categorical data: Is there yet anything "hotter" than one-hot encoding?* ArXiv.org. https://doi.org/10.48550/arXiv.2312.16930