STUDENT PERFORMANCE PREDICTION REPORT

KIET GROUP OF INSTITUTIONS

Department of Computer Science & Engineering (AI)

Title Page

Project Title: Student Performance Prediction

Problem Statement:

Predicting student performance in an **AI exam** based on **study hours, attendance, and past grades** using a **machine learning model**.

Personal Details:

- **Student Name:** Aryan Khokhar
- **Roll Number:** 202401100300068
- **Institution:** KIET Group of Institutions
- **Course:** B.Tech CSE (AI)
- **Date:** 11-03-2025

1. Introduction

Student performance prediction plays a crucial role in **educational data analysis**. By analyzing factors such as **study hours, attendance, and past academic records**, we can build a **predictive model** that categorizes students as either **pass** or **fail** in an AI exam.

This project employs **Logistic Regression**, a widely used **classification algorithm**, to predict student outcomes based on the given input features. The goal is to provide insights to educators, enabling them to take **proactive measures** to support students who are at risk of failing.

2. Methodology

2.1 Data Collection

A **manually created dataset** is used, which includes the following attributes:

- **Study Hours:** Number of hours a student studies per week.
- **Attendance:** Percentage of classes attended.
- **Past Grades:** Previous academic performance.
- **Pass Exam:** Binary label (**1 = Pass, 0 = Fail**).

2.2 Data Preprocessing

- The dataset was **converted into a pandas DataFrame** for easy manipulation.
- Features (**independent variables**) and target labels (**pass/fail outcome**) were separated.

2.3 Data Splitting

- The dataset was **split into training (80%) and testing (20%)** using train_test_split().
- This ensures that the model **learns from one part of the data** and is tested on **unseen data** for evaluation.

2.4 Model Training

- **Logistic Regression** was chosen as the predictive model because it is effective for **binary classification problems**.
- The model was trained on the **training dataset (X_train, y_train)**.

2.5 Prediction & Evaluation

- The trained model was used to make predictions on **X_test**.
- The **accuracy score** was computed using accuracy_score().

2.6 Data Visualization

To better understand the relationships between **study hours, attendance, and past grades**, scatter plots were generated to visualize:

- **Study Hours vs. Attendance**
- **Study Hours vs. Past Grades**
- **Attendance vs. Past Grades**

3. Code Implementation

python

CopyEdit

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

from sklearn.model_selection import train_test_split

from sklearn.linear_model import LogisticRegression

from sklearn.metrics import accuracy_score

```
# Sample dataset for student performance prediction
data = {
    'study_hours': [10, 15, 7, 20, 5, 12, 8, 18, 25, 3],
    'attendance': [90, 85, 75, 95, 60, 80, 70, 92, 98, 50],
    'past_grades': [85, 80, 70, 90, 65, 78, 72, 88, 95, 55],
    'pass_exam': [1, 1, 0, 1, 0, 1, 0, 1, 1, 0]  # 1 = Pass, 0 = Fail
}

# Convert data into a DataFrame
df = pd.DataFrame(data)

# Features and target variable
X = df[['study_hours', 'attendance', 'past_grades']]
y = df['pass_exam']

# Split data into training and test sets (80-20)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Train logistic regression model
model = LogisticRegression()
model.fit(X_train, y_train)

# Make predictions
y_pred = model.predict(X_test)

# Evaluate accuracy
```

```
accuracy = accuracy_score(y_test, y_pred)

print(f'Accuracy: {accuracy:.2f}')


# Visualizing relationships

plt.figure(figsize=(12, 6))


plt.subplot(1, 3, 1)

sns.scatterplot(x='study_hours', y='attendance', hue='pass_exam', data=df)

plt.title('Study Hours vs. Attendance')


plt.subplot(1, 3, 2)

sns.scatterplot(x='study_hours', y='past_grades', hue='pass_exam', data=df)

plt.title('Study Hours vs. Past Grades')


plt.subplot(1, 3, 3)

sns.scatterplot(x='attendance', y='past_grades', hue='pass_exam', data=df)

plt.title('Attendance vs. Past Grades')


plt.tight_layout()

plt.show()
```

4. Output & Results

4.1 Accuracy Score

When the model is executed, it calculates and displays an **accuracy score**, indicating how effectively it predicts student performance.

4.2 Predictions

The model successfully predicts whether a student will **pass or fail** based on study hours, attendance, and past grades.

4.3 Data Visualization

The scatter plots help visualize the impact of different factors on a student's performance:

- **Higher study hours and attendance generally lead to passing.**
- **Low attendance combined with low past grades increases failure probability.**

5. Conclusion

- The **Logistic Regression** model proved to be a simple and effective approach for predicting student performance.
- **Study hours, attendance, and past grades** significantly influence a student's success in exams.
- This model can be **enhanced** by adding additional parameters like **homework completion, extracurricular activities, and learning patterns**.

Future Scope

- **Experiment with other ML models** like **Decision Trees, SVM, or Neural Networks** to improve accuracy.
- **Use real-world student data** for better generalization and insights.
- **Integrate this model into educational platforms** for automated student performance analysis.

6. References & Credits

- **Dataset:** Manually created for this experiment.
- **Libraries Used:** pandas, matplotlib, seaborn, scikit-learn.
- **Machine Learning Model: Logistic Regression** (best for binary classification).
- **Scikit-Learn Documentation:** https://scikit-learn.org/
- **Matplotlib for Visualization:** https://matplotlib.org/