JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

An Analysis of Academic Performance through Educational Data Mining Approach

Preeti Jain	Dr. Neha Jain	Dr. Shankar Ramamoorthy
Research Scholar	HOD Dept of Languages, PU	Guide
SVN University, Sagar	HEI Expert TALA	Banglore

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. In this approach data are sent to Decision Tree, Random forest and logistic Regression to train and test data. The results showed that the proposed approach gives better results.

***Index Terms*: Data mining, Prediction, Student performance, Higher Education, Educational Data Mining.**

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).

Data Mining

To find patterns and trends in data, data mining has evolved into a technique known as data mining (DM). A massive amount of data in the information industry is turned into useful information (Jothi et al., 2015). Aside from that, data mining (DM) is an approach to extracting information from huge amounts of data by performing operations such as cleaning and combining, transformation, mining, estimation, and prediction.

Educational Data Mining

EDM has gotten a lot of attention from academics in the past several decades since there is a wealth of instructional information available from a variety of sources. EDM's primary objective is to improve the effectiveness of DM models in order to preserve a large quantity of instructional material and provide a safe learning environment for students. Diverse DM and analytics models have been used in this technique (Baker et al., 2014). Other models utilized to make predictions included those for Classification, Regression, and Latent Factor Analysis (LFA).

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].

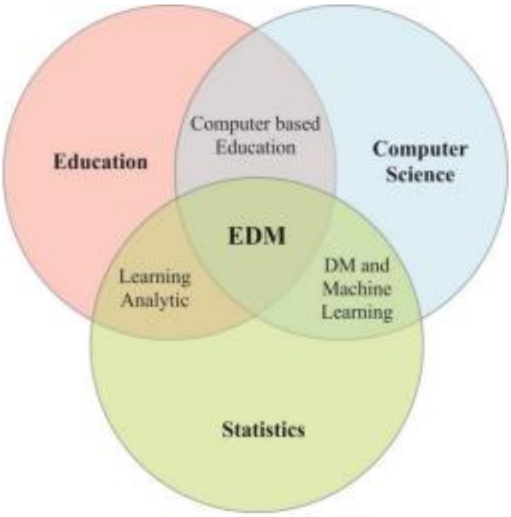


Figure1: Main areas involved in educational data mining

RESEARCH METHODOLOGY:

The data was integrated into two datasets related to Mathematics (with 395 examples) and the Portuguese language (649 records) classes.

During the preprocessing stage, some features were discarded due to the lack of discriminative value. For example: Family income. Almost every student have computer at home.

The target value is G3, which, according to the dataset, can be binned into a passing or failing classification.

If G3 is greater than or equal to 10, then the student passes. Otherwise, she fails. Likewise, the G1 and G2 features are binned in the same manner.

The data can be reduced to 4 fundamental features, in order of importance:

- G2 score
- G1 score
- School
- Absences

In proposed method, decision tree, random forest, and logistic regression methods are used.

The flow chart of proposed method is

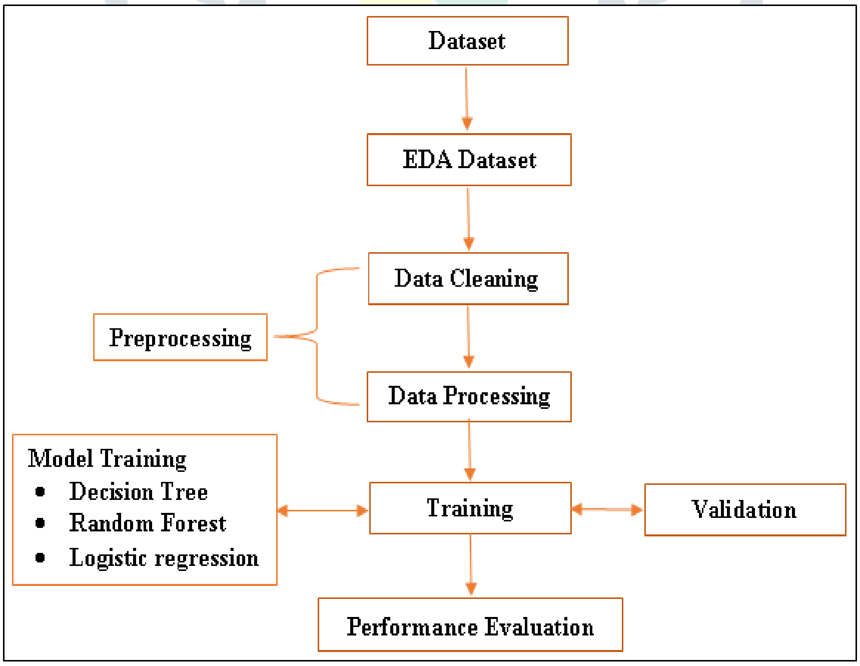


Figure 2: Flow chart of Proposed Method1

Steps of proposed method are as following:

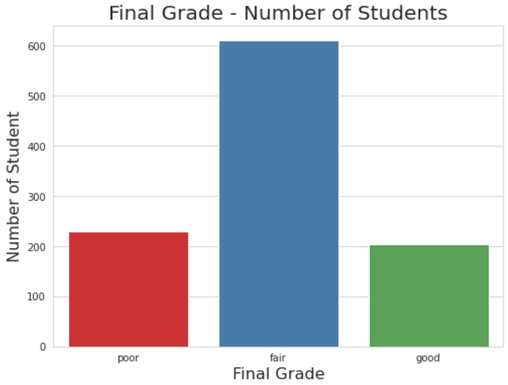
Step 1: Collect the dataset this dataset contains Student academic performance reports.

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

Step 3: Plotting graphs and Performing Exploratory Data Analysis (EDA).

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.

Correlation Heatmap

A correlation heatmap is a heatmap that depicts a two-dimensional correlation matrix between two discrete dimensions, with coloured 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 colour 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 colorbar aids a correlation heatmap, much like a standard heatmap, in making data readily legible and understandable.

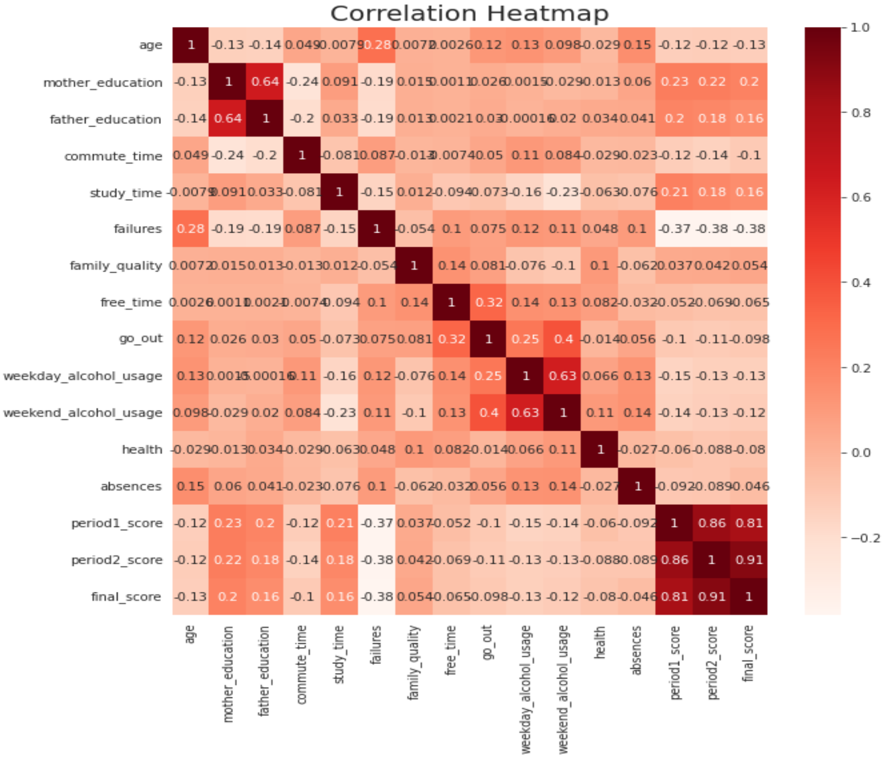


Figure 3 : Correlation Heat maps

All other characteristics are optional except data, which will clearly be the data to be plotted. To build a correlation heatmap, the data must be provided to the corr() function. Furthermore, when constructing a correlation heatmap, corr() excludes columns that will be of little value and picks those that will be useful age, mother education, father education, commute time, study time, failures, family quality, free time, go out, health, absences variables are used.

Step 4: Processing

During the preprocessing stage, some features were discarded due to the lack of discriminative value. For example:
Family income.

- 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, and 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.

RESULT AND DISCUSSION:

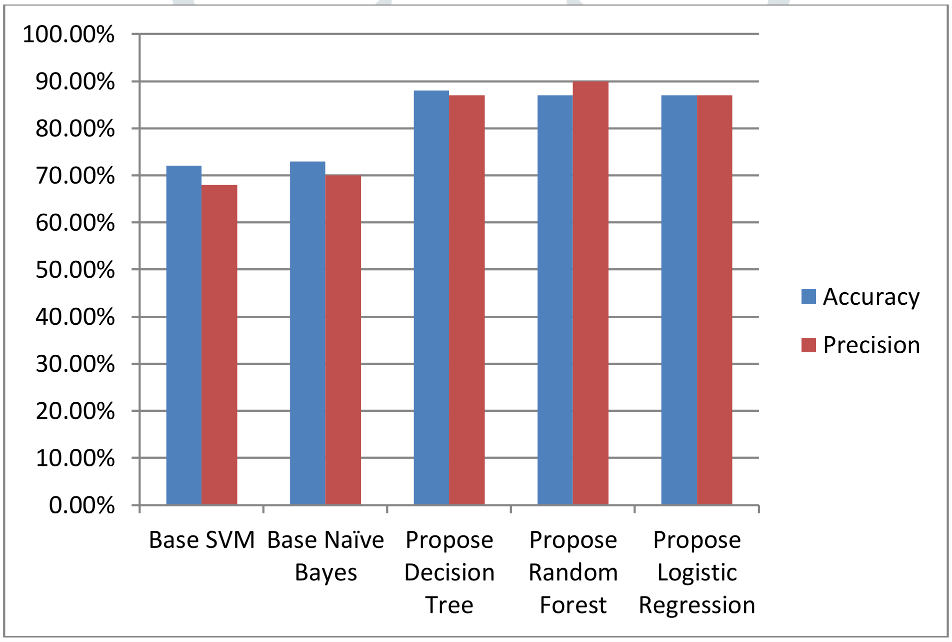
The results of proposed methods are as follows:

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
  -

Model	Accuracy	Precision	Recall	F1
Base SVM	72%	68%	72%	66%
Base Naïve Bayes	73%	70%	72%	71%
Propose Decision Tree	88%	87%	89%	88%
Propose Random Forest	87%	90%	84%	86%
Propose Logistic Regression	87%	87%	87%	87%

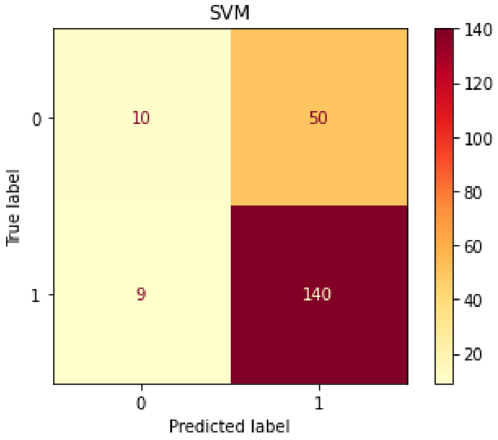
Graphical representation of the result is shown is Graph2 .

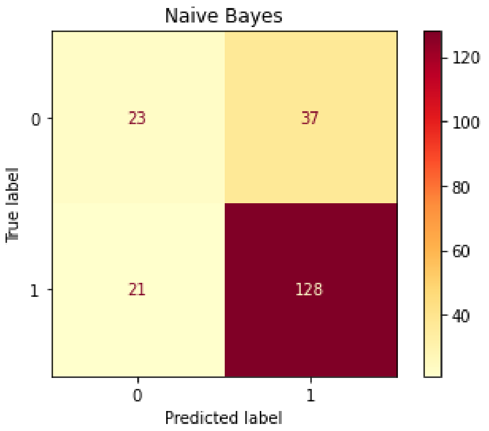


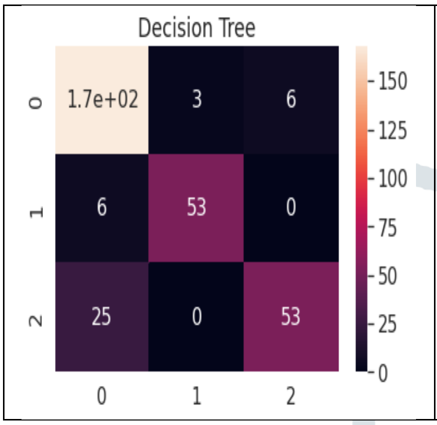
Graph 2 Graphical representation of Proposed method 1

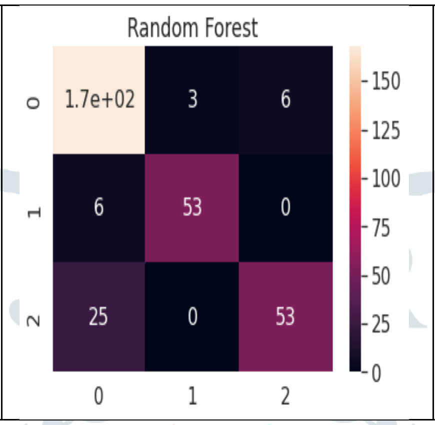
From the above table and graphs it is clear that for Base SVM model the accuracy is 0.72, precision is 0.68, and frequency (F1) is 0.66, for Base Naïve Bayes the accuracy is 0.73, precision is 0.70, and frequency is 0.71, for Propose Decision Tree the accuracy is 0.88, precision is 0.87, and frequency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

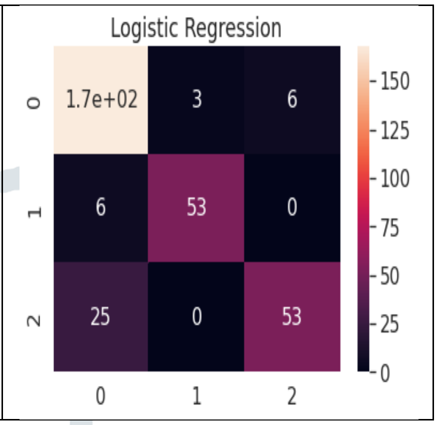
- **Confusion Matrix:**











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.

Students' academic achievement is used in the performance model to calculate their level of competence. Using the model's results will help raise academic success among kids by providing individualized attention, sufficient support, and extra attention.

Even so many research has been conducted to determine the variables that affect student performance, relatively few have produced predictive models for those variables. For students at the upper secondary level, it is especially important to accurately forecast their academic success so that they may get the support they need throughout the learning process. An academic performance model that has two stages of processes, namely identification of best predictive variables in our study by applying different variable selection techniques and prediction of academic performance by higher secondary students using decision tree, Bayesian nets, neural networks and rough set theory, is developed in the current research. This model has been tested and confirmed to be accurate in the current study.

REFERENCE

1. Acharya, A.; Sinha, D. (2014). “Early prediction of students performance using machine learning techniques.” International Journal of Computer Applications (0975 – 8887) Volume 107 – No. 1, DOI:10.5120/18717-9939
2. Chen, M. S., Han, J., & Yu, P. S. (2016). “Data mining: an overview from a database perspective.” IEEE Transactions on Knowledge and data Engineering, Vol. 8, No. 6, 1996, pp. 866-883. doi:10.1109/69.553155
3. E. FrankandI., H. Witten (2005). “Data Mining: Practical Machine Learning Tools and Techniques”, 2nded., SanMateo, CA:Morgan Kaufmann, pp.664.
4. Elena Susnea (2009), “Using Data mining Techniques in Higher Education”, University of Bucharest and “Gh. Asachi” Tehnical University of Iasi. 4th International Conference on Virtual Learning(ICVL) pp. 371-375,2009.
5. Feng, G., Fan, M., & Chen, Y. (2022). “Analysis and Prediction of Students’ Academic Performance Based on Educational Data Mining.” *IEEE Access*, Vol- *10*, 19558-19571.
6. 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, Elsevier, vol. 94(C), pages 335-343, DOI: 10.1016/j.jbusres.2018.02.012
7. Minaei-Bidgoli, B., Kashy, D., Kortemeyer, G. and Punch, W.(2003), “Predicting student performance: an application of data mining methods with an educational web-based system,” Proceedings of IEEE Frontiers in Education, Colorado, USA, 13–18. DOI:10.1109/FIE.2003.1263284
8. Pradeep, A. and Thomas, J. (2015). “Predicting college students dropout using EDM techniques.” International Journal of Computer Applications (0975 – 8887) Volume 123 – No.5, DOI:10.1109/ICSNS.2015.7292372
9. Xu, J.; Moon, K.; Schaar, M.D (2017). “A machine learning approach for tracking and predicting student performance in degree programs.” IEEE J. Sel. Top. Signal Process., Vol-11, pp. 742–753.DOI: https://doi.org/10.1109/JSTSP.2017.2692560.