CSE422 Project Report

Semester: Fall 2024

Course Name: ARTIFICIAL INTELLIGENCE
Course Code: CSE422

Project name: Coronary Heart Disease Prediction

Submitted By:
Samin Haque (21301628)
Khaled Saifullah Karim (24341262)

Section: 11
Group: 11

Date of Submission: 04/01/2025

Table of Contents

Introduction

In the medical sector the ability to implement early interventions against chronic diseases is revolutionary; which brings the purpose of this project; to predict the '10-year risk' of future "Coronary Heart Disease" (CHD) in patients. By analyzing patient data i.e. Demographics, Lifestyle Factors, and Clinical Measurements our project aims to train and develop predictive models that can identify individuals at higher risk for CHD. The main motivation behind this project is to take preventive measures for potential CHD and also to treat it using patients' medical history. The dataset used for this project comprises information on over 4,000 patients and includes 15 attributes, each representing a potential risk factor for CHD. This dataset was thoroughly analyzed and preprossed and was trained and tested on 3 mainstream Machine Learning models for getting comprehensive outcomes.

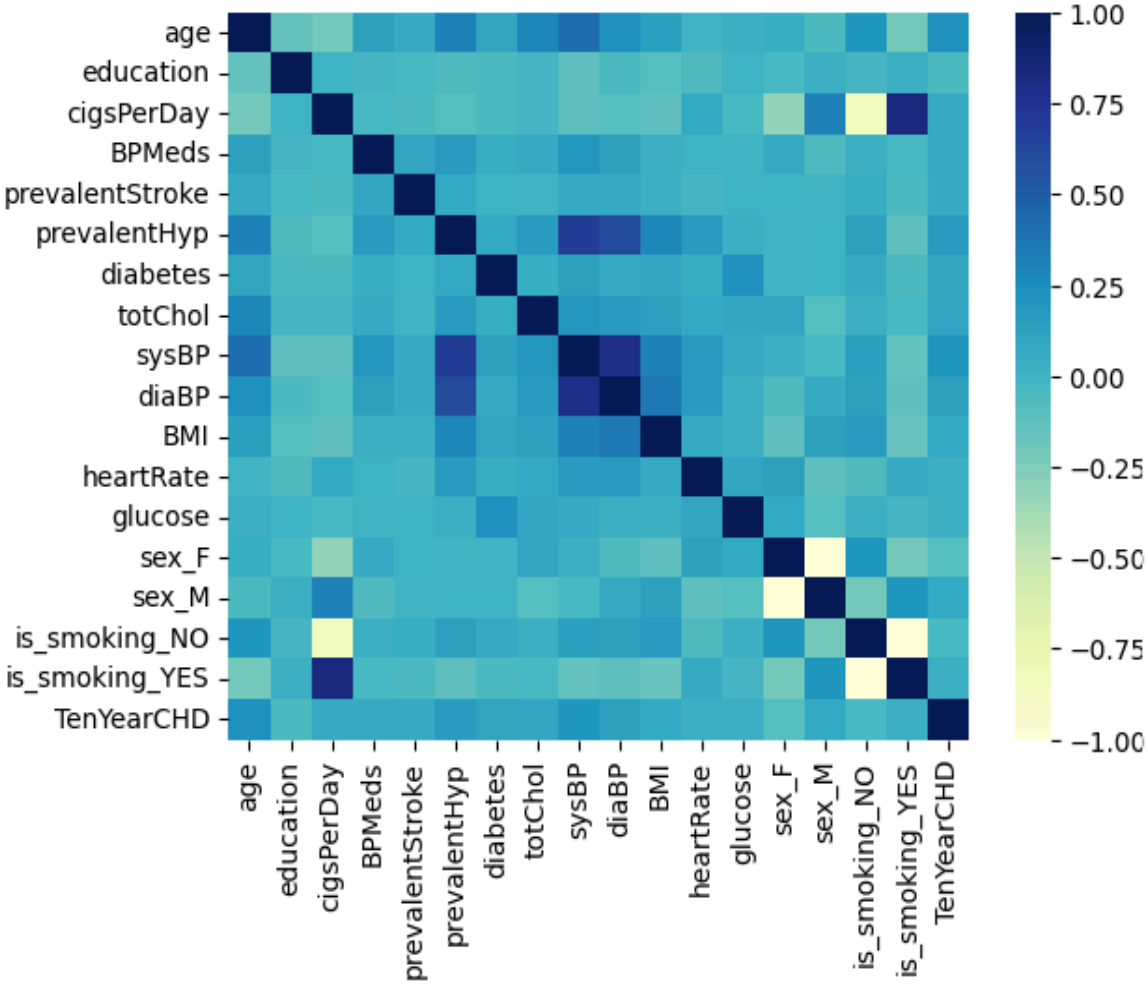
Dataset Description:

Source:

1) **Link:** https://www.kaggle.com/code/faezehbagheri/cardiovascular-risk
2) **Reference: cardiovascular-risk-factor-data. (n.d.). Kaggle.** https://www.kaggle.com/datasets/mamta1999/cardiovascular-risk-data/data

Data Description:

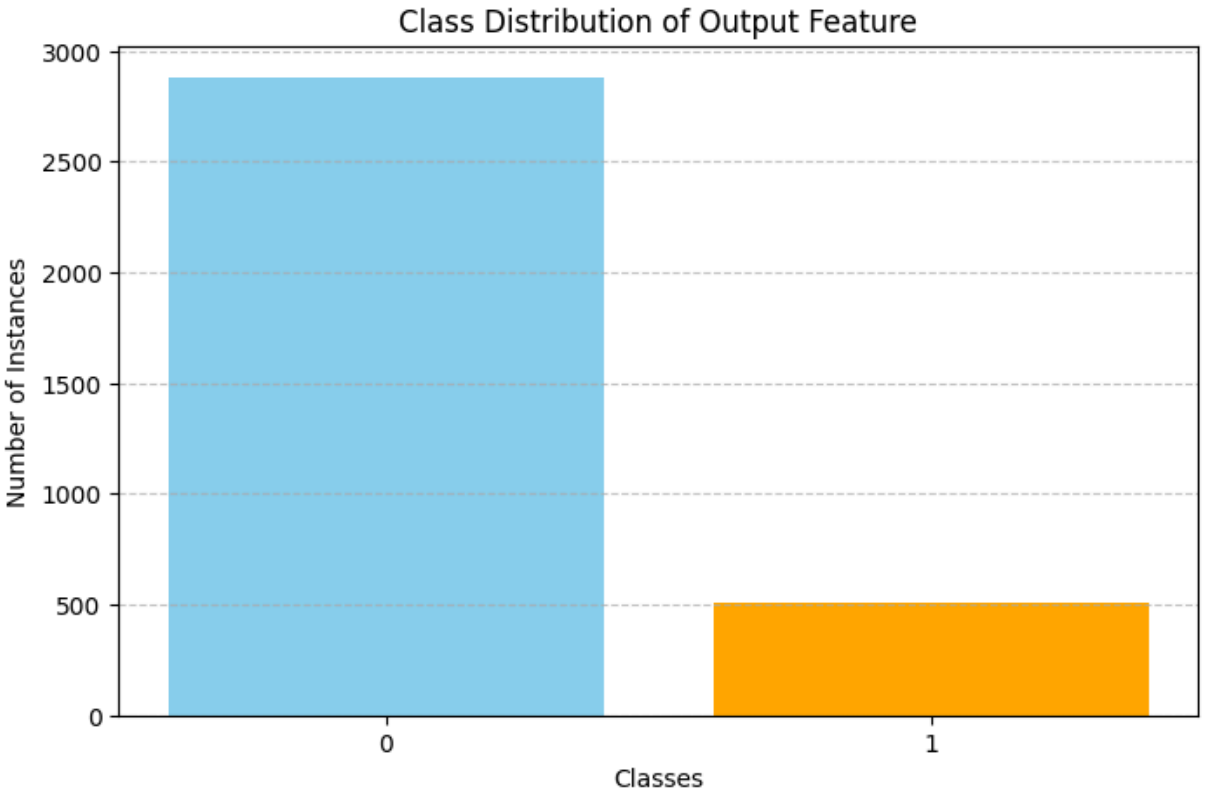
- **The dataset has 17 features:** id, age, education, sex, is_smoking, cigsPerDay, BPMeds, prevalentStroke, prevalentHyp, diabetes, totChol, sysBP, diaBP, BMI, heartRate, glucose, TenYearCHD.
- **Problem Type:** The target column is TnYearsCHD i.e. Ten-year coronary heart disease risk, which is binary(0 or 1). This suggests it is a classification problem.
- **Number of Data Points:** This dataset includes 3390 rows.
- **Kinds of Features:** This dataset contains two types of features: Quantitative Features and Categorical Features.
- **Correlation of all the Features:** The correlation among the numeric features is computed below and a heatmap is being plotted:



Imbalanced Dataset:

- The unique classes in the output feature i.e. TenYearCHD do not have an equal number of instances :
    - Class 0(No CHD): 2879 instances.
    - Class 1(CHD): 511 instances.

- Bar chart of N classes:



Dataset pre-processing:

- **Faults -** Several null values/missing values and categorical values are found in the respective dataset. They significantly impact the performance and accuracy of ML models as many algorithms cannot handle missing values and also ignoring missing values may lead to biased results and reduced accuracy. It also results in loss of information.
  Categorical values in our models are incompatible with algorithms as most machine learning models work with numerical data, not categorical strings or labels.
- **Solutions** - We didn't remove rows and columns with null values because of losing valuable information and some features with high percentages of missing values might contain critical information and dropping these features could harm the model's performance. In the process of the solution, firstly we flagged the missing values with a loop which helped us identify the missing values, and while imputed those missing values an iterative imputer (which itself runs a regression model). This is a multivariate imputer because it uses other column's information while populating the missing values.
  For categorical values, we have used one-hot encoding to create binary separate columns for each category. As this dataset only consists of nominal values that is why one-hot encoding is used.

Feature scaling: Here in this dataset feature scaling doesn't contain much variance and the value of the features are already normalized and categorical features are one-hot encoded which makes scaling irrelevant for them. That's why it is not required to use feature scaling.

Dataset splitting:

In the training set, 70% of the dataset is being used , here the model learns patterns and relationships. And 30% of the dataset is used for testing purposes. This splitting ensures the model doesn't memorize the data but generalises to unseen data. It also evaluates how accurately the model will perform on real-world data.
In this dataset random state 42 is used. This ensures that the same randomised data points are used every time the code has been executed.

Model Training & Testing

The dataset used in this project is designed for classification models; as it includes a target variable that indicates the presence or absence of cardiovascular disease, making it suitable for training models to classify individuals based on their risk. As mentioned previously there were 3 models used.
The models are as follows:

1. Logistic Regression
2. KNN (K-Nearest Neighbors)
3. Naive Bayes

Logistic Regression: To deploy this model we used the "liblinear" solver to train the model.
The Reason we chose it because:
It uses a coordinate descent algorithm for optimization. It is particularly suited for our dataset due to it being medium sized.

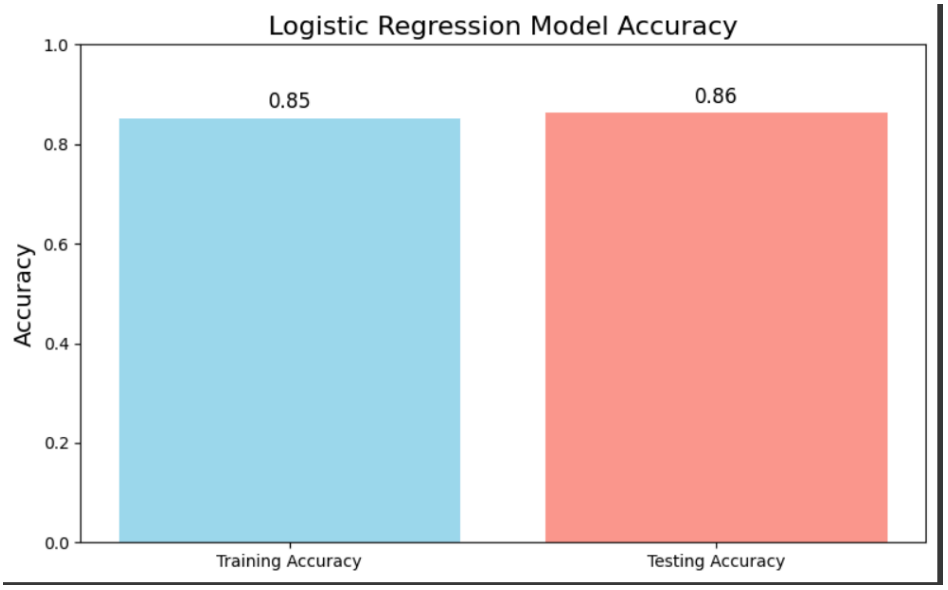
KNN: Since the dataset is catered for classification we chose this model as well. We used n_neibours = 5 as the number of datapoints.

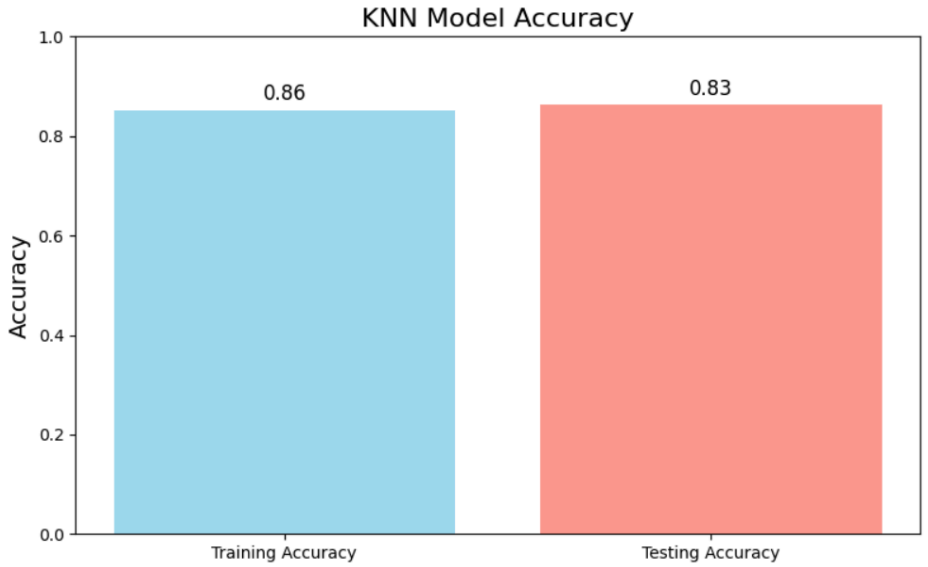
Naive Bayes: Finally for this model we used its “Gaussian variant” assuming it follows gaussian distribution.

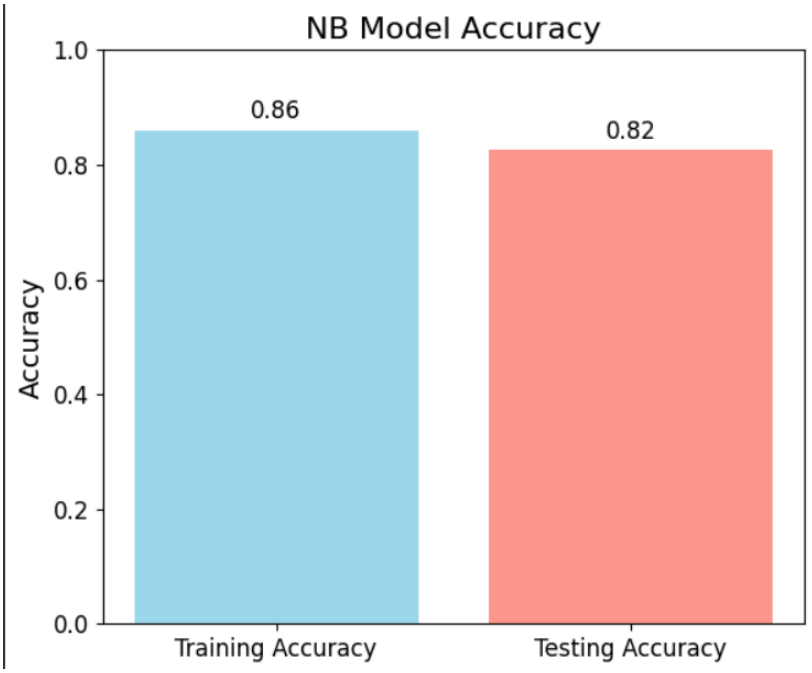
According to the dataset and its preprocessing; Logistic regression was the most accurate in both training and testing. The initial dataset and after it was being pre-processed; it fitted well with the logistic regression model as it assumes a linear relationship.

Model Selection / Comparison analysis

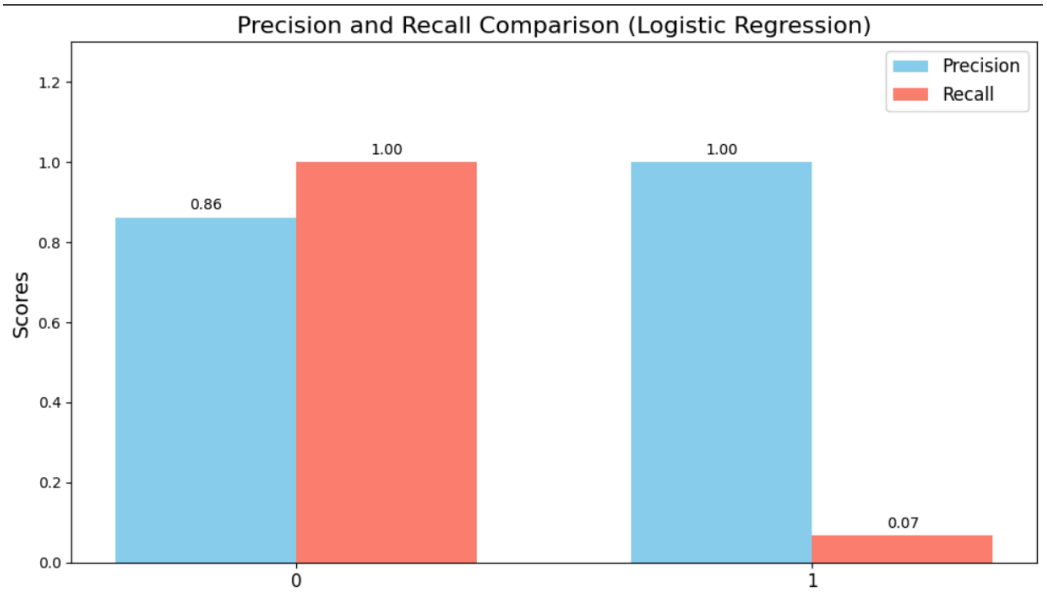
1. Bar charts for Showcasing Prediction Accuracy

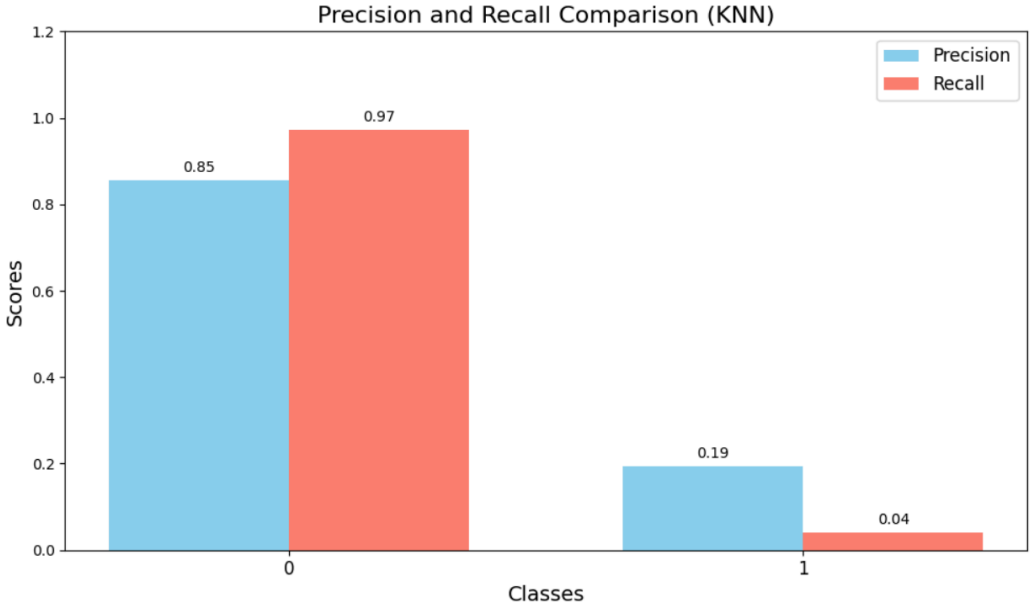


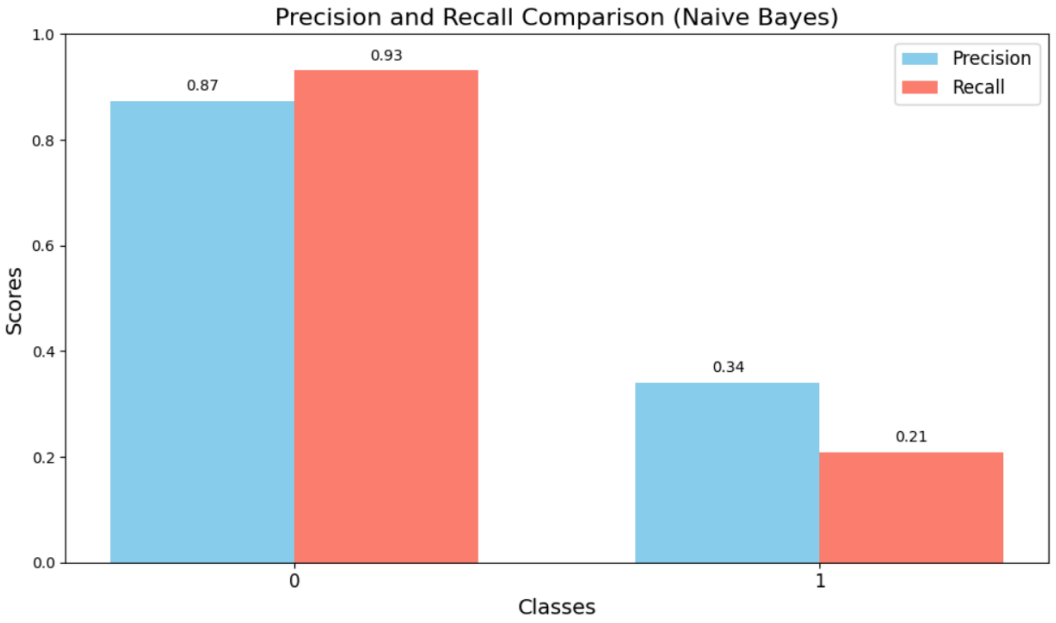

KNN Model Accuracy
Accuracy
1.0
0.8
0.6
0.4
0.2
0.0
0.86
0.83
Training Accuracy
Testing Accuracy


NB Model Accuracy
Accuracy
1.0
0.8
0.6
0.4
0.2
0.0
0.86
0.82
Training Accuracy
Testing Accuracy

2. Precision, Recall Comparison







3. Confusion Matrix

Logistic Regression:

```
Confusion Matrix:
[[868   0]
 [139  10]]
```

KNN:

```
Confusion Matrix:
[[843  25]
 [143   6]]
```

Naive Bayes:

```
Confusion Matrix:
[[808  60]
 [118  31]]
```

Conclusion

There is a huge room for improvement in incorporating machine learning techniques into medical sectors. Due reliability issue utilization of AI in this sector is still at its infancy. Thus much of our work was focused on pre-processing the dataset as much as possible. Three machine learning models Logistic Regression, K-Nearest Neighbors (KNN), and Naive Bayes were trained and evaluated on the dataset. Evaluation matrices such as accuracy, precision, recall, and confusion matrix were used to assess model's performance.

Among the models, Logistic Regression demonstrated the highest accuracy, likely due to the dataset's alignment with its linear assumptions. KNN provided a moderate performance, leveraging its ability to model non-linear patterns but being sensitive to noise and scaling. Naive Bayes exhibited the lowest accuracy, potentially due to the violation of its assumption of feature independence within the dataset.

These evaluations were necessary to highlight the models faced dealing with an imbalanced dataset, which is necessary to jot down the areas of improvement for making more accurate predictions.