ENHANCING STUDENT PERFORMANCE THROUGH DATA ANALYTICS PREDICTION

M.S.Sassirekha[1],
[1]Department of Applied Mathematics and Computational Science, Thiagarajar College of Engineering,
Madurai-625015

Abstract:

It is becoming more and more important to comprehend and maximise student achievement in the quickly changing academic environment of today. In order to solve this problem, educational institutions are using data analytics more and more to put in place predictive algorithms that can recognise students who are at-risk and adjust treatments accordingly, leading to improved retention rates and better academic results overall. Specifically, the application of machine learning algorithms has demonstrated potential in the analysis of diverse student-related data, including demographics, academic background, and behavioural tendencies, for the purpose of predicting performance trajectories and creating customised support plans for individual students. Furthermore, studies have demonstrated that automated systems capable of real-time prediction of student performance can enhance decision-making processes regarding timely educational interventions, thereby fostering greater engagement and motivation among students. Moreover, the integration of these predictive models not only aids in early identification of students who may require additional support but also provides institutions with insights to refine their educational practices and resource allocation, ultimately creating an effective and responsive environment for learning. Through the integration of multiple variables, including socioeconomic status, time management proficiency, and involvement with educational materials, educational institutions can create advanced predictive models that greatly improve their capacity to assist students who may be at risk of performing inadequately or halting their studies.

Keywords: Big data, types of analytics, higher education, prediction

1. INTRODUCTION

Big data analytics is the study and examination of "big data," or massive, complicated data sets. The important insights, patterns, and trends from this study to help make better decisions. Meaningful information from large datasets is processed, managed, and examined using a variety of methods, instruments, and technologies.

When data becomes huge, it is difficult for conventional data processing techniques to handle it effectively, we usually turn to big data analytics. More data means more demands on a wider range of analytical techniques, faster processing speeds, and larger data storage capacities. To extract insights, big data analytics includes multiple phases and procedures.

 https://doi.org/10.5281/zenodo.14271733

Here's a brief summary of how this can appear:

Data collection: Compile information from a set of resources, including databases, websites, social media, questionnaires, and transaction records. The data in question may be semi-structured, unstructured, or structured.

Data storage: Use cloud-based or distributed technologies to store data. These kinds of storage offer fault tolerance and can manage big data volumes.

Preparing data: Cleaning and preparing the raw data is a good idea before doing any analysis. In order to handle missing numbers, standardize formats, deal with outliers, and organize the data into a more appropriate manner, several steps may be necessary.

Integration of data: Data typically originates from multiple sources in disparate formats.

Data processing: Distributed frameworks are a good choice for most organizations when handling large amounts of data [1]. They divide the jobs into manageable portions and split them among several machines so they can be processed in parallel.

Data analysis techniques: You'll probably use a combination of data analysis techniques, depending on the analysis's objectives. These could include machine learning, text mining, exploratory analysis, and other techniques for descriptive, predictive, and prescriptive analytics [2].

Data visualization: Use visual aids like as dashboards, graphs, charts, and other tools to convey the outcomes of the analysis. You may explain complex ideas in a clear and concise manner by using visualization.

Interpretation and decision-making: Use the knowledge gleaned from your analysis to draw inferences and make decisions supported by facts. These choices have an effect on operations, procedures, and corporate strategy.

2. DATASET

The dataset "MCA_Student_Dataset _train_table_1.csv" comprises 396 rows and 11 columns, capturing various academic performance metrics of MCA students. Key columns include "10th overall," "12th overall," "UG overall," "CAT1," "CAT2," "CAT3," "CA," "TR," "Attendance," "Total," and "Result." These columns represent scores from different educational stages and assessments, as well as attendance and overall results [13].

From the first five sample entries, we observe that students generally have high scores in their 10th and 12th grades, with values ranging from 86 to 95 and 61 to 85, respectively. The "UG overall" scores show a slightly lower range, from 66 to 80. Continuous Assessment (CA) and Term Results (TR) scores vary, with CA scores between 31 and 35 and TR scores between 26 and 41. Attendance percentages are relatively high, ranging from 62% to 71%.

 https://doi.org/10.5281/zenodo.14271733

The "Total" column, which likely aggregates various performance metrics, shows scores between 61 and 76. The "Result" column, an enum type, indicates whether a student passed (1) or failed (0). All five sample entries have a "Result" value of 1, indicating that these students passed.

Overall, the dataset provides a comprehensive view of student performance across multiple academic stages and assessments, where patterns can be identified as trends and areas for improvement in educational outcomes.

Based on the first five sample entries, the students generally exhibit strong academic performance. They attain high scores in their 10th and 12th grade evaluations, ranging from 86 to 95 and 61 to 85, respectively. However, their overall undergraduate scores show a slightly lower range, falling between 66 and 80. The Continuous Assessment and Term Result scores vary, with CA scores falling between 31 and 35, and TR scores ranging from 26 to 41. Attendance percentages are relatively high, spanning 62% to 71%.

3. MATERIALS AND METHODS

3.1 DESCRIPTIVE ANALYTICS

Below are the factors that are used for further processing.

Assessment Scores: Lower mean scores in 'CAT1' and 'CAT2' suggest a more rigorous grading system or more challenging course content [3].

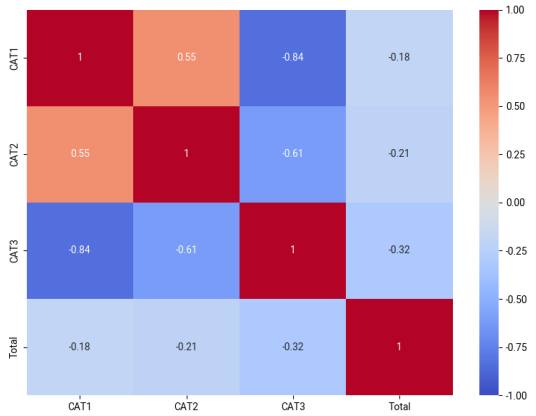
Variability: The relatively higher standard deviation in 'UG overall' indicates greater fluctuation in university grades compared to high school.

Minimum Scores: The significantly lower minimum scores in 'UG overall' and 'CAT1' may point to outliers or instances of exceptionally poor performance. This analysis provides a comprehensive overview of the distribution of scores across diverse educational and assessment categories, highlighting areas of strength as well as potential areas of concern [16].

Result Categories: The 'Result' column categorizes students as either passing ('1') or failing ('0').

Distribution of Results: 386 students passed, while 10 students failed.

The figure 1 visualizes the distribution of results, clearly showing a significant difference between the number of promoted and not promoted students. The area for passing students is substantially higher than the bar for failing students, indicating a much higher pass rate. The distribution of the 'Result' column indicates that the vast majority of students successfully passed, while only a small number failed, suggesting a high success rate among the assessed students.



https://doi.org/10.5281/zenodo.14271733

Fig 1 Contribution of CAT1, CAT2, and CAT3 scores to the Total score

Students demonstrate superior average achievement in '10th overall' and '12th overall', with a gradual decline in scores at the university level. Lower mean scores in 'CAT1' and 'CAT2' suggest either a more rigorous grading system or more challenging course material. The relatively higher standard deviation in 'UG overall' in figure 2 indicates greater fluctuation in university grades compared to high school. The significantly lower minimum scores in 'UG overall' and 'CAT1' may point to outliers or instances of exceptionally poor performance [4]. This analysis provides a comprehensive overview of the distribution of scores across diverse educational and assessment categories, highlighting areas of strength as well as potential areas of concern.

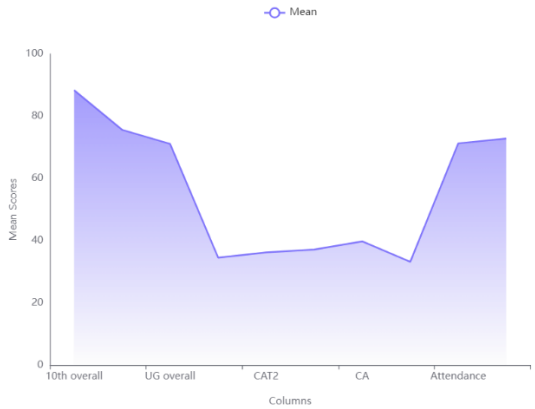


Fig 2 The distribution of scores to each factors: 10th overall, 12th overall, UG overall, CAT1, CAT2, CAT3, CA, TR, Attendance, and Total

3.2 DIAGNOSTIC ANALYTICS

The study employed three machine learning feature selection techniques on a student dataset, which yielded a set of features that play a significant role in the predictive model. While each of the three methods identified the top ten most influential factors based on their respective principles, discrepancies were observed when comparing the selected features across the techniques. To address this issue, 19 unique features were extracted from the full set of 30 features, with their respective frequencies of occurrence considered. Ultimately, the top 10 scoring features were selected as the optimal feature subset for accuracy calculation.

Furthermore, the selected features included demographic factors, academic performance indicators, and behavioral patterns, which collectively enhanced the model's ability to predict students at risk of underperforming, thus facilitating timely interventions to support their academic journey [11].

Using the advantages of both filter and wrapper strategies, a unique hybrid methodology was created and used in this research project to find crucial components that provide more accurate output factor predictions [13]. The fundamental ideas behind the filter and wrapper feature selection strategies are different. Subsets of characteristics are chosen using the filter approach, such as feature importance, according to how they relate to the target variable [13]. On the other hand, the wrapper approach looks for useful feature combinations, just like recursive feature elimination. A descriptive study of the features and the target variable was carried out using scatter plots [13]. In order to determine the optimal feature subset and maintain the benefits of both approaches, a methodology that

https://doi.org/10.5281/zenodo.14271733

For the feature selection process, an approach is described that preserves the benefits of both filter and wrapper methods by integrating their ideas to determine the best feature subset. When using statistical techniques and selection procedures, the univariate method is utilised to forecast independent variables. This method is used for segmentation analysis as well as the training set [5].

By implementing this feature selection approach, the most reliant features are identified. In this study, the recursive feature elimination method is employed to evaluate the quality of the input features. To gauge the uncertainty, the resulting set of conditions can be monitored using a model. The recursive elimination procedure produces a matrix and a vector, in addition to other outcomes such as ranking and prediction. Another step in the elimination process involves selecting one attribute and comparing it to other features. Prior to incorporating features in a predictive model, certain algorithms necessitate the assessment of their relevance. The characteristics were then presented with the frequency generated and aggregated across the three feature selection methods utilized, and ordered based on the frequency obtained through the hybrid approach. Setting a feature frequency threshold is a basic method of feature selection [9]. Any features whose frequency falls below the threshold are eliminated. It assumes, by default, that features with higher frequency hold more meaningful data.

The outcome was the feature set Xi, based on the features selected using three popular strategies: univariate, feature significance, and recursive feature removal technologies, with a cumulative frequency as rank. All of the attributes were arranged in descending order after frequency was determined.

Important features that satisfied the frequency criteria with the highest rank were statistically chosen so that the model could perform more accurately. correlation between the chosen features and the goal utilising the suggested approach [8]. Since the amount of features used and execution time are strongly correlated, optimising for both can yield the maximum accuracy when utilising the fewest features and the shortest execution time.

3.3 PREDICTIVE ANALYTICS

Choosing hybrid features and sifting the important features was the first step in creating a new dataset. As demonstrated in Figure 3, It employs a hybrid feature selection strategy to eliminate significant traits that are shared by the three approaches (univariate, feature importance, and recursive feature removal). To construct a new dataset, some features from the entire feature set were kept, and others were removed. Next, one should find the machine learning model for the suggested framework.
Observations revealed that one of the most widely used supervised machine learning techniques was in use [13]. These techniques included KNN, naive bayes, random forest, logistic regression, support vector machine, linear discriminant analysis, classification and regression tree, and so on. Following the completion of all model runs using the training sample set, It was found that the most recent suggested model varied based on the input attributes after all of the models were run using the training sample set [10]. It depends on several factors, such as the F1 measure, memory, precision, and kappa value. It was discovered that logistic regression and KNN were the most effective classification methods [15].

After that, the final, highly accurate model was set as the real model, and the test dataset was prepared and forecasted. Early interventions could be provided based on the predictions to assist the student in passing the exam with their best effort on the first attempt.
The proposed method computes the relevance of each feature in the feature collection. The importance of each feature is ascertained in the same manner, after which the frequency and rank calculations are made based on the modal existence of each characteristic. For prediction, the feature with the highest ranking was chosen.

 https://doi.org/10.5281/zenodo.14271733

The logistic regression and classification and regression tree models had the highest computed accuracy out of all the models utilised for classification. With likelihood, probabilities are found as a forecast that falls inside the logistic regression. Thus, for every training data point x, the predicted class was y. The optimal fitting algorithm that outperformed every other algorithm was provided by the suggested framework.

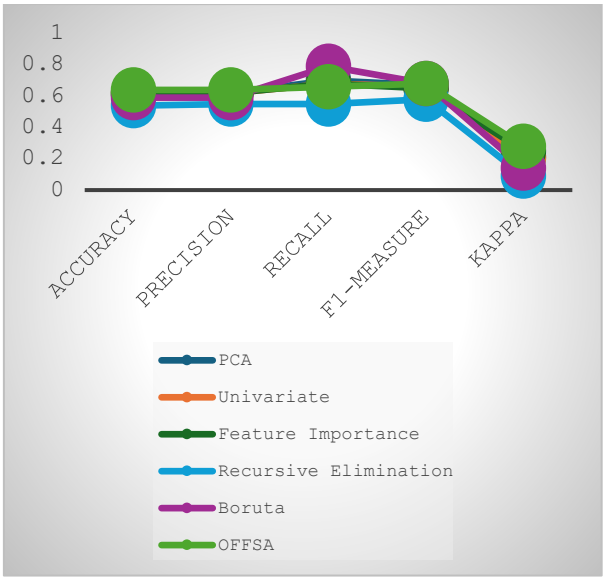


Fig 4 Comparison of various feature selection methods with proposed method

3.4 PRESCRIPTIVE ANALYTICS

There were five experiment groups that were utilized in the ANOVA analysis: 1, 2, 3, 4, and 5. Factors are independent variables, such as treatments. There are five distinct types of experiments, hence there are five levels in the experimental factor. The one-way ANOVA approach was employed for analysis because the X-axis represents the factor or independent variable to evaluate and the Y-axis represents the control and other experimental groups that were previously mentioned.

Various Groups			Group 1	Group 2	Group 3	Group 4	Group 5
The increase in student's academic performance	Control Group(n=50)	Mean	0.63	0.72	0.68	0.75	0.79
		SD	0.12	0.15	0.15	0.13	0.11
	Experiment Group(n=50)	Mean	0.67	0.8	0.69	0.79	0.81
		SD	0.07	0.09	0.09	0.15	0.09

Fig 5 Observation between experiment and control groups

Figure 3 illustrates how the qualities selected to enhance students' academic performance through the experimental group were clearly superior than those learned through the control group, both when taken as a whole and when compared to other groups. This is evident from the previous explanation [11]. Furthermore, as Figure 6 illustrates, learning elements and groups have an effect on a student's improvement in their academic performance when the components that determine the rise are analyzed. In order to support the earlier description of improved student academic performance, statistical studies were done. In particular, the results of the normality test showed that the academic performance data of the students was distributed regularly. This was the basis for the use of the pairwise t-test Analysis of Variance to investigate

https://doi.org/10.5281/zenodo.14271733

variation ns in mean performance and the one-way ANOVA to create simultaneous confidence intervals.

Descriptive analytics was used to identify key properties. After selecting one of these elements, the work of fixing it was started. Exam 2's characteristic was selected as the focal point, and 600 students were randomized to one of five experimental groups: 1, 2, 3, 4, or 5. Making decisions was made easier by applying prescriptive analytics to predict student performance. The ANOVA technique was utilized to show improvement based on the treatments implemented. Instructions on what to do and how to do it correctly the first time are additional benefits [14]. Decision-makers can make decisions that will lead to long-term success and growth by having simultaneous access to projected and present data. It offers suggestions to facilitate decision-making. Furthermore, less effort is spent debugging issues and more time is spent developing the optimal solutions thanks to real-time data analytics and outcome forecasts. This approach also reduces bias and human mistake. Comparatively speaking to individual-based or descriptive methods, predictive analytics employing advanced algorithms and machine learning techniques allows for a wider range of data collection and analysis.

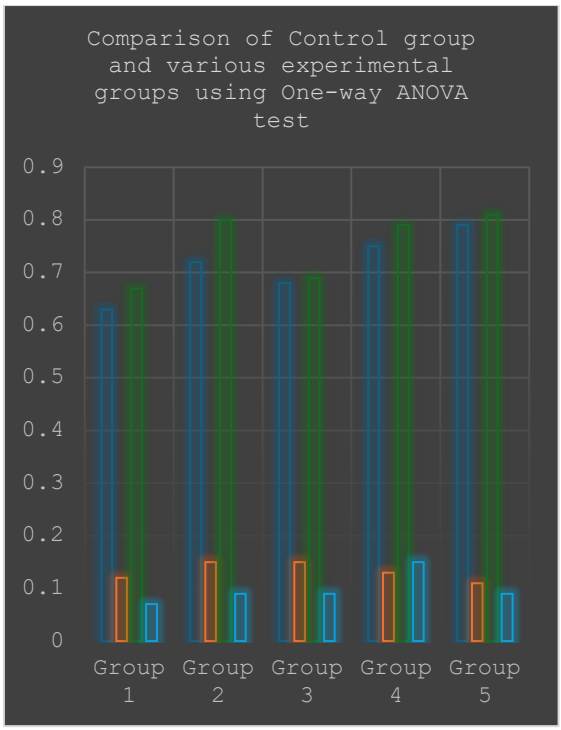


Fig 6 Bar chart for comparison among various group values

4. CONCLUSION

A technique for tracking and forecasting academic progress in postsecondary education is presented in this work. The aim of this methodology was to yield the most precise predictions so that a system for customised learning could be constructed in a subsequent study. This research endeavour determined the current state of the art in student performance prediction.

It also eliminated the traits that were essential for predicting pupils' success in school. It shows that, by employing many performance evaluation criteria, the proposed method confirms the actual combination of features selection technique that predicted the student's academic accomplishment. Additionally, it got rid of the characteristics that were crucial for predicting students' academic progress. It demonstrates how the suggested method validates

 https://doi.org/10.5281/zenodo.14271733

the real feature selection technique that anticipated the student's academic progress by using many performance evaluation criteria. Based on the data collected for this review, supervised learning is the most popular approach for behaviour prediction since it produces reliable and accurate results. Despite the limited amount of data available for more study, the scientists found that the SVM approach was the most effective and generated the most accurate predictions. Apart from SVM, DT, NB, and RF are other well-researched algorithmic concepts that yield useful results.

The likelihood of a student improving academically can be predicted using a variety of supervised machine learning methods. Logistic regression, support vector machines, random forests, naive bayes, and linear discriminant analysis were the six methods applied on the dataset. The real-time dataset was pre-processed using many data preparation techniques to assess the effectiveness of the algorithms. The updated data was then used to build train and test data sets. 90% of the time, the results of applying the six recommended methodologies and characteristics to the dataset showed that the random forest model performed better than all other machine learning models when it used the suggested feature selection methodology within the proposed framework.

REFERENCES:

[1] Abubakar, Y & Ahmad, NBH 2017, ‘Prediction of students’ performance in e-learning environment using random forest’, International Journal of Innovative Computing, vol. 7, no. 2, pp. 1-5.
[2] Abu Saa, A., Al-Emran, M., & Shaalan, K. 2019, ‘Mining student information system records to predict students’ academic performance’, International conference on advanced machine learning technologies and applications, Springer, P. 229–239.
[3] Adekitan, AI & Noma-Osaghae, E 2019, ‘Data mining approach to predicting the performance of first year student in a university using the admission requirements’, Education and Information Technologies, vol. 24, no. 2, pp. 1527-1543.
[4] Buenaño-Fernández, D, Gil, D & Luján-Mora, S 2019, ‘Application of machine learning in predicting performance for computer engineering students: A case study’, Sustainability, vol. 11, no. 10, P. 2833.
[5] Bunkar, K & Tanwani, S 2020, ‘Student performance prediction using C4. 5 decision tree and CART algorithm’, Parishodh Journal, vol. IX, no. II, pp. 1702-1716.
[6] Burman, I & Som, S 2019, ‘Predicting student’s academic performance using support vector machine’, in 2019 Amity International Conference on Artificial Intelligence (AICAI), IEEE, pp. 756-759.
[7] Chung, JY & Lee, S 2019, ‘Dropout early warning systems for high school students using machine learning’, Children and Youth Services Review, vol. 96, pp. 346-353.
[8] Dangi, A & Srivastava, S 2020, ‘An application of student data to forecast education results of student by using classification techniques’, Journal of Critical Reviews, vol. 7, no. 14, pp. 3339-3343.
[9] Khan, A & Ghosh, SK 2021, ‘Student performance analysis and prediction in classroom learning: A review of educational data mining studies’, Education and Information Technologies, vol. 26, no. 1, pp. 205-240.
[10] Khanal, SS, Prasad, PWC, Alsadoon, A & Maag, A 2020, ‘A systematic review: Machine learning based recommendation systems for e-learning’, Education and Information Technologies, vol. 25, no. 4, pp. 2635-2664.

 https://doi.org/10.5281/zenodo.14271733

[11] Rai, S, Shastry, KA, Pratap, S, Kishore, S, Mishra, P & Sanjay, HA 2021, 'Machine learning approach for student academic performance prediction', in Evolution in Computational Intelligence, Springer, Singapore, pp. 611-618.

[12] Rastrollo-Guerrero, JL, Gomez-Pulido, JA & Durán-Domínguez, A 2020, 'Analyzing and predicting students' performance by means of machine learning: A review', Applied Sciences, vol. 10, no. 3, P. 1042.

[13] Sassirekha, MS & Vijayalakshmi, S 2021, 'Student academic performance prediction in higher education by means of machine learning: A review', Design Engineering, pp. 16139-16151, Available from: <http://thedesign engineering.com/index.php/DE/article/view/6761>.

[14] Sekeroglu, B, Dimililer, K & Tuncal, K 2019, 'Student performance prediction and classification using machine learning algorithms', in Proceedings of the 2019 8th International Conference on Educational and Information Technology, pp. 7-11.

[15] Zaffar, M, Hashmani, MA, Savita, KS & Khan, SA 2021, 'A review on feature selection methods for improving the performance of classification in educational data mining', International Journal of Information Technology and Management, vol. 20, no. 1-2, pp. 110-131.

[16] Zaffar, M, Savita, KS, Hashmani, MA & Rizvi, SSH 2018, 'A study of feature selection algorithms for predicting student's academic performance', Int. J. Adv. Comput. Sci. Appl, vol. 9, no. 5, pp. 541-549.

[17] https://hrcak.srce.hr/file/416726

 https://doi.org/10.5281/zenodo.14271733