# FEATURE ENGINEERING FOR STUDENTS ACADEMIC PERFORMANCE AND REIGNITION USING MACHINE LEARNING AND DEEP LEARNING

**BY**

**MD. SAYMON AHAMMAD**
**ID: 202-15-14388**

**AND**

**SADIA AKTER SINTHIA**
**ID: 202-15-14369**

## FINAL YEAR DESIGN PROJECT REPORT

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering

## Supervised By

**Prof. Dr. Syed Akhter Hossain**
Dean and Professor
Department of Computer Science and Engineering
Daffodil International University

## Co-Supervised By

**Professor Md. Monzur Morshed**
Professor
Department of Computer Science and Engineering
Daffodil International University

## DAFFODIL INTERNATIONAL UNIVERSITY

**DHAKA, BANGLADESH**

**July 2024**

# APPROVAL

This Project titled **"Feature Engineering for Students Academic Performance and Reignition using Machine Learning and Deep Learning"**, submitted by **Md. Saymon Ahammad** and **Sadia Akter Sinthia** to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on **15 July, 2024.**

## BOARD OF EXAMINERS

Hossain 15.07.2024

**Mr. Md. Fokhray Hossain (MFH)** — **Board Chairman**
**Professor**
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Firoz Hasan

**Md. Firoz Hasan (FH)** — **Internal Examiner 1**
**Sr. Lecturer**
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Lamia Rukhsara 15.07.24

**Lamia Rukhsara (LR)** — **Internal Examiner 2**
**Sr. Lecturer**
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

15.07.24

**Dr. S. M. Hasan Mahmud (SMH)** — **External Examiner**
**Assistant Professor**
Department of Computer Science
American International University-Bangladesh

## DECLARATION

We hereby declare that this project has been done by us under the supervision of **Prof. Dr. Syed Akhter Hossain, Dean and Professor, Department of Computer Science and Engineering, Daffodil International University**. We also declare that neither this project nor any part of this project has been submitted elsewhere for the award of any degree or diploma.

**Supervised by:**

13.9.24

**Prof. Dr. Syed Akhter Hossain**
Dean and Professor
Department of CSE
Daffodil International University

**Co-Supervised by:**

**Professor Md. Monzur Morshed**
Professor
Department of CSE
Daffodil International University

**Submitted by:**

Saymon

**Md. Saymon Ahammad**
ID: 202-15-14388
Department of CSE
Daffodil International University

Sinthia

**Sadia Akter Sinthia**
ID: 202-15-14369
Department of CSE
Daffodil International University

# ACKNOWLEDGEMENT

First, we express our heartiest thanks and gratefulness to almighty for His divine blessing making it possible for us to complete the final year project/internship successfully.

We are grateful and wish our profound indebtedness to **Prof. Dr. Syed Akhter Hossain**, **Dean and Professor**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of Data Mining and Artificial Intelligence to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts, and correcting them at all stages have made it possible to complete this project.

We would like to express our heartiest gratitude to the **Head of the Department of CSE**, for his kind help in finishing our project and also to other faculty members and the staff of the Department of CSE, Daffodil International University.

We would like to thank our entire course mate in Daffodil International University, who took part in this discussion while completing the course work.

Finally, we must acknowledge with due respect the constant support and patience of our parents.

# ABSTRACT

The research titled "Feature Engineering for Students Academic Performance and Reignition Using Machine Learning and Deep Learning" highlights the transformative potential of modern computational methods in strengthening educational predictive analytics. Utilizing 3028 response collected via a through Google Form survey with 46 queries , the study digs into factors affecting student success, with feature engineering translating raw data into relevant feature for enhanced model accuracy. Among machine learning models Random Forest achieved the best accuracy of 94.61% and the lowest error rates, followed closely by Decision Tree with 93.785 accuracy. Gradient Boost , KNN, AdaBoost had accuracies of 92.29% 88.65% and 81.74% respectively, with AdaBoost performing the lowest. In deep learning models ,ANN stood out with a 94.50% accuracy , while the hybrid CNN-LSTM model followed with 93.48% accuracy. LSTM and CNN scored 93.27% and 93.195 accuracies, respectively with CNN performing the lowest among deep learning models. One essential part of this study was determining the most influential characteristics affecting student performance , with favorably and adversely. Features with correlation value greater than 0.4 or less than -0.4 were considered significant, including ; Gender, Department Choice, Study Hours Per Week, Family Responsibility, Past Failure, Average Sleep Hours, Study Year, Strategy of Measuring Performance_MCQ, Strategy of Measuring Performance of_Oral, Written, Assignment, Presentation. The goal of the research is to improve our knowledge of how feature engineering and deep learning can be used most effectively in this situation. This will help both the academic community and the real world of education, where better predictive models can directly and favorably affect student outcomes and the larger picture of educational success.

# TABLE OF CONTENTS

## LIST OF FIGURES

## LIST OF TABLES

## LIST OF ACRONYMS

| | |
|---|---|
| ML | Machine Learning |
| DL | Deep Learning |
| CNN | Convolutional Neural Network |
| KNN | K-Nearest Neighbor |
| LSTM | Long Short-Term Memory |
| MAE | Mean Absolute Error |
| MSE | Mean Squared Error |
| RMSE | Root Mean Squared Error |

# CHAPTER 1
# INTRODUCTION

## 1.1 Overview

In today's world of education, students' academic performance stands as a pivotal force, shaping not only their current scholastic endeavors but also setting the framework for a future full of possibilities. This essential component is the subject of considerable discussion and reflection among educators, parents, and society at large. Academic success, in addition to grades and exams, may lead to educational possibilities, professional progress, and personal improvement. Good academic performance leads to possibilities such as scholarships and specialized programs. It enhances a student's confidence and drive, thereby improving their future career prospects. On the other hand, poor performance brings several obstacles, including failures and limited prospects. It also lowers self esteem and brings difficulties into a student's life. A student's performance can be influenced by many factors. Their performance can be affected by personal, economic, social, financial, family, and environmental factors. Identifying which factors affect them most is crucial to fostering their academic growth.

## 1.2 Background and Present State

The students' experience is filled with numerous obstacles that can greatly affect both mental and intellectual capabilities. The burden of academic expectations is a significant concern as students frequently experience a sense of being inundated by the requirements of their academic works, examinations and need to get exceptional results. Historically there are many challenges students used to face fixed mindset and external pressures influenced way too much in education and now with the digitalization the challenges are getting broader. Historically , 7% of individuals between the ages of 3 and 17 encountered anxiety problems , although the prevalence of severe depression has shown a gradual rise over time [1] . Students encountered issues pertaining to alcohol usages and substances abouse. Alcohol was prevalent , exhibiting elevated rates of consumption among students. In current situations challenges are different and quite affective on academic performance. Factors like time management , low motivation , lack of concentration, pandemic impact are in high concern [2]. Due to the variety of changes of these factors it is getting difficult for educational institutions or educators to find out what exactly are the main reasons that are influencing students academic performance.

### 1.3 Problem Statement

The current issue in educational research is having an insufficient understanding of the precise components or factors that have significant effects on students' academic performance. While it has been recognized that different factors influenced this result, a lack of complete understanding impedes the creation of effective interventions with specific methods. This knowledge gap is a barrier to improving students' academic results, underscoring the need for targeted study to discover and analyze the precise qualities that are linked to academic success. Addressing this gap is critical for developing evidence-based educational techniques that promote improved academic performance among students.

### 1.4 Objectives

The justification for doing the research on "FEATURE ENGINEERING FOR STUDENT'S ACADEMIC PERFORMANCE AND RETENTION USING MACHINE LEARNING AND DEEP LEARNING" is strongly entrenched in the desire to solve key difficulties in the realm of education and student success. Addressing the knowledge gap concerning the precise determinants that impact students' academic performance is driven by the possibility of sustainability, improving educational results and student triumph and it is the main motivation in this research. Through an examination of these complex factors, policymakers, educators, administrators can formulate interventions and approaches that specifically target the underlying factors contributing to academic difficulties. Acquiring this knowledge can enable academic establishment to develop more efficient learning environments and the object is to

- Identify the influential factors,
- Quantify impact,
- Develop targeted interventions,
- Enhance educational strategies,
- Promote student success,
- Contribute to educational research,
- Improve educational equity

The study's ultimate goal is to lay the groundwork for developments in educational analytics that will significantly improve the accuracy and effectiveness of predictions about student performance and retention. The goal of the research is to improve our knowledge of how feature engineering and deep learning can be used most effectively in this situation. This will help both the academic community and the real world of education, where better predictive models can directly and favorably affect student outcomes and the larger picture of educational success.

### 1.5 Scope and Limitations

This study not only examines the technical aspects of feature engineering, but also broadens its scope to assess the presented models' real world application in educational environments. This research has the potential to have a significant impact by providing educators with actionable information, encouraging a proactive approach to student support, and promoting a culture of continuous improvement. The primary objective of this research is to explore the features that are strongly related to students academic performance by employing both traditional machine learning and advanced deep learning techniques. Additionally this study aims to examine the possibility of rekindling and optimizing student success through the application of machine learning insights. Implementing machine learning models allows predictive analysis for students academic achievement. By training models using historical data we can uncover patterns and links between various academic features, allowing for prediction of future performance. Machine learning algorithms can be used to identify students who are likely to become disengaged or perform poorly. Targeted interventions can be created to rekindling students interest and motivations by detecting early indicators using predictive modelling.it can help to generate personalized learning programs by analyzing individual student characteristic and customizing educational techniques to their specific need .Deep learning architecture , with their ability to automatically build hierarchical representations from data, have the potential to decrease the manual labor necessary for feature engineering. Neural networks , a subset of deep learning , may detect nonlinear connections in academic data. The scope of our research involves leveraging machine learning and deep learning to not only identify the features that affect students performance but also inform reignition strategies. It includes addressing ethical consideration and ensuring the deployment of artificial intelligence in the educational system for academic growth.

Some of the notable limitations of our study are mentioned below:

- **Data Quality and Availability:** The completeness and quality of data have a major impact on the prediction models' accuracy and dependability. Predictions that are erroneous can be caused by missing and inconsistent data.
- **Generalizability:** The model created in this work could work well on the particular training datasets, but they might not translate well to different educational settings or demographics,

- **Ethical Considerations:** Privacy issues arise when personal student data are used for predictive modeling. It is essential to guarantee data protection and acquire appropriate permission from parents and pupils.
- **Implementation Challenge:** Translating predicted findings into actionable solutions might be difficult. Educators may require additional training and assistance to properly incorporate these technologies into their teaching practices.
- **Dynamic Nature of Education:** Changes in curriculum, educational regulations, or socioeconomic situations are just a few examples of the external influences that might cause the elements impacting academic performance to shift over time. The models must be updated and validated on a regular basis in order to remain accurate and relevant.

### 1.6 Report Organization

The suggested report has a thorough format that walks readers through the methodology, conclusions, and implications of the study. Every chapter has a distinct function and adds to the document's overall cohesion and depth.

**Chapter 1: Introduction**

The introduction part of this study demonstrates the identification of the research topic, significance and the impact of this particular study on society. It provides an overview of the research questions, objectives and overall framework of the study.

**Chapter 2: Literature Review**

A thorough analysis of pertinent literature and earlier research is provided in the background chapter. An extensive understanding of the topic, machine learning and deep learning, as well as theoretical underpinnings, approaches, and technical developments is provided in this part. It lays forth the background for the present study and identifies the knowledge gaps.

**Chapter 3: Methodology**

The research methodology chapter describes the details of the overall research procedure through which the whole study has been conducted. It starts from data collection and ends with result analysis. It also describes the data preprocessing steps followed by model selection, model evaluation.

**Chapter 4: Implementation**

This chapter demonstrates the implementation of models of both machine learning and deep learning. A detailed implementation criteria and evaluation is discussed in this chapter of the study.

**Chapter 5: Result And Analysis**

This chapter demonstrates the obtained results from the applied models of both machine learning and deep learning. A detailed analysis of the obtained results are also analyzed in this chapter of the study.

**Chapter 5: Impact On Society, Environment And Sustainability**

This chapter named impact on society describes the social impact of this study, more specifically why this study is important and in which sector the study adds some value to improve social well being. It discussed how an improved student performance prediction system will positively affect society.

**Chapter 6: Summary, Conclusion, Recommendations, and Implication for Future Research**

The whole report is summarized in this last chapter. It provides recommendations for real-world applications and more research in addition to summarizing the main results and restating the research's conclusions. The ramifications for further investigation are explored, offering a guide for researchers who want to build on the work that is being done now.

## 1.7 Summary

Investigating the complex interplay between feature engineering methodologies, machine learning models and students academic performance, this study aims to reduce the distance between theoretical comprehension and practical implementation in educational settings. Through the utilization of both conventional and cutting-edge techniques, this research endeavors to ascertain crucial indicators of scholastic achievement, device focused interventions to rekindle student involvement and establish ethical standards for the responsible application of artificial Intelligence. By means of enhanced predictive models and customized approaches, this study aims to furnish educators, policymakers, and students with actionable insights and resources, thereby ultimately augmenting academic experiences and outcomes in practical educational environments.

# CHAPTER 2
# LITERATURE REVIEW

## 2.1 Overview

This study of the literature critically looks at a number of studies on student performance, pointing out common flaws and uncharted territory. The shortcomings that have been found include inadequate sample representation, insufficient factor investigation, failure to include external variables, and methodological flaws, especially with regard to algorithm selection. One common pattern in earlier research is the application of the findings to particular areas or educational institutions, which limits the generalizability of the results. These studies often overlook the time development of variables impacting student achievement, oversimplify complicated processes, and lack qualitative insights. Future research should use qualitative approaches, have a more varied participant pool, and clearly justify its methodologies to improve quality of findings. The present trend of limiting study to certain topics misses the variety of elements that affect students' performance in a variety of areas. Researchers may produce flexible concepts that can be implemented in many kinds of educational institutions and get a more thorough grasp of the subtleties influencing student performance by expanding their reach.

## 2.2 Related Works

In the study, Hajizadeh et al. [3] used secondary data collected from previous works. They collected 5820 reviews with 32 attributes provided by the students of Gazi University. They used two different algorithms: classification and association rule discovery. The second analysis achieved a result of 84.244% with an average F measure of 0.774, and the third analysis achieved an accuracy of 84.2612% with an average F measure of 0.772. In the future, they aim to analyze the result by collecting more data and creating a new dataset. The data used in the paper of Beckham et al. [4] is quantitative student data collected from Kaggle. They used three algorithms in their paper: MLP-12 Neuron Model, Random Forest, and Decision Tree. The Multi-Layer Perceptron model, with an RMSE value of 4.32, follows Random Forest, with an RMSE value of 4.52, and Decision Tree, with an RMSE value of 5.69, as the ultimate outcome of applying all the algorithms. This implies that the chosen characteristics significantly affect how well students succeed. However, prior research in the study indicated that a mother's education has a positive association (0.217147) with student grades and that past failures have a negative correlation (-0.360415) with student grades. The way these two characteristics correlate with student grades suggests that they may have important effects. Future research ought to focus on gathering more data from around the globe, as the current data variation is still average and only includes students from two Portuguese schools. Additionally, other parametric or advanced machine learning techniques, like deep learning, should be used to explore more options for finding consistency. With the development of artificial intelligence, other statistical techniques, like using mean absolute error (MAE) to calculate error, may be useful in identifying additional hidden factors within a large dataset.

In this study by Alani et al. [5], a survey was conducted among different faculties of Sohar University (SU) using a questionnaire, and primary data was collected from participants and performed a critical analysis using regression analysis. They used Kolmogorov-Smirnov Test, Survey Analysis Tools, Descriptive Statistics, Regression analysis, and ANOVA (Analysis of Variance). Research indicates that personal characteristics significantly affect academic success for Sohar University students. Students believe that institution-related elements significantly impact their performance, preferring a peaceful and suitable university atmosphere. Students see teachers with teaching abilities and diverse teaching methods as having a good impact on their performance.

They had collected student-related dataset from an online questionnaire by using Google Forms in the study of Al-Hagery et al. [6] which was anonymously and without any biasly published in Qassim University students. Initially the size of their dataset was 1143 records. They used two clustering algorithms; K-Means Clustering and X-Means Clustering. The Elbow algorithm revealed three clusters as the ideal number for K-means.Three clusters contained data instances. Cluster 0 had more "Pass" pupils (10.16%) while Cluster 1 had the most "Excellent" students (44.81%). Certain attribute values affected clusters. For instance, "Housewife" was the mother's occupation in Cluster 0, "Service" in Cluster 1, and "Retired" in Cluster 2.With three clusters, X-means was used. Cluster 2 had the most "Good" pupils (54.4%) and Cluster 0 the most "Pass" students (10.3%). Similar to K-means, some characteristic values had a considerable influence on the clusters. For example, "A" as the high school percentage, "Private car" for transport, "Housewife" for mother's occupation, "Retired" for father's occupation, "General education" for both father's and mother's education, "High" for income, "Big" for family size, "Married" for parents' marital status, "Limited" for the number of hours spent with friends, and "High" for the number of friends had influences on the clusters.

Questionnaire was administered in Rasul et al. [7] to the students of Bahauddin Zakariya University Multan, Pakistan and collected data from 200 students;100 from faculty of arts and 100 from faculty of science. They used three methods for analyzing the data; Mean, Standard Deviation and Z-Test. They show their analysis in 5 different comparison tables. In the first table, the presentation reveals that when the mean score surpasses 3.00 (norm), it signifies that a majority of university-level students view numerous psychological, physical, socio-economic, and educational elements to significantly affect their performance in tests. The 2nd table shows the mean score of both male and female students. Hoghest mean for female students was 80.00 in Math and Zoology department while the highest mean for male was 80.5 in English department. The minimum score for female was 66.5 while the score for male was 57.00. 3rd reveals the Mcomb of female students was 74.64 which was bigger than the Mcomb of male students 69.49. 4th indicates the Z value of 4.90 which is bigger than the table value 1.96 and in the last table, It represents the mean of scientific and arts faculties. In arts faculty, the highest mean was 78.24 of English department while in science faculty the highest mean was 76.25 in Math department. The lowest mean of arts faculty was 68.25 of PAK-STUDT department while the lowest mean for scientific faculty was 66.25 in Botany department.

In the study of Hijazi et al. [8] the data was collected using a questionnaire administered by the teachers of the graduate class in the 5th month of 4th year. questionnaires 300 were filled with the response 100% rate of which females were 75 and males were 225 . In that study they got value of 0.72139 in R, 0.520403 in R square by applying , 0.400504 Adjusted R Square, 0.084872 of Standard Error and F Stat 4.340334 by using used multiple regression techniques.

This study of Farooq et al. [9] was conducted to examine different factors influencing the academic performance of secondary school students in a metropolitan city of Pakistan. The respondents for this study were 10th grade students (300 male & 300 female). A survey was conducted by using a questionnaire for information gathering about different factors relating to academic performance of students. The quality of academic performance was determined by their accomplishment results of the 9th grade yearly test validated by the Board of Intermediate and Secondary Education, Lahore and school records. Data on the factors such as parents' education, parents' occupation, SES, urban/ rural belongingness, and students' gender were obtained by utilizing a questionnaire. They used T-Test, Analysis of Variance (ANOVA), Multiple Comparisons.

Primary data that has been used by Singh et al. [10] which obtained by surveys on a five points in the Likert scale ranging from (1and 5 where set into strongly disagree and strongly agree. 200 questionnaires were issued to the management students of management institutes in Haryana state, out of which 175 (87.5%) questionnaires were received. The sample of the present study consisted of 200 management students from ten management institutes of Haryana state. Features are learning facilities, communication skills and proper guidance from parents. Mean, standard deviation and regression analysis were used for the data analysis and interpretation. For further research family income, parent's education and educator these variables can be added.

Kaviyarasi et al. [11] used a total of forty five attributes to be identified and listed in the dataset. They used an Extra Tree Classifier in their study. In that paper, the authors selected the features using feature selection and then perform extra tree classifier algorithm to find the high potential factors among them. The future work of this research is to predict the performance of college Students with high accuracy using a new proposed model

In the study of Mushtaq et al. [12] the data used in the study was collected through assessment from students of a group of private colleges and 175 questionnaires were distributed. They used Regression analysis and ANOVA in the study. Reliability of the scale of total Items is 0.710. The reliability of individual items are; Student performance: 0.716, Communication: 0.497, Learning Facilities: 0.735, Proper Guidance: 0.806, Family Stress: 0.258. And the descriptive Analysis shows that mean of student performance is 3.7903 and standard deviation is .98672.Mean of communication, learning facilities, proper guidance and family stress are 4.1626, 4.2597, 4.1462 and 4.28172 respectively, which shows that respondent are agree that these variables affect student performance and standard deviation for these independent variables are 0.50390, 0.67713, 0.89659 and 0.396398 respectively. In the future, the research can be done using more data and includes more cities rather than 2.

## 2.3 Comparison between existing works

TABLE 2.3.1: COMPARATIVE ANALYSIS WITH PREVIOUS WORKS

| **Author Name** | **Dataset** | **Used Algorithms** | **Best Accuracy with Algorithm** | **Limitations** |
|---|---|---|---|---|
| Hajizadeh et al. [3] | Analyzed 5820 student reviews from Gazi University, focusing on 32 attributes | Classification and Association Rule Mining | Association Rule Mining = 84.2612% | Not well-defined approaches applied |
| Beckham et al. [4] | Analyzed student data from Kaggle | MLP 12 Neuron Model, Random Forest, Decision Tree | ______ | Limited exploration of factors |
| Alani et al. [5] | Surveyed Sohar University, , and primary data was collected from participants | Kolmogorov-Smirnov Test, Survey Analysis Tools, Descriptive Statistics, Regression | ______ | Methodological gaps were found |

| | | | | |
|---|---|---|---|---|
| | | Analysis, and ANOVA (Analysis of Variance) | | |
| Al-Hagery et al. [6] | collected student-related dataset from an online questionnaire by using Google Forms | K-Means Clustering and X-Means Clustering | ______ | Limited dataset |
| Rasul et al. [7] | Questionnaire was administered in to the students of Bahauddin Zakariya University Multan, Pakistan and collected data from 200 students; | Mean, Standard Deviation, and Z-Test | ______ | Processes are not well explained |
| Hijazi et al. [8] | 300 data was collected through questionnaires | Multiple Regression | Multiple Regression = 72.14% | Data set was too short and not well-defined approaches applied |
| Farooq et al. [9] | A survey was conducted by using a questionnaire for information gathering about different factors relating to academic performance of students. | T-Test, Analysis of Variance (ANOVA), Multiple Comparisons | ______ | Limited dataset |
| Singh et al. [10] | Primary data that has been used by which obtained by surveys | Regression Analysis | ______ | Analysis could be more better |
| Kaviyarasi et al. [11] | Used a total of forty five attributes to be identified and listed in the dataset | Extra Tree Classifier | ______ | Incomplete Sample Representation |
| Mushtaq et al. [12] | the data used in the study was collected through assessment from students of a group of private colleges and 175 questionnaires were distributed. | Regression Analysis, ANOVA | ______ | Processes are not well explained |

### 2.4 Open Issues

Across the examined research papers, several common gaps and potential areas for future study emerge.

**Limitations:**

1. **Incomplete Sample Representation**: A common problem in research is inadequate sample representation, which suggests that the variety of the

population of interest may not be sufficiently reflected in the individuals selected. This restriction may make it more difficult to extrapolate the results to larger settings.

2. **Limited Exploration of Factors:** Some studies have a restricted emphasis on particular elements, which may cause them to ignore the complex character of the subject they are studying. To fully comprehend the topic, a more detailed investigation of the many impacting factors is required.
3. **Overlooking External Variables:** Sometimes, external factors that might have a big impact on the phenomena under study are ignored. Ignoring these other influences may result in an incomplete and maybe skewed assessment of the data.
4. **Methodological Gaps (Algorithm Choice):** Another problem is methodological gaps, particularly when it comes to the selection of algorithms or research techniques. The validity and reliability of the study may be jeopardized by the use of out-of-date or incorrect methodologies.

Previous research is only conducted to certain locations or schools, which makes it impossible to extrapolate their conclusions universally. They also tend to simplify things too much, lose out on qualitative insights, and overlook how things develop over time. To enhance future study, we should look at a more varied group of individuals, employ methodologies that incorporate qualitative information, and explain why we picked specific research methods better. Right present, many studies solely concentrate on single topics, neglecting variances in what helps students excel in other industries. By looking at a larger variety of courses, academics may understand better what impacts student performance and come up with concepts that work for lots of various sorts of schools

## 2.5 Summary

Previous works  are only conducted to certain locations or schools, which makes it impossible to extrapolate their conclusions universally. They also tend to simplify things too much, lose out on qualitative insights, and overlook how things develop over time. To enhance future study, we should look at a more varied group of individuals, employ methodologies that incorporate qualitative information, and explain why we picked specific research methods better. Right present, many studies solely concentrate on single topics, neglecting variances in what helps students excel in other industries. By looking at a larger variety of courses, academics may understand better what impacts student performance and come up with concepts that work for lots of various sorts of schools.

# CHAPTER 3
# METHODOLOGY

## 3.1 Overview

The focal point of the research project titled "Feature Engineering for Student's Academic Performance and Retention Using Machine Learning and Deep Learning" is the utilization of sophisticated machine learning methods to improve student retention and academic performance through predictive analytics. The investigation centers on student accomplishment, which is a pivotal concern for academic institutions. The study investigates the difficulties linked to conventional approaches in forecasting academic achievements and student retention. In an effort to surmount these constraints, the research incorporates advanced machine learning and deep learning models with feature engineering. This part will provide an overview of the entire study approach, from data collecting to data preprocessing, data balancing, train-test split,

## 3.2 Proposed Methodology

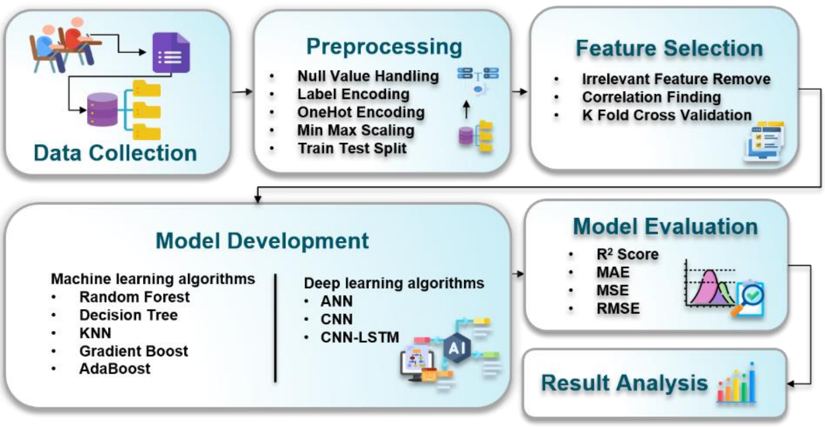



Figure 3.2.1: Methodology of the Study

### 3.2.1 Data Collection Procedure

#### 3.2.1.1 Dataset

The dataset was predominantly obtained through surveys conducted with students and stakeholders, yielding a total 3028 data points. The data has been collected from various universities which includes Daffodil International University, Independent University, BRAC University, Northern University, Prime University, State University, BGMEA University of Fashion Design & Technology and various National Universities. To ensure a thorough and efficient data collection process, a Google Form containing an exhaustive set of 46 queries was designed. By utilizing a cross-questioning approach, every inquiry is deliberately designed to authenticate and bolster the dependability of answers, thereby guaranteeing a comprehensive and nuanced comprehension of the elements influencing student performance. By utilizing the functionalities of ChatGPT, the inquiries on the form were enhanced and streamlined. ChatGPT significantly improved the survey questions' lucidity, pertinence, and efficacy, thereby augmenting the overall accuracy of data acquisition. Furthermore, ChatGPT was employed to conduct initial analysis, which assisted in the identification of possible correlations and patterns in the gathered responses.

#### 3.2.1.2 Interview with Stakeholders

The analysis required a broader knowledge on the field to find out the possible feature that affects students' performance. To do so, before designing the google form to collect data, we went to some teachers, experts as well as students to know what features are affecting the students most in their opinion.

#### 3.2.1.3 Key Features of the Google Form

- **Comprehensive Inquiry:** The form incorporates a diverse range of inquiries, encompassing demographics, study routines, extracurricular engagements, and external influences on academic performance. This breadth attempts to convey a holistic view of the student experience.
- **Cross-Questioning Methodology:** Applying a cross-questioning approach, internal uniformity checks are incorporated to validate responses. This method enhances the robustness of collected data, minimizing inaccuracies.
- **Iterative Design Process:** An iterative design process, supported by observations from ChatGPT, assures continuous refinement and alignment with the goals of the research. This dynamic approach fine-tunes the form for optimal effectiveness.

#### 3.2.1.4 Questionnaires Description

In the pursuit of understanding various aspects of a topic , we crafted a series of questions designed to elicit insightful responses and gather valuable data. Our survey inquiries were meticulously crafted of the subject matters. Here we present an overview of the questions we asked and the rationale behind each query.

TABLE 3.2.1.4.1: QUESTIONNAIRES DESCRIPTION

| Question | Answer Label | Rational |
|---|---|---|
| What is your age ? | — | This question will provide the demographic insight to students development stages |
| How do you identify your gender? | Male<br>Female<br>Other: | This question aims to recognize diversity and provide inclusive support services |
| What is your current grade or level of study year? | 1st Year<br>2nd Year<br>3rd Year<br>4th Year | This question tells about the level of study to understand the academic journey. |
| What is your major or field of study department? | CSE<br>SWE<br>CIS<br>ICE<br>BBA<br>EEE<br>TE<br>English<br>Other: | The student's major field of study provides contextual information to assess how performance varies across different academic disciplines. |
| The department you are currently studying is chosen by you or you are forced to take it? | Personal Choice<br>Forcefully Given | Understand whether students chose their department forcefully or not. |
| What is your current CGPA? ( e.g., 3.25) | —---- | CGPA serves as a target variable in the predictive model, representing overall academic performance. |
| How does your current CGPA compare to your CGPA from the previous semester? | Similar<br>Higher<br>Lower | This comparison helps to identify trends in academic performance and recognize improvement or decline over time. |
| What is your average grade in major courses? | 3.50-4.00<br>2.75-3.25<br>Below 2.75 | Assessing the average grade in major courses specially can provide insights into students performance within their chosen field of study. |
| Is your university question pattern or design understandable? | Yes<br>No | Understanding the clarity of the university assessment methods helps assess potential biases in academic challenges. |
| Which strategies are applied by your | Oral, written, assignment, presentation | The question helps identify the assessment methods used by the university , which can influence academic performance. |

| | | |
|---|---|---|
| university to measure your semester performance ? | Only written MCQ<br>Both MCQ and Written<br>Practical Exam Only<br>Other: | |
| Did any course teacher of a major course use a strict marking strategy while marking? | Yes<br>No | This question aims to gauge students' perceptions of grading rigor within major courses. |
| Do you have any retake courses? | Yes<br>No | This question aims to identify students who have retaken courses , providing insight into potential challenges in their academic journey that may affect performance. |
| How many hours per week do you typically spend on studying? (Hours) | 1<br>2<br>3<br>4<br>5<br>6<br>7<br>8<br>9<br>10 | This provides the insights into their study habits and potential correlation with academic performance and recognition. |
| Where do you prefer to study? | Home<br>Library<br>Other: | This question seeks to understand study locations , providing insight into their study environment preference , which may impact their performance and recognition. |
| What are your preferred study methods? | Reading From Books<br>Group Study<br>Internet Platform<br>Other | This question asks about preferred methods to understand how students learn best , which can affect their academic success and recognition. |
| Are you currently taking any extra tuition or tutoring outside of regular classes? | Yes<br>No | This question asks if a student is receiving extra academic support outside of regular class , which can influence their academic performance and recognition. |
| Are you employed part-time or full-time? | Part Time<br>Full Time<br>Not Employed | This questions asks about students employment status , which can affect their academic commitment |

| | | |
|---|---|---|
| Do you have family responsibilities that impact your study time? | Yes<br>No | This question inquires about students' employment status, which may impact their academic commitments and performance. |
| What is your Family Size? (ex: 3,4,5 members) | 1<br>2<br>3<br>4<br>5<br>6<br>7<br>8<br>9<br>10 | This question asks about the size of the students family , providing insights into their support system and potential influences on academic performance. |
| What is your parents' marital status? | Married<br>Divorced<br>Widowed | This question asks about the marital status of the students' parents , offering insights into family dynamics and their potential impact on academic performance. |
| Father's Occupation ? | Businessman<br>Doctor<br>Engineer<br>Teacher<br>Other: | This question aims to ascertain the occupation of the student's father, providing insights into family background and potential influences on academic performance. |
| Mother's Occupation ? | Housewife<br>Teacher<br>Doctor<br>Engineer<br>Businessman<br>Other | This question aims to ascertain the occupation of the student's mother, providing insights into family background and potential influences on academic performance. |
| Father's education level? | School<br>College<br>Graduate<br>Other | This question seeks to determine the educational level of the student's father , offering insights into family background on academic performance. |
| Mother's education level? | School<br>College<br>Graduate<br>Other | This question seeks to determine the educational level of the student's mother, offering insights into family background on academic performance. |
| What is your living area? | Home<br>University Hall<br>Sublet | This question aims to identify the student's living area , providing context regarding their living environment and |

| | | |
|---|---|---|
| | Other | potential influence on academic performance and recognition. |
| Do you have any financial issue? | Yes<br>No | This question asks if the student is facing financial difficulties , which can affect their academic performance. |
| Do you get support from your family regarding your study? | Yes<br>No | This question inquires whether the student studies related support from their family , providing information on their support system and its influences on academic success. |
| How long does it take you to get to campus from your residence each day? | 10 - 45 min<br>1 - 3 hours<br>4 - 5 hours<br>Other | This question focuses on how long it takes students to commute from home to university , which might have an impact on their family and academic achievement. |
| Do you have access to a computer, internet, and relevant software/tools? | Yes<br>No | This question inquires if students have access to crucial technology which are critical for academic activities and may influence performance. |
| Do you have any past failures? | Yes<br>No | This questions inquiries whether pupils have had previous academic failures , offering information about their academic history and prospective obstacles. |
| Do you face any academic challenges? | Yes<br>No | This question inquires whether students have any academic obstacles , offering insight into potential hurdles to academic progress. |
| Rate your comfort in online learning | 1<br>2<br>3<br>4<br>5 | This question focuses on students' familiarity with online learning , which affects their academic experience and performance. |
| How many hours of sleep do you get on average? | 1<br>2<br>3<br>4<br>5<br>6<br>7<br>8<br>9<br>10 | This question looks for average quantity of sleep students receive , and potential influence on academic performance. |

| | | |
|---|---|---|
| On a scale from 1 to 10, how would you rate your stress levels regarding academics? | 1<br>2<br>3<br>4<br>5<br>6<br>7<br>8<br>9<br>10 | This question asks students to assess their academic stress. It will help to know how much stress they get and how it may affect their performance. |
| Do you participate in extra curricular activities? ( e.g., Sports, Arts, Clubbing etc) | Yes<br>No | This question inquires if students participate in extracurricular activities offering insights into their involvement beyond academics and the possible influences on general well being. |
| Do you engage in regular physical activity? | Yes<br>No | The question asks if students engage in regular physical exercise , which might impact their general well being and academic achievement. |
| On a scale from 1 to 5, how satisfied are you with your overall academic experience? | 1<br>2<br>3<br>4<br>5 | This question asks students to rank their happiness with their academic experience on a scale of 1 to 5 . |
| Give level to your semester fees | Very high<br>Moderate<br>Low | This question asks students if their semester fees are very expensive , average or cheap showing their financial strain. |
| Are you a scholarship holder or getting an academic waiver ? | Yes<br>No | This question inquires whether students receive scholars or academic exemptions , which might impact academic achievement. |
| Are you aware of utilizing support services provided by the institution (e.g., tutoring, counseling)? | Yes<br>No | This question asks if students use institutions that provide support services such as tutoring or counseling , which might have an influence on academic access. |
| How would you rate your current physical health? | 1<br>2<br>3<br>4<br>5 | This question asks students to score their physical health , which reflects their general well being and possible impact on academic achievement. |

| | | |
|---|---|---|
| How many hours have you spent using social media ? (e.g., Facebook, Instagram, YouTube) | 1<br>2<br>3<br>4<br>5<br>6<br>7<br>8<br>9<br>10 | This question focuses on students' social media usages , revealing their digital habits and potential on academic achievement. |
| What is your current relationship status? | Single<br>In a relationship | This question inquires about students' relationships status ,providing insights into their personal life and possible influence on academic achievement. |
| Do you have any kind of addiction? ( e.g. Smoking, Drug) | Yes<br>No | This question asks students about whether they have any addiction that might harm their health or academic performance. |
| Do you have plans to go abroad for higher education? | Yes<br>No | This question asks if students want to study abroad , suggesting their future goals and potential impact on academic achievement. |
| Mention Any Factors that Affect Your Academic Performance- | ------------- | ------------------------ |

**3.2.1.5 Data Validation**

In terms of data validation, we employed cross questioning technique in the Google Form. We checked the inputs given by the students manually and cross checked their answers to find out if there exist any inconsistency in the data. Our dataset was collected from the students of different universities using  a google form. In the google form, we added a total of 45 questionnaires along with the comment sections where the filer can add some more features that he or she thinks might be affecting their academic performances. From those questionnaires, all the questions were not required for our work but we added them for the validation of some sensitive data like CGPA. To validate the CGPA column we added 2 different questions indicating CGPA comparison and Average Grades. Any data having inconsistency with CGPA columns are removed from the dataset.

TABLE 3.2.1.5.1: DATA VALIDATION TABLE

| SL No | CGPA | CGPA Comparison | Average Grades | Retainment |
|---|---|---|---|---|
| 1 | 3.20 | Similar | 3.50-4.00 | No |
| 2 | 3.69 | Lower | 3.50-4.00 | Yes |
| 3 | 3.95 | Lower | 3.50-4.00 | Yes |
| 4 | 3.75 | Higher | 2.75-3.25 | No |

In the table we have visualized 4 students' inputs in the google form from them, we able to take only 2 data for our research because another 2 doesn’t validate their inputs and there are some inconsistencies in their data. For example, the first person said his/her CGPA is 3.20 which is similar to his/her previous CGPA but his/her average Grades are in between 3.50-4.00 that is much higher than his/her CGPA. If his/her average grade is in between 3.50-4.00 then his/her cgpa must be higher than 3.20. In the 4th data, we can see that this student claiming his/her CGPA is 3.75 but his/her average grades are in between 2.75-3.25 along with the higher CGPA than previous semester. It is not possible to have a CGPA of 3.75 with such poor average grades . For this reason we dropped 1 and 4 data. On the other hand 2 and 3 have consistent data, hence we didn’t drop them from the dataset.

### 3.2.2 Dataset Description

After filtering questioners the final dataset contains a total of 3028 data which initially have 43 columns filled with features like ; Age, Gender, Study Year, Department, Department Choice, CGPA, University Question Understanding, Strategy for measuring performance, Strict Marking, Retake Course,Study Hours Per Week,Study Place,Study Methods,Extra Tuition, Employment, Family Responsibility, Family Size, Marital Status, Father's Occupation, Mother's Occupation, Father's education level, Mother's education level, Living Area, Financial Issues, Family Support, Time to Reach Campus,Access to, Computer and Internet, Past Failure ,Academic Challenges ,Comfort in Online Learning, Average, Sleep Hours, Stress Level, Extra Curricular Activities, Physical Activity Engagement, Academic Performance, Semester Fees Level, Scholarship/Waiver, Support Service, Utilization, Physical Health, Hours Spent on Social Media, Relationship Status, Addiction, Plan for Abroad.

### 3.2.3 Data Preprocessing

In that step, the raw data has been preprocessed to make it readily available to feed into the models.

1. **Null value elimination (Row Wise):** It refers to the process of removing an entire row from a dataset where one or more values are missing. This ensures that each row in the dataset contains complete information [13].The raw data contains many null values. We have removed them.
2. **Label Encoding:** In this technique categorical data can be converted into numerical data [14]. The dataset contains some object type features that machines can't understand, By importing LabelEncoder we have changed our label column values in numerical data which are 0,1,2
3. **One-Hot Encoding:** One-Hot encoding is a technique used in machine learning to encode categorical variables into binary matrices , where each unique category is represented by a vector with a single high (1) value and the others are low (0) [15]. The dataset contains some object type features that machines can't understand.
4. **Min-Max Scaling :** Min-max scaling is a normalization approach that translates features to a specific range , often 0 to 1 , by altering the values based on the minimum and maximum values in the dataset [16] After encoding the dataset, we have applied the scaling technique to make all the values range from 0 to 1 as the values in the dataset vary in a wide range.

### 3.2.4 Feature Selection

In the time of data collection using Google Forms, we have employed cross questioning techniques to validate the data. As those cross questions are not useful for our model except validation, we have removed those data from our dataset. Then we calculated the correlation among the CGPA column and all other features and selected the best features that affected the CGPA most, either positive or negative.

The work of feature selection has been completed. We have removed the irrelevant features and used the correlation technique to find out the most relevant features. We have calculated correlation of all features with the CGPA column to select the features and select the features/factors that affect students' performance most.

Table 3.2.4.1: Correlation with Label Column

| **Feature** | **Correlation** | **Feature** | **Correlation** |
|---|---|---|---|
| Age | -0.026 | Physical Activity Engagement | -0.249 |
| Gender | 0.440 | Academic Performance | -0.237 |
| Department Choice | 0.579 | Scholarship/Waiver | -0.232 |
| University Question Understanding | -0.146 | Support Service Utilization | 0.320 |
| Strict Marking | -0.258 | Physical Health | -0.085 |
| Retake Course | -0.136 | Hours Spent On Social Media | -0.242 |
| Study Hours Per Week | 0.463 | Relationship Status | -0.147 |
| Extra Tuition | 0.217 | Addiction | 0.149 |
| Family Responsibility | -0.478 | Plan for Abroad | -0.201 |
| Family Size | 0.129 | Study Year_1st Year | 0.149 |
| Financial Issues | -0.044 | Study Year_2nd Year | 0.230 |
| Family Support | 0.150 | Study Year_3rd Year | -0.486 |
| Access to Computer and Internet | -0.144 | Study Year_4th Year | 0.147 |
| Past Failure | -0.547 | Department_BBA | -0.130 |
| Academic Challenges | -0.335 | Department_Bsc | -0.003 |
| Comfort in Online Learning | 0.074 | Department_Bss | -0.003 |
| Average Sleep Hours | 0.530 | Department_CIS | 0.022 |
| Stress Level | 0.096 | Department_CSE | -0.118 |
| Extracurricular Activities | -0.356 | Department_Chemistry | -0.004 |
| Strategy for measuring performance_Both MCQ and Written | 0.042 | Department_EEE | -0.095 |
| Strategy for measuring performance_MCQ | | Department_English | 0.284 |

| | | | |
|---|---|---|---|
| | -0.708 | | |
| Strategy for measuring performance_Only Written | 0.021 | Department_ICE | 0.013 |
| Strategy for measuring performance_Oral,Written,assignment,presentation | 0.547 | Department_MIC | 0.020 |
| Strategy for measuring performance_Practical Exam Only | 0.014 | Department_Microbiology | -0.016 |
| Strategy for measuring performance_Presentation,oral,assignment,final tests step by step | 0.022 | Department_SWE | 0.020 |
| Study Place_All | -0.008 | Department_TE | -0.029 |
| Study Place_Hall | 0.016 | Living Area_Home | -0.169 |
| Study Place_Home | -0.066 | Living Area_Mess | 0.015 |
| Study Place_Home and library both | -0.067 | Living Area_Sublet | 0.129 |
| Study Place_Library | 0.074 | Living Area_University Hall | 0.085 |
| Study Methods_Group Study | -0.128 | Time to Reach Campus_1-3 hours | -0.093 |
| Study Methods_Internet Platform | -0.248 | Time to Reach Campus_10-3 hours | -0.019 |
| Study Methods_Reading From Books | 0.328 | Time to Reach Campus_10-45 min | 0.255 |
| Study Methods_all | -0.007 | Time to Reach Campus_11-3 hours | -0.019 |
| Employment_Full Time | -0.053 | Time to Reach Campus_12-3 hours | -0.024 |
| Employment_Not Employed | -0.115 | Time to Reach Campus_13-3 hours | -0.035 |
| Employemnt_Part Time | 0.122 | Time to Reach Campus_14-3 hours | -0.369 |

| | | | |
|---|---|---|---|
| Fater's education level_College | 0.252 | Time to Reach Campus_15-3 hours | -0.025 |
| Fater's education level_Graduate | -0.309 | Time to Reach Campus_16-3 hours | -0.025 |
| Fater's education level_School | 0.131 | Time to Reach Campus_17-3 hours | -0.035 |
| Mother's education level_College | 0.226 | Time to Reach Campus_18-3 hours | -0.042 |
| Mother's education level_Graduate | -0.379 | Time to Reach Campus_19-3 hours | -0.028 |
| Mother's education level_School | 0.201 | Time to Reach Campus_2-3 hours | -0.021 |
| Time to Reach Campus_8-3 hours | -0.042 | Time to Reach Campus_20-3 hours | -0.021 |
| Time to Reach Campus_9-3 hours | -0.021 | Time to Reach Campus_3-3 hours | -0.018 |
| Semester Fees Level_Low | -0.004 | Time to Reach Campus_4-3 hours | -0.035 |
| Semester Fees Level_Moderate | 0.139 | Time to Reach Campus_4-5 hours | -0.038 |
| Semester Fees Level_Very High | -0.137 | Time to Reach Campus_5-3 hours | -0.021 |
| Time to Reach Campus_7-3 hours | -0.042 | Time to Reach Campus_6-3 hours | -0.019 |

Here, we got the correlation of every feature with the label column. Having the lowest values in the correlations means the specific feature is not affecting the label column least whereas the highest correlation numbers with the label column indicates that the specific features affect the label column most.

### 3.2.5 Train – Test Split

After preprocessing the dataset is ready to feed into the models. So, before feeding into the models the dataset needs to be split into training and testing parts. The training dataset will be used for training the models while the testing dataset will be used for testing the models. We have splitted the dataset into 80:20 ratio. 80% of the dataset will be used for training and the rest 20% will be used for testing**.**

## 3.3 Hardware & Software Requirement

The education sector is always upgrading, and there is a rising interest in using machine learning and deep learning approaches to analyze and enhance students' academic performance. The goal of this study is to investigate the use of feature engineering in this context to improve predictive modeling and recognition. The requirements for this research,

### 3.3.1 Background Study

- Previous works or related works need to be understood to find out the gaps and analyze them.

### 3.3.2 System Requirements

1. Data collection
2. Feature Selection
3. Data Processing
4. Model Development
5. Feature Engineering
6. Model Evaluation
7. Real world Applicability
8. Ethical Consideration

### 3.3.3 Tools for Data Collection

1. **Google Form**
   - Its easy interface enables the construction of organized questions, ensuring responses from participants.

2. **Google sheet or Excel sheet**
   - These spreadsheet tools offer sophisticated sorting, filtering and early analysis capabilities.

### 3.3.4 Resources

1. **Programming Languages**
    - Python, as a versatile and widely used language for machine learning and data science.
    - Libraries like NumPy, Pandas, Skit Learn, TensorFlow, Matplotlib and Seaborn
    - Jupyter Notebook or Google Colaboratory

2. **Hardware Requirements**
    - CPU
    - RAM
    - Storage

## 3.4 Project Management and Financial Analysis

**Communication Plan:**

**A. Stakeholders Meetings:**

•**Purpose:** For data collections, like what kinds of features they think need to be taken under consideration.

•**Participants:** Students, Teachers, Researchers.

•**Frequency:** For data collections we have to meet with students several times and to design a form we met teachers, scholars and researchers for guldens 7-8 times.

**B. Change Control Meeting:**

•**Purpose:** Changed our project title as per supervisors' guidelines and advice.

•**Participants:** Supervisor and team members.

•**Frequency:** For title requirements, we had several meetings.

- **Risk Management:**


Figure 3.4.1: Risk Management

### Financial analysis

Our project is a machine learning and deep learning project, and as per the requirements, we have to develop a mobile application connected to the project with an aesthetic user interface and host it to the Google Play Store. To do so it will cost some money. We are giving an estimated budget. Please note that it may change in future according to the requirements.

TABLE 3.4.1: ESTIMATED COST FOR PROJECT

| SN | Components | Estimated Cost (BDT) |
|---|---|---|
| 01. | Google Colab Premium | 1100 |
| 02. | Visiting Stakeholders | 1000 |
| 03. | Data Collection | 500 |
| 04. | Documentations and Report Writing | 1000 |
| 05. | Google Play Store Hosting | 3000 |
| 06. | Contingency (10% of total cost) | 660 |
| | **Total Estimated Cost** | **7260** |

### 3.5 Summary

The research titled " Feature Engineering For Student's Academic Performance And Reignition Using Machine Learning And Deep Learning" has effectively illustrated the potential of advanced computational methods to improve predictive analytics in the educational sector. The dataset was predominantly obtained through surveys conducted with students and stakeholders, yielding a total 3028 data points. To ensure a thorough and efficient data collection process, a Google Form containing an exhaustive set of 46 queries was designed. By utilizing a cross-questioning approach, every inquiry is deliberately designed to authenticate and bolster the dependability of answers, thereby guaranteeing a comprehensive and nuanced comprehension of the elements influencing student performance. After that preprocessing steps, data validation and feature selection processes have been applied to the dataset. To implement the data into the model. Requirements and Many risk has been taken under consideration and financial analysis is discussed tas well to complete the proposed methodology of this project.

# CHAPTER 4
# IMPLEMENTATION

## 4.1 Overview

Feature Engineering, a fundamental principle in the field of machine learning, is utilized to extract and enhance significant insights from intricate educational data. The process of converting unprocessed data into valuable features amplifies the prognostic capabilities of diverse deep learning and machine learning models. This research paper utilizes deep learning architectures, including Artificial Neural Network (ANN), Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), and a hybrid model that combines CNN and LSTM, to enhance the predictive process by identifying complex patterns and relationships within the data.

Machine Learning encompasses the application of various models, such as Gradient Boost, Random Forest, Decision Tree, and K-Nearest Neighbors (KNN), to further examine their predictive capabilities. A comparative analysis is conducted to assess the efficacy of these models in forecasting the academic performance and retention of students.

The comparative analysis aspect of the research examines the differences and synergies between these various predictive approaches. While some models concentrate on detecting patterns that predict academic success or risk of attrition, others delve into understanding the factors influencing these outcomes through sophisticated analytical techniques, By scrutinizing these methodologies, the research hopes to offer a comprehensive comprehension of their respective strengths and limitations, influencing the development of more effective and reliable predictive tools.

## 4.2 Train Model

This part will provide an overview of the entire machine learning models, deep learning models and their development process and implementation.

### 4.2.1 Model Selection and Development

Our research is based on machine learning and deep learning. Hence, we have selected some machine learning algorithms; Random Forest, Decision Tree, KNN, Gradient Boost and AdaBoost. Besides, we have selected deep learning algorithms like ANN, CNN, LSTM and CNN-LSTM.

**Gradient Boosting :** An effective boosting approach, Gradient Boosting turns multiple weak learners into a single strong on by training each new model to minimize the loss function, which can be mean squared error or cross entropy used by preceding model.To train a new weak model that minimizes the gradient the approach iteratively calculates the loss function gradient relative to the prediction of the current ensemble, the process in repeated until a stopping threshold has been reached. Gradient boosting assists feature engineering by providing insights into feature relevance, recognizing complex interactions and addressing missing data internally, It automatically performs non-linear transformations and suggests additional features based on decision tree splits. Additionally , it facilitates feature selection by finding the most relevant characteristics and is robust to irrelevant features, enriching overall model efficiency and performance without heavy manual changes [17].

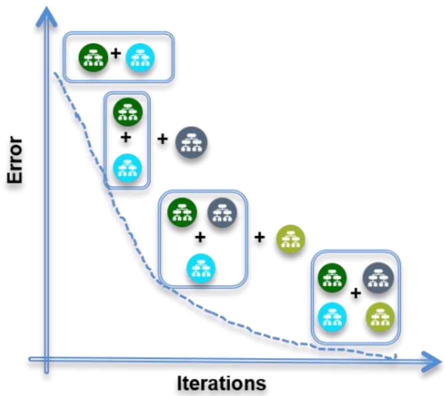



Figure 4.2.1.1: GradientBoost Architecture

**AdaBoost :** AdaBoost , short form Adaptive Boosting, is a cornerstone technique in machine learning developed to boost prediction accuracy. It achieves this by transforming weak learners, mainly decision stumps (single-level decision trees), into a strong combined classifier. AdaBoost distinguishes itself through its sequential training approach, in which each succeeding model focuses on resolving the mistakes of the prior one, thus steadily boosting the overall prediction accuracy. AdaBoost improves the predictive accuracy of a model in feature engineering by continuously improving the features. A weak learner is first trained on the dataset using equal instance weights. After calculating the models errors , weights are raised for the examples that were predicted erroneously , therefore focusing the next weak learner on these difficult problems. Through each iteration, this sequential modification highlights importance aspects that contribute most to prediction accuracy. hence increasing the models overall performance [18].

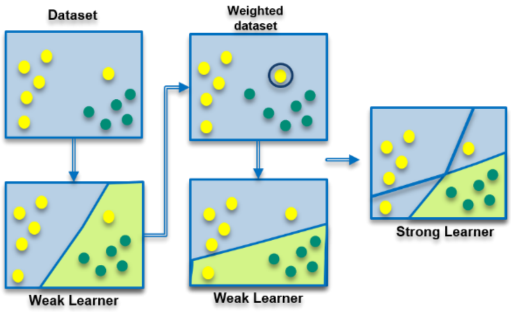



Figure 4.2.1.2: AdaBoost Architecture

**Random Forest :** The Random forest algorithms is a robust tree-based learning technique in machine learning ,which creates multiple decision trees during training. Each tree is created from a random subset of the dataset and assesses a random subset of characteristics at each split. This injection of randomizations promotes heterogeneity among the trees, limiting overfitting and boosting overall predictive accuracy. During prediction, the algorithm consolidates the outcomes of all tree via voting in classification tasks or averaging in regression tasks. This collaborative strategy harnessing the diverse insights from multiple trees, delivers steady and accurate outcomes. Random forests are frequently utilized for both classification and regression problems due to their capacity to manage complex data and offer reliable predictions across multiple settings. this method increases feature engineering by providing feature relevance scores that highlight the most significant characteristics, thereby aiding in feature selection and prioritizing. It captures complicated connections between features , suggests new features based on tree splits and decreases the impact of redundant or irrelevant features. Additionally random forest handles missing values effectively and contributes in dimensionality reduction by focusing on the most relevant features. It ensemble nature provides stability and generalization, making is a helpful tool for managing complicated data and enhancing model performance[19].

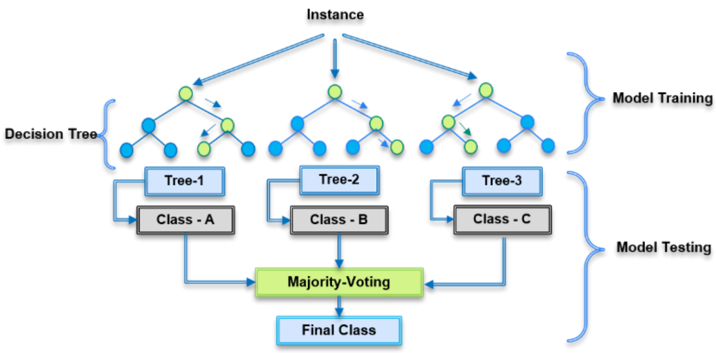



Figure 4.2.1.3: Architecture of Random Forest

**Decision Tree :** A decision tree is an effective flowchart-like structure used for decision-making and predictions. It features nodes that denote tests or judgments based on qualities, branches that demonstrate the outcomes of these tests and leaf nodes that reflects ultimate predictions or outcomes. each internal node corresponds to a specific attribute test and each leaf node stores a class label or a continuous value, resulting to clear and interpretable decision path. It substantially improve feature engineering by providing clear insights into the relevance and interconnections of features. They find essential traits through their dividing criteria, highlighting the most influential features for creating predictions. By evaluating the structure of the tree one can find correlations and interactions between features that may not be visible. Additionally, the splits at various levels of the tree can suggest valuable new features or thresholds for producing binary features. This technique simplifies features selection and construction, ensuring that the most important and impactful simplifies feature selections and construction ensuring that the most important and impactful characteristics are utilized in the model [20] .

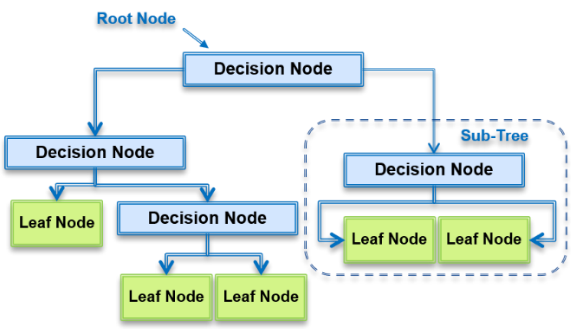



Figure 4.2.1.4: Decision Tree Architecture

**KNN:** K-Nearest Neighbors (KNN) is a basic yet successful classification technique in machine learning. It operates by locating the ‘K’ nearest data points to given input from the training dataset. For classification problems, the algorithm assigns the input to the class most prevalent among its ‘k’ nearest neighbors. For regressing problems, it predicts the value based on the average or median of the ‘k’ nearest points. KNN is non-parametric, meaning it doesn't make assumptions about the data distribution and it’s lazy, meaning it doesn't learn an explicit model but makes decision based on the full dataset at prediction time. This simplicity makes KNN straightforward and easy to implement ,albeit it can be computationally intensive with large datasets. KNN can aid in feature engineering by providing insights into feature relevance through its proximity-based categorization or regression. By analyzing which features contribute most to the nearest neighbors classification or predictions, KNN helps find influential features that significantly affects result [21].

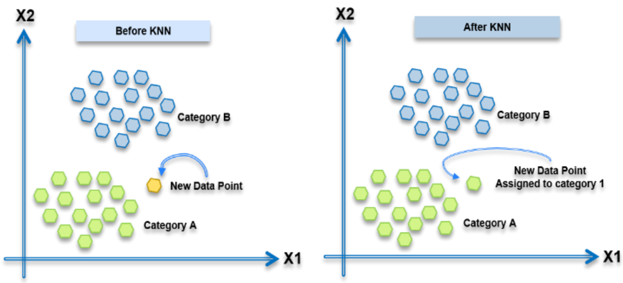



Figure 4.2.1.5: KNN Architecture

**Artificial Neural Network ;** An artificial neural network is a computational model inspired by biological neural networks that seeks to imitate the way the human brain processes information. It consists of interconnected nodes , or artificial neurons, structured in layers. Information flows through the network from input nodes, through hidden layers where computations occur, to output nodes that produce the final results. ANNs are trained using algorithms such as backpropagation, altering weights between neurons to minimize prediction errors. They thrive at complex tasks like picture and speech recognition, natural language processing and pattern detections, utilizing their ability to understand nuanced patterns and relationships from enormous datasets. By automatically learning and extracting pertinent features from raw data through their hidden layers, ANN helps with feature engineering by minimizing the requirement for human feature generation. Their ability to capture intricate, non-linear correlations between features improves the accuracy of predictive modeling because they use the learned features to predict outcomes in tasks in many fields [22].

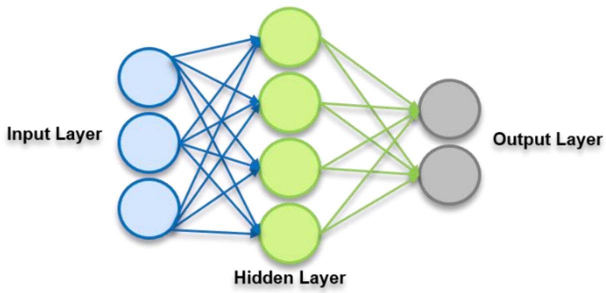



Figure 4.2.1.6: ANN Architecture

**Convolutional Neural Network :** A convolutional neural network is a specialized deep learning method built for data analysis. It plays an important role in featuring engineering by automating the process of feature extraction from raw data , notably images. In CNNs, convolutional layers apply filters to input data to capture crucial features such as edges , textures, and forms ,while pooling layers reduce dimensionality and retain important information. This hierarchical technique allows CNNs to learn and extract increasingly complicated features at each layer, removing the need for manual feature engineering. By doing so, CNN boost the model's ability to recognize detailed patterns and structure , leading to more accurate and robust predictions in tasks like picture categorization, object identification and beyond [23] .

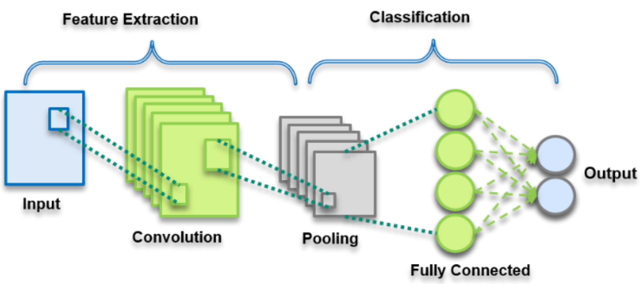



Figure 4.2.1.7: CNN Architecture

**Long Short-Term Memory (LSTM) :** It is a particular sort of recurrent neural network developed to successfully capture long-term dependencies in sequential data. It achieves this by keeping a memory cell with gates that govern the flow of information, enabling it to selectively recall or forget past information over extended sequences. LSTM are frequently utilized in tasks such as natural language processing and in other fields where understanding context and temporal relationships is crucial for correct modeling and prediction. In feature engineering LSTMs automate the extraction of relevant features from sequential data by learning complex patterns and dependencies over time directly from raw inputs, the eradicating the need for manual feature extraction. This capability enhances the model's comprehension of temporal relationships within the data [24].

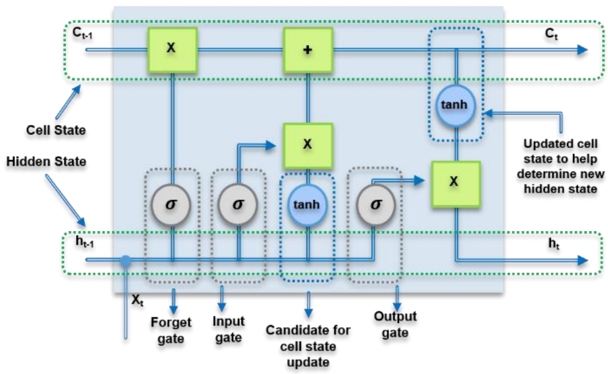



Figure 4.2.1.8: LSTM Architecture

**CNN-LSTM:** The hybrid CNN-LSTM model integrates Convolutional neural networks with Long Short-Term Memory networks to leverage their respective strengths. CNN excel at detecting local patterns, while LSTM are adept at comprehending temporal sequences. By combining these architectures the hybrid model effectively manages sequential data , capturing both local and global properties for more comprehensive analysis and prediction.

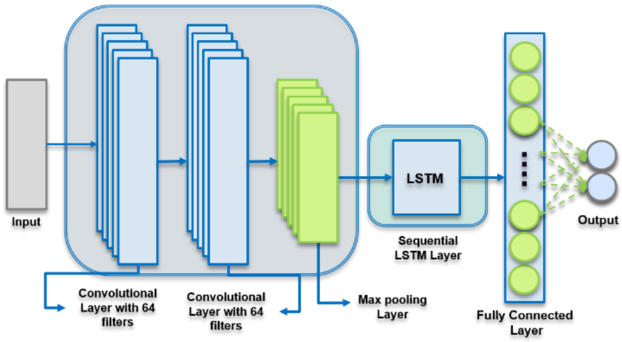



Figure 4.2.1.9: CNN-LSTM Architecture

### 4.2.2 Pseudocode of Model Development

**Step 1:** Import necessary libraries: pandas, numpy, scikit-learn, and TensorFlow.

**Step 2:** Read the dataset from the CSV file into a pandas DataFrame.

**Step 3:** Define a list of columns that are not relevant to the study to drop from the Data Frame. Drop the specified columns from the Data Frame.

**Step 4:** Check for missing values in the Data Frame and drop columns with missing values.

**Step 5:** Define a list of binary columns for label encoding. Encode binary columns using Label Encoder.

**Step 6:** Perform one-hot encoding for categorical columns.

**Step 7:** Convert all non-target columns to integer data type and specify the target variable column name.

**Step 8:** Calculate correlations between features and the target variable. Select features with correlation less than 0.3 and greater than -0.3 with the target variable. Drop the selected features from the Data Frame.

**Step 9:** Separate the target variable (Y) and features (X). Standardize the feature values using Standard Scaler.

**Step 10:** Define a function for k-fold cross-validation with specified model, features, and target variable. Initialize lists to store evaluation metrics for each fold.

**Step 11:** Split the data into k folds using K Fold. Iterate over each fold for training and validation. Fit the model on the training data. Predict the target variable for the validation data. Calculate evaluation metrics: R2 score, Mean Squared Error, Mean Absolute Error, and Root Mean Squared Error for each fold. Calculate average evaluation metrics across all folds.

**Step 12:** Create an instance of Random Forest Regressor model with specified parameters. Perform k-fold cross-validation for the Random Forest Regressor model. Print the average evaluation metrics for Random Forest Regressor.

**Step 13:** Repeat the step 12 for Decision Tree Regressor, K Neighbors Regressor, AdaBoost Regressor, Gradient Boosting Regressor, and Neural Network models.

**Step 14:** Define a function for k-fold cross-validation with the Neural Network model. Then, define a Sequential model architecture with specified layers and activation functions and compile the Sequential model with specified optimizer and loss function.

**Step 15:** Perform k-fold cross-validation for the Neural Network model. Print the average evaluation metrics for Neural Network. Plot average training and validation loss curves for the Neural Network model.

**Step 16:** Define the number of folds for the K Fold cross-validation. Create a K Fold object with a specified number of folds. Initialize lists to store evaluation metrics for each fold and loss values for plotting. Iterate over each fold for training and validation.

**Step 17:** Define a Convolutional Neural Network (CNN) model architecture with specified layers and activation functions. Compile the CNN model with specified optimizer and loss function. Train the CNN model on the training data. Record training and validation loss for plotting.

**Step 18:** Evaluate the CNN model on the validation data. Calculate evaluation metrics: R2 score, Mean Squared Error, Mean Absolute Error, and Root Mean Squared Error for each fold. Calculate mean evaluation metrics across all folds. Print the mean evaluation metrics for the CNN model. Plot the loss curve for the CNN model.

### 4.2.3 Experimental Setup

Python is a well acclaimed programming language for deep learning applications and we used it to develop our machine learning and deep learning models for the study of "Feature Engineering for Student's Academic Performance and Reignition Using Machine Learning and Deep Learning". We ran the whole experiment on Google Colab, which offered a strong computing environment that was vital to our study. Google Colab dramatically accelerated the training of deep learning models by providing over 360 GB of GPU memory and up to 16 GB of RAM using Tesla GPUs. This cloud based platform ensured great performance and accessibility during the development and training of the models. We looked at a number of deep learning and machine learning models as well as a hybrid model that combines various techniques for improved performance. Tesla GPUs' enhanced processing capacity was used to train the models, cutting training time and facilitating quick testing.

## 4.3 Model Evaluation

This step is one of the most important in this research. Model performance was evaluated using evaluation measures such as Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Squared Error (MSE) and R2 Score. We have evaluated everything by applying k fold cross validation technique. We used k = 5 which means we have divided the whole dataset into 5 partitions and checked the model performance by using each partition at a time, and by doing the mean of all, we got the final R2 score, MAE, MSE, and RMSE

**Accuracy ($R^2$ Score):** A statistical indicator showing the degree to which the regression predictions match the actual data points. A model that explains all of the variability in the response data has an $R^2$ value of 1, whereas a score of 0 implies no explanation at all.

$$R^2\ Score = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \quad (1)$$

**MAE:** The mean of the absolute deviations between the actual and projected values. Without taking into account the direction of the mistakes, it gives an estimate of the size of errors in a collection of predictions.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (2)$$

**MSE:** The squared difference between the expected and actual values, averaged. Because greater mistakes are penalized more than smaller ones, it provides an overall impression of the error.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{3}$$

**RMSE:** The square root of the mean squared error. It is easier to read since it shows the average size of the mistakes in the same units as the original data.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{4}$$

**4.4 Summary**

The focal point of the research project titled "Feature Engineering for Student's Academic Performance and Retention Using Machine Learning and Deep Learning" is the utilization of sophisticated machine learning methods to improve student retention and academic performance through predictive analytics. The investigation centers on student accomplishment, which is a pivotal concern for academic institutions. The study investigates the difficulties linked to conventional approaches in forecasting academic achievements and student retention. In an effort to surmount these constraints, the research incorporates advanced machine learning and deep learning models with feature engineering. To do train models data preprocessing, feature selection , data validation process is applied and then a comparative analysis on machine learning and deep learning algorithms is conducted. The effectiveness of contemporary computing resources in carrying out top-notch machine learning and deep learning research was demonstrated by the seamless integration of Python with Google Colab's cloud services. Which enabled effective model training and testing. This configuration emphasizes how crucial it is to use cutting edge programming tools, a strong computing infrastructure, and cloud-based services in order to facilitate productive and successful scientific research.

# CHAPTER 5
# RESULT ANALYSIS

## 5.1 Overview

The analysis and the findings of the investigations carried out to assess how well different deep learning and machine learning models performed in predicting the academic performance of students are presented in this part. The study included four deep learning models (ANN, CNN, LSTM, and a hybrid CNN-LSTM model) in addition to five machine learning models (Random Forest, Decision Tree, KNN, AdaBoost and Gradient Boost). A k fold cross-validation with k = 5 was used to assess each models and performance measures were produced to compare them, including accuracy, MAE, MSE, and RMSE. Finding the top performing models and learning about their predictive power were the main objectives with an emphasis on how effectively these models can support the enhancement of educational outcomes.

## 5.2 Experimental Result

- **Result of Experiment 1: Machine Learning**

In this study, we have employed a total of five machine learning algorithms; Random Forest, Decision Tree, KNN, AdaBoost and GraidentBoost. We calculated the evaluation metrics using k fold cross validation techniques and calculated the R2 Score, MAE, MSE and RMSE for each of the algorithms.

TABLE 5.2.1: PERFORMANCE METRICS OF ML MODELS

| Models | Accuracy | MAE | MSE | RMSE |
|---|---|---|---|---|
| Random Forest | 94.61 | 0.0586 | 0.0139 | 0.1176 |
| Decision Tree | 93.78 | 0.0558 | 0.0158 | 0.1253 |
| GradientBoost | 92.29 | 0.0772 | 0.0198 | 0.1405 |
| KNN | 88.65 | 0.0931 | 0.0290 | 0.1699 |
| AdaBoost | 81.74 | 0.1707 | 0.0471 | 0.2168 |

Table 5.2.1, shows the performances of every machine learning model. The Random Forest outperformed other machine learning models achieving the highest accuracy of 94.61% with the lowest error rates of 0.0586, 0.0139 and 0.1176 for MAE, MSE and RMSE respectively. Decision Tree obtained a outstanding accuracy of 93.78%

which is very close to the Random Forest with good error values of 0.0558, 0.0158 and 0.1253 for MAE, MSE and RMSE followed by GradientBoost, KNN and AdaBoost with accuracy of 92.29%, 88.65% and 81.74%. Here, the worst performing model is AdaBoost having the lowest accuracy and the highest values for all error metrics of 0.1707, 0.0471 and 0.2168 for MAE, MSE and RMSE respectively.

- **Result of Experiment 2: Deep Learning and Hybrid**

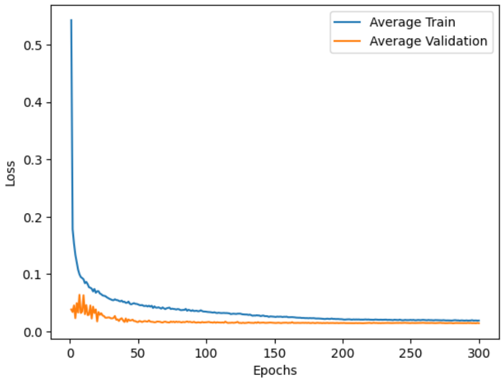
The study employs 3 deep learning algorithms; ANN, CNN and LSTM along with a hybrid model combined of CNN and LSTM. For the evaluation of the models, we also calculated the R2 score, MAE, MSE and RMSE. Obviously, all of them are evaluated using k fold cross validation technique where k = 5.

TABLE 5.2.2: PERFORMANCE METRICS OF DL MODELS

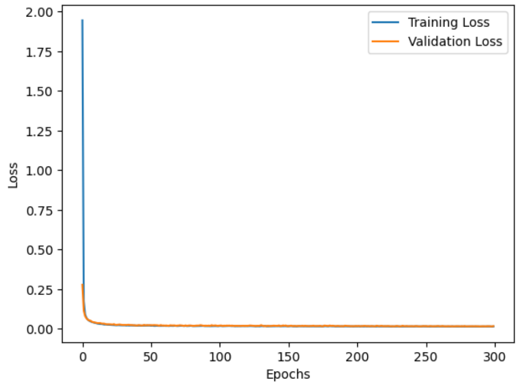
| **Models** | **Accuracy** | **MAE** | **MSE** | **RMSE** |
|---|---|---|---|---|
| ANN | 94.50 | 0.0638 | 0.0143 | 0.1186 |
| CNN-LSTM | 93.48 | 0.0805 | 0.0167 | 0.1293 |
| LSTM | 93.27 | 0.0668 | 0.0174 | 0.1316 |
| CNN | 93.19 | 0.0832 | 0.0175 | 0.1320 |

Here, Table 5.2.2 shows the evaluation metrics for the deep learning models. ANN is the best performing model among all deep learning models achieving the highest accuracy of 94.50% and the lowest error metrics of 0.0638, 0.0143 and 0.1186 for MAE, MSE and RMSE respectively. The hybrid model CNN-LSTM performed very close to the ANN having accuracy of 93.48% followed by LSTM and CNN with accuracy of 93.27 and 93.19% respectively. CNN is the worst performing model among deep learning models having the highest error values of 0.0832, 0.0175 and 0.1320 for MAE, MSE and RMSE.

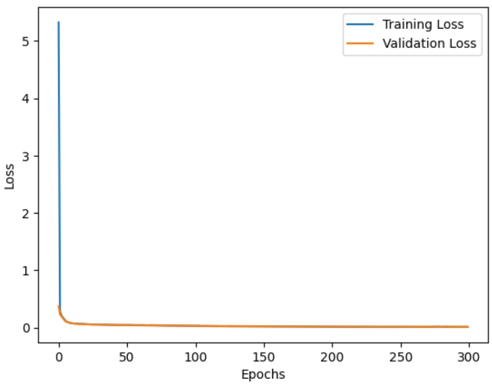
**Training and Validation loss of Deep Learning Models:**



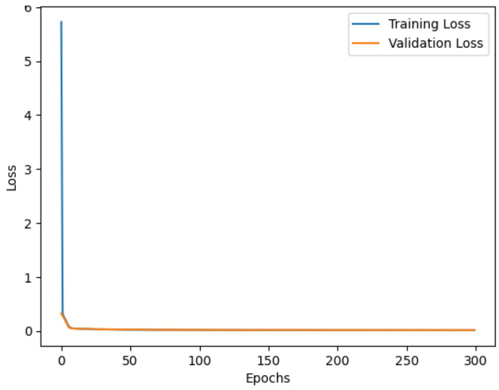

ANN




CNN




LSTM




CNN-LSTM

Figure 5.2.1: Loss Curve of All DL Models

Fig 5.2.1 shows the training and validation loss of all deep learning models. It is clear in the figure that initially or in the first epoch the loss was the highest but as per the epochs increases there is a tendency to reduce the loss rapidly and in the final epoch the loss is the lowest. It is also mentionable that all the models are best fit except the ANN. To overcome the overfit or underfit problem we follows parameter tuning approach and try different parameter values to overcome this problem.

## 5.3 Comparative Analysis

In this study, we have applied both machine learning and deep learning models. A total of 9 models were applied from where 5 were machine learning and 4 were deep learning models. There is a difference between machine learning and deep learning models in terms of performance. Fig 5.2.2. shows the accuracy of different models. Here, it is shown that the best performing model is a machine learning model Random Forest having 94.61% Accuracy, ANN is the second best performer with 94.50% accuracy. All deep learning models have performed exceptionally well but there is an ups and down in the performance of machine learning models

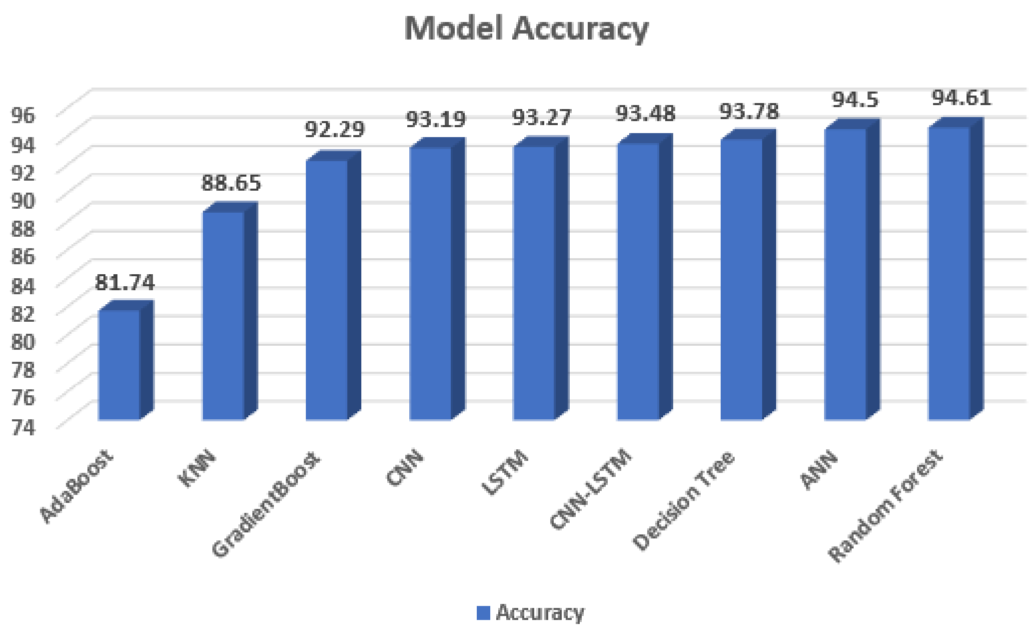



Figure 5.2.2: Accuracy of All Models

The model's performance is also evaluated using error metrics like MAE, MSE and RMSE which gives strong evidence to the accuracy of model performance. The less the error value, the better the accuracy.

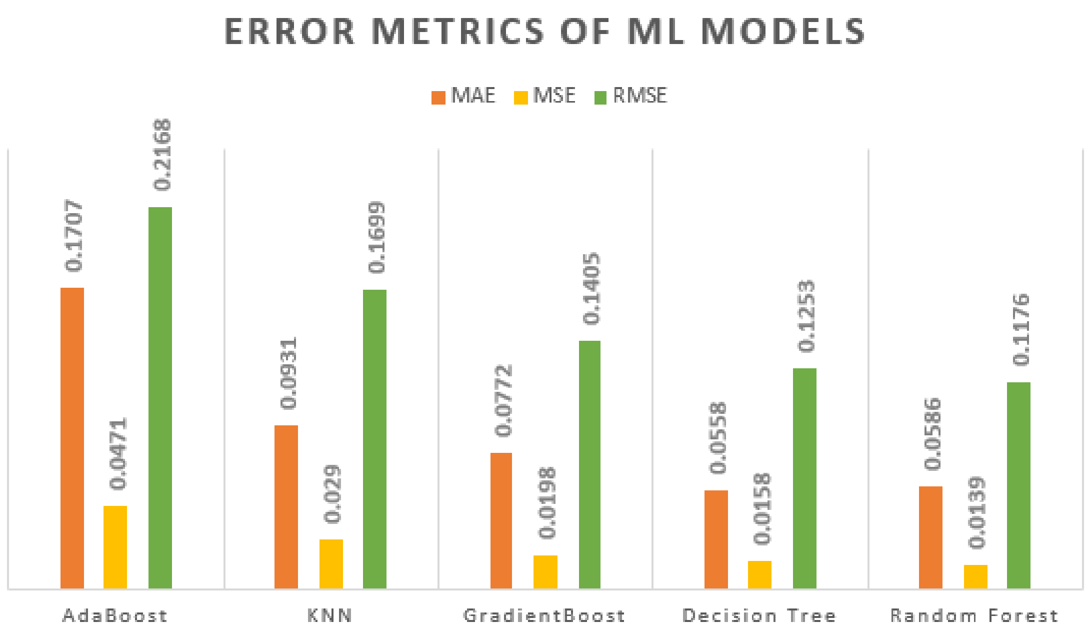



Figure 5.2.3: Error Metrics of All ML Models

The Fig 5.2.3 shows that AdaBoost shows the highest error rate of 0.1707, 0.0471 and 0.2168 for MAE, MSE and RMSE among all machine learning models whereas Random Forest shows the lowest error rate.

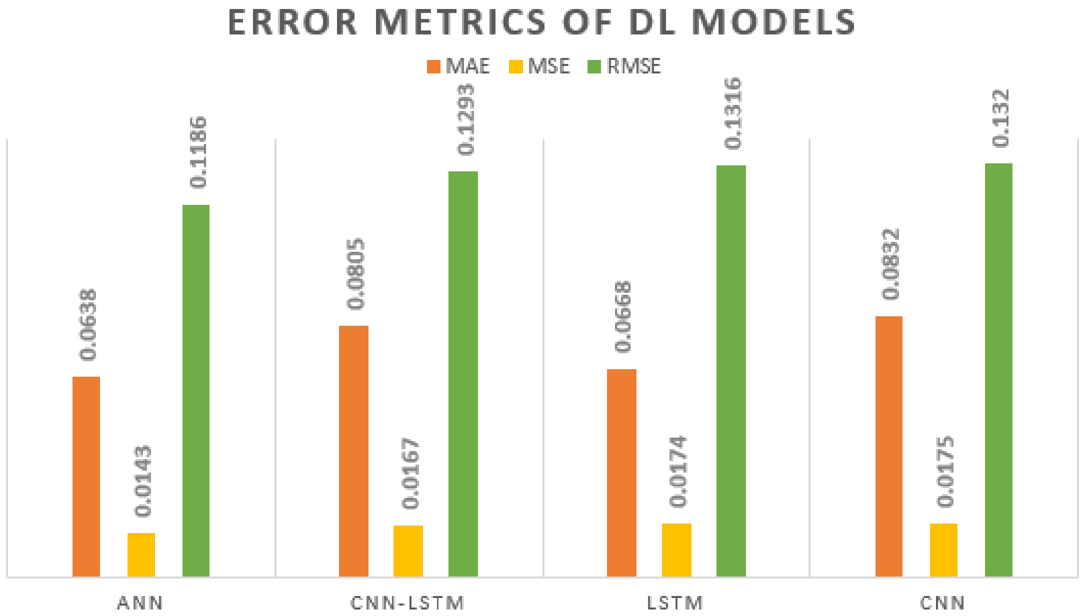


Figure 5.2.4: Error Metrics of All DL Models

Fig 5.2.4 shows the error metrics for the deep learning model and demonstrates that ANN achieved the best performance with lowest error rate of 0.0638, 0.0143 and 0.1186 for MAE, MSE and RMSE whereas ANN has the highest number of errors of 0.0832, 0.0175 and 0.1320 for MAE, MSE and RMSE. In comparison with the machine learning models, all deep learning models performed well with good accuracy.

**5.4 Features Affecting the Performance and Reignition**

One of the most important aspects of this study is to find out the most important features from the dataset that are affecting the performance of students most either positive or negative. We find out the most affecting features by analyzing the correlation between the individual feature and the label feature. We have taken those features in consideration that obtains a correlation value of greater than 0.4 and less than -0.4. The selected features are; Gender, Department Choice, Study Hours Per Week, Family Responsibility, Past Failure, Average Sleep Hours, Study Year, Strategy of Measuring Performance_MCQ, Strategy of Measuring Performance of_Oral, Written, Assignment, Presentation.

TABLE 5.4.1: SELECTED FEATURES THAT AFFECT STUDENTS PERFORMANCE

| Feature | Correlation Value | Relation |
|---|---|---|
| Gender | 0.440479 | It's a binary class. Female indicating 0 while Male 1. The positive correlation indicates if the student is male, it is more likely to have a better CGPA. |
| Department Choice | 0.579605 | It's a binary class where personal choice is 1 whereas Forcefully Given is 0. If the choice is personal then it is more likely to have a better CGPA. |
| Study Hours Per Week | 0.463141 | The increase in Study hour will increase the CGPA |
| Family Responsibility | -0.478851 | If there is Family Responsibility then the CGPA will decrease. |
| Past Failure | -0.547325 | The number of past failures negatively affects the CGPA. |
| Average Sleep Hours | 0.530137 | The higher number of Average Sleep is responsible for increasing the CGPA. |
| Study Year | -0.486282 | As the study year increases the CGPA will decrease. It affects negatively. |
| Strategy Of Measuring Performance_MCQ | -0.708359 | This measurement strategy negatively affects the CGPA |
| Strategy of Measuring Performance_Oral, Written, Assignment, Presentation | 0.546933 | This measurement strategy positively affects the CGPA |

There are some reignitions for students so that the students can overcome the problems they are facing in terms of academic performance and do better performance in the academia;

1. **Gender-Specific Interventions**
   **Observation:** Male students tend to have a higher CGPA.
   **Recommendation:** Create support programs specifically designed to meet the requirements of female students to guarantee that they are given equal

opportunities and encouragement in the classroom. This might involve study groups, courses for boosting confidence, and mentoring programs.

2. **Department Choice**
   **Observation:** Students with personal choice in department selection tend to have better CGPAs.
   **Recommendation:** Students should be encouraged to select their departments according to their areas of interest and strength. Hold counseling sessions to assist students in choosing their academic trajectories.

3. **Study Hours Per Week**
   **Observation:** Increased study hours are associated with higher CGPA.
   **Recommendation:** Encourage efficient time management and study practices. Offer calm, designated study areas as well as workshops and information on creating study plans. Encourage regular study habits as opposed to cramming right before exams.

4. **Family Responsibility**
   **Observation:** Family responsibility negatively impacts CGPA.
   **Recommendation:** Offer assistance to students who have substantial family commitments. This might entail providing access to family support resources, online course alternatives, and flexible scheduling. Develop awareness campaigns to educate families on the value of setting aside time for schoolwork.

5. **Past Failure**
   **Observation:** Past failures negatively affect CGPA.
   **Recommendation:** Provide academic assistance services and counseling to students who have failed in the past. Provide peer mentorship, workshops, and tutoring with an emphasis on resilience and learning from errors made in the past.

6. **Average Sleep Hours**
   **Observation:** Higher average sleep hours are associated with higher CGPA.
   **Recommendation:** Educate students on the significance of appropriate sleep for academic achievement. Create efforts to promote good sleep habits, such as sleep hygiene seminars, and give resources like nap rooms or wellness programs to assist students manage their sleep better.

7. **Study Year**
   **Observation:** CGPA tends to decrease as study year increases.
   **Recommendation:** Provide support systems for upper-year students who may face increased academic pressure. To keep students engaged and goal-focused, offer career counseling, stress management sessions.

8. **Performance Measurement Strategies**
   **Observation:** MCQ-based assessment negatively affects CGPA, whereas performance measured through oral, written, assignment, and presentation methods positively affects CGPA.
   **Recommendation:** Assessing methods should be reevaluated and varied to lessen dependency on MCQs. Increase the number of written assignments , oral presentations and practical projects in the curriculum to effectively assess students' knowledge and abilities.

### 5.5 Summary

The research titled " Feature Engineering For Student's Academic Performance And Reignition Using Machine Learning And Deep Learning" has effectively illustrated the potential of advanced computational methods to improve predictive analytics in the educational sector. total 3028 responses have been collected that offered a comprehensive understanding of the factors that affect student performance by meticulously designing a Google Form that included 46 comprehensive inquiries. Feature engineering was instrumental of unprocessed data into valuable features , thereby enhancing the prognostic capabilities of a variety of machine learning and deep learning. With the lowest error rates (MAE: 0.0586 , MSE: 0.0139 , RMSE: 0.01176) and the maximum accuracy of 94.61%, Random Forest proved to be the most successful machine learning mode evaluated. Closely behind with an accuracy of 93.78% and similarity low error rates (MAE: 0.0558, MSE: 0.0158, RMSE: 0.1253) was the Decision Tree Model. With accuracy of 92.29% , 88.65% and 81.74% respectively, GradientBoost , KNN and AdaBoost growing error metrics suggested their relatively worse performance. Artificial Neural Networks (ANN) outperformed other deep learning models with accuracy of 94.50% and error metrics (MAE: 0.0638,MSE: 0.0143 , RMSE: 0.1186). With accuracy of 93.48%, the hybrid CNN-LSTM model trailed closely behind, while LSTM and CNN, with somewhat higher error rates, each obtained accuracies of 93.27% and 93.19% respectively.

# CHAPTER 6
# IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY

## 6.1 Impact on Life

This research can have significant effect on human lives, promoting personal growth and empowerment. Customizing education to cater to the individual requirements of every student foster self-assurance, improves scholastic achievements and nurtures a passion for occurring knowledge. Adopting this customized method can result in improved job prospects and upward social mobility, enabling individuals to attain financial security and an enhanced standard of lives. At a societal level, this research has the potential to enhance social fairness and integration in schooling. Promoting equal access to high quality education for all students, irrespective of the socio economic backgrounds, can help narrow disparities between different social groups and foster a fair and equitable society. In addition, cultivating an inclusive educational atmosphere can enhance the mental and emotional welfare of students, thereby mitigating stress and anxiety.

## 6.2 Impact on Society & Environment

Exploring feature engineering techniques to understand students academic performance and potential reignition through machine learning and deep learning presents a groundbreaking opportunity to reshape the educational landscape and by extension , society as a whole. By delving into the intricacies nuances of students data , this research endeavors to pioneer a new era of personalized learning system, finely tuned to accommodate the diverse array of learning need and preferences among students. Through such tailored approaches , educational institutions stand poised to revolutionize traditional pedagogical methods, transcending the limitations of one size fits all approach to unlock the full potential of every learner.

Furthermore , the implications go far beyond the classroom, addressing core social concerns like dropout rates and educational inequity. This study sheds light on the fundamental causes of student disengagement and reinvention , paving the door for tailored interventions and support measures. Such discoveries offer the possibilities of not just slowing the flow of dropouts, but also of building a culture of educational inclusion and fairness. From disadvantaged areas to marginalized people , the potential to close socioeconomic gaps in educational success is significant , proving a light of hope for a more just and equal society.

In essence , this study marks a watershed event in the growth of educational procedures and public expectations. By promoting inclusive and effective learning , it lays the groundwork for a future in which all individuals , regardless of background or situation, have the chance to grow and achieve. This research , with its revolutionary potential . demonstrates the lasting power of innovation and limitless possibilities it holds for crafting a brighter , more egalitarian future for all.

The impacts of this research on the environment , though not immediately tangible, holds significant potential to shape long term sustainability practices. While the direct environmental footprint may seem minimal, the indirect effects are far-reaching and profound. As educational outcomes are enhanced and dropout rates reduce through personalized learning system , a more educated populace emerges, often characterized by heightened environmental awareness and a propensity for sustainable behaviors, This shift towards eco-conscious attitudes and actions among students can reverberate across generations , fostering a culture of environmental stewardship that transcends the confines of academia.

In addition , the application of deep learning and machine methods to maximize educational interventions creates opportunities for innovative resource management strategies in educational establishments. Through the analysis of extensive datasets and trends , these technologies enable the creation of effective strategies for allocating resources , leading to the reduction of waste and consumption of energy. Beyond the walls of educational institutions , these innovations have a ripple impact on larger society practices and promote a more sustainable mindset. example of these ripple effects include improved classroom operations and optimized transaction logistics.

In essences, long terms environmental stewardship is sparked by the adoptions of sustainable behaviors, the promotion of sustainable behaviors and the promotions of knowledge , even through the effects on the environment might not be seen right way This research helps to foster a more environmentally conscious society by providing people with the information, abilities and mentality needed to handle environmental concerns. By means of a harmonious fusion of scientific discovery and progressive education, we clear the path for a day when sustainability is not just a target but a way of life that us deeply woven into the fabric of society.

### 6.3 Ethical Aspects

Research involving student data and educational outcomes is fundamentally grounded in ethical considerations , which require strict adherence to principles of privacy , confidentiality, and impartiality. Maintaining the confidentiality of student information requires rigorous compliance with data protection regulations and, when necessary, obtaining informed consent in order to prevent unauthorized access or misuse. Furthermore , it is crucial to exercise caution regarding bias in both data and algorithms, given that skewed models have the potential to sustain inequities and marginalize specific student demographics Transparency is an essential element of ethical research conduct, which requires the unambiguous recording of data collection methods , feature selection procedures, and model interpretations in order to promote responsibility and inspire confidence among interested parties.

In addition , ethical scholars are required to navigate the intricate ethical terrain that pertains to the possible consequences of their decisions on policies and practices in education. Scholars have an obligation to promote fairness , equality and inclusivity throughout the entire research process , in recognition of the profound implications their work may have. This involves a dedication to critically examining the fundamental presumptions and prejudices that are intrinsic in educational system, with aim of deconstructing obstacles to entry and chance , while fostering avenues for empowerment and achievement for every student , irrespective of their socio-economic status or personal history. Educators contribute to the development of a more just and equitable educational environment and maintain the integrity of their work by placing ethical principles at the forefront of the research process.

## 6.4 Sustainability Plan

Several exhaustive strategies can be implemented to guarantee the sustainability of this research, with a particular emphasis on open access, continuous evaluation and collaboration. Initially it is imperative to cultivate partnership with educational institutions and a diverse array of stakeholders. This collaboration can enable the seamless integration of research findings into practical applications, thereby guaranteeing that the insights obtained from the research have a tangible and meaningful impact in real-world educational settings. The relevance and efficacy of the research outcomes can be improved by customizing interventions to suit the unique needs and contexts of diverse educational environments through engagement with teachers, administrators, policymakers and even students.

The efficacy and durability of the proposed interventions are significantly evaluated through longitudinal studies and ongoing monitoring. Researchers can accumulate valuable data regarding the long term effects of these interventions by monitoring their implementation and result over extended periods. The strategies' effectiveness and responsiveness to changing educational challenges are guaranteed by this continuous feedback cycle, which allows for iterative enhancements and refinements. Regular evaluations can assists in the identification of areas that require improvement, thereby guaranteeing that the research remains adaptable and dynamic, thereby resulting in more sustainable and robust educational practices. The capacity to monitor and evaluate interventions over time enables the identification of best practices and the phasing out of less effective methods, thereby ensuring that only the most impactful strategies are retained and scaled up.

Additionally, sustainability promotion necessitates the provision of open access to research findings and resources. The research results are made readily available to promote knowledge sharing and replication, thereby facilitating a more extensive level of management and participation within the research community. Open access encourages a collaborative environment , enabling other researchers to expand upon the findings, verify the results and investigate novel research directions. This collaborative endeavor has the potential to expedite the tempo of innovation and foster a more thorough comprehension of educational practices and outcomes. Moreover , it is imperative to uphold a steadfast dedication to ethical standards and best practices in order to promote responsible research conduct and accountability. This research can make a sustainable contribution to the improvement of educational practice and student outcomes in the long term by emphasizing ethical integrity, collaboration and transparency.

### 6.5 Summary

This study investigates the application of feature engineering in analyzing student feature engineering in analyzing student academic performance and reengagement using machine learning and deep learning techniques. The ultimate goal is to revolutionize educational procedures and foster fairness and long-term viability. By tailoring education to cater to individual learning requirements, it promotes academic achievements , diminishes rates of students leaving school prematurely, and enhances job opportunities, promoting personal development and self-empowerment. At the social level, it tackles the issue of educational inequality by guaranteeing equitable access to high quality education for all pupils.

# CHAPTER 7
# CONCLUSION AND FUTURE WORK

## 7.1 Conclusions

The research titled “ Feature Engineering For Student’s Academic Performance And Reignition Using Machine Learning And Deep Learning” has effectively illustrated the potential of advanced computational methods to improve predictive analytics in the educational sector. total 3028 responses have been collected that offered a comprehensive understanding of the factors that affect student performance by meticulously designing a Google Form that included 46 comprehensive inquiries. Feature engineering was instrumental of unprocessed data into valuable features , thereby enhancing the prognostic capabilities of a variety of machine learning and deep learning. With the lowest error rates (MAE: 0.0586 , MSE: 0.0139 , RMSE: 0.01176) and the maximum accuracy of 94.61%, Random Forest proved to be the most successful machine learning mode evaluated. Closely behind with an accuracy of 93.78% and similarity low error rates (MAE: 0.0558, MSE: 0.0158, RMSE: 0.1253) was the Decision Tree Model. With accuracy of 92.29% , 88.65% and 81.74% respectively, GradientBoost , KNN and AdaBoost growing error metrics suggested their relatively worse performance. Artificial Neural Networks (ANN) outperformed other deep learning models with accuracy of 94.50% and error metrics (MAE: 0.0638,MSE: 0.0143 , RMSE: 0.1186). With accuracy of 93.48%, the hybrid CNN-LSTM model trailed closely behind, while LSTM and CNN, with somewhat higher error rates, each obtained accuracies of 93.27% and 93.19% respectively. This paper’s comparative research emphasizes how well Random Forest and ANN predict academic results and how useful they may be in educational contexts. High accuracy at risk student identification allows educational institutions to quickly and precisely deploy interventions that improve students achievement and retention. This research provides major contributions to the academic sector by delivering robust, accurate and dependable forecasting tools. The insights acquired from this study have potential to affect legislative decisions and instructional strategies, ultimately leading to improved educational outcomes and student retention rates.

## 7.2 Further Suggested Works

We tried to cover all the previous gaps we found while doing literature reviews. In future this study aims to to work on things ,Like ;

1. Data Collection for this study was massive but not from a large amount of data from different institutions. This study gathered a huge amount of data from more institutions in Bangladesh or other countries.
2. Other parameters affecting students' performance can be added according to generational situations , along with more advanced algorithms to find the best of them.

### 7.3 Limitations

In preparation for the next stages of our research on feature engineering for students academic performance and recognition utilizing machine learning and deep learning , we anticipate many critical considerations that may affect the project :

- **Data Quality and Completeness :** We anticipate the need to continuously evaluate data quality and completeness, correcting any unexpected difficulties such as missing data or anomalies in academic performance records.

- **Scalability :** We will evaluate the scalability of our models, particularly if the project is intended for use in large-scale educational contexts. Ensuring that the chosen approaches can handle rising data quantities will be a constant concern.

- **External Factors :** We are aware that external circumstances , such as changes in educational policies , may have an impact on the interpretation or use of our research findings. Flexibility and adaptability will be essential.

By proactively considering these factors , we aim to enhance the robustness and applicability of our research outcomes in the realm of students academic performance and recognition.

# References


[1] D. DeSilver, “The concerns and challenges of being a U.S. teen: What the data show,” Pew Research Center, Apr. 14, 2024. [Online]. available at << https://www.pewresearch.org/short-reads/2019/02/26/the-concerns-and-challenges-of-being-a-u-s-teen-what-the-data-show/ >>last accessed on 27-06-2024 at 12:00 PM.

[2] Amberstudent, “19 Most common student problems and Easiest ways to deal with them,” Amber, Jun. 19, 2024. available at << https://amberstudent.com/blog/post/15-most-common-problems-faced-by-students-and-their-solutions >> last accessed on 27-06-2024 at 12:00 PM.

[3] N. Hajizadeh and M. Ahmadzadeh, "Analysis of factors that affect the students academic performance-Data Mining Approach," *arXiv preprint arXiv:1409.2222*, 2014.

[4] N. R. Beckham, L. J. Akeh, G. N. P. Mitaart, and J. V. Moniaga, “Determining factors that affect student performance using various machine learning methods,” *Procedia Computer Science*, vol. 216, pp. 597–603, Jan. 2023, doi: 10.1016/j.procs.2022.12.174.

[5] F. S. Alani and A. T. Hawas, "Factors affecting students academic performance: A case study of Sohar University," *Psychology and Education*, vol. 58, no. 5, pp. 4624-4635, 2021.

[6] M. A. Al-Hagery, M. A. Alzaid, T. S. Alharbi, and M. A. Alhanaya, "Data mining methods for detecting the most significant factors affecting students’ performance," *International Journal of Information Technology and Computer Science*, vol. 12, no. 5, pp. 1-13, 2020.

[7] S. Rasul and Q. Bukhsh, "A study of factors affecting students’ performance in examination at university level," *Procedia-Social and Behavioral Sciences*, vol. 15, pp. 2042-2047, 2011.

[8] S. T. Hijazi and S. M. M. Naqvi, "Factors affecting students' performance," *Bangladesh e-Journal of Sociology*, vol. 3, no. 1, 2006.

[9] M. S. Farooq, A. H. Chaudhry, M. Shafiq, and G. Berhanu, "Factors affecting students’ quality of academic performance: A case of secondary school level," *Journal of Quality and Technology Management*, vol. 7, no. 2, pp. 1-14, 2011.

[10] S. P. Singh, S. Malik, and P. Singh, "Research paper factors affecting academic performance of students," *Indian Journal of Research*, vol. 5, no. 4, pp. 176-178, 2016.

[11] R. Kaviyarasi and T. Balasubramanian, "Exploring the high potential factors that affect students’ academic performance," *International Journal of Education and Management Engineering*, vol. 8, no. 6, pp. 15-23, 2018.

[12] I. Mushtaq and S. N. Khan, "Factors affecting students’ academic performance," *Global Journal of Management and Business Research*, vol. 12, no. 9, pp. 17-22, 2012.

[13] M. R. Chmielewski and J. W. Grzymala-Busse, “Global discretization of continuous attributes as preprocessing for machine learning,” *International Journal of Approximate Reasoning*, vol. 15, no. 4, pp. 319–331, Nov. 1996, doi: 10.1016/s0888-613x(96)00074-6.

[14] W. Etaiwi and G. Naymat, “The Impact of applying Different Preprocessing Steps on Review Spam Detection,” Procedia Computer Science, vol. 113, pp. 273–279, Jan. 2017, doi: 10.1016/j.procs.2017.08.368.

[15] S. Okada, M. Ohzeki, and S. Taguchi, “Efficient partition of integer optimization problems with one-hot encoding,” Scientific Reports, vol. 9, no. 1, Sep. 2019, doi: 10.1038/s41598-019-49539-6.

[16] S. Sinsomboonthong, “Performance Comparison of New Adjusted Min-Max with Decimal Scaling and Statistical Column Normalization Methods for Artificial Neural Network Classification,” International Journal of Mathematics and Mathematical Sciences, vol. 2022, pp. 1–9, Apr. 2022, doi: 10.1155/2022/3584406.

[17] N. Aziz, E. A. P. Akhir, I. A. Aziz, J. Jaafar, M. H. Hasan, and A. N. C. Abas, "A Study on Gradient Boosting Algorithms for Development of AI Monitoring and Prediction Systems," Oct. 2020, doi: 10.1109/icci51257.2020.9247843.
[18] H. Binder, O. Gefeller, M. Schmid, and A. Mayr, "The Evolution of Boosting Algorithms," *Methods of Information in Medicine*, vol. 53, no. 06, pp. 419–427, Jan. 2014, doi: 10.3414/me13-01-0122.
[19] W. Lin, Z. Wu, L. Lin, A. Wen, and J. Li, "An Ensemble Random Forest Algorithm for Insurance Big Data Analysis," *IEEE Access*, vol. 5, pp. 16568–16575, Jan. 2017, doi: 10.1109/access.2017.2738069.
[20] B. Charbuty and A. Abdulazeez, "Classification Based on Decision Tree Algorithm for Machine Learning," *Journal of Applied Science and Technology Trends*, vol. 2, no. 01, pp. 20–28, Mar. 2021, doi: 10.38094/jastt20165.
[21] A. Moldagulova and R. Bte. Sulaiman, "Using KNN algorithm for classification of textual documents," May 2017, doi: 10.1109/icitech.2017.8079924.
[22] G. P. J. Schmitz, C. Aldrich, and F. S. Gouws, "ANN-DT: an algorithm for extraction of decision trees from artificial neural networks," *IEEE Transactions on Neural Networks*, vol. 10, no. 6, pp. 1392–1401, Jan. 1999, doi: 10.1109/72.809084.
[23] J. Guérin, O. Gibaru, S. Thiery, and E. Nyiri, "CNN features are also great at unsupervised classification," *arXiv (Cornell University)*, Jan. 2017, doi: 10.48550/arxiv.1707.01700.
[24] N. Kadri, A. Ellouze, M. Ksantini, and S. H. Turki, "New LSTM Deep Learning Algorithm for Driving Behavior Classification," *Cybernetics and Systems*, vol. 54, no. 4, pp. 387–405, Apr. 2022, doi: 10.1080/01969722.2022.2059133.

# Appendix A

## Course Outcomes, Complex Engineering Problems (EP) and Complex Engineering Activities (EA) Addressing

**Title:** Feature Engineering For Students Academic Performance And Reignition Using Machine Learning And Deep Learning

**Student ID: 202-15-14388 & 202-15-14369**

### CO Description for FYDP

| CO | CO Descriptions | PO |
|---|---|---|
| **Phase -I** | | |
| **CO1** | Integrate recently gained and previously acquired knowledge to identify a **pneumonia detection** problem for the Final Year Design Project (FYDP) | **PO1** |
| **CO2** | Analyze different aspects of the goals in designing a solution for this FYDP | **PO2** |
| **CO3** | Explore diverse problem domains through a literature review, delineate the issues, and establish this goals for the FYDP | **PO4** |
| **CO4** | Perform economic evaluation and cost estimation and employ suitable project management procedures throughout the development life cycle of the FYDP | **PO11** |
| **Phase -II** | | |
| **CO5** | Design and develop technical solutions and system components or processes that meet specified requirements, ensuring compliance with public health and safety standards, as well as considering cultural, socioeconomic, and environmental factors in this FYDP | **PO3** |
| **CO6** | Choose and apply appropriate methodologies, resources, and contemporary engineering and IT technologies to address complex engineering processes, encompassing prediction and modeling, while adhering to relevant constraints in this FYDP | **PO5** |

| | | |
|---|---|---|
| **CO7** | Analyze societal, health, safety, legal, and cultural considerations, along with associated responsibilities, in the context of professional engineering practice and the resolution of this problem, employing logical reasoning guided by contextual understanding. | **PO6** |
| **CO8** | Comprehend and evaluate the enduring sustainability and impact of professional engineering endeavors in addressing intricate engineering challenges within social and environmental frameworks. | **PO7** |
| **CO9** | Implement ethical principles and adhere to professional standards and norms in this FYDP | **PO8** |
| **CO10** | Capable of operating proficiently both individually and as a team member or leader across diverse teams and interdisciplinary settings in this FYDP. | **PO9** |
| **CO11** | Proficiently communicate with the engineering community and broader society regarding complex engineering endeavors, including the ability to comprehend and generate comprehensive reports and design documentation, as well as provide and receive clear instructions throughout this FYDP. | **PO10** |
| **CO12** | Acknowledge the importance of self-directed and life-long learning within the evolving landscape of technology, and possess the readiness and capability to engage in lifelong learning endeavors. | **PO12** |

# Addressing CO (1 to 8), Knowledge Profile (K), Attainment of Complex Engineering Problems (EP), and Attainment of Complex Engineering Activities (EA)

**Addressing CO (1 to 8), Knowledge Profile (K), Attainment of Complex Engineering Problems (EP):**

| SN | EP Definition | Attainment | CO | Justification (with Knowledge Profile) | References |
|---|---|---|---|---|---|
| 1. | **EP1: Depth of Knowledge required** | Yes | CO1, CO2, CO3, CO5, CO6, CO7 and CO8 | This project demonstrates **fundamental engineering (K3)** principles by employing machine learning and deep learning techniques for predicting students performance. The project demonstrates **specialist knowledge (K4)** by conducting Machine Learning (Random Forest, Decision Tree, KNN, GradientBoo | **Page no:** [1] **Section:** [1.1] **Page no:** [30-38] **Section:** [4.2] |

| | | | | | |
|---|---|---|---|---|---|
| | | | | st and AdaBoost), Deep Learning (ANN, CNN, CNN, LSTM and a hybrid CNN-LSTM) enhancing the prediction of students performance which is crucial for student performance improve and welfare. | |
| | | | | The project applies **engineering practice & design (K5)** by the figure of process of experiments. The project addresses **engineering practice & technology (K6)** by employing machine | **Page no:** [13] **Section:** [3.2] **Page no:** [26-27] **Section:** [3.3] |

| | | | | | |
|---|---|---|---|---|---|
| | | | | learning and deep learning models. | |
| | | | | This project ensures to **K8 (Research Literature)** by synthesizing insights from recent studies, to the analysis and prediction of students performance, showcasing a comprehensive understanding of current methodologies. | **Page no:** [7-11] **Section:** [2.2, 2.3] |

| | | | | | |
|---|---|---|---|---|---|
| **2.** | **EP2: Range of Conflicting Requirements** | Yes | CO2, and CO7 | This project addresses **EP-2** by recognizing the hurdles in feature engineering and performance prediction, including limitations of traditional methods and the complexities of integrating machine learning and deep learning. Through comparative analysis, it confronts challenges in understanding spatial distributions, offering insights for refining predictions. | **Page no:** [11-12] **Section:** [2.4] |

| | | | | | |
|---|---|---|---|---|---|
| **3.** | **EP3: Depth of analysis required** | Yes | CO2, and CO6 | This project addresses **EP-3** by meticulously comparing experimental outcomes, highlighting Machine Learning, Deep Learning along with Hybrid model as the chosen significant solution for enhancing students performance prediction amidst multiple potential approaches. | **Page no:** [40-44] **Section:** [5.2,5.3] |

| | | | | | |
|---|---|---|---|---|---|
| **4.** | **EP4: Familiarity of Issues** | Yes | CO8 | This project's interdisciplinary approach extends beyond computer science and engineering, impacting the educational sectors like students performance and contributing to the advancements of students welfare, public as well as social benefits, and good performance which indicates **EP-4**. | **Page no:** [07-11] **Section:** [2.2, 2.3] |
| **5.** | **EP5: Extends of application codes** | Yes | CO5 | To implement the whole project we have to do python programs along with their libraries which covers **K4** and **K6** | **Page no:** [30-38] **Section:** [4.2] |

| | | | | | |
|---|---|---|---|---|---|
| **6.** | **EP6: Extends of stakeholders involved and conflicting requirements** | Yes | CO8 | As part of our project, we have visited and met with teachers from different institutions as well as their students. The purpose of this visit is to understand the specific features that affect students' performance and design our dataset accordingly. which covers **K4** and **K8** | **Page no:** [14] **Section:** [3.2] |
| **7.** | **EP7: Interdependence** | Yes | CO5 | This project's comprehensive approach addresses high-level problems by integrating various components across data collection, statistical analysis, and proposed methodology, ensuring a holistic | **Page no:** [13-47] **Section:** [3.2, 3.3, 3.4,4.2,4.3,4.4,5.2,5.3,5.4,5.5] |

| | | | | | |
|---|---|---|---|---|---|
| | | | | solution to complex challenges in predicting students performance **EP-7**. | |

**Addressing CO11 with Complex Engineering Activities (EA)]:**

| SN | EA Definition | Attainment | CO | Justification | References |
|---|---|---|---|---|---|
| 1. | **EA1: Range of resources** | Yes | CO 11 | Our project utilizes diverse resources such as high-performance computing infrastructure, GPUs, deep learning frameworks, pre processed dataset, and ethical considerations to ensure systematic research and contribute to advancements in featuring engineering | **Page no:** [13-39] **Section:** [3.2,3.3,3.4,4.2,4.3, 4.4] |
| 2. | **EA2: Level of interaction** | No | | N/A | N/A |
| 3. | **EA3: Innovation** | No | | N/A | N/A |

| | | | | | |
|---|---|---|---|---|---|
| **4.** | **EA4: Consequences for society and the environment** | Yes | | This project contributes to society by finding factors related to students academic performance. This study can help students, while also promoting environmental sustainability by employing efficient computational resources and adhering to ethical guidelines for patient data privacy. | **Page no:** [48-51] **Section:** [6.1, 6.2, 6.3,6.4] |
| **5.** | **EA-5: Familiarity** | Yes | | This project expands upon existing research by examining a novel approach in featuring engineering in students performance demonst rated through preliminary terminologies and a comprehensive comparative analysis, offering new insights into the field. | **Page no:** [7-12] **Section:** [2.1, 2.2,2.3,2.4] |

## Addressing CO (4, 9, 10, and 12):

| SN | COs | Attainment | Justification | References |
|---|---|---|---|---|
| **1** | CO4 | Yes | This project addresses **CO4** by integrating effective project management and financial oversight, ensuring meticulous planning, resource allocation, and budget estimation for optimal resource utilization throughout the research lifecycle. | **Page no:** [27-28] **Section:** [3.4] |
| **2** | CO9 | Yes | The project demonstrates adherence to ethical principles by prioritizing patient privacy, obtaining informed consent, and transparently documenting the research process, ensuring responsible knowledge dissemination and societal well-being through the ethical application of advanced healthcare technologies which comply with **CO9**. | **Page no:** [49] **Section:** [6.3] |
| **3** | CO10 | No | N/A | N/A |
| **4** | CO12 | Yes | The project's dedication to continuous learning **(CO12)** and adaptation within the dynamic technological landscape is reflected in its comprehensive data collection, rigorous statistical analysis, meticulous methodology development, and thorough experimental results and analysis, showcasing a commitment to staying updated and refining techniques to address modern challenges. | **Page no:** [13-47] **Section:** [3.2, 3.3, 3.4, 3.5, 4.2,4.3,5.2,5.3,5.4,5.5] |

# Appendix B
## LIST OF PUBLICATIONS

| Serial No | Details of Publications |
|---|---|
| 1 | Al Ryan, A., Arpita, H. D., Tabassum, A., Ahammad, M. S., & Akram, M. A. (2024, February). A Comparative Analysis Between Deep Learning and Machine Learning Algorithms Based on User Review Sentiment Analysis from Various OTT Applications. In 2024 International Conference on Computer, Electrical & Communication Engineering (ICCECE) (pp. 1-7). IEEE. |
| 2 | Sentiment Analysis of Various Ride Sharing Applications Reviews: A Comparative Analysis between Deep Learning and Machine Learning Algorithms. (n.d.). In Iccids 2024 - 7th International Conference on Computational Intelligence in Data Science. Springer IFIP AICT series |
| 3 | Deep Learning Based Sentiment Analysis of User Generated Reviews of Various AI Powered Mobile Applications"-. (n.d.). In 7th International Conference on Inventive Computation Technologies, scheduled to be held in Lalitpur, Nepal, from April 24th to April 26th, 2024. |
| 4 | Sentiment Analysis of User Generated Reviews of Online Education Applications Using Deep Learning and Transformer Learning Algorithms. In the 15TH International IEEE Conference On Computing, Communication And Networking Technologies (ICCCNT) held in June 24 - 28, 2024, IIT - Mandi, Himachal Pradesh, India. |
| 5 | Forecasting Rainfall in Bangladesh: A sophisticated exploration through machine learning approaches. (n.d.). In the 15TH International IEEE Conference On Computing, Communication And Networking Technologies (ICCCNT) held in June 24 - 28, 2024, IIT - Mandi, Himachal Pradesh, India. |
| 6 | Machine Learning for GDP Forecasting: Enhancing Economic Projections (n.d.). In the 15TH International IEEE Conference On Computing, Communication And Networking Technologies (ICCCNT) held in June 24 - 28, 2024, IIT - Mandi, Himachal Pradesh, India. |
| 7 | Empowering Maternal Health in Bangladesh: Advanced Risk Prediction with Machine Learning (n.d.) In the 15TH International IEEE Conference On Computing, Communication And Networking Technologies (ICCCNT) held in June 24 - 28, 2024, IIT - Mandi, Himachal Pradesh, India. |
| 8 | Enhancing Date Fruit Classification Through Deep Learning Approaches and Real-Time User Interface Integration (n.d.). In the 15TH International IEEE Conference On |

| | |
|---|---|
| | Computing, Communication And Networking Technologies (ICCCNT) held in June 24 - 28, 2024, IIT - Mandi, Himachal Pradesh, India. |

FEATURE ENGINEERING FOR STUDENTS ACADEMIC PERFORMANCE AND REIGNITION USING MACHINE LEARNING AND DEEP LEARNING

ORIGINALITY REPORT

| 13% | 8% | 5% | 7% |
|---|---|---|---|
| SIMILARITY INDEX | INTERNET SOURCES | PUBLICATIONS | STUDENT PAPERS |

PRIMARY SOURCES

| | | |
|---|---|---|
| 1 | www.researchgate.net<br>Internet Source | 2% |
| 2 | Submitted to Daffodil International University<br>Student Paper | 2% |
| 3 | www.journalofbusiness.org<br>Internet Source | 1% |
| 4 | Nicholas Robert Beckham, Limas Jaya Akeh, Giodio Nathanael Pratama Mitaart, Jurike V Moniaga. "Determining factors that affect student performance using various machine learning methods", Procedia Computer Science, 2023<br>Publication | 1% |
| 5 | Saima Rasul, Qadir Bukhsh. "A study of factors affecting students' performance in examination at university level", Procedia - Social and Behavioral Sciences, 2011<br>Publication | 1% |