Heart Disease Prediction Using Machine Learning: A Comprehensive Analysis of Personal Health Indicators

Introduction of Topic/Problem

Cardiovascular diseases (CVDs), especially heart disease, are the world's leading cause of death. The World Health Organization reported that CVDs caused approximately 20.5 million deaths in 2022, and experts expect this number to rise. The size of the public health crisis represented by cardiovascular disease emphasizes the need for an efficient, accessible method for early detection. Suppose we can identify individuals at high risk of heart disease before developing acute or chronic symptoms. In that case, healthcare professionals can provide various preventative strategies, suggest lifestyle changes, or take appropriate medical action. We can do all this to benefit the health of individuals, lessen the future heart disease burden through patient outcomes, and ultimately save lives.

This paper explores the application of predictive modeling based on machine learning algorithms to assess the probability of heart disease based on personal health indicators. Our primary research question is: **"Which machine learning algorithm is most accurate and reliable in predicting heart disease based on personal health indicators?"** The next step would be exploring a valuable model for clinicians assessing cardiovascular risk.

Overview of Data Used

The data for this analysis comes from the "Personal Key Indicators of Heart Disease" dataset. The Centers for Disease Control and Prevention (CDC) created this strong health data source and made it available on the Kaggle platform. The version of the dataset we used for this analysis is called heart_2022_no_nans.csv (the file is free of records with N/A's). Using the cleanest and most complete dataset possible to create accurate predictive models is essential in data science.

The dataset contains typical characteristics associated with **246,022 individual records** and **40 unique features**. We can group these features into the following main categories:

- **Demographics**: Features such as, for example, age and sex
- **Behaviors**: Features that describe behaviors, for example, physical activity and smoking
- **Clinical History**: Clinical history-related features, including whether the person has suffered from diabetes, stroke, and asthma historically.
- **Health Status**: self-rated overall health status

Collectively, these features developed a multi-faceted underpinning of the individual, and overall, are a valuable source of information to establish a predicted model to assess heart disease.

Methods of Analysis

We used various systematic and transparent analytic methods to answer the research question methodically, situating data analysis within a clear framework so others could reproduce and analyze it correctly. The analytic process followed four main phases.

- **Exploratory Data Analysis (EDA)**

  In the first analysis phase, we established EDA as a primary goal, as exploring and understanding the dataset was imperative. The EDA aimed to find underlying patterns, identify significant trends, and gain more understanding of the data structure, and involved:

  - **Distribution analysis:** We produced the distribution for the target variable, the presence and absence of heart disease, and investigated the distribution of critical demographic variables, such as age and sex, to understand the nature of conflict between heart disease cases and non-cases.
  - **Risk factor analysis:** We assessed the number of cases of known cardiovascular risk factors (such as a history of stroke, diabetes, or asthma) in the dataset to quantify the responses to well-known risk factors.

- **Correlation analysis:** A correlation matrix was produced for all numerical features to identify strong linear relationships among the features and understand possible multicollinearity challenges.
- **Statistical Significance Testing:** To go beyond simple observation, we performed Chi-square tests. We assessed the observed associations of categorical features (like "GeneralHealth") and heart disease through these statistical tests and evaluated their statistical significance beyond chance.

- **Data Pre-processing**

When we prepared the data, we undertook essential steps before using the data to train the ML models. The pre-processing step is fundamentally crucial because we want the models to learn as much as they can from our data, and the steps we took included the following:

- **Separation of Features and Target:** We separated the data into the features (39 predictor variables, represented as X) and the target variable (the 'HadHeartAttack' column, defined as y).
- **Encoding:** Machine learning models require data as numbers, so we coded all categorical variables as numbers. The binary target variable ('Yes'/'No') was label encoded as 1 and 0. We converted the multi-class categorical features through one-hot encoding, creating a new binary column for each category within a feature.
- **Scaling:** Some features had some numerical ranges larger than others, so to avoid features excessively impacting the model numbers, we standardized each numerical feature using the `StandardScaler`. The process results in data with a mean of 0 and a standard deviation of 1.
- **Train-Test Split:** We split the dataset into a train set (80% of the data) and a test set (20% of the data). We use the training set to "train" and develop the models, while the

test set allows a final and unbiased evaluation of the model's performance. We used stratified sampling to select members for the training and testing sets while maintaining the proportion of members with and without heart disease. It is an essential consideration for a dataset with class imbalance tendencies.

- **Model Building and Training**

We are ready to build and train the four machine learning models with the prepared data. The selection of the models was intentional in broadly covering a variety of algorithmic approaches for classifying the target outcome.

  - **Logistic Regression:** Researchers widely use this well-known linear model for its interpretability and reliable performance, often adopting it as a benchmark in medical research when assuming a linear relationship. Logistic regression is favored for its simplicity when presenting model results to decision-makers because it is clear and straightforward to understand which factors drove the predictions.
  - **Random Forest:** An ensemble learning method that generates many decision trees during training. It produces a single prediction through averaging all the individual decision tree predictions, which helps improve prediction quality, predict more robustly, and reduce variance from over-fitting.
  - **Support Vector Machine (SVM):** A powerful and flexible model that can capture even complicated non-linear relationships within the data. SVM's working principle is to find an optimal hyperplane separating the different data classes presented in the feature space.
  - **XGBoost (Extreme Gradient Boosting):** A more sophisticated and efficient implementation of the gradient boosting algorithm. It is so successful that one rarely

finds an implementation better than XGBoost, the only implementation that can run just as fast.

In a dataset like this, a key challenge is sample imbalance (i.e., there are far fewer people with heart disease than without heart disease). To minimize this, we set up the models to assign higher weight to the minority data (i.e., people with heart disease) as we trained them (e.g., we set the `class_weight` parameter when using the SVM to 'balanced').

- **Model Evaluation**

  The final and most critical stage was the evaluation of the trained models. We used a two-pronged approach:

  - **Cross-Validation:** We first performed a 5-fold stratified cross-validation on the training set. We split the training data into 5 "folds" and trained our model on four folds, then tested it with the last 5th fold, repeating that process 5 times. This method provides a better estimate of model performance than a single-train-test split.
  - **Holdout Test Set Evaluation:** We evaluated models on the completely unseen test set after cross-validation. It provides the final and accurate measure of predictive capability on new data.

  We used the following metrics to compare the models:

  - **Accuracy:** The overall percentage of correct predictions.
  - **Precision:** Of all the predictions for heart disease, how many were correct? A high precision indicates a low false positive rate.
  - **Recall (Sensitivity):** Of all cases of heart disease, how many did the model correctly predict? A high recall indicates a low false negative rate.
  - **F1-Score:** The harmonic mean of precision and recall. Whatever precision and recall may or may not be valuable, we have a single score that accommodates both.

- **ROC-AUC (Area Under the Receiver Operating Characteristic Curve):** A measure of a model's ability to tell the positive from the negative classes. An AUC of 1 is perfect, 0.5 is random guessing.

Results & Findings Explained

- **Model Performance**

A complete evaluation of the four machine learning models produced interesting results. All four models generally revealed reasonable performance in predicting heart disease; however, the model performance measures varied in different directions.

- The **Logistic Regression** model performed best overall and best for this task. It had the highest composite score based on all metrics' weighted mean and a strong **ROC-AUC** of **0.8857** on the test set. The ROC-AUC also indicated strong separability of positive and negative classes.
- Although the **Support Vector Machine (SVM)** model had the highest accuracy of **0.9479**, it had a critically low recall of **0.2170**. Thus, while the SVM was very good at correctly classifying individuals as **not** having heart disease, it correctly classified only 21.7% of individuals **with** heart disease. In clinical scenarios, a high level of false negatives is unacceptable. In terms of heart disease, a 78% false negative rate means a clinical system is missing opportunities to intervene for many at-risk patients.

The following table provides a detailed breakdown of each model's performance on the unseen test set:

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	0.8304	0.2109	**0.7678**	0.3309	**0.8857**
Random Forest	0.8298	0.2081	0.7540	0.3261	0.8809

SVM	**0.9479**	**0.5595**	0.2170	0.3127	0.8856
XGBoost	0.8363	0.2145	0.7499	0.3335	0.8830

Table 1: We evaluated the performance of machine learning models on the test set, highlighting the best scores for Recall and ROC-AUC.

- **Feature Importance**

A significant benefit of machine learning goes beyond simply forecasting results. It also helps identify the factors that impact those predictions most. The feature importance analysis provided clear and helpful information about this dataset's primary heart disease risk factors.

The most important predictors in each of the four models were the same set of characteristics:

1. The strongest predictor of a future heart attack was a self-reported history of angina, which is chest pain from reduced blood flow to the heart.
2. **Age Group:** Age significantly and steadily predicted heart disease risk, with older age groups strongly linked to positive diagnoses.
3. **Had Stroke:** A previous diagnosis of a stroke was another strong indicator of increased cardiovascular risk.
4. **General Health:** Individuals who rated their general health as "fair" or "poor" were significantly more likely to have heart disease compared to those who rated it as "excellent" or "very good."
5. **Difficulty Walking:** We identified self-reported difficulty in walking or climbing stairs as a key predictive feature. Likely indicates underlying physical limitations related to cardiovascular health.

Conclusion

This analysis shows the significant potential of machine learning for predicting heart disease early. We built and evaluated several predictive models using an extensive and detailed dataset of personal health indicators.

The **Logistic Regression** model stands out due to its strong performance, particularly its high recall and ROC-AUC score. Its interpretability makes it a clear choice for a model clinicians can use effectively in a clinical setting. Although other models like SVM achieved higher accuracy, their failure to identify enough actual positive cases makes them unsuitable for this type of medical application, where missing a diagnosis can have serious consequences.

Moreover, the study's findings on feature importance strengthen our understanding of cardiovascular risk. Identifying factors like a history of angina, age, and previous stroke as top predictors matches established medical knowledge and adds confidence in the model's validity. These insights can help healthcare professionals focus on the most critical risk factors during patient consultations and develop more targeted and effective prevention strategies.

The following steps involve exploring how to integrate this model into a clinical decision support system. Future research could also consider including a broader range of data, such as genetic markers or more detailed clinical measurements, to further improve the predictive accuracy of these models.

References

Centers for Disease Control and Prevention. (2022). Personal key indicators of heart disease - Kaggle. https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease