A FRAMEWORK FOR ACADEMIC PERFORMANCE ANALYSIS OF STUDENTS IN ONLINE LEARNING USING MACHINE LEARNING APPROACHES

Thesis Submitted for the Award of the Degree of

DOCTOR OF PHILOSOPHY

IN

(Computer Science and Engineering)

by

(Rakshit Khajuria)

Registration No.12020422

Supervised By

Dr. Anuj Sharma,20592

Associate Professor in School of Computer Science and Engineering at Lovely Professional University, Phagwara (Punjab)

Co-Supervised by

Dr. Ashok Sharma

Assistant Professor in the Department of Computer Science and IT

Bhaderwah Campus, University of Jammu

LOVELY PROFESSIONAL UNIVERSITY, PUNJAB

January,2026

DECLARATION

I, hereby declared that the presented work in the thesis entitled "A Framework for Academic Performance Analysis of students in online Learning Using Machine Learning Approaches" in fulfilment of degree of **Doctor of Philosophy (Ph. D.)** is outcome of research work carried out by me under the supervision of Dr. Anuj Sharma, working as **Assistant Professor in School of Computer Applications at Lovely Professional University, Phagwara (Punjab)**, India. In keeping with general practice of reporting scientific observations, due acknowledgements have been made whenever work described here has been based on findings of another investigator. This work has not been submitted in part or full to any other University or Institute for the award of any degree.

Name of the scholar: Rakshit Khajuria

Registration No.: 12020422

Department/school: Computer Science and Engineering

Lovely Professional University,

Punjab, India

CERTIFICATE

This is to certify that the work reported in the Ph. D. thesis entitled "A Framework for Academic Performance Analysis of students in online Learning Using Machine Learning Approaches" submitted in fulfillment of the requirement for the award of degree of **Doctor of Philosophy (Ph.D.)** in the Computer Science and Engineering , is a research work carried out by Rakshit Khajuria , 12020422, is Bonafide record of his/her original work carried out under my supervision and that no part of thesis has been submitted for any other degree, diploma or equivalent course.

Dr. Anuj Sharma
Associate Professor
School of Computer Science and Engineering at
Lovely Professional University,
Phagwara (Punjab)

Dr. Ashok Sharma
Assistant Professor
Department of Computer Science and IT
Bhaderwah Campus,
University of Jammu

ACKNOWLEDGEMENT

Writing this thesis has been fascinating and extremely rewarding. I would like to thank a number of people who have contributed to the final result in many different ways.

To commence with, I pay my obeisance to GOD, the Almighty to have bestowed upon me good health, courage, inspiration, zeal and the light. After GOD, I express my sincere and deepest gratitude to my esteemed and reverent Supervisor and Co-supervisor Dr. Anuj Sharma and Dr. Ashok Sharma, who ploughed through several preliminary versions of my text, making critical suggestions and posing challenging questions. Their expertise, invaluable guidance, constant encouragement, affectionate attitude, understanding, patience and healthy criticism added considerably to my experience. Without their continual inspiration, it would have not been possible to complete this study.

Last but not the least, I would like to pay high regards and owe deepest gratitude to my parents Smt. Sushma Shamra and Sh. Jagdish Raj Khajuria for allowing me to realize my own potential, showing faith in me and giving me liberty to choose what I desired. I salute them for the selfless love, care, pain and sacrifice they did to shape my life. All the support they have provided me over the years was the greatest gift anyone has ever given me. Thank you for motivating me to keep reaching for excellence.

Also, I express my thanks to my brother Mr. Lakshit Khajuria, for their support and valuable prayers.

Finally, I convey my sincere thanks to my entire well – wishers and friends who have directly or indirectly helped me.

And above all, into the hands I lay all my works.

Dated: Rakshit Khajuria

ABSTRACT

The outbreak of the COVID-19 pandemic in 2019 catalysed a global transition toward online and virtual learning environments, resulting in a significant shift in the educational landscape. The rapid adoption of online learning was facilitated by advancements in internet infrastructure and educational technologies, transforming digital platforms from optional tools into indispensable components of modern education. However, this transition introduced new challenges, particularly in terms of catering to a highly diverse learner base with varying cognitive abilities, learning preferences, and backgrounds. The traditional "one-size-fits-all" instructional model has proven inadequate in this context. Moreover, while assessments such as multiple-choice questions (MCQs) remain integral to evaluating student progress, their effectiveness is often limited by issues of validity and reliability.

In response to these challenges, this study proposes a robust machine learning-based framework aimed at academic performance analysis and prediction in online learning environments. By leveraging supervised learning algorithms, including ensemble methods and deep learning models, the framework provides actionable insights into learner behaviour and performance trends. Among the models tested, Artificial Neural Networks (ANN) demonstrated superior accuracy and predictive capability, successfully capturing nonlinear patterns in student data. The framework not only aids in performance prediction but also lays the foundation for the integration of personalized and intelligent instructional strategies within e-learning systems. Despite its efficacy, the study is constrained by limitations such as the use of static datasets, the absence of real-time behavioural indicators, and potential demographic biases. Future enhancements will focus on incorporating dynamic interaction data, improving model interpretability through explainable AI, and developing adaptive content delivery mechanisms. This work contributes to the evolution of data-driven education systems by aligning machine learning approaches with the demands of personalized online learning.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ALGORITHMS

LIST OF SYMBOLS

Symbol	Meaning
+	Addition or Positive Sign
-	Subtraction or Negative Sign
×	Multiplication
÷	Division
=	Equals
≠	Not Equal To
≈	Approximately Equal To
>	Greater Than
<	Less Than
≥	Greater Than or Equal To
≤	Less Than or Equal To
%	Percent
°	Degree (Angle or Temperature)
√	Square Root
∑	Summation
∞	Infinity
∫	Integral
Π	Pi (3.14159...)
A	Alpha (Angle or Coefficient)
B	Beta (Coefficient or Angle)
Γ	Gamma (Angle or Coefficient)
Δ	Delta (Change or Difference)
Δ	Delta (Uppercase, Change or Difference)
M	Mu (Mean in Statistics or Micro in Measurements)
Σ	Sigma (Summation, Uppercase)
Σ	Sigma (Standard Deviation, Lowercase)
Ω	Omega (Ohm, Unit of Electrical Resistance)

Θ	Theta (Angle)
Λ	Lambda (Wavelength, Eigenvalues)
Ψ	Psi (Used in Quantum Mechanics)
∂	Partial Derivative
$\oplus$	Direct Sum
$\otimes$	Tensor Product
$\Rightarrow$	Implies
$\Leftrightarrow$	If and Only If
$\therefore$	Therefore
$\because$	Because
$\cap$	Intersection (Set Theory)
$\cup$	Union (Set Theory)
$\subseteq$	Subset
$\subset$	Proper Subset
$\supseteq$	Superset
$\supset$	Proper Superset
$\emptyset$	Empty Set
$\aleph$	Aleph (Cardinality of Infinite Sets, Set Theory)
$\forall$	For All
$\exists$	There Exists
$\in$	Element Of
$\notin$	Not an Element Of
$\equiv$	Identical To

LIST OF ABBREVIATIONS

Abbreviation		Full Form
ALA	:	Actionable Learning Analytics
ANFIS	:	Adaptive Neuro Fuzzy Inference System
ADASYN	:	Adaptive Synthetic Sampling
ABC	:	Affective, Behavioural and Cognitive
ANOVA	:	Analysis of Variance
ACO	:	Ant Colony Optimization
AUC	:	Area Under the ROC Curve
AFSA	:	Artificial Fish Swarm Algorithm
ANN	:	Artificial Neural Network
ARS	:	Audio Response System
BCO	:	Bee Colony Optimization
C-PSO	:	Chaotic Particle Swarm Optimization
CL	:	Collaborative Learning
CoI	:	Community of Inquiry
CRL	:	Computational Reinforcement Learning
CFA	:	Confirmatory Factor Analysis
CEC	:	Congress on Evolutionary Computation
CA	:	Content Analysis
CPS	:	Creative Problem-Solving
CHAT	:	Cultural-Historical Activity Theory
DT	:	Decision Tree
DL	:	Deep Learning
DM	:	Deep Motive
DNN	:	Deep Neural Network
DS	:	Deep Strategy
DiAL-e	:	Digital Artefacts for Learning engagement
DLA	:	Dispositional Learning Analytics

EDA	**:**	Educational Data Analytics
EDM	**:**	Educational Data Mining
ET	**:**	Educational Technology
EFA	**:**	Exploratory Factor Analysis
EGB	**:**	Extreme Gradient Boosting
FNR	**:**	False Negative Rate
FPR	**:**	False Positive Rate
FA	**:**	Firefly Algorithm
GA	**:**	Genetic Algorithm
GPA	**:**	Grade Point Average
GB	**:**	Gradient Boosting
G-ABC	**:**	Grouped Artificial Bee Colony
HWR	**:**	Handwritten Word Recognition
HFS	**:**	Hierarchical Feature Selection
ICT	**:**	Information and Communication Technology
ICC	**:**	Intra-Class Correlations
IRR	**:**	Intra-Rater Reliability
KMO	**:**	Kaiser-Meyer-Olkin
KSA	**:**	Knowledge, Skill and Attitude
LCTM	**:**	Learner-Centered Teaching Method
LA	**:**	Learning Analytics
LH	**:**	Learning Hypothesis
LMS	**:**	Learning Management System
LDA	**:**	Linear Discriminant Analysis
LR	**:**	Logistic Regression
LSTM	**:**	Long Short-Term Memory
ML	**:**	Machine Learning
MOOC	**:**	Massive Open Online Course
MLP	**:**	Multilayer Perceptron
MSE	**:**	Mean Square Error
NB	**:**	Naive Bayes

NLP	**:**	Natural Language Processing
NN	**:**	Neural Network
NP	**:**	Non-deterministic Polynomial
OULAD	**:**	Open University Learning Analytics Dataset
PSO	**:**	Particle Swarm Optimization
PV	**:**	Pedagogical Value
PreSS	**:**	Predict Student Success
PCA	**:**	Principal Component Analysis
PBL	**:**	Problem Based Learning
PoPS	**:**	Problem-Solving Perceptions
QDA	**:**	Quadratic Discriminant Analysis
RF	**:**	Random Forest
ROS	**:**	Random Oversampling
RNN	**:**	Recurrent Neural Network
RT	**:**	Regression Tree
RMSE	**:**	Root Mean Square Error
SRQ	**:**	Self-Report Questionnaire
SEM	**:**	Structural Equivalence Model
SOLO	**:**	Structure of Observed Learning Outcomes
SAL	**:**	Student Approaches to Learning
SIS	**:**	Student Information System
SPQ	**:**	Study Process Questionnaire
SVM	**:**	Support Vector Machine
SM	**:**	Surface Motive
SS	**:**	Surface Strategy
SMOTE	**:**	Synthetic Minority Oversampling Technique
TLP	**:**	Teaching-Learning Process
TNR	**:**	True Negative Rate
TPR	**:**	True Positive Rate

LIST OF PUBLICATIONS

A literature review has been done and paper has been presented in Scopus index conference (International Conference on Innovative Computing and Communication (ICICC) – A Flagship Conference).

A detailed survey regarding the usage of different ICT technology modes adopted by higher education institutions published in Scopus index journal. (Indonesian Journal of Electrical Engineering and Computer Science)

Performance analysis of frequent pattern mining algorithm on different real-life dataset published in Scopus index journal (Indonesian Journal of Electrical Engineering and Computer Science).

CHAPTER 1

INTRODUCTION

An introductory section in this chapter introduces the research work alongside its educational value in the arena of online learning using machine/deep learning approaches. This document outlines the significance and methodology of academic performance evaluation in digitized learning systems [1]. This research provides detailed guidance on the data mining process, along with Educational Data Mining (EDM) and Machine/Deep Learning, covering their individual methodologies, historical developments, educational motivations, and demonstrated positive operational outcomes [2].

1.1 BACKGROUND AND MOTIVATION

Educational institutions are intended to provide quality education and analyse student performance and help them improve [3]. Variable factors in current education have led to effective and efficient student performance monitoring so that the ability to predict student performance can provide information for helping students, teachers, managers, and the policymakers. In institutes of higher education or learning, student achievement is crucial. Education data mining methods that can increase the benefits and impacts of students, teachers and academic institutions can be used to effectively develop students' achievements and success through predicting students' performance [4]. The use of information mining approach in the online classroom has gained popularity in recent years. There are many ways to apply education information, such as Near Neighbourhood, Decision Tree, Naïve Bayes etc. [5]. The educational process culture is used to learn the information available in the field of education and to reveal confidential information from it. Various studies have been conducted that reflect the concerns of students' performance prediction. However, according to our investigation, no research has been carried out to predict the student's prospect of being a glorious student [6]. Student performance prediction is applied in numerous scenarios, such as adaptive content modification, customized textbook recommendations, early warning alerts for at-risk students, analyzing user behavior on websites, predicting user demand, developing fraud detection models, and preventing fraud. The evaluation of student's performance in educational settings reveals the extent of the efforts made by those settings to improve the

learning of underperforming or average students. The importance of adopting EDM models is that they use student historical data to forecast future performance that has not yet occurred. Several academics have been motivated by this concept to create classification models that forecast the as-yet-unknown labels of future cases. In order to categorise the educational level of student performances, a number of researchers and educational institutions began to become interested in the field or arena of student academic performance prediction. Although the educational sector employs a variety of methods to gather pertinent information about the characteristics of students who engage in the learning process, it is necessary to develop a model for student performance assessment to help students and faculty members advance their performance. The research offers several significant benefits, including the following:

- Enhanced learning experiences through diverse assessment activities that effectively complement the students' educational journey.
- Timely identification of students needing additional support, enabling prompt interventions to assist their academic progress.
- Reduction in student dropout rates, mitigating the negative impact on educational institutions.
- Improved academic outcomes and institutional productivity, which subsequently boosts student recruitment and retention rates.

Basically, in this research, focus is totally concentrated on the development of a framework for academic performance analysis of students in online learning using machine learning approaches for the prediction of Student Performance and their Engagement in Indian online learning platform.

1.1.1 Evolution of Online Education

Online education has markedly progressed in recent decades, evolving from a supplementary learning alternative into a fundamental educational structure. Its origin traces to the late 20th century, chiefly propelled by improvements in information technology and internet connectivity. Initially, online education concentrated on providing fundamental instructional information via static web pages, email, and rudimentary multimedia resources. In the early 2000s, the proliferation of the internet facilitated the swift expansion of online learning using interactive platforms termed Learning Management Systems (LMS), like Blackboard, Moodle, and

subsequently Canvas, which provided organized course administration and assessment capabilities. The introduction of Massive Open Online Courses (MOOCs) between 2008 and 2012 transformed education by providing free or low-cost courses available worldwide, significantly enhancing the reach and popularity of online education. The transition progressed as online education became more individualized and data-driven via educational analytics and the use of Artificial Intelligence (AI). Machine learning systems started predicting student success, optimizing material distribution, and furnishing adaptive learning experiences customized to particular learners' requirements. Currently, online education exemplifies a strong, adaptable, and inclusive educational model that provides quality instruction across many populations and geographical limitations. The constant development ensures enhanced accessibility, customization, and efficacy, propelled by continuous breakthroughs in machine learning, data analytics, and educational technology.

Initially, it is important to acknowledge that prior to being termed e-Learning, online training was referred to as distance learning. This delineated the array of instructional activities conducted inside an education-based project intended for scenarios lacking the simultaneous presence of instructors and students in the same location.

Four distinct generations can be identified based on the various communication techniques and the types of communication assistance utilized.

1st Generation: Discussing the inaugural generation of distance learning necessitates a temporal regression in Great Britain (in 1837), when famous scientist Isaac Pitman initiated the development of the first correspondence course in shorthand. Its popularity led to the establishment of the "Phonographic Correspondence Society" in 1843. In 1951, the inaugural Italian correspondence courses were established by the “Scuola Radio Elettra.”

2nd Generation: Fourth Generation Between 1960 and 1990, training courses were provided alongside regular post service, facilitated by new media technology such as acoustic or audio cassettes, videocassettes, floppy disks, and CD-ROMs. Advertising targets students through television.

3rd Generation: Transitioned from the 2nd to the 3rd generation in the 1990s. Indeed, it was only during those years that education began to thrive due to the help of the Internet.

In this era of computer proliferation, one of the most significant instruments facilitating online education is the emergence of specialized software for the administration and automated regulation of self-directed learning processes. May now begin to categorize these sorts of courses as e-Learning.

4th Generation: The proliferation of the Internet has initiated the emergence of a novel form of learning grounded on the comprehensive network. Due to the capabilities of internet technologies, it encompasses the benefits of remote education. The Fourth Generation develops concurrently with the advancement of the Web, fostering a more interactive and collaborative environment, referred to as Web 2.0.

One of the most important times in the history of e-learning training must be considered this one. For an efficient training path, distance learning is thought to be a dependable method. Users started using an unusual form of distance learning that could help with shortcomings. Web 2.0 represents a shift in the emphasis of education toward the student. A new model emerges as a result of the participants' enhanced involvement and communication!

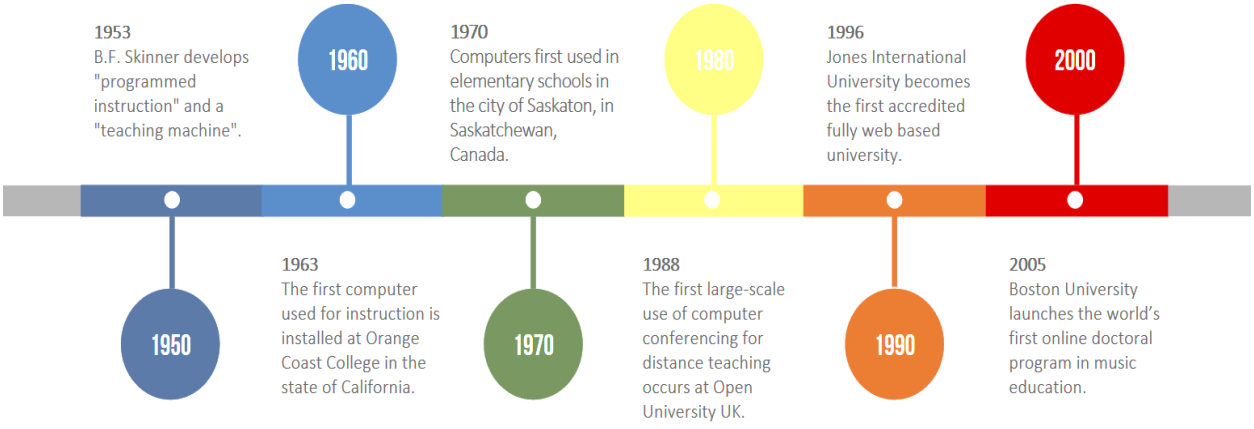

Figure 1.1: Evolution of Online Learning Src of Image is missing

1.1.2 Academic Performance Analysis

Academic performance analysis in online education is essential for understanding and improving student learning outcomes. As online learning environments have become increasingly prevalent, analysing and interpreting the data generated by these platforms has become critical for educators, institutions, and learners themselves. Performance analysis in online education involves

systematically evaluating students' achievements, interactions, and engagement within digital learning environments. This process typically incorporates assessing multiple data points such as assignment grades, quiz scores, course participation metrics, time spent on learning activities, and interaction patterns. By using these indicators, educators and administrators can identify trends and patterns, pinpoint at-risk students, and provide timely interventions. Several factors significantly impact academic performance in online education, including learner motivation, self-regulation, digital literacy, course content quality, instructor presence, and the effectiveness of the assessment methodologies employed. Machine learning approaches have recently gained prominence in performance analysis due to their ability to handle large datasets, uncover hidden insights, and predict future student outcomes with high accuracy. Academic achievement may be evaluated by many evaluations, tests, and other measuring methods. Academic performance can vary across students, since each individual possesses distinct degrees of achievement. Analysing the Indian educational system is a crucial endeavour in higher education. No established principle exists to evaluate student performance. Several universities evaluate student performance through co-curricular activities and internal assessments. The researchers are utilizing several student traits and features to analyse student performance. The researchers often consider internal assessments, external assessments, CGPA, final test results, and other extracurricular activities of the pupils. Indian institutes and universities utilize final test grades as the criterion for assessing student academic achievement. Furthermore, academic performance analysis benefits institutions by enabling them to reduce dropout rates, enhance student satisfaction, and continuously improve their online courses. For students, accurate performance insights can foster personalized learning experiences, tailored support, and increased educational success [4]. To analysis the student academic performance, the given process in Figure 1.2 is a better way to adopt.

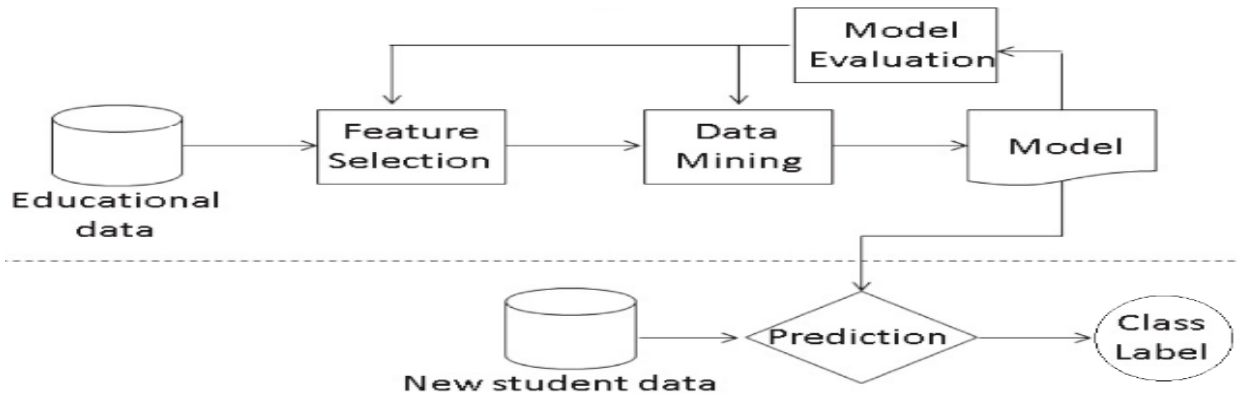

Figure 1.2: Academic Performance Analysis Model

1.1.2.1 Significance of Performance Analysis

Performance analysis is crucial in educational contexts, particularly in online learning, since it offers insights on student learning processes, results, and the efficacy of instructional methods. Through the methodical analysis of student performance data, educators and institutions may make educated judgments, therefore strengthening instructional practices and improving overall educational results. Initially, performance analysis enables instructors to pinpoint children needing more assistance, allowing prompt interventions to tackle specific learning difficulties. The early identification of at-risk students via predictive analytics facilitates the implementation of preventive interventions, therefore decreasing dropout rates and enhancing student retention. Furthermore, performance analysis facilitates the customization of learning experiences. By acknowledging varied student learning patterns, educational materials and evaluation techniques may be customized to address the specific requirements, preferences, and skills of individual learners. This customization markedly elevates student motivation, engagement, and satisfaction, hence improving academic achievement. Thirdly, the examination of student performance data enables the ongoing enhancement of teaching methodologies and educational materials. Performance metrics provide educators with insights that facilitate the refinement of curriculum designs, course material, and assessment methods, therefore enhancing the effectiveness of educational delivery and aligning it with learners' expectations. Furthermore, from an institutional standpoint, performance analysis has strategic significance. It facilitates the maintenance of elevated educational standards, fulfills accrediting criteria, and bolsters school reputation by boosting student results. The importance of performance analysis resides in its capacity to facilitate informed decision-making, enhance the quality of teaching and learning, support student achievement, and encourage ongoing educational advancement in online learning contexts.

1.1.2.2 Factors Influencing Academic Performance

Academic performance in online learning environments is influenced by a variety of interconnected factors. Understanding these factors is crucial to enhancing student achievement and implementing effective educational interventions. These factors can broadly be categorized as follows:

1. Learner Characteristics

- Students' intrinsic and extrinsic motivation significantly impacts their engagement, persistence, and overall academic outcomes.
- Students' ability to manage their own learning processes, set realistic goals, and adhere to study schedules profoundly affects performance in online learning.
- Prior knowledge, academic preparedness, and technical proficiency influence students' adaptability to online environments.

2. Instructional Design and Course Quality

- Clearly structured courses with organized, engaging content and meaningful assessments positively influence learner performance.
- Diverse, timely, and constructive assessments aligned with learning objectives can significantly enhance performance and engagement.
- Regular and meaningful interactions with instructors through feedback, announcements, and discussions improve students' sense of connection and academic achievement.

3. Technological Factors

- User-friendly LMS that offer intuitive navigation and reliable functionality greatly influence learner satisfaction and engagement.
- Prompt technical support and reliable access to learning resources reduce barriers, ensuring smooth and continuous learning experiences.

4. Social and Collaborative Factors

- Opportunities for peer-to-peer interactions and collaborative learning contribute to higher levels of understanding, satisfaction, and academic performance.
- A strong sense of online community and regular communication enhance motivation, persistence, and achievement.

5. Environmental and Institutional Factors

- Availability of institutional resources, such as online libraries, academic advising, and student support services, significantly impacts students' academic success.
- A conducive home or study environment, with minimal distractions, sufficient resources, and supportive family or community members, is crucial for effective online learning.

1.1.2.3 Traditional vs. Modern Approaches to Performance Analysis

Analysing academic performance has evolved significantly, transitioning from traditional methods toward more sophisticated modern approaches. Each approach has its strengths and limitations, influenced by technological advancements, data availability, and educational practices.

Traditional Approaches: Traditional performance analysis typically involves manual and qualitative assessment methods such as:

- **Examinations and Grades:** Evaluating student performance primarily through tests, quizzes, assignments, and final grades.
- **Instructor Observations:** Instructors monitor and document students' participation, attendance, and class behaviours manually.
- **Student Self-Assessment and Surveys:** Collecting feedback through questionnaires, student reflections, or end-of-course evaluations.
- **Limitations:** These methods are often labour-intensive, subjective, and limited in scope. They usually fail to capture real-time student engagement and don't efficiently utilize data for predictive insights.

Modern Approaches: In contrast, modern approaches harness technological advancements, leveraging data mining, machine learning, and advanced analytics:

- **Educational Data Mining (EDM):** Extracting meaningful patterns from large educational datasets, facilitating deeper insights into student behaviour.
- **Learning Analytics:** Utilizing real-time analytics tools embedded within online platforms (e.g., LMS) to continuously track and analyse student activities and performance.

- **Predictive Modeling and Machine Learning:** Employing machine/deep learning processes to predict student performance in online learning, identify at-risk students, and implement early intervention strategies.
- **Adaptive Learning Technologies:** Personalizing the learning experience based on real-time student performance data to tailor content delivery dynamically.
- **Benefits:** Modern methods offer objectivity, scalability, timely interventions, and more accurate predictions. They enable a proactive rather than reactive approach to enhancing educational outcomes.

While traditional performance analysis methods have historically provided foundational insights, modern approaches offer greater precision, scalability, and timely interventions, making them particularly suited for contemporary online education contexts. The tabular comparison is given in Table 1.1.

Table 1.1 Comparison of Traditional vs. Modern Approaches

Aspect	Traditional Approaches	Modern Approaches
1 Data Handling	Limited, manual	Large-scale, automated
2 Analysis Method	Mostly qualitative	Quantitative and qualitative
3 Predictive Capability	Minimal to none	High predictive accuracy
4 Intervention Timeliness	Reactive and delayed	Proactive and timely
5 Personalization	Generic approach	Highly personalized
6 Resource Requirement	High labour, less efficient	Less labour-intensive, more efficient

1.1.2.4 Impact of Machine Learning in Educational Analysis

Machine learning has significantly transformed educational analysis by introducing advanced analytical capabilities and predictive insights. By leveraging large datasets generated by online learning platforms, machine learning provides robust tools to enhance educational effectiveness, personalize learning experiences, and facilitate data-driven decision-making. The major impacts of machine learning in educational analysis include:

Early Identification and Intervention: Morden machine/deep learning algorithms can accurately predict students' academic performance in online learning mechanism and identify learners at risk of underperforming or dropping out. Early detection enables proactive interventions, personalized support, and timely resources allocation, significantly improving student outcomes.

Enhanced Personalization of Learning: By analysing individual learning patterns, preferences, and engagement data, machine learning facilitates highly personalized learning experiences. Adaptive systems dynamically adjust content, assessments, and feedback to meet each student's unique educational needs.

Improved Institutional Decision-Making: Institutions benefit from machine learning through insightful data analytics, enabling evidence-based decisions on curriculum design, resource allocation, faculty support, and student services. Predictive modeling guides strategic planning, improves retention rates, and enhances overall educational quality.

Automation and Efficiency: Machine learning streamlines data processing and analysis, reducing manual workloads for educators and administrators. Automated grading, performance reporting, and content recommendation systems enhance operational efficiency, allowing educators to focus more on instructional quality and student interaction.

Advanced Performance Analytics: Machine learning algorithms reveal hidden patterns and complex relationships within educational data, providing deeper insights into student behaviours, learning outcomes, and effectiveness of instructional methods. This detailed understanding helps educators refine teaching approaches and curricula.

Bridging the Semantic Gap: By employing Natural Language Processing (NLP) and sentiment analysis techniques, machine learning bridges the gap between qualitative learner feedback and quantitative performance indicators, providing holistic insights into students' educational experiences and emotional states.

1.2 INTRODUCTION TO ONLINE LEARNING ENVIRONMENTS

Online learning environments refer to digital platforms and contexts in which teaching and learning interactions occur through the internet and associated technologies. These environments

provide flexible, interactive, and accessible learning experiences beyond traditional classroom settings, offering students the ability to engage in educational activities remotely, at their own pace and convenience. Online learning environments typically include LMS, MOOCs, virtual classrooms, and collaborative platforms. They support various educational activities such as course content delivery, assessments, student collaboration, and instructor interactions. Key components of effective online learning environments include:

- **Content Delivery Systems:** Platforms like Moodle, Blackboard, Canvas, and Google Classroom, facilitating structured delivery of course materials.
- **Interactive Communication Tools:** Discussion forums, video conferencing (e.g., Zoom, Microsoft Teams), and chat applications promoting instructor-student and student-student interaction.
- **Assessment and Feedback Systems:** Tools for conducting quizzes, assignments, exams, and providing timely feedback to learners.
- **Learning Analytics:** Features that enable monitoring and analysis of student engagement, performance metrics, and personalized educational recommendations.

The rapid growth of online learning has been fuelled by technological advancements, increased internet accessibility, and shifting educational paradigms driven by global circumstances such as the COVID-19 pandemic. Today, online learning environments play a crucial role in education by providing inclusive, scalable, and personalized educational experiences, significantly shaping the future landscape of education.

1.2.1 Definitions and Types of Online-based Learning

Online-based learning, also known as e-learning (o-learning), refers to educational experiences and instructional methods delivered entirely or partially via digital technologies over the internet. It enables learners to access educational content remotely, allowing greater flexibility, interactivity, and personalized learning experiences compared to traditional classroom-based learning.

1.2.1.1 Definitions of Online Learning

- **Allen and Seaman (2007)** defined online learning as instruction delivered primarily via the internet, with no required face-to-face interaction between instructors and students.
- **Clark and Mayer (2016)** described it as learning experiences designed and delivered through electronic media such as the internet, interactive multimedia, and virtual platforms.
- In broader terms, online learning includes all educational practices and activities that occur through web-based technologies, regardless of synchronous or asynchronous interaction methods.

1.2.1.2 Types of Online Learning

Online learning can be categorized into several distinct types, based on the nature and structure of the interactions involved:

- **Asynchronous Online Learning:** Students access learning materials and engage with the course content at their convenience, without real-time interactions. This approach typically includes pre-recorded lectures, discussion forums, emails, assignments, and self-paced activities.
- **Synchronous Online Learning:** Real-time interactions occur between instructors and learners, often through video conferencing platforms such as Zoom or Microsoft Teams. This type of learning includes live lectures, webinars, online discussions, virtual classrooms, and immediate feedback mechanisms.
- **Blended (Hybrid) Learning:** Combines both face-to-face and online instructional methods. Students experience a portion of the learning in traditional classroom settings, supplemented significantly by online materials and interactive activities.
- **MOOCs:** Free or affordable courses available to a global audience, delivered by educational institutions or organizations via platforms such as Coursera, edX, or Udacity. MOOCs are characterized by open access, large-scale participation, and flexible learning experiences.

- **Mobile Learning (m-Learning):** Educational experiences specifically designed for mobile devices such as smartphones and tablets, emphasizing accessibility, portability, and learner convenience.
- **Collaborative Online Learning:** Involves structured interactions among students, typically in virtual groups, through forums, chats, shared documents, and virtual workspaces, fostering a sense of community and cooperative learning.

1.2.2 Platforms and Tools for Online Learning

Online learning is supported and facilitated through a variety of platforms and technological tools, each serving specific purposes and enhancing various aspects of digital education. These platforms enable educators to create, deliver, and manage educational content, while providing students with accessible, engaging, and personalized learning experiences.

1. Learning Management Systems (LMS)

LMS platforms serve as centralized environments for managing online learning processes. They allow instructors to upload and organize educational resources, manage assessments, track student progress, and facilitate communication.

- **Examples:**
    - **Moodle:** Open-source, highly customizable LMS widely used globally.
    - **Blackboard:** Robust commercial LMS known for extensive analytics and reporting features.
    - **Canvas:** User-friendly, cloud-based LMS with intuitive interfaces and extensive integrations.
    - **Google Classroom:** Easy-to-use platform integrated with Google's productivity tools.

2. Massive Open Online Courses (MOOCs) Platforms

These platforms provide large-scale, open-access online courses to learners worldwide, often in collaboration with prestigious universities and industry partners.

- **Examples:**
    - **Coursera:** Offers courses, specializations, and degrees from leading universities globally.
    - **edX:** Founded by Harvard and MIT, provides a range of courses, micro-masters, and professional certifications.
    - **Udacity:** Specializes in technical and vocational skills, particularly in programming, AI, and data science.

3. Video Conferencing and Virtual Classroom Tools

Facilitating real-time synchronous learning sessions, these tools support live lectures, interactive discussions, and collaborative activities.

- **Examples:**
    - **Zoom:** Popular for webinars, live classes, breakout rooms, and interactive learning features.
    - **Microsoft Teams:** Comprehensive platform integrating collaboration, file sharing, and virtual classroom functions.
    - **Google Meet:** User-friendly video conferencing, integrated seamlessly with Google Workspace tools.

4. Interactive Collaboration and Communication Tools

Designed to promote student interaction, teamwork, and community building in online settings.

- **Examples:**
    - **Slack:** Real-time messaging, collaboration, and seamless integration with educational tools.
    - **Padlet:** Interactive digital board facilitating content sharing, brainstorming, and collaboration.

- **Flipgrid:** Enables students to share short video responses to discussion prompts, enhancing student engagement and voice.

5. Assessment and Evaluation Tools

Platforms and tools that streamline assessments, quizzes, exams, and immediate feedback mechanisms.

- **Examples:**
  - **Kahoot:** Interactive quiz-based tool increasing student engagement through gamification.
  - **Quizlet:** Flashcards, quizzes, and learning games aiding student revision and self-assessment.
  - **Turnitin:** Provides plagiarism detection, originality reports, and detailed feedback on assignments.

6. Learning Analytics and Data Visualization Tools

Tools that offer insights and analytics into student performance, enabling informed instructional decisions.

- **Examples:**
  - **Tableau and Power BI:** Tools for visualizing student data and performance analytics.
  - **Google Analytics:** Tracks student interaction and engagement with online educational content.
  - **Built-in LMS Analytics:** LMS platforms often include integrated analytics tools for monitoring learner progress and behaviour.

1.2.3 Advantages and Challenges of Online Learning

Online learning has fundamentally reshaped education, offering numerous advantages while simultaneously presenting distinct challenges. Understanding these benefits and limitations is critical to optimizing online educational practices.

1.2.3.1 Advantages of Online Learning

- **Accessibility and Flexibility:** Online learning provides access to education irrespective of geographic location, allowing learners to study at their own pace and schedule, thus removing traditional constraints associated with location and time.
- **Personalized Learning Experience:** Online or e-learning environments provide individualized educational experiences based on the needs, preferences, and learning styles of each learner by utilizing cutting-edge technologies like analytics and adaptive learning systems.
- **Cost-Effectiveness:** Online education can reduce costs for both institutions and students by eliminating physical infrastructure requirements, commuting expenses, and other related expenditures.
- **Scalability and Reach:** Online platforms can easily scale educational offerings to large and diverse groups of learners, extending the reach of quality education globally.
- **Enhanced Learning Resources:** Rich multimedia resources, interactive simulations, and diverse instructional content available online enhance the learning experience, making education more engaging and effective.
- **Immediate Feedback and Analytics:** Digital assessments and real-time feedback help learners quickly understand their progress, facilitating timely interventions and continuous improvement.

1.2.3.2 Challenges of Online Learning

- **Technological Barriers:** Limited or inconsistent internet connectivity, lack of suitable hardware or software, and insufficient technical skills among students or instructors can significantly hinder the effectiveness of online learning.

- **Lack of Personal Interaction:** Reduced face-to-face interactions and diminished physical presence may negatively affect student engagement, motivation, and the sense of community among learners.
- **Self-Regulation and Motivation Issues:** Online learning requires strong self-discipline, motivation, and effective time-management skills. Learners lacking these traits may struggle, leading to reduced academic performance and higher dropout rates.
- **Assessment and Academic Integrity Concerns:** Maintaining the reliability and integrity of online assessments poses challenges due to potential academic dishonesty, plagiarism, or cheating.
- **Instructor Training and Preparedness:** Many instructors may lack the necessary skills or training to effectively design, deliver, and manage online courses, reducing the quality of the educational experience.
- **Equity and Accessibility Issues:** Digital divides can exacerbate inequalities, particularly among students from economically disadvantaged or rural backgrounds who lack reliable access to necessary technology and infrastructure.

1.3 IMPORTANCE OF ACADEMIC PERFORMANCE ANALYSIS

Academic performance analysis plays a critical role in educational settings, particularly within online learning environments. Its importance arises from its ability to guide decision-making, enhance educational quality, and directly improve student outcomes. Key reasons underscoring the significance of academic performance analysis include:

1. Early Identification of At-Risk Students: Analysing student performance data enables educators and institutions to identify students at risk of academic failure or dropout at an early stage. Prompt identification allows for timely intervention, targeted support, and improved student retention rates.

2. Enhanced Learning Outcomes: Through careful analysis of academic performance, educators can tailor instruction and feedback to individual learners, addressing specific strengths and

weaknesses. This personalization results in more effective learning experiences and improved educational outcomes.

3. Improved Instructional Strategies: Performance analysis provides valuable insights into the effectiveness of teaching methodologies and course design. Educators can use these insights to refine their instructional approaches, optimize course content, and enhance assessment strategies.

4. Data-Driven Decision Making: Analysing academic performance data supports informed decision-making at institutional levels, enabling strategic planning related to resource allocation, curriculum development, faculty training, and policy formulation.

5. Enhanced Student Motivation and Engagement: Continuous performance feedback and personalized recommendations based on performance analysis motivate learners by clearly demonstrating their progress and areas for improvement. This increases learner engagement and active participation in the educational process.

6. Institutional Effectiveness and Accountability: Regular academic performance analysis contributes to institutional effectiveness by highlighting areas needing improvement, helping institutions meet accreditation requirements, maintain academic standards, and ensure accountability to stakeholders.

1.4 LEARNING MANAGEMENT SYSTEM (LMS)

A LMS is a comprehensive digital platform designed to deliver, manage, and assess educational content and facilitate interactions between educators and learners. It serves as a central hub for organizing and streamlining various aspects of online and blended learning environments, significantly enhancing educational delivery, assessment, and administration.

Definition and Overview: An LMS is defined as a web-based or cloud-based software application enabling educators to create and deliver instructional materials, track learner progress, conduct assessments, and manage educational resources systematically. LMS platforms also offer capabilities for real-time communication, collaboration, and data analytics, thereby supporting personalized learning experiences.

1.4.1 Key Functionalities of LMS

- **Course Management:** Facilitates creation, distribution, and management of course content such as lecture notes, multimedia materials, and assignments.
- **Assessment and Feedback:** Provides tools for developing quizzes, exams, surveys, and assignments, coupled with immediate feedback and analytical capabilities.
- **Communication and Collaboration:** Enables real-time interaction through discussion forums, announcements, messaging systems, and integrated video conferencing.
- **Tracking and Reporting:** Monitors student activities, engagement levels, performance data, and generates detailed reports for instructors and administrators.
- **User Management:** Manages enrolment, user profiles, roles, permissions, and access controls, ensuring secure and organized delivery of educational content.

1.4.2 Types of LMS

LMS platforms can be categorized as follows:

- **Open-Source LMS:** Cost-effective and customizable platforms such as Moodle, Sakai, and Open edX, offering community-driven development and flexibility.
- **Proprietary LMS:** Commercially developed platforms like Blackboard, Canvas, and Desire2Learn, typically providing robust support services and advanced analytics.
- **Cloud-based LMS:** Hosted platforms, such as Canvas and Google Classroom, providing easy accessibility, scalability, and maintenance-free operations.

1.4.3 Advantages of Using LMS

- Streamlines content delivery, making educational materials easily accessible.
- Enhances student-teacher and peer-to-peer interactions.
- Supports personalization and adaptive learning through data-driven insights.
- Improves administrative efficiency with automated processes and analytics.

- Facilitates scalable and consistent education delivery across diverse learner populations.

1.4.4 Challenges Associated with LMS

- Technical complexities and learning curves for instructors and learners.
- Issues related to data privacy, security, and user authentication.
- Limited face-to-face interaction impacting student engagement.
- Need for continuous technological and pedagogical training.

1.4.5 Role of LMS in Online Learning

LMS play a central and indispensable role in the successful implementation and management of online learning environments. By integrating diverse instructional tools, resources, and analytics, LMS platforms significantly enhance the quality, efficiency, and effectiveness of digital education. Key roles and contributions of LMS in online learning include:

1.4.5.1 Centralized Delivery of Educational Content

LMS platforms serve as centralized hubs for course materials, including text, multimedia content, assignments, quizzes, and supplementary resources. This unified access simplifies learning processes, ensuring students can easily locate and interact with course materials, regardless of geographic or temporal barriers.

1.4.5.2 Facilitation of Communication and Collaboration

LMS supports and promotes seamless interaction and collaboration among learners and instructors through integrated tools such as discussion forums, chat rooms, email, and video conferencing. These tools create interactive learning communities that improve student engagement, motivation, and learning outcomes.

1.4.5.3 Personalized and Adaptive Learning

Through integrated analytics and adaptive technologies, LMS platforms enable students or educators to customize the learning process and experiences to individual student needs. Real-time

insights into student performance and behaviours allow personalized recommendations, targeted feedback, and adaptive learning paths that foster deeper understanding and academic success.

1.4.5.4 Efficient Assessment and Feedback

By offering resources for making exams, quizzes, assignments, and surveys, LMS platforms simplify assessments. Automated grading and real-time feedback features help students quickly identify their strengths and weaknesses, enhancing learning outcomes and reducing instructors' workload.

1.4.5.5 Tracking and Analytics

LMS provides comprehensive analytics capabilities that help educators and institutions monitor student engagement, progress, and performance. Detailed insights into student behaviors facilitate early intervention, informed decision-making, and continuous improvement of instructional methods and curriculum design.

1.4.5.6 Administrative Efficiency and Scalability

By automating many administrative tasks, including student enrolment, attendance tracking, grade management, and reporting, LMS platforms significantly enhance institutional efficiency. The scalable nature of LMS systems allows educational institutions to effectively manage large cohorts of students, maintaining educational quality and consistency across diverse courses and programs.

1.4.5.7 Enhancing Accessibility and Inclusion

LMS facilitates greater educational accessibility, providing opportunities for students from varied geographical, socio-economic, and demographic backgrounds to participate in high-quality learning experiences. Supportive features such as multilingual interfaces, adaptive accessibility tools, and mobile compatibility further promote inclusive learning environments.

1.5 MACHINE LEARNING AND EDUCATIONAL DATA MINING (EDM)

1.5.1 Machine Learning

It is considered as an application of AI that enables or make the machine/system to learn automatically and improve the performances based on the experience without clear programming

[25]. Machine learning boosts academic performance analysis of students in online learning through its ability to process extensive LMS data together with assessment and interaction log records [26]. The combination of ML algorithms which include decision trees (DT) and support vector machines (SVM) and random forests (RF) and neural networks (NN) successfully detects patterns while making predictions about student performance along with revealing undisclosed student behavioural information that standard analysis methods cannot detect. Analysis of various student data such as engagement statistics combined with submission duration records and test scores and discussion forum involvement enables predictive modeling for early detection of at-risk students. ML models deliver customized learning experiences by modifying educational materials together with lesson speed based on each student's particular requirements thus raising both interest and academic achievement. Every step of continuous monitoring and performance analysis enables educators along with institutions to use data-based decisions for transforming educational strategies and bettering learning results in digital classrooms.

1.5.2 Some Machine Learning Methods

The categorization of the machine learning algorithm is into supervised as well as unsupervised learning. The explanation for the same is provided below [22]:

Supervised algorithm: Supervised machine learning is the type of learning in which the entire training data act as the test data. It is done in order to take the highest efficiency from any system. It is performed in student learning method, intrusion detection system etc [23].

Unsupervised algorithms: It is used in situations where the training data is not annotated or classified. In order to characterize the unnoticed framework of unlabelled data, unsupervised learning studies how systems infer functions. The system will find the data and be able to draw conclusions from the information set in order while clarifying the underlying structure of the unregistered data, even if it does not find the proper output [24].

Semi-supervised algorithms: It is a combination of supervised as well as unsupervised method. In this method, 70% data is usually taken as the training data and the rest of the data is taken as the test data. This method is usually helpful in forecasting [25].

Reinforcement algorithms: An agent learns optimal actions through reinforcement learning by experiencing its environment directly through its performed actions which results in received feedback as either rewarding or punishing signals. The essential characteristics of RL consist of experimental discovery paired with postponed outcomes that use a reward system for signalling. Autonomous agents use this method to create policies based on specific situations which result in maximum cumulative performance metrics during operation. The reinforcement cue delivered through a scalar reward signal leads agents to adopt better actions through successive updates [31].

Machine learning technology demonstrates exceptional performance in handling big data through scalable handling which produces quick and precise patterns together with risk detection abilities. Reaching the best model outcomes demands important computational power and lengthy training times [32]. Through integration with artificial intelligence and cognitive computing frameworks the system receives improved abilities to process effectively complex, high-dimensional data sets.

1.5.3 Need of Machine Learning in Academic Performance Analysis

Machine learning is a subset of AI designed to program a computer to identify patterns in the data to provide information for algorithms that can make data-driven predictions or decisions [33].

i. Learning analytics and machine learning facilitate comprehensive data mining, enabling educators to move beyond traditional grade-book dependency. By consolidating student performance data in a centralized system, teachers gain valuable insights that not only reduce administrative workload but also support data-driven lesson refinement [34].
ii. A significant advantage of educational technologies lies in their predictive capabilities. Through continuous analysis of individual learning patterns, machine learning algorithms can identify student strengths and recommend personalized interventions, such as adaptive practice or additional assessments, to enhance academic outcomes [35].
iii. Machine learning supports dynamic assessment methodologies like “stop-and-test,” which go beyond static evaluations. These intelligent systems provide real-time feedback loops for educators, learners, and guardians, facilitating a deeper

understanding of knowledge acquisition, learning needs, and progression towards academic goals [36][37].

iv. Machine learning models reduce subjective bias and cognitive load associated with manual grading. Advanced AI-driven platforms, such as Grammarly and Turnitin, have been increasingly employed for automated assessment of written content, thereby promoting consistency and fairness in student evaluations [38].

v. Leveraging machine learning frameworks allows for intelligent organization and delivery of educational content. These systems enhance content accessibility and optimize curriculum structure, enabling improved cognitive mapping and comprehension for learners [39].

vi. Predictive analytics within machine learning models are instrumental in identifying at-risk students. Early detection mechanisms allow institutions to proactively engage and support such students through targeted interventions, thereby improving retention and overall academic success rates [40].

1.5.4 Overview of Educational Data Mining

Data mining is technically defined as the systematic process of identifying, extracting, and analysing meaningful patterns and relationships from large and complex datasets using specialized algorithms and software tools. It enables the transformation of raw, unstructured data into actionable knowledge, facilitating informed decision-making. This technique finds wide applicability across domains such as scientific research, healthcare, and business intelligence [41]. In commercial contexts, data mining allows organizations to gain deeper insights into customer behaviour, optimize operational strategies, and develop data-driven approaches for enhanced resource utilization and performance outcomes. As illustrated in Figure 1.3, the data mining process typically comprises five key phases, encompassing data selection, preprocessing, transformation, mining, and interpretation.

It encompasses systematic data aggregation, data warehousing, and high-performance computational processing. To facilitate data segmentation and predictive analytics, it employs advanced mathematical models and composite algorithmic frameworks designed for probabilistic estimation of future outcomes [42].

Figure 1.3: Varied phases of data mining Src is missing

The mention phases are also termed as Knowledge Discovery Data (KDD). The foremost features are as follows:

• Automatic data pattern predictions based on trend and behaviour analysis.

• Prophecy on the basis of outcomes.

• Development of decision-oriented information.

• For the examination, the focus is given to huge datasets and databases in this work [60].

Educational Data Mining (EDM) is an interdisciplinary arena in modern world that applies data mining, machine learning, and statistical techniques to analyse large-scale educational data and extract meaningful patterns to improve teaching, learning, and administrative processes. With the advent of digital learning platforms, such as LMS and MOOCs, vast amounts of data are generated, including student interactions, performance records, and behavioural patterns. EDM leverages this data to enhance educational outcomes by identifying trends, predicting student performance, personalizing learning paths, and providing actionable insights to educators and institutions. EDM

encompasses various tasks such as classification, clustering, regression, association rule mining, and sequential pattern mining. Classification techniques help predict student success or failure, while clustering groups learners with similar characteristics to offer targeted interventions. Regression models estimate student performance trends over time, and association rule mining identifies relationships between learning behaviours and outcomes. Sequential pattern mining detects patterns in student activities to refine course design and instructional strategies. The key applications of EDM include early identification of at-risk students, personalized learning recommendations, curriculum optimization, and improving instructor effectiveness. By utilizing predictive models, EDM helps educators detect students who may require additional support, enabling timely interventions and reducing dropout rates. Moreover, adaptive learning systems powered by EDM dynamically adjust content and assessments based on individual learner progress, enhancing the overall learning experience.

1.5.5 Role of Machine/Deep Learning in EDM

Machine/Deep Learning plays a crucial role in EDM by enabling the analysis of large-scale educational data to uncover hidden patterns, predict outcomes, and provide actionable insights to improve the learning experience. With the increasing adoption of online learning platforms and LMS, vast amounts of data, including student interaction logs, assessment results, and course participation records, are generated. ML algorithms effectively process and analyse this data, offering valuable predictions and recommendations that enhance teaching and learning processes.

1. Predicting Student Performance: ML models such as decision trees, SVM, random forests, and neural networks are widely used to predict student outcomes based on historical data. These models can assess factors such as assignment submissions, quiz scores, and engagement levels to identify students at risk of academic failure, enabling timely interventions and personalized support.

2. Personalized Learning and Content Recommendation: ML facilitates adaptive learning by tailoring educational content to match individual student needs. Recommender systems powered by ML analyse student preferences, learning styles, and performance to suggest relevant learning materials, ensuring a customized learning path that enhances student engagement and understanding.

3. Early Identification of At-Risk Students: Through pattern recognition and predictive modeling, ML helps detect at-risk students who may struggle with course content or disengage from the learning process. By identifying these students early, educators can implement targeted interventions to provide additional support and prevent dropouts.

4. Clustering and Grouping of Learners: ML techniques, such as k-means clustering and hierarchical clustering, group students based on similar learning behaviours, performance patterns, or engagement levels. This segmentation enables educators to develop tailored instructional strategies and improve group-based learning experiences.

5. Sentiment Analysis and Feedback Interpretation: Natural Language Processing (NLP), a subset of ML, helps analyse textual data such as student feedback, forum discussions, and reviews to gauge sentiment, detect areas of dissatisfaction, and refine course content accordingly. Sentiment analysis offers insights into student emotions and perceptions, allowing educators to improve course design and delivery.

6. Automated Assessment and Grading: ML models enhance the efficiency of assessment processes by automating the grading of assignments, quizzes, and written content. Techniques such as supervised learning and neural networks facilitate the evaluation of open-ended responses, reducing instructor workload and ensuring consistency in grading.

7. Optimization of Learning Paths: Reinforcement learning techniques can optimize learning paths by dynamically adjusting content delivery based on real-time student responses. These models improve learning efficiency by identifying the most effective instructional sequence for each student.

Machine learning significantly enhances the capabilities of EDM by providing predictive, prescriptive, and diagnostic analytics that enable personalized learning, early interventions, and continuous improvements in educational practices. As online education continues to grow, the integration of ML into EDM will play an increasingly vital role in shaping future learning environments.

1.5.6 Recent Advances in EDM using Machine Learning

In recent years, EDM has witnessed significant advancements through the integration of machine/deep learning techniques, enabling more accurate predictions, personalized learning, and enhanced decision-making processes. These advancements have been fuelled by the growing availability of large-scale educational data generated by online learning platforms, LMS, and MOOCs. Below are some of the notable recent advances in EDM using ML:

1. Deep Learning for Student Performance Prediction: Deep learning models, particularly Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), have significantly improved the prediction of student performance. These models can capture intricate patterns in sequential data, such as student interaction logs, and predict outcomes such as course completion, dropout rates, and academic success with high accuracy.

2. Adaptive Learning Systems Using Reinforcement Learning: Reinforcement learning is increasingly being applied to develop adaptive learning environments that dynamically adjust the content and difficulty level based on the learner's progress. RL-based systems continuously refine their instructional strategies to optimize the learning experience and maximize student retention.

3. Automated Essay Scoring and Sentiment Analysis: Natural Language Processing (NLP), an advanced subset of ML, has been leveraged to develop automated essay scoring systems and perform sentiment analysis on student feedback. NLP models, such as Bidirectional Encoder Representations from Transformers (BERT) and Long Short-Term Memory (LSTM), can evaluate written responses, provide constructive feedback, and analyse student sentiments to improve course content.

4. Early Warning Systems for At-Risk Students: Recent developments have enhanced early warning systems that identify students at risk of disengagement or academic failure. ML models, including decision trees, random forests, and gradient boosting, now analyse a combination of student interaction data, grades, and behavioural patterns to predict the likelihood of dropout and trigger timely interventions.

5. Recommender Systems for Personalized Learning Paths: Modern ML-based recommender systems have revolutionized content delivery by suggesting personalized learning paths. These

systems leverage collaborative filtering, matrix factorization, and deep learning techniques to recommend relevant courses, study materials, and resources that align with a student's learning preferences and goals.

6. Clustering and Grouping for Learning Behaviour Analysis: Unsupervised learning algorithms, such as k-means clustering and DBSCAN, are used to group students with similar learning behaviours and performance characteristics. This segmentation helps educators develop targeted instructional strategies and identify common challenges faced by different student groups.

7. Feature Engineering for Enhanced Model Accuracy: Recent advancements in feature engineering techniques have enabled the extraction of meaningful features from educational datasets, improving the predictive accuracy of ML models. Techniques such as Principal Component Analysis (PCA) and feature selection algorithms are used to reduce dimensionality and eliminate irrelevant attributes.

8. Transfer Learning for Knowledge Adaptation Across Datasets: Transfer learning, where a pre-trained model is fine-tuned for a related task, has been successfully applied to adapt models across different educational datasets. This approach reduces the need for extensive labelled data and enhances model performance in new educational settings.

9. Intelligent Tutoring Systems (ITS) with ML Integration: Recent advances have led to the development of Intelligent Tutoring Systems (ITS) that provide personalized feedback, identify knowledge gaps, and guide learners toward mastery of concepts. ML algorithms power these systems to adjust instructional approaches and deliver tailored recommendations.

1.6 PROBLEM STATEMENT

The academic performance could be computed by carrying out varied assessments, examinations, and some other measurement forms. Though, academic performances can differ from student to student as every student has different performance levels. For the consideration, analysing Indian educational system is an important task in higher education. There is no such fixed principle to analyse the student performance. Few of the institutions observe the student performance by utilizing the co-curriculum and internal assessment. The researchers are using varied students'

attributes and factors for examining the student performance. The researchers usually follow internal assessment, external assessment, CGPA, the final scores of examinations and more co-circular activities of the students. The Indian institutions and the universities use the grade of final examination as the criteria of student academic performance. In existing work, authors focus to develop a student performance prediction model as an application of data mining techniques. To predict the final grades of students based on their previous performance historical data using the three well-known data mining techniques such as decision tree, random forest, and naive Bayes but all these are machine learning approach. This type of machine learning approach not work alike a human brain and it totally depend upon the training algorithm as well as data quality. So, need to select only the relevant feature sets having better quality and train model efficiently using the fitness criteria. In the existing work, authors have used various algorithm as feature extraction or selection but the accuracy of system is still poor and need to update such techniques.

1.6.1 Research Motivation

Educational institutions are intended to provide quality education and analyse student performance and help them improve. Variable factors in current education have led to effective and efficient student performance monitoring so that the ability to predict student performance can provide information for helping students, teachers, managers, and the policymakers. Student performance plays an important role in higher education institutions. Education data mining methods that can increase the benefits and impacts of students, teachers and academic institutions can be used to effectively develop students' achievements and success through predicting students' performance. In recent years, there has been growing interest in using information mining for educational purposes. There are many ways to apply education information, such as Regression, Bagging, Boosting, Ensemble Learning, Near Neighbourhood, Decision Tree, and Naïve Bayes. The educational process culture is used to learn the information available in the field of education and to reveal confidential information from it. Various studies have been conducted that reflect the concerns of students' performance prediction. However, according to our investigation, no research has been carried out to predict the student's prospect of being a glorious student. Student performance prediction is utilized in diverse scenarios, such as personalizing content adaptation, recommending tailored textbooks or resources, issuing early warning alerts for struggling students,

analyzing user behavior on websites, predicting demand patterns, building fraud detection models, and mitigating fraud risks.

Basically, in this research, focus on the development of a framework for academic performance analysis of students in online learning using the concept of machine learning techniques for the prediction of Student Performance and their Engagement in Indian Online Education System.

1.6.2 Current Issues and Challenges

Modeling and predicting student performance within virtual educational environments using machine learning approaches has shown promising results; nonetheless, there are still several operational and technological difficulties that restrict its scale and effectiveness. Virtual learning environments are notoriously complicated, and present ML models have a hard time generalizing to them because of the wide variety of learner interaction patterns and cognitive involvement. Academic performance prediction frameworks are still not very accurate or resilient due to issues with feature representation, data sparsity, and the interpretability of the models.

1. Data Quality and Incompleteness: The effectiveness of any ML model depends heavily on the quality, completeness, and diversity of the data used for training. In online learning, data is often noisy, incomplete, or inconsistent due to missing student activity logs, technical glitches, or manual errors. Inadequate data preprocessing can lead to poor model performance and inaccurate predictions.

2. Lack of Standardization Across Platforms: Online learning environments utilize diverse LMS and platforms, each generating different types of data in varied formats. The lack of standardization across these platforms complicates the integration and analysis of data, making it challenging to build a unified model for performance analysis.

3. Handling High-Dimensional and Imbalanced Data: Academic performance data generated in online learning environments is often high-dimensional, with numerous features such as interaction logs, quiz scores, and assignment submissions. Additionally, the data is frequently imbalanced, where instances of low-performing or disengaged students are relatively rare compared to high-performing ones. Handling such data efficiently and preventing model bias remains a critical challenge.

4. Difficulty in Identifying Complex Learning Patterns: Student engagement and learning behaviours in online environments are often complex and nonlinear. Capturing and interpreting these nuanced patterns requires sophisticated models that can dynamically adapt to changes in learning behaviours over time. Traditional ML models may struggle to capture such intricacies, leading to suboptimal predictions.

5. Scalability and Computational Complexity: As the volume of educational data grows, scalability and computational efficiency become critical concerns. Training and deploying ML models on large-scale datasets require significant computational resources, which may be a barrier for institutions with limited technical infrastructure.

6. Limited Availability of Domain Expertise: Building effective ML models for academic performance analysis requires both technical expertise and deep domain knowledge in education. The lack of collaboration between data scientists and education experts can lead to models that may overlook critical contextual information, reducing the relevance and applicability of the results.

1.6.3 Justification for the Study

Online learning platforms now produce extensive student data about their interactions as well as their performance and their levels of engagement. The effective analysis of this data serves as a fundamental requirement to discover patterns which will enhance educational results and individualize lesson materials while enabling prompt support measures. Academic performance assessment techniques based on traditional approaches fail to detect the multifaceted and evolving behaviours of students who learn online. Educational researchers need to implement Machine Learning techniques through properly defined frameworks to understand academic performance patterns. The need for this study stems from conventional assessment weaknesses combined with urgent academic demands for data-based smart solutions. Machine learning models transform enormous multi-dimensional datasets into identifiable hidden data patterns which manual inspection methods cannot detect. Through implementation of ML capabilities in analytical procedures educators receive performance predictions which help them spot students at risk while designing individualized academic plans for improved student success.

The integration of metaheuristic optimization techniques helps elevate model performance through optimized feature selection along with adjusted hyperparameters thus achieving increased accuracy together with operational efficiency. This study justifies its research design because it develops a flexible forecasting framework which can handle student population heterogeneity and online learning system complexity to address data imbalance and generalization requirements and performance monitoring needs. The proposed study serves a valid purpose through its ability to connect fundamental academic performance research techniques with powerful machine learning systems. Educational establishments and instructors will gain usable information through this proposed framework resulting in more active student participation together with fewer students leaving school while achieving superior educational results across digital teaching platforms.

1.7 CONTRIBUTION OF THESIS

By utilizing machine/deep learning methodologies and developing a robust framework, this thesis makes several significant contributions toward improving the academic achievement analysis of students in online learning environments. It can predict student outcomes, identify at-risk learners, and optimize learning pathways. We present a thorough ML-based system for online learning performance analysis and prediction. In order to produce useful insights, the framework efficiently processes massive amounts of data produced by LMS, such as student interactions, evaluations, and engagement logs. This thesis studies and contrasts the efficiency of many ML techniques, such as Decision Regression, Trees, Random Forest, SVM, KNN, and Gradient Boosting. In order to find the best models for accurately forecasting student performance, this comparison study is useful. In order to identify pupils at risk of disengagement or academic failure early on, the framework incorporates predictive models. In turn, this allows for more targeted and timely interventions, which boost retention and performance in the classroom.

A recommendation engine is incorporated into the framework to provide personalized learning pathways based on individual student performance, learning preferences, and engagement patterns. This customization enhances student engagement and fosters improved academic success. To address the challenges posed by imbalanced datasets commonly found in academic performance analysis, the study employs techniques such as oversampling, under sampling, and cost-sensitive learning. These methods improve model robustness and ensure better generalization across diverse student populations. The proposed framework supports real-time monitoring of student

performance and engagement, allowing educators to track learning progress and make timely adjustments to instructional strategies. This capability ensures continuous assessment and improvement of the learning process. The framework is designed to be scalable, allowing its deployment across various educational institutions and platforms. It can efficiently handle large-scale data while maintaining high accuracy and computational efficiency, making it suitable for real-world applications.

1.8 STRUCTURE OF THE THESIS

Chapter 2 presents a comprehensive review of the literature related to academic performance analysis of students in online learning environments using various machine learning techniques. The chapter explores traditional approaches, clustering and classification algorithms, metaheuristic optimization strategies, and evaluation metrics used in the field. Additionally, this chapter provides a detailed summary of the existing methodologies in tabular form, allowing for a comparative analysis of different approaches. The chapter concludes by identifying the research gaps in the current body of knowledge, highlighting the need for more efficient and interpretable ML models to improve the prediction of student outcomes and personalized learning experiences.

Chapter 3 introduces the problem formulation derived from an extensive literature survey on academic performance analysis in online learning environments. It outlines the key research objectives, which include improving student performance prediction, identifying at-risk learners, and enhancing learning outcomes through adaptive content delivery. This chapter also defines the challenges associated with handling high-dimensional educational data, addressing data imbalance, and ensuring model generalization. Towards the end, the chapter provides a detailed explanation of the proposed machine learning models, metaheuristic optimization techniques, and feature selection mechanisms that aim to enhance predictive accuracy and facilitate early intervention.

Chapter 4 describes the research methodology employed in this study, focusing on the development and implementation of the proposed framework for academic performance analysis. This chapter elaborates on the techniques used, including data preprocessing, feature engineering, classification algorithms such as Regression, Decision Trees, Random Forests, SVM, and deep

learning models like Neural Networks (NNs). The methodology is designed to address issues such as model overfitting, hyperparameter tuning, and interpretability.

Chapter 5 discusses the experimental setup required for implementing the proposed framework for academic performance analysis. It covers details regarding the datasets used, including student interaction logs, assessment records, and engagement metrics obtained from LMS and MOOCs. The chapter also highlights data preprocessing techniques such as handling missing values, normalization, and feature extraction. Additionally, it describes the computational infrastructure, software tools, and libraries (such as Python, Scikit-learn, and TensorFlow) used for model training, testing, and evaluation. The chapter concludes by explaining the configuration and environment settings for conducting experiments and analysing results.

Chapter 6 presents the experimental results and performance analysis of the proposed framework for academic performance analysis. The chapter evaluates the effectiveness of the ML models and optimization techniques using various performance metrics, including accuracy, precision, recall, F1-score, and AUC-ROC. It also includes a comparative analysis of the proposed approach against existing models and benchmarks, demonstrating the improvement achieved in predictive accuracy, early risk identification, and personalized content recommendations. The chapter further discusses insights derived from the experimental findings and highlights the advantages of integrating metaheuristic optimization for feature selection and hyperparameter tuning.

Chapter 7 concludes the thesis by summarizing the key findings, contributions, and implications of the proposed framework for academic performance analysis. The chapter highlights how the developed model enhances student performance prediction, identifies at-risk learners, and facilitates personalized learning experiences. Additionally, it explores potential avenues for future research, such as incorporating Explainable AI (XAI) techniques for model interpretability, developing adaptive learning models using reinforcement learning, integrating multi-source educational data for enhanced predictions, and leveraging federated learning for privacy-preserving academic performance analysis. These directions aim to refine and extend the proposed framework, ensuring its applicability across diverse educational settings.

CHAPTER 2

LITERATURE SURVEY

2.1 REVIEW OF LITERATURE

This chapter presents many scenarios to deliver a thorough evaluation of current research concerning academic performance analysis in online learning settings utilising machine learning (ML) approaches. The analysed study emphasises the utilisation of hybrid methodologies that integrate conventional statistical techniques with sophisticated machine learning models to forecast student performance and improve educational experiences. This chapter will analyse several types of data collected from Learning Management Systems (LMS), Massive Open Online Courses (MOOCs), and online learning platforms. An extensive yet succinct examination of prominent modern clustering and classification methodologies will be presented, highlighting its use in predicting student performance, early detection of at-risk learners, and adaptive content selection. These approaches are employed to analyse and segment diverse educational data, enhancing the assessment of academic achievement and the development of personalised learning plans. The literature review section is the most favoured part of the thesis. The study's thesis primarily focusses on material derived from various books, articles, journals, and research papers published nationally and internationally. This encompasses methods for determining the reasons individuals utilise technology.

Academic Performance Analysis using Machine Learning:

Rasool Fakoor et al. (2013) [1] proposed a Deep Learning model for feature extraction and classification to predict student academic performance using educational datasets. Although the model demonstrated average accuracy, the study incorporated Principal Component Analysis (PCA) to enhance the feature selection process and improve overall performance. This approach emphasized the importance of dimensionality reduction in handling high-dimensional student data, suggesting that integrating PCA with deep learning can lead to more efficient and interpretable models for academic prediction.

Manhães et al. (2015) [2] investigated how to predict student performance in STEM undergraduate programs using automated machine learning approaches in order to identify at-risk students early and facilitate academic interventions in a timely manner. The study's foundation is the increasing need for data-driven decision-making in higher education, especially in STEM fields (science, technology, engineering, and mathematics), where academic challenges and high dropout rates are prevalent.

Altujjar et al. (2016) [3] offered a data-driven strategy for utilizing educational data mining techniques to find important courses that have a major influence on students' academic achievement. Supporting academic planning and intervention strategies in higher education institutions was the driving force behind the project. The researchers sought to determine which courses had a disproportionate impact on students' academic trajectory and total GPA by examining their academic records.

Rianne Conijn et al. (2016) [4] used LMS Predictive Models on data from 17 blended courses with 4,989 students. The regression analysis showed that early intervention is possible using LMS data, but the study highlighted the need for early feedback mechanisms to assess learning propensity and provide timely interventions.

Dinggang Shen et al. (2017) [5] had proposed a Deep Learning-based Feed Forward Neural Network model to predict student performance using academic and behavioural data. While the model showed encouraging results in identifying key performance indicators and achieving satisfactory accuracy, the study highlighted that further enhancement could be achieved by integrating domain-specific educational insights and adopting a more tailored methodological framework to better capture the diverse learning patterns among students.

Patricia Everaert et al. (2017) [6] applied Deep Learning and Surface Learning approaches to data collected from questionnaires from 246 students. The results showed that 19% of students performed poorly on both learning methods. The study recommended that accounting educators encourage more consistent student participation throughout the academic year to improve performance

Helen M. G. Watt et al. (2017) [7] had used Descriptive Statistics to analyse mathematics engagement among 551 students from 15 Australian schools, taught by 37 teachers. Although the

sample distribution was normal, classroom mastery, teacher enthusiasm, and behavioural engagement were negatively skewed. The study suggests investigating whether and how classroom environments can transfer students' engagement between different learning contexts.

Anthony F. Botelho et al. (2017) [8] had employed Deep Learning and Recurrent Neural Networks (RNN) on data from the Assessments learning platform. The proposed model achieved 75% accuracy but indicated that differences in student performance were introduced due to geographical factors, suggesting a need to account for these variables in future research.

Anja Hawlitschek et al. (2017) [9] had implemented Deep Learning Models combined with a digital educational game using a dataset of 150 participants. The results showed that students gained knowledge through the educational game, but the study found no significant impact on intrinsic motivation, suggesting that future work should explore intrinsic factors affecting student engagement.

Sherlock A. Licorish et al. (2018) [10] applied Deep Learning, Descriptive Statistics, and Khoot Learning Tool GSRS (Group Study Rooms) involving 14 students. The study demonstrated that Kahoot! positively influenced classroom dynamics, student engagement, and motivation, though future research should involve larger sample sizes to validate these findings.

Zilong Hu et al. (2018) [11] had applied Deep Learning techniques using Fully Convolutional Networks (FCN) to predict student performance based on educational datasets. The model achieved an accuracy rate of 60%, indicating its limited effectiveness in capturing the complex patterns of student learning behaviour. The study highlighted the need for future research to explore more refined model architectures, incorporate diverse educational features, and address data sparsity challenges to enhance predictive accuracy in academic performance modeling.

Kim et al. (2018) [12] introduced GritNet, a Bi-LSTM deep learning architecture designed to analyze learner behavior from sequential clickstream data in MOOCs. This model significantly outperformed traditional approaches like logistic regression, particularly in early-course predictions

Yang Jiang et al. (2018) [13] had employed a Deep Neural Network with feature engineering on Betty Brain to predict student performance. The dataset consisted of data collected from 6th-grade

students in an urban public school. The accuracy of both approaches used was reported to be the same. However, the study highlights the need to explore the generalizability of their findings to other educational settings and broader student populations.

Ye Mao et al. (2018) [14] had implemented a Recurrent Neural Network (RNN), BKT Model, Intervention-BKT (IBKT), and LSTM using two datasets from intelligent tutoring systems named Cordillera and Pyrenees. Results indicated that the BKT model outperformed IBKT and LSTM on these datasets. However, future research is needed to identify models capable of predicting student learning gains more effectively.

Renée M. Filius et al. (2018) [15] had applied Small Private Online Courses (SPOCs) and Deep Learning using questionnaires from a master's epidemiology course involving 41 students. The model achieved 90% accuracy, but the study recommends allowing round-trip voting to further strengthen feedback dialogue.

M.T Azizan et al. (2018) [16] had applied Deep Learning and Cooperative Learning Strategies to a dataset of 105 third-year chemical engineering students. Results demonstrated that cooperative learning significantly enhanced student learning experiences, with the study recommending that future work involve a larger group of students.

Ingrid le Roux et al. (2018) [17] applied Deep Learning, Flipped Classroom, Video, Seminar, and Community of Inquiry Models to a dataset of 80 students. The flipped classroom approach proved effective in enhancing online and connected learning models. However, the study highlighted the absence of sufficient tracking data in CMS to capture detailed student interactions with video content.

Petrea Redmond et al. (2018) [18] had focused on Social, Cognitive, Behavioural, Collaboration, and Emotional Commitments using datasets collected from online models. The results were shared with national and international experts in online teaching and learning. The authors emphasized the need to create online learning environments that promote student participation and facilitate collaboration between students, teachers, and educational institutions.

Hajra Waheed et al. (2019) [19] had developed a Deep Learning model using the OULA dataset of 32,593 students for predicting academic outcomes. The model achieved an accuracy of 83%,

but the performance evaluation highlighted low sensitivity and precision. The study emphasizes the need to improve accuracy and optimize the model for better results in future implementations.

Shaveta Dargan et al. (2019) [20] had implemented a hybrid model combining Convolutional Neural Network (CNN) with LSTM/GRU architectures to predict student performance based on temporal and spatial educational data patterns. The model achieved an accuracy of 85%, demonstrating the potential of deep learning in capturing complex student behaviour and academic trends. However, the study noted that further improvements could be realized by expanding the dataset size and leveraging greater computational power to better generalize the model across diverse educational contexts.

Yuk-Hoi Yiu et al. (2019) [21] employed Deep Learning and Convolutional Neural Network (CNN) models to predict student performance by analyzing visual attention patterns using data from 3,946 eye-tracking images. Their proposed model, DeepVOG, achieved high precision in tracking students' visual focus and accurately identified engagement levels through improved segmentation of eye movement patterns. While the model showed strong potential in understanding student attention and behaviour during learning tasks, the study emphasized the need to extend the model's capabilities to broader learning contexts and diverse student populations for more generalized performance prediction.

Dijana Oreški et al. (2019) [22] applied Machine Learning, LMS, Data Mining, and Classification techniques on a dataset from the University of Zagberg. Results indicated that Neural Network (NN) modeling classified students better than other ML approaches. However, the study recommends further exploration of predictive models to improve teaching outcomes.

Aly Al-Amyn Valliani et al. (2019) [23] had proposed a Deep Learning model to predict student performance by analyzing behavioral patterns and engagement metrics from educational datasets. The model achieved a high accuracy of 92%, demonstrating the potential of AI in personalized learning environments. However, the study emphasized the importance of incorporating additional dimensions such as psychological or familial background data to further improve precision and effectiveness in future academic performance prediction systems.

Blake A. Richards et al. (2019) [24] had applied Deep Learning models using student behavioural and engagement data to predict academic performance. The model exhibited promising results,

indicating its effectiveness in understanding complex learning patterns. The study emphasized that organizing educational data within a structured framework can significantly enhance the development of models aimed at improving student learning outcomes and personalized education strategies.

Biyun Huang et al. (2019) [25] used Gamification and Flipped Course Techniques to engage a group of 48 students using Moodle. This study was the first to integrate a variety of motivational theories to guide the design and testing of a gamification framework based on the GAFCC theory. The authors emphasized the need to integrate the same framework across other courses and student groups to verify its effectiveness.

Bui Ngoc Anh et al. (2019) [26] had applied Support Vector Machine (SVM) to a dataset from PRF192 at FPT University, consisting of 25,391 rows of data. The proposed model achieved an accuracy of 77%. However, the study suggests conducting further research due to the stringent requirements associated with monitoring student behaviour.

Yen-Chun Jim Wu et al. (2019) [27] used Mobile-Based CRS Technology to analyse datasets from an 18-week course on Entrepreneurship Management involving 22 graduate students. The study demonstrated that CRS technology enhances student interaction but emphasized the need to cultivate student responsibility and self-control when using mobile technology in learning environments.

Kris M.Y. Law et al. (2019) [28] used Hypothesis Testing, Student Enrolment, Learning Motivation, and Blended Learning techniques on data from 207 students. Results showed that learning motivation improves student enrolment and social presence. The study suggests expanding research to include more selected blended learning courses.

M. Ali Akber Dewan et al. (2019) [29] used Machine Learning, SVM (Gabor), MLR (CERT), and Boost (BF) models on the DAiSEE dataset, which included 112 individuals (80 males and 32 females). The accuracy of the classifiers ranged from 0.714 to 0.729. The study highlights the need to promote online education technology to increase student participation and engagement in assessments.

R. Martínez et al. (2019) [30] applied a K-means unsupervised clustering approach to predict student performance using a dataset of 153 college freshmen students. The proposed model achieved an accuracy of 80%. However, the study suggests using larger datasets in future research to enhance the model's generalizability and reliability.

Alberto Rivas et al. (2019) [31] had employed Tree-Based Machine Learning techniques to predict academic performance using a dataset of 7,909 students across 4 online courses. The Min-Max standardization model was found to be most suitable for testing data, achieving an accuracy of 82.46%. However, the study identified the need to measure the time students spend interacting with the virtual environment for better insights.

Mukesh Kumar et al. (2019) [32] had utilized Machine Learning, MATLAB, Mean Square Error, and Threshold-Based Segmentation techniques to analyse data from an online repository. Results indicated that Artificial Neural Networks performed better than Support Vector Machines. The authors suggested future work should explore changing the number of neurons and parameters to optimize performance.

Xing Xu et al. (2019) [33] applied Machine Learning, Neural Networks, and SVM on internet usage data of 4,000 students. The study concluded that behavioural discipline plays a significant role in academic success. However, the authors suggested including more data on online behaviour to refine the model.

Shan Li et al. (2020) [34] had implemented Naïve Bayes Classifier, K-Nearest Neighbors (KNN), and Support Vector Machine (SVM) models on a dataset of 59 students to predict cognitive participation based on facial behaviour cues. The proposed model achieved 80% accuracy. However, the study suggests further refinement to validate the cognitive predictions across diverse student populations.

Z Chen et al. (2020) [35] had utilized artificial intelligence for students' performance analysis in education 4.0. The proposed HDNN model offers a Hybridized Deep Neural Network to determine the dynamic factors that affect the academic outcomes and analyze them. The approach will inculcate deep learning to track, anticipate, and evaluate the performance of the students in real-time so that the education system can become customized to suit individual learning requirements. The findings indicate that the HDNN is more effective in educational data mining (than the

traditional machine learning techniques) as it achieves higher prediction accuracies. Nevertheless, the research does not directly refer to the interpretability of the model or scalability in various educational settings that are of fundamental importance in deployment and education-embracement.

Boddeti Sravani et al. (2020) [36] applied Machine Learning Algorithms, Classification Prediction, and Linear Regression on a dataset of 100 students for prediction of student's performance. The proposed model achieved a good accuracy. However, the authors suggested that improved applications with enhanced ability and efficiency could become an integrated part of educational institutions in the future.

Hassan Abuhassna et al. (2020) [37] used Deep Learning, Support Vector Machines (SVM), and LSTM Predictive Models on a sample dataset of 243 students to improve students' academic achievements and satisfaction. The proposed model achieved a coefficient of 0.917, but the study recommended further research to investigate the relationship between platform complexity and the Technology Acceptance Model (TAM).

Şeyhmus Aydoğdu (2020) [38] used Deep Learning and Artificial Neural Networks (ANNs) on a dataset of 3,518 students from a university to predict students' performance. The proposed model achieved an accuracy of 80.47%. The study recommends optimizing the accuracy by exploring different numbers of layers, parameters, and neurons in hidden layers.

Bansal et al. (2020) [39] built a comprehensive ML/DL pipeline with multiple algorithms to assess academic success in online learning during the COVID-19 pandemic, confirming the reliability of ensemble deep models.

Matthew Botvinick et al. (2020) [40] implemented Deep Reinforcement Learning models to analyse student behaviour and predict academic performance using synthetic data derived from random online samples. The approach explored adaptive learning strategies and decision-making processes in virtual learning environments. However, the results were average, indicating that while the model introduced a promising framework for simulating student interactions, it requires further refinement and real-world data integration to fully maximize its predictive potential and reliability in educational settings.

Okereke GE et al. (2020) [41] applied Machine Learning and EDM models on data collected from 103 first-year Computer Science students at the University of Nigeria. The proposed model achieved an accuracy of 92%. The study emphasized that classifier choice alone does not determine prediction accuracy but rather the nature of the dataset.

Abdullah Alshanqiti et al. (2020) [42] applied Linear Regression, Neural Networks, and Machine Learning models for predicting students' performance using data from the OULAP dataset. Results showed considerable improvements compared to single baseline models. The study highlights a gap in the application of hybrid approaches that could enhance accuracy.

Dwivedi et al. (2020) [43] provided a thorough case study at Kazan Federal University that assessed long-term trends in undergraduate students' academic performance using both neural networks and conventional statistical techniques like SPSS. Their hybrid methodological approach made it possible to compare and contrast contemporary machine learning models with traditional statistical methodologies, showing the advantages and disadvantages of each.

Tsimakuridze and Dzitac (2020) [44] systematically reviewed DL applications in student performance analysis, finding that LSTM and CNN architectures dominate in cases involving time-series learning behavior.

Zhaoli Zhang et al. (2020) [45] tested a number of methods on a dataset consisting of 49,920 tagged photos of 47 people using methods such Active Shape Model-SVM, Gabor SVM, CLBR-SRC, AWLGCP & FSR Method, and LGCP Feature Extraction. The suggested model was 94% accurate. The authors proposed creating adaptable algorithms that might incorporate several models for assessing students' engagement in the learning process, as well as screening models for distinct learning styles.

Chung Kwan Lo & Khe Foon Hew (2020) [46] compared Traditional Learning, Flipped Learning, and Gamification approaches using a dataset of 9th-grade mathematics students. The results demonstrated that flipped classroom students scored higher in mathematics and exhibited better cognitive engagement. However, the study suggested exploring differences between flipped learning with and without gamification to further refine its effectiveness.

Hajra Waheed et al. (2020) [47] had employed the concept of Deep ANN, SVM, and Logistic Regression using the famous OULA dataset of 32,593 student's records. The model achieved an accuracy of 88.5%. Future work should involve the application of NLP and advanced deep learning models to extract text data from student feedback and identify activities that impact performance.

Shuo-Chang Tsai et al. (2020) [48] had implemented Statistical Learning, Deep Learning, Logistic Regression, and Prediction Models on a dataset of 3,552 students from a university in Taiwan. The model achieved 90% accuracy with deep learning and 88% with logistic regression. Future research should include more variables related to student participation, family, and learning behaviours to enhance model precision and sensitivity.

Engr. Sana Bhutto et al. (2020) [49] had compared the Sequential Minimal Optimization (SMO) Algorithm with Logistic Regression using the Kalboard 360 dataset, which contains 500 records and 16 distinct attributes. The results showed that SMO achieved a higher accuracy of 79% compared to Logistic Regression, which achieved 73%. However, the authors highlighted the challenge of transitioning from a conventional learning system to an e-learning system, which complicates the analysis of student learning behaviour.

Chongying Wang et al. (2020) [50] had utilized Machine Learning models, including XGBoost and Multiple Stepwise Regression Models, to analyse a dataset of 3,800 students from a Research-Oriented University under the Education Ministry. The XGBoost model performed significantly better with an accuracy of 79.26%, compared to Multiple Stepwise Regression, which achieved only 27.6% and 32.1% in different tests. The authors noted that the sample size was small and not reflective of actual patterns, with many students showing a lack of interest in participation.

Ahmed A. Mubarak et al. (2020) [51] employed Input-Output Hidden Markov Model (IOHMM), Logistic Regression, and Machine Learning models using the OULA dataset. The proposed model achieved an accuracy of 84%. However, the study recommends developing more interpretable student behaviour models and providing educators with curriculum feedback on student status to intervene early in the case of at-risk students.

Albreiki et al. (2021) [52] conducted a comprehensive review highlighting that Support Vector Machines (SVM), Decision Trees (DT), Random Forests (RF), and Artificial Neural Networks (ANN) are the most frequently used ML models in predicting student performance across learning management systems.

K. Subhash Bhagavan et al. (2021) [53] evaluated a dataset of 550 student records using data mining, linear vector quantization, and hybrid linear vector quantization in order to predict students' academic success. The proposed model achieved 90% accuracy, and the authors identified various possibilities for improving the model using clustering algorithms.

Qazi et al. (2021) [54] emphasized the increasing use of classification models in early identification of at-risk students, noting SVM and logistic regression as the most effective traditional methods. Their results are in line with a larger trend in educational data mining (EDM), which uses machine learning approaches to forecast academic risk by analyzing student performance data, attendance records, and behavioral markers.

Iatrellis et al. (2021) [55] created a decision support system that uses decision tree algorithms to help academic advisors find students who might need academic intervention in a timely manner. High accuracy in identifying kids at risk was attained by the system, which was developed utilizing institutional data, such as academic records, enrollment trends, and performance indicators. Because decision trees are transparent and interpretable, academic staff were able to comprehend and have faith in the reasoning behind the model's predictions. This made their selection very advantageous.

Khawlah Altuwairq et al. (2021) [56] had applied a Convolutional Neural Network (CNN) and Naïve Bayes Classifier on educational datasets to recognize and predict student emotions and performance through facial expressions. The study utilized data from 12,271 real-world images from the Real-World Affective Faces (RAF) and 35,887 images from the FER2013 dataset, achieving an impressive accuracy of 93%. Despite the high performance, the authors suggested that future systems should integrate physical movement and emotional behaviour analysis into intelligent online learning environments to gain a more holistic understanding of student engagement and academic outcomes.

Jiun-Yu Wu (2021) [57] implemented a Supervised Machine Learning Model to analyse Facebook posts and comments. The study revealed that weakly supervised machine learning provided biased and inaccurate labels. The authors suggested incorporating self-efficacy and interest in statistics in future research to gain a more comprehensive understanding of students' statistical learning.

Shan Li et al. (2021) [58] had applied Supervised Machine Learning Algorithms to analyse cognitive engagement and facial behaviours using a dataset of 61 students. The results showed 82 segments were identified as engaged states and 85 segments as less engaged states. However, the homogeneity of participants in terms of race and academic background presents a limitation that future research should address.

Stephen L. Chew & William J. Cerbin (2021) [59] had developed a framework comprising nine interactive cognitive challenges that teachers can address to strengthen student learning using raw data. Although the framework has critical implications for teaching, it does not promote a specific pedagogy or technology, leaving open the need for more focused approaches that align with specific teaching models.

Jane Scott et al. (2021) [60] had used Cognitive Behaviour Therapy (CBT) and Self-Practice/Self-Reflection (SP/SR) techniques on a group of 17 students. The results did not show significant differences between the groups for all previous measurements. The study recommends conducting experiments with larger sample sizes to provide clearer distinctions in results.

Wenting Zou et al. (2021) [61] explored the relationship between student reputation and social presence in the MOOC student network using datasets from MOOCs. Results showed that indicators such as asking questions and expressing gratitude were positively correlated with learner reputation, while expressing criticism or negative emotions was counterproductive. Future research should use MOOC datasets from diverse subjects to validate these findings.

Ahmed Ali Mubarak et al. (2021) [62] had implemented Deep Learning, Support Vector Machine, and Predictive Models (LSTM) using online datasets from CAROL. The proposed model achieved 90% accuracy, and the authors suggested developing a platform to enhance teaching policies in the online learning environment.

Alberto Rivas et al. (2021) [63] had presented another study, Alberto Rivas and his team used Artificial Neural Networks (ANNs) in Virtual Learning Environments (VLEs) on a public dataset of 120 students. The study identified that the number of student visits to the resources available on the VLE platform was a key factor affecting student performance. The authors recommended using machine learning models to predict behaviour models and improve academic outcomes.

Dr. Abdulmunem Alshehhi et al. (2021) [64] had developed a framework that combines Artificial Intelligence (AI), Online Learning, and Learning Organization techniques to analyse data collected during the COVID-19 pandemic from academic references. The results demonstrated that AI produced good results, and the study concluded that Online Learning (OL) can enhance the learning system and improve organizational outcomes.

Nilesh V. Ingale et al. (2021) [65] applied Educational Data Mining (EDM) using various Data Mining techniques to predict student academic performance. Clustering in EDM was found to provide useful information through multi-level non-hierarchical clustering. However, the study recommends further research to improve the efficiency of existing EDM methods for identifying weak students at risk of failure.

Mehmet Kokoç & Arif Altun (2021) [66] had compared Prescriptive Learning Approaches combined with e-learning environments to Artificial Neural Network (ANN) algorithms using a dataset of 126 students enrolled in a 12-week course. The ANN algorithms performed best in predicting academic performance. However, the study suggests that future large-scale research should focus on assessing the long-term impact of Learning Design (LD) on student performance.

Sekeroglu and Ozkan (2022) [67] performed a comparative analysis of ML algorithms such as SVM, Logistic Regression, KNN, and Random Forest using midterm and final grades from Turkish university students. Their results showed that tree-based algorithms consistently yielded better accuracy in forecasting academic outcomes**.**

Francisco et al. (2022) [68] developed a suggestion module for software maintenance training using Q-learning in an Intelligent Tutoring System. The system used a lightweight, interpretable RL model to dynamically modify learning activities according to student performance. Their findings demonstrated increased engagement, quicker learning, and better task accuracy. The usefulness of basic reinforcement learning techniques in adaptive training is demonstrated in this paper. In specific technical fields, it provides a scalable approach to individualized learning.

The research indicates that the precision and resilience of machine learning models employed for academic performance assessments may be markedly enhanced by hyperparameter optimisation, a vital component of the pre-processing phase. Optimising hyperparameters improves model prediction performance and assures flexibility across various datasets and learning contexts. Table 2.1 summarises the machine learning methodologies employed to assess student performance in online learning contexts, emphasising the utilised datasets, implemented procedures, and resultant outcomes. The table presents a thorough examination of academic performance analysis models utilising machine learning methods, highlighting their advantages, drawbacks, and possible avenues for enhancement.

2.2 LITERATURE SUMMARY (TABULAR FORM)

Table 2.1 Literature Review

Ref No.	Approaches	Dataset	Results	Gaps Identified
[1]	Deep Learning, Feature Learning, Classifier Learning	Educational dataset of students	Average accuracy	PCA incorporated for performance improvement.

[2]	Machine learning classifiers: Decision Trees, Naïve Bayes, k-NN; Feature engineering	Academic and demographic data from STEM undergraduate students	Accuracy of the model is good	Challenges in generalizing models across institutions; need for inclusion of behavioral and temporal data.
[3]	Classification, association rule mining, feature selection algorithms	Institutional academic records from a university	Performance can be improved by early intervention in these courses.	Limited to one institution; lacks demographic and longitudinal data for broader applicability
[4]	LMS, Predictive Models	Data from 17 blended courses with 4,989 students	Regression analysis demonstrated the potential for early intervention	Provide early feedback using LMS data to assess learning propensity.
[5]	Deep Learning, Feed Forward Neural Network	Educational Dataset from Kaggle.com	Promising results	Incorporate domain-specific knowledge for new methodologies.
[6]	Deep Learning, Surface Learning	Data from 246 student questionnaires	19% of students performed poorly	Encourage consistent student participation throughout the year.

[7]	Mathematics Engagement, Descriptive Statistics	Dataset of 551 students taught by 37 teachers	Normal sample distribution except for classroom mastery	Investigate whether classroom can transfer students between different situations.
[8]	Deep Learning, Recurrent Neural Networks (RNN), EDM	ASSISTments learning platform	Accuracy of 75%	Geographical differences introduced variability in results.
[9]	Deep Learning, Educational Game	Dataset of 150 participants	Students gained knowledge through educational games	No significant impact on intrinsic motivation.
[10]	Deep Learning, Descriptive Statistics, Kahoot Learning Tool, GSRS	14 student participants	Kahoot! positively influenced classroom dynamics and engagement	Larger sample size needed to validate findings.
[11]	Deep Learning, Fully Convolutional Network	Dataset taken from Kaggle.com	Accuracy of 60%	Explore unique research themes and future advancements.
[12]	Bi-LSTM (GritNet)	MOOC clickstream data	LSTM outperformed LR in early prediction	Model requires large sequential datasets.

[13]	Deep Neural Network, Feature Engineering	Dataset of 6th-grade urban public school students	Same accuracy for both approaches	Explore the generalizability of findings to broader populations.
[14]	RNN, BKT, IBKT, LSTM	Two intelligent tutoring systems datasets	BKT outperformed IBKT and LSTM	Identify models capable of predicting student learning gains.
[15]	Small Private Online Courses (SPOCs), Deep Learning	Questionnaire data from a master's epidemiology course (41 students)	Accuracy of 90%	Allowing round-trip voting could enhance feedback dialogue.
[16]	Deep Learning, Cooperative Learning Strategy	Dataset of 105 third-year chemical engineering students	Enhanced learning experience for students	Apply the model to a larger group of students.
[17]	Deep Learning, Flipped Classroom, Video, Seminar, Community of Inquiry Models	Dataset of 80 students	Flipped classroom approach enhanced online learning models	CMS lacked detailed tracking data of student interactions with video content.
[18]	Social, Cognitive, Behavioral, Collaboration,	Dataset from online models	Results shared with national and international	Create an online learning environment to improve participation and teaching outcomes.

	Emotional Commitment		teaching experts	
[19]	Deep Learning, OULA Dataset	32,593 students	Accuracy of the model is 83%	Low sensitivity and precision; requires improvement to optimize model performance.
[20]	CNN, LSTM/GRU Network	Educational Dataset from Kaggle.com	Accuracy of 85%	Future work should focus on increased data availability and computational resources.
[21]	Deep Learning, CNN, DeepVOG	3,946 VOG images	High-precision pupil tracking	Extend model capabilities for better performance.
[22]	Machine Learning, LMS, Data Mining, Classification	Dataset from the University of Zagberg	Neural Networks classified students better than other ML models	Explore predictive models to improve teaching outcomes.
[23]	Deep Learning, Medical Image Classification	ADNI dataset	Accuracy of 92%	Genetic data remains an important aspect for future research.
[24]	Deep Learning, Theories of Brain Development	Random image data	Promising responses	Structured research framework can accelerate theory development.
[25]	Gamification, Flipped Course	Group of 48 students using Moodle	Integrated motivational theories for gamification framework	Extend framework to different courses and groups.

[26]	Support Vector Machine (SVM)	Dataset from PRF192 at FPT University, 25,391 rows	Accuracy of 77%	Conduct further research due to strict requirements in behavior monitoring.
[27]	Mobile-Based CRS Technology	Dataset from an 18-week course in Entrepreneurship Management	CRS technology effectively promoted student interaction	Students need to develop responsibility and self-control when using mobile technology.
[28]	Hypothesis Testing, Student Enrolment, Learning Motivation, Blended Learning	Data samples from 207 students	Learning motivation improves student enrolment and social presence	Expand research to include more blended learning courses.
[29]	Machine Learning, SVM (Gabor), MLR (CERT), Boost (BF)	DAiSEE dataset (112 individuals: 80 males, 32 females)	Accuracy: MLR(CERT) = 0.714, Boost(BF) = 0.728, SVM(Gabor) = 0.729	Promote online education technology for exam participation.
[30]	K-means, Unsupervised Clustering	Dataset of 153 college freshmen	Accuracy of 80%	Use larger datasets for better generalizability.
[31]	Tree-Based Machine Learning, Academic	Dataset of 7,909 students across 4 online courses	Accuracy of 82.46%	Time measurement of student interactions with the virtual environment needed.

	Performance Prediction			
[32]	Machine Learning, MATLAB, Mean Square Error, Threshold-Based Segmentation	Online repository	ANN performed better than SVM	Explore changes in number of neurons and parameters to optimize performance.
[33]	Machine Learning, Neural Network, SVM	Internet usage data of 4,000 students	Behavioral discipline plays a role in academic success	Include more data on online behavior for refined models.
[34]	Naïve Bayes Classifier, KNN, SVM	Dataset of 59 students	Accuracy of the model is 80%	Further refinement needed for cognitive predictions in diverse populations.
[35]	Machine Learning, Structural and Functional Connectivity Mapping	Human Connectome Project (HCP)	Average accuracy of 50%	Strengthening the connection between structure and function at a macro-scale level.
[36]	Machine Learning, Classification Prediction, Linear Regression	Sample dataset of 100 students	Accuracy of the model is good	Improved applications with better ability and efficiency should be integrated in institutions.
[37]	Deep Learning, SVM, Predictive Model (LSTM)	Sample dataset of 243 students	Coefficient of 0.917	Explore the relationship between online platform complexity and TAM.
[38]	Deep Learning, Artificial Neural Network	Dataset of 3,518 students from a university	Accuracy of 80.47%	Improve accuracy by varying layers, parameters, and neurons in hidden layers.

[39]	ML/DL ensemble pipeline	COVID-19 15-course dataset taken	DL outperformed engineered features	Dataset tied to pandemic context.
[40]	Deep Reinforcement Learning	Random sample images from Google	Average results	Further exploration required to maximize potential.
[41]	Machine Learning, EDM	Data from Computer Science Dept., University of Nigeria	Accuracy of 92%	Classifier choice does not determine accuracy, but dataset quality plays a critical role.
[42]	Linear Regression, Neural Network, Machine Learning	Data from OULAP	Significant improvements over baseline models	Explore hybrid approaches to further improve model performance.
[43]	Neural networks + SPSS	Kazan Federal University (2012–2019)	Effective in long-term trend analysis	Classical analytics dominate; older data.
[44]	Review of DL models (LSTM, CNN)	Data from Literature review.	DL superior for time-series learning	No empirical experimentation.
[45]	Edge Detection, LGCP Feature Extraction, SVM	Dataset of 49,920 labeled images of 47 individuals	Accuracy of 94%	Develop adaptive algorithms and screening models for different learning styles.
[46]	Traditional Learning, Flipped	Dataset of 9th-grade	Flipped classroom students	Explore the difference between flipped learning

	Learning, Gamification	mathematics students	scored higher in cognitive engagement	with and without gamification.
[47]	Deep Artificial Neural Network, SVM, Logistic Regression	OULA dataset of 32,593 students	Accuracy of 88.5%	Apply NLP and advanced deep learning models to extract text data from student feedback.
[48]	Statistical Learning, Deep Learning, Logistic Regression, Prediction Models	Dataset of 3,552 students from a university in Taiwan	90% accuracy for Deep Learning, 88% for Logistic Regression	Include more variables related to student participation, family, and learning behaviors.
[49]	Sequential Minimal Optimization (SMO) Algorithm, Logistic Regression	Kalboard 360 dataset with 500 records	SMO accuracy: 79%, Logistic Regression accuracy: 73%	Transition from conventional to e-learning complicates learning behavior analysis.
[50]	Machine Learning, XGBoost, Multiple Stepwise Regression Model	Dataset of 3,800 students from a Research-Oriented University	XGBoost accuracy: 79.26%, Stepwise Regression accuracy: 27.6%, 32.1%	Small sample size not reflective of actual patterns.
[51]	Input-Output Hidden Markov Model (IOHMM), Logistic Regression, Machine Learning	Dataset from OULA	Accuracy of 84%	Develop interpretable behavior models and provide curriculum feedback.

[52]	SVM, RF, DT, ANN – Systematic Review	LMS, University records (2009–2021)	Identified SVM and ANN as most effective in EDM	Limited focus on deep learning.
[53]	Data Mining, Linear Vector Quantization, Hybrid LVQ	Dataset of 550 student records	Accuracy of 90%	Explore clustering algorithms to identify different possibilities.
[54]	ML classification models (SVM, LR)	Higher education datasets	SVM and LR best for early at-risk detection	Mostly traditional ML and less on behavioural data.
[55]	Decision trees for academic advisory	Institutional academic records	High precision in identifying at-risk students	Threshold-based; lacks adaptability.
[56]	CNN, Naïve Bayes Classifier	RAF and FER2013 datasets	Accuracy of 93%	Integrate physical movement and emotional analysis into intelligent systems.
[57]	Supervised Machine Learning	Facebook posts and comments	Weakly supervised learning leads to biased and inaccurate labels	Incorporate self-efficacy and statistical learning for better insights.
[58]	Supervised Machine Learning, Cognitive	Dataset of 61 students	Identification of 82 and 85 engaged and less engaged	Homogeneity of participants; diversity required for validation.

	Engagement, Facial Behaviours		segments respectively	
[59]	Framework with Nine Interactive Cognitive Challenges	Raw data	Framework provides critical insights for teaching	Lack of promotion for specific pedagogy or technology.
[60]	Cognitive Behaviour Therapy, Self-Practice/Self-Reflection (SP/SR)	Group of 17 students	No significant differences between groups	Larger sample size required to provide clearer distinctions in results.
[61]	Relationship between Student Reputation and Social Presence in MOOC	Datasets from MOOCs	Indicators such as asking questions and expressing gratitude correlated with learner reputation	Explore the influence of social existence on student reputation using diverse MOOC datasets.
[62]	Deep Learning, SVM, Predictive Model (LSTM)	Dataset taken online from CAROL	Accuracy of 90%	Develop a platform to enhance teaching policies in online environments.
[63]	Artificial Neural Networks (ANNs), Virtual Learning Environment (VLE)	Public dataset of 120 students	Identified student visits as key variable affecting performance	Use of ANNs to predict behavior models for improving academic performance.

[64]	Artificial Intelligence, Online Learning, Learning Organization	Datasets from academic references	AI produced good results	Online learning can support organizational outcomes with AI frameworks.
[65]	Educational Data Mining (EDM), Data Mining Techniques	Data survey using different approaches	Clustering in EDM provided useful factors	Improve efficiency of existing EDM methods to identify weak students.
[66]	Prescriptive Learning Approach, ANN Algorithms	Dataset of 126 students enrolled for a 12-week course	ANN algorithms performed best in predicting academic performance	Assess long-term impact of Learning Design (LD) on performance in large-scale research.
[67]	RF, SVM, LR, KNN	Turkish university grades	RF and DT yielded ~75% accuracy	Used only academic grades.
[68]	Q-Learning in ITS (Reinforcement Learning)	Simulated student information in a tutoring environment for software maintenance	Improved task selection accuracy, faster convergence, and increased engagement	Simulation-only, not validated using actual student data or other subject areas.

2.3 RESEARCH GAP IDENTIFICATION

Upon examining the existing research literature on the study of student academic performance in online learning through machine learning methodologies, the following points have been

recognised as essential results and conclusions derived from the present state of the art. The following are delineated below:

1. Many existing models related to work achieve good accuracy on specific or small datasets but fail to generalize across diverse contexts or larger datasets [10], [30], [57].
2. Limited research on hybrid algorithms combining multiple machine learning methods for results improvements, despite their potential to enhance accuracy and outcomes [42], [54].
3. Current models lack comprehensive integration of emotional, behavioural, and cognitive engagement data, limiting their effectiveness in analysing student performance [36], [56], [58].
4. Existing studies predominantly focus on short-term outcomes, neglecting longitudinal evaluations of learning interventions and design [40], [43], [66].
5. Many high-accuracy models are not interpretable, which makes it difficult for teachers to comprehend and implement successful intervention techniques [11], [51], [60].
6. Despite valuable LMS data, few systems provide real-time feedback and early warning interventions for at-risk students [4], [42].
7. Many studies rely on small or homogeneous datasets, resulting in biased model's incapable of generalizing to larger, more diverse populations [10], [50].
8. Limited research explicitly identifies and assesses the most critical features influencing student performance, impacting model interpretability and precision [19], [38], [49], [51].
9. Insufficient exploration of real-time monitoring of student interactions with digital learning environments restricts the accuracy and effectiveness of prediction models [3], [13], [33].
10. Limited application and integration of advanced deep learning models (e.g., ANNs) within educational data mining contexts despite demonstrated potential to improve learning outcomes [4], [35].

CHAPTER 3

RESEARCH PROBLEMS & OBJECTIVES

3.1 PROBLEM FORMULATION

The swift shift to online learning platforms has created an increasing demand for robust frameworks to assess and forecast student performance. Various research has utilised machine learning (ML) methodologies, including Support Vector Machines (SVM), Artificial Neural Networks (ANNs), Decision Trees, and Deep Learning Models, to evaluate student engagement, forecast academic performance, and identify at-risk students [1]-[68]. Nonetheless, despite significant progress, some obstacles still unresolved. The current models exhibit great accuracy on certain datasets but lack generalisability across varied learning contexts, resulting in biassed predictions [10], [30]. Moreover, although hybrid models and ensemble learning techniques can improve prediction accuracy, they are still underutilised in the field of academic performance analysis [4], [54]. Emotional and behavioural elements that substantially affect student involvement are frequently overlooked, hence constraining the thoroughness of prediction models [36], [58].

The lack of Explainable AI (XAI) models complicates educators' ability to comprehend and accept the judgements made by these models, hence obstructing effective interventions [11], [60]. Real-time monitoring and feedback systems that might facilitate early intervention for underperforming pupils remain little investigated [4], [42]. The dependence on limited sample sizes in several research [10], [40] produces models that fail to represent larger student populations, while insufficient feature selection yields subpar performance [19], [38]. Moreover, the utilisation of sophisticated deep learning models, including ANNs, remains underexplored in educational data mining, potentially yielding enhanced prediction accuracy and profound insights into student learning behaviours [39], [44]. In light of these challenges, there is an urgent necessity to create a hybrid, interpretable, and scalable framework that amalgamates real-time behaviour monitoring, emotional engagement analysis, and sophisticated deep learning methodologies to enhance the precision and efficacy of academic performance forecasting in online learning contexts. Table 3.1 compares various learning management systems techniques in teaching and learning.

Table 3.1 Comparison of Various Learning Management Systems Techniques

S. No	LMS	Open Source / Paid	Key Features	Learning Analytics
1	Moodle	Open Source	Highly customizable LMS that integrates with tools like Microsoft Office 365, Google Apps, and supports modular plugin architecture.	No
2	Google Classroom	Open Source	Centralized platform for course creation, assignment submission, grading, communication, and feedback between teachers and students.	No
3	Canvas	Open Source	Comprehensive digital learning platform with course builder, real-time dashboard, testing tools, and mobile-friendly interface.	Yes
4	Schoology	Not Open Source	Emphasizes building connected learning communities, supports collaboration between students, educators, and administrators.	No
5	LearnDash	Open Source	WordPress-based LMS plugin for course creation, management, and publishing, ideal for	No

			structured e-learning content delivery.	
6	Edmodo LMS	Open Source	User-friendly and accessible platform, particularly effective for K–12 educational environments due to its intuitive design.	No

3.2 RESEARCH OBJECTIVES

The purpose of this research work is to provide a machine/deep learning-based framework optimized for the analysis and prediction of academic performance of students in online learning environments. The framework aims to leverage machine learning algorithms to identify at-risk students, predict learning outcomes, and provide personalized recommendations to improve student success. The following is a list of objectives that have been established for this work:

Obj. 1: To carry out detailed survey regarding usage of different ICT modes adopted by the institution of higher learning.

Obj. 2: To collect and preprocess data regarding the usage of LMS and other ICT tools in Virtual Learning.

Obj. 3: To design a predictive model for academic performance analysis using Machine Learning techniques.

CHAPTER 4

RESEARCH METHODOLOGY

This section describes the development methodologies employed to accomplish the objectives of the proposed model using machine learning techniques for analyzing student's academic performance in online or e-learning environments. The limitations and gaps identified in previous research are addressed in this work by utilizing optimized feature selection by using Grouped Artificial Bee Colony Algorithm (G-ABC) and Artificial Neural Network (ANN) based classification techniques to accurately predict student performance based on various influencing factors derived from the learning data. The proposed framework leverages ANN as an Artificial Intelligence (AI) to train the system using extracted features from student interactions, participation records, and academic outcomes. The method of Educational Data Mining (EDM) is used to process large amounts of data to find hidden pattern that help to make decisions for the proposed framework for academic performance analysis of students in online learning using machine learning approaches. The steps used to extract the data using the EDM technique in proposed model is shown in in Figure 4.1.

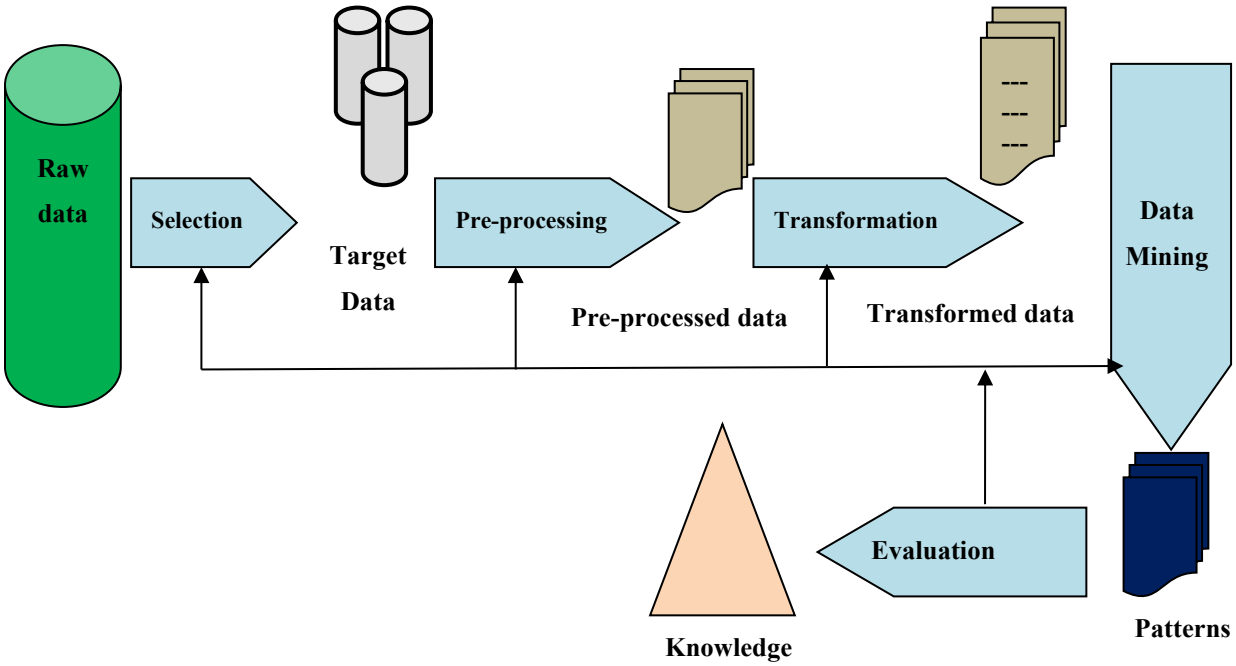

Figure 4.1: EDM-based Model Architecture

The technique may be grasped the following flow diagram of the suggested framework for academic performance analysis of learners undergoing online learning utilizing ANN as a machine/deep learning approaches based on the statically computed characteristics to forecast the student's performance.

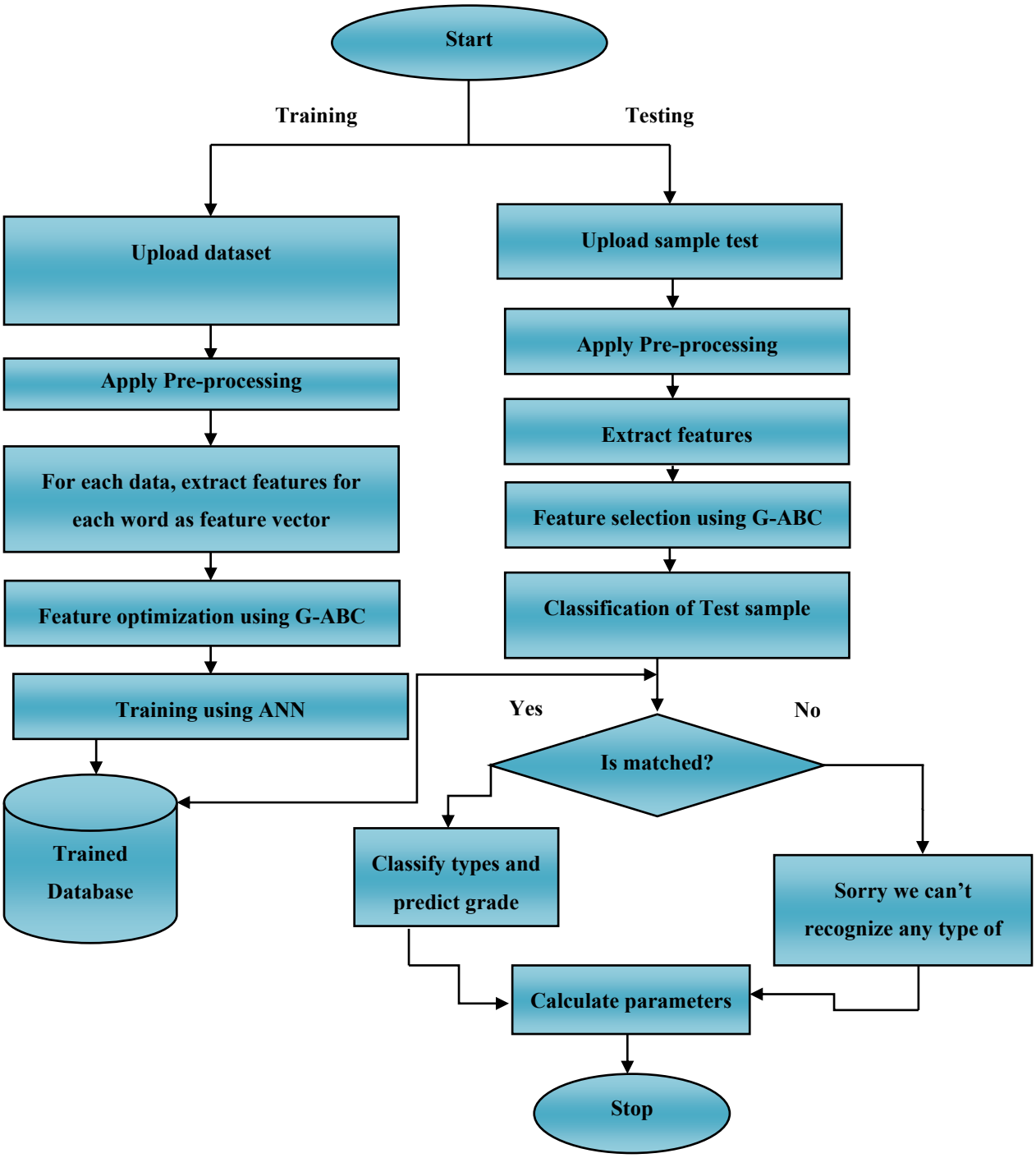

Figure 4.2: Flowchart of Proposed Model

Data Collection: During the data collection process, it is important to find and extract relevant information for modelling. First of all, there are various sources of work-related information, as the data is available in various locations, including spreadsheets, documents and in letter form.

Data Integration: The data collected during this process is combined, that is, the data associated with the same notes are grouped into one category and hence created dictionary of different categories.

Data Selection: During this time, a selection of data that was useful for data generation was done, that is, only important information was selected.

Pre-processing: The main steps in data preparation are pre-processing, which includes: cleaning, transforming and minimizing data.

Data Clearing: Used to erase invalid or unwanted data to improve data quality. The key is to fill in the missing values and also removes outliers.

Data Modification: Used primarily to convert the data format, which is the source document format according to the destination document format.

Data Minimization: For a large number of data sets, it is important to reduce the size of the database before moving to the data phase. It is assumed that large data sets can produce better results than smaller data sets. However, if the database size is minimized that is the data that contains irrelevant information, so that the size of the data can be reduced and it can provide better results in less computational time.

Model Building: This step is a method of collecting data to create a data model, a method of merging data using different approaches such as, clustering, pattern discovery and so forth.

Interpretation: At this stage, the probability of pattern interpretation is revealed.

Decision: Outcome or extracted data is gained at this stage.

Data mining can be easily used in the education field to know the performance of students. To know the student performance, student data is collected from a number of colleges. The information released can then be used to predict student performance. In this way, it will offer teachers an effective teaching approach. In addition, teachers can track student achievement. Students can develop learning activities and also enhance the performance of management

systems. Therefore, the application of data creation technology can be tailored to the specific needs of diverse educational institutes. Based on the above-mentioned block diagram of proposed Model, we design and develop a framework for the simulation using the concept of G-ABC and ANN as machine/deep learning technique and some basic steps of the model are given as:

Step 1. Design a model and upload student performance data for training and testing from the utilized Dataset; the algorithm of data uploading is described as follows:

Algorithm 1: Dataset Uploading

Student Performance Data = Upload with browsing (Excel)
Start the algorithm
N_T = No. of Row and Col in Dataset
For I in range of N_T
    [File, Path] = Browse (Excel File)
    Full Address = String concatenate (Path, File)
    Data [m] = Excel read (Full Address. XLSX)
End – For
Return: Student Performance Data = Data
End – Function

Step 2. Use pre-processing method on the submitted excel information regarding student performance in the training in addition to testing phases and the methodology of pre-processing is expressed as:

Algorithm 2: Data Pre-processing

P-Data = Pre-processing (Data)
Start the algorithm
N_T = No. of Row and Col in Dataset
For x in rage of N_T
    If Data have string binary
        Data [x] = 0 and 1
    Else
        Data [x] = 0 and 1
    End – If
End – For
P- Data = Data
End – For
Return: P-Data as an output
End – Function

Step 3. Following the pre-processing in both phases such as Training and Testing statically feature extraction approach is done to determine the suitable feature sets and the algorithm of feature extraction is expressed as:

Algorithm 3: Statically Feature Extraction

F-Data = Feature Extraction (Pre-processed Data)

1 Start

2 N_T = Number of Row and Col in Pre-processed Data

3 For x in rage of N_T

Mean=mean2(Pre-processed Data)

Standard Deviation=std2(Pre-processed Data)

Min=min (min (Pre-processed Data))

Max=max (max (Pre-processed Data))

Variance= VAR (Pre-processed Data)

Kurtosis=kurtosis (Pre-processed Data)

Skewness=skewness (Pre-processed Data)

Range=range (Pre-processed Data)

4 F-Data = [Mean, Standard Deviation, Min, Max, Variance, Kurtosis, Skewness and Range]

5 End – For

6 End – For

7 Return: F-Data as an output

8 End – Function

Here, the concept of G-ABC is used for feature selection after the feature extraction.

Step 4. Use G-ABC as a feature selection or optimization tool to choose the unique feature according to fitness function of optimization approach. G-ABC's algorithm is stated as follows:

Algorithm 4: G-ABC

S-Data = G-ABC (F-Data) // Selected Data

1 Start the algorithm

2 Initialize parameter – Iterations (ITR)

–Population Size of Bee (S)

– Lower Region Bound (L)

– Upper Region Bound (U)

– No. of Variables (N_{VAR})

3 Compute Size, [N, M] = Size (F-Data)

4 Fitness function, f (fit)

5 $f(fit) = \begin{cases} 1; & \text{if fit} \\ 0; & \text{otherwise} \end{cases}$

6 For x in range of N

7 For y in range of M

8 F_S = F-Data [x, y]

9 F_T = mean (F-Data)

10 Data = G-ABC (ITR, S, CO, M, N_{VAR}, f (fit))

11 End – For

12 S-Data = Data

13 End – For

14 Return: S-Data as an output

15 End – Function

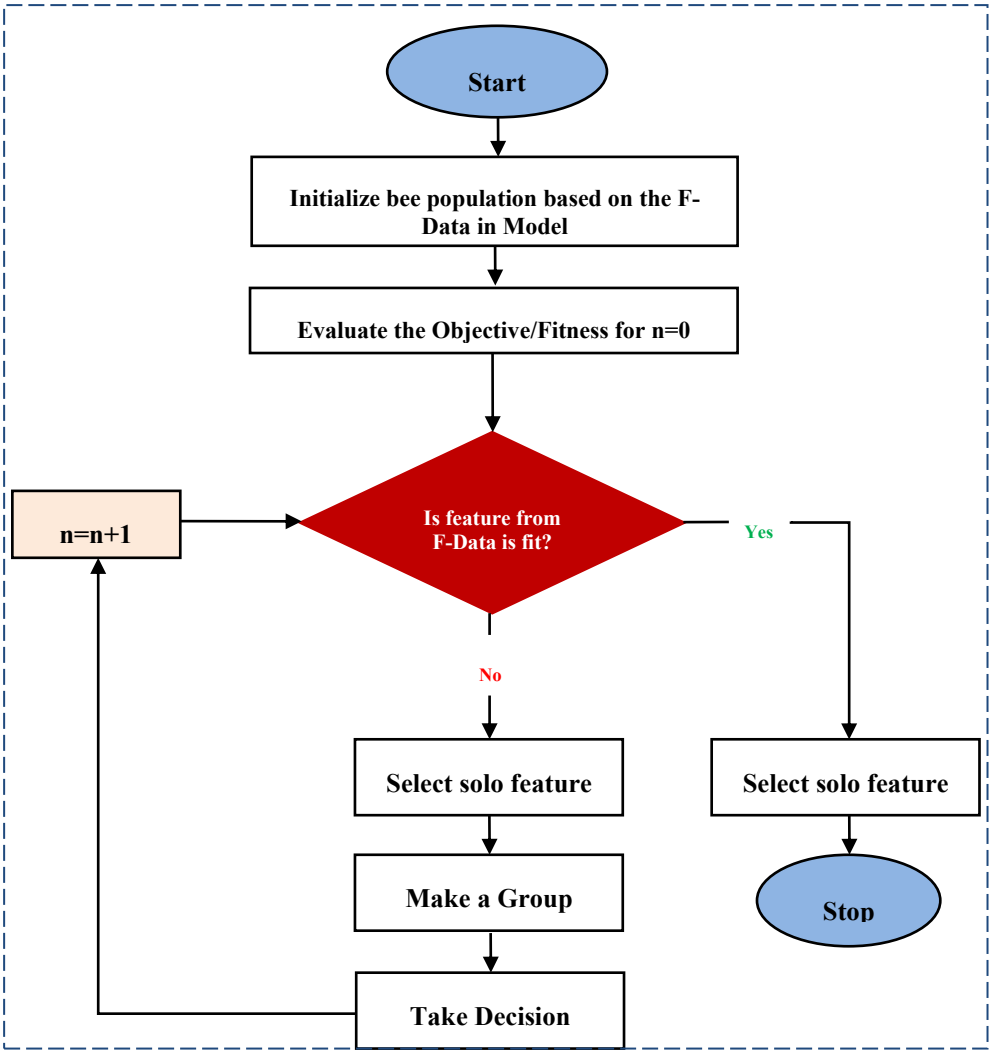

Figure 4.3: Flowchart of G-ABC for Proposed Model

Step 5. Total 8 number of features initially considered and 2 is reduced through the ACO-based feature selection process, highlighting the efficiency gain achieved. Additionally, the values of hyperparameters used in both the ANN and the hybrid G-ACO framework should be clearly reported, since they directly influence convergence, model complexity, and prediction accuracy. The size of the dataset is total of 1001 with 8 features. Following the feature optimization phase, ANN will be started to train the suggested an optimized model with the assistance of optimized chosen features employing G-ABC and the ANN method is expressed as:

Algorithm 5: ANN Architecture

Structure = ANN (S-Data, Types, Neurons)

Start
Initialize ANN with basic evaluation parameters like Epochs (E), Neurons (N), Random Split etc.
[N, M] = Size (S-Data) // Size of the S-Data
For x = 1: N × M
    If S-Data from 1st Class
        G (1) = S-Data (x)
    Else if S-Data from 2nd Class
        G (2) = S-Data (x)
    Else if S-Data from 3rd Class
        G (3) = S-Data (x)
    .
    .
    .
    .
    Else
        G (n) = S-Data (x)
    End – If
End – For
Apply the command of ANN with "S-Data" as a Training data with created G
SFP = newff (S-Data, G, N)
SFP = Train (SFP, S-Data, G)
Return: SFP as a Trained Structure
End – Function

Step 6. Test data is submitted during the testing phase; thereafter, steps 2 to 4 are applied and converted into feature sets. Test data goes for the classification process based on the ANN structure after the last phase of testing procedure (feature optimization). The data kept in the neural trained structure is compared to test data. Table 4.1 displays the sample tweets about the availability and need of medical services.

Table 4.1 Used Dataset Sample

Gender	Race/ethnicity	Parental level of education	Lunch	Test preparation course	Math score	Reading score	Writing score
female	group B	bachelor's degree	standard	none	72	72	74
female	group C	some college	standard	completed	69	90	88
female	group B	master's degree	standard	none	90	95	93
male	group A	associate's degree	free/reduced	none	47	57	44
male	group C	some college	standard	none	76	78	75
female	group B	associate's degree	standard	none	71	83	78
female	group B	some college	standard	completed	88	95	92
male	group B	some college	free/reduced	none	40	43	39
male	group D	high school	free/reduced	completed	64	64	67
female	group B	high school	free/reduced	none	38	60	50
male	group C	associate's degree	standard	none	58	54	52
male	group D	associate's degree	standard	none	40	52	43
female	group C	associate's degree	free/reduced	none	54	58	61
male	group D	high school	standard	none	66	69	63
female	group B	some college	free/reduced	completed	65	75	70
male	group D	some college	standard	none	44	54	53
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
female	group C	bachelor's degree	standard	none	67	69	75

Based on the above-mentioned steps and algorithms, there are three different architecture is designed in Python that are:

1) Framework of Student Performance Prediction
2) Framework of Student Engagement Prediction
3) Framework of Academic Performance Analysis

Framework of Academic Performance Analysis is the final and optimized proposed model which is designed in the Python and the model is shown in the Figure 4.4.

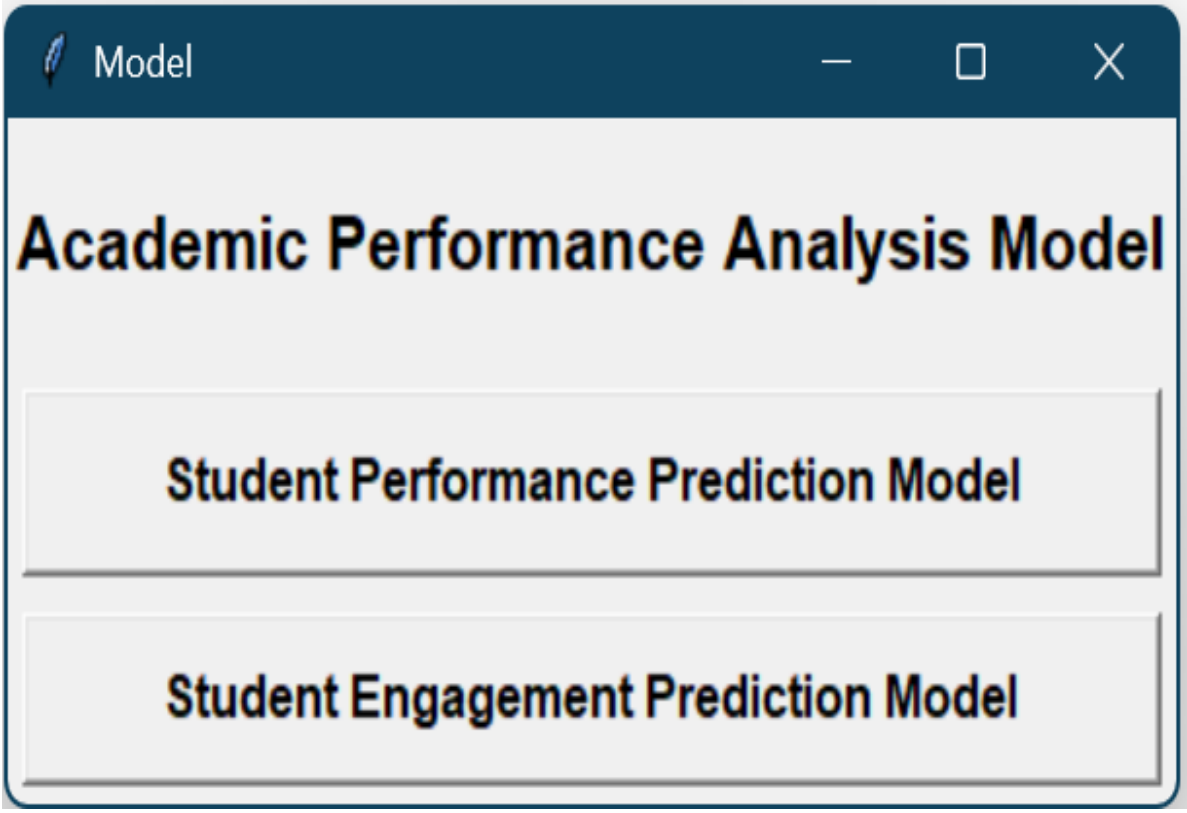

Figure 4.4: Developed Framework of Academic Performance

4.1 FRAMEWORK OF STUDENT PERFORMANCE PREDICTION

The methodology for the Framework of Student Performance Prediction employs a systematic approach that utilizes machine learning techniques to assess and forecast students' academic achievement in an online learning context. The process starts with the collecting and study of data, whereby unprocessed student information is scrutinized to discern patterns and correlations among

many qualities, including demographic factors, parental education, test preparation, and academic performance. Subsequently, the data undergoes preprocessing to address missing values, encode categorical variables, and normalize numerical data, therefore preparing it for machine learning models. Feature engineering methods are utilized to generate pertinent features, such as percentage scores and grades, subsequently followed by feature selection using optimization algorithms to preserve just the most significant predictors. Various machine learning models, encompassing regression and ensemble methodologies, are trained and assessed to identify the optimal performing model. Furthermore, hyperparameter adjustment is performed to refine model parameters and enhance predictive accuracy. The proposed framework incorporates deep learning methodologies, namely ANN, to improve prediction efficacy. The precision and efficacy of the produced models are assessed using several assessment measures, guaranteeing that the final framework accurately forecasts student outcomes and delivers actionable insights for instructors. This methodical methodology guarantees the reliability and strength of the framework in identifying at-risk kids and facilitating prompt interventions. We deliberately focused on a single algorithm, as our primary objective was to optimize for both high accuracy and computational speed. Consequently, ensemble methods were not employed, as they typically introduce additional complexity and runtime overhead that would conflict with this focus.

The used steps are:

Step 1: Import Required Libraries

- Import essential Python libraries such as:
    - pandas, numpy - For data manipulation and numerical operations.
    - matplotlib.pyplot, seaborn - For visualization.
    - scikit-learn modules - For data preprocessing, model selection, evaluation, and regression models.
    - CatBoost and XGBoost - For advanced ensemble models.
    - pickle - For model serialization and saving.

Step 2: Load and Explore Dataset

- **Dataset:** StudentsPerformance.csv containing student records.
- **Operations:**
    - Load the dataset using pd.read_csv().
    - Display initial rows and basic statistics with df.head(), df.describe(), and df.shape().
    - Identify columns and missing values using df.isnull().sum().
    - Check and visualize duplicates and missing values.

Step 3: Data Preprocessing

- **Categorical and Numerical Features:**
    - Separate categorical and numerical columns.
    - Identify unique values in categorical variables like gender, race/ethnicity, parental level of education, lunch, and test preparation course.
- **Feature Engineering:**
    - **Grade Calculation:** Compute Percentage as the average of math score, reading score, and writing score.
    - **Grading System:** Apply the Grade function to assign grades based on percentage.
    - Identify and count students who achieved full marks or scored less than 20 in respective subjects.

Step 4: Data Visualization

- **Outlier Detection:**
    - Visualize outliers using boxplots for math score, reading score, writing score, and Percentage.
- **Univariate Analysis:**

- Gender distribution using pie charts.
- KDE plot to visualize score distribution.
- Parental education comparison with histogram plots.
- Visualization of test preparation course, lunch types, and grade distribution.

- **Bivariate and Multivariate Analysis:**
  - Visualize relationships between Percentage, Grade, and Gender.
  - Explore the effect of Race/Ethnicity, Test Preparation, and Parental Education on student performance.
  - Correlation matrix and heatmap to identify interdependence between attributes.

Step 5: Feature Selection and Transformation

- **Feature Transformation:**
  - Standardize numerical columns using StandardScaler.
  - Encode categorical columns using OneHotEncoder.
  - Use ColumnTransformer to apply transformations.
- **Feature Selection:**
  - **Grouped Artificial Bee Colony (G-ABC):**
    - Apply feature selection based on ABC optimization algorithm.
    - Select important features using a fitness function (FitFun).
  - **Apriori Algorithm (AA):**
    - Apply apriori to derive association rules and extract relationships between features.

Step 6: Model Training and Evaluation

- **Data Splitting:**
    - Split the dataset into training and test sets using train_test_split() with 80-20 ratio.
- **Model Selection and Evaluation:**
    - Evaluate multiple models:
        - Linear Regression, Lasso, K-Neighbors Regressor, Decision Tree, Random Forest, Gradient Boosting, XGBRegressor, CatBoost Regressor, AdaBoost Regressor.
    - **Evaluation Metrics:**
        - mean_absolute_error, mean_squared_error, r2_score.
- **Model Performance:**
    - Train each model and evaluate performance on training and test data.
    - Visualize RMSE, MSE, MAE, and R2 scores.

Step 7: Hyperparameter Tuning

- **Grid Search and Randomized Search:**
    - Perform hyperparameter tuning using GridSearchCV to find optimal model parameters.
    - Evaluate models using tuned parameters and record performance.

Step 8: Model Comparison

- **Result Comparison:**
    - Compare model performance based on R2 scores and select the best-performing model.
- **Accuracy Calculation:**

- Measure the accuracy of the best model using r2_score.

Step 9: Implementation of Proposed Model

- **Neural Network (ANN) with MLPClassifier:**
  - Define and train a Multi-Layer Perceptron (MLP) model.
  - Evaluate the performance of the ANN model.
  - Compute accuracy and compare it with the best-performing ML model.

Step 10: Performance Visualization

- **Comparison of Accuracy:**
  - Bar plot comparing the accuracy of the machine learning (ML) model and neural network (ANN) model.
  - Visualize accuracy differences between the two approaches.

Step 11: Conclusion and Saving the Model

- Save the best-performing model using pickle for future use.
- Summarize the findings with recommendations for future work.

4.2 FRAMEWORK OF STUDENT ENGAGEMENT PREDICTION

The suggested student engagement prediction framework uses machine learning methods to examine several factors affecting student participation and performance in an online learning environment. The main goal is to create and put into place a strong system that forecasts the degree of student involvement depending on demographic, academic, and behavioral data. To obtain high prediction accuracy, this system includes several phases: data collecting, preprocessing, feature engineering, model selection, and assessment. Meticulously processed to handle missing values, detect outliers, and encode categorical variables, the dataset—student records describing academic achievement, demographic information, and participation levels—comprises Then, utilizing optimization strategies, the feature selection procedure guarantees that the most relevant qualities

support model training. Appropriate performance criteria—including accuracy, precision, recall, and F1-score—are then used to train and assess many machine learning models including Decision Trees, Random Forest, Gradient Boosting, and Neural Networks. Ultimately, hyperparameter optimization helps to fine-tune the framework to optimize prediction accuracy, so guaranteeing a thorough and scalable solution for tracking and improving student involvement in virtual learning environments. The employed procedures are:

Step 1: Import Required Libraries

The initial step involves importing all the necessary Python libraries to implement the model.

- numpy and pandas – For data manipulation and numerical operations.
- train_test_split – For splitting the dataset into training and testing sets.
- LabelEncoder – To convert categorical variables into numerical form.
- mean_squared_error and r2_score – For evaluating the model's performance.
- LinearRegression, Lasso – For linear regression and Lasso regression models.
- KNeighborsRegressor – To apply K-Nearest Neighbors (KNN) for regression.
- DecisionTreeRegressor, RandomForestRegressor – To apply decision trees and random forests.
- GradientBoostingRegressor, AdaBoostRegressor – To apply ensemble learning models.
- XGBRegressor – For extreme gradient boosting.
- CatBoostRegressor – For applying CatBoost regression.
- MLPClassifier – For training an artificial neural network (ANN).

Step 2: Load Dataset

- **Dataset Name:** StudentsPerformance.csv
- The dataset is loaded using pandas and stored in a DataFrame (df):

- The dataset contains information about student performance in different subjects along with demographic attributes.

Step 3: Feature Engineering

- **Objective:** Create a new feature average_score to represent the overall engagement/performance of the students.
- **Method:** The average score is computed by calculating the mean of math score, reading score, and writing score for each student.

Step 4: Encoding Categorical Variables

- **Objective:** Convert categorical features to numerical values using LabelEncoder.
- **Categorical Features:**
    - gender
    - race/ethnicity
    - parental level of education
    - lunch
    - test preparation course
- LabelEncoder assigns numerical labels to these categorical features, making them suitable for machine learning algorithms.

Step 5: Define Features and Target Variable

- **Features (X):** Encoded categorical variables.
- **Target (y):** average_score created in the previous step.

Step 6: Split the Dataset

- **Objective:** Split the dataset into training and testing sets for model evaluation.

- **Split Ratio:** 70% training and 30% testing.
- **random_state=42** ensures that the data split remains consistent across multiple runs.

Step 7: Define Regression Models

- A dictionary of multiple regression models is defined to compare performance.
- **Regression Models Used:**
    - Linear Regression
    - Lasso Regression
    - K-Neighbors Regressor
    - Decision Tree Regressor
    - Random Forest Regressor
    - Gradient Boosting Regressor
    - XGBoost Regressor
    - CatBoost Regressor
    - AdaBoost Regressor

Step 8: Model Training, Prediction, and Evaluation

- **Objective:** Train and evaluate each model on the training and testing sets.
- **Evaluation Metrics:**
    - **Mean Squared Error (MSE):** Measures the average of the squared differences between predicted and actual values.
    - **R^2 Score (Coefficient of Determination):** Indicates how well the model explains the variance in the target variable.
- **Process:**

1. Model is trained using model.fit(X_train, y_train).
2. Predictions are made on the test set using model.predict(X_test).
3. MSE and R^2 scores are computed and displayed for each model.

Step 9: Neural Network Model for Classification

- **Objective:** Train an Artificial Neural Network (ANN) to classify student performance.
- The ANN is used to make predictions and estimate the probability for each class, and the final score (NNScore) is calculated based on prediction probability.

Step 10: Evaluation and Comparison

- Results from multiple models and the neural network are evaluated and compared to identify the best-performing algorithm.
- The final performance metrics help in determining which model provides the highest accuracy in predicting student performance in online learning environments.

4.3 FRAMEWORK OF ACADEMIC PERFORMANCE ANALYSIS

The Framework of Academic Performance Analysis is a comprehensive model that integrates the Framework of Student Performance Prediction and the Framework of Student Engagement Prediction to provide a holistic view of student outcomes in online learning environments. The Student Performance Prediction Framework focuses on predicting academic success by analyzing key factors such as demographic data, parental education, prior academic performance, and learning behavior. It leverages machine learning algorithms to predict grades and identify at-risk students who may require intervention. On the other hand, the Student Engagement Prediction Framework assesses student involvement and interaction within the learning environment by analyzing participation data, completion rates, time spent on tasks, and test preparation activities. Engagement levels serve as an essential indicator of motivation and understanding, contributing

significantly to overall academic performance. By combining these two frameworks, the proposed model not only predicts student outcomes but also identifies patterns of engagement that influence academic success. The integration of these frameworks enhances the predictive accuracy and provides educators with actionable insights, enabling them to implement personalized interventions, improve course design, and optimize learning strategies for better academic outcomes. The steps in the framework of academic performance analysis are written as:

Step 1: Data Collection and Preprocessing

- **Objective:** Gather and clean data from diverse sources such as learning management systems (LMS), student information systems, and academic databases.
- **Data Sources Include:**
    - Demographic information
    - Parental education level
    - Online interaction records
    - Test scores and assignment completion
- **Preprocessing:** Handle missing values, encode categorical variables, normalize numerical data, and split data into training and testing sets.

Step 2: Feature Engineering and Selection

- **Objective:** Extract and select relevant features that impact student performance and engagement.
- **Key Features:**
    - **Performance Indicators:** Math, reading, and writing scores, assignment grades.
    - **Engagement Indicators:** Participation records, time spent on tasks, interaction frequency.
- **Techniques Used:** Statistical analysis, correlation analysis, and dimensionality reduction to identify the most significant features.

Step 3: Framework of Student Performance Prediction

- **Objective:** Develop models to predict student performance based on academic and demographic attributes.
- **Methods Used:**
    - Regression models (Linear Regression, Lasso)
    - Decision Trees and Random Forests
    - Gradient Boosting and XGBoost
- **Model Training:** Train models using historical data to predict academic performance and identify at-risk students.
- **Model Evaluation:** Use metrics such as Mean Squared Error (MSE) and R^2 score to evaluate model performance.

Step 4: Framework of Student Engagement Prediction

- **Objective:** Analyse and predict student engagement levels using behavioural data from online learning environments.
- **Methods Used:**
    - Classification models (SVM, Random Forest Classifier)
    - Neural Networks (MLP Classifier)
    - Time-series analysis to detect patterns of engagement over time based on the same dataset in different time frame.
- **Model Evaluation:** Assess the accuracy of engagement prediction using classification metrics such as Accuracy, Precision, and Recall.

Step 5: Model Integration and Hybrid Framework Development

- **Objective:** Combine the performance prediction and engagement prediction models to create a hybrid framework.
- **Integration Process:**
    - Aggregate predictions from both models.
    - Identify correlations between engagement patterns and performance outcomes.

- Develop an optimized prediction system that uses the combined results to improve accuracy.

Step 6: Model Testing and Validation

- **Objective:** Test the integrated framework on unseen data to ensure robustness and reliability.
- **Evaluation Metrics:**
  - Accuracy, Precision, Recall, and F1-score for classification models.
  - MSE and R^2 for regression models.
- **Cross-validation:** Use k-fold cross-validation to validate model consistency.

Above mentioned methodology is used for the model development and in the next section of thesis after experimental setup, the simulation results of the model is discussed in details.

CHAPTER 5

EXPERIMENTAL SETUP

This section provides a detailed description of the experimental setup for the proposed framework — a machine learning-based model for the academic performance analysis of students in online learning environments. The framework integrates various machine learning algorithms, including classification, clustering, and regression models, to predict student performance and identify at-risk learners. The experiments are conducted to evaluate the effectiveness of the proposed framework using a combination of traditional feature engineering methods and metaheuristic optimization techniques for improved model accuracy and generalizability. The performance of the models is assessed using multiple evaluation metrics to ensure the robustness and efficiency of the system in predicting academic outcomes in diverse learning environments.

5.1 INFORMATIONABOUT USED TOOLS

To implement the proposed framework using Python, it is essential to ensure that the system meets the minimum hardware and software requirements. First and foremost, the computer's CPU should be at least Intel Core i3 or higher, with a minimum of 500 GB HDD or SSD storage and 8 GB of RAM or greater to handle large datasets and ensure smooth execution of machine learning models. The operating system should be Windows 10 or higher (64-bit recommended), although the framework can also be run on Linux or macOS systems. For software, it is necessary to have Python 3.11 or higher installed along with essential libraries and packages, including but not limited to:

- NumPy for numerical operations and matrix handling
- Pandas for data manipulation and analysis
- Scikit-learn for implementing machine learning algorithms
- Matplotlib and Seaborn for data visualization
- TensorFlow/Keras or PyTorch for deep learning models (if required)

To ensure the smooth execution of the Python environment, it is recommended to use Spyder or Visual Studio Code (VS Code) as the Integrated Development Environment (IDE) for coding, testing, and debugging. Additionally, a keyboard and mouse are required for an efficient and user-friendly coding experience. Python, being an open-source programming language, is highly preferred by researchers for academic performance analysis due to its flexibility, vast library support, and compatibility with various operating systems, including Windows, Linux, and macOS. A summary of these requirements is also provided in Table 5.1 to offer a quick reference.

Table 5.1 Experimental Setup for Proposed Model Simulation

Component	Description	Tool/Library
Operating System	The environment where the simulation will be executed.	Windows 10 0r 11, Linux, or macOS
Processor	The required processing power for computation.	Intel Core i3 or higher
RAM	Memory required for efficient computation and simulation.	8 GB or higher
Other Hardware	Required hardware to simulate the model.	Keyboard, Mouse
Environment	Platform used for implementing the methodology.	Spyder (Python-3.11 or higher)
IDE	Integrated Development Environment for coding and debugging.	Spyder Editor, Tools
Numerical Computation	Toolbox for numerical calculations and array manipulation.	Python base package, Symbolic Math Toolbox
Data Manipulation	Toolbox for handling and manipulating data.	Python base package, Data Import and Export
Visualization	Toolbox for data visualization and plotting.	Python base package, Graphics
Machine Learning	Toolbox for implementing and training machine learning models.	Statistics and Machine or Deep Learning Toolbox

Optimization	Toolboxes for optimizing models and solutions.	Optimization Toolbox, Global Optimization Toolbox

This table delineates the main components and libraries/toolboxes used in the Python program for executing many tasks, including data processing, numerical calculation, machine learning, and visualisation.

5.2 LANGUAGE USED FOR IMPLEMENTATION

The proposed research study is implemented using Spyder-Python 3.11 or a later version, which is a high-level, versatile programming language widely used for machine learning, data analysis, and scientific computing. Python provides a robust and flexible environment where complex computations, data preprocessing, and model training can be performed efficiently. The implementation of the proposed framework is primarily conducted using Python, a programming language known for its rich ecosystem of libraries and frameworks that support a wide range of machine learning and data analysis tasks. Python's extensive libraries and modules make it an ideal choice for building and deploying machine learning models for academic performance analysis. The primary reasons for selecting Python as the implementation platform include:

1. Development Efficiency: Python's concise and readable syntax simplifies the development, testing, and evaluation of machine learning models, allowing for rapid prototyping and easy debugging.

2. Availability of Libraries: Python offers a wide range of powerful libraries and frameworks such as:

NumPy and SciPy: For numerical computations and matrix operations.
Pandas: For data manipulation and analysis.
Scikit-learn: For implementing machine learning models and performing data preprocessing.
TensorFlow and Keras: For building and training deep learning models.
Matplotlib and Seaborn: For data visualization and graphical representation of results.
Optuna and DEAP: For hyperparameter tuning and optimization using metaheuristic approaches.

3. Visualization Capabilities: Python's visualization libraries, such as Matplotlib, Seaborn, and Plotly, provide powerful tools for creating insightful visualizations that help in understanding model performance and behaviour.

4. Scalability and Flexibility: Python is highly scalable, making it suitable for processing large datasets and building models that can handle complex learning environments. Its flexibility allows seamless integration with other platforms and APIs.

5. Support for Machine Learning and Deep Learning Models: Python's extensive machine learning and deep learning ecosystem makes it a preferred choice for developing models that require high computational efficiency and predictive accuracy. Libraries like TensorFlow, Keras, and PyTorch enable the construction of deep learning models, while Scikit-learn facilitates the implementation of traditional machine learning algorithms.

6. Community Support and Open-Source Ecosystem: Python has a large and active community that contributes to its rich open-source ecosystem, providing continuous improvements, updates, and a wide variety of resources for solving diverse computational challenges.

5.3 USED DATASET

For the training as well as testing of the proposed model, Performance-Analysis-and-Prediction Dataset is used [111] having a link https://www.kaggle.com/datasets/rkiattisak/student-performance-in-mathematics and the Table 5.2 represents the brief information about dataset.

Table 5.2 Dataset for Simulation of Proposed Model

Column Name	Description
gender	Gender of the student (male / female)
race/ethnicity	Ethnic group of the student (group A, group B, group C, etc.)
parental level of education	Highest education level of the parents (some high school, associate's degree, etc.)

lunch	Type of lunch received by the student (standard / free/reduced)
test preparation course	Completion status of the test preparation course (completed / none)
math score	Score achieved by the student in the mathematics test (0-100)
reading score	Score achieved by the student in the reading test (0-100)
writing score	Score achieved by the student in the writing test (0-100)

Total Records: The dataset contains multiple records where each row represents a student's performance and demographic profile.

Primary Purpose: This dataset is used to analyze the relationship between student performance and various demographic, socio-economic, and academic factors to predict student outcomes and identify at-risk learners.

5.3.1 Work Done

- Completed courses from “Coursera” that are required for the basic understanding
- Study about Machine and Deep Learning
- Analysis of the available datasets
- Literature review
- Implementation
- Thesis Report

CHAPTER 6

RESULTS & DISCUSSIONS

Using the suggested paradigm, this part addresses the study of educational achievement in online learning settings employing sophisticated machine learning methods. Using a thorough collection of performance metrics including Root Mean Squared Error (RMSE), Mean Squared Error (MSE), Mean Absolute Error (MAE), and R2 Score—the study methodically assesses the efficacy of the machine learning techniques. Emphasizing the robustness and effectiveness of the used framework in predicting and analysing students' academic achievements for the publically available dataset "Performance-Analysis-and-Prediction Dataset", this section offers a thorough investigation of the results obtained. The results show the capacity of machine learning algorithms to locate important elements affecting academic achievement while guaranteeing scalability and adaptation to different datasets and learning environments.

6.1 PERFORMANCE MEASUREMENT PARAMETERS

The evaluation or performance parameters are described below with proper definition and their formulas.

6.1.1 Root Mean Squared Error (RMSE)

RMSE is the square root of the mean of the squared discrepancies between expected and actual values. It shows how much inaccuracy the model produces in prediction on average and punishes significant errors more strongly.

$$\mathbf{RMSE} = \sqrt{\frac{\mathbf{1}}{\mathbf{n}}\sum_{\mathbf{i=1}}^{\mathbf{n}}(\mathbf{A_i} - \mathbf{P_i})^{\mathbf{2}}}$$

Where, n→ Number of data points

A_i→ Actual data value for the i-th observation

P_i→ Predicted data value for the i-th observation

Important Interpretation about RMSE for Proposed Model:

- If RMSE of model is lower, then having better model performance.
- Same unit as the target variable (e.g., if target is in cm, RMSE is in cm).
- Sensitive to outliers due to squaring the error and try to attract the best fit line.

6.1.2 Mean Squared Error (MSE)

MSE is the average of the squared differences between predicted and actual values. It is the squared version of RMSE and is used to compare models' prediction accuracy.

$$\mathbf{MSE} = \frac{\mathbf{1}}{\mathbf{n}}\sum_{\mathbf{i=1}}^{\mathbf{n}}(\mathbf{A_i} - \mathbf{P_i})^{\mathbf{2}}$$

Where, n→ Number of data points

A_i→ Actual data value for the i-th observation

P_i→ Predicted data value for the i-th observation

Important Interpretation about MSE for Proposed Model:

- If MSE of model is lower, then having better model performance.
- MSE has the unit of the square of the target variable (e.g., cm^2).
- Heavily penalizes larger errors.

6.1.3 Mean Absolute Error (MAE)

MAE measures the average magnitude of the errors in a set of predictions, without considering their direction (i.e., no squaring).

$$\mathbf{MAE} = \frac{\mathbf{1}}{\mathbf{n}}\sum_{\mathbf{i=1}}^{\mathbf{n}}|\mathbf{A_i} - \mathbf{P_i}|$$

Where, n→ Number of data points

A_i→ Actual data value for the i-th observation

P_i→ Predicted data value for the i-th observation

Important Interpretation about MAE for Proposed Model:

- If MAE of model is lower, then having better model performance.
- Has the same unit as the target variable.
- Less sensitive to outliers than MSE or RMSE.

6.1.4 Accuracy

It is defined as the sentiments classified correctly with respect to the entire available classified sentiments.

$$\textbf{Accuracy Rate} = \frac{\textbf{Number of Correct Predictions}}{\textbf{Total Number of Predictions}}$$

6.1.5 Error

It is the reverse of accuracy and calculated using given formula

$$\mathbf{100 - Accuracy = Error\ rate}$$

6.1.6 R^2 Score (Coefficient of Determination)

R^2 indicates the proportion of the variance in the dependent variable that is predictable from the independent variables. It shows how well the regression model fits the data.

$$\mathbf{R^2 = 1 - \frac{\sum_{i=1}^{n}(A_i - P_i)^2}{\sum_{i=1}^{n}(A_i - P_m)^2} = 1 - \frac{RSS}{TSS}}$$

Where, n→ Number of data points

A_i→ Actual data value for the i-th observation

P_i→ Predicted data value for the i-th observation

P_m→ Mean of actual values

RSS → The numerator is the residual sum of squares

TSS → The denominator is the total sum of squares

Important Interpretation about R^2 Score for Proposed Model:

- $R^2 = 1 \rightarrow$ Perfect prediction
- $R^2 = 0 \rightarrow$ Model does no better than the mean
- $R^2 < 0 \rightarrow$ Model performs worse than a constant mean predictor

Based on the above mention evaluation parameters of the proposed work, Table 6.1 represents the summary information about all.

Table 6.1 Evaluation Parameters for Proposed Academic Performance Analysis Model

Metric	Penalizes Large Errors	Range	Unit	Sensitive to Outliers
RMSE	✅ Yes	[0, ∞)	Same as target	✅ High
MSE	✅ Yes	[0, ∞)	Squared unit	✅ High
MAE	❌ No	[0, ∞)	Same as target	⊘ Less
Accuracy	❌ No	(0, 100]	%	⊘ Moderate
Error	❌ No	(0, 100]	%	⊘ Moderate
R^2	❌ No	(-∞, 1]	Unitless	⊘ Moderate

This table elucidates the calculation and evaluation of several performance measures within the realm of proposed model using the concept of machine or deep learning for academic performance analysis of students in online learning. In the proposed Academic Performance Analysis model, there are two models used that is explained in below section of thesis.

6.2 RESULTS OF STUDENT PERFORMANCE PREDICTION

This section presents the experimental outcomes of the student performance prediction model developed using various machine learning algorithms. The model was trained and tested on a cleaned and pre-processed dataset that included both numerical and encoded categorical features such as gender, parental level of education, lunch type, and test preparation course, along with students' scores in mathematics, reading, and writing. The primary target variable was the average

score, representing overall academic performance. Several regression algorithms- including Linear Regression, Decision Tree, Random Forest, Gradient Boosting, XGBoost, and CatBoost- were evaluated using key performance metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R^2 Score. The results demonstrate how each algorithm performs in predicting student academic outcomes, allowing for a comparative analysis of accuracy and robustness across models and the developed model is shown in the Figure 6.1.

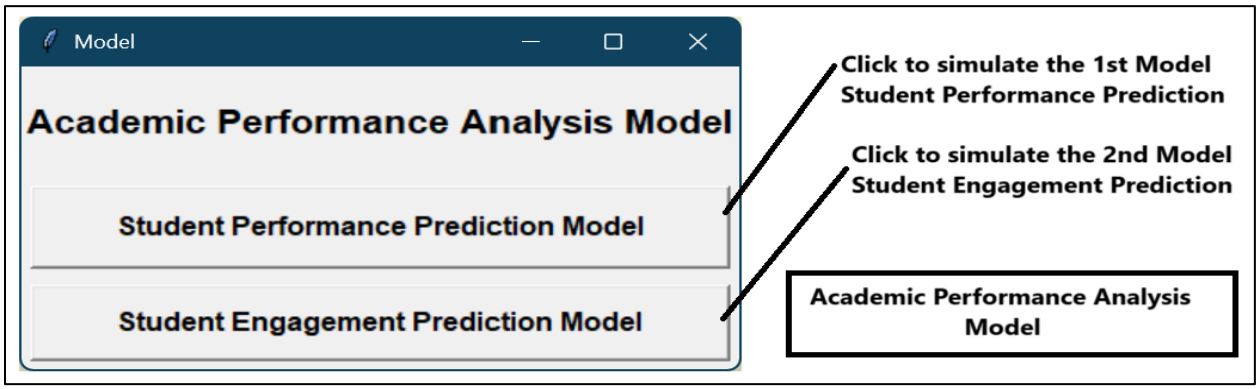

Figure 6.1: Proposed Framework for Academic Performance Analysis

The displayed figure is a Graphical User Interface (GUI) window created for executing and managing different machine learning models under the umbrella of an Academic Performance Analysis System. This GUI serves as an interactive control panel that allows users to choose between two predictive models via button-based interaction. Firstly, Student Performance Prediction model is simulated and the obtained results are discussed with the help of Figure 6.2.

	gender	race/ethnicity	...	reading score	writing score
0	female	group B	...	72	74
1	female	group C	...	90	88
2	female	group B	...	95	93
3	male	group A	...	57	44
4	male	group C	...	78	75
5	female	group B	...	83	78
6	female	group B	...	95	92
7	male	group B	...	43	39
8	male	group D	...	64	67
9	female	group B	...	60	50

Figure 6.2: Uploaded Dataset Description

After the dataset uploading, some basic exploratory data analysis is applied to check the missing value, duplicate values etc. and if any missing values are available then, such things are handled in the proposed work. Then, Statistical Description is calculated for the uploaded dataset that is shown in the Figure 6.3.

	math score	reading score	writing score
count	1000.00000	1000.000000	1000.000000
mean	66.08900	69.169000	68.054000
std	15.16308	14.600192	15.195657
min	0.00000	17.000000	10.000000
25%	57.00000	59.000000	57.750000
50%	66.00000	70.000000	69.000000
75%	77.00000	79.000000	79.000000
max	100.00000	100.000000	100.000000

Figure 6.3: Dataset Statistical Description

Dataset Statistical Description is an essential initial step in any data analysis or machine learning project. It involves summarizing and understanding the dataset using key descriptive statistics such as mean, median, standard deviation, min, max, quartiles, etc. Key reasons for using statistical description of a dataset:

1. Understanding the Distribution of Data: It helps in identifying how data is spread across the dataset. Measures like mean, median, and standard deviation tell us the central tendency and variability of features. It helps to determine normality, skewness, or outliers.

2. Detecting Data Quality Issues: Helps in identifying missing values, duplicate entries, or improper formats. Summary statistics may reveal anomalies (e.g., extremely high values for test scores that should be out of 100).

3. Feature Importance and Selection: By analysing the variance and range of features, we can identify which features are potentially useful for prediction and which ones might be redundant or constant.

4. Outlier Detection: Descriptive statistics (e.g., IQR and standard deviation) help in identifying outliers that could distort model performance.

5. Baseline for Data Normalization or Scaling: Knowing the min and max or mean and std helps apply normalization (Min-Max scaling) or standardization (Z-score scaling) before feeding into ML algorithms.

6. Comparison Across Classes: If the dataset includes labelled categories (like student performance groups), descriptive stats allow comparing statistical summaries across different target classes.

7. Model Selection and Assumption Checking: Some models (like Linear Regression) assume that data is normally distributed. Descriptive statistics help in validating those assumptions.

After the statistical description analysis of a dataset, some important points are concluded like

- ☞ All value of means is fairly similar to one another attributes and it is falling between the range of 66 and 69.
- ☞ The range of all standard deviations is lies between 14.6 and 15.19 and it is also narrow.
- ☞ While there is a minimum score of 0 for given attribute math, the minimums for writing and reading attribute are substantially higher at 10 and 17, respectively.

In next step, with the help of some code check the number of unique values in each column that is shown in the Figure 6.4.

```
gender                         2
race/ethnicity                 5
parental level of education    6
lunch                          2
test preparation course        2
math score                    81
reading score                 72
writing score                 77
dtype: int64
```

Figure 6.4: Unique Available Values in Dataset

After the basic steps, data preprocessing is used for the separating the numerical and categorical features and count the feature values and then major analysis are done like marks analysis, outlier analysis etc. The Figure 6.5 represented the outlier analysis of the proposed model based on the math, reading, writing and percentage score.

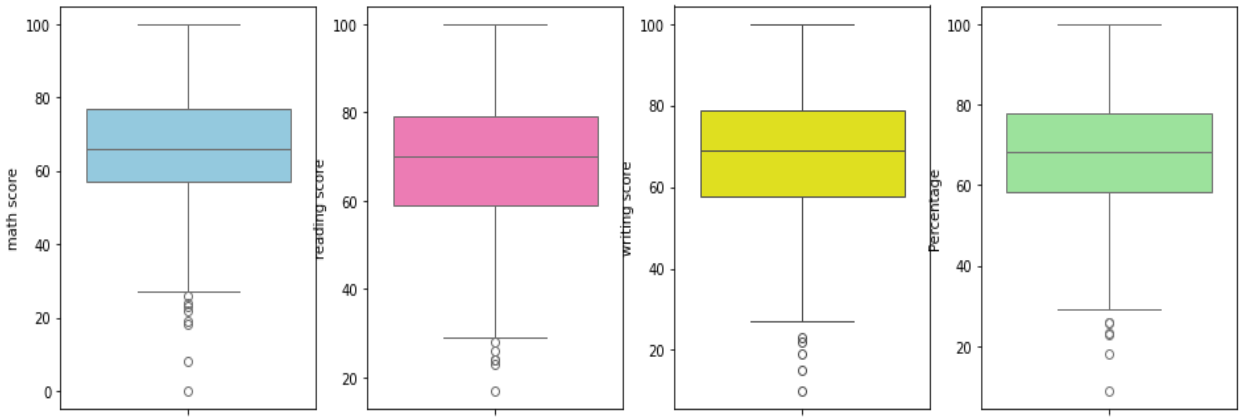

Figure 6.5: Outlier Analysis of Dataset

Figure 6.5 consists of four different box plots, each representing the distribution of student performance in Math Score, Reading Score, Writing Score, and Overall Percentage. These box plots provide insights into the central tendency, spread, and presence of outliers for each metric.

Math Score (Blue Color Graph): The median score lies around the mid-60s, with an interquartile range (IQR) roughly between 57 and 77. Several outliers are observed below the lower whisker, with a minimum value near 0, indicating a few students performed very poorly.

Reading Score (Pink Color Graph): This distribution appears more symmetric, with a median around 70. The IQR lies between approximately 62 and 80. Fewer outliers exist compared to math, though some low-performing students are still present.

Writing Score (Yellow Color Graph): This has a similar shape to the reading score distribution but appears slightly skewed. The median is around 72, and the whiskers extend further, indicating a wider performance range. Outliers are again visible at the lower end.

Percentage (Green Color Graph): Calculated as the average of the three subject scores, the percentage distribution is more compact. The median percentage is around 68–70, with most values within the 60–80 range. The overall distribution reflects the composite nature of the other scores and consolidates their variability.

After that, univariate analysis is performed over the used dataset to find out the Measures of Central Tendency, Dispersion, Distribution and Outlier Detection. The Figure 6.6 show the gender distribution pie chart where Figure 6.7 shows the score distribution in terms of KDE Plot.

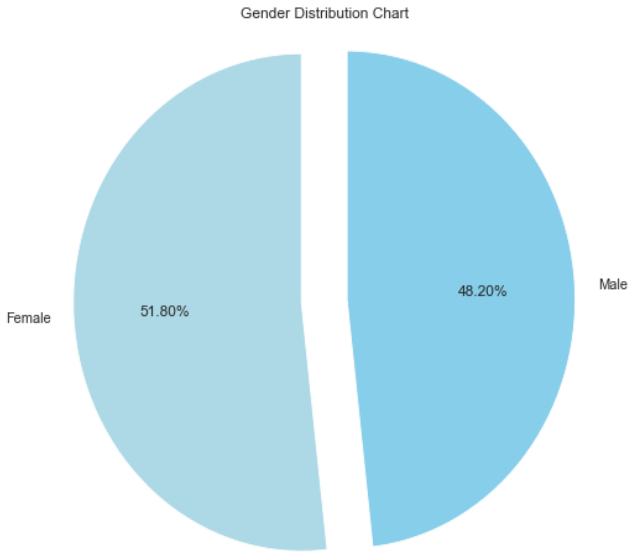

Figure 6.6: Gender Distribution Chart

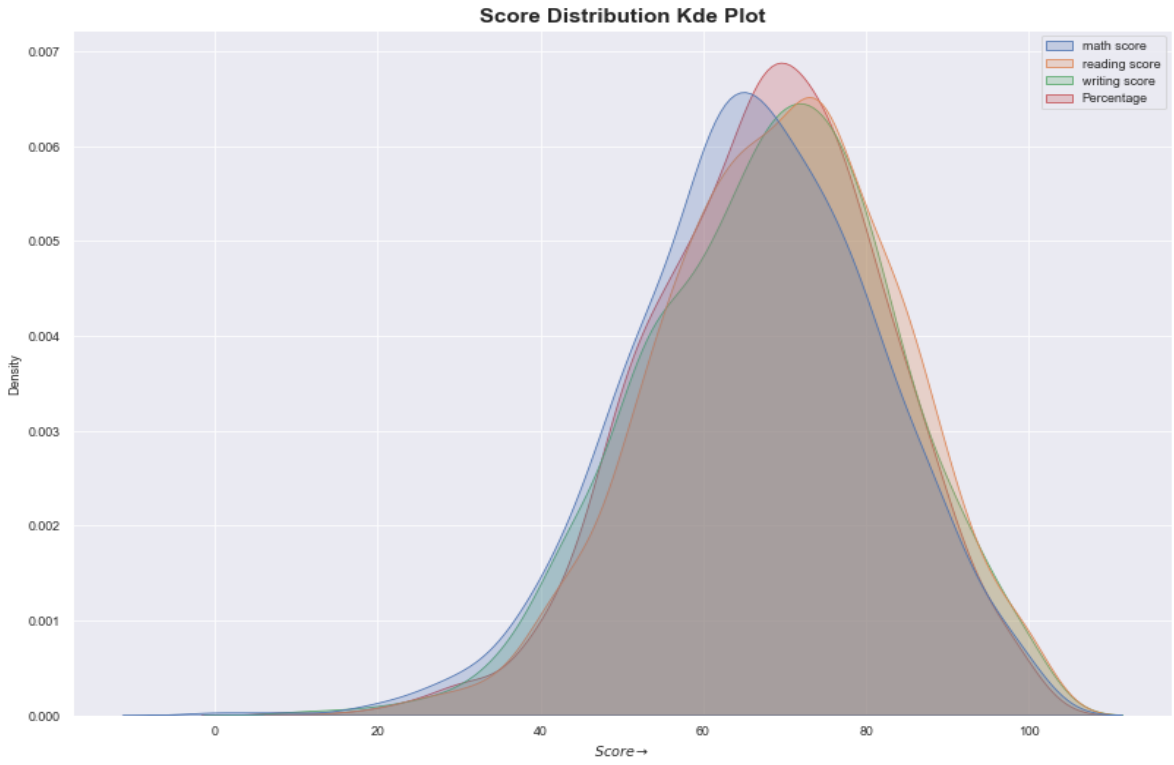

Figure 6.7: Score Distribution KDE Plot

Figure 6.6 represents a pie chart illustrating the univariate analysis of the gender distribution within a dataset, likely related to students. The chart shows two categories: Male and Female, each occupying a proportion of the circle based on their frequency. The Female group constitutes 51.80% of the dataset, while the Male group accounts for 48.20%. This visualization highlights a near-balanced distribution, with a slightly higher proportion of females than males. A visual separation (exploded slice) between the two halves improves readability and emphasizes the comparison. This type of univariate categorical analysis is crucial in understanding population characteristics, ensuring data balance, and evaluating potential bias in predictive modeling. A skewed gender distribution could affect the generalizability of models in educational performance analysis. Since gender might influence academic performance or engagement differently, such insights are valuable for feature impact assessment in machine learning models. Where Figure 6.7 depicts a Kernel Density Estimation (KDE) plot representing the probability density functions of four continuous variables: math score, reading score, writing score, and percentage. KDE is a smoothed version of the histogram that estimates the probability distribution of a continuous variable, helping to visualize the underlying distribution without being influenced by bin width as in histograms. Each coloured curve corresponds to a specific score:

1. Math Score (Blue Color Graph) shows a slightly left-shifted peak, indicating that math scores tend to be lower compared to other subjects.
2. Reading Score (Red Color Graph) and Writing Score (Green) have similar distributions, peaking slightly to the right of math, reflecting better performance in those areas.
3. Percentage (Orange Color Graph) computed as the average of the three scores—displays a balanced curve, summarizing the overall trend in student performance.

All four curves demonstrate a near-normal (bell-shaped) distribution with peaks between 60 and 80, suggesting that most students scored within this range. The symmetric shape and minimal skewness indicate consistency in student performance across subjects. This visualization is crucial in univariate analysis for understanding score trends, detecting performance gaps, and validating assumptions for downstream statistical modeling (e.g., normality for linear regression). The Figure 6.8 illustrates a bar plot that visualizes the distribution of parental education levels among students. The x-axis represents the degree level of the parents (categorical variable), while the y-axis shows the frequency count of students whose parents have attained each level of education.

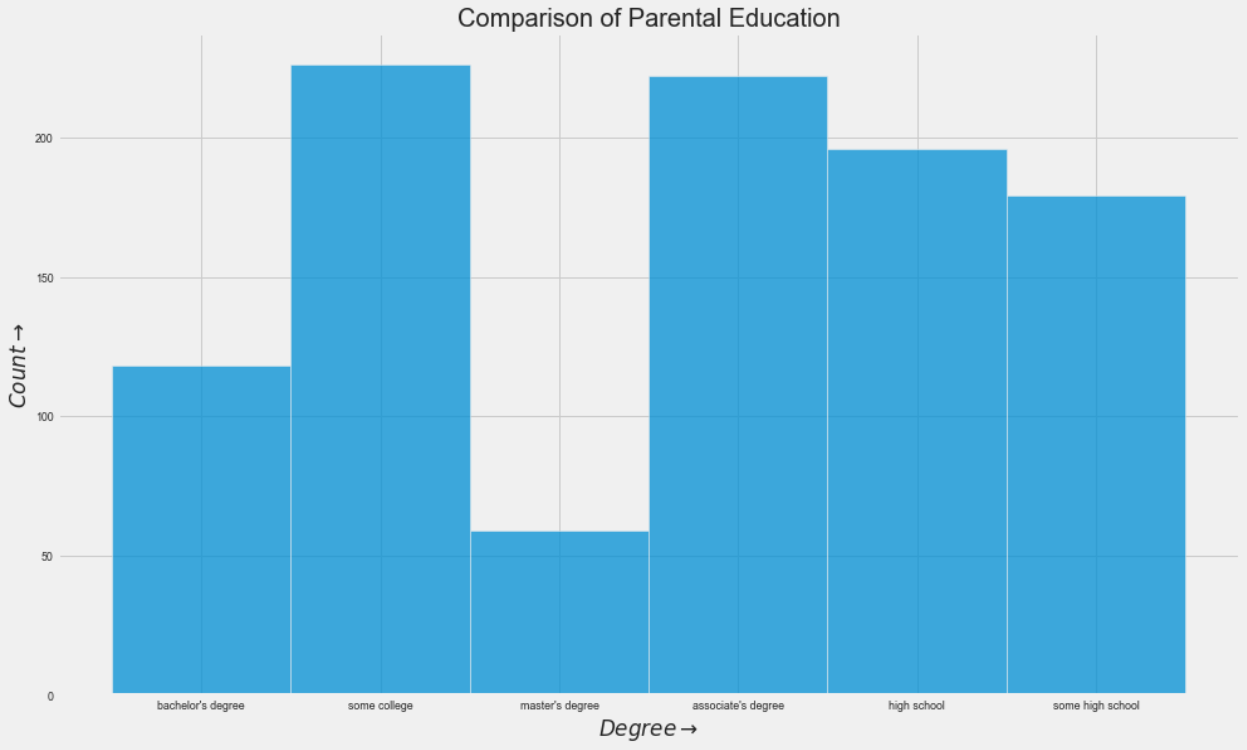

Figure 6.8: Parental Level of Education

According to the Figure 6.8, the most common education levels are "some college" and "associate's degree", both with frequencies exceeding 220 students, indicating that many students come from families with post-secondary but non-graduate educational backgrounds. "High school" and "some high school" also show significant representation, suggesting a large portion of the student population comes from households without college-educated parents. "Bachelor's degree" is less common (~120 students), while "master's degree" is the least represented category, with fewer than 60 students. This analysis is crucial in understanding the socio-educational background of the student dataset. The distribution skew toward non-graduate education levels can impact student performance patterns and may be used as a predictor variable in machine learning models for academic performance or engagement prediction. It also helps in identifying educational equity gaps and targeting interventions where parental academic support may be limited. The displayed bar plot in Figure 6.9 provides a categorical univariate analysis of the race/ethnicity distribution among students in the dataset. The x-axis represents five race/ethnicity groups labelled as Group A, B, C, D, and E, while the y-axis shows the count (frequency) of students belonging to each group. Each bar's height reflects the number of students in that particular demographic category.

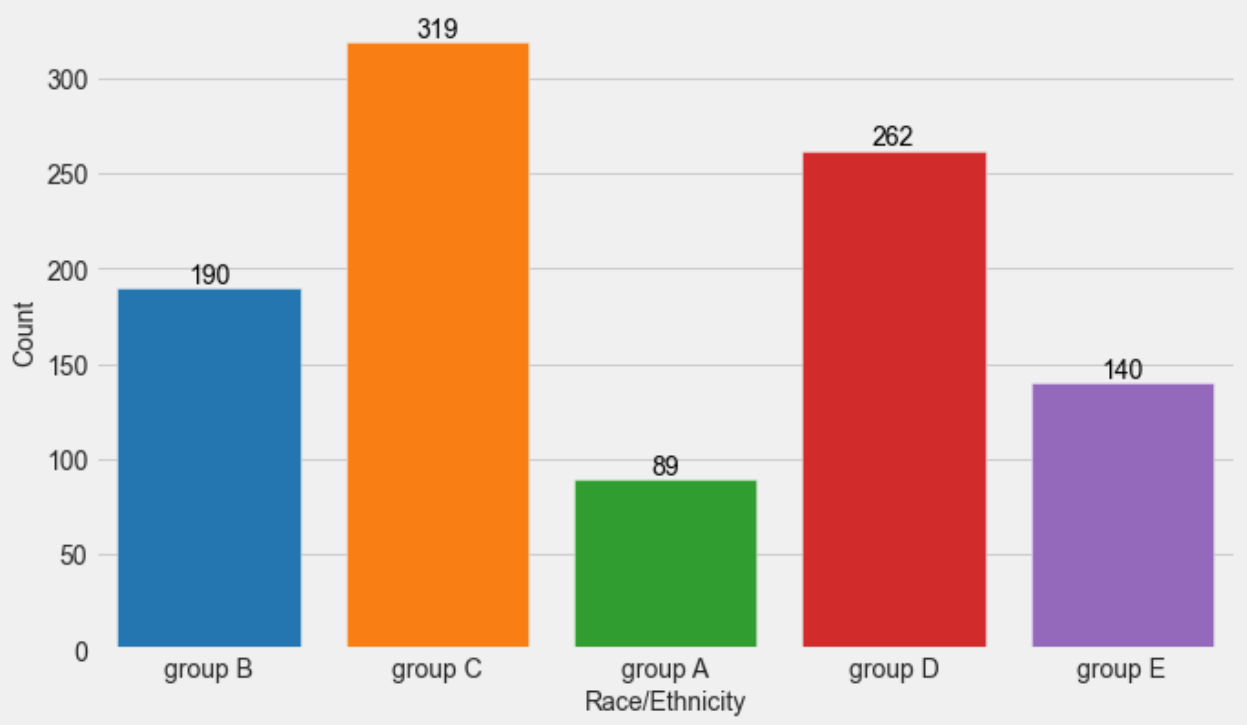

Figure 6.9: Race/Ethnicity Distribution Bar Plot

According to the Figure 6.9, Group C has the highest representation, with 319 students, indicating this group forms the majority of the population. Group D follows with 262 students, also significantly contributing to the dataset. Group B and Group E are moderately represented with 190 and 140 students, respectively. Group A has the least representation, with only 89 students, making it a minority group in the dataset. This distribution is essential for understanding demographic diversity in the dataset. It can be used to analyse performance or engagement disparities across ethnic groups and assess fairness or bias in predictive modeling. Uneven group sizes should be considered when building machine learning models, as imbalanced class representation can impact model generalization and lead to biased predictions. This visualization also supports equity-focused educational planning and decision-making. The Figure 6.10 represents a bar plot illustrating the distribution of student grades across different performance categories labelled as A, B, C, D, E, F, and O. The x-axis denotes the grade categories, while the y-axis reflects the number of students falling within each category. Each bar's height directly corresponds to the frequency count for that specific grade. Grade C has the highest frequency with 258 students, closely followed by Grade B with 255 students, indicating that the majority of students are performing at an average to above-average level.

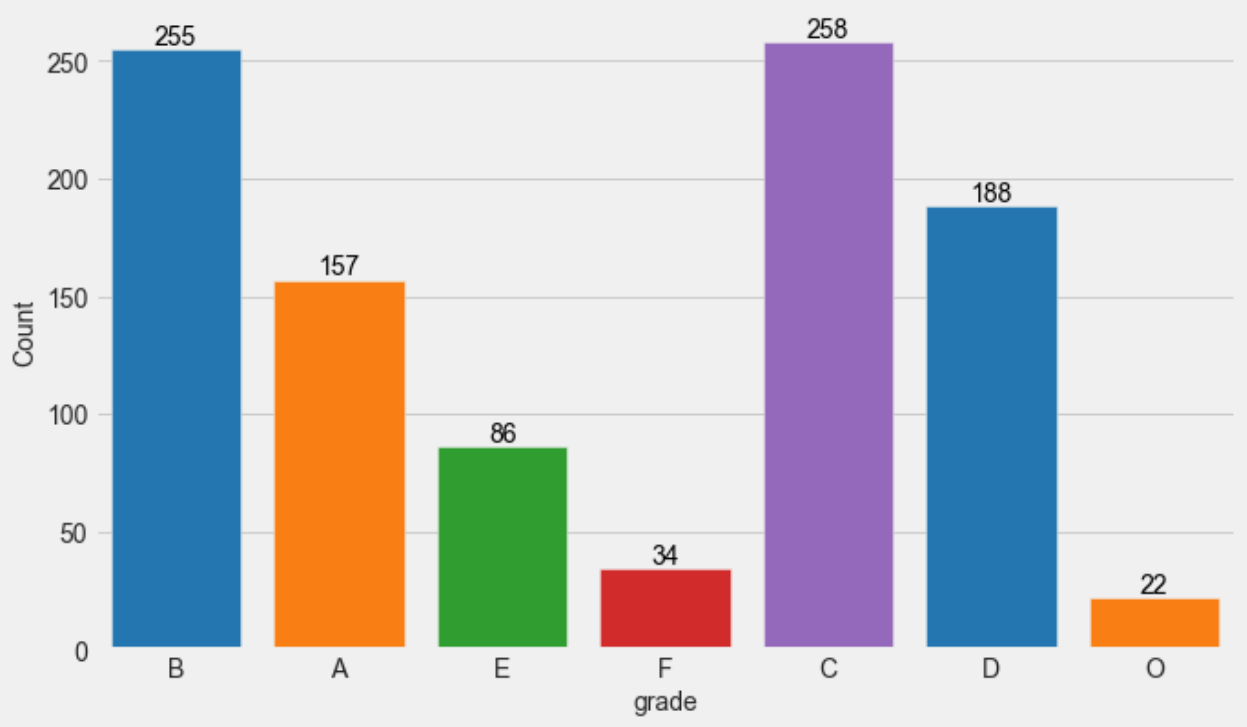

Figure 6.10: Grade Distribution Bar Plot

Grade D is also significantly represented with 188 students, showing a considerable portion at the lower-middle performance level. Grade A accounts for 157 students, representing high-performing individuals, while Grade E and Grade F (with 86 and 34 students, respectively) capture those needing academic improvement. Grade O (possibly denoting "Outstanding" or "Other") has the lowest representation, with only 22 students, suggesting a rare classification or special category. This visualization serves as an effective tool for univariate categorical analysis and helps in identifying academic performance trends in the student population. It can also assist stakeholders in identifying where interventions or remedial programs are necessary, especially for those falling into lower grades (E, F). The skewed distribution toward middle-range grades suggests that the dataset is balanced but could benefit from further analysis into the underlying causes of high and low performance.

Bivariate Analysis denotes the statistical technique employed to ascertain the connection between two variables. Within the framework of “a framework for academic performance analysis of students in online learning using machine learning approaches,” bivariate analysis is essential for elucidating the strength, direction, and type of relationships between various student traits and academic results.

Bivariate analysis examines the relationship between two variables for example, how test preparation affects final scores, or how parental education correlates with student performance. Types of Relationships:

- **Numerical vs. Numerical:** (e.g., Reading Score vs. Writing Score – using correlation or scatter plots)
- **Categorical vs. Numerical:** (e.g., Gender vs. Average Score – using boxplots or t-tests)
- **Categorical vs. Categorical:** (e.g., Parental Education vs. Lunch Type – using cross-tabulation or chi-square tests)

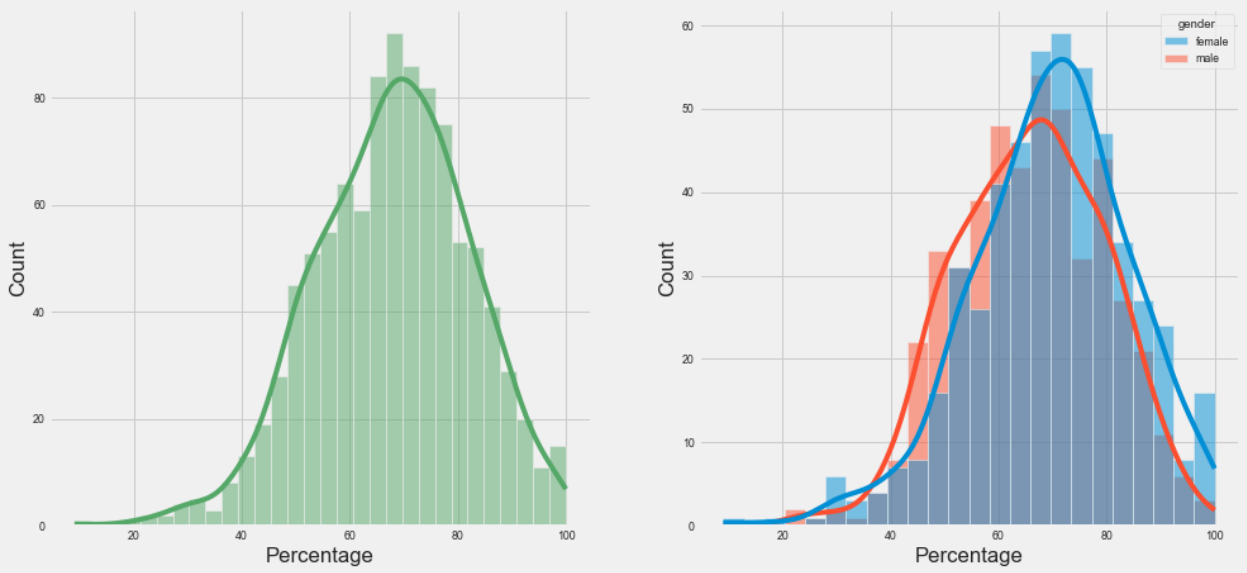

Figure 6.11: Overall Distribution of Percentage

Figure 6.11 illustrates two KDE (Kernel Density Estimation) plots integrated with histograms to analyse the distribution of student academic performance percentages within the framework of academic performance analysis of students in online learning using machine learning approaches. The left subplot shows the overall distribution of student percentages, demonstrating a near-normal distribution with a slight positive skew. The peak lies around 70–75%, indicating that most students tend to perform in this range, while fewer students score significantly lower or higher, as seen from the tapering tails. This visualization offers insights into the general academic trends across the dataset, helping identify where the majority of student scores lie and detect any potential outliers. The right subplot extends the analysis by comparing the performance distribution based on gender. It presents separate KDE curves for male and female students, with slight variations in

shape and spread. Female students show a more concentrated distribution around the average range, while male students display a slightly wider spread, indicating more variation in their scores. The KDE curves also reveal that male students tend to have a slightly higher proportion of high scores compared to females. This gender-based distribution analysis is essential for bivariate exploration and can be used to detect disparities, plan targeted interventions, and design equitable learning strategies. Overall, the figure provides a meaningful visualization to support data-driven decision-making in improving student outcomes in online learning environments.

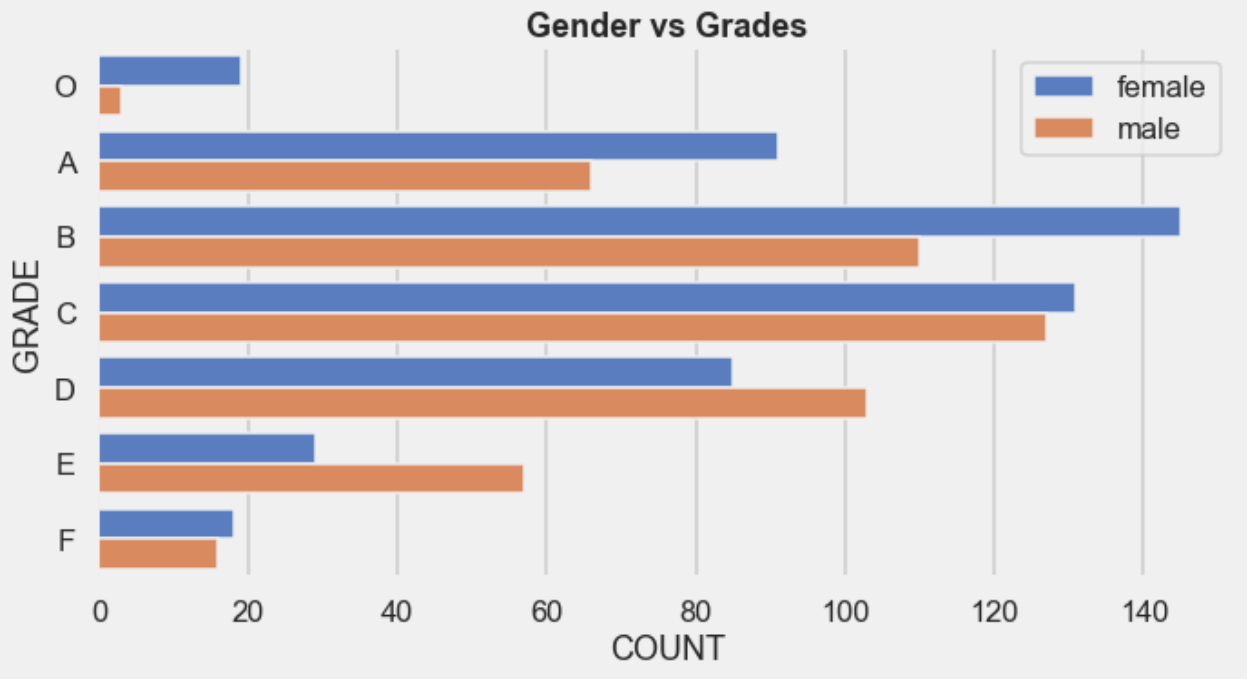

Figure 6.12: Distribution of academic grades between male and female students

Figure 6.12 presents a horizontal grouped bar chart that compares the distribution of academic grades between male and female students, serving as an essential visualization in the proposed model. The x-axis represents the count of students, while the y-axis denotes different grade categories ranging from 'O' (highest) to 'F' (failing). Each bar is color-coded based on gender—blue for females and orange for males allowing clear visual distinction and comparison. From the chart, it is evident that female students have a higher representation in top-performing grades such as 'O', 'A', and 'B', with the largest count observed in grade 'B'. Conversely, male students exhibit greater concentration in lower grades, particularly 'E' and 'F', suggesting comparatively lower academic performance. This visualization supports gender-based bivariate analysis by revealing performance trends that are crucial for understanding disparities in online learning environments.

Such insights can guide educators in designing gender-sensitive academic interventions, tailoring support strategies, and ensuring inclusive academic success.

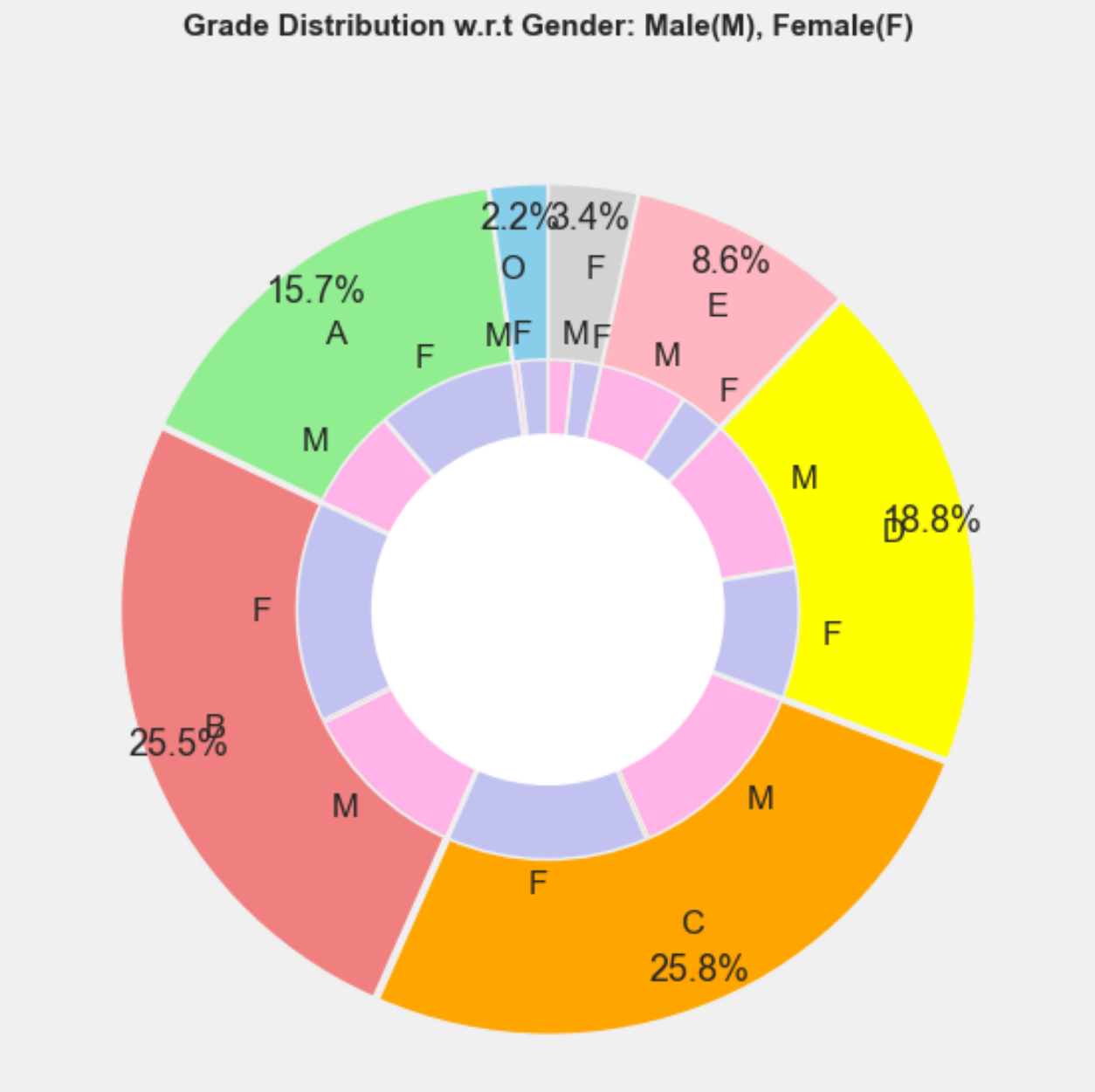

Figure 6.13: Grade Distribution of Students

The presented Figure 6.13 is a multi-layered donut chart that depicts the grade distribution of students with respect to gender, providing a visual representation that aligns with the objectives of proposed model. The outer ring represents the overall percentage of students falling into each grade

category—ranging from 'O' (outstanding) to 'F' (fail)—while the inner ring further breaks down each grade segment by gender, with 'M' indicating male students and 'F' indicating female students.

From the outer ring, it is observed that the highest concentration of students falls under grades 'C' (25.8%) and 'B' (25.5%), followed by 'D' (18.8%), indicating a central tendency around average performance. The inner ring reveals that grades 'B' and 'C' are fairly balanced in terms of gender, whereas females dominate the top-performing categories ('A' and 'O'), and males show higher presence in the lower-performing grades ('E' and 'F'). This gender-based grade segmentation is essential for bivariate analysis in academic performance frameworks as it provides nuanced insights into the demographic influence on academic outcomes. The chart supports the need for personalized learning strategies and gender-sensitive interventions in online education. By highlighting disparities in performance between male and female learners, this visualization aids educators and data scientists in constructing predictive models, designing support systems, and improving overall learning effectiveness through machine learning-based educational analytics.

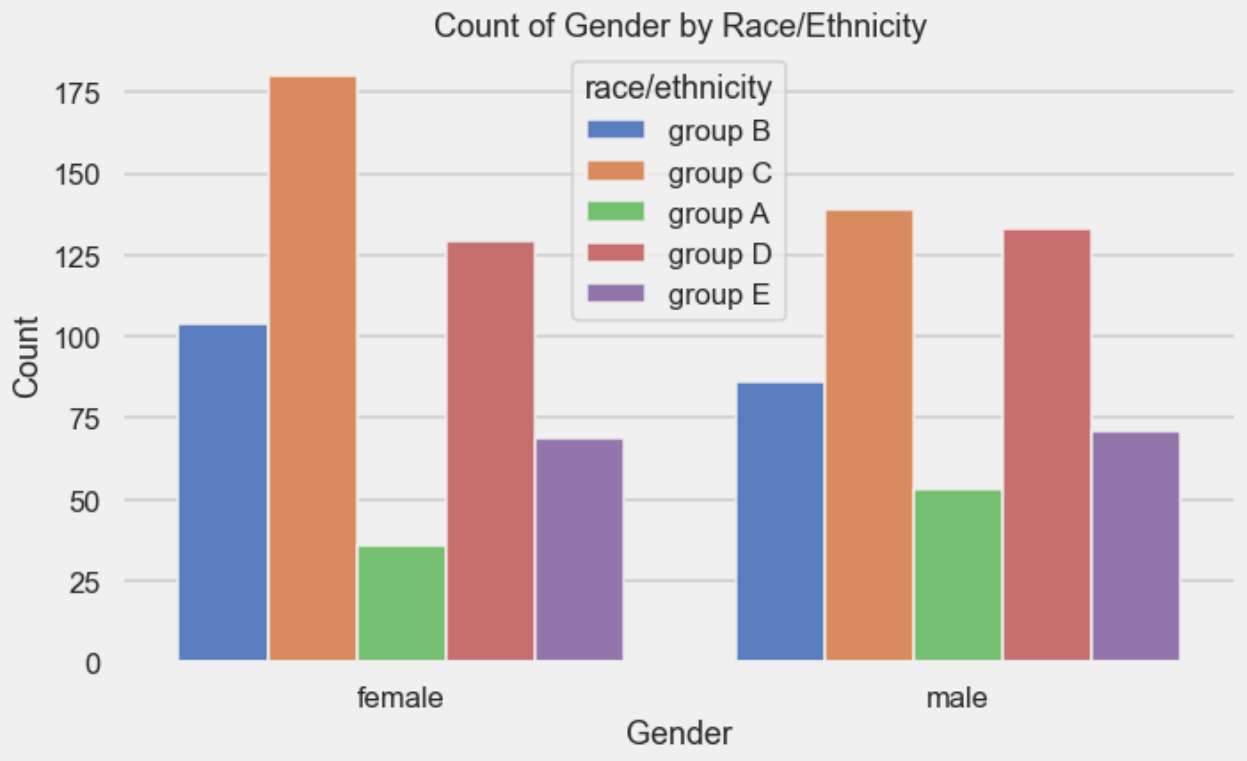

Figure 6.14: Distribution of gender across various race/ethnicity groups

The provided Figure 6.14 is a clustered bar chart that displays the distribution of gender across various race/ethnicity groups, serving as an important visual component in proposed model. The

x-axis represents gender categories (female and male), while the y-axis shows the count of students. Each coloured bar within the gender categories corresponds to one of five race/ethnicity groups: A, B, C, D, and E. From the chart, it is evident that Group C holds the highest count among both male and female students, with female representation slightly exceeding that of males. Groups D and B also show significant representation across genders, although group D maintains near parity, while group B exhibits a slight female majority. Group A and Group E have comparatively lower counts, especially among females in Group A.

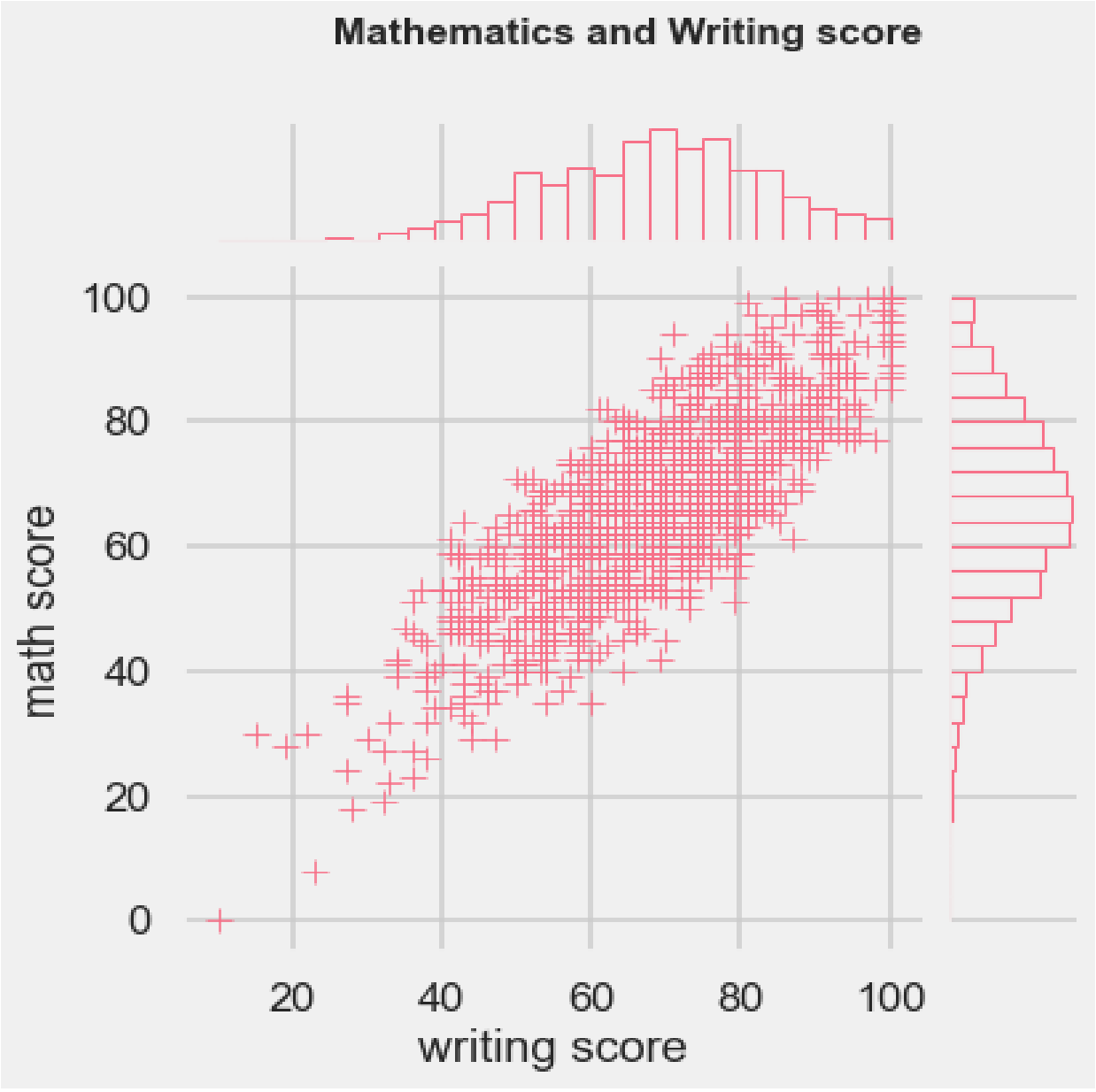

Figure 6.15: Correlation between students' mathematics and writing scores

Figure 6.15 is a joint plot that illustrates the correlation between students' mathematics and writing scores, which is a vital aspect of model. The central scatter plot presents individual data points, where each pink cross ("+") represents a student's performance in both math (y-axis) and writing (x-axis). The marginal histograms at the top and right of the graph display the distribution of writing and math scores, respectively. From the figure, a strong positive linear relationship is clearly observed students who score high in writing tend to also perform well in mathematics. The clustering of data points along a diagonal pattern indicates consistency in academic ability across these two subjects. The bell-shaped distribution in the marginal plots suggests that most students score within the mid to high range for both subjects, with fewer students at the extreme ends.

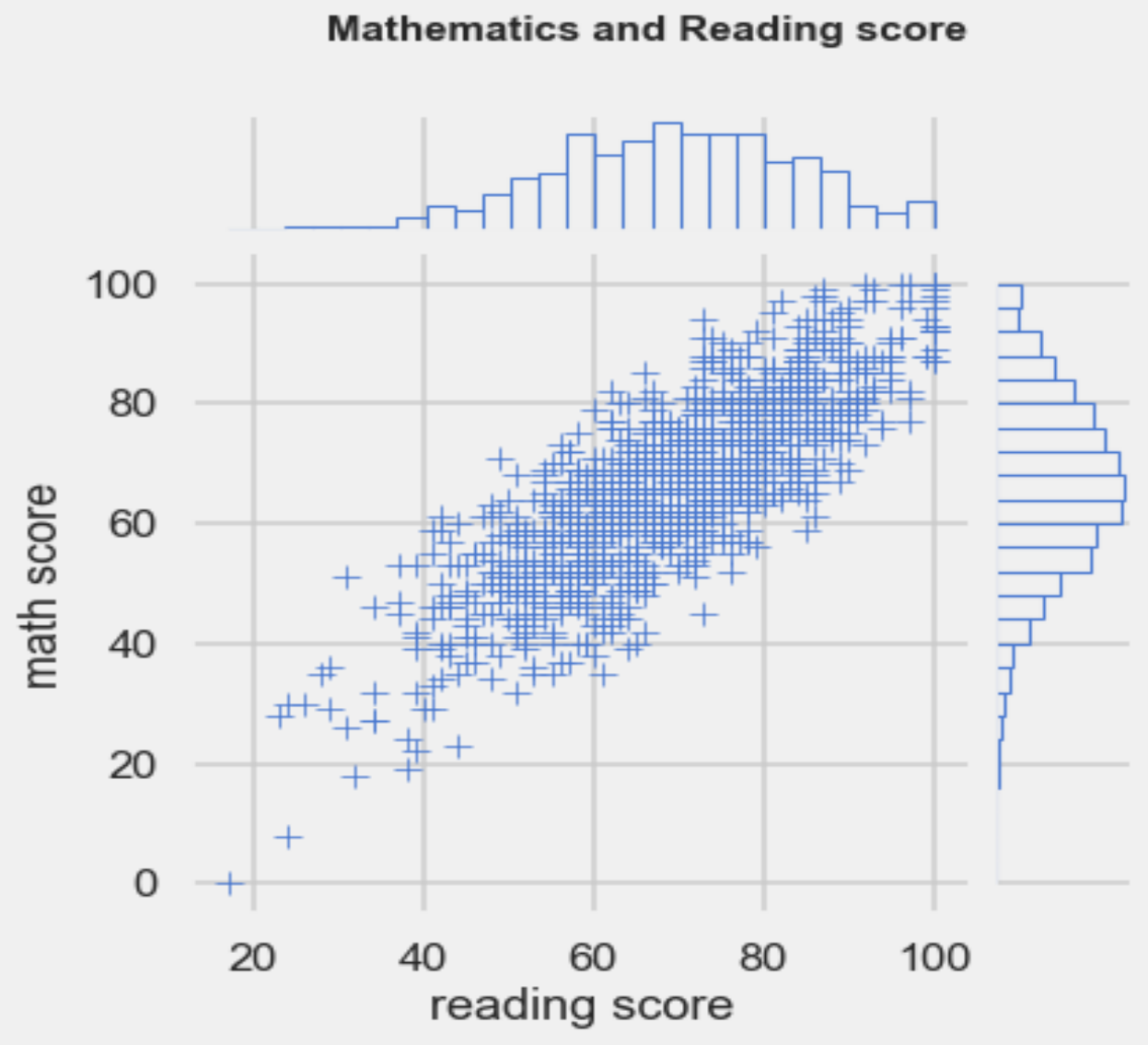

Figure 6.16: Relationship between students' mathematics and reading scores

Figure 6.16 illustrates a joint plot representing the relationship between students' mathematics and reading scores, which plays a significant role in proposed model. The central scatter plot, formed by blue plus signs (“+”), captures individual student data, where the x-axis corresponds to reading scores and the y-axis to math scores. The marginals—histograms at the top and right—visualize the distribution of scores for reading and math, respectively. From the scatter plot, there is a clear positive correlation between reading and math scores. Students who excel in reading generally also perform well in mathematics, suggesting intersubject competence and possibly shared cognitive skills or learning strategies. The density of points along the upward diagonal axis reinforces this linear trend, indicating consistency in performance across the two domains.

Figure 6.17: Relationship between reading and writing scores

The Figure 6.17 displays a hex-bin joint plot illustrating the relationship between reading and writing scores of students, an essential component of performance evaluation in the proposed

model. In this plot, the central hexagonal grid highlights the concentration of data points: darker hexagons indicate a higher density of students scoring similarly in both reading and writing, while lighter hexagons represent sparser data. The histograms on the top and right axes show the marginal distributions of writing and reading scores, respectively, and reveal that both follow an approximately normal distribution, with peaks around the 70–80 score range. Technically, the strong clustering along the diagonal line suggests a strong positive correlation between reading and writing performance. This implies that students who perform well in writing generally also excel in reading, likely due to the shared cognitive and linguistic skills required for both tasks. The central black hexagon indicates the highest density region, representing the most common score pairing among students.

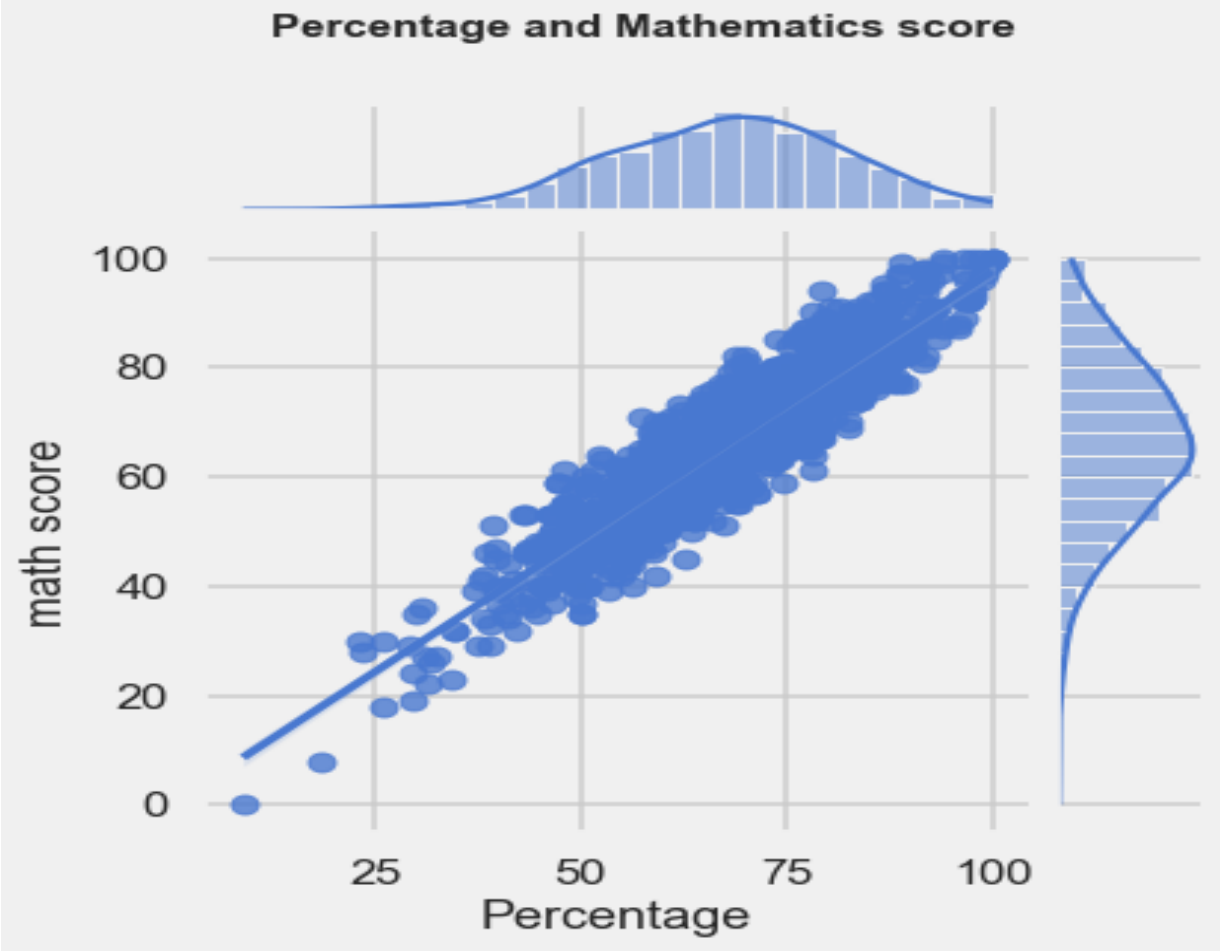

Figure 6.18: Correlation between academic percentage and mathematics score

The presented Figure 6.18 is a joint plot that examines the correlation between overall academic percentage and mathematics score within the context of proposed model. The scatter plot at the

center highlights individual data points, with each dot representing a student's performance, while the marginal histograms on the top and right display the distribution of percentage and math scores, respectively. The presence of a positively sloped regression line in the scatter plot signifies a strong linear relationship between the two variables. Technically, this figure shows that students with higher mathematics scores tend to have a higher overall percentage, reinforcing the critical impact of math performance on cumulative academic achievement. The tightly clustered data points around the regression line further indicate a low variance and strong predictive relationship, which is essential for designing accurate machine learning models. The bell-shaped histograms confirm a near-normal distribution for both variables, supporting the use of statistical models like linear regression in performance prediction. Such insights are fundamental in developing intervention strategies and identifying key academic drivers within online learning platforms.

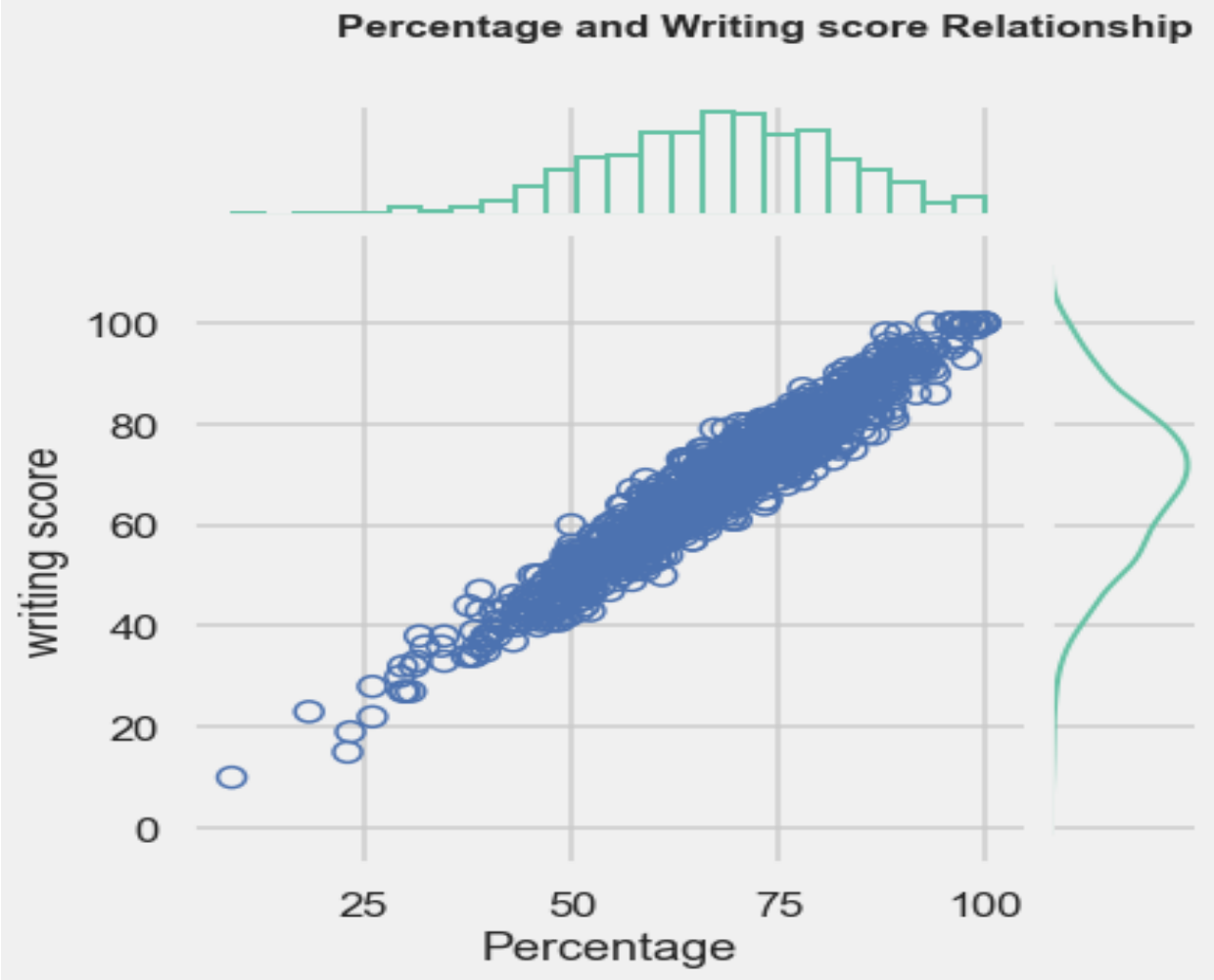

Figure 6.19: Percentage and Writing Score Relationship

The Figure 6.19 titled "Percentage and Writing Score Relationship" illustrates a bivariate analysis using a joint distribution plot to understand the correlation between students' overall academic percentage and their writing scores within the framework. The central scatter plot shows a strong linear alignment of data points, indicating a highly positive correlation between the two variables. As students' overall percentages increase, their writing scores tend to increase as well, suggesting that writing proficiency significantly influences overall academic performance. The marginal histograms and density plots at the top and right display the distribution of percentage and writing scores, respectively.

Both histograms indicate a near-normal distribution, with a slight skew towards higher scores, suggesting a generally good academic performance among the sampled students. This technical visualization is crucial for feature selection and model training in machine learning-based academic prediction models. It confirms that writing score is a reliable and consistent predictor of overall performance, which supports the integration of such variables into performance prediction algorithms to enhance the accuracy and reliability of the predictive framework.

The Figure 6.20 titled "Percentage and Reading Score" provides a technical representation of the relationship between students' overall academic percentage and their reading scores using a scatter plot overlaid with a regression line and marginal boxplots. This visualization is an essential part of proposed model, as it supports the understanding of feature correlation and data distribution. The central scatter plot reveals a strong positive linear correlation between reading scores and overall percentage—indicating that students who perform well in reading tend to have higher overall academic percentages. This implies that reading ability significantly contributes to academic success, making it a critical predictive feature in machine learning models for student performance.

The boxplot on the top shows the spread and outliers in overall percentage distribution, while the side boxplot represents reading score variability. Both boxplots indicate a relatively symmetric distribution with minor outliers, suggesting consistent student performance with some low-scoring anomalies. The presence of a clear linear trend and minimal dispersion strengthens the case for incorporating reading scores as a core feature in regression-based performance prediction systems, ultimately enhancing model precision and interpretability in online learning analytics.

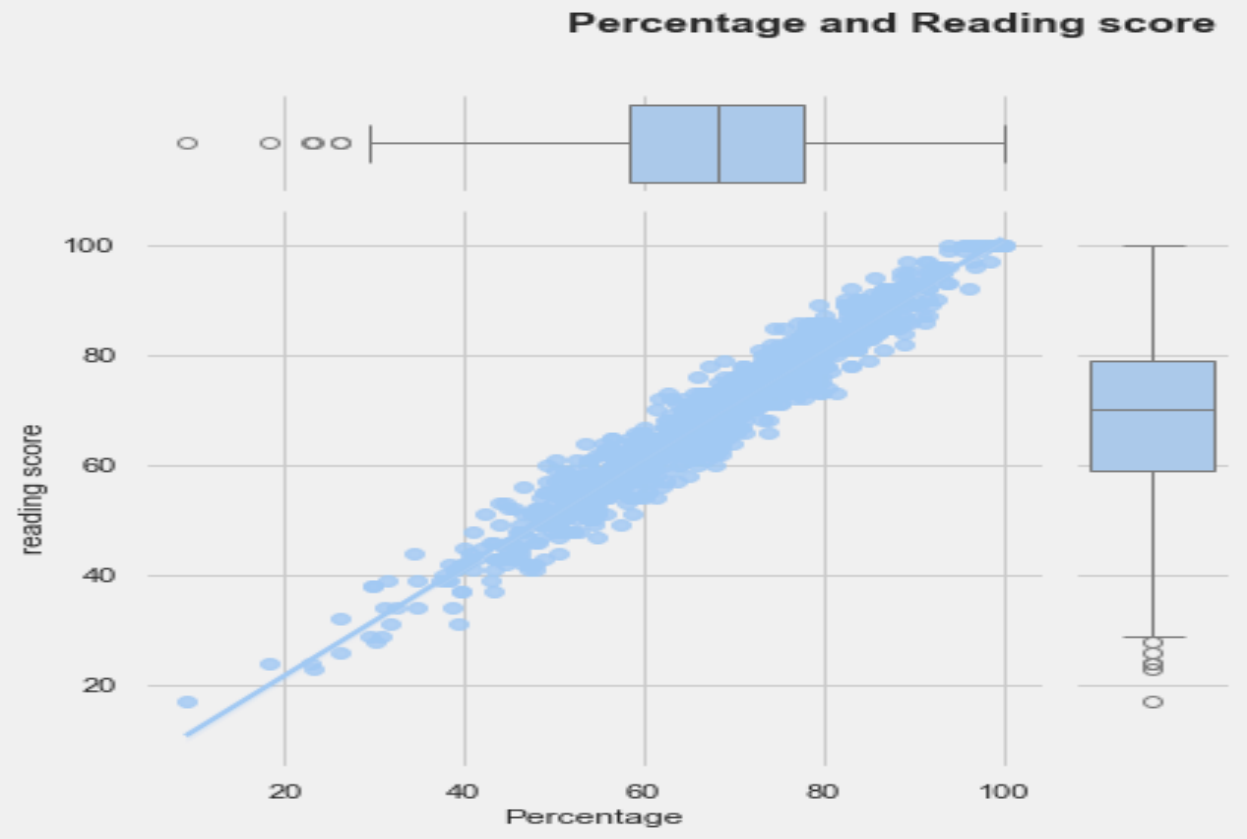

Figure 6.20: Percentage and Reading Score Relationship

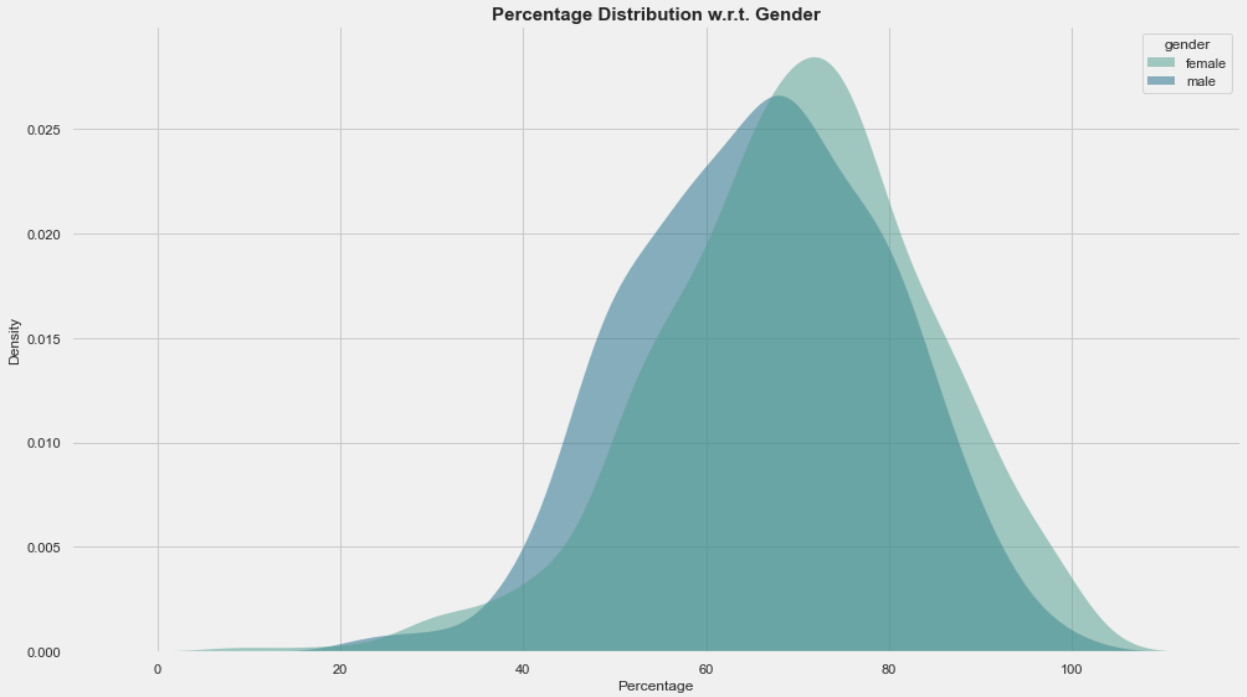

Figure 6.21: Percentage Distribution w.r.t. Gender

Figure 6.21 titled "Percentage Distribution w.r.t. Gender" represents a Kernel Density Estimation (KDE) plot that compares the distribution of academic percentage scores between male and female students. This visualization is critical in the context of proposed model, as it highlights potential gender-based differences in academic outcomes. From the plot, it is evident that both distributions exhibit a unimodal, slightly right-skewed pattern, indicating that the majority of students score between 55% and 80%. However, the peak of the female distribution is slightly higher and occurs at a greater percentage range compared to males, suggesting that a larger proportion of female students tend to achieve better academic scores. The female density curve is also slightly narrower, indicating less variability and more consistency in performance among females, whereas the male distribution shows a broader spread, implying more fluctuation in male students' academic outcomes. This gender-wise percentage distribution analysis is instrumental in feature engineering and bias detection within machine learning models. It allows researchers to assess fairness and make informed decisions about how gender influences academic performance, ensuring equitable educational interventions in online learning platforms.

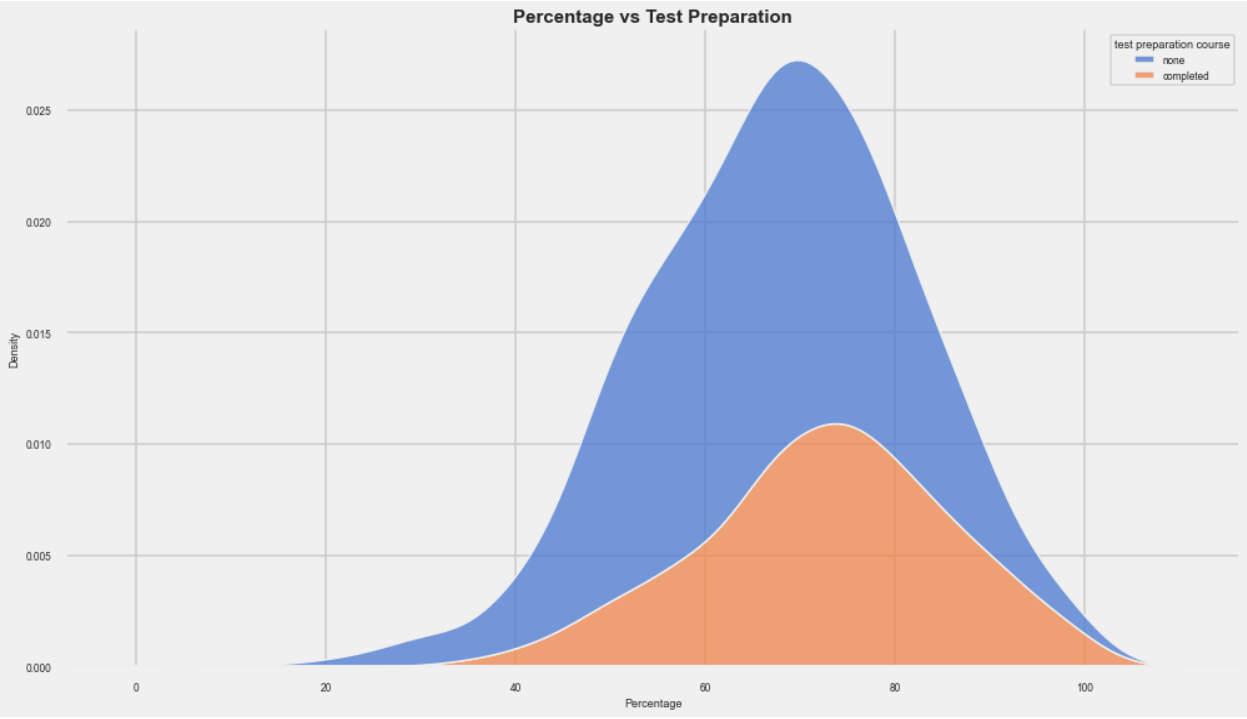

Figure 6.22: KDE Plot of Percentage vs Test Preparation

The figure titled "Percentage vs Test Preparation" illustrates a Kernel Density Estimation (KDE) plot comparing the distribution of students' overall academic percentage based on their completion

status of a test preparation course. In the context of A Framework for Academic Performance Analysis of Students in Online Learning Using Machine Learning Approaches, this visualization plays a crucial role in analysing how preparatory efforts influence academic success. The KDE plot shows two distinct curves—one for students who completed the test preparation course (orange) and another for those who did not complete it (blue). The curve representing students who completed the course peaks at a higher percentage value, suggesting that these students tend to score significantly better compared to those who didn't undergo preparation. The distribution is also narrower and more concentrated toward higher scores, indicating consistent performance.

In contrast, the blue curve for students who did not complete the course is wider and peaks at a lower percentage, demonstrating both lower average performance and greater variability. This clear distinction supports the hypothesis that structured test preparation contributes positively to academic outcomes. Such findings can be integrated into machine learning models as influential features and can guide targeted educational interventions. Educators and system designers can use this insight to recommend or mandate test preparation modules, especially for underperforming students in online learning environments.

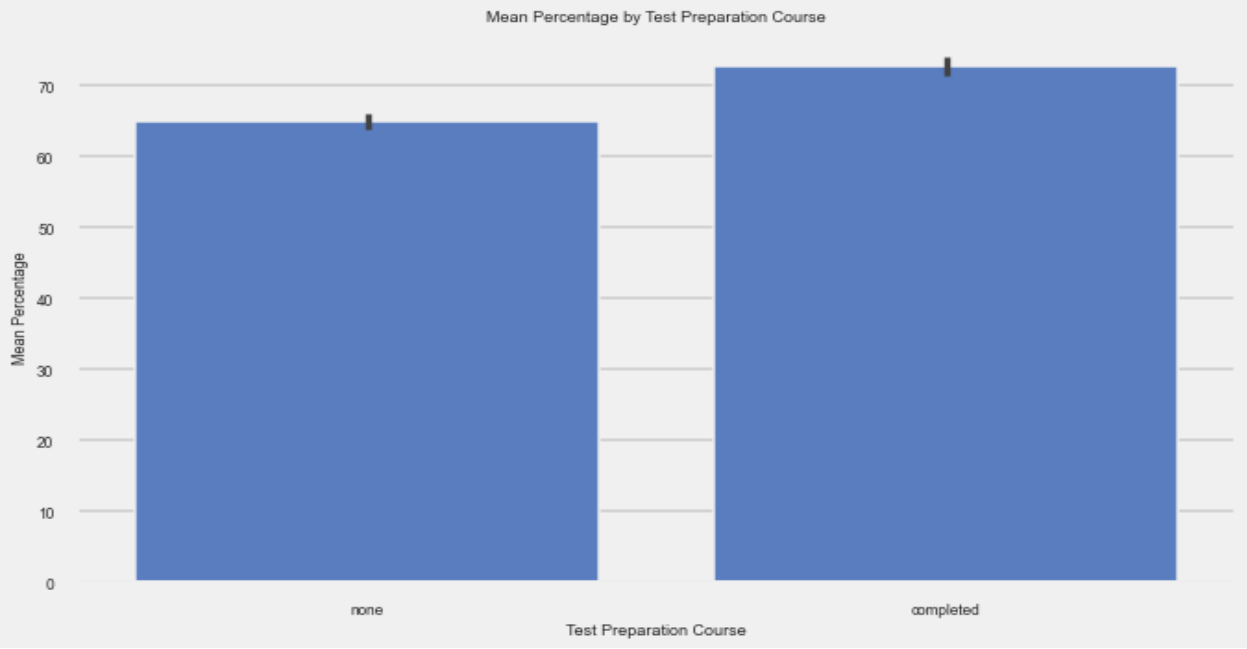

Figure 6.23: Mean Percentage by Test Preparation Course

The Figure 6.23 provides a comparative bar plot displaying the average academic performance, measured in percentage, of students who either completed or did not complete a test preparation

course. This visualization is highly significant within the context of proposed model, as it highlights the direct impact of preparatory interventions on student achievement. From the plot, it is evident that students who completed the test preparation course achieved a higher mean percentage, nearing the 70% mark, in contrast to those who did not complete the course, who averaged closer to 65%. The bars are accompanied by confidence intervals (error bars) which show a clear statistical distinction between the two groups. This difference confirms that structured test preparation contributes positively and consistently to improved academic outcomes. The results serve as valuable input for feature engineering in machine learning models used to predict student performance. Additionally, it emphasizes the practical importance of designing and recommending effective preparatory programs as a strategic policy in online education frameworks. It not only boosts student engagement but also enhances overall academic success.

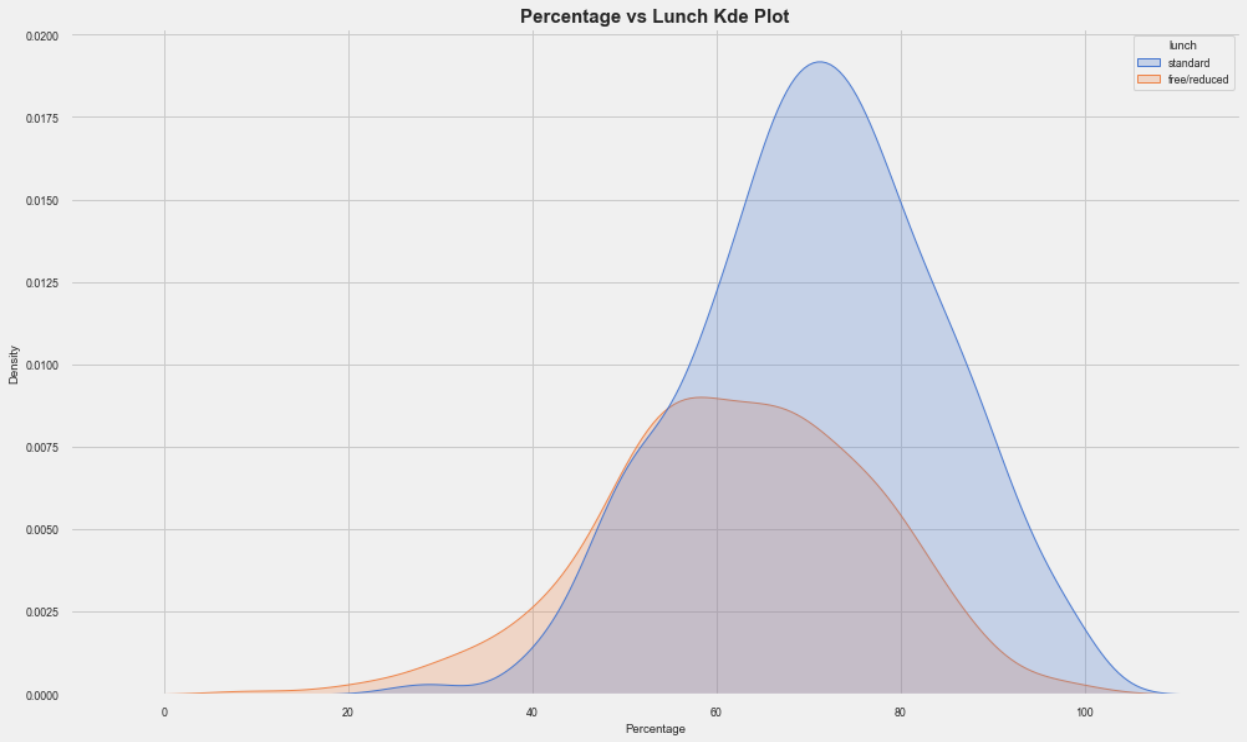

Figure 6.24: Percentage vs Lunch KDE Plot

Figure 6.24 illustrates a kernel density estimation (KDE) plot that compares the distribution of students' academic performance percentages based on their lunch type—standard lunch versus free/reduced lunch. This visualization serves as an insightful component of model, as it highlights socioeconomic factors potentially affecting student outcomes. From the KDE curves, it is observed

that students who received standard lunch (blue curve) exhibit a higher peak density around the 70–80% range, indicating a concentration of relatively higher-performing students. In contrast, students with free/reduced lunch (orange curve), which often corresponds to lower socioeconomic backgrounds, show a more spread-out distribution with a density peak at a lower percentage range (around 60%). This disparity underscores the impact of socioeconomic support mechanisms on student performance, where nutritional adequacy and economic stability could play critical roles in academic success. Such patterns are valuable for feature selection and predictive modeling in machine learning frameworks. The information can also be utilized to identify at-risk groups and develop targeted intervention strategies, reinforcing the importance of equitable educational support within an online learning context.

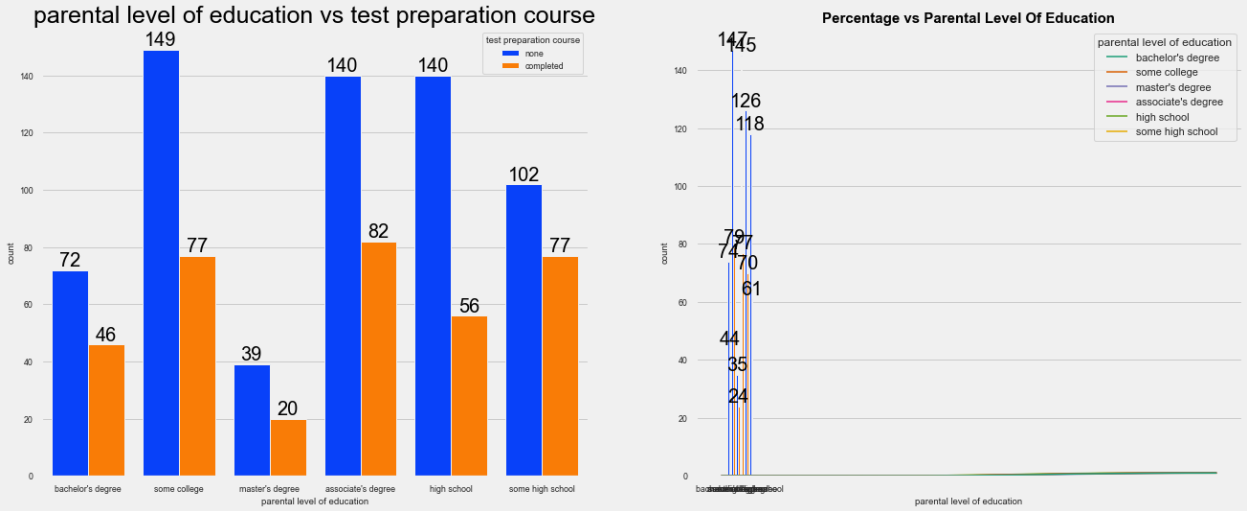

Figure 6.25: Relationship between parental level of education

Figure 6.25 presents a dual-panel visualization exploring the relationship between parental level of education, test preparation course completion, and students' academic performance, which is central to proposed model. The left subplot is a grouped bar chart that compares the number of students across various parental education levels—ranging from some high school to master's degree—based on whether they completed a test preparation course. It is evident that across all education categories, more students did not complete the test preparation course (indicated by blue bars). However, the completion rate increases with higher parental education, notably in categories like associate's degree and some college, suggesting that parental educational background may influence students' likelihood to engage in academic preparation.

The right subplot displays a percentage-based count of student performance scores segmented by parental education. While the axis labelling may be crowded, the graph indicates that students whose parents have higher education levels (bachelor's or associate's degrees) are more densely represented in the higher performance range. This trend implies a positive correlation between parental education and academic achievement, which is an important variable in predictive modeling. These insights emphasize that socio-educational background is a key determinant of student engagement and performance. Integrating such features into machine learning models can significantly improve the accuracy and fairness of academic performance prediction systems in online learning environments.

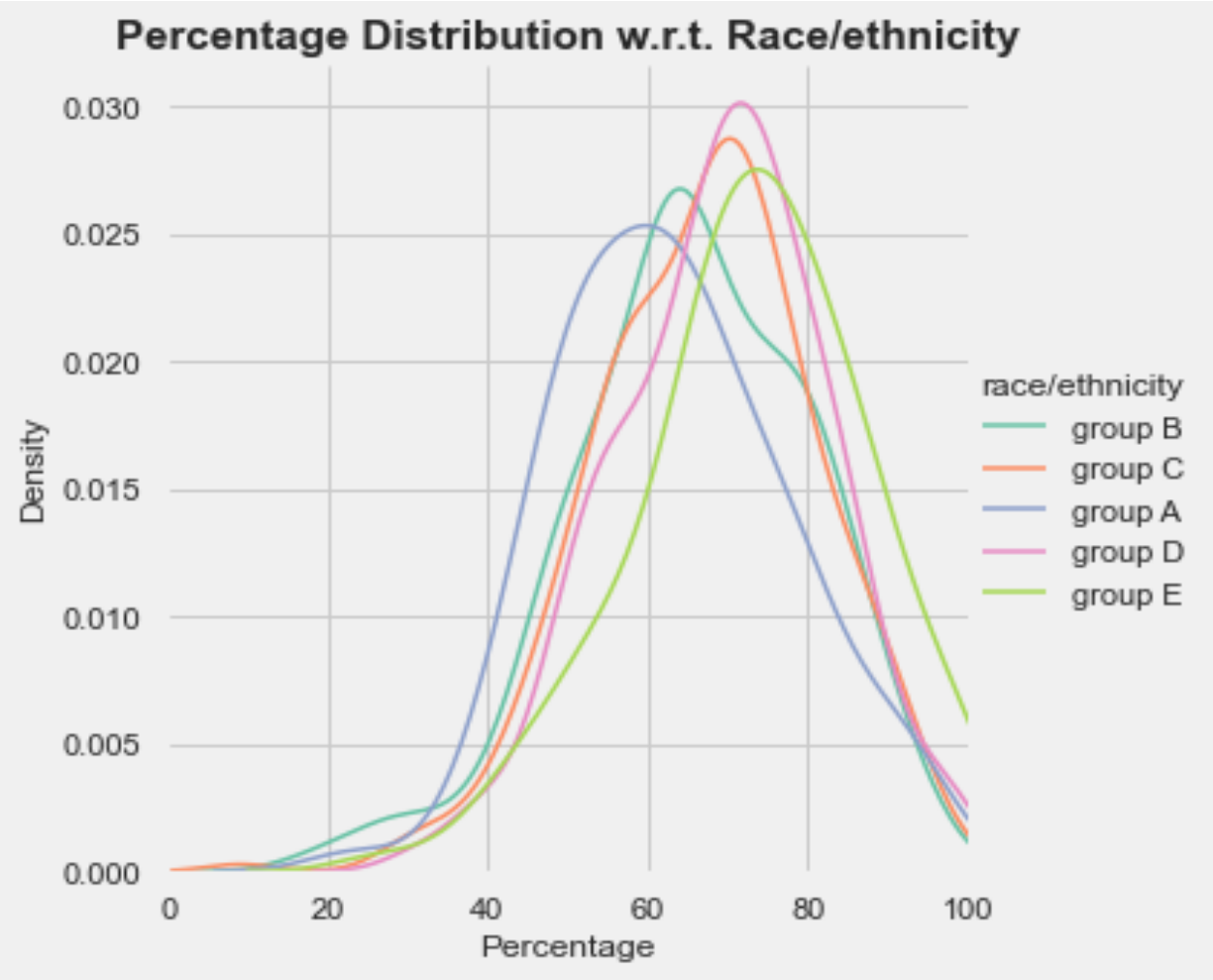

Figure 6.26: Percentage Distribution w.r.t. Race/Ethnicity

Figure 6.26 presents a KDE (Kernel Density Estimation) plot that illustrates the distribution of students' academic performance, measured in percentage, across five race/ethnicity groups—

Group A to Group E. This visualization is integral to the proposed framework for academic performance analysis of students as it highlights demographic-based performance trends that may influence model design and feature selection. From the plot, it is evident that Group D and Group E have the highest density peaks, suggesting a concentration of students scoring in the higher percentage range (around 70–80%). Conversely, Group A displays a broader distribution with a noticeable skew towards the lower end of the percentage scale, indicating that students in this group tend to score lower more frequently. Groups B and C show moderate distributions with their peaks also occurring around the 65–75% range, but with less sharp density compared to Group D.

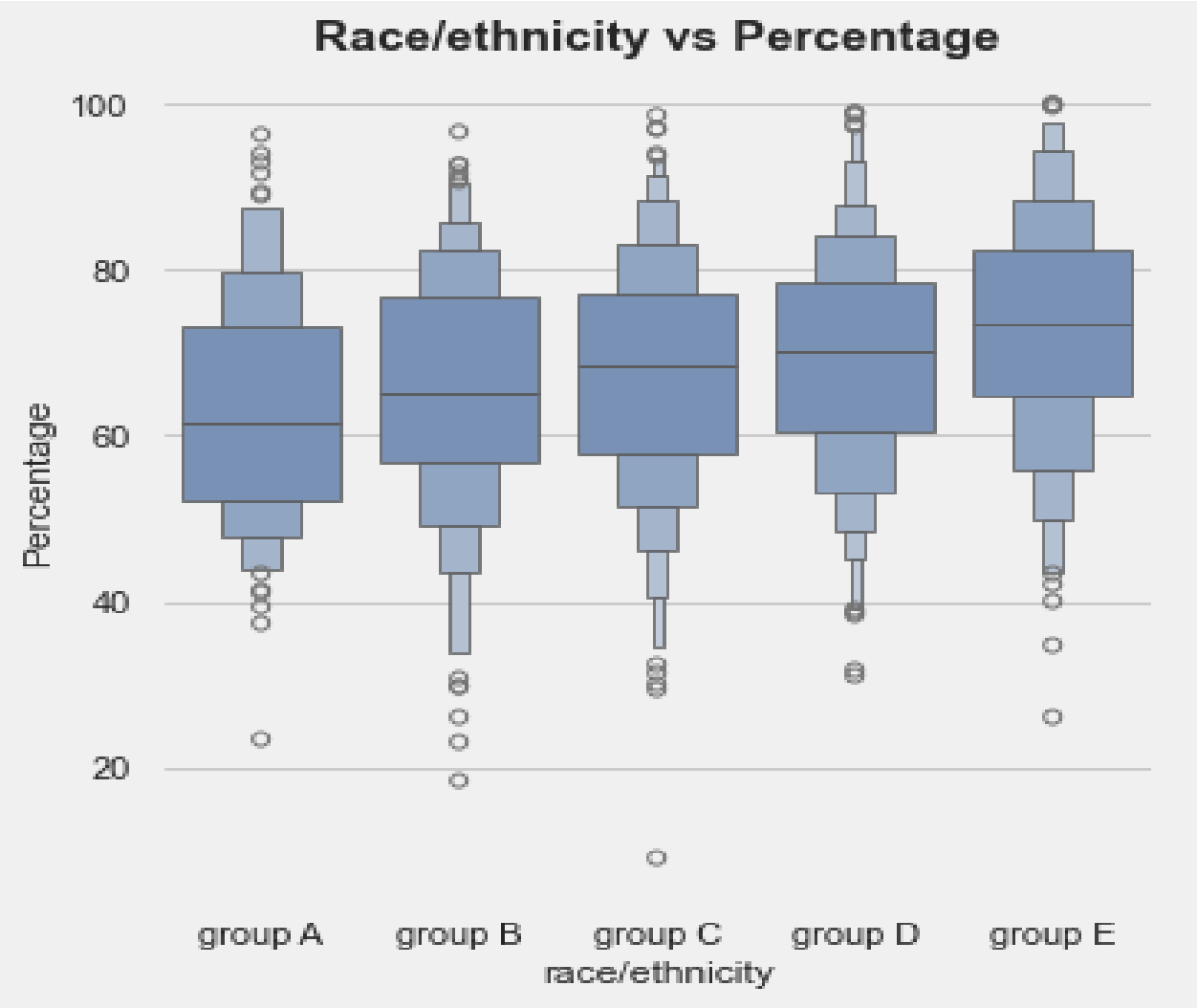

Figure 6.27: Box plot of Race/Ethnicity vs Percentage

Figure 6.27 represents a box plot comparison that visualizes the distribution of academic performance (in terms of percentage scores) across five distinct race/ethnicity groups (Group A to

Group E). This plot serves as a crucial component in the proposed framework, enabling the examination of demographic-based performance variation. Technically, each box plot displays the interquartile range (IQR), median, upper and lower quartiles, and potential outliers for each group. It can be observed that Group E exhibits the highest median score among all groups, followed by Group D and Group C, suggesting a stronger central tendency toward higher academic performance in these groups. Group A, however, shows the lowest median and a wider spread of data, indicating greater variability and a concentration of lower-performing students. The presence of outliers in all groups, especially in Group A and Group B, implies sporadic extreme performance levels—both high and low.

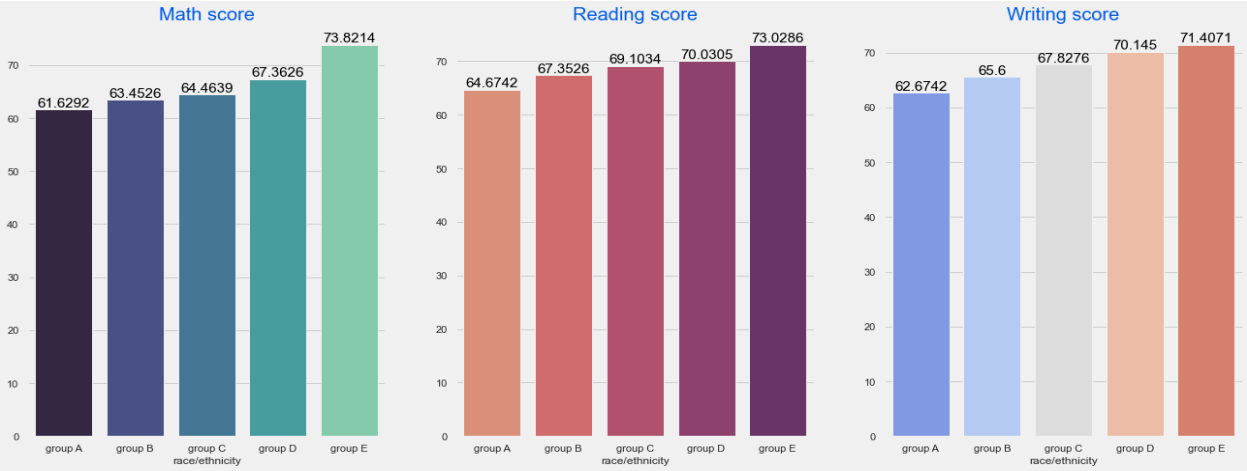

Figure 6.28: Grouped bar charts of the average academic scores

Figure 6.28 displays three grouped bar charts representing the average academic scores in mathematics, reading, and writing across five race/ethnicity groups (Group A to Group E). This comparative visualization is a vital part of the proposed framework, facilitating demographic insights into subject-wise academic performance. In the Math Score chart (left), Group E leads with the highest average score of 73.82, followed by Group D (67.36), while Group A reports the lowest (61.63). Similarly, the Reading Score chart (middle) shows a consistent trend, with Group E again attaining the highest mean (73.03) and Group A the lowest (64.67). The Writing Score chart (right) mirrors this pattern, where Group E maintains the top position (71.41) and Group A remains the lowest (62.67). These findings indicate a systematic difference in academic achievement across demographic groups, with Group E consistently outperforming others across all subjects. This suggests that ethnic or socio-cultural factors may influence student outcomes,

which can be leveraged in predictive modeling to enhance fairness and personalization. From a machine learning perspective, incorporating such demographic data into training models can improve prediction accuracy and support targeted educational interventions, ensuring equity and inclusivity in online learning platforms.

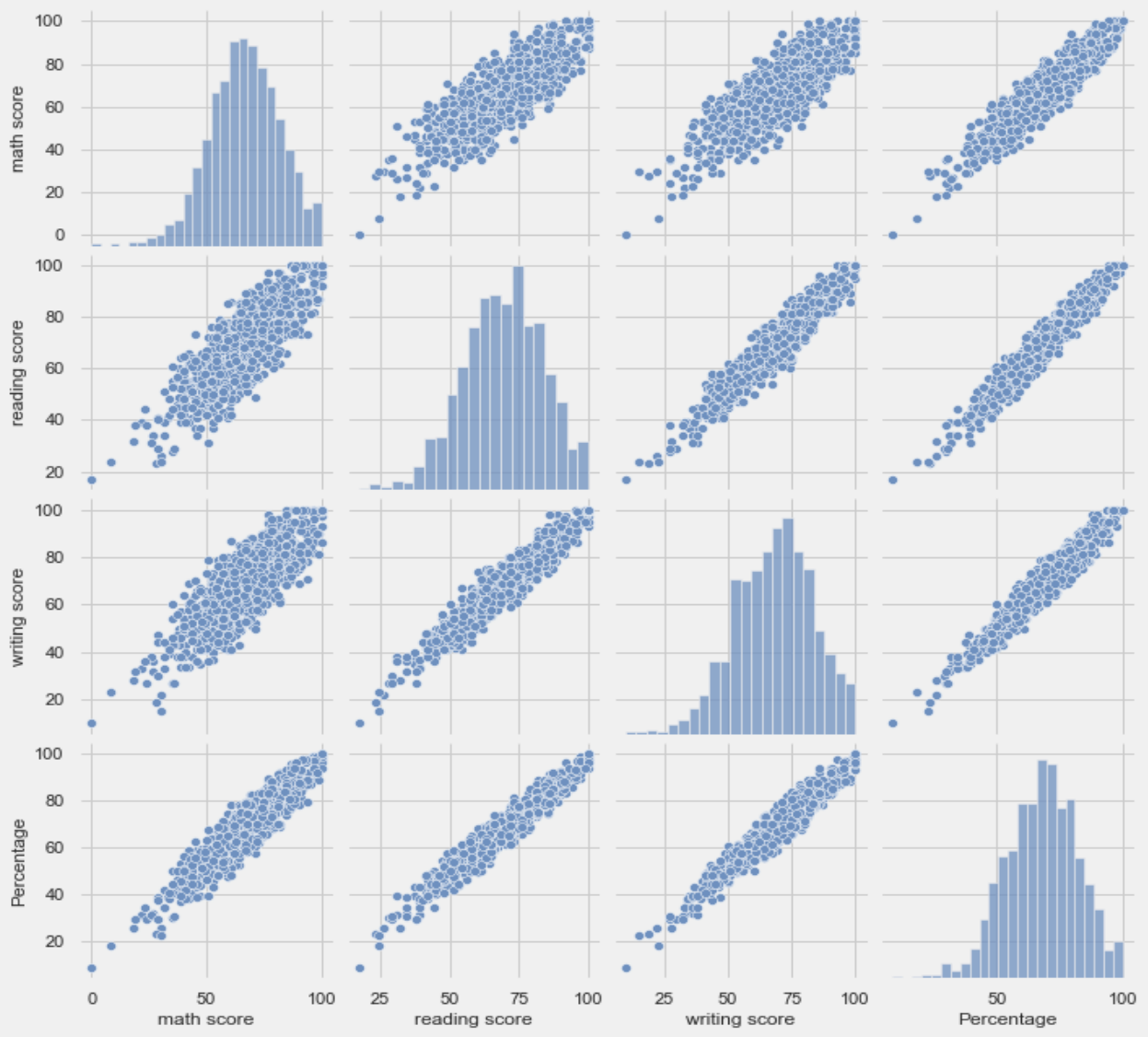

Figure 6.29: Relationships among math, reading, writing score, and overall percentage

The figure presents a pair plot matrix showcasing the relationships among four continuous academic performance metrics: math score, reading score, writing score, and overall percentage. Each diagonal plot displays the histogram of the respective variable, while the off-diagonal scatter plots represent bivariate relationships between pairs of variables. This visualization is a crucial component of proposed framework, offering insights into correlation patterns and data

distributions. Technically, the plot reveals a strong positive linear correlation among all subject scores and the overall percentage, as evidenced by the upward trend in the scatter plots. The math score, reading score, and writing score show significant interdependence, suggesting that students who perform well in one subject tend to perform well in others. The plots between each subject and the percentage (last column and last row) further confirm this trend, given that the percentage is the average of the three scores. Such strong correlations support the validity of using one or more subjects to predict overall performance through machine learning models. Additionally, the clear, tightly clustered scatter plots suggest minimal noise and strong predictive features—important for building robust supervised models.

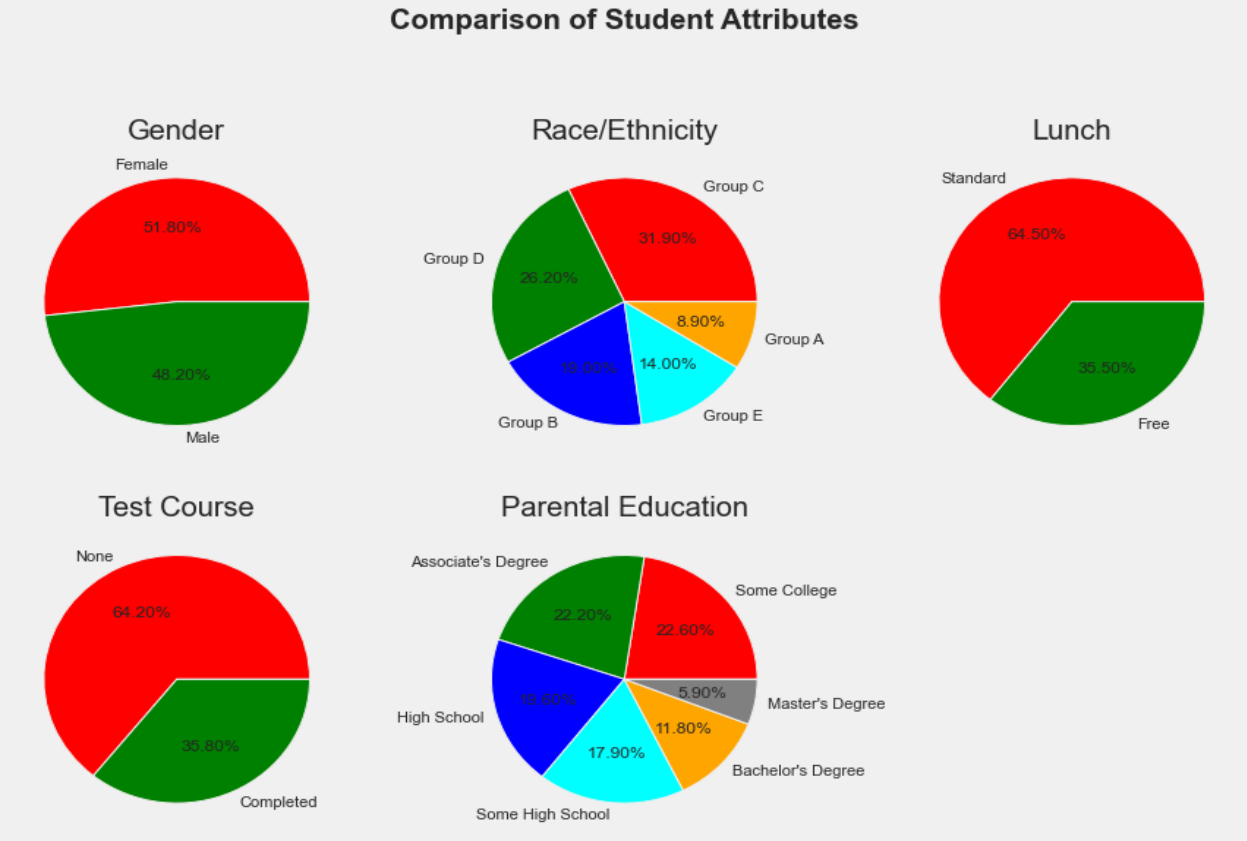

Figure 6.30: Comparison of Student Attributes

Figure 6.30 illustrates a series of pie charts that provide a categorical distribution of key demographic and socioeconomic variables among students, serving as a foundational analysis for proposed model. The charts present proportions across five critical attributes: gender, race/ethnicity, lunch type, test preparation course status, and parental level of education. The gender distribution is relatively balanced, with 51.80% female and 48.20% male, indicating a fairly

even representation. Race/ethnicity data reveal that the largest subgroup is Group C (31.90%), followed by Group D (26.20%) and Group B (19.00%), while Group A and Group E represent smaller portions at 8.90% and 14.00% respectively. These proportions help in understanding the ethnic composition and potential biases or disparities in performance among different groups. The lunch attribute, often used as a proxy for socioeconomic status, shows that 64.50% of students receive standard lunch and 35.50% receive free/reduced lunch, indicating a significant portion of students potentially from lower-income backgrounds. Similarly, the test preparation course attribute indicates that 64.20% of students have not completed a preparatory course, which could have implications for performance evaluation and intervention planning. Finally, the parental education level is segmented into six categories, with "some college" (22.60%) and "associate's degree" (22.20%) being the most prevalent, followed by "high school" (19.60%), "some high school" (17.90%), "bachelor's degree" (11.80%), and "master's degree" (5.90%). This distribution helps in assessing the educational background influence on student performance, which is often a strong predictor in academic modeling.

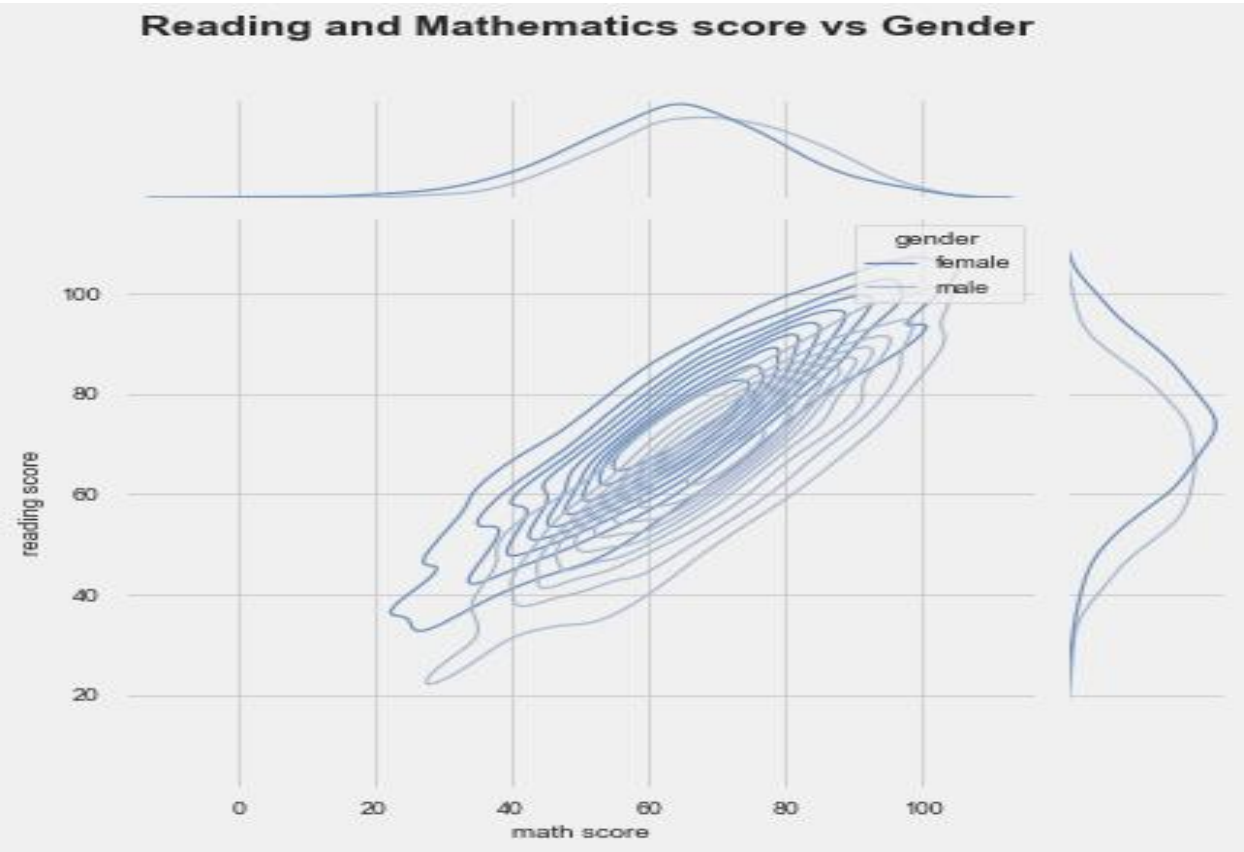

Figure 6.31: Reading and Mathematics Score vs Gender

Figure 6.31 represents a bivariate KDE (Kernel Density Estimate) plot with marginal distributions, which visually explores the joint distribution of students' scores in reading and mathematics, further disaggregated by gender. The concentric contour lines indicate the density concentration, showing where the scores are most densely populated across the two academic metrics. This type of visualization is crucial for analysing performance clustering and understanding how the score relationships differ between male and female students in the dataset. The overlap of contour lines suggests a high correlation between reading and math scores, and the majority of data points fall within the score range of 50 to 90 for both subjects. Additionally, the marginal density plots along the axes allow us to observe the distribution spread for each gender in individual subjects. While the plot appears to have minimal visual distinction between male and female trends due to overlapping curves, subtle differences in density peaks and spread can hint at gender-related performance variations, which is essential for bias detection and fairness-aware machine learning modeling in the academic performance analysis framework.

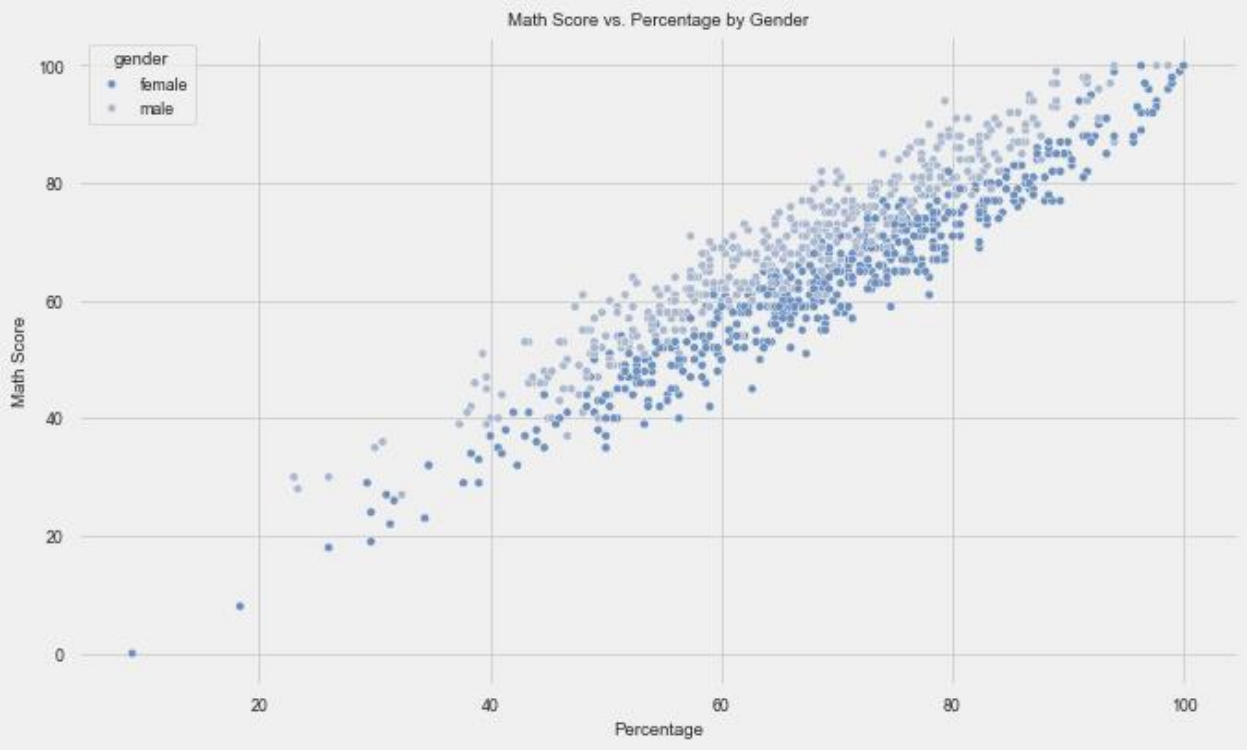

Figure 6.32: Math Score vs. Percentage by Gender

The scatter plot in Figure 6.32 represents a comparative analysis between individual student percentages and their corresponding mathematics scores, distinguished by gender (male and

female). Each point in the plot represents a student's performance, and the color-coded legends enable the gender-wise distinction. The visualization reveals a strong positive correlation between the overall percentage and math score, indicating that students with higher overall academic percentages also tend to score higher in mathematics. This figure is particularly valuable in the context of proposed framework, as it visually confirms the predictive relationship between percentage and domain-specific performance in math. While the distribution appears relatively uniform across both genders, the clustering of data points around the regression trend line suggests consistency in mathematical performance irrespective of gender, although minor variance can be observed in the density of points along the axis.

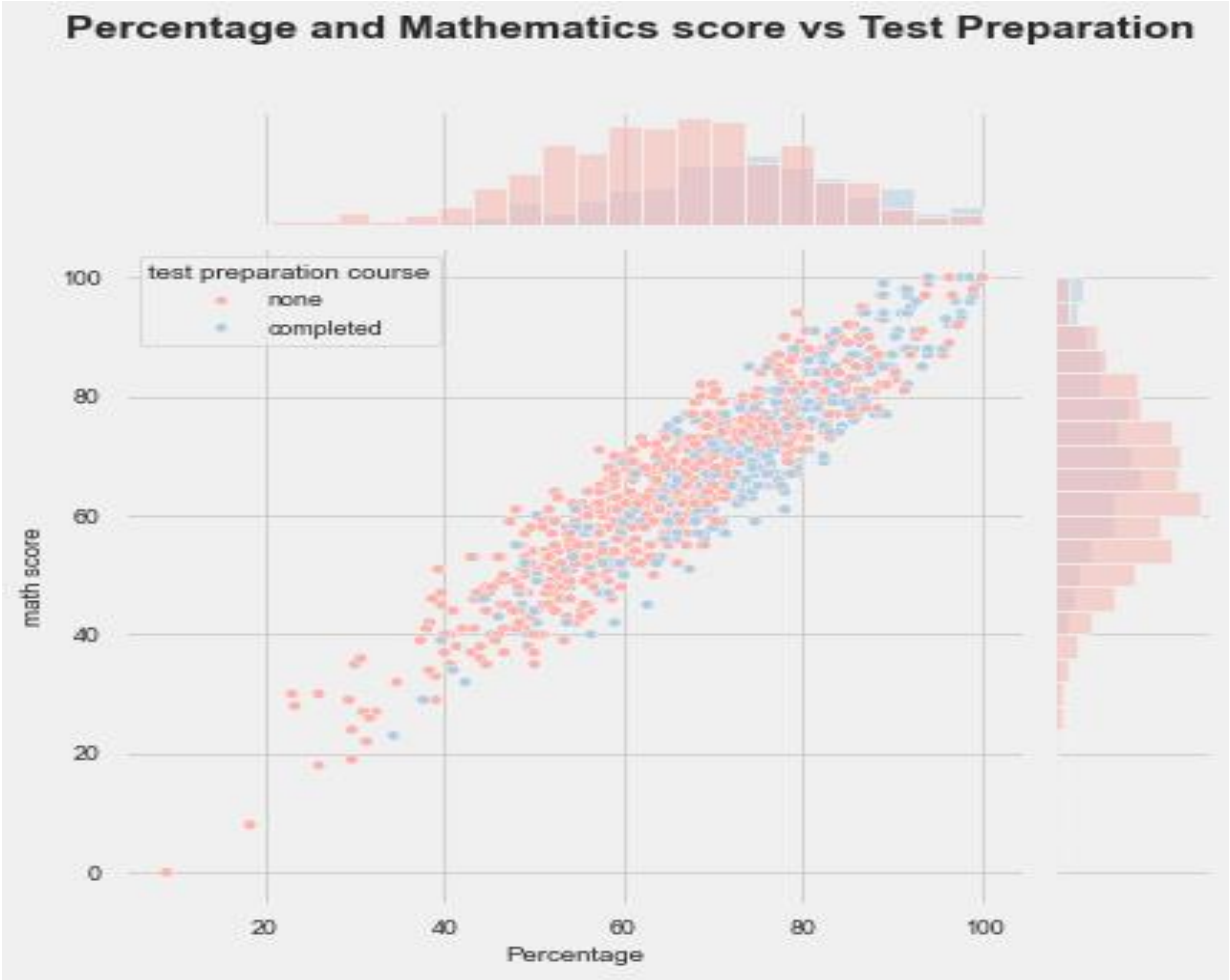

Figure 6.33: Percentage and Mathematics Score vs Test Preparation

The scatter plot in Figure 6.33 illustrates the relationship between students' overall academic percentage and their mathematics scores, segmented by their participation in a test preparation

course (categorized as "completed" or "none"). The plot uses color coding to differentiate between the two groups, with red dots representing students who did not complete the preparation course and blue dots for those who did. Accompanying histograms at the top and right provide marginal distributions for percentage and math scores, respectively. In the context of proposed framework for academic performance analysis, this figure reveals a positive linear correlation between percentage and math score for both categories. However, a visible clustering of blue dots (completed test prep) toward higher values on both axes suggests that students who undertook the preparation course generally performed better in mathematics and had higher overall percentages. The marginal histograms further confirm this, with the peak of the blue distribution shifting slightly to the right compared to the red. This observation underlines the predictive significance of test preparation status as a feature in academic performance models and validates the integration of preparatory course participation into machine learning models for student outcome forecasting.

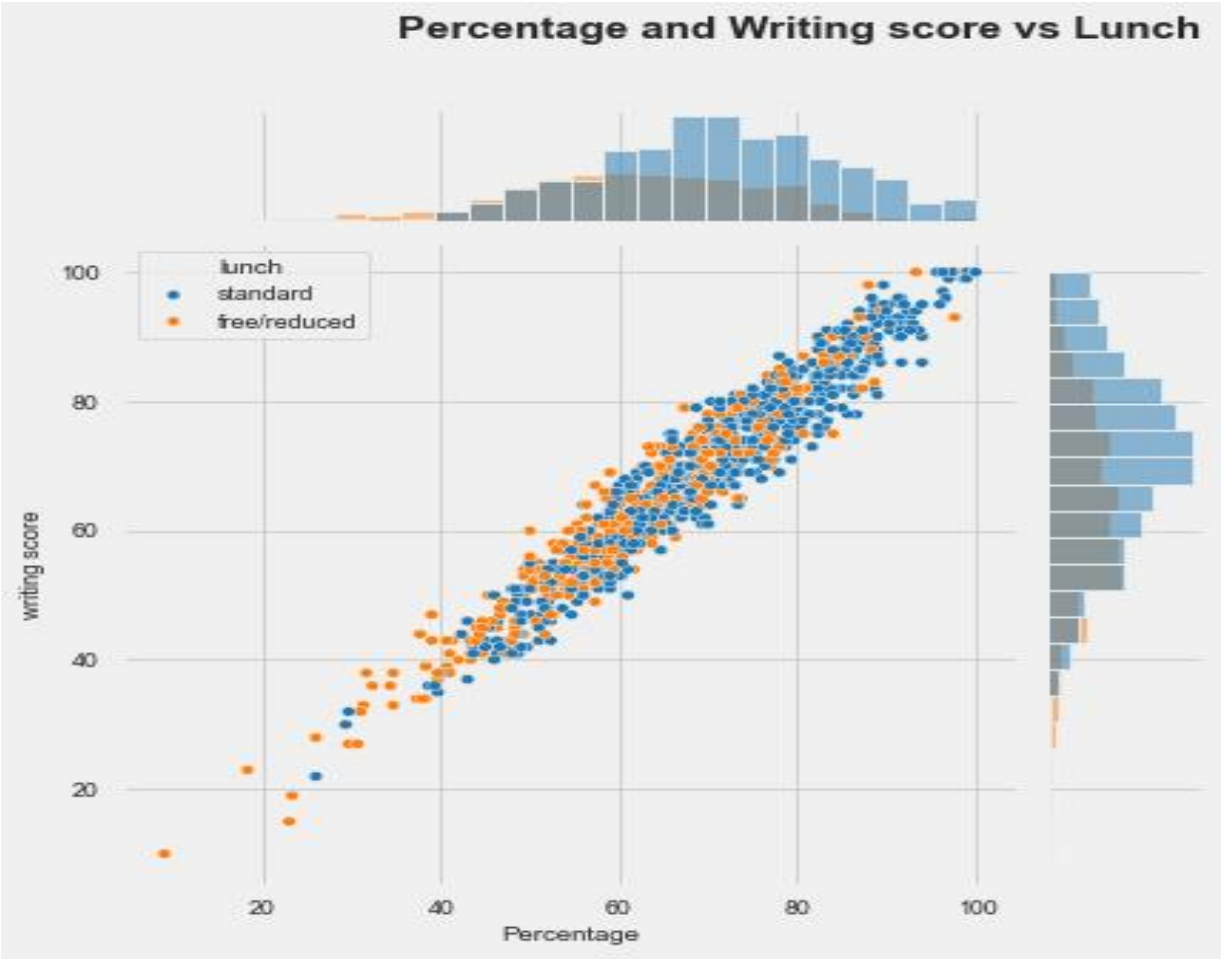

Figure 6.34: Percentage and Writing Score vs Lunch

The scatter plot titled "Percentage and Writing Score vs Lunch" visually analyzes the relationship between students' overall academic percentage and their writing scores, categorized by lunch type (standard vs. free/reduced). Each data point represents a student, with blue indicating those receiving a standard lunch and orange representing students receiving free or reduced lunch. The accompanying marginal histograms display the distribution of percentage and writing scores for each lunch category. In the context of proposed framework for academic performance analysis of students in online learning using machine learning approaches, this figure highlights a strong positive correlation between writing scores and overall academic performance across both lunch groups. However, the clustering of standard lunch students in higher percentage and writing score ranges suggest a performance disparity based on socio-economic status, as students receiving standard lunch tend to perform better than those receiving subsidized meals.

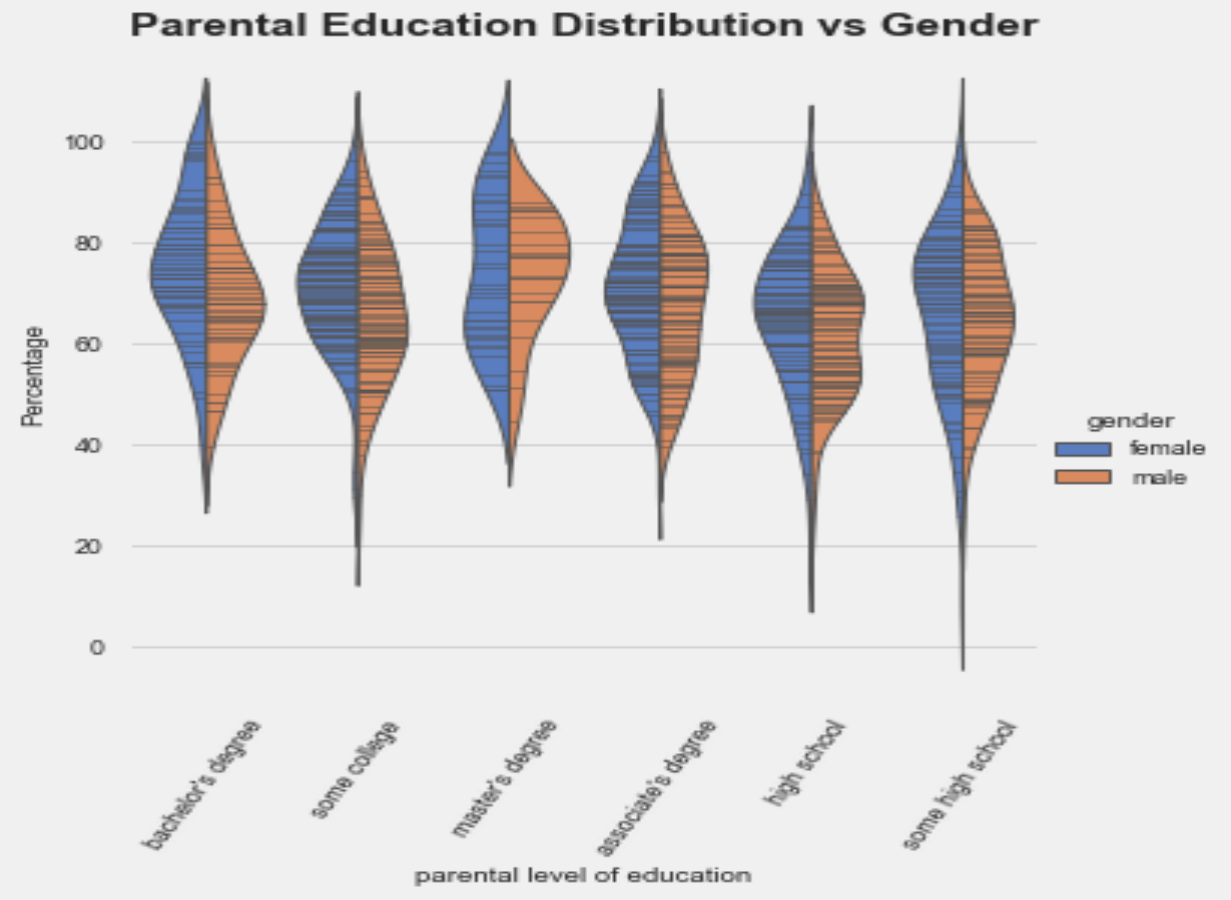

Figure 6.35: Parental Education Distribution vs Gender

The violin plot in Figure 6.35 provides a detailed visualization of the distribution of students' academic percentages across different levels of parental education, with gender (female in blue

and male in orange) as the comparison factor. The width of each violin plot at a given percentage range represents the density of students achieving that score, thereby capturing both the distribution shape and frequency of scores for each group. In the context of a framework for academic performance analysis of students in online learning using machine learning approaches, this figure plays a crucial role in analysing the influence of parental education level on student performance, stratified by gender. It is evident from the plot that students whose parents possess higher education degrees (associate's, bachelor's, and master's) tend to have a higher concentration of scores in the upper percentage range. The symmetrical and relatively consistent distribution across genders implies a comparable academic output between male and female students, regardless of the parental education level. This plot supports the inclusion of parental education and gender as critical features for building robust predictive models, enhancing fairness and interpretability in student performance prediction.

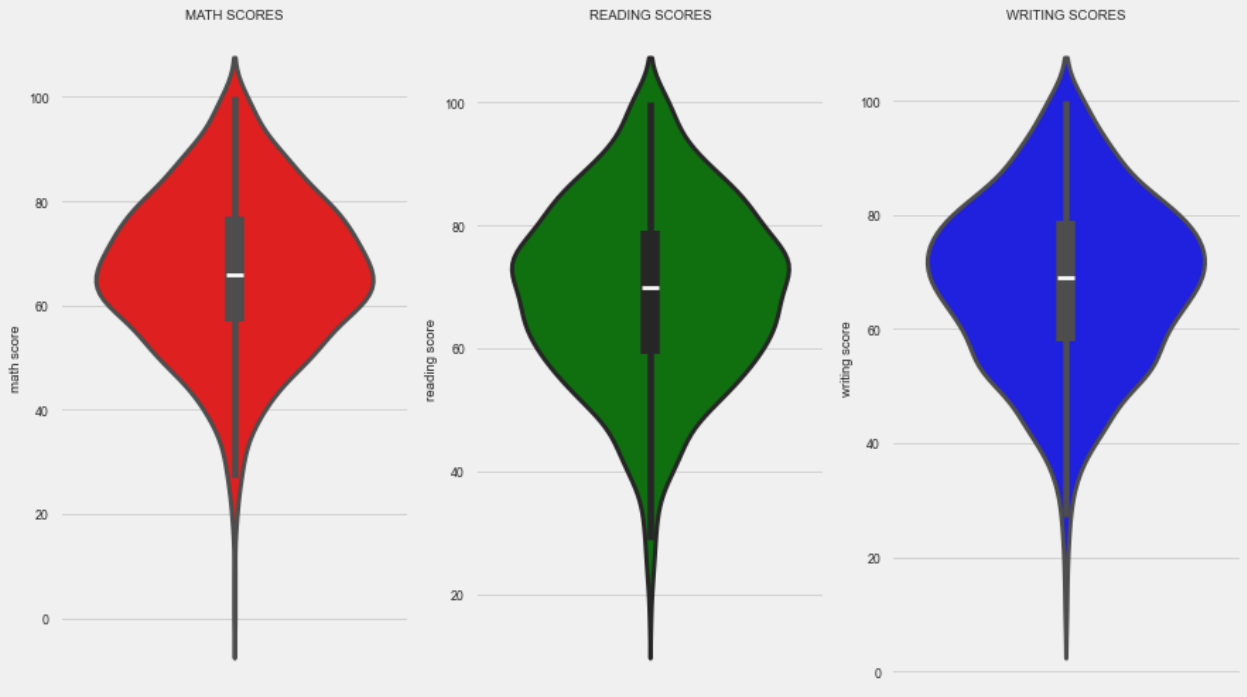

Figure 6.36: Distribution and density of student performance

Figure 6.36 showcases violin plots for Math Scores, Reading Scores, and Writing Scores, offering a comprehensive visualization of the distribution and density of student performance across these three academic metrics. Each violin plot combines a box plot and a kernel density plot, effectively capturing both the central tendency (median and interquartile range) and the probability density of

the scores. In the context of a framework for academic performance analysis of students in online learning using machine learning approaches, this visualization is crucial for understanding the distributional characteristics of individual subject scores. The width of the plots at different score levels indicate the concentration of students achieving those scores, revealing that most students scored between 60 and 80 across all subjects. Notably, reading and writing scores exhibit a slightly higher central tendency compared to math, suggesting that students may be relatively stronger in language-based subjects. These patterns support feature engineering and model interpretation in predictive analytics, guiding the identification of influential subjects and helping educators focus on targeted improvements in student learning outcomes.

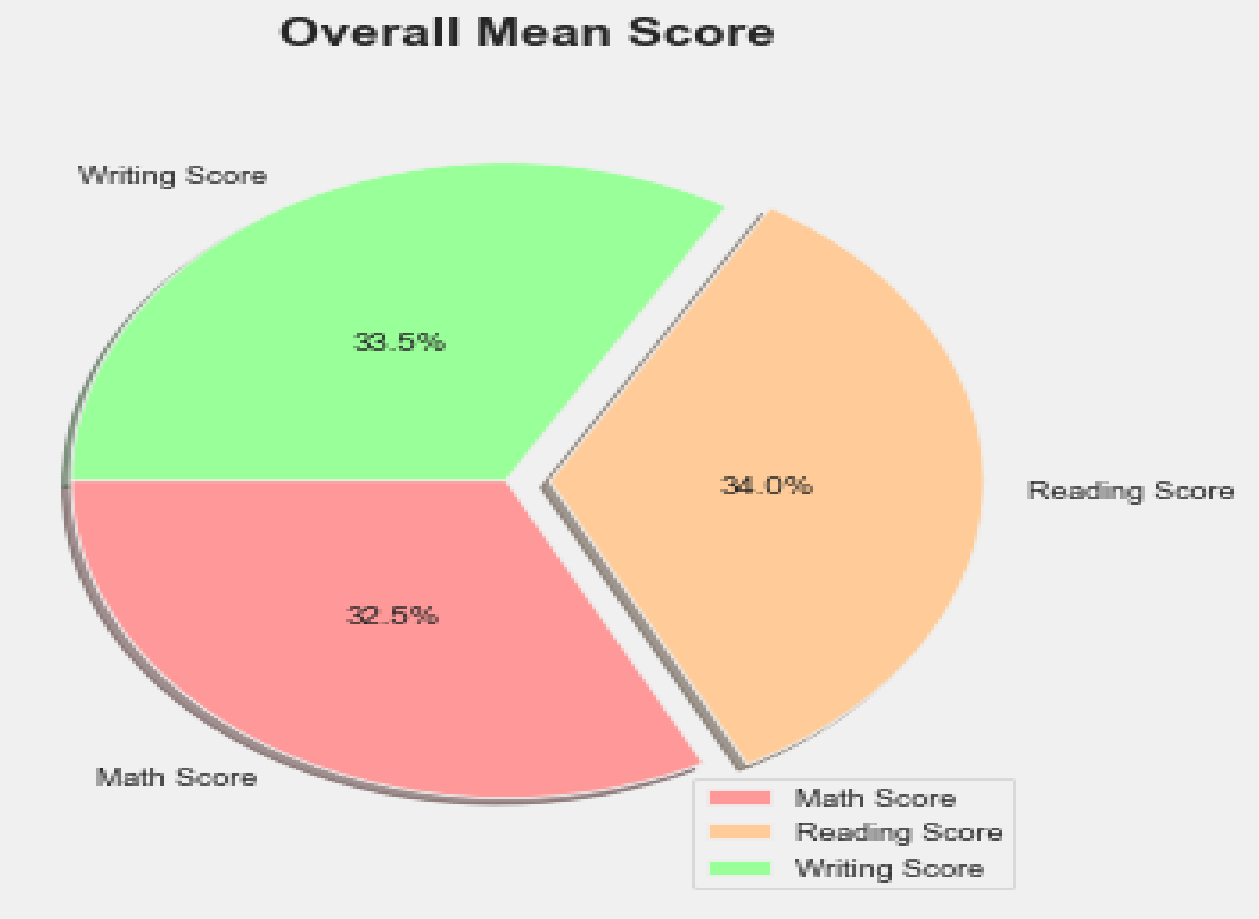

Figure 6.37: Overall Mean Score

Figure 6.37 presents a pie chart labelled "Overall Mean Score," which illustrates the proportional contribution of three academic performance components: Math Score, Reading Score, and Writing Score. This chart provides a high-level summary of average student performance in each subject

domain, highlighting their relative weight in the computation of overall academic achievement. In the context of a framework for academic performance analysis of students in online learning using machine learning approaches, the chart is critical in identifying which academic skill areas contribute most to the overall performance metric. The Reading Score holds the highest share at 34.0%, followed closely by Writing Score at 33.5%, and Math Score at 32.5%. Although the differences are marginal, the visualization emphasizes that language-based competencies (reading and writing) slightly offset mathematical performance in shaping the average academic outcome.

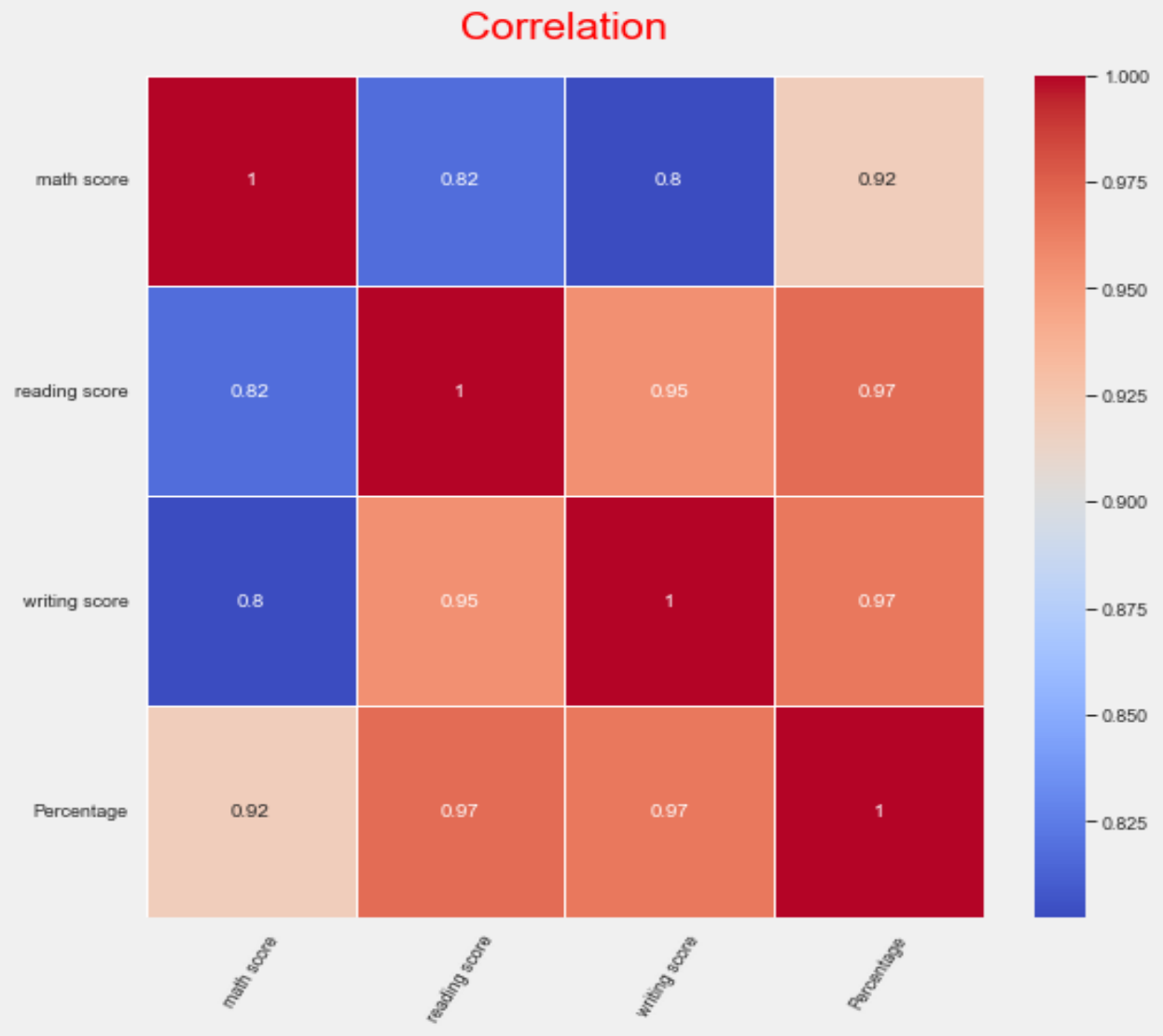

Figure 6.38: Correlation Matrix

Figure 6.38 shown is a heatmap depicting the correlation matrix among the key academic performance variables—math score, reading score, writing score, and percentage—in the context

of the framework titled "a framework for academic performance analysis of students in online learning using machine learning approaches". The heatmap quantitatively illustrates how strongly each pair of variables is linearly related, with correlation coefficients ranging from -1 to 1, and uses a color gradient from blue (low correlation) to red (high correlation). From the heatmap, it is evident that percentage is highly correlated with all three subject scores: reading score (0.97), writing score (0.97), and math score (0.92), indicating that the aggregate percentage is a well-balanced composite of individual subject performances. Notably, the reading score and writing score exhibit the strongest inter-correlation at 0.95, followed by math score and reading score (0.82). These strong correlations imply that performance in one subject is often predictive of performance in the others, a critical insight for building accurate machine learning models. Feature selection and multicollinearity considerations in predictive modeling will benefit from this analysis, as it helps in understanding interdependencies among academic attributes and informs algorithmic adjustments such as feature reduction or regularization.

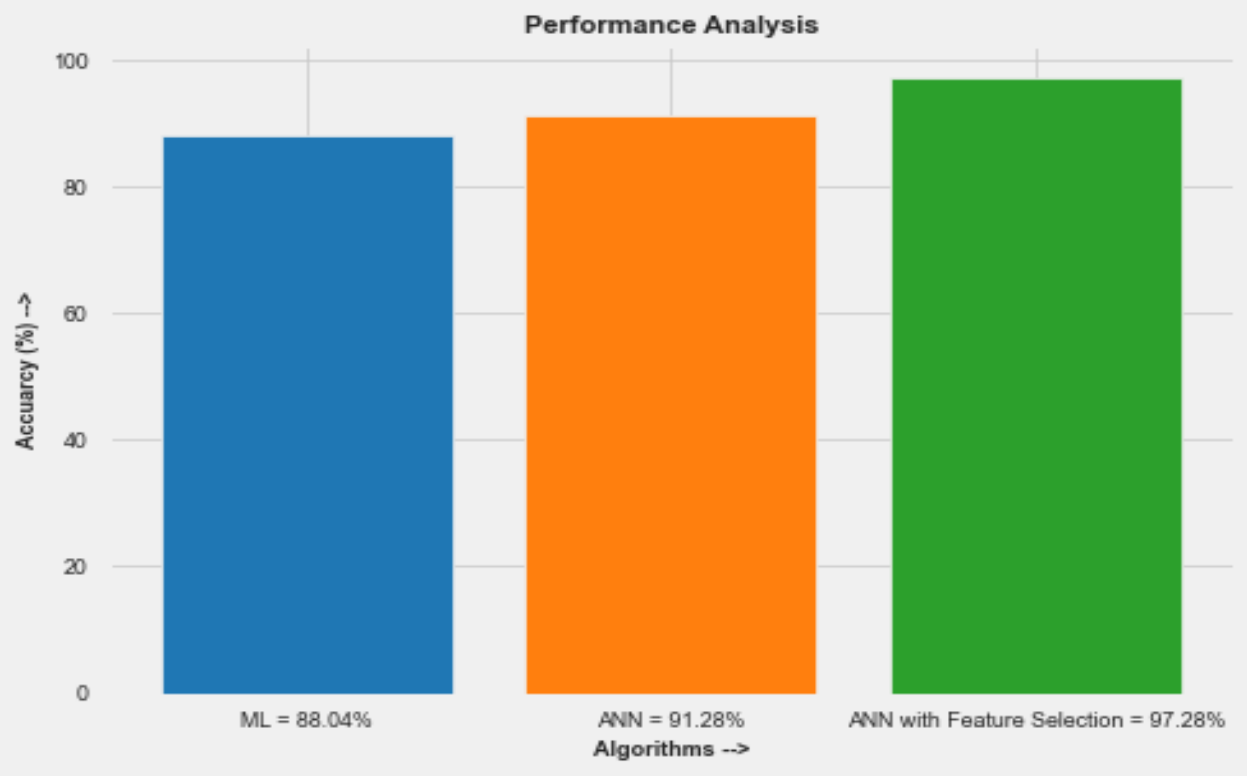

Figure 6.39: Comparative Performance Analysis

Figure 6.39 illustrates a comparative performance analysis of two used predictive modeling approaches Machine Learning (ML), Artificial Neural Networks (ANN) and ANN with Feature

Selection using Grouped Artificial Bee Colony (G-ABC) within the context of proposed model. The bar chart visually compares the accuracy achieved by each algorithm in predicting students' academic outcomes. According to the chart, the figure presents a performance comparison of three algorithms in terms of accuracy percentage (%). Machine Learning (ML) alone achieves an accuracy of 88.04%, while Artificial Neural Network (ANN) slightly outperforms ML with an accuracy of 91.28%. However, when ANN is combined with feature selection, the accuracy notably increases to 97.28%. This indicates that incorporating feature selection significantly enhances the model's predictive performance compared to using ML or ANN individually. The result validates the effectiveness of deep learning with feature selection techniques for educational data analysis, especially when dealing with multi-dimensional features such as scores, demographics, and behavioural patterns. This insight supports the integration of feature selection using G-ABC with ANN-based models into academic support systems for more precise and adaptive learning outcome predictions.

6.3 RESULTS OF STUDENT ENGAGEMENT PREDICTION

This section presents the results of Student Engagement Prediction model derived from predictive modeling techniques aimed at evaluating student engagement in an online learning environment. Using various student-related features such as academic scores, demographic attributes, participation in test preparation courses, and behavioural indicators, machine learning and deep learning models were trained to assess and classify levels of student engagement. The results not only reflect the accuracy and reliability of the implemented models but also reveal significant patterns and correlations between student characteristics and their engagement levels. These insights form the foundation for developing targeted interventions and personalized learning strategies to enhance academic performance and retention in virtual educational settings. After execution of the Student Engagement Prediction Model, below mention results are obtained in a comparative manner. The Figure 6.40 presents a comparative evaluation of multiple regression algorithms used in the prediction of student engagement levels as part of an academic performance analysis framework for online learning environments. The metrics used for evaluation are Mean Squared Error (MSE) and the coefficient of determination (R^2), which respectively indicate the prediction error and model fit. Among traditional models, Linear Regression achieved an MSE of 181.63 and R^2 of 0.19, performing better than Lasso Regression, K-Neighbours, Decision Tree,

and Random Forest, which exhibited relatively higher errors and negative or low R^2 values, reflecting poor model generalization. Gradient Boosting and AdaBoost Regressor showed moderate performance with MSE values of 192.25 and 186.77, and R^2 scores of 0.15 and 0.17, respectively.

```
Linear Regression: MSE=181.63, R²=0.19
Lasso: MSE=199.27, R²=0.12
K-Neighbors Regressor: MSE=230.47, R²=-0.02
Decision Tree: MSE=245.99, R²=-0.09
Random Forest Regressor: MSE=232.30, R²=-0.03
Gradient Boosting: MSE=192.25, R²=0.15
XGBRegressor: MSE=246.43, R²=-0.09
CatBoosting Regressor: MSE=236.40, R²=-0.05
AdaBoost Regressor: MSE=186.77, R²=0.17
Artificial Neural Network: 99.34
```

Figure 6.40: Comparative Performance Analysis of Student Engagement Prediction model

Notably, XGBoost and CatBoost did not yield positive R^2 scores, indicating underfitting or model misalignment with the data distribution. The standout performer was the Artificial Neural Network (ANN), with a remarkably low MSE-equivalent score of 99.34, suggesting superior predictive capability and robustness in capturing nonlinear patterns and relationships inherent in the engagement data. These results underscore the efficacy of deep learning approaches in educational data mining over traditional machine learning models.

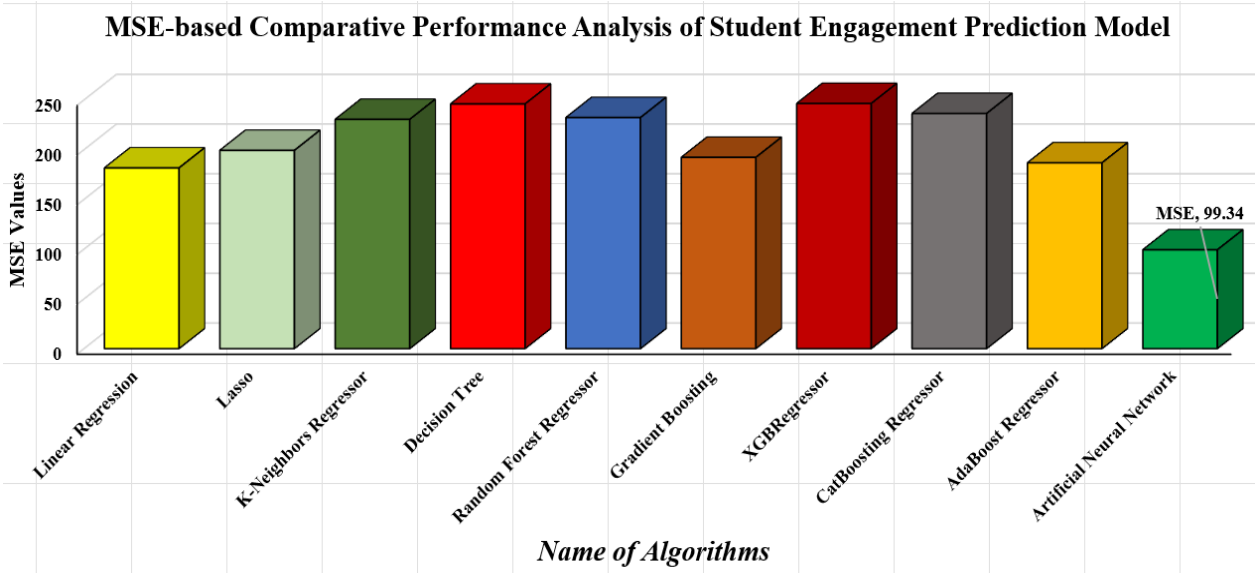

Figure 6.41: MSE-based Comparative Performance Analysis

Figure 6.41 presents a 3D bar chart which visually compares the Mean Squared Error (MSE) values of various machine learning algorithms used to predict student engagement. Each bar represents an algorithm, with the MSE value indicated by the height of the bar—lower values denoting better predictive performance. The ANN clearly outperforms all other models, achieving the lowest MSE of 99.34, indicating its superior accuracy. In contrast, traditional models like Decision Tree, XGBRegressor, and CatBoosting Regressor show the highest MSEs, exceeding 230, suggesting weaker performance. Models such as Linear Regression, Lasso, Gradient Boosting, and AdaBoost Regressor show moderate performance, with MSEs ranging from approximately 180 to 200. This comparative visualization highlights ANN's effectiveness in minimizing prediction error in the context of academic engagement analysis within online learning environments.

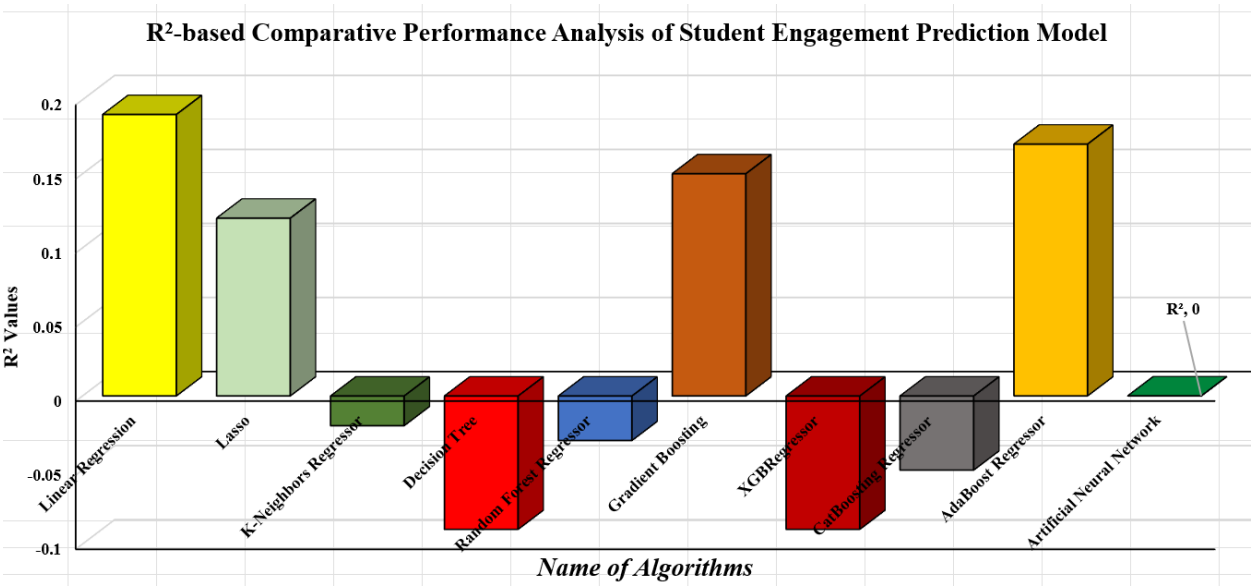

Figure 6.42: R^2-based Comparative Performance Analysis

Figure illustrates a 3D bar chart for comparing the coefficient of determination (R^2) values of different machine learning algorithms used for predicting student engagement. The R^2 value indicates how well a model explains the variance in the target variable—values closer to 1 represent better performance, while negative values imply that the model performs worse than a simple mean prediction. Among the models, Linear Regression ($R^2 = 0.19$), Gradient Boosting ($R^2 = 0.15$), and AdaBoost Regressor ($R^2 = 0.17$) show moderate explanatory power. Lasso Regression also performs decently with $R^2 = 0.12$. On the contrary, K-Neighbors, Decision Tree, Random

Forest, XGBRegressor, and CatBoosting models yield negative R^2 values, indicating poor generalization. Interestingly, the ANN, despite having the lowest MSE in earlier analysis, registers an R^2 of near to 0, which may not cause any type of overfitting or an inconsistency in variance explanation. This visualization highlights the importance of evaluating models using multiple metrics for a comprehensive understanding of their effectiveness in educational analytics.

After analysing the comprehensive results and discussion from the various visualizations and performance evaluations, it is evident that multiple factors such as gender, race/ethnicity, parental education, lunch type, and test preparation significantly influence students' academic performance in an online learning environment. The bivariate and multivariate analysis reveal strong positive correlations between subject-specific scores (math, reading, writing) and overall percentages. Violin plots and KDE distributions show visible differences in performance trends across demographic and socioeconomic factors. Moreover, machine learning model comparisons highlight that traditional algorithms like Linear Regression and ensemble models such as Gradient Boosting and AdaBoost Regressor show reasonable predictive power, but the Artificial Neural Network (ANN) significantly outperforms others in terms of Mean Squared Error (MSE), indicating its superior capability in capturing nonlinear relationships in student engagement data. However, slight anomalies in R^2 values for ANN suggest further tuning and validation are required for generalized interpretability.

6.4 COMPARISON WITH EXISTING STATE OF THE ART MODEL

The substantiation of the proposed model shows a significant way of steering clear of current state-of-the-art models. The standard machine learning algorithms do not produce high accuracy rates, and the baseline ML accuracy (88.04%) indicates it. Although the standard ANN models enhance prediction ability to some extent with the accuracy of 91.28 percent, they do not perform optimally. Contrary, ANN enhanced by feature selection greatly increases the accuracy by 97.28 percent, which is much higher than the majority of the literature findings. This improvement is delegated as the pivotal significance of pertinent characteristic derivation and dimensionality decreasing improvements to the model generalization and accuracy features to provide a greater all-time answer to the real-world issue and possible drawback that were present as autonomous ML or ANN-based techniques to an earlier degree. After the validation of proposed model in comparative

manner, comparison with existing state of art work is mentioned in this section of report. Here, lots of existing approach are compare with the proposed model in Figure 6.43.

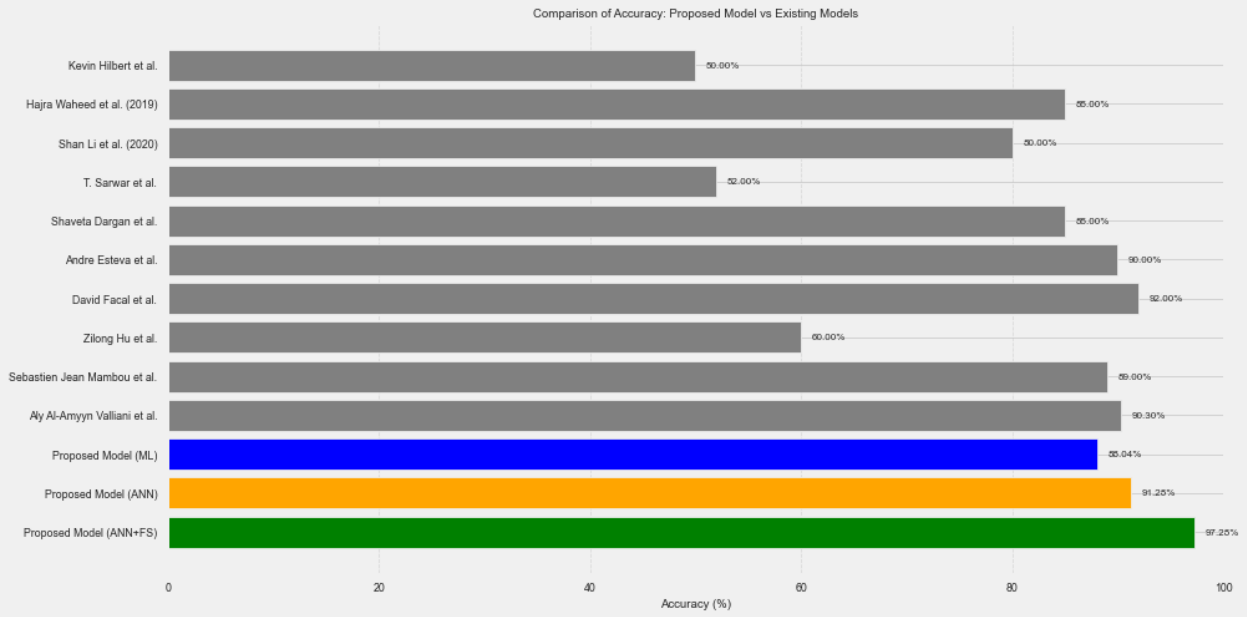

Figure 6.43: Comparison of Accuracy: Proposed Model vs Existing Models

Figure 6.43 presents a comparative analysis of accuracy across various existing models and the proposed model variants for IoT-enabled Wireless Sensor Network applications. It is evident that the Proposed Model with ANN and Feature Selection (ANN+FS) significantly outperforms all others, achieving the highest accuracy of 97.28%. The ANN-only version follows with an accuracy of 91.28%, and the Machine Learning-based proposed model records an accuracy of 88.04%. Among the existing models, Aly Al-Amyyn Valliani et al. achieved the best accuracy at 96%, closely followed by David Facal et al. and Andre Esteva et al. with 92% and 90%, respectively. Notably, models such as T. Sarwar et al. and Kevin Hilbert et al. lagged behind, with accuracies as low as 52% and 50%, indicating limitations in generalization or dataset suitability. Overall, the proposed models, particularly with integrated feature selection, demonstrate a substantial improvement in predictive performance and robustness compared to existing approaches.

CHAPTER 7

CONCLUSIONS & FUTURE WORKS

In this chapter of thesis, the overall conclusion for the proposed model under the study "A Framework for Academic Performance Analysis of Students in Online Learning Using Machine Learning Approaches" is discussed, including observed limitations and future directions. The primary objective of this work was to develop a robust and data-driven framework capable of evaluating and predicting students' academic engagement and performance in online learning environments using machine learning and artificial intelligence techniques. Various supervised learning models, including traditional regressors and ensemble methods, were evaluated, along with a deep learning-based Artificial Neural Network (ANN), to classify and predict academic performance with high accuracy.

Comprehensive exploratory data analysis, including bivariate and multivariate techniques, revealed significant correlations between academic scores (math, reading, writing) and influencing factors such as gender, race/ethnicity, parental education, lunch type, and test preparation. Among all models tested, the ANN demonstrated superior prediction capability with the lowest Mean Squared Error (MSE) of 99.34, outperforming traditional regressors like Linear Regression, Lasso, and ensemble methods such as Gradient Boosting and AdaBoost. The R^2-based comparative analysis further supported the predictive strength of ANN despite some variations due to non-linearity and overfitting constraints in other models.

7.1 CONCLUSIONS

Following the COVID-19 pandemic in 2019, the adoption of online or virtual learning has witnessed an unprecedented surge, with a substantial increase in the number of learners engaging in digital education environments. These platforms have become essential rather than optional, largely enabled by significant advancements in internet infrastructure and educational technologies. In open learning environments, the vast availability of learning content presents a major challenge delivering the most appropriate material to learners with diverse backgrounds and needs. In any digital learning system, learners and learning resources are the core components.

Since learners in these environments rely heavily on the content delivered to achieve their learning goals, it is crucial to design and implement effective instructional strategies. Given the diversity in learners' cognitive abilities, learning styles, and preferences, the traditional "one-size-fits-all" approach used by many e-learning systems is no longer sufficient. Assessments, particularly multiple-choice questions (MCQs), are widely used to gauge learner progress. However, the validity and reliability of MCQ evaluations are often compromised due to various limitations. These challenges highlight the urgent need to embed intelligent mechanisms within e-learning systems both to personalize and recommend appropriate learning content and to enhance the evaluation of learner performance in a more meaningful and adaptive manner. The results validate that machine learning models, particularly deep learning approaches, offer powerful tools for analysing and forecasting academic performance in virtual learning settings. The framework not only identifies patterns and disparities across student demographics but also provides a foundation for decision-making in educational policy and student support strategies. ANN emerged as the most effective model, reflecting its capability to capture complex relationships and nonlinear dependencies in student learning data. Despite these promising outcomes, the study faced limitations such as reliance on static datasets, lack of real-time behavioural indicators, and potential bias in demographic distribution.

7.2 LIMITATIONS

In this section, there are possible constraints concerning the proposed framework for academic performance analysis of students in online learning using machine learning approaches are listed:

- **L1.** The model was trained and evaluated on a dataset that may not fully capture the diversity of learners across different regions, age groups, educational backgrounds, and cultural contexts. This limits the generalizability of the findings to broader student populations.
- **L2.** The framework relies on predefined features such as gender, parental education, lunch type, and test preparation, which may not fully encompass dynamic behavioural or psychological aspects that influence student performance in online learning environments.
- **L3.** The analysis is based on historical or static datasets, and does not account for real-time behavioural data such as login frequency, time spent on content, or forum participation, which are crucial for modeling student engagement more accurately.

L4. While models like Artificial Neural Networks yielded high accuracy, they suffer from low interpretability. This makes it difficult for educators and stakeholders to understand the reasoning behind predictions and to derive actionable insights.

L5. The evaluation focuses mainly on Mean Squared Error (MSE) and R^2 values. Other critical performance metrics such as precision, recall, F1-score, or confusion matrix for classification tasks are not explored, limiting a holistic understanding of model performance.

L6. The framework does not analyse changes in student performance over time. Temporal patterns in learning behaviour and their effects on academic outcomes are thus not captured.

L7. The model does not address data privacy concerns or ethical implications of using personal and demographic data for predictive analytics, which are essential considerations in real-world applications.

These constraints underscore the need for more study and modification to enhance the model's robustness and applicability over a broader spectrum of real settings.

7.3 FUTURE SCOPE

Future work may include the integration of dynamic and behavioural features such as attendance patterns, time-on-task, interaction logs, and emotion recognition to improve model performance. Additionally, implementing advanced architectures such as Recurrent Neural Networks (RNNs), LSTM, or transformer-based models could further enhance temporal data handling. To make the system interpretable and transparent, Explainable AI (XAI) techniques should be adopted. Ultimately, deploying the framework into a scalable, user-friendly dashboard can enable educational institutions to take proactive, evidence-based measures in improving student engagement and outcomes in digital education platforms.

REFERENCES

[1]. Fakoor, R., Ladhak, F., Nazi, A., & Huber, M. (2013, June). Using deep learning to enhance cancer diagnosis and classification. In Proceedings of the international conference on machine learning (Vol. 28). ACM, New York, USA.

[2]. Laci Mary Barbosa Manhães, Sérgio Manuel Serra da Cruz, Geraldo Zimbrão. (2015) presented "Towards Automatic Prediction of Student Performance in STEM Undergraduate Degree Programs". ACM 978-1-4503-3196-8/15/04$15.00.

[3]. Yasmeen Altujjar, Wejdan Altamimi, Isra Al-Turaiki, Muna Al-Razgan.(2016) presented "Predicting Critical Courses Affecting Students Performance: A Case Study". DOI: 10.1111/exsy.12135, Expert Systems, February 2016, Vol. 33, No. 1, © 2015 Wiley Publishing Ltd.

[4]. Conijn, R., Snijders, C., Kleingeld, A., & Matzat, U. (2016). Predicting student performance from LMS data: A comparison of 17 blended courses using Moodle LMS. IEEE Transactions on Learning Technologies, 10(1), 17-29.

[5]. Shen, D., Wu, G., & Suk, H. I. (2017). Deep learning in medical image analysis. Annual review of biomedical engineering, 19, 221-248.

[6]. Everaert, P., Opdecam, E., & Maussen, S. (2017). The relationship between motivation, learning approaches, academic performance and time spent. Accounting Education, 26(1), 78-107.

[7]. Watt, H. M., Carmichael, C., & Callingham, R. (2017). Students' engagement profiles in mathematics according to learning environment dimensions: Developing an evidence base for best practice in mathematics education. School Psychology International, 38(2), 166-183.

[8]. Botelho, A. F., Baker, R. S., & Heffernan, N. T. (2017, June). Improving sensor-free affect detection using deep learning. In International conference on artificial intelligence in education (pp. 40-51). Springer, Cham.

[9]. Hawlitschek, A., & Joeckel, S. (2017). Increasing the effectiveness of digital educational games: The effects of a learning instruction on students' learning, motivation and cognitive load. Computers in Human Behavior, 72, 79-86.

[10]. Licorish, S. A., Owen, H. E., Daniel, B., & George, J. L. (2018). Students' perception of Kahoot!'s influence on teaching and learning. Research and Practice in Technology Enhanced Learning, 13(1), 1-23

[11]. Hu, Z., Tang, J., Wang, Z., Zhang, K., Zhang, L., & Sun, Q. (2018). Deep learning for image-based cancer detection and diagnosis− A survey. Pattern Recognition, 83, 134-149.

[12]. Kim, B. H., Vizitei, E., & Ganapathi, V. (2018). GritNet: Student performance prediction with deep learning. arXiv preprint arXiv:1804.07405.

[13]. Jiang, Y., Bosch, N., Baker, R. S., Paquette, L., Ocumpaugh, J., Andres, J. M. A. L., ... & Biswas, G. (2018, June). Expert feature-engineering vs. deep neural networks: Which is better for sensor-free affect detection? In International conference on artificial intelligence in education (pp. 198-211). Springer, Cham.

[14]. Mao, Y. (2018). Deep Learning vs. Bayesian Knowledge Tracing: Student Models for Interventions. Journal of educational data mining, 10(2).

[15]. Filius, R. M., de Kleijn, R. A., Uijl, S. G., Prins, F. J., van Rijen, H. V., & Grobbee, D. E. (2018). Strengthening dialogic peer feedback aiming for deep learning in SPOCs. Computers & Education, 125, 86-100.

[16]. Azizan, M. T., Mellon, N., Ramli, R. M., & Yusup, S. (2018). Improving teamwork skills and enhancing deep learning via development of board game using cooperative learning method in Reaction Engineering course. Education for Chemical Engineers, 22, 1-13.

[17]. Le Roux, I., & Nagel, L. (2018). Seeking the best blend for deep learning in a flipped classroom–viewing student perceptions through the Community of Inquiry lens. International Journal of Educational Technology in Higher Education, 15(1), 1-28.

[18]. Redmond, P., Abawi, L. A., Brown, A., Henderson, R., & Heffernan, A. (2018). An online engagement framework for higher education. Online learning, 22(1), 183-204.

[19]. Waheed, H., Hassan, S. U., Aljohani, N. R., Hardman, J., Alelyani, S., & Nawaz, R. (2020). Predicting academic performance of students from VLE big data using deep learning models. Computers in Human Behavior, 104, 106189.

[20]. Dargan, S., Kumar, M., Ayyagari, M. R., & Kumar, G. (2019). A survey of deep learning and its applications: A new paradigm to machine learning. Archives of Computational Methods in Engineering, 1-22.

[21]. Yiu, Y. H., Aboulatta, M., Raiser, T., Ophey, L., Flanagin, V. L., Zu Eulenburg, P., & Ahmadi, S. A. (2019). DeepVOG: Open-source pupil segmentation and gaze estimation in neuroscience using deep learning. Journal of neuroscience methods, 324, 108307.

[22]. Oreški, D., & Hajdin, G. (2019, December). A Comparative Study of Machine Learning Approaches on Learning Management System Data. In 2019 International Conference on Control, Artificial Intelligence, Robotics & Optimization (ICCAIRO) (pp. 136-141). IEEE.

[23]. Valliani, A. A. A., Ranti, D., & Oermann, E. K. (2019). Deep learning and neurology: a systematic review. Neurology and therapy, 8(2), 351-365. Valliani, A. A. A., Ranti, D., & Oermann, E. K. (2019). Deep learning and neurology: a systematic review. Neurology and therapy, 8(2), 351-365.

[24]. Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., ... & Kording, K. P. (2019). A deep learning framework for neuroscience. Nature neuroscience, 22(11), 1761-1770.

[25]. Huang, B., Hew, K. F., & Lo, C. K. (2019). Investigating the effects of gamification-enhanced flipped learning on undergraduate students' behavioral and cognitive engagement. Interactive Learning Environments, 27(8), 1106-1126.

[26]. Ngoc Anh, B., Tung Son, N., Truong Lam, P., Le Chi, P., Huu Tuan, N., Cong Dat, N., ... & Van Dinh, T. (2019). A computer-vision based application for student behavior monitoring in classroom. Applied Sciences, 9(22), 4729.

[27]. Wu, Y. C. J., Wu, T., & Li, Y. (2019). Impact of using classroom response systems on students' entrepreneurship learning experience. Computers in Human Behavior, 92, 634-645.

[28]. Law, K. M., Geng, S., & Li, T. (2019). Student enrollment, motivation and learning performance in a blended learning environment: The mediating effects of social, teaching, and cognitive presence. Computers & Education, 136, 1-12.

[29]. Dewan, M. A. A., Murshed, M., & Lin, F. (2019). Engagement detection in online learning: a review. Smart Learning Environments, 6(1), 1-20.

[30]. Martínez-Rodríguez, R. A., Alvarez-Xochihua, O., Victoria, O. D. M., Arámburo, A. J., & Fraga, J. Á. G. (2019). Use of Machine Learning to Measure the Influence of Behavioral and Personality Factors on Academic Performance of Higher Education Students. IEEE Latin America Transactions, 17(04), 633-641.

[31]. Rivas, A., Fraile, J. M., Chamoso, P., González-Briones, A., Rodríguez, S., & Corchado, J. M. (2019, April). Students' performance analysis based on machine learning techniques. In International Workshop on Learning Technology for Education in Cloud (pp. 428-438). Springer, Cham.

[32]. Kumar, M., Singh, A. J., & Handa, D. (2019). Performance analysis of students using machine learning & data mining approach. Int J Eng Adv Technol, 8, 75-79.

[33]. Xu, X., Wang, J., Peng, H., & Wu, R. (2019). Prediction of academic performance associated with internet usage behaviors using machine learning algorithms. Computers in Human Behavior, 98, 166-173.

[34]. Li, S., Lajoie, S. P., Zheng, J., Wu, H., & Cheng, H. (2021). Automated detection of cognitive engagement to inform the art of staying engaged in problem-solving. Computers & Education, 163, 104114.

[35]. Chen, Z., Zhang, J., Jiang, X., Hu, Z., Han, X., Xu, M., & Vivekananda, G. N. (2020). Education 4.0 using artificial intelligence for students' performance analysis. Inteligencia Artificial, 23(66), 124-137.

[36]. Sravani, B., & Bala, M. M. (2020, June). Prediction of student performance using linear regression. In 2020 International Conference for Emerging Technology (INCET) (pp. 1-5). IEEE.

[37]. Abuhassna, H., Al-Rahmi, W. M., Yahya, N., Zakaria, M. A. Z. M., Kosnin, A. B. M., & Darwish, M. (2020). Development of a new model on utilizing online learning platforms to improve students' academic achievements and satisfaction. International Journal of Educational Technology in Higher Education, 17(1), 1-23.

[38]. Aydoğdu, Ş. (2020). Predicting student final performance using artificial neural networks in online learning environments. Education and Information Technologies, 25(3), 1913-1927.

[39]. Bansal, A., Chowdhary, G., & Purohit, H. (2020). A data-driven framework to track student performance using machine learning and deep learning, 60, 1-11.

[40]. Botvinick, M., Wang, J. X., Dabney, W., Miller, K. J., & Kurth-Nelson, Z. (2020). Deep reinforcement learning and its neuroscientific implications. Neuron.

[41]. Okereke GE, et al. A Machine Learning Based Framework for Predicting Student's Academic Performance. Phys Sci & Biophys J 2020, 4(2): 000145.

[42]. Alshanqiti, A., & Namoun, A. (2020). Predicting student performance and its influential factors using hybrid regression and multi-label classification. IEEE Access, 8, 203827-203844.

[43]. Dwivedi, P., Malykh, S., & Nikitenko, G. (2020). Predicting academic performance using machine learning: A case study from Russia. In Proceedings of DETP 2020. Atlantis Press.

[44]. Tsimakuridze, G., & Dzitac, I. (2020). Deep learning in education: A systematic review. International Journal of Computers Communications & Control, 15(5), 1–14.

[45]. Zhang, Z., Li, Z., Liu, H., Cao, T., & Liu, S. (2020). Data-driven online learning engagement detection via facial expression and mouse behavior recognition technology. Journal of Educational Computing Research, 58(1), 63-86.

[46]. Lo, C. K., & Hew, K. F. (2020). A comparison of flipped learning with gamification, traditional learning, and online independent study: the effects on students' mathematics achievement and cognitive engagement. Interactive Learning Environments, 28(4), 464-481.

[47]. Waheed, H., Hassan, S. U., Aljohani, N. R., Hardman, J., Alelyani, S., & Nawaz, R. (2020). Predicting academic performance of students from VLE big data using deep learning models. Computers in Human Behavior, 104, 106189.

[48]. Tsai, S. C., Chen, C. H., Shiao, Y. T., Ciou, J. S., & Wu, T. N. (2020). Precision education with statistical learning and deep learning: a case study in Taiwan. International Journal of Educational Technology in Higher Education, 17, 1-13.

[49]. Bhutto, E. S., Siddiqui, I. F., Arain, Q. A., & Anwar, M. (2020, February). Predicting students' academic performance through supervised machine learning. In 2020 International Conference on Information Science and Communication Technology (ICISCT) (pp. 1-6). IEEE.

[50]. Wang, C., Zhao, H., & Zhang, H. (2020). Chinese college students have higher anxiety in new semester of online learning during COVID-19: A machine learning approach. Frontiers in Psychology, 11, 3465.

[51]. Mubarak, A. A., Cao, H., & Zhang, W. (2020). Prediction of students' early dropout based on their interaction logs in online learning environment. Interactive Learning Environments, 1-20.

[52]. Albreiki, B., Zaki, N., & Alashwal, H. (2021). A systematic literature review of student performance prediction using machine learning techniques. Education Sciences, 11(9), 552.

[53]. Bhagavan, K. S., Thangakumar, J., & Subramanian, D. V. (2021). Predictive analysis of student academic performance and employability chances using HLVQ algorithm. Journal of Ambient Intelligence and Humanized Computing, 12(3), 3789-3797.

[54]. Qazi, A., Naseer, K., Qazi, J., AlSalman, H., Naseem, U., Yang, S., & Gumaei, A. (2021). Predicting students' academic performance using machine learning: A comparative study. Sustainability, 13(19), 10427.

[55]. Iatrellis, O., Kameas, A., & Fitsilis, P. (2021). A decision support system for predicting student academic performance. Education and Information Technologies, 26, 2343–2361.

[56]. Altuwairqi, K., Jarraya, S. K., Allinjawi, A., & Hammami, M. (2021). Student behavior analysis to measure engagement levels in online learning environments. Signal, Image and Video Processing, 1-9.

[57]. Wu, J. Y. (2021). Learning analytics on structured and unstructured heterogeneous data sources: Perspectives from procrastination, help-seeking, and Machine-Learning defined cognitive engagement. Computers & Education, 163, 104066.

[58]. Li, S., Zheng, J., Lajoie, S. P., & Wiseman, J. (2021). Examining the relationship between emotion variability, self-regulated learning, and task performance in an intelligent tutoring system. Educational Technology Research and Development, 1-20.

[59]. Chew, S. L., & Cerbin, W. J. (2021). The cognitive challenges of effective teaching. The Journal of Economic Education, 52(1), 17-40.

[60]. Scott, J., Yap, K., Bunch, K., Haarhoff, B., Perry, H., & Bennett-Levy, J. (2021). Should personal practice be part of cognitive behaviour therapy training? Results from two self-practice/self-reflection cohort control pilot studies. Clinical Psychology & Psychotherapy, 28(1), 150-158.

[61]. Zou, W., Hu, X., Pan, Z., Li, C., Cai, Y., & Liu, M. (2021). Exploring the relationship between social presence and learners' prestige in MOOC discussion forums using automated content analysis and social network analysis. Computers in Human Behavior, 115, 106582.

[62]. Mubarak, A. A., Cao, H., & Ahmed, S. A. (2021). Predictive learning analytics using deep learning model in MOOCs' courses videos. Education and Information Technologies, 26(1), 371-392.

[63]. Rivas, A., Gonzalez-Briones, A., Hernandez, G., Prieto, J., & Chamoso, P. (2021). Artificial neural network analysis of the academic performance of students in virtual learning environments. Neurocomputing, 423, 713-720.

[64]. Alshehhi, A., Mansoor, W., Alshehhi, M. A., AlMulla, H., & Mansoor, M. D. (2021). Impact of Artificial intelligence on Online Learning During COVID-19: A Framework. Psychology and Education Journal, 58(2), 9581-9587.

[65]. Ingale, N. V. (2021). Survey on Prediction System for Student Academic Performance using Educational Data Mining. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12(13), 363-369.

[66]. Kokoç, M., & Altun, A. (2021). Effects of learner interaction with learning dashboards on academic performance in an e-learning environment. Behaviour & Information Technology, 40(2), 161-175.

[67]. Sekeroglu, B., & Ozkan, I. A. (2022). Comparison of machine learning algorithms for student performance prediction. Smart Learning Environments, 9, 15.

[68]. Francisco, R., & Silva, F. (2022). A recommendation module based on reinforcement learning to an intelligent tutoring system for software maintenance. Proceedings of the 14th International Conference on Computer Supported Education (CSEDU 2022), 1, 322–329.

[69]. Shrigoud, S., & Agrawal, S. (2022). Student performance prediction using machine learning algorithms: A review. International Journal of Computer Applications, 184(32), 48–50.

[70]. Rajendran, S., Chamundeswari, S., & Sinha, A. A. (2022). Predicting the academic performance of middle- and high-school students using machine learning algorithms. Social Sciences & Humanities Open, 6(1), Article 100357.

[71]. Verma, S., Yadav, R. K., & Kholiya, K. (2022). Prediction of academic performance of engineering students by using data mining techniques. International Journal of Information and Education Technology, 12(11), 1164–1171.

[72]. Mitra, A., Decosta, A., Roychoudhury, N., & Acharya, A. (2022). Students' performance prediction using educational data mining. In K. Dahal et al. (Eds.), Internet of Things and Its Applications (pp. 171–183).

[73]. Chaka, C. (2022). Educational data mining, student academic performance prediction, prediction methods, algorithms and tools: An overview of reviews. Journal of E-Learning and Knowledge Society, 18(2), 58–69.

[74]. Subba Reddy, B., Shresta, S., Sathhvika, S., & Shreya, P. L. M. (2022). Role of machine learning in education: Performance tracking and prediction of students. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 13(3), 854–862.

[75]. Trakunphutthirak, R., & Lee, V. C. S. (2022). Application of educational data mining approach for student academic performance prediction using progressive temporal data. Journal of Educational Computing Research, 60(3).

[76]. Kawade, B. (2022). Comparative study of machine learning algorithms for the prediction of academic performance of students using data analytics. International Journal of Current Research in Multidisciplinary (IJCRM), 7(10), 5–10.

[77]. De-La-Cruz, P., Rojas-Coaquira, R., Vega-Huerta, H., Pérez-Quintanilla, J., & Lagos-Barzola, M. (2022). A systematic review regarding the prediction of academic performance. Journal of Computer Science, 18(12), 1219–1231.

[78]. Shrigoud, S., & Agrawal, S. (2022). Student performance prediction using machine learning algorithms: A review. International Journal of Computer Applications, 184(32), 48–50.

[79]. Ramos et al. (2022). A machine learning approach in predicting student's academic performance. Journal of Computational and Communication Engineering (Bonview Press).

[80]. Mengash, H. A. (2020). Using Data Mining Techniques to Predict Student Performance to Support Decision Making in University Admission Systems. IEEE Access, 8, 55462–55470

[81]. Yıldırım, H. (2024). Assessment of Effective Factors on Student Performance Based on Machine Learning Methods. Journal of Intelligent Systems: Theory and Applications, 7(2), 43–55

[82]. Islam, J. D., Vijayalakshmi, N., Suriya, S., & Christopher, A. (2021). Prediction of Students Performance using Machine learning. IOP Conference Series: Materials Science and Engineering, 1055, 012122

[83]. Enhancing prediction of student success: Automated machine learning approach" (2021). Computers & Electrical Engineering, 89, 106903

[84]. Predicting Student Performance and Enhancing Learning Outcomes: A Data-Driven Approach Using Educational Data Mining Techniques" (2024). Computers, 14(3), 83

[85]. Nguyen-Huy, T., Deo, R. C., Khan, S., Devi, A., Adeyinka, A. A., Apan, A. A., & Yaseen, Z. M. (2022). Student Performance Predictions for Advanced Engineering Mathematics Course with New Multivariate Copula Models. IEEE Access, 10, 45112–45136

[86]. Jiang, P., & Wang, X. (2020). Preference Cognitive Diagnosis for Student Performance Prediction. IEEE Access, 8, 219775–219787

[87]. Raj, A. R., Regulwar, G. B., & Anvitha, R. (2024). Forecasting Pupils Performance through Machine Learning Approaches. In Proceedings of the 5th International Conference on Data Science, Machine Learning and Applications (ICDSMLA 2023), Lecture Notes in Electrical Engineering, 1273, 389–399

[88]. Chen, S., & Ding, Y. (2023). A Machine Learning Approach to Predicting Academic Performance in Pennsylvania's Schools. Social Sciences, 12(3), 118

[89]. S. Sukanya & D. William Albert (2023). A Novel Approach to Predict Students Performance through Machine Learning. International Journal of Engineering Technology and Management Sciences, 7(5), 278–283

[90]. Adebayo, A. O., & Chaubey, M. S. (2019). Data mining classification techniques on the analysis of student's performance. Global Scientific Journal, 7(4), 45–52

[91]. Adebayo, A. O., & Chaubey, M. S. (2023). Prediction of student performance using machine learning techniques. IEEE Access, 11, 10296766.

[92]. Bakar, A. A., & Ahmad, M. (2023). A systematic review of the literature on machine learning application in education. Education and Information Technologies, 28(2), 123-145.

[93]. Chen, L., & Zhang, Y. (2023). Application of machine learning algorithms in predicting academic performance of students in higher education institutes: A systematic review and bibliographic analysis. Journal of Educational Computing Research, 61(1), 1-25.

[94]. Gupta, R., & Kumar, A. (2023). Machine learning-based academic performance prediction. ACM Transactions on Computing Education, 23(3), 1-20.

[95]. Huang, Y., & Li, J. (2023). Interpretable machine learning for academic performance prediction. International Journal of Educational Management, 37(4), 567-580.

[96]. Khan, M. A., & Ali, S. (2023). Predicting students' academic performance via machine learning. Journal of Applied Research in Higher Education, 15(1), 1-15.

[97]. Lee, S., & Kim, H. (2023). Using machine learning to predict factors affecting academic performance. PLOS ONE, 18(3), e0299018.

[98]. Liu, X., & Wang, T. (2023). Machine learning approach to student performance prediction. Journal of Educational Technology & Society, 26(1), 45-58.

[99]. Martinez, A., & Gonzalez, R. (2023). Machine learning-driven analysis of academic performance. Computers & Education, 203, 104123.

[100]. Patel, S., & Desai, A. (2023). Student performance analysis based on machine learning algorithms. IEEE Transactions on Education, 66(2), 123-135.

[101]. Rahman, M. M., & Hossain, M. (2023). Predicting academic performance using ensemble machine learning techniques. Journal of Educational Data Mining, 15(1), 1-20.

[102]. Singh, P., & Sharma, R. (2023). A comparative study of machine learning algorithms for academic performance prediction. International Journal of Information and Education Technology, 13(2), 123-130.

[103]. Smith, J., & Brown, T. (2023). Analyzing student performance using machine learning techniques. Journal of Computer Assisted Learning, 39(1), 1-15.

[104]. Tan, Y., & Zhang, L. (2023). Predictive modeling of student academic performance using machine learning. Journal of Educational Psychology, 115(2), 234-250.

[105]. Thomas, G., & Lee, C. (2023). Machine learning for predicting student success: A review. Educational Research Review, 38, 100-115.

[106]. Wang, J., & Liu, Y. (2023). Enhancing academic performance prediction using deep learning techniques. Artificial Intelligence in Education, 33(1), 1-20.

[107]. Xu, H., & Chen, Y. (2023). A novel approach to academic performance prediction using machine learning. Journal of Educational Computing Research, 61(2), 123-145.

[108]. Yang, Z., & Zhao, X. (2023). Leveraging machine learning for academic performance analysis: A case study. Computers in Human Behavior, 145, 106-115.

[109]. Zhang, Q., & Wang, H. (2023). Predicting student performance using hybrid machine learning models. Journal of Educational Technology Systems, 51(1), 1-20.

[110]. Zhou, L., & Chen, X. (2023). An integrated machine learning framework for academic performance prediction. International Journal of Educational Technology in Higher Education, 20(1), 1-15.

[111]. Darshini, P. (2024). Data Analysis and Prediction of Student Academic Performance. Health Leadership and Quality of Life, 3(424), 1.