Contents

Final Project Report 1

Title: Student Performance Prediction and Analysis 1

Section 1: Survey 3

1(a): Paper No.1: Using Data Mining to Predict Secondary School Student Performance 3

1(b): Paper No.2: An analysis of students' performance using classification algorithms. 3

1(c): Paper No.3: Data Mining approach for predicting Student performance. 4

1(d): Paper No.4: A review on predicting Students' performance using Data Mining Techniques.. 5

Section 2: Implementation.. 6

Requirements: 6

Finding the correlation for each column:- 6

Results...12

References: 13

Section 1: Survey

1(a): Paper No.1: Using Data Mining to Predict Secondary School Student Performance

Objective: The main objective of the paper was to build an efficient student grade prediction tool for two core classes (i.e. Mathematics and Portuguese) to improve the quality of education and enhance school resource management. This was done by using Business Intelligence (BI)/Data Mining (DM) techniques.

Contribution: This paper presents an education tool for student achievement in secondary education using DM techniques. The significant contribution of this work is that the data of two core classes are modeled using DM models (i.e., Decision Tree, Random Forest, Neural Network and Support Vector Machines) and the PCC(Percentage Correct Classification) of these models are compared with a Naive Classifier. And also, the regression results are compared with each DM model. This gave the model with the highest accuracy to build an efficient education tool.

Method: They have taken the data of student grades in two core classes from 2 schools and modeled it under 3 DM goals: 1) Binary Classification (pass/fail) 2) Five level Classification(I - very good to V - insufficient) 3) Regression: Numeric output ranges between zero (0%) and twenty (100%). They have used the threshold as G3(final grade)>10 is passed. Specifically, they have used RMiner for classification and regression tasks (i.e., rpart (DT), randomForest (RF), neural network (NN), and kernlab (SVM) packages) [1]. They have created the data into 3 main input configurations. First, with all features except G3 (which is the output). Second, similar to A but without G2 (the second-period grade) and third - similar to B but without G1 (the first-period grade). The DM models are applied to these data to get the best model that generates the highest accuracy.

Performance: The RF model gave the best accuracy with the least regression result and best PCC for A input configuration with 95% confidence intervals.

1(b): Paper No.2: An analysis of students' performance using classification algorithms.

Objective: To perform analysis and evaluation of the students' performance by applying data mining classification algorithms in the Weka tool.

Contribution: The noteworthy contribution of the report is to use data mining procedures to research and determine the students' work. Data mining gives many tasks that could be utilized to review the students' performance. In this report, the distribution task is applied to determine students' performance and deals with the accuracy, confusion matrices and the execution time taken by the various classification data mining algorithms using WEKA tool [2].

Method: This work has used WEKA which is an open source software system that enforces a large selection of machine learning algorithms. The student dataset is uploaded to WEKA Explorer. Then the student dataset is analyzed and organized using decision tree classifier C4.5 (J48), Random Forest, Neural Network (Multilayer Perceptron) and Lazy based classifier (IB1) Rule based classifier (Decision Table) were implemented in WEKA. Under the "Test options", the 10 fold cross validation was adopted [2].

Performance: After obtaining the information gain and gain ratio for the attributes it was found that students' attendance had a considerably high significance in students' performance. This work also found that the Random Forest classifier took the least build time and had the most correctly classified data when compared to other classifiers. And also, with the error rate (Root Mean Square Error and Mean Absolute Error), Random Forest has a smaller error rate related to other classifiers. Therefore, this work establishes that Random Forest is an efficient classification method when matched to other classifiers.

1(c): Paper No.3: Data Mining approach for predicting Student performance.

Objective: To build a prediction model that can derive the conclusion on students' academic success that is very critical for students' higher education institutions.

Contribution: The significant contribution of this paper is that a model was built to predict students' academic success by comparing the data with different methods and techniques of data mining(Bayesian classifier, neural networks (Multi-Layer Perceptron) and decision trees(J48)) [3]. This being the crucial indicators for higher education institutions because the kind of the teaching process is the capability to meet pupils' requirements.

Method: To assess the stability of the classifiers above, cross-validation is done on the classifiers. 3-fold cross-validation was used to separate data set randomly into 3 subsets of equal size. Two subsets were used for training, one subset for cross-validating, and one for estimating the predictive accuracy of the final established structure. This procedure was done 3 times so that each subset was confirmed once. Test results were averaged over 3-fold cross-validation runs. Tests were also conducted to know the importance of input variables (Chi-square test, One R-test, Info Gain test, and Gain Ratio test) [3]. They used the Weka software toolkit to calculate all these performance metrics. Thus, the prediction accuracy of the models was analyzed.

Performance: For the tests conducted on input variables the average value of all the algorithms was taken as the final result of an attribute (input variables) ranking. This analysis aimed to determine the importance of each attribute and was found that the PO (GPA) had the most significance. Considering all the evaluation criteria i.e., Build time, correctly classified data, and errors it was found that Naïve Bayes predicts better than other algorithms.

1(d): Paper No.4: A review on predicting Students' performance using Data Mining Techniques

Objective: The primary purpose of this report was to provide an analysis of the data mining approaches that have been adopted to anticipate students' performance []. This report also concentrates on how the prediction algorithm can be employed to determine the most significant attributes in students' data.

Contribution: The significant contribution of the paper is that they have provided a systematic review to support the objectives of their study, which were:1. To study and identify the gaps in existing prediction methods. 2. To study and identify the variables used in analyzing students' performance. 3. To study the existing prediction methods for predicting students' performance [4]. Hence to find suitable methods for existing parameters to improve the research in the educational data mining field.

Method: To improve better prediction on students' performance this work has focused on research questions like Q1: What are the important attributes used in predicting students' performance? Q2: What are the prediction methods used for students' performance? So, to address these questions a well-planned search strategy was used so that every relevant piece of work can be found in the search results[4]. The search terms used in this systematic review were developed using Kitchenham et al. (2010) [5].

Performance: The search strategies applied to different research works unmasked that the attributes that have been frequently used are cumulative grade point average (CGPA) and internal assessment. CGPA was the best attribute because it has tangible value for future educational and career mobility[]. It was also found that the algorithms used by most of the research works were Decision tree, Artificial Neural Networks, Naive Bayes, K-Nearest Neighbor, and Support Vector Machine. In most of the exploration works Neural Network has the greatest prediction accuracy by (98%) followed by Decision Tree by (91%). It should also be noted that the result of prediction accuracy is depending on the attributes or features that were used during the prediction process and their respective datasets.

Section 2: Implementation

From the survey that we had done, the prediction of grades was done by different classifiers algorithms, but the methodologies used were very outdated, and in our project, we tried to implement new methodologies. In our project, we tried to filter out those attributes that were affecting the proper prediction of the grades and applied various machine learning models to unmask good prediction accuracy from the several models.
For the project, we have used a dataset from Kaggle, which has 2 files of Maths and Portuguese subjects.

Requirements:

• scikit-learn (sklearn) • Pandas • NumPy • matplotlib.pyplot and figure from matplotlib.pyplot • seaborn • mean_squared_error, mean_absolute_error from sklearn.metrics • scipy

Finding the correlation for each column:-

Since the dataset has many columns, we need to initially find the correlation between each column and the G3 column(final grade), which is the value that we have to predict. Thus, the following is the output for the correlation of each of the columns with the other columns for the math data. The darker colors indicate a higher correlation and lighter colors indicate a lower correlation. Thus, we can see that the grades of period 1(G1) and period 2(G2) have a larger correlation with the final grades. Mother and father education have a large correlation index with each other. Alcohol consumption for each day is correlated to each other and going out factor.

Thus, based on this map, we remove the columns with higher correlation values i.e. G1 and G2. From the remaining data, we get the most dominant columns with the least correlation and try to understand their effects on the G3 score. The following graphs display the effects of various attributes on G3.

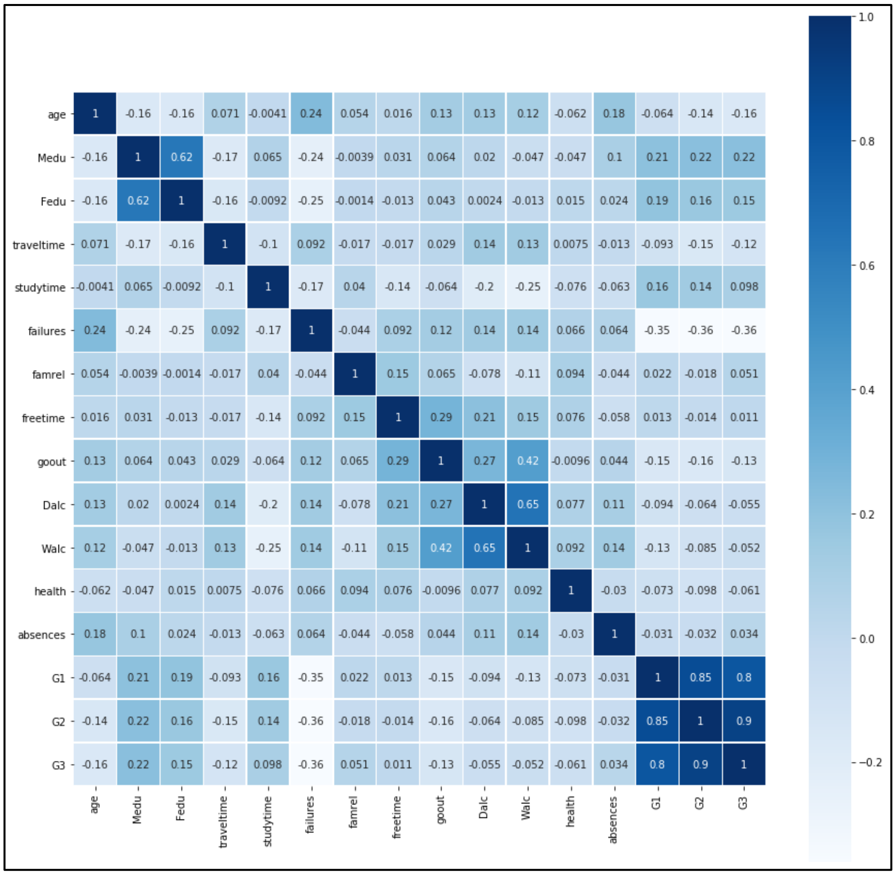
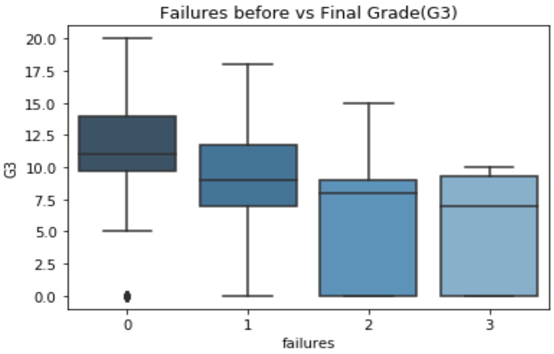
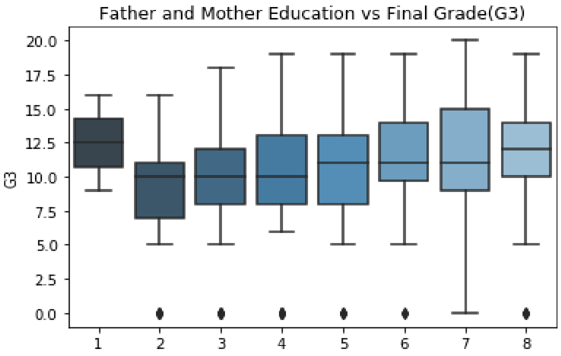


Fig 1: Heat map for Math subject to show correlation.

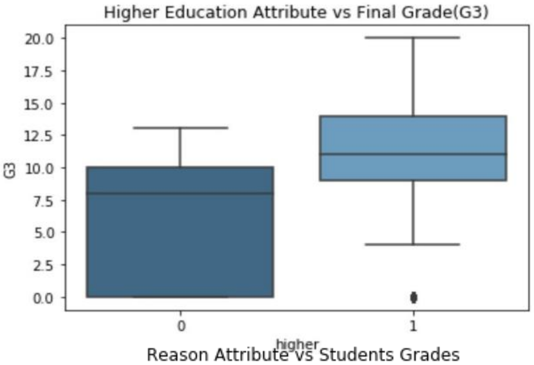


1. We can see that there are many students with the number of failures as 0 and G3 below 10 are very less. The one outlier may also be the reason for a lower median. Thus, students who do not fail usually do good. The students having 1 fail mostly have their G3 between 7.5 and 10. For 2 fails and 3 fails, we see that the number of students with higher than median values is larger than the students with a lower score.

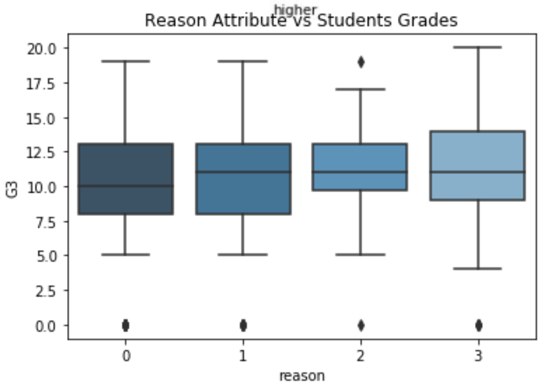


2. For mother and father education, we see that the value of G3 is not much affected as the median for most of them is almost equal.

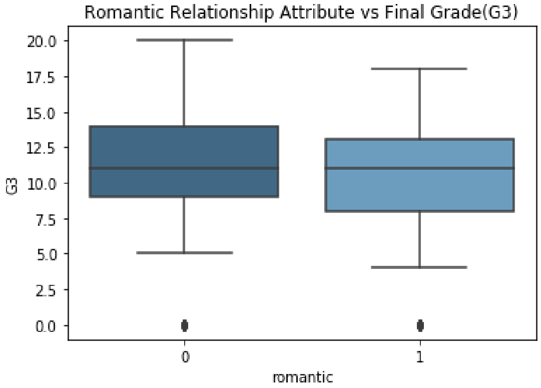
3. The median for G3 is greater in the higher education boxplot for people who wish to go for higher education. This means that a student works better if they wish to pursue higher education.



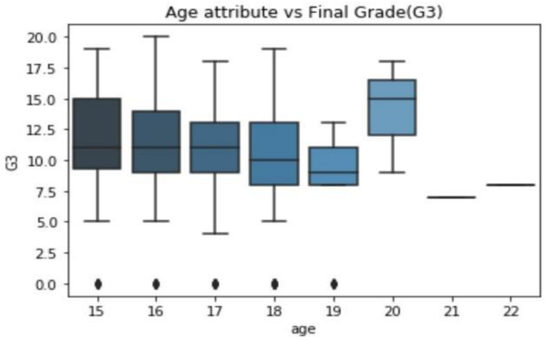
4. The reason for absence also does not affect the G3 medians much.



5. The median for G3 is also not much different for whether a person has romantic relationships or does not. Here, 1 represents a romantic relationship and 0 represents no romantic relationships.

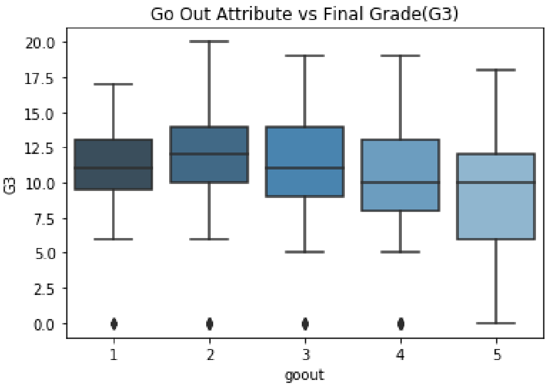


6. The people with the age of 20 and 15 score the highest grades.



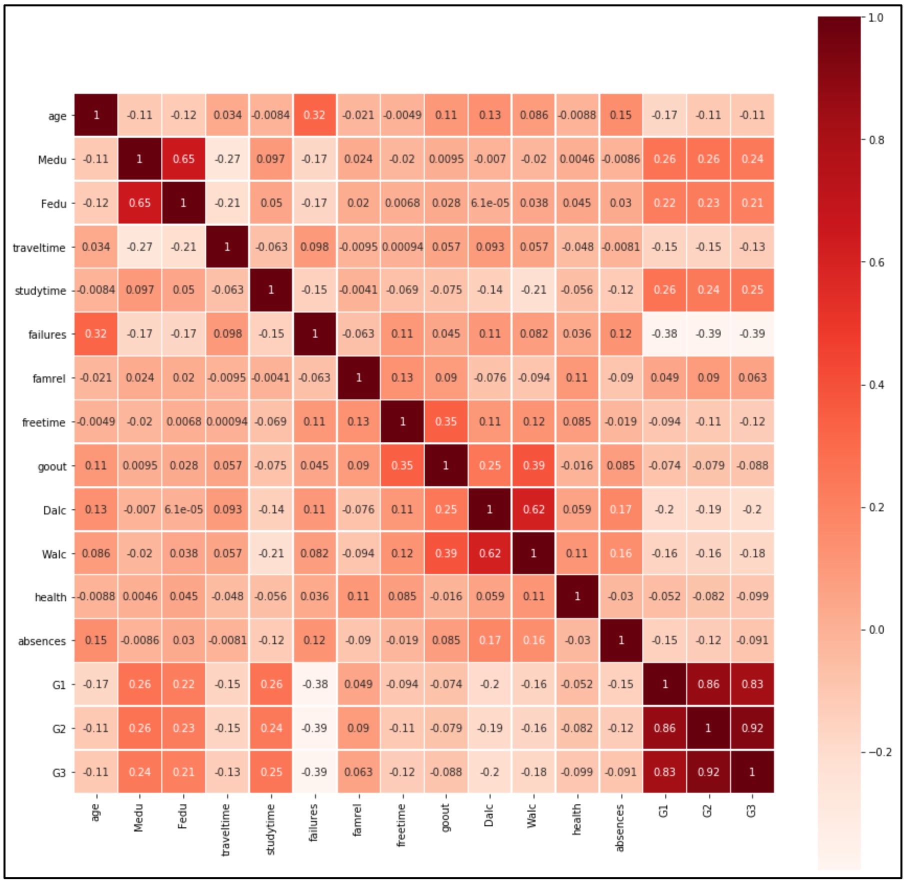
7. The going out attribute shows that the more a student goes out, the fewer marks he/she scores.

.



The following is the output for the correlation of each of the columns with the other columns for the Portuguese data. The darker colors indicate a higher correlation and lighter colors indicate a lower correlation.

Thus, we can see that the grades of period 1 and period 2 have a larger correlation with the final grades. Mother and father education have a large correlation index with each other. Alcohol consumption for each day is correlated to each other and going out factor. But the correlation of these values with the Portuguese data is not as high as in the math data. Also, in Portuguese data, the grades of period 2 have a slightly higher correlation with G3 than the grades of period 1.

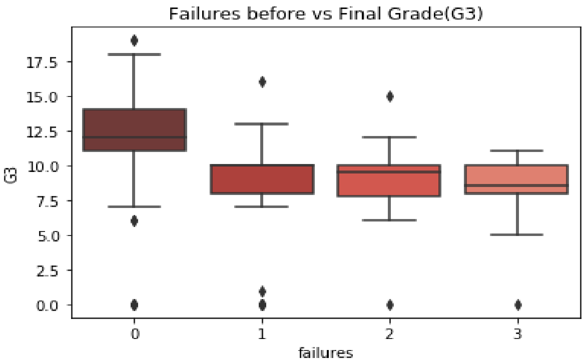


	age	Medu	Fedu	traveltime	studytime	failures	famrel	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3
age	1	-0.11	-0.12	0.034	-0.0084	0.32	-0.021	-0.0049	0.11	0.13	0.086	-0.0088	0.15	-0.17	-0.11	-0.11
Medu	-0.11	1	0.65	-0.27	0.097	-0.17	0.024	-0.02	0.0095	-0.007	-0.02	0.0046	-0.0086	0.26	0.26	0.24
Fedu	-0.12	0.65	1	-0.21	0.05	-0.17	0.02	0.0068	0.028	6.1e-05	0.038	0.045	0.03	0.22	0.23	0.21
traveltime	0.034	-0.27	-0.21	1	-0.063	0.098	-0.0095	0.00094	0.057	0.093	0.057	-0.048	-0.0081	-0.15	-0.15	-0.13
studytime	-0.0084	0.097	0.05	-0.063	1	-0.15	-0.0041	-0.069	-0.075	-0.14	-0.21	-0.056	-0.12	0.26	0.24	0.25
failures	0.32	-0.17	-0.17	0.098	-0.15	1	-0.063	0.11	0.045	0.11	0.082	0.036	0.12	-0.38	-0.39	-0.39
famrel	-0.021	0.024	0.02	-0.0095	-0.0041	-0.063	1	0.13	0.09	-0.076	-0.094	0.11	-0.09	0.049	0.09	0.063
freetime	-0.0049	-0.02	0.0068	0.00094	-0.069	0.11	0.13	1	0.35	0.11	0.12	0.085	-0.019	-0.094	-0.11	-0.12
goout	0.11	0.0095	0.028	0.057	-0.075	0.045	0.09	0.35	1	0.25	0.39	-0.016	0.085	-0.074	-0.079	-0.088
Dalc	0.13	-0.007	6.1e-05	0.093	-0.14	0.11	-0.076	0.11	0.25	1	0.62	0.059	0.17	-0.2	-0.19	-0.2
Walc	0.086	-0.02	0.038	0.057	-0.21	0.082	-0.094	0.12	0.39	0.62	1	0.11	0.16	-0.16	-0.16	-0.18
health	-0.0088	0.0046	0.045	-0.048	-0.056	0.036	0.11	0.085	-0.016	0.059	0.11	1	-0.03	-0.052	-0.082	-0.099
absences	0.15	-0.0086	0.03	-0.0081	-0.12	0.12	-0.09	-0.019	0.085	0.17	0.16	-0.03	1	-0.15	-0.12	-0.091
G1	-0.17	0.26	0.22	-0.15	0.26	-0.38	0.049	-0.094	-0.074	-0.2	-0.16	-0.052	-0.15	1	0.86	0.83
G2	-0.11	0.26	0.23	-0.15	0.24	-0.39	0.09	-0.11	-0.079	-0.19	-0.16	-0.082	-0.12	0.86	1	0.92
G3	-0.11	0.24	0.21	-0.13	0.25	-0.39	0.063	-0.12	-0.088	-0.2	-0.18	-0.099	-0.091	0.83	0.92	1

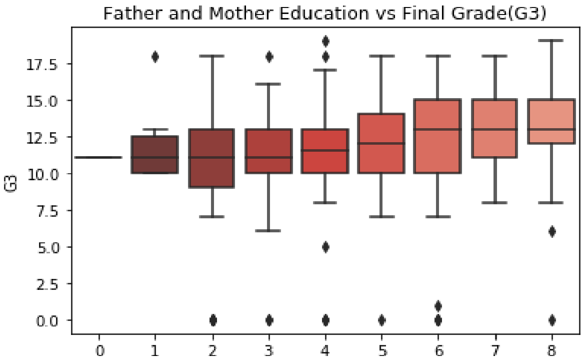
Fig 2: Heat map for Portuguese subject to show correlation

Thus, on the basis of this map, we remove the columns with higher correlation values i.e. G1 and G2. From the remaining data, we get the dominant with the least correlation and try to understand their effects on the G3 score. The following graphs display the effects of various attributes on G3.

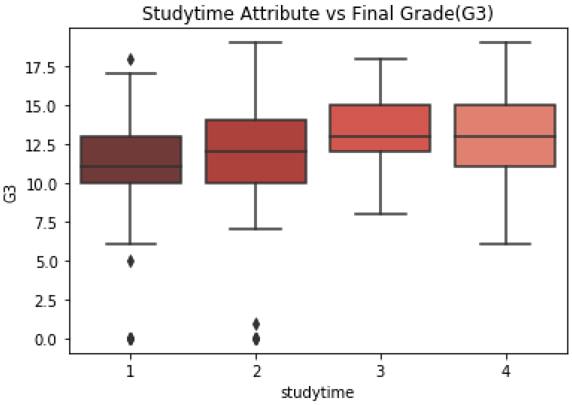
1. We can see that there are many students with the number of failures as 0 and they usually score very high grades than the other categories except a few outliers. The median may be high in the 1 and 2 fails due to the few outliers.



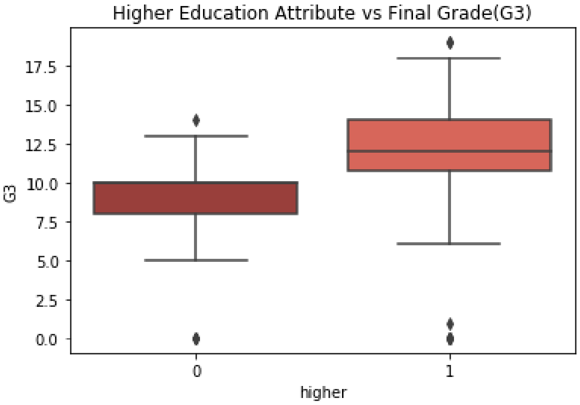
2. For mother and father education, we see that the value of G3 is affected and the median is gradually increasing with the education levels of mother and father.



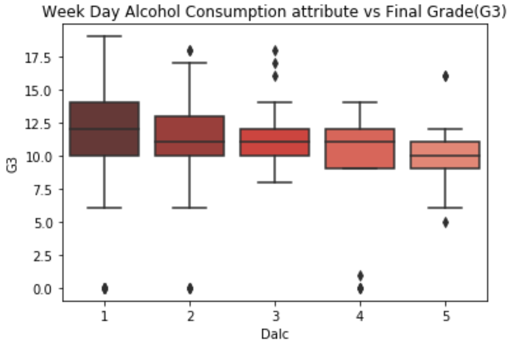
3. The score also increases with the increase in the study times. But it does not affect the grades after 3 hours of studies.



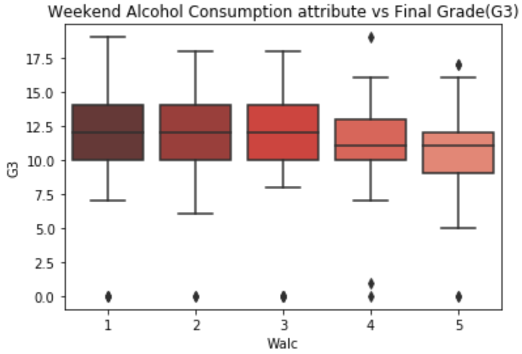
4. The median for G3 is greater in the higher education boxplot for people who wish to go for higher education. This means that a student works better if they wish to pursue higher education.



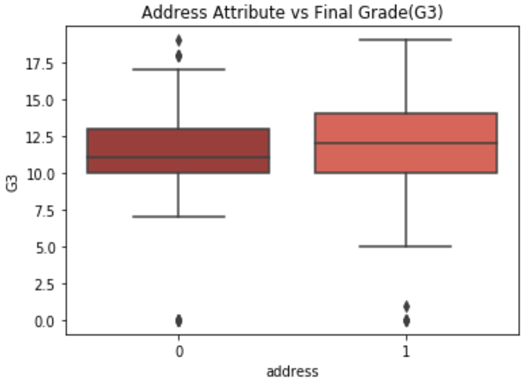
5. As expected, the students who consume less alcohol during a weekday, tend to score higher grades.



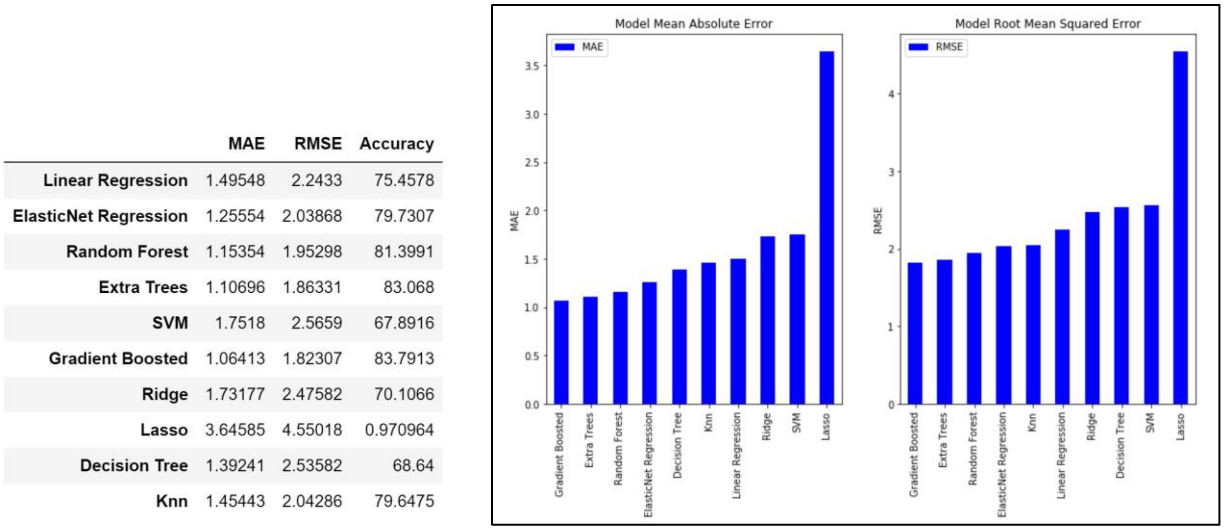
6. The students who consume less alcohol during the weekend, tend to score higher grades.



7. It does not affect the grades much whether the student lives in an urban area or a rural area.

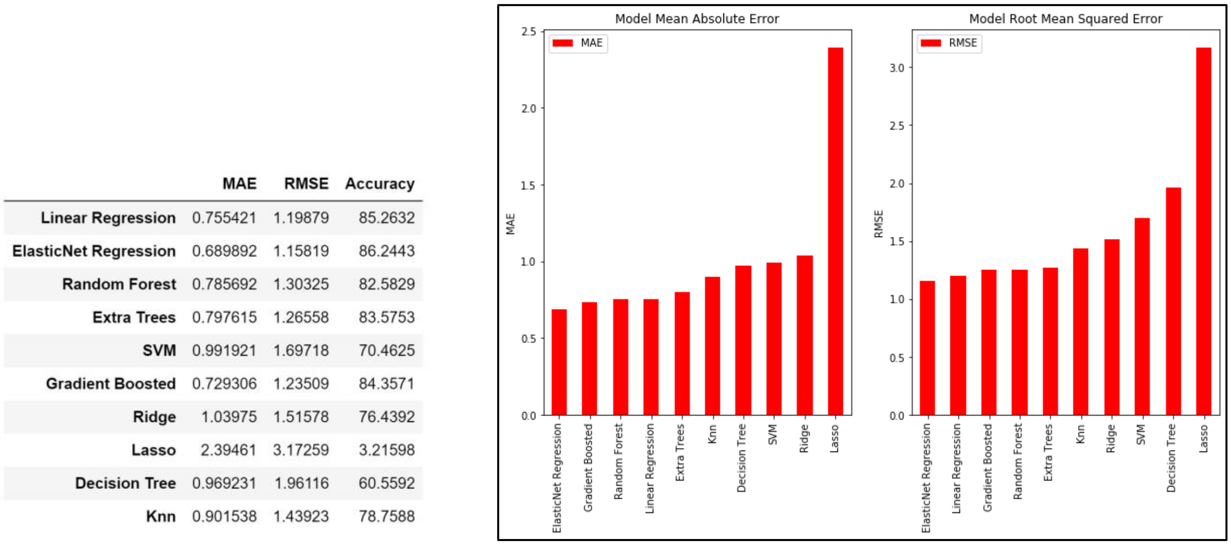


Results



	MAE	RMSE	Accuracy
Linear Regression	1.49548	2.2433	75.4578
ElasticNet Regression	1.25554	2.03868	79.7307
Random Forest	1.15354	1.95298	81.3991
Extra Trees	1.10696	1.86331	83.068
SVM	1.7518	2.5659	67.8916
Gradient Boosted	1.06413	1.82307	83.7913
Ridge	1.73177	2.47582	70.1066
Lasso	3.64585	4.55018	0.970964
Decision Tree	1.39241	2.53582	68.64
Knn	1.45443	2.04286	79.6475

Fig 3: Results for predicting the best accuracy for Math subject using different models



	MAE	RMSE	Accuracy
Linear Regression	0.755421	1.19879	85.2632
ElasticNet Regression	0.689892	1.15819	86.2443
Random Forest	0.785692	1.30325	82.5829
Extra Trees	0.797615	1.26558	83.5753
SVM	0.991921	1.69718	70.4625
Gradient Boosted	0.729306	1.23509	84.3571
Ridge	1.03975	1.51578	76.4392
Lasso	2.39461	3.17259	3.21598
Decision Tree	0.969231	1.96116	60.5592
Knn	0.901538	1.43923	78.7588

Fig 4: Results for predicting the best accuracy for Portuguese subject using different models

As, we know that, that to achieve accuracy we need to feed the model with a bigger amount of data, that is why, we have used our whole dataset, and not just the attributes affecting the final grades attribute, and mapped the dataset to integers using LabelEncoder(). Then we trained and tested the data and split the data into 80-20 pattern 80 for train and 20 for the test.

As seen in the above plots the error value for Math data (blue plot) (both MAE and RMSE) is low for Gradient Boosted when compared to other models which infer that this model has the highest accuracy. And the error value for Port data (red plot) is low for ElasticNet Regression when compared to other models which infer that this model has the highest accuracy, also it could be highly likely to be inferred that MAE and RMSE both works almost equally.

We can also observe that the average error in the math data is more than in the Portuguese data. Thus, for future work we can try to implement a method of which gives less error for math data. We can also try to improve the accuracy for the models by increasing the dataset size for training.

References:

[1] Using Data Mining to Predict Secondary School Student Performance by Paulo Cortez and Alice Silva Dep. Information Systems/Algoritmi R&D Centre University of Minho.

[2] An analysis of students' performance using classification algorithms by Mrs. M.S. Mythili, Dr. A.R.Mohamed Shanavas, Ph.D Research Scholar, Bharathidasan University & Assistant Professor.

[3] Data Mining approach for predicting Student performance by Osmanbegović E., Suljić M.

[4] A review on predicting Students' performance using Data Mining Techniques by Amirah Mohamed Shahiri, WahidahHusain, Nur'aini AbdulRashid.

[5] B. Kitchenham, R. Pretorius, D. Budgen, O. Pearl Brereton, M. Turner, M. Niazi, S. Linkman, Systematic literature reviews in software engineering - a tertiary study, Inf. Softw. Technol. 52 (8) (2010) 792–805. doi:10.1016/j.infsof.2010.03.006.

[6]https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html

[7]https://towardsdatascience.com/how-to-perform-lasso-and-ridge-regression-in-python-3b3b75541ad8

[8] https://www.statisticssolutions.com/what-is-linear-regression/

[9] https://towardsdatascience.com/an-introduction-to-support-vector-regression-svr-a3ebc1672c2

[10]https://www.analyticsvidhya.com/blog/2018/08/k-nearest-neighbor-introduction-regression-python/

[11]https://medium.com/human-in-a-machine-world/mae-and-rmse-which-metric-is-better-e60ac3bde13d

[12]https://towardsdatascience.com/https-medium-com-lorrli-classification-and-regression-analysis-with-decision-trees-c43cdbc58054