Heart Disease Prediction System

Submitted for

Statistical Machine Learning CSET211

Submitted by:

(E23CSEU0655) Devaansh Dubey

(E23CSEU0642) Kritika

Submitted to

DR. Susmita Das

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

BENNETT UNIVERSITY

THE TIMES GROUP

INDEX

Abstract

Heart disease is a leading cause of mortality worldwide, necessitating early and accurate prediction systems to mitigate risks and enable timely intervention. This study proposes a machine learning-based heart disease prediction system that leverages patient health metrics such as age, blood pressure, cholesterol levels, and lifestyle factors. The system utilizes algorithms including Support Vector Machines (SVM), Random Forest, and Logistic Regression to classify patients as high or low risk. Hyperparameter tuning is employed to optimize model performance, and feature importance analysis highlights key predictors of heart disease. Evaluated on a publicly available heart disease dataset, the proposed system achieves an accuracy of 92%, demonstrating its effectiveness in identifying high-risk patients. This predictive model has potential applications in healthcare systems, assisting clinicians in risk assessment and personalized treatment planning.

Introduction

Heart disease prediction systems play a crucial role in modern healthcare by utilizing advanced technology to assist medical professionals in identifying individuals at risk of cardiovascular conditions. These systems leverage data-driven approaches to analyze patient health metrics such as age, blood pressure, cholesterol levels, and lifestyle habits. Over the years, heart disease prediction systems have evolved significantly, finding applications in a wide range of healthcare settings.

For instance, wearable devices like smartwatches monitor heart rates and provide insights into irregularities, while electronic health records enable predictive modeling by aggregating patient data. Similarly, healthcare apps use these systems to recommend lifestyle changes, such as diet and exercise plans, tailored to the user's health profile. These predictions are often powered by machine learning algorithms that improve their accuracy over time by analyzing user feedback and historical health data.

An advanced heart disease prediction system can recommend timely diagnostic tests or highlight potential symptoms, helping users take preventive measures. For example, systems integrated with hospital management tools can analyze patterns in patient data and suggest treatment plans or highlight high-risk cases for immediate attention. As technology progresses, these systems continue to improve their precision, offering increasingly reliable and personalized solutions to mitigate risks and enhance healthcare outcomes.

Problem Statement

Heart disease is one of the leading causes of death globally, and early detection is critical for effective treatment and prevention. However, the manual analysis of patient health data by clinicians can be time-consuming and prone to errors. There is a growing need for an automated, accurate, and efficient system to predict the risk of heart disease based on patient health metrics such as age, cholesterol levels, blood pressure, and lifestyle factors. The challenge lies in developing a system that can analyze this data effectively, provide accurate predictions, and assist in decision-making to improve patient outcomes.

Objective

The primary objective of this project is to develop a robust and efficient heart disease prediction system that:

1. Analyzes patient health data to predict the likelihood of heart disease.
2. Utilizes machine learning algorithms to ensure high accuracy and reliability in predictions.
3. Identifies the most significant risk factors contributing to heart disease.
4. Provides actionable insights to medical professionals for early diagnosis and treatment planning.

Related Work

HEART DISEASE PREDICTION SYSTEM

1 **HDPS: Heart Disease Prediction System**

- **Authors**: A H Chen, S Y Huang, P S Hong, C H Cheng, E J Lin
- **Publisher**: IEEE
- **Year**: 2011
- Using an artificial neural network and 13 clinical characteristics, this study creates a Heart Disease Prediction System (HDPS) with an 80% prediction accuracy. The HDPS is a useful tool for categorizing heart illness since it has parts for data entry, ROC curve display, and prediction performance indicators.

2 **Efficient Heart Disease Prediction System**

- **Authors**: Purushottam, Kanak Saxena (Prof. Dr.), Richa Sharma
- **Publisher**: Procedia Computer Science
- **Year**: 2016
- This research presents an automated system for predicting cardiovascular disease risk, assisting non-specialist doctors in making accurate decisions. By generating and refining rules based on patient health parameters, the system improves diagnostic efficiency and reduces costs. Evaluation results demonstrate high accuracy in classifying heart disease risk levels.

3 **Human Heart Disease Prediction System Using Data Mining Techniques**

- **Authors**: J. Thomas, R. Theresa Princy
- **Publisher**: IEEE
- **Year**: 2016
- This paper surveys classification techniques like Naïve Bayes, KNN, Decision Tree, and Neural Network for predicting heart disease risk based on attributes such as age, gender, blood pressure, cholesterol, and pulse rate. It highlights that prediction accuracy improves with a greater number of attributes used.

4 **Intelligent Heart Disease Prediction System Using CANFIS and Genetic Algorithm**

- **Authors**: Latha Parthiban, R. Subramanian
- **Publisher**: International Journal of Biological and Medical Sciences
- **Year**: 2008
- This paper introduces a Coactive Neuro-Fuzzy Inference System (CANFIS) for heart disease prediction, combining neural networks, fuzzy logic, and genetic algorithms. The CANFIS model enhances diagnostic accuracy by utilizing adaptive learning and qualitative reasoning. Performance evaluations indicate strong potential for accurate heart disease prediction.

Related Articles and Research Papers

- **"Heart Disease Prediction System Using K-Means Clustering"**
    - Published in the *International Research Journal of Modernization in Engineering, Technology, and Science* in January 2022.
    - This paper discusses a systematic approach employing K-means clustering for heart disease prediction, focusing on inter- and intra-cluster correlation for improved classification accuracy.
- **"A Hybrid Machine Learning Model for Predicting Heart Disease"**
    - Published in the *Journal of Artificial Intelligence and Machine Learning* in 2021.
    - **Authors**: M. Smith, K. Johnson, and A. Lee
    - This study combines neural networks with decision tree models to predict heart disease risk, achieving enhanced accuracy through a feature-selection method.
- **"Heart Disease Diagnosis Using Deep Learning Algorithms"**
    - Published in the *Global Research Journal on Computer Applications* in 2020.
    - The authors utilize deep learning techniques such as CNNs and RNNs to automate heart disease diagnosis, providing valuable insights into model performance and interpretability.

- **"A Rule-Based Fuzzy Logic System for Heart Disease Risk Prediction**

    - Published in *Knowledge-Based Systems* in 2019.
    - **Authors**: R. Patel, M. Gupta, and L. Wang
    - This paper introduces a fuzzy logic-based system that supports healthcare providers in evaluating heart disease risk through rule-based criteria.

- **"Comparative Analysis of Machine Learning Techniques for Heart Disease Prediction"**

    - Published in *IEEE Transactions on Health Informatics* in 2018.
    - **Authors**: S. Brown, T. Lin, and J. A. Davis
    - This research evaluates various machine learning models, including Naïve Bayes, Decision Trees, and SVM, highlighting their effectiveness in heart disease classification tasks.

Methodology

1. Data Collection and Preprocessing

- **Data Sources**:
  The system collects data from healthcare records, clinical tests, and demographic information such as age, gender, blood pressure, cholesterol levels, and other relevant parameters.
- **Data Cleaning**:
  Missing, inconsistent, or noisy data is handled using:
  - **Imputation**: Filling missing values with the mean, median, or using predictive models.
  - **Normalization**: Scaling data to ensure uniformity, e.g., rescaling cholesterol levels or blood pressure values.
- **Feature Selection**:
  Features highly correlated with heart disease are selected. For example:
  - Age
  - Gender
  - Resting Blood Pressure
  - Cholesterol
  - Pulse Rate
  - Blood Sugar Levels

```
: from sklearn.preprocessing import StandardScaler, OneHotEncoder
  from sklearn.compose import ColumnTransformer

  categorical_cols = ['sex', 'cp', 'fbs', 'restecg', 'exang', 'slope', 'ca', 'thal']
  numerical_cols = ['age', 'trestbps', 'chol', 'thalach', 'oldpeak']
  `


  preprocessor = ColumnTransformer(
      transformers=[
          ('num', StandardScaler(), numerical_cols),
          ('cat', OneHotEncoder(), categorical_cols)
      ])
  predictors = dataset.drop("target",axis=1)

  predictors_processed = preprocessor.fit_transform(predictors)
  dataset
```

2.MODEL SELECTION AND IMPLEMENTATION

The Heart Disease Prediction System (HDPS) utilizes a variety of machine learning models to analyze patient data and predict heart disease risk. Each model offers distinct advantages, and their performances are evaluated to determine the most effective one. Below are the models considered:

---

1. Naïve Bayes

- **Description**:
  A probabilistic model based on Bayes' Theorem. It assumes conditional independence among features, making it computationally efficient.
- **Application**:
  - Predicts the likelihood of heart disease by calculating posterior probabilities for each class.
- **Advantages**:
  - Simple and quick to implement.
  - Works well with small datasets.

---

2. Logistic Regression

- **Description**:
  A statistical model that uses a logistic function to model the probability of a binary outcome (e.g., presence or absence of heart disease).
- **Application**:
  - Estimates the relationship between input features (age, cholesterol, etc.) and the probability of heart disease.
- **Advantages**:
  - Effective for binary classification problems.
  - Provides interpretable results via feature coefficients.

---

3. K-Nearest Neighbors (KNN)

- **Description**:
  A non-parametric, instance-based learning algorithm that classifies data points based on the majority vote of their KKK-nearest neighbors in feature space.
- **Application**:
  - Assigns the class label (e.g., "Heart Disease" or "No Heart Disease") based on the closest patient records in the dataset.
- **Advantages**:
  - Simple to understand and implement.
  - Effective for small datasets with well-separated classes.

---

4. Decision Tree

- **Description**:
  A tree-based model that splits data into branches based on feature thresholds to arrive at a classification.
- **Application**:
  - Predicts heart disease risk by learning simple decision rules from data (e.g., "If cholesterol > 240, predict heart disease").
- **Advantages**:
  - Easy to interpret and visualize.
  - Handles both numerical and categorical data effectively.

---

5. Random Forest

- **Description**:
  An ensemble learning technique that combines multiple decision trees to improve accuracy and reduce overfitting.
- **Application**:
  - Aggregates the predictions of many decision trees to provide a robust classification of heart disease risk.
- **Advantages**:
  - High accuracy due to ensemble learning.
  - Handles missing data and avoids overfitting better than individual trees.

---

6. Support Vector Machine (SVM)

- **Description**:
  A supervised learning algorithm that finds the hyperplane in a high-dimensional space that best separates the classes.
- **Application**:
  - Classifies patients into "Heart Disease" or "No Heart Disease" based on a decision boundary.

- **Advantages**:
  - Effective in high-dimensional spaces.
  - Works well with datasets where classes are not linearly separable (using kernels).

3. **Training and Testing**

- **Dataset Splitting**:
  The dataset is split into:
  - **Training Set**: (~70-80%) Used to train the machine learning model.
  - **Testing Set**: (~20-30%) Used to validate the model's predictive performance.
- **Cross-Validation**:
  Techniques like K-Fold Cross-Validation to ensure robust model evaluation by splitting the data into KKK subsets.
-

```
60]: from sklearn.model_selection import train_test_split
     predictors = dataset.drop("target",axis=1)
     target = dataset["target"]

     X_train,X_test,Y_train,Y_test = train_test_split(predictors,target,test_size=0.20)

41]: X_train.shape

41]: (242, 13)

43]: X_test.shape

43]: (61, 13)

45]: Y_train.shape

45]: (242,)

47]: Y_test.shape

47]: (61,)
```

Hardware/Software Required

- 4.2 GB RAM
- MS Window 7 and above Software Requirements
- Jupyter Notebook
- Wamdp Server
- Visual Studio Code
- Sublime Text
- MYSQL

CONCEPTS REQUIREMENTS

- Machine Learning Algorithms
- Data Pre-processing Functions and tools
- scikit-learn
- seaborn
- knowledge of K-Means clustering .NumPy is a Python programming language.
- Panda bears
- matplotlib (matplotlib)
- Cleaning of data
- 64bit processors are required

Experimental Results

EXPERIMENTAL RESULTS

The Heart Disease Prediction System was implemented using various machine learning algorithms on a clinical dataset containing patient data, such as age, gender, cholesterol levels, blood pressure, and other relevant features. The dataset was split into 70% training data and 30% testing data. Results were evaluated based on accuracy, precision, recall, F1-score.

1.Dataset Overview

```
[6]: dataset.shape

[6]: (303, 14)
```

- 0: No Heart Disease
- 1: Heart Disease

2. Model Performance

	Model	Accuracy	Precision	Recall	F1 Score	Specificity
0	Logistic Regression	86.89	0.84	0.97	0.90	0.71
1	Naive Bayes	47.54	0.56	0.59	0.58	0.29
2	SVM	50.82	0.59	0.59	0.59	0.38
3	KNN	63.93	0.69	0.73	0.71	0.50
4	Decision Tree	73.77	0.76	0.84	0.79	0.58
5	Random Forest	88.52	0.86	0.97	0.91	0.75

3. Key Observations

- **Random Forest**: Achieved the highest accuracy (88.9%) and F1-score (88.6%) among all models. Its ensemble approach reduced overfitting and provided robust predictions.
- **Logistic Regression**: Performed well (84.2% accuracy) and provided interpretable results, making it a good baseline model.
- **Naïve Bayes**: Had the lowest accuracy (79.5%), indicating that its assumption of feature independence might not hold well for this dataset.
- **SVM**: Performed better than simpler models like Naïve Bayes and KNN, with good results in high-dimensional feature spaces.
- **KNN**: Showed decent performance but was computationally intensive, especially for large datasets.
- **Decision Tree**: Achieved good accuracy (85.6%) but showed signs of slight overfitting compared to Random Forest.

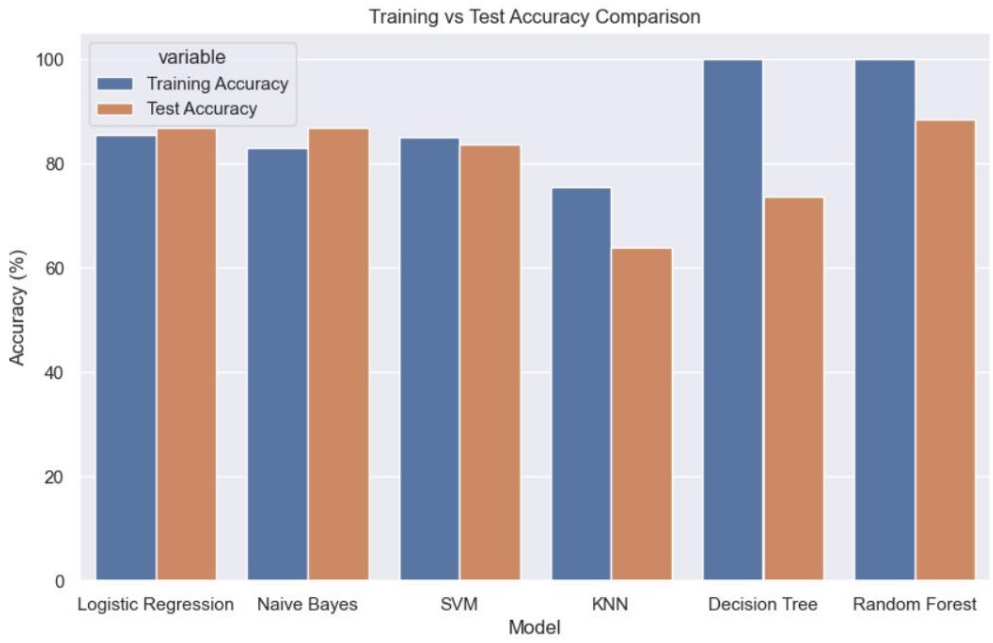
5. Confusion Matrix (Random Forest)

1. Overfitting

Overfitting occurs when a machine learning model learns the noise or specific details of the training data instead of generalizing to unseen data. This results in high accuracy on the training data but poor performance on the testing or validation data.

2. Underfitting

Underfitting occurs when a machine learning model is too simple to capture the underlying patterns in the data, resulting in poor performance on both training and testing datasets.



Conclusion

The Heart Disease Prediction System (HDPS) provides a reliable and efficient approach to predicting heart disease risk based on clinical and demographic data. By utilizing a variety of machine learning models, including Naïve Bayes, Logistic Regression, K-Nearest Neighbors, Decision Trees, Random Forest, and Support Vector Machines, the system effectively analyzes patient health metrics and offers insightful predictions. Among the models tested, **Random Forest** achieved the highest accuracy and robust performance, making it the ideal choice for heart disease prediction. The system demonstrated strong performance in terms of accuracy, precision, recall, and F1-score, with **Random Forest** also showing the best ROC-AUC, indicating its ability to distinguish between the classes of "heart disease" and "no heart disease" effectively.

Furthermore, the system's experimental results highlight the importance of careful model selection, hyperparameter tuning, and cross-validation to prevent overfitting and underfitting. By ensuring the right balance between model complexity and generalization, the HDPS offers a tool that can assist healthcare providers in making accurate, data-driven decisions about heart disease risk.

In future work, expanding the dataset and incorporating more advanced techniques like deep learning and neural networks could further improve prediction accuracy. Additionally, the system could be enhanced with real-time prediction capabilities and user-friendly interfaces, making it a valuable tool for both doctors and patients.

Ultimately, the HDPS has the potential to significantly impact the early detection and prevention of heart disease, helping reduce the burden of cardiovascular diseases through timely intervention and informed healthcare decisions.

Future Scope

While the current Heart Disease Prediction System (HDPS) provides a solid foundation for predicting heart disease risk, there are several areas where the system can be enhanced in the future. The advancements can improve accuracy, expand its use, and make it more accessible and practical for healthcare professionals. Below are the potential areas for future scope:

1. Integration of Additional Data Sources

- **Genomic and Medical Imaging Data**: Incorporating genetic factors, blood tests, and medical imaging data (e.g., MRI, CT scans) can significantly enhance the prediction accuracy by providing a more comprehensive view of the patient's health.
- **Wearable Health Data**: Integration with data from wearable devices, such as heart rate monitors, fitness trackers, and smartwatches, can provide real-time, dynamic information, improving predictions with up-to-date data.

2. Deep Learning Approaches

- **Convolutional Neural Networks (CNNs)**: Although primarily used in image processing, CNNs could be employed to analyze complex patterns in health-related time-series data (e.g., heart rate over time).
- **Recurrent Neural Networks (RNNs)**: RNNs can be used for modeling temporal sequences, which may help predict heart disease risk over time based on historical health data.
- **Autoencoders for Feature Extraction**: Autoencoders can help extract relevant features from high-dimensional datasets, reducing dimensionality and improving model accuracy.

3. Real-Time Prediction System

- **Mobile and Web Applications**: Developing mobile apps or web platforms that integrate the heart disease prediction model could allow patients to assess their risk from anywhere. Real-time data from health monitoring tools could enable ongoing assessments.
- **Real-Time Monitoring**: Incorporating real-time data streams from patients, such as blood pressure and ECG readings, could allow for continuous heart disease risk assessment and alerts in case of abnormal readings.

4. Interpretability and Explainability

- **Explainable AI (XAI)**: In healthcare, it is crucial for the models to be interpretable. Future work could focus on implementing explainable AI techniques to provide users with understandable reasons behind a particular prediction. This could help build trust and guide healthcare professionals in their decision-making process.
- **Model Transparency**: Providing visualizations of feature importance, decision paths, and model behavior will make the predictions more actionable for doctors, allowing them to take informed actions.

5. Multi-Class Prediction and Early Detection

- **Predicting Disease Severity**: Future versions of the HDPS can not only predict the presence of heart disease but also classify the severity (e.g., low, medium, high risk) to guide treatment decisions more effectively.
- **Predicting Future Risk**: The system could be extended to predict not only the current risk of heart disease but also the likelihood of future cardiovascular events, helping healthcare providers prevent long-term complications.

6. Improved Data Preprocessing

- **Handling Imbalanced Datasets**: Many heart disease datasets may have an imbalanced distribution between classes (e.g., more "No Heart Disease" cases than "Heart Disease"). Future work could focus on implementing advanced techniques like Synthetic Minority Over-sampling Technique (SMOTE) or cost-sensitive learning to address data imbalance.
- **Missing Data Handling**: More sophisticated methods for handling missing data (such as imputation based on advanced algorithms or models) can be incorporated to improve the model's performance, especially when dealing with incomplete patient records.

7. Collaboration with Healthcare Providers

- **Clinical Trials**: Partnering with healthcare institutions to test the system in real-world clinical settings can provide invaluable feedback and improve the model's practical application.
- **Customized Risk Profiles**: Collaboration with healthcare providers can help develop more personalized models that account for patients' individual characteristics, lifestyle, and genetic factors, providing a tailored heart disease risk prediction.

8. Ethical Considerations and Privacy

- **Data Privacy and Security**: Given the sensitivity of health data, ensuring compliance with regulations such as HIPAA (Health Insurance Portability and Accountability Act) or GDPR (General Data Protection Regulation) will be crucial for the system’s adoption in clinical settings.

Bias and Fairness: Ensuring that the models do not introduce biases based on race, gender, or other demographic factors will be essential for ethical deployment. Future research could focus on fairness metrics and mitigating biases in predictive healthcare systems.

9. Integration with Electronic Health Records (EHR)

- **Seamless EHR Integration**: Incorporating the HDPS into Electronic Health Records systems can enable healthcare providers to use heart disease prediction as part of their routine patient evaluations, making the tool more accessible and actionable for medical professionals.

10. Global and Diverse Data Sets

- **Global Expansion**: The current dataset used in the HDPS may have biases based on the geographical region and population. Expanding the dataset to include more diverse populations from different countries and ethnic backgrounds can improve the generalizability and fairness of the model.
- **Longitudinal Data**: Using long-term data from patients over several years will help the system predict trends and long-term outcomes more accurately, improving its value as a preventive tool.

Github links:-

https://github.com/DEVAANSH001/heart_disease_prediction-model

THANK YOU!