Cardiovascular Disease Prediction

Machine Learning Development Assessment Report

Nanak Shrestha

November 16, 2024

Table of Contents

List of Figures

Introduction

Cardiovascular disease (CVD) encompasses a broad range of conditions affecting the heart and blood vessels, including coronary artery disease, stroke, and hypertension. It remains one of the leading causes of death worldwide, accounting for nearly 18 million deaths annually, according to the World Health Organization (WHO). This immense burden highlights the critical need for early detection and preventive interventions to reduce morbidity, mortality, and healthcare costs.

1. Problem Statement

The challenge lies in accurately predicting CVD risk in individuals, especially before symptoms manifest. Traditional risk assessment methods often rely on clinical expertise and extensive diagnostic tests, which can be resource-intensive and inaccessible in many settings. Machine learning (ML) presents an opportunity to address this challenge by leveraging health data to create predictive models, providing efficient and scalable risk assessments.

2. Objective

The primary objective of this project is to develop a machine learning model that predicts the risk of cardiovascular disease based on key health indicators. The model will use a dataset sourced from Kaggle, which includes indicators such as age, blood pressure, cholesterol levels, smoking status, physical activity, and diabetes.

3. Value of the Model

While the model cannot fully address the complex challenges of cardiovascular health, it serves as a valuable decision-support tool for healthcare professionals. By providing risk assessments, the model can help prioritize patients for early medical evaluation and guide preventive strategies.

4. Scope

The dataset used for this project includes several well-recognized predictors of CVD, such as age, blood pressure, cholesterol levels, and lifestyle factors. However, it does not encompass all risk factors identified by the WHO, such as air pollution or genetic predisposition. Therefore, the model's predictions should be interpreted with caution and seen as supportive insights rather than definitive diagnostics.

Data Exploration

The dataset contains 70,000 patient records with 11 features and a target variable. It includes medical and lifestyle information gathered during routine check-ups. The data combines objective measurements, examination results, and self-reported details, making it a valuable resource for predicting the risk of cardiovascular disease (CVD).

Feature	Data Type	Description
Age	Integer	Age of the patient in days.
Height	Integer	Height of the patient in centimeters (cm).
Weight	Float	Weight of the patient in kilograms (kg).
Gender	Categorical	Gender of the patient (1 = Female, 2 = Male).
Systolic Blood Pressure (ap_hi)	Integer	Upper value of the patient's blood pressure measured in mmHg.
Diastolic Blood Pressure (ap_lo)	Integer	Lower value of the patient's blood pressure measured in mmHg.
Cholesterol	Categorical	Cholesterol levels: 1 = Normal, 2 = Above Normal, 3 = Well Above Normal.
Glucose	Categorical	Glucose levels: 1 = Normal, 2 = Above Normal, 3 = Well Above Normal.
Smoking	Binary	Indicates whether the patient is a smoker (1 = Yes, 0 = No).
Alcohol Intake (alco)	Binary	Indicates alcohol consumption (1 = Yes, 0 = No).
Physical Activity (active)	Binary	Indicates whether the patient is physically active (1 = Active, 0 = Inactive).
Cardio	Binary	Presence or absence of cardiovascular disease (1 = Disease Present, 0 = None).

Table 1 Dataset Overview

id	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio
0	18393	2	168	62.0	110	80	1	1	0	0	1	0
1	20228	1	156	85.0	140	90	3	1	0	0	1	1
2	18857	1	165	64.0	130	70	3	1	0	0	0	1
3	17623	2	169	82.0	150	100	1	1	0	0	1	1
4	17474	1	156	56.0	100	60	1	1	0	0	0	0

Figure 1 Dataset Sample

Key Insights from Exploration

Category	Feature	Key Insights
Objective Features	Age	Recorded in days but may need transformation into years for better interpretability.
Examination Features	Blood pressure (ap_hi, ap_lo)	Critical for analysis but prone to outliers and data entry errors (e.g., maximum ap_hi = 16020).
Subjective Features	Smoking Alcohol Intake	Based on self-reported data, encoded in binary format for simplicity.
Target Variable	Cardio	Binary classification task to predict the presence of cardiovascular disease.

Figure 2 Key Insights from data

Understanding Individual Features

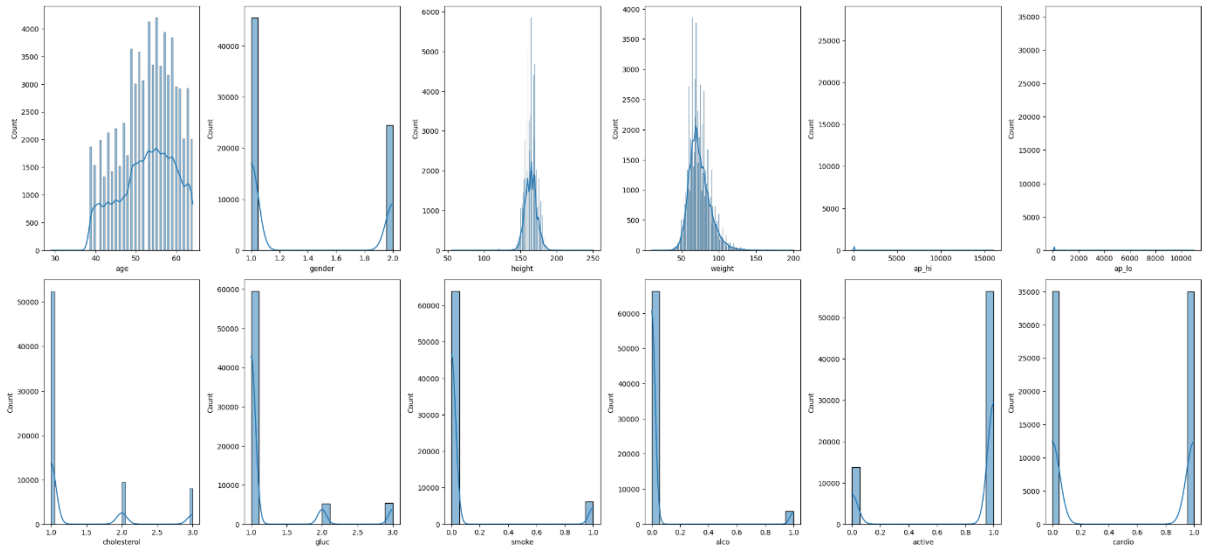


Figure 3 Features selection

Through univariate analysis, the distribution of each feature was examined. Numerical features like age, weight, and blood pressure showed varying levels of skewness and potential outliers, which required careful attention during modelling. Categorical features, such as gender, smoking status, alcohol consumption, and physical activity, highlighted the frequency of each category.

Target Variable Distribution (Cardiovascular Status)

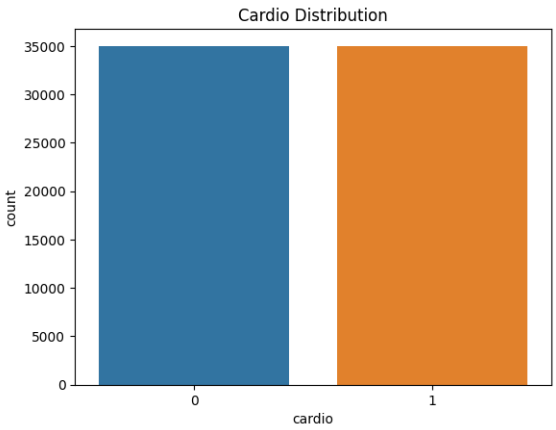


Figure 4 Cardio Distribution

The target variable, Cardio, reflects the presence or lack of cardiovascular disease. After analyzing the variable's distribution it is initially appearing balanced with almost equal counts of '0' (low risk) and '1' (high risk) cases, closer inspection reveals a subtle imbalance that requires addressing. This is very important because even a minor imbalance can lead to model bias toward the majority class during training and thus impair the model's ability to accurately identify high-risk individuals, which is a primary goal of this project.

Data Cleaning

Before proceeding with the development of predictive models, it is crucial to ensure that the dataset is clean and accurate. Several steps were taken to address issues such as missing values, inconsistencies, and outliers. These actions help ensure the dataset is in optimal condition for analysis and modelling.

Handling Missing value:

Despite the absence of apparent missing values in the dataset, a thorough inspection was carried out to verify the absence of null values. This step is crucial to ensure the data is thorough, without any missing pieces, necessary for building a trustworthy model.

Data Transformation

During the process of data exploration, it was found that certain features needed to be transformed to improve interpretability and align with machine learning models:

- **Age** was initially measured in days, but converting it into years would enhance its clarity and ease of interpretation.

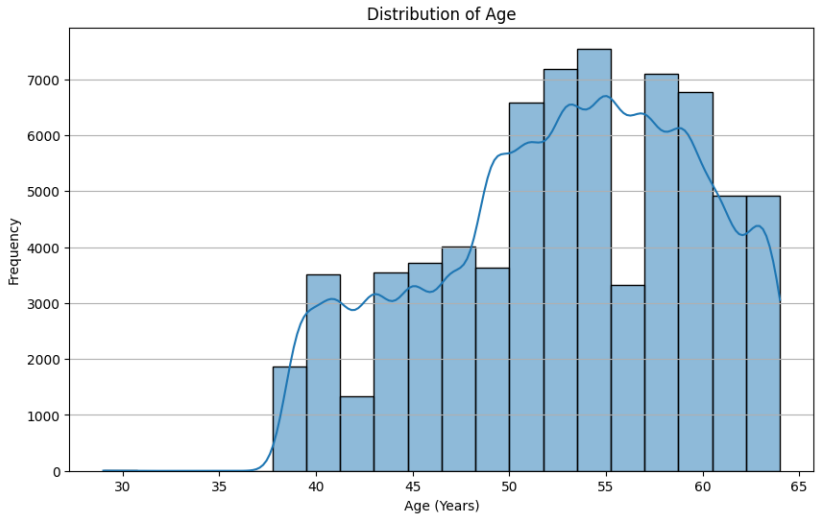


Figure 5 Age Distribution

Handling Outliers:

Outliers can have a major impact on the performance of machine learning models. In this dataset, some features needed to be corrected to prevent any distortion in the results.

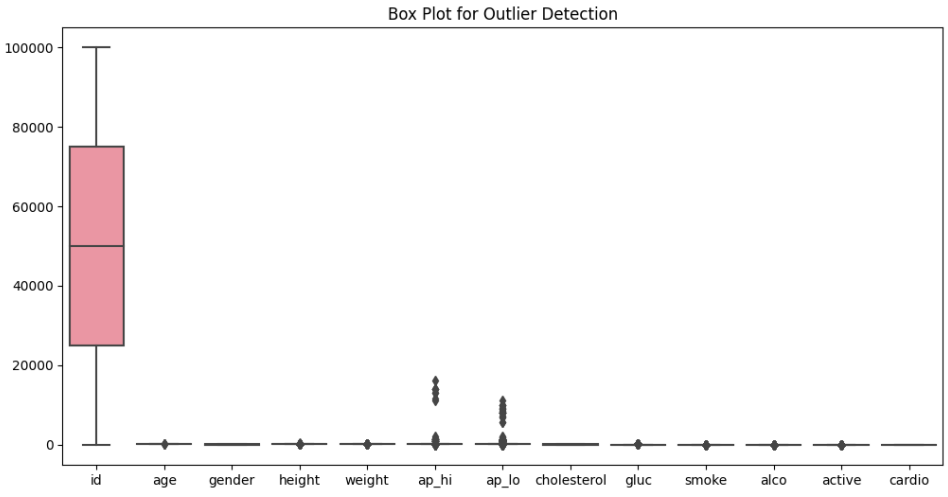


Figure 6 Before Handling Outliers

The above boxplots confirm our suspicion that there are outliers in the dataset. There appear to be several outliers in the systolic and diastolic variables (ap_hi and ap_lo). These outliers might be explained by human error when entering data into the .csv format. Our prediction model might benefit if these outliers were removed from the data.

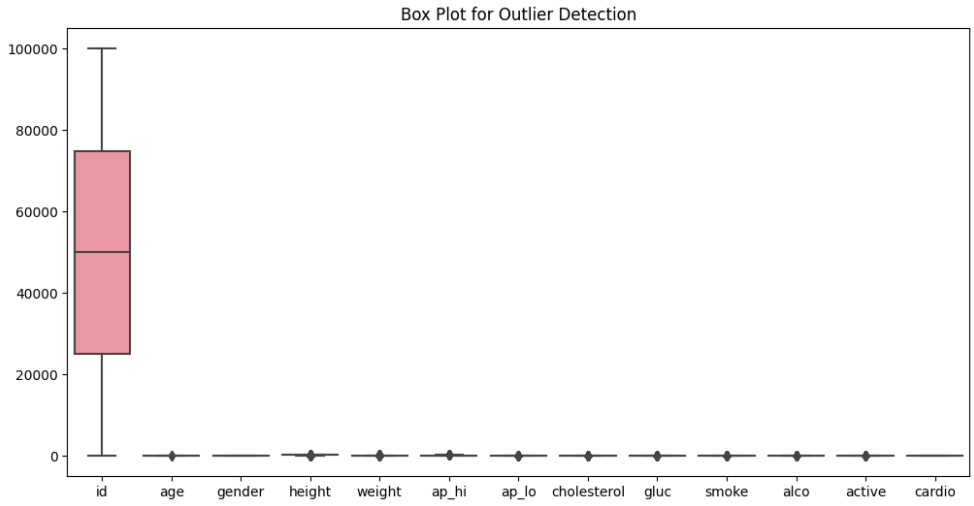


Figure 7 After Handling Outliers

The outliers in 'ap_hi' and 'ap_lo' have been greatly reduced or removed, which is a good indication that those issues have been addressed. As for the other features, there are only a few outliers, suggesting that the rest of the data is mostly clean.

Feature Engineering

The success of machine learning models heavily depends on the quality of the features used to train them.

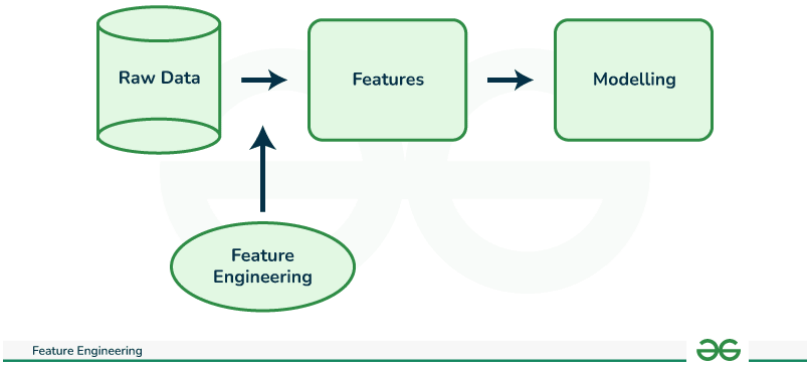


Figure 8 Feature Engineering

To enhance the capabilities, new features were created to capture key health indicators closely linked to cardiovascular disease. These additional features aim to provide a meaningful information and improve the model's ability to identify patterns.

- **Body Mass Index (BMI)**

  BMI, as a single measure, would not be expected to identify cardiovascular health or illness, the same is true for cholesterol, blood sugar, or blood pressure as a single measure. BMI may be more useful at prediction future rather than current health (Harvard Health Publishing, 2016).

  It is calculated as weight in kilograms divided by height in meters squared.

  Formula:

$$\boldsymbol{BMI = \frac{Weight\ (kg)}{(Height\ (m)^2)}}$$

  Rows with BMI values outside the reasonable range of 10 to 60 were removed to ensure the data contains meaningful and accurate information.

- **Blood Pressure Category (BP Category)**

  Blood pressure is a key indicator of cardiovascular health. Based on clinical guidelines, patients were categorized into different stages of hypertension using their systolic (ap_hi) and diastolic (ap_lo) blood pressure readings.

Normal	**Systolic**	**Diastolic**
Elevated	< 120	< 80
Hypertension Stage 1	130-139	80-89
Hypertension Stage 2	140-180	90-120
Hypertension Crisis	> 180	> 120

Table 2 Blood Pressure Category

The new features, BMI and Blood Pressure Category, provide clinically relevant information that enhances the dataset and supports a more comprehensive analysis of cardiovascular disease.

BMI Distribution by Blood Pressure Category and Gender

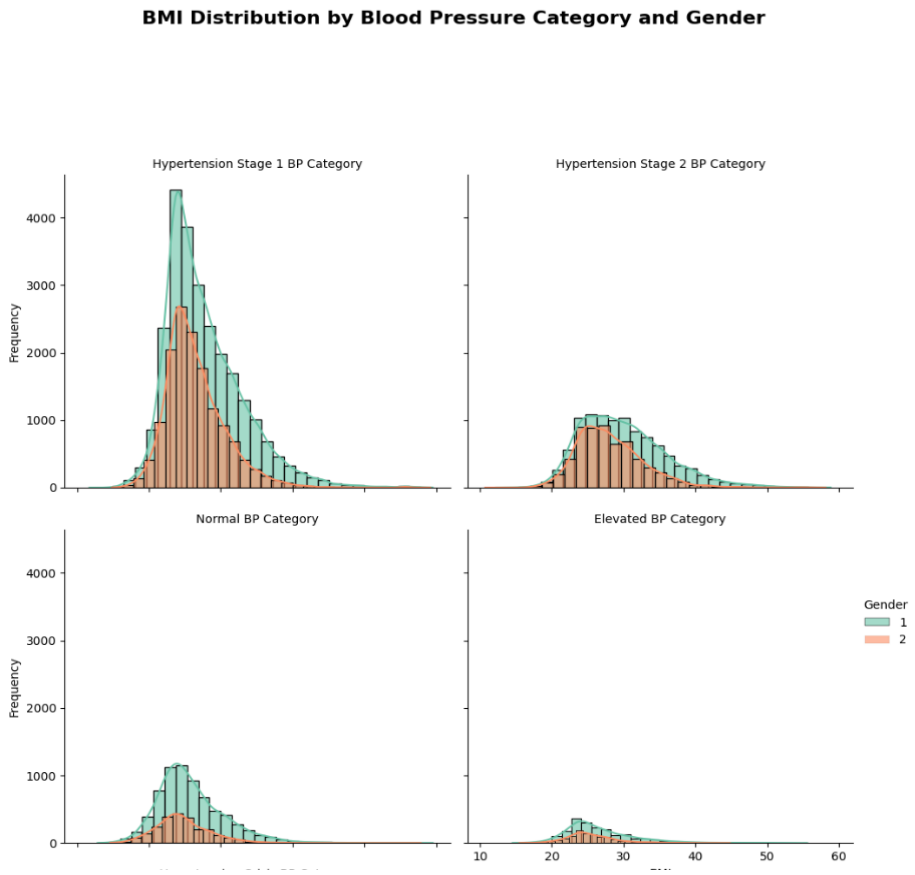


Figure 9 BMI Distribution By Blood Pressure And Gender

A clear relationship between blood pressure category and BMI, with higher blood pressure categories generally correlating with higher BMI. Gender differences also seem to influence the distribution of BMI within these categories, and while the 'Hypertensive Crisis' category suggests a stronger link between high blood pressure and elevated BMI, caution is needed due to the smaller sample size in this group.

Relation between height, weight, blood pressure category and gender

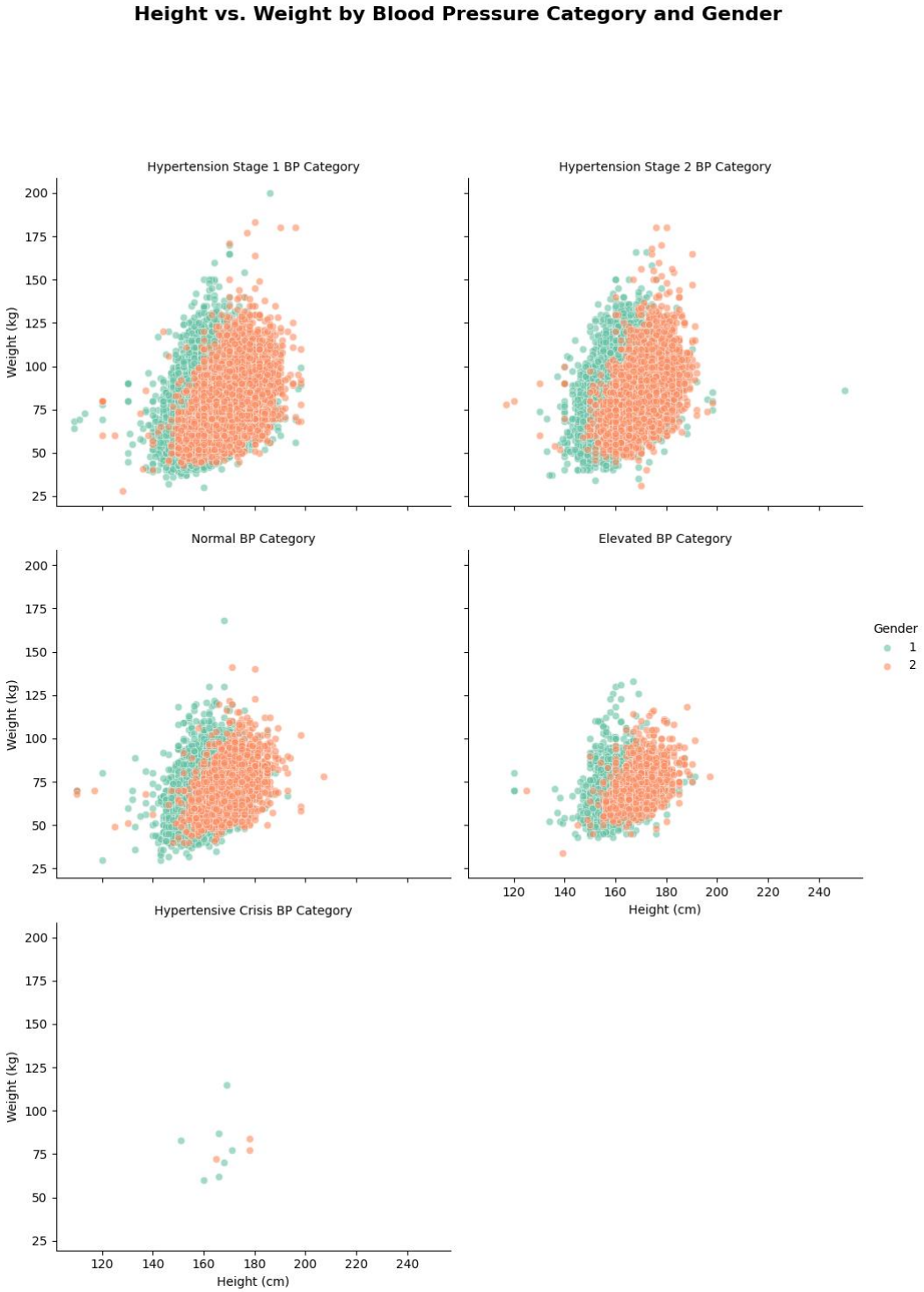


Figure 10 Relation between height, weight, blood pressure and gender

Taller individuals tend to weight more, showing a positive correlation between height and weight. Higher blood pressure categories (e.g Hypertension Stage 1 and Stage 2) shows a wider range of weights for a given height, with individuals generally weighting more than those in lower blood pressure groups. Male typically have higher weights than females at similar heights, particularly in Hypertension Stage 1. People in the 'Hypertensive Crisis' category appear to weight more than those in other blood pressure categories, though the small sample size limits the reliability of this observations. Height, weight, blood pressure category, and gender are important variables to consider in cardiovascular disease prediction models, as their interactions may significantly affect the model's accuracy.

Blood Pressure Category Distribution

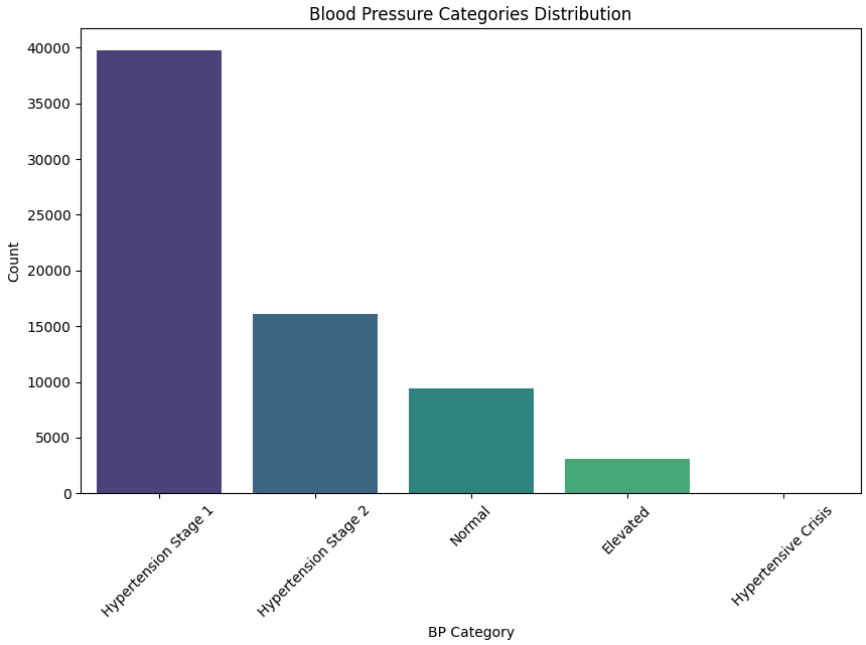


Figure 11 Blood Pressure Categories Distribution

Across all the individuals blood pressure categories Hypertension Stage 1 have the larger portion among blood pressure category followed by Hypertension Stage 2, Normal and Elevated while Hypertensive Crisis have fewest individuals.

Cholesterol and Glucose Levels Distribution

The group cholesterol and glucose levels are in three categories. Higher levels of cholesterol or glucose are much less common.

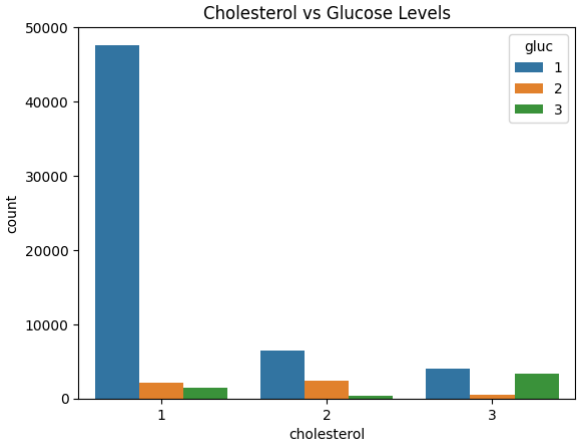


Figure 12 Cholesterol and Glucose Levels

Cardiovascular Risk and Age

Age is a well-known traditional risk factor, which is generally considered to be non-modifiable for obvious reasons (Nair, 2012) which plays a vital role as a factor in health also in cardiovascular disease, as the data shows the individuals with cardiovascular conditions are generally older than those without. The risk of cardiovascular issues may increase with age.

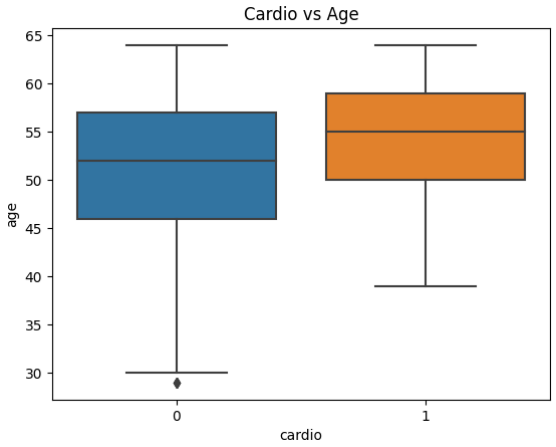


Figure 13 Disease Risk with Age

Impact of Activity on Disease

Both active and cardio are represented as binary values, with active being (1) and cardio being (0). Active indicates if someone is engaging in physical activity (1) or not (0), while cardio indicates the presence (1) or absence (0) of cardiovascular disease. For both activity levels number of individuals with and without CVD are relatively close. However, who do not have cardiovascular disease has a large proportion of group which is physically active.

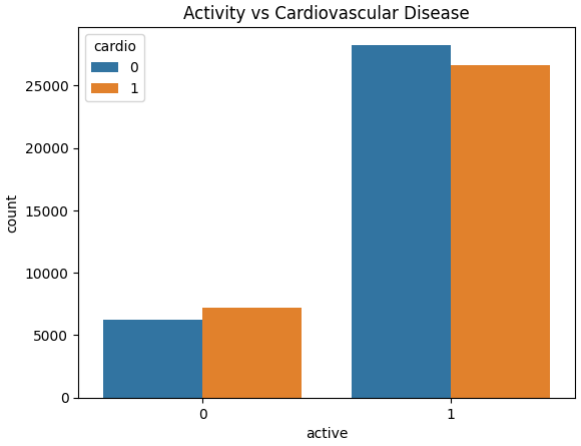


Figure 14 Impact of activity on disease

Correlation Matrix

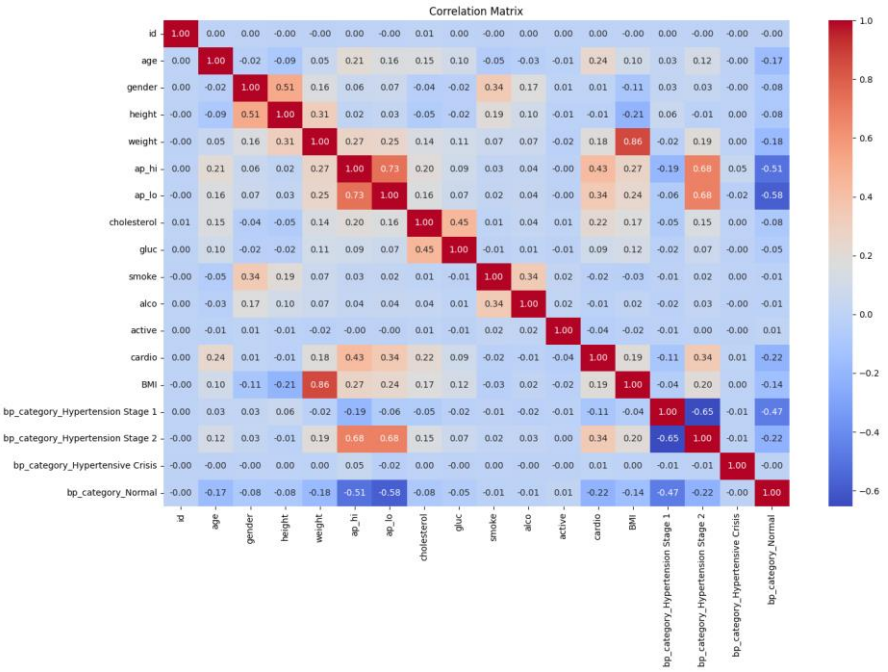


Figure 15 Correlation Matrix

Systolic Blood pressure and diastolic blood pressure are strongly positively correlated, with a correlation which indicates as one increases or decreases, the other tends to do the same. Likewise the in Age and cardio (presence of CVD) shows that older individuals are more likely to have CVD.

Model Development and Evaluation

Data pre-processing

Here it starts by removing the id column, as it does not contribute to model training. The bp_category column, which contains categorical data, is then converted into numerical values using one-hot encoding. This process creates a separate binary column for each category, making it easier for model to interpret the data. To prevent multicollinearity, one category is dropped.

The dataset is divided into features and the target variable. The target variable is the cardio column, which indicates whether the individual has cardiovascular disease or not. The feature include all other columns.

Data Splitting

Once the features and target are separated, the data is split into training and testing sets, data is divided into 80% for training and 20% for testing, ensuring that the model has enough data to learn from while still being evaluated on unseen data. After that features are scaled, which standardizes the data which is the crucial step for models that are sensitive to the scale of the data.

Model Selection

Different types of machine learning models are chosen for evaluation –

1. Logistic Regression
2. Random Forest
3. Support Vector Machine (SVC)
4. Naïve Bayes
5. Gradient Boosting
6. LightGBM
7. CatBoost
8. Decision Tree
9. K-Nearest Neighbors
10. XGBoost

These models were chosen because they are well-suited to managing various types of data during categorization tasks.

- Random Forest and Gradient Boosting use decision trees to detect complicated pattern in data.
- Logistic regression is a simple yet effective model for binary classification problems.

- XGBoost and LightGBM are advanced gradient boosting frameworks that are fast and highly efficient, making them popular choices.
- CatBoost is also another gradient boosting model that works particularly well with categorical data.

Moreover the models like SVM, KNN, Native Bayes and Decision Trees provide additional approaches which offers variety and helps to compare the results across different techniques.

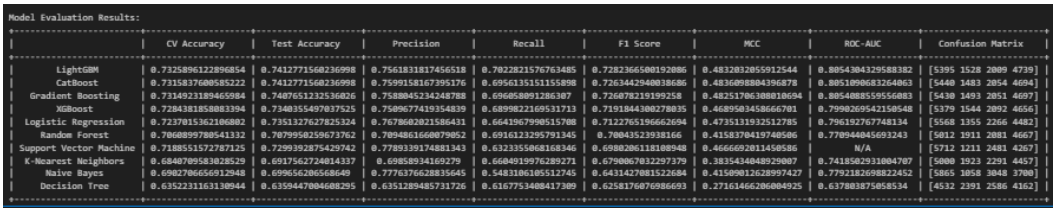
Cross-Validation

One method for assessing a model's performance on unseen data is cross validation. It entails splitting the given data into several folds or subsets, training the model on the remaining folds, and using one of these folds as a validation set.

So, to evaluate the performance of each model robustly, Stratified K-Fold Cross-Validation is used which splits the data into 5 folds. This technique helps in reducing bias during training because each model trained and test on different folds of the data.

Model Evaluation Metrics

A summary of model performance on the dataset, via cross validation.



Model Evaluation Results:

	CV Accuracy	Test Accuracy	Precision	Recall	F1 Score	MCC	ROC-AUC	Confusion Matrix
LightGBM	0.7325896122896854	0.7412771560236998	0.7561831817456518	0.7022821576763485	0.7282366500192086	0.4832032055912544	0.8054304329588382	[5395 1528 2009 4739]
CatBoost	0.7315837600585222	0.7412771560236998	0.7599158167395176	0.6956135151155898	0.7263442940038686	0.4836088804396878	0.8051890683264863	[5440 1483 2054 4694]
Gradient Boosting	0.7314923189465984	0.7407651232536026	0.7588045234248788	0.6960580091286307	0.7260782319199258	0.4825170630010694	0.8054085559550883	[5430 1493 2051 4697]
XGBoost	0.7284381858083394	0.7340355497037525	0.7509677419354839	0.6899822169531713	0.7191844300278035	0.4689503458666701	0.7990269542150548	[5379 1544 2092 4656]
Logistic Regression	0.7237015362106802	0.7351327627825324	0.7678602021586431	0.6641967990515708	0.7122765196662694	0.4735131932512785	0.7961927677481134	[5568 1355 2266 4482]
Random Forest	0.7060899780541332	0.7079950259673762	0.7094861660079052	0.6916123295791345	0.70043523938166	0.4158370419740506	0.7709440456932343	[5012 1911 2081 4667]
Support Vector Machine	0.7188551572787125	0.7299392875429742	0.7789339174881343	0.6323355068168346	0.6980206118108948	0.4666692011450586	N/A	[5712 1211 2481 4267]
K-Nearest Neighbors	0.6840709583028529	0.691756272401433 7	0.69858934169279	0.6604919976289271	0.6790067032297379	0.3835434048929007	0.7418502931004707	[5000 1923 2291 4457]
Naive Bayes	0.6902706656912948	0.699656206568649	0.7776376628835645	0.5483106105512745	0.6431427081522684	0.4150901262899742 7	0.7792182698822452	[5865 1058 3048 3700]
Decision Tree	0.6352231163130944	0.6359447004608295	0.6351289485731726	0.6167753408417309	0.6258176076986693	0.2716146620600492 5	0.6378838750585 34	[4532 2391 2586 4162]

Figure 16 Model Evaluation Metrics

The models such as Booting delivered better performance compared to other algorithms. Simpler models like Decision Tree and Naïve Bayes performed worse, with lower recall and F1 scores, notifying a clear poor performance.

Performance of the Model

The closer the curve is top-left corner, the better the model's performance represents a perfect classifer (100% TPR and 0% FPR). A model which performs worse have a curve below the diagonal dashed line. AUC represents the probability that the model, AUC 1.0 represents a perfect classifer while 0.5 represents a random classifer. The models LightGBM, CatBoost, and Gradient Boosting exhibit the highest AUC values in the given plot, therefore showing better classification performance compared to the other models presented here.

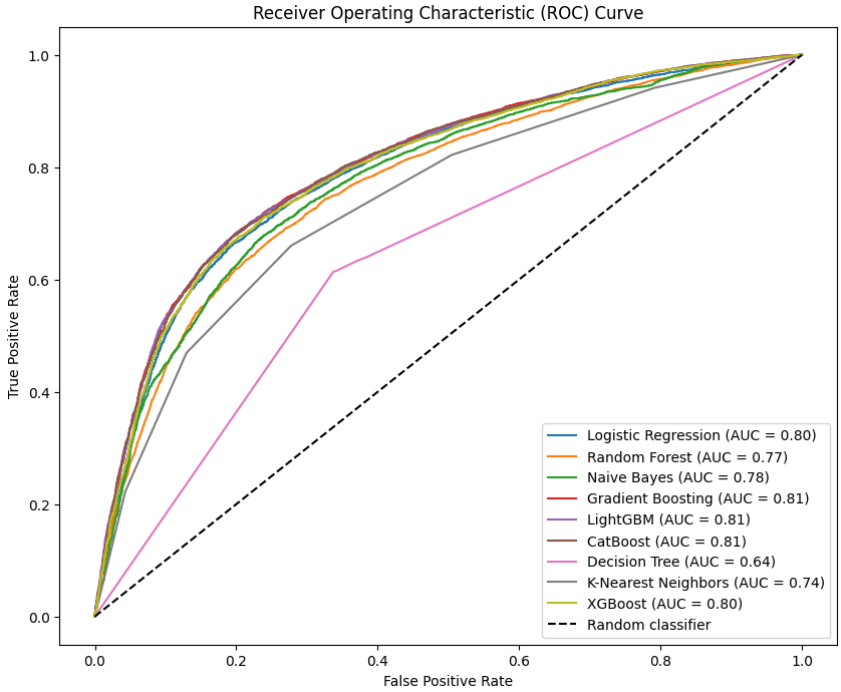


Figure 17 ROC Curve for model performance

Hyperparameter Tuning

A process for finding the best combination of hyperparameters for a model. In this case, Optuna hyperparameter optimization is used to find the best values for parameters of models. After optimizing each model using Optuna, the best hyperparameters and the accuracy score are recorded. So the final results show the optimal hyperparameters for each model and accuracy scores, which is used to compare the models performance after tuning.

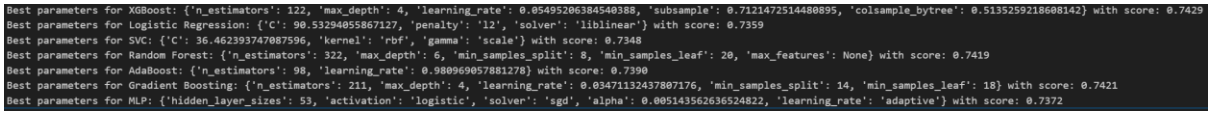


Figure 18 Best parameters

After identifying the best parameters for each model, it's trained and evaluated them.

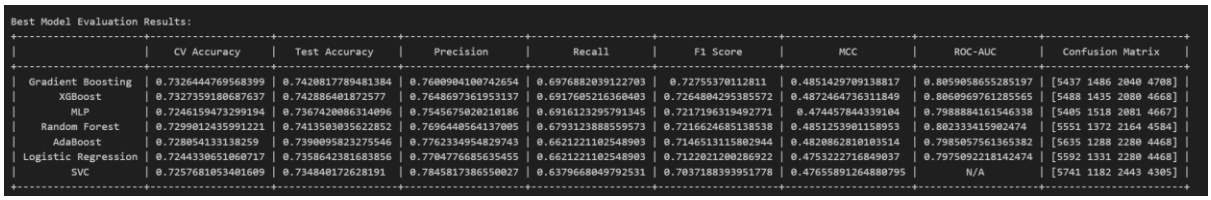


Figure 19 Best Model Evaluation Results

Ensemble Model with Voting Classifier

Ensemble modelling is a process where a multiple diverse base models are used to predict an outcome to reduce the error off the prediction or to obtain better predictive performance than could be obtained from any of the constituent learning algorithms alone. So, Here a Voting classifier which combines the predictions of several base classifiers (weak learners) to make a final better prediction.

Base Estimators-

- Gradient Boosting Classifier
- XGBoost
- Random Forest Classifier
- AdaBoost Classifier

These classifiers were chosen to provide better results on the classification task, with each model contributing its own strengths.

In Voting Classifier a soft voting method were used, which balances the predicted probabilities and outputs the class label with the highest average probabilities and also models outputs the probability that can be combined to make a better overall prediction accuracy. The Voting Classifer achieved an accuracy on **73.59%** on the test set.

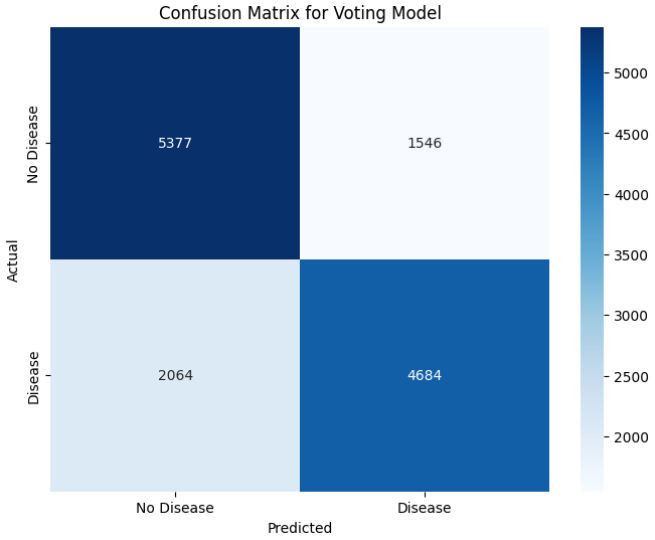


Figure 20 Confusion Matrix for Voting Model

The confusion matrix shows a reasonable balance between True Positive (4684) and False Positive (1546), with fewer False Negatives (2062), indicating that the model is performing adequately in terms of identifying positive cases (Disease).

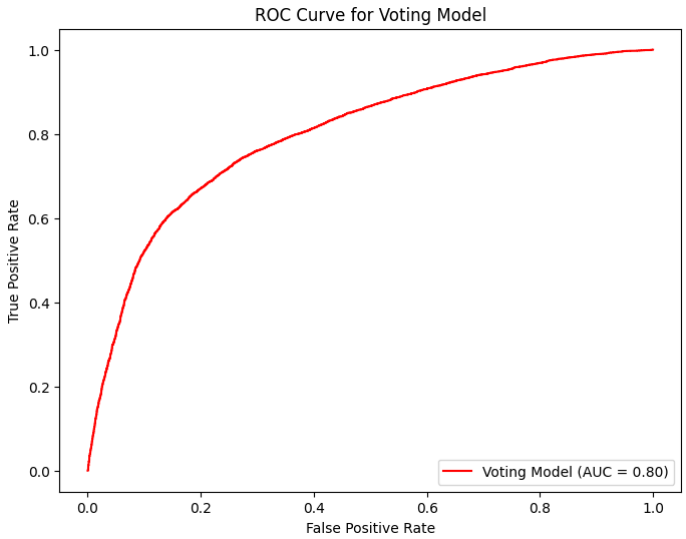


Figure 21 ROC Curve for Voting Model

Voting Classifier achieved an AUC of 0.80 on ROC, which indicates a good performance between 'Disease' and 'No Disease' classes, but still there is not a better improvement in performance as compared to other models which are tested before. So, although Voting Classifier performs well, but does not substantially outperform other models.

Lime Model Interpretation

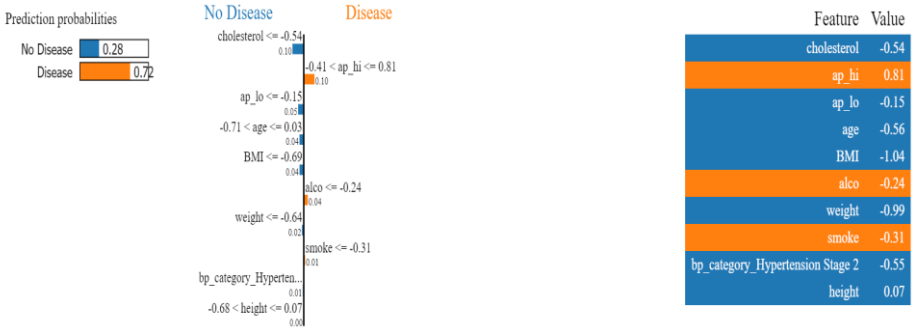


Figure 22 LIME model one of the insights

The LIME model interpretations helps to understand how the certain conditions factors influence the prediction of whether an individual has disease or not. For each case, the model shows the features that had the greatest impact on its decision. While the most instances are systolic blood pressure and cholesterol levels are key factors. When the systolic blood pressure is low, the model tends to predict "No Disease" and vice versa. Other factors like age, BMI play a smaller role, but they still influence the outcome. The Lime model is amazing that it's even provides probabilities, showing how confident is prediction. 72% probability of having the disease suggest the model is quietly confident in that prediction but still need a lot of improvements. So, the model's prediction are largely shaped by cholesterol and blood pressure, with other factors showing a lesser, but still notable.

Deployment Strategy

To deploy the model, the trained Voting Classifier was saved. For deployment, Streamlit is used, a powerful, and easy-to-use framework for creating interactive web applications for machine learning models. In streamlit the saved model is loaded and used to make prediction based on user inputs and the model provides real-time predictions regarding disease.

Conclusion

In short, the developed predictive model shows a strong ability to predict an individual's chance of having a cardiovascular disease based on health key factor. By using a voting classifier that combines multiple algorithms, the model achieved a robust performance indicating a strong balance to identifying individuals having disease or not disease. And, the LIME model allows for transparency in the mode's decision where the transparency build trust in the model. Moreover, the model was successfully deployed which access interaction with the model.

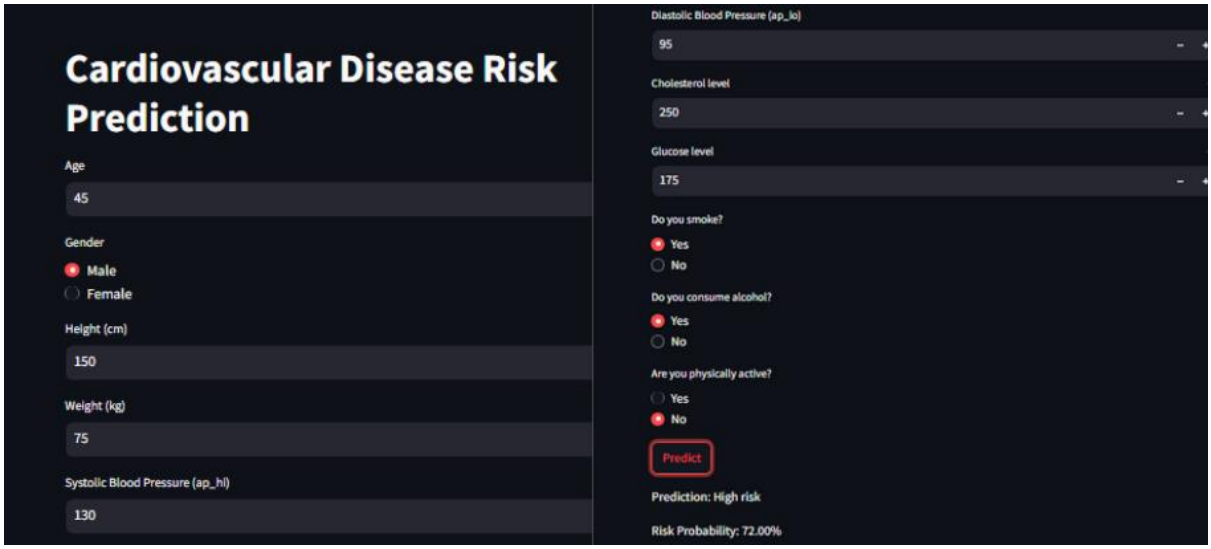


Figure 23 Deployment

Further Improvements

- Additional features, such as lifestyle factors could be enrich the model's predictive power.
- Alternative ensemble techniques could improve performance.
- For advanced interpretability techniques such as SHAP could help deeper insights in model decision.
- Correlation only measure strength and direction of linear relationship, therefore two variables may perform a strong non-linear relationship but show a weak or no correlation. So, other techniques can implement to capture non-linear dependencies between features.
- To validate the differences between genders within each blood pressure category using statistical tests like t-tests or ANOVA may help to determine deeper insights.

Reference

Harvard Health Publishing. (2016, March 30). *How useful is the body mass index (BMI)?* Harvard Health Blog. Retrieved from https://www.health.harvard.edu/blog/how-useful-is-the-body-mass-index-bmi-201603309339

Nair, R. (2012). *Ageing and cardiovascular health*. *National Center for Biotechnology Information*. https://pmc.ncbi.nlm.nih.gov/articles/PMC3297980/