Analysis of 10-Year Coronary Heart Disease Risk

TaeWoo Kim, Yuxi Guo, JunChang Song

1 Introduction

Our research aims to contribute to the medical field. During our exploration of various online resources, we discovered an intriguing dataset on the Kaggle platform. This dataset `https://www.kaggle.com/datasets/captainozlem/framingham-chd-preprocessed-data?select=CHD_preprocessed.csv`, which is publicly accessible on Kaggle, originates from a continuous cardiovascular study involving the residents of Framingham, Massachusetts. The primary objective of this classification task is to determine the 10-year risk of future coronary heart disease (CHD) in patients. The dataset encompasses information on over 4,000 individuals and includes 15 attributes, providing a comprehensive basis for our analysis. The detailed description of variables in this data can be found in Appendix.

2 Motivation

The global burden of Coronary Heart Disease (CHD), a leading cause of death worldwide, underscores the urgent need for improved early detection and prevention strategies. Traditional risk assessment models, limited by the scope of variables they consider, often fall short in accurately predicting CHD risk. The emergence of machine learning (ML) in healthcare presents an unprecedented opportunity to enhance CHD risk prediction by analyzing complex datasets to identify intricate patterns not detectable by conventional methods.
This research is propelled by the necessity to overcome the limitations of existing models through the development and comparison of advanced ML models for 10-Year CHD Risk prediction. By embracing the complexity of CHD risk factors and employing sophisticated ML techniques, this study aims to contribute to more precise, personalized, and early CHD risk assessment. Ultimately, the goal is to inform better preventative and treatment strategies, significantly impacting public health and patient care in the realm of CHD.

3 EDA: Exploratory Data Analysis

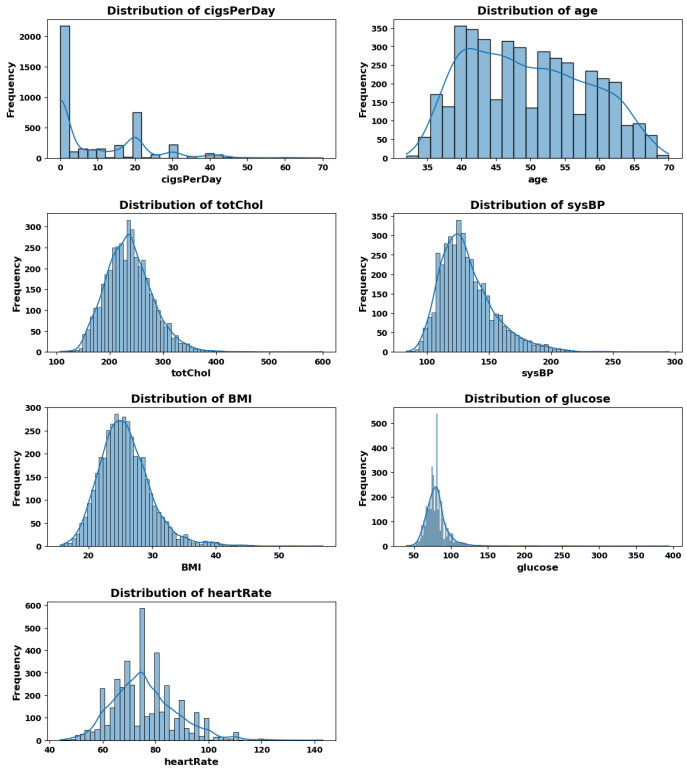

Figure 1: Numerical Predictor Variables

Upon examining the distribution of numerical variables, it was observed that while some variables appear to follow a normal distribution, others do not, displaying a skewed distribution.

- **Cigarettes Per Day (*cigsPerDay*)**: Most individuals are non-smokers, but for the majority of smokers, 20 cigarettes per day is an average.
- **Age**: Most individuals are middle-aged, with a small skew towards older ages.
- **Total Cholesterol (*totChol*)**: Levels are normally distributed with a peak in the desirable range, showing a healthy profile for the majority.
- **Systolic Blood Pressure (*sysBP*)**: The distribution is right-skewed, with a number of individuals showing higher than average systolic blood pressure, which could be a health concern.

- **Body Mass Index (*BMI*)**: Mostly normal to overweight range with a slight right skew, indicating some individuals are in the obese range.
- **Glucose**: High right skewness indicates most individuals have normal glucose levels, but a subset has elevated levels, which might suggest diabetes risk.
- **Heart Rate**: Distribution is close to normal with a slight skew towards higher rates, but generally centered around a normal resting range.

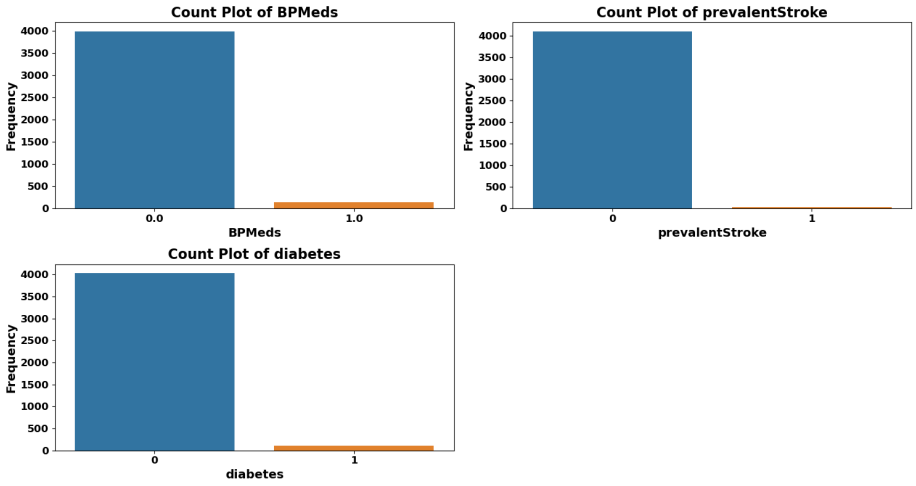

Figure 2: Categorical Predictor Variables

- **Count Plot of BPMeds**:
  - Majority of the individuals are not on blood pressure medication (`BPMeds=0`).
  - A small number are taking blood pressure medication (`BPMeds=1`).
- **Count Plot of prevalentStroke**:
  - Large majority of individuals have not had a stroke (`prevalentStroke=0`).
  - A very small number have had a stroke (`prevalentStroke=1`).
- **Count Plot of diabetes**:
  - Most individuals do not have diabetes (`diabetes=0`).
  - A minority of individuals have diabetes (`diabetes=1`).

For all three conditions—blood pressure medication, stroke prevalence, and diabetes—the number of individuals without the condition (`0`) significantly outweighs the number with the condition (`1`). This suggests that these health issues are not widespread within the sample population.

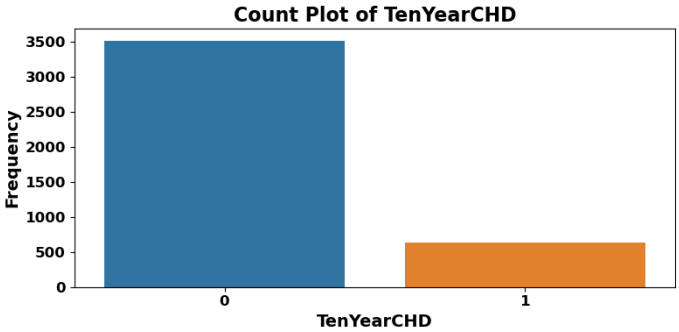

Figure 3: Target Variable

Upon examination of the target variable within the dataset, it is observed that the distribution is substantially imbalanced. Such an imbalance in the target variable distribution could potentially introduce bias into the predictive modeling process, leading to less reliable predictions for the minority class. Consequently, this necessitates the consideration of resampling methodologies. These methods, such as oversampling the minority class, undersampling the majority class, or synthesizing new data, are pivotal to mitigate the effects of imbalance and enhance the robustness of the ensuing predictive models.

Variable	**Association**	**p-value**
Male	Significant	8.45×10^{-08}
Education	Not Significant	0.087
Current Smoker	Not Significant	0.308
BPMeds	Significant	3.04×10^{-09}
Prevalent Stroke	Significant	0.00018
Prevalent Hyp	Significant	0.00103
Diabetes	Significant	8.34×10^{-10}

Table 1: Chi Square Test

The chi-square tests reveal that male, BPMeds, diabetes, prevalentStroke, and prevalentHyp have a statistically significant association with the target variable, whereas education and currentSmoker do not.

male	**age**	**education**	**currentSmoker**	**cigsPerDay**
1.198498	1.356255	1.02634	2.561237	2.699904
BPMeds	**prevalentStroke**	**prevalentHyp**	**diabetes**	**totChol**
1.115104	1.020592	2.058312	1.582344	1.111583
sysBP	**diaBP**	**BMI**	**heartRate**	**glucose**
3.741963	2.971938	1.225777	1.096229	1.60437

Table 2: VIF Table

Examining the VIF result table above, all the values are below 10, indicating that no multicollinearity exists.

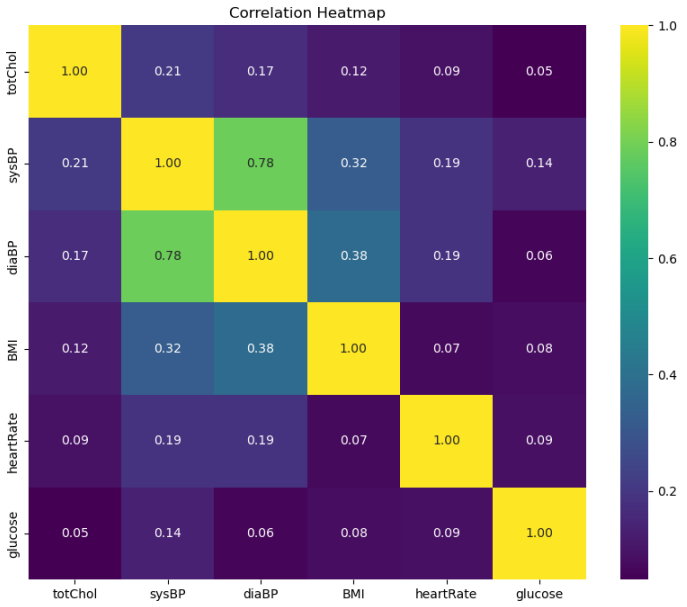

Figure 4: Correlation Heatmap

The heatmap shows that systolic and diastolic blood pressure are strongly correlated with each other. There's a moderate relationship between body mass index (BMI) and blood pressure, but cholesterol, heart rate, and glucose levels show little to no linear relationship with the other variables in this dataset. Overall, most variables do not have strong linear correlations with each other.

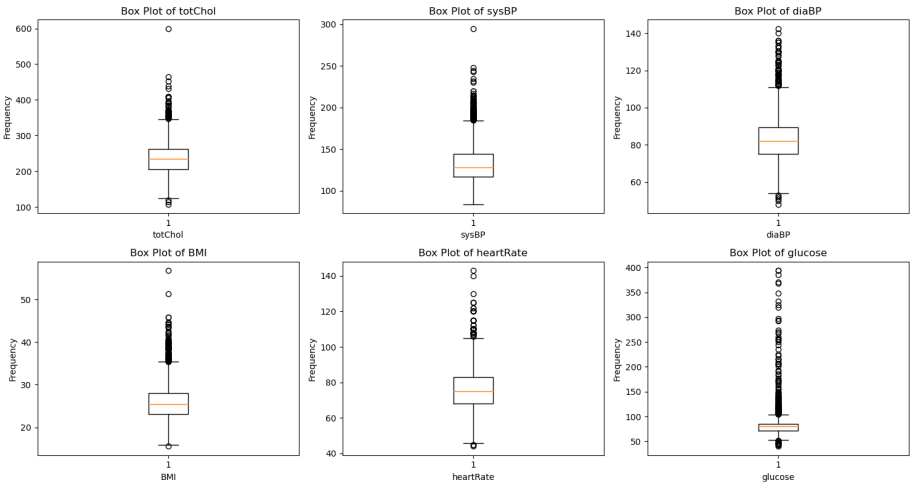

Figure 5: Box Plots

- **Total Cholesterol (totChol)**: The median cholesterol level is centrally located within the interquartile range, with numerous outliers indicating significantly higher cholesterol levels in some individuals.
- **Systolic Blood Pressure (sysBP)**: The median is situated closer to the lower quartile, suggesting a skewness towards lower systolic blood pressure values. However, there is a presence of many outliers on the higher end.
- **Diastolic Blood Pressure (diaBP)**: A similar pattern to systolic blood pressure is observed, with the median towards the lower quartile and several high-end outliers, reflecting a skewness towards lower values within the central 50% of the data.
- **Body Mass Index (BMI)**: The median is closer to the upper quartile, indicating a slight skew towards higher BMI values, along with some extremely high outliers.
- **Heart Rate**: The median heart rate is approximately central in the box, indicating a symmetric distribution, albeit with some high-end outliers.
- **Glucose**: The median is observed near the lower end of the box, showing a distribution skewed towards lower glucose readings, complemented by a substantial number of high outliers.

Inspection of the boxplots for the numerical predictor variables indicates the presence of potential outliers. To address this, the z-score method was employed, identifying data points that deviate markedly from the mean. Specifically, data points with a z-score exceeding 3 or falling below -3 were considered outliers and subsequently excluded from the analysis. This criterion led to the removal of 219 observations, resulting in a refined dataset comprising 3913 data points suitable for further analysis.

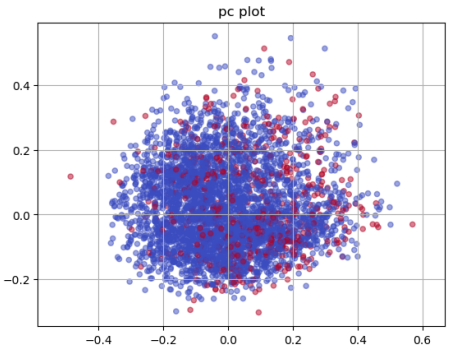

Figure 6: 2D PCA plot

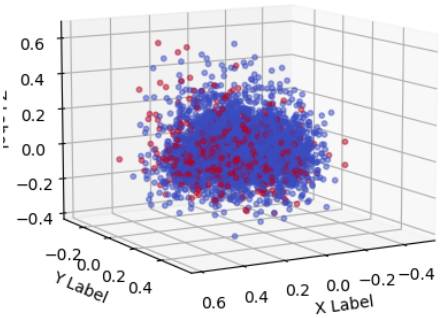

Figure 7: 3D PCA plot

The Principal Component Analysis (PCA) plots presented herein depict the projection of data points onto the principal components' axes. Observations denoted in red signify individuals identified with a 10-Year Coronary Heart Disease (CHD) Risk. The proximity of red points to other data points in the PCA plots intimates potential challenges in effectively classifying the individuals with regard to their 10-year CHD risk. Furthermore, the first three principal components cumulatively account for 40% of the variance within the dataset, which provides a substantial yet incomplete representation of the data's complexity. This partial variance explanation suggests that additional factors might be influential in the classification and should be considered.

4 Method

4.1 Random Forest

The Random Forest algorithm embodies an ensemble learning strategy, particularly leveraging the bagging technique, to accomplish the classification of 10-Year CHD Risk. This method involves the generation of numerous decision trees at the time of training, which collectively determine the outcome by identifying the most frequently occurring class among the individual trees.

Following the development of these trees, Random Forest evaluates new data by inputting the features into each tree and compiling their respective class predictions. The ultimate classification is determined by the class that garners the majority votes across all trees.

Thanks to its ability to average the predictions of numerous trees, Random Forest is renowned for its high accuracy across various tasks, effectively minimizing the overfitting risk. This model's strength also lies in its capacity to embrace diversity through random feature selection and the bootstrapping of data samples, thereby enhancing its ability to generalize.

Compared to a singular decision tree, Random Forest exhibits a higher tolerance to data noise and outliers, making it a more robust option for predictive modeling.

4.2 Gradient Boosting

Gradient Boosting builds a series of decision trees over time. Each new tree attempts to correct the errors or mistakes made by the previous trees. It's like playing a team sport where, after each game, the team focuses on improving the areas where they didn't perform well in the last game.

Initially, the model makes a simple guess for the classification of each instance (like guessing whether an email is spam or not). As it builds more trees, each one learns from the shortcomings of the previous trees, gradually improving the model's predictions. This is done by focusing more on the instances that were harder to classify correctly in earlier rounds.

Instead of relying on a single tree's prediction, Gradient Boosting combines the predictions from all the individual trees it has created. Each tree's vote is weighted based on its accuracy, and the combined result is used to make the final classification. This process is a bit like asking a group of experts to weigh in on a decision and then taking an average of their opinions, with more weight given to the more knowledgeable experts.

4.3 Support Vector Machine (SVM)

The Support Vector Machine (SVM) is a supervised learning algorithm used for classification tasks. Its main objective is to find the optimal decision boundary, known as the hyperplane, with the maximum margin between classes in a high-dimensional space. SVM achieves this by identifying support vectors, which are data points closest to the decision boundary.

In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces. SVMs can also be used for regression tasks. SVM performs well even with small datasets, often achieving high classification accuracy with limited data.

In our problem, the appropriate kernel will be decided by using `GridsearchCV`.

4.4 Logistic Regression

Logistic Regression is a statistical model used to solve classification problems. It predicts the probability of a sample belonging to a specific category based on given input variables.

While Logistic Regression is based on linear regression, it is utilized when the dependent variable is categorical. It is suitable for binary or multi-class outputs. The model applies a linear combination of input variables to a logistic function to output probabilities.

The logistic function, also known as the sigmoid function, produces an S-shaped curve and transforms input values into the range $[0, 1]$, representing probabilities. However, the decision boundary it generates is linear. Due to this characteristic, Logistic Regression outputs probabilities in binary classification and classifies samples based on a threshold.

Logistic Regression is relatively easy to interpret, simple to implement, and performs well on small datasets.

4.5 Multilayer Perceptron (MLP)

A Multilayer Perceptron (MLP) is a type of artificial neural network that consists of multiple layers, including input, hidden, and output layers. Each layer is composed of multiple neurons.

In an MLP, each neuron receives inputs, computes a weighted sum along with biases, and applies an activation function to produce an output. Typically,

nonlinear activation functions (such as sigmoid or ReLU) are used in the hidden and output layers, allowing the network to capture complex patterns in the data.

MLPs use the backpropagation algorithm to adjust the weights and biases in order to minimize the error between the predicted outputs and the actual targets. Backpropagation works by propagating the error backward through the network, updating the weights based on the gradients of the error with respect to each weight.

One of the key features of MLPs is their ability to learn complex relationships between inputs and outputs, making them suitable for various tasks such as classification and regression. MLPs can have multiple hidden layers, and the number of layers and neurons can be adjusted based on the problem complexity.

4.6 Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis (LDA) is a powerful technique for classification, assuming a normal distribution of predictors and modeling the distribution separately for each response class. It seeks to find a linear combination of features to form a decision boundary effectively separating classes.

LDA assumes each class follows a Gaussian distribution and shares the same covariance, simplifying computation and enabling effective classification even with limited data. It uses Bayes' rule to find posteriors for each class from data points and decides the label by comparing these posteriors.

4.7 K-Nearest Neighbors (KNN)

The K-Nearest Neighbors (KNN) algorithm predicts the 10-Year CHD Risk by employing a straightforward method: it examines the 'K' nearest labeled data points to the new data point requiring classification, and allocates to this new point the most frequently occurring class among those neighbors. Through cross-validation, the optimal value of 'K' determined for classifying 10-Year CHD Risk is 26.

5 Result

5.1 Performance

Table 3: Test Accuracy Comparison

Model	Test Accuracy (w/o CV)	Test Accuracy (w/CV=50)
Random Forest	0.872232	0.860204
Gradient Boosting	0.873935	0.860902
Logistic Regression	0.875639	0.862964
SVM	0.869676	0.860596
LDA	0.878194	0.860562
KNN	0.869676	0.861345
MLP	0.869676	0.858415

Table 4: Logistic Regression vs Others Comparison

Model	T-Test Result	W-Test Result
Random Forest	0.121463	6.45×10^{-05}
Gradient Boosting	0.274544	0.000496
SVM	0.187386	0.000110
LDA	0.192216	3.89×10^{-07}
KNN	0.354933	0.005190
MLP	0.015624	1.23×10^{-06}

The variables selected from the feature selection process were utilized. Numerical variables underwent scaling prior to being incorporated into the models. Parameter tuning was conducted using GridSearchCV. It's important to note that results may vary depending on the performance metrics used and the number of cross-validation folds.

The p-values of comparing Logistic Regression with other models are summarized in the Table 4. Two-sample test such as T-Test and W-Test were utilized, and the result indicates that the Logistic Regression shows reasonable significance compared to all other models. Here, T-Test is for parametric test when the data is normally distributed, and W-Test is for non-parametric test when the data might not normally distributed. From the result, we reject the null hypothesis of two models having similar performance because at least the result from W-Test clearly reject null hypothesis.

5.2 Model Selection & Interpretation

Logistic Regression emerged as the top-performing model among all evaluated models. The accompanying table outlines the coefficients derived from the Logistic Regression analysis. The findings highlight 'age' as the predominant factor influencing the 10-year risk of Coronary Heart Disease (CHD), with 'sysBP', 'cigsPerDay', 'male', and 'prevalentHyp' also playing significant roles.

Notably, 'cigsPerDay' is a factor within an individual's control, suggesting that reducing cigarette consumption or quitting smoking altogether could effectively lower the risk of 10-years risk of CHD. Additionally, understanding that lowering 'sysBP' can also reduce the 10-year risk of CHD, individuals can benefit from medical treatments or physical activities aimed at decreasing systolic blood pressure, potentially further reducing the risk of CHD.

From the perspective of healthcare businesses, this presents an opportunity to elevate awareness regarding the link between smoking, systolic blood pressure and CHD. This could allow them to enhance offerings in anti-smoking campaigns and physical activities. Additionally, focusing efforts more towards male demographics could be beneficial, underlining a potential association between gender and CHD risk.

Table 5: Logistic Regression Coefficients

age	sysBP	cigsPerDay	male	prevalentHyp	prevalentStroke
0.544997	0.273672	0.244173	0.223103	0.123665	0.077285

totChol	glucose	diabetes	BMI	BPMeds	diaBP
0.042633	0.019573	0.017461	0.000759	-0.005869	-0.070794

6 Finding

- **Imbalanced Data**

Due to the significantly low proportion of our minority class, we opted for the oversampling method. Specifically, we employed SMOTENC (Synthetic Minority Over-sampling Technique for Nominal and Continuous data). Unlike other oversampling techniques, SMOTENC does not simply duplicate or repeat existing samples of the minority class. The algorithm is designed to handle mixed data types, particularly datasets that include both continuous and categorical features. This is beneficial for learning a more generalized decision boundary. Overall, SMOTENC serves as a potent tool in addressing class imbalance within datasets featuring mixed types of attributes, enabling a more nuanced and effective modeling of complex data structures. After applying SMOTENC, we observed an approximate 5% increase in performance.

- **Evaluating Model Performance**

Accuracy, while a valuable and commonly used indicator, must be considered alongside a more comprehensive array of metrics that offer nuanced insights

Precision: Crucial in scenarios where false positives carry significant consequences, precision is of particular relevance in healthcare to prevent unnecessary procedures or tests.

Recall: The role of recall is fundamental in medical diagnostics, as missing a diagnosis can lead to severe outcomes. Hence, minimizing false negatives is crucial.

Specificity: This metric is key in screening processes to accurately identify individuals who do not have the condition, thus preventing mislabeling of healthy individuals.

F1 Score: An important metric for balancing precision and recall, especially in datasets where class imbalance is prevalent.

AUC-ROC: The AUC-ROC curve provides insights into the model's ability to distinguish between classes. A model with a high AUC indicates strong classification capabilities.

- **Challenges in Scaling Medical Data**

Scaling our medical dataset introduced several challenges. One major concern was how to handle outliers which could be significant in a medical context but may negatively impact model performance. Smoothing techniques were considered to reduce the influence of outliers without excluding them completely.

Furthermore, the dataset contained binary attributes like gender and smoker status, which posed unique scaling challenges. We standardized numerical attributes for Principal Component Analysis (PCA) to discover patterns. However, this standardization process complicates the interpretation of medical data within our analysis, necessitating a detailed and nuanced explanation.

7 Conclusion

Our analysis has successfully identified critical factors and built multiple models, such as Random Forest, Gradient Boosting, LDA, SVM, KNN, Logistic Regression, and Multilayer Perceptron (MLP), to predict the risk of CHD. Several important findings were discovered, including factors that individuals could benefit from to reduce the risk of CHD, and implications for businesses operating in the healthcare industry. By providing valuable and actionable insights into the risk factors associated with Coronary Heart Disease (CHD) and potential interventions, our models can help healthcare providers, insurance companies, and patients make more informed decisions about managing and reducing the risk of CHD. This can ultimately lead to reduced healthcare costs and improved patient health management.

Furthermore, our model's potential value to businesses in the healthcare industry becomes even more apparent when combined with a study conducted in the USA. According to the study, the annual incidence of Cardiovascular Disease (CVD) is estimated at over 600,000 cases. The total mean direct medical care costs for patients with established CVD were $17,532 per patient per year, with inpatient costs making up 42.8% $(7,503) of total expenses. By reducing the risk of CHD, our model can help healthcare businesses improve patient outcomes, reduce costs, and stay competitive in an increasingly data-driven market.

However, with the limited number of attributes, our model still has its limitations in classifying patients with a 10-year risk of CHD. Additionally, based on the results of coefficient degrees, we can assume that not one single factor is influencing the 10-year risk of CHD. Therefore, further research will be needed to explore various attributes that can potentially impact CHD risk to build a more robust model.

In conclusion, our machine learning-based models can serve as important technical tools to facilitate decision-making for healthcare industry stakeholders and potential cardiovascular disease (CVD) patients, helping to reduce the future healthcare costs and improve patient health management. In future work, with the possible addition of new attributes and more data, we can explore ways to leverage data and enhance machine learning techniques to further advance our goals.

8 Appendix - Data Dictionary

Variable Name	Description
Male	binary: “1”, means “Male”, “0” means “Female”
age	Age of the patient; Continuous - Although the recorded ages have been truncated to whole numbers, the concept of age is continuous
education	0: Less than High School and High School degrees, 1: College Degree and Higher
currentSmoker	whether or not the patient is a current smoker (binary: “1”, means “Yes”, “0” means “No”)
cigsPerDay	the number of cigarettes that the person smoked on average in one day. (can be considered continuous)
BPMeds	whether or not the patient was on blood pressure medication (binary: “1”, means “Yes”, “0” means “No”)
prevalentStroke	whether or not the patient had previously had a stroke (binary: “1”, means “Yes”, “0” means “No”)
prevalentHyp	whether or not the patient was hypertensive (binary: “1”, means “Yes”, “0” means “No”)
diabetes	whether or not the patient had diabetes (binary: “1”, means “Yes”, “0” means “No”)
totChol	total cholesterol level (Continuous)
sysBP	systolic blood pressure (Continuous)
diaBP	diastolic blood pressure (Continuous)
BMI	Body Mass Index (Continuous)
heartRate	heart rate (Continuous)
glucose	glucose level (Continuous)
TenYearCHD	10 year risk of coronary heart disease CHD (binary: “1”, means “Yes”, “0” means “No”)

Table 6: Description of Variables in the Dataset