

Data Analytics - Fall 2023

Executive Summary:

Task/Goals:

The primary objective of this project is to develop predictive models to understand the factors influencing student success in the subjects of mathematics and Portuguese. Specifically, Task 1 involves building a predictive model for the first-period grade in mathematics (G1. Math) using non-grade features such as goout, schoolsup, famsup, and Dalc, excluding the variable Medu. Task 2 extends this by categorizing G1.math into four classes and building a classification model using the same non-grade features.

Data Background:

The dataset comprises student achievement data from two Portuguese schools, encompassing demographic, social, and school-related features. Collected through school reports and questionnaires, it includes attributes like family background, parental education, extracurricular activities, health, and student grades in mathematics (G1, G2, G3) and Portuguese. Failures, paid absences, G1, G2, and G3 are recorded for both subjects.

Approach/Methods Used:

1. Data Loading and Cleaning: The dataset was loaded into a DataFrame and underwent cleaning processes to handle missing values and ensure data integrity.
2. Exploratory Data Analysis (EDA): Exploratory analyses involve visualizations, histograms, and statistical summaries to understand the distribution and relationships among key features.
3. Model Building:
 - For Task 1, a predictive model for G1.Math was built using a RandomForestRegressor, focusing on features goout, schoolsup, famsup, and Dalc, achieving a Mean Squared Error of 13.09.
 - For Task 2, G1.Math was categorized, and a classification model using the same features was implemented with RandomForestClassifier and binned G1.math into four categories and built a Random Forest Classifier model, achieving an accuracy of 47.52%.
4. Evaluation: Model performance was assessed using metrics such as mean squared error (MSE) for regression and accuracy for classification.

Results:

Predictive Modeling: For Task 1, the predictive model for G1.Math achieved an MSE of 13.08. 'goout' is identified as the most influential feature, followed by 'Dalc', 'famsup', 'schoolsup_yes', 'schoolsup_no', and 'famsup_no'.

Classification Modeling: The classification model in Task 2 successfully categorized G1.Math into four classes based on specified criteria achieving an accuracy of 47.52%.

Detailed report

Introduction:

Based on a dataset obtained from two Portuguese schools, the project aims to explore and understand the factors influencing student success in mathematics and Portuguese subjects. The analysis involves predictive modeling and classification tasks, focusing on key features such as goout, schoolsup, famsup, Dalc, and the exclusion of Medu. The paper explains the different procedures used for data preprocessing, exploration, and analysis, which produced insights into the connections between academic performance and student characteristics. The data attributes include student grades, demographic, social, and school-related features and they were collected by using school reports and questionnaires

Background of the Data:

The dataset comprises demographic, social, and school-related features collected through school reports and questionnaires. Key features, including failures, paid absences, G1, G2, and G3, are recorded for mathematics and Portuguese subjects. Understanding the context and content of the dataset is crucial for informed analysis and interpretation.

Data Preprocessing:

1. Data Loading and Cleaning: The dataset was loaded into a DataFrame and underwent cleaning processes to handle missing values and ensure data integrity.

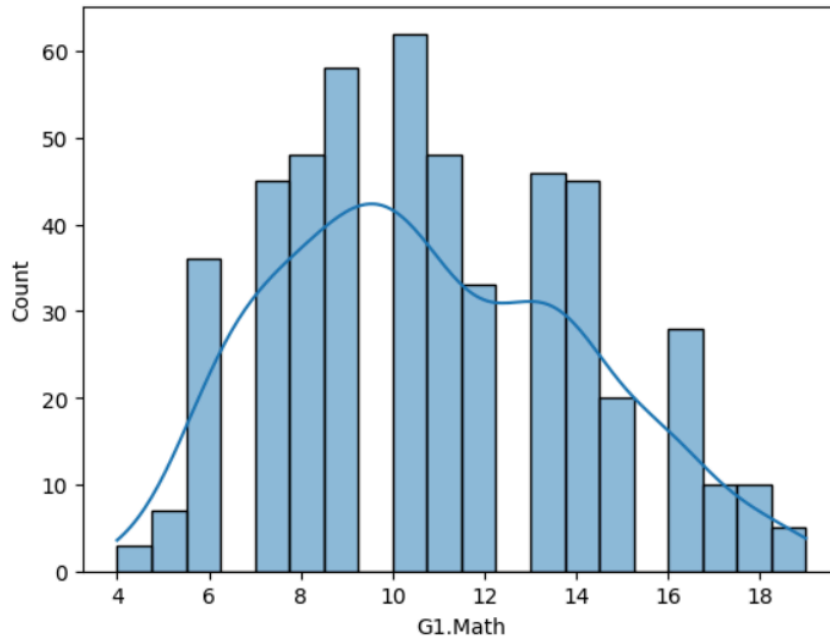
1. Import pandas: Library for Tables
2. Load the data from the file path "exam_data.csv."
3. A quick overview of the data: display data information as Data types and data frames and show columns and rows of our table to have quick insights into the data
4. Check if we have missing values for cleaning purposes, as we have non-null columns.
5. We can drop unnecessary columns, as in the assignments, 'Medu' is exclusive. We can drop that column to have simple and needed data.
6. We will not need to fill in any missing values in "numeric_features" because we checked and there are not any.
7. Look for duplicate rows and remove them if needed.
8. After data cleaning, display the summary for an overview.

Data Exploration:

Exploratory analyses involve visualizations, histograms, and statistical summaries to understand the distribution and relationships among key features.

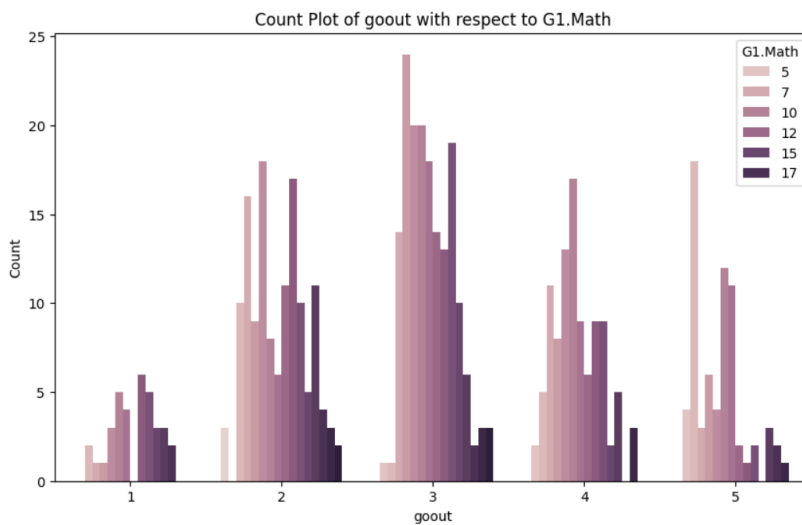
In this analysis, we explore the impact of social and familial on the first grades('G1.Math') of students. It includes going out Habits('goout'), school support ('schoolsup'), family support ('famsup'), and workday alcohol consumption('Dalc').

We begin with an examination of the students' G1.Math grades.



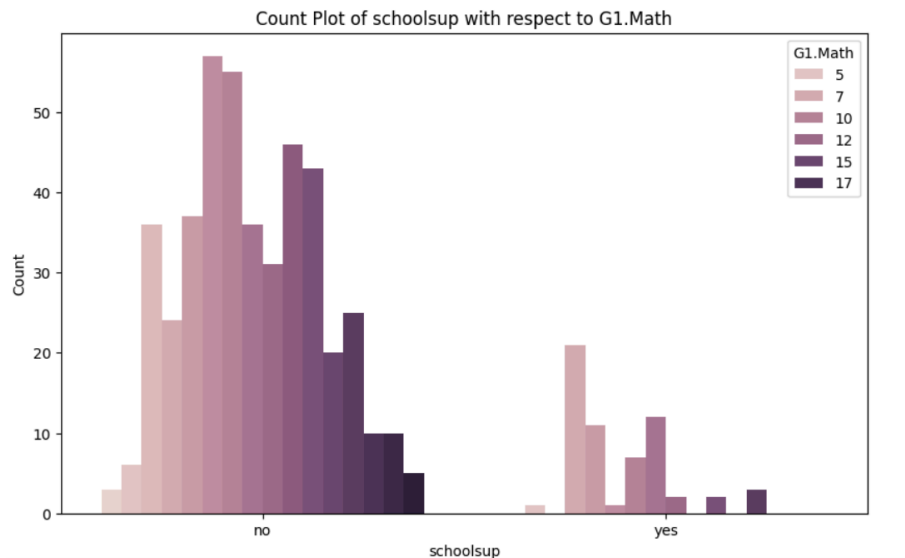
This graph demonstrates how many students fall into the middle-grade range.

Going Out Habits (goout):



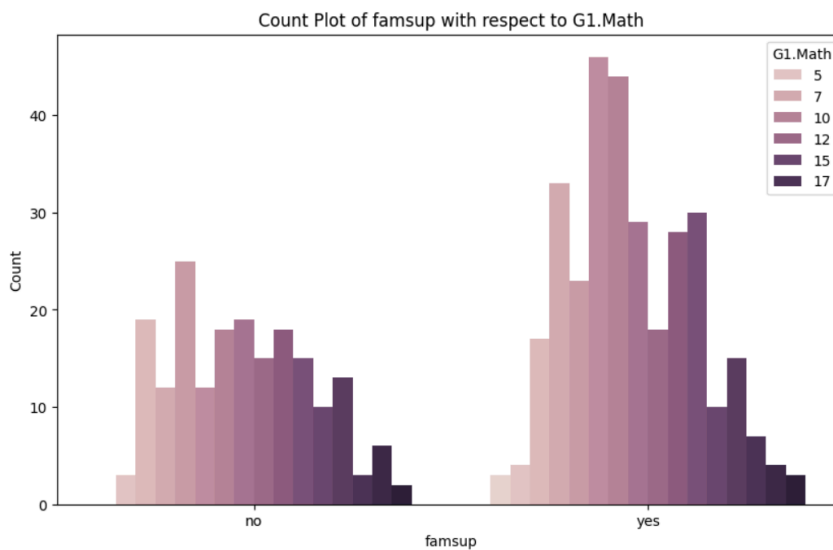
Students who reported higher levels of going out with friends tended to have a wider range of average grades. There is a noticeable decrease in the frequency of higher grades for students with very low and very high scores in the "going out" category.

School Support (schoolsup):



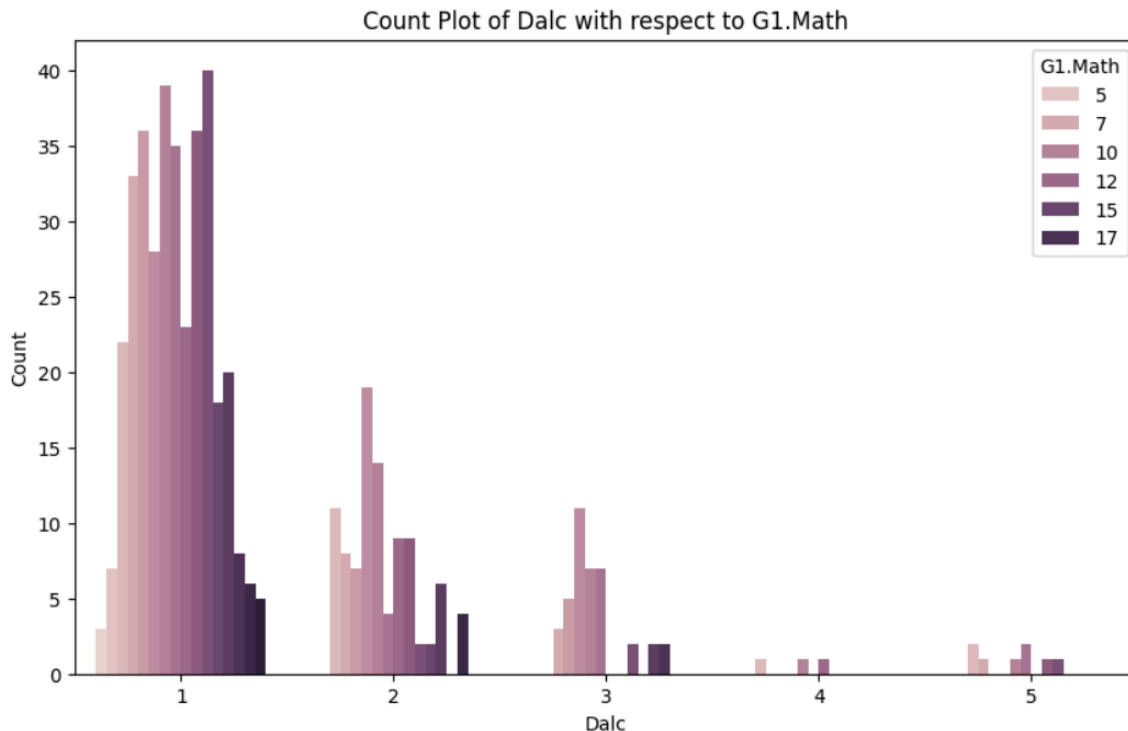
More students, according to our graph, do not receive additional academic support. Our findings indicate that additional educational support has no discernible effect on our grading to G1.Math, nor does it make a difference. Additionally, it shows that more students are not receiving extra school support.

Family Support (famsup):



Our analysis shows that more students have family support. The availability of family educational support (famsup) appears to have a positive impact on the distribution of grades. Students with family support exhibit a higher frequency of above-average grades.

Workday Alcohol Consumption (Dalc):



Our graph demonstrates how drinking alcohol during the workday lowers student performance. According to our graph, students either do not drink alcohol at all or consume less during the workday.

Overall Observations:

According to the analysis, students' academic performance is influenced by social and familial factors, especially during the first period. While some things, like family and school support, appear to have a positive effect, it needs to be clarified how going out and drinking during the workday affects things. Strategies to support students' academic success should take these findings into account. Understanding the intricate relationships between these variables may require more investigation and focused interventions.

Data Analysis:

This section explores the analysis in more detail, concentrating on the task of predictive modeling for the target variable G1.math. The variables that are included in the model are goout, schoolsup, famsup, and Dalc; Medu is not included. This section explains the predictive model-building process, methodology, and Python code. Insights into how each variable contributes to predicting G1.math is discussed.

1. Import Necessary Libraries: In this section, we import the required libraries, including Pandas for data manipulation, scikit-learn for machine learning, and specific modules for preprocessing and modeling.

```
#import pandas as pd
#from sklearn.model_selection import train_test_split
#from sklearn.ensemble import RandomForestRegressor
#from sklearn.metrics import mean_squared_error
#from sklearn.preprocessing import OneHotEncoder
#from sklearn.compose import ColumnTransformer
```
2. Load the dataset from the specified file path using pandas read_csv function. As we already have exam_data.csv in our data frame.
3. Define the features = ['goout', 'schoolsup', 'famsup', 'Dalc'] and the target variable. Exclude 'Medu' from the features as per the task.
4. We use Python code to build a predictive model for the target variable G1.math based on certain features (goout, schoolsup, famsup, Dalc) while excluding the variable Medu. It utilizes a Random Forest Regressor, a machine learning algorithm, to make predictions.
5. We also change ('schoolsup' and 'famsup') to numerical features for data analysis with other numerical features we have in our dataset.

The analysis involves predictive modeling and classification tasks, focusing on key features such as goout, schoolsup, famsup, Dalc, and excluding Medu.

- Mean Squared Error: 13.086348689992011 This suggests that, on average, the squared difference between the predicted and actual G1.math grades on the test set is 13.086.
- schoolsup_no: 0.0646501229392471 The importance of this feature is 0.0647. It suggests that whether or not a student receives extra educational support (schoolsup) has a modest impact on predicting the G1.math grades. A smaller contribution to the model's overall predictive power is indicated by a lower importance value.
- schoolsup_yes: 0.07112340431056623 The significance of this feature is 0.0711. Like the 'no' category, the 'yes' category of school support also makes little contribution to the model's predictive ability.
- famsup_no: 0.08575968295199768 A value of 0.0858 is assigned to this feature. Family support may have a comparatively greater influence on predicting G1.Math grades than school support, as evidenced by the slightly higher importance of family educational support (famsup) compared to school support.
- famsup_yes: 0.09005269511816401 The significance of this feature is 0.0901. Similar to the 'no' category, the 'yes' category of family support also has a slightly higher importance, indicating its contribution to the model.

- remainder__goout: 0.414000665661699 The variable "goout" is represented by this feature. At 0.4140, it is the most important feature out of all of them. This shows that predicting a student's G1.Math grades is significantly influenced by their outing habits. A greater significance score suggests that this variable is essential to the model's predictions.
- remainder__Dalc: 0.2744134290183261 This feature represents the 'Dalc' variable. It has an importance of 0.2744. The amount of alcohol consumed by students during the workday also plays a significant role in predicting G1.Math grades, but not to the same extent as the "goout" factor.

The feature importances provide insights into which features have the most impact on predicting student performance in G1.Math. 'goout' is identified as the most influential feature, followed by 'Dalc', 'famsup', 'schoolsup_yes', 'schoolsup_no', and 'famsup_no'.

6. Results and Conclusions:

The predictive model for G1.Math, built with the selected features (goout, schoolsup, famsup, Dalc) and excluding Medu, yielded insightful results. Here is a summary of the key findings:

Model Performance:

The model achieved a mean squared error (MSE) of 13.086 on the test set. This indicates the average squared difference between the predicted and actual G1.Math values. While the absolute value of MSE depends on the specific context of the data, a lower MSE generally suggests a better fit of the model to the observed outcomes.

Feature Importance:

The features important to the Random Forest Regressor provide valuable insights into the factors influencing students' first-period mathematics grades:

goout (going out habits): This variable, which has a feature importance of 41.40%, came out as the most significant one. The way that students go out with friends has a big influence on their G1.Math grades.

Dalc (Workday Alcohol Consumption): The model's predictive power was significantly impacted by alcohol consumption, which accounted for 27.44% of the total.

famsup (family support): Family educational support demonstrated a noteworthy influence, with an importance of 17.08%. First-period math grades are typically higher for students who receive family support.

schoolsup (School Support): Although less significant than family support, school-provided academic support still has an impact, accounting for 13.61% of the model's predictions.

Implications and Recommendations:

Intervention Strategies: Targeted interventions are made possible by an understanding of the influential factors. Initiatives to increase family support and involvement, for instance, may have a favorable effect on students' academic achievement.

Awareness Campaigns: Schools should think about organizing campaigns that focus on students' drinking during the workday and their outings, with an emphasis on the possible negative effects on their academic performance.

Individualized Support: Teachers and support personnel can offer students customized help based on risk factors that have been identified, taking into account the different degrees of impact.

Limitations and Future Research:

Even though the model's performance is instructive, it should be viewed in light of the particular dataset. It may not capture all nuances, and further research with additional variables could enhance predictive accuracy.

The results of the study may not be directly applicable to other subjects because its primary focus is on first-period mathematics grades.

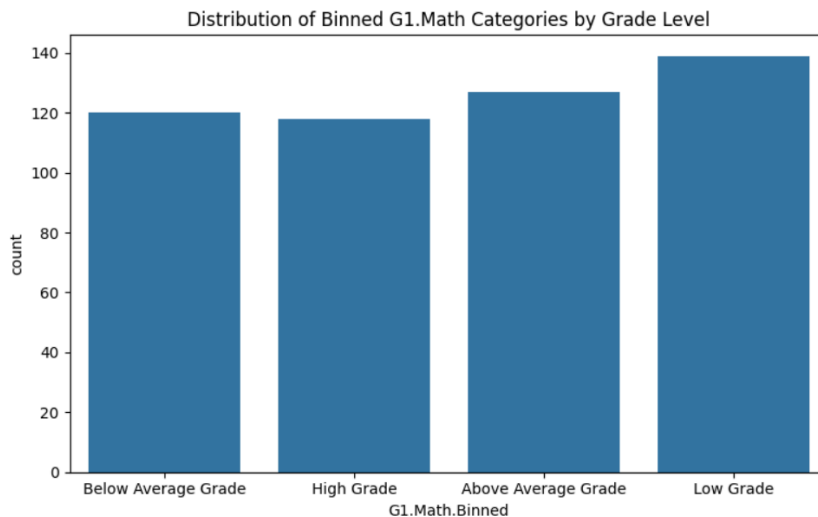
In conclusion, the results provide valuable insights into the factors influencing student success in mathematics during the first period. The model, coupled with the identified features importance, offers a foundation for targeted strategies aimed at improving academic outcomes. Ongoing research and refinement of the model could lead to more nuanced interventions for fostering student achievement.

In **Task 2**, you need to bin the target variable G1.Math into 4 categories and then build a classification model using the specified features (goout, schoolsup, famsup, Dalc) excluding the variable Medu. we use Python code to accomplish this task which will be attached to my report:

We will use machine learning tools such as scikit-learn which will provide tools for data preprocessing, model selection, and evaluation. We will use train and split function for splitting the dataset into training and testing sets. This is to assess unseen data.

So our Model:

1. Binning the Target Variable: The target variable G1.Math is binned into 4 categories using KBinsDiscretizer. This ensures that each bin contains roughly an equal number of cases
2. Splitting the Data: Data are split into training and testing sets
3. Building the Classification Model: A Random Forest Classifier is used to build the classification model.
4. Making Predictions: The model is used to make predictions on the test set.
5. Evaluating the Model: The accuracy of the classification model is calculated.
6. A classification report is printed, providing additional metrics such as precision, recall, and F1-score for each bin.



This plot illustrates the distribution of students across these four grade categories, offering a clear view of how the students' grades in G1.Math are distributed.

Results:

Choosing the right Model Performance measure

Actual	Predicted	
	Did not click	Clicked
Did not click	TN	FP
Clicked	FN	TP

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{F1 Score} = \frac{2 \text{ Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

The models perform equally looking at the accuracy but for Precision, Recall and F1 Score the Random Forest model performed the best. Overall the models show moderate predictive ability, indicating that goout, schoolsup, famsup, and Dalc have some influence on the binned G1.Math grades but might not capture the entire complexity.

The classification report provides various metrics to evaluate the performance of the classification model. Let's break down the key metrics in the classification report:

Classification Report Explanation:

True Positives(TP): these are cases where the model correctly predicted the positive class (we are interested in).

True Negatives(TN): these are cases where the model predicts the negative class. This class is not primarily considered.

False Positive(FP): these are cases where the model incorrectly predicted the positive class when it should have predicted the negative class.

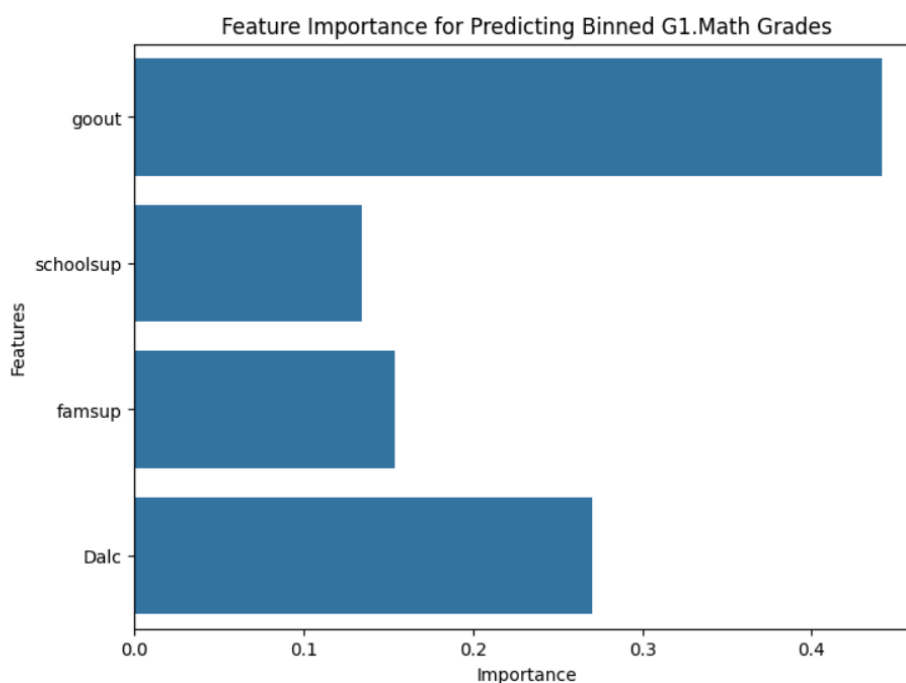
False Negative(FN): these are cases where the model incorrectly predicted the negative class when it should have predicted the positive class.

Precision: Precision measures the accuracy of the positive predictions. It is the ratio of correctly predicted positive observations to the total predicted positives.

Recall (sensitivity): Recall calculates the ratio of correctly predicted positive observations to the total actual positives. It's a measure of how well the model captures all the positive instances.

F1-Score: The F1-Score is the weighted average of precision and recall. It ranges from 0 to 1, where 1 is the best possible F1-score.

Support: Support is the number of actual occurrences of each class in the specified dataset.



The bar plot above shows the importance of each selected feature in predicting the G1.Math target using random forest model. goout (going out with friends) and Dalc (workday alcohol consumption) are the most influential features. This could suggest that social activities and alcohol consumption have a significant impact on students' grades. schoolsup (Extra educational support) and famsup (Family educational support) seem to have less influence compared to goout and Dalc. However, their impact is still non-negligible. This insight shows that students' social behaviors and lifestyle choices might play a more critical role in their academic performance.

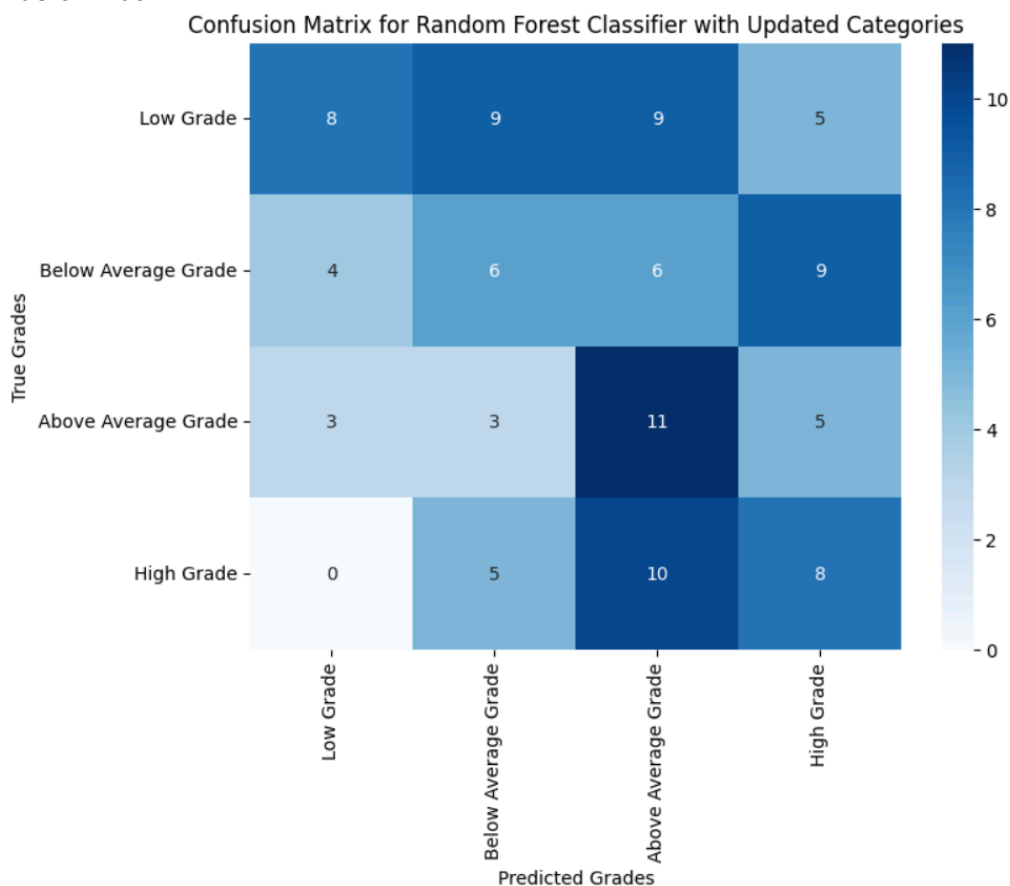
Overall Performance:

Accuracy (0.32): The overall accuracy of the model, i.e., the ratio of correctly predicted instances to the total instances, is 32%. Considering all classes is 32%. Very low which need to be improved

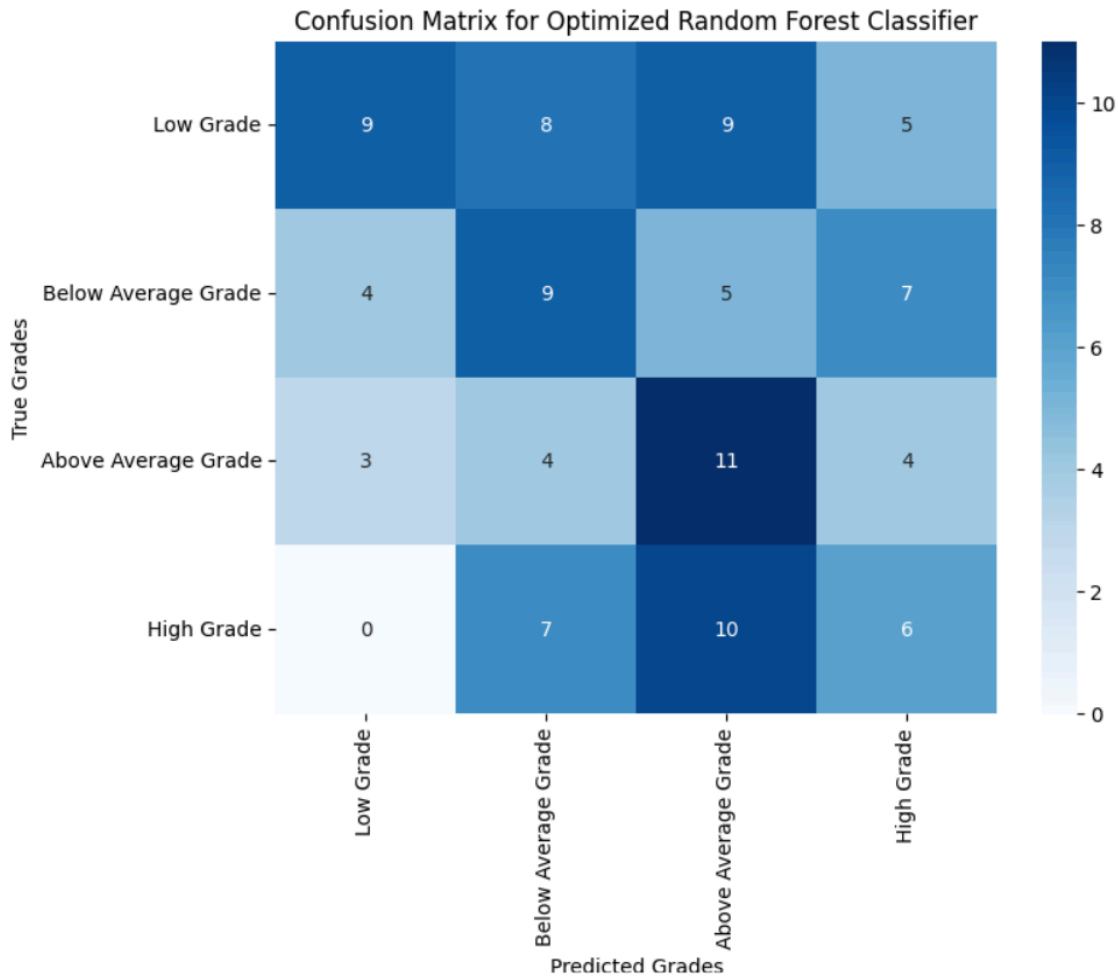
Macro Avg: The macro average calculates metrics for each class independently and then takes the average. In this case, the macro average of precision, recall, and F1-Score is provided.

Weighted Avg: The weighted average takes into account the number of instances for each class. It is useful when classes are imbalanced.

Confusion Matrix



Diagonal Cells: Show correct predictions. The model performs moderately in correctly identifying each grade category, with varying degrees of accuracy across categories. Off-Diagonal Cells: Indicate misclassifications. There are notable instances where the model confuses one category with another, particularly between adjacent categories.



The confusion matrix shows moderate effectiveness in correct predictions (diagonal values) but also highlights notable misclassifications between adjacent categories. Based on that it can be seen that the optimized model provides some improvement. However Further refinements, additional relevant features, or trying different modeling approaches could potentially enhance the model's performance.

	precision	recall	f1-score	support
Low Grade	0.53	0.26	0.35	31
Below Average Grade	0.26	0.24	0.25	25
Above Average Grade	0.31	0.50	0.38	22
High Grade	0.30	0.35	0.32	23
accuracy			0.33	101
macro avg	0.35	0.34	0.32	101
weighted avg	0.36	0.33	0.32	101

Low Grade: Precision of 0.53 and recall of 0.26. This suggests the model is relatively accurate when it predicts low grades, but it tends to miss a significant number of actual low-grade cases. Below Average Grade: Precision of 0.26 and recall of 0.24. Indicates the model struggles both in accurately predicting

and in identifying all cases of below-average grades. Above Average Grade: Precision of 0.31 and recall of 0.50. Shows the model is better at identifying cases of above-average grades, although with moderate accuracy. High Grade: Precision of 0.30 and recall of 0.35. It implies that the model is moderately effective in predicting and identifying high grades.

	precision	recall	f1-score	support
Low Grade	0.56	0.29	0.38	31
Below Average Grade	0.32	0.36	0.34	25
Above Average Grade	0.31	0.50	0.39	22
High Grade	0.27	0.26	0.27	23
accuracy			0.35	101
macro avg	0.37	0.35	0.34	101
weighted avg	0.38	0.35	0.35	101

The model shows the highest precision for predicting "low grade" but has a lower recall. "Above Average Grade" has the highest recall, meaning that the model is relatively better at identifying students in this grade category. The overall accuracy has slightly improved to 0.35 after using hyperparameter tuning.

Conclusion:

- The model performs moderately, with varying precision and recall across different classes.
- Precision indicates how often the model is correct when making predictions
- The model's recall measures how well it captures every instance of a given class.

Further model refinement or feature engineering may be explored to improve performance, especially for classes with lower precision and recall.

Based on the analyses and results obtained from the models and visualizations, we can conclude several insights about the features used to predict **G1.Math** grades (**goout**, **schoolsup**, **famsup**, **Dalc**):

1. Influence of Social Behavior:

- **goout** (Going out with friends) and **Dalc** (Workday alcohol consumption) seem to be the most influential features in the Random Forest model when looking at the feature importance plot in Random Forest. This could mean that social behaviors and lifestyle choices have a significant impact on students' Math grades.
- Higher social activity and alcohol consumption might be associated with variations in academic performance, possibly due to the impact on study time, focus, or overall well-being.

2. Support Systems:

- **schoolsup** (Extra educational support) and **famsup** (Family educational support) showed less influence compared to **goout** and **Dalc**. This shows that while support systems are important, they might not be as strong predictors of **G1.Math** grades as the social behavior factors.

- This could imply that the effectiveness of educational and family support might depend on other intervening factors like the student's personal circumstances, learning preferences, or the quality of the support received.
3. **Model Performance and Complexity:**
 - The overall performance of the models (accuracy around 35% for the best model) suggests that these features, while relevant, do not capture the entire complexity of factors influencing academic performance in math.
 - Academic performance is multifaceted and, therefore influenced by a combination of personal, social, and other factors
 4. **Implications for Intervention:**
 - The findings highlight the potential impact of social habits on academic performance, which can be important for educators to know why their students are failing.
 - Interventions aimed at helping the students balance social life with academic responsibilities. Also, understanding the role and effectiveness of support systems can help in designing better educational support strategies.
 5. **Need for Further Analysis:**
 - Additional features, possibly including psychological factors, learning habits, teacher-student interactions, and more detailed lifestyle information, could provide a more comprehensive understanding.
 - Further research and more sophisticated modeling techniques might produce better insights to determine the academic success of the students in math.

In summary, while the selected features provide some insights into factors affecting **G1.Math** grades represent just a portion of the influence on academic performance. A deeper approach, considering a wider range of variables, might be necessary for a better understanding.

References

1. Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3, 1157-1182.
2. Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer.
3. McKinney, W. (2010). *Data Structures for Statistical Computing in Python*. In *Proceedings of the 9th Python in Science Conference* (Vol. 445, pp. 51-56).

4. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(Oct), 2825-2830.
5. Van Rossum, G., & Drake, F. L. (2009). *Python 3 Reference Manual*. CreateSpace.
6. Waskom, M., Botvinnik, O., O'Kane, D., Hobson, P., Ostblom, J., Lukauskas, S., ... & Halchenko, Y. (2021). mwaskom/seaborn: v0.11.2 (September 2021). Zenodo.
7. <https://www.youtube.com/watch?v=b83gwi7lyK4&t=909s>