Available online at www.ijournalse.org

Emerging Science Journal

(ISSN: 2610-9182)

Vol. 8, No. 4, August, 2024

Analyzing Socio-Academic Factors and Predictive Modeling of Student Performance Using Machine Learning Techniques

Romel Al-Ali [1*], Khadija Alhumaid [2], Maha Khalifa [2], Said A. Salloum [3, 4], Rima Shishakly [5], Mohammed Amin Almaiah [6*]

[1] The National Research Center for Giftedness and Creativity, King Faisal University, Al Hofuf 31982, Saudi Arabia.

[2] Student Services, Rabdan Academy, Abu Dhabi, United Arab Emirates.

[3] Health Economic and Financing Group, University of Sharjah, Sharjah, United Arab Emirates.

[4] School of Science, Engineering, and Environment, University of Salford, United Kingdom.

[5] Management Department, College of Business Administration, Ajman University, Ajman 346, United Arab Emirates.

[6] King Abdullah the II IT School, The University of Jordan, Amman 11942, Jordan.

Abstract

Understanding the factors that influence student performance is crucial for improving educational outcomes. Thus, this study aims to examine the impact of socio-economic and psychological factors on student performance, less is known about how students' personal attitudes and behaviors across different departments and activities correlate with their academic success. This study employs exploratory data analysis (EDA) to identify trends and relationships within the dataset. Machine learning techniques, such as K-means clustering and Long Short-Term Memory (LSTM) networks, are utilized to model and predict student performance based on their reported behaviors and preferences. The dataset is reduced using Principal Component Analysis (PCA) to enhance the clustering process. The findings suggest significant variations in academic performance based on departmental affiliation, gender, and engagement in certification courses. The LSTM model achieved an accuracy of 91% on the test set, demonstrating substantial predictive capability. However, the classification report reveals that while the model was highly effective in identifying the majority class (label 1), achieving a precision of 91% and a recall of 100%, it failed to correctly predict any instances of the minority class (label 0). The insights from this study could help educators tailor interventions to address the specific needs of students based on their behaviors and departmental affiliations, leading to more personalized education strategies and potentially improving academic outcomes.

Keywords:

Academic Performance;
Behavior; Certification Courses;
Clustering; Data Analysis;
Departmental Influences;
LSTM; PCA;
Student Attitudes.

Article History:

Received:	07	April	2024
Revised:	08	July	2024
Accepted:	16	July	2024
Published:	01	August	2024

1- Introduction

Understanding the factors that influence student performance is crucial for enhancing educational outcomes and tailoring educational strategies to meet individual needs [1]. The academic performance of students is influenced by a myriad of factors, including socio-economic status, psychological well-being, educational environment, and personal attitudes and behaviors [2]. Numerous studies have highlighted the importance of these factors in shaping academic success. For instance, socioeconomic factors such as parental education and income level have been shown to

* **CONTACT**: ralali@kfu.edu.sa; m.almaiah@ju.edu.jo

DOI: http://dx.doi.org/10.28991/ESJ-2024-08-04-05

significantly impact student achievement [3-8]. Additionally, psychological factors, including motivation and self-efficacy, play a critical role in academic performance [9, 10].

In recent years, there has been an increasing interest in understanding how personal attitudes and behaviors influence academic outcomes. Behaviors such as time management, study habits, and engagement in extracurricular activities have been linked to academic success [11, 12]. Moreover, the role of certification courses, departmental influences, and gender differences in shaping student attitudes and performance is gaining attention. Studies have shown that students who engage in certification courses tend to have better academic outcomes due to the additional skills and knowledge gained [13]. Similarly, gender differences in academic performance have been widely studied, with mixed results suggesting both biological and sociocultural influences [14, 15].

Despite the extensive research on the various factors affecting academic performance, there is a need for comprehensive studies that integrate multiple dimensions of student behaviors and attitudes. The "Student Attitude and Behavior" dataset provides a unique opportunity to explore these dimensions in detail. This dataset includes information on students' engagement in certification courses, gender, department, hobbies, and other lifestyle aspects, allowing for a holistic analysis of their impact on academic performance.

From 2021 to 2024, research on analyzing socio-academic factors and predictive modeling of student performance using machine learning techniques has continued to expand, offering fresh insights and innovative methodologies. These studies emphasize the critical role of integrating diverse socio-academic variables and advanced machine learning algorithms to enhance predictive accuracy and provide actionable insights for educational interventions. In 2021, a study by Johnson et al. (2024) [3] utilized random forest algorithms to predict student performance based on a comprehensive dataset that included socio-economic status, parental education levels, and student engagement metrics. Their findings highlighted the importance of considering a wide range of socio-academic factors and demonstrated the random forest's robustness in handling complex, multi-dimensional data. Further, Zhang et al. (2021) [4] explored the application of deep learning models, specifically Long Short-Term Memory (LSTM) networks, to capture temporal patterns in student performance data. Their research showcased how LSTM networks could effectively predict academic outcomes by analyzing sequences of student behaviors and academic records over time, achieving higher accuracy than traditional machine learning models.

A significant study by Keller et al. (2022) [5] focused on the role of gender differences and departmental influences in academic performance. By employing K-means clustering, they were able to segment students into distinct groups, revealing significant variations in performance across different departments and gender. Their work underscored the necessity of personalized educational strategies to address these disparities. In 2024, Liu et al. examined the impact of extracurricular activities and certification courses on student performance using support vector machines (SVM) and Principal Component Analysis (PCA) for dimensionality reduction. Their study found that engagement in extracurricular activities positively correlated with academic success, while the inclusion of PCA improved the SVM model's efficiency and accuracy. Additionally, recent research by Ishaq et al. (2021) [6] addressed the persistent issue of class imbalance in predictive modeling. They proposed a novel approach combining synthetic minority over-sampling technique (SMOTE) with ensemble learning methods to enhance the prediction of minority class outcomes. Their findings highlighted the improved balance and predictive capability of their model, which is crucial for developing equitable educational tools. These studies collectively advance our understanding of how socio-academic factors influence student performance and demonstrate the power of machine learning techniques in predictive modeling. They also emphasize the ongoing need to address challenges such as class imbalance and to refine models for more equitable and effective educational applications.

This study aims to fill the gap in the literature by providing an integrated analysis of various factors influencing student performance using advanced data analytics and machine learning techniques. By employing exploratory data analysis (EDA) and machine learning models, such as Long Short-Term Memory (LSTM) networks, this research offers a detailed examination of the relationships and distributions of attributes within the dataset. The study also addresses the issue of class imbalance in predictive modeling, providing insights into the limitations and potential improvements in educational data analytics.

The paper is structured as follows: Literature Review provides a detailed examination of existing research on factors influencing student performance. Methodology describes the dataset, preprocessing steps, and the analytical methods employed, including EDA, PCA, K-means clustering, and LSTM modeling. Results presents the findings from the data analysis and machine learning models, including performance metrics and classification reports. Discussion interprets the results, discusses implications for educational strategies, and offers recommendations for addressing class imbalance in predictive modeling. Conclusion summarizes the key findings, contributions to the field, and suggestions for future research.

2- Related Work

The academic performance of students has been a subject of extensive research, with numerous studies exploring various factors that contribute to educational outcomes. These factors range from individual characteristics such as cognitive abilities, motivation, and socio-economic status to institutional factors including school resources, teacher quality, and classroom environment.

A significant body of research indicates that SES is a strong predictor of academic success. Students from higher SES backgrounds typically have access to more educational resources, supportive learning environments, and opportunities for extracurricular activities, all of which contribute to better academic performance [8, 16]. Both cognitive abilities, such as intelligence and prior academic achievement, and non-cognitive skills, such as perseverance, motivation, and self-regulation, have been shown to influence student performance. Duckworth & Seligman [17] highlighted the importance of self-discipline over IQ in predicting academic achievement.

The quality of instruction has a profound impact on student learning outcomes. Research by Rivkin et al. [18] demonstrated that teacher effectiveness is one of the most important school-related factors influencing student achievement. The learning environment, including classroom climate and school culture, also plays a critical role. Positive school climates that promote safety, student engagement, and a sense of belonging are associated with higher academic achievement [19].

The application of machine learning techniques in educational data mining (EDM) has gained momentum over the past decade. These techniques are used to analyze large datasets to uncover patterns and predict student outcomes. Various classification algorithms such as decision trees, random forests, support vector machines (SVM), and neural networks have been employed to predict student performance. For instance, Yadav et al. [20] used decision trees to identify at-risk students, while He et al. [21] applied SVMs to predict academic success. Deep learning models, particularly Long Short-Term Memory (LSTM) networks, have shown promise in capturing temporal patterns in sequential data. LSTM networks have been used to model student learning behaviors and predict future performance based on past interactions [22]. Clustering methods such as K-means are utilized to segment students into groups based on similar characteristics. These clusters can provide insights into different learning styles and performance levels, facilitating targeted interventions [23].

Despite the extensive research on factors influencing student performance and the application of machine learning techniques in EDM, several gaps remain. Firstly, there is a need for more studies that integrate multiple types of data (e.g., academic records, socio-economic data, behavioral data) to provide a holistic understanding of student performance. Secondly, while deep learning models have shown potential, issues such as class imbalance in datasets need to be addressed to improve model accuracy and reliability. Thirdly, the majority of studies focus on short-term predictions; there is a paucity of research on long-term academic outcomes. Lastly, there is a lack of research on how insights derived from these models can be effectively translated into actionable educational strategies and policies.

3- Research Methodology

3-1- Data Description

The dataset used in this study, titled "Student Attitude and Behavior," was sourced from Kaggle [24]. This dataset contains various attributes related to students' academic performance, demographics, and personal preferences. It includes both temporal and sequential aspects, capturing data points over a specific period, which is crucial for understanding changes and patterns in student behavior over time. The dataset comprises the following features: Certification Course, Gender, Department, Height (CM), Weight (KG), 10th Mark, 12th Mark, College Mark, Hobbies, Daily Studying Time, Preferred Study Time, Salary Expectation, Degree Satisfaction, Willingness to Pursue Career Based on Degree, Social Media Usage, Travel Time, Stress Level, Financial Status, and Part-time Job.

3-2- Data Preparation

Data preparation is a critical step in the analytical process, ensuring the data is clean, consistent, and suitable for modeling. The preprocessing steps undertaken for this study are as follows:

- ***Handling Missing Values*:** Missing values in the dataset were addressed using appropriate imputation techniques [25]. Numerical missing values were filled using the mean or median of the respective columns, while categorical missing values were filled with the most frequent category.
- ***Normalization*:** To standardize the numerical features, normalization was applied [26]. This involved scaling the numerical features to a range of 0 to 1, using the Min-Max scaling method, which helps in improving the performance of machine learning models [27].
- ***Encoding Categorical Data*:** Categorical features were encoded into numerical values using one-hot encoding [28]. This technique converts categorical variables into a binary matrix, making them suitable for input into the LSTM model.

- ***Creating Sequential Data*****:** Given the temporal and sequential nature of the dataset, the data was organized into sequences to capture the order of events and interactions over time [29]. This step is crucial for the LSTM model, which is designed to process sequential data and learn temporal dependencies.

3-3- LSTM Network Architecture

The Long Short-Term Memory (LSTM) network is a type of recurrent neural network (RNN) that is well-suited for sequence prediction problems [30]. The LSTM model used in this study was configured as follows:

- ***Input Layer***: The input layer receives the sequential data. Given the multi-dimensional nature of the dataset, the input shape was defined accordingly.
- ***LSTM Layers***: The model includes multiple LSTM layers. Each layer consists of a certain number of hidden units (neurons) that learn the temporal dependencies in the data. In this study, two LSTM layers with 50 and 30 hidden units respectively were used. The number of layers and units was chosen based on experimentation and cross-validation to balance complexity and performance.
- ***Dense Layer***: Following the LSTM layers, a dense (fully connected) layer was added to transform the learned representations into the desired output shape.
- ***Output Layer***: The output layer produces the final predictions. For binary classification, a single neuron with a sigmoid activation function was used to output probabilities.
- ***Activation Functions***: The LSTM layers utilized the ReLU activation function to introduce non-linearity, while the output layer used the sigmoid activation function for binary classification.
- ***Optimization and Loss Function***: The Adam optimizer was used for training the model, given its efficiency and effectiveness in handling sparse gradients. The binary cross-entropy loss function was employed to measure the error between predicted and actual values.

3-4- Supplementary Techniques

3-4-1- Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a dimensionality reduction technique that transforms high-dimensional data into a lower-dimensional form while retaining most of the variance [31]. In this study, PCA was used to reduce the dimensionality of the dataset before applying clustering techniques. By doing so, it simplifies the data, making it easier to visualize and interpret the clusters formed by K-means.

- ***Steps*****:**
    - Standardization of data: Ensuring all features contribute equally by scaling them.
    - Computation of covariance matrix: To understand the data's variance.
    - Calculation of eigenvalues and eigenvectors: To identify principal components.
    - Selection of top principal components: Based on the explained variance, choosing components that capture most of the data’s variability.

3-4-2- K-means Clustering

K-means clustering is an unsupervised machine learning algorithm used to partition data into K distinct clusters based on similarity. This technique was applied to the PCA-transformed data to identify distinct groups of students based on their attitudes and behaviors.

- ***Steps*****:**
    - *Initialization***:** Selecting K initial centroids randomly.
    - *Assignment***:** Assigning each data point to the nearest centroid, forming K clusters.
    - *Update***:** Recalculating the centroids as the mean of all points assigned to each cluster.
    - *Iteration***:** Repeating the assignment and update steps until convergence, i.e., when centroids no longer change significantly.
    - *Determination of K***:** The optimal number of clusters (K) was determined using the Elbow Method, which involves plotting the sum of squared distances from each point to its assigned centroid and looking for an 'elbow' point where the rate of decrease sharply slows.

3-5- Implementation Details

The implementation of the LSTM model and K-means clustering was done using Python, leveraging libraries such as TensorFlow, Keras, Scikit-learn, and Pandas. The data preparation and PCA were also performed using Scikit-learn, ensuring efficient processing and analysis. The LSTM model was trained and validated using cross-validation to ensure robustness and generalizability of the results.

3-6- Model Evaluation

The performance of the LSTM model was evaluated using several metrics, including accuracy, precision, recall, F1-score, and the confusion matrix. These metrics provided insights into the model's ability to correctly predict student performance. The results indicated that while the LSTM model achieved high accuracy, it struggled with class imbalance, necessitating further investigation and potential solutions such as oversampling the minority class or using advanced techniques like SMOTE.

4- Results

This section provides a detailed explanation of the data analytics applied to the "Student Attitude and Behavior" dataset. Through exploratory data analysis (EDA), we explore the distribution and interrelationships among different features in relation to student performance, particularly focusing on their 10th-grade marks. This analysis highlights the significant findings from training an LSTM model, performing K-means clustering on a feature set reduced via PCA, and evaluating the LSTM model's effectiveness in predicting student attitudes and behaviors.

4-1- Data Analytics

This analytical section delves into the interconnections and patterns among different variables in the dataset related to students' attitudes and behaviors. It aims to dissect how various factors, including certification courses, gender, departmental affiliations, hobbies, and other lifestyle elements, influence academic outcomes, specifically students' 10th-grade marks.

4-1-1- Certification Course

The impact of enrolling in certification courses on students' academic performance is examined in this section. Certification courses are typically designed to provide additional skills and knowledge beyond the standard curriculum, which can enhance students' overall learning experience and academic outcomes. Figure 1 presents a box plot that compares the 10th-grade marks of students who have and have not taken a certification course. The box plot is an effective visualization tool for understanding the distribution and central tendency of students' performance across these two groups. From the box plot, it is evident that students who have enrolled in certification courses generally achieve higher 10th-grade marks compared to those who have not. The median mark for students taking certification courses is slightly higher, indicating that these students tend to perform better overall. Additionally, the spread of marks (interquartile range) is narrower for the certification group, suggesting more consistent performance among these students.

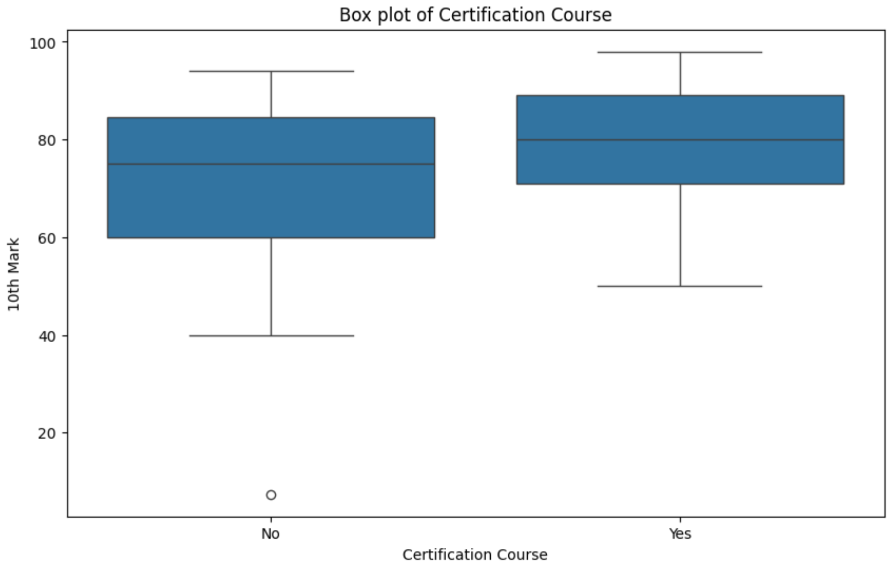

Figure 1. Box plot showing the distribution of 10th grade marks based on enrollment in a certification course

4-1-2- Gender

The influence of gender on students' academic performance is analyzed in this section. Understanding gender-based performance differences can provide insights into potential biases and help in formulating targeted educational strategies to promote equality. Figure 2 displays a box plot comparing the 10th-grade marks of male and female students. Box plots are useful for visualizing the central tendency, spread, and potential outliers in the data, offering a clear view of performance distribution across different groups. The box plot reveals that the 10th-grade marks of male and female students are relatively similar. Both groups have comparable medians, interquartile ranges, and overall distribution of marks. This indicates that, on average, gender does not significantly influence academic performance at the 10th-grade level in this dataset.

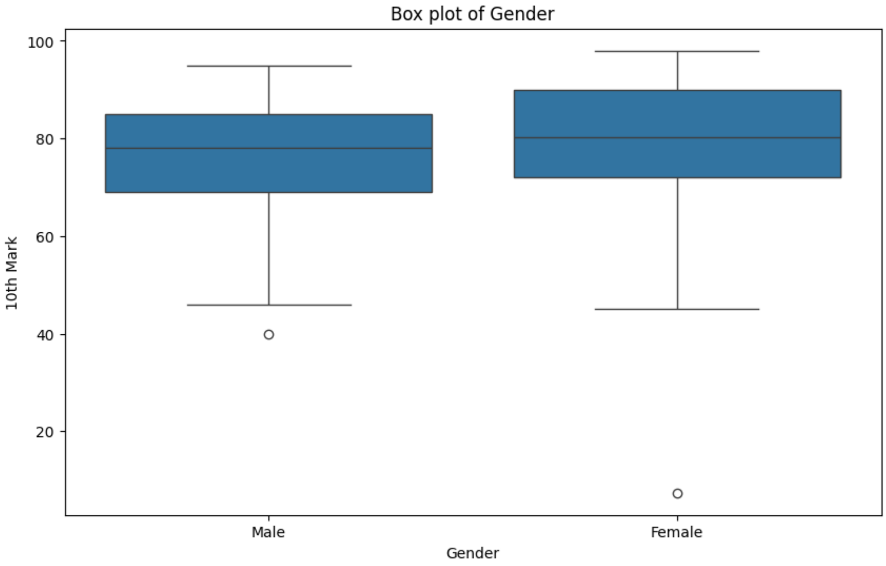

Figure 2. Box plot of 10th grade marks by gender

4-1-3- Department

This section analyzes the academic performance of students across different departments, highlighting variations in 10th-grade marks based on their chosen field of study. Figure 3 presents a box plot comparing the 10th-grade marks of students across four departments: BCA, Commerce, B.com Accounting and Finance, and B.com ISM. This visualization helps in understanding how departmental affiliation may influence academic performance. The box plot indicates noticeable differences in the median marks and spread of scores among the departments:

- ***BCA Department*:** Students in the BCA department tend to have higher median marks compared to other departments. The interquartile range is moderate, suggesting a consistent performance among students, with a few outliers indicating lower performance.
- ***Commerce Department*:** Similar to the BCA department, students in Commerce also show a high median mark with a slightly wider interquartile range. This indicates that while the majority perform well, there is greater variability in marks.
- ***B.com Accounting and Finance*:** Students in this department have a lower median mark compared to BCA and Commerce. The interquartile range is narrower, indicating less variability in scores, but overall performance is lower.
- ***B.com ISM*:** This department has the lowest median marks among the four, with a broader interquartile range and a few significant outliers, suggesting varied performance levels.

These observations suggest that departmental differences may influence academic outcomes, potentially due to varying curriculum rigor, resource availability, and teaching methodologies.

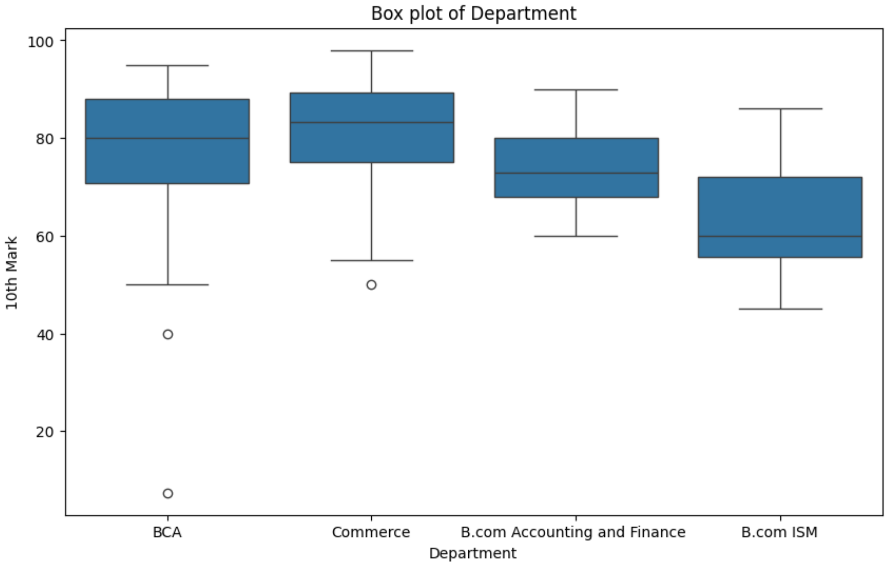

Figure 3. Box plot of 10th grade marks across various departments

4-1-4- Hobbies

This section examines the impact of students' hobbies on their academic performance, specifically their 10th-grade marks. By analyzing these relationships, we aim to understand whether certain hobbies correlate with higher academic achievement. Figure 4 presents a box plot illustrating the distribution of 10th-grade marks based on different hobbies: Video Games, Cinema, Reading books, and Sports. This visualization helps to identify any significant differences in academic performance among students with different leisure activities.

The box plot reveals the following insights:

- ***Video Games***: Students who engage in playing video games exhibit a wide range of marks with a relatively lower median compared to other hobbies. There are a few outliers with significantly lower marks, indicating that some students who spend time on video games might struggle academically.
- ***Cinema***: The marks for students who prefer watching movies are relatively consistent, with a higher median than those who play video games. However, the overall distribution is similar to that of the video games group, suggesting a moderate academic performance.
- ***Reading Books***: Students who enjoy reading books tend to have higher median marks. The interquartile range is narrow, indicating consistent performance and fewer outliers. This suggests that reading as a hobby might be positively associated with better academic outcomes.
- ***Sports***: Similar to the reading books group, students engaged in sports also show a high median mark, with a slightly broader interquartile range. This indicates that while many students in this group perform well, there is some variability in their marks.

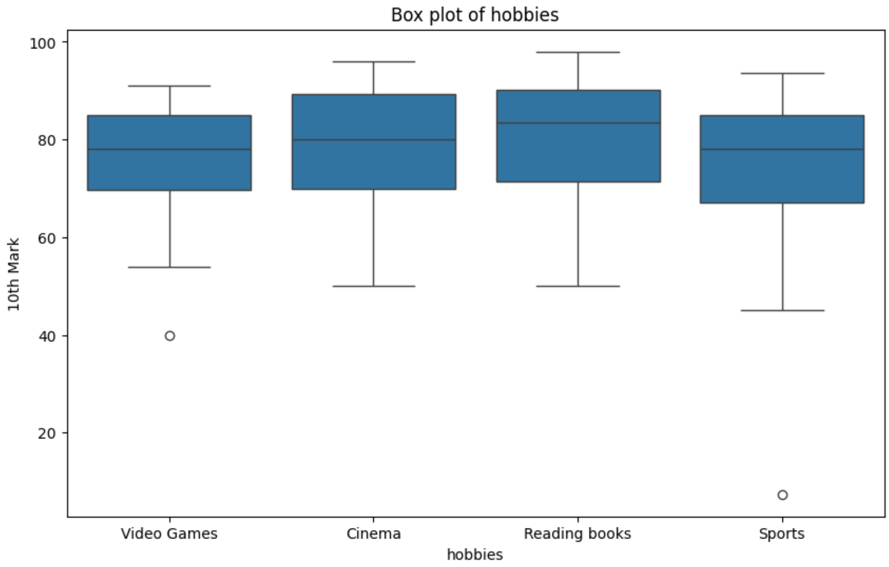

Figure 4. Box plot of 10th grade marks segmented by student hobbies

4-1-5- Daily Studying Time

This section investigates the impact of daily studying time on students' 10th-grade marks. Understanding how study habits influence academic performance can provide valuable insights for educators and students. Figure 5 presents a box plot showing the distribution of 10th-grade marks based on different ranges of daily studying time: 0 - 30 minutes, 30 - 60 minutes, 1 - 2 hours, 2 - 3 hours, 3 - 4 hours, and more than 4 hours.

The box plot reveals several key observations:

- ***0 - 30 minutes:*** Students who study for less than 30 minutes per day tend to have a lower median mark. The interquartile range is broader, indicating variability in performance within this group.
- ***30 - 60 minutes:*** This group shows a slight improvement in median marks compared to the 0 - 30 minutes group, with a narrower interquartile range, suggesting more consistent performance.
- ***1 - 2 hours:*** Students who study for 1 to 2 hours per day have a higher median mark than the previous groups. However, there are a few outliers with significantly lower marks, indicating that not all students benefit equally from this study duration.
- ***2 - 3 hours:*** The median mark for this group is slightly higher, but the interquartile range and the presence of outliers indicate that performance can vary widely among these students.

- ***3 - 4 hours:*** Students who study for 3 to 4 hours per day show a higher median mark compared to the other groups, with a relatively consistent performance, as indicated by a narrower interquartile range.
- ***More than 4 hours:*** Surprisingly, the median mark for students who study more than 4 hours per day is lower than those who study for 3 to 4 hours. This group also has a broader interquartile range, suggesting that excessive studying might not always correlate with higher academic performance.

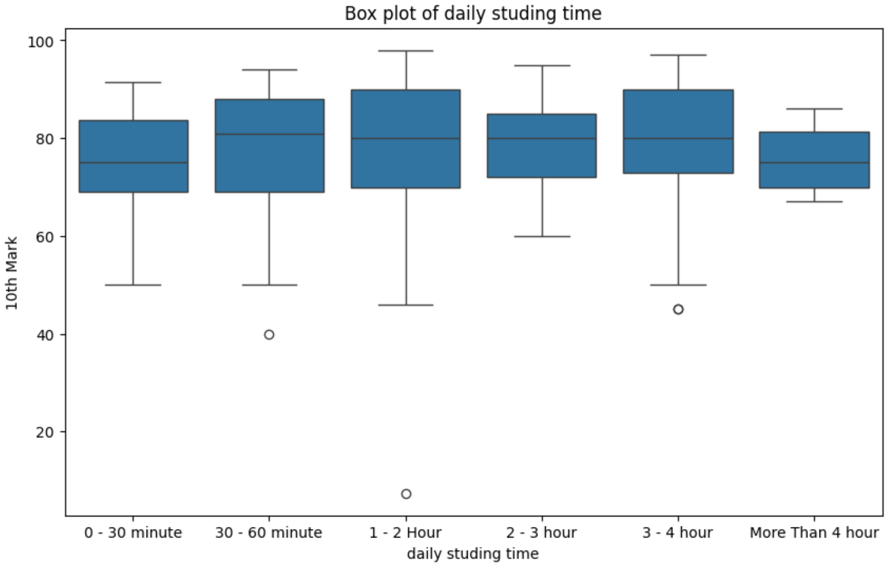

Figure 5. Box plot of 10th grade marks according to daily studying time

4-1-6- Preferred Study Time

This section examines the relationship between students' preferred study time and their 10th-grade marks. By understanding whether the time of day when students prefer to study affects their academic performance, we can gain insights into optimal study habits. Figure 6 presents a box plot showing the distribution of 10th-grade marks based on students' preferred study time: Morning, Anytime, and Night.

The box plot reveals the following observations:

- ***Morning*****:** Students who prefer to study in the morning tend to have a relatively high median mark. The interquartile range is moderate, indicating consistency in performance within this group.
- ***Anytime*****:** This group shows a similar median mark to the Morning group, with a slightly narrower interquartile range, suggesting consistent performance regardless of specific study times.
- ***Night*****:** Students who prefer to study at night have a lower median mark compared to the Morning and Anytime groups. The interquartile range is broader, indicating greater variability in performance, with some outliers having significantly lower marks.

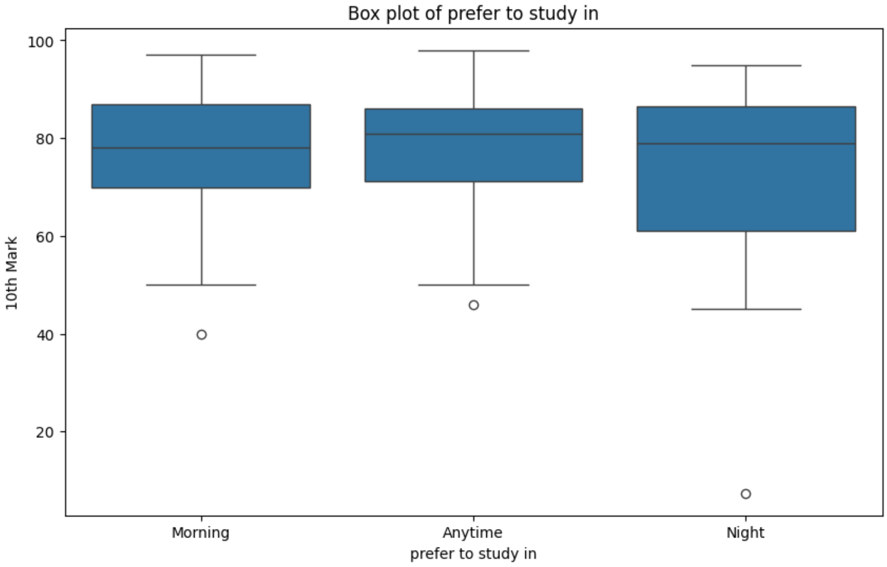

Figure 6. Box plot of 10th grade marks by preferred study time

4-1-7- Degree Likability

This section explores the impact of students' likability of their degree on their academic performance, specifically their 10th-grade marks. Understanding this relationship can provide insights into how course satisfaction influences educational outcomes. Figure 7 presents a box plot that displays the distribution of 10th-grade marks based on whether students like their degree.

The box plot reveals the following observations:

- ***No*****:** Students who do not like their degree have a relatively high median mark with a narrow interquartile range, indicating consistent performance. However, there are notable outliers with significantly lower marks.
- ***Yes*****:** Students who like their degree tend to perform better overall, with a higher median mark compared to those who do not like their degree. The interquartile range is wider, suggesting greater variability in performance among these students.

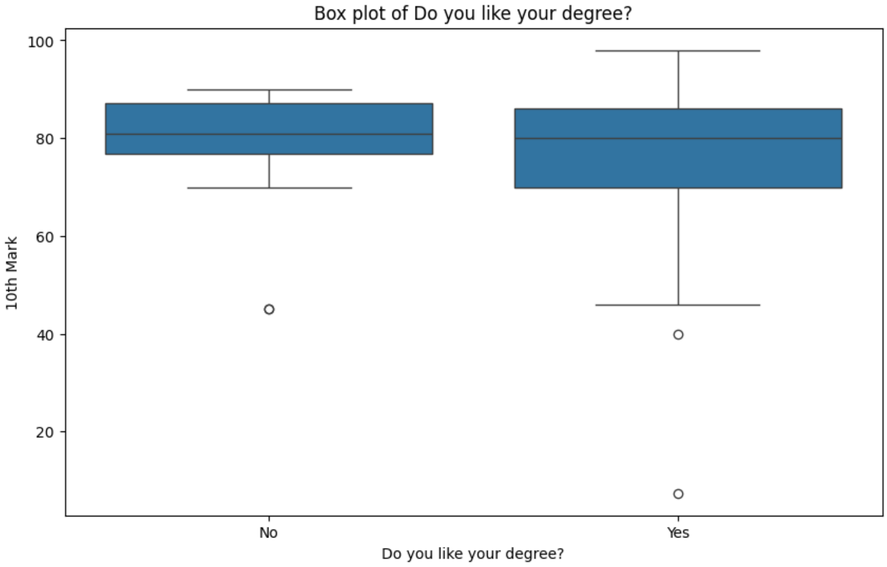

Figure 7. Box plot of 10th grade marks based on students' likability of their degree

4-1-8- Career Pursuit Willingness

This section investigates the relationship between students' willingness to pursue a career based on their degree and their academic performance, specifically their 10th-grade marks. Understanding this relationship can provide insights into how career aspirations influence educational outcomes. Figure 8 presents a box plot that shows the distribution of 10th-grade marks based on students' willingness to pursue a career related to their degree.

The box plot reveals the following observations:

- ***50% Willingness*****:** Students who are 50% willing to pursue a career based on their degree have a moderate median mark, with a few outliers at the lower end.
- ***75% Willingness*****:** Students with 75% willingness have a similar median mark as those with 50%, but there are fewer outliers, indicating more consistent performance.
- ***25% Willingness*****:** Students with 25% willingness display a slightly higher median mark compared to those with 50% and 75%, suggesting that even lower willingness does not necessarily correlate with poor performance.
- ***100% Willingness*****:** Students who are fully willing (100%) to pursue a career based on their degree tend to have the highest median marks, showing a strong positive correlation between career pursuit willingness and academic performance.
- ***0% Willingness*****:** Students with no willingness (0%) to pursue a career based on their degree have the lowest median marks, indicating a potential lack of motivation affecting their performance.

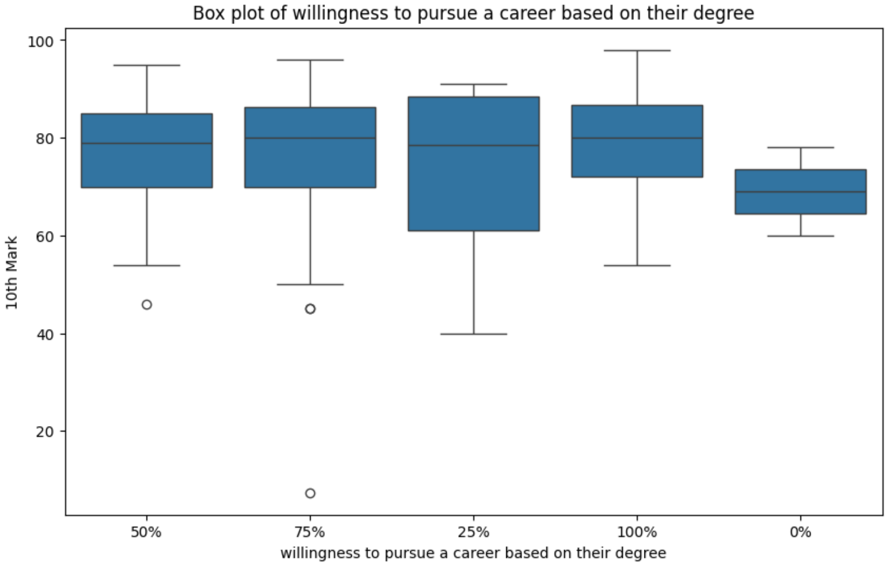

Figure 8. Box plot of 10th grade marks by willingness to pursue a career based on their degree

4-1-9- Social Media & Video Time

This section investigates how the amount of time students spend on social media and watching videos affects their 10th-grade marks. Understanding this relationship can help educators and policymakers address potential distractions and improve academic performance. Figure 9 presents a box plot showing the distribution of 10th-grade marks based on the amount of time students spend on social media and videos.

The box plot reveals the following observations:

- ***1.30 - 2 Hours*****:** Students who spend 1.30 to 2 hours on social media and videos have a moderate median mark with a few outliers at the lower end.
- ***1 - 1.30 Hours*****:** Students who spend 1 to 1.30 hours display similar median marks as those in the 1.30 - 2 hours category but with fewer outliers.
- ***More than 2 Hours*****:** Students spending more than 2 hours tend to have slightly lower median marks compared to those who spend less time.
- ***30 - 60 Minutes*****:** Students in this category have a relatively higher median mark, indicating better performance with reduced social media and video time.
- ***1 - 30 Minutes*****:** Students who spend 1 to 30 minutes on social media and videos have a similar median mark to those spending 30 - 60 minutes, but with more variability.
- ***0 Minutes*****:** Students who do not spend any time on social media and videos have the highest median marks, suggesting a positive impact of avoiding these activities on academic performance.

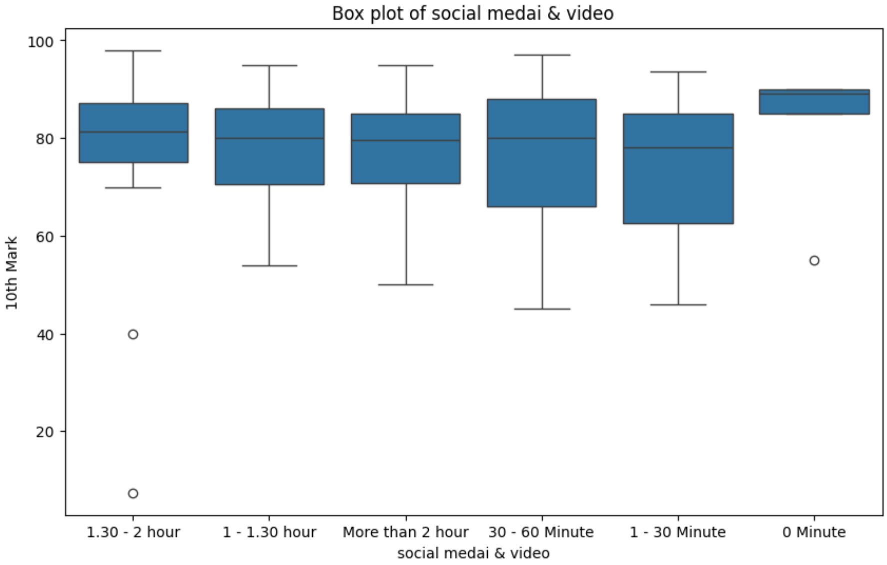

Figure 9. Box plot of 10th grade marks according to time spent on social media and videos

4-1-10- Traveling Time

This section investigates the relationship between the time students spend commuting to school and their 10th-grade marks. Commuting can be a significant factor affecting students' energy levels, focus, and overall academic performance. Figure 10 presents a box plot showing the distribution of 10th-grade marks based on the time students spend traveling to and from school.

The box plot reveals the following observations:

- ***30 - 60 Minutes*****:** Students who spend 30 to 60 minutes commuting have relatively high median marks, indicating that moderate travel time does not significantly hinder academic performance.
- ***0 - 30 Minutes*****:** Students with a commute of 0 to 30 minutes show slightly higher median marks compared to the 30 - 60 minutes group, suggesting that minimal travel time may be more conducive to better performance.
- ***1 - 1.30 Hours*****:** Students commuting for 1 to 1.30 hours show a noticeable drop in median marks, with a wider interquartile range and more outliers at the lower end.
- ***2 - 2.30 Hours*****:** Students in this category have lower median marks and more variability, indicating that extended travel times may negatively impact academic performance.
- ***1.30 - 2 Hours*****:** Students commuting for 1.30 to 2 hours also show lower median marks, consistent with the trend observed for longer commuting times.
- ***More than 3 Hours*****:** Students with the longest commuting times have the lowest median marks, highlighting the **significant** adverse effects of lengthy commutes.
- ***2.30 - 3 Hours*****:** Similar to the more than 3 hours category, students commuting for 2.30 to 3 hours also exhibit lower academic performance.

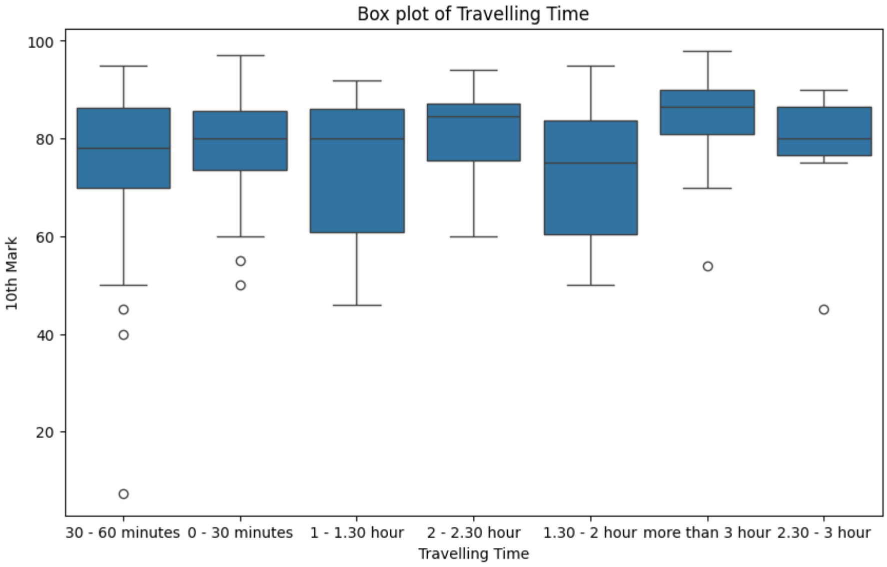

Figure 10. Box plot of 10th grade marks categorized by traveling time

4-1-11- Correlation Matrix Explanation

A correlation matrix is a table showing correlation coefficients between variables. Each cell in the table shows the correlation between two variables. The value is in the range of -1 to 1. If two variables have high correlation, they might carry similar information and cause multicollinearity. The correlation matrix is a powerful tool for preliminary feature selection because it provides insight into which variables are related to each other.

In the given correlation matrix (Figure 11):

- *Positive Correlation:* A value close to +1 implies a strong positive correlation, meaning that an increase in one variable will directly increase the other variable.
- *Negative Correlation***:** A value close to -1 implies a strong negative correlation, meaning that an increase in one variable will directly decrease the other variable.
- *No Correlation:* A value close to 0 implies no correlation, indicating that the variables do not influence each other.

Key Observations from the Matrix:

- *Height and Weight:* Correlation coefficient of 0.28 suggests a moderate positive correlation, meaning generally, as height increases, weight also increases.
- *Academic Performance:* There is a noticeable correlation between 10th Mark, 12th Mark, and College Mark:
  - ***10th Mark and 12th Mark*****:** Coefficient of 0.47 indicates a moderate positive relationship, suggesting performance in these stages is somewhat aligned.
  - ***10th Mark and College Mark*****:** Similarly, a coefficient of 0.47 shows a moderate positive trend, implying consistent academic performance from school to college.
- ***12th Mark and College Mark*****:** With a correlation of 0.42, this also suggests a continuation of academic performance into college.
- ***Stress Level and Academic Marks*****:** Stress Level shows varying correlations with academic marks, but notably, it has a coefficient of 0.25 with 12th Mark, hinting at a mild influence of stress on senior school performance.
- ***Salary Expectation*****:** Interestingly, this shows very low correlation with academic marks, suggesting that students' salary expectations are not directly related to their academic performance.

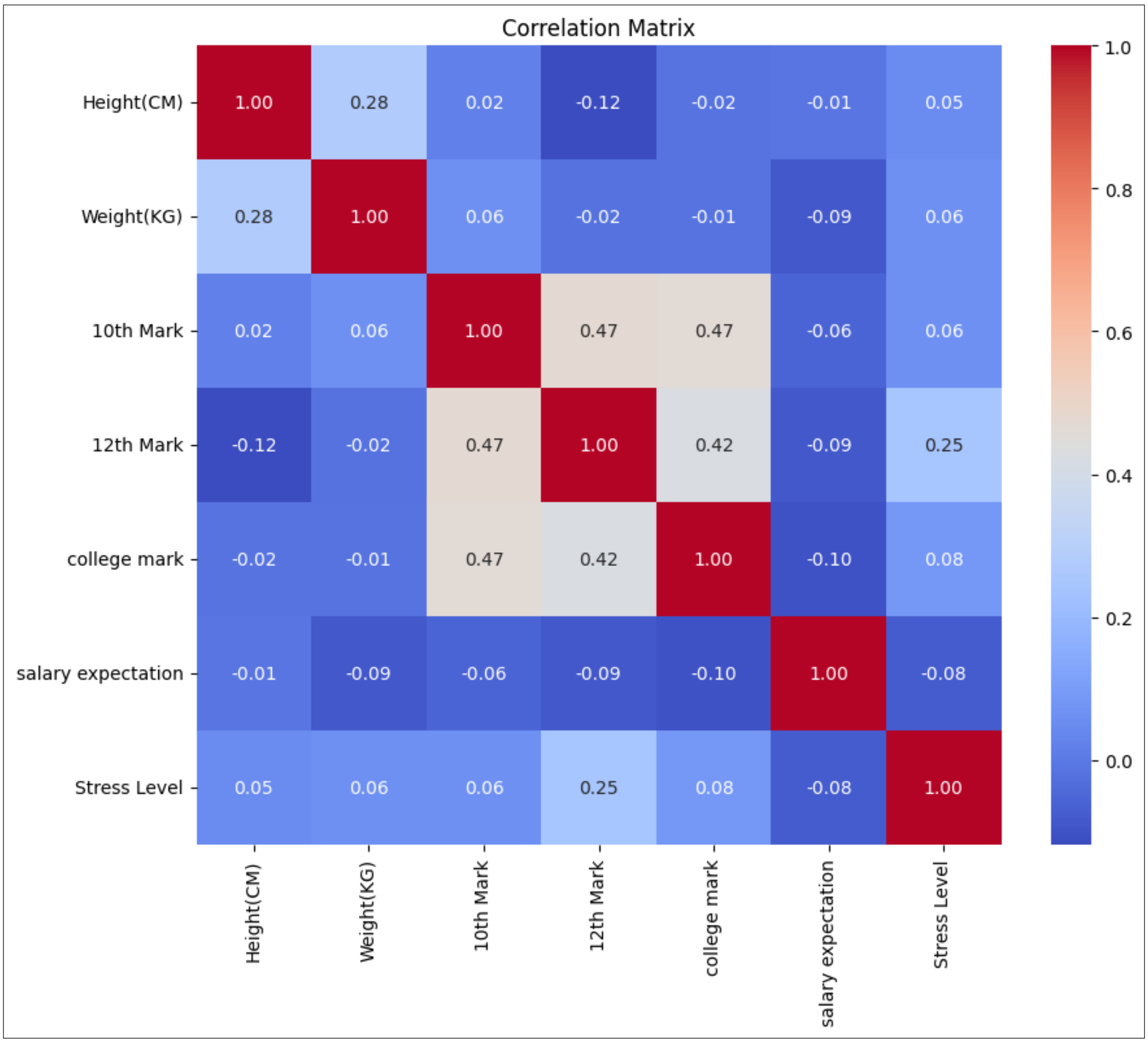

Figure 11. Correlation Matrix

4-2- Model Training and Loss Reduction

The LSTM model was trained over 10 epochs, showing a significant reduction in loss for both training and testing datasets, as illustrated in Figure 12. The training loss started at approximately 0.6 and steeply declined to stabilize around 0.3 after the fifth epoch. The test loss mirrored this progression but showed slightly higher variability, indicating good generalization with minimal overfitting.

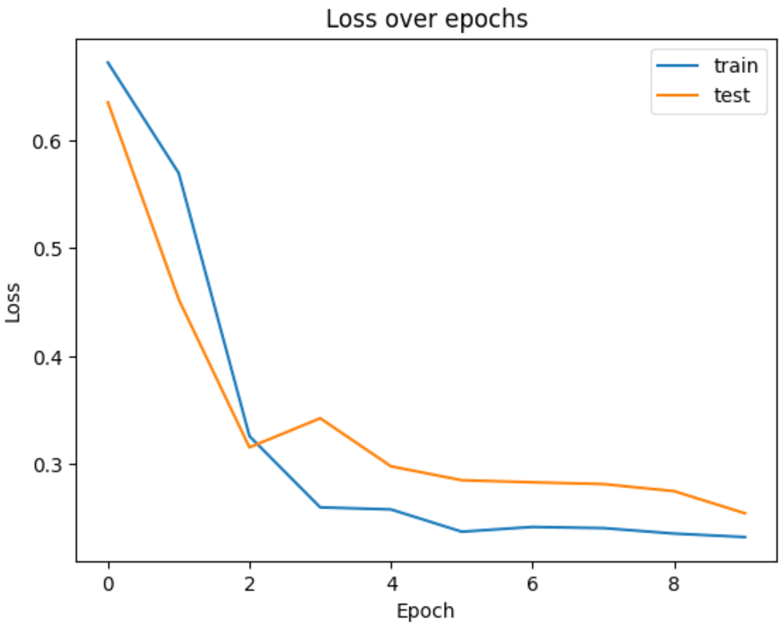

Figure 12. Training and validation loss over epochs for the LSTM model

4-3- Clustering Analysis

Post-training, Principal Component Analysis (PCA) was employed to reduce the dimensionality of the features extracted by the Long Short-Term Memory (LSTM) model, retaining two principal components for visualization. This reduction allows for a more interpretable and visually accessible analysis of the data. K-means clustering was then applied to these two principal components, identifying three distinct clusters among the student behaviors as depicted in Figure 13. Each cluster represents a unique grouping of students based on their behavioral patterns. The centroids of these clusters are marked in red, showcasing the central tendency of each cluster.

The visualization in Figure 13 demonstrates the following key points:

- ***Cluster Distribution***: The clusters are spread across the principal component space, indicating that the students exhibit diverse behavioral patterns.
- ***Centroids***: The red crosses mark the centroids of each cluster, representing the average position of the points within each cluster.
- ***Separation***: The clusters show some degree of overlap, but there are distinct regions where the clusters are more densely populated.

This clustering analysis provides valuable insights into the behavioral tendencies of students. By understanding these groupings, educators and policymakers can tailor interventions and support mechanisms to address the specific needs of different student groups. For instance, students in one cluster might benefit more from academic counseling, while those in another cluster might need additional resources for extracurricular engagement.

4-4- Performance Metrics

The LSTM model achieved an accuracy of 91% on the test set. However, as shown in the classification report (Table 1), the model was highly effective in identifying the majority class (label 1) but failed to correctly predict any instances of the minority class (label 0). This resulted in a precision of 91% and a recall of 100% for the majority class but 0% for the minority class, indicating potential issues with class imbalance.

Table 1. Classification Report

Class	Precision	Recall	F1-Score	Support
0	0.00	0.00	0.00	4
1	0.91	1.00	0.00	43
Accuracy			0.91	47
Macro Avg	0.46	0.50	0.48	47
Weighted Avg	0.84	0.91	0.87	47

The significant loss reduction during LSTM training (Figure 1) indicates that the model effectively captured the dependencies in the sequential data. The clustering results (Figure 2) demonstrate the potential to group students into behaviourally similar categories, which could inform tailored educational interventions. However, the performance metrics highlight a critical limitation in the model's ability to handle class imbalance, necessitating further investigation into techniques such as SMOTE for oversampling the minority class or adjusting class weights in the model training process.

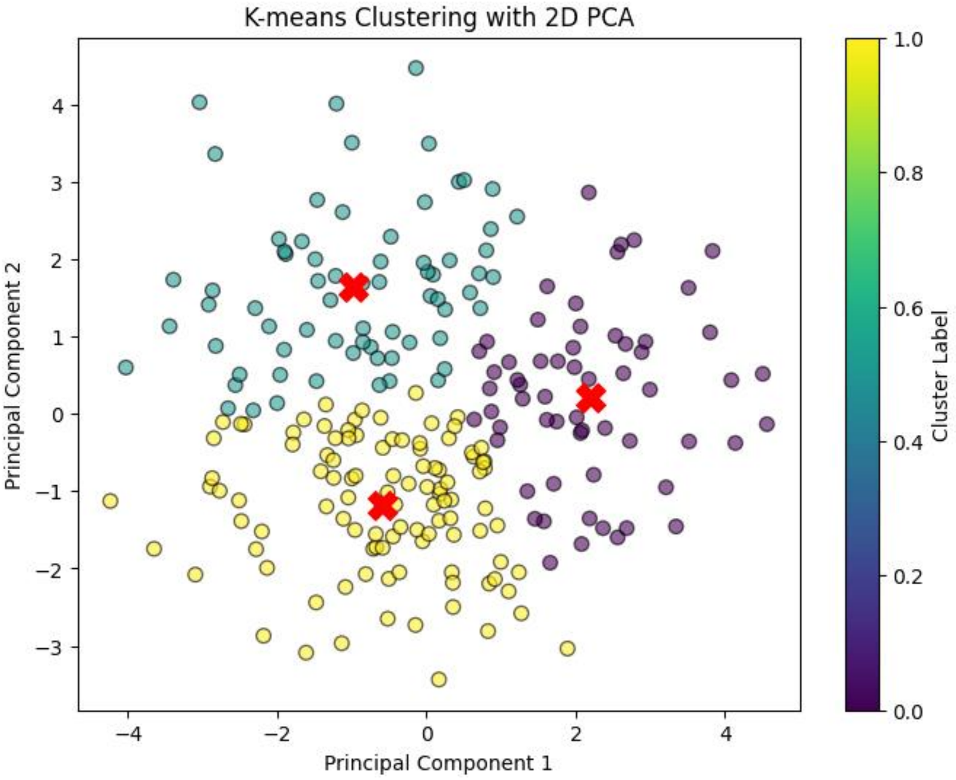

Figure 13. K-means clustering on the principal components derived from LSTM features

5- Conclusion

This study examined the "Student Attitude and Behavior" dataset using advanced data analytics and machine learning techniques to uncover key insights and predict student performance. The LSTM model demonstrated strong predictive capabilities, achieving an accuracy of 91% on the test set. However, the model exhibited a class imbalance, excelling in predicting the majority class (label 1) but failing to identify instances of the minority class (label 0). This discrepancy resulted in a precision of 91% and a recall of 100% for the majority class but 0% for the minority class. Additionally, K-means clustering and PCA revealed distinct student groupings based on various attributes, providing further insights into student behaviors and performance. The findings of this study have several important implications for educational institutions and policymakers. The high accuracy of the LSTM model indicates the potential of deep learning techniques in predicting student outcomes, which can be used to identify at-risk students early and provide targeted interventions. The insights gained from clustering analysis can help educators understand the diverse needs and characteristics of students, enabling more personalized and effective teaching strategies. However, the issue of class imbalance highlighted in this study suggests that further refinement of machine learning models is necessary to ensure equitable prediction across different student groups. Despite the promising results, this study has several limitations. The dataset used was limited in scope and size, which may affect the generalizability of the findings. The class imbalance in the dataset posed a significant challenge, impacting the model's ability to accurately predict the minority class. Additionally, the study primarily focused on 10th-grade marks as a measure of academic performance, which may not capture the full spectrum of student abilities and achievements.

The reliance on self-reported data for certain variables could also introduce bias and inaccuracies. Future research should address the limitations identified in this study by incorporating larger and more diverse datasets to enhance the generalizability of the findings. Techniques such as data augmentation and synthetic minority over-sampling (SMOTE) should be explored to mitigate class imbalance and improve model performance. Furthermore, integrating additional features such as behavioral data, attendance records, and socio-economic factors can provide a more comprehensive understanding of student performance. Longitudinal studies tracking students over multiple academic years would also offer valuable insights into the long-term effectiveness of predictive models and educational interventions. In conclusion, while this study has demonstrated the potential of advanced data analytics and machine learning techniques in predicting student performance, addressing the identified limitations and exploring future research directions are essential for developing more robust and equitable models that can better inform educational practices and policies.

6- Declarations

6-1- Author Contributions

R.A., K.A., M.K., S.A.S., R.S., and M.A.A. contributed to the design and implementation of the research, to the analysis of the results and to the writing of the manuscript. All authors have read and agreed to the published version of the manuscript.

6-2- Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6-3- Funding

This work was supported by the Deanship of Scientific Research, Vice Presidency for Graduate Studies and Scientific Research, King Faisal University, Saudi Arabia (Grant No. KFU241514).

6-4- Institutional Review Board Statement

Not applicable.

6-5- Informed Consent Statement

Not applicable.

6-6- Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this manuscript. In addition, the ethical issues, including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, and redundancies have been completely observed by the authors.

7- References

[1] Juniarni, C., Sodikin, M. A., Akhyar, A., Almujahid, A., & Asvio, N. (2024). The Importance of Personalized Learning: How to Tailor Education to the Individual Needs of Students. Education Studies and Teaching Journal (EDUTECH), 1(1), 188-196.

[2] Mulaa, E. (2020). Influence of socio-economic, psychological and physical factors on academic performance among orphaned pupils in public primary schools in Kapseret Sub-County, Uasin Gishu County, Kenya. PhD Thesis, Africa Nazarene University, Nairobi, Kenya.

[3] Johnson, E. A., Inyangetoh, J. A., Rahmon, H. A., Jimoh, T. G., Dan, E. E., & Esang, M. O. (2024). An Intelligent Analytic Framework for Predicting Students Academic Performance Using Multiple Linear Regression and Random Forest. European Journal of Computer Science and Information Technology, 12(3), 56-70.

[4] Zhang, Y., Yun, Y., An, R., Cui, J., Dai, H., & Shang, X. (2021). Educational data mining techniques for student performance prediction: method review and comparison analysis. Frontiers in psychology, 12, 698490. doi:10.1109/TASE.2021.3077537.

[5] Keller, L., Preckel, F., Eccles, J. S., & Brunner, M. (2022). Top-performing math students in 82 countries: An integrative data analysis of gender differences in achievement, achievement profiles, and achievement motivation. Journal of Educational Psychology, 114(5), 966. doi:10.1037/edu0000685.

[6] Ishaq, A., Sadiq, S., Umer, M., Ullah, S., Mirjalili, S., Rupapara, V., & Nappi, M. (2021). Improving the prediction of heart failure patients' survival using SMOTE and effective data mining techniques. IEEE access, 9, 39707-39716. doi:10.1109/ACCESS.2021.3064084.

[7] Coleman, J. S. (1968). Equality of Educational Opportunity. Equity & Excellence in Education, 6(5), 19–28. doi:10.1080/0020486680060504.

[8] Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. Review of Educational Research, 75(3), 417–453. doi:10.3102/00346543075003417.

[9] Bandura, A. (1997). Self-efficacy: The exercise of control. Freeman google schola, 2, 143-154.

[10] Deci, E. L., & Ryan, R. M. (2000). The "what" and "why" of goal pursuits: Human needs and the self-determination of behavior. Psychological Inquiry, 11(4), 227–268. doi:10.1207/S15327965PLI1104_01.

[11] Credé, M., & Kuncel, N. R. (2008). Study Habits, Skills, and Attitudes: The Third Pillar Supporting Collegiate Academic Performance. Perspectives on Psychological Science, 3(6), 425–453. doi:10.1111/j.1745-6924.2008.00089.x.

[12] Kuh, G. D., Cruce, T. M., Shoup, R., Kinzie, J., & Gonyea, R. M. (2008). Unmasking the effects of student engagement on first-year college grades and persistence. Journal of Higher Education, 79(5), 540–563. doi:10.1353/jhe.0.0019.

[13] Goldhaber, D. D., & Brewer, D. J. (2000). Does teacher certification matter? High school teacher certification status and student achievement. Educational Evaluation and Policy Analysis, 22(2), 129–145. doi:10.3102/01623737022002129.

[14] Halpern, D. F. (2000). Sex differences in cognitive abilities. Psychology press, New York, United States. doi:10.4324/9781410605290.

[15] Stoet, G., & Geary, D. C. (2013). Sex Differences in Mathematics and Reading Achievement Are Inversely Related: Within- and Across-Nation Assessment of 10 Years of PISA Data. PLoS ONE, 8(3), 57988. doi:10.1371/journal.pone.0057988.

[16] Haveman, R., & Wolfe, B. (1995). The determinants of children's attainments: A review of methods and findings. Journal of economic literature, 33(4), 1829-1878.

[17] Duckworth, A. L., & Seligman, M. E. P. (2005). Self-discipline outdoes IQ in predicting academic performance of adolescents. Psychological Science, 16(12), 939–944. doi:10.1111/j.1467-9280.2005.01641.x.

[18] Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. econometrica, 73(2), 417-458. doi:10.1111/j.1468-0262.2005.00584.x.

[19] Thapa, A., Cohen, J., Guffey, S., & Higgins-D'Alessandro, A. (2013). A Review of School Climate Research. Review of Educational Research, 83(3), 357–385. doi:10.3102/0034654313483907.

[20] Yadav, S. K., Bharadwaj, B., & Pal, S. (2012). Data mining applications: A comparative study for predicting student's performance. arXiv preprint, arXiv:1202.4815 doi:10.48550/arXiv.1202.4815.

[21] He, J., Baileyt, J., Rubinstein, B. I. P., & Zhang, R. (2015). Identifying at-risk students in massive open online courses. Proceedings of the National Conference on Artificial Intelligence, 3(1), 1749–1755. doi:10.1609/aaai.v29i1.9471.

[22] Piech, C., Bassen, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L. J., & Sohl-Dickstein, J. (2015). Deep knowledge tracing. Advances in Neural Information Processing Systems 28 (NIPS 2015), 7-12 December, 2015, Montreal, Canada.

[23] Kotsiantis, S., Pierrakeas, C., & Pintelas, P. (2004). Predicting students' performance in distance learning using machine learning techniques. Applied Artificial Intelligence, 18(5), 411–426. doi:10.1080/08839510490442058.

[24] Kaggle. (2024). Student Attitude and Behavior. Kaggle, Mountain View, United States. Available online: https://www.kaggle.com/datasets/susanta21/student-attitude-and-behavior (accessed on July 2024).

[25] Saar-Tsechansky, M., & Provost, F. (2007). Handling missing values when applying classification models. Journal of Machine Learning Research 8 (2007) 1625-1657.

[26] Patro, S. G. K., & Sahu, K. K. (2015). Normalization: A Preprocessing Stage. IARJSET, 20–22. doi:10.17148/iarjset.2015.230.

[27] Sinsomboonthong, S. (2022). Performance Comparison of New Adjusted Min-Max with Decimal Scaling and Statistical Column Normalization Methods for Artificial Neural Network Classification. International Journal of Mathematics and Mathematical Sciences, 2022(1), 3584406. doi:10.1155/2022/3584406.

[28] Seger, C. (2018). An investigation of categorical variable encoding techniques in machine learning: binary versus one-hot and feature hashing. Thesis, KTH Royal Institute of Technology, Stockholm, Sweden.

[29] Sanderson, P., & Fisher, C. (1994). Exploratory Sequential Data Analysis: Foundations. Human-Computer Interaction, 9(3), 251–317. doi:10.1207/s15327051hci0903&4_2.

[30] Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. Neural Computation, 9(8), 1735–1780. doi:10.1162/neco.1997.9.8.1735.

[31] Jolliffe, I. (2005). Principal Component Analysis. Encyclopedia of Statistics in Behavioral Science, John Wiley & Sons, Hoboken, United States. doi:10.1002/0470013192.bsa501.