Model for Cardiovascular Disease Prediction

1. Introduction

Cardiovascular disease (CVD) remains a leading cause of mortality globally, making early risk identification a critical objective for public health intervention. Utilizing a comprehensive dataset of approximately 70,000 anonymized patient records, this study aims to construct and evaluate machine learning models capable of effectively predicting individual disease probability. The dataset encompasses a multi-dimensional array of health indicators, integrating objective clinical measurements—such as blood pressure, cholesterol levels, and Body Mass Index (BMI)—with subjective lifestyle features, including smoking habits, alcohol consumption, and physical activity frequency. Specifically, to enhance clinical interpretability, raw age data was processed to investigate the correlation between demographic progression and disease risk.

To ensure predictive accuracy and stability, this study implemented a systematic modeling pipeline progressing from linear baselines to advanced ensemble methods. We initially employed Logistic Regression not only as a baseline classifier but also to perform a linearity check, assessing whether the relationship between risk factors and disease presence follows a simple linear pattern. To capture more complex, non-linear interactions within the data, the study subsequently incorporated Random Forest and Gradient Boosting algorithms. Furthermore, addressing the cross-sectional nature of the dataset, we conducted an aged-based trend analysis. By grouping individuals by age, this approach maps how cardiovascular risk evolves over time, revealing latent patterns that static analysis might overlook.

Through a systematic comparison of algorithms and feature importance analysis, this study identified the optimal predictive strategies. Experimental results demonstrate that while the XGBoost model achieved robust performance (Average Accuracy: 73.59%), its advantage over the Logistic Regression baseline (Testing Accuracy: 73.39%) was marginal. This comparable performance provides a critical insight: the primary cardiovascular risk factors, particularly Systolic Blood Pressure and Age, exhibit strong linear predictive power. Feature importance analysis confirmed these findings, highlighting Age, ap_hi, and BMI as the dominant predictors. These results validate that for this specific dataset, well-calibrated linear models are highly effective, offering a balance between predictive accuracy and clinical interpretability without the need for excessive model complexity.

2. Data

2.1.Data features

Age | Objective Feature | age | int (days)
Height | Objective Feature | height | int (cm) |
Weight | Objective Feature | weight | float (kg) |
Gender | Objective Feature | gender | categorical code |
Systolic blood pressure | Examination Feature | ap_hi | int |
Diastolic blood pressure | Examination Feature | ap_lo | int |
Cholesterol | Examination Feature | cholesterol | 1: normal, 2: above normal, 3: well above normal |

Glucose | Examination Feature | gluc | 1: normal, 2: above normal, 3: well above normal |
Smoking | Subjective Feature | smoke | binary |
Alcohol intake | Subjective Feature | alco | binary |
Physical activity | Subjective Feature | active | binary |
Presence or absence of cardiovascular disease | Target Variable | cardio | binary |

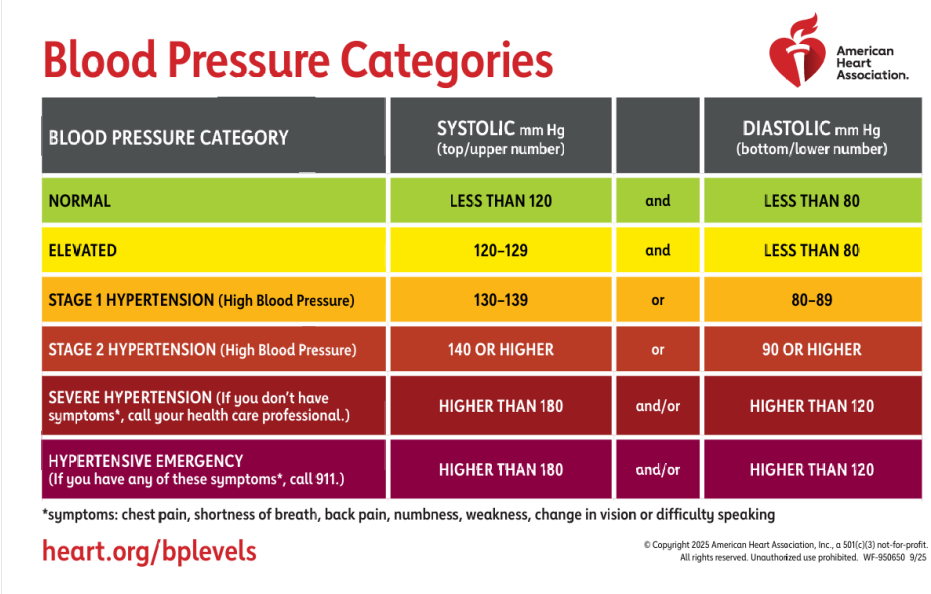
2.2.Data preprocessing and cleaning

After removing duplicate values and outliers with BMI greater than 60 or less than 15, the available data amounts to 68,602.

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 68602 entries, 0 to 69999
Data columns (total 16 columns):
 #   Column          Non-Null Count  Dtype  
---  ------          --------------  -----  
 0   age             68602 non-null  float64
 1   female          68602 non-null  int64  
 2   male            68602 non-null  int64  
 3   height          68602 non-null  int64  
 4   weight          68602 non-null  float64
 5   bmi             68602 non-null  float64
 6   ap_hi           68602 non-null  int64  
 7   ap_lo           68602 non-null  int64  
 8   bp_cat          68602 non-null  object 
 9   cholesterol     68602 non-null  int64  
 10  gluc            68602 non-null  int64  
 11  smoke           68602 non-null  int64  
 12  alco            68602 non-null  int64  
 13  active          68602 non-null  int64  
 14  cardio          68602 non-null  int64  
 15  cardio_percent  68602 non-null  int64  
dtypes: float64(3), int64(12), object(1)
memory usage: 8.9+ MB
```

Based on the data from the [American Heart Association | To be a relentless force for a world of longer, healthier lives](), the stages of blood pressure are classified by Ferri.



Blood Pressure Categories

American Heart Association.

BLOOD PRESSURE CATEGORY	SYSTOLIC mm Hg (top/upper number)		DIASTOLIC mm Hg (bottom/lower number)
NORMAL	LESS THAN 120	and	LESS THAN 80
ELEVATED	120–129	and	LESS THAN 80
STAGE 1 HYPERTENSION (High Blood Pressure)	130–139	or	80–89
STAGE 2 HYPERTENSION (High Blood Pressure)	140 OR HIGHER	or	90 OR HIGHER
SEVERE HYPERTENSION (If you don't have symptoms*, call your health care professional.)	HIGHER THAN 180	and/or	HIGHER THAN 120
HYPERTENSIVE EMERGENCY (If you have any of these symptoms*, call 911.)	HIGHER THAN 180	and/or	HIGHER THAN 120

*symptoms: chest pain, shortness of breath, back pain, numbness, weakness, change in vision or difficulty speaking

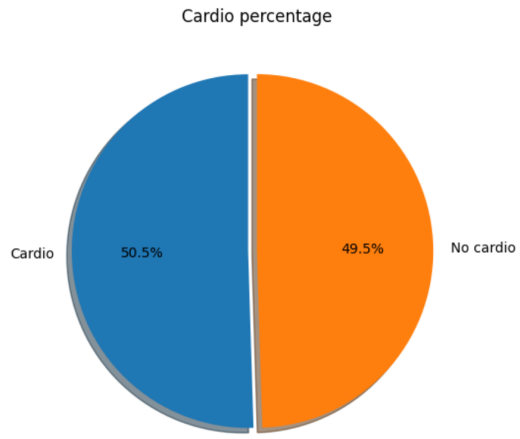
heart.org/bplevels

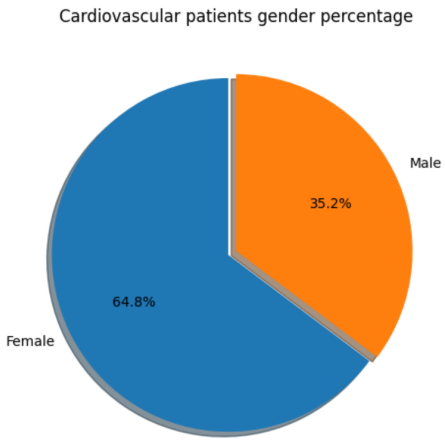
Remove the outliers from the blood pressure variable

ap_hi >220 or ap_lo >180 or ap_hi<40 or ap_lo<40

3. Data analysis

3.1.Analysis of the patient population



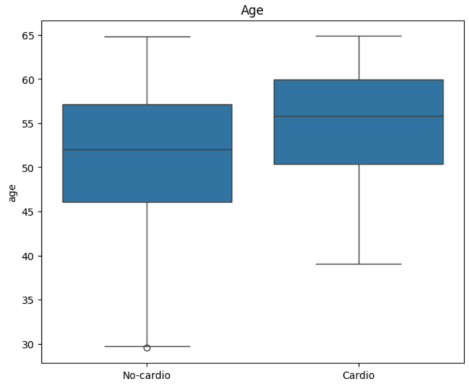


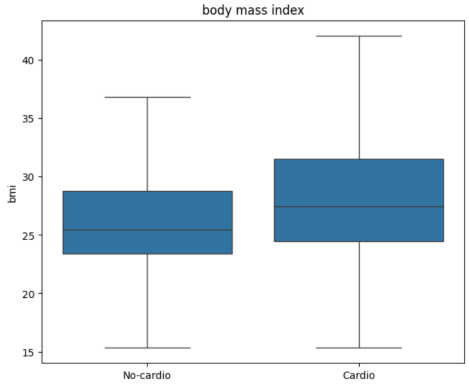
Based on the analysis of the charts, we can draw the following conclusions:

First, the data reveals that 50.5% of the sample population has cardiovascular diseases, while 49.5% does not. This indicates that cardiovascular diseases affect a significant portion of the population, with nearly half of the individuals in the sample suffering from the condition. This high prevalence suggests the importance of addressing cardiovascular health in the studied population and highlights the need for effective prevention and intervention measures.

Second, regarding the gender distribution of individuals with cardiovascular diseases, the charts show that 64.8% of the affected individuals are female, while 35.2% are male. This significant difference in the gender ratio indicates that females are more likely to suffer from cardiovascular diseases compared to males. The higher proportion of women with the condition suggests that gender-specific factors, such as hormonal differences or lifestyle habits, may play a role in the increased susceptibility of women to cardiovascular diseases.

3.2.The conditions of heart patients and non-heart patients





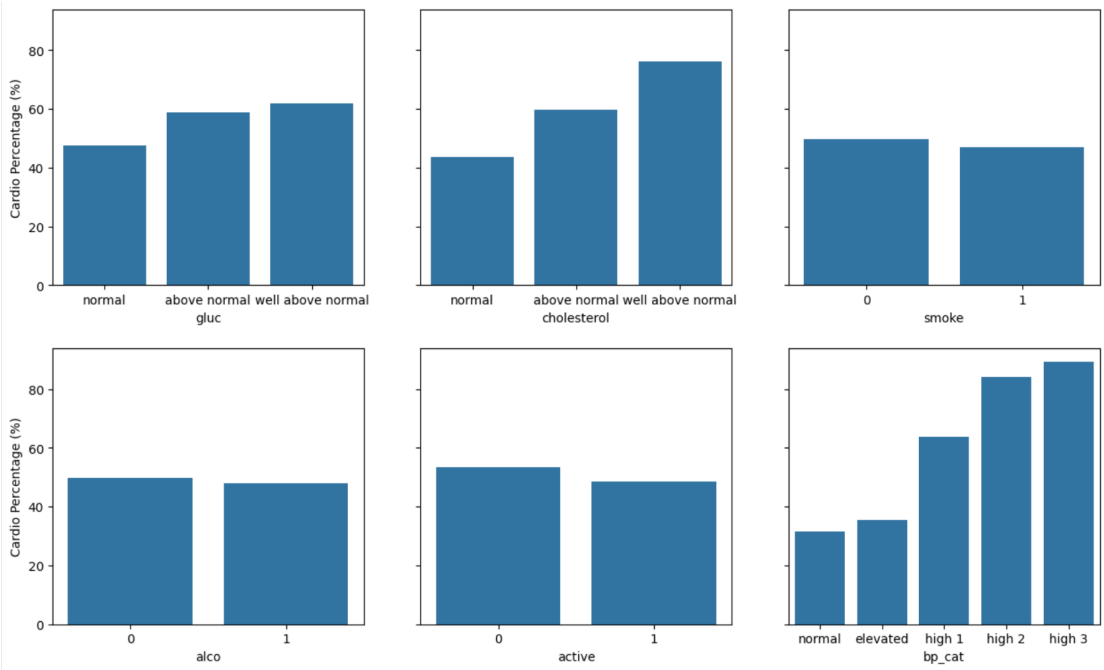
Based on the analysis of the boxplots, we can draw the following conclusions:

First, the age distribution of people with and without cardiovascular diseases shows a clear relationship between age and the presence of cardiovascular conditions. Individuals with cardiovascular diseases tend to be older, with the median age for

those with the disease being higher compared to those without. This suggests that as people age, they are more likely to develop cardiovascular diseases, highlighting the importance of age as a risk factor for these conditions.

Second, the body mass index (BMI) distribution also shows a clear relationship with cardiovascular diseases. People with higher BMIs are more likely to have cardiovascular diseases, as indicated by the higher median BMI in the group with cardiovascular diseases compared to those without. This reinforces the link between obesity and the risk of cardiovascular issues, emphasizing the need for weight management as part of cardiovascular disease prevention strategies.

3.3.The relationships among the variables in the dataset



Based on the analysis of the bar charts, we can conclude the following:

First, regarding glucose levels, the percentage of people with cardiovascular diseases is highest among those with well above normal glucose levels, followed by those with above normal glucose levels, and lowest for those with normal glucose levels. This indicates a clear relationship between elevated glucose levels and the increased risk of cardiovascular diseases.

Second, for cholesterol levels, a similar trend is observed. The percentage of people with cardiovascular diseases is highest among those with well above normal cholesterol levels, followed by those with above normal levels, and lowest among those with normal cholesterol levels. Elevated cholesterol levels appear to be a significant risk factor for cardiovascular diseases.

Third, the smoking status chart shows no significant difference between smokers and non-smokers, as the percentage of cardiovascular patients is nearly identical between those who smoke (1) and those who do not smoke (0). This suggests that smoking may not be as strongly linked to cardiovascular diseases in this particular dataset.

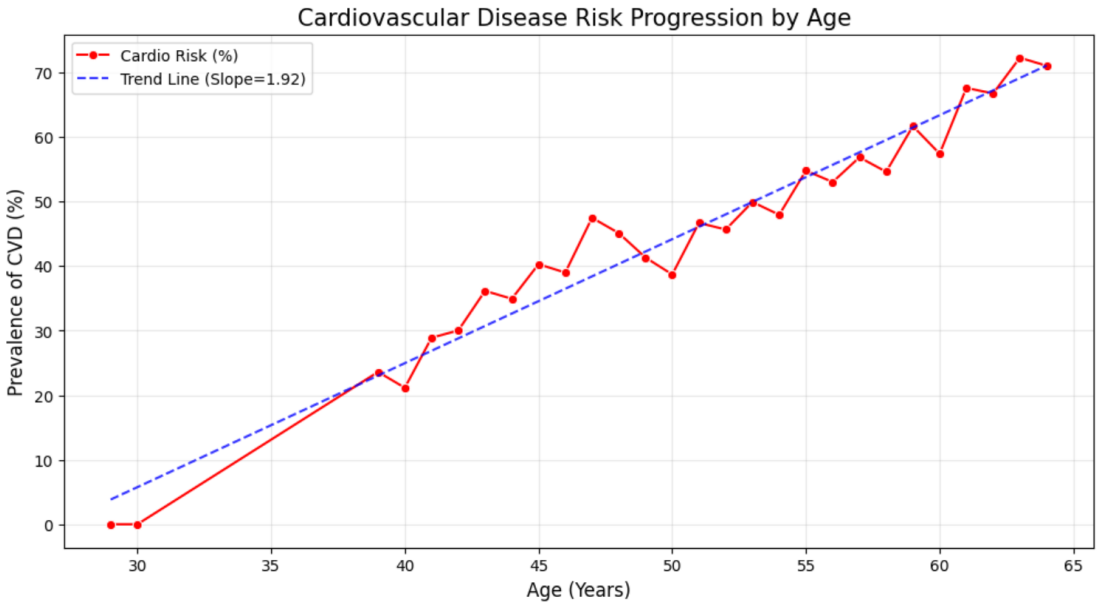
Fourth, for alcohol consumption, there is also no noticeable difference in the cardiovascular disease percentage between those who drink alcohol (1) and those who

do not (0). This suggests that alcohol consumption might not have a substantial impact on the likelihood of developing cardiovascular diseases in this dataset.

Fifth, in terms of physical activity, the cardiovascular disease percentage is almost the same for both active (1) and inactive (0) individuals. This suggests that physical activity might not be a significant differentiator for cardiovascular diseases in this dataset.

Finally, the blood pressure category (bp_cat) chart shows a clear trend: individuals with higher blood pressure categories (from elevated to high 3) have progressively higher percentages of cardiovascular diseases. This indicates that higher blood pressure is a strong risk factor for cardiovascular diseases.

3.4.Age-Trend Analysis



This graph depicts the progression of cardiovascular disease risk as a function of age. The data points, shown in red, represent the percentage of people with cardiovascular disease risk at different ages ranging from 30 to 65 years. As age increases, the prevalence of CVD risk also rises, indicating a positive correlation between age and cardiovascular disease risk.

Additionally, the graph includes a trend line (represented by a blue dashed line) with a slope of approximately 1.92, suggesting a steady increase in CVD risk as age advances. This line helps emphasize the overall upward trend in the data, visually supporting the notion that older individuals are more likely to experience higher cardiovascular risks.

In conclusion, this graph highlights the increasing risk of CVD as individuals grow older, which is a crucial factor in health assessments and preventive healthcare strategies.

4. Predicting using Machine Learning

4.1.Baseline Modeling Setup

Using multiple health features from the dataset, this study aimed to predict

whether an individual has cardiovascular disease. The process began by constructing a feature matrix from the raw data, where columns not intended as input features were removed. This included the target variable cardio itself, as well as columns such as bp_cat, id, and cardio_percent to prevent information leakage and redundancy. The target variable was defined as the cardio status. Subsequently, the dataset was split into training and test sets. The training set was utilized to train a Random Forest Classifier with a fixed random seed of 0 to ensure experimental reproducibility. Upon evaluation on the test set, the model achieved an Accuracy of 71.53%. This result indicates that the model correctly predicted the disease status in approximately 71.53% of the cases, showcasing a reasonable classification capability. To ensure a more comprehensive evaluation—particularly given the clinical importance of minimizing misdiagnosis—additional metrics such as the confusion matrix, Precision, Recall, F1-score, and AUC were also considered.

4.2.Logistic Regression Analysis

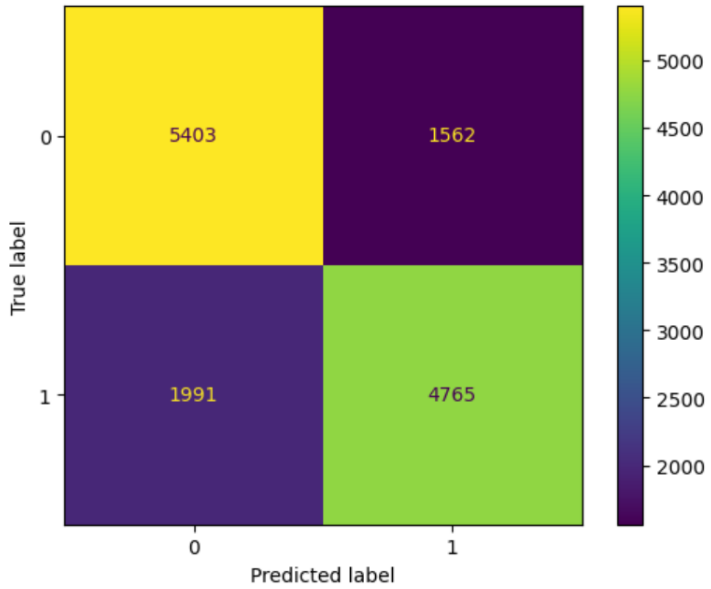
As the designated baseline model, Logistic Regression demonstrated robust stability, achieving a Testing Accuracy of ~73.39% alongside a Training Accuracy of ~72.55%. The minimal divergence (less than 1%) between these metrics indicates that the model generalizes well to unseen data without suffering from overfitting, validating the linear separability of the primary risk factors. Feature correlation analysis revealed that Systolic Blood Pressure (ap_hi) is the absolute dominant predictor of cardiovascular disease, exhibiting the largest positive coefficient in the logistic regression model. This finding quantitatively confirms that hypertension is the most critical linear determinant of risk in this dataset. Secondary physiological indicators, including Cholesterol (0.34) and Age (0.17), also showed significant positive contributions, aligning with the study's hypothesis regarding age-related progression. Conversely, physical activity (active) displayed a negative correlation (-0.09), providing statistical evidence for the protective effect of exercise. Notably, lifestyle factors such as smoking (smoke) and alcohol consumption (alco) exhibited negligible or slightly negative weights, suggesting that their predictive power is overshadowed by the profound impact of direct physiological markers like blood pressure and BMI in a linear modeling context.

4.3.Decision Tree Analysis

Following the linear baseline, a Decision Tree classifier was implemented to explore potential non-linear patterns within the dataset. The model, constrained with a maximum depth of 10, achieved a Testing Accuracy of 72.63% and an F1 Score of 0.7152. Interestingly, this performance was slightly inferior to the Logistic Regression baseline (73.39%). This result suggests that while decision trees are easy to interpret, a single tree structure struggled to efficiently capture the simple, linear relationships present in the data. Unlike the linear model which fits a smooth trend, the decision tree imposes rigid split points, potentially breaking down a straightforward pattern into too many complex rules. This caused a slight loss in generalization performance, reinforcing the observation that the primary risk factors (like blood pressure) follow a direct and continuous trend. Consequently, the inability of the single decision tree to surpass the baseline highlights the limitations of using just one estimator and

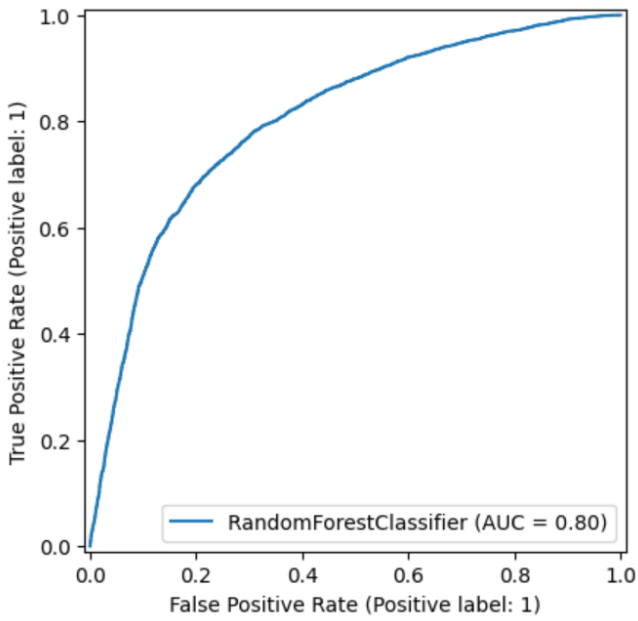
underscores the necessity of employing ensemble techniques, such as Random Forest, to improve stability and accuracy.

4.4.Random Forest Model



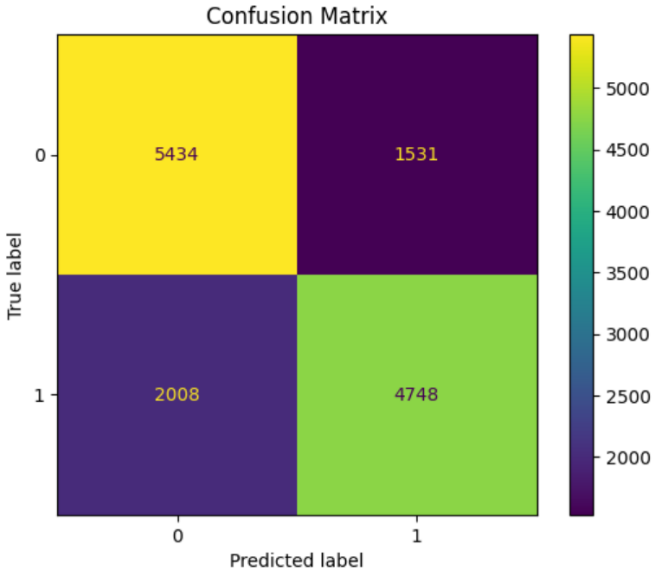
The Random Forest model in the current analysis shows a testing accuracy of approx. 71%, meaning the model correctly predicts whether an individual has cardiovascular disease about 71% of the time. The F1 score is 0.7284, indicating a reasonable balance between precision and recall, which is crucial when handling imbalanced datasets like this one. Additionally, the average cross-validation accuracy is 73.39%, reflecting the model's consistency across different subsets of the data.

Looking at the confusion matrix, we observe that the model correctly predicts 5,403 true negatives, which are healthy individuals identified correctly. It also identifies 4,765 true positives—people with cardiovascular disease who were accurately predicted to be at risk. However, the model still makes 1,562 false positive predictions, where healthy individuals are wrongly identified as having cardiovascular disease, and 1,991 false negatives, where individuals with the disease are missed. False negatives are particularly concerning in healthcare, as they can delay diagnoses and treatment for those who need immediate care.



The ROC curve shown in the image illustrates the performance of the Random Forest Classifier model. The AUC is 0.80, indicating that the model performs reasonably well at distinguishing between the positive and negative classes. An AUC value of 0.80 means that the model has a high ability to correctly identify individuals with and without cardiovascular disease.

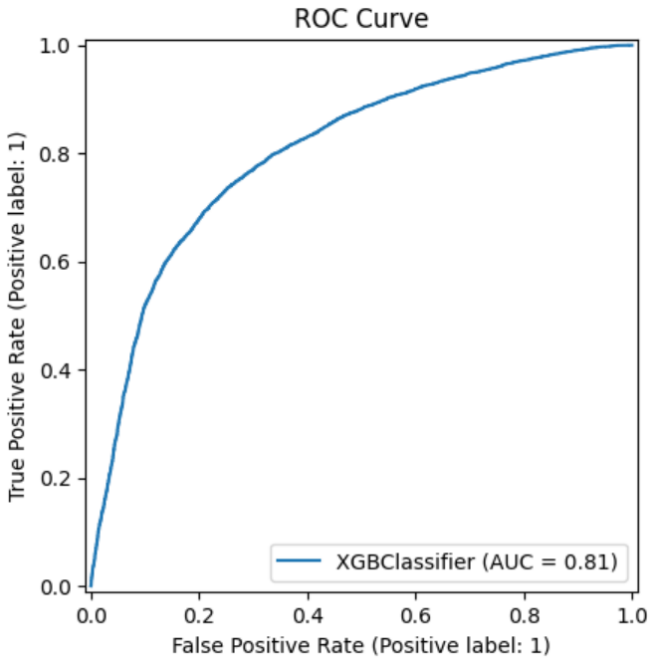
4.5.Gradient Boosting



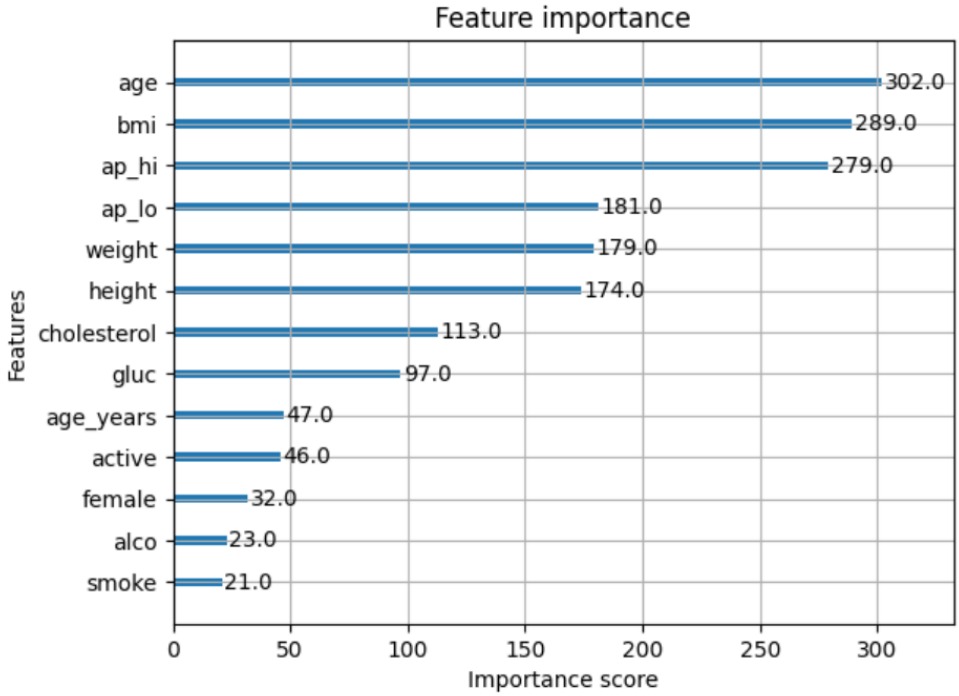
The model has a testing accuracy of approx. 73% and an F1 score of 0.7285, which indicates that it performs reasonably well in balancing precision and recall. Using cross-validation, the average testing accuracy came out to be ~73.59%, which shows consistent performance across different subsets of the data.

Looking at the confusion matrix, we can see that the model correctly identified 5,434 healthy individuals (true negatives), and it also correctly predicted 4,748 individuals who had cardiovascular disease (true positives). However, the model also made some mistakes. There were 1,531 false positives, meaning the model predicted

cardiovascular disease in healthy individuals. Additionally, there were 2,008 false negatives, where individuals with cardiovascular disease were incorrectly predicted as healthy. This is a significant issue, as false negatives could result in missed diagnoses and delays in treatment.



The XGBoost model performs well with an AUC of 0.81, indicating strong discriminatory power. The ROC curve also suggests that the model is good at distinguishing between those with and without cardiovascular disease.



The feature importance chart provides insight into the variables that most strongly influence the model's predictions of cardiovascular disease risk. From the chart, we can see that age is the most important feature with an importance score of 302, followed closely by BMI (289) and ap_hi (279), which likely refer to systolic

blood pressure measurements. These features have significantly higher importance scores compared to others, suggesting they have a greater impact on predicting cardiovascular disease risk.

Following these, ap_lo , weight, and height are also crucial, with importance scores of 181, 179, and 174 respectively. Cholesterol and glucose levels come next, with scores of 113 and 97, highlighting their relevance in predicting health risks.

Other variables such as age_years, activity level, gender (female), and lifestyle factors like alcohol consumption (alco) and smoking (smoke) have comparatively lower importance scores. However, they still contribute to the model, but their influence is weaker than the primary health metrics like age, BMI, and blood pressure.

In summary, the model relies heavily on age, BMI, and blood pressure readings, with secondary importance given to weight, height, cholesterol, and glucose levels, while lifestyle factors like smoking and alcohol consumption play a smaller role. This emphasizes the importance of physiological measurements in predicting cardiovascular risk.

5. Summary of the performance of the prediction model

Rank	Model	Testing Accuracy (%)	F1	Key Observation
1	XGBoost	73.63%	0.7285	Achieved the highest accuracy, but the improvement over the Logistic Regression baseline was very
2	Logistic Regression	73.39%	0.7159	Performed very well for a simple baseline. This indicates that the relationship between risk factors and the disease is mostly linear.
3	Decision Tree	72.63%	0.7152	Scored slightly lower than the baseline. The single tree structure was less effective than the linear model for
4	Random Forest	71.53%	0.7284	Lowest accuracy among all models. The ensemble approach did outperform the

6.

As shown in Table , XGBoost and Logistic Regression demonstrated the highest

stability. The minimal performance gap between the linear baseline and the complex ensemble model reinforces the finding that cardiovascular risk factors in this dataset exhibit strong linear characteristics.

5.1.Limitations

While the models in this study demonstrate reasonable predictive performance, several limitations should be acknowledged. First, the dataset is cross sectional, meaning each observation represents a single point in time rather than tracking individuals longitudinally. As a result, the analysis cannot capture causal relationships or true disease progression over time.

Second, several lifestyle variables, including smoking, alcohol consumption, and physical activity, are self reported, which introduces potential reporting bias and measurement error. This may partially explain why these variables appear less predictive compared to objective clinical measurements.

Additionally, although extensive cleaning was performed, the dataset may still contain unobserved confounders such as family history, medication use, or socioeconomic factors that influence cardiovascular risk but are not included in the data.

Finally, model performance is evaluated using accuracy and F1 score, which may not fully reflect clinical usefulness, particularly in settings where false negatives carry a higher cost than false positives.

5.2.Conclusion and Next Steps

This study demonstrates that machine learning models can effectively predict cardiovascular disease risk using routinely collected clinical and lifestyle data.

Across all models tested, Logistic Regression and XGBoost achieved the most stable and consistent performance, with testing accuracies around 73 to 74 percent and strong AUC values. The minimal performance gap between the linear baseline and more complex ensemble models suggests that the dominant risk factors in this dataset, particularly age, systolic blood pressure, and BMI, exhibit largely linear relationships with disease outcomes.

These findings highlight the value of simpler, more interpretable models in clinical risk prediction tasks where transparency is important. Future work could focus on incorporating longitudinal data to better capture disease progression, expanding the feature set to include medication history or genetic risk factors, and optimizing decision thresholds to prioritize the reduction of false negatives in real world healthcare settings.