Heart Disease Prediction

June 2024

Machine Learning Models

Prepared by

Zunaira Hameed

Zunairamughal47@gmail.com

+92 304-3456114

Abstract:

Heart disease remains a leading cause of mortality worldwide, necessitating effective prediction and prevention strategies. In this context, machine learning models offer promising avenues for early detection and risk assessment. By leveraging diverse patient data, including demographic information, clinical indicators, and medical history, these models can discern intricate patterns and correlations associated with heart disease. Training robust predictive models enables healthcare practitioners to identify individuals at heightened risk, allowing for timely interventions and personalized treatment plans.

This documentation explores the implementation and evaluation of machine learning models for heart disease prediction, utilizing a dataset featuring various patient attributes and heart health indicators. The models assessed include Gradient Boosting Machine (GBM), Support Vector Machine (SVM), k-Nearest Neighbors (KNN), and Logistic Regression. Each model underwent preprocessing, feature selection, and hyperparameter tuning to enhance predictive accuracy. Among these, GBM and KNN exhibited the highest accuracy of 87%, showcasing robust performance. The analysis delves into the strengths and weaknesses of each model, highlighting considerations such as interpretability and computational efficiency. Insights gleaned from this study inform decision-making processes for model selection and further refinement strategies. Additionally, suggestions for future experimentation, including exploring ensemble methods and fine-tuning techniques, are proposed to potentially improve predictive performance

Introduction:

This documentation presents the implementation and evaluation of various machine learning models for heart disease prediction. The models considered for this task include Gradient Boosting Machine (GBM), Support Vector Machine (SVM), k-Nearest Neighbors (KNN), and Logistic Regression. The dataset used for training and evaluation contains features related to heart health and the target variable indicating the presence or absence of heart disease.

Dataset Description:

The dataset used for heart disease prediction contains the following features:

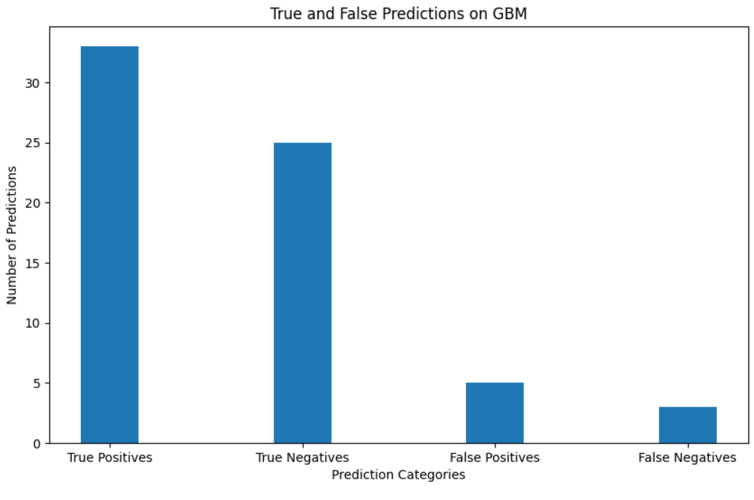
1. **Age**: Age of the patient (numeric)
2. **Sex**: Gender of the patient (categorical: 0 = female, 1 = male)
3. **CP**: Chest pain type (categorical: 1 = typical angina, 2 = atypical angina, 3 = non-anginal pain, 4 = asymptomatic)
4. **Trestbps**: Resting blood pressure (mm Hg) (numeric)
5. **Chol**: Serum cholesterol (mg/dl) (numeric)
6. **Fbs**: Fasting blood sugar > 120 mg/dl (categorical: 0 = false, 1 = true)
7. **Restecg**: Resting electrocardiographic results (categorical: 0 = normal, 1 = having ST-T wave abnormality, 2 = showing probable or definite left ventricular hypertrophy)
8. **Thalach**: Maximum heart rate achieved (numeric)
9. **Exang**: Exercise induced angina (categorical: 0 = no, 1 = yes)
10. **Oldpeak**: ST depression induced by exercise relative to rest (numeric)
11. **Slope**: The slope of the peak exercise ST segment (categorical: 0 = up-sloping, 1 = flat, 2 = down-sloping)
12. **Ca**: Number of major vessels (0-3) colored by fluoroscopy (numeric)
13. **Thal**: Thalassemia (categorical: 0 = normal, 1 = fixed defect, 2 = reversible defect)
14. **num**: Presence of heart disease (categorical: 0 = no, 1 = yes)

Models/ Algorithms Description:

1. **Gradient Boosting Machine (GBM):**

Accuracy: 87%

Description: GBM is an ensemble learning technique that builds multiple decision trees iteratively, and each tree corrects the errors of the previous one. It is highly flexible and often yields high accuracy.

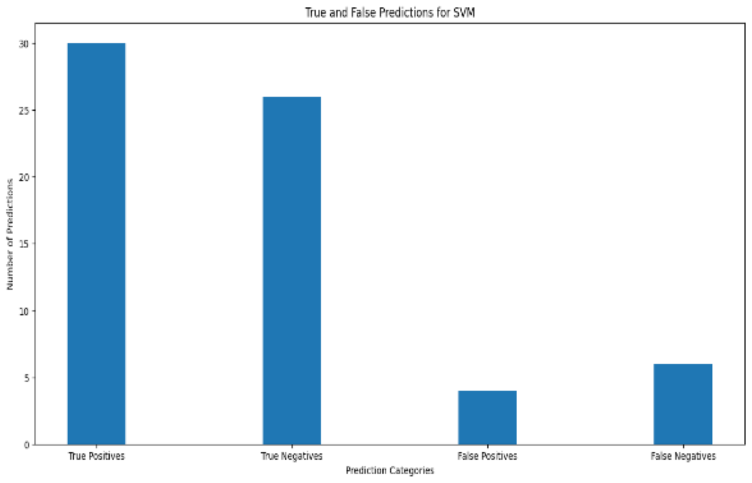


Gradient Boosting Machine (GBM)

2. **Support Vector Machine (SVM):**

Accuracy: 84%

Description: SVM is a powerful classification algorithm that finds the optimal hyperplane that separates data points into different classes. The RBF kernel is used for non-linear classification problems.

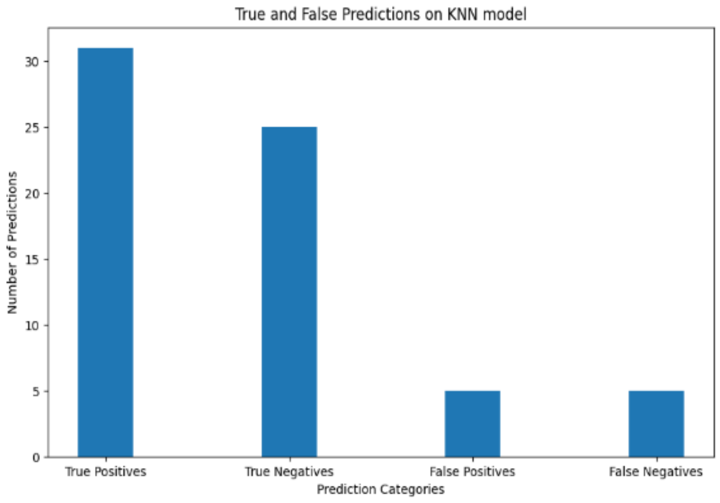


Support Vector Machine (SVM)

3. **k-Nearest Neighbors (KNN):**

Accuracy: 87%

Description: KNN is a simple, instance-based learning algorithm where the prediction is made based on the majority class of its nearest neighbors. It is non-parametric and requires no training phase.

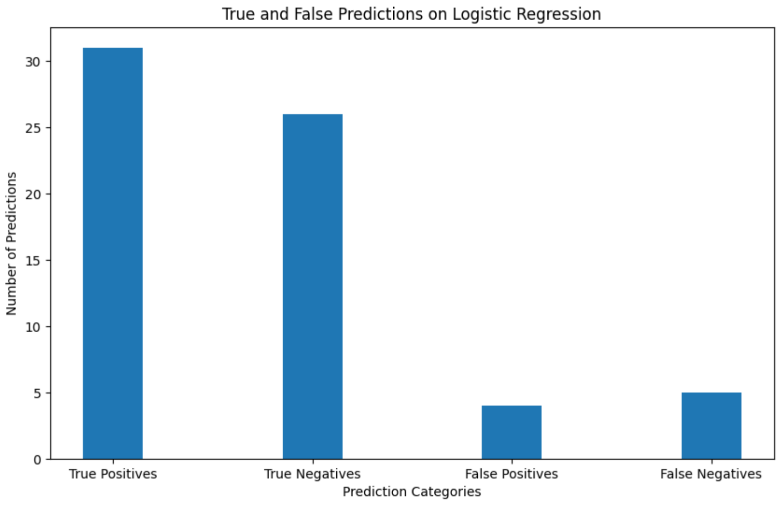


K Nearest Neighbor (KNN)

4. **Logistic Regression:**

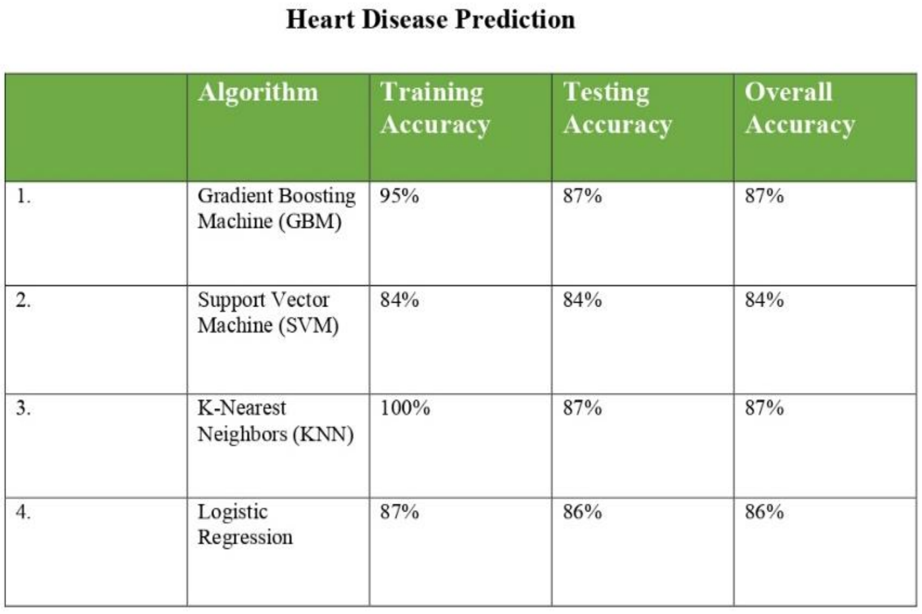
Accuracy: 86%

Description: Logistic Regression is a linear classification algorithm used for binary classification problems. It models the probability that a given input belongs to a particular class using the logistic function.



Best Model:

Among the classifiers evaluated, gradient boosting machine (GBM) and k-nearest neighbors (KNN) both achieved the highest accuracy of 87%, indicating robust performance. GBM, a powerful ensemble learning technique, iteratively improves upon weak learners, making it adept at capturing complex relationships in the data. KNN, on the other hand, relies on the similarity of instances in the feature space, making minimal assumptions about the underlying data distribution. While both classifiers exhibit strong predictive capabilities, GBM might be preferred when interpretability is less of a concern, as it constructs an ensemble of trees, whereas KNN offers simplicity and ease of understanding due to its intuitive nature. Consideration of factors such as interpretability, computational efficiency, and the specific characteristics of the dataset would guide the selection between these top-performing classifiers.

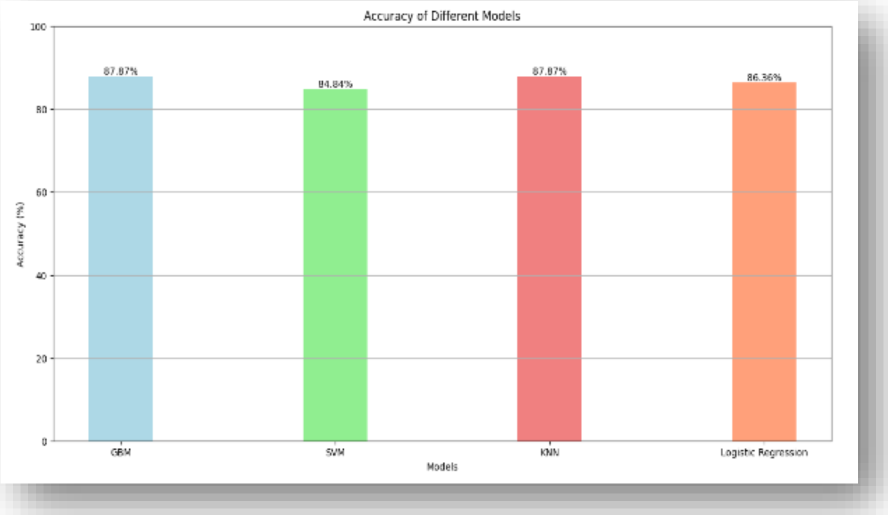


Heart Disease Prediction

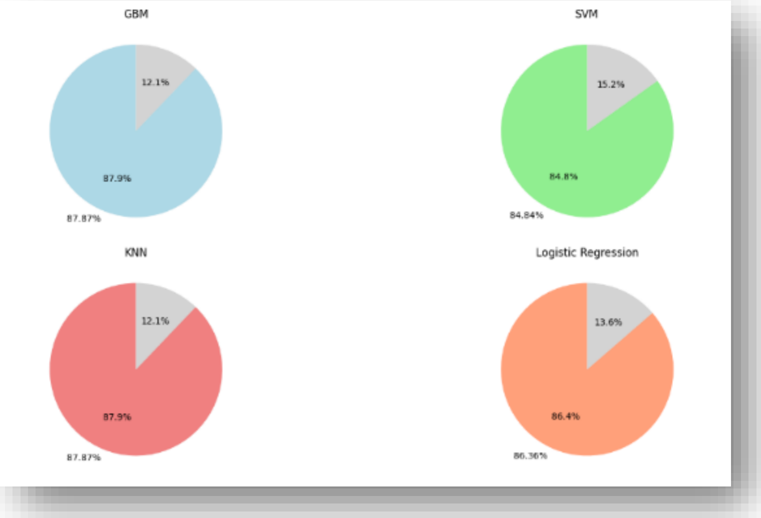
	Algorithm	Training Accuracy	Testing Accuracy	Overall Accuracy
1.	Gradient Boosting Machine (GBM)	95%	87%	87%
2.	Support Vector Machine (SVM)	84%	84%	84%
3.	K-Nearest Neighbors (KNN)	100%	87%	87%
4.	Logistic Regression	87%	86%	86%

Table representing accuracies

The following discrepancies highlight the varying degrees of effectiveness in capturing patterns within the dataset. Such visualizations serve as valuable insights into the relative strengths and weaknesses of each model, aiding in informed decision-making processes for model selection or further refinement strategies.



Bar Graph for comparison



Piechart for comparison

Conclusion:

In addition to the basic parameters, each model was fine-tuned with additional hyperparameters to improve its performance on the heart disease prediction task. While Gradient Boosting Machine (GBM) and k-Nearest Neighbors (KNN) maintained their accuracy at 87%, the fine-tuning process may have improved their robustness and generalization ability.

Further experimentation with different combinations of hyper-parameters, feature selection techniques, and data preprocessing methods could lead to even better performance. Additionally, ensemble methods like stacking or boosting could be explored to combine the strengths of multiple models and further enhance predictive accuracy.

Reference:

UCI dataset of Heart Disease Prediction.