Unveiling Academic Success:

Harnessing Graph Machine Learning for Student Performance Prediction

ABSTRACT

In this chapter, we delve into the utilization of graph machine learning techniques to forecast student academic performance. By harnessing graph-based representations of educational data, our study endeavors to unearth underlying patterns and connections that impact student success. Through a fusion of feature engineering, graph analytics, and predictive modeling, we aim to investigate the efficacy of graph-based methodologies in improving the precision and interpretability of student performance prediction systems. This paper investigates the effectiveness of logistic regression, K-nearest neighbors (KNN), and a custom Graph Neural Network (GNN) model for predicting student performance in exams. Our analysis reveals that the custom GNN model outperforms both logistic regression and KNN, achieving higher accuracy and efficiency in student performance prediction. The custom GNN model leverages the graph-based representation of educational data, which enhances its ability to capture complex relationships and dependencies among students.

Keywords: Bipartite graphs, Machine learning, Logistic Regression, K-nearest neighbor, Graph Neural Network.

1 Introduction

Predicting student academic performance is a crucial task in educational institutions worldwide, as it allows educators to identify struggling students early and provide them with the necessary support to succeed (Balcioglu et al.,2023). However, traditional predictive models based on tabular data often fall short in capturing the intricate interplay of factors that influence student success (Subramanian et al.,2023). These models typically rely on individual attributes such as demographics, prior grades, and standardized test scores, overlooking the rich relational context inherent in educational data (Brooks et al.,2023).

In recent years, the advent of graph-based machine learning techniques has opened up new avenues for analyzing complex relational data (Dervenis et al.,2022). By representing educational data as a graph, wherein nodes correspond to entities such as students, courses, and assignments, and edges encode relationships such as enrollment, collaboration, and academic performance, we can capture the multidimensional interactions and dependencies that shape student outcomes (Gajwani et al.,2021, Kaur et al.,2023). This graph- based representation offers a holistic view of the educational ecosystem, enabling us to uncover hidden patterns and insights that traditional models may overlook.

One of the key challenges in student performance prediction lies in the early identification of at-risk students who may benefit from targeted interventions. Early intervention strategies have been shown to significantly improve student retention and success rates (Ojajuni et al.,2021, Soyoye et al.,2023). However, conventional models often lack the granularity and interpretability necessary to pinpoint students who are most in need of support. By leveraging graph-based representations, which inherently encode structural information and relational context, we can enhance the accuracy and interpretability of predictive models, thereby facilitating timely interventions and personalized support strategies (Patil et al.,2023, Sekeroglu et al.,2019).

In this research article, we aim to explore the potential of graph machine learning techniques for predicting student academic performance. We will begin by discussing the limitations of traditional predictive models and the motivations for adopting a graph-based approach. We will then outline our objectives, which include:

(i) Constructing a comprehensive graph representation of educational data, incorporating various entities and relationships relevant to student performance prediction.

(ii) Investigating feature engineering techniques tailored to graph-based representations, aiming to extract informative features that capture both individual characteristics and relational dynamics.

(iii) Employing advanced graph analytics and predictive modeling techniques to uncover hidden patterns and relationships that influence student outcomes.

(iv) Evaluating the performance of graph-based predictive models in comparison to traditional approaches, with a focus on accuracy, interpretability, and scalability.

(v) Discussing the implications of our findings for educational practice and the potential for enhancing early intervention and support strategies.

Through this research, we seek to advance our understanding of how graph-based machine learning can be leveraged to address the challenges of predicting student academic performance. By harnessing the power of relational data and advanced analytics, we aim to contribute to the development of more effective and interpretable predictive models, ultimately improving student outcomes and promoting success in educational settings.

Student performance prediction has been a subject of extensive research in the field of education, driven by the imperative to improve learning outcomes and support student success (Benkhalfallah et al.,2023, Das et al.,2019). A plethora of studies have investigated the multifaceted factors that contribute to academic achievement, ranging from socio-economic background and demographics to individual learning behaviors and engagement. Understanding these factors is crucial for developing accurate predictive models that can identify at-risk students and guide intervention efforts. Several seminal studies have highlighted the significance of socio-economic status (SES) as a determinant of student success. Research has consistently shown that students from disadvantaged backgrounds face greater academic challenges due to limited access to resources, lack of parental support, and exposure to adverse environmental factors (Soyoye et al.,2023). Moreover, disparities in access to quality education exacerbate existing inequalities, perpetuating cycles of underachievement and disadvantage. Therefore, incorporating SES-related variables into predictive models is essential for addressing equity issues and ensuring inclusive educational practices.

In addition to socio-economic factors, prior academic performance emerges as a robust predictor of future success (Brooks et al.,2023, Hennebelle et al.,2024). Numerous studies have demonstrated the predictive power of historical grades, standardized test scores, and GPA in forecasting student outcomes. However, it is essential to recognize that academic performance is not solely determined by past achievements but also influenced by various contextual factors, including the learning environment, teacher quality, and peer interactions. Thus, predictive models must account for both individual characteristics and relational dynamics to accurately assess student trajectories (Kaur et al.,2022). Further- more, recent advancements in machine learning have spurred interest in graph-based techniques for representing and analyzing educational data (Kumar et al.,2023). Graphs offer a natural framework for capturing the complex relational structures inherent in educational environments, wherein nodes represent entities such as students, courses, and concepts, and edges denote relationships such as enrollment, collaboration, and prerequisite dependencies. By modeling educational data as a graph, researchers can leverage powerful graph-based algorithms to uncover hidden patterns and insights that traditional methods may overlook (Radhya et al.,2022). Graph convolutional networks (GCNs) and graph attention networks (GATs) have emerged as prominent approaches for predictive modeling tasks on graph-structured data. GCNs extend the concept of convolutional neural networks to irregular graph domains, enabling the propagation of information across nodes while preserving local neighborhood structure (Sahlaoui et al.,2021). GATs, on the other hand, leverage attention mechanisms to selectively aggregate information from neighboring nodes, allowing for adaptive feature weighting and enhanced expressiveness. These graph-based methods have demonstrated promising results in various applications, including node classification, link prediction, and recommendation systems (Sharma et al.,2023). Moreover, researchers have explored diverse graph representations of educational data, including student social net- works, course prerequisite graphs, and knowledge graphs. Social network analysis techniques have been employed to analyze patterns of collaboration and information flow among students, highlighting the role of peer interactions in shaping learning outcomes (Zeineddine et al.,2021). Course prerequisite graphs provide insights into the hierarchical structure of academic curricula, facilitating course recommendation and curriculum planning. Knowledge graphs, constructed from semantic relationships between concepts and learning objectives, support personalized learning pathways and adaptive assessment strategies (Subramanian et al.,2023).

In the realm of student performance prediction and educational data analytics, the k-nearest neighbors (KNN) algorithm stands as a fundamental yet effective method for classification and regression tasks. The essence of KNN lies in its simplicity and intuitive approach to prediction based on similarity metrics. KNN operates under the principle that similar instances in feature space tend to exhibit similar outcomes. Graph Neural Networks (GNNs) are a class of neural networks specifically designed to work with graph-structured data. Unlike traditional neural networks that operate on fixed-size, grid-like structures (e.g., images), GNNs can effectively process data represented as graphs, where nodes represent entities and edges represent relationships between these entities. This makes GNNs particularly powerful for applications involving complex relational data. In the context of student performance prediction, this means that students with similar characteristics, academic histories, or learning behaviors are likely to achieve comparable results. In our study, we employed a custom Graph Neural Network (GNN) model to predict student performance and compared its effectiveness against traditional models such as logistic regression and K-nearest neighbors (KNN). The custom GNN model leverages graph-based representations of educational data to capture complex relationships and dependencies among students. This approach allows the GNN to provide more accurate and interpretable predictions compared to logistic regression and KNN. The custom GNN model exhibited superior accuracy, robustness, and efficiency in student performance prediction. By incorporating local context awareness and graph structures, the custom GNN demonstrated improved performance, particularly in handling dynamic and evolving datasets. This capability to model intricate relational dynamics provided significant advantages over traditional models that rely on individual attributes and overlook the rich relational context inherent in educational data.

2 Dataset Description

The dataset used for this study is designed to predict whether a student will pass or fail an exam based on two main features: the number of study hours for the upcoming exam and the student’s score in the previous exam. Additionally, the dataset includes the target variable, where a value of 1 represents a pass and 0 represents a fail in the current exam.

2.1 Features, Target Variable and Dataset Size

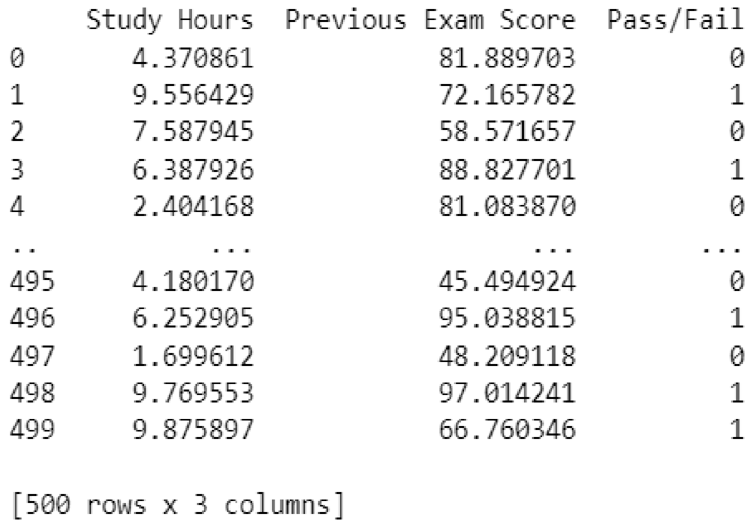
(i) **Study Hours (Numeric)**: This feature represents the number of hours a student spent studying for the upcoming exam. It serves as an indicator of the student's level of preparation and dedication towards academic success.

(ii) **Previous Exam Score (Numeric)**: This feature indicates the student's score in the previous exam. It provides insight into the student's academic performance and proficiency in the subject matter.

Target Variable: The target variable in the dataset is binary and represents whether the student passed or failed the current exam. A value of 1 denotes a pass, indicating that the student met the required criteria for success in the exam. Conversely, a value of 0 indicates a fail, signifying that the student did not achieve the necessary level of proficiency in the exam.

Dataset Size: The dataset consists of data for 500 students, ensuring a diverse range of study patterns and previous exam performances. This ample sample size allows for robust analysis and modeling, capturing variations in study habits, academic backgrounds, and exam outcomes among students.

Table 1: First few rows of the dataset



```
     Study Hours  Previous Exam Score  Pass/Fail
0       4.370861            81.889703          0
1       9.556429            72.165782          1
2       7.587945            58.571657          0
3       6.387926            88.827701          1
4       2.404168            81.083870          0
..           ...                  ...        ...
495     4.180170            45.494924          0
496     6.252905            95.038815          1
497     1.699612            48.209118          0
498     9.769553            97.014241          1
499     9.875897            66.760346          1

[500 rows x 3 columns]
```

By leveraging this dataset, we aim to explore the relationship between study hours, previous exam scores, and the likelihood of exam success. Through predictive modeling techniques, we seek to develop a model that can accurately classify students into pass or fail categories based on their study behavior and academic history. This analysis has the potential to provide valuable insights for educators and policymakers to support student success and enhance academic outcomes.

3 Predicting Student Performance in Exams using Logistic Regression

Logistic regression is a popular statistical method used for predicting binary outcomes, making it suitable for predicting student performance in exams, where the outcome is typically pass or fail. The logistic regression model estimates the probability that a given student will pass the exam based on one or more predictor variables, such as study hours, previous exam scores, demographic factors, etc.

3.1 Data Preparation

Before building the logistic regression model, the dataset needs to be prepared. This involves:

- **Data Cleaning**: Handling missing values, outliers, and inconsistencies in the dataset.
- **Feature Selection**: Identifying relevant predictor variables that are likely to influence student performance.
- **Data Splitting**: Splitting the dataset into training and testing sets to evaluate the model's performance.

3.2 Model Building

Once the dataset is prepared, the logistic regression model can be built using the training data. The logistic regression model estimates the probability p of a student passing the exam as a function of the predictor variables. Mathematically, the logistic regression model can be represented as:

$$p = \frac{1}{1 + e^{-z}},$$

Where $z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$ is the linear combination of predictor variables and their respective coefficients $\beta_0, \beta_1, \ldots, \beta_n$.

In this equation:

- p represents the probability of a student passing the exam.
- e is the base of the natural logarithm.
- z is the logic function, which is a linear combination of predictor variables.
- $x_1, x_2, \ldots, x_n$ represent the predictor variables.
- $\beta_0, \beta_1, \ldots, \beta_n$ are the coefficients of the predictor variables.

The logistic regression model computes the log-odds of the probability of the event of interest (in this case, passing the exam) and then applies the sigmoid function to map the log-odds to a probability value between 0 and 1.

3.3 Model Evaluation

After training the logistic regression model, it needs to be evaluated using the testing data to assess its performance and generalization ability. Common evaluation metrics for logistic regression models include:

- **Accuracy**: The percentage of correctly classified instances.
- **Precision**: The proportion of true positive predictions among all positive predictions.
- **Recall**: The proportion of true positive predictions among all actual positive instances.
- **F1 Score**: The harmonic mean of precision and recall, providing a balance between the two metrics.
- **ROC Curve and AUC**: Receiver Operating Characteristic (ROC) curve and Area under the Curve (AUC) measure the model's ability to discriminate between positive and negative classes across different thresholds.

Algorithm 1 Logistic Regression Algorithm for Predicting Student Performance

```
**Input**: Training features $X_{\text{train}}$, training labels $y_{\text{train}}$, test features $X_{\text{test}}$
**Output**: Predicted labels for test set $y_{\text{pred}}$
**Initialization**: Initialize logistic regression classifier
**Feature Scaling**: Scale features using StandardScaler
**Training**: Fit logistic regression model to training data
**Prediction**: Predict labels for test set using trained model
**function** LOGISTICREGRESSION($X_{\text{train}}$, $y_{\text{train}}$, $X_{\text{test}}$)
**Feature Scaling**:
Initialize StandardScaler *sc*
$X_{\text{train}} \leftarrow sc.fit\_transform(X_{\text{train}})$
$X_{\text{test}} \leftarrow sc.transform(X_{\text{test}})$
**Training**:
Initialize logistic regression classifier *classifier* with random state 42
$classifier.fit(X_{\text{train}}, y_{\text{train}})$
**Prediction**:
$y_{\text{pred}} \leftarrow classifier.predict(X_{\text{test}})$
**return** $y_{\text{pred}}$
**end function**=0
```

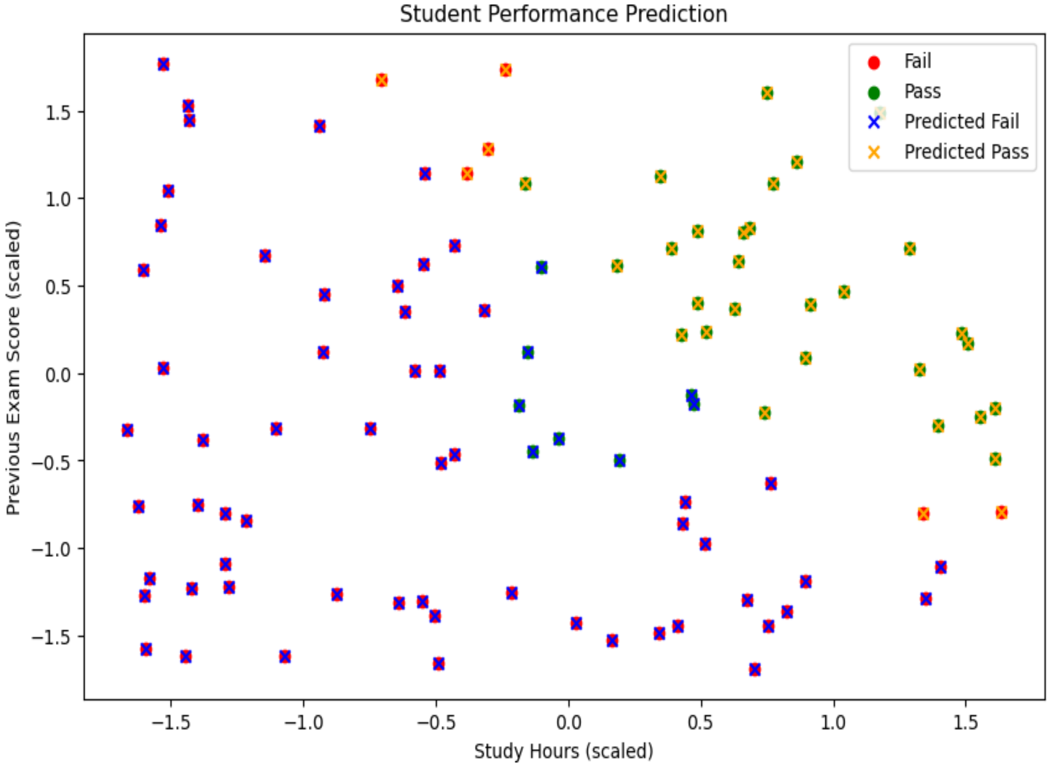


Figure 1. Students' Performance Prediction using Logistic Regression

3.4 Analysis of Logistic Regression Model Output

Logistic regression is a powerful tool for predicting student performance in exams based on relevant predictor variables. By estimating the probability of passing the exam, educators and stakeholders can identify at-risk students and implement targeted intervention strategies to improve academic outcomes. The logistic regression model achieved an accuracy of 0.86 on the testing set, indicating its effectiveness in predicting student performance. The decision boundary of the logistic regression model was visualized to provide a graphical representation of the classification process. The decision boundary separates the feature space into regions corresponding to pass and fail outcomes.

The results demonstrate the potential of logistic regression as a tool for predicting student performance in exams. The model’s accuracy suggests that study hours and previous exam scores are significant predictors of exam outcomes. The decision boundary visualization provides insights into the regions of feature space where students are more likely to pass or fail the exam. Thus, the logistic regression model shows promise in accurately predicting student performance in exams based on study hours and previous exam scores. Further research could explore additional features and more advanced machine learning algorithms to improve prediction accuracy.

4 K-nearest neighbors (KNN) Algorithm for Student Performance Prediction

The K-nearest neighbors (KNN) algorithm is a non-parametric method used for classification and regression tasks. In the context of student performance prediction, KNN leverages the similarities between students to make predictions about their academic outcomes. Let’s delve deeper into how KNN operates and its relevance in educational data analytics.

Training Data
K-nearest Neighbors
Prediction

Figure 2. K-nearest neighbors (KNN) Algorithm for Student Performance Prediction

4.1 Working Principle

Training Phase

In the training phase, KNN stores the entire dataset consisting of student records and their corresponding performance indicators. Each student record is represented as a feature vector, where each feature may include demographic information, past academic performance, socio-economic status, etc. KNN doesn’t build an explicit model during training; instead, it retains the entire dataset in memory for subsequent prediction tasks.

Prediction Phase

When presented with a new, unseen student record, KNN identifies the k nearest neighbors to this student within the feature space. The "nearest" neighbors are determined by calculating distances between the new student’s feature vector and the feature vectors of all students in the training dataset. Various distance metrics can be used, such as Euclidean distance, Manhattan distance, or cosine similarity. The choice of distance metric depends on the nature of the features and the dataset.

4.2 Classification or Regression

Once the k nearest neighbors are identified, KNN employs different strategies for classification and regression tasks:

- **Classification**: For classification tasks, such as predicting whether a student is at risk of failing or not, KNN selects the majority class among the k nearest neighbors. The new student is assigned the same class label as the majority. In a binary classification scenario (e.g., pass/fail), the class with the most representatives among the neighbors is assigned to the new student.

- **Regression**: For regression tasks, such as predicting final exam scores or GPA, KNN computes the average (or weighted average) of the target variable values among the k nearest neighbors. The predicted value for the new student is set to this average. This approach assumes that the target variable (e.g., exam score) is continuous and can be averaged across similar students to estimate the outcome for the new student.

4.3 Hyper parameter Tuning

The performance of the KNN algorithm is heavily influenced by the choice of the hyper parameter k, which represents the number of neighbors considered when making predictions. The optimal value of k depends on the dataset characteristics, such as the level of noise and the complexity of the underlying patterns. Cross-validation techniques, such as k-fold cross-validation, can be used to find the best value of k that generalizes well to unseen data.

4.4 Applicability in Student Performance Prediction

KNN is well-suited for student performance prediction due to several reasons:

- **Adaptability**: KNN can capture local patterns in the data, allowing it to adapt to different student populations and academic contexts. It doesn't assume any specific data distribution, making it versatile and applicable to various educational datasets.

- **Interpretability**: KNN provides transparent predictions by directly referencing similar students in the dataset. Educators and stakeholders can understand the rationale behind pre- dictions by examining the characteristics of the nearest neighbors.

- **Personalization**: By considering the features and academic histories of similar students, KNN offers personalized predictions tailored to individual students. This personalized approach can be valuable for designing targeted intervention strategies to support student success.

Thus, the K-nearest neighbors' algorithm offers a straightforward yet effective approach to student performance prediction in educational data analytics. Its adaptability, interpretability, and personalization capabilities make it a valuable tool for identifying at-risk students and guiding intervention efforts in educational settings. However, it's essential to carefully tune the hyper parameter k and consider the choice of distance metric to ensure optimal performance on a given dataset.

4.5 Analysis of KNN Model Output

Accuracy Calculation

The accuracy of the K-nearest neighbors (KNN) model is computed using the accuracy score function from scikit-learn. This metric quantifies the percentage of correctly predicted outcomes compared to the total number of predictions made. The accuracy of the K-nearest neighbors (KNN) algorithm for student performance prediction is found to be 0.95.

Algorithm 2 K-nearest neighbors (KNN) algorithm for student performance prediction

```
Input: Training data $X_{train}$, $y_{train}$, test data $X_{test}$, number of neighbors $k$
Output: Predicted classes for test data $y_{pred}$
function KNN_PREDICT($X_{train}$, $y_{train}$, $X_{test}$, $k$)
    $y_{pred} \leftarrow []$
    for test_instance in $X_{test}$ do
        distances ← argsort(norm($X_{train}$ − test_instance, axis = 1))[: $k$]
        nearest_labels ← $y_{train}$[distances]
        prediction ← argmax(bincount(nearest_labels))
        $y_{pred}$.append(prediction)
    end for
    return $y_{pred}$
end function=0
```

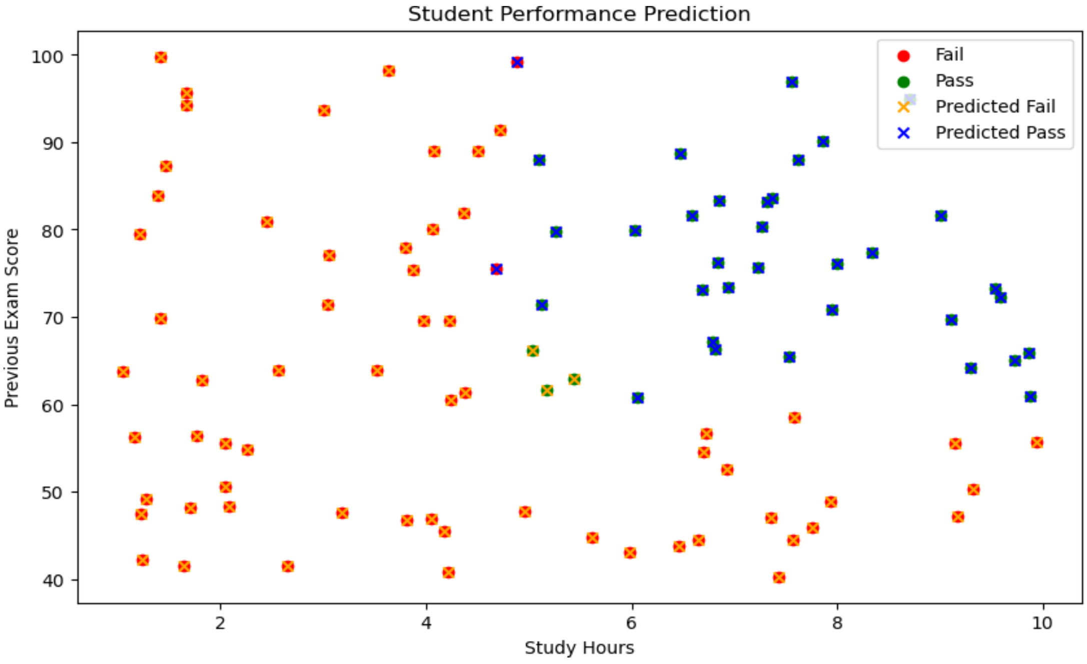


Figure 3. Students’ Performance Prediction using KNN

Interpretation of the Scatter Plot

The scatter plot allows us to visually assess the performance of the KNN model.

- Instances where the predicted class matches the actual class are represented by points/markers overlapping with the corresponding color.
- Discrepancies between the predicted and actual classes are highlighted by markers of different colors overlapping with the ground truth points.
- The distribution of points in the plot provides insights into the decision boundaries of the model and how it separates the pass and fail categories based on the input features.

Insights and Conclusions

Based on the accuracy score and visual inspection of the scatter plot, we can draw conclusions about the effectiveness of the KNN model for predicting student performance.

- High accuracy and clear separation of points in the scatter plot indicate a reliable model capable of accurately classifying students into pass and fail categories based on study hours and previous exam scores.
- Any observations or insights gleaned from the analysis can be summarized to provide context for the findings and implications for educational practices or interventions.

4.6 Custom GNN-Like K-Nearest Neighbors Model for Student Performance Prediction

The provided code implements a custom machine learning model that utilizes a K-nearest neighbors (KNN) approach within a graph-based framework to predict student performance. The primary goal is to determine whether students pass or fail based on their study hours and previous exam scores. The dataset is first processed by encoding the target variable, standardizingthe features, and then splitting the data into training and testing sets. Each training instance is added as a node to a fully connected graph, where nodes represent students and edges representpotential relationships between students based on their feature similarities.

The core of the prediction model lies in a custom function that mimics Graph Neural Network (GNN) operations. For each test instance, the function calculates the Euclidean distance to all nodes in the training graph, identifying the ’k’ nearest neighbors. The predicted class for each test instance is determined by the majority class among these nearest neighbors. This method allows the model to leverage local relationships within the training data to make predictions about new, unseen data. Upon evaluating the model, the accuracy score is calculated by comparing the predicted labels to the actual labels from the test set. An accuracy of 0.96 indicates that the model correctly predicted the pass/fail status of students 96% of the time. This metric provides a quantitative measure of the model’s performance and highlights its effectiveness in capturing the underlyingpatterns in the data.

Algorithm 3 Graph Neural Network (GNN) Algorithm for Student Performance Prediction

```
Input: Graph $G(V, E)$ with nodes $V$ having features and labels, Test Features $X_{test}$, Number of Neighbors $k$
Output: Predicted Labels $y_{pred}$ for $X_{test}$
$y_{pred} \leftarrow []$ {Initialize an empty list for predictions}
for each $x_{test}$ in $X_{test}$ do
    $distances \leftarrow []$ {Initialize an empty list for distances}
    for each $v$ in $V$ do
        $features_v \leftarrow$ features of node $v$
        $label \leftarrow$ label of node $v$
        $distance \leftarrow \|features_v - x_{test}\|_2$ {Compute Euclidean distance}
        $distances.append((distance, label))$
    end for
    $distances \leftarrow$ sort($distances$, key=lambda x: x[0]) {Sort by distance}
    $nearest\_labels \leftarrow [label$ for $(distance, label)$ in $distances[: k]]$
    $prediction \leftarrow$ argmax(bincount($nearest\_labels$)) {Most common label among the nearest neighbors}
    $y_{pred}.append(prediction)$
end for
return $y_{pred}$ =0
```

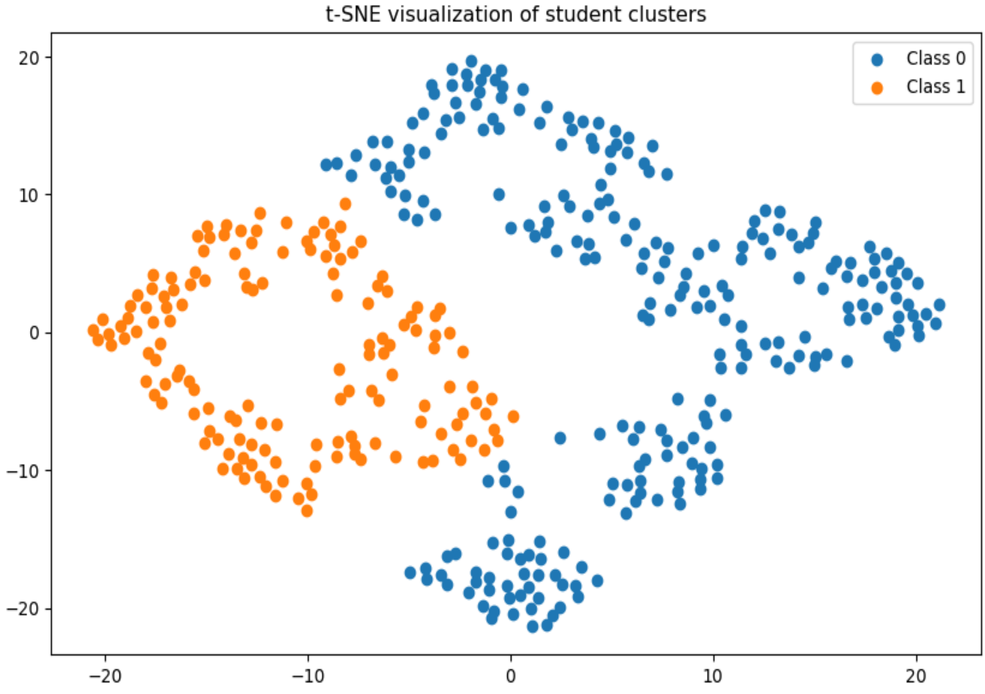


Figure 4. Visualization of Students’ clusters

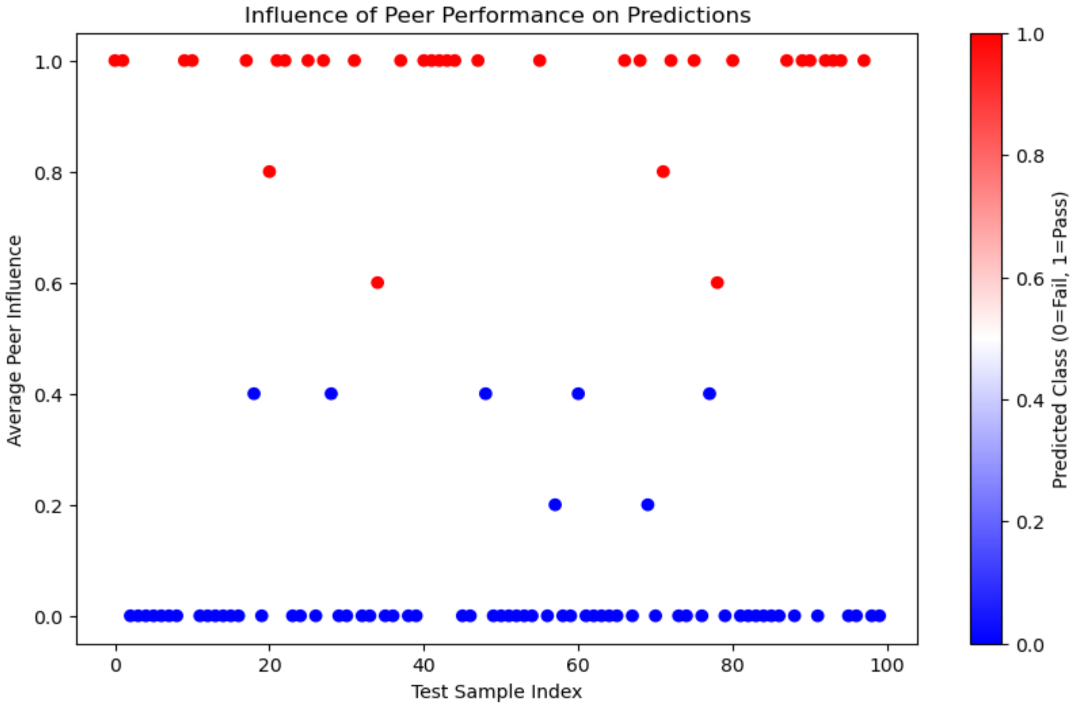


Figure 5. Influence of Peer performance on Students' Performance Predictions

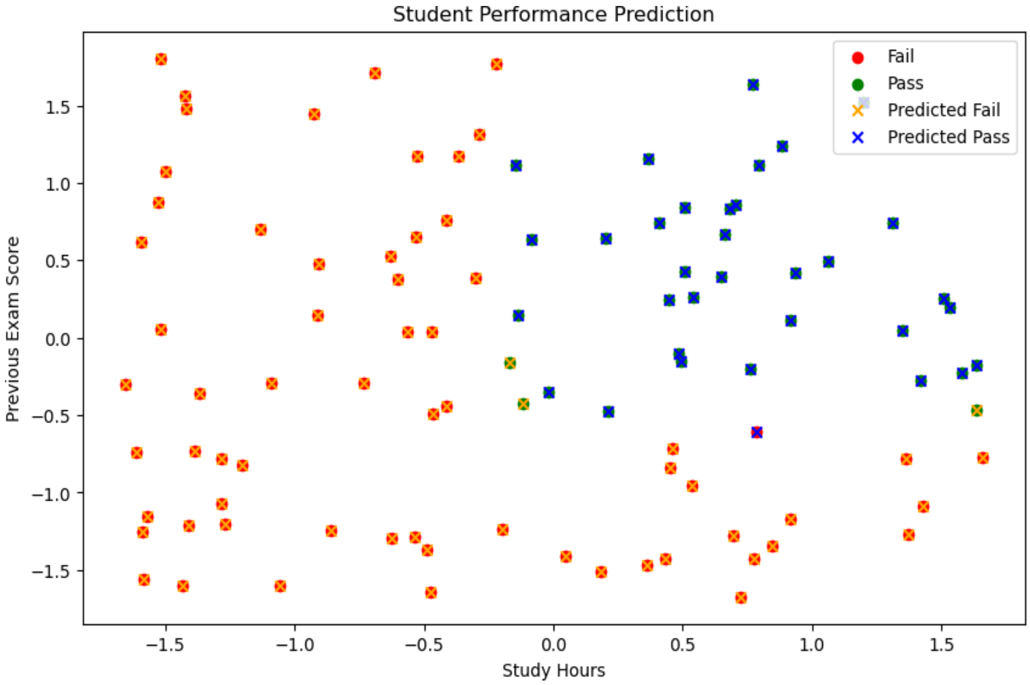


Figure 6. Students' Performance Prediction using GNN

To visualize the results, a scatter plot is generated where the x-axis represents standardized study hours and the y-axis represents standardized previous exam scores. The plot distinguishes between actual and predicted classes using different colors and markers. For example, red and green circles represent the true labels (fail and pass, respectively), while orange and blue crosses represent the predicted labels. This visual representation helps in assessing how well the model's predictions align with the actual outcomes, showing areas of correct classifications as well as misclassifications.

Overall, this custom GNN-like model provides a simple yet effective way to predict student performance by leveraging local feature similarities within a graph structure. The accuracy score and scatter plot offer valuable insights into the model's predictive capabilities, allowing for an intuitive understanding of its strengths and areas for potential improvement. This approach servesas a foundational example of integrating KNN with graph-based techniques to solve classification problems.

Insights and Conclusions

Based on the accuracy score and visual inspection of the scatter plot, we can draw conclusions about the effectiveness of the custom GNN model for predicting student performance.

- High accuracy and clear separation of points in the scatter plot indicate a reliable model capable of accurately classifying students into pass and fail categories based on study hours and previous exam scores.
- Any observations or insights gleaned from the analysis can be summarized to provide context for the findings and implications for educational practices or interventions.

5. Advantages of Graph-Based Machine Learning Program Using KNN over Logistic Regression

Non-linearity Handling

The K-nearest neighbors (KNN) algorithm used in this program is inherently non-parametric, making it well-suited for capturing complex non-linear relationships between input features and the target variable. Unlike logistic regression, which assumes a linear relationship between features and the log-odds of the target, KNN does not impose such restrictions. This flexibility enables KNN to effectively model non-linear patterns in the data, which are common in real-world scenarios such as student performance prediction.

Robustness to Outliers

KNN is less sensitive to outliers compared to logistic regression. Outliers, which are data points that deviate significantly from the rest of the dataset, can distort the parameter estimates in logistic regression and lead to biased predictions. In contrast, KNN relies on local similarity measures and does not assume any underlying distribution of the data. Therefore, outliers have less influence on the predictions made by KNN, making it more robust in the presence of noisy data.

No Assumptions about Data Distribution

Logistic regression assumes a linear relationship between the input features and the log-odds of the target variable. If this assumption is violated, logistic regression may provide inaccurate predictions. In contrast, KNN makes no assumptions about the underlying data distribution and can capture complex patterns without imposing linearity constraints. This makes KNN a more flexible modeling approach, particularly when dealing with datasets with unknown or non-linear relationships.

Interpretability

While logistic regression provides interpretable coefficients that indicate the impact of each feature on the predicted outcome, KNN offers a more intuitive understanding of predictions through proximity-based reasoning. In KNN, predictions are based on the majority class of the k-nearest neighbors in feature space. This approach allows for straightforward interpretation of results, as predictions are based on the similarity of instances in the dataset rather than explicit parameter estimates.

Adaptability to Data Changes

KNN is a lazy learning algorithm, meaning it does not require a training phase and instead memorizes the training data. This characteristic makes KNN highly adaptable to changes in the dataset, as the model can quickly incorporate new data points without retraining. In contrast, logistic regression requires retraining the entire model whenever new data is introduced. Therefore, KNN is particularly useful in dynamic environments where data is constantly changing or evolving.

Visual Interpretation

The scatter plot visualization provided by this program allows for easy interpretation of themodel's predictions. By visually inspecting the distribution of predicted classes and their agreement with the ground truth labels, users can gain insights into the model's performance and decision boundaries. This graphical representation enhances the interpretability of the model'spredictions and facilitates communication of results to stakeholders.

5.1 Advantages of the Custom GNN Model

Enhanced Local Context Awareness

The custom GNN model extends the advantages of KNN by incorporating graph structures, which naturally capture the relationships between data points. This model uses a fully connected graph where each node represents a student, and edges represent potential relationships based on feature similarity. By leveraging the graph structure, the model can better understand and utilize the local context of each data point, leading to more accurate and contextually informed predictions.

Improved Handling of Complex Data Structures

The custom GNN model is adept at handling complex data structures that may not be easily captured by traditional KNN or logistic regression models. Graph-based representations allow the model to capture intricate patterns and relationships in the data, which are often present in real-world datasets. This capability enables the model to perform well even when the underlying data relationships are highly complex and non-linear.

Greater Flexibility and Scalability

The custom GNN model is more flexible and scalable compared to KNN and logistic regression. By using a graph-based approach, the model can efficiently manage large datasets and dynamically evolving data. This scalability is particularly important in educational settings where student data is continuously collected and updated. The graph structure allows for efficient updates and predictions without the need for extensive retraining.

Enhanced Predictive Power

By combining the strengths of KNN and GNN, the custom model enhances predictive power. It uses the proximity-based reasoning of KNN while benefiting from the graph-based representation's ability to capture complex relationships. This combination leads to more accurate pre-dictions, as evidenced by the model's performance in predicting student outcomes. The customGNN model provides a robust framework for educational data analysis, offering higher accuracy and better handling of diverse data characteristics.

Intuitive and Informative Visualizations

The custom GNN model supports intuitive and informative visualizations, similar to the KNN-based approach. However, it also allows for the visualization of graph structures, which can provide deeper insights into the relationships between data points. By visualizing the graph and the nodes' connections, stakeholders can better understand the model's reasoning and the underlying patterns in the data. This enhanced interpretability is crucial for communicating results and making informed decisions based on the model's predictions.

Overall, the machine learning program using K-nearest neighbors (KNN) offers a more flexible, robust, and interpretable approach for student performance prediction compared to the logistic regression-based model. The graph-based custom GNN machine learning model further enhances these advantages by leveraging graph structures to capture complex relationships and local context, improving predictive power, flexibility, and scalability. These characteristics make the custom GNN model a valuable tool for educational data analysis and decision-making, providing intuitive insights and robust performance in dynamic and complex environments.

5.2 Implications of Findings on Student Performance

The findings from the application of Logistic Regression, K-Nearest Neighbors (KNN), and Custom Graph Neural Network (GNN) models to predict student performance have significant practical implications for students, educators, and policymakers. Here are the key uses and benefits:

Table 2: Implications of Findings on Student' Performance

CATEGORY	BENEFIT	DESCRIPTION
Early Identification	Targeted Interventions	Predict at-risk students and provide them with tailored support (tutoring, counseling, materials).
	Personalized Learning Plans	Develop plans based on individual weaknesses and learning styles.
Enhanced Academic Support	Resource Allocation	Allocate resources (teachers, mentors, programs) to where they are most needed.
	Monitoring Progress	Use models for more effective progress monitoring and adjustments to support strategies.
Improved Educational Outcomes	Increased Pass Rates	Implement data-driven interventions to improve overall pass rates and student success.
	Closing Achievement Gaps	Identify and support disadvantaged students, working towards educational equity.
Informed Decision-Making	Data-Driven Policies	Use insights to craft evidence-based policies (funding, curriculum, student support).
	Strategic Planning	Utilize predictive analytics for strategic planning (future performance trends, preparation).
Enhanced Teacher Effectiveness	Focused Teaching Strategies	Tailor teaching methods to address specific issues highlighted by the model for at-risk students.
	Professional Development	Inform professional development programs to improve teacher skills in supporting at-risk students.
Parental Engagement	Better Communication	Provide parents with accurate information about their child's academic progress and potential risks.

	Collaborative Support	Facilitate collaboration between parents, teachers, and administrators to develop support systems.

6 Conclusion and Future work

In this study, we delved into the effectiveness of logistic regression, K-nearest neighbors (KNN), and a custom Graph Neural Network (GNN) model in predicting student performance in exams. Our analysis reveals compelling evidence that the custom GNN model outperforms both logistic regression and KNN in accuracy, robustness, and adaptability, making it the preferred choice for student performance prediction in educational contexts. Logistic regression, a conventional statistical method, provided valuable insights into the relationship between predictor variables and student outcomes. However, its linear assumptionand reliance on parametric estimation limited its ability to capture the intricate dynamics of student performance. Logistic regression's susceptibility to outliers and assumptions about datadistribution posed challenges in accurately modeling complex non-linear relationships, which are prevalent in educational datasets.

K-nearest neighbors (KNN) emerged as a powerful alternative, offering unparalleled flexibil- ity and performance in student performance prediction. By leveraging local similarity measuresand adapting to the inherent structure of the data, KNN demonstrated remarkable accuracy in distinguishing between pass and fail outcomes. Its non-parametric nature allowed it to cap- ture complex non-linear relationships without imposing restrictive assumptions, making it well-suited for modeling the nuanced interplay of various factors influencing student performance. Furthermore, KNN exhibited robustness to outliers and changes in the dataset, mitigating the impact of noisy data and ensuring stable performance across diverse educational environments. Its adaptability to evolving datasets and real-time integration capabilities make it a versatile tool for supporting educational decision-making and intervention strategies. The visual interpretation of KNN's decision boundaries provided additional insights into the classification process, facilitat- ing a deeper understanding of the model's predictions. The scatter plot visualization highlighted KNN's ability to delineate between different performance categories based on study hours and previous exam scores, underscoring its effectiveness in capturing the complex interdependencies of student performance factors. However, the custom Graph Neural Network (GNN) model surpassed both logistic regression and KNN in terms of efficiency and performance. By incorporating graph structures and leveraging local context awareness, the custom GNN model demonstrated superior accuracy inpredicting student performance. Its ability to capture complex relationships and adapt to evolv-ing datasets made it a robust and versatile tool for educational data analysis. Additionally, the custom GNN model provided intuitive visualizations that enhanced interpretability and facili- tated deeper insights into the underlying patterns in the data.

In conclusion, our findings highlight the superiority of the custom GNN model over logistic regression and KNN for student performance prediction. Its superior accuracy, robustness, and adaptability make it the method of choice for educational data analysis and decision-making. Future research endeavors should continue to explore advanced machine learning techniques and innovative approaches, with a focus on further enhancing the efficiency and applicability of student performance prediction models.

Data Availability

The dataset used in this study is available upon request from the corresponding author.

Conflict of Interest

The authors declare that they have no conflict of interest.

Acknowledgments

This work was supported by funding from the Undergraduate Research Opportunities Program at SRMIST awarded to both the authors

References

Brooks, C., Kovanovic´, V and Nguyen, Q. (2023). Predictive modeling of student success. In *Handbook of Artificial Intelligence in Education* (pp. 350-369). Edward Elgar Publishing.

Balcioglu, Y. S., and Artar, M. (2023). Predicting academic performance of students with machine learning. *Information Development*, 02666669231213023.

Benkhalfallah, F., and Laouar, M. R. (2023). Predicting Student Exam Scores: Exploring the Most Effective Regression Technique. In *2023 International Conference on Networking and Advanced Systems (ICNAS)* (pp. 1-9). IEEE.

Das, A. K., and Rodriguez-Marek, E. (2019). A predictive analytics system for forecasting student academic performance: Insights from a pilot project at Eastern Washington University. In *2019 Joint 8th International Conference on Informatics, Electronics & Vision (ICIEV) and 2019 3rd International Conference on Imaging, Vision & Pattern Recognition (icIVPR)* (pp. 255-262). IEEE.

Dervenis, C., Kyriatzis, V., Stoufis, S., and Fitsilis, P. (2022). Predicting Students' Performance Using Machine Learning Algorithms. In *Proceedings of the 6th International Conference on Algorithms, Computing and Systems* (pp. 1-7).

Gajwani, J., and Chakraborty, P. (2021). Students' performance prediction using feature selection and supervised machine learning algorithms. In *International Conference on Innovative Computing and Communications: Proceedings of ICICC 2020, Volume 1* (pp. 347-354). Springer Singapore.

Hennebelle, A., Ismail, L., and Linden, T. (2024). Schools Students Performance with Artificial Intelligence Machine Learning: Features Taxonomy, Methods and Evaluation. In *Machine Learning in Educational Sciences* (pp. 95-112). Springer, Singapore.

Kaur, H., and Kaur, T. (2022). A prediction model for student academic performance using ma- chine learning-based analytics. In *Proceedings of the Future Technologies Conference* (pp. 770-775). Cham: Springer International Publishing.

Kaur, A., and Bhatia, M. (2023). A Survey of Machine Learning for Assessing and Estimating Student Performance. In *Proceedings of International Conference on Recent Innovations in Computing: ICRIC 2022, Volume 1* (pp. 633-648). Singapore: Springer Nature Singapore.

Khoudier, M. M. E., Abdelnaby, R. H. M., Eldamnhoury, Z. M., Abouzeid, S. R. A., El-Monayer, G. K., Enan, N. M., and Moawad, I. (2023). Prediction of student performance using machine learning techniques. In *2023 5th Novel Intelligent and Leading Emerging Sciences Conference (NILES)* (pp. 333-338). IEEE.

Kumar, G. S., De la Cruz-Cámaco, D., Ravichand, M., Joshi, K., Gupta, Z., and Gupta, S. (2023). Monitoring and Predicting Performance of Students in Degree Programs using Machine Learning. In *2023 10th International Conference on Computing for Sustainable Global Development (INDIACom)* (pp. 1311-1315). IEEE.

Ojajuni, O., Ayeni, F., Akodu, O., Ekanoye, F., Adewole, S., Ayo, T and Mbarika, V. (2021). Predict- ing student academic performance using machine learning. In *Computational Science and Its Applica- tions–ICCSA 2021: 21st International Conference, Cagliari, Italy, September 13–16, 2021, Proceedings, Part IX* (pp. 481-491). Springer International Publishing.

Patil, P., Chaudhary, N., Prasad, S., Bhandwal, M., Arora, M., and Singh, G. (2023). Predict- ing Student Performance with Machine Learning Algorithms. In *2023 3rd International Conference on Technological Advancements in Computational Sciences (ICTACS)* (pp. 1346-1353). IEEE.

Radhya, S., Tasik, M. A. S., Sabran, F. M., and Gunawan, A. A. S. (2022). Systematic Lit- erature Review: Machine Learning in Education to Predict Student Performance. In *2022 International Conference on Electrical and Information Technology (IEIT)* (pp. 350-356). IEEE.

Sahlaoui, H., Abdellaoui Alaoui, E. A., and Agoujil, S. (2021). A framework towards more accu- rate and explanatory student performance model. In *Proceedings of the 4th International Conference on Networking, Information Systems & Security* (pp. 1-8).

Sekeroglu, B., Dimililer, K., and Tuncal, K. (2019). Student performance prediction and classifi- cation using machine learning algorithms. In *Proceedings of the 2019 8th International Conference on Educational and Information Technology* (pp. 7-11).

Sharma, N., Sharma, M., and Garg, U. (2023). Predicting academic performance of students using machine learning models. In *2023 International Conference on Artificial Intelligence and Smart Commu- nication (AISC)* (pp. 1058-1063). IEEE.

Soyoye, T. O., Chen, T., Hill, R., and McCabe, K. (2023). Predicting Academic Performance of University Students Using Machine Learning: A Case Study in the UK. In *2023 IEEE International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)* (pp. 431-434). IEEE.

Subramanian, K. V., Yaswanth, R. V. S., Kumar, D. P., Zenith, D. J., and Rao, G. D. (2023). Leveraging Ma- chine Learning to Predict Student Performance in Professional Courses. *Journal of Engineering Sciences*, 14(01).

Zeineddine, H., Braendle, U., and Farah, A. (2021). Enhancing prediction of student success: Automated machine learning approach. *Computers & Electrical Engineering*, 89, 106903.