International Journal of Engineering Trends and Technology
ISSN: 2231–5381 / https://doi.org/10.14445/22315381/IJETT-V72I11P116

Volume 72 Issue 11, 131-143, November 2024
© 2024 Seventh Sense Research Group®

Original Article

Exploratory Data Analysis and Feature Selection for Predictive Modeling of Student Academic Performance Using a Proposed Dataset

Hardik I. Patel[1], Dharmendra Patel[2]

[1,2]*Faculty of Computer Science and Applications, Smt. Chandaben Mohanbhai Patel Institute of Computer Applications, Charotar University of Science and Technology, Gujarat, India.*

[1]*Corresponding Author : hardikipatel.mca@charusat.ac.in*

Received: 01 June 2024 Revised: 07 October 2024 Accepted: 24 October 2024 Published: 29 November 2024

Abstract - *Academic performance prediction is vital for numerous applications. The previous research adopted methodologies that were lacking in scientific approach. Most researchers have predicted student academic performance by focusing on academic parameters. However, social and economic factors also influence academic outcomes. The proposed research encompasses both academic and socio-economic parameters to predict student academic performance more comprehensively. The careful steps of data collection, cleansing, and use of Exploratory Data Analysis (EDA) to improve model prediction accuracy are described in depth in this study. This article depicted research gaps in previous research. This article focuses mainly on exploratory data analysis. Exploratory data analysis is vital for understanding data thoroughly before applying a prediction model. In data science, understanding data is more important than applying predictive algorithms. This proposed research has designed a novel exploratory data analysis algorithm and applied it to the proposed dataset. The article also decided on features that are essential for better prediction of predictive algorithms. This research aims to improve the predictive modeling of student academic success by using a large dataset that includes academic and socioeconomic characteristics. By going beyond conventional academic measures, this study fills a significant research vacuum by acknowledging the variety of factors impacting educational results. Through presenting a comprehensive viewpoint, the study seeks to advance the comprehension of the factors that influence academic achievement. The technique presented here provides a solid foundation for further studies in this area, highlighting the significance of taking a variety of factors into account for a more thorough assessment of student performance.*

Keywords - *Exploratory data analysis, Predictive model, Feature selection, Data preprocessing, Machine learning.*

1. Introduction

A student's academic performance plays a crucial role in their educational journey and has a significant impact on their prospects in the future. Academic performance is a crucial determinant of a student's abilities and accomplishments in the context of higher education. It helps educational institutions evaluate and improve the calibre of education they offer in addition to reflecting individual strengths. Institutions may promote their students' overall development and put strategies for continual improvement into place by keeping a careful eye on academic achievement. (Tadese, Yeshaneh, and Mulu, 2022) (York, Gibson, and Rankin, 2015) It not only reflects their understanding of the subject matter but also their ability to learn, apply, and communicate their knowledge effectively. It is also an indicator of their discipline and dedication to their studies. (Kimbark, Peters, and Richardson, 2017) Higher education institutions use various methods to evaluate student performance, such as exams, assignments, case studies, and presentations. (Bonney, 2015) (Guo, Saab, Post, and Admiraal, 2020). These assessments provide students with opportunities to demonstrate their knowledge and skills and help educators identify areas where they can improve their teaching methods. (Sawant, Gupta, Sharma, and Singh, 2021) Moreover, institutions also use performance data to measure the effectiveness of their programs and make necessary adjustments to improve the quality of education they provide. Early identification of students who are at risk of dropping out is a critical step in reducing dropout rates and improving educational outcomes. (Pati, Hashim, Brown, Fiks, and Forrest, 2011) When schools identify students who are struggling early on, they can provide targeted support to help them overcome challenges and stay on track. (Bañeres, Rodríguez González, Guerrero Roldán, and Cortadas, 2023) Interventions such as tutoring, counselling, and mentoring can help students improve their academic skills, build resilience, and develop a sense of belonging to the school community.

(Oluwayemisi, Akin, and Israel, 2021) These interventions can also help students address personal and social issues that may be affecting their academic performance and provide them with the resources and support they need to succeed (Dagdag, Cuizon, and Bete, 2019). Early identification and intervention enable schools to allocate resources more efficiently, directing support toward students who require it the most. This targeted approach not only benefits individual students by providing tailored assistance but also contributes to the overall enhancement of the educational system. To achieve accurate predictions of student academic performance, it is essential to consider both academic and non-academic parameters. Incorporating these diverse factors ensures a more holistic understanding of the elements that influence student success.

Academic parameters refer to the student's performance in their classes and on assessments, such as grades, test scores, and attendance records. (Khan and Ghosh, 2021) These parameters provide a snapshot of the student's understanding of the subject matter and their ability to apply their knowledge (Jacob, John, and Gwany, 2020). Non-academic parameters refer to factors outside the classroom that can impact a student's academic performance, such as socio-economic status, family background, family income, education of parents, etc (Li and Qiu, 2018) (Rodríguez Hernández, Cascallar, and Kyndt, 2019). These parameters can have a significant impact on a student's ability to succeed academically and should be taken into account when making predictions about their performance (O.K, A.O, M.A, and R.H, 2013). By considering both academic and non-academic parameters, educators can gain a more comprehensive understanding of a student's academic performance and the factors that contribute to it. This information can be used to provide targeted support and resources to help students overcome obstacles and achieve their academic goals.

2. Related Work

Prior research has predominantly focused on utilizing academic parameters such as GPA, test scores, and attendance records to predict student outcomes. While these parameters are undeniably crucial, this study also considers the need to acknowledge the influence of diverse non-academic factors on academic performance.Data mining techniques help to determine different educational effects and outcomes. The outcomes mentioned include student performance (Xing, 2018), retention (Parker, Hogan, Eastabrook, Oke, and Wood, 2006), success (Martins, Migu´eis, Fonseca, and Alves, 2019) satisfaction (Alqurashi, 2018), achievement (Willems, Coertjens, Tambuyzer, and Donche, 2018), and dropout rate (P´erez, Castellanos, and Correal, 2018). These findings demonstrate the potential for data mining to be applied in the educational field to predict and understand important student outcomes (Umamaheswari, Vanitha, Devi, Thephoral and Basha, 2023). Academic histories and baseline data are employed to forecast a student's academic success. Initial academic data comprises the tenth and twelfth results, among other data. Each parameter is given a certain amount of weight, and the admission score is computed using the nominated marks obtained. The writers regarded attendance, scores on the two sessional exams, grades from assignments, and other internal scores as academic records. If a pupil is at risk or not, a forecast is provided. The findings will help educational institutions target those pupils specifically and appropriately for motivational and counselling support. (Mustapha, 2023) (Narayana Swamy and Hanumanthappa, 2012). The academic performance of students in both bachelor's and master's programs was predicted independently for each subject using a decision tree algorithm and fuzzy genetic algorithm. (Mehdi and Nachouki, 2023)

The decision tree algorithm identified a more significant number of students in the "at-risk" category, which prompted lecturers to take additional care of these students (Hasan, Palaniappan, Raziff, Mahmood, and Sarker, 2018). This makes it easier to anticipate improved and nearly perfect scores from the final tests. Because the fuzzy genetic algorithm considers individuals who are between risk and safe to be in a safe condition, the results provide more passed pupils with a sense of accomplishment. Nonetheless, the professors will indirectly draw attention to them. As a result, professors and students will interact cordially. (Hashmia Hamsha, 2016) The student dataset was utilized in combination with a classification model to predict student performance. To improve the model's predictive accuracy, several Data Pre-processing (DPP) Techniques were applied. These included discretization, correction of invalid entries, removal of missing values, and the use of the Synthetic Minority Oversampling Technique (SMOTE) to address class imbalance.

These pre-processing steps were essential in preparing the dataset for effective modeling and ensuring robust classification outcomes. (Liang, Jiang , Li, Xue, and Wang, 2020) (Pamela Chaudhary, 2016) The feature selection method is used on the dataset attributes, which were analysed to determine if it was possible to forecast student performance using a smaller collection of features while retaining system accuracy (Zaffar, Savita, Hashmani, and Rizvi, 2018). In the student dataset, attribute reduction was primarily done to increase predictor precision, decrease computational cost, and make handling a smaller dataset easier (Asselman, Khaldi, and Aammou, 2021). By gathering a small number of variables from each student and using them as input for the prediction model, this method would allow educational organisations to forecast student academic achievement(Hussain, Akbar, Hassan, Aziz & Urooj, 2024). From the dataset, each feature selection algorithm chose a subset of student qualities. Various subsets of attributes were employed as inputs for the classifier during the experiments. Subsequently, the accuracy of the classifier or predictor was assessed for each of these subsets.

Table 1. Student academic performance prediction summary of research results

Paper Reference	Algorithm Used	Model	Sample Size	Best Accuracy
(Hamoud, Hashim, and Awadh, 2018)	J48, REPTree, RT	Classification	161	REPTree-62.3%
(Asif, Merceron, and Pathan, 2015)	NB, KNN, NN, DT, RI	Classification	347	NB-83.65%
(Al-Barrak and Al-Razgan, 2016)	J48	Classification	236	-
(Aluko, Daniel, Oshodi, Aigbavboa, and Abisuga, 2018)	LR, SVM	Classification, Regression	101	SVM-76.67%
(Adekitan and Salau, 2019)	PNN, RF, DT, NB, TE, LR	Classification, Regression	1841	LR-89.15%
(Asif, Merceron, Ali, and Haider, 2017)	NB, K-NN, RF, NN, DT, RI, X means	Classification, Clustering	210	NB-83.65%
(P´erez, Castellanos, and Correal, 2018)	DT, LR, NB, RF	Classification, Regression	762	RF-91%
(Imran, Latif, Mehmood, and Shah, 2019)	J48, NNge, MLP	Classification	1044	J48-95.78%
(Obsie and Adem, 2018)	NN, LR, SVR	Classification, Regression	1340	LR-98.05%
(Ahadi, Lister, Haapala, and Vihavainen, 2015)	DT, Bayesian, RF	Classification	296	RF-90%
(Hirokawa, 2018)	SVM	Classification	480	80%
(Okubo, Yamashita, Shimada, and Ogata, 2017)	NN, MLR	Classification, Regression	108	NN-93%
(Amazona and Hernandez, 2019)	NB, DT, DL	Classification	300	DL-95%
(Aziz, Ismail, Ahmad, and Hassan, 2015)	NB	Classification	245	NB-68.5%S
(Saa, Al-Emran, and Shaalan, 2020)	DT, RF, GBT, DL, NB, LR, GLM	Classification, Regression	56000	RF-75.52%

The experimentation was conducted using both Weka simulation software and MATLAB, with default parameters applied for all tests (Pamela Chaudhury, 2017). The research work presents a novel approach for student scholastic performance prediction, which involves utilizing specific attributes such as family expenditure, family income, student personal information, and family assets (K.R. and Blessie, 2018). Notably, the method incorporates an enhanced attribute subset selection technique to identify the most crucial features contributing to accurate student scholastic performance prediction (Sokkhey and Okazaki, 2020). Based on the comparative analysis, it is evident that the proposed attributes serve as effective predictors, demonstrating a significant impact on the scholastic performance of real-life undergraduate students.

The achieved F1-score on the dataset substantiates the predictive capability of these attributes. From the conclusive findings, it is evident that attributes such as family expenditure and personal information play a crucial role in influencing student performance, likely due to their inherent and intuitive significance in shaping academic outcomes. (Daud et al., 2017) The research culminated in the realization that students' performance is influenced not only by academic factors but also by various personal, social, and extracurricular activities (Fujiyama, Kamo, and Schafer, 2021).

To achieve this understanding, the researcher employed the Naïve Bayes algorithm in conjunction with three decision tree algorithms for data classification (Aziz, Ismail, Ahmad, and Hassan, 2015) (Matzavela and Alepis, 2021). The process began with conducting a comprehensive survey to gather relevant students' data, which was then utilized in data mining tasks. Subsequently, the data mining algorithms were applied to the dataset, resulting in the creation of classification models capable of accurately predicting students' performance based on the identified attributes. (Saa, 2016).

3. Research Gap Identification

Predicting how well students will perform academically is a complicated task influenced by many different factors. One crucial factor is the data from students' past academic records, which offers significant insights into their educational journey (Saa, Al-Emran, and Shaalan, 2020). Using this academic data to forecast outcomes is based on the idea that looking at past achievements like grades (Yagci, 2022) (Al-Barrak and Al-Razgan, 2016), chosen courses (Cagliero, Canale, Farinetti, Baralis, and Venuto, 2021), and academic patterns can help predict how well they will do in the future (Imran, Latif, Mehmood, and Shah, 2019). It is crucial to highlight that forecasting a student's academic success is a collaborative process. Academic data on students is essential, but it is only one aspect of the larger picture. A wide range of factors, both academic and non-academic, influences academic results.

These determinants include things like family income, the distance from the place of residence to the educational institution, and the education and employment of family members. The intricate relationship between each of these components gives the prediction process more nuance and complexity. While many researchers have predominantly concentrated on academic parameters within student datasets (Obsie and Adem, 2018), this research uniquely emphasizes the inclusion of both academic and socio-economic parameters.

This approach acknowledges the broader spectrum of influences on student performance. By considering social and economic factors, the study aims to provide a more comprehensive understanding of the multifaceted aspects affecting students' educational outcomes. This holistic perspective recognizes that a student's performance is shaped not only by academic capabilities but also by socio-economic circumstances, which can play a pivotal role in shaping their educational journey.

This research holds the potential to unveil previously overlooked insights and contribute to a more refined understanding of student success and its determinants. Many authors in the field have historically relied on relatively small and limited datasets with only a few parameters when researching student academic performance prediction. However, this research recognizes that to achieve more accurate and impactful predictions, a more extensive and more comprehensive dataset is essential.

A substantial dataset with a wide range of meaningful parameters allows for a more nuanced and thorough analysis. There exists an opportunity to enhance the quality of academic data by refining and optimizing it before proceeding with subsequent processing. This strategic refinement process seeks to increase the accuracy and significance of the data, rendering it more suitable for in-depth analysis and further application. Achieving this enhancement is attainable through the application of Exploratory Data Analysis (EDA) techniques to the dataset.

4. Identification of Parameters

A myriad of factors intricately influences student academic performance. Among these, academic parameters emerge as significant determinants in shaping student success. Simultaneously, social factors exert a considerable influence on educational outcomes, often operating in tandem with academic elements. Moreover, economic variables can exert indirect yet discernible impacts on student academic performance.

In cognizance of these multifaceted dynamics, this study has meticulously identified and delineated both academic and socio-economic parameters that collectively contribute to the intricate fabric of student academic performance, as stated in Table 2. Through this comprehensive approach, the intricate interplay of these diverse elements in shaping students' scholastic achievements endeavours to be unravelled, as shown in Figure 1.

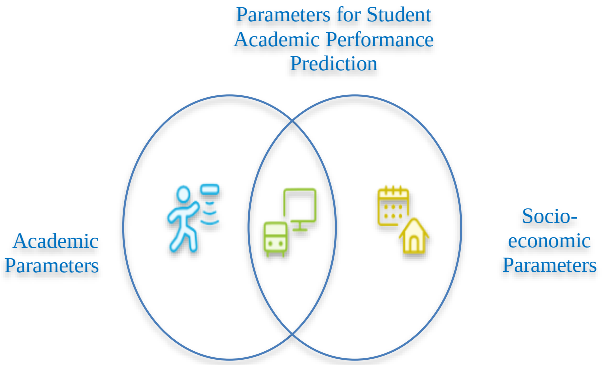

Parameters for Student Academic Performance Prediction
Academic Parameters
Socio-economic Parameters

Fig. 1 Parameters for the student data set

Table 2. Parameters for student academic performance prediction

Academic Parameters	Socio-Economic Parameters
10th Result	Parents' Qualification
12th Result	Parents' Occupation
Attendance	Stay at Hostel/Home
Unit Test Result	Family Monthly Income
Assignment Marks	
Case Study Marks	
Sessional Exam Result	
Backlog	

4.1. Justification of Parameters Identified

4.1.1.Academic Parameters

10th Result

One crucial indicator of a student's early achievements and intellectual background is their tenth grade. It shows the student's general academic ability, study methods, and foundational knowledge.

The student's performance in the tenth grade may indicate how effectively they comprehend and apply fundamental ideas to a range of subjects. Additionally, it might provide details on a student's dedication to and consistency in their academic pursuits, two factors that are significant indicators of future academic achievement.

12th Result

The results of the twelfth grade, like those of the tenth, are an essential measure of a student's academic standing and eligibility for advanced coursework. It serves as a summary of the student's secondary education and offers insights into their aptitude and accomplishments in the classroom. In comparison to the 10th grade, the 12th grade score shows the student's competency in the disciplines studied at a higher level. It also indicates the student's potential for success in more demanding academic settings and their preparedness for higher education.

Attendance

Since attendance has a direct influence on a student's capacity to actively participate in classes, interact with professors, and keep current on course material, it is an important factor in predicting student academic achievement. Students who attend class regularly can take advantage of interactive learning activities such as demonstrations, debates, and other in-class activities that are essential to the learning process. Additionally, it makes it easier for students to communicate with professors, ask for clarification on subjects, and get quick feedback on their progress. Consistent attendance also shows a student's dedication to learning and readiness to put in time and effort for their academic goals, two traits that are predictive of future academic success.

Unit Test Result

Unit examinations evaluate a student's knowledge and comprehension of specific courses or themes, making them useful measures of academic achievement. Unit examinations usually cover a certain area of the curriculum and ask students to answer multiple-choice, short answer, and problem-solving questions to show that they have mastered the topic. Teachers can assess a student's understanding, retention, and application of the information presented in class by analysing how well the student performed on these examinations. Additionally, unit examinations give teachers and students feedback by pointing out areas of strength and those that could need more time or consideration.

Assignment Marks

Assignments are useful instruments for evaluating students' writing, analysis, and research skills. Students are usually required to conduct extensive research, compile pertinent material, evaluate arguments or statistics, and present their conclusions in a logical and organized way as part of their assignments.

Students exhibit their ability to think critically, apply theoretical principles to real-world circumstances, and communicate effectively by completing projects. In addition to promoting autonomous study, assignments aid in the development of critical abilities in students, including organization, time management, and attention to detail.

Case Study Marks

Case studies are useful evaluation tools because they force students to apply their theoretical knowledge to actual circumstances, which develops their analytical and critical thinking abilities. Through case studies, students are exposed to intricate, real-world situations that call for information analysis, issue identification, option evaluation, and well-informed decision-making or recommendation-making.

Students get a greater comprehension of course themes and learn how to apply theoretical knowledge to real-world scenarios by actively participating in case studies. Because students frequently analyse and debate case scenarios in groups, case studies also help students develop their teamwork and communication skills.

Sessional Exam Result

Comprehensive tests, known as semester examinations, are used to assess a student's comprehension of all material presented in a given term, such as a semester or academic year. These tests usually include a broad range of subjects and depend on candidates to show that they are knowledgeable about, comprehend, and can apply important ideas and concepts. Sessional exams are intended to evaluate the breadth and depth of a student's learning, offering a thorough assessment of their academic success and development.

Backlog

Courses that a student has failed or not successfully finished in prior terms, known as backlogs, offer important insights into areas where a student may have previously struggled or encountered challenges. They highlight certain subjects or ideas that the learner might have found difficult or that call for more guidance and help. Teachers can better understand a student's academic strengths and weaknesses by detecting backlogs, which enables them to modify their education and assist in efficiently addressing these areas.

4.1.2. Socio-Economic Parameters

Parents' Qualification

A parent's level of education has a significant impact on how successful their child is in school and how they feel about learning. Higher-educated parents are more likely to place high importance on education and give their kids' academic success a priority. They often offer a more encouraging learning atmosphere at home, complete with access to learning tools like computers, literature, and educational activities. Furthermore, parents who have completed more education may be more aware of the value of education and take a more active role in their kids' education by offering direction, support, and academic help.

Parents' Occupation

The employment of their parents can significantly impact a student's academic performance. The student's access to emotional and financial assistance may vary depending on the nature of their job. Having a reliable source of income may allow parents to devote more funds to their kids' education, including extracurricular activities, educational materials, and tutoring. The degree to which parents are involved in their children's education can also be influenced by their line of work; some jobs that need long hours or travel may leave little time for parents to be involved in their children's academic affairs.

Stay at Hostel/Home

A student's academic performance can be significantly affected by whether or not they live at home or in a dorm. A student's living situation has a significant impact on their general well-being, study circumstances, and resource availability. Hostel residents may have access to resources that might improve their educational experience, such as study spaces, libraries, and academic support services. However, students who live at home could gain from a more accustomed and cosy setting, which might enhance their motivation and emotional health.

Family Monthly Income

One important aspect that might have a significant impact on a student's academic achievement is their family's monthly income. Higher family incomes often give students better access to extracurricular activities, educational resources, and support systems—all of which can improve their academic performance. Higher-income households, for instance, might be able to purchase technology that might enhance classroom instruction, educational resources, and private tutors. Additionally, if they have the money, they might be able to sign their kids up for extracurricular activities like music or athletics, which can help with both their general growth and academic achievement.

5. Data Set

The required information for the research can be obtained from multiple sources. Academic data can be readily accessed from the widely used Student Information System (SIS) that is prevalent in educational institutions today. The SIS can offer valuable insights into students' academic performance, and it may also provide data related to students' demographics and certain socio-economic factors. However, it is important to note that specific information regarding students' socio-economic status might not be explicitly available through the SIS. In such cases, there might be an option to deduce socio-economic indicators from the existing data available in the SIS or acquire this information directly from students through surveys. A comprehensive review of various research papers reveals a prevailing trend wherein authors predominantly focus on utilizing student academic parameters within their datasets. However, an evident gap emerges as socio-economic parameters have largely remained unexplored.

To attain a heightened degree of precision in predicting student academic performance, a holistic approach necessitates the inclusion of both academic and socio-economic dimensions within the dataset. This research endeavour diligently addresses this void by meticulously incorporating both academic and socio-economic attributes during the formulation of the student dataset. Through this integrated approach, the predictive accuracy of student academic performance is enriched, while a more comprehensive understanding of the multifaceted factors influencing scholastic outcomes is attained. By combining data from the Student Information System and conducting surveys to gather additional socio-economic data, it is possible to obtain a comprehensive dataset to analyze and explore the relationship between academic performance and socio-economic factors in their research on student scholastic prediction. As part of the experimental process, data pertaining to undergraduate students was gathered from one of the institutions of the University of Gujarat. The data collection was carried out using various modes, including questionnaire surveys and access to the student database, which is an integral part of the university's student information system. The combination of these data collection methods allowed for a comprehensive and diverse dataset to be compiled for analysis and experimentation in the research study. In this study, a dataset comprising 16 distinct academic and socio-economic parameters was utilized to analyse student performance. The dataset includes 12 numerical parameters, represented across 84 numerical fields, and 4 categorical parameters, distributed across 11 categorical fields. The dataset consists of a total of 20,995 records, providing a comprehensive data pool for analysis. It contains 221 rows and 95 columns, enabling a robust exploration of the relationships between academic achievements and socio-economic factors. The detailed structure of the dataset ensures a thorough analysis, contributing to the accuracy and reliability of the study's predictive models.

5.1. Data Preparation

The process of cleaning, translating, and organising raw data into a format appropriate for analysis, modelling, and other data-driven activities is known as data preparation. Inconsistencies, missing numbers, mistakes, and other difficulties that might impede accurate analysis are common in raw data as it is acquired from multiple sources. The goal of data preparation is to overcome these difficulties and prepare the data for subsequent processing. The prepared dataset has been submitted for exploratory data analysis to improve its meaning and precision.

Why is Exploratory Data Analysis required?

EDA involves visually and statistically analyzing the data to gain insights, identify patterns, and understand the underlying structure of the dataset, as shown in Figure 2.

- Data Understanding: EDA helps you familiarize yourself with the dataset. You can understand the variables, their types, distributions, and relationships. This understanding is essential for making informed decisions throughout the analysis process.
- Feature Selection: EDA can help you identify which features (variables) are most relevant for predicting student performance. By visualizing correlations or performing statistical tests, you can pinpoint which features have a stronger association with the target variable.
- Identifying Patterns and Trends: EDA allows you to discover patterns and trends within the data. For instance, you might uncover trends in academic performance based on certain demographic variables like gender, ethnicity, or socioeconomic background.

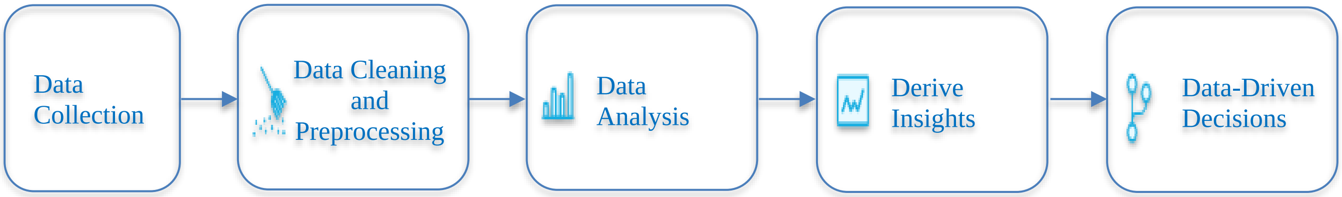


Fig. 2 EDA on the proposed dataset

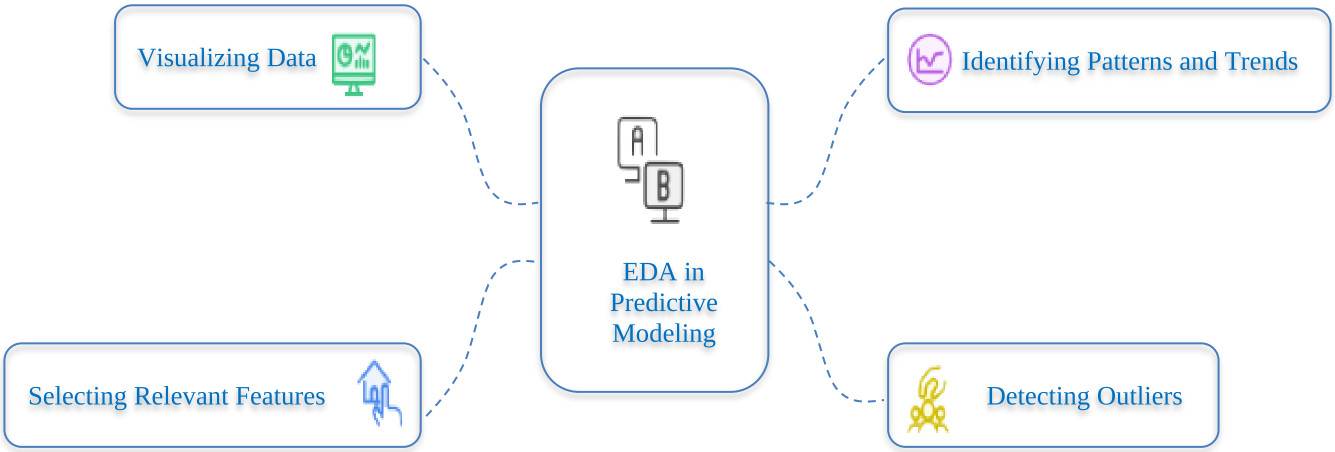

Visualizing Data
Identifying Patterns and Trends
A
B
EDA in Predictive Modeling
Selecting Relevant Features
Detecting Outliers

Fig. 3 How EDA helps in data preprocessing

- Outlier Detection: EDA helps identify outliers—data points that deviate significantly from the rest of the data. Outliers can impact model training and performance, and understanding their presence is crucial.
- Handling Missing Data: Through EDA, you can assess the extent of missing data in different variables. This understanding guides decisions on how to handle missing values, such as imputation or removal.
- Data Distribution: EDA provides insights into the distribution of variables. Understanding whether the data is normally distributed, skewed, or has other patterns can inform your choice of modeling techniques.
- Validation of Assumptions: EDA helps validate assumptions you might make about the data. For instance, you can check if variables meet the assumptions of certain statistical methods you plan to use.
- Visualization of Results: EDA produces visualizations that can effectively communicate your findings to others. Visualizations like histograms, scatter plots, and box plots can convey complex relationships more understandably.
- Hypothesis Generation: Exploring the data might lead to the formulation of hypotheses about factors influencing student performance. These hypotheses can guide further analysis and model building.
- Model Selection and Building: EDA helps you understand whether the relationship between features and the target variable is linear, nonlinear, or more complex. This insight influences the choice of appropriate modeling techniques.
- Avoiding Bias: EDA can help identify potential bias in the data, which is crucial when predicting academic performance. For example, you can investigate whether certain groups are underrepresented or if bias is present in variables like teacher ratings.

5.2. Application of EDA to Data Set

In this proposed research, the methodologies adopted for Exploratory Data Analysis are mentioned as an algorithm available in Table 3.

In dataset preparation, several key steps were executed: the systematic naming of variables and fields, the integration of columns, the treatment of missing values through either imputation or the elimination of tuples, and the application of label encoding to handle categorical data, as shown in Figure 5.

5.2.1. Naming of Variables and Fields

In the context of preparing data for predicting student academic performance, a fundamental consideration revolves around the naming of variables and fields.

The selection of appropriate names and labels for these components holds substantial importance, as it directly impacts the dataset's clarity, consistency, and ease of understanding.

The systematic naming conventions utilized in this research facilitate a comprehensive comprehension of the dataset's content.

Table 3. Proposed methodology for exploratory data analysis on the data set

Step 1: Collect data from sources such as ERP and Google Forms and prepare a consolidated Excel sheet
Repository of Data in Excel Sheet ES = {Source1, Source 2….}
Step 2: if names of Variables and fields are missing
{
Assign meaningful names manually
}
else **{Perform Step-3**}
Step 3: Join Columns if electives courses are available
Join Columns JC= {Elective1, Elective2, ….}
Step 4: Treating Missing Values either Imputation or Elimination
Impute numerical values: mean or median (x1, x2, x3….)
Where x1, x2, x3… are relevant data
Impute Categorical values: mode (y1, y2, y3….)
Where y1, y2, y3… are relevant data
Eliminate Tuples {t1, t2, t3….)
Where t1, t2, t3 are different tuples where students have left the programme
Step 5: Label Encoding for Categorical Data
LEC = {C1= { Un1}, C2={Un2},….}
Where C1, C2,…Categorical Data
Un1,Un2,…. Unique Numerical Values

5.2.2. Joining of Several Columns (Elective Subject records)

In the pursuit of a comprehensive understanding of the academic parameters within this dataset, it became evident that the elective subject records, dispersed across multiple columns, necessitated a concerted effort to consolidate and streamline the information. Elective subject records, which encompass a diverse array of courses chosen by students, were distributed across distinct columns, each corresponding to a specific elective subject.

5.2.3. Treating of Missing Values

Missing data can significantly affect the integrity of analysis and modelling efforts. To address this issue, a combination of methods has been employed for treating missing values:

5.2.4. Imputation (Missing records are inserted)

For variables with missing values, imputation techniques have been applied to estimate and replace these missing entries, as shown in Figure 4.

The choice of imputation method depended on the nature of the data. For numeric variables, mean or median imputation has been employed, while for categorical variables, mode imputation has been used. Imputed values were clearly marked for transparency, as shown in Figure 5. For mean imputation, the proposed work has adopted arithmetic mean as described in Equation 1

$$\bar{d\iota} = (\sum \mathrm{di})/\mathrm{N} \quad (1)$$

Where di is ith value

- N is the total number of values
- ∑ is the summation of values

Median is also used as an imputation technique. If the total number of values is odd, then Equation 2 is used

$$Median = \left(\frac{(N+1)}{2}\right) th\ Value \quad (2)$$

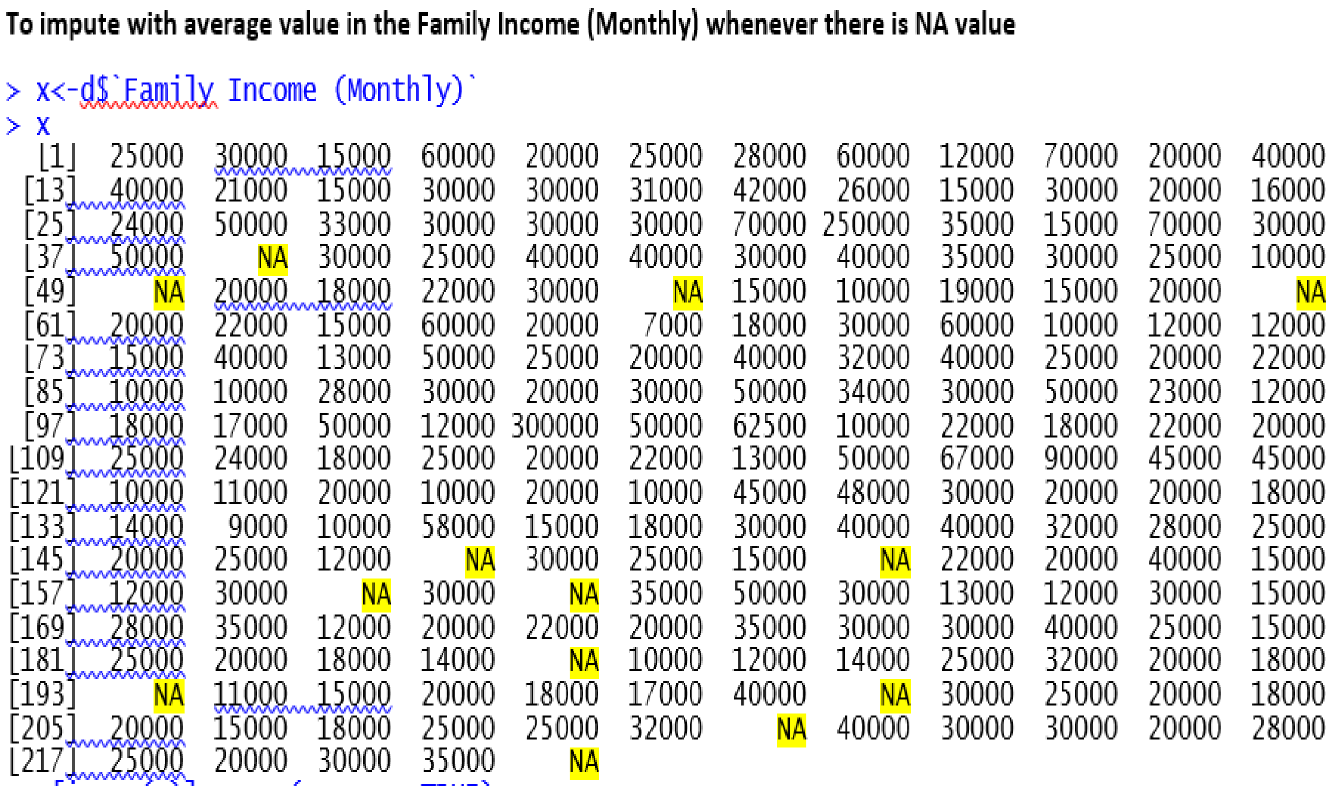
If the total number of values is an even number, then equation 3 is used.

$$Median = \frac{\left(\left(\frac{N}{2}\right) th\ Value + \left(\frac{N}{2}\right) th + 1st\ Value\right)}{2} \quad (3)$$

Where N is the total number of observations For the categorical data, the mode is the most frequently occurring value. Elimination of Tuples (Students who pursued for few days in a semester and left college before completion of a Semester):Certain tuples within the dataset represented students who initiated their studies but subsequently withdrew before the completion of a semester. To maintain the integrity of the academic performance prediction model and focus on students with meaningful academic records, these tuples have been eliminated. The criteria for tuple elimination included a minimum period of enrollment in a semester, typically set to a meaningful threshold such as four weeks. The decision to remove these tuples is grounded in the desire to focus on students with sufficiently substantial academic records for meaningful analysis and prediction. It ensures that the model is trained and evaluated on data that captures meaningful academic engagement. The bar chart illustrates the proportions of different pre-processing methods applied during Exploratory Data Analysis (EDA) for data preparation. The analysis reveals that data integration is the most significant pre-processing step, accounting for approximately 10% of the overall effort, highlighting the importance of combining data from multiple sources to enhance the dataset. Imputation with relevant data follows, representing around 9%, which suggests a focus on handling missing values to maintain data completeness. Manual rectification of data, at about 6%, indicates efforts to correct errors manually, ensuring data accuracy. The elimination of data accounts for only 3%, reflecting a minimal focus on discarding data during pre-processing. This distribution underscores the prioritization of integrating and imputing data over other methods, providing a cleaner, more comprehensive dataset for subsequent analysis in EDA.

5.2.5. Converting Categorical Data into Numerical Data

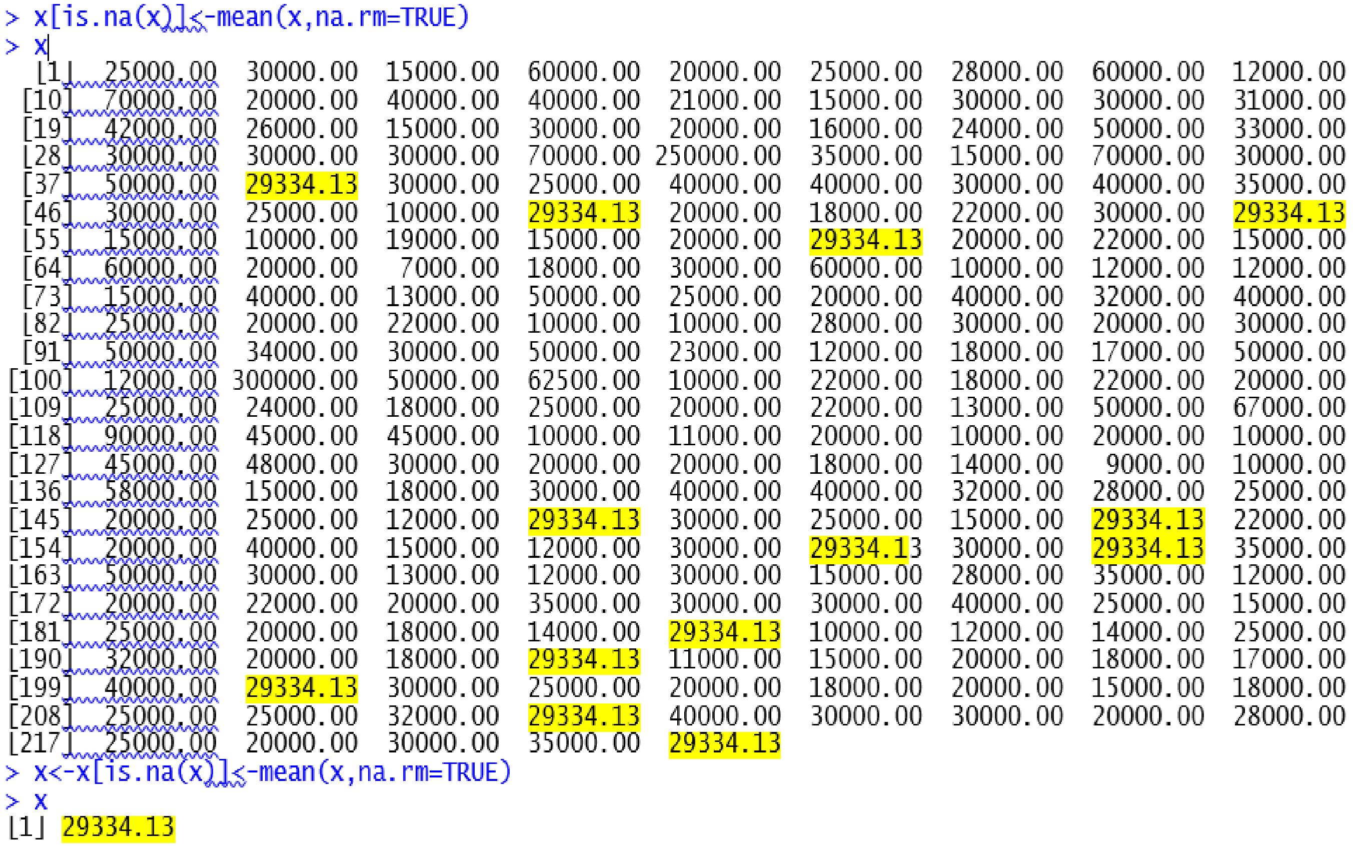
Categorical data, an inherent component of a dataset, required transformation into numerical format to render it amenable to mathematical and statistical analysis. One of the methods employed for this purpose was label encoding, a systematic approach that assigns unique integer values to each category within a categorical variable.



To impute with average value in the Family Income (Monthly) whenever there is NA value

```
> x<-d$`Family Income (Monthly)`
> x
  [1]  25000  30000  15000  60000  20000  25000  28000  60000  12000  70000  20000  40000
 [13]  40000  21000  15000  30000  30000  31000  42000  26000  15000  30000  20000  16000
 [25]  24000  50000  33000  30000  30000  30000  70000 250000  35000  15000  70000  30000
 [37]  50000     NA  30000  25000  40000  40000  30000  40000  35000  30000  25000  10000
 [49]     NA  20000  18000  22000  30000     NA  15000  10000  19000  15000  20000     NA
 [61]  20000  22000  15000  60000  20000   7000  18000  30000  60000  10000  12000  12000
 [73]  15000  40000  13000  50000  25000  20000  40000  32000  40000  25000  20000  22000
 [85]  10000  10000  28000  30000  20000  30000  50000  34000  30000  50000  23000  12000
 [97]  18000  17000  50000  12000 300000  50000  62500  10000  22000  18000  22000  20000
[109]  25000  24000  18000  25000  20000  22000  13000  50000  67000  90000  45000  45000
[121]  10000  11000  20000  10000  20000  10000  45000  48000  30000  20000  20000  18000
[133]  14000   9000  10000  58000  15000  18000  30000  40000  40000  32000  28000  25000
[145]  20000  25000  12000     NA  30000  25000  15000     NA  22000  20000  40000  15000
[157]  12000  30000     NA  30000     NA  35000  50000  30000  13000  12000  30000  15000
[169]  28000  35000  12000  20000  22000  20000  35000  30000  30000  40000  25000  15000
[181]  25000  20000  18000  14000     NA  10000  12000  14000  25000  32000  20000  18000
[193]     NA  11000  15000  20000  18000  17000  40000     NA  30000  25000  20000  18000
[205]  20000  15000  18000  25000  25000  32000     NA  40000  30000  30000  20000  28000
[217]  25000  20000  30000  35000     NA
```

Fig. 4 Finding missing values



```
> x[is.na(x)]<-mean(x,na.rm=TRUE)
> x
  [1]  25000.00  30000.00  15000.00  60000.00  20000.00  25000.00  28000.00  60000.00  12000.00
 [10]  70000.00  20000.00  40000.00  40000.00  21000.00  15000.00  30000.00  30000.00  31000.00
 [19]  42000.00  26000.00  15000.00  30000.00  20000.00  16000.00  24000.00  50000.00  33000.00
 [28]  30000.00  30000.00  30000.00  70000.00 250000.00  35000.00  15000.00  70000.00  30000.00
 [37]  50000.00  29334.13  30000.00  25000.00  40000.00  40000.00  30000.00  40000.00  35000.00
 [46]  30000.00  25000.00  10000.00  29334.13  20000.00  18000.00  22000.00  30000.00  29334.13
 [55]  15000.00  10000.00  19000.00  15000.00  20000.00  29334.13  20000.00  22000.00  15000.00
 [64]  60000.00  20000.00   7000.00  18000.00  30000.00  60000.00  10000.00  12000.00  12000.00
 [73]  15000.00  40000.00  13000.00  50000.00  25000.00  20000.00  40000.00  32000.00  40000.00
 [82]  25000.00  20000.00  22000.00  10000.00  10000.00  28000.00  30000.00  20000.00  30000.00
 [91]  50000.00  34000.00  30000.00  50000.00  23000.00  12000.00  18000.00  17000.00  50000.00
[100]  12000.00 300000.00  50000.00  62500.00  10000.00  22000.00  18000.00  22000.00  20000.00
[109]  25000.00  24000.00  18000.00  25000.00  20000.00  22000.00  13000.00  50000.00  67000.00
[118]  90000.00  45000.00  45000.00  10000.00  11000.00  20000.00  10000.00  20000.00  10000.00
[127]  45000.00  48000.00  30000.00  20000.00  20000.00  18000.00  14000.00   9000.00  10000.00
[136]  58000.00  15000.00  18000.00  30000.00  40000.00  40000.00  32000.00  28000.00  25000.00
[145]  20000.00  25000.00  12000.00  29334.13  30000.00  25000.00  15000.00  29334.13  22000.00
[154]  20000.00  40000.00  15000.00  12000.00  30000.00  29334.13  30000.00  29334.13  35000.00
[163]  50000.00  30000.00  13000.00  12000.00  30000.00  15000.00  28000.00  35000.00  12000.00
[172]  20000.00  22000.00  20000.00  35000.00  30000.00  30000.00  40000.00  25000.00  15000.00
[181]  25000.00  20000.00  18000.00  14000.00  29334.13  10000.00  12000.00  14000.00  25000.00
[190]  32000.00  20000.00  18000.00  29334.13  11000.00  15000.00  20000.00  18000.00  17000.00
[199]  40000.00  29334.13  30000.00  25000.00  20000.00  18000.00  20000.00  15000.00  18000.00
[208]  25000.00  25000.00  32000.00  29334.13  40000.00  30000.00  30000.00  20000.00  28000.00
[217]  25000.00  20000.00  30000.00  35000.00  29334.13
> x<-x[is.na(x)]<-mean(x,na.rm=TRUE)
> x
[1] 29334.13
```

Fig. 5 Imputation for missing values

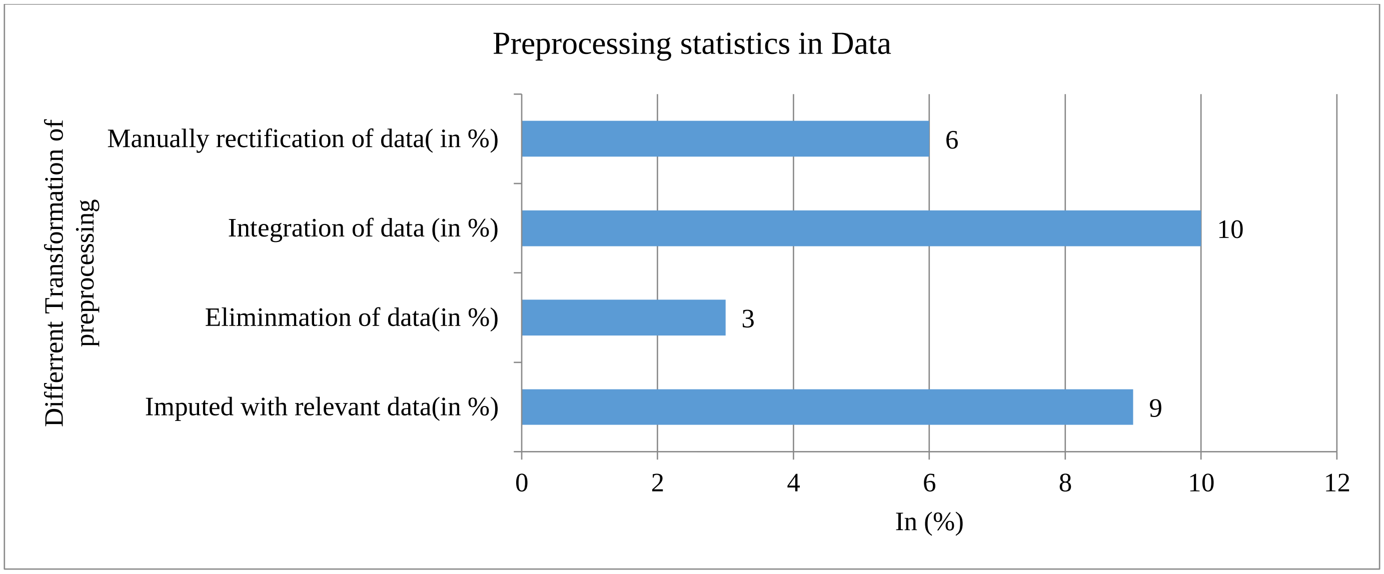


Fig. 6 Data Preprocessing on the proposed dataset

5.2.6. Label Encoding Methodology

In the label encoding process, each category within a categorical variable is assigned a distinct integer value. The assignment of these integers is performed logically and consistently, often based on the inherent ordinality or hierarchy of the categories. For example, consider the categorical variable "ClassSem1," which records student performance categories as "Distinction," "First," "Second," "Pass," and "Fail." Through label encoding, this variable is transformed as follows:

"Fail" is assigned the integer value 0.
"Pass" is assigned the integer value 1.
"Second" is assigned the integer value 2.
"First" is assigned the integer value 3.
"Distinction" is assigned the integer value 4.

This encoding scheme captures the inherent order among the categories, with higher integer values representing superior performance categories.

6. Future Research Direction

By using machine learning methods, including logistic regression, linear regression, decision trees, random forests, SVM, and neural networks, future studies will improve the prediction of student performance. Model accuracy will be maximized by fine-tuning using methods like grid search and cross-validation. The efficacy of the suggested methods will be evaluated by comparing them with cutting-edge models utilizing measures such as accuracy and F1 score. By growing the dataset, the models' generalizability across various educational systems will be examined to guarantee a more comprehensive application. These initiatives will advance academic understanding and provide useful strategies for improving student performance.

7. Conclusion

This proposed research has focused mainly on exploratory data analysis. The research has given novel algorithms for exploratory data analysis and concluded that exploratory data analysis is vital for the effective prediction of students' academic performance.

It is concluded that the previous research related to the prediction of academic performance lacked a scientific approach and did not emphasize understanding of the data. The previous research also used only academic-related features, which were not sufficient for the effective prediction of academic performance. Many authors in the field have historically relied on relatively small and limited datasets with only a few parameters when researching student academic performance prediction. It is concluded that the small data set and limited feature sets are not enough for effective prediction. It is also concluded that an opportunity exists to enhance the quality of academic data by refining and optimizing it before proceeding with subsequent processing. This strategic refinement process seeks to increase the accuracy and significance of the data, rendering it more suitable for in-depth analysis and further application. Achieving this enhancement is attainable through the application of Exploratory Data Analysis (EDA) techniques to the dataset.

References

[1] Mesfin Tadese, Alex Yeshaneh, and Getaneh Baye Mulu, “Determinants of Good Academic Performance Among University Students in Ethiopia: A Cross Sectional Study,” *BMC Medical Education*, vol. 22, no. 1, pp. 1-9, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[2] Travis T. York, Charles Gibson, and Susan Rankin, “Defining and Measuring Academic Success,” *Practical Assessment, Research & Evaluation*, pp. 1-20, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[3] Kris Kimbark, Michelle L. Peters, and Tim Richardson, “Effectiveness of the Student Success Course on Persistence, Retention, Academic Achievement, and Student Engagement,” *Community College Journal of Research and Practice*, vol. 41, no. 2, pp. 124-138, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[4] Kevin M. Bonney, “Case Study Teaching Method Improves Student Performance and Perceptions of Learning Gains,” *Journal of Microbiology & Biology Education*, vol. 16, no. 1, pp. 21-28, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[5] Pengyue Guo et al., “A Review of Project-Based Learning in Higher Education: Student Outcomes and Measures,” *International Journal of Educational Research*, vol. 102, pp. 1-13, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[6] Poonam Sawan et al., “Classification Approach for Evaluating Students Performance in Covid 19 Pandemic,” *International Journal of Engineering and Advanced Technology*, vol. 10, no. 4, pp. 110-113, 2021. [Google Scholar] [Publisher Link]

[7] Susmita Pati et al., “Early Identification of Young Children at Risk for Poor Academic Achievement: Preliminary Development of a Parent-Report Prediction Tool,” *BMC Health Services Research*, vol. 11, no. 1, pp. 1-13, 2011. [CrossRef] [Google Scholar] [Publisher Link]

[8] David Bañeres et al., “An Early Warning System To Identify and Intervene Online Dropout Learners,” *International Journal of Educational Technology in Higher Education*, vol. 20, no. 1, pp. 1-25, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[9] Alebiosu, Eunice Oluwayemisi, Akintoke, Victor Akin, and Oginni, Omoniyi Israel, “Implications of Counselling, Psychological and Social Services on Academic Performance of Primary School Pupils in Southwest, Nigeria,” *Contemporary Research in Education and English Language Teaching*, vol. 3, no. 2, pp. 1-8, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[10] Januard D. Dagdag, Hydee G. Cuizon, and Aisie O. Bete, “College Students’ Problems and their Link to Academic Performance: Basis for Needs-Driven Student Programs,” *Journal of Research, Policy & Practice of Teachers &Teacher Education*, vol. 9, no. 2, pp. 54-65, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[11] Anupam Khan, and Soumya K. Ghosh, “Student Performance Analysis and Prediction in Classroom Learning: A Review of Educational Data Mining Studies,” *Educational and Information Technologies*, vol. 26, no. 1, pp. 205-240, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[12] Filgona Jacob, Sakiyo John, and D. M. Gwany, “Teachers' Pedagogical Content Knowledge and Students' Academic Achievement: A Theoretical Overview,” *Journal of Global Research in Education and Social Science*, vol. 14, no. 2, pp. 14-44, 2020. [Google Scholar] [Publisher Link]

[13] Zhonglu Li, and Zeqi Qiu, “How Does Family Background Affect Children’s Educational Achievement? Evidence From Contemporary China,” *Journal of Chinese Sociology*, vol. 5, no. 1, pp. 1-21, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[14] Carlos Felipe Rodríguez-Hernández, Eduardo Cascallar, and Eva Kyndt, “Socio-economic Status and Academic Performance in Higher Education: A Systematic Review,” *Educational Research Review*, vol. 29, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[15] O.K Osonwa et al., “Economic Status of Parents, a Determinant on Academic Performance of Senior Secondary Schools Students in Ibadan, Nigeria,” *Journal of Educational and Social Research*, vol. 3, no. 1, pp. 115-122, 2013. [Google Scholar] [Publisher Link]

[16] Wanli Xing, “Exploring The Influences of Mooc Design Features on Student Performance and Persistence,” *Distance Education*, vol. 40, no. 1, pp. 1-16, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[17] James D.A. Parker et al., “Emotional Intelligence and Student Retention: Predicting the Successful Transition From High School to University,” *Personality and Individual Differences*, vol. 47, no. 7, pp. 1329-1336, 2006. [CrossRef] [Google Scholar] [Publisher Link]

[18] Maria P.G. Martins et al., “A Data Mining Approach for Predicting Academic Success: A Case Study,” *Information Technology and Systems*, pp. 45-56, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[19] Emtinan Alqurashi, “Predicting Student Satisfaction and Perceived Learning Within Online Learning Environments,” *Distance Education*, vol. 40, no. 1, pp. 133-148, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[20] Jonas Willems et al., “Identifying Science Students at Risk in the First Year of Higher Education: The Incremental Value of Non-Cognitive Variables in Predicting Early Academic Achievement,” *European Journal of Psychology of Education*, vol. 34, pp. 847-872, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[21] Boris Pérez, Camilo Castellanos, and Darío Correal, “Predicting Student Drop-Out Rates Using Data Mining Techniques: A Case Study,” *Applications of Computational Intelligence*, vol. 833, pp. 111-125, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[22] S.M.F.D. Syed Mustapha, “Predictive Analysis of Students’ Learning Performance Using Data Mining Techniques: A Comparative Study of Feature Selection Methods,” *Applied System Innovation*, vol. 6, no. 5, pp. 1-24, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[23] M. Narayana Swamy, and M. Hanumanthappa, “Predicting Academic Success from Student Enrolment Data using Decision Tree Technique,” *International Journal of Applied Information Systems (IJAIS)*, vol. 4, no. 3, pp. 1-6, 2012. [CrossRef] [Google Scholar] [Publisher Link]

[24] Riyadh Mehdi, and Mirna Nachouki, “A Neuro-Fuzzy Model for Predicting and Analyzing Student Graduation Performance in Computing Programs,” *Education and Information Technologies*, vol. 28, pp. 2455-2484, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[25] Raza Hasan et al., "Student Academic Performance Prediction by using Decision Tree Algorithm," *4th International Conference on Computer and Information Sciences*, Kuala Lumpur, Malaysia, pp. 1-5, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[26] Hashmia Hamsa, Simi Indiradevi, and Jubilant J. Kizhakkethottam, "Student academic performance Prediction Model using Decision Tree and Fuzzy Genetic Algorithm," *Procedia Technology*, vol. 25, pp. 326-332, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[27] X.W. Liang et al., "LR-SMOTE-An Improved Unbalanced Data Set Oversampling Based on K-Means and SVM," *Knowledge Based Systems*, vol. 196, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[28] Pamela Chaudhary et al., "Enhancing the capabilities of Student Result Prediction System," *ICTCS '16: Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies*, pp. 1-6, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[29] Maryam Zaffar et al., "A Study of Feature Selection Algorithms for Predicting Students Academic Performance," *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 5, pp. 541-549, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[30] Amal Asselman, Mohamed Khaldi, and Souhaib Aammou, "Enhancing the Prediction of Student Performance Based on the Machine Learning XGBoost Algorizthm," *Interactive Learning Environments*, vol. 31, no. 6, pp. 3360-3379, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[31] Pamela Chaudhury, and Hrudaya Kumar Tripathy, "An Empirical Study on Attribute Selection of Student," *International Journal of Learning Technology*, vol. 12, no. 3, pp. 241-252, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[32] K.R. Vineetha, and E. Chandra Blessie, "Efficient Prediction of Student Performance Using Hybrid SVM Classifier," *International Journal of Computer Science and Engineering Technology*, vol. 9, no. 3, pp. 32-39, 2018. [Google Scholar] [Publisher Link]

[33] Phauk Sokkhey, and Takeo Okazaki, "Study on Dominant Factor for Academic Performance Prediction using Feature Selection Methods," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 11, no. 8, pp. 492-502, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[34] Ali Daud et al., "Predicting Student Performance using Advanced Learning Analytics," *WWW '17 Companion: Proceedings of the 26th International Conference on World Wide Web Companion*, Perth, Australia, pp. 415-421, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[35] Hideki Fujiyam, Yoshinori Kamo, and Mark Schafer, "Peer Effects of Friend and Extracurricular Activity Networks on Students' Academic Performance," *Social Science Research*, vol. 97, pp. 1-37, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[36] Azwa Abdul Aziz et al., "A Framework for Students' Academic Performance Analysis using Naïve Bayes Classifier," *Journal of Technology*, vol. 75, no. 3, pp. 13-19, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[37] Vasiliki Matzavela, and Efthimios Alepis, "Decision tree learning through a Predictive Model for Student Academic Performance in Intelligent M-Learning Environments," *Computer and Education: Artificial Intelligence*, vol. 2, pp. 1-12, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[38] Amjad Abu Saa, "Educational Data Mining and Student's Performance Prediction," *International Journal of Advanced Computer Science and Applications*, vol. 7, no. 5, pp. 212-220, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[39] Alaa Hamoud, Ali Salah Hashim, and Wid Akeel Awadh, "Predicting Student Performance in Higher Education Institutions Using Decision Tree Analysis," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 5, no. 2, pp. 26-31, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[40] Raheela Asif, Agathe Merceron, and Mahmood K. Pathan, "Predicting Student Academic Performance at Degree Level: A Case Study," *International Journal of Intelligent Systems and Applications*, vol. 7, no. 1, pp. 49-61, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[41] M. Al-Barrak, and M. Al-Razgan, "Predicting Students Final GPA Using Decision Trees: A Case Study," *International Journal of Information and Education Technology*, vol. 6, no. 7, pp. 528-533, 2016.[Google Scholar] [Publisher Link]

[42] Ralph Olusola Aluko, "Towards Reliable Prediction of Academic Performance of Architecture Students Using Data Mining Techniques," *Journal of Engineering, Design and Technology*, vol. 16, no. 3, pp. 385-397, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[43] Aderibigbe Israel Adekitan, and Odunayo Salau, "The Impact of Engineering Students' Performance in the First Three Years on their Graduation Result Using Educational Data Mining," *Heliyon*, vol. 5, no. 2, pp. 1-21, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[44] Raheela Asif et al., "Analyzing Undergraduate Students' Performance Using Educational Data Mining," *Computers & Education*, vol. 113, pp. 177-194, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[45] Muhammad Imran et al., "Student Academic Performance Prediction Using Supervised Learning Techniques," *International Journal of Emerging Technologies in Learning*, vol. 14, no. 14, pp. 92-104, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[46] Efrem Yohannes Obsie, and Seid Ahmed Adem, "Prediction of Student Academic Performance Using Neural Network, Linear Regression and Support Vector Regression: A Case Study," *International Journal of Computer Applications*, vol. 180, no. 40, pp. 39-47, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[47] Alireza Ahadi et al., "Exploring Machine Learning Methods to Automatically Identify Students in Need of Assistance," *Proceedings of the Eleventh Annual International Conference on International Computing Education Research*, USA, pp. 121- 130, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[48] Sachio Hirokawa, "Key Attribute for Predicting Student Academic Performance," *Proceedings of the 10th International Conference on Education Technology and Computers*, pp. 308-313, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[49] F. Okubo et al., "A Neural Network Approach for Students' Performance Prediction," *Proceedings of the 7th International Learning Analytics & Knowledge Conference*, Canada, pp. 598-599, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[50] Mayreen V. Amazona, and Alexander A. Hernandez, "Modelling Student Performance Using Data Mining Techniques: Inputs for Academic Program Development," *Proceedings of the 2019 5th International Conference on Computing and Data Engineering*, pp. 36-40, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[51] Amjad Abu Saa, Mostafa Al-Emran, and Khaled Shaalan, "Mining Student Information System Records to Predict Students' Academic Performance," *Proceedings the International Conference on Advanced Machine Learning Technologies and Applications*, pp. 229-239, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[52] Mustafa Yağcı, "Educational Data Mining: Prediction of Students' Academic Performance Using Machine Learning Algorithms," *Smart Learning Environments*, vol. 9, no. 11, pp. 1-19, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[53] Luca Cagliero et al., "Predicting Student Academic Performance by Means of Associative Classification," *Applied Sciences*, vol. 11, no. 4, PP. 1-21, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[54] Eyman Alyahya, and Dilek Düştegör, "Predicting Academic Success in Higher Education: Literature Review and Best Practices," *International Journal of Educational Technology in Higher Education*, vol. 17, pp. 1-21, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[55] P. Umamaheswari et al., "Student Success Prediction using a Novel Machine Learning Approach based on Modified SVM," *Multidisciplinary Science Journal*, vol. 5, no. 15, pp. 1-7, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[56] Muhammad Mubashar Hussain et al., "Prediction of Student's Academic Performance through Data Mining Approach," *Journal of Informatics and Web Engineering*, vol. 3, no. 1, pp. 241-251, 2024. [CrossRef] [Google Scholar] [Publisher Link]