Machine Learning Models for Disease Type Classification: CSE422 Lab Project

Muntaha Fatema Tahiat, 23341038, Raiyan Zakir Ayiman, 23301211
CSE 422L, Section 11, Group 1

***Abstract*—This lab project focuses on classifying three types of patients using a dataset with nine features, both numerical and categorical. Various machine learning algorithms were implemented, including K-Nearest Neighbors (KNN), Random Forest, Neural Networks, and K-Means clustering. Data preprocessing involved handling missing values, encoding categorical variables, and applying Min-Max scaling. Model performance was evaluated using accuracy, precision, recall, F1-score, confusion matrices, and ROC curves. The results indicate that KNN achieved the highest F1-score and sensitivity, making it the most suitable model for disease prediction where minimizing false negatives is crucial.**

***Index Terms*—Disease Type, Machine Learning, KNN, Random Forest, Classification, Sensitivity, F1-score**

I. Introduction

We have a dataset about three types of patients, which falls under the classification problem in Machine Learning. We will be implementing different types of algorithms, and evaluating the algorithms.

II. Dataset Description

The dataset contains 1800 data points with nine features:

- Numerical: Age, BMI, Blood Pressure, Cholesterol, Heart Rate
- Categorical: Smoking Habit, Physical Activity Level, Family History, Disease Type

Data URL: https://drive.google.com/uc?export=download&id=1inGFudTwX6uZTGhcyVhCIZykKmqJMLuS

A. Numerical Feature Overview

Feature	Count	Mean	Std	Min	Max
Age	1800	49.14	17.40	20	79
BMI	1710	25.11	4.92	7.46	41.77
Blood Pressure	1800	118.67	22.89	80	159
Cholesterol	1710	223.37	43.66	150	299
Heart Rate	1800	89.66	17.07	60	119

TABLE I
NUMERICAL FEATURE SUMMARY OF THE DATASET.

B. Categorical Feature Overview

Feature	Count	Unique	Top	Freq
Smoking Habit	1710	2	Smoker	861
Physical Activity Level	1800	3	Low	636
Family History	1800	2	No	910
Disease Type	1800	3	Type B	634

TABLE II
CATEGORICAL FEATURE SUMMARY OF THE DATASET.

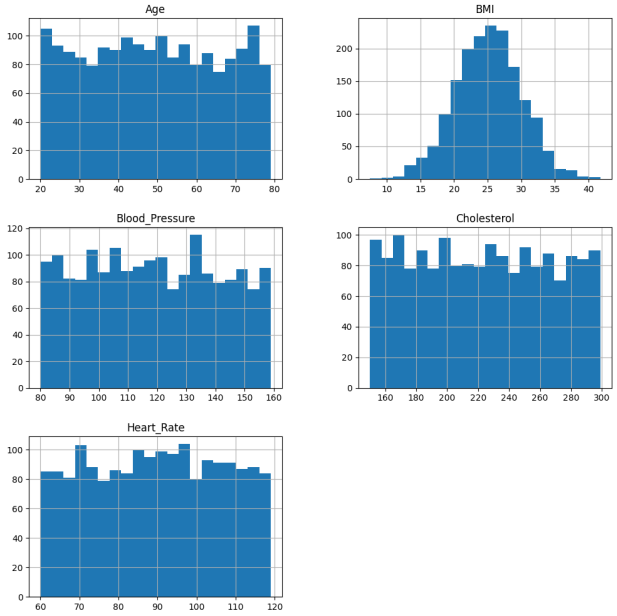

Fig. 1. Histogram of numerical features from the disease type dataset.

III. Data Preprocessing

Three features contained missing values: BMI, Cholesterol, and Smoking Habit.

- BMI and Cholesterol were replaced with the mean.
- Smoking Habit was replaced with the mode.

Categorical variables were encoded as follows:

- Smoking Habit, Family History, and Disease Type: LabelEncoder (0 or 1)
- Physical Activity Level: mapped to 0 (Low), 1 (Moderate), 2 (High)

As we can see in the correlation table attached below, no features are overly correlated with each other. As a result, it was not necessary to drop any features.

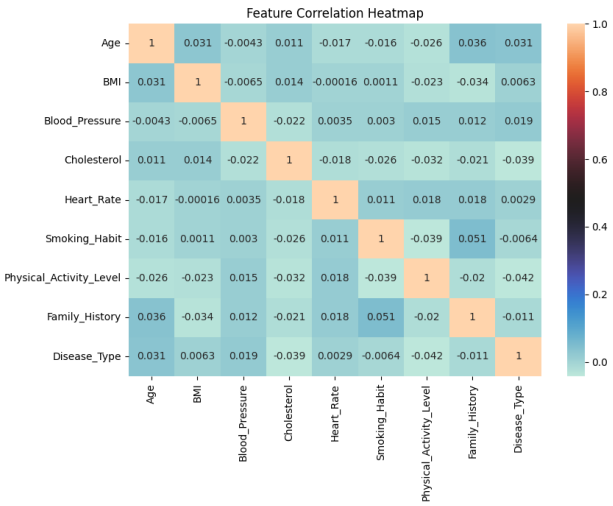

Fig. 2. Correlation matrix of numerical and categorical features in the disease type dataset.

Scaling using Min-Max normalization was applied to all numerical features to ensure comparable ranges. Feature correlation analysis indicated no highly correlated features, so all features were retained.

IV. Model Training

We trained four models:

1) K-Nearest Neighbors (KNN)
2) Random Forest
3) Neural Network
4) K-Means clustering (unsupervised)

KNN was chosen because it is non-parametric and captures local patterns effectively. Random Forest was selected because it does not assume linearity and can model complex feature interactions. Neural Networks were also tested, but due to the small dataset and weak correlations, their performance was limited. K-Means was used as an unsupervised benchmark.

V. Model Evaluation

A. Performance Metrics

Model	Accuracy	Precision	Recall	F1-Score
KNN	0.37	0.36	0.37	0.36
Random Forest	0.35	0.35	0.35	0.35
Neural Network	0.31	0.31	0.30	0.29
K-Means	0.33	0.34	0.34	0.33

TABLE III
Comparison of model performance metrics.

B. Bar Chart of Accuracy and F1-Score

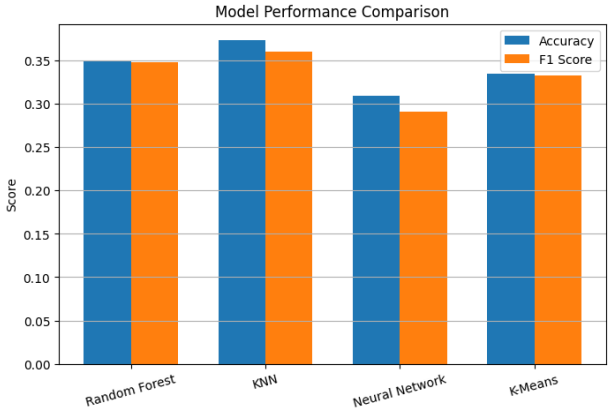

Fig. 3. Bar chart comparing Accuracy and F1-Score for all models.

C. Confusion Matrix

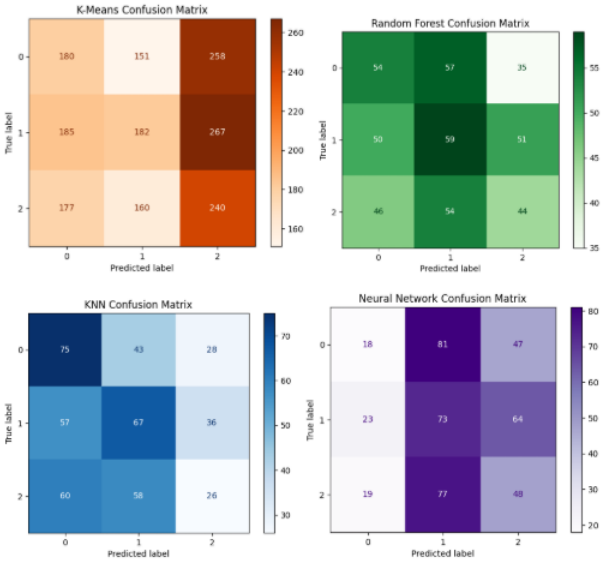

Fig. 4. Confusion matrix for the evaluated models.

D. ROC Curve Comparison

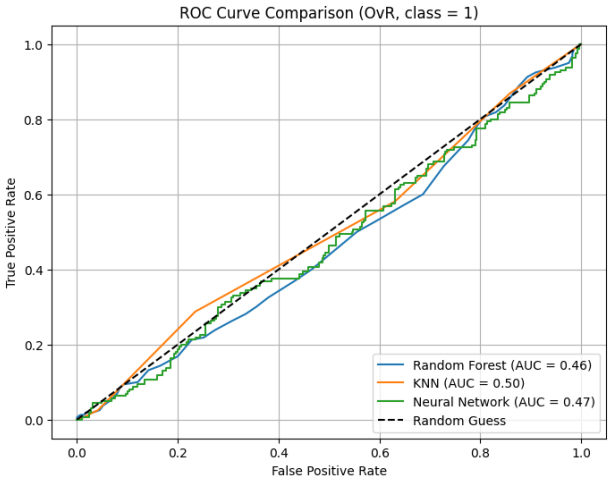

Fig. 5. ROC curve comparison for KNN, Random Forest, and Neural Network models.

VI. Conclusion

Based on the model evaluation, we can see that KNN has the highest f1 value, which means it gives an overall better precision and sensitivity compared to the rest of the model. Also, it has the highest AUC for the ROC curve, meaning it's the best model in this scenario. For a disease prediction, we should emphasize on the sensitivity as we should minimize the number of false negatives which would result in patients going undiagnosed. KNN has the most sensitivity. Therefore, we can conclude by saying, KNN is the best model based on the model evaluations.