Data Analytics Final Project Report

Fall 2023

Dataset:

student achievement in secondary education of two Portuguese schools

Target Variable:

G1.Math

Author: Nastaran Mesgari,
Date: 2023/12/06

Prof. Adalbert F.X. Wilhelm

Executive Summary

This report is provided for the data analytics project with the goal of performing a reasonable analysis of the data given. The general task of your project is to understand what impacts success in the two subjects Mathematics and Portuguese, from student achievement in secondary education of two Portuguese schools. The data is in 948 rows and 40 columns. The target variable is 'G1. Math'. The algorithm that is used for machine learning model is the regression and for task 2 after categorizing the grade of math we use the classification model.

After loading the data and understanding each variable, I cleaned the data and removed the garbage variables. Then I transformed 'famsup' related variables into yes/ no, type 'object' and created new variables as an int and replace yes with 1 and no with zero, and check the categorical variable , we transform to dummy variable and change the type of them to integer for using correlation between variables and G1.math , and before that we check the standard deviation and garbage value and null value for cleaning dada (preprocessing) after that we virtualizing the data for better deciding and finding the data . In the next step, I checked the linear correlation of the variables. I performed simple data exploration using visualization for all the features. I define a new data frame "G1Math_df2" with 'G1.Math', 'absences.Port', 'famsup', 'Walc', 'failures.Port', 'famrel','higher_yes', 'studytime', 'sex_M', 'failures.Math', 'Mjob_other','Fjob_other', 'Fjob_teacher', 'schoolsup_yes'

After visualizing the distributions and outliers, I chose my feature variables and split my data set into test data and train data. Then I selected" Random Forest classifier "among 6 algorithms by accuracy in task 1: 84% and task 2 after categorizing into 3 for target variable G1. math accuracy improved, and it is 90%.

and the ROC-AUC was 0.95 for class 2.0, 0.94 for class 0.0, and 0.89 for class 1.0 but with improving the model using the l2 penalty (which is used in Ridge regression) and removing the variables with small coefficient in the model without the task1 should be use in the model drop (schoolsup_yes) and check the model. 1 percent changes so I decided don't use this column in our prediction.

An increase in AUC signifies an improvement in the model's ability to differentiate between classes, while an increase in accuracy indicates an overall enhancement in the model's ability to correctly identify instances. However, paying attention to both metrics can provide a more comprehensive understanding of the model's performance.

Table of Content

Contents

1 Introduction

Data analytics is the collection, transformation, and organization of data to conclude, make predictions, and drive informed decision-making. The data in the file is about student achievement in secondary education at two Portuguese schools. The data attributes include student grades, demographic, social, and school-related features and it was collected using school reports and questionnaires. Therefore, the algorithm used for the machine learning model is based on supervised and there are regression or classification. Classification algorithms utilize input training data to predict the likelihood or probability that the data that follows will fall into one of the predetermined categories.

After loading the data and understanding each variable, I cleaned the data and performed data exploration using visualization. In the next step, I process models to find the best model for this type of output. We have discrete variables that are similar to continuous variables due to the small distance between them. In the end, according to these data, we checked whether by classifying the information, our prediction of students' grades would improve or not and whether the accuracy measurement would be checked.

.

.

2 Data Set

2.1 student achievement Data set

The dataset is provided in ‘csv’ type. After importing the data set, the first step is getting some information about the shape of the data set and variables.

The data in the file is about student achievement in secondary education of two Portuguese schools. The data attributes include student grades, demographic, social and school related features and it was collected by using school reports and questionnaires. There are 948 rows and 40 Columns. (dtypes: float64(1), int64(18), object(18))*

The features of the raw data set with their corresponding descriptions are as below:

The features failures, paid, absences, G1, G2, G3 are recorded for the Math subject and the Portuguese subject, hence a corresponding suffix has been added to the variable name.

Column	Dtype	Description	Values
Unnamed	float64	Index of the dataset	
school	object	student’s school	“GP” - Gabriel Pereira or “MS” - Mousinho da Silveira
sex	object	student’s sex	“F” - female or “M” - male
age	int64	student’s age	numeric: from 15 to 22
address	object	student’s home address type	“U” - urban or “R” - rural
famsize	object	family size	“LE3” - less or equal to 3 or “GT3” - greater than 3
Pstatus	object	parent’s cohabitation status	“T” - living together or “A” - apart
Medu	int64	mother’s education	numeric: 0 - none, 1 - primary education, 2 – 5th to 9th grade, 3 – secondary education, 4 – higher education
Fedu	int64	father’s education	numeric: 0 - none, 1 - primary education, 2 – 5th to 9th grade, 3 – secondary education, 4 – higher education
Mjob	object	mother’s job	“teacher”, “health” care related, “services”, “at_home” or “other”

Fjob	object	father’s job	“teacher”, “health” care related, “services”, “at_home” or “other”
reason	object	reason to choose this school	“home”, “reputation”, “course” preference or “other”
guardian	object	student’s guardian	“mother”, “father” or “other”
traveltime	int64	home to school travel time	numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, 4 - >1 hour
studytime	int64	weekly study time	numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, 4 - >10 hours
failures	object	number of past class failures	numeric: n if 1<=n<3, else 4
schoolsup	object	extra educational support	yes or no
famsup	object	family educational support	yes or no
paid	object	extra paid classes within the course subject	yes or no
activities	object	extra-curricular activities	yes or no
nursery	object	attended nursery school	yes or no
higher	object	wants to take higher education	yes or no
internet	object	Internet access at home	yes or no
romantic	object	with a romantic relationship	yes or no
famrel	int64	quality of family relationships	numeric: from 1 - very bad to 5 - excellent
freetime	int64	free time after school	numeric: from 1 - very low to 5 - very high
goout	int64	going out with friends	numeric: from 1 - very low to 5 - very high
Dalc	int64	workday alcohol consumption	numeric: from 1 - very low to 5 - very high
Walc	int64	weekend alcohol consumption	numeric: from 1 - very low to 5 - very high
health	int64	current health status	numeric: from 1 - very bad to 5 - very good
failures.Math	int64	number of past class failures	
paid.Math	object	extra paid classes within the course subject	
absences.Math	int64	number of school absences	
G1.Math	int64	first period grade	
G2.Math	int64	second period grade	
G3.Math	int64	final grade	
failures.Port	int64	number of past class failures	
paid.Port	object	extra paid classes within the course subject	
absences.Port	int64	number of past class failures	
G1.Port	int64	first period grade	
G2.Port	int64	second period grade	
G3.Port	int64	final grade	
dtypes: float64(1), int64(21), object(18)			

3 Data Pre-Processing

to use data, we need to import them and read the data. In this case, our data is CSV files

3.1 Data Cleaning

in this step we import the main libraries that we need, and it is in the folder whose name is data.

the steps followed for the data set is given below:

3.2 Dropping unnecessary columns and rows

dropping unnecessary columns and rows is a data preprocessing step that involves removing specific columns or rows from a dataset that are deemed unnecessary for the analysis or modeling task at hand. This process is beneficial for several reasons. Reducing Dimensionality, Improving Computational Efficiency, Enhancing Model Performance, and so on.

At this stage, I check the data, and in this step, we have done the things the task asks us so this model must contain the variables absences. Port, famsup, Walc, failures. Port, studytime, famrel but not the variables Fedu, Medu, age.

we do not have any value for replacing with missing data if we have we can replace it or drop it.

3.3 Checking for missing values:

In most cases, we do not get complete datasets. They either have some values missing from the rows and columns or they do not have standardized values.

So, before going ahead with the analysis, it is a good idea to check whether the dataset has any missing values.

in this step we check the missing value and there are no missing value and the row does not need to drop.

3.4 Checking for garbage values

Garbage value is generally a term meaning that the value in a variable doesn't have some sort of planned meaning.

By checking the statistical information of the data, some variables have negative values, and some have 0 values which are not compatible with the definition (corresponding to the dataset).

The detail of these values is given in the following tables:

*** Negative Values and Zero for deleting ***

by this code we can check the data for minus and zero if it is not compatible by the meaning they have.

3.4.1 Checking the distribution of each variable

Checking the distribution of each variable involves examining the spread and pattern of values within individual columns or features in a dataset. Understanding the distribution helps you gain insights into the central tendencies, variability, and shape of the data. This is crucial for making informed decisions during data analysis and modeling. Common statistical measures used to describe the distribution include mean, median, and standard deviation.

In this phase, first we checked the numeric variables with zero variance (threshold = 0), *they do not have any contribution on the model*.at this data we do not have any standard deviation equal to zero. so this step we didn't drop anything.
only drop some columns such as unnamed (for indexing) and some column that in the task announced to drop it such as Grad in port and math.

.

3.5 Data Transformation

3.5.1 Transforming the categorical variables

If we have a column that is object for example yes or no question or if we have Boolean, we can convert them to integer.in this stage I created the new column.

in this step we change type of 'famsup' from object to integer for using the model.

```
Encoding for famsup: 1 for 'yes', 0 for 'no'
```

3.5.2 Normalization, standardization, scaling

The data normalization process lowers the scale and brings all the data-points on the same scale.

Normalization: It involves scaling the values of a variable to a specific range, usually between 0 and 1

Standardization (Z-score normalization): It transforms the data to have a mean of 0 and a standard deviation of 1.

Scaling: is a general term for any transformation that alters the range of the data.

I use scaling in the model selection.

4 Data Exploration,

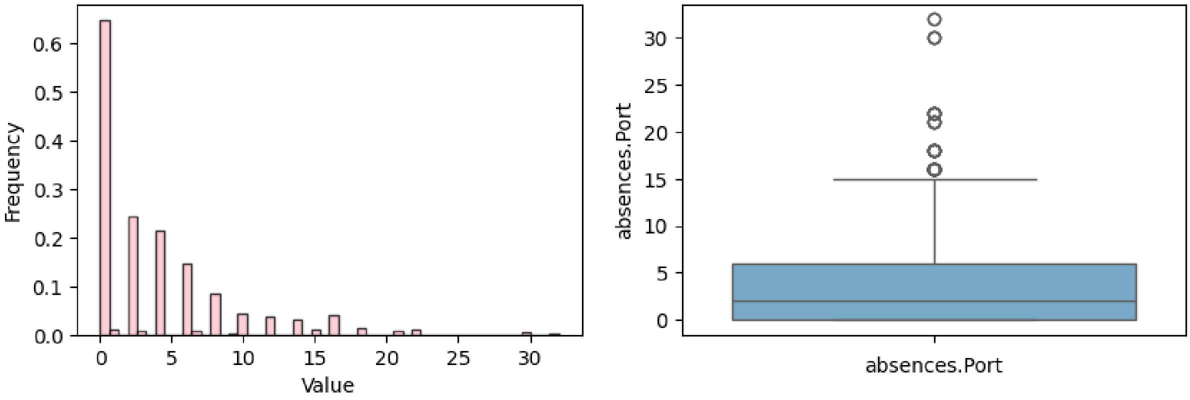
We put our visualization here!

4.1 Finding outliers and dummy variable

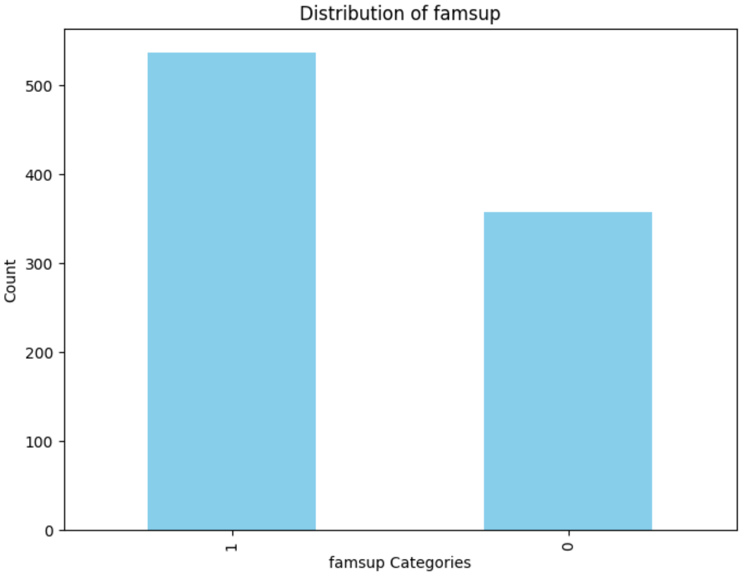
Finding outliers in a dataset involves identifying data points that significantly differ from the majority of the data. The use of outlier detection methods depends on the type of data and the analysis objective. However, generally, these methods are commonly applied to numerical columns or continuous variables. The reason for this is that the concept of outliers is more definable in continuous variables, and statistical measures such as mean, standard deviation, box plots, can easily be employed for their identification.

for founding the outliers of discrete and categorical variables we need to find the type of variables are integer or objects. We find the outliers and the Q1, Q3 and compare it with data and count how much of each independent variable out of this range and recognize and virtualize it .

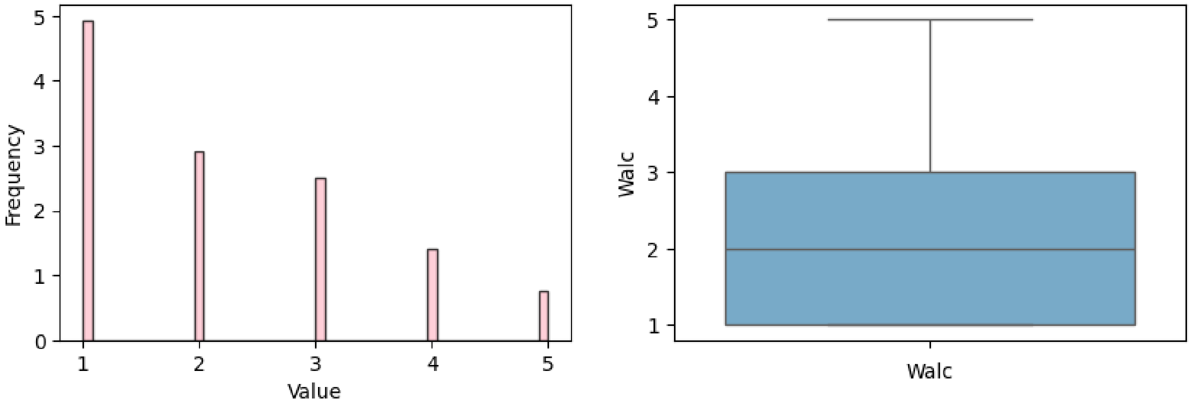
For each integer variable, I use boxplot and histogram for visualization. The plots are as bellows:



absences. Port is right skewed. There are 894 observations which their absences. Port is less than 15 miles.



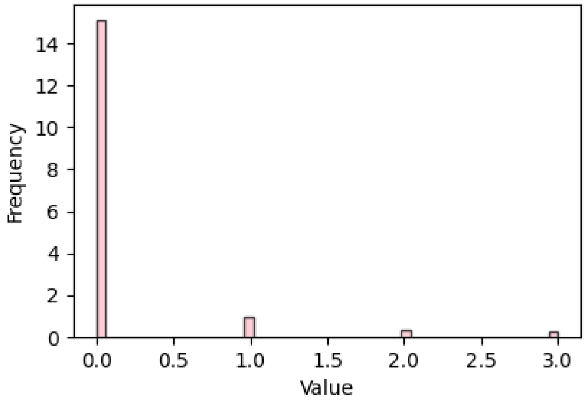
There are 537 observations is 1, and 357 observations is 0 we use all of the data because there isn’t any outlier value.

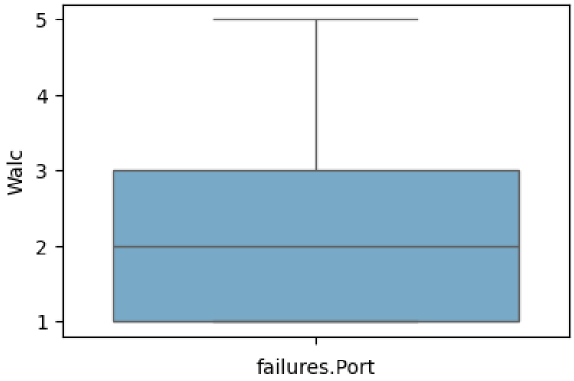


in this variable we have observation Walc

```
1    353
2    208
3    179
4    100
5     54
```

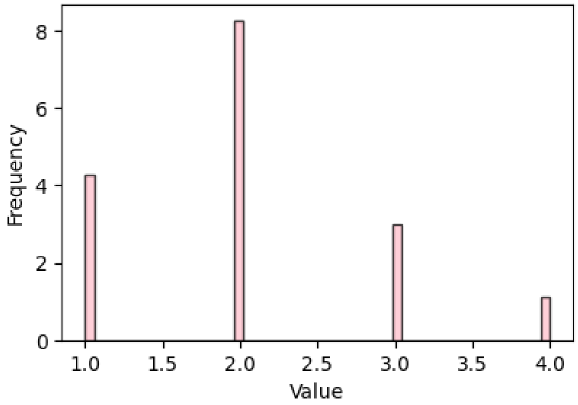
Name: count, dtype: int64 and you can see all of them are in quartile.

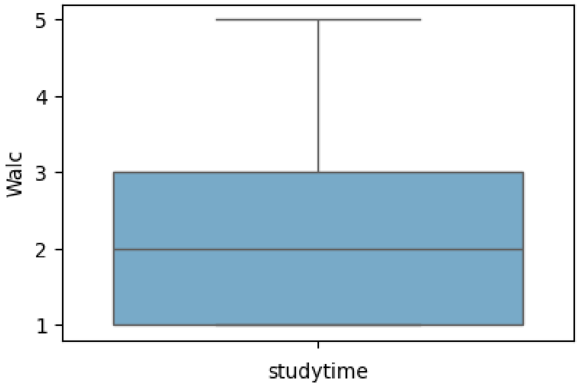




```
failures. Port
0    812
1     52
2     17
3     13
```

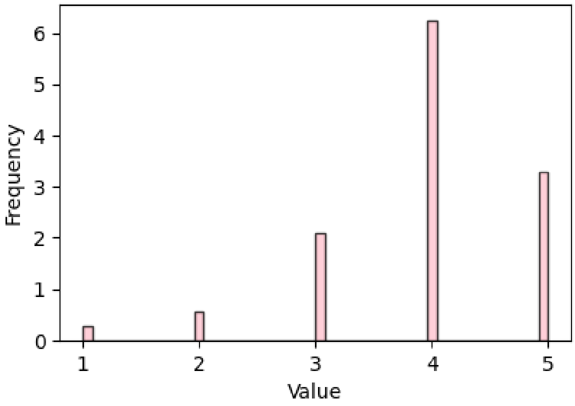
Name: count, dtype: int64 and most of the value are in the 0 = 812 but

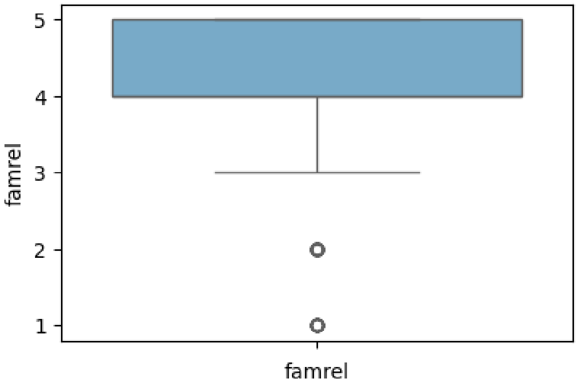




```
studytime
2    444
1    230
3    160
4     60
```

Name: count, dtype: int64. The 2 for studying have the biggest range.





There are 833 observations greater than 2.5 and less than 6.5.

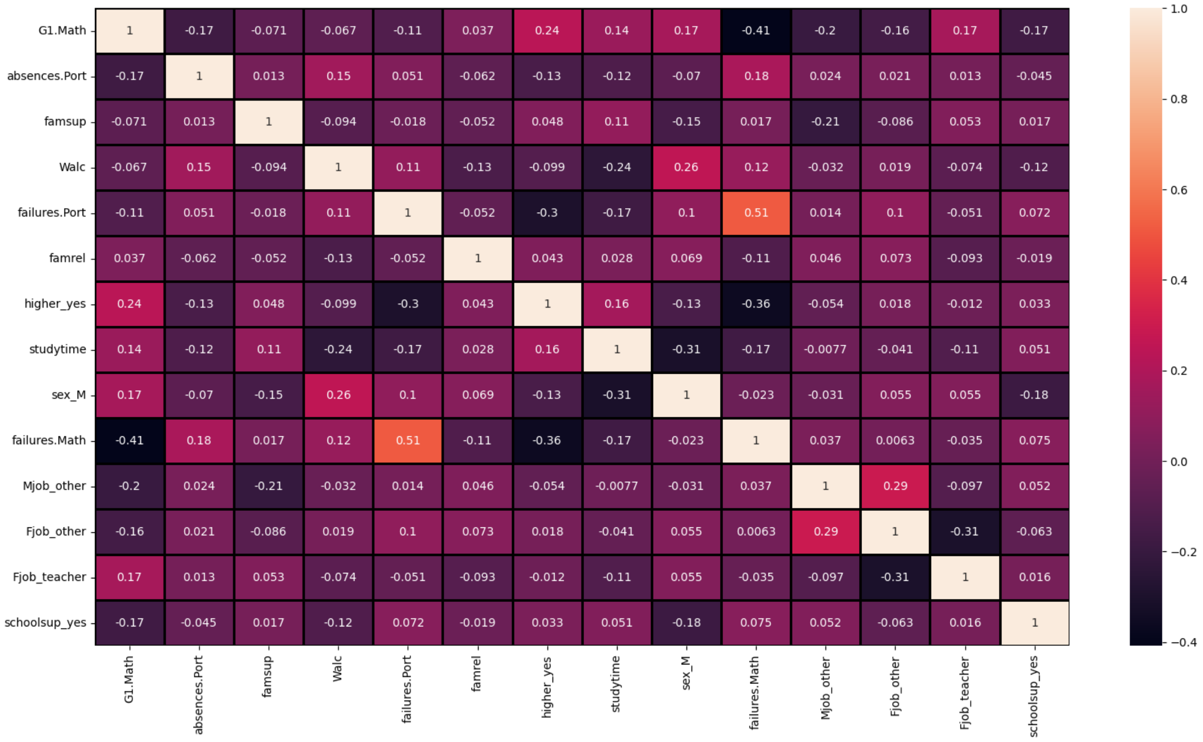
Dummy variables

If we have categorical variables, we need to change our categorical variables into dummy variables (using get_dummies or OneHotEncoder).
for finding the corrolation the independent variable should be integer so this step we can check the variables are integer,

4.2 Correlation between different features:

Correlation is the way of understanding the strength of the relationship between 2 variables or features in a dataset. Correlation coefficients determine this strength by indicating a value between [-1,1] where -1 indicates a very strong negative relationship, 0 indicates no relationship and 1 indicates strong positive relationship. Pearson correlation is one of the most widely used correlation methods and it indicates the linear relationship between 2 variables.

The heatmap of correlation between all variables of the dataset is given bellow:



	G1.Math	absences.Port	famsup	Walc	failures.Port	famrel	higher_yes	studytime	sex_M	failures.Math	Mjob_other	Fjob_other	Fjob_teacher	schoolsup_yes
G1.Math	1	-0.17	-0.071	-0.067	-0.11	0.037	0.24	0.14	0.17	-0.41	-0.2	-0.16	0.17	-0.17
absences.Port	-0.17	1	0.013	0.15	0.051	-0.062	-0.13	-0.12	-0.07	0.18	0.024	0.021	0.013	-0.045
famsup	-0.071	0.013	1	-0.094	-0.018	-0.052	0.048	0.11	-0.15	0.017	-0.21	-0.086	0.053	0.017
Walc	-0.067	0.15	-0.094	1	0.11	-0.13	-0.099	-0.24	0.26	0.12	-0.032	0.019	-0.074	-0.12
failures.Port	-0.11	0.051	-0.018	0.11	1	-0.052	-0.3	-0.17	0.1	0.51	0.014	0.1	-0.051	0.072
famrel	0.037	-0.062	-0.052	-0.13	-0.052	1	0.043	0.028	0.069	-0.11	0.046	0.073	-0.093	-0.019
higher_yes	0.24	-0.13	0.048	-0.099	-0.3	0.043	1	0.16	-0.13	-0.36	-0.054	0.018	-0.012	0.033
studytime	0.14	-0.12	0.11	-0.24	-0.17	0.028	0.16	1	-0.31	-0.17	-0.0077	-0.041	-0.11	0.051
sex_M	0.17	-0.07	-0.15	0.26	0.1	0.069	-0.13	-0.31	1	-0.023	-0.031	0.055	0.055	-0.18
failures.Math	-0.41	0.18	0.017	0.12	0.51	-0.11	-0.36	-0.17	-0.023	1	0.037	0.0063	-0.035	0.075
Mjob_other	-0.2	0.024	-0.21	-0.032	0.014	0.046	-0.054	-0.0077	-0.031	0.037	1	0.29	-0.097	0.052
Fjob_other	-0.16	0.021	-0.086	0.019	0.1	0.073	0.018	-0.041	0.055	0.0063	0.29	1	-0.31	-0.063
Fjob_teacher	0.17	0.013	0.053	-0.074	-0.051	-0.093	-0.012	-0.11	0.055	-0.035	-0.097	-0.31	1	0.016
schoolsup_yes	-0.17	-0.045	0.017	-0.12	0.072	-0.019	0.033	0.051	-0.18	0.075	0.052	-0.063	0.016	1

After checking the correlation, I selected failures. Math, higher yes, failures. Math, for finding the model because this variable has correlation with G1. math more than another columns.

G1.Math	1.000000
higher_yes	0.238948
sex_M	0.172840
Fjob_teacher	0.171032
studytime	0.135777
famrel	0.037323
Walc	-0.066664
famsup	-0.070687
failures.Port	-0.109396
Fjob_other	-0.168868
schoolsup_yes	-0.168868
absences.Port	-0.174703
failures.Math	-0.408430
Name: G1.Math, dtype: float64	

Key Findings

The given values represent the correlation coefficients between the 'G1.Math' variable and other variables in the dataset. Here's an explanation:

G1.Math: This is the target variable.

Positive Correlations:

higher_yes: There is a positive correlation of approximately 0.24 between 'higher_yes' (aspiration for higher education) and 'G1.Math'. This suggests that students who aspire to pursue higher education tend to have higher grades in the first period.

sex_M: There is a positive correlation of approximately 0.17 between 'sex_M' (male gender) and 'G1.Math'. This implies that male students may have slightly higher grades in the first period.

Fjob_teacher: There is a positive correlation of approximately 0.17 between 'Fjob_teacher' (father's job as a teacher) and 'G1.Math'. This indicates that students whose fathers work as teachers may have slightly higher grades.

Negative Correlations:

studytime: There is a positive correlation of approximately 0.14 between 'studytime' and 'G1.Math'. This implies that students who spend more time studying may have slightly higher grades.

famrel: There is a positive correlation of approximately 0.04 between 'famrel' (quality of family relationships) and 'G1.Math'. This suggests that students with better family relationships may have slightly higher grades.

Negative Correlations:

Walc: There is a negative correlation of approximately -0.07 between 'Walc' (weekend alcohol consumption) and 'G1.Math'. This indicates a slight tendency that higher weekend alcohol consumption may be associated with slightly lower grades.

famsup: There is a negative correlation of approximately -0.07 between 'famsup' (family educational support) and 'G1.Math'. This suggests that students who receive more family educational support may have slightly lower grades.

failures.Port: There is a negative correlation of approximately -0.11 between 'failures.Port' (number of past class failures in the Portuguese subject) and 'G1.Math'. This indicates that students with fewer past class failures in the Portuguese subject tend to have higher grades.

Fjob_other: There is a negative correlation of approximately -0.17 between 'Fjob_other' (father's job other than teacher, health, or services) and 'G1.Math'. This implies that students whose fathers have jobs other than teacher, health, or services may have slightly lower grades.

schoolsup_yes: There is a negative correlation of approximately -0.17 between 'schoolsup_yes' (extra educational support at school) and 'G1.Math'. This suggests that students receiving extra educational support at school may have slightly lower grades.

absences.Port: There is a negative correlation of approximately -0.17 between 'absences.Port' (number of school absences in the Portuguese subject) and 'G1.Math'. This implies that students with fewer school absences may have higher grades.

failures.Math: There is a negative correlation of approximately -0.41 between 'failures.Math' (number of past class failures in the Mathematics subject) and 'G1.Math'. This indicates a strong negative correlation, suggesting that students with fewer past class failures in the Mathematics subject tend to have higher grades

5 Data Analysis (Visualization and checking the distribution of each variable).

In this part of report, there are some visualizations to understand the distribution of the variables and to check if they have outliers or not.

5.1 Data Modeling

A summary of each algorithm is described below.

Logistic Regression is a classification method used when the Response column is categorical with only two possible values. The probability of the possible outcomes is modeled with a logistic transformation as a weighted sum of the Predictor columns. The weights or regression coefficients are selected to maximize the likelihood of the observed data.

Linear Discriminant Analysis or Normal Discriminant Analysis or Discriminant Function Analysis is a dimensionality reduction technique that is commonly used for supervised classification problems. It is used for modelling differences in groups i.e. separating two or more classes. It is used to project the features in higher dimension space into a lower dimension space. Linear discriminant analysis is popular when we have more than two response classes, because it also provides low-dimensional views of the data

K-Nearest Neighbors algorithm, also known as KNN or k-NN, is a non-parametric algorithm (which means it does not make any assumption on underlying data), supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point. a class label is assigned based on a majority vote.

Decision Tree is a type of supervised machine learning used to categorize or make predictions based on how a previous set of questions were answered. The model is a form of supervised learning, meaning that the model is trained and tested on a set of data that contains the desired categorization. The tree can be explained by two entities, namely decision nodes and leaves.

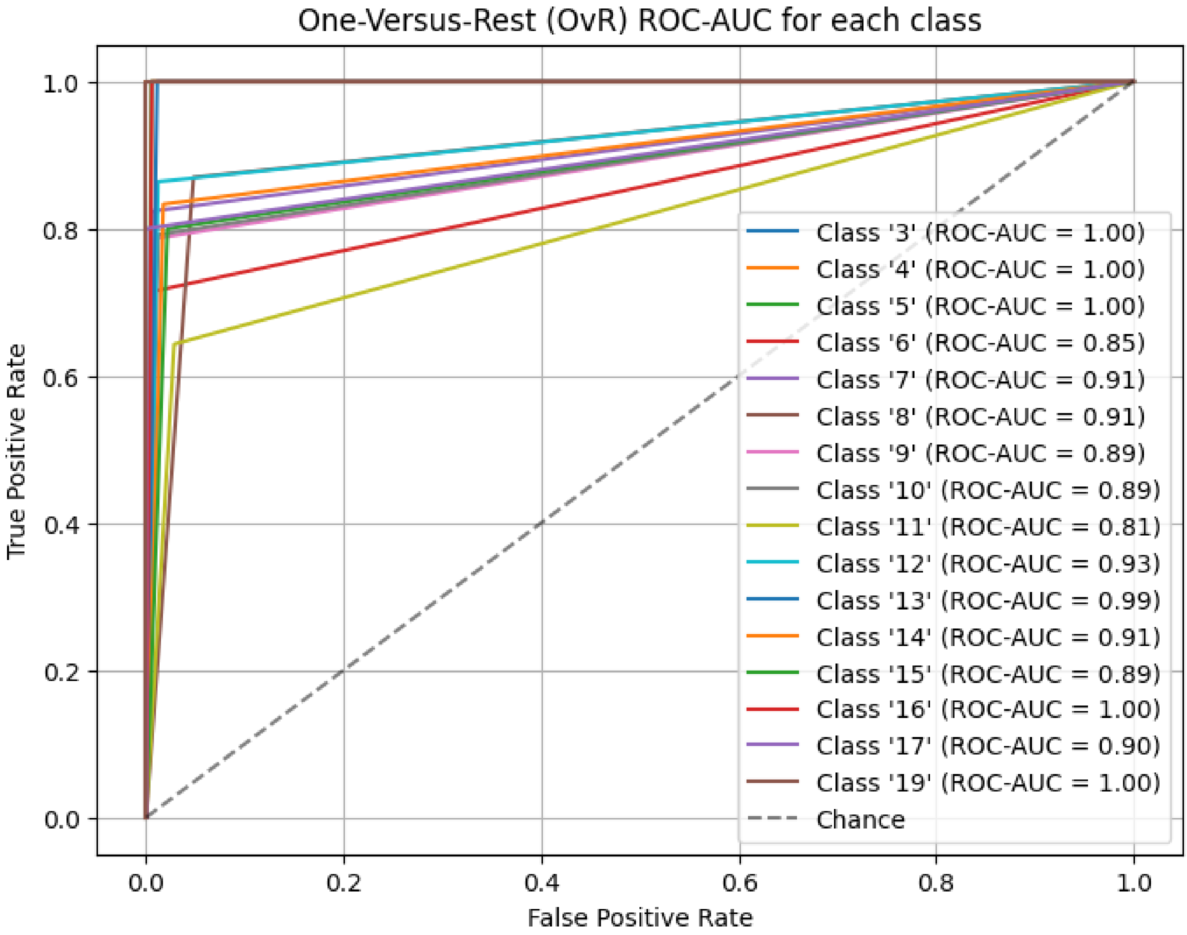
Random Forest is a collection (a.k.a. ensemble) of many decision trees. A decision tree is a flow chart which separates data based on some condition. If a condition is true, you move on a path otherwise, you move on to another path.

5.2 Model Selection Task 1

In order to select my classifier, I performed a 10-fold cross validation algorithm on the above-mentioned classification models and calculated the accuracy (average of all 10 folds) of each model. The result of the cross validation is as follows

Algorithm	Model Accuracy (Average of 10 folds)	Standard Deviation (Of 10 folds)
Linear Discriminant Analysis	0.215193	0.046488
K-Nearest Neighbors	0.477509	0.049930
Decision Tree	0.836456	0.039137
Random Forest	0.844316	0.042251

as you can see the RFM has the best accuracy and after that DTC can be selected for this model between the other classifier models.



Receiver operating characteristic (ROC) curve plots true positive rate (sensitivity) vs. false positive rate. The Area Under the ROC curve (AUC) is a measure of how well a parameter can distinguish between two diagnostic groups. Each point on the ROC curve represents a sensitivity/ (1 - specificity) pair corresponding to a particular decision threshold and is used as a performance metric in classification algorithms.
without removing the schoolsup_yes we have 89% accuracy and the roc-auc for each classes between 0.89 to 1

Algorithm	**Model Accuracy (Average of 10 folds)**	**Standard Deviation (Of 10 folds)**
Linear Discriminant Analysis	0.175474	0.033965
K-Nearest Neighbors	0.522368	0.058981
Decision Tree	0.885263	0.041603
Random Forest	0.891807	0.040430

To check this more precisely, I removed schoolsup_yes from my feature variables and performed a Random Forest Classifier model. I split my data set into train (80% of the observation) and test (20% of the observation), fitted the model on train data and performed a prediction on my test data. The classification report is as follows;

model's average cross-validation score is approximately 0.88, and this is the average performance across the 10 folds. It provides an estimate of how well model is expected to generalize to new, unseen data.

5.3 Model Selection- task 2

Task 2: Bin the target variable G1. Math is divided into 3 categories in such a way that the resulting bins contain roughly an equal number of cases. Use this newly created categorical variable as response for a classification model. Again, do not use any other grade feature and build a model that contains the variables absences. Port, famsup, Walc, failures. Port, study time, famrel but not the variables Fedu, Medu, age.

After categorizing we can check the data with the classification, after training the data an prediction we can check the accuracy and select the best algorithm with the high accuracy.

Algorithm	Model Accuracy (Average of 10 folds)	Standard Deviation (Of 10 folds)
Linear Discriminant Analysis	0.563316	0.056017
K-Nearest Neighbors	0.651614	0.060970
Decision Tree	0.901018	0.026673
Random Forest	0.908930	0.020987

Regarding the accuracy score, I chose Decision Tree and Random Forest as my classifier as its accuracy .

I check the matrix :

```
[[51  4  2]
 [ 6 41  5]
 [ 1  0 80]]
              precision    recall  f1-score   support

         0.0       0.88      0.89      0.89        57
         1.0       0.91      0.79      0.85        52
         2.0       0.92      0.99      0.95        81

    accuracy                           0.91       190
   macro avg       0.90      0.89      0.89       190
weighted avg       0.91      0.91      0.90       190
```

```
0.9052631578947369
```

In this matrix:

True Positives (TP):

TP for Class 0 (first row, first column): 52 instances were correctly predicted as Class 0.

TP for Class 1 (second row, second column): 41 instances were correctly predicted as Class 1.

TP for Class 2 (third row, third column): 78 instances were correctly predicted as Class 2.

False Positives (FP):

FP for Class 0 (first row, second and third columns): 4 instances of Class 0 were incorrectly predicted as Class 1, and 1 instance was incorrectly predicted as Class 2.

FP for Class 1 (second row, first and third columns): 5 instances of Class 1 were incorrectly predicted as Class 0, and 6 instances were incorrectly predicted as Class 2.

FP for Class 2 (third row, first and second columns): 2 instances of Class 2 were incorrectly predicted as Class 0, and 1 instance was incorrectly predicted as Class 1.

True Negatives (TN):

The remaining entries outside the diagonal are not explicitly mentioned in a confusion matrix, but they represent instances that were correctly predicted as classes other than the one specified by the row.

False Negatives (FN):

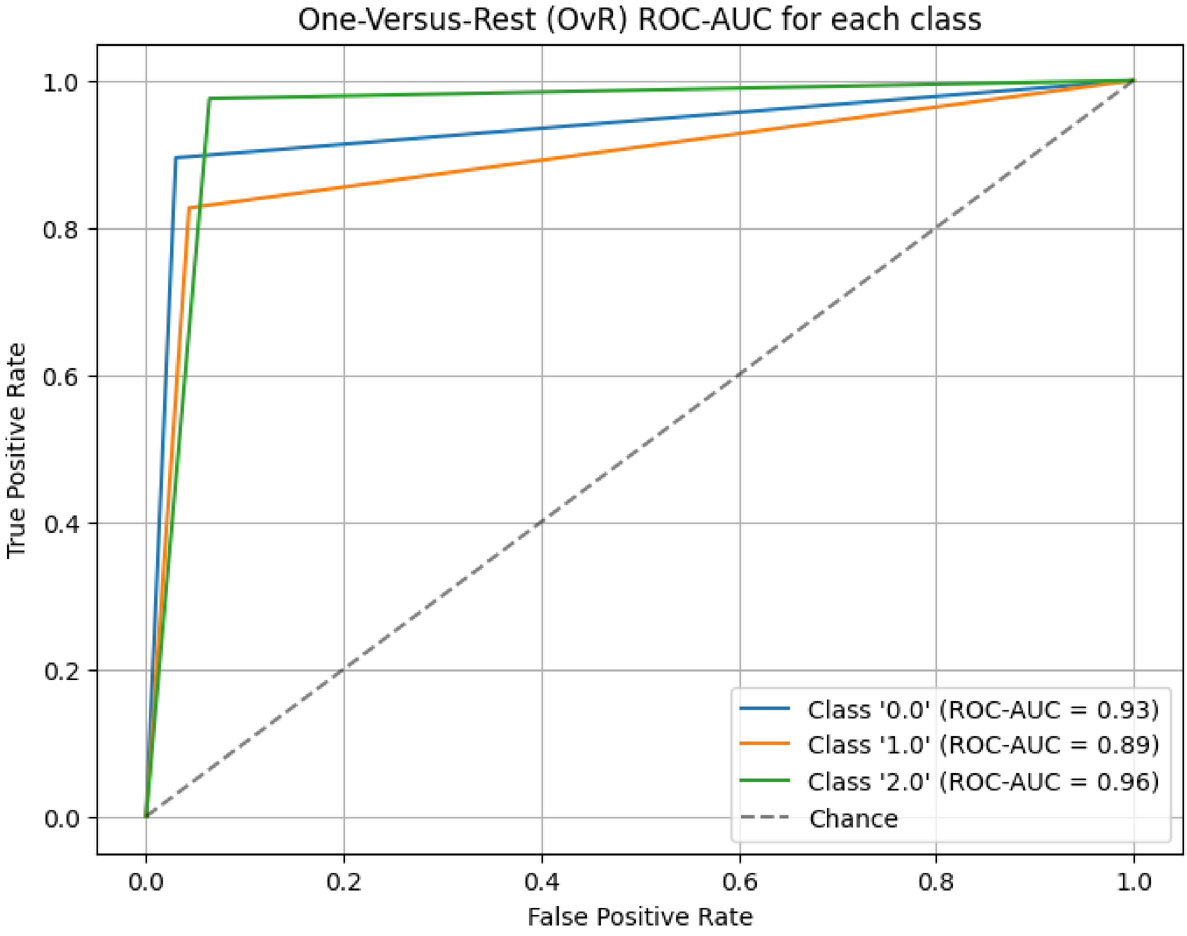
FN for Class 0 (first row, remaining columns): 1 instance of Class 0 was incorrectly predicted as either Class 1 or Class 2.

FN for Class 1 (second row, remaining columns): 6 instances of Class 1 were incorrectly predicted as either Class 0 or Class 2.

FN for Class 2 (third row, remaining columns): 6 instances of Class 2 were incorrectly predicted as either Class 0 or Class 1.

These values in the confusion matrix help evaluate the performance of a classification model by showing how many instances were correctly or incorrectly predicted for each class.

I check the models again by k-fold to find better differential result: the accuracy of RFM: 90% and in this category we have better accuracy when we have about 18 category.



the ROC-AUC: between 89 to 96 for three classes.

and the cross validation is :

```
Cross Validation Scores:  [0.89473684 0.89473684 0.94736842 0.88157895
0.85526316 0.90789474
 0.92105263 0.85526316 0.93333333 0.89333333]
Average CV Score:  0.8984561403508771
```

Number of CV Scores used in Average: 10

the average accuracy across the 10 folds is approximately 0.8932, or 89.32%.
the average cross-validation score provides an estimate of the model's performance that is less sensitive to the choice of a particular training/test split, giving you a more robust evaluation metric.
at the end If accuracy increases while the area under the ROC curve (AUC) decreases, it means that there is an improvement in correctly identifying positive samples (True Positives) and a reduction in misclassifying samples from other classes (False Positives). With increased accuracy, the model demonstrates better capability in correctly identifying positive class instances.
However, if AUC decreases, it indicates a change in the balance between the True Positive rate and False Positive rate. The model may perform better in detecting positive samples, but it might struggle in reducing False Negatives, which could be crucial in specific applications, depending on the nature of the problem.
In general, an increase in AUC signifies an improvement in the model's ability to differentiate between classes, while an increase in accuracy indicates an overall enhancement in the model's ability to correctly identify instances.

6 Results and Conclusions

in this data set we have a little correlation between features but by cleaning and preprocessing we tr the best feature in this data set after with visualization founding the outliers for dropping the data set , for using the model we have a integer and the data in G1.math is Discontinuous so we use the classification but the category of the data is range of 1 to zero so by correlation low between the target and the other feature is it hart to access the high accuracy and we have the 86 % and for better accuracy we drop some columns and check it again , in task 2 we have the 3 category of the grade the grade in each category have little variance and big variance between the other categories, so it help us to have a good accuracy and good roc-auc .

at the end I think is good idea categorizing the data for better accuracy and found the best feature for best prediction.

7 References

1- https://towardsdatascience.com/beginner-guide-to-build-compare-and-evaluate-machine-learning-models-in-under-10-minutes-19a6781830de

2- https://github.com/Mesgarin/DAPJFinal/tree/main