# Analyzing the Mathematical Performance of Secondary School Students in Portugal: A Statistical Approach

Ankita Sharma
*Dyson School of Design Engineering*
*Imperial College London*
as4021@imperial.ac.uk

Maryam Fetanant
*Dyson School of Design Engineering*
*Imperial College London*
mf1420@imperial.ac.uk

Kang Yang
*Dyson School of Design Engineering*
*Imperial College London*
ky321@ic.ac.uk

Avyay Jamadagani
*Dyson School of Design Engineering*
*Imperial College London*
asj21@imperial.ac.uk

Site Ma
*Dyson School of Design Engineering*
*Imperial College London*
sm2921@imperial.ac.uk

## Contents



## 1. Abstract

The report aims to support the student performance in a Portuguese secondary school with statistical analysis and predictions. Five predictions were performed with different machine learning techniques (Logistic regression, Decision trees, Random Forest and Support Vector Machines). Recall was prioritized over precision as false negatives were to be minimized. Pass and Fail for first and second term exams were predicted using Decision Trees and Random Forest, with final recall results being 0.696 and 0.750 respectively. Support Vector Machines were used with rbf kernels to predict the effect of social activities, family relationships, and financial background on student performance in final exams, with a resultant recall of 0.765 for both. Lastly, a Logistic Regression was used to determine the probability of obtaining above a B in grade average to provide enrichment activities, resulting in a recall of 0.857.

Logistic Regression was determined to result in the highest accuracy of 89.9% and highest recall of 85.7%. It was concluded that all models must be used conservatively and in conjunction with each other to provide a holistic result for the student, but also to compensate for the moderate accuracies of the model. It was also concluded that more data is required in the dataset to achieve more accurate results by mitigating overfitting.

## 2. INTRODUCTION

As a school, we are deeply committed to the well-being and success of each and every student. We recognize that every student is unique, and we strive to provide the necessary support to ensure their academic journey is fulfilling and rewarding. To achieve this, we are continually exploring innovative approaches and leveraging the power of machine learning to gain valuable insights to better understand the factors that impact student performance and to identify effective strategies for supporting their individual needs.

Leveraging these insights, we can identify early warning signs for struggling students, predict academic outcomes, and implement targeted interventions and support programs. Our aim is to provide proactive assistance so each student can reach their full potential and thrive academically.

## 3. RELATED WORK

The original work on this dataset, "Using Data Mining to Predict Secondary School Student Performance" from Cortez, P., & Silva, A. (2008) analyzed the performance of different models (NN, DT, RF, SVM). In the report, it was discussed that there were strong correlations between grades, but it would be also interesting to look at the other factors. Another noteworthy study is "The influence of social and economic disadvantage in the academic performance of school students in Australia" [ Considine, & Zappalà. (2002)]][1]. This study delves into the impact of social and economic factors on the academic performance of students in Australia. It highlights the significance of considering socio-economic disadvantages when analyzing student performance, as these factors can have a profound effect on educational outcomes. The study suggests that addressing social and economic disparities is crucial for enhancing the academic performance of students. It also suggests that both "nature" and "nurture" factors are probabilistic, not deterministic. This research aims to build upon these studies by not only considering the academic factors but also exploring the various non-academic factors such as demographic, social, and financial aspects in predicting student grades.

## 4. METHODOLOGY

### *3.1 Our data*

Our dataset, sourced from the UCI Machine Learning Repository [2], consists of 396 samples and 33 attributes encompassing grades, demographic, social and school related features. Notably, the data pertains to the secondary education of two Portuguese schools. The dependent variable focuses on the mathematics grades attained by students in terms 1, 2, and 3. Our models are designed to predict student performance.

### *3.2 Feature Engineering*

To clean a dataset for analysis and eliminate biases and outliers, we initially addressed missing data and duplicates, and also reviewed for spelling errors. Class imbalances were handled, and data normalization was performed individually as required based on the chosen methods. In the dataset, there were 350 entries for Gabriel Pereira school and only 46 entries for Mousinho da Silveira school. To ensure a more representative dataset and avoid outliers, the entries for Mousinho da Silveira school were removed. Consequently, the column indicating the school's name attribute was eliminated. Finally, hot encoding was applied to transform categorical variables into numerical features that can be utilized in data analysis methods.

## 5. PREDICTING STUDENT PASS AND FAIL GRADE IN FIRST TERM EXAMS

***Maryam Fetanat***

### *5.1 Aim*

This section focuses on predicting student pass and fail grades in first term exams and identifying the factors that contribute to student failure. Early identification of struggling students is crucial for providing targeted assistance. A reliable model for predicting potential failures is vital for timely intervention and support.

### *5.2 Model Selection*

A decision tree is a supervised learning algorithm that is commonly used for classification problems. It creates a tree-like model with nodes representing features, branches representing decisions, and leaf nodes representing outcomes. In the context of predicting student pass and fail grades, a decision tree can be trained using contextual data in the data set such as absences, past failures, and family background. The algorithm learns how different features contribute to the classification of pass or fail.

Gini impurity is a metric used in decision trees to measure the impurity of a set of data points. It quantifies the probability of misclassifying a randomly selected element in the dataset if it were randomly labelled according to the class distribution of the data. The goal of a decision tree algorithm is to find the feature that minimizes Gini impurity, as it indicates the best feature for creating pure splits and improving the overall classification accuracy of the tree.

The tree can be interpreted to understand which features have the most significant impact on the prediction, providing valuable insights for intervention and support strategies. To prioritize the prediction of student fail grades in this evaluation, minimizing the occurrence of falsely predicting a student's passing status is crucial. Therefore, recall was selected as the primary metric to minimize false negatives.

### *5.3 Data Preparation*

#### *5.3.1 Feature Engineering*

As explained in the *Methodology* section, the dataset was cleaned to eliminate biases and outliers. To address the classification problem focused on pass and fail outcomes, the numerical grades were transformed into a binary format. In the Portuguese education system, grades range from 1 to 20, with grades above 10 considered a pass.[3] Therefore, a binarization process was applied, where a grade of 1 represents a fail and a grade of 0 represents a pass.

#### *5.3.2 Data Balance*

Class distribution balance is a crucial factor influencing results. Imbalanced datasets, with one class having significantly fewer instances, can negatively impact recall. In this case, there were 349 valid data points, with only nine sample difference in classification. Therefore, there was no need to under sample the majority class, as it could introduce inaccuracies by reducing the overall sample size.

#### *5.3.3 Data split*

To avoid overfitting, the dataset was randomly divided into three sets: 80% for training, 10% for validation, and 10% for testing. The training set was utilized to construct the decision tree model, while the validation set facilitated fine-tuning of the model. Lastly, the performance of the model was evaluated using the testing set.

### *5.4 Hyper-parameter optimization*

Hyperparameters are user-defined settings that influence the structure and prevent overfitting in decision trees, playing a vital role in the learning process and model behavior. The optimal values for maximum depth and minimum impurity decrease in decision trees were determined by using the training and validation set and plotting them against recall.

### *5.4.1 Maximum Depth*

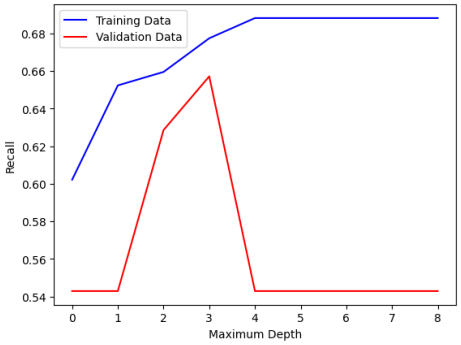



*Figure 1: Recall vs Maximum depth*

As the model becomes more complex, the recall on the training set increases, but overfitting occurs simultaneously, as evidenced by the validation set. Therefore, a maximum depth of 3 was chosen.

### *5.4.2 Minimum Impurity Decrease*

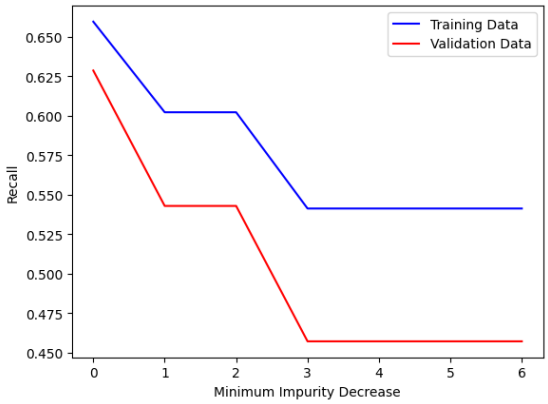



*Figure 2: Minimum Impurity Decrease*

*Figure 2* demonstrates that the model overfits the training data, but only to a minimal extent initially. Hence, the value of Minimum Impurity Decrease was set to 0, as it corresponds to a higher recall level.

### *5.5 Results*

The decision tree illustrated in Figure 3 was constructed using the identified hyperparameters. The leaf nodes had Gini coefficients ranging from 0 to 0.5. On the validation set, the model exhibited a satisfactory recall of 70.0%, which is the most significant metric for this dataset.

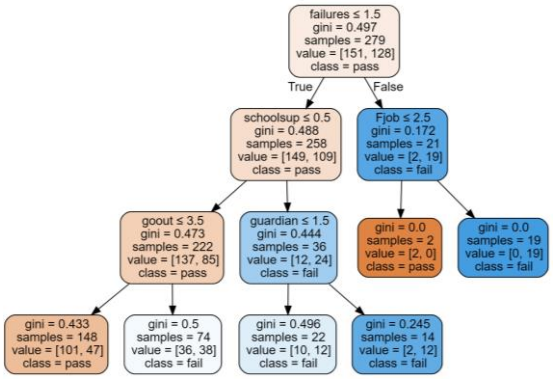



*Figure 3- Decision tree*

*Table 1: Model Performance Metrics*

| | **Accuracy** | **Precision** | **Recall** |
|---|---|---|---|
| **Training** | 0.659 | 0.628 | 0.633 |
| **Validation** | 0.629 | 0.688 | 0.579 |
| **Testing** | 0.714 | 0.842 | 0.696 |

Table 1 illustrates that the accuracy and recall on the test set surpassed those of the training set. This indicates that the model is a reliable predictor without overfitting, allowing for generalization to unseen data for students in this school. A confusion matrix was also plotted, showing a majority of true positives and true negatives, indicating accurate predictions by the model.

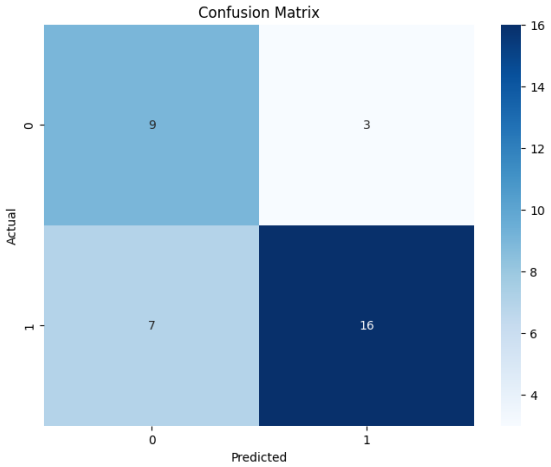



*Figure 4- confusion matrix*

*5.6 Discussion*

If a student has more than one past failure and their father works in service jobs, at home or others, the model indicates a high probability of failing. The Gini impurity of this node is 0, indicating no chance of a false positive. This suggests a correlation between these features, but not a causality. On the other hand, students who have failed less than two times, do not receive extra support, and socialize with friends above a medium level also tend to be predicted as failures. However, this data is less reliable as it has a high Gini impurity of 0.5, indicating a significant chance of being incorrect.

The reliability of the results is also influenced by sample sizes. In the case of students who have failed more than once and have fathers working in health and teaching roles, there are only two samples. Although they were classified as pass with a Gini impurity of 0, the limited sample size reduces the reliability of this result when making predictions or drawing conclusions.

A feature importance plot of the decision tree model shows the most significant features are number of past class failures, extra educational support, going out with friends, father's job, and the student's guardian.

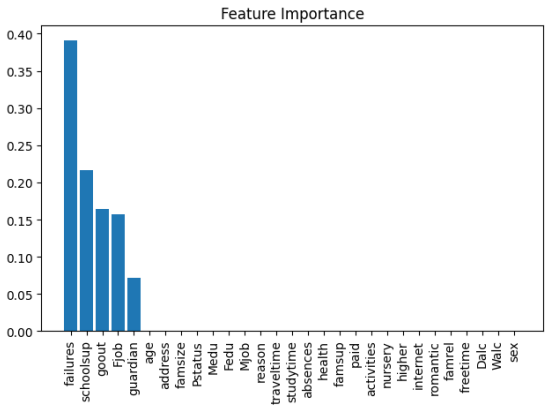



*Figure 5: Feature importance*

The number of past class failures emerged as the most important feature in other models, indicating a strong correlation that can assist the school in providing better support to these students. The overall model achieved a high precision of 0.84 and a recall of 0.70 on the testing data, demonstrating its effectiveness. To enhance the fit of the decision tree model in predicting student pass and fail outcomes, it is beneficial to obtain a larger and more diverse dataset for comprehensive training and mitigating overfitting. Additionally, since the data was collected from only one school, it is important to note that the model should not be extrapolated to larger populations.

# 6. PREDICTING PASS OR FAIL RATE FOR SECOND PERIOD GRADE

***Kang Yang***

## *6.1 Aim*

This study aims to use the Random Forest algorithm to predict student performance in second-period mathematics (G2). The goal is to identify students at risk of underperforming, enabling timely interventions before the final evaluation (G3).

## *6.2 Model Selection*

Random Forest, an ensemble learning method [4], was chosen for its high accuracy, ability to handle diverse data, robustness to overfitting, and feature importance estimation. The model was built by bootstrapping the data, selecting features, building multiple trees, and making predictions based on majority voting.

## *6.3 Data Preparation*

The dataset was inspected for missing values, which were either dropped or imputed. Categorical variables were converted into numerical format using one-hot encoding. The target variable, G2, was transformed into a binary variable (pass or fail). The dataset was divided into training (80%), validation (10%), and testing(10%) subsets[5].

## *6.4 Hyperparameter Optimization*

The model was fine-tuned on four features (max_depth, min_sample_split, min_impurity decrease, and n_estimators). The focus was on the recall of the training and validation datasets to select parameters that limited overfitting, where a model learns the training data too well and performs poorly on unseen data.

The emphasis on recall, which measures the model's ability to correctly identify all actual instances of a certain class, is particularly important in this context. Since the goal is to predict the pass or fail rate for the second period grade, it's crucial to minimize false negatives, i.e., students who are predicted to pass but actually fail. By identifying students who are at risk of failing, the school can intervene early and provide additional support to help them improve their performance for the final G3 test. The final model's performance was evaluated using the test dataset.

Model 1 had the starting variables max_depth = 2, min_samples_split = 0.1 and random_state = 0.

## *6.5 Maximum Depth*

The maximum depth of a tree, denoted as max_depth, represents the longest distance from the root node to a leaf node. As this depth increases, the model's capacity to capture more data also increases. However, this can lead to overfitting if the tree grows too complex due to excessive splits. Therefore, this parameter can be adjusted to control the complexity of each tree [6].

In the model under consideration, a notable discrepancy was observed between the accuracy on the training set and the validation set, with the latter being significantly lower. This discrepancy suggests that the model may not be effectively generalizing to unseen data. To rectify this, the max_depth parameter was adjusted to 7 in the second iteration of the model, as this value provided the highest accuracy on the validation set.

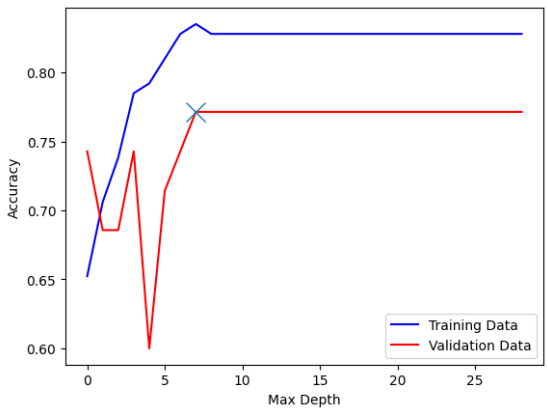



*Figure 6: Max depth vs Accuracy*

## *6.6 Minimum Sample Split*

The parameter 'minimum sample split' refers to the smallest quantity of observations required at a node for it to be eligible for splitting. This parameter plays a crucial role in controlling the level of constraint at each node and determines

the number of subdivisions that will occur, which could potentially lead to overfitting [7].

It was observed that as the value of the minimum split increased, the accuracy for both the training and validation sets declined. To optimize the model's performance and prevent overfitting, the 'minimum sample split' parameter was adjusted to 2 in the third iteration of the model. This adjustment was made as it resulted in the highest accuracy for the model without overfitting [8].

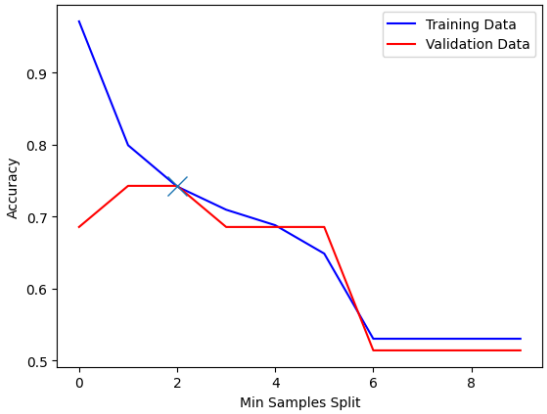



*Figure 7: Min Samples Split vs Accuracy.*

## *6.7 Minimum Impurity Decrease*

The 'minimum impurity decrease' parameter in a decision tree model sets a threshold for the reduction in impurity that must be achieved for a node to be split. This parameter helps control the complexity of the model and prevent overfitting [9].

In the model under consideration, there was a corresponding decrease in accuracy for both the training and validation sets. To optimize the model's performance, the 'minimum impurity decrease' parameter was adjusted to 0 in the fourth iteration of the model. This adjustment was made as it was found that the accuracy declined rapidly beyond this point [10].

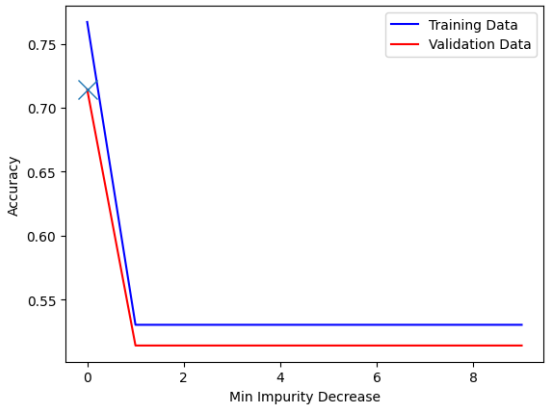



*Figure 8: Min Impurity Decrease vs Accuracy*

## *6.8 N Estimators*

The 'n_estimators' parameter signifies the count of trees in the model. Generally, a larger number of trees can enhance the model's ability to learn from the data. However, an excessive increase in the number of trees can lead to a slower training process without a corresponding significant improvement in accuracy once a saturation point is reached [11].

In the model under consideration, it was observed that the accuracy of both the training and validation sets continued to increase until a certain point. Upon reaching an 'n_estimators' value of 7, the accuracy ceased to increase. Consequently, this value was incorporated into the fourth iteration of the model [11].

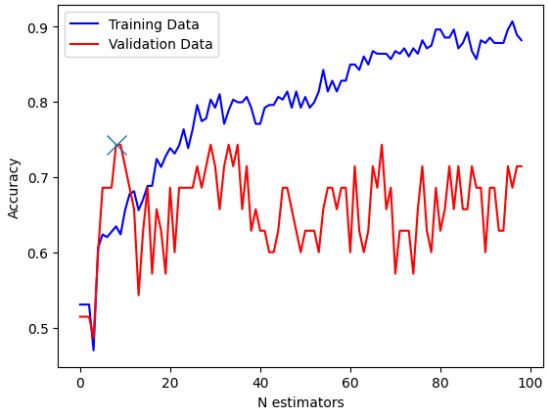



*Figure 9: N estimators vs Accuracy*

## *6.9 Results*

The final model had the following parameters for the Random Forest Classifier: max_depth = 7, min_samples_split = 2, random_state = 0, min_impurity_decrease = 0, n_estimators = 7.

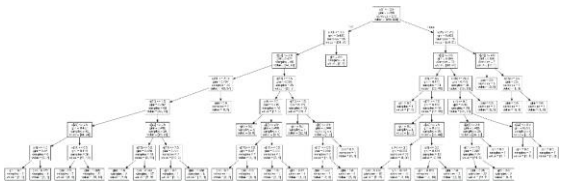


*Figure 10: Example tree from random forest (iteration 4, refer to appendix for larger picture)*

*Table 2: Comparison of recall with different hyperparameters*

| Iteration | Training | Validation | Test |
|---|---|---|---|
| **1** | 0.571 | 0.471 | 0.600 |
| **2** | 0.690 | 0.630 | 0.663 |
| **3** | 0.701 | 0.692 | 0.705 |
| **4** | 0.723 | 0.702 | 0.750 |

*Table 5: Comparison of recall with different hyperparameters*

| | Accuracy | Precision | Recall |
|---|---|---|---|
| **Training** | 0.799 | 0.893 | 0.723 |
| **Validation** | 0.713 | 0.863 | 0.702 |
| **Test** | 0.827 | 0.912 | 0.750 |

## *6.10 Discussion*

The recall on the test data showed a significant improvement, exceeding 70%, indicating that the model is effective at identifying students who are at risk of failing. This is crucial in an educational context, as early identification of students who may struggle allows for timely interventions and support, potentially improving their final grades.

The Random Forest algorithm, with its inherent randomness in generating a multitude of decision trees and aggregating their predictions, may exhibit some variability in recall values for the same hyperparameter settings across different runs. This variability, far from being a drawback, is a strength of the Random Forest algorithm. It contributes to the creation of a more robust model that can generalize well to unseen data, making it a reliable tool for predicting student performance in diverse and changing educational environments.

Iteration 4 was selected as the final model, a decision based on its overall performance and the steps taken to prevent overfitting. This model, with its high recall rate, serves as a valuable tool for educators and educational stakeholders. It can help in identifying students at risk of underperforming in their second-period mathematics grade, enabling targeted interventions. This could lead to improved educational outcomes, benefiting not only the students themselves but also the broader educational community.

# 7. PREDICTING THE EFFECT OF FAMILY AND FINANCIAL BACKGROUND STUDENT WILL PASS OR FAIL IN THEIR FINAL EXAM

## *Site Ma*

### *7.1 Aim*

The aim of this analysis is to develop a prediction model to assess students' performance in the final evaluation (G3) at a Portuguese secondary school, with a specific focus on investigating the influence of students' family and financial backgrounds. The primary objective of the prediction model is to identify students who may require additional support without knowledge of their performance (G1&G2), aiding in early intervention strategies to enhance academic outcomes.

### *7.2 Model Selection*

For this analysis, the Support Vector Machine (SVM) model is selected. The SVM was chosen due to its effectiveness to model non-linear relationships with complex variables. It is particularly suitable for classification problems and could be adapted for regression. SVM aims to find an optimal hyperplane that maximally separates classes while minimizing the influence of individual data points, making it robust to overfitting. [12]

In this analysis, G3 will be predicted for pass/ fail as a binary classification. Though there is linear relationships between G1, G2, and G3, the dataset contains more complex variables indicating students' background and family/social/financial status. SVM would be handling these non-linear inputs and capture the relationships influenced by various factors.

### *7.3 Data Preparation*

#### *7.3.1 Data Observation*

No missing values were found, and all data was usable. The dataset includes nominal data such as mother jobs (MJob), ranked ordinal data family relationships (famreln), and continuous data (G3). Further preparations were needed for proceeding categorical datatypes.

#### *7.3.2 Feature Engineering*

All binary features were transformed into 0 and 1. Nominal data, including Mjob, Fjob, and guardian were transformed into numerical format using hot encoding, making them suitable for SVM. [13] Target variables, G1, G2, and G3 were transformed into 0 and 1, representing pass and fail respectively. A threshold of 10 was set according to the original paper. [14]

#### *7.3.3 Attribute Removal*

This analysis focused on the impact of family and financial status. Attributes not directly related, such as absences and study time, were removed. Additionally, G2 was excluded as per the analysis requirement. Mjob_1 and Fjob_1 were also arbitrarily removed to avoid multicollinearity, so that high correlation between the encoded variables would be eliminated. Some were removed during iterations of the training, due to irrelevance to the result. The final ones are:

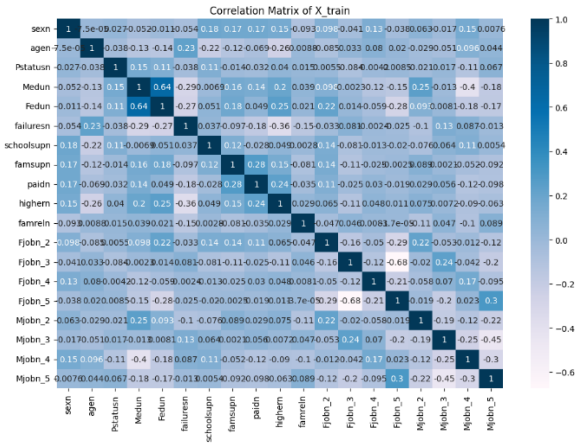



*Figure 11: Correlation Matrix*

#### *7.3.4 Data Balance*

The dataset was observed to be imbalanced with respect to the target variable G3. 190 samples were negative and 151 were positive. The majority class was undersampled to match the number of positive cases.

### *7.3.5 Preprocessing*

The data was shuffled and split into 80% training, 10 % validation, and 10% testing. Training data was standardized for SVM model.

### *7.4 Hyperparameter optimization*

The goal is to aid all students needed for help, so a high recall is required. The cost is that the prediction may lead to putting resources onto those who are fine, but this is acceptable as a school. Models were trained with different hyperparameters and compared by the scores on the validation samples.

### *7.4.1 Kernel Type Selection*

Different kernel functions, such as linear, polynomial, radial basis function (RBF), and sigmoid, offer various ways to capture and represent the underlying patterns and structures in the data. Table S1 shows the validation scores for each kernel type under default settings.

*Table 6: Model Validation scores of each Kernel type*

| | **Linear** | **Poly** | **RBF** | **sigmoid** |
|---|---|---|---|---|
| **Accuracy** | 0.75 | 0.687 | 0.687 | 0.875 |
| **Precision** | 0.909 | 0.769 | 0.733 | 1.0 |
| **Recall** | 0.588 | 0.588 | 0.647 | 0.764 |

Sigmoid had the best scores, however, it appeared to be overfitting on the validation data. RBF was then selected for its performance.

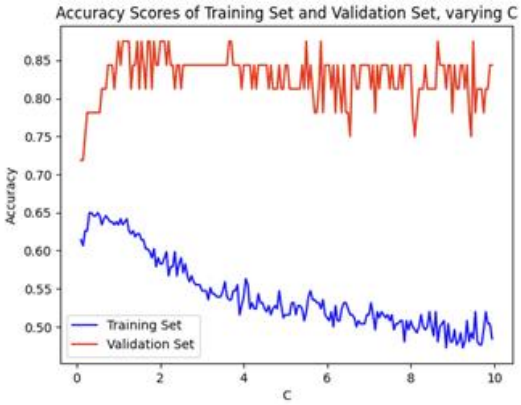



*Figure12: Accuracy scores of hyperparameters (Kernel = sigmoid, C = 0 - 10)*

### *7.4.2 C-Value Selection*

C value controls the regulation of the data. A larger C value imposes a stricter margin, aiming to minimize the training error by penalizing misclassifications more heavily. This can potentially lead to a more complex decision boundary, allowing for a better fit to the training data. However, it also increases the risk of overfitting, where the model becomes overly sensitive to the noise in the training data, leading to poor generalization on new data.[15] The recall score was checked with C ranging from 0 to 10, compared to the precision score to determine the trade-off

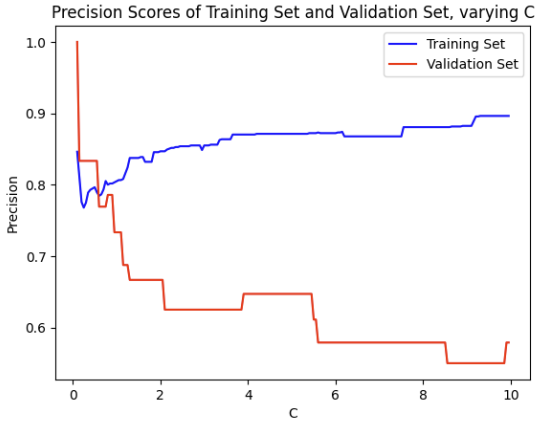



*Figure 13: Precision scores of varying CFigure: Precision scores of varying C*


*Figure14: Recall scores of hyperparameters (Kernel = sigmoid, C = 0 - 10)*

Larger C values wouldn't contribute to the recall scores but lowering the precision and increase the chance of overfitting. C = 1.2 was picked with recall = 0.647.

### *7.4.3 Gamma Value Selection*

A similar method was used for gamma selection. Gamma in SVM determines the proximity of points in the input space. Higher gamma values make points closer together, potentially leading to overfitting.

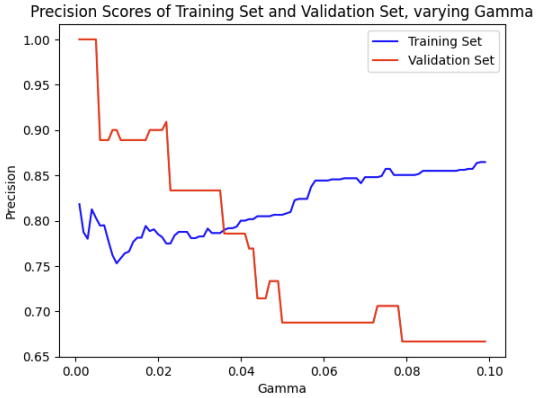



*Figure 15: Precision scores of varying gamma*

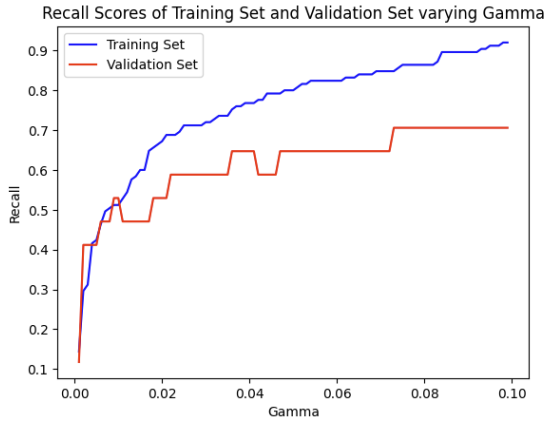



*Figure 16: Recall scores of varying gamma*

### *7.5 Results*

The optimized SVM model was trained on the prepared dataset.

*Table 7: Model Performance Metric data*

| | Training | Validation | Testing |
|---|---|---|---|
| **Accuracy** | 0.862 | 0.688 | 0.656 |
| **Precision** | 0.857 | 0.706 | 0.684 |
| **Recall** | 0.864 | 0.706 | 0.765 |

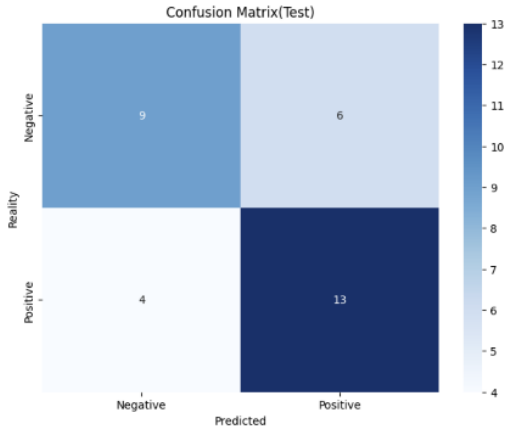



*Figure 17: Confusion Matrix Testing data*

### *7.6 Discussion*

The SVM model had moderate performance in predicting the final performance of students based on family and financial backgrounds, with recall = 0.765 and precision = 0.684 for testing data. However, there is room for improvement, such as finding a better balance between precision and recall using other scoring methods, including F1 and AUC. The data set is also relatively insufficient for training and validation, larger sample size would increase the functionality of the model.

SVM could also be difficult to interpret the model, as SVM does not provide feature importance scores inherently, but insights can still be derived from the feature selection and scores.

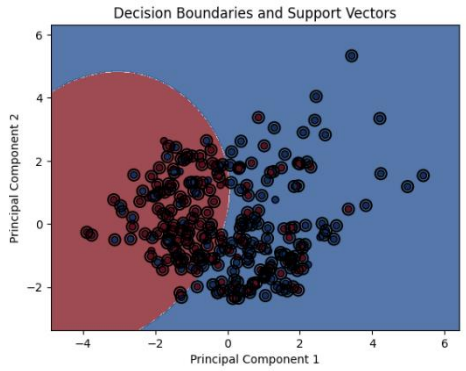



*Figure18: Decision Boundaries and Support vector*

It is also noticed that it was difficult to predict students' performance with only given backgrounds. The prediction would be much more accurate with one grade, such as G1 (accuracy = 0.88, precision = 0.81, recall = 0.96), indicating that the school could have better interventions after first term

examinations. However, the current model can provide insights for earlier interventions.

In conclusion, while the SVM model provides valuable insights into the impact of family and financial backgrounds on student performance, it should be used in conjunction with other factors and models for a more comprehensive analysis.

# 8. PREDICTING WHO GETS ABOVE B IN GRADE AVERAGE TO PROVIDE ENRICHMENT ACTIVITIES

***Ankita Sharma***

## *8.1 Aim*

It is crucial to ensure that every student in the school is adequately supported to reach their full potential. This recognising and nurturing the talents and abilities of the top achieving students. By offering super curricular activities, such as advanced academic programs and research opportunities we can further enhance their educational journey.

We would like to predict students likely to get above an A/B as an average of their G1, G2 and G3 mathematic exams.

## *8.2 Model Selection*

Logistic regression was chosen as it is a good model to predict whether a student gets an A, it is well-suited for binary classification problems, where the target variable has two categories. In this case, the two categories would be "A" and "Not A". Logistic regression is generally robust to outliers compared to other algorithms like linear regression maintaining the model's stability.

## *8.3 Data Preparation*

An 'A' according to national Portuguese standards are for students who achieve on average above 18/20, in the dataset presented only 5 out of the 350 samples achieved an A. Henceforth the boundary was brought down to a B grade at 15/20. This improved the distribution from 5/350 to 43/350.

### *8.3.1 Feature Engineering*

A new attribute was created called G_Average, which was a total sum of the 3 grades, this predictor variable was in the form of continuous data ranging from 0-20, hence a threshold value of 15/20 was chosen to binarise the data into 0s and 1s, where 1 was a student achieving B and higher.

### *8.3.2 Under sampling the majority class*

Generally, logistic regression can still perform reasonably well even with moderate class imbalance. However, this dataset would come under an extreme case where the minority class of receiving a B or higher is severely underrepresented. The model may struggle to generalize and predict the minority class accurately and the model risks becoming biased towards the majority class. Hence the majority class of receiving less than B was under sampled randomly to 107.

## *8.4 Attribute Removal*

### *8.4.1 Correlation Heatmap*

Performing this correlation analysis helped in assessing the interdependencies between features and removing highly correlated variables to avoid multicollinearity such as G1, G2, and G3, which were removed from the list of attributes as G_average is comprised of those 3 values.

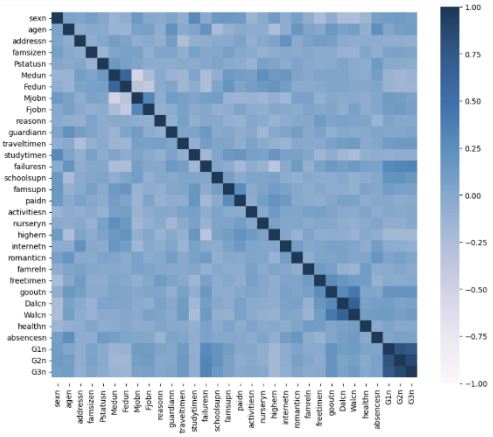



*Figure19: Correlation Map of features*

### *8.4.2 LASSO Coefficients*

After removal, there were 43 attributes, LASSO regression was used to filter through all the features and understand those that had the highest predictive power.

LASSO performs both feature selection and coefficient shrinkage simultaneously. The strength of the penalty term in LASSO is controlled by the hyperparameter alpha. Increasing the value of alpha results in more aggressive feature selection, where more coefficients are shrunk to zero.

At alpha = 0.01, there was only one prominent feature visually, which was Mother's Job

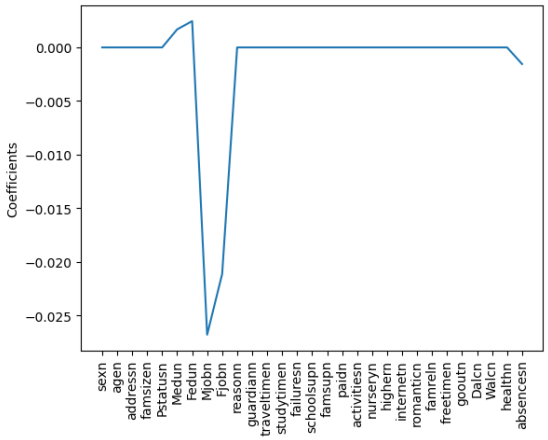



*Figure 20: alpha = 0.1 LASSO coefficients*

## *8.4 Hyperparameter Optimisation*

The Alpha value was further refined to a smaller value, to see what other attributes also had a strong predictive power. In figure 3, alpha has been brought down to 0.02. The alpha parameter controls the regularization strength applied to the model, a small alpha value in Lasso regression implies that the model may risk in being less constrained and be more likely to fit the training data closely, also known as overfitting.

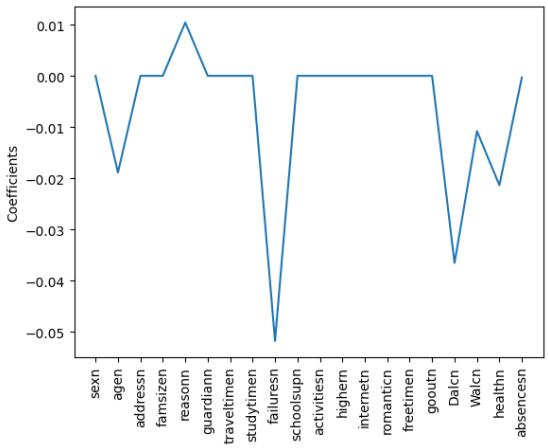



*Figure 21: Alpha= 0.2 LASSO Coefficients*

Features were selected based on the highest magnitude of coefficients as it represents the strength of the relationship between the feature and the target variable.

Further feature selection was carried out by calculating the p-values associated with each feature's coefficient, which provide information about the statistical significance of

*Table 8: LASSO and p values for each feature*

| Feature | LASSO Coefficient | P value |
|---|---|---|
| School Support | -0.240608 | 0.041 |
| Paid | -0.21853 | 0.004 |
| Romantic | -0.09919 | 0.142 |
| Failures | -0.06860 | 0.165 |
| Study time | 0.06294 | 0.083 |
| Mother education | 0.05085 | 0.258 |
| Father education | 0.05001 | 0.179 |
| Mother job | -0.24831 | 0.576 |
| Father job | -0.03742 | 0.181 |
| Health | -0.02808 | 0.040 |
| Free time | 0.01650 | 0.654 |
| Age | -0.00125 | 0.669 |
| Higher | 0.00000 | 0.574 |

It is evident that some of the features chosen in the model have a non-significant p value, however, in exploratory data analysis, including non-significant features can help uncover potential relationships or patterns that were not initially evident.

## *8.5 Data Splitting*

To prevent overfitting, the dataset was split at random into training, validation and testing subgroups. This was done at a ratio of 60:20:20. Initially an 80:10:10 ratio was tested however this resulted in a very small sample size for the validation and testing groups.

## *8.6 Recall as the chosen parameter.*

Recall was chosen as the parameter to maximise, to minimise the False Negatives. We would like to minimise falsely predicting saying someone will get above B when they are likely to get lower.

There have been many studies and research that show that *Mixed-ability classrooms can negatively affect students with low academic achievement*.[16] Rather than the possibility of causing more mental stress and a negative impact on these students, we would rather give the correct support so students can flourish in school.

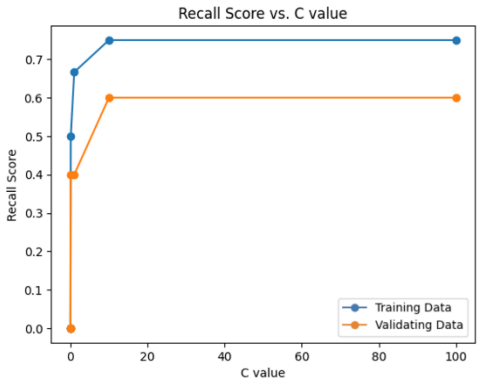



*Figure 22: Recall Score vs C value*

By fine tuning the C value, it is possible to reduce the risk of overfitting by controlling the complexity of the model. In the graph above different C values were iterated through to see which C-value was optimum, C values above 20 was at risk of over-fitting, whereas below 0.1 was at risk of under fitting, hence, a C value of 0.01 was chosen to improve the recall of the overall model.

## *8.7 Results*

*Table 9: Model Performance Metric Data*

| | **Accuracy** | **Precision** | **Recall** |
|---|---|---|---|
| **Training** | 0.700 | 0.621 | 0.645 |
| **Validation** | 0.911 | 0.800 | 0.821 |
| **Testing** | 0.889 | 0.800 | 0.857 |

## *8.8 Discussion*

The performance metrics in testing data indicate high accuracy, precision, and recall, maximizing the desired recall. This shows that the model has a high predicative power on which student is likely to get above a B.

However, reducing features led to a stark decrease in recall, despite selecting parameters with high LASSO coefficients. This suggests high correlation among parameters, causing underfitting and a loss of complexity. Adding features significantly increased recall from 0.023 to 0.621, leading to the decision of retaining a larger feature set.

The noticeable recall variation among training, validation, and testing sets may stem from the small dataset size, potential data distribution discrepancies, or variations in sample characteristics. Collecting more data is recommended to identify prominent features and gain a better understanding of factors affecting student achievement.

## *8.9 Ethical Considerations*

Considerations should be made regarding the strong predictive power of features associated with privileged backgrounds, such as access to schooling and employment opportunities, in determining student achievement. It is crucial for schools to avoid perpetuating systemic divisions among students based on their backgrounds. Pushing privileged students further ahead solely due to their background would be unjust and inequitable, undermining the efforts and potential of less privileged students who are hardworking and equally determined.

It is highly recommended that this model be used as a guide rather than a tool to form academic ability sets within the classroom. Additional data collection should prioritize attributes related to students' attitudes and work ethic, independent of their family backgrounds. This approach would foster fairness and equal opportunities for all students, acknowledging their individual potential and contributions.

# 9. PREDICTING THE EFFECT OF 'SOCIAL ACTIVITIES ON WHETHER A STUDENT WILL PASS OR FAIL IN THEIR FINAL EXAM

***Avyay Jamadagni***

## *9.1 Aim*

Social activities like 'going out' have dramatic impacts on their academic performance. Lavy. V. et al [17] suggested that the first circle of friends provides a supportive atmosphere for the student, but Ansari. W. et al [18] suggested that alcohol consumption, a common activity when 'going out', results in a decline in academic performance. To classify whether a student will pass or fail, it is essential that Gabriel Pereira School considers the social factors around the student. As social factors influence the student year-round, the question aims to ask how 'going out' and other social factors influence whether the student will pass or fail in their final exams. In this study, false negatives are minimized as it would be detrimental to a student if their Failure is incorrectly deemed to be a Pass, and thus does not receive additional support. Hence, the recall is prioritized over the precision.

## *9.2 Model Selection*

The problem requires classification into Pass or Fail, should be efficient with 12 independent features, and should support non-linear correlations, so a kernelized-Support Vector Machine (SVM) model may be used. SVMs are supervised machine learning algorithms that draw a hyperplane separating different classes with a margin relating to the allowed error between classes. SVMs are more computationally efficient for higher dimensions than Random Forests and Decision Trees and are robust to outliers [19]. As the number of samples is significantly larger than the number of features in this case, SVMs are suitable.

SVMs have three main hyperparameters, as described below:

**Kernel:** Describes the method to draw and shape of the hyperplane. Linear kernels are default which are used for linearly correlated data, but Radial Basis Function (rbf), Polynomial (poly), and Sigmoid are used for non-linear data.

**C:** The penalty for misclassifications, which is inversely proportional to the margin size, indicating that a higher C value draws a more complex decision boundary and increases the chance of overfitting.

**Gamma:** Applicable for non-linear kernels and dictates the influence of individual training points on the model. A higher gamma draws a more complex decision boundary and increases the chance of overfitting.

## *9.3 Data Preparation*

### *9.3.1 Data Observation*

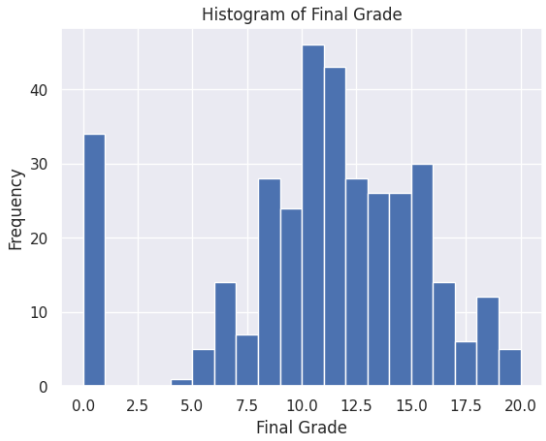



*Figure 23: Histogram of final grade to assess distribution and standardization.*

The data appears to be almost normally distributed. The mean is 10.5, median is 11, and standard deviation is 4.63. With the pass condition as achieving 11 or above, the dataset should be close to balanced. This will be confirmed later through a confusion matrix. The 0 grades will not be disregarded as they are not an outlier and belong to the 'Failure' set of interest.

### *9.3.2 Feature engineering*

The Final Grade data was converted from continuous data to binary (0 and 1), with the threshold being 11, as indicated above. If a student fails, it will be considered by the model as a 'positive' result as the model aims to

predict Failures. Feature removal (Intuition and Logistic Regression)

### *9.3.3 Feature removal (Intuition and Logistic Regression)*

As the focus is on features related to social aspects, features related to demographics, parental education, and financial support were removed. The remaining features 12 features consisted of interpersonal relationships, personal habits, and family support. However, after performing a logistic regression to view p-values, it was noted that 5 features had negligible correlation to the final grades, so they were removed too. The remaining 7 features are listed below:

- *Going out*
- *Daily alcohol consumption*
- *Weekly alcohol consumption*
- *Extracurricular activities*
- *Romantic relationship*
- *Access to internet*
- *Free time*

### *9.3.4 Data balance*

Through a confusion matrix, it was determined that the total failures were 159 and the total passes were 190. Given that the model is to be built to detect failures, this bias must be removed to provide an accurate model. The passes were randomly under-sampled to create a balanced dataset of 318 samples.

### *9.3.5 Data split*

Out of the 318 samples, an 80-10-10 split was used because the number of samples is relatively small, so more samples are to be used for training. Hence, 254 samples were allocated for training, and 32 samples for validation and testing. The 286 samples for training and validation were standardized to reduce the chance of misclassification by the penalty term.

## *9.4 Hyperparameter optimization*

A correlation heatmap was plotted for the training data to gain an intuition for the type of possible correlations to exist.

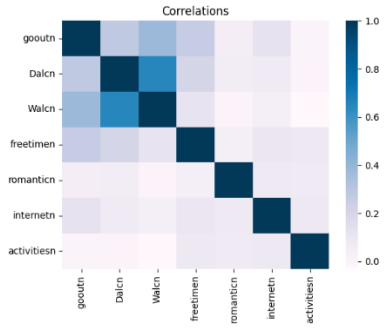



*Figure 24: Correlation heatmap of chosen features*

Only ‘Going out’, ‘Daily alcohol consumption’, and ‘Weekly alcohol consumption are linearly correlated with each other, implying that all other relationships are non-linear. Hence, it is expected that a non-linear kernel will be optimal, and higher C and Gamma values must be considered to achieve a more accurate decision boundary.

### *9.4.1 Kernel Optimization*

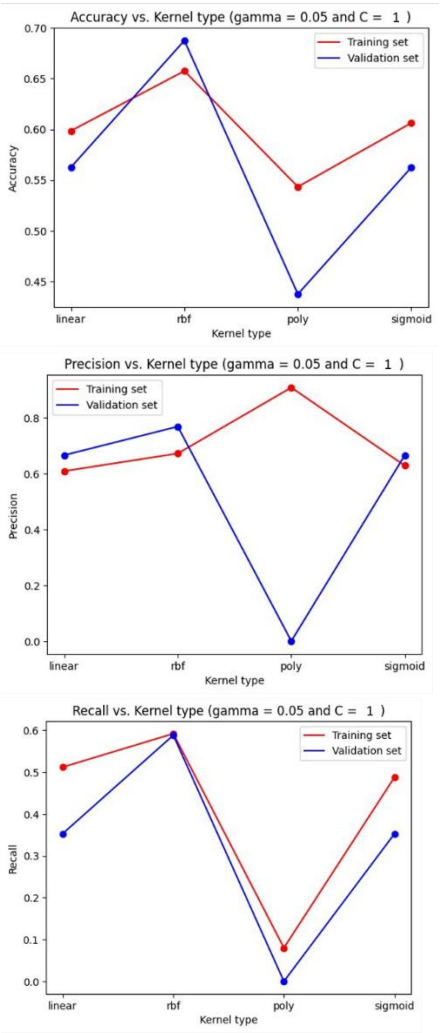


*Figure 25: Kernel optimization graphs for (from top to bottom) accuracy, precision, and recall*

The **rbf kernel** has the highest accuracy and recall with a consistently small discrepancy between the training and validation values, which shows that it is likely to be the most accurate kernel.

### *9.4.2 C-value Optimization*

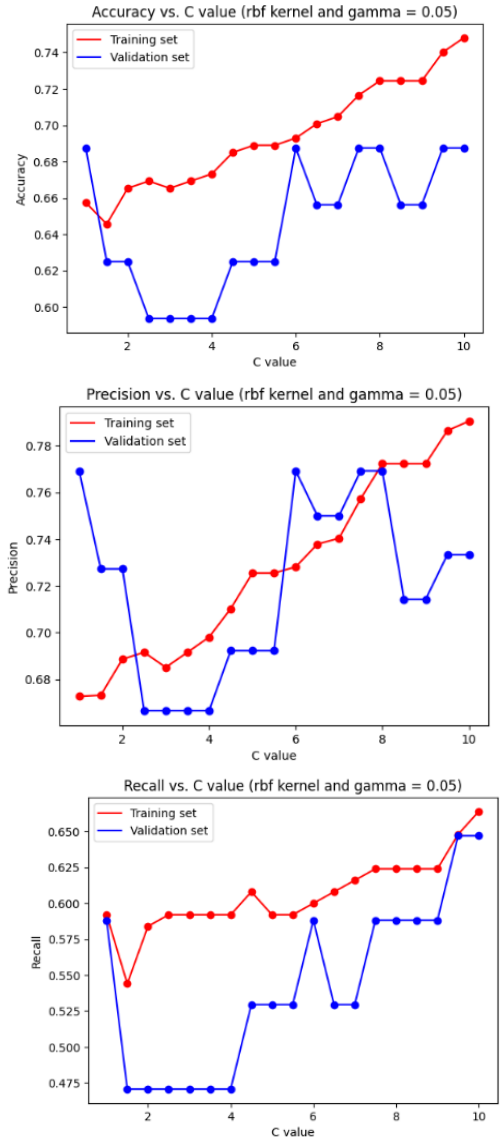



*Figure 26: C-value optimization graphs for (from top to bottom) accuracy, precision, and recall*

The highest recall and high accuracy was determined at C = 9.5 and 10. Hence, the lower value of **C = 9.5** will be chosen to reduce the possibility of overfitting when combined with a better gamma.

### *9.4.3 Gamma Optimization*

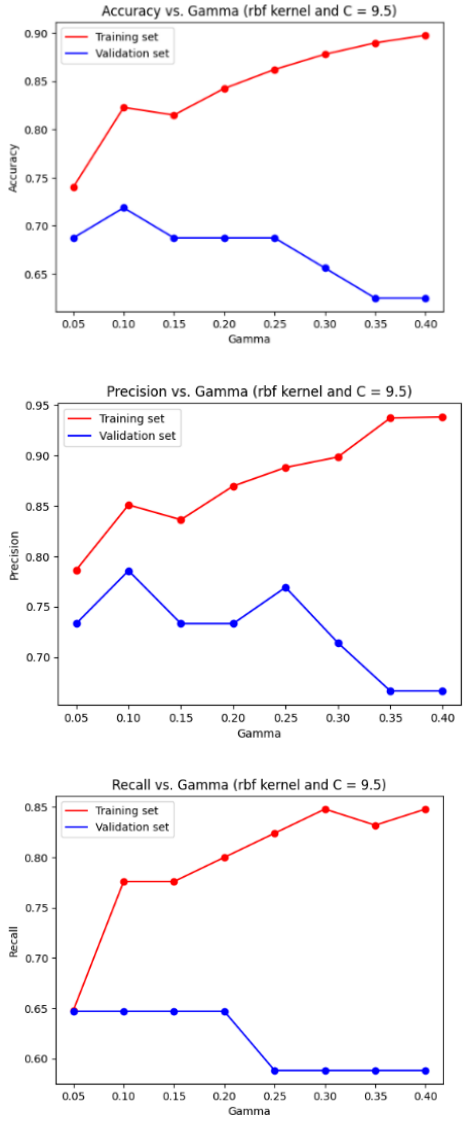



*Figure 27:: Gamma optimization graphs for (from left) accuracy, precision, and recall*

Lower values of gamma lead to higher metric values and less overfitting. The local accuracy and precision maxima at **Gamma = 0.1** indicate that this is a suitable value to choose.

### *9.5 Results*

*Table 10: Final results for performance metrics for training, validation, and testing*

| | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| **Training** | 0.823 | 0.851 | 0.776 | 0.812 |
| **Validation** | 0.719 | 0.786 | 0.647 | 0.710 |
| **Testing** | 0.625 | 0.619 | 0.765 | 0.684 |

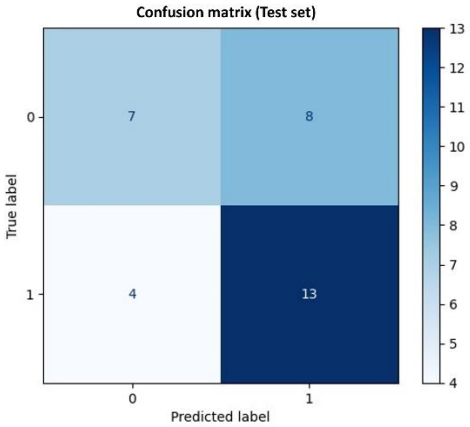



*Figure 28: Confusion matrix for test set*

### *9.6 Discussion*

The SVM model performed to a satisfactory extent with the test set. The accuracy of 0.625 is 24% lower than that of the training set, which is a considerable decrease, perhaps due to overfitting. However, the recall decreased by only 1.4%, indicating that it has a similar level of reliability in detecting when a student has failed compared to the training set. This is also seen in the confusion matrix of the test set, as the number of false negatives is half of that of false positives. Although the recall score increased by 18% from the validation set, the precision decreased by 17%, which resulted in a lower F1-score for the test set, further emphasizing its reduced reliability.

Within the school, this model should be used in tandem with other models to develop a holistic understanding of the student, because of this model's focus on social factors and its lower accuracy, the latter being significant given that it predicts Pass or Fail of their final year grade.

Increasing the number of samples would dramatically increase the reliability of this model as it would reduce the possibility of overfitting and hence would increase the values of the metrics. This would also amend the rapid fluctuation observed in the C-value optimization as it would create a smoother curve at higher values due to the reduction in overfitting. The kernel could be changed to the sigmoid kernel as it showed promise in the optimization. The dataset itself could be improved by increasing the number of entries from different schools, as that would reduce the bias and allow for better generalization of features.

## 10. CONCLUSION

In this study, Logistic Regression, Support Vector Machines, Decision Trees, and Random Forests were used to predict different grades for students based on factors affecting their performance. Among the models, Logistic Regression performed the best with the highest accuracy and recall of 88.9% and 85.7% respectively. However, generally, all models performed to a satisfactory extent, with metric values between 0.6 – 0.75. This can be explained through the correlation heatmap plotted for the entire dataset below, with most correlation values around 0.4 - 0.7.

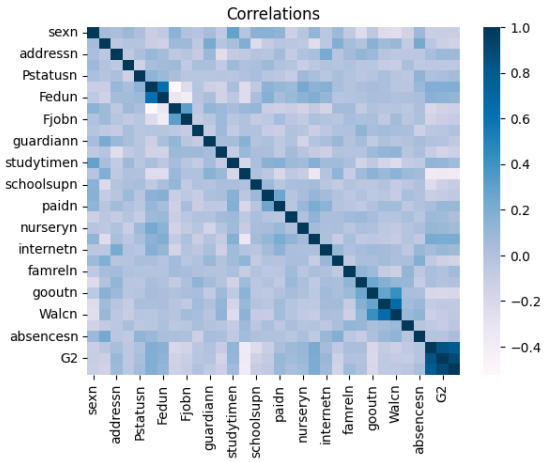



*Figure 29: Confusion matrix for test set*

Although it was determined that 'School support' and 'Going out' are key factors in affecting whether a student passes or fails, it was concluded that all models must be used in conjunction with each other to give a holistic

picture, given that each model has moderate accuracies and recalls.

An improved data set with a higher quantity and more diverse data from different schools would be imperative to reducing the possibility of overfitting. SVM models could be changed to use a Sigmoid kernel given the high performance of the logistic regression and tuned to achieve higher results.

## 11. REFERENCES


[1] *Considine, G., & Zappalà, G. (2002). The influence of social and economic disadvantage in the academic performance of school students in Australia. Journal of Sociology, 38(2), 129–148.* https://doi.org/10.1177/144078302128756543

[2] *Student performance* (no date) *UCI Machine Learning Repository*. Available at: https://archive.ics.uci.edu/dataset/320/student+performance (Accessed: 20 June 2023).

[3] europe, S. in (no date) *The Portuguese grading system*, *Study in Europe*. Available at: https://www.studyineurope.eu/study-in-portugal/grades (Accessed: 20 June 2023).

[4]. [Citation: Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5-32]

[5]. [Citation: Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In Ijcai (Vol. 14, No. 2, pp. 1137-1145)].

[6] . L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, "Classification and Regression Trees," Wadsworth, 1984.

[7] . J. H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001.

[8]. T. Hastie, R. Tibshirani, and J. Friedman, "The Elements of Statistical Learning: Data Mining, Inference, and Prediction," Springer, 2009.

[9] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[10]. G. James, D. Witten, T. Hastie, and R. Tibshirani, "An Introduction to Statistical Learning: with Applications in R," Springer, 2013.

[11]. T. Hastie, R. Tibshirani, and J. Friedman, "The Elements of Statistical Learning: Data Mining, Inference, and Prediction," Springer, 2009.

[12] Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. Machine Learning, 20(3), 273-297.

[13] Hsu, C. W., Chang, C. C., & Lin, C. J. (2003). A practical guide to support vector classification

[14] Cortez, P., & Silva, A. (2008). Using Data Mining to Predict Secondary School Student Performance. In Proceedings of 5th Annual Future Business Technology Conference. Porto, Portugal.

[15] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media.

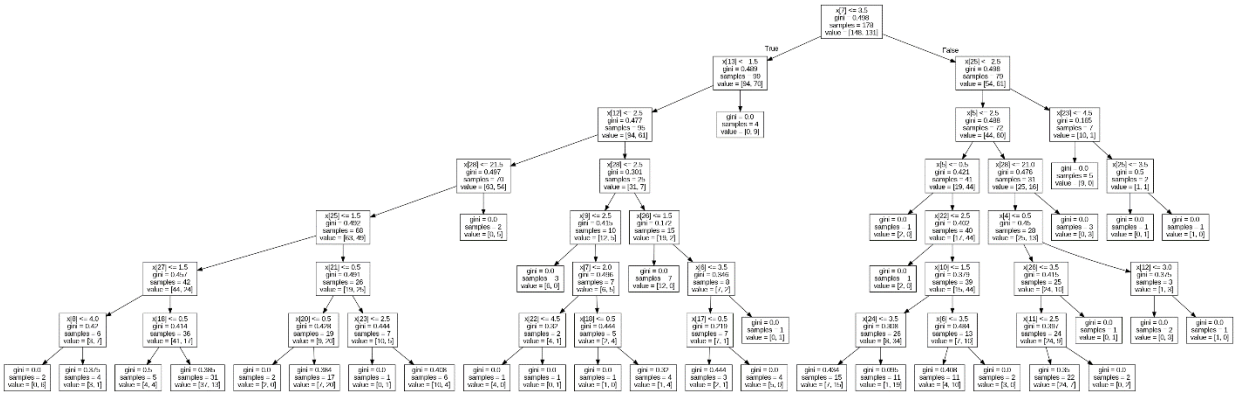
[16 ]Fleischmann, M., Hübner, N., Nagengast, B., & Trautwein, U. (2023). The dark side of detracking: Mixed-ability classrooms negatively affect the academic self-concept of students with low academic achievement. Learning and Instruction, 86, 101753. https://doi.org/10.1016/j.learninstruc.2023.101753

[17] Lavy, V. (2014) Social Networks and Human Capital. *University of Warwick.* https://warwick.ac.uk/fac/soc/economics/staff/vlavy/text_and_tables_social_netweorks_and_human_capital_april_28_2014.pdf

[18] Ansari, W. (2013) Is Alcohol Consumption Associated with Poor Academic Achievement in University Students? *IJPM*. https://www.csus.edu/faculty/m/fred.molitor/docs/is%20alcohol%20and%20academic%20achievement.pdf

[19[ Raj, Ashwin. (2022). *Everythin About Support Vector Classification – Above and Beyond.* https://towardsdatascience.com/everything-about-svm-classification-above-and-beyond-cc665bfd993e

## 12. APPENDIX



Example tree from random forest (iteration 4)