A Multiple Linear Regression-Based Approach to Predict Student Performance

Ouafae El Aissaoui[1(✉)], Yasser El Alami El Madani[2], Lahcen Oughdir[1], Ahmed Dakkak[1], and Youssouf El Allioui[3]

[1] LSI, FPT, University of Sidi Mohammed Ben Abdellah, Taza, Morocco
ouafae.elaissaouil@usmba.ac.ma
[2] ENSIAS, Mohammed V University, B.P.: 713, Agdal, Rabat, Morocco
[3] LS3M, FPK, USMS University, B.P.: 145, 25000 Khouribga, Morocco
yelallioui@gmail.com

Abstract. Predicting students' academic outcome is useful for any educational institution that aims to ameliorate students' performance. Based on the resulted predictions, educators can provide support to students at risk of failure. Data mining and machine learning techniques were widely used to predict students' performance. This process called Educational data mining. In this work, we have proposed a methodology to build a student' performance prediction model using a supervised machine learning technique which is the multiple linear regression (MLR). Our methodology consists of three major steps, the first step aims to analyze and preprocess the students' attributes/variables using a set of statistical analysis methods, and then the second step consists in selecting the most important variables using different methods. The third step aims to construct different MLR models based on the selected variables and compare their performance using the k-fold cross-validation technique. The obtained results show that the model built using the variables selected from the Multivariate Adaptive Regression Splines method (MARS), outperforms the other constructed models.

Keywords: Supervised machine learning technique · Educational data mining · Prediction · Students' performance · Multiple linear regression · K-fold cross-validation technique

1 Introduction

The quality of teaching in educational institutions is considered as one of the development keys of any country, thus, it is essential for educational institutions to come up with strategies that enhance their performance system. Those strategies can be planned after analyzing the students' performance, since the advanced estimation of failure rate can help educational institutions to take preventive decisions to decrease this rate.

One way to predict students' performance is by applying data mining techniques on data that comes from educational databases, this process called educational data mining [1]. Educational data mining field aims to apply machine learning algorithms and data mining techniques on data that can be extracted from educational devices in order to analyze the learners' behaviors and improve the learning process.

M. Ezziyyani (Ed.): AI2SD 2019, AISC 1102, pp. 9–23, 2020.
https://doi.org/10.1007/978-3-030-36653-7_2

The data attributes used to predict students' performance can include many features, such as student grades in some materials which were studied previously, demographic information such as sex, age and address, and social information such as parents cohabitation status, mother's and father's job, family size and so on.

One of the methods that was wildly used by researchers to predict students' performance is regression. Regression is a supervised machine learning technique that shares the same concept as classification in using a training dataset to make a prediction. The difference between them is that the output variable in classification is categorical while in regression is numerical.

In regression analyses, a crucial problem that can be emerged is the presence of variables that don't contribute significantly to explain the dependent variable and build an effective prediction model. Thus, when building a regression model it's essential to determine the most important variables [2].

In this work, we are interested to investigate various methods for quantifying variables' importance, those methods are already implemented with the R language. The aim of our work is to select among those methods which one can determine the most important variables that contribute in building a Student's Performance Prediction model. The dataset used in this study is a Student Performance Dataset that is extracted from the University of California Irvine (UCI) Machine Learning Repository [3].

The rest of this paper is organized as follows. Section 2 gives a literature review of related works. Section 3 describes our methodology. The experiments and results are presented in Sect. 4; Finally, Sect. 5 presents our conclusions and future works.

2 Related Works

Student modeling is one of the educational data mining field objectives. In student modeling, we distinguish between two mains tasks which are prediction and structure discovery. Also in the prediction task, we differentiate between two applications, predicting the undesirable students' behaviors, and predicting the students' characteristics such as learning styles and performance [1, 4–7].

Various data mining techniques have been applied to build students' performance prediction model, such as classification and regression [8]. Classification technique can be applied when the outcome variables are categorical (or discrete), while the regression technique is applied when the outcome variables are numerical (or continuous). Classification is the most commonly applied data mining technique in higher education [9]. Many algorithms under classification technique can be used to predict students' performance, among those algorithms there are, Naïve Bayes, K-Nearest Neighbor, and Decision Tree.

Decision tree technique is widely used by researchers to predict students' performance, for example, authors in [10] applied decision tree technique to predict the drop out feature of students. Authors in [11] used different implementations of decision tree technique to build performance prediction model based on students' social integration, academic integration, and various emotional skills which have not been considered so far. In [12] Authors applied a decision tree algorithm to predict a suitable career for a student based on their behavioral patterns.

Naïve Bayes algorithm is also widely used by researchers to make predictions. In [13] authors used different classification methods including the naïve Bayes to create different classification models to predict student performance, using data collected from an Australian university. Authors in [14] also applied three supervised data mining algorithms on the preoperative assessment data to predict success in a course (either passed or failed), they found that the Naïve Bayes classifier outperforms the decision tree and neural network methods. The same, Authors in [15] presented a technique to improve the accuracy of the students' final grade prediction model for a particular course. After comparing the built models they found that two of the models which have the highest accuracy of about 75% are Neural Network and Naive Bayes.

K-Nearest Neighbor is another technique that was used by researchers to predict students' performance. Authors in [16] compared the efficiency of some classification techniques in predicting students' performance. Experimentation results showed that Multi-labeled K-Nearest Neighbor had taken less time in classification when compared to other techniques.

Against the great number of researchers who used classification techniques to make predictions, there are some researchers who preferred to use Linear Regression. Authors in [17] presented a comparison study between Artificial Neural Network (ANN) and Linear Regression (LR) in predicting the academic performance. They found that both prediction models indicated similar results as far as Mean square Error is concerned. In [18], authors combined a multiple linear regression and principal component analysis to establish a more accurate prediction model. Authors in [19] aimed to provide the prediction of students' performance in final examination by applying linear regression and multilayer perceptron, the result showed that multilayer perceptron provides better prediction results of final examination than linear regression.

Even if many works have been proposed to predict students' academic outcome, they tend to use all students' attributes to generate regression models. In this study, we have built a multiple linear regression model that takes into account only the most important variables.

3 Methodology

3.1 Problem Description

Prediction is the task of estimating the value of an unknown output variable based on the values of input variables. In education the output variables that refer to students' performance can be in the form of marks, numeric values (regression task) or decisions, categorical values (classification task).

In linear regression task, we distinguish between two types of methods, a simple linear regression method that aims to find the relationship between a dependent variable and one independent variable, and a multiple linear regression method that aims to find the relationship between a dependent variable and many independent variables.

Multiple linear regression aims to model the relationship between two or more independent variables and a dependent variable by fitting a linear equation to the observed data. The theoretical assumption behind the MLR is that every unit change in

the independent variable causes an uniform change in the dependent variable. A Multiple linear regression model can be written as [20]:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

where:

Y is the output variable, also called response and dependent variable.
X_i for $i = 1, 2 \ldots P$, are the dependent variables, also called regressors and predictors.
β_0 is the value of Y when each X_i equals to 0. Also called an intercept.
β_j for $i = 1, 2 \ldots P$ are the regression coefficients, β_j is the change in Y based in a unit change in X_i.
ε is a random error term that represents the difference in the linear model and a particular observed value for y.

In order to build an efficient Student's Performance Prediction model, we have to select the most important attributes that will be used in building the model, especially when we have a dataset that contains a huge number of attributes, we have to select among them which ones are more significant [2, 17, 18, 21].

The dataset used in our work was extracted from The UCI Machine Learning Repository [1, 3]. This data contains students' math achievement in secondary education of two Portuguese schools. The dataset includes 32 attributes and 395 records. The following table describes the dataset attributes (Table 1):

Table 1. Dataset attributes

Attribute	Description/values	Attribute	Description/values
School	student's school (binary: 'GP'—Gabriel Pereira or 'MS'—Mousinho da Silveira)	Famsup	Family educational support (binary: yes or no)
Sex	student's sex (binary: 'F'—female or 'M'—male)	Paid	extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
Age	student's age (numeric: from 15 to 22)	Activities	extra-curricular activities (binary: yes or no)
Address	student's home address type (binary: 'U'—urban or 'R'—rural)	Nursery	attended nursery school (binary: yes or no
Famsize	family size (binary: 'LE3'—less or equal to 3 or 'GT3'—greater than 3)	Higher	wants to take higher education (binary: yes or no)
Pstatus	parent's cohabitation status (binary: 'T'—living together or 'A'—apart)	Internet	Internet access at home (binary: yes or no)
Medu	mother's education (numeric: 0—none, 1—primary education (4th grade), 2–5th to 9th grade, 3—secondary education or 4—higher education)	Romantic	with a romantic relationship (binary: yes or no)

(*continued*)

Table 1. (*continued*)

Attribute	Description/values	Attribute	Description/values
Fedu	father's education (numeric: 0—none, 1—primary education (4th grade), 2–5th to 9th grade, 3—secondary education or 4—higher education)	Famrel	quality of family relationships (numeric: from 1—very bad to 5—excellent)
Mjob	mother's job (nominal: 'teacher', 'health' care related, civil 'services', 'at_home' or 'other')	Freetime	free time after school (numeric: from 1—very low to 5—very high)
Fjob	father's job (nominal: 'teacher', 'health' care related, civil 'services', 'at_home' or 'other')	Goout	going out with friends (numeric: from 1—very low to 5—very high)
Reason	reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')	Dalc	workday alcohol consumption (numeric: from 1—very low to 5—very high)
Guardian	student's guardian (nominal: 'mother', 'father' or 'other')	Walc	weekend alcohol consumption (numeric: from 1—very low to 5—very high)
Traveltime	home to school travel time (numeric: 1—<15 min., 2—15–30 min., 3–30 min. to 1 h, or 4—>1 h)	Health	current health status (numeric: from 1—very bad to 5—very good)
Studytime	weekly study time (numeric: 1—<2 h, 2—2–5 h, 3—5–10 h, or 4—>10 h)	Absences	number of school absences (numeric: from 0 to 93)
Failures	number of past class failures (numeric: n if 1 <= n < 3, else 4)	G3	final grade (numeric: from 0 to 20, output target)
Schoolsup	extra educational support (binary: yes or no)		

The aim of this work is to select among all the previous described attributes the most important ones to build a multiple regression linear model that enables the prediction of the student's final grade, to do that, we have used a set of methods that have been already implemented with The R language. The R language and environment for statistical computing contains many packages and methods that can be used to find the most important variables that contribute more significantly to explain the dependent variable.

The following schema resumes our methodology (Fig. 1):

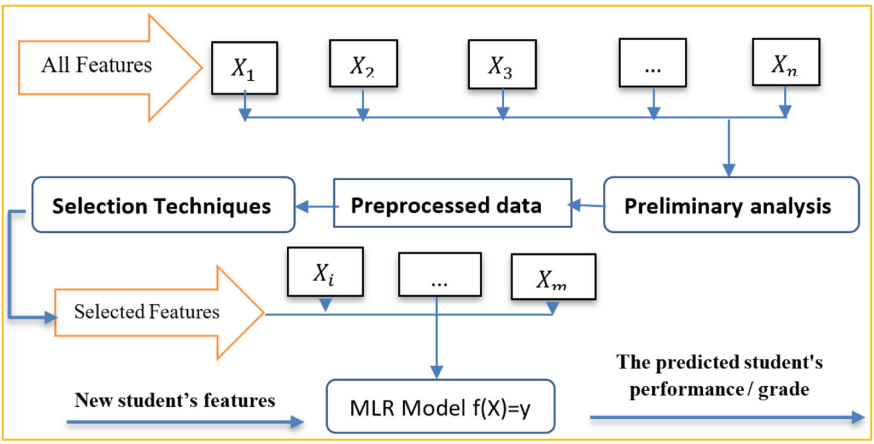

Fig. 1. Our methodology

In this work we will compare the performance of some methods in selecting the most important variables, the selected variables will be used to create models, and the model with the heist performance will be chosen as the best one. The following subsections describe the different steps that we have done in our analysis.

3.2 Preliminary Analysis

Before moving on to determine the most important variables, a preliminary analysis has to be done in order to preprocess the data and check the assumptions of the Multiple Linear Regression method.

Exploratory analysis

First of all, an exploratory analysis has to be done in order to check if the relationship between the independent and dependent variables is linear. This linearity assumption can be tested with a scatter plot. We have also to check if the variables are normally distributed using a density plot and a box plot. If a variable is highly skewed, we have to determine if there are any outliers to analyze them.

In this step we have run an R code that enables the creation of the density plot, scatter plot and box plot, all within one frame, for all the variables in input data. For example, the below picture displays the three plots for the variable "absences" (Fig. 2).

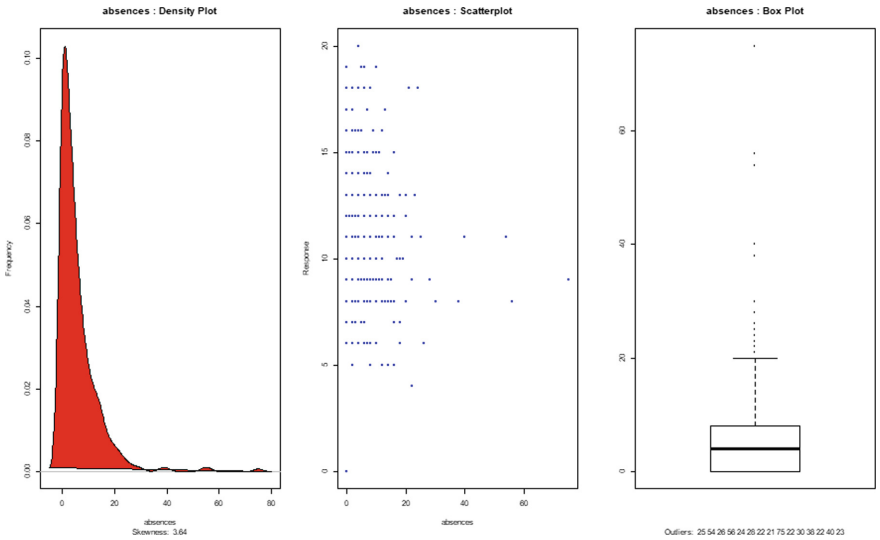

Fig. 2. The density-plot, scatter-plot and box-plot for absences' attribute

The above picture shows that the variable is highly skewed with a very low negative correlation with the dependent variable. Also, it is noticeable that there are many outliers within this variable that must be treated in the next step.

Outliers analysis

After creating the density plot, the scatter plot and the box plot for all the variables, we have noticed that the input data contains outliers that have to be analyzed. To analyze the outliers we have run an R code to replace the outliers with missing values (NAs), and then the missing values were handled using an imputation approach that consists in replacing the missing values in a particular variable by the variable's mean.

Testing the null and Alternate hypotheses

After analyzing the outliers, we have moved to test for each variable the two followings hypotheses:

Null Hypothesis H0: The population correlation coefficient isn't significantly different from zero. There is no significant linear relationship between the dependent and independent variables.

Alternative Hypothesis Ha: The population correlation coefficient is significantly different from zero. There is a significant linear between the dependent and independent variables.

The two above hypotheses can be tested using the p-value, if the p-value is less than a pre-determined significance level; the null hypothesis can be rejected safely.

In this step, we have built a simple regression model for each variable. If the value of p was less than 1 we decided to keep the variable. After running the R code specific to this step we found that the following variables were the ones that have a p-value less than 1:

Address, Mjob, schoolsup, paid, romantic, Medu, gout, age, Fedu, traveltime, studytime.

The above determined variables will be used in the next steps.

3.3 Features Selection

Building a multiple linear regression model is not only about using a dataset to create the mapping function from the input variables (X) to the output variable (Y), It's more about feeding the most important features into the training model. In the next subsections, we will present seven methods for selecting important variables. Those methods are already implemented in R packages.

Random Forest Method

Random forest can be considered as an effective method to select among all predictors the ones that best explain the variance in the response variable. There are two packages in R that implement the Random Forest Method.

- Package *randomForest* with the *randomForest*() method to build the model and the function *importance*() to measure the relative importance of predictors fed into the built model.
- Package party with the *cforest* method to build the model and the function *varimp*() to measure the relative importance of predictors fed into the built model.

After using the two packages, and plotting the results of the two methods *randomForest*() and *cforest*, we have obtained the two followings plots that display the ranking of variables from the less important to the most important (Fig. 3).

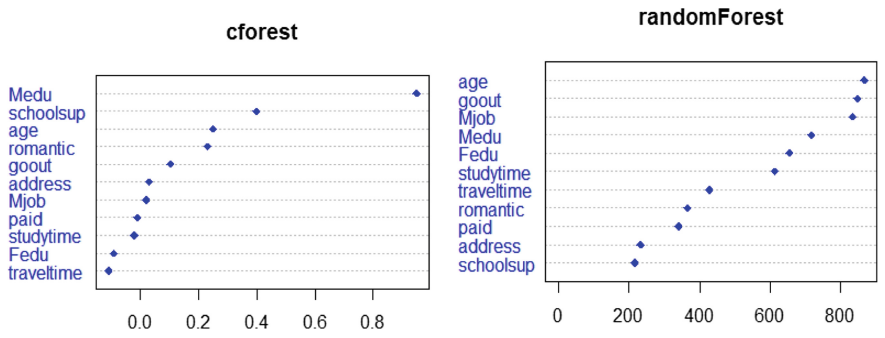

Fig. 3. The ranking of most important variables using cforest and randomForest methods

3.3.1 Relative Importance Package

The *R* package *relaimpo* implements six different metrics that measure the relative importance of repressors in a multiple regression linear model. In this work, we have used just three of them which are: *LMG*, *First* and *Last*. The three others metrics: *pmvd*,

betasq and *pratt* haven't been used because they deal only with a numerical dataset, while our data contains categorical variables.

First method

One way to determine the relative importance of variables is by comparing the impact of each one on the response, i.e., to build P simple regression models using the P regressors, and then compare their univariate R^2—values. The univariate R^2—values of each regressor is identical to the squared correlation of that regressor with the response.

The '*First*' method ranks the repressors' importance based on their univariate R^2—values.

Last method

Another way to determine the relative importance of variables is by comparing what each regressor is able to explain when all the other regressors exist in the model. The '*last*' method evaluates the importance of a regressor by measuring the increase in the total R^2 when adding this regressor as the last one.

LMG method

The '*LMG*' method demands more computational effort compared to the two methods discussed above. It decomposes R^2 into non-negative contributions that sum to the total R^2. The relative importance measure of the kth regressor can be given as:

$$Lmg(X_k) = \frac{1}{P!}\sum_{r \in P} R^2(\{X_k\}|S_k(r))$$

where:

- $S_k(r)$ is the set of regressors entered into the model before the regressor X_k corresponding to the order $r = (r_1, \ldots, r_p)$.
- $R^2(\{X_k\}|S_k(r)) = R^2(\{X\}_k \cup S_k(r)) - R^2(S_k(r))$
- P denotes the set of all permutations of $= (r_1, \ldots, r_p)$.

From the above definition, we can notice that the '*LMG*' metric uses both direct effects (orders where X_k enters first in the model) and effects adjusted to other regressors (X_k enters last).

The following diagram presents the relative importance variables resulted from each of the above-described methods after performing them with *calc.relimp*{*relaimpo*} (Fig. 4).

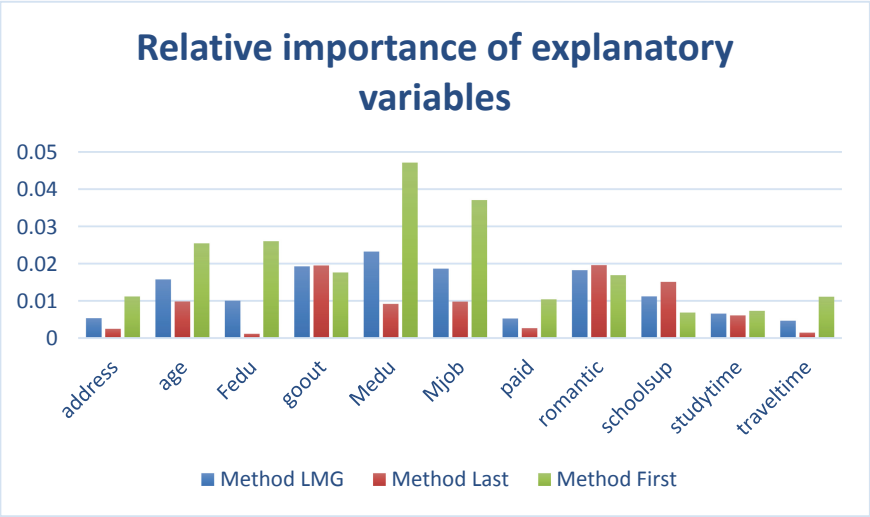

Fig. 4. The ranking of most important variables using 'LMG', 'Last' 'First' methods

Multivariate adaptive regression splines (MARS) Method

The '*Mars*' method, is implemented under the earth package, this latter uses a generalized cross-validation (GCV) statistic to determine the contribution (or variable importance score) for each predictor.

With the earth package, we use the method *earth*() to build the model, and the function *evimp*() to estimate variables' importance.

After using the earth package and plotting the results, we have obtained the following plot that displays the top five most important variables (Fig. 5).

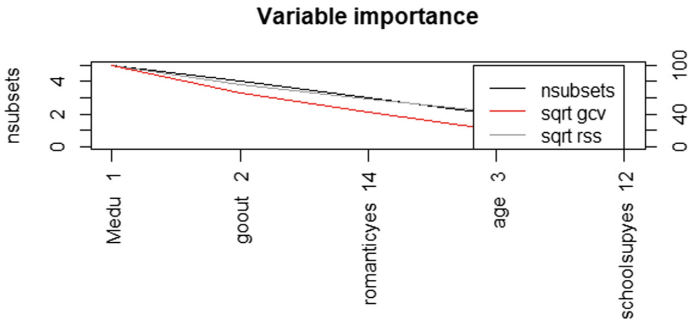

Fig. 5. The ranking of most important variables using 'MARS' method

Boruta Method

The 'Boruta' method can be used to decide among a set of variables which ones are important. This method is implemented in R under the name Boruta, It receives as input the dependent and independent variables, and it gives as output the set of important variables. After performing this method, we have found that the followings variables are the most important:

"Medu" "goout" "age" "Mjob" "schoolsup" "romantic".

4 Results and Discussion

4.1 Building the Models

After determining the most important variables resulted from each method in the previous section, we will use those selected variables to build multiples linear regression models, and then we will compare the performance of the built models. The model with the highest performance accuracy refers to the method that can select the most important variables.

The following table presents for each method, the top five most important variables and the built model (Table 2):

Table 2. The built models

Heading level	Methods	The top five important variables	The built model
Model 1	cforest	Medu, schoolsup, age, romantic, goout	17.7935 + 0.8627 * Medu − 1.6562 * schoolsupyes − 0.4423 * age − 1.2458 * romanticyes − 0.5603 * goout
Model 2	randomForest	age, goout, Mjob, Medu, Fedu	15.69329 − 0.37087 * age − 0.57549 * gout + 1.55061 * Mjobhealth + 0.17745 * Mjobother + 1.00327 * Mjobservices + 0.09585 * Mjobteacher + 0.65238 * Medu + 0.17046 * Fedu
Model 3	LMG	Medu, goout, Mjob, romantic, age	15.06200 + 0.83527 * Medu − 0.58406 * goout + 1.41263 * Mjobhealth + 0.09578 * Mjobother + 0.83088 * Mjobservices − 0.17181 * Mjobteacher − 1.20858 * romanticyes − 0.30396 * age
Model 4	First	Medu, Mjob, Fedu, age, goout	15.69329 + 0.65238 * Medu + 1.55061 * Mjobhealth + 0.17745 * Mjobother + 1.00327 * Mjobservices + 0.09585 * Mjobteacher + 0.17046 * Fedu − 0.37087 * age − 0.57549 * goout
Model 5	Last	romantic, goout, schoolsup, age, Mjob	19.9301 − 1.0979 * romanticyes − 0.5392 * goout − 1.6442 * schoolsupyes − 0.5021 * age + 2.7360 * Mjobhealth + 0.6203 * Mjobother + 1.7147 * Mjobservices + 1.4480 * Mjobteacher

(*continued*)

Table 2. (*continued*)

Heading level	Methods	The top five important variables	The built model
Model 6	MARS	Medu, goout, romantic, age, schoolsup	17.7935 + 0.8627 * Medu − 0.5603 * gout − 1.2458 * romanticyes − 0.4423 * age − 1.6562 * schoolsupyes
Model 7	Boruta	Medu, goout, age, Mjob, schoolsup	18.21514 + 0.76773 * Medu − 0.57590 * goout − 0.49550 * age + 1.24235 * Mjobhealth + 0.08792 * Mjobother + 0.91238 * Mjobservices − 0.17065 * Mjobteacher − 1.58944 * schoolsupyes

In the following subsection, we will describe the method used to assess the performance of the built models, and we will interpret the results of the validation metrics to identify the best model.

4.2 Testing the Performance of the Built Models

After building the above models, we have interested in determining the accuracy of those models in predicting the students' grade. The model with the highest accuracy refers to the method that determines the most important variables.

There are many methods that can be used to assess the performance of a regression model, in this work, we have used the k-fold cross-validation method because it's a widely recommended and used in regression and classification settings, and also due to its ability to give an accurate estimation of the test error rate.

K-fold cross-validation is described as follows:

1. The entire dataset is randomly split into k-subsets of approximately equal size
2. A model is trained on K − 1 of these subsets and tested on the remaining subset. While testing the model, a measure of the model error is obtained.
3. Repeat this process K times until each of the k subsets has used as the test set.
4. Compute the average of the K model errors. This is called the cross-validation error using to estimate the performance of the model.

When performing the K-fold using the caret R package, the followings regression-model-accuracy-metrics are computed. Those metrics can be used to measure the overall quality of regression models.

- R—squared (R^2): measures the squared correlation between the actual outcome values and the values predicted by the model. The higher the adjusted R2, the better the model.

$$R^2 = \frac{Model\ sum\ of\ squared}{Total\ sum\ of\ squared} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

where $\widehat{y}_i$ is the predicted outcome, y_i is the observed outcome and $\bar{y}$ *is the average of observed outcomes*

- Root Mean Squared Error (RMSE), which measures the average magnitude error made by the model while predicting an observation's outcome. It's the square root of the average of squared residuals. Residuals are the difference between the actual values and the predicted values. The lower the RMSE, the better the model.

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (\widehat{y}_i - y_i)^2}{N}}$$

- Mean Absolute Error (MAE), is a measure of average absolute differences between observed and predicted outcomes. The lower the MAE, the better the model

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} |y_i - \widehat{y}_i|$$

After testing the performance of the built models using the K-fold cross-validation method, we have obtained for each model the values of the metrics RMSE, R-square and MAE. The following table presents those values (Table 3).

Table 3. Regression models accuracy metrics

	RMSE	R-squared	MAE
Model 1	4.375388	0.1012048	3.305773
Model 2	4.479863	0.05837953	3.379078
Model 3	4.395619	0.09086479	3.355966
Model 4	4.442338	0.08832177	3.358106
Model 5	4.445559	0.08029582	3.348512
Model 6	4.368274	0.1025792	3.301121
Model 7	4.398278	0.08766335	3.304367

As we can see from the above table, the model that has the heist R-squared and lowest RMSE and MAE Is the model 6, so we can say that this model is more performant than the others models in predicting students' grads. From Table 1 we can notice that the method used to select model 1's variables is 'MARS' and the selected variables are: *Medu, gout, romantic, age, schoolsup*. Therefore, we can conclude that the best method for selecting important variables in our experiment is 'MARS' method, and the students' attributes that can affect their final grade results more than others are: The mother education level, the student's age, the student' romantic situation, the time spent with friends and the extra educational support.

As we can notice from model 6, all the variables have negative correlation coefficients except mother-education attribute, which means that the increase in the age and goout variables will cause a decrease in the final grade. Also, when the value of the two categorical variables romantic and *schoolsup* is yes, the outcome variable decreases. While when the value of mother-education attribute increases the outcome value increases too.

5 Conclusion

Determining the factors that affect the students' performance in academic institutions is a very interesting task since it will help educators to enhance their learning and teaching process. In this context, we have proposed a methodology that well examines the students' attributes and selects among them the most important to build a prediction model. Our methodology consists in applying different methods for selecting the most important variables and then using the selected variables to build different multiples linear regression models. After comparing the performances of the built models, we have found that the most performant is the one created using 'MARS' method.

In future work, we would like to have used a dataset that records Moroccan university students' attributes, in order to identify the factors that influence their performance in Moroccan universities. Also, we would like to have compared the performance of multiple regression technique with others regression and classification techniques.

References

1. Bakhshinategh, B., Zaiane, O.R., ElAtia, S., Ipperciel, D.: Educational data mining applications and tasks: a survey of the last 10 years. Educ. Inf. Technol. **23**(1), 537–553 (2018)
2. Grömping, U.: Relative importance for linear regression in R: the package relaimpo. J. Stat. Softw. **17**(1), 1–27 (2006)
3. Cortez, P., Silva, A.: Using data mining to predict secondary school student performance. In: Proceedings of 5th Annual Future Business Technology Conference (FUBUTEC 2008), pp. 5–12 (2008)
4. El Aissaoui, O., El Alami El Madani, Y., Oughdir, L., El Allioui, Y.: A fuzzy classification approach for learning style prediction based on web mining technique in e-learning environments. Educ. Inf. Technol. 1–17 (2018)
5. El Aissaoui, O., El Alami El Madani, Y., Oughdir, L., El Allioui, Y.: Combining supervised and unsupervised machine learning algorithms to predict the learners' learning styles. Procedia Comput. Sci. **148**, 87–96 (2019)
6. El Aissaoui, O., El Madani El Alami, Y., Oughdir, L., El Allioui, Y.: Integrating web usage mining for an automatic learner profile detection: a learning styles-based approach. In: 2018 International Conference on Intelligent Systems and Computer Vision (ISCV), pp. 1–6 (2018)
7. El Aissaoui, O., El Madani El Alami, Y., Oughdir, L., El Allioui, Y.: A hybrid machine learning approach to predict learning styles in adaptive E-learning system. Adv. Intell. Syst. Comput. **915**, 772–786 (2019)
8. Shahiri, A.M., Husain, W., Rashid, N.A.: A review on predicting student's performance using data mining techniques. Procedia Comput. Sci. **72**, 414–422 (2015)
9. Aldowah, H., Al-Samarraie, H., Fauzy, W.M.: Educational data mining and learning analytics for 21st century higher education: a review and synthesis. Telemat. Inf. **37**, 13–49 (2019)
10. Quadri1, M.M., Kalyankar, N.V.: Drop out feature of student data for academic performance using decision tree techniques. Glob. J. Comput. Sci. Technol. **10**(2), 2–5 (2010)

11. Mishra, T., Kumar, D., Gupta, S.: Mining students' data for prediction performance. In: 2014 Fourth International Conference on Advanced Computing & Communication Technologies, pp. 255–262 (2014)
12. Parack, S., Zahid, Z., Merchant, F.: Application of data mining in educational databases for predicting academic trends and patterns. In: 2012 IEEE International Conference on Technology Enhanced Education (ICTEE), pp. 1–4 (2012)
13. Helal, S., et al.: Predicting academic performance by considering student heterogeneity. Knowledge-Based Syst. **161**, 134–146 (2018)
14. Osmanbegovic, E., Suljic, M.: Data mining approach for predicting student performance. Econ. Rev. J. Econ. Bus. **10**(1), 3–12 (2012)
15. Jishan, S.T., Rashu, R.I., Haque, N., Rahman, R.M.: Improving accuracy of students' final grade prediction model using optimal equal width binning and synthetic minority over-sampling technique. Decis. Anal. **2**(1), 1 (2015)
16. Mayilvaganan, M., Kalpanadevi, D.: Comparison of classification techniques for predicting the performance of students academic environment. In: 2014 International Conference on Communication and Network Technologies, pp. 113–118 (2014)
17. Arsad, P.M., Buniyamin, N., Manan, J.A.: Prediction of engineering students' academic performance using artificial neural network and linear regression: a comparison. In: 2013 IEEE 5th Conference on Engineering Education (ICEED), pp. 43–48 (2013)
18. Yang, S.J.H., Lu, O.H.T., Huang, A.Y.Q., Huang, J.C.H., Ogata, H., Lin, A.J.Q.: Predicting students' academic performance using multiple linear regression and principal component analysis. J. Inf. Process. **26**, 170–176 (2018)
19. Widyahastuti, F., Tjhin, V.U.: Predicting students performance in final examination using linear regression and multilayer perceptron. In: 2017 10th International Conference on Human System Interactions (HSI), pp. 188–192 (2017)
20. Dietrich, D., Heller, R., Yang, B., EMC Education Services: Data science and big data analytics : discovering, analyzing, visualizing and presenting data
21. Hoffman, J.I.E., Hoffman, J.I.E.: Multiple regression. In: Basic Biostatistics for Medical and Biomedical Practitioners, 2nd edn. pp. 525–560. Academic Press, Cambridge (2019)