Mathematical Statistician and Engineering Applications
ISSN: 2094-0343
2326-9865

Student Grade Prediction Using R Language

Kiran Ingale

Department of Electronic and Telecommunication Engineering

Vishwakarma Institute of Technology

Pune, India

kiran.ingale@vit.edu

Tanushri Bhuruk
Department of Artificial Intelligence and Data Science
Vishwakarma Institute of Technology

Pune, India
tanushri.bhuruk20@vit.edu

Ritesh Pokarne

Department of Artificial Intelligence and Data Science

Vishwakarma Institute of Technology

Pune, India

ritesh.pokarne.20@vit.edu

Nikita Punde
Department of Artificial Intelligence and Data Science
Vishwakarma Institute of Technology

Pune, India
nikita.punde20@vit.edu

Mahalakshmi Phaldesai

Department of Artificial Intelligence and Data Science

Vishwakarma Institute of Technology

Pune, India mahalakshmi.phaldesai.20@vit.edu

Prachi Kumar
Department of Artificial Intelligence and Data Science
Vishwakarma Institute of Technology

Pune, India
Prachi.kumar20@vit.edu

Article Info
Page Number: 6471-6479
Publication Issue:
Vol. 71 No. 4 (2022)

Article History
Article Received: 25 March 2022
Revised: 30 April 2022
Accepted: 15 June 2022
Publication: 19 August 2022

Abstract— In today's world all major companies store their records together in some kind of database or an excel sheet and it is very important to keep this data cleaned and up to date to avoid any confusion and be safe, secure .This is where the concept of data science comes into picture which not only used the data we stored in excel sheets but is used to find patterns within our dataset. Then we use these patterns to make valuable insights and some data derived decisions like predicting the values for the dataset or classifying the data into some parts and then comparing with the known values to check the accuracy of the built model .There are various processes like data extraction, preprocessing, cleaning through which our dataset has to go for getting a final result which is necessary for a better prediction and then our main model code is being implemented in the way we want to work with the data and get the output. This paper deals with the student grade prediction system predicting the grade of a student.

Keywords— data science. predicting , classifying , accuracy, preprocessing , dataset, grade prediction .

I. INTRODUCTION

Nowadays the importance of studies and getting good grades for a child is very important be in any class for their progress and getting admission in higher divisions .This is where the student grade prediction system is used which uses many different factors that are responsible in the final score of the student and build the model according to it .The dataset is of two schools of Germany Gabriel Pereira [GP] and Mousinho da Silveira [MS] together in a random order with records of both male and female .It finds the final grade [G3] of the student and the dataset takes into consideration many different factors that might or might not affect the grade and there are 357 entries of students with which we work in this project and determine the most important factors.

The process used in the project to make the prediction is prediction using regression and not classification as in classification the complications increase drastically having to find out the grade from 1-20 which means for every entry the machine learning model developed has to calculate the 20 different types meaning a 20 dimensional graph which is not possible and even if we divide the dataset into three types taking into consideration the grades like A,B,C like in [1] and so on still the classification doesn't give the accuracy required to build a reliable model. The first method used for predicting is the support vector machine(SVM) regression in which graphically when all factors are plotted, it gives a margin of tolerance equal length from the regression line giving a cushion for the line while predicting and the rest method is done same as the multiple linear regression but with this margin helps to increase the accuracy of the model.

The second method is the decision tree classification which builds the model in a tree like structure to determine each and every condition. This tree like structure is basically the breaking down of the dataset into smaller sets while the regression model will work simultaneously on the dataset to predict the marks. The Last and final method based on the accuracy every model gives is the random forest which is a further upgradation of decision tree where it makes decision tress on different samples rather than one tree and then uses the method of average for in case

of regression like in our model thus predicting the grade of a student using those multiple variables.

II. Related Work

"Multiclass Prediction Model for Student Grade Prediction Using Machine Learning". In this paper we studied about the multiclass prediction model for student grade prediction. It entails a multi-attribute analysis and data sampling from a range of sources in order to predict student grades in various outcomes. However, in education areas, the performance of predictive models for unbalanced datasets is still infrequently studied. Accuracy performance of six well-known machine learning techniques namely Decision Tree (J48), Support Vector Machine (SVM), Naïve Bayes (NB), K-Nearest Neighbor (in), Logistic Regression (LR) and Random Forest (RF) was down using the students course grade dataset. In this paper a multiclass prediction model was proposed in order to minimize the overfitting and misclassifications caused duo to imbalance multi classification which is based upon SMOTE or known as the Synthetic Minority Oversampling Technique with two feature selection methods. The highest accuracy was achieved by using Random forest method i.e. 99.5%.1

"Machine Learning Based Student Grade Prediction: A Case Study".In this paper we studied Collaborative Filtering (CF), Matrix Factorization (MF), and Restricted Boltzmann Machines (RBM) methods to analyze the student data. After calculating the performance of all the mentioned techniques, RBM i.e. Restricted Boltzmann Machines was found to be better than the other techniques.2

"A Review on Predicting Student's Performance using Data Mining Techniques". This paper's main goal is to give an overview of the data mining approaches that have been utilized to forecast student performance. The prediction method is also utilized in this research to find the most relevant features in a student's data. Using educational data mining approaches, we could truly enhance student attainment and success more effectively and efficiently. It has the potential to benefit students, instructors, and academic institutions.3

"Linear Regression Analysis Using R for Research and Development". In this paper we studied about linear regression and how to predict the outcome using historical data and how to find a linear relationship. We also studied how to implement linear regression using R which is a statical computing language.4

III. Methodology

A. Discription of Dataset

Table 1 Table of all variable names with types

Variable name	Type
1) school	Categorical
2) sex	Categorical
3) age	Numerical
4) address	Categorical

5) famsize	Categorical
6) Pstatus	Categorical
7) Medu	Numerical
8) Fedu	Numerical
9) Mjob	Categorical
10) Fjob	Categorical
11) reason	Categorical
12) guardian	Categorical
13) traveltime	Numerical
14) studytime	Numerical
15) failiures	Numerical
16) schoolsup	Categorical
17) famsup	Categorical
18) paid	Categorical
19) activities	Categorical
20) nursery	Categorical
21) higher	Categorical
22) internet	Categorical
23) romantic	Categorical
24) famrel	Numerical
25) freetime	Numerical
26) goout	Numerical
27) Dalc	Numerical
28) Walc	Numerical
29) Health	Numerical
30) Abscences	Numerical
31) G3	Numerical

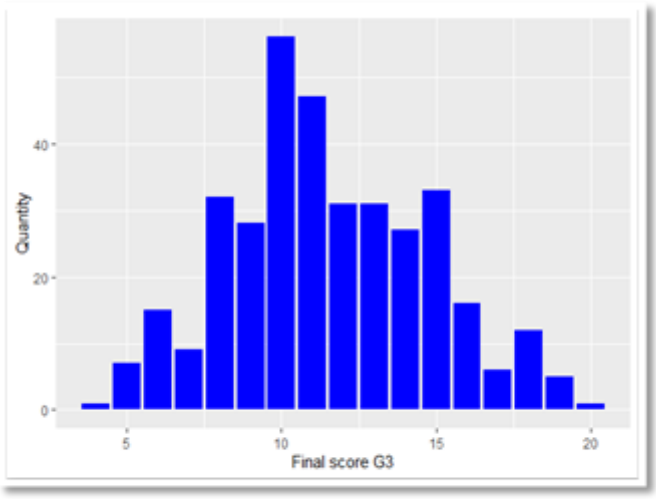

Figure 1 graph of final grades of students in dataset

The dataset is a record of the research done on 2 schools of Portugal Gabriel Pereira [GP] and Mousinho da Silveira [MS] which has the records of 357 students .There are 30 independent variables on which the whole machine learning takes place. Students are taken both male and female and the age group is 15-19 years of age because they are in college at that time. The next data is address telling whether the person is from rural or urban area . famsize which tells whether the size of the family is greater than or less than 3. Pstatus or parent status telling whether the parents are together or apart. Medu and Fedu ranking the education of both mother and father with 0-none,1-primary(4 standard),2- 5-9 standard, 3-secondary or 4 – higher . Mjob and Fjob giving information about the education of the parent's teacher ,health services, civil services, at home or other. Reason to mention the reason of choosing the particular school. Guardian to tell between mother and father who is their guardian. Traveltime to show how much time the child takes to go to school from their home.

Studytime data shows the time for which the child studies divided into categories like less than 2 hours,2-5 hours,5-10 hours ,and greater than 10 hours. Failures to give the data how many times the person has failed. Activities for recording any extra co-curricular activities done or not. Nursery for attended nursery school or not. Higher to record if the child wants to take higher studies or not. Internet data gives information if the student uses internet at home . romantic to tell if the person is in a relationship or not. Famrel to tell the relationship between the family from a counter of 1 to 5. freetime to denote the freetime student gets after the school . goout to record the time student goes out with the friends .Dalc to tell the weekdays alcohol consumption and Walc to tell the weekends alcohol consumption both from a counter of 1 to 5. Health to tell the current health status from 1 very bad to 5 excellent . absences denotes the number of school absences from 0 to 93.Then in the end 3 target variable to work with namely G1,G2,G3 being the grades for each term from 0-20.

B. Data Cleaning

So after importing the dataset in our model, we understood that the value of G3 is dependent on G1 and G2 also so it will be better if we use only G3 as our target variable out of the three then we will have a better result and only one final grade predicted rather than 3 reducing the confusion around all 3 variables. As shown in **Error! Reference source not found.**.

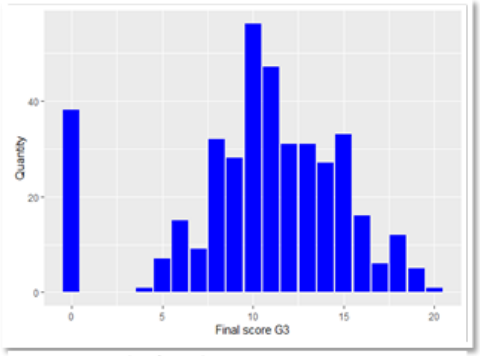

Figure 1 Graph after cleaning Data

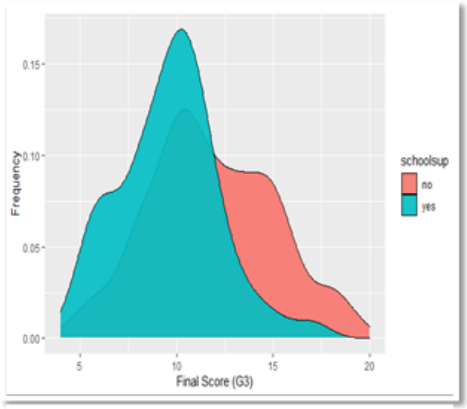

Figure 2 graph of final score with schoolsup

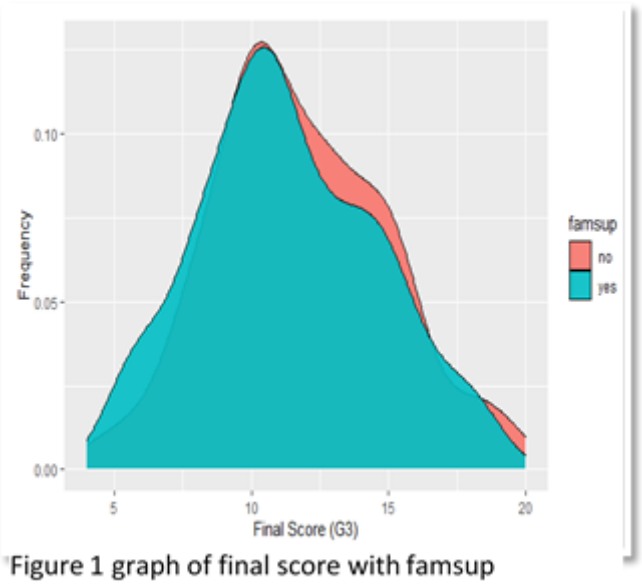

Figure 1 graph of final score with famsup

Then the independent variable G3 which is our final output had some student's data showing their final grades as 0 which could be not attending the exams due to some unavoidable reasons so we removed those values using the code filter and then giving the exact data frames who values we wanted to remove which further cleaned our data and is also shown with the help of a graph. As shown in **Error! Reference source not found.**.

C. Encoding the categorical variable

As there were many categorical variables in our dataset it was not possible to directly include them into the regression procedure as there were many chances of the accuracy being decreased. So all the categorical variables were encoded using the as.factor code which converts the variables internally making the r studio know that the variables are known given some numeric value and the process of regression analysis can be used on them now. All the categorical variables visible from the dataset like school , address , famsize and so on were converted using this method itself.

D. Plotting the graphs for better visualization

We used two major graphs for comparison first one being the schoolsup variable denoting the school support for the student with the variable g3 of the dataset giving a density graph (**Error! Reference source not found.**). The next one is famsup denoting the family support explaining the support of the family for the student and the density graph with the final grade G3 of the students(**Error! Reference source not found.**).

E. Splitting the dataset into training and test set

This is the final process before starting to build the exact regression model in which we divide the dataset into a training set and the test set done randomly by the r studio using an external library known as catools which enables the splitting feature. The syntax for writing is very simple .split and then the data frame is being specified and the ratio in which we want to divide

our data is to be mentioned there. Normally for optimum results all the split ratios are 80:20 or 75:25.

So we have used the first one at the moment and will change if the accuracy changes drastically. Now the dataset has been divided into 2 new subsets trainset and test set for making the regression model for the student grade prediction system . After checking the relations between all the variables we decided to use these variables as the independent ones rather than using all of them to increase the accuracy of the model namely studytime, failiures ,schoolsup, famsup ,gout and absences of the student.

D. Regression Model

So , the first algorithm used for making the model is support vector machine(SVM) Regression model which is an upgradation of the linear regression model where in linear regression the graph has a straight line ,in SVR it uses the hyperplane as its line So, the svr tries to fit in the best hyperplane on a graph with a boundary line used to create a maximized boundary and all the points within this range are used for predicting the values. In the r studio the package we had to install for making the svm model work was e1071 .

Then the regressor was built with variable l2 and the code was svm() where in the parenthesis the target variable was entered followed by the independent variables then entering the training set to give the regressor all variables to work with and then in the end the type of svm nu regression .Then in pred2 the grades were stored with syntax pred() and the test set with the regressor were entered. The last thing accuracy was calculated by importing a package forecast which calculates the accuracy with the same keyword and entering the variables predicted and the one with which it has to be checked. The accuracy is then printed on the console.

Next is the decision tree algorithm in which it acts like a tree structure where every condition is in a node continuing the tree. So in a dataset the model checks for every variable and makes a separate branch for every answer it gives and then when next variable is checked it continues the same process until all conditions are satisfied. In the code we again have installed the package rpart which enables to use the decision tree algorithm. A regressor is defined with syntax rpart() where the target variable is mentioned and followed by all independent variables and then training set. Then the regressor will check all the conditions and the prediction will be made with syntax predict() with entering the regressor and the test set in the regressor. Then accuracy is checked in the same way using library forecast giving the means absolute percentage error.

The last algorithm used is random forest algorithm that is a better version of the decision tree in which it takes the average or mean of the outputs of various trees and is used to increase the accuracy of the regression model. For our code the method is same as of the decision tree where we have to declare the regressor variable ,then specify the variable which is to be predicted followed by the dependent variables with which the regression will happen then declaring the training set .After that the values are predicted using predict and entering the test set with the regressor .In the end accuracy is again checked using forecast library.

IV. RESULTS

Table 2 all the 3-regression model MAPE

Type of Error	SVR	Decision Tree	Random Forest
ME	0.0923037	-0.04799539	-0.1769306
RMSE	2.744317	2.749662	2.778834
MAE	2.245656	2.137061	2.273419
MPE	-5.648984	-6.722818	-8.101714
MAPE	21.53653	20.68576	21.98222

V. CONCLUSION

So a student grade prediction system has been designed by predicting the grade G3 of students including most important independent variables for making the regressors and using three regression methods namely SVM(Support vector Machine)regression which gives an accuracy of 79 % then decision tree regression giving an prediction accuracy of 80% and random forest regression which gives the prediction accuracy of 79 %. These accuracies are 100-MAPE (Mean Absolute Percentage Error) from all of the regression methods.

VI. FUTURE SCOPE

Now, there are many prospects in which the project can be made more advanced and efficient. We can try to segregate which variables are the most important for our regression which would not only decrease our burden to work with so many variables but also make our model more efficient.

We can try using classification algorithms in our model because categorical variables are best suited for classification itself based on their properties and then check which among the regression or classification gives us the best results so that we can build a stronger model and whoever would ever want to use the mode with both types of algorithm will get the best result of them.

REFERENCES

1. S. D. A. Bujang et al., "Multiclass Prediction Model for Student Grade Prediction Using Machine Learning," in IEEE Access, vol. 9, pp. 95608-95621, 2021, doi: 10.1109/ACCESS.2021.3093563.
2. Zafar Iqbal, Junaid Qadir, Adnan Noor Mian, and Faisal Kamiran, “Machine Learning Based Student Grade Prediction: A Case Study”.

3. Amirah Mohamed Shahiri, Wahidah Husain, Nur'aini Abdul Rashid, "A Review on Predicting Student's Performance using Data Mining Techniques".
4. Anjali Pant, R S Rajput, "Linear Regression Analysis Using R for Research and Development".
5. E. Alyahyan and D. Düştegör, ''Predicting academic success in higher education: Literature review and best practices,'' Int. J. Educ. Technol. Higher Educ., vol. 17, no. 1, Dec. 2020
6. V. L. Miguéis, A. Freitas, P. J. V. Garcia, and A. Silva, ''Early segmentation of students according to their academic performance: A predictive modelling approach,'' Decis. Support Syst., vol. 115, pp. 36–51, Nov. 2018.
7. K. L.-M. Ang, F. L. Ge, and K. P. Seng, ''Big educational data & analytics: Survey, architecture and challenges,'' IEEE Access, vol. 8, pp. 116392–116414, 2020.
8. X. Zhang, R. Xue, B. Liu, W. Lu, and Y. Zhang, ''Grade prediction of student academic performance with multiple classification models,'' in Proc. 14th Int. Conf. Natural Comput., Fuzzy Syst. Knowl. Discovery (ICNC-FSKD), Jul. 2018, pp. 1086–1090.