Effect of Non-Academic Parameters on Student's Performance.

Shantanu Lokhande
Department Of Information Technology
G.H. Raisoni College Of Engineering
Nagpur, Maharashtra
shantanulokhande5501@gmail.com

Vedant Bahel
Department Of Information Technology
G.H. Raisoni College Of Engineering
Nagpur, Maharashtra
vbahel@ieee.org

***Abstract*—With the exponential growth in today's technology and its expanding areas of application it has become vital to incorporate it in education. One such application is Knowledge Discovery in Databases (KDD) which is a subset of data mining. KDD deals with extracting useful information and meaningful patterns from the database that were not known before. This study is a detailed application of KDD and focuses on analyzing why a particular set of students performed better than others and what factors influenced the results. The study is conducted on a dataset of 480 students and across 16 different features. The authors implemented 4 major classification techniques namely Logistic Regression, Decision Tree, Random Forest and XGB classifier. Obtaining the key features from the top performing ML algorithms that have a major impact on the performance of the student, the study takes these features as a baseline for further analysis. Further data analysis highlights patterns in the data. The study concludes that there are a lot of non-academic factors that influence the overall performance of a student and should be taken into consideration by universities and other relevant bodies.**

***Index Terms*—Learning Analytics, Education, Educational Data Mining, Pattern Recognition, Data Visualization.**

I. Introduction

Educational Data Mining (EDM) is a small but significant subset in the trend of Data Mining and Knowledge Discovery in Databases. EDM concentrates on the development of techniques to exploit the data from educational databases. The main aim is to detect useful patterns from student databases. These databases contain information such as academic performance, gender, financial conditions, etc. Previously unknown patterns, relationships, mathematical algorithms, and statistical models are generated from the data for implementation in the educational system for the overall betterment of the student. As the dataset is specific to the educational area and thereby having intrinsic semantic information, relationships with other data, and multiple levels of meaningful hierarchy [2]. Researchers in this field focus on discovering useful knowledge either to help the educational institutes manage their students better or to help students to manage their education and deliver better and enhance their performance [1]. The target outcome of EDM can be roughly classified into 4 major categories namely; improving student performance models, improving domain models, studying and strengthening the pedagogical support provided by the learning software or mentor, and scientific research into learning and teaching [2]. A lot of work has been done regarding EDM on web-based learning and distance education and recent trends are more focused on intelligent platforms that evaluate and guide students on what to learn based on their interests [3]. The primary goal has always been to equip students with the knowledge and skills needed to transition into successful areas within a specific period and EDM helps achieve that [5]. A similar approach of ensemble methods was implemented by the author in [4] that used Random Forest, Artificial Neural Networks, and Naïve Bayes. In [6] the author used a wrapper-based feature selection method called Boruta.

II. Related Work

In the past years, researchers have tried to implement the developed algorithms for EDM purposes. Over time, these algorithms and techniques have evolved for better performance. In a paper published in 2003, the authors [8] showed considered factors affecting students' dropout rate. These factors are conditions related to the students before admission, factors related to the students during the study periods in the university, and all factors including the target value to be predicted for factor analysis. The authors used a tree-based classification algorithm, J48 or C4.5, and Naïve Bayes to analyze the data. A study conducted by Ibtissem Daoudi Et al. [15] in 2021 using Crisis Management Serious Games (CMSG) has shown its potential for teaching people both technical and soft skills related to managing crises in a safe environment while reducing training costs. In summary, various researches [9] [10] [12] investigated to solve the educational problems using data mining techniques. However, very few researches shed light on student's behavior during the learning process and its impact on the student's academic success [16].

III. Dataset

The dataset in this paper has been taken from Kaggle [17] [18] which is a huge repository of datasets that is available for training machine learning algorithms. The database features can be roughly divided into 3 categories that are : (i) Demographic features such as gender and nationality,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 August 2021 doi:10.20944/preprints202108.0256.v1

(ii) Academic background features such as educational stage, grade level, section, etc, and (iii) Behavioral features such as raised hands, visited resources, answering survey by parents and school satisfaction. In this paper, the main analysis will be focused on the "class" feature of the dataset. Apart from that, various parameters will be analyzed with respect to gender and factors influencing the "class" i.e the overall performance of the student.

The first step in pre-processing the dataset and preparing it for analysis is checking for null values (Fig.1) and then converting the dataset into a machine-readable format as the algorithms won't be able to generate results from non-numerical data and will fail to converge.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 480 entries, 0 to 479
Data columns (total 17 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   gender                    480 non-null    object
 1   NationalITy               480 non-null    object
 2   PlaceofBirth              480 non-null    object
 3   StageID                   480 non-null    object
 4   GradeID                   480 non-null    object
 5   SectionID                 480 non-null    object
 6   Topic                     480 non-null    object
 7   Semester                  480 non-null    object
 8   Relation                  480 non-null    object
 9   raisedhands               480 non-null    int64
 10  VisITedResources          480 non-null    int64
 11  AnnouncementsView         480 non-null    int64
 12  Discussion                480 non-null    int64
 13  ParentAnsweringSurvey     480 non-null    object
 14  ParentschoolSatisfaction  480 non-null    object
 15  StudentAbsenceDays        480 non-null    object
 16  Class                     480 non-null    object
dtypes: int64(4), object(13)
memory usage: 63.9+ KB
```

Fig. 1. A table showing null values in the dataset. Gives an overall picture of the dataset and shows that there are no null values to work on.

The data has been converted into a working format using various encoding techniques such as binary encoding, ordinal encoding, and one hot encoding. For such encodings to work we first had to select the features that best fit these categories for encoding. The features that had non-numeric values were selected and broadly classified into 3 categories. The first category was binary that contained the features whose value varied into 2 values like "gender" (M/F) or "semester" (S/F). The second category was ordinal that contained the features in which order of the data mattered like "StageID" or "GradeID". The third and final category was nominal that contained nominal features in which there are more than 2 values but the order does not matter like "Nationality", "PlaceofBirth" etc. And the target column was selected as "class" (Fig.2).

```
binary_features = ['gender', 'Semester', 'Relation', 'ParentAnsweringSurvey', 'ParentschoolSatisfaction', 'StudentAbsenceDays']

ordinal_features = ['StageID', 'GradeID']

nominal_features = ['NationalITy', 'PlaceofBirth', 'SectionID', 'Topic']


target_column = 'Class'
```

Fig. 2. (This figure shows a division of features for encoding. Binary features are for 0,1 features, ordinal for order of important features and nominal for more than one variation in the feature.)

The encoding functions used are shown in Fig.3 and some sample encoding results are shown in Fig.4. In Fig.4 "gender" feature was encoded according to the binary encoding, similarly, "StageID" and "GradeID" were encoded according to ordinal encoding.

```
def binary_encode(df, column, positive_value):
    df = df.copy()
    df[column] = df[column].apply(lambda x: 1 if x == positive_value else 0)
    return df

def ordinal_encode(df, column, ordering):
    df = df.copy()
    df[column] = df[column].apply(lambda x: ordering.index(x))
    return df

def onehot_encode(df, column, prefix):
    df = df.copy()
    dummies = pd.get_dummies(df[column], prefix=prefix)
    df = pd.concat([df, dummies], axis=1)
    df = df.drop(column, axis=1)
    return df
```

Fig. 3. A code for encoding features according to different classes i.e binary, ordinal and one-hot encoding

	gender	StageID	GradeID	Semester	Relation
0	1	0	1	0	1
1	1	0	1	0	1
2	1	0	1	0	1
3	1	0	1	0	1
4	1	0	1	0	1
...	...	...	...	...	...
475	0	1	5	1	1
476	0	1	5	0	1
477	0	1	5	1	1
478	0	1	5	0	1
479	0	1	5	1	1

480 rows × 56 columns

Fig. 4. Sample encoding results. Shows how the binary, ordinal and one hot encoding changed the values of the features.

After proper preparation of the dataset for analysis, it's important to analyze the dependence of features on one another. There are various methods to analyze the interdependence of features but the most widely used is plotting a heat map. Fig.5 is a heat map that shows the relationship between various features and makes it better to visualize features of the dataset. It also gives a broad and basic understanding of the database and highlights the key features for further analysis.

IV. Machine Learning Analysis

There are various techniques available for data mining which are also used in knowledge discovery in databases (KDD) such as classification, clustering, association rule learning, A.I., etc. Classification is one of the most important and widely used data mining techniques. Researchers use and study classification because it is easy to use [1]. This paper is going to focus on 4 classification algorithms namely Logistic Regression, Decision Tree, Random Forest, and XGB Classifier. For evaluating the performance of the algorithms we have used 5 parameters and compared the results of various classification algorithms based on these parameters.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 August 2021 doi:10.20944/preprints202108.0256.v1

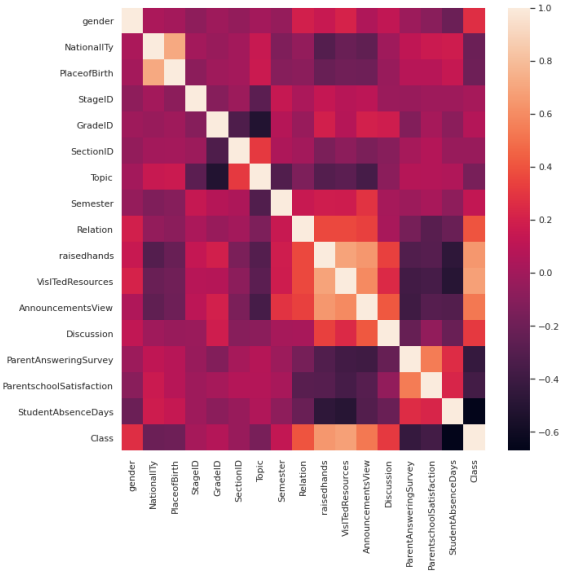

Fig. 5. A heatmap highlighting key features of the dataset. A key tool for visualizing relationships among features and selecting important features

The parameters used are (i) Precision (refers to the fraction of relevant instances among the retrieved instances), (ii) Recall (refers to the fraction of relevant instances that were retrieved), (iii) F-1 Score (refers to the weighted average of precision and recall), (iv) Support (refers to the number of actual occurrences of the class in the specified dataset) and (v) Accuracy (refers to the percentage of correct prediction of test data).

A. Logistic Regression

Logistic regression is a Machine Learning algorithm that is used for classification problems. For the dataset used in this paper, the training set and the test set were from the same database and were divided in a ratio of 70:30 respectively. This criterion remains the same throughout different algorithms. Selecting logistic regression as the base algorithm the results obtained are shown in Fig.6.

```
Results for: Logistic Regression

[[27  9  0]
 [ 7 39 10]
 [ 1 15 36]]
              precision    recall  f1-score   support

           0       0.77      0.75      0.76        36
           1       0.62      0.70      0.66        56
           2       0.78      0.69      0.73        52

    accuracy                           0.71       144
   macro avg       0.72      0.71      0.72       144
weighted avg       0.72      0.71      0.71       144

accuracy is 0.7083333333333334
```

Fig. 6. Logistic Regression results shows that the accuracy of our model is 70.83

B. Decision Tree

A decision tree is a supervised learning technique that can be used for both classification and regression problems but mostly used for solving classification problems. The results for running the decision tree algorithm on our dataset are depicted in Fig.7.

```
Results for: Decision Tree

[[28  5  3]
 [ 6 40 10]
 [ 0 20 32]]
              precision    recall  f1-score   support

           0       0.82      0.78      0.80        36
           1       0.62      0.71      0.66        56
           2       0.71      0.62      0.66        52

    accuracy                           0.69       144
   macro avg       0.72      0.70      0.71       144
weighted avg       0.70      0.69      0.70       144

accuracy is 0.6944444444444444
```

Fig. 7. Decision Tree results shows the accuracy to be around 69.44

C. Random Forest

Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. The results for running the random forest classifier on our dataset are depicted in Fig.8. For the parameters, the random state was chosen to be 52 and the number of estimators was 150.

```
Results for: Random Forest

[[30  6  0]
 [ 4 50  2]
 [ 0 14 38]]
              precision    recall  f1-score   support

           0       0.88      0.83      0.86        36
           1       0.71      0.89      0.79        56
           2       0.95      0.73      0.83        52

    accuracy                           0.82       144
   macro avg       0.85      0.82      0.83       144
weighted avg       0.84      0.82      0.82       144

accuracy is 0.8194444444444444
```

Fig. 8. Random Forest results shows that model performed with an accuracy of 81.94.

D. XGB Classifier

XGB is a decision tree-based ensemble ML algorithm. It works on a gradient boosting framework. For our dataset the parameters used were max-depth=4 ; learning-rate=0.10 ; n-estimators=50 and seed=52. The results are depicted in Fig.9.

E. Results

Table I shows the results (accuracy) of all the algorithms used. From the comparison of accuracies of different algorithms, it is clear that Random Forest and XGB Classifier performed best. The accuracies of Random Forest and XGB are remarkably similar. This gives the pathway for selecting two algorithms for further analysis.

After the generation of accuracies, in order to analyze the key values in the dataset, let's plot the feature importance graph for both the best-performing algorithms and compare them. Fig.10 shows the feature importance for the Random

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 August 2021 doi:10.20944/preprints202108.0256.v1

```
[[31  5  0]
 [ 4 48  4]
 [ 0 13 39]]
              precision    recall  f1-score   support

           0       0.89      0.86      0.87        36
           1       0.73      0.86      0.79        56
           2       0.91      0.75      0.82        52

    accuracy                           0.82       144
   macro avg       0.84      0.82      0.83       144
weighted avg       0.83      0.82      0.82       144

accuracy for XGB is 0.8194444444444444
```

Fig. 9. XGB Classifier results show the model performance to be 81.94. Note that for Random Forest and XGB classifier the value of the confusion matrix (i.e. the upper left hand matrix) is almost same.

TABLE I
COMPARISON BETWEEN PERFORMANCE OF IMPLEMENTED ML MODELS

Algorithm	Accuracy
Logistic Regression	0.708
Decision Tree	0.694
Random Forest	0.819
XGB Classifier	0.819

Forest Classifier and Fig.11 shows the feature importance for the XGB Classifier.

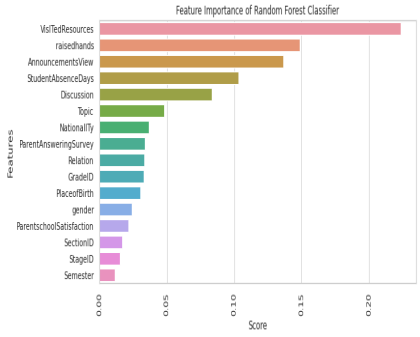

Fig. 10. Feature importance for random forest classifier. Note how the top 5 features have a major impact on the score

Table II shows the comparison between the top 5 features for both algorithms. Most of the features are roughly the same but for better visualization purposes we will take the features that are common in both and understand their impact on the overall grades i.e the "class". The features that will be taken into consideration are: Visited resources, Raised hands, Discussion, and Announcement views.

TABLE II
COMPARISON BETWEEN TOP FEATURES OF RF AND XGB

Ranking	Random Forest	XGB
1	Visited Resources	Visited Resources
2	Announcement View	Raised Hands
3	Raised Hands	Announcement View
4	Discussion	Student Absence Days
5	Topic	Discussion

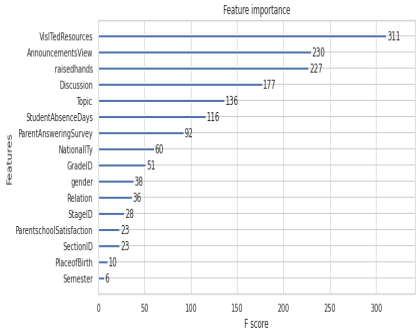

Fig. 11. Feature importance for XGBClassifier. Note how the top 5 features have a major impact on the F- score

V. DATA ANALYSIS

Over the years researchers have used various techniques for data analysis and visualization. There are lots of different methods for visualizing data like swarm pots, bar plots, graphs, heatmaps, etc. As the ML analysis showed the important features that impacted the overall performance i.e. the "class", we are going to keep them as the baseline for our analysis. Also we are going to keep our study focused on class "2" because that's the highest class and consists of top performing students.

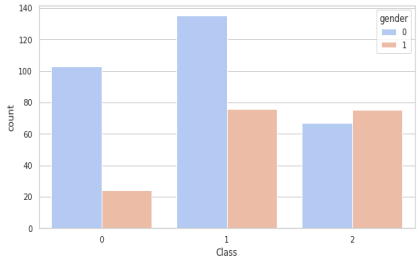

Fig. 12. A gender wise performance analysis plot. Gives an overall performance of the students (0-Male ; 1-Female)

Fig.12 shows a plot of performance gender wise. It gives a clear picture of who performed better. As it is evident from the graph that for classes "0" and "1" male students performed better compared to the female students. But in the final class "2" i.e. the highest one, females performed better than males. This highlights the path for further analysis. We need to analyse the reason behind this result and why female students scored more than male students.

Starting with one of the important features of our dataset i.e. the number of "visited resources" let's have a look at the visualization results.

Fig.13 is a swarm plot that gives a rough picture of the distribution of students class wise and gender wise over the feature of visited resources. As we can see for class "0" most students are concentrated towards the lower part of the graph and for classes "1" and "2" students are concentrated towards the upper part of the graph. When examined closely we can see that for class "2" the concentration of females (depicted

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 August 2021 doi:10.20944/preprints202108.0256.v1

Fig. 13. Comparison on visited resources. Note the concentration of students in the upper bound of the distribution for classes 1 and 2

by orange dots) is high in the upper bound. This plot depicts that if the number of resources visited by the students are more, then their overall performance will increase and they will score a better class.

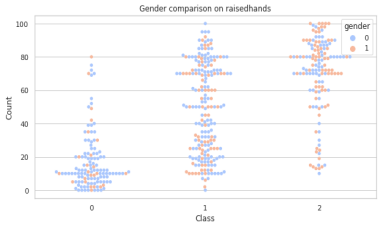

Fig. 14. Comparison on raised hands. Note the even distribution of students across class 1.

Let's now look at another key feature that affects the class of our students i.e. "raised hands". Fig.14 is a swarm plot that shows the distribution of the number of students class wise as well as gender wise relative to raised hands. It is clear from the plot that distribution across class "1" was roughly even but concentration of male students was more compared to female students. But if we look at class "2" the distribution is highly uneven and concentrated more in the upper part of the distribution. Also the frequency of female students is more compared to male students. We noticed a similar relationship in the previous result of visited resources.

Let us study the remaining two key factors that are "Discussions" and "Announcements View" and comment on their results. Fig.15 shows a swarm plot of "Discussions". Fig.16 describes another swarm plot of "Announcement View". Both of these swarms plots show a basic difference between the performance of both the genders. If views both display similar features i.e. more frequency of females in the higher class. But there is a high similarity between class "2" of both the features.

An interesting observation on the results "Discussions" and "Announcements View" is that, both of them have roughly the same frequency of male and female students in class "2" as compared to the previous results. In the case of "raised hands" and "visited resources" the frequency of distribution of female students was fairly more compared to the male students in class "2". Apart from that the pattern of distribution is roughly the same and that makes it an unique observation. Let's look at these features from another perspective of visualization.

Fig. 15. Comparison of discussions. Note the high concentration of students scoring grade 0 and are present in the lower bound of the plot

Fig. 16. Comparison of Announcements View. Notice the frequency of female students in class 2

Fig.17 and Fig.18 display a joint plot that shows the distribution via contour maps as well as frequency distribution across the range. Fig.17 refers to the "Announcement View" feature and Fig.18 refers to the "Discussion" feature. If we examine the "Announcement View" plot we can see that in the horizontal graph, towards the end i.e. from 1.5 - 2.5 the frequency distribution of male and female students overlap. This is the same case for the "Discussions" feature. In the horizontal plot, towards the end, the distribution overlaps for both genders. But when we take a look at the contour map of "Announcement View" we can see that for class "2" the number of female students is more compared to class "1" of the same distribution. Also for the contour map of "Discussions", we can see similar results.

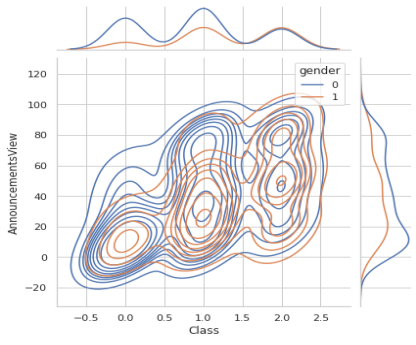

Fig. 17. Joint plot for comparison of announcement views and class. Look at the graph above for a better view. In the late section note how the graphs overlap for male and female students

From all the results obtained by comparing the key features of the dataset it is fair to say that there is an emerging pattern across the features and their impact on the resulting class. This result required further comparison and a more deeper

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 August 2021 doi:10.20944/preprints202108.0256.v1

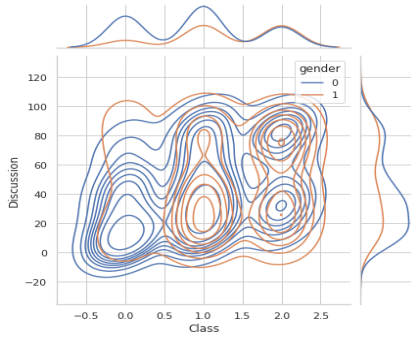

Fig. 18. Joint plot for comparison of discussions and class. Compare the similarities between this plot and Fig.17 and the overlapping of graphs but the differences in the contour maps for class 2.

analysis. Authors conducted a detailed study with all the key features and compared their results among each other. For all the features authors found a common trait that if the student is a high performer then the values of key features were bound to be high. Throughout the study authors saw that the number of females were high in the upper bound of the key features. This resulted in female students performing better and achieving a higher class compared to their male counterparts.

VI. Conclusion

After the completion and a thorough analysis of the study we can conclude that for being able to perform better and score a good, students need to focus on the key aspects of their fields. In the current study it was also found that student's performance also depends on some other non-academic factors such as "student absence days", "parents answering survey", "Nationality" etc. To conclude, the main aim of this study was to focus on various factors that affect students and their performance so that they can improve. Also the study aimed at motivating more research in the field of data mining and finding interesting patterns which could help the students and universities in a constructive way.

Using this database and this study one can find more information and better patterns. A more interesting approach would be to include Convolutional Neural Networks for training and then generation of results. Similarly other A.I. oriented approaches could be used for analysis. Universities can make data mining an integral part of their evaluation scheme which will grant them the ability to make correct decisions in favor of the students.

References

[1] Abu Saa, Amjad. (2016). Educational Data Mining & Students' Performance Prediction. International Journal of Advanced Computer Science and Applications. 7. 10.14569/IJACSA.2016.070531.

[2] Romero, Cristóbal, and Sebastián Ventura. "Educational data mining: a review of the state of the art." IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 40, no. 6 (2010): 601-618.

[3] Bahel, Vedant & Bajaj, Preeti & Thomas, A.. (2019). Knowledge Discovery in Educational Databases in Indian Educational System: A Case Study of GHRCE, Nagpur. 235-239.10.1109/ICCIKE47802.2019.9004421.

[4] Amrieh, Elaf & Hamtini, Thair & Aljarah, Ibrahim. (2016). Mining Educational Data to Predict Student's academic Performance using Ensemble Methods. International Journal of Database Theory and Application. 9. 119-136. 10.14257/ijdta.2016.9.8.13.

[5] Algarni, Abdulmohsen. (2016). Data Mining in Education. International Journal of Advanced Computer Science and Applications. 7. 10.14569/IJACSA.2016.070659.

[6] Shrestha, Sushil & Pokharel, Manish. (2021). Educational data mining in moodle data. International Journal of Informatics and Communication Technology (IJ-ICT). 10. 9. 10.11591/ijict.v10i1.pp9-18.

[7] Khalilia, Hadi & Samar, Thaer & Sleet, Yazeed. (2020). Predicting Students Performance Based on Their Academic Profile.

[8] Yathongchai, Wilairat, Chusak Yathongchai, Kittisak Kerdprasop and Nittaya Kerdprasop. "Factor Analysis with Data Mining Technique in Higher Educational Student Drop Out." (2012).

[9] Bahel, Vedant, and Achamma Thomas. "Text similarity analysis for evaluation of descriptive answers." arXiv preprint arXiv:2105.02935 (2021).

[10] Bahel, Vedant, Shreyas Malewar, and Achamma Thomas. "Student Interest Group Prediction using Clustering Analysis: An EDM approach." In 2021 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE), pp. 481-484. IEEE, 2021.

[11] Merceron, Agathe & Yacef, Kalina. (2005). TADA-Ed for educational data mining. 7.

[12] Baker, Ryan Shaun, and Paul Salvador Inventado. "Educational data mining and learning analytics." In Learning analytics, pp. 61-75. Springer, New York, NY, 2014.

[13] Mining Educational Data to Analyze Students' Performance arXiv:1201.3417 [cs.IR]

[14] Alejandro Peña-Ayala, Educational data mining: A survey and a data mining-based analysis of recent works, Expert Systems with Applications, Volume 41, Issue 4, Part 1, 2014, Pages 1432-1462, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2013.08.042. (https://www.sciencedirect.com/science/article/pii/S0957417413006635)

[15] Ibtissem Daoudi, Raoudha Chebil, Erwan Tranvouez, Wided Lejouad Chaari, Bernard Espinasse, Improving Learners' Assessment and Evaluation in Crisis Management Serious Games: An Emotion-based Educational Data Mining Approach, Entertainment Computing, Volume 38, 2021, 100428, ISSN 1875-9521, https://doi.org/10.1016/j.entcom.2021.100428. (https://www.sciencedirect.com/science/article/pii/S1875952121000252)

[16] Amrieh, Elaf & Hamtini, Thair & Aljarah, Ibrahim. (2015). Preprocessing and Analyzing Educational Data Set Using X-API for Improving Student's Performance. 10.1109/AEECT.2015.7360581.

[17] Amrieh, E. A., Hamtini, T., & Aljarah, I. (2016). Mining Educational Data to Predict Student's academic Performance using Ensemble Methods. International Journal of Database Theory and Application, 9(8), 119-136.

[18] Amrieh, E. A., Hamtini, T., & Aljarah, I. (2015, November). Preprocessing and analyzing educational data set using X-API for improving student's performance. In Applied Electrical Engineering and Computing Technologies (AEECT), 2015 IEEE Jordan Conference on (pp. 1-5). IEEE.