**Predicting Student Performance Based on Study Hours and Attendance**

**1. Introduction**

**Project Goals and Significance:**

* The objective of this project is to predict student exam performance based on **two key factors:** study hours and class attendance.
* As education systems increasingly rely on data-driven insights, understanding the relationship between student behavior (such as study habits and attendance) and academic performance becomes crucial for both educators and students.
* This study aims to explore and model the effect of study hours and attendance on exam scores. By building a predictive model, the project seeks to offer insights into how these factors can help predict a student's performance in exams. This has significant potential for improving academic outcomes, providing early warnings for underperforming students, and supporting personalized educational interventions.
* The results of this project could help educators, administrators, and students themselves make informed decisions on how to optimize study hours and improve attendance to boost exam performance.

**2. Methodology**

**Data Collection:**

For the analysis, we use a dataset that includes the following features for each student:

* **Student ID**: Unique identifier for each student.
* **Study Hours**: The total number of hours spent studying before the exam.
* **Attendance (%)**: The percentage of classes attended by the student.
* **Exam Score**: The student's score on the exam.

An example of the dataset is :

| **Student ID** | **Study Hours** | **Attendance (%)** | **Exam Score** |
| --- | --- | --- | --- |
| 1 | 5 | 80 | 75 |
| 2 | 10 | 90 | 85 |
| 3 | 15 | 70 | 65 |
| 4 | 12 | 85 | 80 |
| 5 | 8 | 60 | 70 |

**Data Preprocessing:**

1. **Data Cleaning**: The dataset was checked for missing values or anomalies. No missing values were found, and the data was considered clean for analysis.
2. **Feature Scaling**: The features (study hours, attendance, and exam score) were scaled if necessary to ensure they are on a similar scale and to improve model performance.
3. **Correlation Analysis**: Before modeling, the correlation between the features was examined to understand how study hours, attendance, and exam scores relate to each other. This helps identify which features are most predictive of the exam score.

**A computer screen shot of a test

Description automatically generatedPseudo Code for Model Implementation:**

**Modeling Techniques:**

* To predict exam scores, a **multiple linear regression model** was chosen.
* The model uses study hours and attendance as independent variables and the exam score as the dependent variable.
* Linear regression is suitable for this problem since we aim to model the relationship between continuous predictors and a continuous outcome.

**The steps in the modeling process included:**

1. **Splitting the data**: The data was split into a training set (80%) and a testing set (20%).
2. **Fitting the model**: A linear regression model was trained on the training data to learn the relationship between study hours, attendance, and exam scores.
3. **Model Evaluation**: The model's performance was evaluated using Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R²).

**3. Results**

**Analysis and Model Performance:**

* **Descriptive Statistics**:  
  The dataset consists of 100 students. The average study time is **10.69 hours**, with an average attendance of **79.35%**. The average exam score is **80.39%**, indicating that most students performed relatively well on the exam.
* **Correlation Analysis**:  
  The correlation matrix revealed that study hours and attendance have a positive relationship with exam scores, with **attendance** showing a stronger correlation (**0.91**) compared to study hours (**0.76**). This indicates that both study hours and attendance are significant predictors of exam performance.
* **Model Coefficients**:  
  The multiple linear regression model yielded the following coefficients:
  + **Intercept**: 31.36
  + **Study Hours Coefficient**: 0.47
  + **Attendance Coefficient**: 0.56

These coefficients suggest that for every additional hour of study, the exam score increases by **0.47 points**, and for every 1% increase in attendance, the exam score increases by **0.56 points**.

* **Evaluation Metrics**:
  + **Mean Absolute Error (MAE)**: 1.96
  + **Mean Squared Error (MSE)**: 5.93
  + **R-squared (R²)**: 0.89  
    The R-squared value of **0.89** indicates that the model explains approximately **89%** of the variance in exam scores, which is a strong result for predictive modeling.

**4. Conclusion**

**Findings:**

* The analysis and regression model reveal strong predictive relationships between study hours, attendance, and exam scores.
* The model demonstrates that both study hours and class attendance are significant predictors of student performance, with attendance having a slightly stronger impact on exam scores.
* **Study Hours**: An increase in study time leads to a moderate increase in exam scores, confirming that consistent study habits positively impact performance.
* **Attendance**: Regular class attendance has a more significant impact on exam scores, underscoring the importance of being present in class for academic success.

**Future Scope:**

1. **Additional Factors**: Future models could include additional features such as participation in class activities, time management skills, or access to study resources. These factors may provide an even more comprehensive understanding of student performance.
2. **Non-linear Models**: Given the complexity of human learning, exploring non-linear models (such as decision trees, random forests, or neural networks) could offer improved accuracy and more nuanced insights.
3. **Longitudinal Studies**: A longer-term study that tracks students' performance over multiple years could provide deeper insights into the impact of study habits and attendance over time.
4. **Personalized Recommendations**: The model could be extended to provide personalized recommendations for students to optimize their study hours and attendance based on their predicted performance.