A Predictive Analysis of Coronary Heart Disease(CHD) Using Health and Lifestyle Variables

Anteneh G. Yitayal 256983

2025-03-28

Introduction

Coronary Heart Disease (CHD), also known as coronary artery disease (CAD) is a condition in which the coronary arteries (the blood vessels that supply oxygen-rich blood to the heart) become narrowed or blocked. This happens due to the buildup of plaque, a combination of fat, cholesterol, calcium, and other substances, on the artery walls, a process known as atherosclerosis.

CHD is one of the leading causes of death worldwide. It increases the risk of heart attacks, heart failure, and other cardiovascular complications [1]. The condition often develops over decades, and many people may not realize they have it until they experience symptoms like chest pain (angina), shortness of breath, or a heart attack.

Objective

The objective of this analysis is to identify key risk factors associated with Coronary Heart Disease (CHD) and compare the predictive performance of logistic regression and k-Nearest Neighbors (KNN). The study aims to assess the impact of demographic, health, and lifestyle factors on CHD risk while evaluating the differences between a parametric model (logistic regression) and a non-parametric model (KNN) in terms of recall, AUC-ROC, and accuracy.

Data set

The data set consists of 4,238 observations and 13 attributes related to cardiovascular health, specifically focusing on risk factors for Coronary Heart Disease (CHD). The attributes include demographic information such as sex and age, lifestyle factors like smoker and cigarettes per day (cpd), and clinical data such as hypertension (HTN), diabetes, cholesterol (chol), diastolic blood pressure (DBP), body mass index (BMI), and heart rate (HR). The data set also includes education level and a target variable CHD, indicating whether the individual has Coronary Heart Disease ("Yes" or "No"). This data set provides a comprehensive set of features that can be used for predictive modeling and analysis of factors influencing CHD risk.

The data set contains a total of 204 missing values, with education level accounting for 51% of them. This indicates that a significant portion of the missing data is concentrated in the education attribute. To ensure the accuracy of the analysis, handling these missing values is essential. Possible approaches include imputation (estimating and replacing missing values) or removal of affected rows. However, since the number of missing values is relatively small, the affected rows have been removed for this analysis.

Exploratory Data Analysis

There is a significant class imbalance in the target variable, with approximately 85% of the data belonging to the "No CHD" category as shown in figure 1. This imbalance is common in medical data sets and can affect model performance by favoring the majority class. To address this, specialized evaluation metrics such as AUC-ROC and recall were used instead of re-sampling techniques. These metrics are commonly used to provide better assessment of model performance in imbalanced scenarios [2].

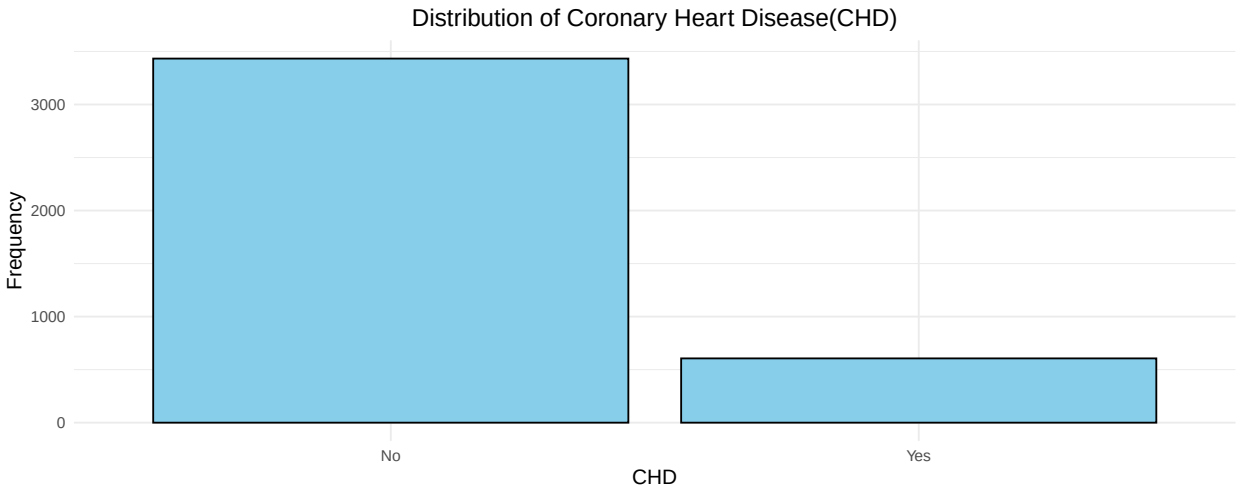

Figure 1: Distribution of Coronary Heart Disease(CHD).

The box plot in figure 2 illustrates the distribution of age across individuals with and without coronary heart disease (CHD), categorized by sex. Overall, individuals diagnosed with CHD tend to be older than those without the condition, as evidenced by the higher median age in the "Yes" CHD group. Additionally, the boxes (representing the interquartile range) are larger for the CHD group, indicating more variability in age.

For both CHD "Yes" and "No" groups, the age distributions between males and females are fairly similar. However, there is a slight tendency for females with CHD to have a higher median age compared to males with CHD, which might lined with menopause. Overall, the relationship between age and CHD appears to hold true regardless of sex.

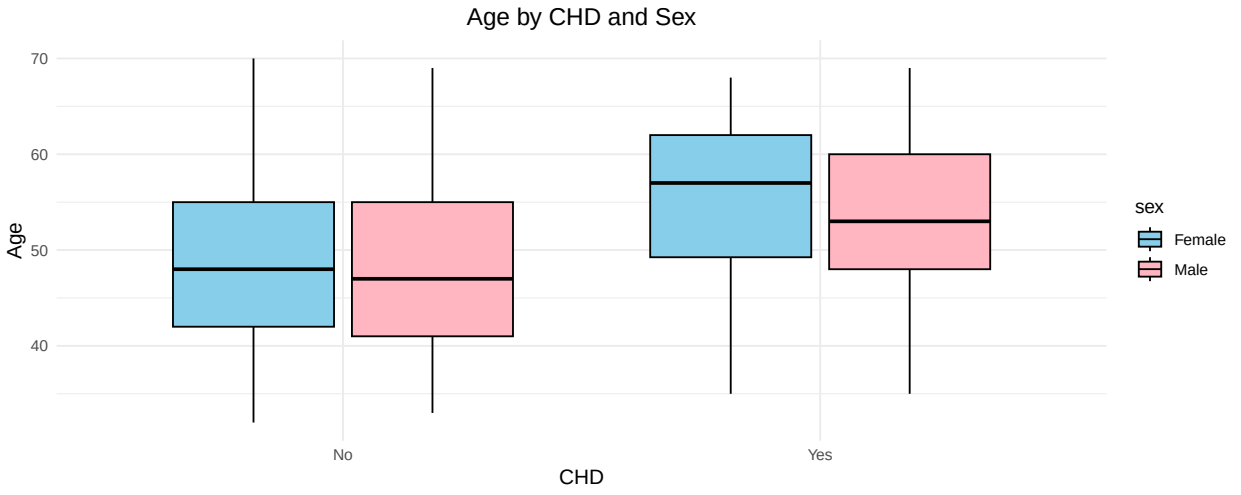

Figure 2: Distribution of Coronary Heart Disease(CHD) with age between male and female.

Another interesting insight from the data is that CHD tends to be more prevalent among individuals with lower education levels as shown in figure 3. This may be linked to factors such as income and standard of living—typically, the more education a person has, the higher their income, which in turn may lead to better overall health and reduced risk of conditions like CHD. The effect of smoking appears insignificant based on the current data, seeking further investigation using logistic regression or other appropriate statistical tests. Figure 4 suggests a potential association between slightly elevated diastolic blood pressure and an increased risk of Coronary Heart Disease (CHD). Additionally, while extreme blood pressure levels were observed in both the "Yes" and "No" CHD groups, the distribution of cholesterol levels and body mass index appeared similar across both groups.

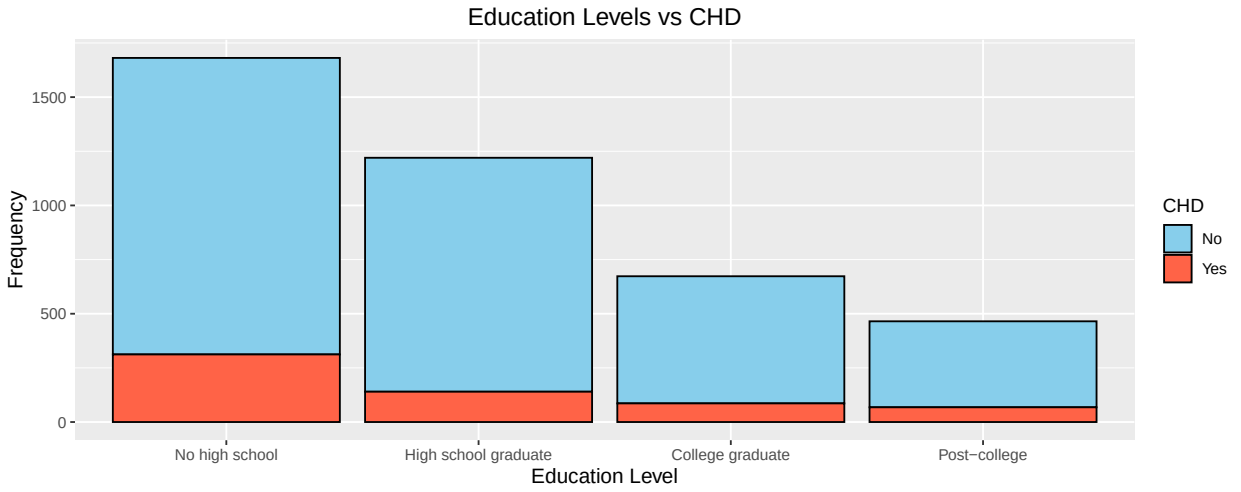

Figure 3: Distribution of Coronary Heart Disease(CHD) with education level.

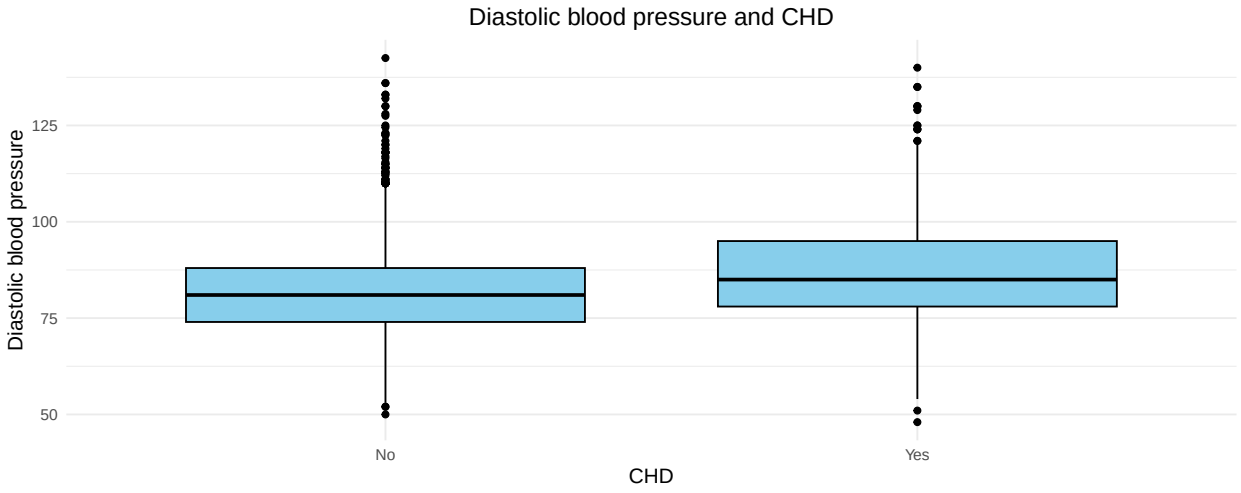

Figure 4: Distribution of Coronary Heart Disease(CHD) with Diastolic blood pressure.

Methods

The data set is split into training and testing sets using stratified sampling, ensuring that the class distribution of the target attribute is maintained. The training set comprises 70% of the data, while the testing set accounts for the remaining 30%.

Logistic Regression

The logistic regression model predicts the likelihood of Coronary Heart Disease (CHD) using various risk factors, as shown in Equation (1). Significant predictors include male sex, age, cigarettes per day (cpd), hypertension (HTN1), diabetes (diabetes1), and diastolic blood pressure (DBP), all of which increase CHD risk.

The model shows that age ($\beta = 0.07$, $p < 0.001$) and male sex ($\beta = 0.45$, $p < 0.001$) are strong predictors, indicating that older individuals and males are at higher risk. While smoking status is not statistically significant, the number of cigarettes per day (cpd) ($\beta = 0.02$, $p = 0.0026$) has a significant positive effect, suggesting that smoking intensity matters more than simply being a smoker.

Other variables such as education level, cholesterol, BMI, and heart rate do not significantly impact CHD risk in this model. The model demonstrates a good fit, as indicated by the reduction in deviance compared to the null model. These findings suggest that controlling factors like smoking intensity, hypertension, and diabetes could be key strategies in CHD prevention.

$$\text{logit}(\mathbb{E}(\text{CHD})) = \beta_0 + \beta_1(\text{sex}) + \beta_2(\text{age}) + \beta_3(\text{education}) + \cdots + \beta_{12}(\text{HR}) \tag{1}$$

Since the data is imbalanced, adjusting the classification threshold for predicting CHD is necessary. This can be done by plotting the true positive rate (i.e., recall) against a range of possible threshold values. The goal is to maximize the correct classification of CHD cases while avoiding classifying all samples as CHD.

```
#fit with all features
log_reg_all <- glm(CHD ~ .,
    data = train_set,
    family = binomial
)

summary(log_reg_all)
```

```
##
## Call:
## glm(formula = CHD ~ ., family = binomial, data = train_set)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -7.757625   0.786695  -9.861  < 2e-16 ***
## sexMale      0.453489   0.123524   3.671 0.000241 ***
## age          0.070010   0.007446   9.403  < 2e-16 ***
## education2  -0.098396   0.138861  -0.709 0.478575
## education3  -0.044298   0.168128  -0.263 0.792182
## education4   0.037595   0.185200   0.203 0.839136
## smoker1      0.124940   0.174744   0.715 0.474617
## cpd          0.020422   0.006792   3.007 0.002640 **
## stroke1      0.882155   0.537760   1.640 0.100917
## HTN1         0.414611   0.144534   2.869 0.004123 **
## diabetes1    0.842581   0.264373   3.187 0.001437 **
## chol         0.001312   0.001301   1.008 0.313296
## DBP          0.016822   0.005691   2.956 0.003120 **
## BMI          0.006931   0.014319   0.484 0.628333
## HR          -0.002291   0.004771  -0.480 0.631094
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2393.9  on 2828  degrees of freedom
## Residual deviance: 2144.5  on 2814  degrees of freedom
## AIC: 2174.5
##
## Number of Fisher Scoring iterations: 5
```

K-Nearest Neighbor (KNN)

KNN is applied to compare its performance with logistic regression in terms of maximizing the true positive rate and minimizing false negatives. Missing CHD "Yes" cases could result in misdiagnosing individuals with CHD as healthy, potentially leading to severe health risks or even loss of life. In contrast, false positives, while undesirable, are less costly in comparison.

All available attributes have been used in the model. Binary categorical variables, such as sex, are mapped to numerical values (e.g., 0 and 1). However, for education level, one-hot encoding is applied since it has more than two categories, preventing the model from assuming an ordinal relationship between education levels and the likelihood of developing CHD, as no such relationship is established. Continuous variables are standardized to mitigate the impact of different numerical scales on distance calculations in KNN. Without standardization, features with larger numerical ranges could disproportionately influence the model. To address this, the training set is standardized, and the same mean and standard deviation are applied to the test set to ensure consistency, as machine learning models generally assume that training and testing data come from the same distribution.

Given the imbalanced nature of the data, accuracy or error rate alone does not provide meaningful insights into the model's ability to identify CHD "Yes" cases. Instead, recall is prioritized to select the optimal K, ensuring that at least 80% of CHD "Yes" cases are correctly identified while allowing some tolerance for false positives.

Results and Discussion

The primary evaluation metrics for this task are recall and AUC-ROC, with accuracy also considered. Given the class imbalance in the dataset, recall is crucial for correctly identifying CHD cases, while AUC-ROC provides a comprehensive measure of the model's ability to distinguish between classes. I trained both models, evaluated them on the test set, and tested a range of K values for KNN (Figure 6) and threshold values for logistic regression (Figure 7). The optimal values were selected based on achieving an 80% true positive rate for the CHD "Yes" class, resulting in a threshold, within the range indicated by two dotted red vertical lines which give 80% true positive rate, 0.1039 for logistic regression and K=7 for KNN.

The main concern when comparing models is not only correctly predicting the positive class but also minimizing false positives without compromising the true positive rate. The performance comparison between KNN and logistic regression highlights key differences in their ability to classify CHD cases. KNN, with K=7, achieved an accuracy of 19.1%, misclassifying a significant number of negative ("No") cases, with 960 false positives and 162 true positives. Its AUC-ROC score of 0.4953 suggests that the model's performance is close to random guessing. In contrast, logistic regression, with a threshold of 0.1039, demonstrated better performance, achieving an accuracy of 57.2% and an AUC-ROC score of 0.738, as shown in Figure 5. It correctly identified 145 true positives while reducing false positives to 482, compared to KNN, as shown in Table 1. Although logistic regression performs better overall, both models still struggle with class imbalance, as indicated by their relatively low accuracy and AUC-ROC scores.

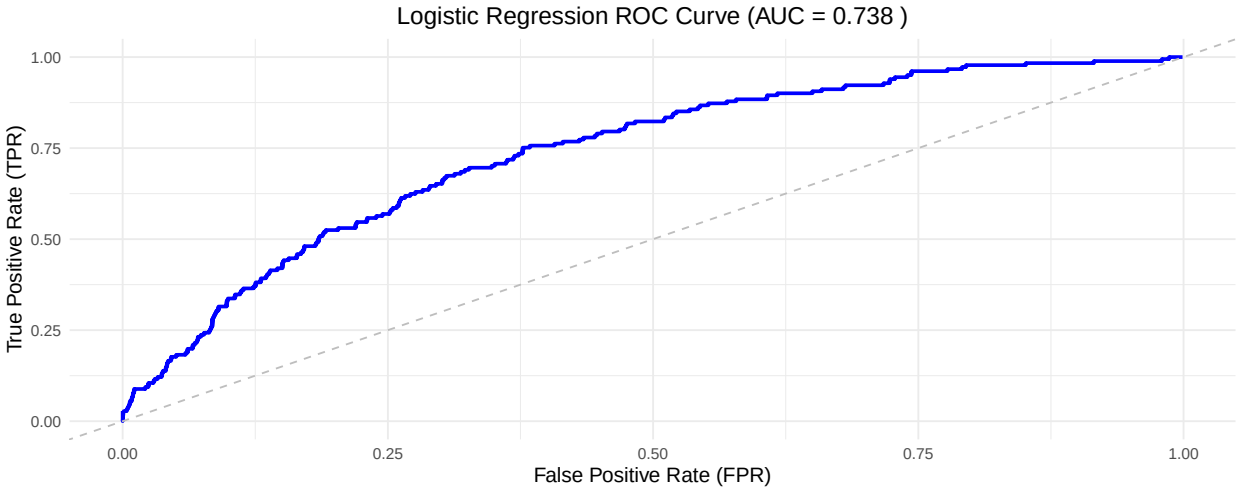

Figure 5: Logistic Regression ROC Curve.

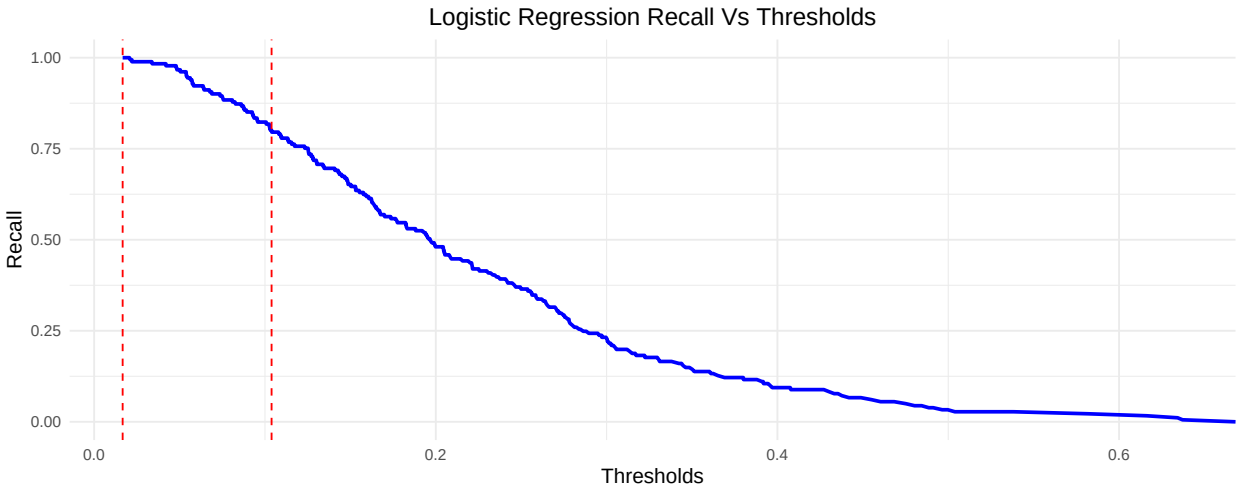

Figure 6: Recall vs Possible values of thresholds

Table 1: Confusion Matrices for Logistic Regression vs KNN

Actual \ Prediction	KNN-No	KNN-Yes	Logistic Regression-No	Logistic Regression-Yes
No	66	18	547	36
Yes	963	163	482	145

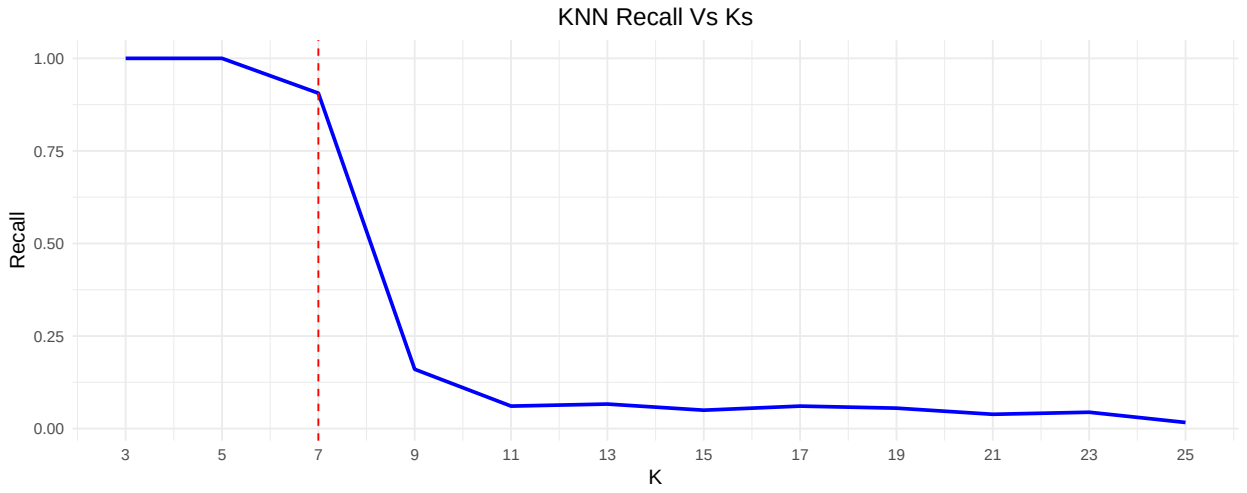

Figure 7: Recall vs Possible values of K.

Conclusion

The analysis aimed to identify key risk factors for Coronary Heart Disease (CHD) and compare the performance of logistic regression and k-Nearest Neighbors (KNN) models in predicting CHD. Key insights from the data suggest that demographic and health factors, such as age, male sex, cigarettes per day (cpd), hypertension (HTN), diabetes, and diastolic blood pressure, have a significant impact on CHD risk.

Despite the better performance of logistic regression, both models struggled with class imbalance, which can lead to high false positives and misclassification of CHD cases. The use of recall and AUC-ROC metrics helped mitigate this challenge, providing a more accurate assessment of model performance in imbalanced datasets. Logistic regression's threshold adjustment and KNN's optimal K value (7) were crucial in maximizing true positive rates.

In conclusion, while logistic regression proved to be a better model for predicting CHD in this analysis, further improvements, such as handling the class imbalance through more sophisticated techniques (e.g., resampling, penalization), may enhance model performance and help achieve more accurate predictions for CHD.

References

1. National Center for Health Statistics. Multiple Cause of Death 2018–2022 on CDC WONDER Database. Accessed May 3, 2024. https://wonder.cdc.gov/mcd.html

2. Eve Richardson, Raphael Trevizani, Jason A. Greenbaum, Hannah Carter, Morten Nielsen, Bjoern Peters, The receiver operating characteristic curve accurately assesses imbalanced datasets, Pat-

terns, Volume 5, Issue 6, 2024, 100994, ISSN 2666-3899, https://doi.org/10.1016/j.patter.2024.100994. (https://www.sciencedirect.com/science/article/pii/S2666389924001090)