Guidance System for Scrutinizing the Students Performance using Random Forest Classifier

P. Satya Shekar Varma, P. Shyam Sunder, Koppula Sri Vasuki Reddy

Abstract**:** ***For today's education leaders, the ongoing challenge is to assess the student's academic performance that could probably affect their organization's potential. So, the emergence of Educational Data Mining [EDM] became the solution. By utilizing the data mining methods, infused with a theory of understanding the application and elucidation of the education and learning experience, EDM practitioners are able to generate training models that interpret the results and spot the students who may show poor performance so as to help tutors to offer effective learning environment. This paper proposes a guidance system which aims to analyze student's demographic data, academic details and extract all possible knowledge through surveys from students, parents and teachers with regard to latter state to configure whether the student is on the proper course of achieving the goals using the random forest classification algorithm. This model pursues highest possible accuracy comparison to the other previously related models proposed by authors. Furthermore, Anaconda3 data mining tool is used to develop this model which flourishes to draw the attention towards the pupils functioning based on their interests. In this study, we have accumulated the records of 480 students with 16 attributes. After contemplating all records of factors considered earlier for forecasting student's academic participation, we cull the most consistent featured based on their hypothesis and association with the performance.***

Keywords **:** ***Anaconda3, Classification, Component, Decision Tree, EDM, Guidance System, Prediction, Random Forest.***

I. INTRODUCTION

Rapid growth in the education sector led to the enormous amount of student's data which in-turn led to the need for the more sophisticated algorithms to mine such data. Student performance is the essential requirement for the today's educational institutions as they depict the potential of the organizations. The traditional methods cannot be directly applied to the data as they are pertinent to specific objective. So the emergence of Educational data mining concerns the advance practices dealing with prospecting the distinctive and wide-range data attained from the education domain.

Educational Data Mining (EDM) alludes to work, research tools build to impulsively grasp knowledge from the vast data repositories to gain insight on the critical aspects of student's academic performance using the primary essence of data mining techniques.

EDM exploits academic data to cater educational institutions to draft academic strategies, in order to enhance quality of service in education [1]. The data mining is crucial in scrutinizing the students performance by considering all the factors such as gender, nationality, stageid, no of days absent, Parents attending survey, visited resources, etc. The classification algorithm classifies and predicts the performance in a specific fashion to obtain maximum accuracy. Mythili and Mohamed Shanavas [2], evaluated the students performance in the Weka tool by applying all the classification algorithms namely Multilayer perceptron, random forest, j48, IB1 and Decision Table. The results obtained through the above classification model review the execution time, confusion matrices and also the accuracy level. Their research further explores the cogency of machine learning algorithms in determining the impact of various factors like gender, economy, result, parental education and the locality within the study and analyze of school student's academic performance. Manzoor Ahmed Hashmani, Maryam Zaffar and K.S.Savita [3], compared all the feature selection algorithms by applying on the students dataset. The results obtained from this are used to explain the importance of the selection of relevant features or attributes for the upcoming researchers in combining the classifiers and feature selection algorithms best suitable for student prediction model which is one of the crucial issues for the education leaders. Ghadeer Mobasher, Ahmed Shawish and Osman Ibrahim [4], proposed a EDM framework in form of system that filters the user's choices to predict their preferences and ratings. This model is utilized for analyzing the students performance based on the study related, psychological, demographic attributes. This work was done considering 200 students and this realistic case guaranteed the marvelous proficiency in their academic performance prediction, in-depth knowledge and skills extraction, and great benefit of recommendations provided by the proposed chassis. Shovon and Haque [5], presented a procedure to predict student's GPA using the decision tree of the data mining and data clustering kNN algorithm. This prediction helps the academicians to take the appropriate measures and steps to enhance the academic performance of the students by also reducing the failing ratio. Maria Koutina et al [6], mainly looked for the best techniques to analyze the final grades of postgraduate students of Lonian University Informatics, Greece.

Revised Manuscript Received on April 30, 2020.
***** Correspondence Author

P. Satya Shekar Varma*, Department of Computer Science and Engineering at Mahatma Gandhi Institute of Technology, Hyderabad, India. E-mail: pssvarma_cse@mgit.ac.in.

P. Shyam Sundar, Department of Computer Science and Engineering at Mahatma Gandhi Institute of Technology, Hyderabad, India. E-mail: pshyamsunder_cse@mgit.ac.in

Koppula Sri Vasuki Reddy, Department of Computer Science and Engineering at Mahatma Gandhi Institute of Technology, Hyderabad, India. E-mail: vasuki.koppula6@gmail.com

By undergoing many previous works, they considered age, gender, occupation, marital status, number of children, jobs correlated with bachelors, computers, computer literacy, Bachelor in informatics, another master.

Earlier, a significant research work has been done to group the student's data and assess the academic performance of student's using various feature selection techniques, clustering and classification algorithms to retrieve the best possible prediction. In this paper, the research work of all the authors is considered and the stimulation results reveal that the random forest classification algorithm exhibit maximum accuracy among the others. Random forest classification algorithm creates a large number of decision trees from all the attributes that function as an aggregate. Then each individual tree in the random forest cleaves out a class prevision and the class with the bulk elects turns out to be complete model's prediction. Anaconda is precise to build, train, and deploy the model as it provides open source tools in centralized, collaborative, and version-controlled environment.

The rest of the paper is organized as follows; Section 2 demonstrates the basic structure of the system. Section 3 elaborates methodologies used in this study work. Section 4 the findings of the intended working model are revealed in form of confusion matrix. Section 5 gives the conclusion and the future research purview.

II. SYSTEM ARCHITECTURE

The proposed model primarily focuses on consolidating the various demographic, knowledge from surveys, study related characteristics as dataset. Many existing systems focused purely on the performance of the student regarding the data from the time of admission. Whereas this system gathers the data through pretest EPQ. Extended project qualification [EPQ] is a course taken by many of the students of England and Wales to gain a lot of information regarding the academic content and explore their interests. This data in the form of questionnaire which drive towards student's interests and greater value placed on results achieved. In addition to that the data is also collected from the academic performance and also library through resource scale. Resource scale refers to the type and the number of resources accumulated. Resources visited, discussions attended, announcements viewed are all acquired for this scaling. The final result dataset is formed consisting of 16 attributes of 480 students after EPQ analysis and resources utilization analysis. In this paper we aim to use random forest classifier as a classification algorithm for the result dataset. Random Forest constructs the decision trees for all possible attributes and predicts the class label with maximum votes from the decision trees. Finally, the classifier predicts the accuracy of the final predicted model and visualize the confusion matrix. Fig. 1 shows the work flow of the proposed model.

III. METHODOLOGY

The setup used for this model is Anconda3, Jupyter notebook. Anaconda is a data mining software which provides the data mining tools such as jupyter, spyder, orange which manage environments with conda consisting of all the packages predefined. Jupyter notebook allows to run pythonscript displaying input and output code. The process flow of the illustrated model is in the four main steps starting with the dataset preparation and validating the model. These steps are explained as A, B, C, D.

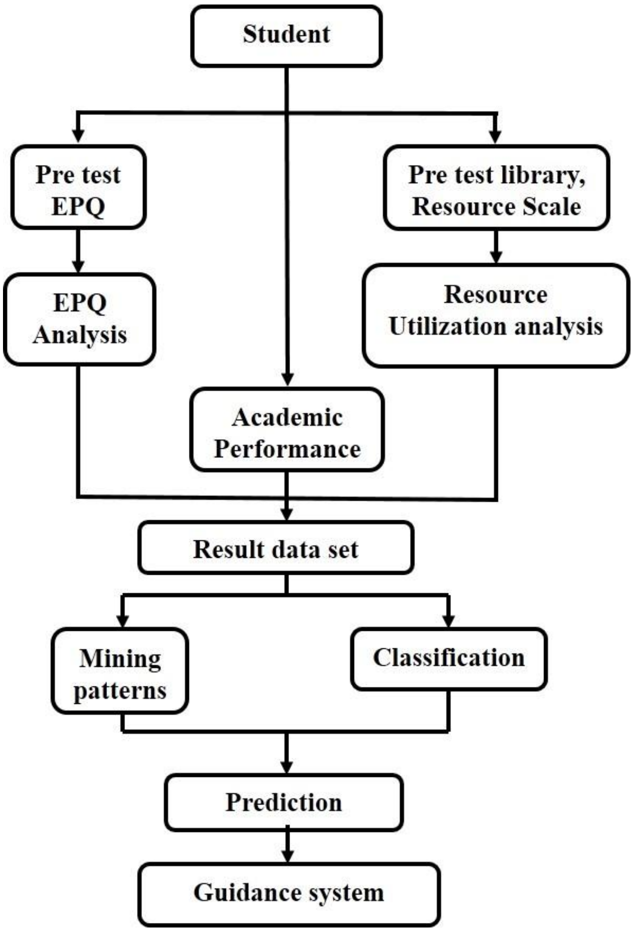


Fig. 1. Pattern of the proposed guidance system.

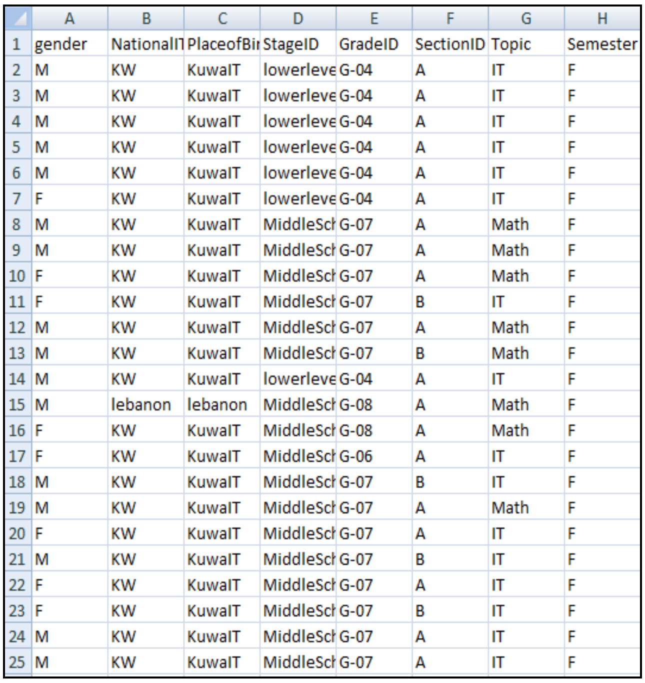
A. Dataset collection and preprocessing

The dataset includes the features gathered related to mainly four categories. Firstly, the demographic data, which includes variables i.e., gender, NationalITy, PlaceofBirth, StageID, SectionID. The second category specifies the data related to academics such as GradeID, Semester, Topic, StudentAbsenceDays. The next one includes parents and teachers survey data i.e., Relation, ParentsAnsweringSurvey, ParentsschoolSatisfaction, Discussions. The final one constitutes the data from Resource utilization analysis and Pre-test EPQ analysis i.e., raisedhands, visITedResources, AnnouncementsView. Table 1 specifies the catalogue of attributes with their possible affiliated values and the type. Data processing is a crucial phase of data mining. The real-world data is unclear, incomplete and noisy so it should be cleaned and complete for further process. Data Processing includes a set of activities, so the data after the procedure can be directly induced to the model to obtain results [7]. The chosen data is clear and there are no missing values.

Fig. 2 displays the screenshot of the actual dataset used for this model in csv file. It shows the few attributes mentioned in the Table 1.

Table- I: List of Attributes

Attributes	Possible values and their attribute type
Gender	M, F (binary)
NationalITy	KW, Jordon, Palestine, Iraq, lebanon, Tunis, SaudiArabia, Egypt, Syria, Lybia, Iran, USA, Morocco, venzuela (nominal)
PlaceofBirth	KuwaIT, Jordon, Palestine, Iraq, lebanon, Tunis, SaudiArabia, Egypt, Syria, Lybia, Iran, USA, Morocco, venzuela (nominal)
StageID	Lowerlevel, MiddleSchool, HighSchool (nominal)
GradeID	G-02 to G-012 (ordinal)
SectionID	A, B, C (ordinal)
Topic	IT, Math, Arabic, Science, English, Quran, Spanish, French, History, Biology, Chemistry, Geology (nominal)
Semester	F, S (binary)
Relation	Father, Mum (binary)
raisedhands	Variable between 1 to 100 (numeric)
VisITedResources	Variable between 1 to 100 (numeric)
AnnouncementsView	Variable between 1 to 100 (numeric)
Discussion	Variable between 1 to 100 (numeric)
ParentAnsweringSurvey	Yes, No (binary)
ParentsschoolSatisfaction	Good, Bad (binary)
StudentAbsenceDays	Under-7, Above-7 (binary)
Class label	L, M, H (nominal)



	A	B	C	D	E	F	G	H
1	gender	NationalIT	PlaceofBi	StageID	GradeID	SectionID	Topic	Semester
2	M	KW	KuwaIT	lowerleve	G-04	A	IT	F
3	M	KW	KuwaIT	lowerleve	G-04	A	IT	F
4	M	KW	KuwaIT	lowerleve	G-04	A	IT	F
5	M	KW	KuwaIT	lowerleve	G-04	A	IT	F
6	M	KW	KuwaIT	lowerleve	G-04	A	IT	F
7	F	KW	KuwaIT	lowerleve	G-04	A	IT	F
8	M	KW	KuwaIT	MiddleSch	G-07	A	Math	F
9	M	KW	KuwaIT	MiddleSch	G-07	A	Math	F
10	F	KW	KuwaIT	MiddleSch	G-07	A	Math	F
11	F	KW	KuwaIT	MiddleSch	G-07	B	IT	F
12	M	KW	KuwaIT	MiddleSch	G-07	A	Math	F
13	M	KW	KuwaIT	MiddleSch	G-07	B	Math	F
14	M	KW	KuwaIT	lowerleve	G-04	A	IT	F
15	M	lebanon	lebanon	MiddleSch	G-08	A	Math	F
16	F	KW	KuwaIT	MiddleSch	G-08	A	Math	F
17	F	KW	KuwaIT	MiddleSch	G-06	A	IT	F
18	M	KW	KuwaIT	MiddleSch	G-07	B	IT	F
19	M	KW	KuwaIT	MiddleSch	G-07	A	Math	F
20	F	KW	KuwaIT	MiddleSch	G-07	A	IT	F
21	M	KW	KuwaIT	MiddleSch	G-07	B	IT	F
22	F	KW	KuwaIT	MiddleSch	G-07	A	IT	F
23	F	KW	KuwaIT	MiddleSch	G-07	B	IT	F
24	M	KW	KuwaIT	MiddleSch	G-07	A	IT	F
25	M	KW	KuwaIT	MiddleSch	G-07	A	IT	F

Fig. 2. Screenshot of dataset in excel sheet.

B. Unpruned decision trees construction

Random forest is a form of assemblage learning algorithm which applies classification on the randomly selected dataset forming a forest i.e. bulk of decision trees with their predictive class label. The final model is predicted by measuring the prediction of the most of the trees [8]. Initially, the un pruned decision trees are built but can be pruned by reducing the size of trees that boost the power for classifying instances. The data is considered as two-dimensional data which has three class labels (L, M, H). A decision tree is built for every feature of the data by splitting the values, classes along one or the other axis, and for each value, the count of its repletion is plotted against the graph. For the first split, counts for the possible values of the feature are plotted and then count for the values according to their respective class labels. Fig. 3 shows the decision trees for Semester, ParentsSchoolSatisfaction attributes with respect to their splitting criteria. Fig. 4 shows the decision trees for ParentAnsweringSurvey, StudentAbsenceDays, Topic attributes. Fig. 5 and Fig. 6 shows the trees for NationalITy, GradeID, StageID attributes and Gender, Relation, SectionID attributes respectively.

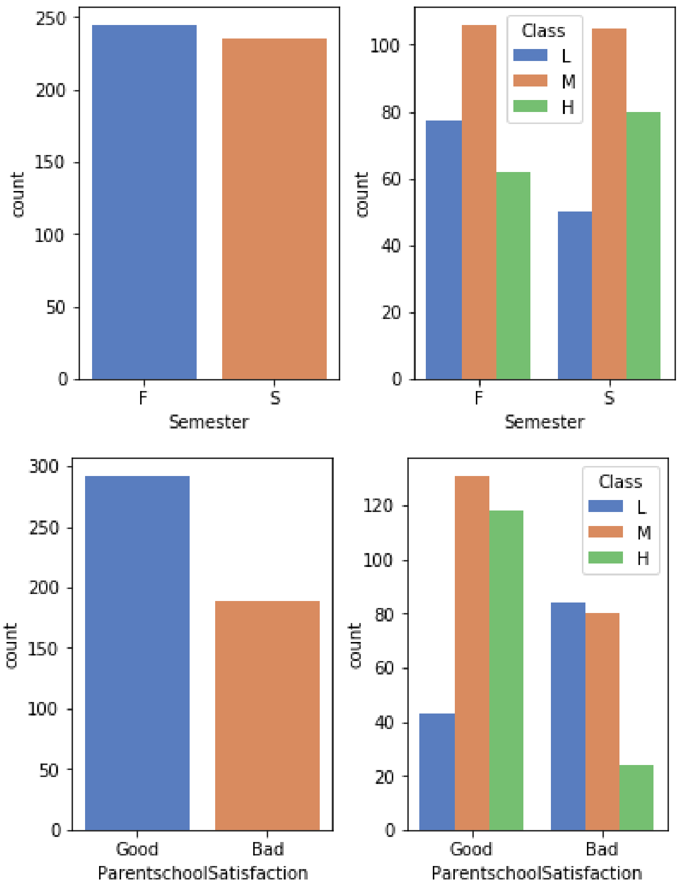


Fig. 3. Plot for Semester, ParentsSchoolSatisfaction attributes.

C. Preparation of the dataset

Unfortunately, the real-life data is collected from variant sources with different ranges. So, we cannot directly feed this data i.e. raw data, into the model. Data has to be prepared such that it can be fed to the model obtaining precise accuracy. Normalization is often considered a part of data preparation of machine learning before feature selection method. It transforms features to be on same scale. All the datasets do not require normalization. If the normalization (normalize=True) is done to a particular feature, it will return an object which shows the relative frequencies of the unique values of that attribute. Fig. 7 gives the normalized values for the topic and PlaceofBirth attributes. Also, it is considered that the raw data is designed in the human readable form or categorical values i.e. names or words. Machine learning algorithms can work in a better way if the labels are of machine-readable form, which refers to numbers. So, the technique of encoding came into picture. This is also considered as an important part of dataset preparation.

There are a wide range of encoding techniques such as label encoding, original encoding, one hot encoding, binary encoding, sum encoding, helmert encoding etc [9]. Choosing the encoder has a great impact on the prediction of the model, so right encoder for the right scenario may advance the accuracy by great number. For our input dataset we have chosen label encoding.

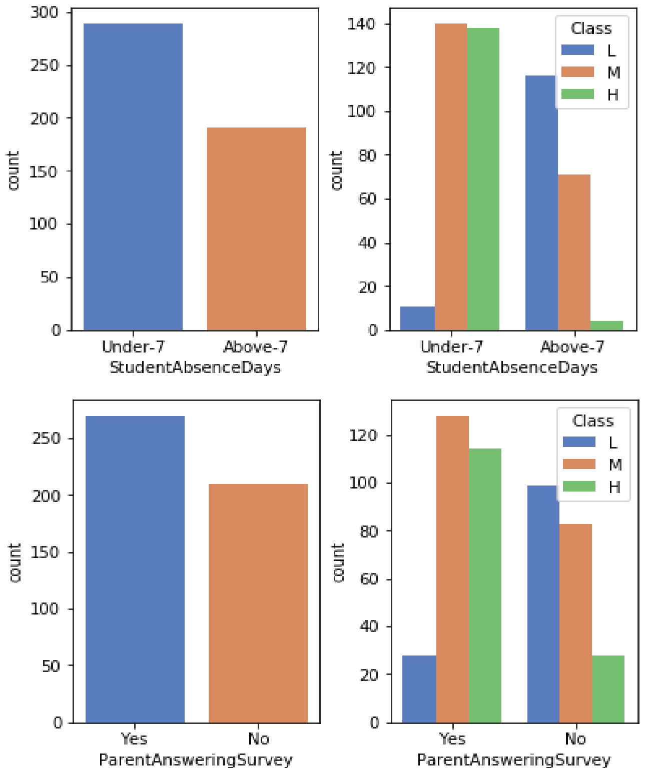


Fig. 4. Plot for StudentAbsenceDays, ParentAnswering Survey, Topic, attributes.

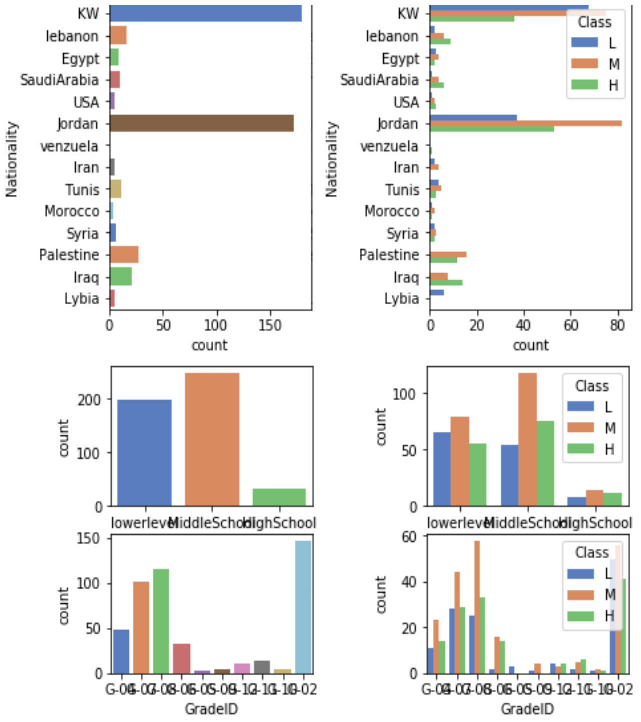


Fig. 5. Plot for NationalITy, StageID and GradeID attributes.

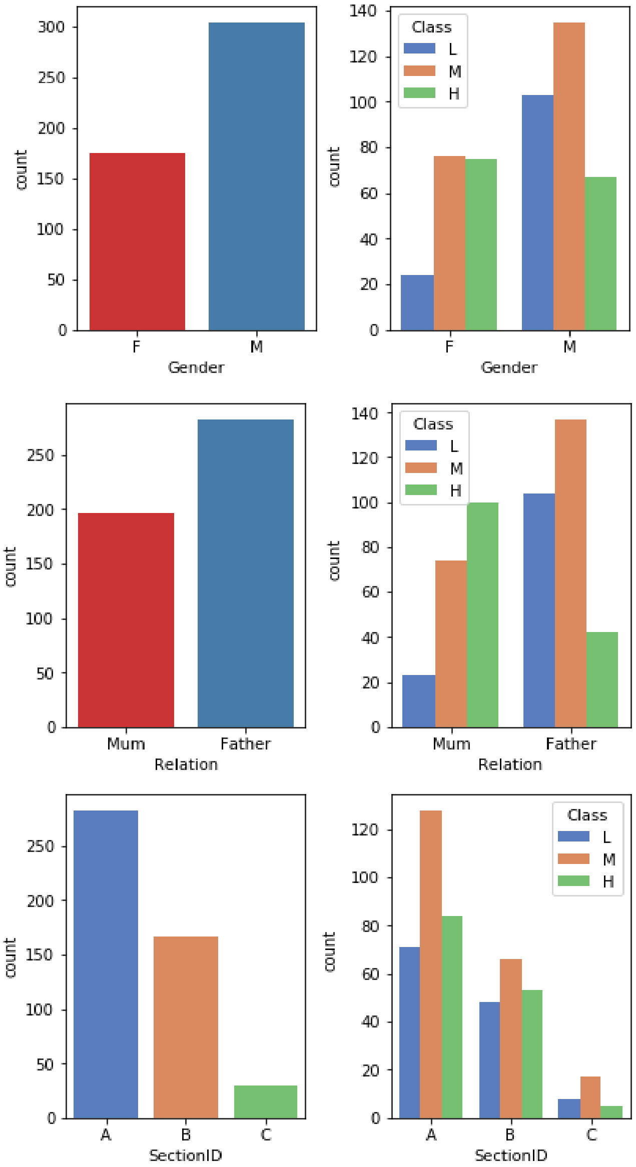


Fig. 6. Plot for Gender, Relation, SectionID attributes.

In ML, encoders convert textual to numeric representation of attributes. Label encoder encodes the labels with values from 0 to n-1 values, where n is the number of distinct labels used in the dataset. If the label repeats the same value assigned earlier is reassigned. Sklearn library provides a efficient tool, LabelEncoder for this preprocessing step. The corresponding encoded values of the dataset are shown in below Fig. 8.

D. Training the model

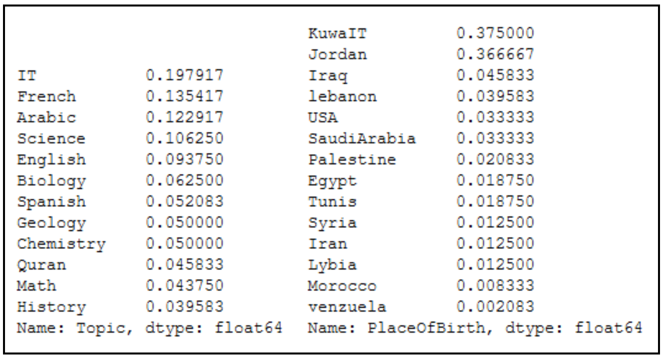
After all the initial steps of data collection, cleaning and pre-processing we need to fit the data into the proposed ML model. In order to train the model, the following 3 steps are performed on the encoded dataset:

1) Dividing the data into features and target: Firstly, to train our model we need the training set of data. The train data contains complete set of feature variables (independent variables) and target variable (dependent variable). So, we need to separate all the features that are used for predicting the target variables in our dataset.

Class label is the only target variable that is predicted for all the feature variables.

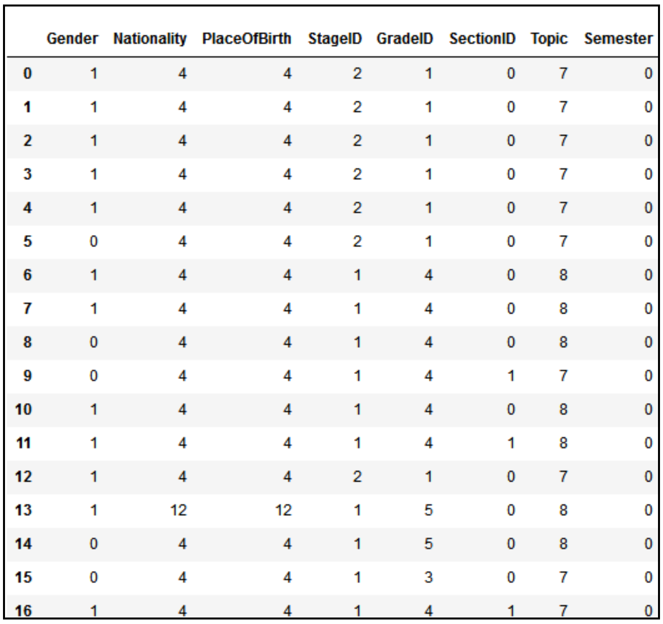
2) Partitioning the data into training and testing set: For a large dataset, the maximum part of the dataset is considered for training and the remaining is for testing purpose, which is not considered while training the model. If X% of the data is for training purpose then the remaining (100-X)% is determined for the testing . Both the features and target of training set are used to train the model and compare the test values for the same to predict the model. There are many ways for the train-test splitting ratio but one of the common ways is cross validation [10]. Cross validation prevents over fitting of the data. For our model, the train set is composed of all the 384 data values and test set is of 96 values.

3) Fit the model: After splitting the data, the proposed model is build using the random forest classifier. Then the training set is supplied to the model. The model learns the relationships from the training set and train itself accordingly. Now, the time for testing the model starts where the test set features are fitted in the model and the target variable value is predicted. Fig. 9 depicts the count of test as well as training data.



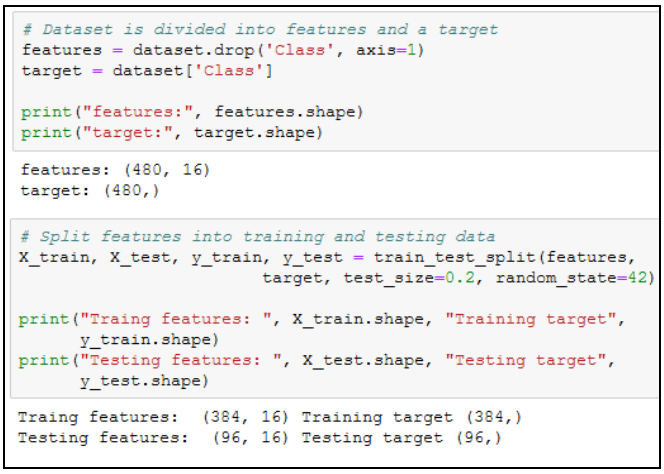
```
                                 KuwaIT          0.375000
                                 Jordan          0.366667
IT            0.197917           Iraq            0.045833
French        0.135417           lebanon         0.039583
Arabic        0.122917           USA             0.033333
Science       0.106250           SaudiArabia     0.033333
English       0.093750           Palestine       0.020833
Biology       0.062500           Egypt           0.018750
Spanish       0.052083           Tunis           0.018750
Geology       0.050000           Syria           0.012500
Chemistry     0.050000           Iran            0.012500
Quran         0.045833           Lybia           0.012500
Math          0.043750           Morocco         0.008333
History       0.039583           venzuela        0.002083
Name: Topic, dtype: float64      Name: PlaceOfBirth, dtype: float64
```

Fig. 7. Normalized values of attributes.



	Gender	Nationality	PlaceOfBirth	StageID	GradeID	SectionID	Topic	Semester
0	1	4	4	2	1	0	7	0
1	1	4	4	2	1	0	7	0
2	1	4	4	2	1	0	7	0
3	1	4	4	2	1	0	7	0
4	1	4	4	2	1	0	7	0
5	0	4	4	2	1	0	7	0
6	1	4	4	1	4	0	8	0
7	1	4	4	1	4	0	8	0
8	0	4	4	1	4	0	8	0
9	0	4	4	1	4	1	7	0
10	1	4	4	1	4	0	8	0
11	1	4	4	1	4	1	8	0
12	1	4	4	2	1	0	7	0
13	1	12	12	1	5	0	8	0
14	0	4	4	1	5	0	8	0
15	0	4	4	1	3	0	7	0
16	1	4	4	1	4	1	7	0

Fig. 8. Normalized values of attributes. Sample encoded label values of dataset.



```
# Dataset is divided into features and a target
features = dataset.drop('Class', axis=1)
target = dataset['Class']

print("features:", features.shape)
print("target:", target.shape)
```

```
features: (480, 16)
target: (480,)
```

```
# Split features into training and testing data
X_train, X_test, y_train, y_test = train_test_split(features,
                        target, test_size=0.2, random_state=42)

print("Traing features: ", X_train.shape, "Training target",
      y_train.shape)
print("Testing features: ", X_test.shape, "Testing target",
      y_test.shape)
```

```
Traing features:  (384, 16) Training target (384,)
Testing features:  (96, 16) Testing target (96,)
```

Fig. 9. Train and test data.

IV. RESULTS AND PREDICTION

After the final model is formed it is used for predicting the results. The test set is unseen to the model during its training. After the training, the test set is given and the outputs upheld by the training are compared with test set to compute the performance of the guided model. The analysis and performance measurement is a rigid process and require statistical knowledge. Classification accuracy is best fit metric to gauge the performance of proposed model. The results obtained illustrates the accuracy score of the classifier with the report of main classification metrics. Precision, recall, f-score, support are computed for each class to obtain the resultant report. These parameters are called the confusion metrics. Accuracy is the intuitive measure that is defined as the fraction of appropriately envisaged observations to the total number of observations. Table 2 gives the account of terms required for calculated these values.

Table- II: List of Attributes

	Predicted class		
Actual Class		***True***	***False***
	True	True Positive	False Negative
	False	False Positive	True Negative

- *Accuracy*: It is a metric apt for the performance analysis of the any model. But in case of imbalanced classification problem accuracy is not satisfactory. In such problem other metrics are considered [11].
- *Precision:* Fraction of correctly predicted positive instances to the total number of positive instances.

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{True Positives} + \text{False Positives} + \text{False Negatives} + \text{True Negatives}} \qquad (1)$$

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \qquad (2)$$

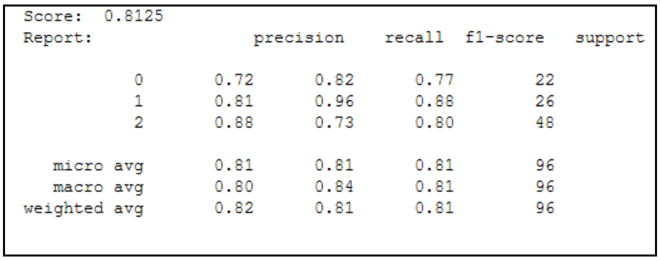
- *Recall:* Recall is the proportion of rightly predicted positive instances to the all instances in definite class.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \qquad (3)$$

- *F1 score:* It addresses into account both the precision and recall metrics by computing their harmonic mean.

$$\text{F1 Score} = 2 \times \frac{(\text{Recall} \times \text{Precision})}{(\text{Recall} + \text{Precision})} \qquad (4)$$

- *Support:* specifies the count of the actual contingencies of the class label in the resultant dataset.



```
Score:  0.8125
Report:                precision    recall  f1-score   support

           0       0.72      0.82      0.77        22
           1       0.81      0.96      0.88        26
           2       0.88      0.73      0.80        48

   micro avg       0.81      0.81      0.81        96
   macro avg       0.80      0.84      0.81        96
weighted avg       0.82      0.81      0.81        96
```

Fig. 10. Output results.

Fig. 10 summarizes the results of the prediction of the model in the form of confusion matrix. From the observation, it is shown that the exactness paradigm is approximately 82%. Precision, recall, f1 score measures are computed for each of the instance of the class label for analyzing the results more precisely. The micro average, macro average, weighted average of the three class instances are summed up together

- *Code:* The source code with the dataset developed for the proposed work is available athttps://drive.google.com/drive/u/0/folders/1EDzD9qUMFAzp5aisdfsPA8ZyhOlfOW7g.

V. CONCLUSION

Our education system is developing and has vast interest in students performance, which reflects the organization's potential. In order to incorporate this need, many techniques has been invented and experimented considering various features to analyze the student grades, interests. In this paper, the main emphasis is made to compute the students performance from the surveys by parents and teachers, academic data and student's demographic data that influence the result and excavates the potential of the classifier used. Considering these features extended the accuracy of the model built using random forest classifier. This classifier proved to be the best among the other classifiers to predict accurate results by many other researchers as well.

Our study is restricted to only 16 attributes but there may be scope in future to include other relevant attributes such as hours spent on social media, playing games, sleeping timings, stress level, family status, which may considerably impact the performance of the students of future generation. Other regression, clustering, association rule data mining techniques or the fusion of two or more of them can also be used for this evaluation purpose and this can be included in the future studies.

REFERENCES

1. M. Ramaswami and R. Bhaskaran, "A study on feature selection techniques in educational data mining" arXiv preprint arXiv:0912.3924, 2009.
2. Mythili and Mohamed Shanavas, "An Analysis of students' performance using classification algorithms," IOSR Journal of Computer Engineering (IOSR-JCE), vol.16, iss. 1, ver. III, Jan. 2014, pp.63-69.
3. Maryam Zaffar, Manzoor Ahmed Hashmani, K.S. Savita, "Performance Analysis of Feature Selection Algorithm for Educational Data Mining," 2017 IEEE Conference on Big Data and Analytics (ICBDA), 2017.
4. Ghadeer Mobasher, Ahmed Shawish and Osman Ibrahim, "Educational Data Mining Rule based Recommender Systems," In Proceedings of the 9th International Conference on Computer Supported Education (CSEDU 2017), vol. 1, pp. 292-299, 2017.
5. Shovon and Haque, "An Approach of Improving Student's Academic Performance by using K-means clustering algorithm and Decision tree," International Journal of Advanced Computer Science and Applications 3(8), Nov. 2012.
6. Maria Koutina and Katia LidaKermanidis," Literature Survey on Student's Performance Prediction in Education using Data Mining Techniques," International Journal of Education and Management Engineering · Nov. 2017.
7. Dwivedi, S. K., & Rawat, B," A review paper on data preprocessing: A critical phase in web usage mining process". 2015 International Conference on Green Computing and Internet of Things (ICGCIoT).
8. Available online at - https://medium.com/analytics-vidhya/machine-learning-decision-trees-and-random-forest-classifiers-81422887a544.
9. KedarPotdar, Taher S. Pardawala, Chinmay D. Pai," A Comparative Study of Categorical Variable Encoding Techniques for Neural Network Classifiers," International Journal of Computer Applications (0975 – 8887), vol. 175, no.4, Oct. 2017.
10. Available online at- https://towardsdatascience.com/train-validation-and-test-sets-72cb40cba9e7.
11. Available online at - https://blog.exsilio.com/all/accuracy-precision-recall-f1-score-interpretation-of-performance-measures.

AUTHORS PROFILE

P. Satya Shekar Varma received his Master's Degree in Computer Science from Jawaharlal Nehru Technological University, Hyderabad. He is currently an Assistant Professor in the Department of Computer Science and Engineering at Mahatma Gandhi Institute of Technology, Hyderabad. He has 13 years of teaching experience. He has 14 publications in international journals and conferences. His research areas include data mining, cloud computing, big data and data security.

P. Shyam Sundar received his Master's Degree in Computer Science from Jawaharlal Nehru Technological University, Hyderabad. He is pursuing his Ph.D at Osmania University, Hyderabad. He is currently an Assistant Professor in the Department of Computer Science and Engineering at Mahatma Gandhi Institute of Technology, Hyderabad. He has 13 years of teaching experience. He has 12 publications in international journals and conferences. His research areas include data mining, cloud computing, big data and data security.

Koppula Sri Vasuki Reddy is currently pursuing bachelor's degree in Computer Science and Engineering at Mahatma Gandhi Institute of Technology, Hyderabad, India. Her research areas include data mining, cloud computing, big data and data security.