Student Grade Prediction

Definition: Predict student performance in secondary education (high school).

Description:

This data approach 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. Two datasets are provided regarding the performance in two distinct subjects: Mathematics (mat) and Portuguese language (por). Here, G3 is the final year grade (issued at the 3rd period), while G1 and G2 correspond to the 1st and 2nd period grades.

Data Set Characteristics:	Multivariate
Attribute Characteristics:	Integer
Number of Instances:	649
Number of Attributes:	33
Missing Values?	N/A

Objective:

- Predicting student's first period grade, second period grade and final grade.
- Finding out major factors affecting Students' grades.

Data Source Link:

https://archive.ics.uci.edu/ml/datasets/Student+Performance

Introduction:

Data is becoming the new oil of the 21st century, and the fields of Business Intelligence (BI)/Data Mining (DM) offer interesting automated tools which could enable us to know how to drill and refine it, that is, how to produce data and turn them into wisdom, information and knowledge. Using BI/DM techniques, the student achievement in secondary education have been analyzed. Student details (e.g. student grades, demographic, social and school related features) were collected by using school reports and questionnaires. The two core classes (i.e. Mathematics and Portuguese) were taken into consideration under binary/five-level classification and regression tasks. Also, one DM models (i.e. Decision Trees) and three input selections (e.g. with and without previous grades) were tested.

Attributes' Detail:

Following is the list of all the attributes, its datatypes and its possible values that we have in our data.

Number	Attribute Name:	Datatype	Possible Values
1	school - student's school	Binary	GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira
2	sex - student's sex	Binary	F' - female or 'M' - male
3	age - student's age	Numeric	from 15 to 22
4	address - student's home address type	Binary	U' - urban or 'R' - rural
5	famsize - family size	Binary	LE3' - less or equal to 3 or 'GT3' - greater than 3
6	Pstatus - parent's cohabitation status	Binary	T' - living together or 'A' - apart
7	Medu - mother's education	Numeric	0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education
8	Fedu - father's education	Numeric	0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education
9	Mjob - mother's job	Nominal	teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other'
10	Fjob - father's job	Nominal	teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other'
11	reason - reason to choose this school	Nominal	close to 'home', school 'reputation', 'course' preference or 'other'
12	guardian - student's guardian	Nominal	mother', 'father' or 'other'
13	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
14	studytime - weekly study time	Numeric	1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours
15	failures - number of past class failures	Numeric	n if 1<=n<3, else 4
16	schoolsup - extra educational support	Binary	yes or no
17	famsup - family educational support	Binary	yes or no

18	paid - extra paid classes within the course subject (Math or Portuguese)	Binary	yes or no
19	activities - extra-curricular activities	Binary	yes or no
20	nursery - attended nursery school	Binary	yes or no
21	higher - wants to take higher education	Binary	yes or no
22	internet - Internet access at home	Binary	yes or no
23	romantic - with a romantic relationship	Binary	yes or no
24	famrel - quality of family relationships	Numeric	from 1 - very bad to 5 - excellent
25	freetime - free time after school	Numeric	from 1 - very low to 5 - very high
26	goout - going out with friends	Numeric	from 1 - very low to 5 - very high
27	Dalc - workday alcohol consumption	Numeric	from 1 - very low to 5 - very high
28	Walc - weekend alcohol consumption	Numeric	from 1 - very low to 5 - very high
29	health - current health status	Numeric	from 1 - very bad to 5 - very good
30	absences - number of school absences	Numeric	from 0 to 75
	# these grades are related with the course subject, Math or Portuguese:		
31	G1 - first period grade	Numeric	from 0 to 20
32	G2 - second period grade	Numeric	from 0 to 20
33	G3 - final grade	Numeric	from 0 to 20, output target

Initial Analysis:

Given data contains two different datasets which includes various social, demographic and school related factors of students and their performances in two different subjects - Mathematics and Portuguese.

Initially, the analysis has been provided for the student data based upon the Mathematics subject. Analysis will be done on Portuguese data also in the future, and we will try to find out relations between both the datasets.

Observations

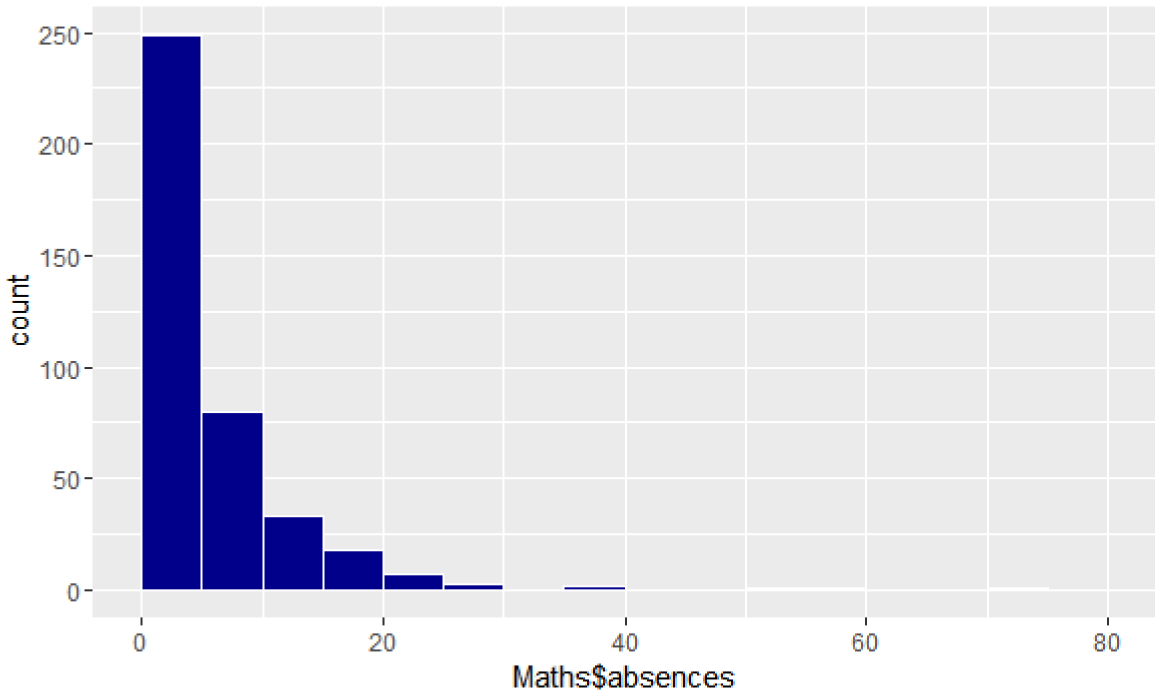
Observation 1:

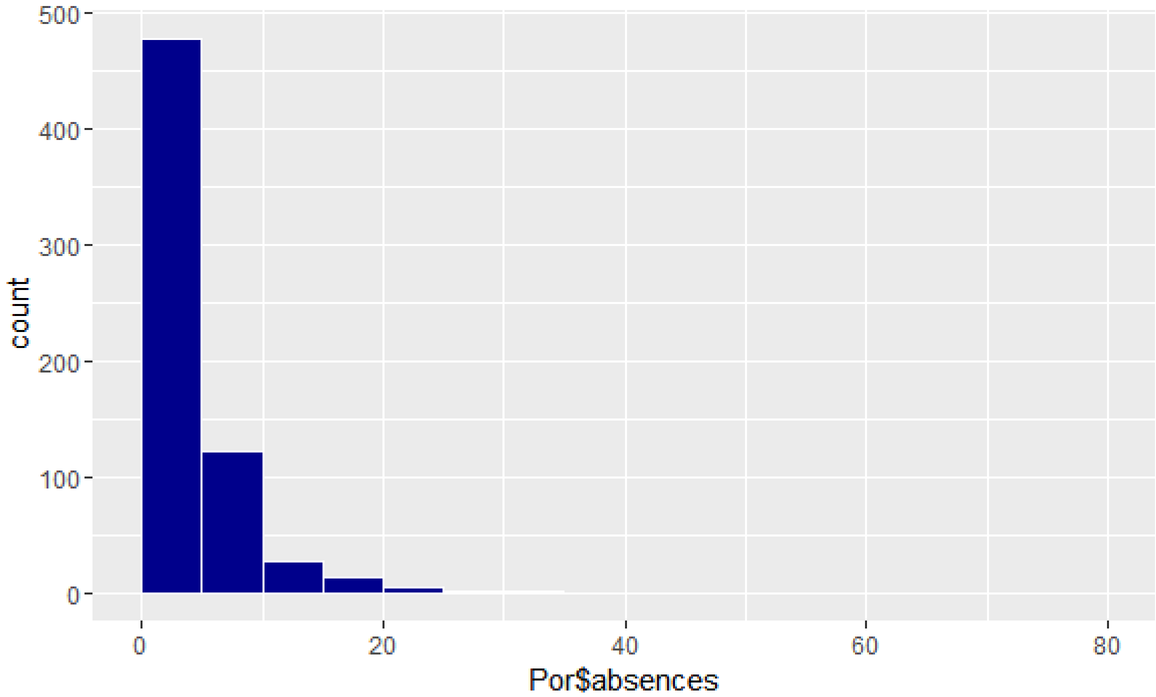
Missing Data: After observing the data it can be inferred that there are no missing values have been present.

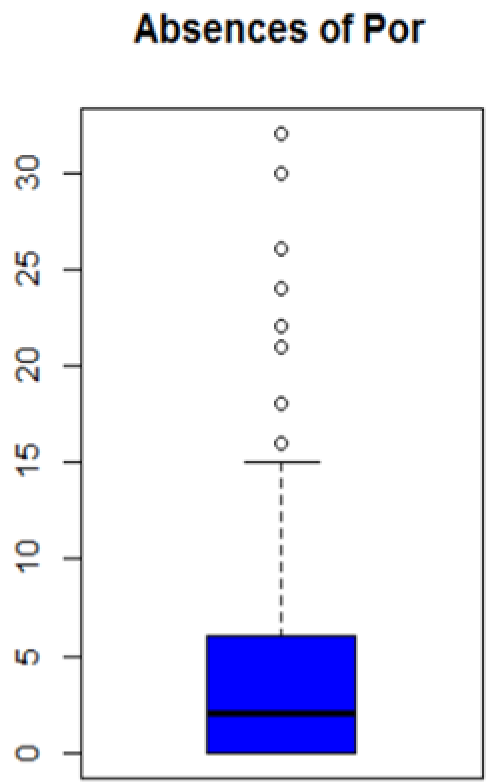
Observation 2:

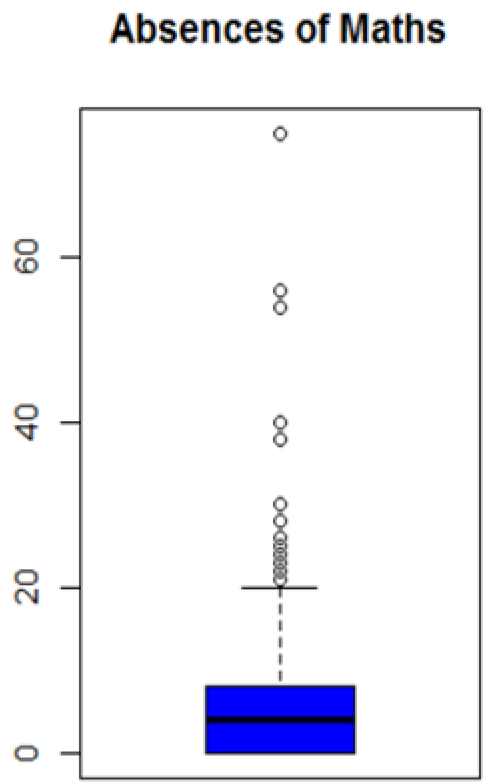
Outlier Analysis (Unusual data Values):

Below is the outlier analysis of students based on their absences in the class. The result shows that most of the students have their absence rate between 0 and 20 but there are a few students who have an absence rate more than 50. Those students can be considered as outliers.


250
200
150
100
50
0
count
0
20
40
60
80
Maths$absences


500
400
300
200
100
0
count
0
20
40
60
80
Por$absences





Observation 3:

Correlation Between G1, G2 (Independent Variable) and G3 (Predicted Variable)

Correlation	G1	G2	G3
G1	100%	85%	80%
G2	85%	100%	90%
G3	80%	90%	100%

By looking at the values in the above table, we can conclude that G1 and G2 (the marks of the previous semesters) are highly correlated with our target variable G3 (marks of the current semester). So, it should get reflected in the outcome of the prediction model.

Data Analysis

Classification algorithm Data Preparation

Divide the prediction variable into 5 categories. The count shows the distribution of the grades.

Grade Distribution					
Score	0-9	10-11	12-13	14-15	16-20
Result	fail	D	C	B	A
Count	130	103	62	60	40

In order to run the classification algorithm on our data, we must convert our output variable into categorical data. So, we have divided G3 into the grades shown in the table above.

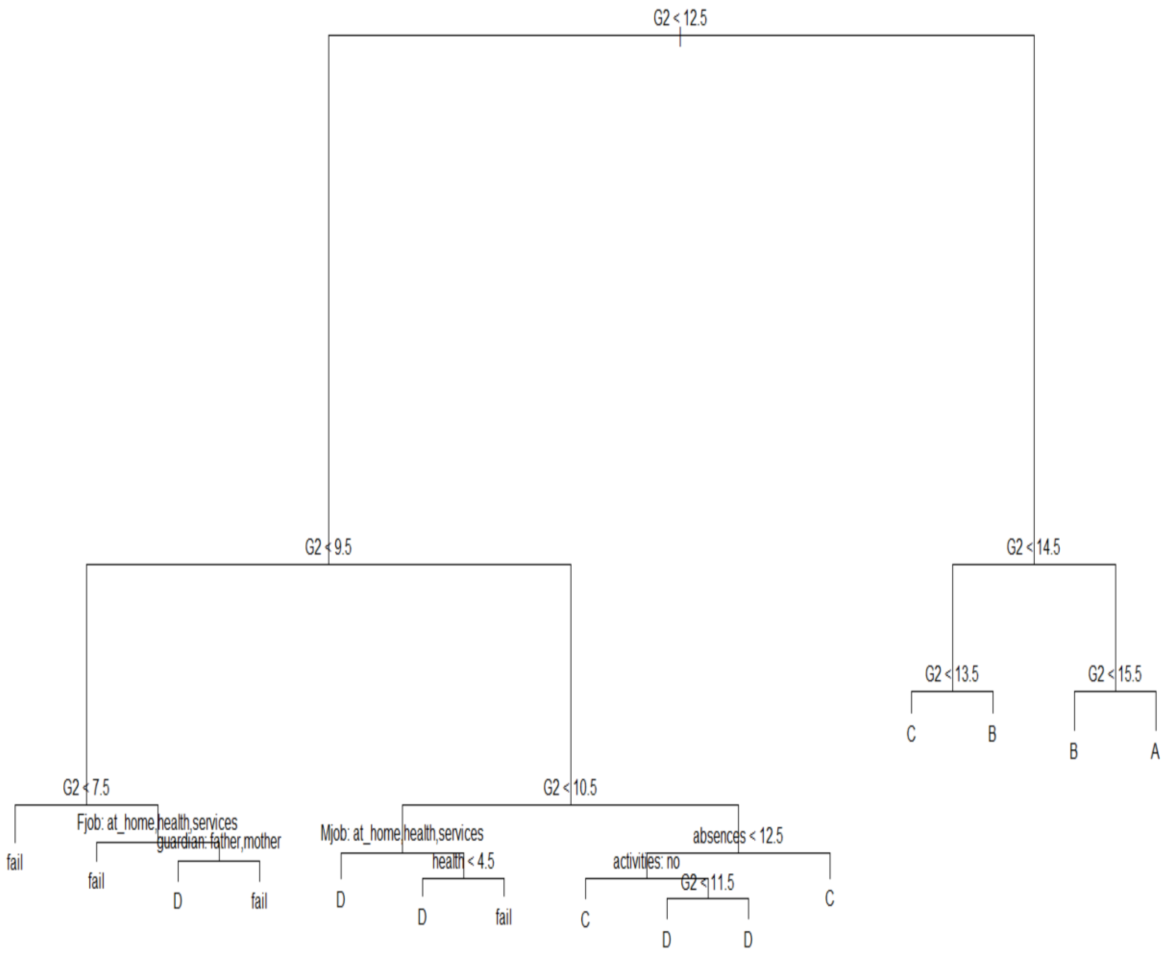
Generating the Test and training data sets

We took 50% of the sample for training our model and other 50% for testing the model.

We can now proceed to apply classification algorithms on the training data.

Algorithm: Decision tree

We applied the decision tree model on the data by giving G3 as a dependent variable and all other variables as independent variables. The generated output of the decision tree is shown below.



So our model has shown the decision tree as shown in the graph. So we can predict the grades of the testing data from this tree and check the accuracy of the predicted grades.

$$\text{Predicted Accuracy} = \frac{\text{number of correctly predicted observation}}{\text{total number of observations}} * 100$$

The accuracy of the predicted grades is approximately 72%. The distribution of the predicted grades and the actual grade is shown in below table.

	Actual Grade				
Predicted Grade	A	B	C	D	fail
A	13	1	0	0	0
B	4	24	2	0	0
C	0	7	20	15	0
D	0	0	10	32	11
fail	0	0	0	6	53

The Green color shows the correct number of prediction of the result. The yellow color shows the error in the prediction value.

Analysis With G1 and G2:

Description:

Predicting the student's performance in the final exam (G3) including G1 and G2. (Performance of the student in 1st and 2nd Exam).

This analysis can be useful for the school if they want to predict the grades or the performance of existing student in the final exam.

Models Implemented

1. Decision Tree
2. Gradient boosting Machine
3. Random Forest

Algorithms:

1) Gradient Boosting Machine Algorithm:

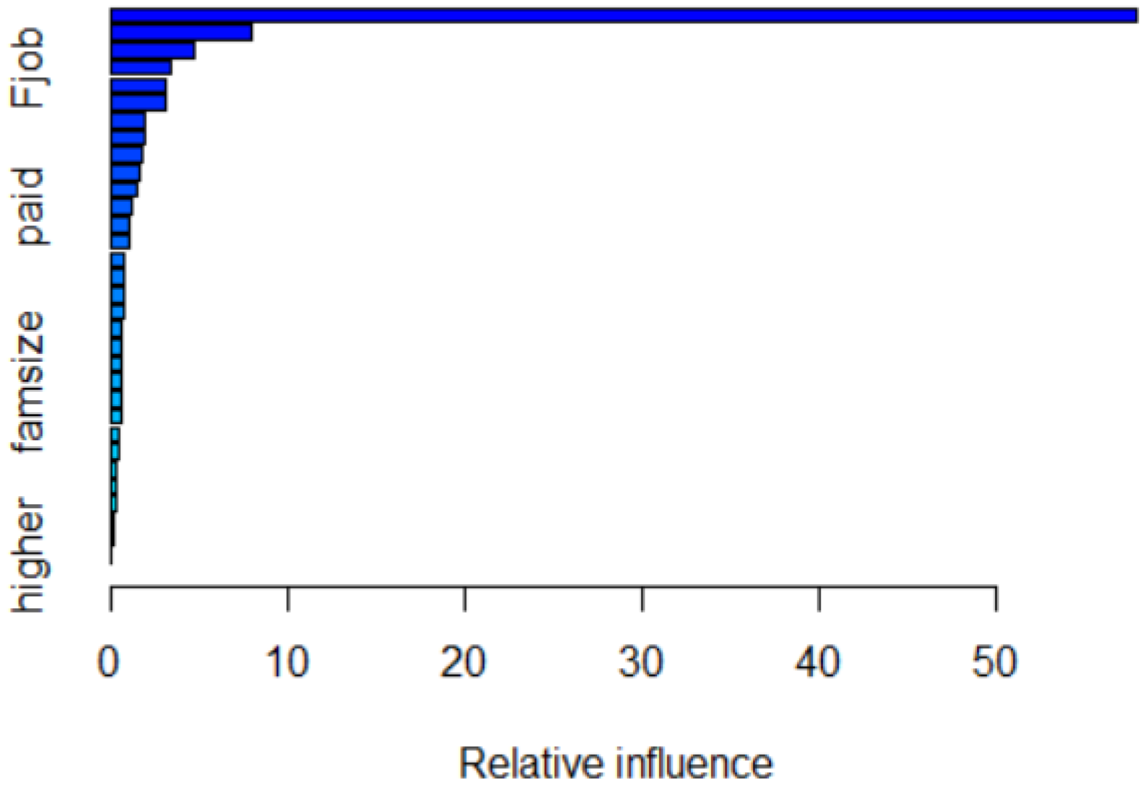
Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function.

The idea of gradient boosting originated in the observation that boosting can be interpreted as an optimization algorithm on a suitable cost function. This functional gradient view of boosting has led to the development of boosting algorithms in many areas of machine learning and statistics beyond regression and classification.

Accuracy of this model is 81.09% which means that it predicts the grades of 81% of the students correctly.

	A	B	C	D	Fail
A	13	1	0	0	0
B	4	23	2	0	0
C	0	2	5	3	0
D	0	0	7	30	2
Fail	0	0	0	10	62

Gradient boosting algorithm gives us the list of significant variables also. According to this model 'Fjob' is the significant variable here.

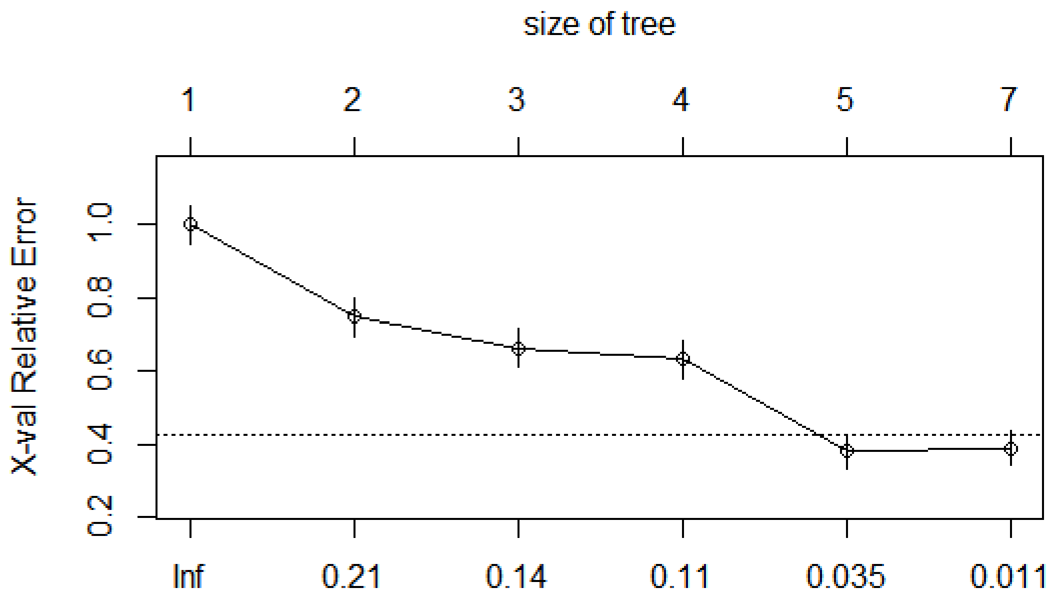


2) R part

R part is a type of Decision Tree. Decision tree is a graph to represent choices and their results in form of a tree. The nodes in the graph represent an event or choice and the edges of the graph represent the decision rules or conditions.

Before pruning

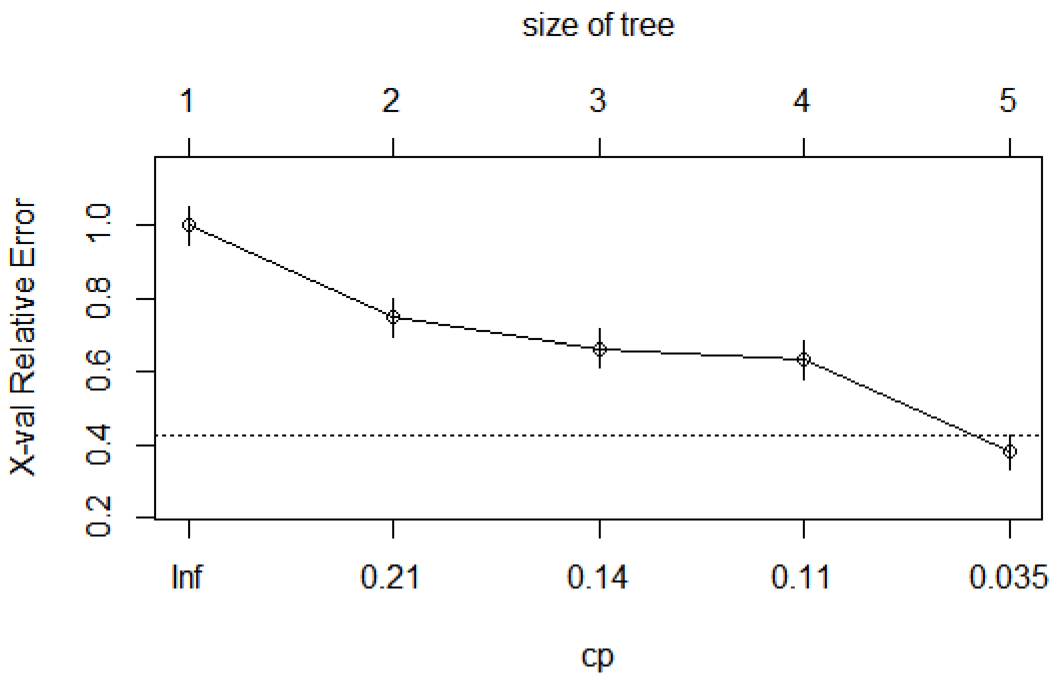
	A	B	C	D	Fail
A	13	1	0	0	0
B	4	24	2	0	0
C	0	7	28	8	0
D	0	0	2	35	5
Fail	0	0	0	10	59



Accuracy: 80.30%

After Pruning

	A	B	C	D	Fail
A	13	1	0	0	0
B	4	24	2	0	0
C	0	7	28	8	0
D	0	0	2	33	2
Fail	0	0	0	12	62



Accuracy: 80.80%

3) Random Forest Algorithm

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set.
The first algorithm for random decision forests was created using the random subspace method, which, in Ho's formulation, is a way to implement the "stochastic discrimination" approach to classification proposed by Eugene Kleinberg.

	A	B	C	D	Fail
A	13	2	0	0	0
B	4	24	1	0	0
C	0	4	22	10	0
D	0	2	9	33	2
Fail	0	0	0	10	62

Accuracy: 78.28%

Hence, the best predictive model for our analysis is **Gradient Boosting Machine Algorithm** because it predicts students' grades with highest possible accuracy.

Conclusion:

Above models give the students' performance prediction when we are including parameters G1 and G2 in our analysis. These models will help the school to predict grades of the students. So that the school officials can be prepared for the results of final exam which is G3 in our data and they can take necessary actions.

Analysis Without G1 and G2

Description:

Predicting the Student's performance (G3) without including G1 and G2.

This Analysis can be useful to predict the performance of the incoming new students by just observing the background of the student.

Models Implemented

1. Decision Tree
2. Gradient boosting Machine
3. Random Forest
4. Logistic regression

Algorithms:

1) **Gradient Boosting Machine (GBM) Algorithm:**

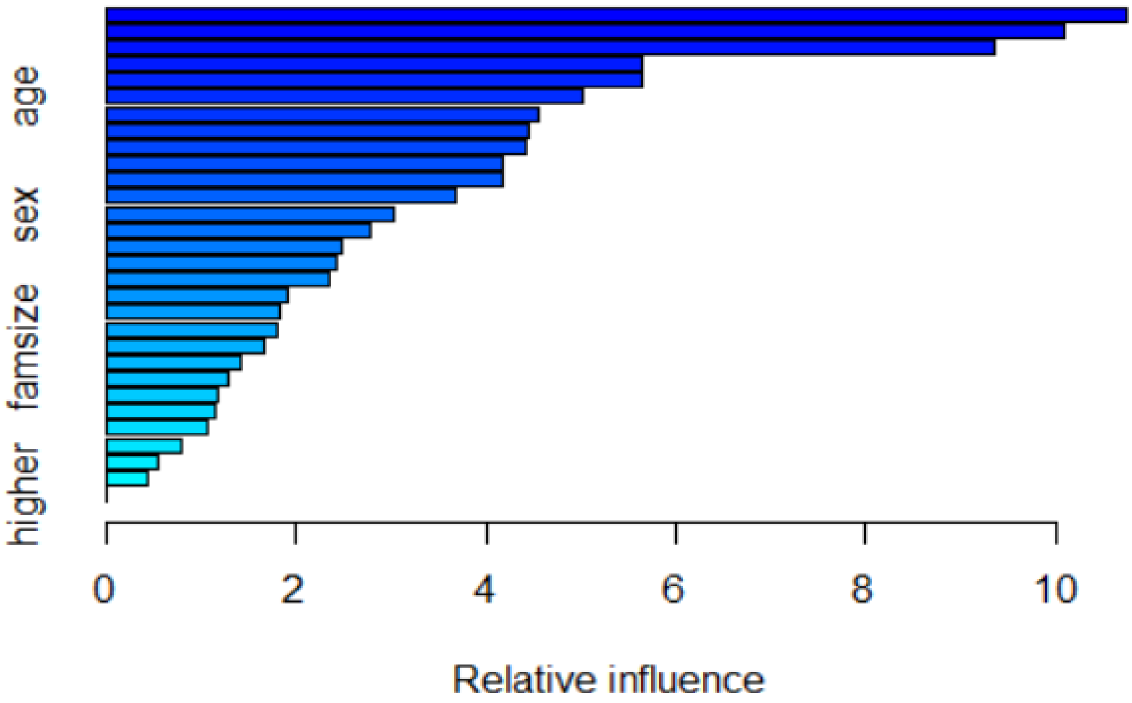
Here, we are converting the result to binary in 'Pass' and 'Fail'. Accuracy for the prediction whether the student will pass or fail in this algorithm is 72.22 %.

The model correctly shows that 26 students will fail and they are getting failed. 117 students will pass and they are getting passed. These are the correct predictions. But there are few incorrect results also which shows that 17 students will fail but they are getting passed and 38 students will pass but they are getting failed.

Confusion Matrix:

	Fail	Pass
Fail	26	17
Pass	38	117

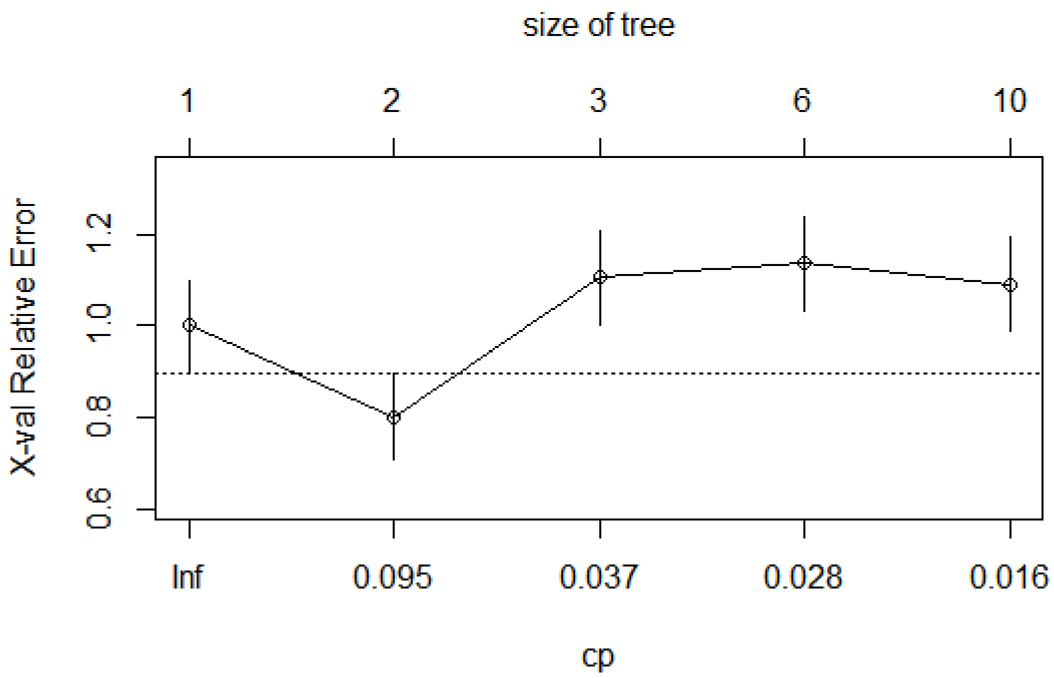
Gradient Boosting algorithm gives us the list of significant variables also. According to this algorithm, 'age', 'sex' and 'famsize' are the significant variables.



2) **Rpart Binary**

Without Pruning

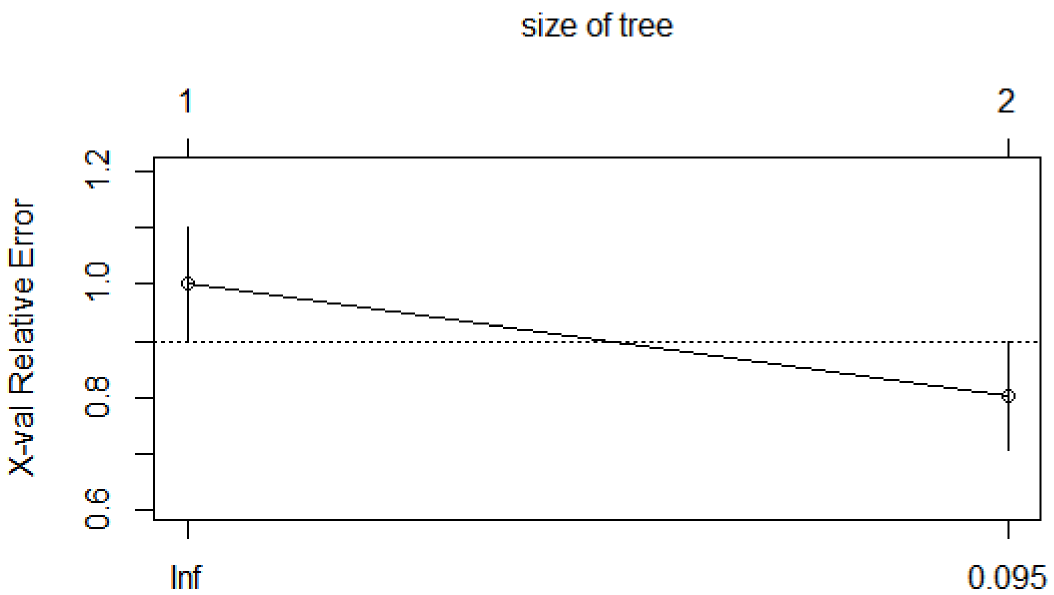
	Fail	Pass
Fail	28	34
Pass	36	100



Accuracy: 64.64%

After Pruning

	Fail	Pass
Fail	24	16
Pass	40	118



3) Random Forest Algorithm

		Actual	
		Fail	Pass
Predicted	Fail	18	14
	Pass	46	120

Accuracy: 69.69%

4) **Logistic Regression**

	Fail	Pass
Fail	27	20
Pass	37	114

Accuracy: 71.21%

Like the analysis done above, here also the best predictive model for our analysis is **Gradient Boosting Machine Algorithm** because it predicts students' grades with highest possible accuracy.

Conclusion:

Above models give the students' performance prediction when we are not considering parameters G1 and G2 in our analysis. These models will help the school to predict grades of the new students. So that the school officials can be prepared for the results of final exam and they can take necessary actions.

Future Steps:

1) Analysis of student data based upon the Portuguese Language.
2) Analysis of student data based upon the Mathematics and Portuguese Language combined
3) Improving the results, getting insights based on other classification algorithms and regression algorithms and box plots
4) Checking accuracy after removing outliers
5) Put the summary of the function, and put the out put of cp and pruning charts