EDUCATIONAL DATA MINING PREDICTION MODEL USING DECISION TREE ALGORITHM

Tejashree U. Sawant
Ph. D. Research Scholar
Dept of Computer Science
Shivaji University, Kolhapur

Dr. Urmila R. Pol
Asst. Professor
Dept of Computer Science
Shivaji University, Kolhapur

Dr. Pratibha S. Patankar
Professor and Head
Dept of Education
Shivaji University,
Kolhapur

Abstract: This paper focused on improving student academic performance, based on their personal and academic performance features. This approach of prediction student result at the final semester will notify the teachers to know about student performance and it gives as a chance to improve their performance in future. The dataset used for the research purposes includes data about students' performance from the academic and other factors which may affect the performance. Educational data mining decision tree algorithm is used to predict the student performance.

Index Terms: Performance, prediction model, Educational data mining algorithms, decision tree

I. INTRODUCTION

EDM is defined as an amalgamation of techniques involved in analyzing student data to extract hidden knowledge regarding students. The result of EDM can be used to make better decisions to improve teachers and student performance, which will directly help to enhance the overall performance of the university. EDM helps in predicting student performance, which is the main area of our research. Predicting student performance is an activity which highlights inferring knowledge from student performance data to understand. The result of prediction of student performance is to improve higher education effectiveness, when the number of predicted failures are high, it might be because of ineffective syllabus or because of the other factors leading to failures that can be reassessed and improved. This provides better ability to predict student performance at an early enough stage to potentially find ways to increase the success rate.

II. LITERATURE REVIEW

Classification task is used on student database to predict the students division on the basis of previous database and the decision tree method is used. Information's like Attendance, Class test, Seminar and Assignment marks were collected from the student's previous database, to predict the performance at the end of the semester (Baradwaj & Pal, 2011).

Educational data mining is used to study the data available in the educational field and bring out the hidden knowledge from it. Classification methods like decision trees, Bayesian network etc can be applied on the

educational data for predicting the student's performance in examination. This prediction will help to identify the weak students and help them to score better marks. The C4.5, ID3 and CART decision tree algorithms are applied on engineering student's data to predict their performance in the final exam (Yadav & Pal, 2012).

Authors applied data mining techniques to discover association, classification, clustering and outlier detection rules to improve graduate students' performance, and overcome the problem of low grades of graduate students (Tair & El-Halees, 2012).

The Student - Teacher Model for Enhancement of higher education System is used as a guideline for the application of data mining in higher educational system. With the help of model increased ratio of admissions by counseling students for better selection of courses. Also, it helps to enhance the evaluation and performance of students as well teachers.focuses on capabilities of data mining in context of education by presenting a student –teacher conceptual model for students as well as teachers (Bhanap & Kulkarni, 2013).

Authors in this study focused on data mining techniques to improve the efficiency of academic performance in the educational institutions and evaluated the performance of student by the four selected classification algorithms based on Weka. The best algorithm based on the placement data is IB1 Classification with an accuracy of 82.00% (Pal & Pal, Data Mining Techniques in EDM for Predicting the Performance of Students, 2013).

III. METHOD

Data mining is concerned with the analysis of data and they use different software techniques to find the hidden and unexpected patterns and their relationships in data set. The work of mining the data is to extract the information that is unknown and unpredicted. Generally data mining contains several algorithms and techniques for finding out interesting patterns from large size of data sets. The techniques of data mining are classified into two groups: they are supervised learning and unsupervised learning. In supervised learning, a model is built earlier for the analysis and then it applies the algorithm to the data in order to estimate the parameters for the model. Classification, Association Rule Mining ,Decision Tree, Bayesian Classification, Neural Networks, etc. are some examples of supervised learning. Various data mining techniques such as Classification, Decision Tree, Bayesian Classification, Neural Networks, Clustering, Association Rule Mining, Prediction, Time Series Analysis, Sequential Pattern and Genetic Algorithm and Nearest Neighbor have been used for knowledge discovery from large data sets.

IV. DATASET DESCRIPTION

Dataset contains 262 student records with 40 columns. A data set is prepared in the form of csv file to give training to the machine and testing it. The data has split for data transformation. First 70% of the data can be used for training the model and the remaining 30% of the data is used for testing the model.

V. EXPERIMENTAL RESULTS AND DISCUSSIONS

A decision tree is a flowchart-like tree structure where an internal node represents feature (or attribute), the branch represents a decision rule, and each leaf node represents the outcome. The topmost node in a decision tree is known as the root node. It learns to partition on the basis of the attribute value. It partitions the tree in recursively manner call recursive partitioning. This flowchart-like structure helps in decision making. It's visualization like a flowchart diagram which easily mimics the human level thinking. That is why decision trees are easy to understand and interpret.

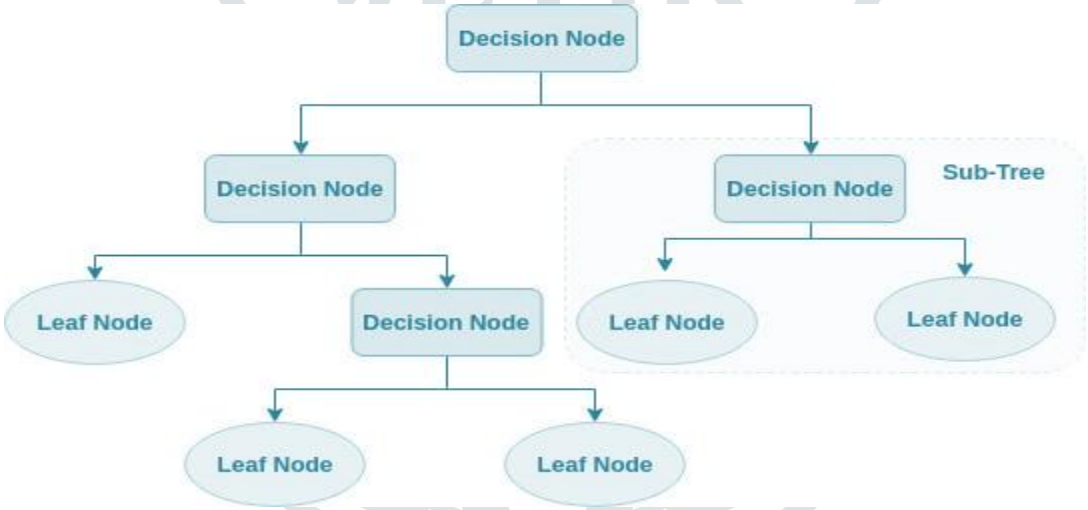

Figure 1. Decision Tree Process

Decision Tree is a white box type of ML algorithm. It shares internal decision-making logic, which is not available in the black box type of algorithms such as Neural Network. Its training time is faster compared to the neural network algorithm. The time complexity of decision trees is a function of the number of records and number of attributes in the given data. Decision trees can handle high dimensional data with good accuracy.

a. Decision Tree

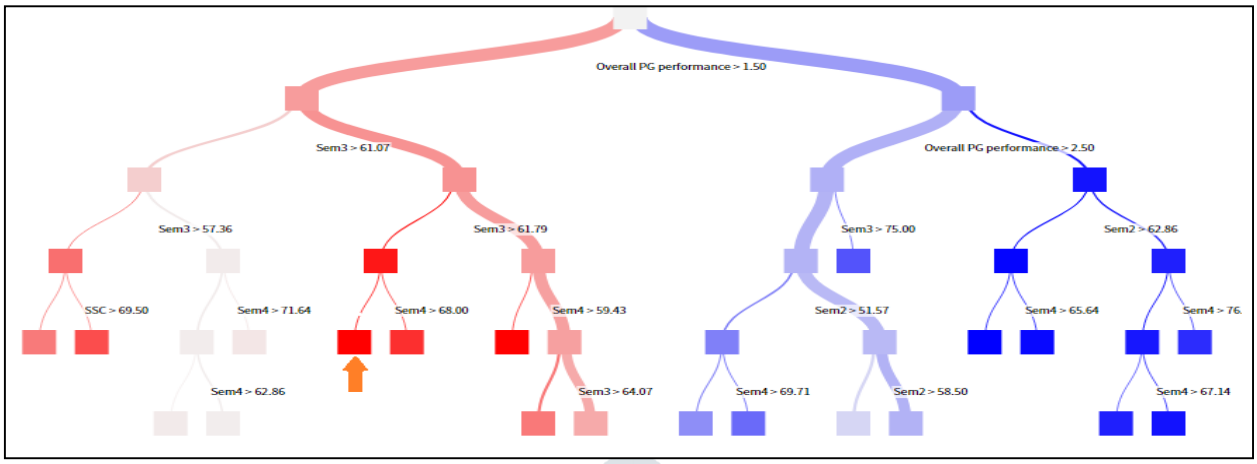

Figure 2. Decision Tree

Above figure is a decision tree obtained having 81% of prediction probability. Also it is observed that feature like overall performance, marks of sem3, Sem4 and Sem 2 has higher importance in predicting Sem 5 marks respectively.

b. Feature Importance

The importance of a feature is the increase in the prediction error of the model after we permuted the feature's values, which breaks the relationship between the feature and the true outcome.

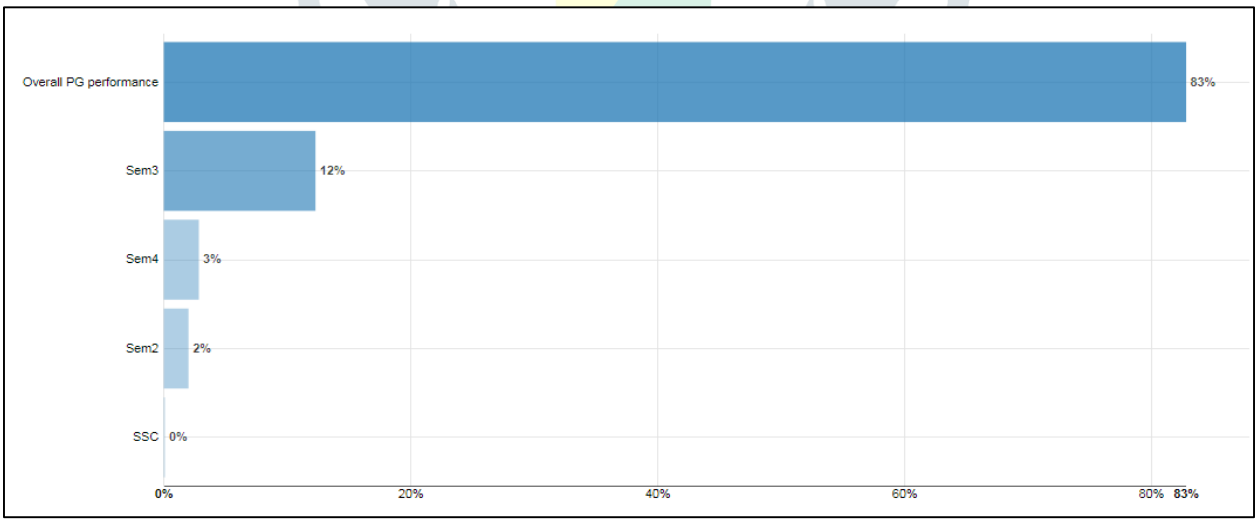

Figure 3. Feature Importance

The Feature importance for decision tree model is displayed in figure.

We can interpret from the above figure that the Overall PG performance features from the dataset has higher importance to predict the target variable with 83% followed by Sem3 with 12%, Sem4 with 3% and Sem2 with 2%.

c. Scatter Plot for Predicted value Vs Actual value

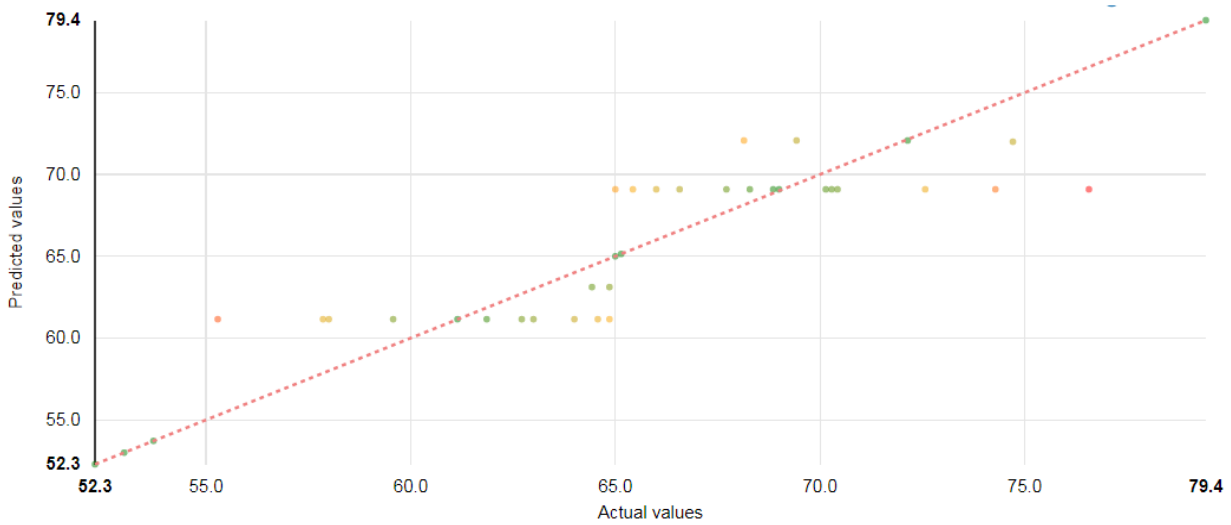

Figure 4. Predicted value Vs Actual value

The above figure shows the scatter plot for Predicted value and actual values. Almost 81% of the values are correctly predicted by the decision tree algorithm. The dotted red line is the perfect matched values for predicted and actual.

d. Error Distribution

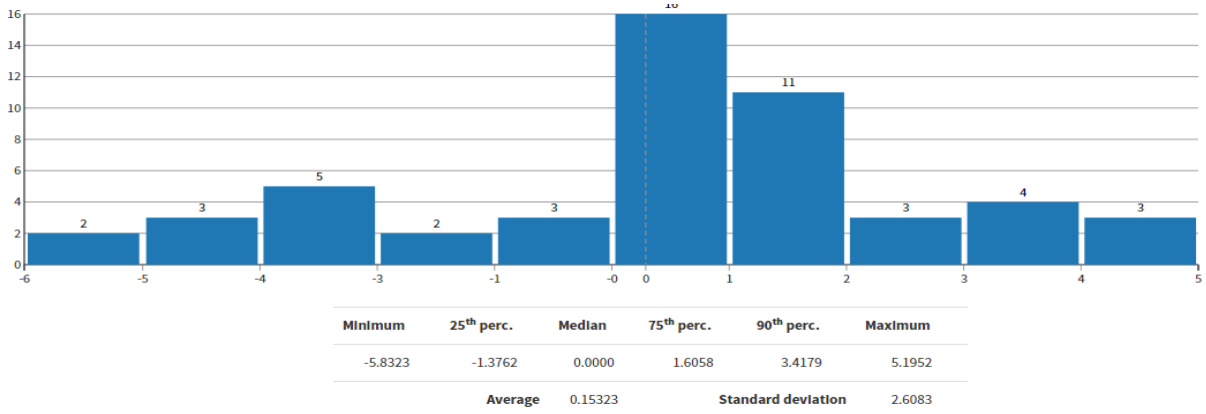

Figure 5. Error Distribution chart

This figure shows the error distribution with minimum -5.83 to Maximum 5.19 value. The error value should lie in the centre to obtain the best fit model.

e. Partial Dependency

The partial dependence plot (short PDP or PD plot) shows the marginal effect one or two features have on the predicted outcome of a machine learning model (Friedman & H., 2001).

A partial dependency plot shows the dependence of the predicted response on a single feature. The log-odds for a probability p are defined as log(p / (1 - p)). They are strictly increasing, i.e. higher log odds mean higher probability.

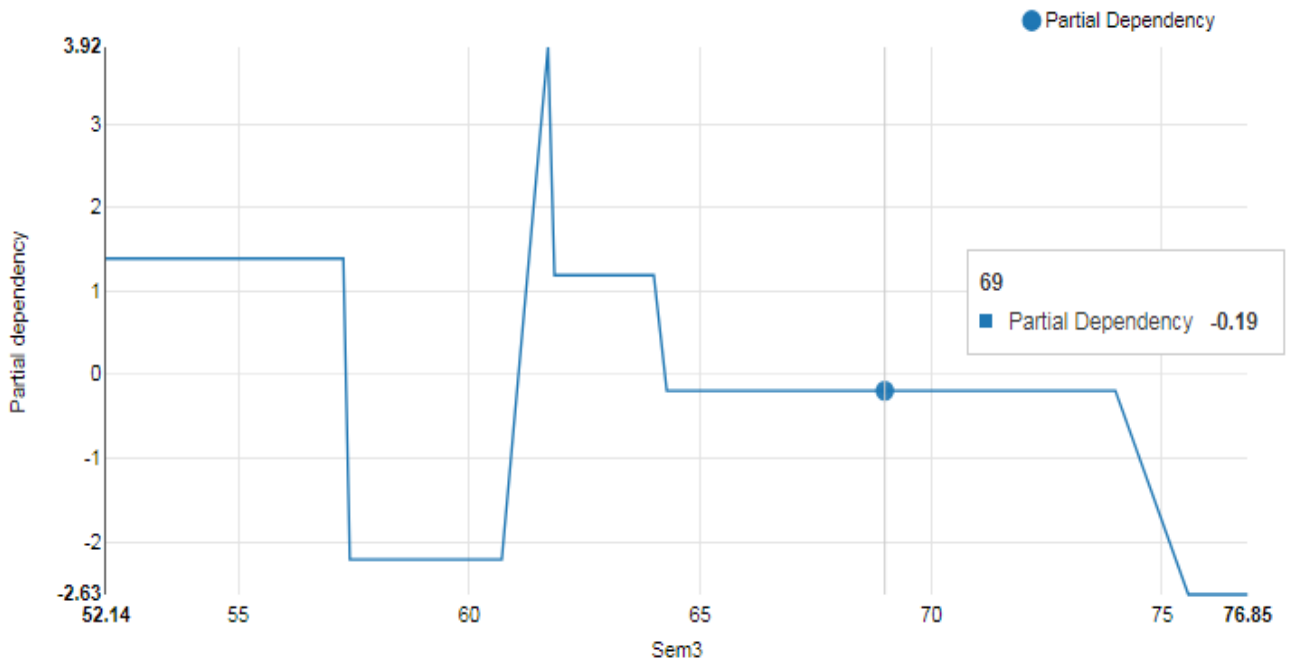

Figure 6. Partial Dependency chart

In above Fig.6 x axis displays the value of the selected feature here sem3 feature is selected to check the partial dependency on the sem5 feature which is a target or predicted variable and the y axis displays the partial dependence. The value of the partial dependence is by how much the log-odds are higher or lower than those of the average probability. When the value of sem3 feature is 69 then partial dependence value is -0.19 respectively.

f. Predicted Dataset with Decision Tree

exam stress	Overall PG performance	prediction	error	error_decile	abs_error_decile
Integer	Integer	Decimal	Decimal	Integer	Integer
2	2	61.1535181237	-0.0106609808102	5	0
1	1	69.0904761905	1.19523809524	6	2
1	1	69.0904761905	5.19523809524	9	8
1	1	69.0904761905	-1.37619047619	4	2
2	2	61.1535181237	-5.832257285	0	9
1	1	72.0761904762	-2.64761904762	2	4
1	1	69.0904761905	1.05238095238	6	1
1	1	65.1428571429	0.0	5	0
1	1	69.0904761905	-0.233333333333	5	0
1	1	69.0904761905	-1.37619047619	4	2
1	1	69.0904761905	-3.09047619048	2	5
1	1	65.0	0.0	5	0
1	1	69.0904761905	5.19523809524	9	8
1	1	69.0904761905	1.3380952381	6	2

Figure 7. Dataset with New Prediction feature

This figure shows dataset with predicted data values and error values for each student record. The new feature gets added into the dataset i.e. prediction and error between actual and predicted values.

g. Evaluation of model:

This table describes the performance of built model for student performance.

Performance Metrics of Decision Tree Algorithm	
Explained Variance Score	**0.81247**
Mean Absolute Error (MAE)	**1.9840**
Mean Average Percentage Error	**3.01%**
Mean Squared Error (MSE)	**7.2690**
Root Mean Squared Error (RMSE)	**2.6961**
Root Mean Squared Logarithmic Error (RMSLE)	**0.040364**
Pearson Coefficient	**0.90137**
R2 Score	**0.81148**

We can observe that our R2 score and Pearson Coefficient are both very good. This means that we have found a good fitting model to predict the student performance. There can be a further improvement to the metric by doing some preprocessing before fitting the data. With all the performance metrics, it was clear from the results, that the enhancement made with Decision Tree is successful.

VI. CONCLUSION AND FUTURE WORK

This work implements decision tree algorithm that improve the student performance prediction. Through this study we conclude that the decision tree algorithm finds to be 81% of accuracy. Our future work would to implement other data mining techniques and compare the algorithm results to find the best data mining technique for prediction of student academic performance in higher education.

VII. REFERENCES

1. Baradwaj, B. K., & Pal, S. (2011). Mining Educational Data to Analyze Students' Performance. (IJACSA) International Journal of Advanced Computer Science and Applications , 63-69.
2. Yadav, S. K., & Pal, S. (2012). Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification. World of Computer Science and Information Technology Journal (WCSIT) , 51-56.
3. Tair, M. M., & El-Halees, A. M. (2012). Mining Educational Data to Improve Students' Performance:A Case Study. International Journal of Information and Communication Technology Research , 140-146.
4. Bhanap, M. S., & Kulkarni, M. R. (2013). Student - Teacher Model for Higher Education System. Current Trends in Technology and Science , II (III), 258-261.
5. Pal, A. K., & Pal, S. (2013). Data Mining Techniques in EDM for Predicting the Performance of Students. International Journal of Computer and Information Technology (ISSN: 2279 – 0764) , 1110-1116.