UNIVERSITY OF THE
WITWATERSRAND,
JOHANNESBURG

Exploration of the usability of explainable and counterfactual machine learning for students at risk of failing a subject

Course Code: MECN4006A
Course: Research Project

Supervisor: Dr Bevan Smith
Course Coordinator: Mr Sjouke Schekman

A project report submitted to the Faculty of Engineering and the Built Environment, University of the Witwatersrand, Johannesburg, in partial fulfilment of the requirements for the degree of Bachelor of Science in Engineering.

Johannesburg, 3 November 2023

Abstract

This research paper explores the potential of explainable and counterfactual machine learning in the field of education. The research question of this study is to investigate the usability of using explainable and counterfactual machine learning techniques to improve the academic performance of students at risk of failing a subject. The research aims to investigate whether machine learning approaches can provide interpretable explanations regarding predictions for at-risk students while evaluating their effectiveness in suggesting personalised interventions to improve the students' final grades. To accomplish this, the research employs model-agnostic methods like SHAP and LIME, which allows for a deep analysis of the results, ensuring a rigorous examination of the consistency of feature importance across various machine learning models. The results reveal that machine learning offers a promising avenue to assist educators and students alike in navigating the complex landscape of education, steering away from failure and towards success. The study holds the promise of improving student academic performance, but it is crucial to utilise these techniques with careful consideration. Through this investigation, the education landscape may find a valuable tool to address this educational concern effectively.

SCHOOL OF MECHANICAL,
INDUSTRIAL & AERONAUTICAL
ENGINEERING

UNIVERSITY OF THE WITWATERSRAND
JOHANNESBURG

INDIVIDUAL DECLARATION WITH JOINT TASK SUBMITTED FOR ASSESSMENT

I (name) Castello Chetty (student number) 2081719,

am registered for (course code) MECN4006A (course name) Research project

in the year 20 23.

I herewith submit the following task, "(title) Research Project - Exploration of the usability of explainable and counter-factual machine learning for students at risk of failing a subject"

in partial fulfilment of the requirements of the above course.

I hereby declare the following:

○ I am aware that plagiarism (the use of someone else's work without their permission and / or without acknowledging the original source) is wrong;

○ I confirm that the work submitted herewith for assessment in the above course is my own unaided work except where I have explicitly indicated otherwise;

○ This task has not been submitted before, either individually or jointly, for any course requirement, examination or degree at this or any other tertiary educational institution;

○ I have followed the required conventions in referencing the thoughts and ideas of others;

○ I understand that the University of the Witwatersrand may take disciplinary action against us if it can be shown that this task is not our own unaided work or that we have failed to acknowledge the sources of the ideas or words in our writing in this task.

Signature________ **Date** 02/11/2023

List of content

List of figures

List of tables

Nomenclature

AI	-	Artificial intelligence
LIME	-	Local interpretable model-agnostic explanations
ML	-	Machine learning
MSE	-	Mean squared error
RBF	-	Radial Basis Function
SHAP	-	SHapley Additive exPlanations
UCI	-	University of California Irvine
XAI	-	Explainable artificial intelligence

1. Introduction

1.1 Background and motivation

In the dynamic landscape of education, where the quest for knowledge and success unfolds, there exists a persistent challenge that has remained for generations: the educational prospects of students that are on the brink of failing a subject. The pursuit to improve the academic performance of students remains a central focus in the ever-changing landscape of the education sector.

Ensuring that students pass their academic endeavours holds vital importance. Passing signifies personal achievement and paves the students' way for a brighter future. It bolsters an individual's self-esteem and confidence, reducing the stress and anxiety associated with academic struggles. Passing empowers individuals to pursue better opportunities and leads to improved overall well-being. It fosters a sense of accomplishment and potential, instilling the belief that they can overcome challenges and fulfil their aspirations.

When students perform well academically, they tend to be more productive and innovative in the workforce, contributing to economic growth. Higher-achieving students also have higher earning potential, which not only improves their individual financial well-being but also leads to increased consumption and economic activity. Moreover, better student performance can help reduce social costs related to lower academic achievement, including crime, unemployment, and social welfare expenditures. Countries with highly educated workforces are more competitive in the global market, attracting investments and enhancing international trade, which ultimately benefits the economy. Hence, this reiterates the importance of improved academic performance not only for students but for the economy as well.

The motivation behind this research topic stemmed from the need to address educational discrepancies and empower educators and students with actionable insights. Students face an array of challenges, from varying socioeconomic backgrounds to diverse learning abilities, that can significantly impact their academic performance. Identifying at-risk students is an essential first step, but it is insufficient when isolated. To genuinely effect change, one must delve into the underlying mechanisms driving academic performance and explore actionable interventions capable of altering the predicted path of failure, thus serving as counterfactual explanations for student academic performance.

Machine learning is an ideal tool for tackling the challenge of improving the academic performance of at-risk students for several reasons. Machine learning excels in handling complex and diverse datasets, making it well-suited to process large volumes of student-related data, including demographic, social, and academic information. This capability surpasses the constraints of traditional manual analysis by humans.

Machine learning offers powerful predictive capabilities. These models can forecast a student's academic progress and identify potential risks early. This predictive element allows educators and institutions to intervene proactively, providing timely support to at-risk students. Additionally, machine learning models have the potential to provide personalized interventions. By leveraging a student's individual characteristics and performance data, these models have the potential to tailor recommendations through counterfactual explanations to address specific needs. This level of personalization is often difficult and time-consuming to achieve with conventional methods.

Explainable machine learning algorithms provide a critical lens through which the decision-making processes of predictive models can be interpreted. These algorithms unearth the critical features that contribute most significantly to model predictions. However, mere interpretation, while valuable, may fall short in addressing the core issue. At this juncture, counterfactual machine learning comes into play, permitting educated interventions to ensure that the predicted outcome is changed. Consequently, it can reveal actionable insights for directing students away from the predicted path of failure towards improved academic performance.

This research travels down a path that combines the art of teaching with the science of machine learning by investigating the application of explainable and counterfactual machine learning in the field of student academic performance. Through understanding why students are at risk of failing and uncovering strategies to modify their academic outcomes, this research aims to equip educators and students with the tools to foster the potential within each student, ultimately ensuring a brighter future for all.

1.1.1 Problem statement

As educators and institutions grapple with the complex task of identifying and providing optimal support for at-risk students, it becomes increasingly evident that relying solely on conventional approaches, such as observations by educators or involvement of students' parents', is no longer adequate. The intricate nature of academic challenges, entangled with innumerable personal, demographic, and socioeconomic factors, demands a more refined and data-driven approach. As a result, there is a growing recognition that innovative approaches are needed to identify, understand, and address the specific needs of at-risk students.

1.1.2 Critical research question

How can the integration of explainable and counterfactual machine learning techniques generate academic support and interventions for students at risk of failing, providing an effective approach to improve their academic performance and outcome?

1.1.3 Objectives

- Investigate the use of machine learning models in providing interpretable explanations to users regarding predictions concerning students that are at risk of failing a subject.
- Explore the application of counterfactual explanations in suggesting tailored interventions aimed at improving the academic performance of students at risk of failing a subject.

1.2 Literature review

Machine learning and artificial intelligence have garnered considerable interest in the field of education, especially when it comes to enhancing students' academic performance. A crucial factor in employing machine learning in education is ensuring that models are transparent and interpretable. Explainable artificial intelligence and counterfactual machine learning have emerged as promising methods for enhancing comprehension and decision-making abilities in the effort to boost student achievement.

1.2.1 Prediction of student academic performance

Being among the foremost concerns within the realm of academia, the academic achievement of students holds significant importance for educational institutions. According to Doctor, early signs indicating a student's likelihood to successfully complete a course enable educators to promptly intervene by offering additional educational support, thereby aiding in the enhancement of student performance [1]. This analogy is also applicable to students that are at risk of failing a course. Early awareness allows educational institutions to intervene promptly. They can provide these students with the necessary support, such as tutoring and additional resources, to help them succeed. Institutions can allocate resources more efficiently by targeting students who need the most assistance. This ensures that limited resources, such as time and funding, are used effectively to support those who are at risk of failing.

Singh's research on the impact of extracurricular activities on student academic performance determined that these activities had a favourable effect on academic achievement [2]. They complemented the primary educational objectives by positively influencing students' behaviour. It was found that co-curricular activities, such as participating in sports and clubs, did not hinder academic performance. Instead, they enhanced knowledge acquisition and cultivated a competitive spirit that strengthened students' determination in examinations. [2]

Chavez et al. created a precise machine learning model that exclusively focused on behaviours susceptible to intervention [3]. Their research revealed that students who had previously been at risk of failing but adhered to the intervention recommendations outperformed their peers who were on a similar path but did not heed the advice [3]. This reiterates the need to develop accurate prediction models which paves the way for tailored interventions to address specific behaviours or challenges contributing to the students’ risk of failure.

In an academic performance prediction study conducted by Orji and Vassileva, a comparison of various machine learning models showed that tree-based models, such as the random forest, outperformed linear, support vector, and k-nearest neighbours models, demonstrating superior results for that specific dataset [4]. It was found that comparing models is a crucial practice as it allowed for a thorough assessment of how well different models perform in solving specific problems, aiding in the selection of the most appropriate one. Model comparison helps optimize model configurations and hyperparameter tuning for better performance on a given dataset. This literature showed that an optimised, proficient model is critical before attempting to provide interventions based on the model’s predictions.

According to research on student academic performance conducted by Shahiri et al., the most precise classification machine learning model was a neural network when applied to a Malaysian educational institution dataset [5]. The neural network outperformed models such as the decision tree, support vector, k-nearest neighbour, and naïve bayes. One benefit of the neural network was its capacity to effortlessly capture non-linear relationships [5]. Shahiri et al. noted that the neural network model exhibited adaptability by being able to update historical data similar to the way the human brain does [5]. However, the study also found that the neural network model struggled to make accurate predictions based on qualitative data such as psychometric factors [5].

In a research paper published by Niyogisubizo et al., a cross validation approach was used on machine learning models to reduce the risk of over-fitting [6]. The paper was based on the accurate prediction of student dropouts in university. To further improve accuracy of the models, a grid search method was utilised to optimise hyperparameters for each model [6]. However, the study did not solely focus on optimisation of model accuracy as it could have led to misinterpretation. Evaluation metrics such as precision and F1-score were used to support the selection of models [6]. Niyogisubizo et al. found that a limitation for the machine learning models was the limited size of data. In the field of education data often needs to represent individual students' learning outcomes or behaviour, resulting in datasets that are typically smaller than what machine learning algorithms ideally require [6]. This posed a limitation to the predictive performance of the tested machine learning models.

Although prediction of student academic performance is important, explainability of the prediction is more important. Explainability allows humans to interpret the models and provide reasoning as to why machine learning models made certain predictions.

1.2.2 Explainable artificial intelligence in education

Miller found that explanations are inherently comparative in nature, as they are prompted by specific counterfactual scenarios [7]. Individuals do not request why an event occurred but rather why the event occurred instead of some other event [7]. He mentioned that this characteristic has significant implications for both the social and computational aspects of explainable artificial intelligence [7]. According to Miller, probabilities may hold limited significance. Although truth and probabilities play essential roles in providing explanations, the utilization of statistical connections is not as impactful as invoking explanations [7]. Relying solely on statistical generalizations to account for the occurrence of events is insufficient unless there exists an underlying explanation for the generalization itself. [7]

Certain models, like rule-based models and decision trees, are inherently interpretable due to their relatively straightforward structures, enabling easy explanations of their predictions [8]. For instance, in the case of a basic decision tree, the rules can be articulated in a way that humans can comprehend and replicate the model's decision-making process. Conversely, models like tree ensembles, support vector machines, and deep neural networks possess intricate structures that are not immediately understandable. To render these models interpretable to humans, Khosravi et al. have dedicated considerable effort to developing post-hoc explainability techniques. These techniques aim to clarify how a complex model generates its predictions without revealing the model's underlying structure. [8]

The necessity for explanations arises from the responsibility of educators to be accountable, whether it be to students, parents, or educational authorities. This requirement is evident when giving individual feedback to students, providing teachers with diagnostic feedback to identify areas where a class of students requires more attention, and during parental consultations to assist parents in supporting their child's learning [8]. Feedback given to students and teachers represents a form of educational explanation crucial to achieving educational objectives. In the case of students, these explanations most commonly originate from teachers and encompass aspects such as evaluating student performance on specific tasks, offering suggestions for improvement, providing guidance for self-monitoring and direction, as well as offering emotional feedback [8]. The primary goals of this feedback are to support learning, foster the acquisition of domain knowledge, and to provide a sense of self-worth. However, it is vital to note that useful feedback cannot be obtained without interpretation of model predictions.

SHAP (Shapley Additive exPlanations) is an explainability method that is applicable to global and grouped data. Shalih et al. mentioned that in SHAP, the components analogous to players and payouts are replaced by features and the resulting model outcome [9]. SHAP computes a score for each feature within the model, signifying its significance in influencing the model's output [9]. To compute these scores, it takes into account all possible combinations of features, encompassing scenarios where all features are present or when only a subset of features is included in the model [9]. According to Shalih et al., it is vital to accompany SHAP results with easily understandable explanations of the findings and the underlying assumptions of SHAP, such as the independence of features and model-specific factors [9]. Additionally, it is advisable for the end user to employ various machine learning models when dealing with collinear features. This approach allows for a comparative analysis of SHAP outcomes from each model, facilitating a more comprehensive understanding of the results [9].

LIME (local interpretable model-agnostic explanations) is a method for providing local explanations that is not tied to a specific model. Its purpose is to explain the impact of each feature on the outcome for an individual instance [9]. It visually represents how each feature contributes to the target outcome. However, LIME achieves this by transforming any model into a local linear model, revealing coefficient values that indicate the feature weights within the model [9]. If the user employs models that consider complex relationships between features and outcomes, LIME's explanations may overlook this complexity as it simplifies the model into a linear form [9]. This was a limitation identified by Shalih et al. An important characteristic to note is that the interpretation provided by LIME pertains solely to an individual instance and cannot be extrapolated or regarded as a universal explanation for the entire model.

1.2.3 Counterfactual explanations in education

An explanation generated by an artificial intelligence model conveys the inner workings of an algorithm responsible for the system's decision [10]. In the context of machine learning, these decisions often involve discrete labels or classes (or, in regression tasks, numeric values). While explanations centred on the algorithm's internal logic can be valuable to ML researchers, they might be less practical for end-users who are more concerned with how they can alter their current circumstances to achieve a desired outcome in the future. This highlights the need for explainable AI methods that focus on identifying relationships among input dependencies that lead to the system's decision. [10]

According to Spreitzer et al., a counterfactual explanation suggests the smallest alterations to the input data resulting in a new model outcome, addressing the fundamental query: "Would altering a specific feature have yielded a different outcome?" [11]. These explanations can be regarded as suggestions for what adjustments to make to attain a preferred model outcome. In the context of this research paper, it is important to know what minimal factors can be changed in order to change the outcome grade of a student from a fail to a pass. Although Spreitzer et al. did not apply counterfactual explanations to the education field, the fundamentals of counterfactuals still apply to this research paper. The counterfactual explanations for this research paper are limited to logical, actionable insights.

Barocas et al. demonstrated that the computation of counterfactual explanations often rests upon assumptions that are frequently overlooked but are crucial for these explanations to be applicable in real-life scenarios [12]. The primary issue is that features are not independent, as actions are likely to simultaneously impact other features. Barocas et al. pointed out that many generated counterfactual explanations rely solely on the distribution of training data, neglecting the fact that not everything considered equivalent in the training data necessarily holds true for individual data instances [12]. This meant that specific counterfactual explanations may be practical for some individuals but not necessarily suitable for others. This is a vital consideration in this research paper when applying counterfactual explanations with the intent of improving student academic performance as interventions for one student may not be applicable to another student.

2. Methodology

The raw dataset, obtained from the UC Irvine Machine Learning Repository, focuses on assessing student academic performance in mathematics in secondary education within two Portuguese schools. For context, the secondary education system in Portugal consists of grades 10 to 12. The dataset encompassed a range of features, including student grades, demographics, social factors, and school-related features. It is worth noting that for the purposes of this research the sex of a student referred to the sex documented on their birth certificate. Data collection was carried out through the utilization of surveys and school reports [13].

Jupyter Notebook was utilised to code, debug and visualise the developed algorithms. Jupyter was used due to its interactive and versatile nature. It allows users to write and execute code in a cell-based format, making it easy to experiment, iterate, and troubleshoot code step by step [14]. Python libraries, namely numPy, pandas and scikit-learn, were utilised for techniques such as data manipulation, model training and evaluation, and data visualisation. These libraries were key for data-driven research.

Before implementing machine learning algorithms, the dataset was pre-processed since the raw dataset may have contained errors, inconsistencies, or missing values. Data pre-processing helped identify and address these issues, ensuring that the data used for training and testing the machine learning models was accurate and reliable. A well-developed machine learning model performing on an erroneous dataset will result in inconsistent predictions. Consequently, inconsistent predictions will result in flawed interpretations and counterfactual explanations of the machine learning models.

The associated task for this research was a regression analysis. This was because the target outcome was a numerical value depicting students' final grade for mathematics. Six different regression machine learning models were trained on 80% of the dataset and tested on the remainder of 20%. This split ratio allowed for the evaluation of the models' performance on unseen data and prevention of overfitting. Testing the models on an independent data subset ensured that the models generalize well and provide a benchmark for assessing its reliability and suitability for completing the objectives of this research. The models' hyperparameters were tuned to maximise model performance and ensure robustness, resulting in more accurate and effective machine learning models. The accuracy of all developed models was compared and ranked.

The top three best-performing models were used for the explainability and counterfactual segment of the research. Explainability techniques, including global, grouped and local, were applied to these models to investigate the most influential factor that contributed to students' final grade. This was done to unravel the black-box nature of the machine learning models by providing insights into their decision-making processes behind student grades.

The local explainability technique formed the basis of the counterfactual explanations as it provided the dominant features that affected academic performance for each student. To ensure reliable results, the overlapping explanations from all models were considered as shown in Figure 1.

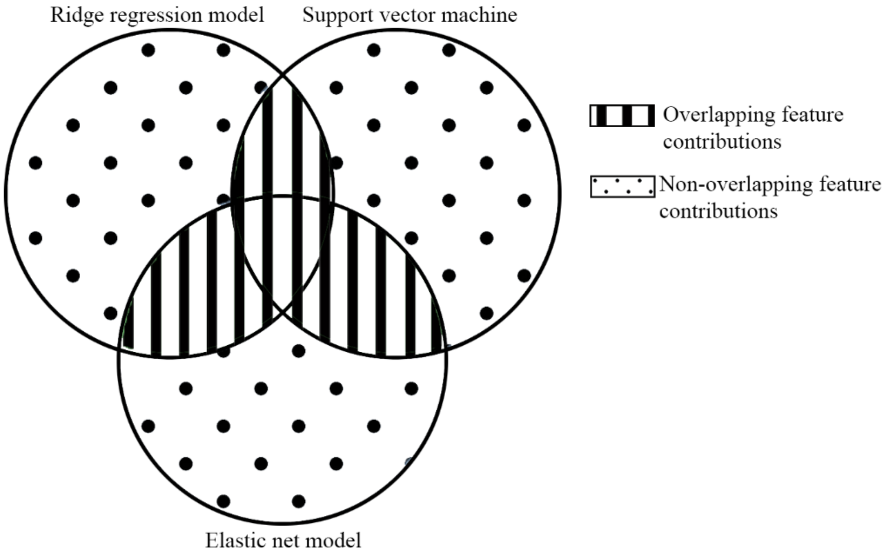

Ridge regression model
Support vector machine
Overlapping feature contributions
Non-overlapping feature contributions
Elastic net model

Figure 1 Depiction of overlapping and non-overlapping feature contributions for all models

Model-agnostic methods, such as SHAP and LIME, were favoured over post-hoc methods in this research due to their flexibility, interpretability, and focus on local explanations. These methods accommodated the diverse range of machine learning models used in the study (ridge regression, support vector machine, elastic net) while providing understandable, detailed insights into prediction rationale. Their ability to generate counterfactual explanations, which was crucial for understanding how to improve students' academic performance, was a standout feature. Model-agnostic methods also demonstrated robustness across different data types and problem domains, making them ideal for educational research seeking to enhance student outcomes.

The counterfactual explanations were limited to actionable insights. Non-actionable insights such as age and sex were not considered for the counterfactual explanations. These features were, majority of the time, immutable and did not provide actionable information for educational interventions. Instead, it was more meaningful to focus on features that could be influenced, like study habits or engagement in specific educational programs.

3. Data processing

The raw data consisted of 395 instances. Duplicate instances were removed. This was done to maintain data consistency and quality, prevent biases, and improve the efficiency of the machine learning models. Duplicate instances can introduce errors and bias, affect computational resources, and possibly lead to overfitting for machine learning models. When duplicate instances are present, the model may fit excessively to repeated data. This would have resulted in the model capturing the training dataset's anomalies rather than the genuine patterns, reducing its ability and performance to generalize to new, unseen data.

Several features were removed from the dataset for varying reasons. Table 1 summarised the features that were removed from the raw data coupled with the reason for the removal.

Table 1 Summary of features that were removed and why they were removed

Feature name	Feature description	Reason for removal
School [13]	Student's school [13]	To prevent overfitting, where the model may become overly specialized in school-specific patterns that may not generalize to different datasets.
Famsize [13]	Student's family size [13]	This feature may not directly contribute to the model's goal and was not actionable. Instead, the model's focus can be directed toward features closely tied to academic performance and intervention, enabling more effective and relevant insights for at-risk students.
Address [13]	Student's home address type [13]	The distinction between rural and urban home addresses may not have a direct, meaningful impact on a student's academic performance. Including this feature in the model could introduce unnecessary complexity without significantly contributing to the model's explanatory or predictive power. The requirement for a student to change their home address was considered non- actionable.
Medu [13]	Student's mother's education [13]	A student's mother's education level may not have a direct impact on academic performance or the risk of subject failure. Focusing on academic, behavioural, and demographic factors that directly affect student performance is more relevant and actionable for the model's goal. Removing this feature also addressed privacy concerns from an ethics perspective.
Fedu [13]	Student's father's education [13]	A student's father's education level may not have a direct impact on academic performance or the risk of subject failure.

		Focusing on academic, behavioural, and demographic factors that directly affect student performance is more relevant and actionable for the model's goal. Removing this feature also addressed privacy concerns from an ethics perspective.
Reason [13]	Reason the student chose this school [13]	This feature was considered as non-actionable. It cannot be altered for a counterfactual explanation as it stated why the student was currently enrolled at the school.
Guardian [13]	Student's guardian [13]	The distinction between a student's mother and father as their guardian does not have a direct bearing on their risk of failing a subject. Including this feature may have introduced model biases.
Failures [13]	Number of past class failures [13]	This was a non-actionable feature. It may have also introduced biases due to the model developing a relationship between the student's risk of failing an independent subject and the number of past class failures.
Higher [13]	Depiction of whether the student wants to purse a higher education [13]	This may not have had an influence on the academic outcome of a student. A student may struggle with academic material but still possess the intention to pursue higher education.
Romantic [13]	Depiction of whether the student is in a romantic relationship [13]	Some students may have not fully disclosed their relationship status due to privacy concerns. This feature may also have no direct bearing on the risk of failure for a student. A student may be in a healthy, supportive relationship but struggle with subject material independently.
Famrel [13]	Student's quality of family relationships [13]	Although this feature was captured from surveys, it is subjective and varies for different students. Hence, the feature may not reflect a true depiction of the students' quality of family relationships.
Health [13]	Student's current health status [13]	A student's health status does not always affect their academic performance. Some students may be healthy but still have poor academic performance and vice versa. This feature is partially non-actionable as some students may not be able to change their health status. For example, a student with influenza can consume medication to fully recover, but a student with stage 4 cancer cannot do the same.

As part of pre-processing the dataset, the numerical features were normalised while the categorical features were converted into numerical values. Normalisation of features equalized the scales of features, ensuring that no single feature dominated the machine learning process. This leads to faster convergence during model training, and potentially improving model performance. [15]

The conversion of categorical features represented categories in a numerical format, which was a requirement for the machine learning algorithms. This was crucial in machine learning as it allowed algorithms to process and derive insights from these converted features effectively. By transforming categorical data into a numerical format, the algorithm was prevented from mistakenly concluding ordinal relationships or biases among categories, promoting fairness and reducing errors. This ensured reliability when training machine learning models on the dataset.

Normalisation of features was done using the *MinMaxScaler* class from the scikit-learn library. The normalised features were scaled to between 0 and 1 to ensure that features from the raw data with different scales did not disproportionately influence the behaviour of the machine learning algorithms. The effect of data normalisation was visualised in Figure 2.

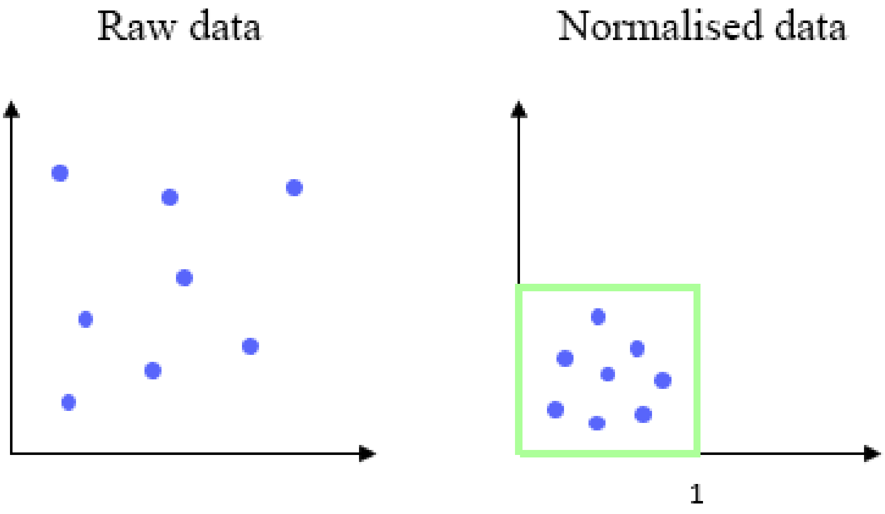
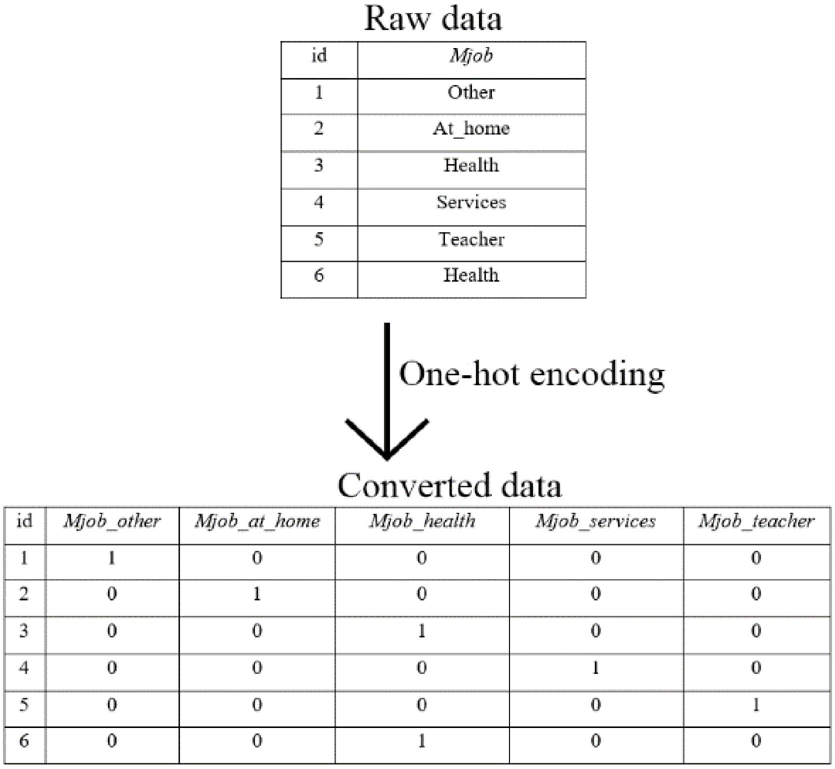


Figure 2 Visualisation of the effect of data normalisation

The *get_dummies* function from the Pandas library was utilised to convert categorical features into numerical values using one-hot encoding. One-hot encoding was used because it preserves all information within categorical variables by creating binary columns for each category, ensuring no ordinal relationships are assumed [16]. This avoided biases and inaccuracies introduced by other encoding methods. The transparency of one-hot encoding, with human-interpretable labels, enhances model interpretability and eases model results explanation. A schematic of how one-hot encoding was used was shown in Figure 3.



Raw data

id	*Mjob*
1	Other
2	At_home
3	Health
4	Services
5	Teacher
6	Health

One-hot encoding

Converted data

id	*Mjob_other*	*Mjob_at_home*	*Mjob_health*	*Mjob_services*	*Mjob_teacher*
1	1	0	0	0	0
2	0	1	0	0	0
3	0	0	1	0	0
4	0	0	0	1	0
5	0	0	0	0	1
6	0	0	1	0	0

Figure 3 Depiction of how one-hot encoding was used for pre-processing the data

The data was separated into the input features and the target outcome, *G1*. After removing duplicates and applying one-hot encoding, there were 34 input features. This was only two more than the initial 32 input features from the raw data. The *train_test_split* function from the scikit-learn library was used to split the data into train and test subsets. 80% of data was allocated for training while the remaining 20% was used for evaluating the model. The 80:20 ratio is commonly used in the machine learning industry and provides a good balance between training and testing data, which aids in reducing the likelihood of overfitting and underfitting [17]. The *random_state* parameter was utilised to ensure reproducibility. Since a random state was specified, the data splitting process always produced the same random split when executed multiple times. This was essential for consistent and reproducible results which essentially drives consistent interpretations of the machine learning models.

Several regression models, namely ridge regression, lasso regression, random forest, support vector regression, elastic net, and k-nearest neighbours, were trained and subjected to hyperparameter tuning to identify the most effective models in terms of predictive performance. To optimise the models using hyperparameter tuning, the *RandomizedSearchCV* class was used from the scikit-learn library. This class was used due to its significantly reduced computation time, ability to conduct a broad exploration of the hyperparameter space, and its likelihood of finding optimal hyperparameters. It is particularly practical in scenarios where computational resources are limited, as it efficiently samples hyperparameter combinations at random [18]. The element of randomness also mitigates overfitting to the training data and encourages model generalization to new data. A schematic of how the *RandomizedSearchCV* class works in comparison to a grid search was shown in Figure 4.

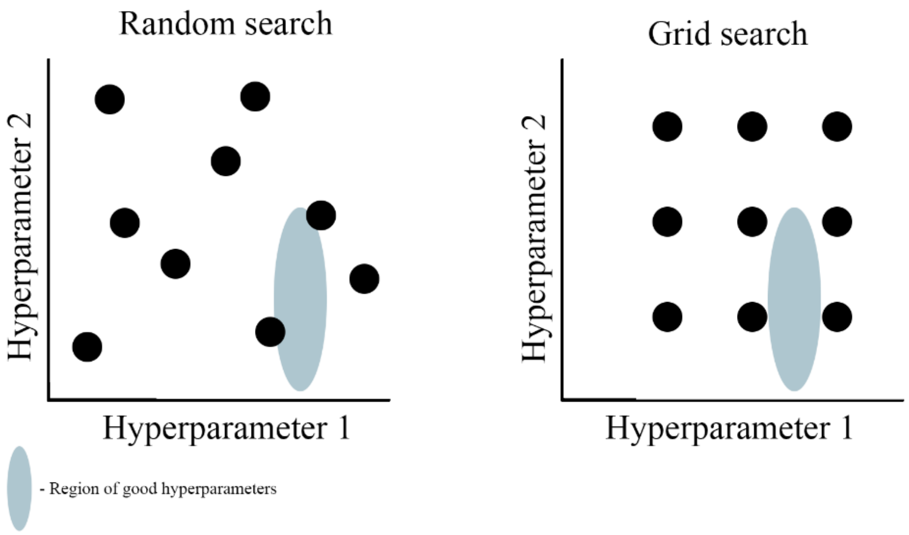


Figure 4 Comparison of random search and grid search for hyperparameter tuning

The predictive performance for each model was calculated using the mean squared error obtained from cross-validation. Cross-validation was used for evaluating model mean squared errors due to its ability to provide a more reliable and comprehensive assessment of a model's performance on unseen data. By systematically partitioning the dataset into multiple folds, it reduces the variance in performance estimates and ensures that all data points contribute to the evaluation [19]. It is an essential tool in machine learning for robust and informed model evaluation and comparison. Figure 5 shows a comparison of the mean squared errors for the models.

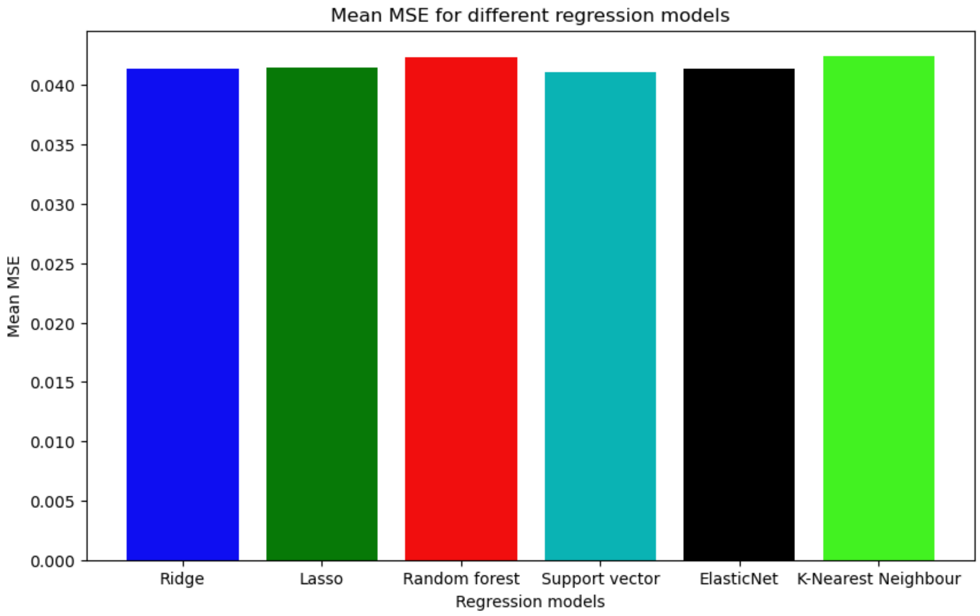


Figure 5 Comparison of mean squared errors for the regression models to evaluate performance

The best-performing algorithms, obtaining the lowest mean squared errors, were ridge regression, support vector machine, and elastic net models.

SHAP was applied to the three best-performing models for global and grouped interpretability. SHAP was used because it provided insights into individual feature contributions and global feature importance. It promotes model transparency and accountability by quantifying the impact of each feature on predictions [20]. For global interpretability, SHAP was applied to the entire dataset for all three models to determine which features dominate the models' machine learning process. This provided an average explanation for the dataset. For grouped interpretability, SHAP was applied to all three models but only students that were at risk of failing mathematics. The criteria for this were students with final grades, *G1*, between 40% and 49% (normalised values between 0.3125 and 0.4375). This was a good range to specify students that were at risk of failing the subject because it was close to the pass mark of 50%. Students in this range were in danger of failing the module if they did not make some counterfactual change to alter their final grade. Grouped interpretability was done to obtain a granular understanding of how features affect the final grade of students that were at risk of failing mathematics. This was in line with the aim of this research and provided an average explanation for at-risk students.

LIME was applied to all three models for scenarios where students were at risk of failing. This was primarily done to obtain a local explanation of feature contributions to specific students' final grade. LIME was used because it provided tailored explanations for individual predictions, making it a powerful tool for understanding why each model made a specific decision for a particular student. LIME's model-agnostic nature allowed it to be applied universally across various machine learning models, promoting transparency. Overlapping feature contributions for the three models were examined and visualised to obtain meaningful, actionable insights and counterfactual explanations.

4. Results and discussion

4.1 Global interpretability

On a global level, the overlapping feature contributions were obtained and analysed. This was done for various reasons. When comparing overlapping feature contributions, the consistency of feature importance was checked across the three machine learning models. If a feature consistently appeared in the summary plots for multiple models, it added confidence to the interpretation and suggested that the feature was dominant and influential in contributing to the models' predictions. Overlapping features were a common ground for model-agnostic insights. Since they appeared consistently across machine learning models, the importance of these features was understood and communicated without reliance on the specific machine learning modelling algorithms used.

SHAP analysis was conducted on the three machine learning models to achieve an average explanation of all students. A SHAP summary plot, displaying the top five feature contributions, was generated for each model. Figures 6 to 8 show the global interpretation SHAP summary plots for each model.

There were six features, namely *Mjob_other; studytime; goout; Fjob_other; Fjob_services; schoolsup_no*, that appeared more than once across the three models. However, only the top three overlapping features were considered. The selection was based on the SHAP values to establish the top three dominant features since a larger absolute SHAP value had a stronger impact on the prediction. The top three overlapping features were *Mjob_other*, *studytime*, and *goout*. The fact that these three features consistently appeared in various models demonstrated their consistency and robustness in influencing predictions. This suggested that they were not model-specific but played a fundamental role in the prediction of students' final grades, *G1*. The contribution of these features was geared toward the entire dataset as it entailed global interpretability. These features exhibited a high degree of predictive power globally. Their consistent impact on model predictions implied that changes in these features had a substantial influence on the outcomes, making them crucial factors in understanding and improving predictions.

The *Mjob_other* feature was created due to one-hot encoding. The feature was a binary type with a value of 1 indicating that the student's mother had a job that was not a teacher, healthcare related, civil services or at home. A value of 0 indicated that the student's mother had a job that was categorised as 'other'.

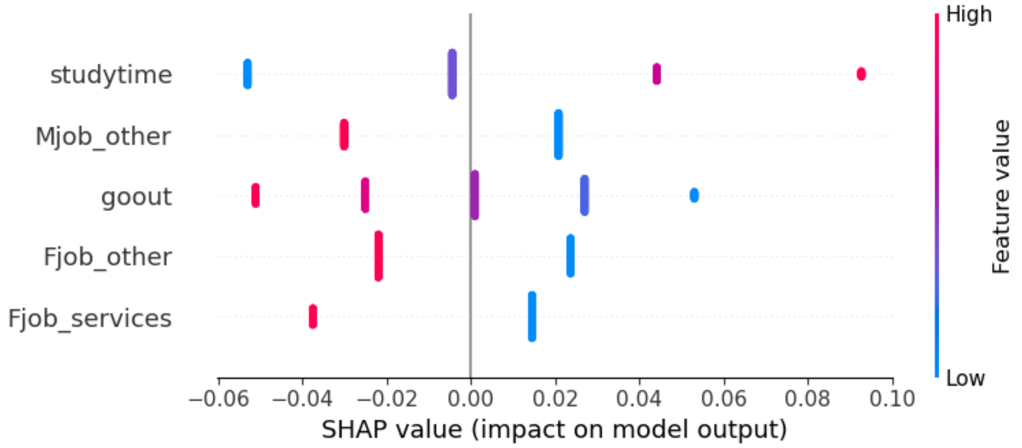


Figure 6 SHAP summary plot for ridge regression model applied globally

Figure 6 showed that the *Mjob_other* feature had a maximum contribution of 2.1% for a low feature value. However, this was an inversely proportional relationship. Since the feature was a binary type a low feature value meant that the value was 0, indicating that the student's mother did not have a job categorised as 'other'. According to Figure 6, when the students' mother had a job not categorised as 'other', it contributed 2.1% toward increasing their final grade. On the other hand, when the students' mother did have a job categorised as 'other', it contributed 3% toward decreasing their final grade. The implication of this relationship was that the students' mothers' job had an effect on the students' final mathematics grade. The nature of the jobs categorized as 'other' may have involved long hours, irregular schedules, or high-stress levels, which could have negatively impacted the students' academic performance. Jobs not categorized as 'other' may have been more conducive to a stable home environment and support for the students' education. These were only possible factors and may have not necessarily been the reason for the relationships developed. It is important to note that these results suggested a correlation between the students' mothers' job category and students' final mathematics grade, but they did not necessarily imply causation. Other unobserved factors may have been at play. The concept of causation was not in the scope of this research.

The *goout* feature stated the frequency that the students' go out with friends. From the SHAP summary plot in Figure 6, the ridge regression model developed a correlation between the *goout* feature and the target outcome, *G1*. A high *goout* value was correlated to a maximum feature contribution of 5.1% toward decreasing the students' final grade. On the contrary, a low *goout* value contributed a maximum of 5.3% toward increasing the students' final grade. This indicated that students' social activities may have had an impact on their final grade. Students that go out with friends more frequently had a lower final grade than those that go out with friends less frequently. There were many implications that made this correlation logical. Students who go out frequently may spend a significant amount of their time socializing with friends or engaging in extracurricular activities. This leaves them with less time for studying and academic commitments. In contrast, students who prioritize their studies and spend more time on academic tasks at the sacrifice of socializing and going out with friends may be likely to obtain higher grades. According to the data, achieving a balance between social life and academics was essential for student success. Going out with friends and socializing is a valuable aspect of personal development, but when it is excessive, it may lead to neglect of academic responsibilities [21]. However, it must be noted that every student is unique, and individual differences in study habits, learning styles, and time management play a significant role. Some students may thrive academically despite frequent social outings, while others may need more focused study time. The global interpretability method did not account for this because it was an average explanation of the entire dataset.

From Figure 6, the SHAP summary plot based on the ridge regression model showed that the *studytime* feature had a maximum contribution of 9.2% for a high feature value. This meant that a high *studytime* value (which stated the number of weekly study hours) contributed a maximum of 9.2% toward increasing the outcome, *G1*. Conversely, as seen in Figure 6, a low *studytime* value contributed a maximum 5.3% toward decreasing the outcome. This interpretation was logical and showed that students with an increased number of study hours subsequently obtained higher grades. This relationship developed by the ridge regression model was rational since, in general, students that study for more hours usually attain higher grades and vice versa.

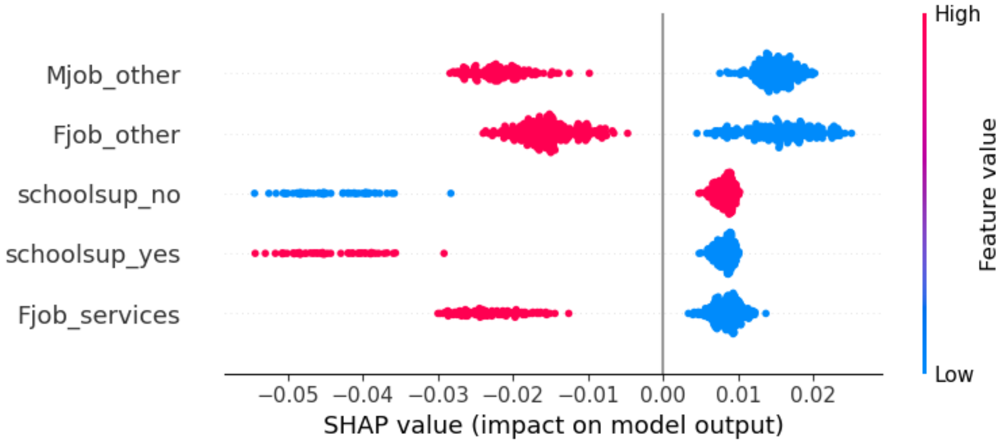


Figure 7 SHAP summary plot for support vector model applied globally

The SHAP summary plot, generated from the support vector machine, in Figure 7 showed a similar relationship to the ridge regression model for the *Mjob_other* feature. When the students’ mother had a job that was not categorised as ‘other’, the feature contribution varied between 0.8% and 1.9% toward increasing the students’ final grade. This distribution was seen on Figure 7 with the majority of low feature values at a contribution of 1.3%. When the students’ mother had a job that was categorised as ‘other’, the feature contribution varied between 1.1% and 2.8% toward decreasing the students’ final grade. This correlation was similar to that developed by the ridge regression model. However, the correlation generated by the support vector machine had a wider distribution. The reason for this difference could have been due to the fact that ridge regression employed L2 regularisation, which encouraged a balance between feature coefficients, preventing any single feature from dominating the predictions [22]. This may have tended to produce more consistent and stable feature contributions across instances, resulting in a narrower SHAP value distribution. After tuning the hyperparameters for the support vector machine, the kernel that was utilised was RBF (Radial Basis Function). The choice of this non-linear kernel in the support vector machine may have led to more complex and non-linear relationships between features and the prediction. Hence, the influence of the *Mjob_other* feature on the final grade may not have been linear or uniform, resulting in a wider distribution of SHAP values.

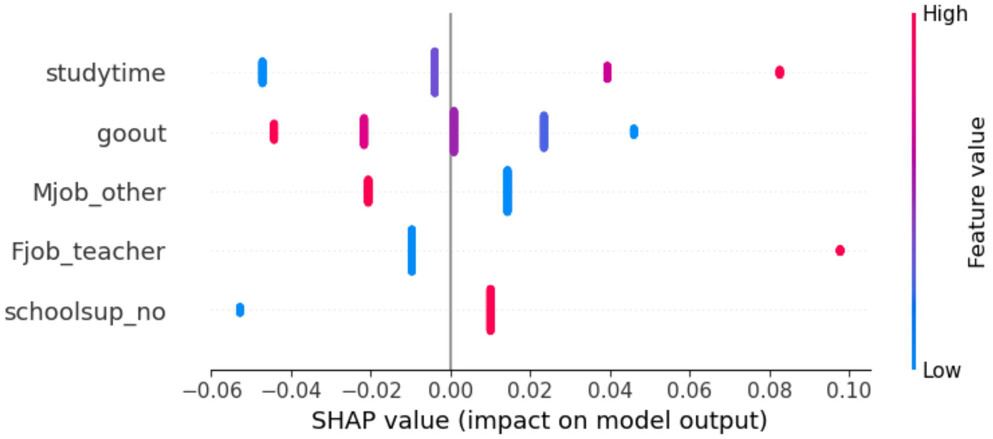


Figure 8 SHAP summary plot for elastic net model applied globally

The SHAP summary plot from Figure 8 showed the feature contributions for the elastic net model. When the students' mother had a job that was categorised as 'other', the feature contribution was 1.9% toward decreasing the students' final grade. The SHAP values were not distributed as seen in Figure 7 for the support vector machine. When the students' mother had a job that was not categorised as 'other', the feature contribution for *Mjob_other* was 1.8%. This showed that the feature was more dominant in reducing the students' final grade since it contributed 1.9% in doing so. The reason the SHAP values were not widely distributed as in the support vector machine was likely because the elastic net model was a linear regression technique that combined L1 (Lasso) and L2 (Ridge) regularization. These regularization terms add a penalty to the loss function, encouraging sparsity (L1) and controlling the magnitude of coefficients (L2) [23]. This regularization effect often leads to more consistent and stable feature contributions, resulting in a narrower distribution of SHAP values similar to that seen in Figure 6.

The elastic net model from Figure 8 generated a similar correlation between the *goout* feature and the target outcome as the ridge regression model. A high *goout* value was linked to a maximum feature contribution of 4.5% toward decreasing the students' final grade. A low *goout* value contributed a maximum of 4.7% toward increasing the students' final grade. This correlation was analogous to the relationship developed by the ridge regression model. Therefore, it was logical as previously mentioned. Although there were similar relationships for the *goout* feature developed by the ridge regression and elastic net models, the support vector machine did not did not generate a relationship for the feature in the top five contributions. This may be due to the fact that ridge regression, elastic net and support vector machines are different modelling techniques with distinct approaches to capturing relationships between features and the target outcome. Ridge regression and elastic net are linear models that seeks to fit a linear relationship between features and the target outcome, while support vector machines aim to find the optimal hyperplane that maximally separates data points into different classes [24]. This inherent modeling dissimilarity may have led to variations in feature importance rankings, as seen on the SHAP summary plots. The complexity of the model can also play a role. Ridge regression is a relatively simple linear model, while support vector machines can have varying degrees of complexity depending on the choice of kernel and hyperparameters [24]. More complex models may exhibit more intricate feature relationships.

The SHAP summary plot from Figure 8, based on the elastic net model, showed that the *studytime* feature had a maximum contribution of 9.2% for a high feature value. This correlation was similar to that developed from the ridge regression model – on average, students that put in more weekly study hours obtain higher final grades. From Figure 8, low *studytime* values contributed a maximum of 3.8% toward decreasing the final grades. Hence, the elastic net model developed a logical relationship between the students' weekly study hours and their final grade, *G1*. Despite the relatively strong feature contributions for the *studytime* feature created by the ridge regression and elastic net models, the support vector machine's top five feature contributions did not include the *studytime* feature. As mentioned previously, it may be due to the variances in model complexities and inherent modelling differences.

Despite the varying distribution of SHAP values across the three models for the *Mjob_other* feature, the correlations between *Mjob_other* and the target outcome, *G1* were consistent. Consistency in these SHAP correlations was likely a sign of robust and reliable patterns in the dataset. It suggested that the importance of these features relative to each other was globally dependable and not influenced by specific modelling algorithms. This was a desirable quality for model-agnostic insights.

The *studytime* feature did not account for 0 or no study hours. Hence, the correlations that were generated by the models for *studytime* may have not necessarily been true. Due to the *studytime* feature not accounting for 0 study hours, it meant that students who study for less hours tended to have lower grades. Logically, this meant there was a threshold of weekly study hours required in order for the student to actually increase their final grade. As a result, students that did not study at all during the week were not accounted for.

As a global interpretation method, a counterfactual explanation based on altering these features will likely result in a change in predicted outcome overall. However, this is an average interpretation of all students in the dataset. A counterfactual explanation based on global interpretability is sufficient for an average explanation but may not be applicable to individual students. Global interpretability techniques provide general rules that apply to the entire dataset. Counterfactuals, on the other hand, aim to find specific scenarios where the outcome can be changed significantly. Hence, counterfactuals based on global interpretability methods will likely not maximise the change in predicted outcome. Although global interpretability does not provide the level of granularity needed to achieve this, it was still valuable to identify an overview of the feature contributions to the students' final grade target outcome.

4.2 Grouped interpretability

From a grouped perspective, the overlapping feature contributions were captured and analysed. Identifying features that consistently appeared as important across different models enhanced the robustness and reliability of the insights. If multiple models, each with its unique approach and assumptions, highlighted the same features on a grouped level, it strengthened the case for the importance and contribution of those features. This was useful on a grouped level because the models' unravelled relationships between features and the target outcome for at-risk students only. The aim of this research was linked to at-risk students and how they could improve their academic position to alter their predicted outcome. Hence, the grouped interpretation obtained from this was used to attain an average explanation that applied only to students at risk of failing and revealed factors that were common amongst at-risk students. Overlapping features helped identify common characteristics and challenges shared by this specific group of students. It highlighted the aspects of their academic performance were universally affected, regardless of the modeling approach. When examining overlapping features, focus was placed on the features that consistently and collectively influenced the group of students at risk of failing the subject. This provided a more holistic understanding of the factors contributing to their academic challenges. Grouped interpretability with overlapping features helped in making data-driven decisions about where to allocate resources for supporting at-risk students and improving overall educational outcomes. This was directly related to the aim of this research to establish counterfactual explanations required to improve the academic performance of students that were at risk of failing.

Additional SHAP summary plots were generated for grouped interpretability. This provided an overview interpretation of only students that were at risk of failing. Figures 9 to 11 show the grouped interpretation SHAP summary plots for each model.

All features from each of the three models appeared more than once across all models. However, only the top three overlapping features were considered. In the same way as for global interpretability, the selection was based on the SHAP values to obtain the top three contributing features since a larger absolute SHAP value had a stronger impact on the model prediction. The top three contributing overlapping features were *Mjob_other*, *studytime*, and *Fjob_other*. The consistent presence of these features across the three models indicated that they were likely fundamental to understanding the academic performance and risk of failure for at-risk students. These features represented core elements that affected the group universally.

The features *Mjob_other* and *studytime* both appeared as top contributors for global and grouped interpretability. This suggested that these two features had a strong correlation with the target outcome on a global level as well as on a grouped level applied only to students at risk of failing. Just as the *Mjob_other* feature, t*he Fjob_other* feature was created due to one-hot encoding. The feature was a binary type with a value of 1 indicating that the student's father had a job that was categorised as 'other'. A value of 0 indicated that the student's father did not have a job that was categorised as 'other'. The remaining categories were the same as that for *Mjob_other* – teacher; healthcare related; civil services; at home.

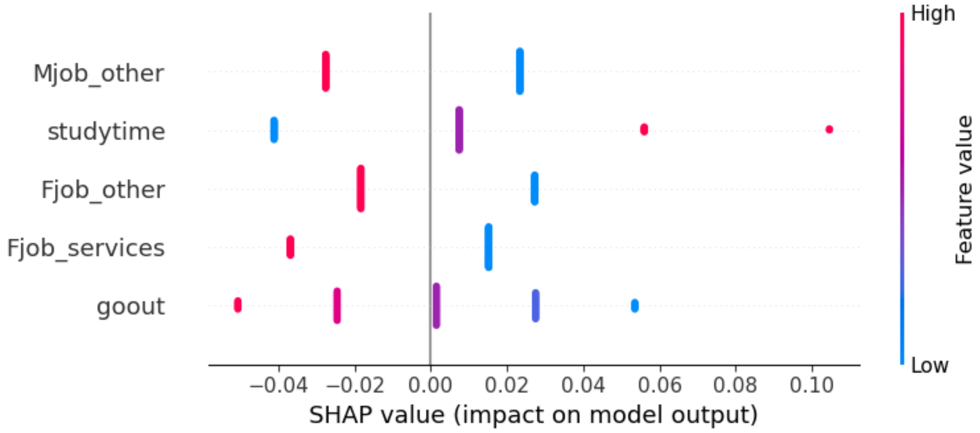


Figure 9 SHAP summary plot for ridge regression model applied only to at-risk students

From the SHAP summary plot for the ridge regression model in Figure 9, when the students' mother had a job that was not categorised as 'other' it contributed 2.3% toward increasing the students' final grade. Conversely, when the students' mother did have a job categorised as 'other' it contributed 2.9% toward decreasing the students' final grade. These values were close to that obtained for global interpretability by the ridge regression model. Hence, it can be deduced that for the ridge regression model the feature contribution of *Mjob_other* held high predictive value both globally and grouped.

From Figure 9, the grouped interpretability SHAP summary plot showed that the *studytime* feature had a maximum contribution of 10.2% for a high feature value. Hence, an at-risk student that put in more weekly study hours contributed a maximum of 10.2% toward increasing their final grade. An at-risk student that put in less weekly study hours contributed a maximum of 4.1% toward decreasing their final grade. Hence, an at-risk student was affected more by higher weekly study hours than less. This correlation was logical and similar to that generated for global interpretability.

The *Fjob_other* feature appeared on the grouped interpretability SHAP summary plots for the ridge regression and support vector machine models. For the ridge regression model in Figure 9, there was a contribution of 2.7% toward increasing the final grade when the at-risk students' father had a job that was not categorised as 'other'. On the contrary, when the at-risk students' father had a job that was categorised as 'other', the feature contributed 1.9% toward decreasing their final grade.

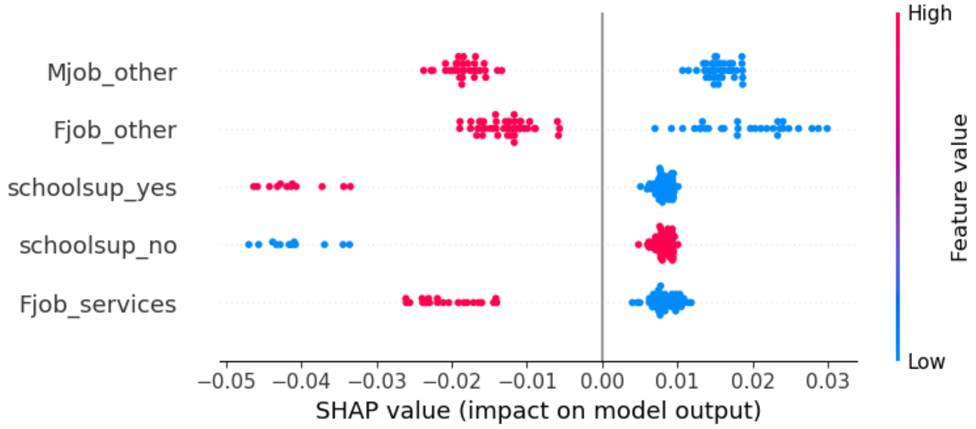


Figure 10 SHAP summary plot for support vector model applied only to at-risk students

The SHAP summary plot from Figure 10 showed the grouped feature contributions for the support vector machine. When the students' mother had a job that was categorised as 'other', the *Mjob_other* feature contributed between 1.3% and 2.4% toward decreasing the students' final grade. Similarly, when the students' mother had a job that was not categorised as 'other', the feature contributed between 1.2% and 1.8% toward increasing the students' final grade. This was contrasting to the global interpretation SHAP summary plot for the support vector machine. For the top five feature contributions, there was no global correlation between *Mjob_other* and the target outcome for the support vector machine. This meant that on a global level the support vector machine did not find a significant relationship between the students' mothers' job and the students' final grade. However, this relationship was significant when looking at only students that were at risk of failing. Therefore, for the support vector machine only students that were at risk of failing had final grades that were affected by the students' mothers' job. The support vector machine's feature importance may not have prioritized the *Mjob_other* feature in the global model, possibly due to the presence of other features that overshadowed its contribution. However, within the at-risk student group, this feature might have gained importance.

From the SHAP summary plot in Figure 10, the support vector machine generated a correlation between the *Fjob_other* feature and the target outcome, *G1*. When the at-risk students' fathers' job was not categorised as 'other', the feature contributed between 0.7% and 2.8% toward increasing the students' final grade. On the other hand, when at-risk students' fathers' job was categorised as 'other', the feature contributed between 0.6% and 1.9% toward decreasing the students' final grade. The support vector machine had a wide distribution of SHAP values for the feature contribution of *Fjob_other* while for the ridge regression there was no distribution. This was due to the non-linear, RBF kernel used for the support vector machine while the ridge regression was a linear algorithm.

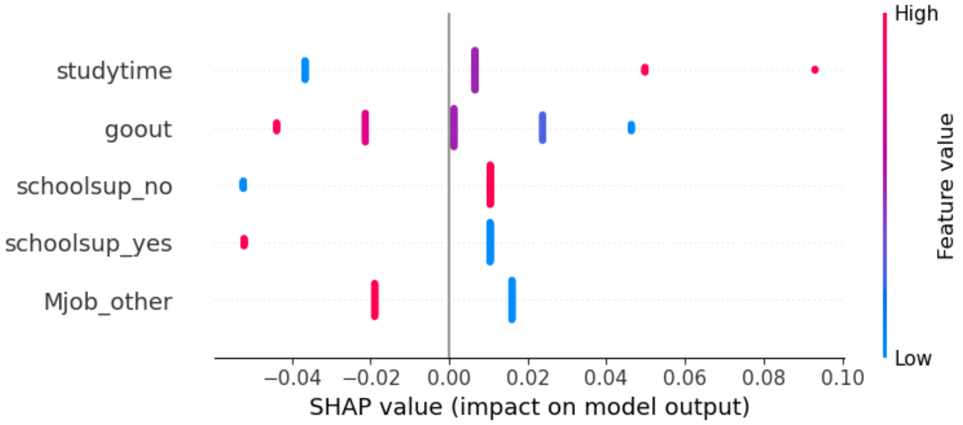


Figure 11 SHAP summary plot for elastic net model applied only to at-risk students

From Figure 11, the *Mjob_other* feature had a contribution of 1.9% toward decreasing the students' final grade when the students' mothers' job was categorised as 'other'. The feature contributed 1.7% toward increasing the outcome when the students' mothers' job was not categorised as 'other'. These feature contribution SHAP values were near to that obtained for global interpretability by the elastic net model. On a grouped level, the *Mjob_other* feature held high predictive value since it was a top contributor for the ridge regression, elastic net and support vector machine when analysing the group of at-risk students.

The SHAP summary plot from Figure 11 showed that the studytime feature had a maximum contribution of 9.3% for a high feature value. Low studytime values contributed a maximum of 3.8% toward decreasing the at-risk students' final grades. Hence, the elastic net model developed a logical relationship between the at-risk students' weekly study hours and their final grade, *G1*. There was no significant relationship developed by the support vector machine for the *studytime* feature for grouped interpretability. However, this was not alarming as the model also did not generate a correlation for the feature for global interpretability. The reason for this could therefore be attributed to the support vector model complexities.

The inference drawn from these correlations was that the occupation of the at-risk students' fathers' job had an influence on their final grades. In a similar way to the *Mjob_other* feature, occupations classified as 'other' might have involved factors such as extended working hours or high stress levels, which could potentially have had a detrimental effect on the academic performance of at-risk students. Occupations falling outside the 'other' category, such as teacher and civil services, may have provided a more stable home environment and better support for at-risk students' education. It is important to recognize that these are plausible factors but not necessarily the sole reasons for the observed relationships. It is crucial to emphasize that these results indicated a correlation between the category of the students' fathers' jobs and their final grades, but they do not establish a causal relationship as with the *Mjob_other* feature. Tables 1 and 2, available in Appendix A, summarise the top three dominant features with their contributions amongst the best-performing models for global and grouped interpretability, respectively. This was done to supportively illustrate the similarity in maximum feature contribution values across the three models.

4.3 Local Interpretability

LIME was applied to each model for local interpretability. Since LIME was a local method, it was applied to a random student that was at risk of failing to understand why that student was at risk of failing. A LIME explanation provided the feature contributions for a specific instance. The LIME explanations for each model were summarised and compiled into bar graphs to compare the feature contributions for each model based on the randomly selected at-risk student.

Local interpretability was particularly important when it came to counterfactual explanations because it provided fine-grained insights into the behaviour and decision-making of machine learning models for individual students. This provided students with counterfactual explanations that had the potential to maximise the academic performance of each student individually. Local interpretability enables the understanding of how a model arrived at a particular prediction for that student. It highlighted the features and their impact for that individual, which was essential for tailoring interventions through the use of counterfactual explanations.

Counterfactual explanations often involve making small changes to input features to achieve a different prediction, in this case it was ensuring that a previously at-risk student was not at risk anymore. Local interpretability helped assess the sensitivity of the model's predictions for that specific student. It revealed which features were most influential in driving the prediction change. While global and grouped interpretability offered valuable insights at a higher level, they did not capture the nuances of individual students. They offered, at best, average explanations of all students in the dataset. Local interpretability complemented these higher-level insights by addressing specific, case-by-case considerations.

The dataset used for this research consisted of 395 students. From the 395 students, there were 72 students that faced the risk of failure since their final grade was between 40% and 49% (a normalised value between 0.3125 and 0.4375). Due to the number of at-risk students and the time constraints to conduct this research, it was not possible to investigate and provide local counterfactual explanations for all students that were at risk of failing. Hence, only one randomly chosen at-risk student was investigated from the perspective of each of the three machine learning models.

The randomly chosen student was at index 202 in the dataset. The student was at risk because their predicted final grade by all three models were between 40% and 49%. This showed credibility in the outcomes as all predictions resulted in the student being at risk of failure. The prediction values were shown in Figure 12.

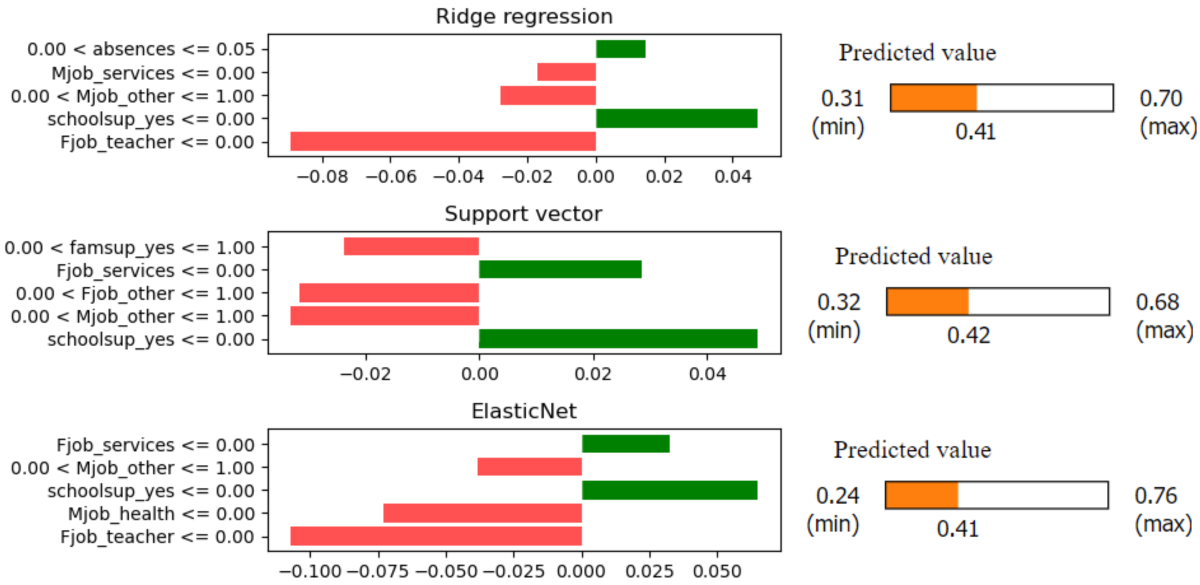


Figure 12 Comparison of LIME explanations of each model for an at-risk student

From Figure 12, it was seen that most feature contributions for the ridge regression model tended to drive the prediction down. The biggest feature contributor was *Fjob_teacher* – a binary feature that stated whether or not the student’s father was a teacher by profession. This feature contributed approximately 8.4% toward decreasing the student’s final grade because the student’s father was not a teacher. This correlation may be logical since teachers generally have a strong understanding of the educational system and may be more actively engaged in their child's academic development. They can provide valuable support, guidance, and resources that contribute to a student's success. Since this binary feature was the biggest contributor, a counterfactual explanation was generated to increase the student’s final grade. This was done by altering the student’s father’s job in the dataset to a teacher (by assigning the *Fjob_teacher* feature to a value of 1). The counterfactual result was shown in Figure 13.

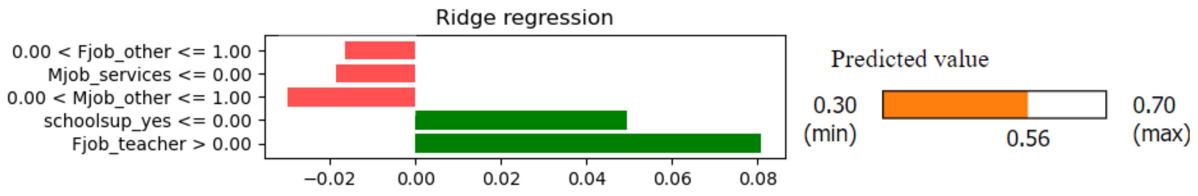


Figure 13 Counterfactual result and prediction for the ridge regression model

After applying the counterfactual explanation, the students' normalised predicted final grade was greater than 0.4375 and no longer at risk of failure. The feature now contributed 8% toward increasing the student's final grade. The logic of whether the *Fjob_teacher* feature was actionable depended on various factors. Since the actionability of the feature varied, the feature was considered to be partially actionable. The counterfactual explanation assumed a direct causal relationship between the father's job and the student's grade. While it could be a plausible factor, it may not be the sole determinant of academic performance. Additionally, a logical question must be asked: Is it possible for the student's father to change their occupation to that of a teacher? This will depend on the student's father and in some cases might not be within the control of the student or their family.

For the support vector machine in Figure 13, the biggest contributor to the target outcome was the binary feature *Mjob_other*. The effect of this feature on the target outcome was previously discussed. Just as the *Fjob_teacher* feature, this feature was also partially actionable as the actionability was dependent on various factors. The feature's value was changed from 1 to 0 in order to develop a counterfactual explanation. The result was shown in Figure 14.

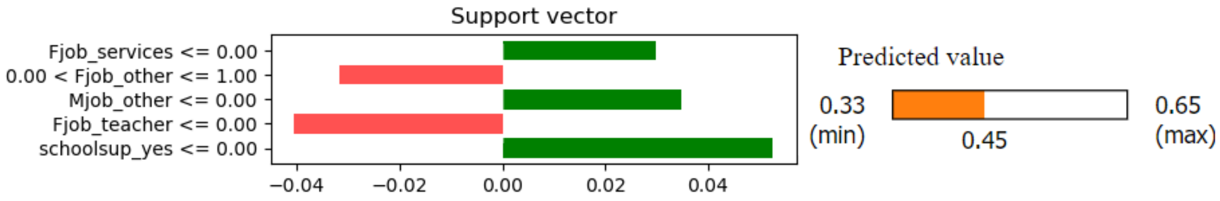


Figure 14 Counterfactual result and prediction for the support vector machine

The student's normalised predicted final grade was greater than 0.4375 and no longer at risk of failure. The feature now contributed approximately 3.5% toward increasing the student's final grade. In a similar way to the *Fjob_teacher* feature, the counterfactual explanation assumed a direct causal relationship between the mother's job and the student's grades. It may not be the exclusive driver of academic performance. Furthermore, this counterfactual assumed that other features in the dataset remained constant. This may not be the case as the student's family and other factors may be affected by the mother's change of occupation.

The biggest feature contributor for the elastic net model was *Fjob_teacher*. This aligned with the model prediction by the ridge regression, possibly due to the fact that both models were linear in nature. This feature contributed approximately 11% toward decreasing the student's final grade because the student’s father was not a teacher. This correlation was reasonable as teachers are well-positioned to offer educational support and guidance to enhance a student's achievement. A counterfactual explanation was generated to increase the student’s final grade by altering the student’s father’s job in the dataset to a teacher. The counterfactual result was shown in Figure 15.

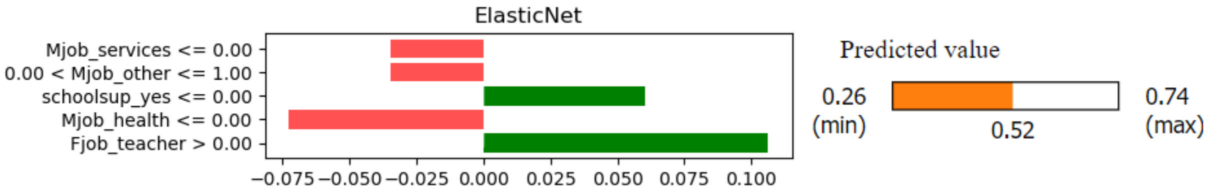


Figure 15 Counterfactual result and prediction for the elastic net model

The counterfactual resulted in the student not being at risk of failure anymore since the normalised predicted grade was 0.52. Due to the student’s father now being a teacher, the student may have enough educational support and guidance to no longer face the risk of failure. Instead of decreasing, this increases the student’s final grade by a contribution of 11%. As mentioned previously, this feature was partially actionable. However, due to the feature being a top contributor across both models it meant that a counterfactual based on this feature will likely increase the student’s final grade due to credibility across two different models.

4.4 Limitations

Global interpretability methods can lead to overgeneralization. They provide insights that apply on average but may not be suitable for individual, highly specific counterfactual scenarios. Hence, maximizing the change in a prediction often requires fine-tuning at the individual student level.

Training the machine learning models on student data has inherent limitations, primarily due to the subjectivity and potential lack of complete reflection in the dataset. The student data contained subjective information, such as qualitative data about their personal experiences. This subjectivity can introduce bias, as students may interpret questions differently, have varying levels of honesty, or provide answers influenced by social desirability bias [25]. This subjectivity could have led to inaccuracies in the training data and production of biased machine learning models.

The student data used for this research was by no means exhaustive. There were only 31 features that were considered. Hence, the data may have not captured the full spectrum of factors influencing student academic performance. It included a finite set of features, such as demographic information and school-related data. Important factors like a student's motivation, family dynamics, mental health, and external stressors were not extensively represented in the dataset.

The data used in this research was at a fixed point in time and may not have accounted for changing circumstances. A student's academic performance could be influenced by evolving factors, such as changes in family dynamics or evolving personal goals. The training data did not capture these changes.

Societal, cultural, and economic factors play a significant role in shaping students' experiences and academic outcomes. These factors can vary significantly from one country to another or even among schools within the same country. Educational systems, teaching methods, family expectations, and socioeconomic conditions can differ widely across countries. Hence, the results obtained may not be directly applicable to students residing in other countries or attending other schools. The limitation is that the results may be only applicable to the dataset that was considered. However, the machine learning model is applicable.

The age of the data, in this case being captured in 2014, presents a significant limitation in the context of understanding student learning and educational strategies. This limitation stemmed from the fact that the field of education is constantly evolving, and the strategies and approaches used by students and educators can change over time. Learning methodologies, curriculums, and teaching techniques evolve to adapt to the needs of each generation of students. What was effective in 2014 may not be the best approach for students in subsequent years. Thus, the data captured in 2014 may not reflect the current and future state of education. Each generation of students may have distinct learning preferences and behaviours. The COVID-19 pandemic, which began in December 2019, disrupted traditional educational practices and accelerated the adoption of remote and online learning. This global event had a profound impact on educational strategies, which was likely not reflected in the 2014 data used for this research.

The machine learning models used can establish correlations, but they do not inherently imply causation. The models highlight relationships between features and student grades, but establishing causal links requires additional research.

The size of the dataset was a significant limitation since it only contained 395 instances. With this small dataset, statistical tests and analyses may have reduced power. This means that it's more challenging to detect statistically significant relationships or effects, even if they exist in the dataset. The results may be less reliable due to the limited dataset size. This small dataset may not capture the full range of variability present in the student population. The limited variability can lead to overfitting, where the model learns to fit the idiosyncrasies of the dataset rather than general patterns.

5. Conclusion

The exploration into explainable machine learning models has revealed the significance of transparent, interpretable predictions in the educational domain. The ability to provide understandable explanations for predictions related to student failure empowers educators and administrators to make informed interventions from counterfactual explanations, fostering a data-informed educational ecosystem. Factors affecting student performance are often countless and complex. The usability of AI through investigations has highlighted the importance of bridging the gap between complex algorithms and actionable insights for educators and students.

The central question that guided the investigation was whether these machine learning models could offer a definitive solution to improving the academic performance of at-risk students. The research first explored the realm of explainable machine learning models, designed to provide interpretable explanations to users regarding predictions for students at the risk of failing. The resounding outcome was clear: machine learning models possess a significant capacity to bridge the gap between complex algorithms and human understanding. By offering transparent, actionable insights, they empower educators and students themselves to make informed decisions that bolster academic success.

The application of counterfactual explanations, which suggests tailored interventions for at-risk students, has proven to be a transformative step in shaping the educational landscape. The holistic evaluation of this research affirms that counterfactual machine learning not only identifies what might have been but also guides the way to effective interventions. It lays the foundation for personalizing academic support, addressing individual needs, and propelling students away from the brink of failure toward the path of success. Although this was advantageous for students, the method did have limitations which were explored. As a result, the use of the counterfactual explanations for at-risk students should be approached with caution and knowledge of the inherent data and model limitations.

It is important to recognize that there is no single solution in education, and the success of these models is contingent on ethical data considerations, data quality, and their successful integration into the educational environment. However, this research provides a definitive answer: The utilization of explainable and counterfactual machine learning techniques for students at risk of failing a subject requires a cautious approach and thorough examination of the findings despite personalised results. Machine learning models, as shown in this research, are not inherently sentient and require human expertise in education, data preparation, bias prevention, and decision-making to be effective in real-world contexts. The collaboration between machine and human intelligence is essential for addressing these challenges.

5.1 Recommendations

The machine learning models used in the research have primarily uncovered correlations between various factors and student academic performance, signifying associations but not causation. To embark on original research, the focus can shift towards exploring causation, seeking to establish the cause-and-effect relationships between specific factors and students' educational outcomes. Researchers can also delve into the effectiveness of educational interventions, the role of mediating and moderating variables, and the ethical considerations surrounding causation studies. Combining machine learning techniques with causal inference can yield new insights, and counterfactual analyses can shed light on how changes in factors would causally impact student performance.

Exploring the potential of hybrid machine learning models in the context of student academic performance offers a promising avenue for original research. Researchers can investigate innovative approaches, such as integrating explainable AI techniques for enhanced model interpretability, transfer learning adapted to educational data, multi-modal learning to harness the power of diverse data types, and hybrid recommendation systems for personalized interventions.

Researchers can consider continuous data to track students' progress and academic performance over time. They can explore how machine learning models can utilize continuous academic data to develop early warning systems and effective intervention strategies. By delving into this area, researchers can not only improve the understanding of students' long-term academic progress but also contribute to the development of practical tools and strategies for supporting at-risk students throughout their educational journeys.

6. References

[1] A. Doctor, “A Predictive Model using Machine Learning Algorithm in Identifying Student’s Probability on Passing Semestral Course,” International Journal of Computing Sciences Research, vol. 7, pp. 1830–1856, Jan. 2023, doi: 10.25147/ijcsr.2017.001.1.135.

[2] A. Singh, “Effect of Co-Curricular Activities on Academic Achievement of Students,” IRA International Journal of Education and Multidisciplinary Studies (ISSN 2455-2526), vol. 6, no. 3, p. 241, Mar. 2017, doi: 10.21013/jems.v6.n3.p4.

[3] M. L. Bernacki, M. M. Chavez, and P. M. Uesbeck, “Predicting achievement and providing support before STEM majors begin to fail,” Computers &#38;amp; Education, vol. 158, p. 103999, Dec. 2020, doi: 10.1016/j.compedu.2020.103999.

[4] F. A. Orji and J. Vassileva, “Machine Learning Approach for Predicting Students Academic Performance and Study Strategies based on their Motivation,” arXiv.org, Oct. 15, 2022. https://arxiv.org/abs/2210.08186

[5] A. M. Shahiri, W. Husain, and N. A. Rashid, “A Review on Predicting Student’s Performance Using Data Mining Techniques,” Procedia Computer Science, vol. 72, pp. 414–422, 2015, doi: 10.1016/j.procs.2015.12.157.

[6] J. Niyogisubizo, L. Liao, E. Nziyumva, E. Murwanashyaka, and P. C. Nshimyumukiza, “Predicting student’s dropout in university classes using two-layer ensemble machine learning approach: A novel stacked generalization,” Computers and Education: Artificial Intelligence, vol. 3, p. 100066, 2022, doi: 10.1016/j.caeai.2022.100066.

[7] T. Miller, “Explanation in artificial intelligence: Insights from the social sciences,” Artificial Intelligence, vol. 267, pp. 1–38, Feb. 2019, doi: 10.1016/j.artint.2018.07.007.

[8] H. Khosravi et al., “Explainable Artificial Intelligence in education,” Computers and Education: Artificial Intelligence, vol. 3, p. 100074, 2022, doi: 10.1016/j.caeai.2022.100074.

[9] A. Salih, Z. Raisi-Estabragh, and P. Radeva, “Commentary on explainable artificial intelligence methods: SHAP and LIME,” May 2023.

[10] N. Wiratunga, A. Wijekoon, I. Nkisi-Orji, K. Martin, C. Palihawadana, and D. Corsar, “DisCERN:Discovering Counterfactual Explanations using Relevance Features from Neighbourhoods,” arXiv.org, Sep. 13, 2021. https://arxiv.org/abs/2109.05800

[11] N. Spreitzer, H. Haned, and I. van der Linden, “Evaluating the Practicality of Counterfactual Explanations,” vol. 3277, no. 3, 2022.

[12] S. Barocas, A. D. Selbst, and M. Raghavan, “The Hidden Assumptions Behind Counterfactual Explanations and Principal Reasons,” arXiv.org, Dec. 10, 2019. https://arxiv.org/abs/1912.04930

[13] P. Cortez, “Student Performance,” UCI Machine Learning Repository, Nov. 26, 2014. https://archive.ics.uci.edu/dataset/320/student+performance (accessed Oct. 16, 2023).

[14] D. Vorotyntsev, “Should I Use Jupyter Notebooks or Python Scripts for My Next ML Project?” Medium, Dec. 05, 2022. Accessed: Oct. 16, 2023. [Online]. Available: https://medium.com/@tearth/should-i-use-jupyter-notebooks-or-python-scripts-for-my-next-ml-project-7be0ab2ae57e

[15] A. Soni, “Feature Scaling in Machine Learning,” Medium, Jun. 20, 2023. Accessed: Oct. 16, 2023. [Online]. Available: https://medium.com/@soniaman809/feature-scaling-in-machine-learning-regularization-and-normalization-40d1091a45f8

[16] Deepchecks, “What is One-hot Encoding,” Deepchecks, May 29, 2021. https://deepchecks.com/glossary/one-hot-encoding/ (accessed Oct. 23, 2023).

[17] Ahmed, “The Motivation for Train-Test Split - Ahmed,” Medium, Mar. 21, 2023. Accessed: Oct. 23, 2023. [Online]. Available: https://medium.com/@nahmed3536/the-motivation-for-train-test-split-2b1837f596c3

[18] C. Ayuya, “Using Random Search to Optimize Hyperparameters,” Engineering Education (EngEd) Program | Section, Mar. 30, 2021. https://www.section.io/engineering-education/random-search-hyperparameters/ (accessed Oct. 24, 2023).

[19] A. Sharma, “Cross Validation in Machine Learning,” GeeksforGeeks, Nov. 21, 2017. Accessed: Oct. 24, 2023. [Online]. Available: https://www.geeksforgeeks.org/cross-validation-machine-learning/

[20] C. Molnar, Interpretable Machine Learning. Lean Publishing, 2019. Accessed: Oct. 24, 2023. [Online]. Available: https://christophm.github.io/interpretable-ml-book/index.html

[21] BestColleges, “How to Maintain Work-Life-School Balance,” BestColleges.com, Oct. 29, 2021. https://www.bestcolleges.com/resources/work-life-school-balance/ (accessed Oct. 29, 2023).

[22] N. Tyagi, “L2 vs L1 Regularization in Machine Learning,” Ridge and Lasso Regularization, Mar. 01, 2021. https://www.analyticssteps.com/blogs/l2-and-l1-regularization-machine-learning (accessed Oct. 28, 2023).

[23] B. Giba, “Elastic Net Regression Explained, Step by Step,” Machine Learning Compass, Jun. 26, 2021. https://machinelearningcompass.com/machine_learning_models/elastic_net_regression/

[24] A. Jain, “Support Vector Machines (S.V.M) — Hyperplane and Margins,” Medium, Sep. 25, 2020. Accessed: Oct. 29, 2023. [Online]. Available: https://medium.com/@apurvjain37/support-vector-machines-s-v-m-hyperplane-and-margins-ee2f083381b4

[25] C. A. Latkin, C. Edwards, M. A. Davey-Rothwell, and K. E. Tobin, “The relationship between social desirability bias and self-reports of health, substance use, and social network factors among urban substance users in Baltimore, Maryland,” Addictive Behaviors, vol. 73, no. 1, pp. 133–136, Oct. 2017, doi: 10.1016/j.addbeh.2017.05.005.

7. Appendices

7.1 Appendix A (Summary of top dominant features)

Table 2 Summary of top three dominant features with their contributions for global interpretability

Feature	Maximum feature contribution (%)					
	Low feature value			High feature value		
	Ridge	Support vector	Elastic net	Ridge	Support vector	Elastic net
Mjob_other	2.1	1.9	1.8	3	2.8	1.9
studytime	5.3	-*	3.8	9.2	-*	9.2
goout	5.3	-*	4.7	5.1	-*	4.5

* The correlation between the feature and target outcome was not in the top five contributors for the model.

Table 3 Summary of top three dominant features with their contributions for grouped interpretability

Feature	Maximum feature contribution (%)					
	Low feature value			High feature value		
	Ridge	Support vector	Elastic net	Ridge	Support vector	Elastic net
Mjob_other	2.3	1.8	1.7	2.9	2.4	1.9
studytime	4.1	-*	3.8	10.2	-*	9.3
Fjob_other	2.7	2.8	-*	1.9	1.9	-*

* The correlation between the feature and target outcome was not in the top five contributors for the model.