Student Grade Prediction

Name: Venkata Raghavendra Karthik Kanna
UIN: [redacted]

Introduction

Overview

In Portugal, the secondary education consists of 3 years of schooling, preceding 9 years of basic education and followed by higher education. Most of the students join the public and free education system. There are several courses that share core subjects such as the Portuguese Language and Mathematics. Like several other countries a 20-point grading scale is used, where 0 is the lowest grade and 20 is the perfect score. During the school year, students are evaluated in three periods and the last evaluation (G3 of Table 1) corresponds to the final grade. This study considers data collected during the 2005- 2006 school year from two public schools, from the Alentejo region of Portugal. Although there has been a trend for an increase of Information Technology investment from the Government, the majority of the Portuguese public school information systems are very poor, relying mostly on paper sheets. Hence, the database was built from two sources: school reports, based on paper sheets and including few attributes (i.e. the three period grades and number of school absences); and questionnaires, used to complement the previous information.

Goal

The present work intends to approach student achievement in secondary education for a school in Portugal, Europe. The aim is to predict student achievement and if possible to identify the key variables that affect educational success/failure, using recent real-world data (e.g. student grades, demographic, social and school related features) collected from reports and questionnaires.

Proposed Architecture

From the two core classes (i.e. Mathematics and Portuguese), one will be modeled under the following methodologies:

i) Binary classification (pass - 1 / fail - 0)
ii) Regression, with a numeric output that ranges between zero (0%) and twenty (100%)

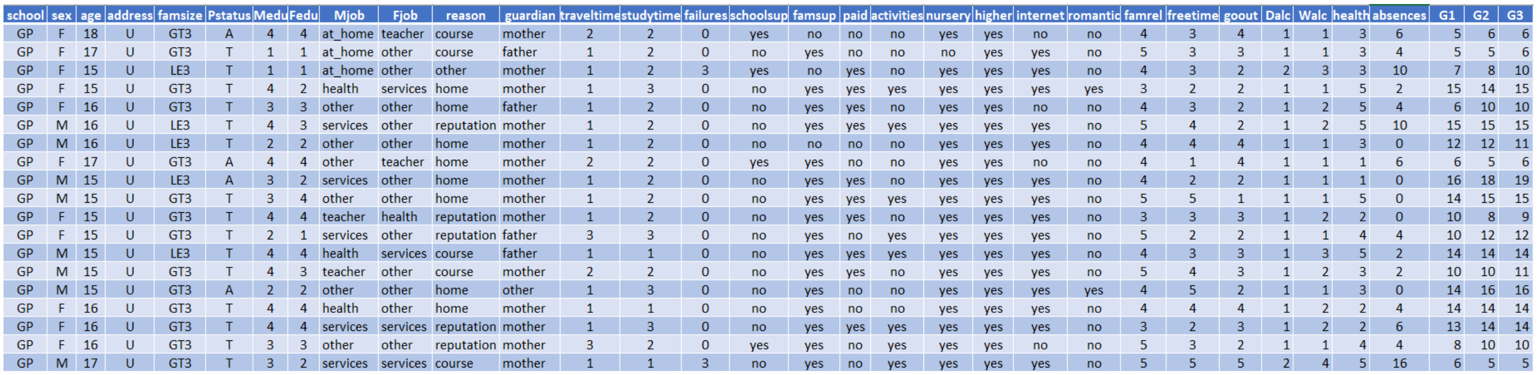
For each of these approaches, three input setups (e.g. with and without the school period grades) and five DM algorithms such as Decision Trees, Random Forest, Linear, Ridge and SVM will be tested. Moreover, an explanatory analysis will be performed over the best models, in order to identify the most relevant features.

Dataset

This dataset contains information about student achievement in secondary education of two Portuguese schools. The data attributes include student grades, demographic, social and school-related features collected through school reports and questionnaires.

Two datasets are extracted for evaluating the performance in two distinct subjects: Mathematics (mat) and Portuguese language (por), from which one is considered for prediction.

It is important to note that the target attribute G3 has a strong correlation with attributes G2 and G1. This occurs because G3 is the final year grade (issued at the 3rd period), while G1 and G2 correspond to the 1st and 2nd period grades. We hypothesize that it is more difficult to predict G3 without G2 and G1, but such prediction is much more useful.



school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	reason	guardian	traveltime	studytime	failures	schoolsup	famsup	paid	activities	nursery	higher	internet	romantic	famrel	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3
GP	F	18	U	GT3	A	4	4	at_home	teacher	course	mother	2	2	0	yes	no	no	no	yes	yes	no	no	4	3	4	1	1	3	6	5	6	6
GP	F	17	U	GT3	T	1	1	at_home	other	course	father	1	2	0	no	yes	no	no	no	yes	yes	no	5	3	3	1	1	3	4	5	5	6
GP	F	15	U	LE3	T	1	1	at_home	other	other	mother	1	2	3	yes	no	yes	no	yes	yes	yes	no	4	3	2	2	3	3	10	7	8	10
GP	F	15	U	GT3	T	4	2	health	services	home	mother	1	3	0	no	yes	yes	yes	yes	yes	yes	yes	3	2	2	1	1	5	2	15	14	15
GP	F	16	U	GT3	T	3	3	other	other	home	father	1	2	0	no	yes	yes	no	yes	yes	no	no	4	3	2	1	2	5	4	6	10	10
GP	M	16	U	LE3	T	4	3	services	other	reputation	mother	1	2	0	no	yes	yes	yes	yes	yes	yes	no	5	4	2	1	2	5	10	15	15	15
GP	M	16	U	LE3	T	2	2	other	other	home	mother	1	2	0	no	no	no	no	yes	yes	yes	no	4	4	4	1	1	3	0	12	12	11
GP	F	17	U	GT3	A	4	4	other	teacher	home	mother	2	2	0	yes	yes	no	no	yes	yes	no	no	4	1	4	1	1	1	6	6	5	6
GP	M	15	U	LE3	A	3	2	services	other	home	mother	1	2	0	no	yes	yes	no	yes	yes	yes	no	4	2	2	1	1	1	0	16	18	19
GP	M	15	U	GT3	T	3	4	other	other	home	mother	1	2	0	no	yes	yes	yes	yes	yes	yes	no	5	5	1	1	1	5	0	14	15	15
GP	F	15	U	GT3	T	4	4	teacher	health	reputation	mother	1	2	0	no	yes	yes	no	yes	yes	yes	no	3	3	3	1	2	2	0	10	8	9
GP	F	15	U	GT3	T	2	1	services	other	reputation	father	3	3	0	no	yes	no	yes	yes	yes	yes	no	5	2	2	1	1	4	4	10	12	12
GP	M	15	U	LE3	T	4	4	health	services	course	father	1	1	0	no	yes	yes	yes	yes	yes	yes	no	4	3	3	1	3	5	2	14	14	14
GP	M	15	U	GT3	T	4	3	teacher	other	course	mother	2	2	0	no	yes	yes	no	yes	yes	yes	no	5	4	3	1	2	3	2	10	10	11
GP	M	15	U	GT3	A	2	2	other	other	home	other	1	3	0	no	yes	no	no	yes	yes	yes	yes	4	5	2	1	1	3	0	14	16	16
GP	F	16	U	GT3	T	4	4	health	other	home	mother	1	1	0	no	yes	no	no	yes	yes	yes	no	4	4	4	1	2	2	4	14	14	14
GP	F	16	U	GT3	T	4	4	services	services	reputation	mother	1	3	0	no	yes	yes	yes	yes	yes	yes	no	3	2	3	1	2	2	6	13	14	14
GP	F	16	U	GT3	T	3	3	other	other	reputation	mother	3	2	0	yes	yes	no	yes	yes	yes	no	no	5	3	2	1	1	4	4	8	10	10
GP	M	17	U	GT3	T	3	2	services	services	course	mother	1	1	3	no	yes	no	yes	yes	yes	yes	no	5	5	5	2	4	5	16	6	5	5

The dataset in consideration includes records of 395 students from two public schools studying the subject Mathematics.

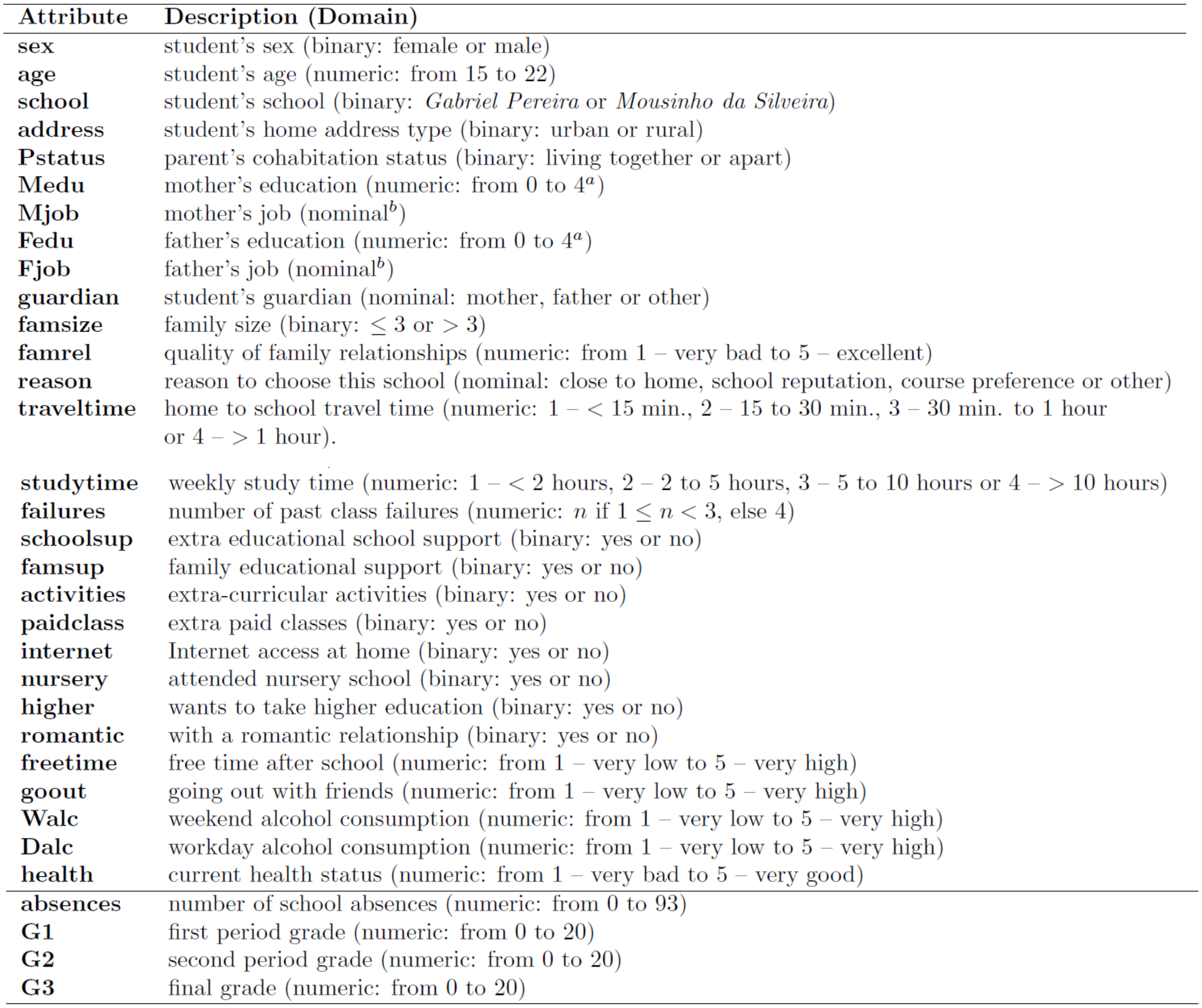
The data consists of 32 predictors as well as 1 response variable (G3).

The attributes are divided into:

- 16 Categorical Features
- 16 Continuous Features
- 1 Target Variable

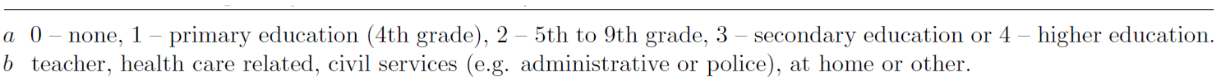
Attribute Information:

The description of the predictors are as follows:



Attribute	Description (Domain)
sex	student's sex (binary: female or male)
age	student's age (numeric: from 15 to 22)
school	student's school (binary: *Gabriel Pereira* or *Mousinho da Silveira*)
address	student's home address type (binary: urban or rural)
Pstatus	parent's cohabitation status (binary: living together or apart)
Medu	mother's education (numeric: from 0 to 4[a])
Mjob	mother's job (nominal[b])
Fedu	father's education (numeric: from 0 to 4[a])
Fjob	father's job (nominal[b])
guardian	student's guardian (nominal: mother, father or other)
famsize	family size (binary: ≤ 3 or > 3)
famrel	quality of family relationships (numeric: from 1 – very bad to 5 – excellent)
reason	reason to choose this school (nominal: close to home, school reputation, course preference or other)
traveltime	home to school travel time (numeric: 1 – $<$ 15 min., 2 – 15 to 30 min., 3 – 30 min. to 1 hour or 4 – $>$ 1 hour).
studytime	weekly study time (numeric: 1 – $<$ 2 hours, 2 – 2 to 5 hours, 3 – 5 to 10 hours or 4 – $>$ 10 hours)
failures	number of past class failures (numeric: n if $1 \leq n < 3$, else 4)
schoolsup	extra educational school support (binary: yes or no)
famsup	family educational support (binary: yes or no)
activities	extra-curricular activities (binary: yes or no)
paidclass	extra paid classes (binary: yes or no)
internet	Internet access at home (binary: yes or no)
nursery	attended nursery school (binary: yes or no)
higher	wants to take higher education (binary: yes or no)
romantic	with a romantic relationship (binary: yes or no)
freetime	free time after school (numeric: from 1 – very low to 5 – very high)
goout	going out with friends (numeric: from 1 – very low to 5 – very high)
Walc	weekend alcohol consumption (numeric: from 1 – very low to 5 – very high)
Dalc	workday alcohol consumption (numeric: from 1 – very low to 5 – very high)
health	current health status (numeric: from 1 – very bad to 5 – very good)
absences	number of school absences (numeric: from 0 to 93)
G1	first period grade (numeric: from 0 to 20)
G2	second period grade (numeric: from 0 to 20)
G3	final grade (numeric: from 0 to 20)

Key Index:

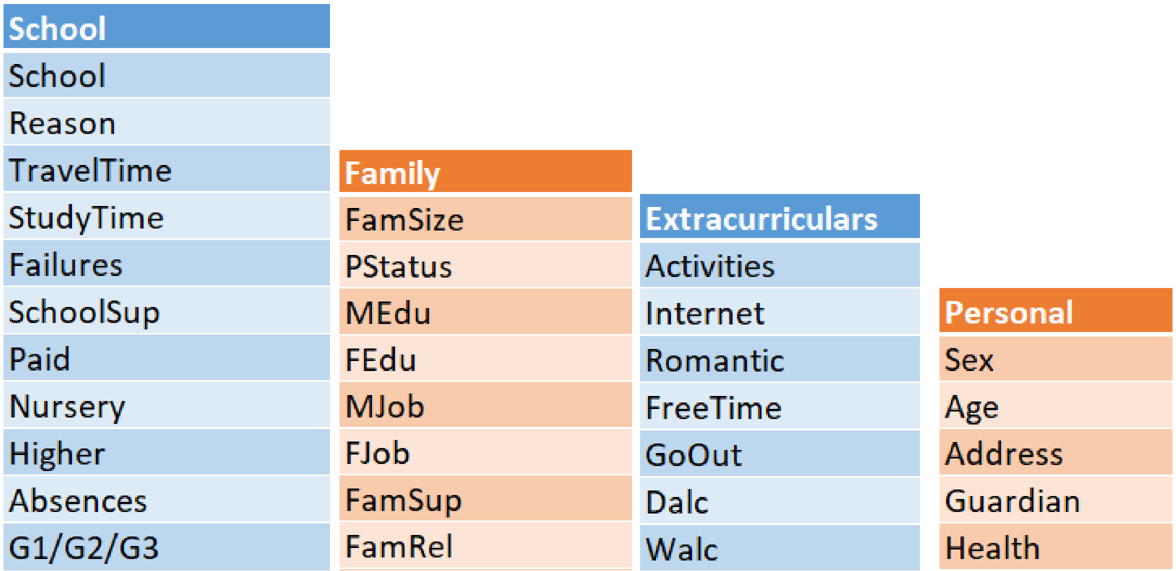


[a] 0 – none, 1 – primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education.

[b] teacher, health care related, civil services (e.g. administrative or police), at home or other.

Attribute Categorizing

To help us better understand the nature of the attributes that may be significant in predicting a student's grade, we categorized the feature variables into four different categories which are as follows:



School
School
Reason
TravelTime
StudyTime
Failures
SchoolSup
Paid
Nursery
Higher
Absences
G1/G2/G3

Family
FamSize
PStatus
MEdu
FEdu
MJob
FJob
FamSup
FamRel

Extracurriculars
Activities
Internet
Romantic
FreeTime
GoOut
Dalc
Walc

Personal
Sex
Age
Address
Guardian
Health

School:
These attributes involve the information related to a student's school life, including the reason why they chose a particular school, how much time they spend studying, their previous grades as well as the number of absences from school.

Family:
These features consist of each students' family information like their parents' education, parents' jobs, how much they support their children and how their relationship is with their children etc.

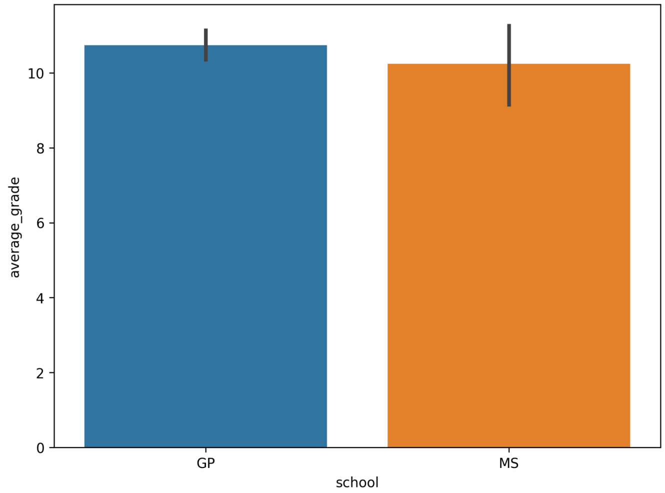
Extracurriculars:
The predictors related to a student's free time, romantic relationships, how much they go out or how much alcohol they consume have been grouped together under the extracurriculars category.

Personal:
The attributes consist of an individual's demographics including their age, sex, address and health.

Exploratory Data Analysis

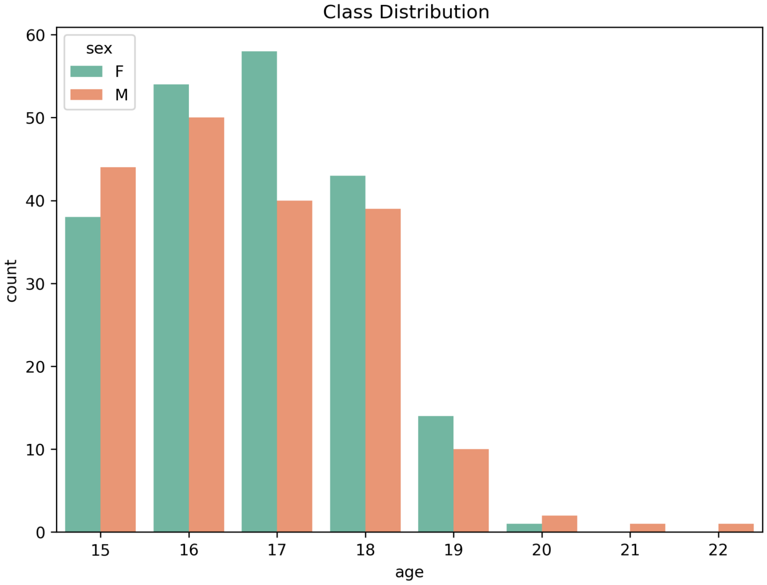
Here are some of the insights we have got from the dataset:

Average Grades by School



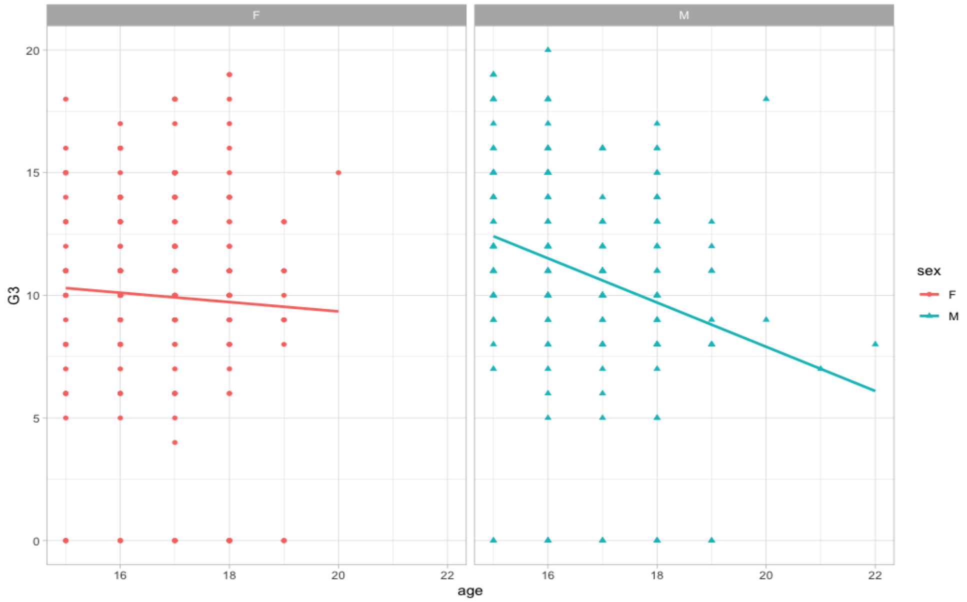
From the graph we can see that school does not play a significant role for a student's grade prediction as they have a similar trend of grades overall.

Average Grades by School



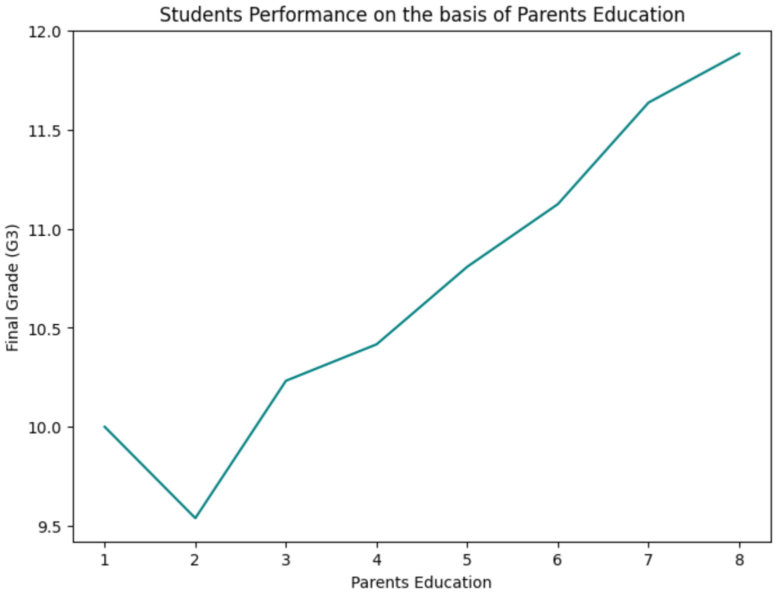
From the graph we can understand that students' age range is from 14 to 22 years. The histogram shows that most of the students are aged between 15 and 18, which makes sense since most students start high school around the age of 15 and graduate by 18 given the fact that generally, high schools around the world last 3-4 years. However, there are 29 students older than 18 years of age. On average, female students are more than male students.

Do girls perform better than boys?



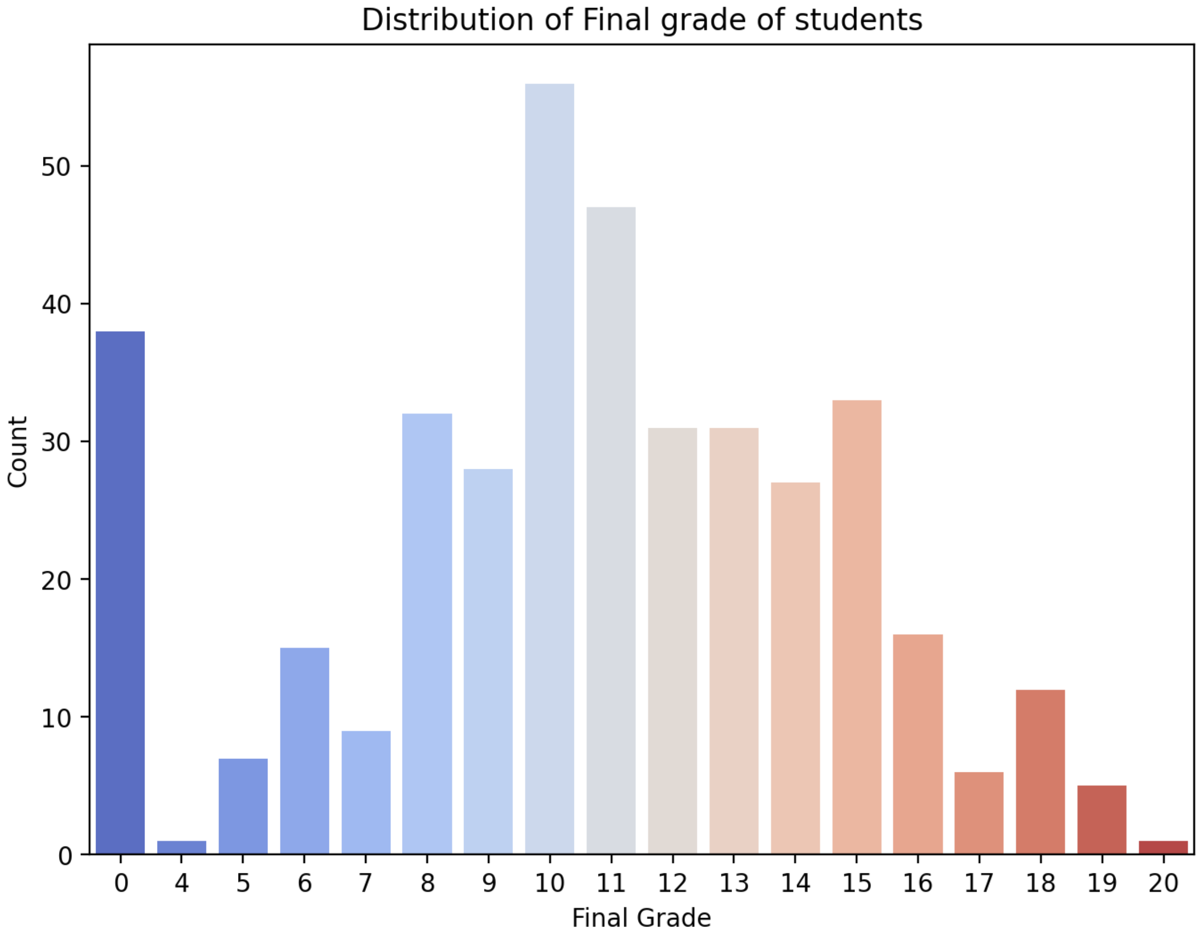
It can be seen that girls' performance improves with age, however, a decrease in the boys' performance can be detected in the graph.

Student's Performance on the basis of Parents' Education



As expected, the higher the level of parents' education is, the higher their children's score at school. Students whose parents have higher levels of education may have an enhanced regard for learning, more positive ability and beliefs, a stronger work orientation, and may use more effective learning strategies.

Distribution of Final Grade of Students

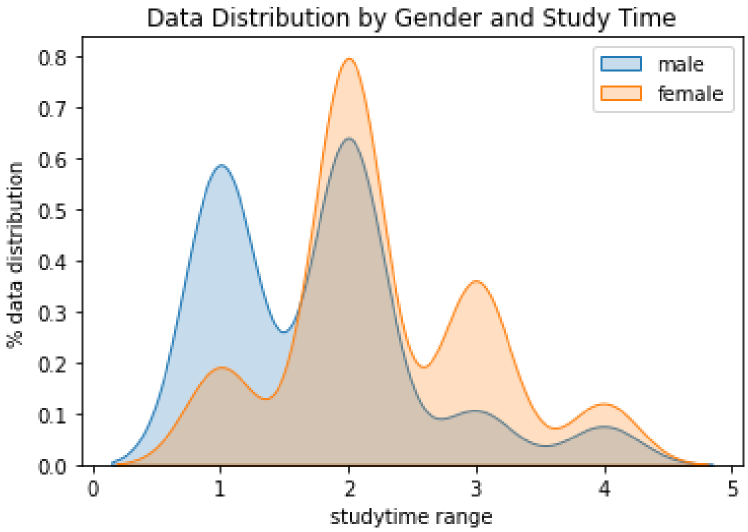


Most students can be seen to have received a borderline passing score of 10 or 11.

While many students received a zero, which was the 3rd highest count in this case, i.e 9.62% of the students.

Very few students received a full score of 20 which is 0.2% of the class size.

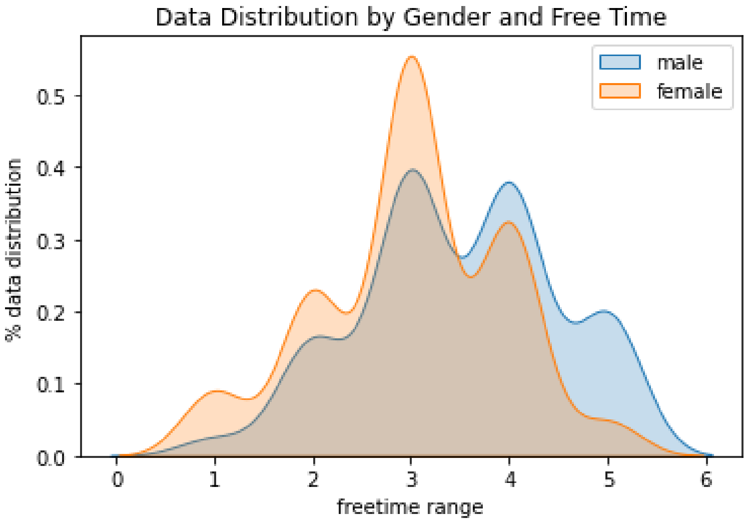
Study Time vs Free Time



Weekly study time:
1 < 2 hours **3** 5 to 10 hours
2 2 to 5 hours **4** > 10 hours

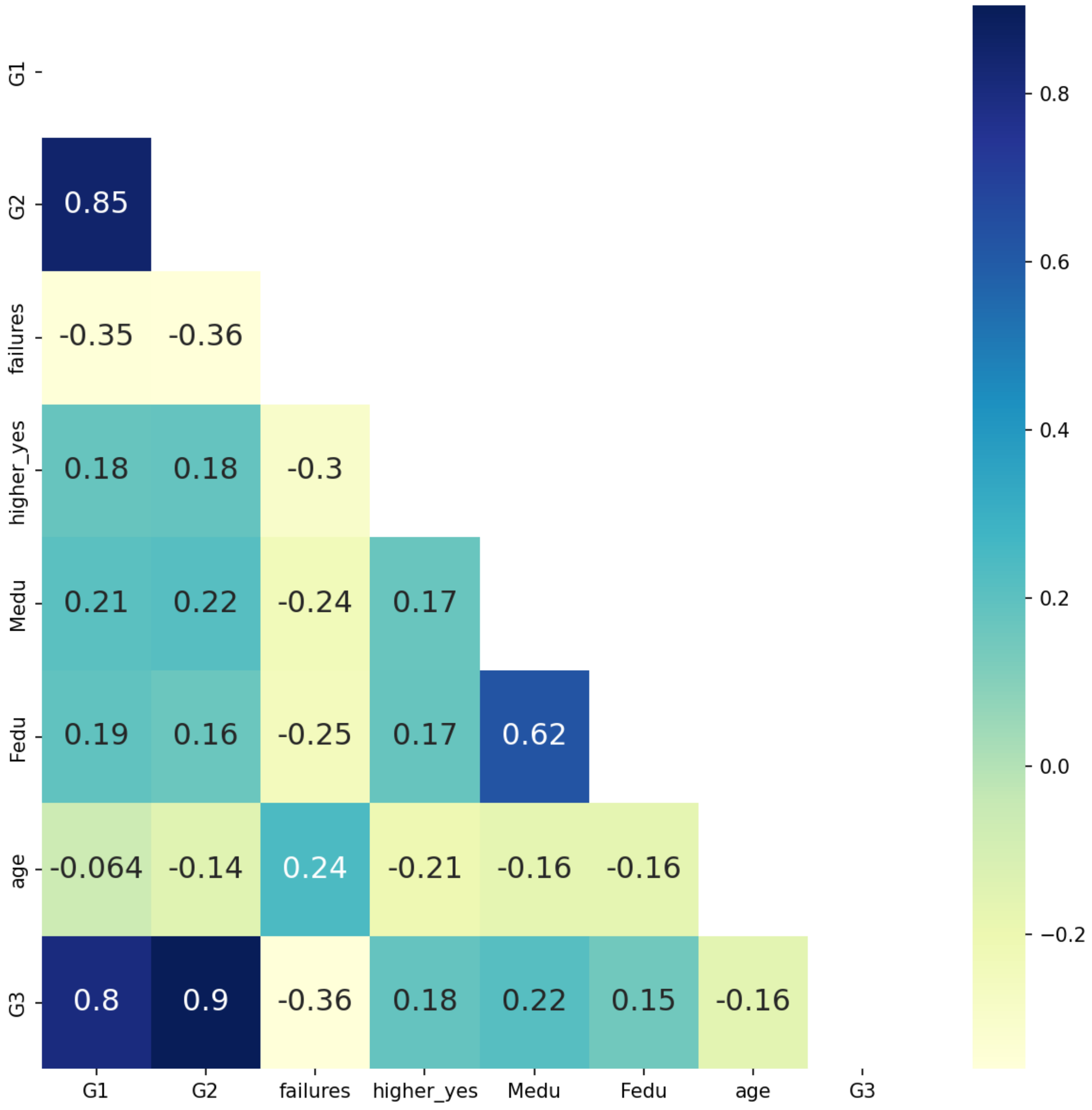
It can be seen that male students hardly studied for 2 hours a day, however most girls studied between 2.5 hrs to 5.5 hrs. Moreover, few percent of female students have also studied for more than 5 hours a day.

Free time after school:
from **1** (very low)
to **5** (very high)



While female students study for more hours a day than male counterparts, as a result they get less free time than male students.

Correlation



Our target variable is ‘G3’, so we check the correlation of other variables with it and find that ‘G1’, ‘G2’, ‘failures’, ‘higher_yes’, ‘Medu’, ‘Fedu’, ‘age’ are amongst the top most correlated variables with ‘G3’.

Preprocessing

i) Removing Null Values

Searched for NULL values to remove or impute but did not find any in this dataset.

ii) One-hot Encoding

In order to use the machine-learning libraries we must convert the categorical data variables into dummy variables. This is called a One-Hot encoding technique where unique values in each categorical variable get a separate column.

iii) Drop Redundant Columns

We dropped the redundant columns which represent the same type of information such as romantic_no, famsize_LE3 as the info is captured in the columns romantic_yes, famsize_GT3.

iv) Train-Test Split

We have used *’train_test_split’* to split the dataset such that 75% of the data is used for training and 25% for testing.

v) Normalization

Rescaling of the data from the original range so that all values are within the new range of 0 and 1 is called Min-Max normalization. We used min-max scalar so that all the features are with-in [0,1] range.

Model

We worked on 3-scenarios:

- Setup-A : When the grades ‘G1’ and ‘G2’ both were excluded.
- Setup-B : When the grade ‘G1’ is only included.
- Setup-C : When the grades ‘G1’ and ‘G2’ both were included.

We have addressed both **Regression** and **Classification** problems.

The main difference is set in-terms of output representation, the output variable is continuous in a regression problem and discrete for a classification setting.

Regression

We implemented the following regression models.

1. Linear Regression
2. Ridge Regression
3. Support Vector Regression
4. Decision Tree Regression
5. Random Forest Regression

We have used “Mean Squared Error” (MSE) as the metric to measure the performance of the regression models. The lower the value the better the performance.

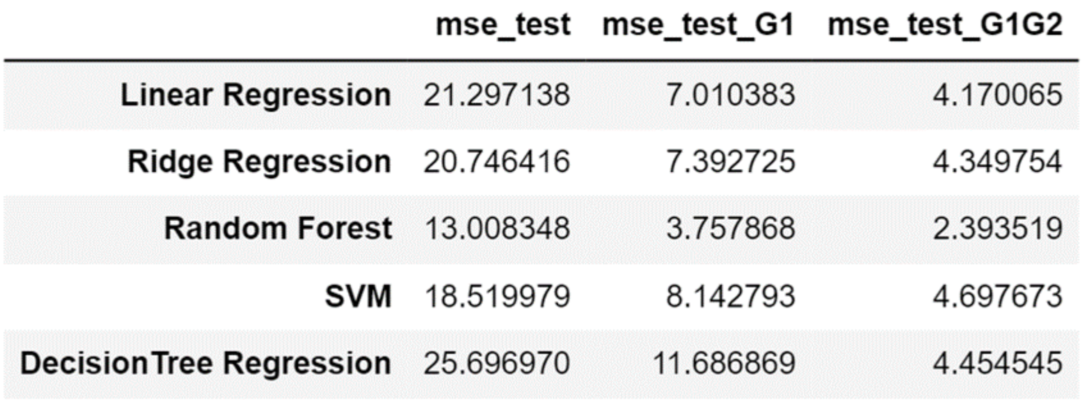
$$MSE = \frac{\sum(y_i - \hat{y}_i)^2}{n}$$

where n = the number of observations, $\hat{y}_i$ = estimated value of response, y_i = true value of the response variable

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \text{ , } 0 \le R^2 \le 1$$

where:
SSR = amount of variation in y that is systematic and can be explained via the linear model, i.e. the linear relationship between y and x
SSE = amount of variation in y that is unsystematic i.e. not explainable through x and is purely due to random error or noise unrelated to x
SST = Total Sum of Squares

Comparison of Mean Squared Error values on test set for various models and setups



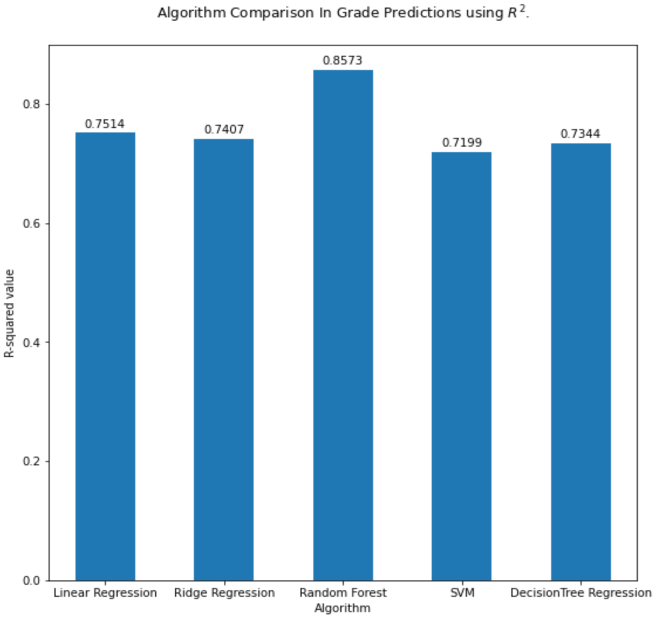
	mse_test	mse_test_G1	mse_test_G1G2
Linear Regression	21.297138	7.010383	4.170065
Ridge Regression	20.746416	7.392725	4.349754
Random Forest	13.008348	3.757868	2.393519
SVM	18.519979	8.142793	4.697673
DecisionTree Regression	25.696970	11.686869	4.454545

From above we can see that:

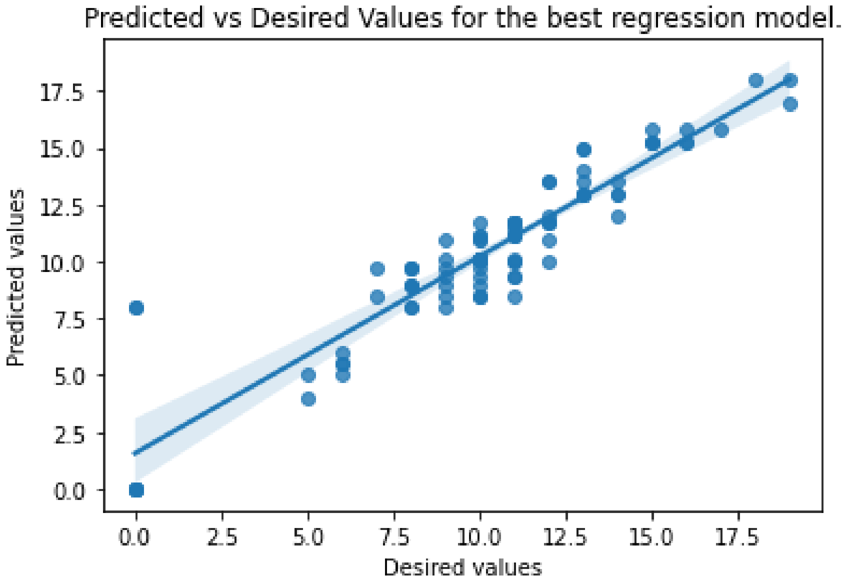
Inclusion of either G1 or both G1 and G2 variables have reduced the error values in all the algorithms.

It can be seen that the Random Forest Regression model out-performed other models.

We also checked the Coefficient of Determination (R-squared value) as well and found out that Random Forest has the highest R-squared value.



After the comparison of R-squared values for various regression models, it was found that random forest algorithms performed better. This can be seen from above graph.

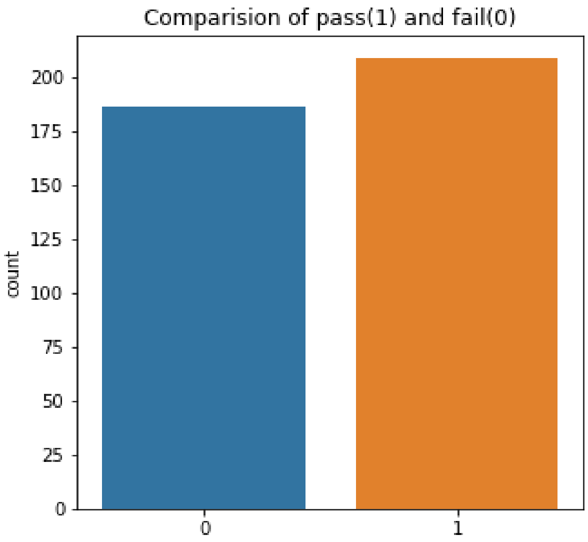


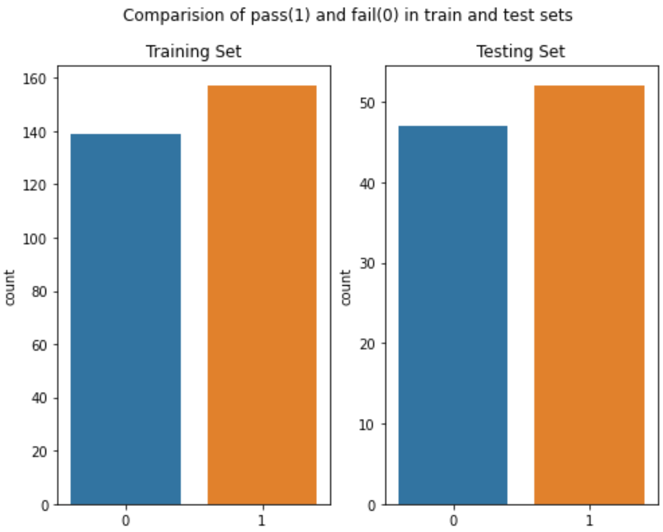
After finding the best regression model, we have plotted a graph between predicted values and desired (true) values of response.

Classification

We have addressed the classification problem for this dataset as well due to its dual nature of variables and existence of categorical features.

For this first of all we need to have our response variable to be in terms of labels.
If a student scores greater than 10 in ‘G3’ he passes the exam, else he fails.
So, we encoded the response variable to 1 if ‘G3 > 10’, else 0.





We used stratified ‘train_test_split’ to split the training and testing sets so that they are representative of the original data.

Receiver Operating Characteristic Curve

An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters:

- True Positive Rate
- False Positive Rate

True Positive Rate (TPR) is a synonym for recall and is therefore defined as follows:

TPR = TP / TP+FN

False Positive Rate (FPR) is defined as follows:

FPR = FP / FP+TN

Area Under the Curve

AUC stands for "Area Under the ROC Curve".

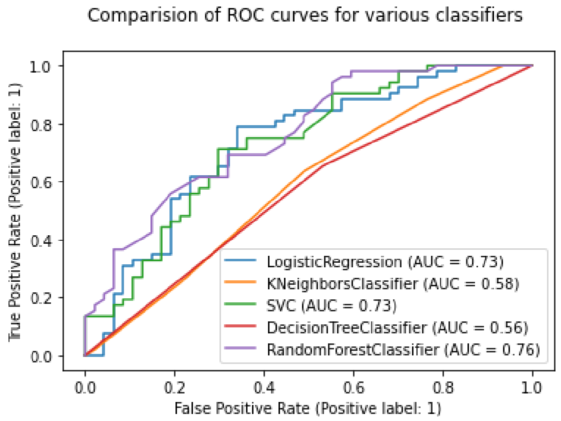
AUC measures the entire two-dimensional area underneath the entire ROC curve (think integral calculus) from (0,0) to (1,1).

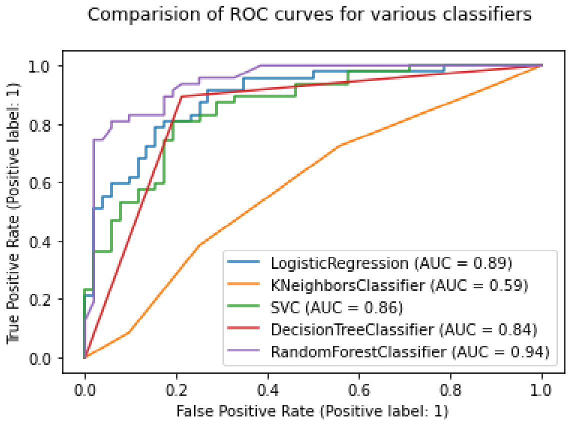
Results

We tried various classifications algorithms namely:

1. LogisticRegression
2. KNeighborsClassifier
3. Support Vector Classifier (SVC)
4. DecisionTreeClassifier
5. RandomForestClassifier

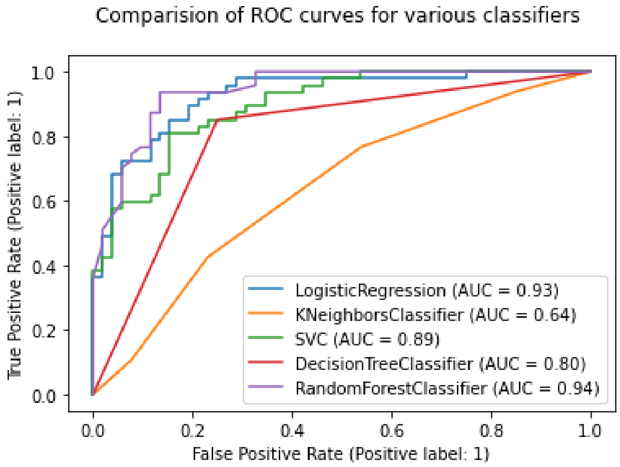
We measured the distribution of the classes in the output variable:



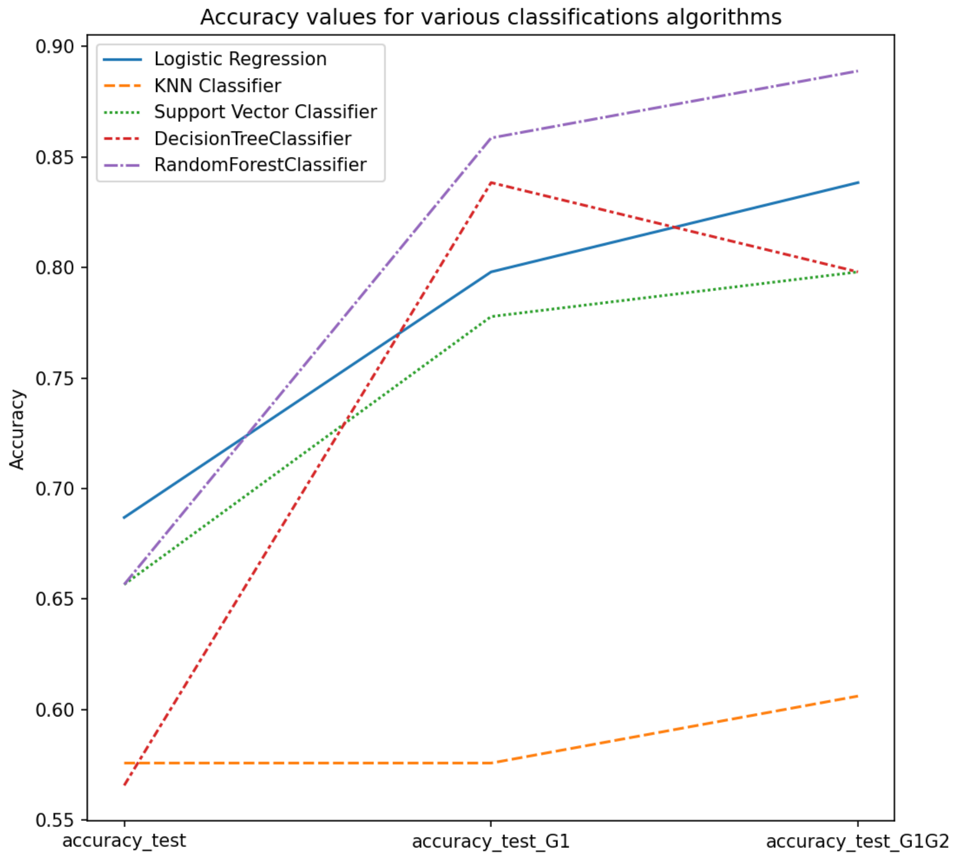


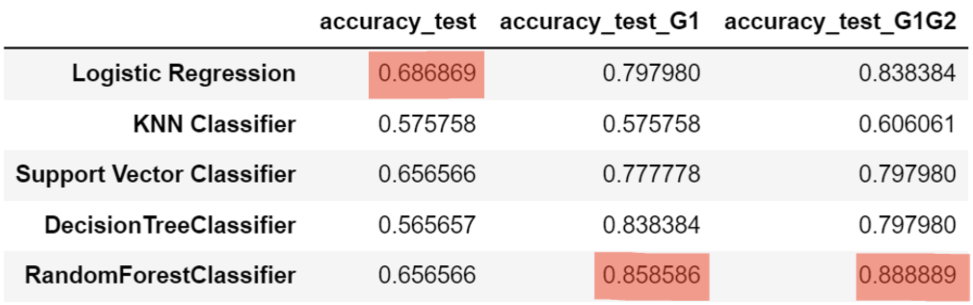
Setup-A: When the grades 'G1' and 'G2' both were excluded.

Setup-B: When the grade 'G1' is only included



Setup-C: When the grades 'G1' and 'G2' both were included.





	accuracy_test	accuracy_test_G1	accuracy_test_G1G2
Logistic Regression	0.686869	0.797980	0.838384
KNN Classifier	0.575758	0.575758	0.606061
Support Vector Classifier	0.656566	0.777778	0.797980
DecisionTreeClassifier	0.565657	0.838384	0.797980
RandomForestClassifier	0.656566	0.858586	0.888889

By adding G1 only, or G1 and G2 both variables increases the accuracy score in all algorithms except in the Decision Tree Classifier.

Also, excluding both G1 and G2 the accuracy of the Logistic Regression model is best.

The Random Forest model gives better accuracy in all other scenarios.

Improving the Results

1) **Bagging (Bootstrap Aggregation)**

Decisions trees are very sensitive to the data they are trained on — small changes to the training set can result in significantly different tree structures. Random forest takes advantage of this by allowing each individual tree to randomly sample from the dataset with replacement, resulting in different trees. Instead of the original training data, it takes a random sample of size N with replacement.

2) **Hyperparameter tuning:**

Hyperparameters are used in random forests to either enhance the performance and predictive power of models or to make the model faster. Following hyperparameters increased the predictive power for our model:

i) n_estimators– number of trees the algorithm builds before averaging the predictions.

ii) max_features– maximum number of features random forest considers splitting a node.

iii) mini_sample_leaf– determines the minimum number of leaves required to split an internal node.

We used **stratifiedKFold** for tuning the hyperparameters such as max_depth, n_estimators.We used 10-Fold CV for tuning and used **cross_val_score** with scoring metric set to “accuracy”. We found the max_depth of 6 gives better accuracy for this setting and we plotted the tree from it.

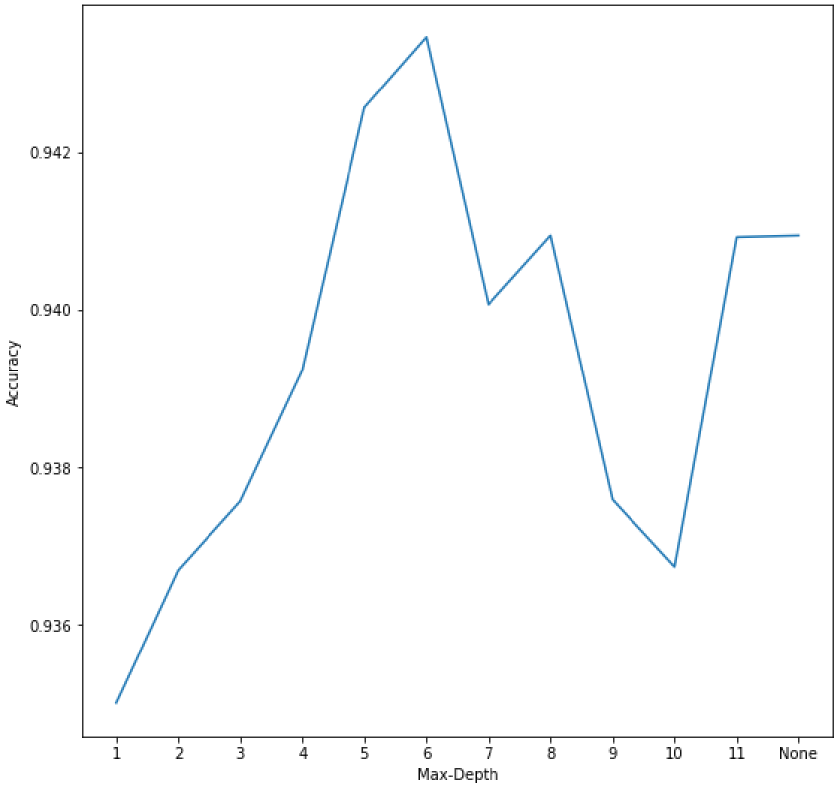
3) **Splitting Criteria:**

The decision-trees from the random-forest classifier uses **gini-index** for the splitting criteria:

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2$$

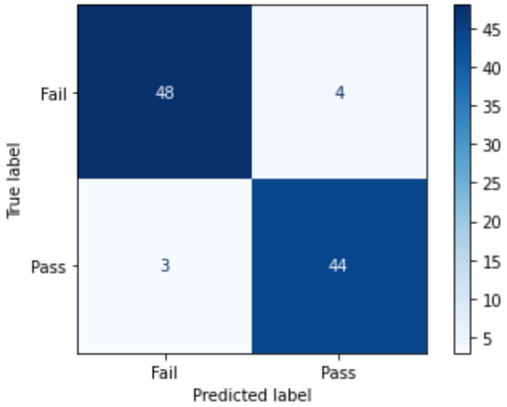
where Pi denotes the probability of an element being classified for a distinct class.

Mean Classification Accuracy at various max_depth values



For the **max_depth = 6** and **n_estimators = 51**, the random forest produced the accuracy of **0.92929**.

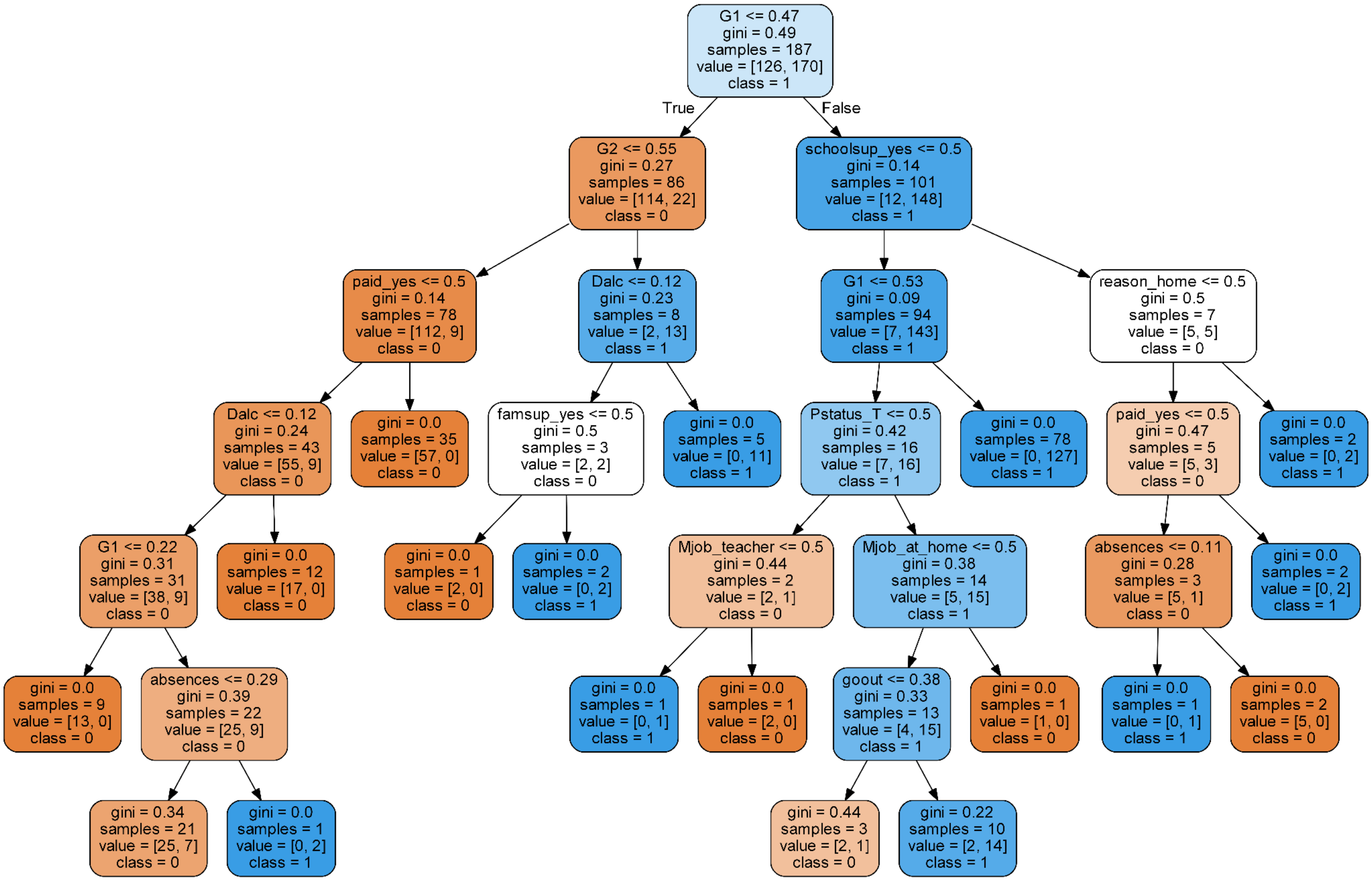
Confusion Matrix for the Random Forest



From the above confusion matrix, we can see that our model's mis-classifications are minimal.

The resulting tree shows one of the decision-trees from the random-forest classifier which uses gini-index for the splitting criteria.

Therefore, the best random forest is plotted below:



G1 <= 0.47 gini = 0.49 samples = 187 value = [126, 170] class = 1
True
False
G2 <= 0.55 gini = 0.27 samples = 86 value = [114, 22] class = 0
schoolsup_yes <= 0.5 gini = 0.14 samples = 101 value = [12, 148] class = 1
paid_yes <= 0.5 gini = 0.14 samples = 78 value = [112, 9] class = 0
Dalc <= 0.12 gini = 0.23 samples = 8 value = [2, 13] class = 1
G1 <= 0.53 gini = 0.09 samples = 94 value = [7, 143] class = 1
reason_home <= 0.5 gini = 0.5 samples = 7 value = [5, 5] class = 0
Dalc <= 0.12 gini = 0.24 samples = 43 value = [55, 9] class = 0
gini = 0.0 samples = 35 value = [57, 0] class = 0
famsup_yes <= 0.5 gini = 0.5 samples = 3 value = [2, 2] class = 0
gini = 0.0 samples = 5 value = [0, 11] class = 1
Pstatus_T <= 0.5 gini = 0.42 samples = 16 value = [7, 16] class = 1
gini = 0.0 samples = 78 value = [0, 127] class = 1
paid_yes <= 0.5 gini = 0.47 samples = 5 value = [5, 3] class = 0
gini = 0.0 samples = 2 value = [0, 2] class = 1
G1 <= 0.22 gini = 0.31 samples = 31 value = [38, 9] class = 0
gini = 0.0 samples = 12 value = [17, 0] class = 0
gini = 0.0 samples = 1 value = [2, 0] class = 0
gini = 0.0 samples = 2 value = [0, 2] class = 1
Mjob_teacher <= 0.5 gini = 0.44 samples = 2 value = [2, 1] class = 0
Mjob_at_home <= 0.5 gini = 0.38 samples = 14 value = [5, 15] class = 1
absences <= 0.11 gini = 0.28 samples = 3 value = [5, 1] class = 0
gini = 0.0 samples = 2 value = [0, 2] class = 1
gini = 0.0 samples = 9 value = [13, 0] class = 0
absences <= 0.29 gini = 0.39 samples = 22 value = [25, 9] class = 0
gini = 0.0 samples = 1 value = [0, 1] class = 1
gini = 0.0 samples = 1 value = [2, 0] class = 0
goout <= 0.38 gini = 0.33 samples = 13 value = [4, 15] class = 1
gini = 0.0 samples = 1 value = [1, 0] class = 0
gini = 0.0 samples = 1 value = [0, 1] class = 1
gini = 0.0 samples = 2 value = [5, 0] class = 0
gini = 0.34 samples = 21 value = [25, 7] class = 0
gini = 0.0 samples = 1 value = [0, 2] class = 1
gini = 0.44 samples = 3 value = [2, 1] class = 0
gini = 0.22 samples = 10 value = [2, 14] class = 1

The random forest uses gini-index criteria for making a decision at a node. Similarly, for each split, we will calculate the Gini impurities and the split producing minimum Gini impurity will be selected as the split. The minimum value of Gini impurity means that the node will be purer and more homogeneous.

Assumptions & Limitations

We assume that all other unknown parameters or factors that could affect grades are the same.

Students who got zero as their final grade were 9.62% of the total, which is a rare case as not many students get zero in their final exams and it may be due to external factors.

Since 9.87% students were over the age of 18 therefore according to Portugal's minimum legal drinking age only they were alcohol consumers in the dataset.

We know that the dataset has a very small size with too many features which may possibly cause overfitting.

Conclusion

The results show that a good predictive accuracy can be achieved, provided that the first and/or second school period grades are available. Although student achievement is highly influenced by past evaluations, an explanatory analysis has shown that there are also other relevant features (e.g. performance by gender, parent's job and education, time spent studying vs time spent free).

As a direct outcome of this project, more efficient student prediction tools can be developed, improving the quality of education and enhancing school resource management.