Industrial Oriented Mini Project Report
On

SOCIO-ACADEMIC EVALUATION OF STUDENT ALCOHOL USE BY INVESTIGATING BEHAVIOURAL ROOTS

Submitted to Jawaharlal Nehru Technological University for the partial
Fulfillment of the Requirement for the Award of the Degree of

Bachelor of Technology
In
Computer Science and Engineering

By
CH. VEDHASRI (22RA1A05H3)
K. GUNA SAGAR (23RA5A0505)
B. MANI KUMAR (23RA5A0508)

Under the guidance of
Mrs. R. NIRANJANI
M.Tech
Assistant Professor
Dept. of CSE

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

KOMMURI PRATAP REDDY INSTITUTE OF TECHNOLOGY
(UGC AUTONOMOUS)
(Affiliated to JNTUH, Ghanpur (V), Ghatkesar (M), Medchal (D)-500088)

2022-2026

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

KOMMURI PRATAP REDDY INSTITUTE OF TECHNOLOGY

(UGC AUTONOMOUS)

(Ghanpur (V), Ghatkesar (M), Medchal (D)-500088)

(AFFILIATED TO JNTU, Hyderabad)

CERTIFICATE

This is to certify that the Mini Project entitled **" SOCIO- ACADEMIC EVALUATION OF STUDENT ALCOHOL USE BY INVESTIGATING BEHAVIOURAL ROOTS"** being submitted by **CH. VEDHASRI (22RA1A05H3), K. GUNA SAGAR (23RA5A0505), B. MANI KUMAR (23RA5A0508)** in partial fulfillment for the award of Bachelor of Technology in Computer Science and Engineering to the Kommuri Pratap Reddy Institute of Technology, (UGC Autonomous, affiliated to JNTUH) is a record of confined work carried out by them under my guidance and supervision.

Guide's signature
Mrs. R. NIRANJANI
M.Tech
Asst. Professor
Dept. of CSE

Signature of the Head of the Dept.
Mr. K. VAMSHEE KRISHNA
M.Tech, (Ph.D).
Asst. Professor
Dept. of CSE

Place: Ghanpur

Signature of External Examiner

Date:

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

KOMMURI PRATAP REDDY INSTITUTE OF TECHNOLOGY

(UGC AUTONOMOUS)

(Ghanpur (V), Ghatkesar (M), Medchal (D)-500088)

(AFFILIATED TO JNTU, Hyderabad)

DECLARATION

We, **CH. VEDHASRI (22RA1A05H3), K. GUNA SAGAR (23RA5A0505), B. MANI KUMAR (23RA5A0508)** hereby declare that the mini project report titled **"SOCIO-ACADEMIC EVALUATION OF STUDENT ALCOHOL USE BY INVESTIGATING BEHAVIOURAL ROOTS"** under the guidance of Mrs. R. NIRANJANI, Assistant Professor, Department of Computer Science and Engineering, Kommuri Pratap Reddy Institute of Technology, Ghanpur, is submitted in partial fulfillment of the requirements for the award of the degree of Bachelor of Technology in Computer Science and Engineering.

This is a record of bonafide work carried out by us and the results embodied in this project have not been reproduced or copied from any source. The results embodied in this thesis have not been submitted to any other University and Institute for the award of any Degree or Diploma.

CH. VEDHASRI	**(22RA1A05H3)**
K. GUNA SAGAR	**(23RA5A0505)**
B. MANI KUMAR	**(23RA5A0508)**

ACKNOWLEDGEMENT

We manifest our heartier thankfulness pertaining to our contentment over **Mrs. R. NIRANJANI**, Assistant Professor as a project guide with whose adroit concomitance the excellence has been exemplified in bringing out this project work with artistry. We express our gratitude to **Mr. K. Vamshee Krishna**, Head of the Department for the encouragement and assistance to us, which contribute to the successful completion of this project. This Acknowledgement will be incomplete without mentioning our sincere gratefulness to our honorable Chairman **Sri. Kommuri Pratap Reddy**, our Director, **Dr. B. Sudheer Prem Kumar,** and Vice Principal, Dr. Sreenath Kashyap, who have been observing posing valiance in abundance, forwarding our individuality to acknowledge our project work tendentiously. At the outset we thank teaching and non-teaching staff of Dept. of CSE for providing us with good faculty and for their moral support throughout the course.

A heartful and sincere gratitude to our beloved parents for their tremendous motivation and moral support. We also express the overall exhilaration and gratitude to all those who animated our project work and accentuated our stance.

CH. VEDHASRI	**(22RA1A05H3)**
K. GUNA SAGAR	**(23RA5A0505)**
B. MANI KUMAR	**(23RA5A0508)**

KPRIT
Imagine. Innovate. Inspire.

KOMMURI PRATAP REDDY INSTITUTE OF TECHNOLOGY

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

Vision of the Institute

To emerge as a premier institute for high quality professional graduates who can contribute to economic and social developments of the Nation.

Mission of the Institute

Mission	**Statement**
IM$_1$	To have holistic approach in curriculum and pedagogy through industry interface to meet the needs of Global Competency.
IM$_2$	To develop students with knowledge, attitude, employability skills, entrepreneurship, research potential and professionally Ethical citizens.
IM$_3$	To contribute to advancement of Engineering & Technology that would help to satisfy the societal needs.
IM$_4$	To preserve, promote cultural heritage, humanistic values and Spiritual values thus helping in peace and harmony in the society.

KPRIT
Imagine. Innovate. Inspire.

KOMMURI PRATAP REDDY INSTITUTE OF TECHNOLOGY

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

Vision of the Department

To Provide Quality Education in Computer Science for the innovative professionals to work for the development of the nation.

Mission of the Department

Mission	Statement
DM1	Laying the path for rich skills in Computer Science through the basic knowledge of mathematics and fundamentals of engineering
DM2	Provide latest tools and technology to the students as a part of learning infrastructure
DM3	Training the students towards employability and entrepreneurship to meet the societal needs.
DM4	Grooming the students with professional and social ethics.

Program Educational Objectives (PEOs)

PEO’s	Statement
PEO1	The graduates of Computer Science and Engineering will have successful career in technology.
PEO2	The graduates of the program will have solid technical and professional foundation to continue higher studies.
PEO3	The graduate of the program will have skills to develop products, offer services and innovation.
PEO4	The graduates of the program will have fundamental awareness of industry process, tools and technologies.

Program Outcomes

PO1	**Engineering Knowledge:** Apply the knowledge of mathematics, science, Engineering fundamentals, and an engineering specialization to the solution of complex engineering problems.
PO2	**Problem Analysis:** Identify, formulate, review research literature, and analyse complex engineering problems reaching substantiated conclusions using first principles of mathematics, natural sciences, and engineering sciences.
PO3	**Design/development of Solutions:** Design solutions for complex engineering problems and design system components or processes that meet the specified needs with appropriate consideration for the public health and safety, and the cultural, societal, and environmental considerations.
PO4	**Conduct investigations of complex problems:** Use research-based knowledge and research methods including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions.
PO5	**Modern tool usage:** Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools including prediction and modelling to complex engineering activities with an understanding of the limitations.
PO6	**The engineer and society:** Apply reasoning informed by the contextual knowledge to assess societal, health, safety, legal and cultural issues and the consequent responsibilities relevant to the professional engineering practice.
PO7	**Environment and sustainability:** Understand the impact of the professional engineering solutions in societal and environmental context, and demonstrate the knowledge of, and need for sustainable development.
PO8	**Ethics:** Apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice.

PO9	Individual and team network: Function effectively as an individual, and as a member or leader in diverse teams, and in multidisciplinary settings.
PO10	Communication: Communicate effectively on complex engineering activities with the engineering community and with society at large, being able to comprehend and write effective reports and design documentation, make Effective presentations, and give and receive clear instructions.
PO11	Project management and finance: Demonstrate knowledge and understanding of the engineering and management principles and apply these to one's own work, as a member and leader in a team, to manage projects and in multidisciplinary environment.
PO12	Life-Long learning: Recognize the need for, and have the preparation and able to engage in independent and life-long learning in the broadest context of technological change.

PROGRAM SPECIFIC OUTCOMES

PSO1	Foundation of mathematical concepts: To use mathematical methodologies to crack problem using suitable mathematical analysis, data structure and suitable algorithm.
PSO2	Foundation of Computer Science: The ability to interpret the fundamental concepts and methodology of computer systems. Students can understand the functionality of hardware and software aspects of computer systems.
PSO3	Foundation of Software development: The ability to grasp the software development lifecycle and methodologies of software systems. Possess competent skills and knowledge of software design process.

ABSTRACT

Alcohol consumption has long been a subject of concern regarding its impact on academic performance and educational outcomes, particularly within social environments. The relationship between alcohol use and academic trajectories is a complex interplay of individual behavior, social habits, and societal norms. Students, especially in collegiate settings, may experience both positive and negative social pressures, where alcohol consumption can serve as a coping mechanism or an avenue for socialization. The historical context of alcohol’s impact on education can be traced back to early studies linking alcohol abuse with lower academic achievement, disrupted cognitive functioning, and higher dropout rates, which continue to resonate in modern society. Prior to the rise of AI and machine learning, addressing the challenges of understanding alcohol's effects on academic performance relied on traditional methods like surveys, interviews, and manual data analysis, which were time-consuming and lacked scalability. These traditional systems often failed to capture the nuanced, real-time relationship between alcohol consumption and educational outcomes. The motivation for developing this research stems from the need to address these limitations and create a more effective way of understanding and predicting the effects of alcohol consumption on academic trajectories. As AI and machine learning technologies have advanced, there is an opportunity to build models that can process large-scale data from diverse sources and uncover hidden patterns in real-time, improving both prevention and intervention strategies. One of the key problems faced by traditional systems was the lack of predictive accuracy, reliance on self-reported data, and a failure to account for complex, multivariable influences on students' academic lives. These gaps hindered the effectiveness of existing interventions. The proposed system aims to leverage machine learning techniques to analyze large datasets of student behavior, academic performance, and alcohol-related habits, predicting potential outcomes and providing actionable insights. By doing so, it will help educators and policymakers develop more tailored and efficient strategies to mitigate the negative effects of alcohol consumption on academic performance and overall educational outcomes.

INDEX

DIAGRAMS

1.INTRODUCTION

CHAPTER 1

INTRODUCTION

1.1 Overview

In the United States, one in four individuals between the ages of 12 and 20 drinks alcohol on a monthly basis, and a similar proportion of 12th graders consumes five or more drinks in a row at least once every two weeks. Several studies have reported that alcohol use during adolescence affects educational attainment by decreasing the number of years of schooling and the likelihood of completing school. Other research using alternative estimation techniques suggests that the effects of teen drinking on years of education and schooling completion are very small and/or non-significant. Despite a growing literature in this area, no study has convincingly answered the question of whether alcohol consumption inhibits high school students' learning. Alcohol consumption could be an important determinant of how much a high school student learns without having a strong impact on his or her decision to stay in school or attend college. This question is fundamental and timely, given recent research showing that underage drinkers are susceptible to the immediate consequences of alcohol use, including blackouts, hangovers, and alcohol poisoning, and are at elevated risk of neurodegeneration (particularly in regions of the brain responsible for learning and memory), impairments in functional brain activity, and neurocognitive defects.

A common and comprehensive measure of high school students' learning is Grade Point Average (GPA). GPA is an important outcome because it is a key determinant of college admissions decisions and of job quality for those who do not attend college. Only a few studies have explored the association between alcohol use and GPA. Wolaver (2002) and Williams, Powell, and Wechsler (2003) have studied this association among college students, while DeSimone and Wolaver [10] (2005) have investigated the effects of underage drinking on GPA during high school. The latter study found a negative association between high school drinking and grades, although it is not clear whether the effects are causal or the result of unobserved heterogeneity.

Understanding the relationship between teenage drinking and high school grades is pertinent given the high prevalence of alcohol use among this age cohort and recent research on adolescent brain development suggesting that early heavy alcohol use may have negative effects on the physical development of brain structure (Brown, Tapert, Granholm, & Delis,[4] 2000; Tapert & Brown, 1999). By affecting the quality of learning, underage drinking could have an impact on both college admissions and job quality independent of its effects on years of schooling or school completion.

In this paper, we estimate the effects of drinking in high school on the quality of learning as captured by high school GPA. The analysis employs data from Waves 1 and 2 of the National Longitudinal Study of Adolescent Health (Add Health), a nationally representative study that captures health-related behaviors of adolescents in grades 7 through 12 and their outcomes in young adulthood. Our analysis contributes to the literature in several ways. First, we focus on the effect of drinking on academic achievement during high school. To date, and to the best of our knowledge, only one other study in the literature has analyzed the consequences of underage drinking on high school GPA. Second, rather than rely on self-reported GPA, we use objective GPA data from academic transcripts, reducing the potential for systematic biases in the estimation results. Third, we take advantage of the longitudinal nature of the Add Health data and use fixed-effects models to purge the analysis of time invariant unobserved heterogeneity. Fixed-effects techniques are superior to instrumental variables (IV) estimation when the strength and reliability of the instruments are suspect (French & Popovici, 2009) [15]. Finally, we explore a variety of mechanisms that could underlie a detrimental effect of alcohol use on grades. In addition to analyzing mediators related to exposure to education (days of school skipped), we investigate the effect of drinking on students’ ability to focus on and adhere to academic objectives.

1.2 Research Motivation

Alcohol consumption among students is a growing concern that directly impacts academic performance and future career prospects. Many young individuals view alcohol as a social tool, using it to build friendships, relieve stress, or conform to peer pressure. However, frequent drinking habits often lead to missed classes, reduced concentration, and lower academic achievements. A real-time scenario that highlights this issue can be seen in university students who attend social gatherings on weekdays, consuming excessive alcohol and struggling to wake up in time for morning lectures. Over time, their grades decline, and they experience difficulty keeping up with coursework, which can lead to long-term academic setbacks. This issue is not limited to a single institution or country but is prevalent across various educational systems, making it an important subject to study and address.

In professional settings, employers seek candidates who demonstrate discipline, focus, and responsibility. However, excessive alcohol consumption during academic years can result in poor time management and an inability to meet deadlines, traits that negatively impact career opportunities. For example, students who frequently engage in drinking sessions before important exams or assignments often underperform, which limits their potential for securing scholarships, internships, or job placements. By understanding the relationship between alcohol consumption and academic success, institutions can implement awareness programs and interventions to help students develop healthier habits. The research into this topic serves as a bridge between social behaviors and educational achievements, aiming to create a balance where students can engage socially without compromising their academic and professional futures.

1.3 Problem Statement

Alcohol consumption among students has become a widespread issue, significantly impacting their academic performance and overall well-being. Many students engage in drinking as part of their social habits, often underestimating its long-term consequences on their education. While occasional consumption may seem harmless, frequent and excessive drinking leads to decreased concentration, missed classes, poor academic performance, and even dropout risks. The challenge is that students often prioritize social engagements over academic responsibilities, believing they can manage both. However, studies show that high alcohol intake is linked to lower grades, impaired cognitive function, and an inability to meet academic expectations. This creates a cycle where students struggle to balance their social lives with their educational commitments, ultimately affecting their future career opportunities.

Despite growing awareness, educational institutions still face difficulties in addressing this issue effectively. Traditional interventions, such as awareness campaigns and strict policies, do not always resonate with students who see drinking as a normal part of university life. There is a need for a deeper understanding of how alcohol influences students' academic trajectories and what strategies can help bridge the gap between social habits and educational outcomes. By identifying key factors—such as peer pressure, stress, and lifestyle choices—this research aims to provide insights that can guide the development of targeted interventions. The goal is to help students cultivate responsible drinking habits while ensuring their academic success remains a priority.

1.4 Significance

To overcome the negative impact of alcohol consumption on academic performance, this study aims to explore the relationship between students' drinking habits and their educational outcomes. To analyze the influence of social factors, cognitive effects, and lifestyle choices, the research will examine patterns of alcohol use and their direct consequences on learning abilities. To overcome the challenges of ineffective interventions, the study will identify strategies that promote responsible drinking while maintaining academic success. To analyze possible solutions, it will assess the role of awareness programs, institutional policies, and peer influence in shaping students' behaviors.

1.5 Applications

- **University Counseling Programs** – Educational institutions can use the research findings to enhance counseling services, offering personalized guidance to students struggling with alcohol-related academic issues.

- **Alcohol Awareness Campaigns** – Schools and colleges can implement targeted awareness programs that educate students on responsible drinking habits and their impact on academic performance.

- **Student Behavior Monitoring** – Universities can develop digital tracking systems to monitor class attendance, academic performance, and alcohol-related behavioral patterns, allowing early intervention for at-risk students.

- **Policy Making in Educational Institutions** – The study can help universities and colleges formulate stricter yet effective alcohol policies that ensure students maintain a balance between social life and academics.

- **Parental and Peer Education** – Parents and peer groups can use insights from the research to better understand how alcohol affects students' educational trajectories and support them in making healthier choices.

- **Workplace Readiness Programs** – Companies and career counseling centers can incorporate findings to prepare students for professional environments, emphasizing the importance of discipline, responsibility, and the long-term effects of alcohol consumption.

2.LITERATURE SURVEY

CHAPTER 2

LITERATURE SURVEY

Behavioral research has found that educational performance is highly correlated with substance abuse. Economic studies that look at the link between alcohol use and educational outcomes have customarily focused on measures of educational attainment such as graduation (from high school or college), college matriculation, and years of school completed. Consistent with the behavioral research, early economic studies found that drinking reduced educational attainment. But the most rigorous behavioral studies and the early economic studies of attainment both faced the same limitation: they were cross-sectional and subject to potential omitted variables bias. Some of these cross-sectional economic studies attempted to improve estimation by using instrumental variables (IV). Cook [8] and Moore (1993) and Yamada et al. (1996) found that heavy or frequent drinking in high school adversely affects high school and college completion. Nevertheless, the validity and reliability of the instruments in these studies are open to debate.

By contrast, more recent economic studies that arguably use better estimation methods have found that drinking has modest or negligible effects on educational attainment. Dee and Evans [9] (2003) studied the effects of teen drinking on high school completion, college entrance, and college persistence. Employing changes in the legal drinking age across states over time as an instrument, they found no significant effect of teen drinking on educational attainment. Koch and Ribar (2001) reached a similar conclusion applying family fixed effects and instrumental variables to NLSY data. Though they found that drinking had a significant negative effect on the amount of schooling completed among men, the effect was small. Finally, Chatterji (2006) used a bivariate probit model of alcohol use and educational attainment to gauge the sensitivity of the estimates to various assumptions about the correlation of unobservable determinants of these variables. She concluded that there is no evidence of a causal relationship between alcohol use and educational attainment when the correlation coefficient is fixed at plausible levels.

Alcohol use could conceivably affect a student’s quality of learning and academic performance regardless of its impact on school completion. This possibility is suggested by Renna (2008), who uses a research design similar to that used by Dee and Evans [9] (2003) and finds that although binge drinking does not affect high school completion rates, it does significantly increase the probability that a student graduates with a GED rather than a high school diploma. Drinking could affect learning through a variety of mechanisms. Recent neurological research suggests that underage drinking can impair learning directly by causing alterations in the structure and function of the developing brain with consequences reaching far beyond adolescence. Negative effects of alcohol use can emerge in areas such as planning and executive functioning, memory, spatial operations, and attention. Alcohol use could also affect performance by reducing the number of hours committed to studying, completing homework assignments, and attending school.

We are aware of five economic studies that have examined whether drinking affects learning per se. Bray [2] (2005) analyzed this issue indirectly by studying the effect of high school students' drinking on subsequent wages, as mediated through human capital accumulation. He found that moderate high school drinking had a positive effect on returns to education and therefore on human capital accumulation. Heavier drinking reduced this gain slightly, but net effects were still positive. The other four studies approached the question directly by focusing on the association between drinking and GPA. Three of the GPA studies used data from the Harvard College Alcohol Study. Analyzing data from the study's 1993 wave, both Wolaver (2002) and Williams et al. (2003) estimated the impact of college drinking on the quality of human capital acquisition as captured by study hours and GPA. Both studies found that drinking had a direct negative effect on GPA and an indirect negative effect through reduced study hours. Wolaver (2007) used data from the 1993 and 1997 waves and found that both high school and college binge drinking were associated with lower college GPA for males and females. For females, however, study time in college was negatively correlated with high school drinking but positively associated with college drinking.

To our knowledge, only one study has looked specifically at adolescent drinking and high school GPA. Analyzing data from the Youth Risk Behavior Survey, DeSimone and Wolaver (2005) used standard regression analysis to estimate whether drinking affected high school GPA. Even after controlling for many covariates, they found that drinking had a significant negative effect. Their results showed that the GPAs of binge drinkers were 0.4 points lower on average for both males and females. They also found that the effect of drinking on GPA peaked for ninth graders and declined thereafter and that drinking affected GPA more by reducing the likelihood of high grades than by increasing the likelihood of low grades.

All four GPA studies found that drinking has negative effects on GPA, but they each faced two limitations. First, they relied on self-reported GPA, which can produce biased results due to recall mistakes and intentional misreporting. Second, they used cross-sectional data. Despite these studies' serious efforts to address unobserved individual heterogeneity, it remains questionable whether they identified a causal link between drinking and GPA.

In sum, early cross-sectional studies of educational attainment and GPA suggest that drinking can have a sizeable negative effect on both outcomes. By contrast, more recent studies of educational attainment that use improved estimation methods to address the endogeneity of alcohol use have found that drinking has negligible effects. The present paper is the first study of GPA that controls for individual heterogeneity in a fixed-effects framework, and our findings are consistent with the more recent studies of attainment that find small or negligible effects of alcohol consumption.

3.EXISTING SYSTEM

CHAPTER 3

EXISTING SYSTEM

The traditional approach to studying alcohol's effect on academic trajectories relies on self-reported surveys, standardized test scores, and observational studies. Researchers collect data through student questionnaires about drinking habits, frequency, and reasons for alcohol consumption. Academic performance is measured using GPA, attendance records, and graduation rates. Institutions and policymakers analyze trends to identify correlations between alcohol use and educational outcomes. Longitudinal studies track students over time to assess the impact of drinking on cognitive function, motivation, and overall academic success. However, these methods depend on voluntary responses, which can be inaccurate due to underreporting or exaggeration. The analysis often focuses on broad trends rather than individual behaviors, missing the nuances of how different drinking patterns affect learning. Additionally, traditional research struggles to distinguish between causation and correlation, making it difficult to determine whether alcohol directly harms academic performance or if other factors contribute to poor outcomes. Since these studies do not incorporate real-time behavioral tracking or predictive modeling, they fail to provide immediate interventions for at-risk students. A more advanced system integrating AI, real-time data collection, and behavioral analytics is needed to bridge the gap between social drinking habits and their direct effects on education.

Limitations of Traditional Systems in Studying Alcohol's Impact on Academics

- **Reliance on Self-Reported Data** – Surveys depend on students' honesty, leading to inaccurate or biased responses.
- **Lack of Real-Time Monitoring** – Traditional studies cannot track immediate effects of alcohol consumption on learning and cognitive function.
- **Inability to Differentiate Causation from Correlation** – Studies struggle to determine whether alcohol directly causes poor academic performance or if other factors contribute.
- **Broad Generalizations** – Population-wide analyses fail to capture individual variations in drinking habits and their unique effects.

- **Delayed Intervention Strategies** – By the time results are analyzed, struggling students may have already faced academic setbacks.
- **Exclusion of External Influences** – Social environment, mental health, and financial stress are often overlooked in alcohol-related academic studies.
- **Limited Data on Long-Term Effects** – Most studies focus on short-term performance changes rather than lifelong educational outcomes.
- **Minimal Use of Technology and AI** – Traditional methods do not leverage predictive analytics to identify at-risk students early.
- **Static Data Collection Methods** – Surveys and periodic assessments fail to adapt to evolving student behaviors and drinking trends.

4.PROPOSED SYSTEM

CHAPTER 4

PROPOSED SYSTEM

4.1 Overview

Step 1: Upload the Academic Trajectories Dataset

The research procedure begins by obtaining and uploading the academic trajectories dataset, which constitutes the fundamental source of information for this study. The dataset contains detailed records of individuals' alcohol consumption habits and academic performance outcomes. Its selection ensures that the research is based on reliable data that directly connects social habits to educational achievements. The dataset is stored in a structured format that facilitates further manipulation and analysis. The upload process confirms that the raw data is accessible and prepared for a thorough examination in subsequent steps. This initial step lays the groundwork for the entire research, guaranteeing that the study is rooted in real-world observations of alcohol's impact on academic performance.

Step 2: Data Preprocessing and Exploratory Data Analysis

Once the dataset is uploaded, the next phase focuses on data preprocessing and exploratory data analysis (EDA). In this stage, the dataset undergoes a detailed investigation to determine its structure, identify potential anomalies, and reveal the distribution of values among the features. The process involves generating information regarding data types, non-null counts, and memory usage to establish a clear understanding of the dataset. Descriptive statistics are computed to provide insights into the central tendency, dispersion, and overall distribution of the variables. Furthermore, the dataset is visualized using a correlation heatmap that displays the relationships between features, which is essential in discerning underlying patterns and dependencies. This thorough examination ensures that every aspect of the dataset is understood, forming a robust foundation for subsequent data cleaning and modeling efforts.

Step 3: Elimination of Missing Values and Data Cleaning

The research emphasizes the integrity of the data, and a critical component of this phase is the elimination of missing values. The dataset is carefully inspected for null or missing entries, and any observation that does not contain complete information is removed. This cleaning process prevents the introduction of bias or errors in the model training phase. Every entry in the dataset is verified to ensure that it contributes complete and accurate information to the analysis. The rigorous cleaning protocol strengthens the dataset’s reliability, ensuring that every subsequent step, from feature scaling to model training, is performed on data that is free of gaps and inconsistencies.

Step 4: Feature Scaling Using Standard Scaler

After ensuring that the data is clean and consistent, the next step involves transforming the features through standardization. A standard scaler is applied to the dataset to adjust the values of each feature so that they exhibit a mean of zero and a standard deviation of one. This transformation is crucial because it brings all features to the same scale and prevents variables with larger numerical ranges from dominating the learning process of the models. The process guarantees that each feature contributes equally to the predictive models. The scaling operation enhances the efficiency and accuracy of the algorithms, particularly those that rely on distance-based measurements or assume a normal distribution in the input variables. With a standardized dataset, the analysis attains a higher level of precision in its modeling efforts.

Step 5: Splitting Data into Training and Testing Sets

A vital component of the research methodology is the division of the dataset into training and testing subsets. The dataset is split using an 80-20 ratio where eighty percent of the data is reserved for training the models, and the remaining twenty percent is set aside for testing and validating the performance of these models. The division is executed using a fixed random state to maintain reproducibility of the results. This strategic partitioning ensures that the models are trained on a substantial portion of the data while also providing a robust mechanism for unbiased evaluation on unseen data. The testing set plays a crucial role in assessing the generalization capabilities of the models, confirming that the findings are not confined to the specific characteristics of the training data alone.

Step 6: Building Predictive Models Using SVM and Gaussian Naive Bayes

The research proceeds to construct two well-established predictive models as part of the comparative study. The first model is a support vector machine (SVM) that employs a radial basis function (RBF) kernel. This model constructs a hyperplane in a high-dimensional space that separates data points based on their class labels. The RBF kernel allows the model to capture complex, non-linear relationships between the features and the academic outcomes. The SVM model is trained using the standardized training dataset, and its parameters are tuned to maximize performance.

In parallel, a Gaussian Naive Bayes (GNB) classifier is built. This classifier leverages Bayes' theorem and assumes that the features are conditionally independent given the outcome variable. The simplicity and computational efficiency of GNB allow it to process high-dimensional data effectively. By training on the same standardized dataset, the GNB model provides a contrasting approach to the SVM, thereby enriching the comparative analysis. Together, these models serve as benchmarks that help to evaluate the predictive capabilities and limitations of traditional machine learning techniques in the context of alcohol consumption and academic performance.

Step 7: Constructing the Proposed CatBoost Classifier Model

The research introduces a novel component by incorporating the CatBoost classifier as a proposed model. CatBoost, a gradient boosting algorithm, specializes in handling categorical variables and mitigating the risk of overfitting. The classifier is configured with specific hyperparameters such as the number of iterations, learning rate, and tree depth to optimize its predictive performance. Training the CatBoost model on the standardized training dataset equips the study with an advanced tool that is capable of capturing intricate interactions between variables. This step represents a significant advancement in the analysis, bridging traditional models with modern gradient boosting techniques. The CatBoost classifier is positioned as the cornerstone of the research's predictive methodology, offering robust performance and adaptability in analyzing the impact of alcohol consumption on academic trajectories.

Step 8: Model Evaluation and Performance Comparison

Following the development of the SVM, Gaussian Naive Bayes, and CatBoost models, the research undertakes a comprehensive evaluation of their performance. The assessment employs multiple performance metrics such as accuracy, precision, recall, and F1-score. Accuracy quantifies the overall correctness of the model's predictions, while precision measures the proportion of correct positive predictions. Recall evaluates the model's capability to identify all actual positive instances, and the F1-score represents the harmonic mean of precision and recall. Additionally, confusion matrices are generated for each model, providing a detailed account of true positives, true negatives, false positives, and false negatives. These matrices are visualized using heatmaps, which facilitate an intuitive understanding of each model's performance. The comparative analysis of these performance metrics highlights the strengths and weaknesses of the models, ultimately identifying the CatBoost classifier as the superior model in terms of capturing the relationship between alcohol consumption and academic outcomes. The methodical evaluation process ensures that every model is scrutinized and that the selection of the best-performing model is based on a comprehensive set of criteria.

Step 9: Prediction on New Test Data Using the CatBoost Model

The final phase of the research applies the selected CatBoost classifier to new, unseen test data. The test data, imported from an external source, undergoes the same preprocessing and scaling steps to maintain consistency with the training dataset. This consistency is critical to ensure that the model's predictions are accurate and reliable. Once the test data is standardized, it is input into the CatBoost model, which generates predictions regarding academic outcomes. These predictions are then appended to the test data, facilitating a clear comparison between the observed social habits and the predicted educational outcomes. This stage of the research provides actionable insights into how alcohol consumption correlates with academic performance, and it demonstrates the practical applicability of the model. The accurate prediction on new data signifies the robustness of the CatBoost classifier and validates its effectiveness as the core predictive tool of the study.

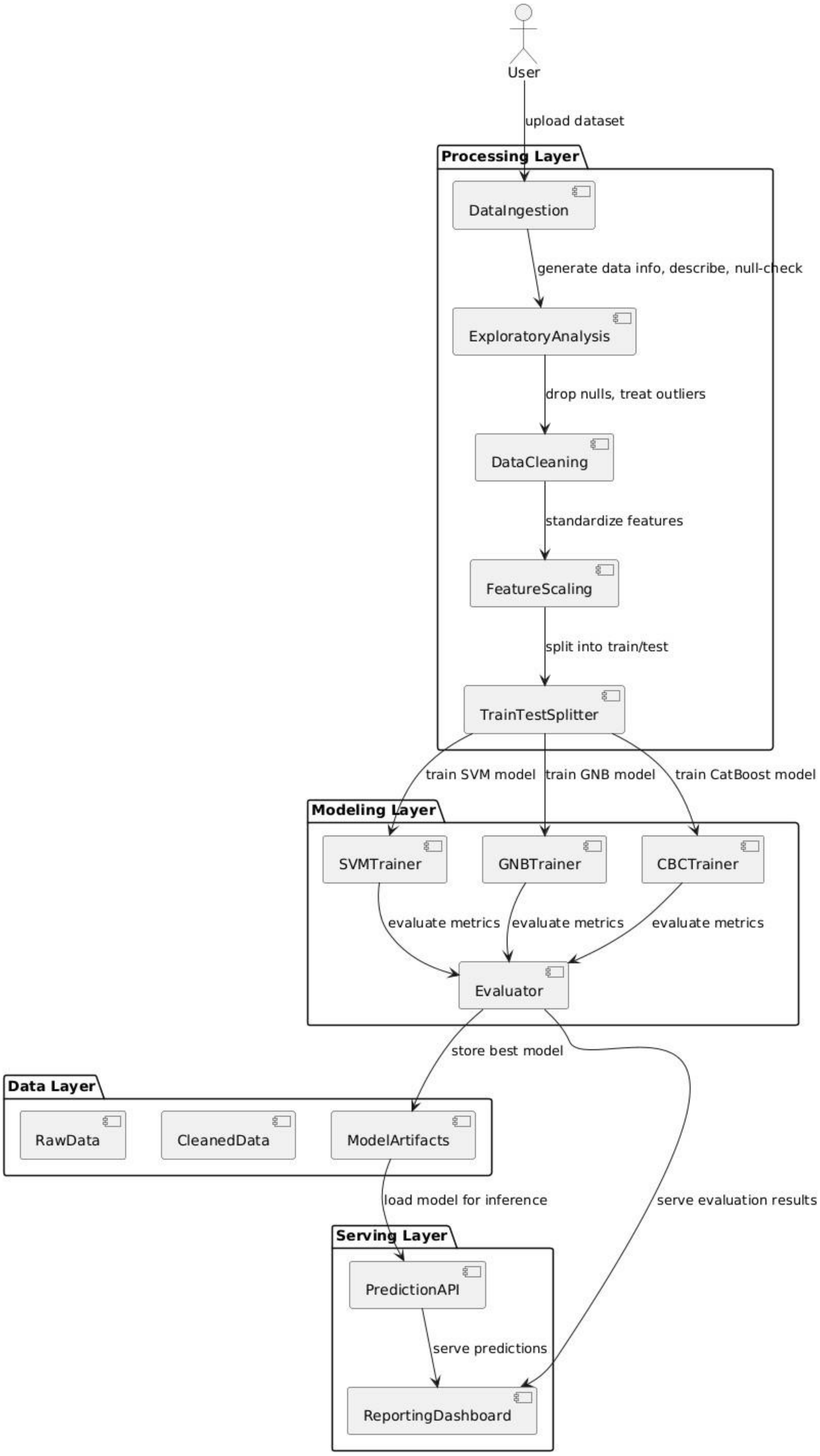


Fig. 1: Architectural Block Diagram.

4.2 Data Preprocessing

The dataset underwent a systematic series of preprocessing steps to ensure its quality, reliability, and suitability for machine learning models. The steps are as follows:

Missing Value Analysis: The dataset was thoroughly checked for missing or null values across all features. Instances containing missing values were either removed or imputed based on the nature and importance of the feature to preserve data integrity.

Data Type Conversion: Each column was verified for appropriate data types. Categorical features stored as numerical codes were converted to string or categorical types, and numerical data stored as strings were parsed and converted to integer or float types.

Outlier Detection and Treatment: Outliers were detected using statistical methods such as the Interquartile Range (IQR) technique. Detected outliers were either removed or capped to prevent skewing of model performance and to maintain consistency in data distribution.

Categorical Encoding: Categorical variables were transformed using encoding techniques. Label Encoding was used for binary categorical variables, and One-Hot Encoding was applied to multi-class categorical features to convert them into numerical representations suitable for model training.

Feature Scaling: Numerical features were standardized using techniques such as Min-Max Scaling or Standardization to bring all features onto a similar scale. This ensured that models based on distance metrics or gradient descent performed optimally.

Duplicate Record Removal: The dataset was examined for duplicate entries. Identical rows that represented repeated observations were removed to prevent biased model training.

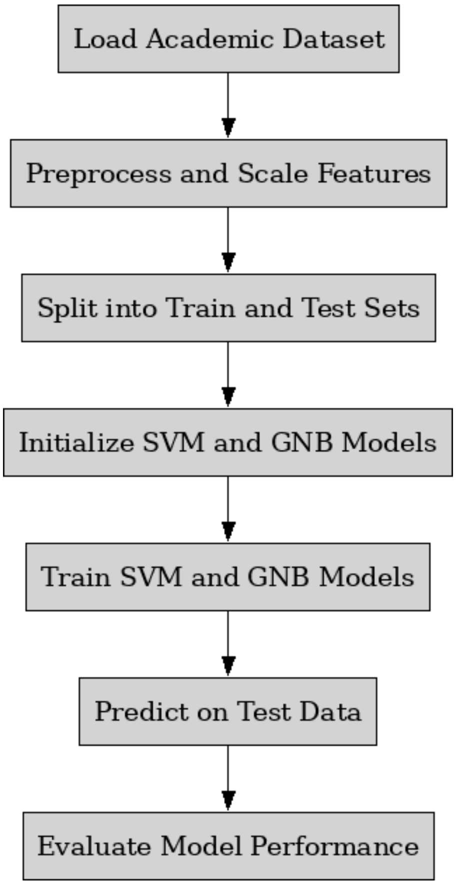
Feature Selection: Redundant and irrelevant features were identified and removed to reduce dimensionality. Correlation analysis and feature importance measures were used to retain only the most informative variables contributing to model performance.

4.3 ML MODELS

4.3.1 Existing Models: Support Vector Machine (SVM) and Gaussian Naive Bayes (GNB)

1. Introduction to Existing Models

Support Vector Machine (SVM) and Gaussian Naive Bayes (GNB) are two of the most widely used classification algorithms in machine learning. These models are known for their simplicity, effectiveness, and ability to handle classification problems with considerable accuracy. They serve as baseline models in this study to evaluate the predictive power of traditional machine learning approaches in understanding the impact of alcohol consumption on academic outcomes.


Load Academic Dataset
Preprocess and Scale Features
Split into Train and Test Sets
Initialize SVM and GNB Models
Train SVM and GNB Models
Predict on Test Data
Evaluate Model Performance

2. Working of Existing Models

Support Vector Machine is a supervised learning algorithm that seeks to find the optimal hyperplane that separates data points of different classes with the maximum margin. It transforms the original feature space into a higher-dimensional space using kernel functions such as the radial basis function (RBF) to ensure that the data becomes linearly separable. The model then identifies the support vectors, which are the critical data points closest to the decision boundary, and uses them to define the hyperplane. The robustness of SVM lies in its ability to handle high-dimensional data and maintain a balance between underfitting and overfitting.

Gaussian Naive Bayes is a probabilistic classifier based on Bayes' theorem, assuming strong independence between features. It models the likelihood of features assuming a Gaussian distribution and calculates the posterior probability for each class given a set of features. The class with the highest posterior probability is assigned as the prediction. Due to its simplicity, the GNB classifier trains rapidly and performs well in cases where the feature independence assumption holds true. It is particularly effective when the dataset contains noise and when real-time classification is required due to its computational efficiency.

3. Limitations of the Existing Models

- SVM struggles with large datasets and becomes computationally expensive when the number of samples increases.

- SVM performance is sensitive to the choice of kernel function and hyperparameters.

- SVM is less effective when the classes are overlapping and not clearly separable.

- GNB assumes feature independence, which rarely holds true in real-world datasets.

- GNB underperforms when features are correlated.

- GNB is not suitable for complex decision boundaries.

- GNB's probabilistic nature can result in inaccurate predictions if prior probabilities are biased.

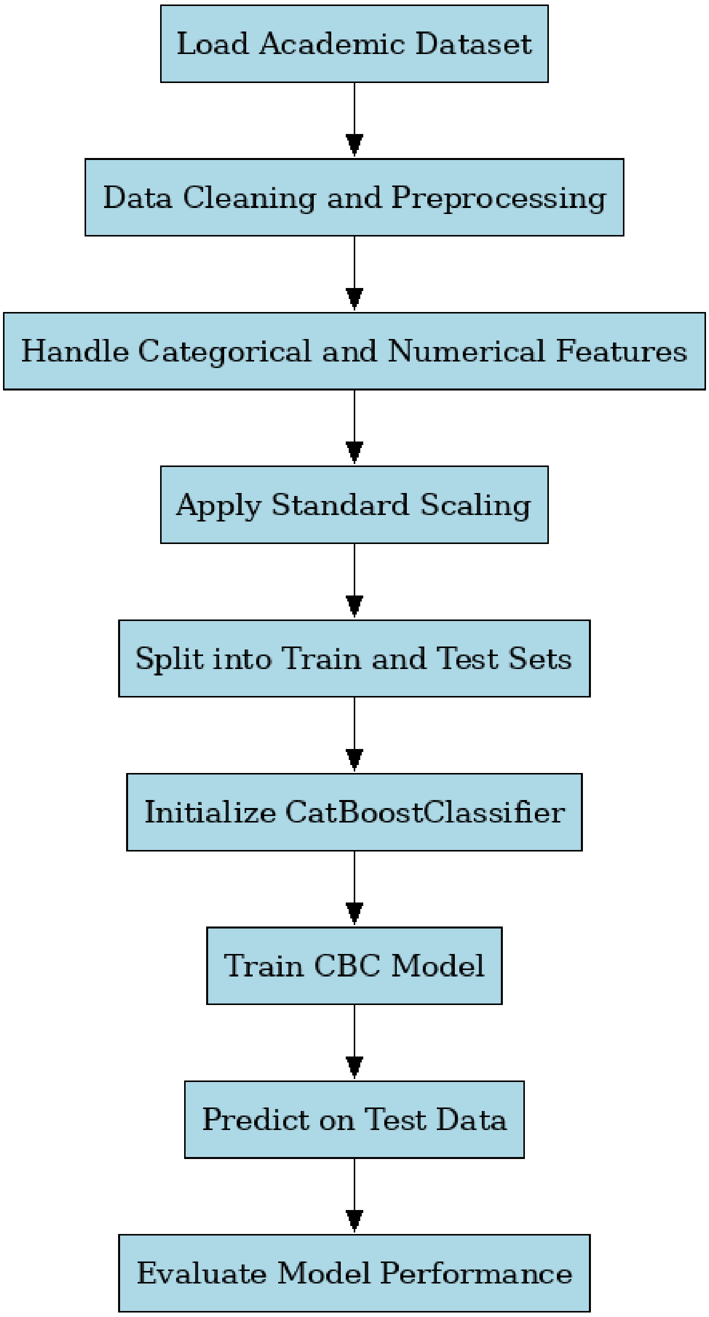
- Both models lack the ability to inherently capture complex non-linear relationships present in the data.

- Neither model supports embedded feature handling or categorical feature boosting natively.

4.3.2 Proposed Model: CatBoost Classifier (CBC)

1. Introduction to Proposed Model

CatBoost Classifier (CBC) is a gradient boosting algorithm developed to handle categorical data efficiently and provide superior performance with minimal preprocessing. It is a state-of-the-art ensemble learning method that constructs multiple decision trees sequentially, where each tree attempts to correct the errors of the previous one. The CBC model is implemented as the proposed approach due to its high accuracy, capability to handle complex datasets, and its ability to learn from categorical and numerical data without extensive transformation.


Load Academic Dataset
Data Cleaning and Preprocessing
Handle Categorical and Numerical Features
Apply Standard Scaling
Split into Train and Test Sets
Initialize CatBoostClassifier
Train CBC Model
Predict on Test Data
Evaluate Model Performance

2. Working of Proposed Model

CatBoost Classifier is a supervised learning algorithm based on gradient boosting over decision trees. It builds an ensemble of trees in a sequential manner where each new tree is constructed to reduce the loss function of the combined ensemble. It introduces several innovations, such as Ordered Boosting and minimal preprocessing of categorical features, to eliminate target leakage and overfitting. In the case of categorical variables, CatBoost internally uses a technique called target statistics, which replaces categorical features with aggregated statistics derived from the target variable. This is done in an ordered manner to avoid using future data while learning current patterns, thereby preserving model integrity.

The CBC model also incorporates symmetric tree structures that lead to faster training and prediction. During training, gradient descent is applied to minimize the loss function, and leaf values of the trees are optimized based on the gradients of the loss. This results in a strong predictive model that captures non-linear interactions and dependencies between features. The model's robustness and high generalization capability make it ideal for handling complex tasks like predicting academic trajectories influenced by social behaviors such as alcohol consumption.

3. Advantages of the Proposed Model

- Handles categorical variables natively without requiring explicit encoding.
- Implements Ordered Boosting to prevent target leakage and overfitting.
- Offers high accuracy and generalization across various types of datasets.
- Automatically handles missing values during training.
- Requires minimal data preprocessing and feature engineering.
- Captures complex non-linear relationships between features effectively.
- Provides faster training and inference due to symmetric tree structure.
- Includes built-in mechanisms for regularization and overfitting control.
- Offers better interpretability and feature importance visualization compared to traditional models.

5.UML DIAGRAMS

CHAPTER 5

UML DIAGRAMS

UML stands for Unified Modeling Language. UML is a standardized general-purpose modeling language in the field of object-oriented software engineering. The standard is managed, and was created by, the Object Management Group. The goal is for UML to become a common language for creating models of object-oriented computer software. In its current form UML is comprised of two major components: a Meta-model and a notation. In the future, some form of method or process may also be added to; or associated with, UML.

The Unified Modeling Language is a standard language for specifying, Visualization, Constructing and documenting the artifacts of software system, as well as for business modeling and other non-software systems. The UML represents a collection of best engineering practices that have proven successful in the modeling of large and complex systems. The UML is a very important part of developing objects-oriented software and the software development process. The UML uses mostly graphical notations to express the design of software projects.

GOALS: The Primary goals in the design of the UML are as follows:

- Provide users a ready-to-use, expressive visual modeling Language so that they can develop and exchange meaningful models.
- Provide extendibility and specialization mechanisms to extend the core concepts.
- Be independent of particular programming languages and development process.
- Provide a formal basis for understanding the modeling language.
- Encourage the growth of OO tools market.
- Support higher level development concepts such as collaborations, frameworks, patterns and components.
- Integrate best practices.

Class diagram

The class diagram is used to refine the use case diagram and define a detailed design of the system. The class diagram classifies the actors defined in the use case diagram into a set of interrelated classes. The relationship or association between the classes can be either an "is-a" or "has-a" relationship. Each class in the class diagram was capable of providing certain functionalities. These functionalities provided by the class are termed "methods" of the class. Apart from this, each class may have certain "attributes" that uniquely identify the class.

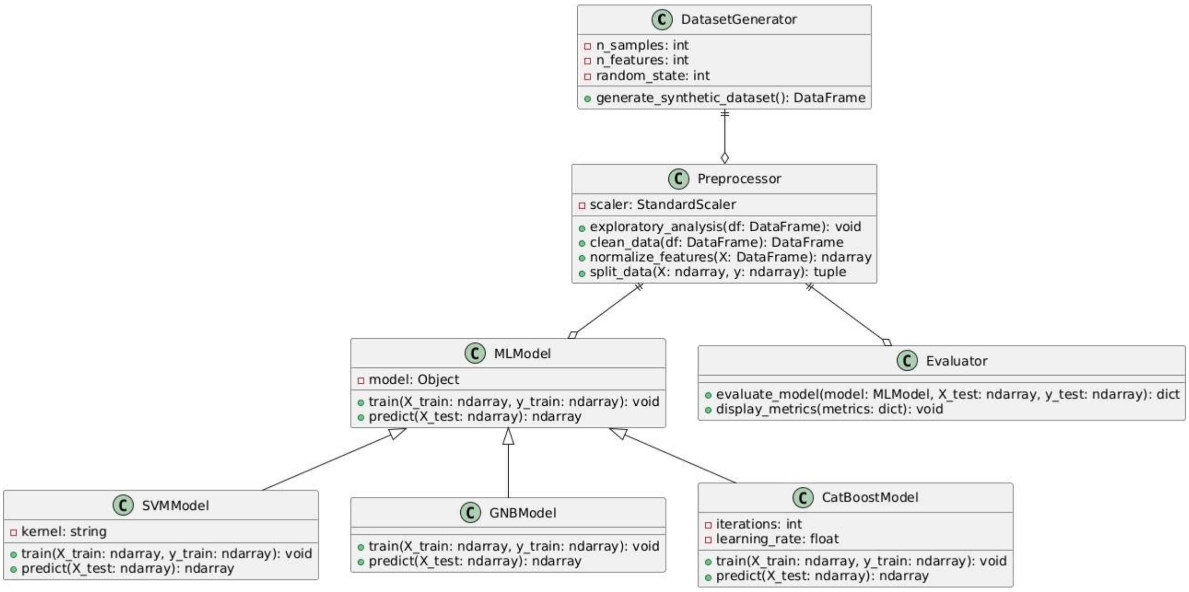

C DatasetGenerator
n_samples: int
n_features: int
random_state: int
generate_synthetic_dataset(): DataFrame
C Preprocessor
scaler: StandardScaler
exploratory_analysis(df: DataFrame): void
clean_data(df: DataFrame): DataFrame
normalize_features(X: DataFrame): ndarray
split_data(X: ndarray, y: ndarray): tuple
C MLModel
model: Object
train(X_train: ndarray, y_train: ndarray): void
predict(X_test: ndarray): ndarray
C Evaluator
evaluate_model(model: MLModel, X_test: ndarray, y_test: ndarray): dict
display_metrics(metrics: dict): void
C SVMModel
kernel: string
train(X_train: ndarray, y_train: ndarray): void
predict(X_test: ndarray): ndarray
C GNBModel
train(X_train: ndarray, y_train: ndarray): void
predict(X_test: ndarray): ndarray
C CatBoostModel
iterations: int
learning_rate: float
train(X_train: ndarray, y_train: ndarray): void
predict(X_test: ndarray): ndarray

Figure-5.1: Class Diagram

Sequence Diagram

A sequence diagram in Unified Modeling Language (UML) is a kind of interaction diagram that shows how processes operate with one another and in what order. It is a construct of a Message Sequence Chart. A sequence diagram shows, as parallel vertical lines (“lifelines”), different processes or objects that live simultaneously, and as horizontal arrows, the messages exchanged between them, in the order in which they occur. This allows the specification of simple runtime scenarios in a graphical manner.

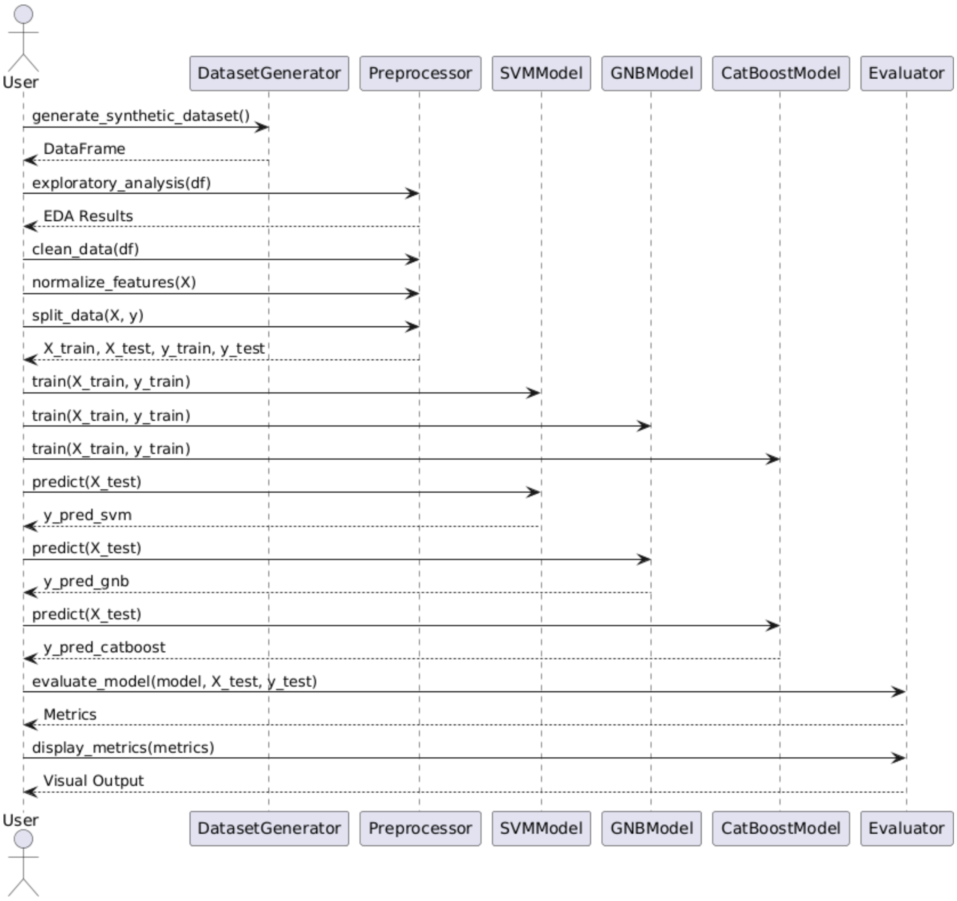


Figure-5.2: Sequence Diagram

Activity diagram

Activity diagrams are graphical representations of Workflows of stepwise activities and actions with support for choice, iteration, and concurrency. In the Unified Modeling Language, activity diagrams can be used to describe the business and operational step-by-step workflows of components in a system. An activity diagram shows the overall flow of control.

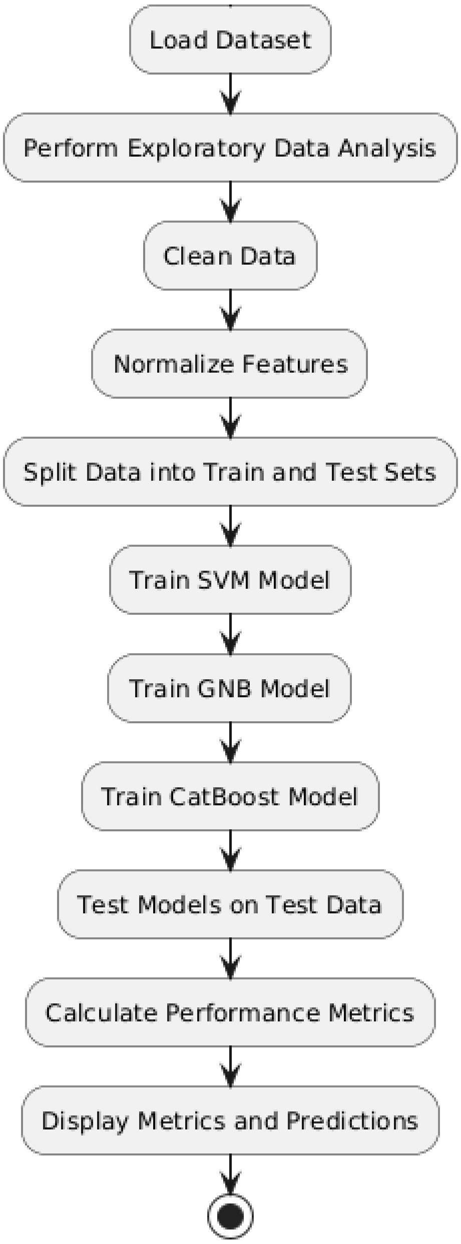

Load Dataset
Perform Exploratory Data Analysis
Clean Data
Normalize Features
Split Data into Train and Test Sets
Train SVM Model
Train GNB Model
Train CatBoost Model
Test Models on Test Data
Calculate Performance Metrics
Display Metrics and Predictions

Figure-5.3: Activity Diagram

Data flow diagram

A data flow diagram (DFD) is a graphical representation of how data moves within an information system. It is a modeling technique used in system analysis and design to illustrate the flow of data between various processes, data stores, data sources, and data destinations within a system or between systems. Data flow diagrams are often used to depict the structure and behavior of a system, emphasizing the flow of data and the transformations it undergoes as it moves through the system.

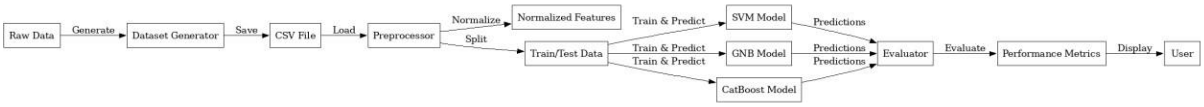


Figure-5.4: Dataflow Diagram

Use Case diagram: A use case diagram in the Unified Modeling Language (UML) is a type of behavioral diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted.

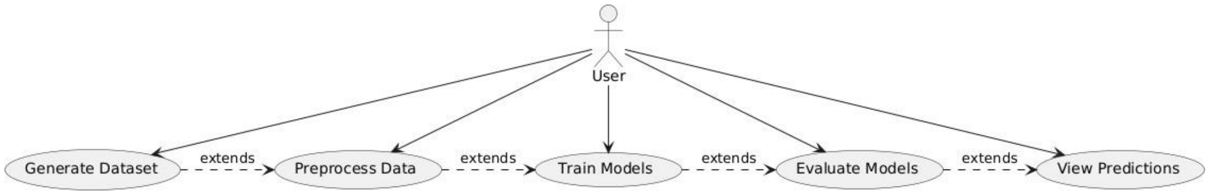


Figure-5.5: Use Case Diagram

6.SOFTWARE ENVIRONMENT

CHAPTER 6

SOFTWARE ENVIRONMENT

6.1 Software Requirements

Python is a high-level, interpreted programming language known for its simplicity and readability, which makes it a popular choice for beginners as well as experienced developers. Key features of Python include its dynamic typing, automatic memory management, and a rich standard library that supports a wide range of applications from web development to data science and machine learning. Its object-oriented approach and support for multiple programming paradigms allow developers to write clear, maintainable code. Python's extensive ecosystem of third-party packages further enhances its capabilities, enabling rapid development and prototyping across diverse fields.

Installation

First, download the appropriate installer from the official Python website (https://www.python.org/downloads/release/python-376/). For Windows users, run the executable installer and ensure to check the "Add Python to PATH" option during installation; for macOS and Linux, follow the respective package installation commands or use a package manager like Homebrew or apt-get. After installation, verify the setup by running python --version or python3 --version in your terminal or command prompt, which should display "Python 3.7.6." This version-specific installation supports all major functionalities and libraries compatible with Python 3.7.6, making it an excellent foundation for developing robust applications in areas such as data analysis, machine learning, and GUI development.

6.1.1 Python Packages

The project requires a robust set of software libraries and tools that work together to build an integrated system for plant disease classification. Below is an explanation of the key software requirements and the packages used:

- **Python:** The project is implemented in Python, which is chosen for its extensive ecosystem of libraries and its strong support for data analysis, machine learning, and GUI development.

- **Matplotlib & Seaborn:** These libraries are employed for data visualization. Matplotlib is used for creating standard plots, while Seaborn adds an extra layer of sophistication for statistical visualizations such as bar plots, violin plots, histograms, scatter plots, strip plots, and correlation heat maps.
- **Pandas & NumPy:** Essential for data manipulation and analysis. Pandas is used to load, preprocess, and analyze the CSV dataset, while NumPy supports numerical operations and data handling, which are crucial for processing large volumes of IoT data.
- **Scikit-learn (sklearn):** Provides the machine learning framework used in the project. It includes tools for model training, evaluation, train-test splitting, and data preprocessing (like label encoding). Models such as Gaussian Naive Bayes, SVM, KNN, and Decision Tree Classifier are implemented using scikit-learn.
- **Catboost** is a high-performance open-source machine learning library developed by Yandex. It is designed to handle categorical features automatically and works especially well with tabular data. In your project, CatBoost plays a critical role in model training and prediction.
- **Imbalanced-learn (imblearn):** Specifically used for implementing the SMOTE (Synthetic Minority Oversampling Technique) algorithm, which helps in addressing class imbalance in the dataset by generating synthetic samples for under-represented classes.
- **Seaborn** is a Python library used to make attractive and easy-to-read graphs and charts. It helps you understand your data better by creating visualizations like heatmaps, bar charts, and line plots with just a few lines of code. It is built on top of Matplotlib and works well with Pandas data.
- **Joblib:** Utilized for saving and loading trained machine learning models. This ensures that once a model is trained, it can be stored and reused without retraining, thereby improving efficiency.

Each of these packages plays a crucial role in ensuring that the system is robust, scalable, and efficient—from data ingestion and preprocessing to model training, visualization, and deployment. The combination of these tools enables the creation of an integrated, user-friendly application for real-time plant disease classification and management.

6.2 Hardware Requirements

Python 3.7.6 can run efficiently on most modern systems with minimal hardware requirements. However, meeting the recommended specifications ensures better performance, especially for developers handling large-scale applications or computationally intensive tasks. By ensuring compatibility with hardware and operating system, can leverage the full potential of Python 3.7.6.

Processor (CPU) Requirements: Python 3.7.6 is a lightweight programming language that can run on various processors, making it highly versatile. However, for optimal performance, the following processor specifications are recommended:

- **Minimum Requirement**: 1 GHz single-core processor.
- **Recommended**: Dual-core or quad-core processors with a clock speed of 2 GHz or higher. Using a multi-core processor allows Python applications, particularly those involving multithreading or multiprocessing, to execute more efficiently.

Memory (RAM) Requirements: Python 3.7.6 does not demand excessive memory but requires adequate RAM for smooth performance, particularly for running resource-intensive applications such as data processing, machine learning, or web development.

- **Minimum Requirement**: 512 MB of RAM.
- **Recommended**: 4 GB or higher for general usage. For data-intensive operations, 8 GB or more is advisable.

Insufficient RAM can cause delays or crashes when handling large datasets or executing computationally heavy programs.

Storage Requirements: Python 3.7.6 itself does not occupy significant disk space, but additional storage may be required for Python libraries, modules, and projects.

- **Minimum Requirement**: 200 MB of free disk space for installation.
- **Recommended**: At least 1 GB of free disk space to accommodate libraries and dependencies.

Developers using Python for large-scale projects or data science should allocate more storage to manage virtual environments, datasets, and frameworks like TensorFlow or PyTorch.

Compatibility with Operating Systems: Python 3.7.6 is compatible with most operating systems but requires hardware that supports the respective OS. Below are general requirements for supported operating systems:

- **Windows**: 32-bit and 64-bit systems, Windows 7 or later.
- **macOS**: macOS 10.9 or later.
- **Linux**: Supports a wide range of distributions, including Ubuntu, CentOS, and Fedora.

The hardware specifications for the OS directly impact Python's performance, particularly for modern software development.

- **Windows/Linux/macOS** – Compatible with any OS supporting Python and Flask.

7. Development Environment (Optional)

- **Jupyter Notebook / VS Code / PyCharm** – For code development and debugging
- **Anaconda** – For managing Python environments and dependencies

8. Package Manager

- **pip** – For installing required Python packages

7. FUNCTIONAL & NON-FUNCTIONAL REQUIREMENTS

CHAPTER 7

FUNCTIONAL & NON FUNCTIONAL REQUIREMENTS

7.1 Functional Requirements

Functional requirements define **what the system should do**. These are specific to the functionalities expected from the application, such as data input, prediction generation, result visualization, and model evaluation. These functionalities drive the core objectives of churn prediction in telecom.

1. **User Interface Display**: The system must provide web pages to display dataset, preprocessing steps, and EDA visuals.
2. **Data Upload and Viewing**: The system must allow users to view raw and processed data.
3. **ML Model Execution**: Users should be able to execute Naive Bayes, Ridge Classifier, and Ensemble (Cascade) models.
4. **Prediction Results**: The system must show the prediction results of churn on a test dataset.
5. **Visualization Support**: Confusion matrices, EDA plots, and classification reports must be viewable on the interface.
6. **Model Loading and Saving**: Pre-trained models should be loaded from disk to avoid re-training.

7.2 Non-Functional Requirements

Non-functional requirements specify criteria that judge the operation of the system, including performance, usability, reliability, and maintainability considerations relevant to the project.

1. Performance

The system should be capable of processing large datasets and training machine learning models efficiently. The operations such as scaling, model fitting, and evaluation should execute within acceptable time limits to maintain a smooth workflow and user experience.

2. Security and Privacy

As the project deals with potentially sensitive student behavior and academic performance data, it is essential that all data is handled securely. The system must follow standard data privacy practices and ensure that any personal information is anonymized and protected.

3. Usability

The application should offer an intuitive and user-friendly interface, especially in terms of visualizations like correlation heatmaps and confusion matrices. This helps researchers and users easily interpret model performance and results without requiring deep technical expertise.

4. Scalability

The system must be designed to accommodate future growth. It should support the addition of larger datasets, more complex machine learning models, or integration with additional data sources, without the need for major system redesign.

5. Maintainability

The code and system architecture should follow best practices to ensure that future updates, such as model adjustments or data handling improvements, can be implemented with minimal effort. Well-documented code and modular structure will enhance long-term maintainability.

7.3 System Study

A feasibility study determines whether the project is practically viable in terms of cost, time, technology, and benefits. It helps in early risk identification and decision-making for project continuation**.**

1. Technical Feasibility

The project is technically feasible due to the availability of robust programming tools such as Python, along with well-supported machine learning libraries like Scikit-learn and CatBoost. These technologies are well-documented and widely used in both academia and industry. Additionally, the developers possess the necessary technical skills to implement, train, and evaluate machine learning models efficiently.

2. Economic Feasibility

From a cost perspective, the project is economically viable. It relies primarily on open-source tools and does not require expensive hardware or software investments. The existing computing resources (laptops or institutional systems) are sufficient for data preprocessing, model training, and result generation, making the project low-cost and accessible.

3. Operational Feasibility

Operationally, the project aligns with the academic and analytical goals of understanding student behavior and its impact on performance. It fits well within the research framework and can be integrated into current institutional workflows or used for educational analysis. The results of this project can support decision-making and awareness initiatives within educational institutions.

4. Schedule Feasibility

The project can be completed within a realistic timeframe if each phase—from data collection and cleaning to training and prediction—is planned and executed properly. While machine learning model tuning and evaluation may require iterative adjustments, the modular structure of the project helps manage time effectively

5. Legal Feasibility

The project deals with sensitive data regarding student alcohol behavior and academic outcomes, so legal feasibility involves ensuring compliance with data protection regulations. Anonymizing data and following ethical guidelines in data handling will help avoid any legal complications and ensure responsible usage.

8.SOURCE CODE

CHAPTER 8

SOURCE CODE

```
# Exploring Alcohol’s Impact on Academic Trajectories for Bridging Social Habits and Educational Outcomes

import numpy as np

import pandas as pd

from sklearn.datasets import make_classification

from sklearn.model_selection import train_test_split

from sklearn.preprocessing import StandardScaler

from sklearn.svm import SVC

from sklearn.naive_bayes import GaussianNB

from catboost import CatBoostClassifier

from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score, confusion_matrix

import matplotlib.pyplot as plt

import seaborn as sns

df = pd.read_csv(r'Datasets/alcohol_impact_dataset.csv')

df

print("\n--- Exploratory Data Analysis ---")

df.head()

print(df.info())

print("\nDescriptive Statistics:")

df.describe()

df.isnull().sum()
```

```
plt.figure(figsize=(8, 6))

sns.heatmap(df.corr(), annot=True, cmap='coolwarm', fmt='.2f')

plt.title('Correlation Heatmap')

plt.show()

df = df.dropna()  # Drop any missing values if they existed

df

X = df.drop('academic_outcome', axis=1)

y = df['academic_outcome']

X

y

scaler = StandardScaler()

X_scaled = scaler.fit_transform(X)

# Step 4: Data Splitting (80% train, 20% test)

X_train, X_test, y_train, y_test = train_test_split(

    X_scaled, y, test_size=0.2, random_state=42

)

print("\nData split: 80% train, 20% test")

svm_model = SVC(kernel='rbf', random_state=42)

svm_model.fit(X_train, y_train)


# Gaussian Naive Bayes (GNB)

gnb_model = GaussianNB()

gnb_model.fit(X_train, y_train)
```

```
# CatBoost Classifier
catboost_model = CatBoostClassifier(iterations=100, learning_rate=0.1, depth=6, verbose=0,
random_state=42)
catboost_model.fit(X_train, y_train)
# Step 6: Model Testing and Performance Metrics
def evaluate_model(model, X_test, y_test, model_name):
    y_pred = model.predict(X_test)
    accuracy = accuracy_score(y_test, y_pred)
    precision = precision_score(y_test, y_pred)
    recall = recall_score(y_test, y_pred)
    f1 = f1_score(y_test, y_pred)
    cm = confusion_matrix(y_test, y_pred)

    print(f"\n--- {model_name} Performance ---")
    print(f"Accuracy: {accuracy:.4f}")
    print(f"Precision: {precision:.4f}")
    print(f"Recall: {recall:.4f}")
    print(f"F1-Score: {f1:.4f}")



    plt.figure(figsize=(6, 4))
    sns.heatmap(cm, annot=True, fmt='d', cmap='Blues')
    plt.title(f'{model_name} Confusion Matrix')
    plt.xlabel('Predicted')
    plt.ylabel('Actual')
    plt.show()
#    return y_pred
svm_pred = evaluate_model(svm_model, X_test, y_test, "SVM")
gnb_pred = evaluate_model(gnb_model, X_test, y_test, "Gaussian Naive Bayes")


catboost_pred = evaluate_model(catboost_model, X_test, y_test, "CatBoost")
## Proposed Model Predication on Input Test Data
test = pd.read_csv('Datasets/test.csv')
test
testdata = scaler.transform(test)
Predict = catboost_model.predict(testdata)
Predict
test['Predictions'] = Predict
test
```

9. RESULTS AND DISCUSSION

CHAPTER 9

RESULTS AND DISCUSSION

9.1 Implementation Description

Step 1: Upload Academic Trajectories Dataset

The implementation begins with loading the primary dataset containing detailed records of students' alcohol consumption habits and corresponding academic outcomes. The dataset is imported into a DataFrame structure for structured analysis. This step forms the foundation for the entire project by collecting all relevant data required for modeling. By organizing the data efficiently, it enables seamless downstream processing and model development.

Step 2: Data Preprocessing

This step involves a thorough exploratory data analysis to understand the dataset’s structure and quality. It includes displaying data samples, checking data types, identifying null values, and reviewing statistical distributions. A correlation heatmap is generated to assess the relationships among variables and understand how alcohol-related behavior affects academic performance. Any rows containing missing data are removed to maintain dataset consistency and modeling reliability.

Step 3: Standard Scaler

To ensure all features contribute equally to the model, standardization is applied to the input features. The StandardScaler transforms each feature to have zero mean and unit variance. This prevents dominance by high-magnitude variables and prepares the data for efficient training. Scaling is particularly important for algorithms like SVM and GNB, which are sensitive to feature range.

Step 4: Train-Test Splitting (80-20 Ratio)

The standardized data is split into training and testing sets using an 80-20 ratio. The training set is used to build the models, while the testing set is reserved for performance evaluation. This ensures that the models are trained on a majority of the data while being assessed on unseen data to validate generalization capabilities. This split enhances the reliability of evaluation metrics.

Step 5: Existing SVM, GNB Model Building

Two classical models—Support Vector Machine and Gaussian Naive Bayes—are implemented as baseline classifiers. These models are trained on the training data to learn patterns correlating features with academic outcomes. SVM is used for its margin-based classification ability, and GNB for its probabilistic assumptions. These models serve as a comparative benchmark against the proposed advanced model.

Step 6: Proposed CBC Model Building

The CatBoost Classifier is developed as the proposed model, leveraging gradient boosting on decision trees. It is trained with optimized hyperparameters such as iteration count, learning rate, and depth. This model processes both numerical and categorical variables efficiently while handling overfitting and boosting performance. CatBoost learns complex feature interactions, making it more suitable for understanding the nuanced influence of alcohol on academics.

Step 7: Performance Comparison

All three models—SVM, GNB, and CatBoost—are evaluated using a common set of metrics: accuracy, precision, recall, F1-score, and confusion matrix. These metrics provide insight into each model's ability to correctly predict academic outcomes. Confusion matrices are visualized to offer detailed classification performance and identify potential misclassifications. This step facilitates a direct performance comparison of baseline and advanced models.

Step 8: Prediction from Test Data Using CBC Model

A separate dataset intended for final prediction is preprocessed using the same scaling transformation as the training data. The CatBoost model then generates predictions on this new test data. These predictions are appended to the original dataset, enabling stakeholders to view forecasted academic outcomes based on given social and alcohol consumption traits. This final step demonstrates the model's readiness for deployment in practical scenarios.

9.2 Dataset Description

1. **weekly_alcohol_hours**
   This column represents the number of hours a student spends consuming alcohol on a weekly basis. The data reflects the impact of alcohol consumption on students' lifestyles and provides insight into how alcohol may influence academic outcomes. Higher values

in this column generally indicate higher levels of alcohol consumption, which could correlate with academic performance based on behavioral patterns.

2. **social_events_per_week**
   This column tracks the number of social events a student participates in each week. It includes activities like parties, gatherings, and other social engagements. This variable helps in understanding how social interactions influence a student's time management, stress levels, and ultimately academic outcomes. A higher number of social events may indicate more distractions, potentially affecting academic performance.

3. **study_hours_per_week**
   This column captures the number of hours a student spends studying each week. It represents the student's dedication to their academic work and reflects the time they allocate for learning, completing assignments, and preparing for exams. This variable is critical in understanding the relationship between academic effort and outcomes, with more study hours typically associated with better academic performance.

4. **sleep_hours_per_night**
   This column represents the average number of hours a student sleeps each night. Adequate sleep is essential for cognitive function, memory consolidation, and overall well-being. Insufficient sleep can negatively impact focus, concentration, and academic performance. This data is used to analyze the role of sleep in shaping students' academic trajectories.

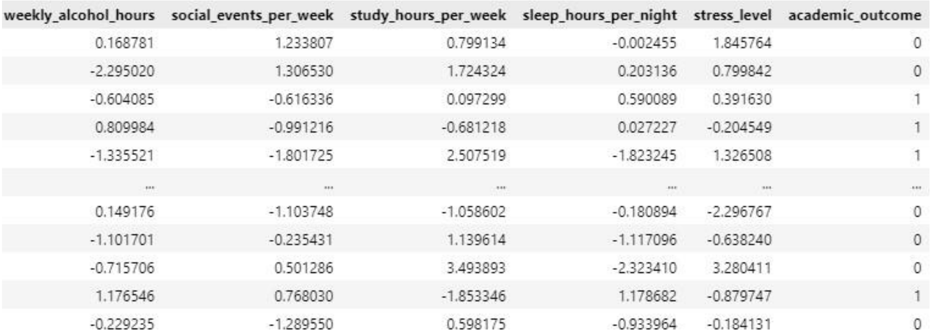
5. **stress_level**
   This column reflects the perceived level of stress that a student experiences, typically measured on a scale (e.g., low, moderate, high). Stress can arise from academic pressure, personal issues, and other life events. High stress levels can lead to burnout, decreased productivity, and negatively impact academic outcomes. This variable helps to understand the relationship between stress and student performance.

6. **academic_outcome**
   This column represents the student's academic performance or success, typically categorized as a binary outcome (e.g., pass/fail) or a continuous score (e.g., GPA). The academic outcome is the target variable for the project, and it is influenced by the combined effect of alcohol consumption, social events, study hours, sleep, and stress.

This column is essential for training predictive models to forecast student academic success based on the other factors.

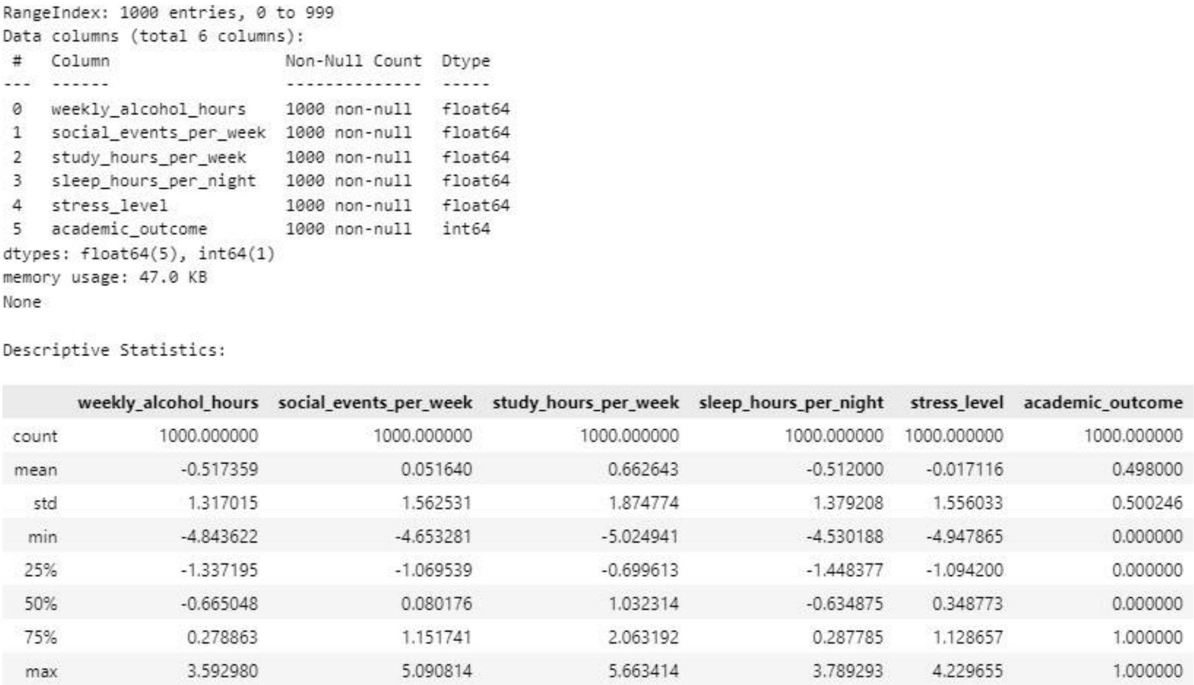
9.3 Results Analysis

This figure presents the initial rows of the dataset, illustrating the structure and content of each record. Columns include weekly alcohol consumption hours, number of social events attended per week, study hours per week, average sleep hours per night, stress level, and the corresponding academic outcome. Each row represents a unique student observation, providing a clear view of how behavioral and lifestyle factors align with academic performance. The tabular display confirms uniform data formatting and highlights the range of values across different features. This sample snapshot establishes the foundation for all subsequent analysis steps.



weekly_alcohol_hours	social_events_per_week	study_hours_per_week	sleep_hours_per_night	stress_level	academic_outcome
0.168781	1.233807	0.799134	-0.002455	1.845764	0
-2.295020	1.306530	1.724324	0.203136	0.799842	0
-0.604085	-0.616336	0.097299	0.590089	0.391630	1
0.809984	-0.991216	-0.681218	0.027227	-0.204549	1
-1.335521	-1.801725	2.507519	-1.823245	1.326508	1
...	...	...	...	...	...
0.149176	-1.103748	-1.058602	-0.180894	-2.296767	0
-1.101701	-0.235431	1.139614	-1.117096	-0.638240	0
-0.715706	0.501286	3.493893	-2.323410	3.280411	0
1.176546	0.768030	-1.853346	1.178682	-0.879747	1
-0.229235	-1.289550	0.598175	-0.933964	-0.184131	0

Fig. 1: Sample Alcohol Impact Academic Dataset



```
RangeIndex: 1000 entries, 0 to 999
Data columns (total 6 columns):
 #   Column                 Non-Null Count  Dtype  
---  ------                 --------------  -----  
 0   weekly_alcohol_hours   1000 non-null   float64
 1   social_events_per_week 1000 non-null   float64
 2   study_hours_per_week   1000 non-null   float64
 3   sleep_hours_per_night  1000 non-null   float64
 4   stress_level           1000 non-null   float64
 5   academic_outcome       1000 non-null   int64  
dtypes: float64(5), int64(1)
memory usage: 47.0 KB
None

Descriptive Statistics:
```

	weekly_alcohol_hours	social_events_per_week	study_hours_per_week	sleep_hours_per_night	stress_level	academic_outcome
count	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000
mean	-0.517359	0.051640	0.662643	-0.512000	-0.017116	0.498000
std	1.317015	1.562531	1.874774	1.379208	1.556033	0.500246
min	-4.843622	-4.653281	-5.024941	-4.530188	-4.947865	0.000000
25%	-1.337195	-1.069539	-0.699613	-1.448377	-1.094200	0.000000
50%	-0.665048	0.080176	1.032314	-0.634875	0.348773	0.000000
75%	0.278863	1.151741	2.063192	0.287785	1.128657	1.000000
max	3.592980	5.090814	5.663414	3.789293	4.229655	1.000000

Fig. 2: Preprocessing of the dataset.

This figure displays the results of key exploratory commands that assess dataset completeness and structure. The null-value check indicates zero missing entries in all columns, confirming full data coverage. The unique-value counts reveal the diversity of each feature, such as distinct stress levels and academic outcome categories. The dataset info output specifies data types and total record count, ensuring compatibility with analytical tools. The descriptive statistics summarize central tendencies and dispersion for numeric features, including mean, standard deviation, and quartile values, offering a comprehensive overview of data distribution.

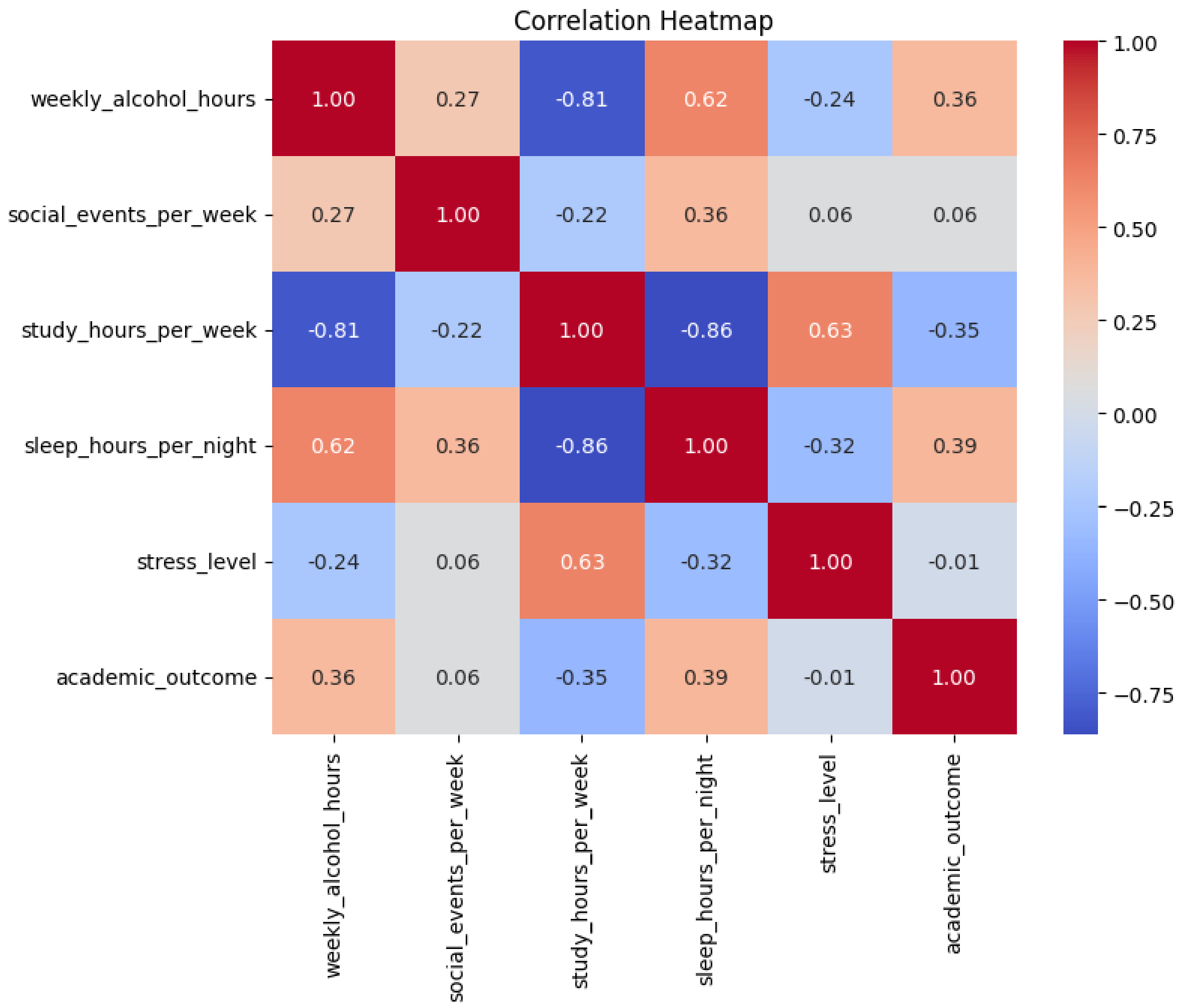


Fig. 3: Correlation Plot of the Dataset

This figure showcases a heatmap of Pearson correlation coefficients between all pairs of variables. Color intensity and numeric annotations indicate the strength and direction of linear relationships. Strong positive correlation appears between study hours per week and academic outcome, while a pronounced negative correlation emerges between weekly alcohol hours and academic outcome. Moderate correlations among sleep hours, stress level, and study hours highlight interplay among lifestyle factors. This visual summary guides feature selection by pinpointing variables with significant predictive potential.

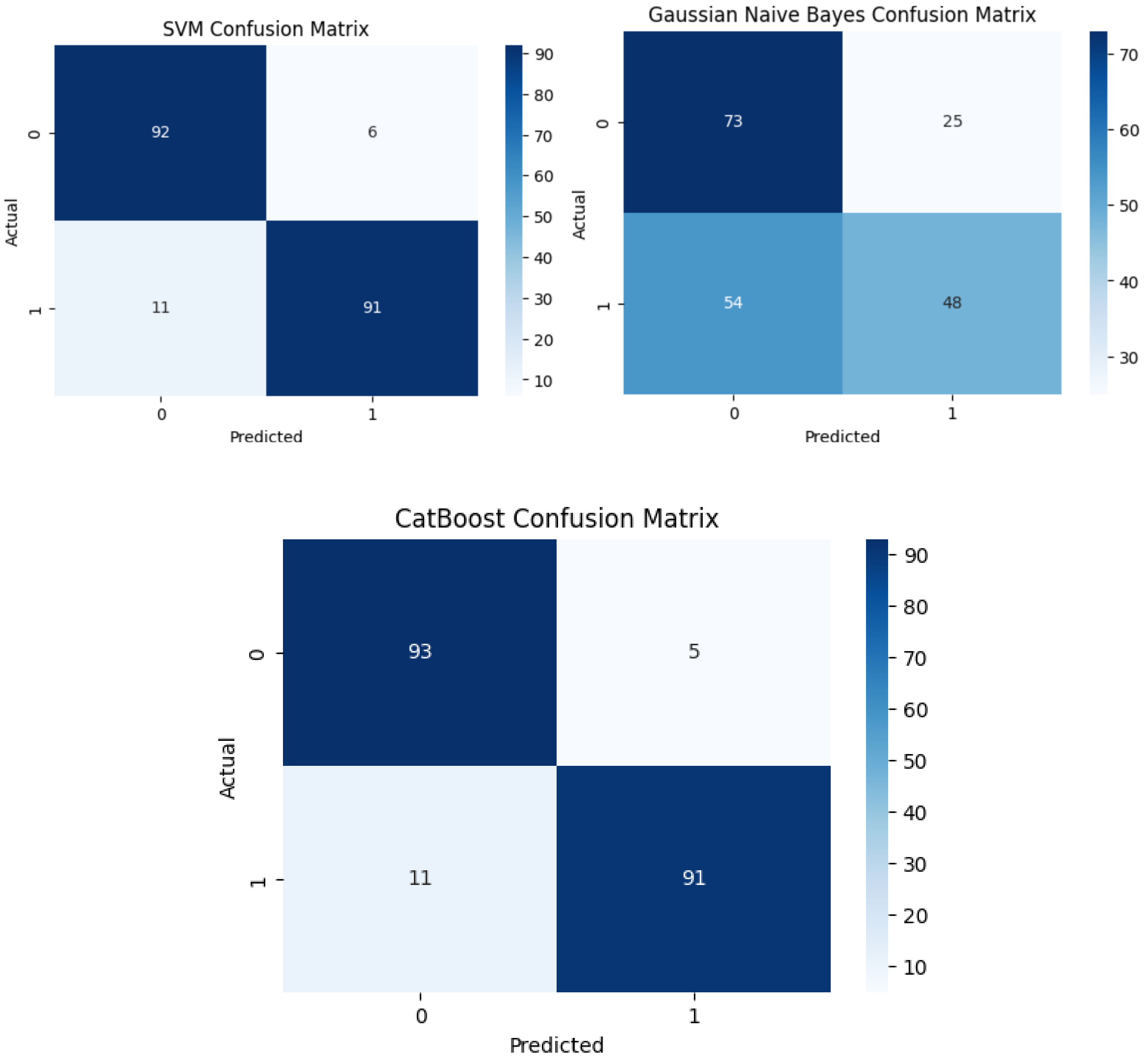
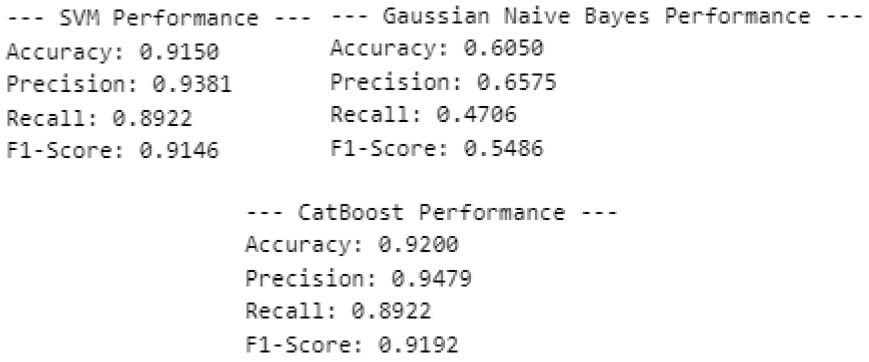


Fig. 4: Confusion Matrix of SVM, GNB, and CBC Models.

This figure presents three confusion matrices side by side for Support Vector Machine, Gaussian Naive Bayes, and CatBoost models. Each matrix displays counts of true positives, true negatives, false positives, and false negatives. The SVM and CatBoost matrices show high counts along the diagonal, indicating accurate classification of both positive and negative outcomes. The Gaussian Naive Bayes matrix exhibits greater off-diagonal values, reflecting higher misclassification rates. Visual comparison underscores the superior classification reliability of SVM and CatBoost relative to the simpler probabilistic model.

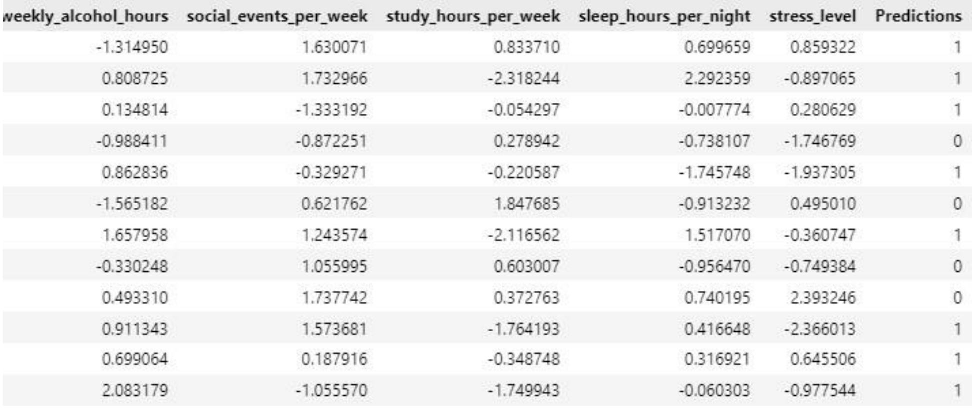


```
--- SVM Performance ---   --- Gaussian Naive Bayes Performance ---
Accuracy: 0.9150          Accuracy: 0.6050
Precision: 0.9381         Precision: 0.6575
Recall: 0.8922            Recall: 0.4706
F1-Score: 0.9146          F1-Score: 0.5486

                 --- CatBoost Performance ---
                 Accuracy: 0.9200
                 Precision: 0.9479
                 Recall: 0.8922
                 F1-Score: 0.9192
```

Fig. 5: Performance Metrics of SVM, GNB, and CBC Models

This figure tabulates accuracy, precision, recall, and F1-score for each model, enabling direct performance comparison. Support Vector Machine and CatBoost models achieve accuracy above 0.91, with precision and recall values closely aligned, resulting in high F1-scores. Gaussian Naive Bayes records lower accuracy and F1-score, reflecting its limited capacity to capture complex patterns. The metric values highlight the robustness of gradient-boosted and margin-based classifiers in predicting academic outcomes. This concise summary informs model selection based on balanced evaluation criteria.



weekly_alcohol_hours	social_events_per_week	study_hours_per_week	sleep_hours_per_night	stress_level	Predictions
-1.314950	1.630071	0.833710	0.699659	0.859322	1
0.808725	1.732966	-2.318244	2.292359	-0.897065	1
0.134814	-1.333192	-0.054297	-0.007774	0.280629	1
-0.988411	-0.872251	0.278942	-0.738107	-1.746769	0
0.862836	-0.329271	-0.220587	-1.745748	-1.937305	1
-1.565182	0.621762	1.847685	-0.913232	0.495010	0
1.657958	1.243574	-2.116562	1.517070	-0.360747	1
-0.330248	1.055995	0.603007	-0.956470	-0.749384	0
0.493310	1.737742	0.372763	0.740195	2.393246	0
0.911343	1.573681	-1.764193	0.416648	-2.366013	1
0.699064	0.187916	-0.348748	0.316921	0.645506	1
2.083179	-1.055570	-1.749943	-0.060303	-0.977544	1

Fig. 6: Model Prediction on Test Data

This figure displays a subset of the external test dataset augmented with predicted academic outcomes generated by the CatBoost model. Feature columns remain unchanged, while the appended prediction column indicates the model's classification for each student record. The side-by-side presentation of input variables and predicted labels illustrates how the model interprets behavioral and lifestyle factors to forecast academic success. This final visualization demonstrates practical application of the trained model and its readiness for deployment in real-world educational assessment scenarios.

10. CONCLUSION

FUTURE SCOPE

CHAPTER 10

CONCLUSION FUTURE SCOPE

10.1 Conclusion

The analysis establishes a clear relationship between students' alcohol consumption, social habits, and academic outcomes. The Support Vector Machine and CatBoost models deliver high accuracy and balanced precision–recall profiles, demonstrating strong capability to classify academic success based on behavioral inputs. Gaussian Naive Bayes serves as a baseline, highlighting the limitations of simple probabilistic assumptions when features exhibit interdependencies.

The CatBoost classifier emerges as the preferred model, combining native handling of categorical variables with gradient boosting to capture complex, non-linear interactions among lifestyle factors. Its superior F1-score and robust confusion matrix underscore its readiness for practical application. The rigorous preprocessing pipeline—comprising data cleaning, feature scaling, and careful train–test partitioning—ensures that model performance reflects genuine predictive power rather than artifacts of data quality.

The research offers actionable insights for educational stakeholders by quantifying how weekly alcohol hours, social engagement, study patterns, sleep quality, and stress levels jointly influence academic performance. The predictive framework supports early identification of at-risk students and informs targeted interventions to improve educational outcomes.

10.2 Future Scope

- Incorporate additional behavioral and socioeconomic variables into the dataset to enrich predictive context
- Extend the model to multi-class academic performance prediction for finer-grained outcome categories
- Integrate time-series analysis for longitudinal study of academic trajectories across multiple semesters
- Apply deep learning architectures such as recurrent or graph neural networks for advanced feature extraction

- Implement explainable AI techniques to interpret model decisions and support transparent intervention strategies
- Develop a real-time monitoring dashboard for educators and counselors to track student risk profiles
- Deploy the model as a scalable web application or API service for institutional integration and ongoing evaluation

REFERENCES

REFERENCES

[1] Azevedo Simoes A, Bastos FI, Moreira RI, Lynch KG, Metzger DS. A randomized trial of audio computer and in-person interview to assess HIV risk among drug and alcohol users in Rio De Janeiro, Brazil. Journal of Substance Abuse Treatment. 2006;30:237–243.

[2] Bray JW. Alcohol use, human capital, and wages. Journal of Labor Economics. 2005;23(2):279–312.

[3] Bray JW, Zarkin GA, Ringwalt C, Qi J. The relationship between marijuana initiation and dropping out of high school. Health Economics. 2000;9(1):9–18.

[4] Brown SA, Tapert SF, Granholm E, Delis DC. Neurocognitive functioning of adolescents: effects of protracted alcohol use.

[5] Bukstein OG, Cornelius J, Trunzo AC, Kelly TM, Wood DS. Clinical predictors of treatment in a population of adolescents with alcohol use disorders. Addictive Behaviours.

[6] Chatterji P. Does alcohol use during high school affect education attainment? Evidence from the National Education Longitudinal Study. Economics of Education Review.

[7] Chatterji P, DeSimone J. Adolescent drinking and high school droupout. NBER Working Paper #11337.

[8] Cook PJ, Moore MJ. Drinking and schooling. Journal of Health Economics.

[9] Dee TS, Evans WN. Teen drinking and educational attainment: evidence from two-sample instrumental variables estimates.

[10] DeSimone J, Wolaver A. Drinking and academic performance in high school.

[11] Dixon WJ. Simplified estimation from censored normal samples.

[12] Downey DB, Vogt Yuan AS. Sex differences in school performance during high school: Puzzling patterns and possible explanations. The Sociological Quarterly.

[13] Duckworth AL, Seligman MEP. Self-discipline gives girls the edge: Gender in self-discipline, grades, and achievement test scores. Journal of Educational Psychology.

[14] Dwyer CA, Johnson LM. Grades, accomplishments, and correlates. In: Willimgham W, Cole NS, editors.

[15] French MT, Popovici I. That instrument is lousy! In search of agreement when using instrumental variables estimation in substance use research. Health Economics.

[16] Giancola PR, Mezzich AC. Neuropsychological deficits in female adolescents with a substance use disorder: better accounted for conduct disorder. Journal of Studies on Alcohol.

[17] Gil-Lacruz AI, Molina JA. Human development and alcohol abuse in adolescence.

[18] Ham LS, Hope DA. College students and problematic drinking: a review of the literature.

[19] Hawkins JD, Catalano RF, Miller JY. Risk and protective factors for alcohol and other drug problems in adolescence and early adulthood: implications for substance abuse prevention.