STUDENT PERFORMANCE PREDICTION
USING DATA MINING TECHNIQUES

Aswin Sivam Ravikumar
x16134621
National College of Ireland
Masters in data analytics

Abstract— **This electronic document is about the student performance prediction for the Portuguese school under different measures, student performance it remains a hard and unpredictable one here the data is collected under the school reports and their corresponding questionnaires here the objective is to build a different data mining techniques and different algorithms to predict their grades. Some of the most efficient algorithms like a Random forest, linear regression, gradient boost, bagging is used here under the KDD model on par as the comparison is done between them.**

Keywords – Data mining, Random forest, gradient boost, KDD, linear regression, R-Studio, Rapid Miner, Weka

INTRODUCTION

What will be the students' performance in future? The most asked and one popular questions in the world now is why we need to predict the student performance what is the use of that, it may be an interesting question but this question can be answered globally around the world here the real motivation behind the prediction of student performance is to enhance the student learning especially with online assessment we can trace their progress and so on and helps to reduce the dropout rates as well as to identify the weak and average students in an early manner and we can intervene them so I believe that these are some of the potential motivations.**[1]**

Here the objective is to study the existing techniques and tools for the better understanding of concepts of education and to evaluate the student's data to determine their performance in their courses or the semesters and to improve their quality.

Most of the people in the world say that there is a no possibility of predicting the student's performance but it's possible by the data related to the students every step we can help the students to improve their performance that is from the cradle to their career there is quote told by the henry ford that if you don't measure it then you don't prove it. When it comes to the students the fact is the data can be measured in so many ways the individual experts related to the educational field or the sectors they are using to improve their performance so by having the individual student's data it helps to know that what an induvial student needs what the reason for a lag so by improving it they can thrive and succeed.

If we use the data, then it means we are empowered so that every school will have the responsibility to use the data and they need to support every student for their success.

The main goal of the proposed research work is to identify the high influencing predictive variables on their academic performance throughout their different semesters, to find out the best classification and regression algorithms on the student data and finding the relationship between the semesters if the future semester grades related to the past results attained by the students.

LITERATURE REVIEW

Before predicting the student performance, I have done some literature review on the previous researches done by different scholars. Even in the early 60's number of apparent works were done they were using the structured equation model one of the oldest model with the help of the structured equation they are getting the correlations for every pair of the variable, so the equation will tell the correlated variable with each other one of the drawback in this method they don't have much data or the data related platforms that present in the today world.

[2] Chemer and Garcia these authors try to understand the problem, so they interviewed the students the survey taken from the students they met one of the main hypothesis from the previous works which is known as the academic self-efficacy it is the single important variable that could predict the students. the self-efficacy can be defined as the belief in one's capability to organize and to execute courses required to produce given attainments from our case it will be the student performance. The model proposed by them is a little bit complicated. As per the paper, it tells that self-efficacy is related to the optimism suggesting that optimism would influence the self-efficacy to my perspective instead of using the full model we can use the simplified model. There is also an academic expectation where the future performance will influence their academic outcome as suggested in the chemer, hu and Garcia's model.

[3] Rossi and Montgomery's model its one of the study that took place united states of America the author says that here the main hypothesis the student performance depend on the two factors the first one is the quality of home and community environment and the second one quality of the school and the curriculum suggesting that for some student the home provides a better baseline in their studies and school is one of the major influence when comes to the students, these two factors will influence the students perception, incentives and potentially pressures to obtain the good performance many factors like readiness to learn, level and the quality of academic investment and many others factors are also correlated with the student performance.

[4] The next approach is the Schmitt's approach they use the non-cognitive abilities like the personality, motivation, experience and students background such as their interests, hobbies also their type of behaviors here the author suggests that these are some potential factors that could predict the student performance and the author also suggested to conduct the situational judgment test under the academic as well as the social conditions to predict the better part of students with their respected performances.

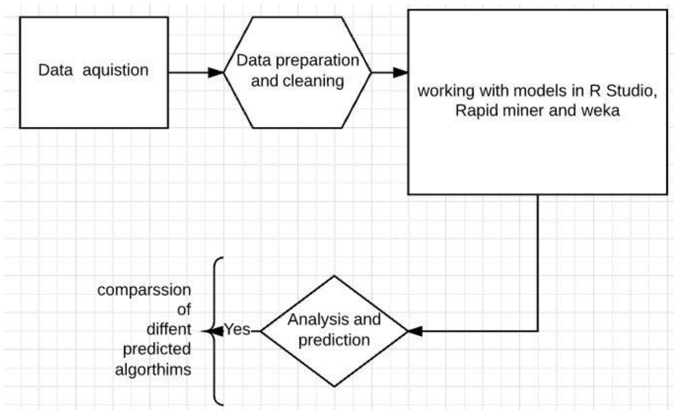
[5] The Lounsbury suggested a new type of approach called the big traits and the narrow traits here the author classifies the big traits into five types namely agreeableness, conscientiousness, emotional stability, extraversion and openness and these terms are well known in the psychological studies and the narrow threats are of four types namely aggression, optimism, tough-mindedness and work drive

SUMMARY

By conducting the literature review with the different authors many models and factors have been suggested by them so I am trying to put all together so there will be lots of factors to measure if there is a possibility measuring the all the factors we could potentially use these variables to predict the student performance the main factor consider on this prediction should be the students past academic performance.

I have implemented some models using different data mining techniques and different and I tried to capture every possible relevant experience.

METHODOLOGY



The above flowchart represents the complete methodology of the project I have followed the KDD.

TOOLS USED

R STUDIO- open source development environment mostly used for statistics and machine learning.

RAPID MINER- data science software or a platform to do the data preparation, machine learning, deep learning, predictive analysis.

WEKA- machine learning software mainly focused on the data mining techniques.

METHOD

KDD

knowledge discovery in databases it has following patterns like data cleaning, which is used to remove unwanted data, data integration, to multiple or to add the data sources, selection of data ,transform and with the help of data mining to get the data patterns for the prediction variables, with the help of the patterns we can derive the knowledge using different techniques and we can represent them.

DATA COLLECTION

I have used the student performance dataset taken from the UCI machine learning repository the data is collected from one of the Portuguese school it describes the student achievement in their mathematics exams on different conditions and this data represents the secondary school students, the **attributes** in the dataset are given below.

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 â€" 5th to 9th grade, 3 â€" secondary education or 4 â€" higher education)
8 Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 â€" 5th to 9th grade, 3 â€" secondary education or 4 â€" 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 93)

these grades are related with the course subject, Math or Portuguese:
31 G1 - first period grade (numeric: from 0 to 20)
31 G2 - second period grade (numeric: from 0 to 20)
32 G3 - final grade (numeric: from 0 to 20, output target)

The above screenshot shows the attribute information the dataset totally consists of 33 attributes 30 general attributes and 3 main attributes G1, G2 and G3 these are the grade that related to the mathematics core subject.

DATA CLEANING AND PREPARATION

The data taken from the machine learning repository have some inconsistencies and the null values before loading the data it should clean to get a more accurate prediction, so I have used the R program to clean the data and certain conversations are made to load the data. The data is the comma separated file, but the **sample data** is shown below.

	A
1	school;sex;age;address;famsize;Pstatus;Medu;Fedu;Mjob;Fjob;reason;guardia
2	GP;"F";18;"U";"GT3";"A";4;4;"at_home";"teacher";"course";"mother";2;2;0;"y
3	GP;"F";17;"U";"GT3";"T";1;1;"at_home";"other";"course";"father";1;2;0;"no";
4	GP;"F";15;"U";"LE3";"T";1;1;"at_home";"other";"other";"mother";1;2;3;"yes";
5	GP;"F";15;"U";"GT3";"T";4;2;"health";"services";"home";"mother";1;3;0;"no";
6	GP;"F";16;"U";"GT3";"T";3;3;"other";"other";"home";"father";1;2;0;"no";"yes'
7	GP;"M";16;"U";"LE3";"T";4;3;"services";"other";"reputation";"mother";1;2;0;"
8	GP;"M";16;"U";"LE3";"T";2;2;"other";"other";"home";"mother";1;2;0;"no";"nc
9	GP;"F";17;"U";"GT3";"A";4;4;"other";"teacher";"home";"mother";2;2;0;"yes";'
10	GP;"M";15;"U";"LE3";"A";3;2;"services";"other";"home";"mother";1;2;0;"no";

To convert this, I have loaded the data into **R** with help of this command

```
Adm<-read.csv("G:/data/student/student-mat.csv",header=T,sep = ";")
names(Adm)
ttr<-Adm[,]
```

```
> Adm<-read.csv("G:/data/student/student-mat.csv",header=T,sep = ";")
> names(Adm)
 [1] "school"     "sex"        "age"        "address"    "famsize"    "Pstatus"    "Medu"       "Fedu"       "Mjob"       "Fjob"
[11] "reason"     "guardian"   "traveltime" "studytime"  "failures"   "schoolsup"  "famsup"     "paid"       "activities" "nursery"
[21] "higher"     "internet"   "romantic"   "famrel"     "freetime"   "goout"      "Dalc"       "Walc"       "health"     "absences"
[31] "G1"         "G2"         "G3"
> ttr<-Adm[,]
```

the tools used here for cleaning the data is R studio, and excel.

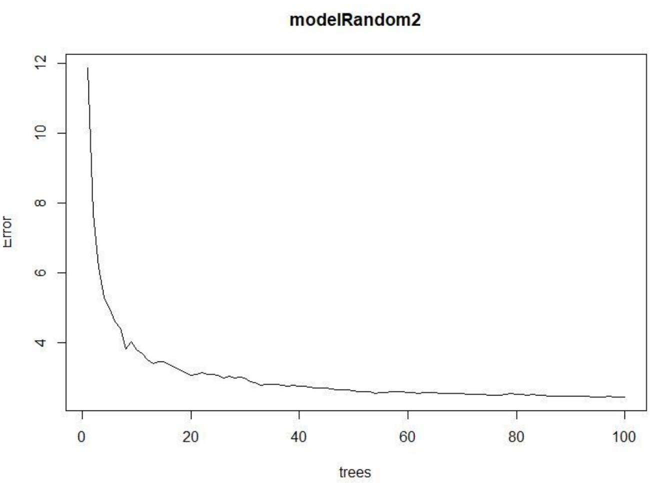
IMPLEMENTATION

R STUDIO

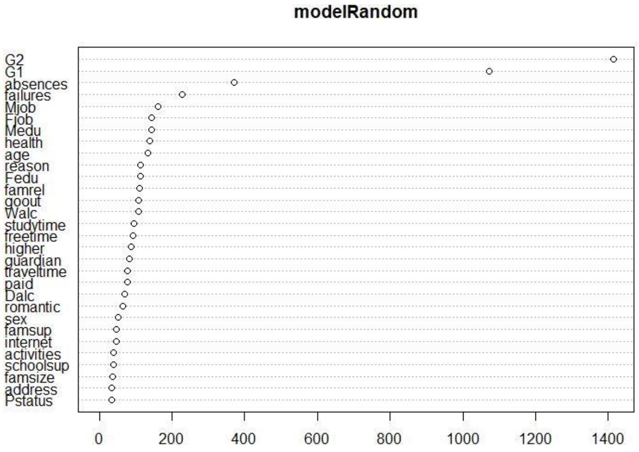
RANDOM FORSET REGRESSION (supervised learning)

The random forest algorithm is one of the most powerful supervised learning algorithm which can perform both the regression and the classification the name itself that suggests that it creates a number of forests within the decision trees. If there are more trees in the forest the prediction will be more robust and more accurate, and the accuracy rate will be high.

After loading the data, I have I found the number of observation and I have portioned the data with help of random seed I can do this analysis reputably the partition data percentage is like 75/25, 75 percent of data is used for the training data and the other remains for testing and I have stored it and seen the observations I have predicted the G3 with the help of all attributes and the number trees used here is 100.



Here I have identified the variable that which gives more contribution in predicting or in other words which attribute has the more influence in predicting the G3 grade.



HIGH VARIABLE CONTRIBUTION

Here we can see that Grades G1, G2, absences, and failures are the most important variables for predicting the G3. By calculating RMSE we can find the error percentage so for G3 it's around 2.17 so the accuracy percentage will be around 85 % to 95%.

```
ttr2<-Adm[,-c(33,32)]
set.seed(124)
smp_size2 <- floor(0.75 * nrow(ttr2))
train_ind2 <- sample(seq_len(nrow(ttr2)), size = smp_size2)

trainAdm2 <- ttr2[train_ind2, ]
testAdm2 <- ttr2[-train_ind2, ]
dim(testAdm2)

modelRandom2<-randomForest(G1~.,data = trainAdm,mtry=3,ntree=200,na.action =na.exclude)
modelRandom2
rest2<-predict(modelRandom,testAdm)
summary(modelRandom2)
sqrt( sum( (rest2 - testAdm2$G1)^2 , na.rm = TRUE ) / nrow(testAdm2) )
```

R CODE RANDOM FORSET

The above table represents some follow conditions the semester three marks that is G3 it is predicted using every possible information's which means variables, but it also can be predicted without the G1 and G2 grades the difference will be the accuracy level changes.

FINDING G1 WITHOUT G2 ANDA G3

The semester one grades G1 are predicted without using the semester two G2 and semester three G3 because there will no need of including them semester one results will not depend on the semester 2 and 3, here the absences and the failures will be the more important predicting variables for G1. Here the RMSE is 4.97 so the accuracy percentage will be around 70%. As same, the G2 can be calculated with and without the G1. The procedure is as same for the G3.

Conditions	Error percentage RMSE	Accuracy percentage
G3 with all variable	1.772	91.14%
G3 without G2, G1	5.577	72.115%
G2 without G3	4.821	75.89%
G1 without G2, G3	4.937	75.31%

The above table represents the following conditions that I have used as I said in the literature review the old academics will helpful in predicting the future ones as we compare G3 without the old academic performance G2, G1 the error percentage is high and accuracy percentage will be less.

LINEAR REGRESSION

Linear regression first it was developed under the statistics but now its also used in machine learning it is a statistical algorithm also the machine learning algorithm it is also known as the linear model it helps to find the relationship between the input variable and the single output variable if there are more input variables it will be known as the multiple linear regression.

variable	dependency
G3 semester 3 marks	G1,G2,other variables
G2 semester 2 marks	G1, other variables not G3
G1 semester 1 marks	Other variables, not G2, G3

```
> rest2<-predict(model,testAdm)
> rest2
        2         4         5        12        13        14        15        17        19        22        23        29        30 
 4.029681 13.129122  8.966265 11.852036 14.147689 10.654263 15.717947 13.297237  4.716717 15.944072 15.655496 11.323364 11.534339 
       34        35        37        39        41        46        53        59        62        70        74        79        84 
 9.420268 14.671345 16.193910 12.333891  9.205111  8.311945 12.704827  9.496811  8.364508 17.480015 12.576483  6.836189 15.683325 
       91        96       100       104       108       111       115       120       121       134       140       158       164 
 5.899906  8.075619  8.929823  6.890226 18.089962 20.308289  8.427340 13.339630 14.871541 11.428327 16.141854  8.029435  9.283331 
      167       169       174       178       183       188       196       197       198       205       222       225       226 
 9.628018  6.248133  4.371380  4.313134 16.982017 15.048664 13.431326 15.996434  8.274739  9.817367  3.179378 12.561021  7.472971 
      228       231       232       233       237       245       248       250       254       258       264       271       283 
11.443355 12.196632 10.249643  9.549660 12.448463 -1.027626  6.380385 15.697746  8.335861 11.112717  7.674405  7.359411 11.478827 
      284       286       289       290       292       294       298       306       308       316       324       330       331 
 8.242698 10.205876 14.013956 13.719496 14.878153 18.650205  7.482986 11.471490  9.011668 11.156816 13.420718 14.152078  7.322638 
      332       338       341       343       346       350       353       357       358       360       362       365       371 
12.945264  6.599612 10.054436 15.035137 12.940872 13.050914  5.714516 12.966371 11.501441 16.580838 12.383914 10.458546  4.353871 
      372       373       377       384       387       388       389       393 
11.853143 10.443984 14.132129  4.474650  4.270142  3.907591  7.969594  7.406391 
```

G3 PREDICTION LINEAR REGRESSION

Here I have done as same as the random forest I have split the data for training and testing as the same percentage between 75 % to training data and 25% data after predicting it I have also calculated its RMSE.

```
model<-lm(G3~.,data = ttr)
model
rest2<-predict(model,testAdm)
rest2
sqrt( sum( (rest2 - testAdm$G3)^2 , na.rm = TRUE ) / nrow(testAdm) )
```

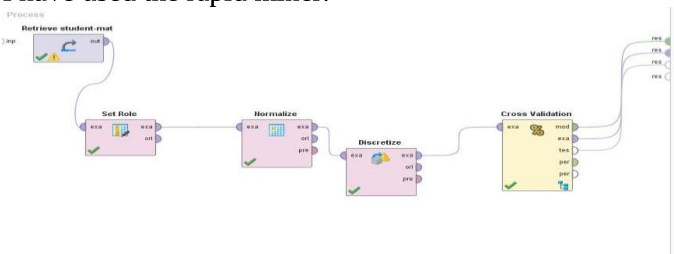
The above picture represents the linear regression code I have used by calculating the RMSE it's about 1.73 so the accuracy percentage will be more.

RAPID MINER

GRADIENT BOOSTED MACHINES(GBM)

The term boosting it ensemble the weak learners it fits consecutive trees where each solves for the net loss of the prior tress the results of the new trees will be applied partially to the entire solution some of the advantages using the gradient boosted is its one of the best model and it’s a robust and helps to directly optimize a cost function. On the other hand, it also has some cons like overfitting, so we need to find the proper stopping cost and the lack of transparency.

Here the following picture shows the flow of that how I have used the rapid miner.



REPRESENTATION OF WORKFLOW

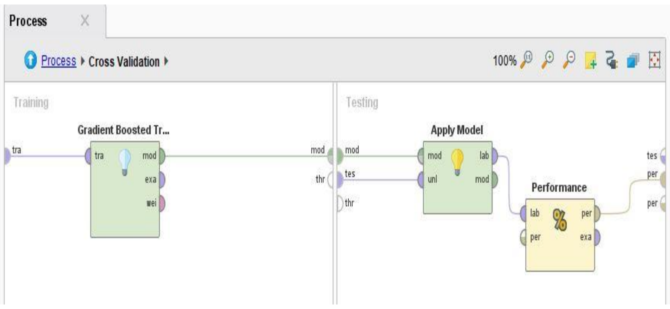
The data is stored in the local repository as the data consists of some null values I have used the function called replace null values here you can in the design pan I have selected the data and connected to the set role, the set role functions helps to label the special attribute the variable which needs to be predicted will change the role as label after changing the role I have connected the data to the normalize function.

NORMALIZE AND DISCRETIZE

Normalize function is nothing but a type of preprocessing technique which helps to keep the variable value in a specific range if every variable or an attribute has the same scale the performance will be good, there will not be a domination of other variables, so I have used the normalize function. The discretize by entropy I have selected this function to change the numerical variable into the nominal variable, only there will not be errors and I have connected to the cross-validation.

CROSS-VALIDATION

To improve the model performance, we are using the cross-validation here the test data process starts the test data is nothing but the segregating part of an original data the below image shows that the process inside the cross-validation



PROCESS OF CROSS-VALIDATION

Here you can see that inside the training model I have selected the gradient boosted trees it’s one of the predictive modelings and at last, the performance regression and the performance classification is given to obtain the different outputs respectively. After the connections are accurate to execute the process and the predicted value will be displayed.

ExampleSet (395 examples, 1 special attribute, 32 regular attributes)

Row No.	G3	age	Medu
1	-0.964	1.022	1.142
2	-0.964	0.238	-1.598
3	-0.091	-1.329	-1.598
4	1.001	-1.329	1.142
5	-0.091	-0.546	0.229
6	1.001	-0.546	1.142
7	0.128	-0.546	-0.685
8	-0.964	0.238	1.142
9	1.874	-1.329	0.229

So, the output is displayed using the performance regression and RMSE is calculated and it is about 0.379 so this algorithm is more efficient.

PERFORMANCE CLASSIFICATION

Performance classification helps to obtain the relationship between the attributes which attributes have more influence in predicting the G3 the below image will show the highly influenced variable.

```
     16              1.00000000          0.066257  0.00000
Variable Importances:
  Variable Relative Importance Scaled Importance Percentage
        G2          401.788727          1.000000   0.847579
        G1           21.620583          0.053811   0.045609
traveltime            9.467589          0.023564   0.019972
      Fjob            7.862660          0.019569   0.016586
  guardian            6.491365          0.016156   0.013694
      Mjob            5.792557          0.014417   0.012219
   nursery            3.147806          0.007834   0.006640
    famrel            2.806885          0.006986   0.005921
      Medu            2.793905          0.006954   0.005894
      Walc            2.467169          0.006140   0.005205
---
      paid            0.129673          0.000323   0.000274
    school            0.123830          0.000308   0.000261
```

So now its clear that G2 and G1 have the more influence in predicting the G3 variable when compared to the other attributes the output of the cross-validation is represented below.

root_mean_squared_error

```
root_mean_squared_error: 0.379 +/- 0.081 (mikro: 0.387 +/- 0.000)
```

RMSE

The gradient boost root mean squared error is 0.379 so it means this model gives us more accuracy in predicting the student performance.

WEKA

BOOTSTRAP AGGREGATING

It is also known as the bagging algorithm these algorithms are the extension of classification and the regression problems. Here the main aim is to create a sample from train data, for the prediction here it uses the trained algorithm and its one of the simplest approach. The cleaned data has been imported into the weka after loading the data its been split into three types test, train data and cross-validation the reason for the data split to get the efficient and more accurate prediction.

The data totally consists of 33 attributes so the one will have predicted we are going to predict the G3 semester three marks with the help of other attributes. The below diagram will show split of data and the attributes were taken.

```
Attributes:   33
              ï»¿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
Test mode:    split 66.0% train, remainder test

=== Classifier model (full training set) ===

Bagging with 10 iterations and base learner

weka.classifiers.trees.REPTree -M 2 -V 0.001 -N 3 -S 1 -L -1 -I 0.0

Time taken to build model: 0.05 seconds
```

TEST TRAIN SPLIT AND ATTRIBUTE INFO

In the above picture, the train data split is about 66% and the remaining data is used as the test data and the time taken to build this model is 0.05 seconds.

The below image shows the prediction values of G3 and the output

```
inst#,actual,predicted,error
1,9,7.255,-1.745
2,13,11.898,-1.102
3,13,12.29,-0.71
4,10,9.775,-0.225
5,8,5.533,-2.467
6,13,12.948,-0.052
7,11,9.924,-1.076
8,13,12.997,-0.003
9,10,11.411,1.411
10,0,3.156,3.156
11,0,5.534,5.534
12,12,11.754,-0.246
13,12,12.309,0.309
14,12,8.821,-3.179
15,11,9.907,-1.093
16,0,0.584,0.584
17,0,0.584,0.584
18,11,11.143,0.143
19,6,7.803,1.803
20,10,9.389,-0.611
21,13,14.272,1.272
22,7,1.944,-5.056
23,19,18.079,-0.921
24,8,6.117,-1.883
25,11,11.484,0.484
26,8,8.734,0.734
27,8,8.6,0.6
```

PREDICTED VALUES

```
=== Evaluation on test split ===

Time taken to test model on test split: 0.03 seco

=== Summary ===

Correlation coefficient                  0.9144
Mean absolute error                      1.0967
Root mean squared error                  1.6663
Relative absolute error                 36.2019 %
Root relative squared error             40.9544 %
Total Number of Instances              134
```

OUTPUT RMSE

The above shows the correlations between the attributes here RMSE is about 1.0967 so the predicted variable will have the accuracy of 94%.

EVALUATION

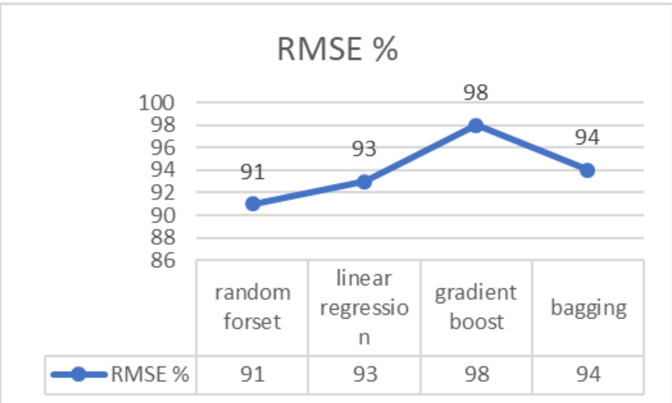
ALGORITHM	RMSE	ACCURACY %
Random forest	**1.772**	**91 %**
Linear regression	**1.733**	**93 %**
Gradient boost	**0.379**	**98 %**
Bagging	**1.0967**	**94 %**

The above table shows the evaluation of the algorithms I have used the formula for the accuracy as

ACCURACY = G3(COUNT RANGE) – RMSE * 5

Here the G3 its maximum value the maximum mark is 20 so I have used it in the formula to calculate the accuracy. So by comparing each algorithm, it is clear for predicting the student performance gradient boost is the best algorithm for predicting the GS grades or the student marks.

COMPARISION OF ALGORITHMS



CONCLUSION AND FUTURE WORK

By conducting the different techniques and the various algorithms to the student performance datasets it's clear that the marks can be predicted using their previous characteristics. The main usage is we can help the students who are weak in their studies or we can collaboratively work with them to outcome their lags. In my perspective, if there more attributes the prediction level and accuracy can be increased by these feedbacks can be given daily to the students by conducting some seminars or the tutorials after every exam or after every assessment because feedbacks also play a major role in the student life to upcome their riddles and make a successful career. So in future, if every school or university follow a systematic way of predicting the student performance it would be helpful for both the students and universities.

REFERENCES

[1] Baker, R.S., and Yacef, K., 2009. The state of educational data mining in 2009: A review and future visions. *JEDM| Journal of Educational Data Mining*, *1*(1), pp.3-17.

[2] Chemers, M.M., Hu, L.T. and Garcia, B.F., 2001. Academic self-efficacy and first-year college student performance and adjustment. *Journal of Educational Psychology*, *93*(1), p.55.

[3] Rossi, R., and Montgomery, A., 1994. Educational reforms and students at risk: A review of the current state of the art. *Washington, DC: US Department of Education, Office of Educational Research and Improvement, Office of Research. Retrieved May*, *25*, p.2007.

[4] Norman, G.T. and Schmidt, H.G., 1992. The psychological basis of problem-based learning: a review of the evidence. *Academic medicine*, *67*(9), pp.557-565.

[5] Lounsbury, J.W., Sundstrom, E., Loveland, J.L. and Gibson, L.W., 2003. Broad versus narrow personality traits in predicting academic performance of adolescents. *Learning and individual differences*, *14*(1), pp.65-75.