**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 grad 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 dataframe “G1Math\_df2” with absences.Port, famsup, Walc, failures.Port, studytime, famrel,higher\_yes,failure.math and G1.Math’*

*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”* RandomForestRegressor *“among 5 algorithms by performing r2 and mse. I trained my model with the training data and made a prediction on the test data using and evaluated the model.*

*After categorizing the model I can use the classification model: RandomForestClassifier with accuracy 82% and it shown that this type of data by using categorizing we can better accuracy .*

**Table of Content**

Contents

[1 Introduction 4](#_Toc152803736)

[2 Data Set 4](#_Toc152803737)

[2.1 student achievement Data set 4](#_Toc152803738)

[3 Data Pre-Processing 6](#_Toc152803739)

[3.1 Data Cleaning 6](#_Toc152803740)

[3.1.1 Dropping unnecessary columns 6](#_Toc152803741)

[3.1.2 Checking for missing values: 6](#_Toc152803742)

[3.1.3 Checking for garbage values 6](#_Toc152803743)

[3.1.4 Checking the distribution of each variable 6](#_Toc152803744)

[3.2 Data Transformation 7](#_Toc152803745)

[3.2.1 Transforming the categorical variables 7](#_Toc152803746)

[3.2.2 Normalization, standardization, scaling 7](#_Toc152803747)

[4 Data Exploration 7](#_Toc152803748)

[4.1 Finding outliers 7](#_Toc152803749)

[4.2 Correlation between different features: 7](#_Toc152803750)

[5 Visualization and checking the distribution of each variable. 9](#_Toc152803751)

[5.1.1 Continues Variables 9](#_Toc152803752)

[5.1.2 Discrete / Categorical Variables 9](#_Toc152803753)

[6 Results and Conclusions 10](#_Toc152803754)

[7 References 10](#_Toc152803755)

# 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. We have discrete variables that are similar to continuous variables due to the small distance between them, and at the end, according to these data, we checked whether by classifying the information, our prediction of students' grades would improve or not, and we check the accuracy.

.

# 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.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **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) | | | |  |

# Data Pre-Processing

## Data Cleaning

in this step we import the main libraries that we need, 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.

In this phase, we can also check the numeric variables with zero variance (threshold = 0), they do not have any contribution on the model. so, all columns that are selected can be used.

### 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 all data check and there didn’t find any missing value.

### 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).

Data has been checked and there aren’t any null or minus values. and zero has been a meaning in this feature.

### 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 :0 (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.

If we have time series in our data, it is crucial to convert date and time into date time objects.

Machine learning models are quite sensitive to the scale of data. They give more importance to the larger values while learning the properties of data. Hence, it becomes crucial for us to remove this bias by bringing down all the data points to the same scale.

# Data Exploration, dummy variable

We put our visualization here!

## Finding outliers

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 virtualized 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 miles.

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 aren’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

A comparison of a graph

Description automatically generated

studytime

2 444

1 230

3 160

4 60

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

A comparison of a graph

Description automatically generated with medium confidence

there are 833 observations greater than 2.5 and less than 6.5

## 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:

A screenshot of a computer screen

Description automatically generated

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 | |
| studytime | 0.135777 | |
| famrel | 0.037323 | |
| Walc | -0.066664 | |
| famsup | -0.070687 | |
| failures.Port | | -0.109396 |
| absences.Port | | -0.174703 |
| failures.Math | | -0.408430 |
| Name: G1.Math, dtype: float64 | | |

\*\*Key Findings\*\*

Positive Correlation:

studytime: There is a positive correlation of approximately 0.14 between 'studytime' and 'G1.Math'. This suggests that students who spend more time studying tend to have higher grades in the first period.

Walc: There is a positive but weak correlation of approximately 0.06 between 'Walc' (weekend alcohol consumption) and 'G1.Math'. This indicates a slight tendency that higher weekend alcohol consumption may be associated with slightly higher grades.

famrel: There is a positive correlation of approximately 0.03 between 'famrel' (quality of family relationships) and 'G1.Math'. This implies that students with better family relationships may have slightly higher grades.

Negative Correlation:

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 in the first period.

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

absences.Port: There is a negative correlation of approximately -0.16 between 'absences.Port' (number of school absences the Portuguese subject) and 'G1.Math'. This implies that students with fewer school absences may have higher grades.

As you can see the G1.math distributed 3 to 19 and the frequency of 10 is the more than another grades. and when we want to show the study time with Gi.Math with 19 has 3 studytimes but the time study 2 G1.math more distributed.

# 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

Linear Regression:

Description: Linear Regression is a simple and widely used regression algorithm. It models the relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data.

Lasso Regression (L1 Regularization):

Description: Lasso Regression is a linear regression technique with L1 regularization. It adds a penalty term proportional to the absolute values of the coefficients, encouraging sparsity and feature selection.

Ridge Regression (L2 Regularization):

Description: Ridge Regression is another form of linear regression with L2 regularization. It adds a penalty term proportional to the square of the coefficients, preventing overfitting by shrinking the coefficients.

Support Vector Regression (SVR):

Description: Support Vector Regression is a regression algorithm that uses support vector machines to find a hyperplane that best fits the data while minimizing deviations within a specified margin (epsilon).

Random Forest Regressor:

Description: Random Forest is an ensemble method that builds a collection of decision trees and merges their predictions. It provides better accuracy and reduces overfitting compared to a single decision tree.

Neural Network (MLP - Multi-Layer Perceptron):

Description: A Neural Network, specifically MLP, is a type of artificial neural network with multiple layers (input layer, hidden layers, and output layer). It can learn complex patterns and relationships in data.

## Model Selection

for using the models, we should have data for testing and training so we create it after check the models and find the MSE and r-squared and check the predictions,

the columns in this step are.

'G1.Math', 'absences.Port', 'famsup', 'Walc', 'failures.Port', 'famrel',

'higher\_yes', 'failures.Math', 'studytime'

|  |  |  |  |
| --- | --- | --- | --- |
| Algorithm | R squared (%) | MSE | Prediction new value |
| *Linear Regression* | 23 | 8.61 | 5.05 |
| *Lasso* | 3 | 10.92 | 11.29 |
| *Ridge* | 23 | 8.61 | 5.10 |
| *SVR* | 18 | 9.14 | 9.21 |
| *Random Forest Regressor* | 66 | 3.77 | 6.21 |
| *MLP Regressor* | 22 | 8.7 | 3.01 |

**Linear Regression**

Mean Squared Error (MSE):

The Mean Squared Error is a measure of the average squared difference between the actual and predicted values. It provides a way to quantify how well the model is predicting the target variable. A lower MSE indicates better model performance, with zero representing a perfect fit (the predicted values match the actual values). In this case, an MSE of 8.66 means, on average, the squared difference between the actual and predicted ‘G1. Math’ values is 8.66.

R-squared (R2):

R-squared is a measure of how well the independent variables explain the variance in the dependent variable. It ranges from 0 to 1, where 1 indicates that the model perfectly predicts the dependent variable based on the independent variables. A higher R2 value suggests a better fit. In this case, an R2 of 0.23 means that approximately 23% of the variance in ‘G1. Math’ is explained by the independent variables in this model. The remaining 77% of the variance is not captured by the model. R2 should be interpreted in the context of specific application, and it’s essential to consider other factors like the nature of data and the complexity of model.

*so, we selected Random Forest Regressor because the MSE is good and the R-squared 66 is better than the other algorithms.*

## Model Selection- task 2

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.

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, studytime, 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** |
| Linear Discriminant Analysis | 0.51 |
| K-Nearest Neighbors | 0.65 |
| Decision Tree | 0.82 |
| Random Forest | 0.82 |

Regarding the accuracy score, I chose Decision Tree and Random Forest as my classifier as its accuracy score surpassed Logistic Regression Classifier and K-Nearest Neighbors, Linear Discriminant Analysis.

# Results and Conclusions

In this part of the report, I explain the steps I took for the algorithm:

**Step 1:** After deciding what variables to choose, I split my data into train (80% of observations) and test (20% of observations) dataset.

**Step 2:** I selected following hyper parameters for my model:

**Step 3:** After fitting the model on my tarin data set and check the accuracy for model classification and MSE and R-square for regression models, then select the best model.

at the end I found it when the data is very similar to each other, and the correlation is not very big between the dependence variable and independence variables we can use categorizing and use the classification model for better accuracy.

when I use the K flod for testing the accuracy of model I got this accuracy:

|  |  |  |
| --- | --- | --- |
| **Algorithm** | **Model Accuracy**  **(Average of 10 folds)** | **Standard Deviation**  **(Of 10 folds)** |
| Logistic Regression | 0.430070 | 0.056090 |
| Linear Discriminant Analysis | 0.171439 | 0.062116 |
| K-Nearest Neighbors | 0.544825 | 0.050729 |
| Decision Tree | 0.792789 | 0.043093 |
| Random Forest | 0.813912 | 0.048154 |

However, I again used 10-fold class validation on the train data set again and compared other classification algorithms.

The results show that the accuracy metric for all other algorithms (except Linear Discriminant Analysis Random forest, the accuracy metric is 81.

[[46 4 7]

[ 6 38 8]

[ 4 6 71]]

Accuracy: 0.8157894736842105

Classification Report:

precision recall f1-score support

0.0 0.82 0.81 0.81 57

1.0 0.79 0.73 0.76 52

2.0 0.83 0.88 0.85 81

accuracy 0.82 190

macro avg 0.81 0.80 0.81 190

weighted avg 0.82 0.82 0.81 190

In this matrix:

Rows represent the actual (true) classes.

Columns represent the predicted classes.

For each combination of actual and predicted classes, the corresponding entry in the matrix shows the count of instances. For example, the entry in the first row and first column (46) indicates that the model correctly predicted 46 instances of class 0.

# References

https://towardsdatascience.com/

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