Journal of Engineering and Applied Sciences 14 (22): 8254-8260, 2019
ISSN: 1816-949X

Prediction of Student's Academic Performance using k-Means Clustering and Multiple Linear Regressions

[1]Oladele Tinuke Omolewa, [1]Aro Taye Oladele, [2]Adegun Adekanmi Adeyinka and
[2]Ogundokun Roseline Oluwaseun
[1]Department of Computer Science, University of Ilorin, Ilorin, Nigeria
[2]Department of Computer Science, Landmark University, Omu-Aran, Kwara State, Nigeria
ogundokun.roseline@lmu.edu.ng, +2347036261504

Abstract: In today's educational system, performances of students are mainly based on tests, assignments, attendance, quizzes and final examination. It is at the end of this exercise that a minimum mark is determined on which promotion will be based. There is need to identify factors that lead to a student's success or failure. This will allow the teacher to provide appropriate counselling and focus more on such factors. Hence, a model for forecasting student's performance academically is of a pronounced significance, therefore, data mining techniques in classifying and forecasting the academic performance of students was put into application in this research study. k-means clustering and Multiple Linear Regression (MLR) were used for assessing student's performance. The results showed that student's test scores, quiz and assignment were the major factors that could be used in predicting academic performance of students. Also, two clusters were derived with the use of elbow method to group all the students into clusters.

Key words: Data mining, k-means, cluster, multiple linear regression, academic performance, academic performance

INTRODUCTION

Educational Data Mining (EDM) is fast becoming a fascinating research area which allows researchers to find useful, previously unknown patterns from the educational database for better understanding (Thakar *et al*., 2015), improved educational performance and assessment of the student learning process (Verma *et al*., 2016). EDM is pertained with the development and modelling approaches that discovers knowledge from data originating from educational environments (Chaudhari *et al*., 2017). The benefits of teaching and learning has made the forecasting of student's academic performance as important research for, so long (Yassein *et al*., 2017). Tutors could have utilised the predicted outcomes to identify the variables that enable students perform excellent, averagely or fairly in the lecture rooms, so as to make the lecturers to be proactive. There are numerous data like courses offered by students, the academic performance (Ogundokun *et al*., 2018) and so on that were collected and gathered into the repositories over time are being maintained by educational institution (Pandeeswari and Rajeswari, 2014). Enormous amount of data about students (Marion *et al*., 2019) been collected by institutions remains unutilized and doesn't assist in policy and decision making to improve the student's performance academically. If factors for low rate of student's performance can be detected at the early stage by the universities system, then, the knowledge deduced cab help in taking proactive actions, so as to help improve the students of any student in such case.

This study employs techniques in data mining for student's performance. Two algorithms: k-means clustering and multiple linear regressions were used to predict student's academic performance.

Litrature review: A few researchers have worked in the implementation of data mining to EDM. Patil *et al*. (2017) developed a predictive system with the help of data mining techniques for academic success of students in term of rankings and quitter for a subject. Various classification data mining techniques such as Naive Bayes, LibSVM, C4.5, Random Forest and ID 3 were

Corresponding Author: Ogundokun Roseline Oluwaseun, Department of Computer Science, Landmark University Omu Aran, Kwara State, Nigeria, ogundokun.roseline@lmu.edu.ng, +2347036261504

compared and by overcoming the flaws of existing techniques. Based on the rules obtained from the developed technique, the system derived the key factors influencing student performance.

Asif *et al.* (2017) used data mining techniques for predicting the student's graduation performance in final year at university using only pre-university marks and examination marks of early years at university, no socio-economic or demographic features. The result showed a reasonable accuracy for the prediction of the graduation performance in a 4 years university program using only pre-university marks and marks of first and second year courses, no socio-economic or demographic features. The model makes the implementation of a performance support system in a university simpler because from an administrative point of view it is easier to gather marks of students than their socio-economic data. The result also, revealed that decision trees could also be used to identify the courses that act as indicator of low performance. By identifying these courses, warning can be given to students earlier in the degree program.

Ahmad *et al.* (2015) presented a framework for predicting student's academic performance for first year bachelor students in computer science course. The data were gathered from July 2006/2007 until July 2013/2014 which contained the student's demographics, previous academic records and family background information. Decision Tree, Naive Bayes and Rule Based classification techniques were applied to the student's data in order to give the best student's academic performance prediction model. Experimental result showed the rule based out performed other techniques by recording the highest accuracy value of 71.3%.

MATERIALS AND METHODS

Normalized data consist of data in rows and column. There are nine columns, respectively. The nine columns are quality of family relationship, health status, attendance, age, medu, fedu, first grade, second grade and final grade. The last column (final grade) serves as the Y dependent variable for multiple linear regression analysis. Nine initial k-mean clusters were selected randomly to test k-mean clustering on all 395 rows of data to group them into classes. The result of k-mean analysis showed a plotted graph displaying 2 cluster set from the student data sets. After performing k-means clustering on the nine variables the output gotten was not well clustered, so, 2 clusters were later used. Multiple linear regression analysis was also used to get the final predictor model. Scikit-learn library was used to solve the multiple linear regression based on the aforementioned data. The coefficients and intercept were found from the trained set and the corresponding R^2 value.

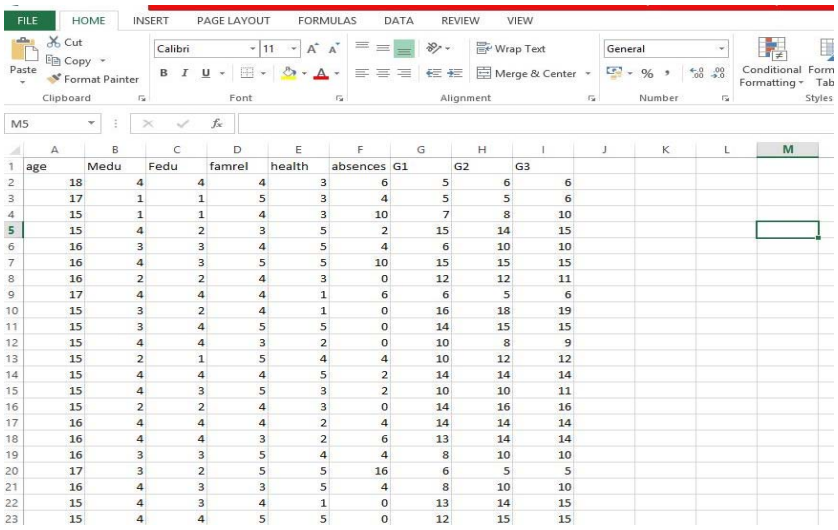
k-means: k-mean clustering technique is a technique of clustering which is widely used. This algorithm is the most popular clustering tool that is used in scientific and industrial applications. The main idea of using k-means is to define k centroids, one for each cluster. These centroids should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid.

When no point is pending, the first step is completed and an early group age is done. At this the centre of the clusters resulting from the previous step need to be updated. After these k new centroids are obtained a new binding has to be done between the same data set points and the nearest new centroid. A loop has been generated. As a result of this loop we may notice that the k centroids change their location step by step until no more changes are done. In other words, centroids do not move any more. Finally, this algorithm aims at minimizing an objective function in this case a squared error function. The objective function:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2 \quad (1)$$

where, $\|x_i^{(j)}-c_j\|^2$ is a chosen distance measure between a data point $x_i^{(j)}$ and the cluster centre is an indicator of the distance of n data points from their respective clustering centres. The basic algorithm for k-means is composed of the following steps:

- Place k points into the space represented by the objects that are being clustered. These points represent initial group centroids
- Assign each object to the group that has the closest centroid
- When all objects have been assigned, recalculate the positions of the k centroids
- Repeat steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated



	age	Medu	Fedu	famrel	health	absences	G1	G2	G3
2	18	4	4	4	3	6	5	6	6
3	17	1	1	5	3	4	5	5	6
4	15	1	1	4	3	10	7	8	10
5	15	4	2	3	5	2	15	14	15
6	16	3	3	4	5	4	6	10	10
7	16	4	3	5	5	10	15	15	15
8	16	2	2	4	3	0	12	12	11
9	17	4	4	4	1	6	6	5	6
10	15	3	2	4	1	0	16	18	19
11	15	3	4	5	5	0	14	15	15
12	15	4	4	3	2	0	10	8	9
13	15	2	1	5	4	4	10	12	12
14	15	4	4	4	5	2	14	14	14
15	15	4	3	5	3	2	10	10	11
16	15	2	2	4	3	0	14	16	16
17	16	4	4	4	2	4	14	14	14
18	16	4	4	3	2	6	13	14	14
19	16	3	3	5	4	4	8	10	10
20	17	3	2	5	5	16	6	5	5
21	16	4	3	3	5	4	8	10	10
22	15	4	3	4	1	0	13	14	15
23	15	4	4	5	5	0	12	15	15

Fig. 1: Overview of proposed data attributes

Multiple linear regressions: Linear regression is an excellent, simple method for numeric prediction and it has been widely used in statistical applications for decades. Multiple linear regression is applicable to numerous data mining situations. Examples are: predicting customer activity on credit cards from demographics and historical activity patterns, predicting the time to failure of equipment based on utilization and environment conditions, predicting expenditures on vacation travel based on historical frequent flier data, predicting staffing requirements at help desks based on historical data and product and sales information, predicting sales from cross selling of products from historical information and predicting the impact of discounts on sales in retail outlets. Multiple linear regression equation to be used for this study is as follows:

$$\hat{Y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + , \ldots, \ \beta_i x_i + \varepsilon \qquad (2)$$

Where:

$\hat{Y}$ = Independent variable (Final grade in Mathematics)

β_0 = Mean value of Y when all independent variables (X) are zero

β_i = Regression Co-efficient of X_i

ε = Residual which is difference between $Y-\hat{Y}$

The term 'linear' is used because in multiple linear regressions, we assume that y is directly related to a linear combination of the explanatory variables.

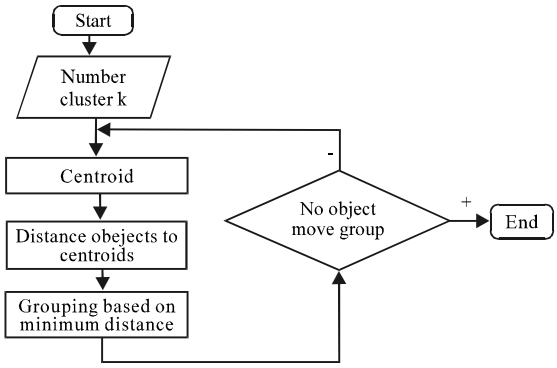


Fig. 2: k-means flowchart

Data collection: Data was collected from UCI (University of California, Irvine) machine learning repository. Dataset is based on the performance of student in mathematics. Data attributes from raw data includes student grades demographic, social and school related features.

Data model: The dataset was modelled under binary/five-level classification and regression tasks as shown in Fig. 1. The attribute are: the target attribute G3 has a strong correlation with attributes G2 and G1. This occurs because G3 is the final year grade (issued at the 3rd period) while G1 and G2 correspond to the 1st and 2nd period grades. It is more difficult to predict G3 without G2 and G1 but such prediction is much more useful.

Data attributes: The following data attributes were selected from the dataset for the test system:

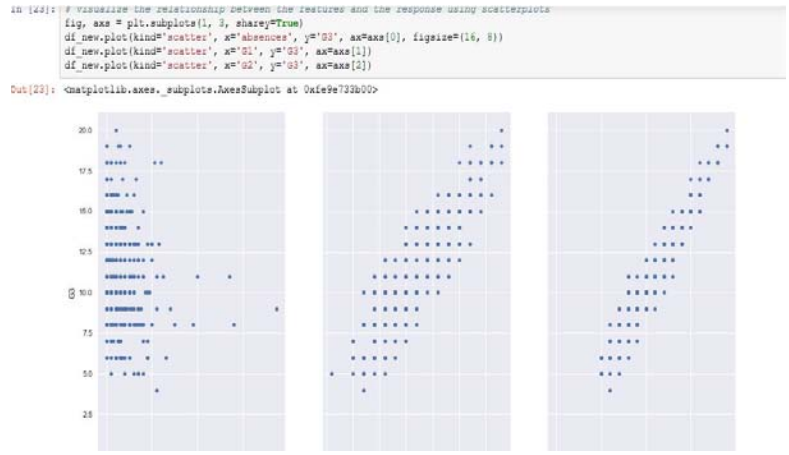
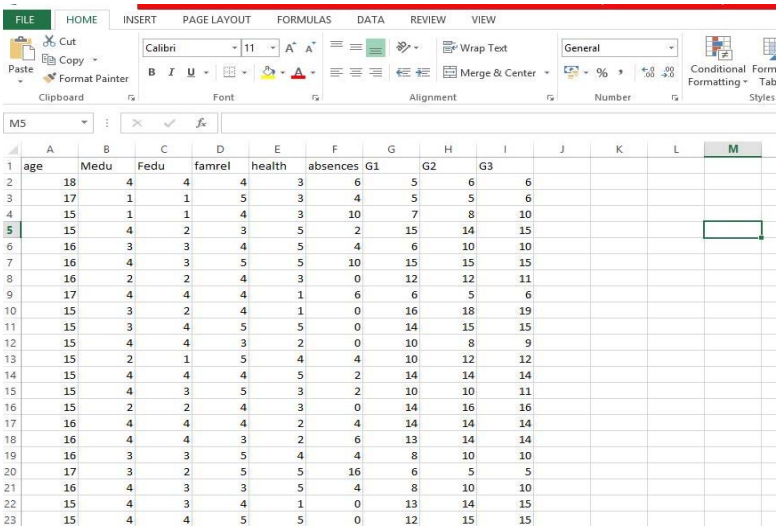


Fig. 3: Screen shot of scatter plot of predictors vs. target variable



	A	B	C	D	E	F	G	H	I
1	age	Medu	Fedu	famrel	health	absences	G1	G2	G3
2	18	4	4	4	3	6	5	6	6
3	17	1	1	5	3	4	5	5	6
4	15	1	1	4	3	10	7	8	10
5	15	4	2	3	5	2	15	14	15
6	16	3	3	4	5	4	6	10	10
7	16	4	3	5	5	10	15	15	15
8	16	2	2	4	3	0	12	12	11
9	17	4	4	4	1	6	6	5	6
10	15	3	2	4	1	0	16	18	19
11	15	3	4	5	5	0	14	15	15
12	15	4	4	3	2	0	10	8	9
13	15	2	1	5	4	4	10	12	12
14	15	4	4	4	5	2	14	14	14
15	15	4	3	5	3	2	10	10	11
16	15	2	2	4	3	0	14	16	16
17	16	4	4	4	2	4	14	14	14
18	16	4	4	3	2	6	13	14	14
19	16	3	3	5	4	4	8	10	10
20	17	3	2	5	5	16	6	5	5
21	16	4	3	3	5	4	8	10	10
22	15	4	3	4	1	0	13	14	15
23	15	4	4	5	5	0	12	15	15

Fig. 4: Normalized data

- Age
- Father's educational status (Fedu)
- Quality of family relationship (Femrel)
- Health status (Health)
- Mothers educational status (Medu)
- School attendance (Absences)
- Test Grade 1 (G1)
- Test Grade 2 (G2)
- Final year Grade (G3)

Flow chart for k-mean algorithm: Figure 2 represents the k-means flowchart which explains how data clustering is being carried out, it shows basic procedure involved in k-means. While the screen shot of scatter plot of predictors vs. target variable is shown in Fig. 3.

Cluster and prediction: k-means and multiple linear regression algorithms were implemented using python programming language. Python packages/libraries like scikit-learn, pandas and numpy were crucial in solving both k-means and multiple linear regressions. The experiment was performed to implement k-mean and multiple linear regressions for classifying and predicting student performance as shown in Fig. 4. The output in Fig. 4 contains the normalized data showing various selected variable to design the model. Figure 5 mentions the summary of the captured variables. The count function counts all the data set that was used and returns an output of 395 data sets were available. It continued with other functions such as mean, standard

	G1	G2	AGE	MEDU	FEDU	FAMREL	HEALTH	ABSENCES	G3
count	395.000000	395.000000	395.000000	395.000000	395.000000	395.000000	395.000000	395.000000	395.000000
mean	10.908861	10.713924	16.696203	2.749367	2.521519	3.944304	3.554430	5.708861	10.415190
std	3.319195	3.761505	1.276043	1.094735	1.088201	0.896659	1.390303	8.003096	4.581443
min	3.000000	0.000000	15.000000	0.000000	0.000000	1.000000	1.000000	0.000000	0.000000
25%	8.000000	9.000000	16.000000	2.000000	2.000000	4.000000	3.000000	0.000000	8.000000
50%	11.000000	11.000000	17.000000	3.000000	2.000000	4.000000	4.000000	4.000000	11.000000
75%	13.000000	13.000000	18.000000	4.000000	3.000000	5.000000	5.000000	8.000000	14.000000
max	19.000000	19.000000	22.000000	4.000000	4.000000	5.000000	5.000000	75.000000	20.000000

Fig. 5: Summary of numerical fiels

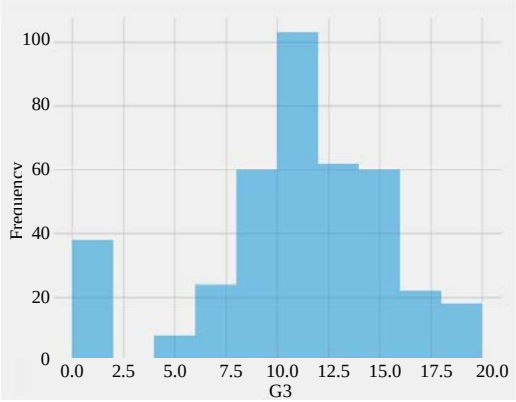


Fig. 6: Histogram representation of frequency variable against G3

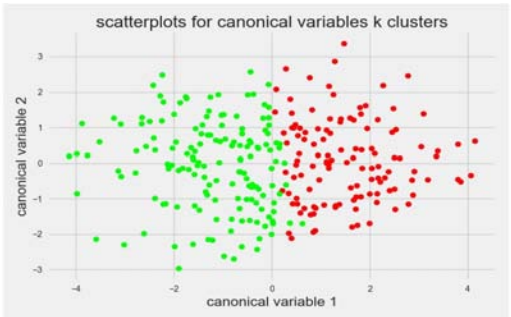


Fig. 7: k-means clustering output (scatterplots for canonical variables k clusters)

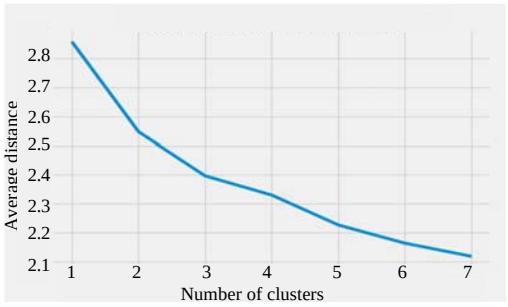


Fig. 8: Output of determining the number of clusters using 'elbow method' (selecting k with the elbow method)

deviation, minimum, maximum value and so on. This summary was performed on all the selected variables. While Fig. 6 shows the histogram representation of frequency variable against G3, all values of G3 are distributed in a histogram.

k-means analysis result: Figure 7 shows result of k-means analysis. k-means being an unsupervised learning algorithm. Figure 8 shows the accurate and efficient output of determining the number of clusters. The method is called 'elbow method'. The elbow curve falls at 2 which deduced indicates that clustering the data set into 2 clusters give the best clustering output.

The output in Fig. 7 displays two unique clusters of all the student data set. The first cluster presented

its result using the green colour code and the second cluster presented its result using red colour code.

RESULTS AND DISCUSSION

The section discusses results obtained for the developed prediction of student's academic performance using k-means clustering and multiple linear regression. The following subsections give the detail results (Fig. 9).

Regression analysis result: The result shows the correlated values for all the selected variables as shown in Fig. 10. It can be deduced that variable G1 and G2 have a very strong correlation value because it falls between the range of 0.5-1. Other variables does not fall in between this range, thus, making variable G1 and G2 more efficient and accurate to perform prediction while variable G3 is perfectly correlated which equals 1. The output of Fig. 11 shows the regression line for G2 and it shows that there is minimum number of outliers.

Result of regression analysis coefficient: The output in Fig. 12 shows the intercept which is denoted as B_0 = -1.8300, B_1, X_1 where, B_1 = 0.1532 X_1 is chosen by the user as well as X_2. This model can be used to perform prediction and this model can be tested by inserting values from the training set and there after checking, if it

```
Out[9]: G1          0.801468
        G2          0.904868
        AGE        -0.161579
        MEDU        0.217147
        FEDU        0.152457
        FAMREL      0.051363
        HEALTH     -0.061335
        ABSENCES    0.034247
        G3          1.000000
        Name: G3, dtype: float64
```

Fig. 9: Display all correlation values for the above variables

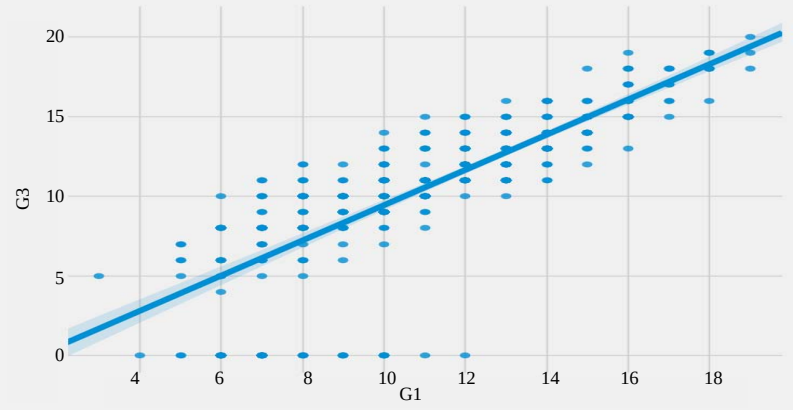


Fig. 10: Scatter plots showing regression line

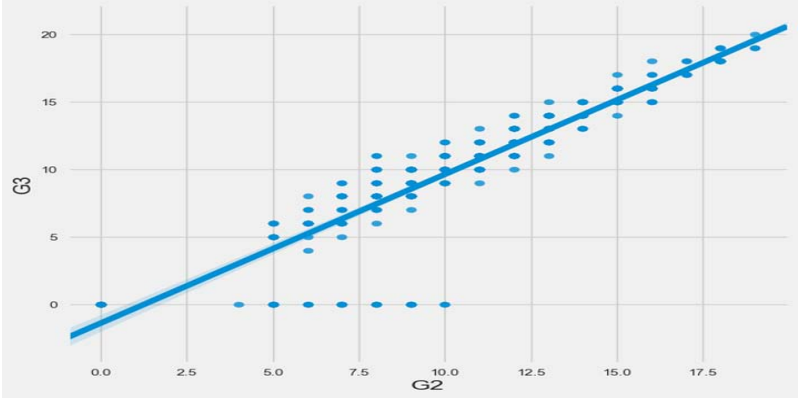


Fig. 11: Scatter plots showing regression line

Intercept: -1.83001214058

Co-efficient for X1 - G1: 0.15326858528068094

Co-efficient for X2 - G2: 0.9868668387417142

Y (G3) = -1.83001214058 + 0.15326858528068094X1 + 0.9868668387417142X2

Fig. 12: Multiple linear regression model

tally with the target variable. For the multiple linear regression the values of variables that was obtained are stated as follows:

- Intercept: -1.83001214058
- Co-efficient for X1-G1: 0.153268585281
- Co-efficient for X2-G2: 0.986866838742
- Y (G3) = -1.83001214058+0.153268585281X1+ 0.986866838742X2
- R^2 = 0.822

The above result shows the intercept B_0 = -1.830012, B_1 = 0.153268 and B_2 = 0.98686. The R^2 = 0.822 which shows the model fits well and the error gotten is minimal R^2 values are between 0-100.

CONCLUSION

In this study, a model was developed to make prediction of student's performance using multiple linear regressions. Clustering of data was also achieved using k-means clustering. Online data was collected from UCI (University of California, Irvine) machine learning repository. When k-means clustering was performed, two distinct clusters were obtained. These clusters helped in grouping all the three-hundred and ninety-five students. R^2 for this study was 0.822 showing that the final model fit well into the given data set. Also, the correlation test indicates that variable G1 and G2 played a vital role in designing this model. The flask framework which serves as the GUI (Graphic User Interface) was created to serve as the interface between the user and the model.

REFERENCES

Ahmad, F., N.H. Ismail and A.A. Aziz, 2015. The prediction of students academic performance using classification data mining techniques. Appl. Math. Sci., 9: 6415-6426.

Asif, R., S. Hina and S.I. Haque, 2017. Predicting student academic performance using data mining methods. Intl. J. Comput. Sci. Netw. Secur., 17: 187-191.

Chaudhari, K.P., R.A. Sharma, S.S. Jha and R.J. Bari, 2017. Student performance prediction system using data mining approach. Intl. J. Adv. Res. Comput. Commun. Eng., 6: 833-839.

Marion, O.A., O.O. Florence, O.A. Daramola, O.O. Roseline and A.A. Emmanuel, 2019. RFID-based human tracking system in tertiary institution. J. Eng. Applied Sci., 14: 2345-2351.

Ogundokun, R.O., M.O. Adebiyi, O.C. Abikoye, T.O. Oladele and A.F. Lukman *et al*., 2018. Performance evaluation: Dataset on the scholastic performance of students in 12 programmes from a private university in the South-West geopolitical zone in Nigeria. F1000 Res., 8: 1-7.

Pandeeswari, L. and K. Rajeswari, 2014. Student academic performance using data mining techniques. Intl. J. Comput. Sci. Mob. Comput., 3: 726-731.

Patil, V., S. Suryawanshi, M. Saner, V. Patil and B. Sarode, 2017. Student performance prediction using classification data mining techniques. Intl. J. Sci. Dev. Res., 2: 163-167.

Thakar, P., A. Mehta and Manisha, 2015. Performance analysis and prediction in educational data mining: A research travelogue. Intl. J. Comput. Appl., 110: 60-68.

Verma, K., A. Singh and P. Verma, 2016. A review on predicting student performance using data mining method. Intl. J. Curr. Eng. Sci. Res., 3: 127-132.

Yassein, N.A., R.G.M. Helali and S.B. Mohomad, 2017. Predicting student academic performance in KSA using data mining techniques. J. Inf. Technol. Software Eng., 7: 1-5.