Final Report – KNN Classification on Student Performance Dataset

1. Introduction

This project aims to predict student academic success using the Student Performance dataset. The dataset includes demographic and academic factors such as gender, age, study hours, GPA, major, part-time job status, and extracurricular activities. The target variable is a binary outcome: whether a student is successful (Pass) or not (Fail), determined using a GPA threshold. The primary algorithm used is K-Nearest Neighbors (KNN), with Logistic Regression included for comparison.

2. Data Preprocessing

- Dropped irrelevant column: StudentID.
- Encoded categorical variables (Gender, Major, PartTimeJob, ExtraCurricularActivities) using LabelEncoder.
- Created binary target variable: Pass (GPA ≥ 2.5) vs. Fail (GPA < 2.5).
- Scaled features with StandardScaler.
- Split data: 60% training, 20% validation, 20% testing.

3. KNN Model

- Implemented with scikit-learn's KNeighborsClassifier.
- Tuned K from 1 to 20 using validation accuracy.
- Selected optimal K based on highest validation score.
- Evaluated final model on the test set.

4. Cross-Validation

- Performed 5-Fold Cross-Validation on the training set.
- Validation Accuracy: ___%.
- Test Accuracy: ___%.
- Cross-Validation Accuracy: ___%.
- Cross-validation provided a more stable performance estimate and helped detect overfitting.

5. Confusion Matrix Analysis

- Computed confusion matrix on the test set and visualized as a heatmap.
- Derived metrics: Accuracy, Precision, Recall, F1-score.
- Insights: Model performance on both classes and any misclassification patterns.

6. Overfitting Discussion

- Compared training, validation, and test accuracies to assess overfitting.

• No severe overfitting observed due to optimal K selection and cross-validation.
• Techniques applied: increased K, feature scaling, and cross-validation.

7. Visualizations

• Accuracy vs. K plot to identify best K.
• Confusion matrix heatmap for interpretability.
• Optional PCA visualization (2D/3D) to show class separation in feature space.

8. Conclusion

• KNN provided solid results for predicting student outcomes.
• Logistic Regression offered a comparable baseline.
• Future improvements: include additional features (e.g., attendance), try advanced models (e.g., Random Forest), and perform feature importance analysis for interpretability.