OPEN ACCESS

Middle East Research Journal of Engineering and Technology
ISSN: 2789-7737 (Print) ISSN: 2958-2059 (Online)
Frequency: Bi-Monthly
DOI: 10.36348/merjet.2021.v01i01.013

KUWAIT SCHOLARS PUBLISHER
Website: http://www.kspublisher.com/
Email: office@kspublisher.com

An Analysis of Academic Performance through Educational Data Mining Approach

Dr. Ajay Goyal[1*], Dr. Vadivel. G[2], Dr. Pawan Kumar[2], Er. Abhinash Singla[3]
[1]Professor IT, Bhai Gurdas Institute of Engineering & Technology, Sangrur, Punjab
[2]Associate Professors IT, Bhai Gurdas Institute of Engineering & Technology, Sangrur, Punjab
[3]Assistant Professors IT, Bhai Gurdas Institute of Engineering & Technology, Sangrur, Punjab

Abstract: It is essential to improve the value of education by accurately predicting student academic achievement. Some studies have been done that concentrate primarily on predicting students' success in college. In contrast, research on secondary-level performance prediction has been sparse, yet the secondary level is often used as a standard to define students' educational progress at higher levels of education. The goal of this study was to identify the most important influences on secondary school students' academic performance and to develop an effective classification model for the prediction of academic performance by combining single and ensemble-based classifiers.

Keywords: Data mining, Prediction, Student performance, Higher Education, EDM, Ensemble Learning.

Research Paper

***Corresponding Author:**
Dr. Ajay Goyal
Professor IT, Bhai Gurdas Institute of Engineering & Technology, Sangrur, Punjab

How to cite this paper:
Ajay Goyal, Vadivel. G, Dr. Pawan Kumar, Er. Abhinash Singla (2021). An Analysis of Academic Performance through Educational Data Mining Approach. *Middle East Res J. Eng. Technol, 1*(1): 81-85.

Article History:
| Submit: 22.10.2021 |
| Accepted: 27.11.2021 |
| Published: 31.12.2021 |

INTRODUCTION

Predicting student performance is becoming increasingly important in today's world because of the critical role it plays in the growth of nations because it is entirely dependent on the educational process that produces a generation capable of leading this country and its progress toward development in all areas of life (scientific, economic, social and military, etc) [Minaei-Bidgoli et al., 2003]. Also, the evaluation of students' performance is a reflection of the efficacy of educational institutions which is responsible for developing successive generations in line with the different stages of the lives of people in every country. Therefore, emphasizing on the expansion of the educational process is one of the major necessities that motivate governments represented via educational institutions to make vast and painful efforts to push the educational process towards continuous and rising development.

Future knowledge may be gathered through prediction. The higher the amount of data is, such in enormous databases, the better the prediction is made; this process is known as data mining which is used to identify hidden information by reviewing numerous data sources connected to diverse domains such as commercial, social, medical, and educational. The knowledge offered by numerous resources of educational data may be examined to gain needed information. A new area named as Educational Data Mining (EDM) was formed as a technique of obtaining important information [Pradeep et al., 2015. The relevance of EDM has expanded swiftly in the present day because of the growth in the acquired data, according to the educational data received from different e-learning systems, as well as the development of traditional educational systems. The power of EDM is shown by several facts in different sectors and how they are connected together. It concerned with the extraction of features to aid the development of educational process from huge data offered by institution. Unlike the examination of traditional database, which can answer questions, such as who is the student who failed in the exam? EDM can answer more sophisticated issues such as the prediction of the result of the student (whether he will pass or fail in the exam).

Educational Data Mining

Academic institutions seek to construct their student's model to anticipate both attributes and performance of each student individually. Therefore, the researchers that are engaged with the EDM area utilize various methods of data mining in order to evaluate lecturers, to perform their educational organization [Elena Susnea et al., 2009]. Because existing educational institutions do not place enough emphasis on predicting students' success, they are inefficient.

Raising educational efficiency is made possible by anticipating what classes a student will find interesting and tracking his activities in educational institutions. Many educational institutions constantly evaluate their pupils using machine learning and EDM methods. Students' performance and the educational process as a whole may be improved using these assessment methods [E. Frank et al., 2005].

METHODOLOGY

In proposed method, decision tree, random forest, and logistic regression methods are used. Steps of proposed method are as following:

Step 1: Collect the dataset this dataset contains Student academic performance reports.

Step 2: The dataset contains two different CSVs based on two different subject maths and protégées.

Step 3: Plotting graphs and Preforming EDA.

Step 4: Processing

- Merging both Dataset into one.
- Renaming columns
- Removing null values by removing the whole row.
- convert final score to categorical variable # Good:15~20 Fair:10~14 Poor:0~9

Step 5: Creating multiple ensemble models like: Decision Tree, Logistic Regression, Random Forest.

Step 6: Evaluation of the model, testing the model on the test set and measuring the performance in terms of precision, recall and F1-Score.

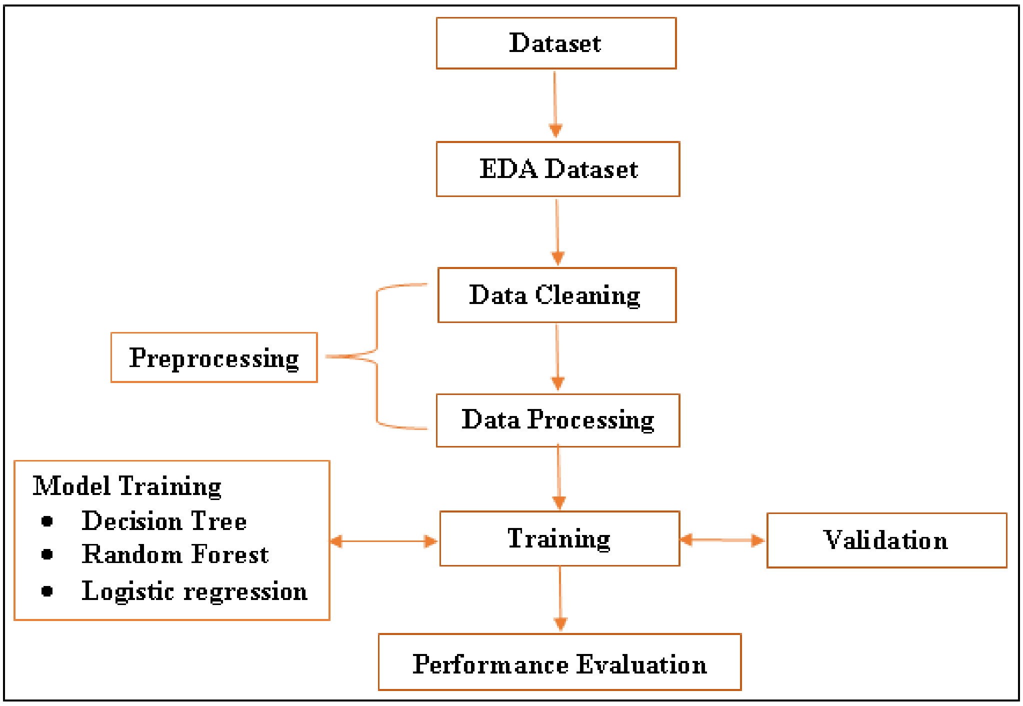


Figure-1: Flow chart of Proposed Method

Model Summary

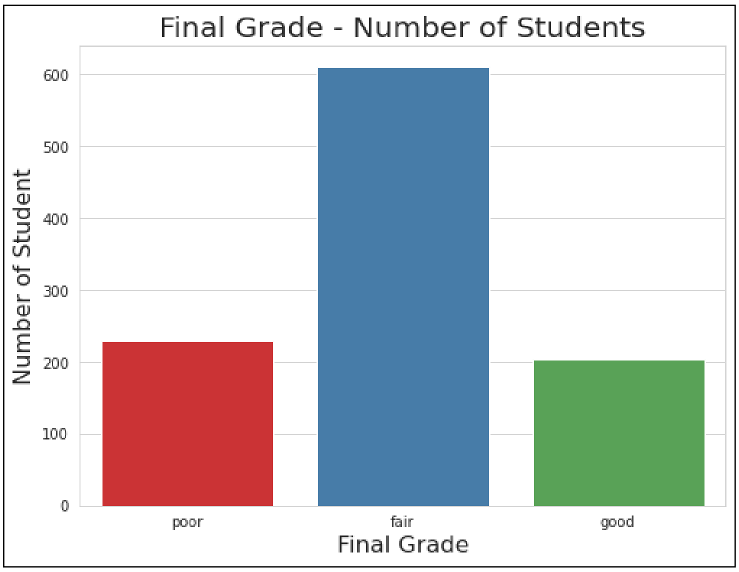
- Decision Tree
  - Min_samples_leaf =17
- Random Forest
  - n_estimators=36
  - min_samples_leag=2
- Logistic Regression
  - Multi_class=Multinomial
  - Solver=newton-cg
  - Fit_intercept=True

RESULT AND DISCUSSION

The results of Proposed methods are as follows:

Table 1: Final Grade - Number of Students

S. no.	Final grade	Number of students
1.	Poor	215
2	Fair	610
3.	Good	200



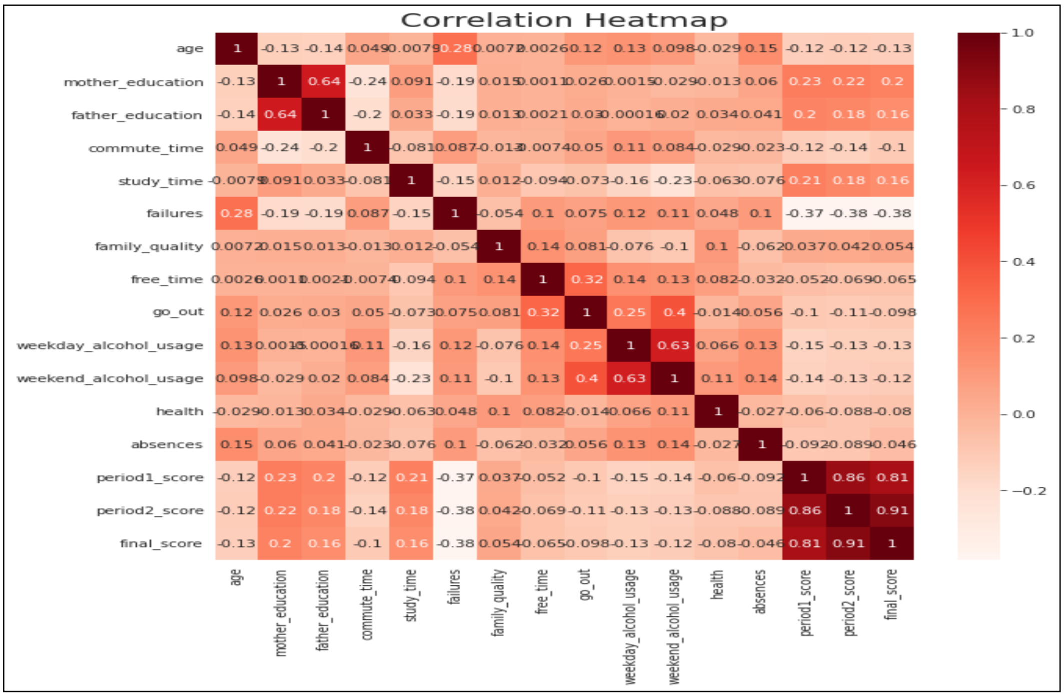
Graph 1: Final Grade Number of students

From the above graph we found the final grade of the students, in which 200 students have come in good grade category, <600 students have come in fair grade category, and 200>300 students have come in poor grade category.

Model	Accuracy	Precision	Recall	F1
Base SVM	0.72	0.68	0.72	0.66
Base Naïve Bayes	0.73	0.70	0.72	0.71
Propose Decision Tree	0.88	0.87	0.89	0.88
Propose Random Forest	0.87	0.90	0.84	0.86
Propose Logistic Regression	0.87	0.87	0.87	0.87

Correlation Heatmap

A correlation heatmap is a heatmap that depicts a two-dimensional correlation matrix between two discrete dimensions, with colored pixels representing data on a monochromatic scale. The first dimension's values display as rows in the table, while the second dimension's values appear as columns.



The amount of measurements that match the dimensional value determines the color of the cell. This makes correlation heatmaps great for data analysis since they show differences and variance in the same data while making patterns clearly visible. A color bar aids a correlation heatmap, much like a standard heatmap, in making data readily legible and understandable.

Confusion Matrix:

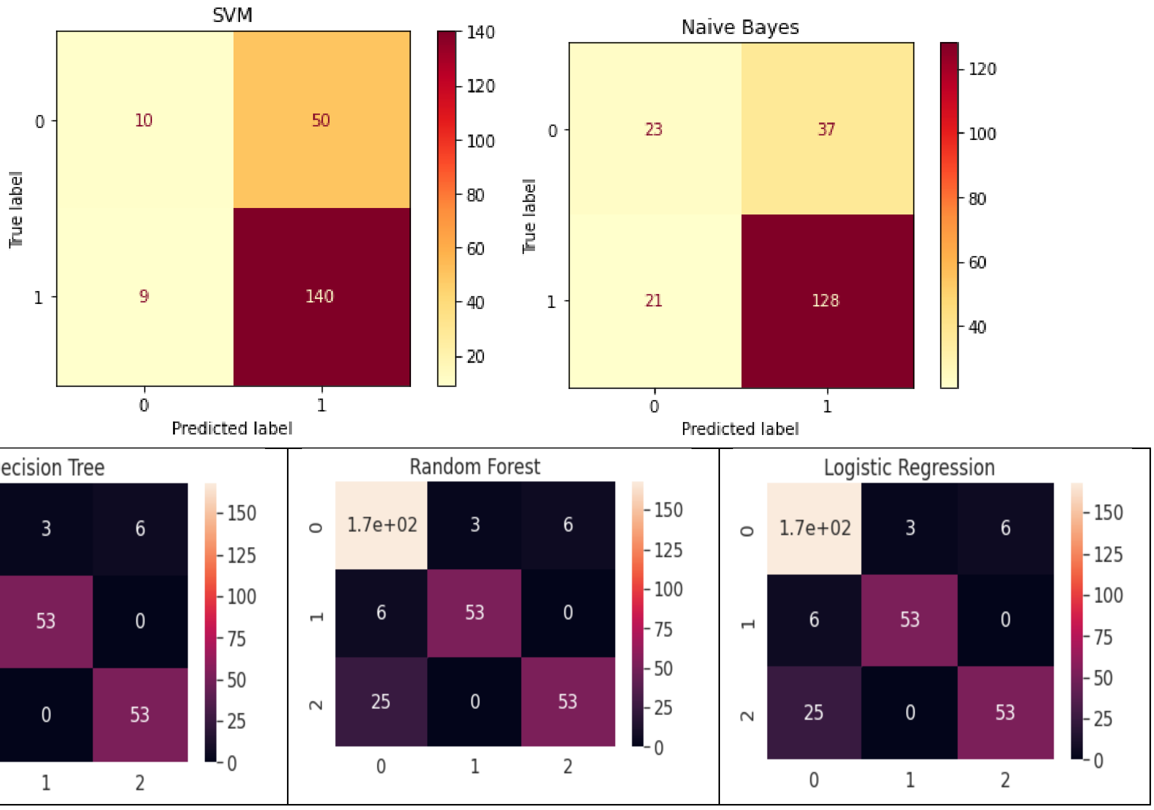


Figure: 1 Correlation Heatmaps

From the above table or graphs it is clear that for Base SVM model the accuracy is 0.72, precision is 0.68, and frequency (F1) is 0.66, for Base Naïve Bayes the accuracy is 0.73, precision is 0.70, and frequency is 0.71, for Propose Decision Tree the accuracy is 0.88, precision is 0.87, and frequency is 0.88, for Propose Random Forest the accuracy is 0.87, precision is 0.90, and frequency is 0.86, for Propose Logistic Regression the accuracy is 0.87, precision is 0.87, and frequency is 0.87.

CONCLUSION

Academic accomplishment of pupils is a major issue for educational institutions all over the world. We found the final grade of the students, in which 200 students have come in good grade category, <600 students have come in fair grade category, and 200>300 students have come in poor grade category. Proposed methods are giving better results as compare to Base SVM and Naïve Bayes methods. By using proposed methods, results of students can be improved.

REFERENCES

1. Acharya, A., & Sinha, D. (2014). Early prediction of students performance using machine learning techniques. *International Journal of Computer Applications*, *107*(1), 37-43.
2. Chen, M. S., Han, J., & Yu, P. S. (1996). Data mining: an overview from a database perspective. *IEEE Transactions on Knowledge and data Engineering*, *8*(6), 866-883.
3. Witten, I. H., Frank, E., Hall, M. A., Pal, C. J., & Data, M. (2005, June). Practical machine learning tools and techniques. In *Data mining* (Vol. 2, No. 4, pp. 403-413). Amsterdam, The Netherlands: Elsevier.
4. Şuşnea, E. (2009). Using data mining techniques in higher education. In *The 4th international conference on virtual learning* (p. 373).
5. Fernandes, E., Holanda, M., Victorino, M., Borges, V., Carvalho, R., & Van Erven, G. (2019). Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil. *Journal of business research*, *94*, 335-343.
6. Minaei-Bidgoli, B., Kashy, D. A., Kortemeyer, G., & Punch, W. F. (2003, November). Predicting student performance: an application of data mining methods with an educational web-based system. In *33rd Annual Frontiers in Education, 2003. FIE 2003.* (Vol. 1, pp. T2A-13). IEEE.

7. Pradeep, A., Das, S., & Kizhekkethottam, J. J. (2015, February). Students dropout factor prediction using EDM techniques. In *2015 International Conference on Soft-Computing and Networks Security (ICSNS)* (pp. 1-7). IEEE.

8. Xu, J., Moon, K. H., & Van Der Schaar, M. (2017). A machine learning approach for tracking and predicting student performance in degree programs. *IEEE Journal of Selected Topics in Signal Processing*, *11*(5), 742-753.