Students Performance Prediction by Using Data Mining Algorithm Techniques

Mohammed Nasih Ismael

Geography Department, College of Arts, Kirkuk University, Iraq
mohammednasih82@gmail.com

ABSTRACT

The use of educational data mining with different techniques aims to solve various problems of the educational environment, especially educators to help learners avoid failure in the study. In this proposal, information on extracting educational data and how to extract knowledge from it to help the educational environment is presented. In this context, previous studies aimed at predicting student performance and the traits that influence their performance have been extensively presented in this research. However, each researcher used different attributes with different techniques to reach the same goal which is to predict the level of students based on data collected from educational institutions. In the context of data preprocessing, the selected data is worked on and configured to work correctly with WEKA. Then, in order to better understand the data, data attributes and data format were introduced. In addition to that we propose in our study to use a learning dataset from the UCI repository and analyze it using Waikato Environment for Knowledge Analysis (WEKA), classification and regression techniques were applied. Then, the characteristics that affect the student's performance, and predict the student's academic performance on the other hand, were determined. Finally, the Random Forest algorithm showed a superiority over the Linear Regression algorithm in the prediction of values only in training set mode.

Keywords: Data mining, Classification, Prediction, WEKA, student performance, Attributes selection.

1. Introduction

Educating data mining is an area which utilizes data mining algorithms over different data types, such as data mining algorithms, machine learning, and statistical data. Its main aim is to analyze such data to solve issues of educational research. EDM is working to develop ways to explore the unique data types in educational settings and to understand students and their learning environments more effectively using these ways [1]. Furthermore, Educational data mining (EDM) uses data mining (DM), machine learning, information retrieval, psycho-pedagogy, cognitive psychology and statistics to solve various educational problem areas through the application of machine-learning methods and technological advice systems. Also, the EDM is defined as an “Emerging Discipline” by the International Educational Data Mining Society. It seeks to develop methods to determine unique kinds of information from educational environments and better understand students and their learning environments [2]. The

number of studies in the field of educational data mining EDM has increased significantly, with the aim of providing a better learning environment and helping both parties (Educators and Learners) to use the best methods to reach the main goal, which is learning and success. In this regard, a number of literatures that deals with this field, methodology, data, and results will be presented. The researcher Shakir Khan [3] tested the effect of factors on student performance, based on the used dataset, which contains many factors that can affect the final score. The tests are done by using a linear regression model on the response variable for the final grade G3, and the result obtained was only absence and study time, which can affect students' performance. The methodology comprises of the following steps:

1. Data collection.
2. Data pre-processing, training, and testing dataset generation.
3. Model development.
4. Prediction in different phases of the model.

Similarly, Pallathadka et al [4] use the same dataset for the purpose of classifying students based on their performance, showing the possibility of evaluating question papers to determine levels of difficulty. Machine learning algorithms such as SVM, ID3, Naive Bayes, and C4.5 were used and tested by evaluating some metrics, such as error rate and accuracy.

This study demonstrates that SVM has more accuracy in classifying a dataset of student performance. The framework of this study consisted of the following points:

1. UCI student performance Dataset.
2. Preprocessing of dataset.
3. Classification using machine learning algorithms.
4. Classification result.
5. Prediction of student performance.

Also, Nahar et al [5] classified and predicted the results of a student using two types of datasets. Two models were generated from the result of the analysis. The first model is based on the decision tree, and it was generated according to the first dataset (known as AI Prerequisites dataset). The model provided an accuracy of 64.3% in the test dataset. The following model is based on the Naive Bayes built by analyzing the second dataset (known Theory Performance). The accuracy of this model on the test dataset reached (75%). The methodology consists of data collection, preprocess data, apply various classifiers and build model, and evaluate model. While MENGASH [6] conducted a study to support higher education institutions to make good decisions in the admission process. The study anticipates applicants' academic performance before they are accepted. The performance accuracy levels in the ANN models have been reached around 79.22%. The implementation of the new weighting system demonstrated the following results in the first year: 1) 18% decrease of the proportion of students with acceptable or poor CGPAs. 2) 31% increase in the proportion of students with excellent or very good CGPAs. The methodology in this study depends on answering a number of questions using data mining techniques such as (Is it possible to determine an admission criterion that accurately predicts the future academic performance of applicants?). On the other hand, Harvey & Kumar [7] used prediction ratings to analyze data in K-12 education. Three classifiers have been utilized to develop the productive models, including Naive Bayes techniques, decision tree, and linear regression. The highest accuracy has been obtained by Naive Bayes techniques to predict SAT Math scores for high school students. Methodology of this study

consists of Input data, clean data, create training/Testing subset, develop model using training data, Test model using testing data, calculate accuracy of model using confusion matrix. Besides, Khan et al [8] presented a study whose primary goal was to anticipate failure, decline, and improvement before the start of the semester. The study proposed a scalable classifier, called the random wheel that works with categorical attributes only. The proposed classifier correctly predicts success and failure of more than 80%. Furthermore, prior to starting the course 2 out of 3 performance improvements were expected. In this context, Tsiakmaki et al [9] introduced an active fuzzy-based learning method for dictating a priori students' academic performance that, in a modular way, combines autoML practices. Many experiments have been conducted which demonstrate the effectiveness of the method proposed to accurately predict the risk of failure of students. An important tool for identifying underperforming higher education students may be the proposed fuzzy-based method. The study methodology proposed a hybrid method that takes advantage of active learning potential and incorporates the fuzzy approach to learning.

Moreover, Dabhade et al. [10] looked at anticipating students' academic performance at a technical institution in India. To determine the academic performance, the machine learning algorithms, such as support vector-linear regression, support vector-Poly regression, support vector-Boolean regression, and multiple linear regression, were utilized. The supportive vector regression algorithm has provided a superior prediction. The linear model achieved the best fit with 83.44% accuracy. The results obtained demonstrated the relationship between the students' academic performance and their behavior traits. Aggarwal et al. [11] compared two models: One was built according to academic criteria only. In contrast, the latter was based on academic and non-academic (Demographic) criteria. The models were built using eight classification algorithms. According to the results, a set of academic and non-academic parameters only provided the best prediction model. This was demonstrated by comparing the F1-score, which improved in almost all classification models if non-academic parameters were also considered with academic criteria. A summary of all this literature and information on the data and technique used are shown in Table 1.

Table1: Summary of works of literature.

Ref.	Dataset	Software	Technique	No. of instance	Repository
[3]	University of Minho Portugal	R software	Regression Model	649	UCI
[4]	University of Minho Portugal	Not mentioned	Naive Bayes, Decision Tree, Bagging, Boosting, Part, RF	649	UCI
[5]	Notre Dame University Bangladesh	WEKA	Decision tree, Naive Bayes	80	Real data collected by researcher
[6]	PNU Saudi public university	WEKA	Naïve Bayes, Support Vector Machines, Decision Trees and	2039	Real data collected by researcher

			ANN		
[7]	Massachusetts Public Schools	WEKA	Naive Bayes techniques, decision tree Linear regression	1861	Massachusetts Department of Elementary and Secondary Education website
[8]	Institute of national importance in India (ACAD-INST)	WEKA	Random wheel	14264	Academic information management system (AIMS)
[9]	Aristotle University of Thessaloniki (AUTh)	Scikit learn	Fuzzy algorithm	866	Real data collected from compulsory courses
[10]	Technical institution in India	Python3	MLR, SVR (Rbf, Linear, Poly)	85	Questionnaire-based & GPA from academic section of the institution
[11]	Uttar Pradesh Technical college in India	WEKA	LR, SVM, MLP, J48, RF, AdaBoost, Bagging & Voting	6807	Real data collected from the college

2. Basic Concepts

The available data in the UCI repository consisting of 649 samples suitable for classification and regression purposes will be use [12]. To complete the analysis, the Waikato Knowledge Analysis Environment (WEKA) will be used. In addition, we will divide the data set into two groups based on the attributes that will be used for classification and regression purposes and point out the parameters that affect student performance. Below (Table 2) is information about the data to be used:

Table 2: The student achievement of two Portuguese schools (in secondary education).

Data Set Characteristics:	Multivariate	**Area:**	Social	**Number of Instances:**	649
Attribute Characteristics:	Integer	**Date Donated:**	2014-11-27	**Number of Attributes:**	33
Associated Tasks:	Classification, Regression	**Number of Web Hits:**	1039759	**Missing Values?**	N/A

3. Data Preprocessing

According to [13] the preprocessing is the first step in obtaining knowledge from data using data mining techniques as shown in Figure.1. So, at this stage the data is prepared for data mining technique. The data that was selected and downloaded from the UCI repository was in csv extension, when trying to load it into WEKA it did not work, and the error shown in Figure.2 occurred. After checking the contents of the data, all the words and numbers were surrounded by double quotation (") and after replacing them with a space, the file was uploaded by the program. After that, data still contained an error, as the number of attributes in the original data according to the metadata was 33 as shown in table 2, but the data was loaded with 649 samples and one attribute, meaning that the program did not recognize the number of attributes correctly, as shown in Figure.3.

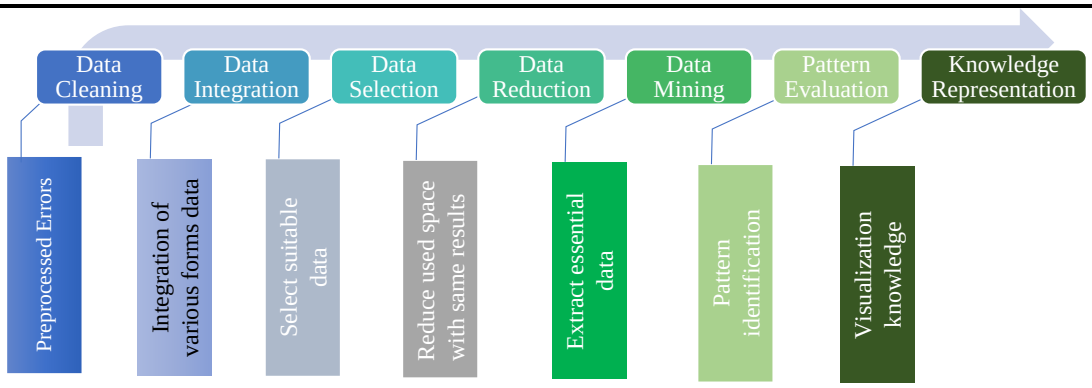

Fig.1: Steps of Extraction Knowledge

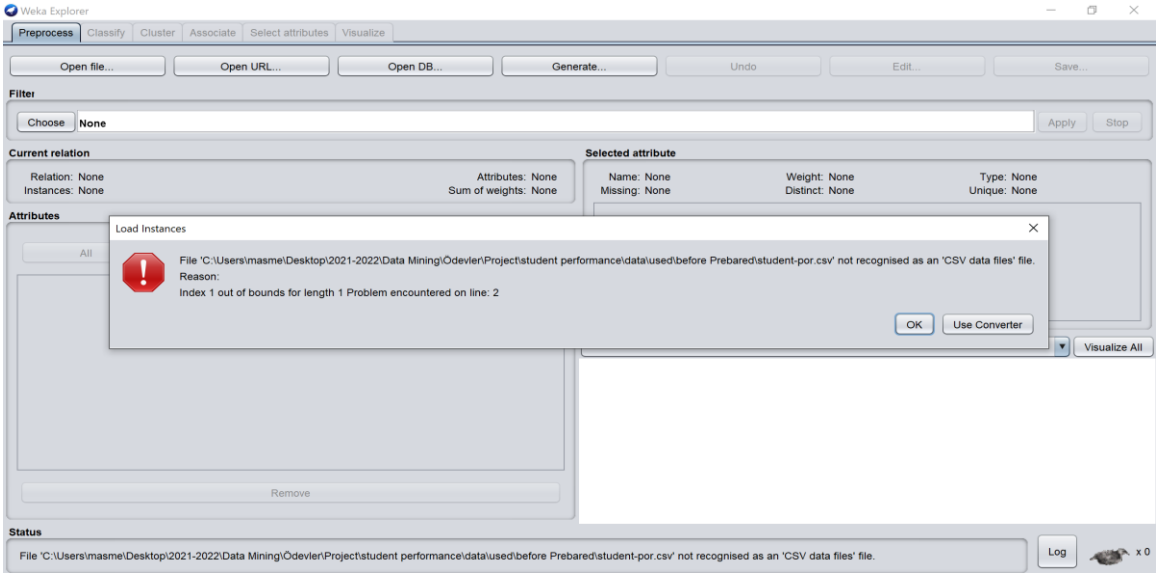

Figure 2: Error when loading data in WEKA

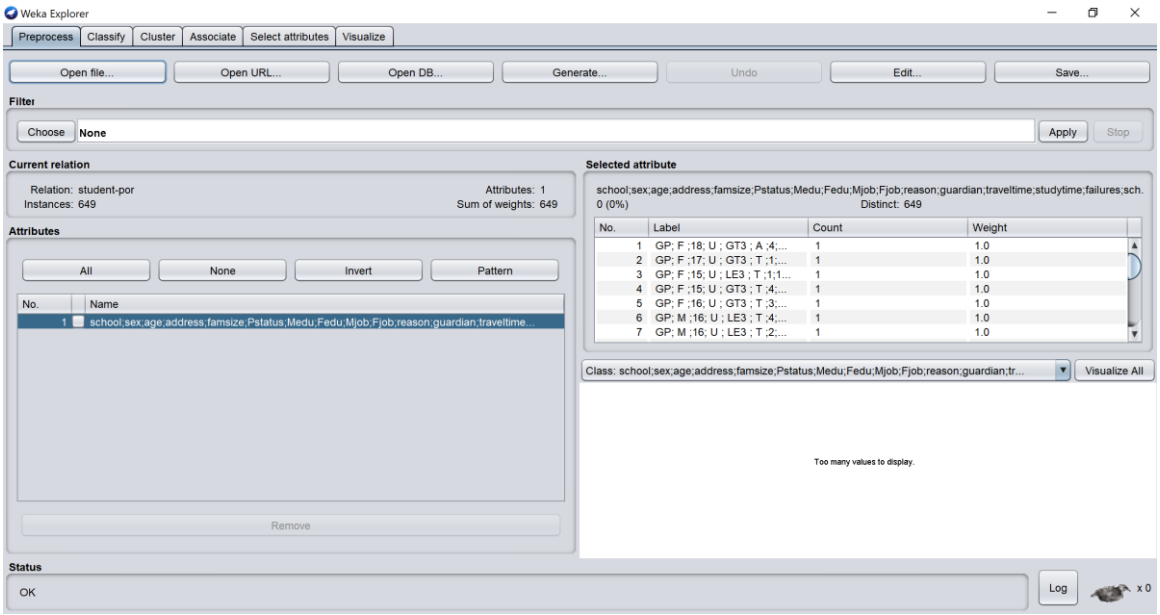

Figure 3: Loading data after the first processing

In order to solve this error, Microsoft Excel was used, after importing the data, the number of original attributes was restored, shown in Figure 4.

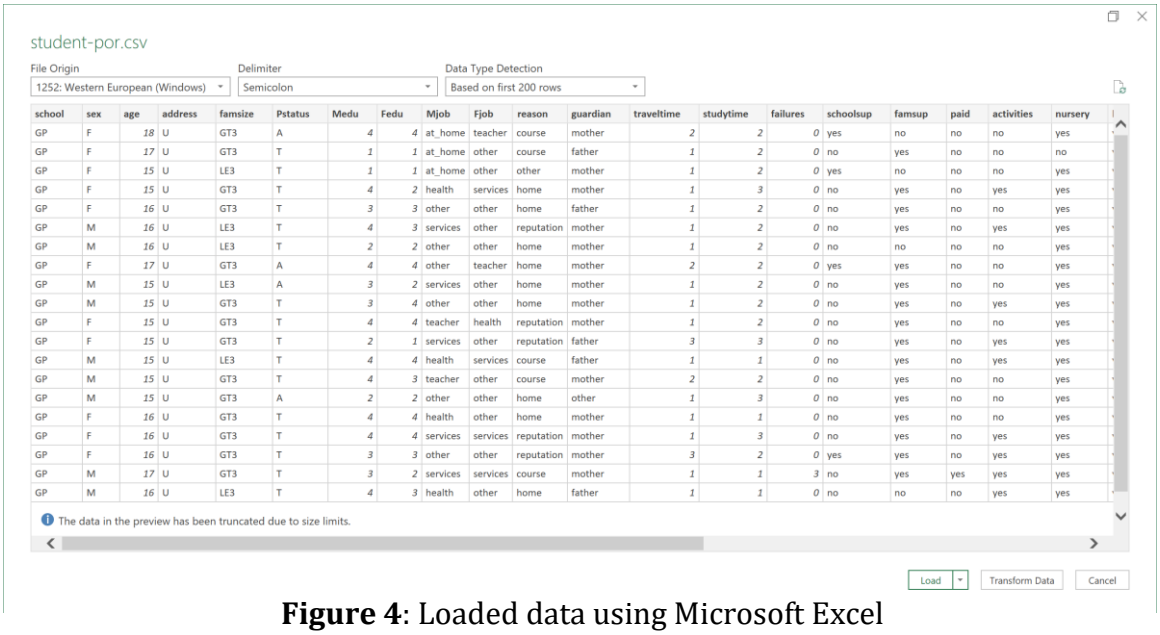

Figure 4: Loaded data using Microsoft Excel

Then, the data was saved in the current format with csv extension again, and then re-uploaded using WEKA where the attributes were uploaded correctly as shown in Figure 5.

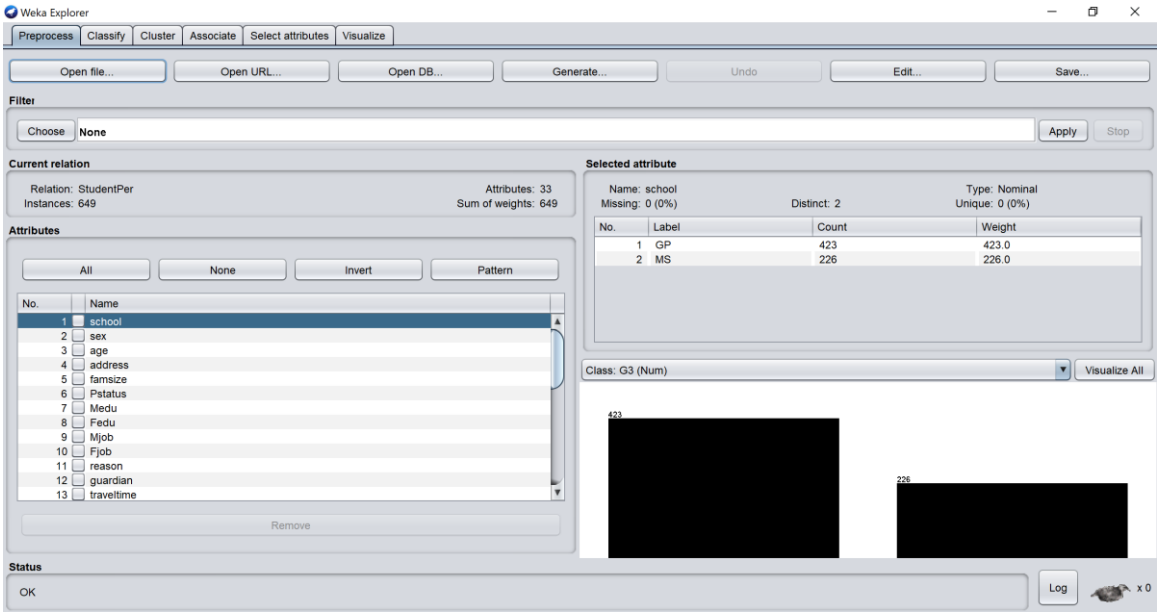

Figure 5: Loaded data after second processing

After the data upload procedure to WEKA was completed correctly, it was verified that there were no missing data in all the attributes. Therefore, the data is ready for the stage of its use in data mining techniques.

4. Data analysis

In order to work with data, better understanding provides more accurate results. Therefore, Table No. 3 shows the variables related to the student[14].

Table 3: The preprocessed student related attributes

Attributes	Type	Description
sex	Binary	Sex of the student
age	Numeric	Age of the student
address	Binary	The address type of the student's home
school	Binary	School of the student
Pstatus	Binary	The cohabitation status of the Parent
Fedu	Numeric	The education of the Father
Medu	Numeric	The education of the mother
Fedu	Numeric	The education of the Father
Mjob	Nominal	The job of the mother
guardian	Nominal	The guardian of the student
famsize	Binary	Family size
famrel	Numeric	Quality of family relationships
reason	Nominal	Reason to choose this school
traveltime	Numeric	The travel time from home to school
studytime	Numeric	Weekly study time
failures	Numeric	Number of past class failures
schoolsup	Binary	Extra educational school support
Famsup	Binary	Family educational support
activities	Binary	Extra-curricular activities
paidclass	Binary	Extra paid classes
internet	Binary	Internet access at home
nursery	Binary	Attended nursery school
higher	Binary	Needs for higher education
romantic	Binary	With a romantic relationship
freetime	Numeric	Free time after school
goout	Numeric	Going out with friends
Walc	Numeric	Consumption of alcohol on the weekend
Dalc	Numeric	Consumption of alcohol in Workday
health	Numeric	Current health status
absences	Numeric	Number of school absences
G1	Numeric	The grade of the first period
G2	Numeric	The grade of the second period
G3	Numeric	Final grade

The features mentioned in the above table (Table 3), some of which may not be closely related to the desired outcome or have little effect. Therefore, classifiers can be used in WEKA to make the results clearer and more accurate. In the next section, we will discuss how to choose the most important attributes and use appropriate algorithms to generate models.

5. Experimental results

Using the algorithms available in WEKA at the stage of data preprocessing, the most important influencing attributes can be selected from used dataset (Table 3). After applying the ***AttributeSelectedClassifier*** classifier the obtained results, according to the selected parameters, as shown in table 4, and the visualization of those attributes shown in figure 6.

Table 4: Applied (*AttributeSelectedClassifier)* classifier to obtain most important attributes

The most important remaining attributes	***AttributeSelectedClassifie***	**Settings**	
school {GP, MS}	Before being passed on to a classifier, the dimensionality of training and test data is decreased	**Classifier**	**RandomForest**: Class for constructing a forest of random trees.
studytime numeric		**Evaluator**	**CfsSubsetEval**: Evaluates the worth of a subset of attributes by considering each feature's individual predictive ability and the degree of redundancy between them.
		Search	**BestFirst:** Searches the space of attribute subsets by greedy hill climbing augmented with a backtracking facility.
failures numeric	**Summary of applying classifier**		
higher {' yes ',' no '}	Correlation coefficient	0.9847	
freetime numeric	Relative absolute error	16.0428 %	
Dalc numeric	Mean absolute error	0.386	
health numeric	Root relative squared error	17.9033 %	
G2 numeric	Root mean squared error	0.5779	
G3 numeric	Total Number of Instances	649	

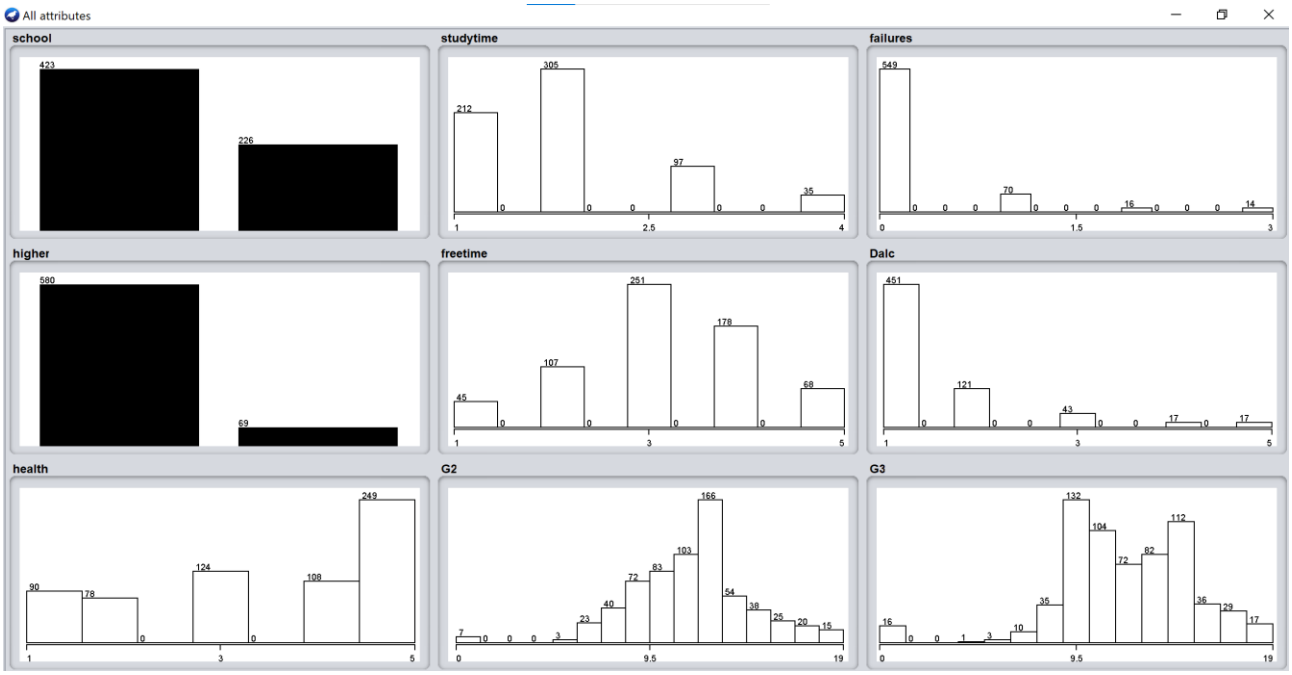

Figure 6: Visualization of remaining attributes

In addition, first period grade G1 will be included in the model because it is difficult to predict final grades without it. Also, according to the plot matrix visualization which shown in Figures 7,8 there is a strong positive linear correlation between them (G1, G3 and G2, G3). In a sense, the correlation coefficient is a filtering technique. It is the Pearson correlation coefficient method that is most commonly used.

The following is the equation for the Pearson correlation coefficient method[15]:

$$P_{X,Y} = \frac{cov(X,Y)}{\sigma^X \sigma^Y} \qquad \text{Eq.1}$$

Where $cov(X,Y)$ is the covariance of X and Y. $\sigma^X \sigma^Y$ are respectively the standard deviation of X and Y.

Figure 7: Correlation of G1, G3

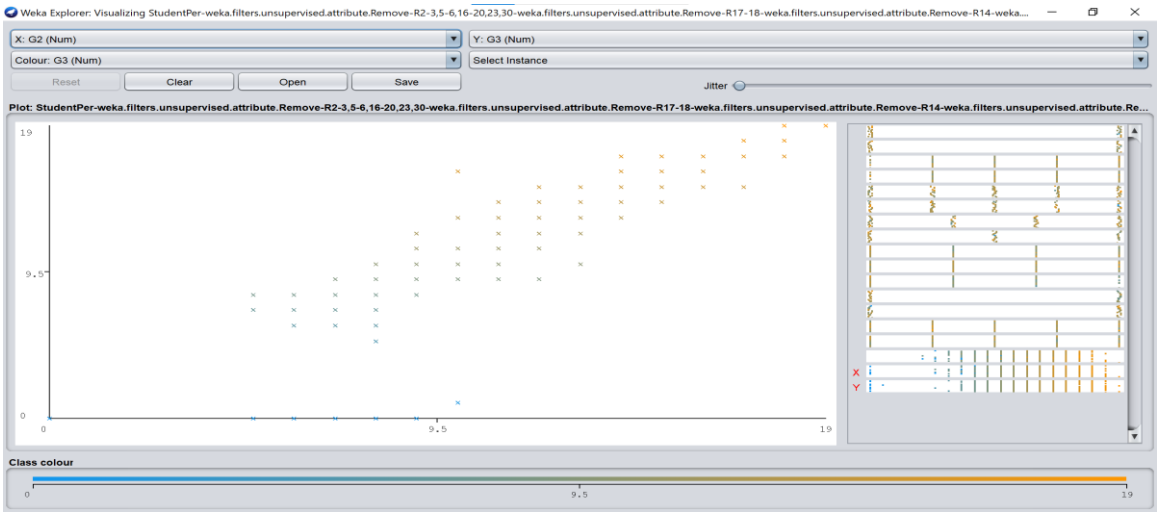

Figure 8: Correlation of G2, G3

On the other hand, in order to ensure that the features selected include the most influential features (G1 and G2) shown in Figures (8,9) other algorithms can be used. These algorithms are available in the used platform. So, algorithms for evaluating attributes were used and the results as shown in the following table 5:

Table 5: Applied methods to obtain most important attributes

No.	Attribute evaluation method	Search method	Attribute selected mode
1.	Classifier feature evaluator	Attribute ranking	Use full training set, No class

Obtained attributes (first 9 from 33 attributes)		
G2, G1, Fjob, reason, guardian, traveltime, studytime, Mjob, Fedu		
2. ReliefF Ranking Filter	Attribute ranking	Use full training set, No class
Obtained attributes (first 9 from 33 attributes)		
G2, G1, failures, higher, school, address, goout, Medu, freetime		
3. CFS Subset Evaluator	Greedy Stepwise (forwards)	Use full training set, No class
Obtained attributes (8 from 33 attributes)		
School, studytime, failures, higher, freetime, Dalc, health, G2		
4. Correlation Ranking Filter	Attribute ranking	Use full training set, No class
Obtained attributes (first 9 from 33 attributes)		
G2, G1, higher, school, studytime, Medu, Fedu, address, internet		

Depending on the results obtained in the table above, the common and the most effective characteristics between the used methods are G1, G2, on this basis we will choose all the characteristics in the results (1, 2, 4) from the above table, where the number of selected characteristics will be (17) from (33), which are (*G1, G2, higher, school, studytime, Medu, Fedu, failures, freetime, address, internet, goout, Fjop, Mjob, reason, guardian, traveltime*). Moreover, only the results that shared the attributes (G1, G2) were adopted, so the results of Table No. 4 were not taken into consideration.

The procedures implemented in the previous three sections (3, 4, 5) and the following procedures can be summarized in the following figure No. 9:

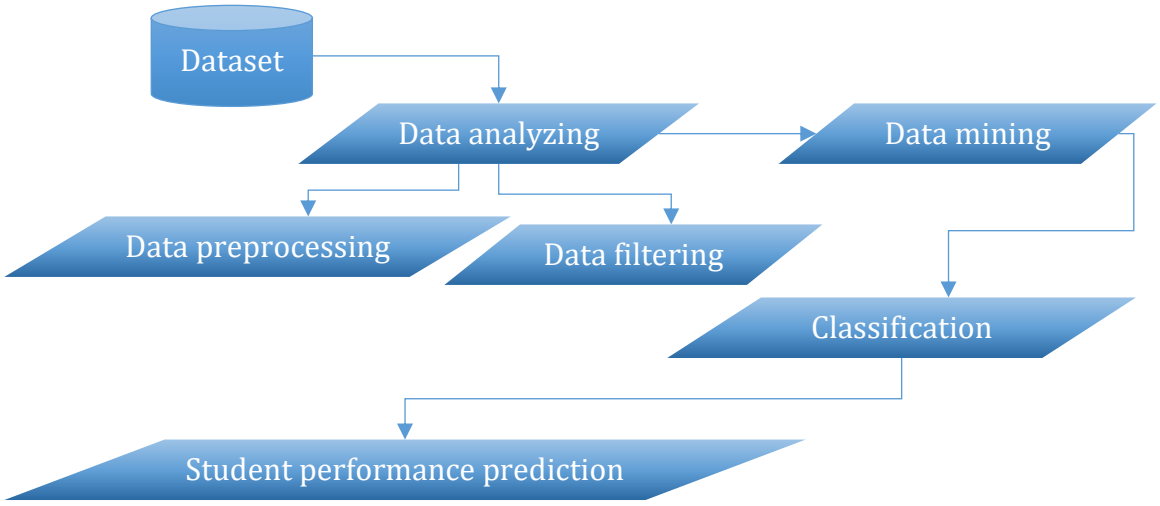

Figure 9: Procedures used to reach the final result

At this point, the dataset is ready for applying the classification algorithms, *Linear Regression* and *Random Forest* were chosen to work with the selected dataset because these algorithms deal with various numeric and nominal data. Also, due to the fact that they are commonly used in prediction.

5.1 Linear Regression

In machine learning, there are two basic types of methods, supervised and unsupervised technique of machine learning [16]. The linear regression algorithm, which is one of supervised learning algorithms, represents by the following equation [17]:

$$Y = a * X + b \qquad Eq.2$$

Where Y: Dependent variable, a: slope, X: Independent variable, b: Intercept

After applying the algorithm to the selected data with the most important influencing characteristics, the results were as shown in Table 6, and Figures 10, 11,12.

Table 6: classifier output (Linear Regression)

Evaluation on training set		**Cross-validation (10-Folds)**
Correlation coefficient	0.9246	0.9163
Relative absolute error	32.1579 %	33.5661 %
Mean absolute error	0.7737	0.8095
Root relative squared error	38.0891 %	39.9801 %
Root mean squared error	1.2296	1.2935
Total Number of Instances	649	649
Evaluation on test split 70%		
Correlation coefficient	0.8896	
Relative absolute error	37.4864 %	
Mean absolute error	0.8881	
Root relative squared error	46.4903 %	
Root mean squared error	1.4845	
Total Number of Instances	195	

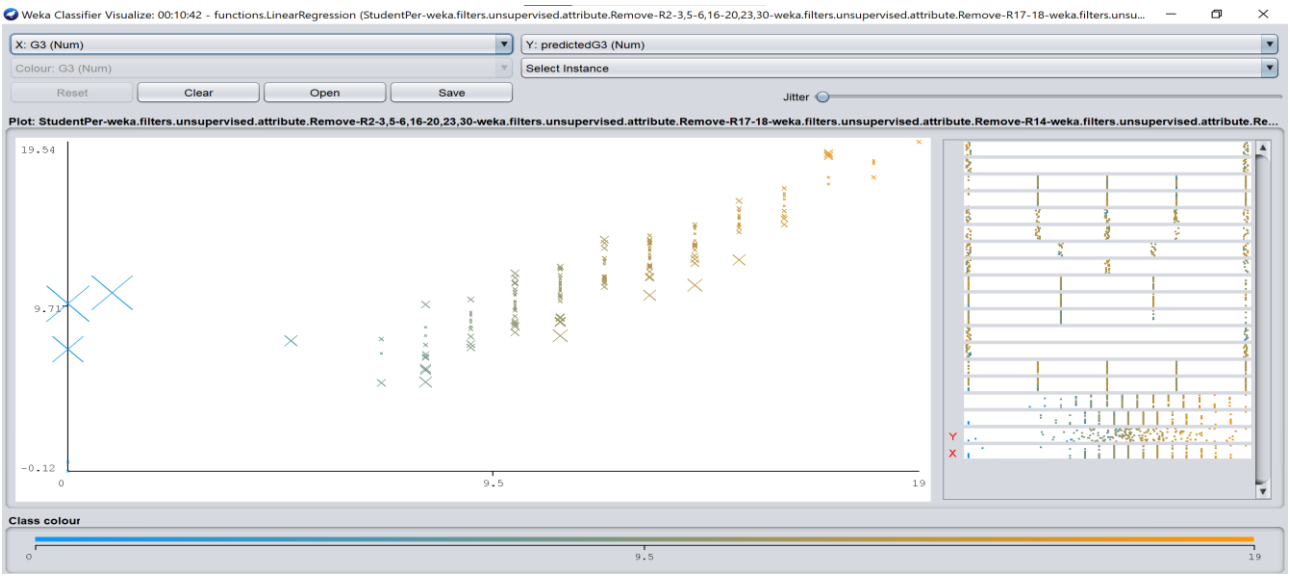

Figure 10: Visualization of classifier errors (*195 instances, Test split 70%*)

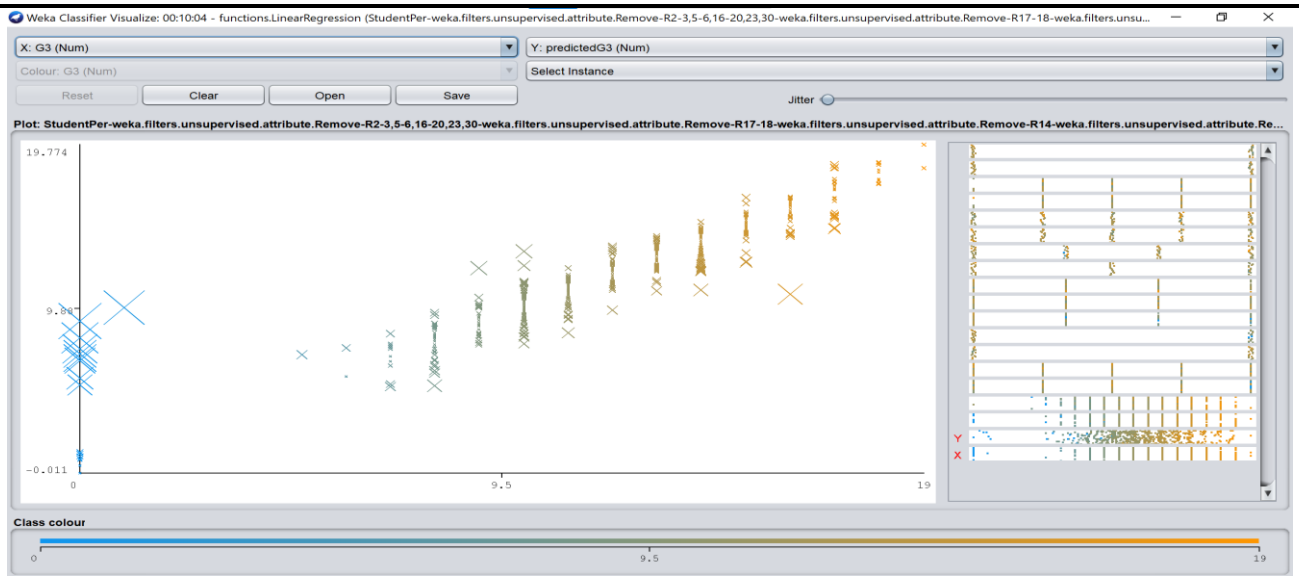

Figure 11: Visualization of classifier errors (*649 instances, Training set*)

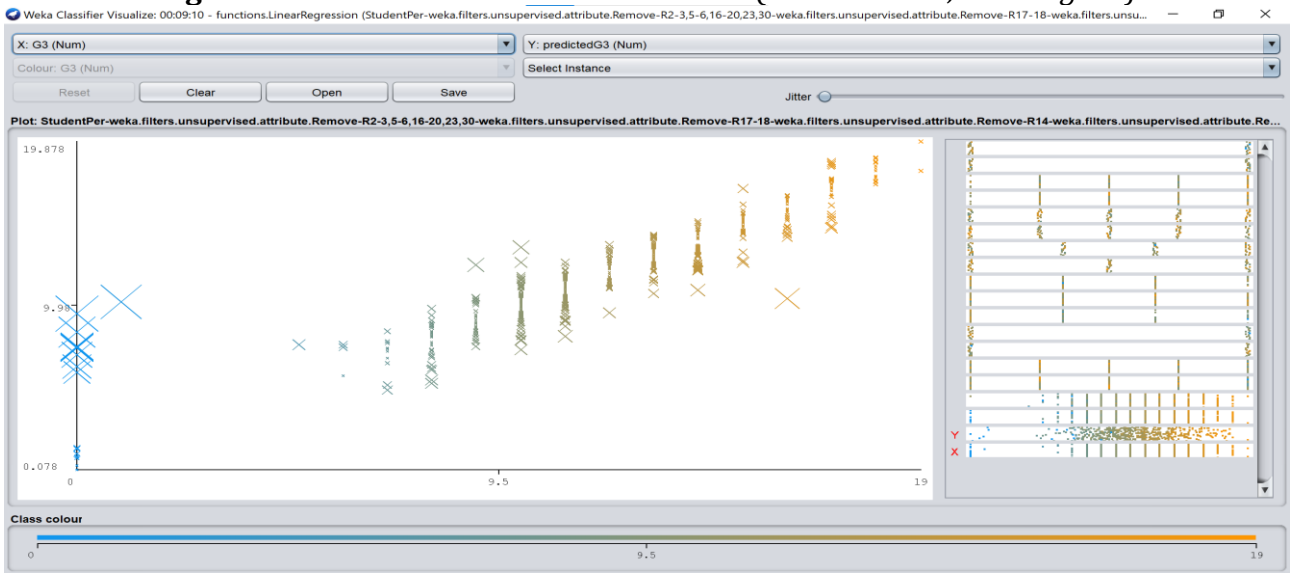

Figure 12: Visualization of classifier errors (*649 instances, Cross-validation*)

5.2 Random Forest

Different samples from the dataset are used to build several different decision trees in Random Forest. A random selection of each variable is used for data splitting at each node in this process. As a result of the implementation, the method now works by building multiple trees and combining their predictions [16]. After applying the algorithm to the selected data with the most important influencing characteristics the results were as shown in Table 7, and Figures 13, 14,15.

Table 7: classifier output (Random Forest)

Evaluation on training set		**Cross-validation (10-Folds)**
Correlation coefficient	0.9898	0.8975
Relative absolute error	13.7213 %	36.7403 %
Mean absolute error	0.3301	0.8861
Root relative squared error	15.8195 %	44.7576 %
Root mean squared error	0.5107	1.448
Total Number of Instances	649	649

Evaluation on test split 70%	
Correlation coefficient	0.8757
Relative absolute error	38.9795 %
Mean absolute error	0.9235
Root relative squared error	48.819 %
Root mean squared error	1.5589
Total Number of Instances	195

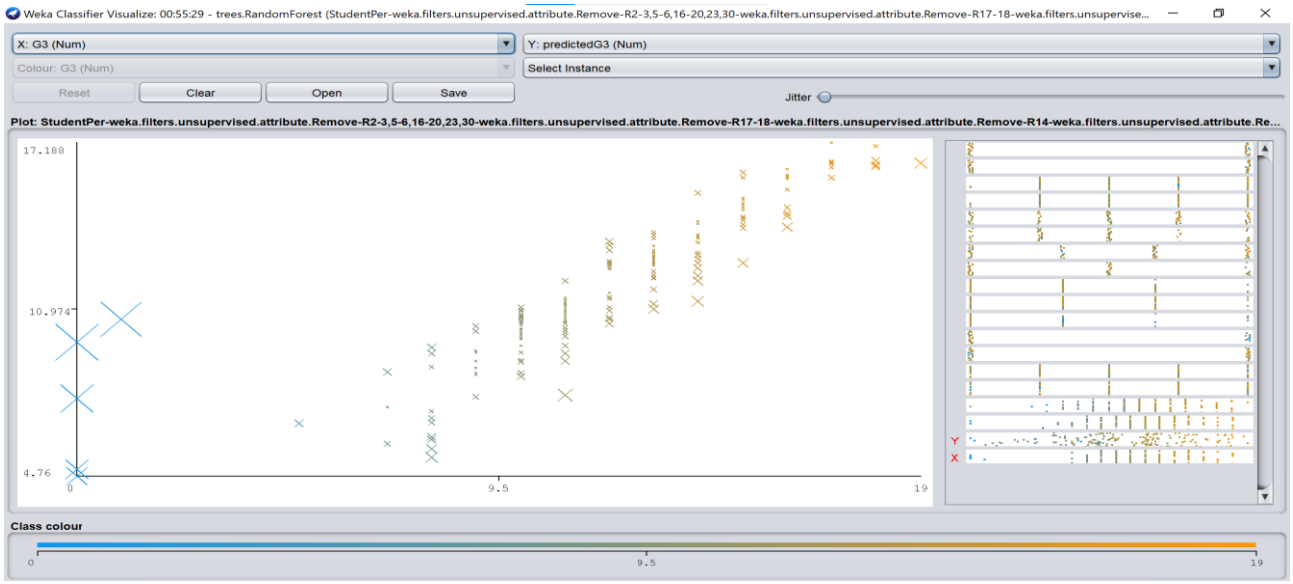

Figure 13: Visualization of classifier errors (*195 instances, Test split 70%*)

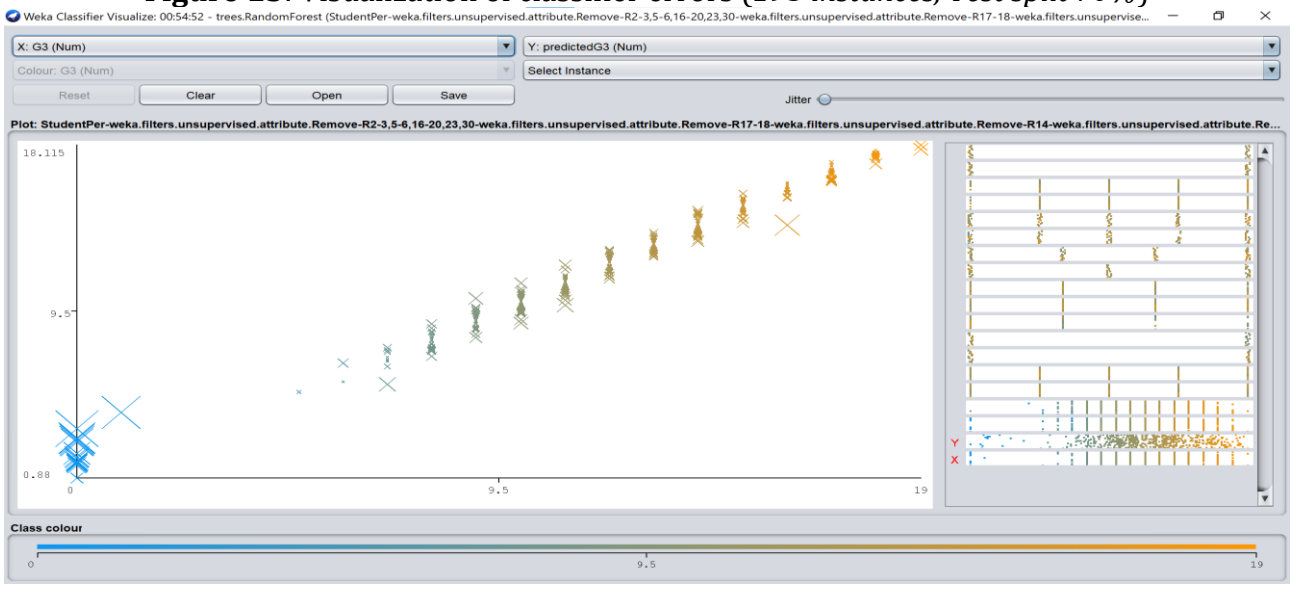

Figure 14: Visualization of classifier errors (*649 instances, Training set*)

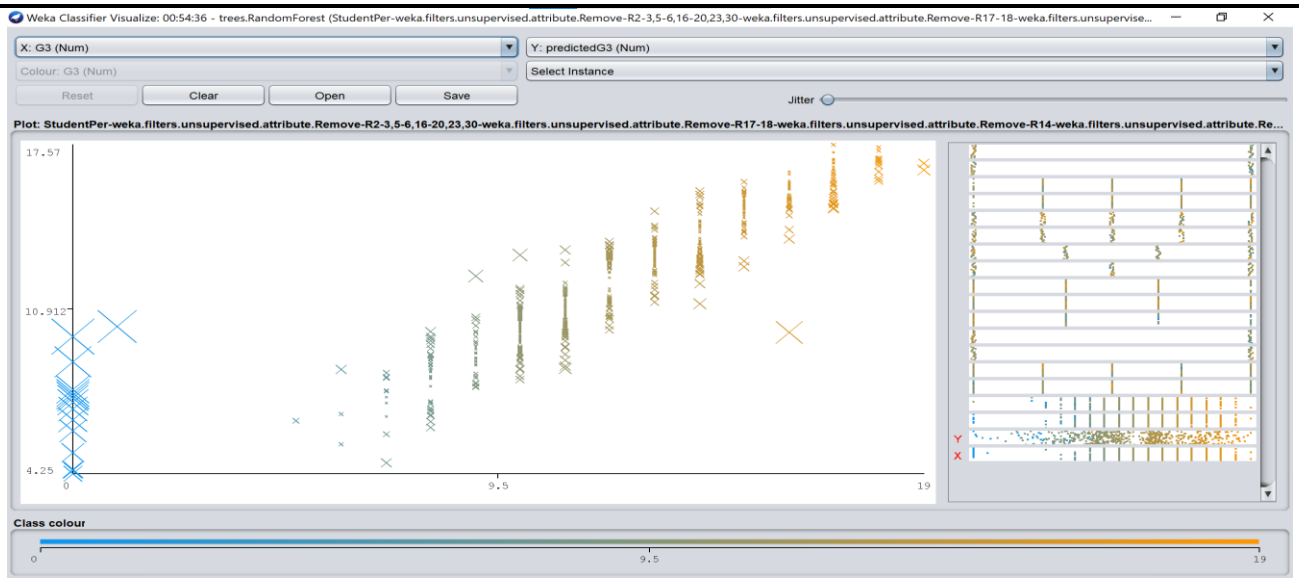

Figure 15: Visualization of classifier errors (*649 instances, Cross-validation*)

6. Discussion

Getting back to the students' dataset, the analysis of the dataset revealed that students' performance (G3 class) is primarily affected by 17 attributes (Table 5). In all cases, 70 percent of all data are used to train the model, while 30 percent are used to test it. For linear regression, the RMSE is 1.2296 in training set mode, 1.4845 in percentage split mode, and 1.2935 in cross-validation (Table 6). Also, for Random Forest, the average RMSE of 0.5107 in the training set, 1.5589 in the percentage split mode, and 1.448 in cross-validation (Table 7). Therefore, Linear regression model performed worse than Random Forest model according to the value of RMSE in the training set mode but did not outperform it in other modes (Percentage split, Cross-validation modes). According to [16] Among all other methods, Random Forests outperformed, but in this study the experiments approved that Random Forest is not outperformed in all models. Actually, the used datasets and its attributes have the main role in the performance of the algorithms.

7. Conclusion

In order to help the educational staff to provide better solutions in helping students to improve their performance, data mining techniques were used to achieve this goal. In this context, previous studies aimed at predicting student performance and traits that influence their performance were extensively presented in this research. In the course of data preprocessing, the selected data has been worked on and configured to work correctly with WEKA. Then, in order to better understand the data, data attributes and data format were introduced. Then the characteristics that affect the student's performance and predict the student's academic performance on the other hand were identified. Finally, the Random Forest algorithm showed superiority over the linear regression algorithm in predicting values only in the training set mode with the used dataset.

References

[1] C. Romero and S. Ventura, "Educational data mining: a review of the state of the art," *Ieee Trans. Syst. Man, Cybern. C Appl. Rev.*, vol. 40, no. 6, pp. 601–618, 2010.

[2] A. Dutt, M. A. Ismail, and T. Herawan, "A Systematic Review on Educational Data Mining," *IEEE Access*, vol. 5, pp. 15991–16005, 2017, doi: 10.1109/ACCESS.2017.2654247.

[3] S. Khan, "Study Factors for Student Performance Applying Data Mining Regression Model Approach," *IJCSNS Int. J. Comput. Sci. Netw. Secur.*, vol. 21, no. 2, pp. 188–192, 2021.

[4] H. Pallathadka, A. Wenda, E. Ramirez-Asís, M. Asís-López, J. Flores-Albornoz, and K. Phasinam, "Classification and prediction of student performance data using various machine learning algorithms," *Mater. Today Proc.*, no. xxxx, 2021, doi:

10.1016/j.matpr.2021.07.382.

[5] K. Nahar, B. I. Shova, T. Ria, H. B. Rashid, and A. H. M. S. Islam, "Mining educational data to predict students performance: A comparative study of data mining techniques," *Educ. Inf. Technol.*, vol. 26, no. 5, pp. 6051–6067, 2021, doi: 10.1007/s10639-021-10575-3.

[6] H. A. Mengash, "Using data mining techniques to predict student performance to support decision making in university admission systems," *IEEE Access*, vol. 8, pp. 55462–55470, 2020, doi: 10.1109/ACCESS.2020.2981905.

[7] J. L. Harvey and S. A. P. Kumar, "A Practical Model for Educators to Predict Student Performance in K-12 Education using Machine Learning," *2019 IEEE Symp. Ser. Comput. Intell. SSCI 2019*, pp. 3004–3011, 2019, doi: 10.1109/SSCI44817.2019.9003147.

[8] A. Khan, S. K. Ghosh, D. Ghosh, and S. Chattopadhyay, "Random wheel: An algorithm for early classification of student performance with confidence," *Eng. Appl. Artif. Intell.*, vol. 102, no. May, p. 104270, 2021, doi: 10.1016/j.engappai.2021.104270.

[9] M. Tsiakmaki, G. Kostopoulos, S. Kotsiantis, and O. Ragos, *Fuzzy-based active learning for predicting student academic performance using autoML: a step-wise approach*, no. 0123456789. Springer US, 2021.

[10] P. Dabhade, R. Agarwal, K. P. Alameen, A. T. Fathima, R. Sridharan, and G. Gopakumar, "Educational data mining for predicting students' academic performance using machine learning algorithms," *Mater. Today Proc.*, vol. 47, pp. 5260–5267, 2021, doi: 10.1016/j.matpr.2021.05.646.

[11] D. Aggarwal, S. Mittal, and V. Bali, "Significance of Non-Academic Parameters for Predicting Student Performance Using Ensemble Learning Techniques," *Int. J. Syst. Dyn. Appl.*, vol. 10, no. 3, pp. 38–49, 2021, doi: 10.4018/ijsda.2021070103.

[12] "UCI." https://archive.ics.uci.edu/ml/datasets/Student+Performance.

[13] S. Umadevi and K. S. J. Marseline, "A survey on data mining classification algorithms," *Proc. IEEE Int. Conf. Signal Process. Commun. ICSPC 2017*, vol. 2018-Janua, no. July, pp. 264–268, 2018, doi: 10.1109/CSPC.2017.8305851.

[14] P. Cortez and A. Silva, "Using data mining to predict secondary school student performance," *15th Eur. Concurr. Eng. Conf. 2008, ECEC 2008 - 5th Futur. Bus. Technol. Conf. FUBUTEC 2008*, pp. 5–12, 2008.

[15] S. Wang and L. Zhang, "A Supervised Correlation Coefficient Method : Detection of Different Correlation," *12th Int. Conf. Adv. Comput. Intell.*, pp. 19–22, 2020.

[16] W. ALmadhoun, "Predictive modelling of student academic performance – the case of higher education in Middle East," University of East London, 2020.

[17] B. Sravani, "Prediction of Student Performance Using Linear Regression," in *International Conference for Emerging Technology*, 2020, pp. 1–5.