**Final Project Report**

Fall 2023

**Dataset:**

|  |  |  |
| --- | --- | --- |
| Author: Nastaran Mesgari,  Sanaz Achik,  Fatemeh Rajaeian  Date: 2023/12/22 |  |  |

**Executive Summary**

*This report is provided for the data analytics project with the goal of performing a reasonable analysis of the given data. The general task of 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 Random Forest and for task 2 after categorizing the grade of math we use Random Forest, but the accuracy is better than before.*

*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 virtualize 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 defined 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 " among 4 algorithms by accuracy in task 1: 84%,for improving the algorit I check eagin correlation and remove 'schoolsup\_yes' and check again the accuracy it better than befor and it equal to 89%  ROC-AUC  between 0.81 t0 1 for 16 classes , and task 2 after categorizing into 3 for target variable G1. math accuracy improved I check automaticly category and manualy , for better resualt in accuracy and Roc-Auc I prefer manualy cattegorizing and it set it in code by ownself, and it is 94%. the ROC-AUC was 1 for class 0.0, 0.94 for class 1.0, and 0.94 for class 2.0 in AUC signifies an improvement in the model's ability to differentiate between classes but for improving the model.*

*.*

**Table of Content**

Contents

[1 Introduction 4](#_Toc152975563)

[2 Data Set 4](#_Toc152975564)

[2.1 student achievement Data set 4](#_Toc152975565)

[3 Data Pre-Processing 6](#_Toc152975566)

[3.1 Data Cleaning 6](#_Toc152975567)

[3.1.1 Dropping unnecessary columns and rows 6](#_Toc152975568)

[3.1.2 Checking for missing values: 6](#_Toc152975569)

[3.1.3 Checking for garbage values 6](#_Toc152975570)

[3.1.4 Checking the distribution of each variable 6](#_Toc152975571)

[3.2 Data Transformation 7](#_Toc152975572)

[3.2.1 Transforming the categorical variables 7](#_Toc152975573)

[3.2.2 Normalization, standardization, scaling 7](#_Toc152975574)

[4 Data Exploration, 7](#_Toc152975575)

[4.1 Finding outliers and dummy variable 7](#_Toc152975576)

[4.2 Correlation between different features: 10](#_Toc152975577)

[5 Data Analysis (Visualization and checking the distribution of each variable). 13](#_Toc152975578)

[5.1 Data Modeling 13](#_Toc152975579)

[5.2 Model Selection Task 1 13](#_Toc152975580)

[5.3 Model Selection- task 2 15](#_Toc152975581)

[6 Results and Conclusions 19](#_Toc152975582)

[7 References 19](#_Toc152975583)

# 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.

.

. Research data

The aim of this project is to diagnose heart attack from patient test data. There are many symptoms of heart disease, finding patterns of heart disease data helps in diagnosing the future causes of this disease. This data set obtained from the Kaggle site contains 1024 samples that represent the number of lines and 13 columns that represent attributes or features and one column represents the class label. This data set dates back to 1988 and consists of four databases: Cleveland, Hungary, Switzerland, and Long Beach, which includes 14 target field characteristics indicating the presence of heart disease in the patient. which are divided into two groups having a heart attack with the number 1 and 0 for people who have not had a heart attack.

# Data Set

## 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.

age, gender, type of chest pain (cp), resting blood pressure (trestbps), cholesterol (chol), fasting blood sugar (fbs), resting electrocardiography results (restecg), maximum Acquired heart rate (thalach), exercise-induced angina (exang), exercise-induced ST depression compared to rest (old peak), slope of heart rate increase during exercise (slope), number of damaged vessels (ca), impairment Blood thalassemia (thal).

|  |  |  |  |
| --- | --- | --- | --- |
| **Column** | **Dtype** | **Description** | **Values** |
| Unnamed | float64 | Index of the dataset |  |
| trestbps | Int64 | student’s school | “GP” - Gabriel Pereira or “MS” - Mousinho da Silveira |

# Data Pre-Processing

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

## 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:

## 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.

## Checking for missing values:

In most of the 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.

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 .

So, before going ahead with the analysis, it is a good idea to check whether the dataset has any missing values.As you can see in the output of the block below, there is no missing value in this dataset

## 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.

### 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.

.

## Data Transformation

### 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'

### 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.

# Data Exploration,

We put our visualization here!

## 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:

A comparison of a graph

Description automatically generated

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

A graph with blue rectangular bars

Description automatically generated with medium confidence

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

A comparison of a graph

Description automatically generated

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.

A graph of a bar and a bar

Description automatically generated with medium confidence

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

failures. Port right skewed. There are 82 observations which are their failures.Port is more than 812.

A comparison of a graph

Description automatically generated

studytime

2 444

1 230

3 160

4 60

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

Walc is right skewed. But all data is in quartiles.

A comparison of a graph

Description automatically generated with medium confidence

famrel is left skewed. There are 82 observations which are their famrel There are 833 observations greater than 2.5 and less than 6.5.

.

A graph and diagram of a graph

Description automatically generated with medium confidence

G1.Math is not skewed. and we do not have any outliers. and the frequency in 10 is more than the other numbers.

**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,

## 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:

after checking the correlation I selected, 'G1.Math','absences.Port', 'famsup', 'Walc','failures.Port','famrel', 'higher\_yes','studytime','sex\_M','failures.Math','Mjob\_other', 'Fjob\_other', 'Fjob\_teacher','schoolsup\_yes'

for finding the model becouase this variable has corrolation with G1.math more than another columns.

A colorful squares with different shades of purple and white

Description automatically generated with medium confidence

there are a good relation between failler.port and failler.math

|  |  |  |
| --- | --- | --- |
| 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

# 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.

## Data Modeling

A summary of each algorithm is described below.

**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.

## 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.

A graph with different colored lines

Description automatically generated

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 87% 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 | 0.871807 | 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 0.87, and this is the average performance across the 10 folds. It provides an estimate of how well the model is expected to generalize to new, unseen data.

## 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.

I check the correlation again:

A screenshot of a computer screen

Description automatically generated

After categorizing we can check the data with the classification, after training the data and 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 | 0.671614 | 0.073235 |
| K-Nearest Neighbors | 0.753263 | 0.068585 |
| Decision Tree | 0.934070 | 0.021956 |
| Random Forest | 0.935368 | 0.027229 |

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

I check the matrix :

[[ 3 0 0]

[ 0 121 7]

[ 0 4 55]]

precision recall f1-score support

0 1.00 1.00 1.00 3

1 0.97 0.95 0.96 128

2 0.89 0.93 0.91 59

accuracy 0.94 190

macro avg 0.95 0.96 0.96 190

weighted avg 0.94 0.94 0.94 190

0.9421052631578948

In this matrix:

True Positives (TP):

TP for Class 0 (first row, first column): 3 instances were correctly predicted as Class 0.

TP for Class 1 (second row, second column): 121 instances were correctly predicted as Class 1.

TP for Class 2 (third row, third column): 80 instances were correctly predicted as Class 2.

False Positives (FP):

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

FP for Class 1 (second row, first and third columns): 0 instances of Class 1 were incorrectly predicted as Class 0, and 7 instances were incorrectly predicted as Class 2.

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

A graph of a number of classes

Description automatically generated

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

and the cross validation is :

Cross Validation Scores: [0.94736842 0.94736842 0.94736842 0.94736842 0.90789474 0.94736842

0.93421053 0.85526316 0.97333333 0.93333333]

Average CV Score: 0.9340877192982455

Number of CV Scores used in Average: 10

 the average accuracy across the 10 folds is approximately 0.9340, or 93.40%.

 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.

# Results and Conclusions

in this data set we have a little correlation between features but by cleaning and preprocessing we try 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 hard to access the high accuracy and we have the 84 % and for better accuracy we drop some columns and check it again , in task 2 we have the 3 category of the grade. categorizing the data for better accuracy 93 and found the best feature for best prediction.

# References

1. https://towardsdatascience.com/

beginner-guide-to-build-compare-and-evaluate-machine-learning-models-in-under-10-minutes-19a6781830de

1. https://github.com/Mesgarin/DAPJFinal/tree/main