CSE422 - Artificial Intelligence

Disease Type Predictor Using Machine Learning

Group Member 1: Zarin Tasnim Vega

ID: 22201039

Group Member 2:Mahir Asaf Khan

ID: 22299187

Submission Date: 4thJanuary, 2026

1. Introduction

What the project aims to do:

The goal of this project is to develop a classification model that predicts the type of disease (Type A, B, or C) a patient may have based on physiological data such as BMI, Blood Pressure, and Cholesterol levels. By automating this process, we aim to assist healthcare providers in making faster, data-driven diagnostic decisions.

Problem statement:

Accurately identifying disease types (Type A, B, or C) is difficult because patients often show similar physical symptoms. Manual diagnosis is slow and prone to errors. There is a need for a fast, automated system that uses medical data to help doctors make correct diagnostic decisions.

Motivation:

This project aims to improve patient care by using machine learning to find hidden patterns in health data. By analyzing factors like BMI, Blood Pressure, and Heart Rate, we can provide early and more reliable disease detection.

2. Dataset Description

Source:

- Link: Disease_dataset

Dataset Details:

- **Number of Features:8**
  ['Age', 'BMI', 'Blood_Pressure', 'Cholesterol', 'Heart_Rate', 'Smoking_Habit', 'Physical_Activity', 'Family_History']
- **Problem Type:** This is a Classification problem because we are predicting discrete classes (Disease Types) rather than continuous numerical values.
- **Number of Data Points:1800 rows**
- **Feature Types:**A mix of Quantitative (Age, BMI, etc.) and Categorical (Smoking_Habit, Family_History).

3. Dataset Pre-processing

1. **NULL Values:**
   a. Columns like BMI, Cholesterol, and Smoking_Habit contained null values.

      Solution:

      We used Imputation. Numerical nulls were filled with the Mean to maintain distribution, and categorical nulls were filled with the Mode (most frequent value).

2. **Categorical Values:**
   a. Machine Learning models require numerical input, but our data contained strings like "Smoker" and "High".

      Solution:

      We applied One-Hot Encoding for independent features and

      Label Encoding for the target variable.

3. **Feature Scaling**
   a. Problem: Features like age (20-80) and cholesterol (150-300) have different scales, which can cause some features to dominate others.

      Solution:

      We applied Standard Scaler to normalize the numerical features. This ensures that features with larger ranges (like Cholesterol) do not dominate those with smaller ranges (like Age), which is critical for distance-based models like KNN and Neural Networks.

4. Dataset Splitting

The dataset was split into training and testing sets:
Training Set: 80% of the data.
Testing Set: 20% of the data.

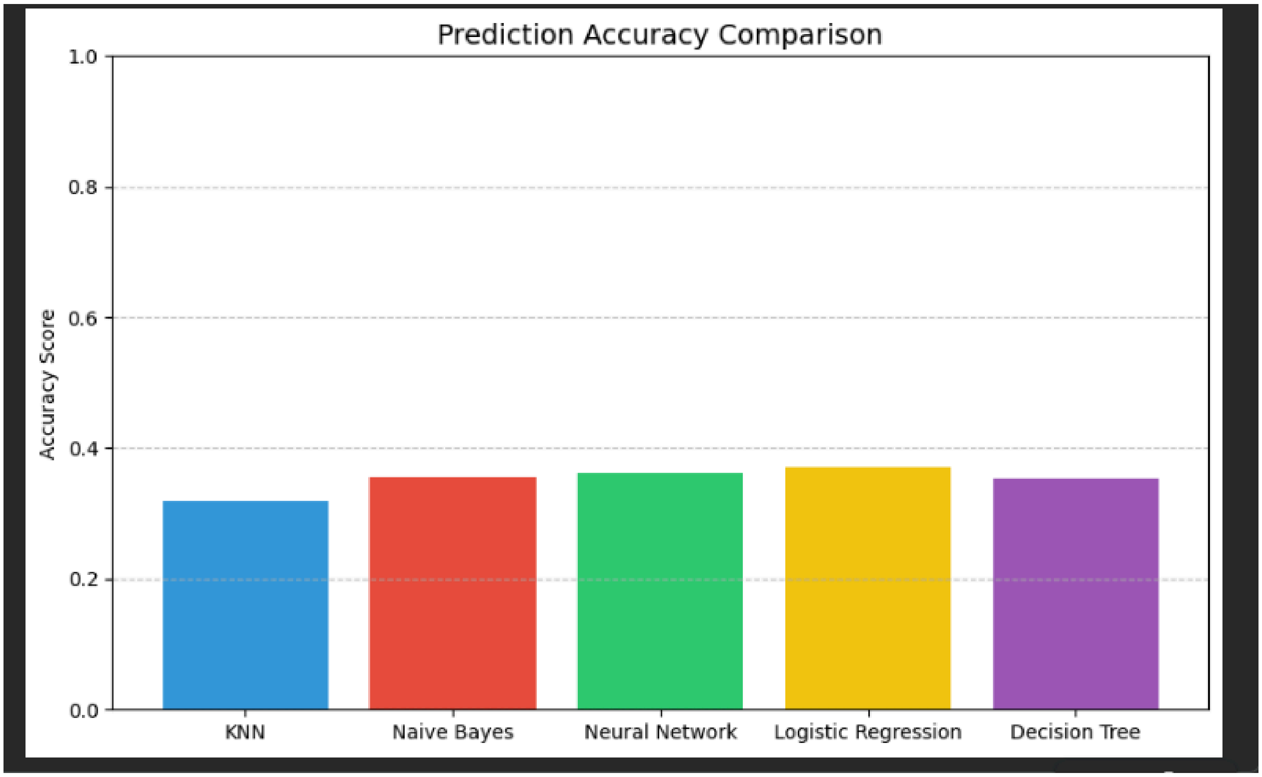
5. Model Training & Testing

We implemented the following models:

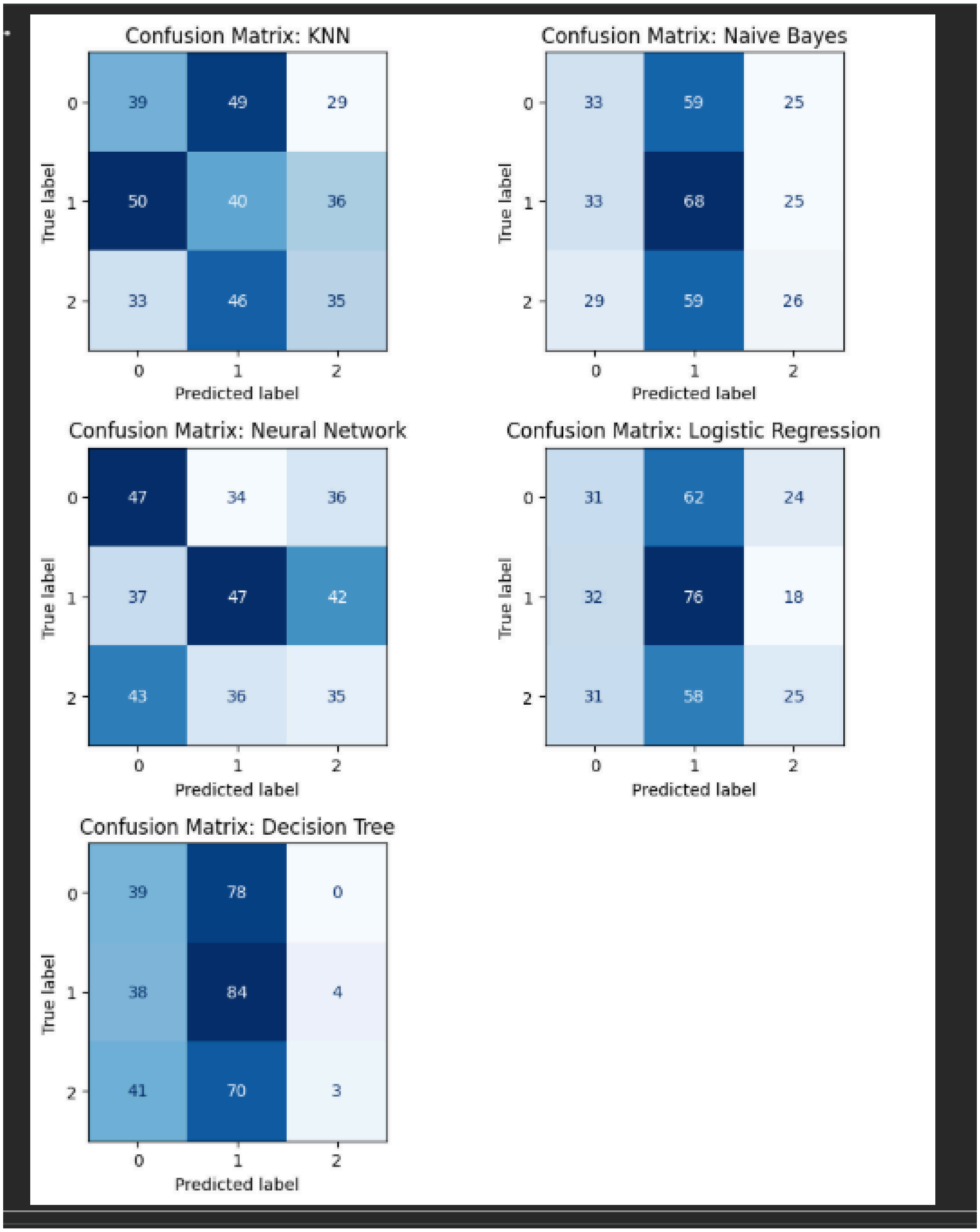
1. **Neural Network (MLPClassifier):** A multi-layer perceptron was used to capture complex non-linear patterns.
2. **Logistic Regression:** Used as a baseline linear classifier.
3. **Decision True:**Used with a max_depth=5 to prevent overfitting.
4. **KNN (K-Nearest Neighbors):** A distance-based classifier to see how patients group together.
5. **Neural Network (MLPClassifier):** A model with two hidden layers to capture complex patterns.
6. **K-Means Clustering (Unsupervised):** We applied K-Means with 3 clusters to see if the data naturally separates into the three disease types without labels.

6. Model Selection/Comparison Analysis

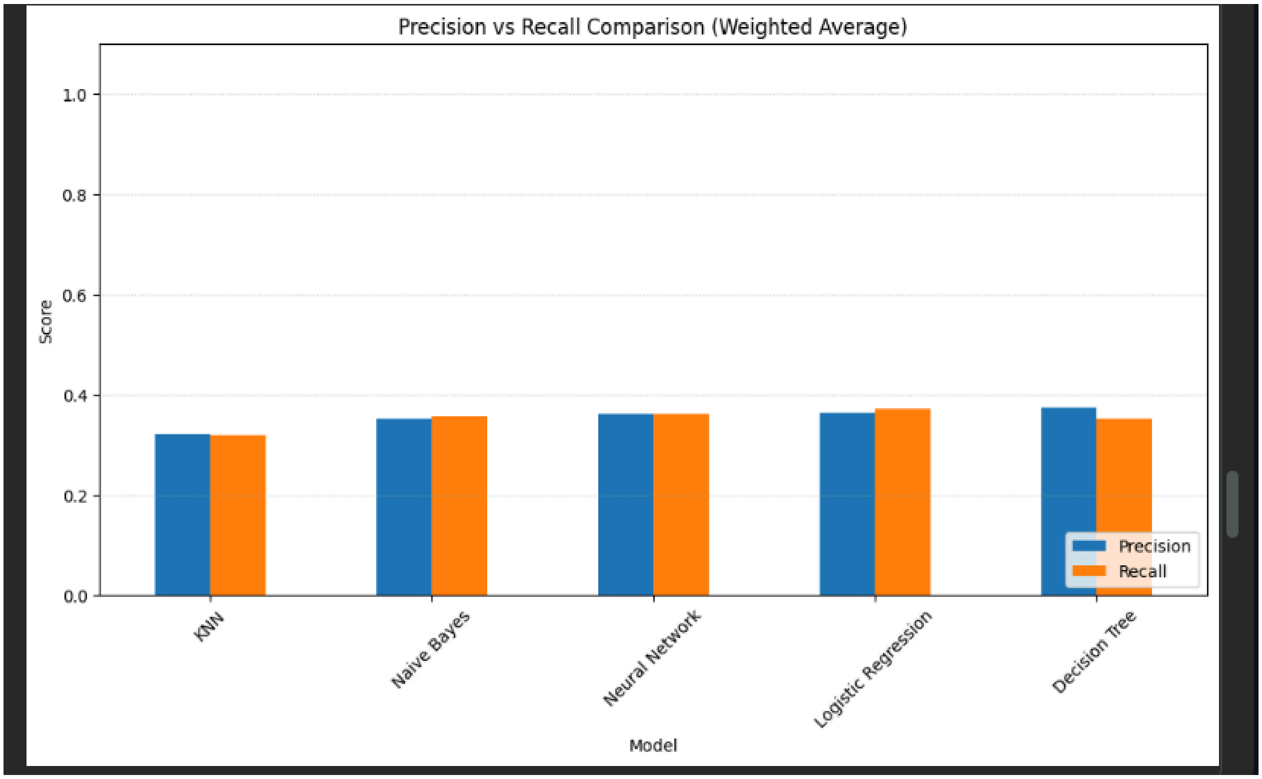
Accuracy: The percentage of correct predictions.



Confusion Matrix: Used to identify specific misclassifications. For example, the model occasionally confused Type A with Type B due to overlapping physiological markers.

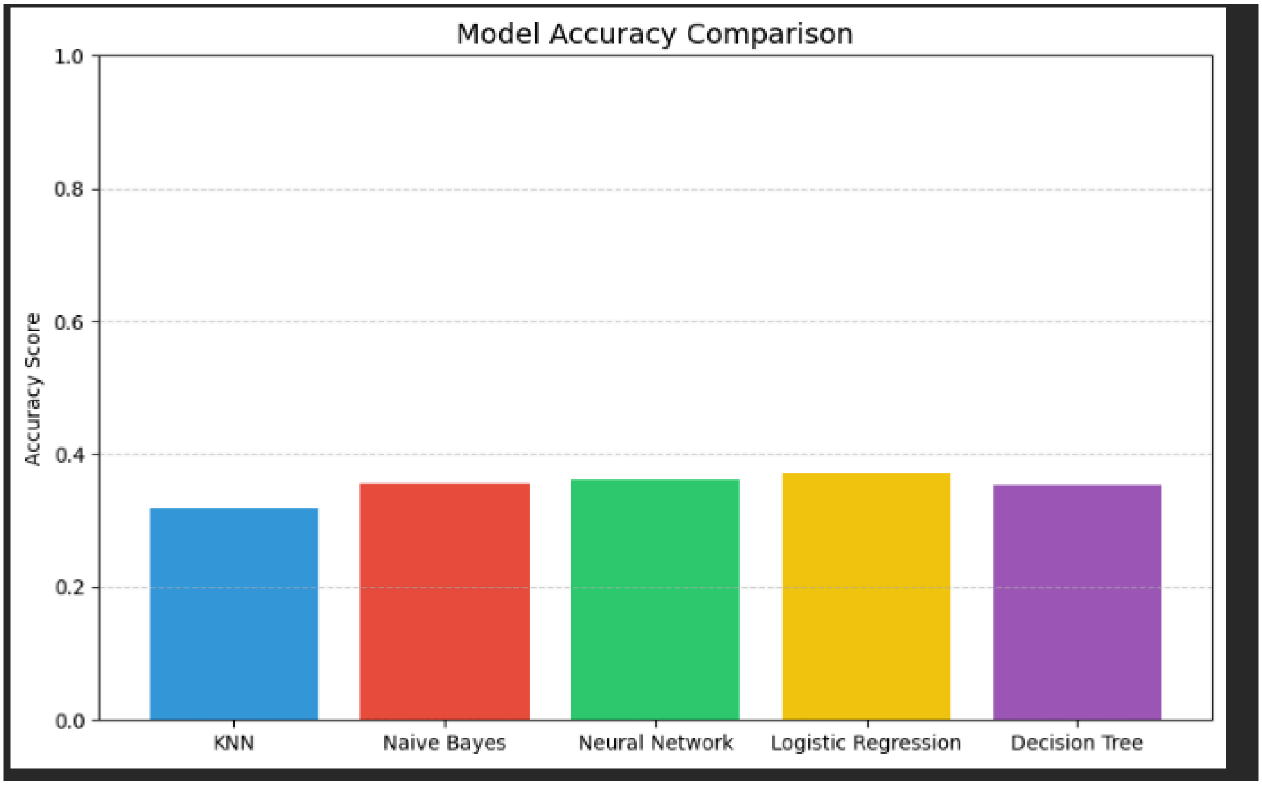


Precision & Recall: We evaluated these to ensure we aren't missing many positive cases (High Recall) and that our positive predictions are reliable (High Precision).

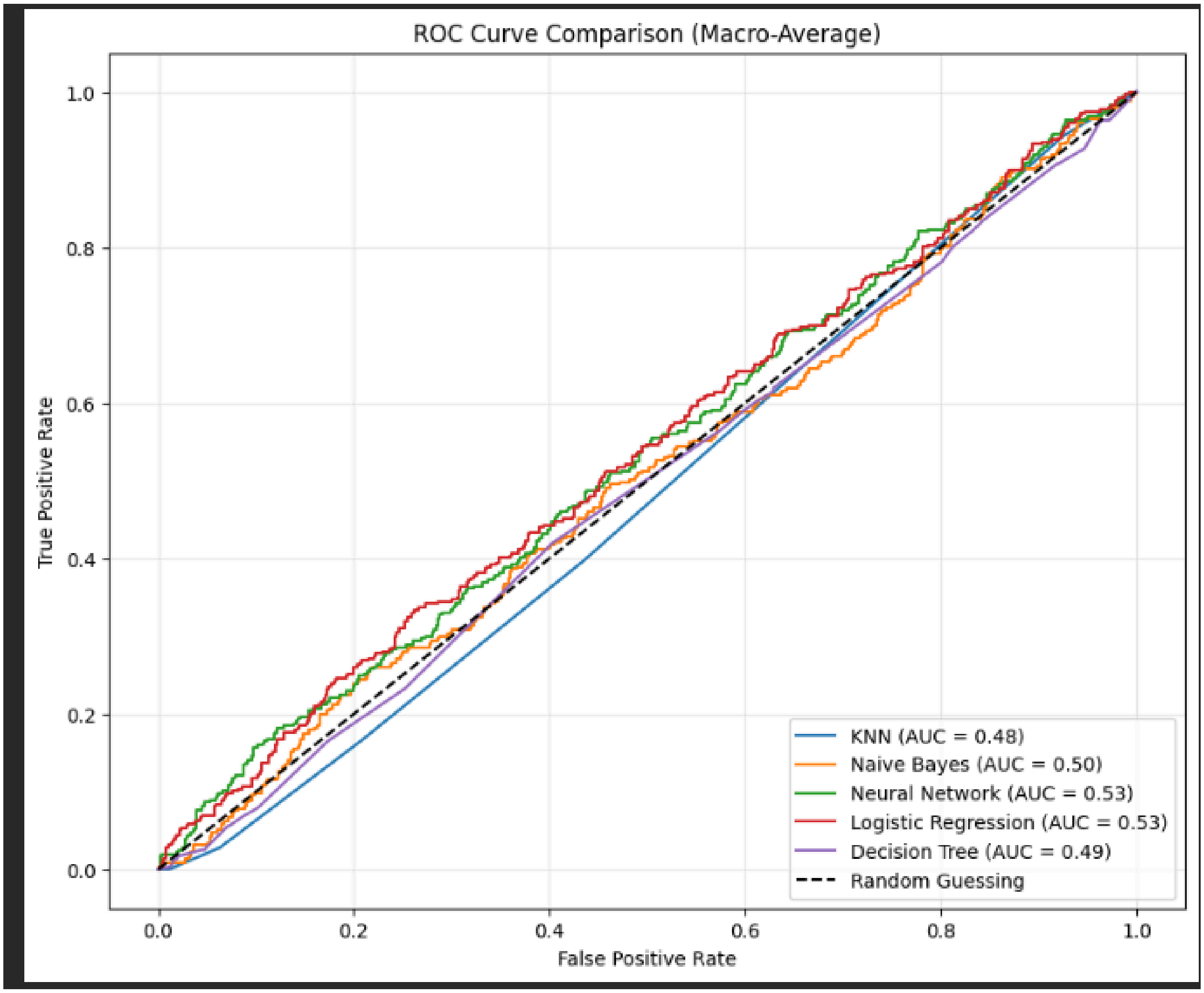


Bar Chart:

A bar chart was generated to compare the accuracy across all models, identifying the best-performing classifier.



AUC score and ROC curve:



7. Conclusion

The project developed a system to predict disease categories. The **Logistic Regression** and **Neural Network** models demonstrated moderate performance after proper feature scaling and outlier removal since the dataset did not have good correlation.

Limitations:

- Very poor correlation between the target and the values.
- Accuracy scores are low due to few data points and low correlation.
- Model accuracy depends heavily on the precision of clinical measurements.
- Some disease types have overlapping symptoms, making them harder to separate perfectly.

Future Improvements:

- Include more diverse data like patient diet, sleep patterns, or genetic history.
- Test "Ensemble" models like Random Forest to see if they provide higher accuracy than single models.

Dataset Challenges:

The dataset had only 1800 data points which isn't enough for forming a proper relationship with high accuracy models.