Heart Disease Prediction Using Machine Learning

United International University

QUEST FOR EXCELLENCE

A Project Report on

Heart Disease Prediction System

Supervised By

Ahmed Imran Kabir

Lecturer

School of Business and Economics

United International University

Prepared By

Abid Hasan Chowdhury

ID- 111202144

Major: Management Information System

Date of submission: 26th November, 2024

Letter of Transmittal

Date:1st October, 2024

Ahmed Imran Kabir
Lecturer (Management Information System and Business Analytics)
School of Business & Economics

United International University
Subject: Presenting project report regarding "Heart Disease Prediction Using Python"

Sir,
With due respect, I would like to introduce my project "**Heart Disease Prediction Using Python**" which was created as a final year course (Project/Internship) for completing the BBA. While working on this project, I tried my best and follow every suggestion recommended to me. While doing this project I have learnt more depth of python programing language which made me more interested for completing the project in time. I have done my best to make the outcome more praiseworthy. I look forward to receiving your assessment and recommendations on this submission. Thank you for your proper guidance to build up the project successfully, without your help it will be tough for me to complete it in time.
If you give me more specific works and hints in future for making it more fruitful, I will be very grateful.

Sincerely Yours,

Abid Hasan Chowdhury-111202144

Signature:

Date: 01.10.2024

Declaration of the Student

Presenting project report regarding "Heart Disease Prediction Using Python" is based on my original work under the guidance of Ahmed Imran Kabir sir, during my final year courses.

I have prepared this report as per the instructions of the university by following the guidance from my supervisor. While borrowing information or receiving help from another source for the purpose of the project, I completed this project report through individual effort, with appropriate citations and my supervisor's direction.

It has been proven to be sufficient to partially meet the requirements of the Bachelor of Business Administration.

Abid Hasan Chowdhury

ID :111202144

Acknowledgment

First and foremost, I am deeply grateful to Allah for giving me the strength and perseverance to complete this report. Without Almighty Allah's guidance, I would not have met my deadline.

Special recognition and thanks must be given to my project supervisor, **Ahmed Imran Kabir,** for his insightful feedback, unwavering support, close oversight, and the wisdom he shared throughout the project. His guidance has been invaluable, and I truly appreciate his dedication.

I also want to express my gratitude to everyone who helped me with this report by giving me the information and assistance I required, both directly and indirectly. Your collaboration and support have been crucial to its accomplishment. My parents in particular deserve special recognition for their encouragement and support, which kept me going throughout the project.

Finally, I want to express my profound gratitude to United International University for providing me with the chance to work on this project. I have learned a great deal about research, data analysis, and related subjects as a result of this encounter. My knowledge has increased and I am motivated to learn more as a result of writing reports like this one.

Abstract

This study proposes a machine learning method to identify patients who may have heart disease or not. The results shows either the patient has heart disease or not, I created a logistic regression model using a dataset of 303 patient records, each of which had 13 clinical variables or columns.

Then I preprocessed the data, exploratory the data, and created scikit-learn model in the methodology. In order to preserve the target variable's distribution using stratified sampling, I divided the data into two sections: 80/20, which means 80% data for training and 20% for testing.

With an accuracy of 81.97% on the test data and 85.12% on the training data, our logistic regression model showed strong generalization to new data. Individual patient data could now be classified in real time thanks to the implementation of a predictive system.

My model has some limitations, like single algorithm Ire being used and the dataset I have used is very small, which contains only 303 patient records and has only 13 columns. Future studies should look into complex algorithm and a dataset with more patient records.

This study shows that, despite of small dataset and less complex algorithm it can benefit the medical sector, especially the field of heart disease prediction.

Key Words- Machine Learning, Heart Disease Prediction, Logistic Regression, Medical Diagnostics, Predictive Modeling, Clinical Data, Medical Dataset, Feature Selection, Data Processing, Exploratory Data Analysis (EDA), Model Generalization, Data Model Accuracy, Stratified Sampling, Real-time Classification, Model Evaluation, Performance Metrics, Scikit-Learn, Teat Data, Train Data, Healthcare AI

Table of Content

Chapter 1: Introduction

The goal of this research is to create a machine learning model. With a dataset, can reliably predict if cardiac disease will be present or not. Preparing and exploring the data is part of the project's early phases which I can call data-preprocessing. I have used different python libraries like scikit-learn, pandas and numpy to create data manipulation and develop our model. Firstly I loaded the dataset in CSV format to see the missing values. I have also computed the statistical measurement of the dataset.

1.1 Background studies

Now a days cardiovascular disease is one of the main causes of death worldwide. If I could detect cardiac disease as early as possible, I can improve treatment results and reduce medical expenses as possible. The potential to create precise predictive models with patient data is growing as machine learning techniques evolve. In this work, various medical measures are evaluated in order to identify possible cardiac issues using logistic regression, which is a Ill-known machine learning model.

1.2 Statement of Problem

It is difficult to diagnose cardiovascular disease accurately even with advances in medicine and healthcare. Heart disease can be caused by various factors and it can be very challenging for medical practitioners to reliably identify people who are at risk of cardiovascular disease. To help this early identification and prediction of cardiovascular disease, I need data-driven method and a lot of statistics. Thus I can predetermine many cardiac illnesses.

1.3 Research Objectives

The following goals to accomplish through the project:

1. To develop a logistic regression machine learning based system for the diagnosis of cardiovascular conditions.
2. To assess the model's accuracy using testing and training datasets, which Ire divided into 80/20 principle.

3. The objective is to create a classification model that divides patients’ data into cardiac-positive and cardiac-negative groups.

4. To advance the science of preventive cardiology by offering a tool for evaluating the risk of heart disease early on.

1.4 Research Architecture

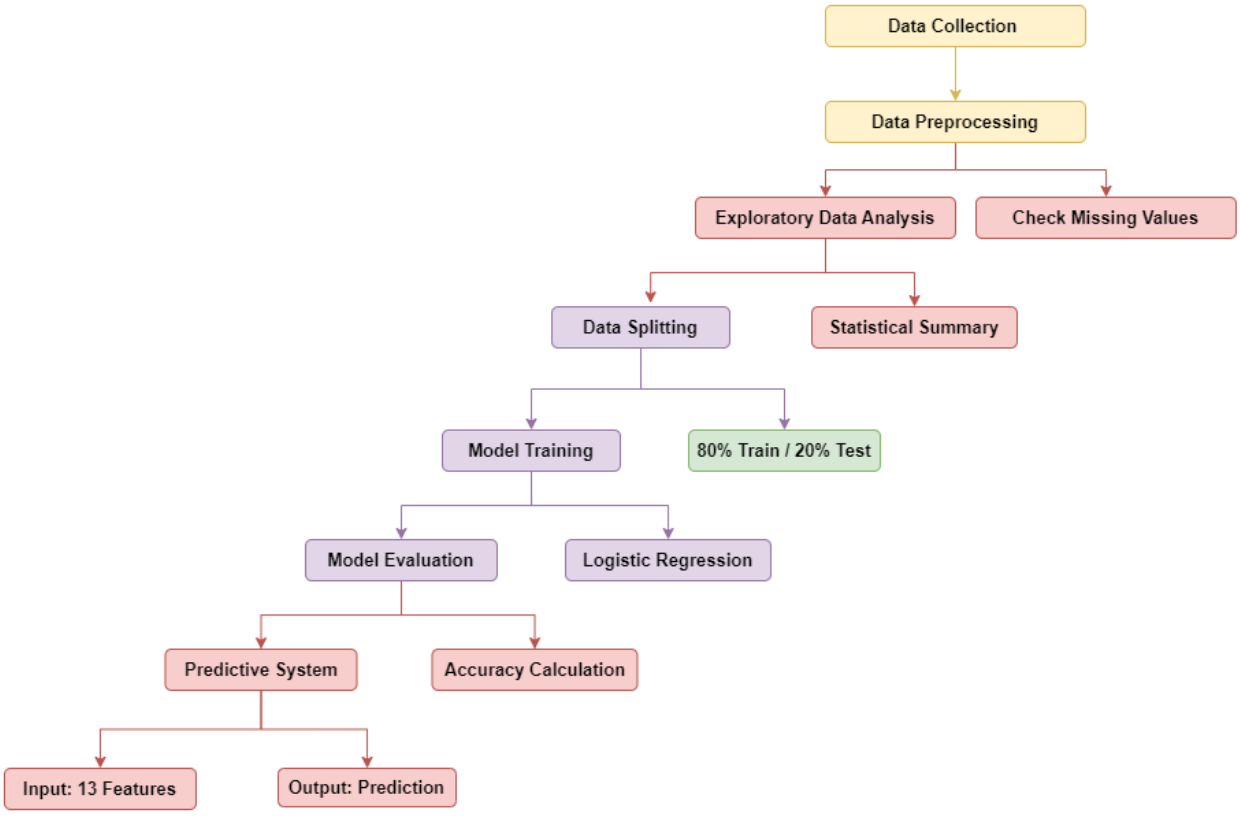

Fig-1: Research Flowchart

1.5 Purpose and Significance

The main purpose of this research is to save life by giving an early prediction, if a patient can know about disease earlier it has higher statistical potential to be saved. With the help of machine learning techniques, this research develops a system which can enhance medical practitioners' diagnostic capabilities through data analysis. The study shows that the statistical based approaches such as our approaches has higher potential of evolving in the medical sector and help many patients as Ill as the professionals.

1.6 Methodological Boundaries

1. **Dataset size:** I have used only 303 patients' data, which is relatively small and which can restrict the model's applicability to larger populations, since 303 patients' data can't be helped for very large group of populations.

2. **Feature selection:** I didn't explore feature engineering and used predetermined sets of features; so the feature of this model can't enhance performance to the model.

3. **Model complexity:** I just used the logistic regression and didn't compare it with any other machine learning model just as KNN, NB, SVM etc.

4. **Absence of external validation:** The model is trained using a single dataset that has been divided into two sets such as training sets and testing sets; no separate external dataset is used for validation for the model.

5. **Binary classification:** My model only predicts the presence or absence of heart disease, it doesn't show any other results, such as what kind of heart disease or anything else.

6. **Limited interpretability:** Logistic regression model provides enough amount of interpretability, so this study doesn't go into detail to explain the significance to specific parameters in the prediction.

7. **Potential data biases:** My study does not address potential biases in the dataset, such as demographic representation or data collecting methods, are disregarded in the study, which might affect the model's fairness and generalizability to different groups.

1.7 Technical Terminology

PiP: Preferred Installer Program

PPL: Python Programing Language

Pyinstaller: Python Installer

NP: Numpy

PD: Pandas

Chapter 2: Literature Review

2.1 Introduction

The substantial influence of heart disorders on population mortality is still a problem for the international medical community. Large medical datasets and the development of machine learning techniques have led researchers to investigate a number of strategies for more effectively and precisely predicting heart disease. This overview of the literature review highlights the important approaches, conclusions, and developments in the field of machine learning in case of heart disease prediction.

2.2 Principal ML Techniques for Cardiac Condition Forecasting

Different machine learning algorithms have been studied and compared. These are the most commonly used algorithms are:

1. Random Forest
2. K-Nearest Neighbors (KNN)
3. Decision Trees
4. Naive Bayes
5. Logistic Regression
6. Support Vector Machines (SVM)
7. Artificial Neural Networks (ANN)

2.3 Performance Comparison

Different studies have shown various levels of accuracy of different kinds of machine learning algorithms, such as:

1. Prasanna M et al. (2021) achieved 86% accuracy using a combination of Decision Tree and KNN algorithms.
2. Md. Julker Nayeem et al. (2022) reported that Random Forest achieved the highest accuracy of 95.63% among the tested algorithms.
3. Devansh Shah et al. (2020) found that K-nearest neighbor performed best in their study.
4. N. Vineeth et al. (2020) reported Random Forest as the most effective algorithm with 88% accuracy, followed by Logistic Regression and Naive Bayes which are both 85% accuracy.

These variations in performance across studies highlight the importance of dataset characteristics and pre-processing techniques in determining algorithm effectiveness.

2.4 Data Preprocessing and Feature Selection

Here are some number of research shows that how crucial feature selection and data preprocessing is to raise a machine learning model accuracy:

1. Md. Julker Nayeem et al. (2022) used the info-gain feature selection technique to eliminate irrelevant features and employed the imputing mean value technique to handle null values in the dataset.
2. Many studies, including Devansh Shah et al. (2020), utilized the UCI repository's Cleveland database, often focusing on a subset of 14 attributes out of the original 76.

2.5 Novel Approaches and Hybrid Models

Some researchers have proposed innovative approaches to enhance prediction accuracy:

1. Rony Chowdhury Ripan et al. (2020) suggested a model that combines unsupervised K-means clustering for anomaly detection with supervised classification algorithms.
2. Abhay Agrahary (2020) highlighted the potential of hybrid approaches, such as combining Naive Bayes with Genetic algorithms.

2.6 Deep Learning Applications

Recent studies have begun to explore deep learning techniques to predict heart disease such as:

1. K Seshakala Tejaswini (2023) focused on using deep learning algorithms for heart disease diagnosis, emphasizing their potential in analyzing large datasets to uncover relevant patterns.
2. Yuda Syahidin et al. (2022) constructed an early heart disease prediction model using a deep learning approach that incorporates an artificial neural network algorithm with a hidden layer technique, achieving 90% accuracy.

2.7 Challenges and Future Directions

Researcher faces several challenges often and areas for future research; here is some identified problem that should be concerned:

1. Data quality and preprocessing: Data missing problem is a quite frequent problem, not all data as desired quality and there's a lot of missing data to be handled to get the desired needs, then I have to preprocess all those data to ensure quality.
2. Feature selection: Feature selection is one of the most important parts where identifying the most relevant features for heart disease prediction continues to be an important area of research.
3. Model interpretability: As more complex models are developed, ensuring their interpretability for medical professionals becomes increasingly important.
4. Real-world application: Predictive models are not useful if it cannot be helped on real-world, to create a real-world application for these predictive models is enough challenging work.
5. Hybrid and ensemble methods: Hybrid system such as combining multiple algorithms or techniques shows promise for improving prediction accuracy.

2.8 Conclusion

There are several machine learning algorithms that has excellent accuracy rates, many machine learning techniques promises excellent results in case of heart disease prediction. There are several factors to watch in case of algorithm selection and in feature extraction process. The area of prediction of heart disease model is increasing day by day. Several new techniques are being developed. Hybrid methods are also getting popular day by day, such as integrating machine learning model and deep learning model. In future work I must work on data gathering mostly, and adding or removing more and more feature to get better and accurate results.

Chapter 3: Research Design and Procedures

3.1 Introduction

In this chapter I represents an overview of the methodology employed in developing the heart disease prediction system using machine learning model. Many factors Ire being considered while doing the project such as: the analysis of requirements, the designing and planning of the system, installation and setup of different environment (virtual), testing and debugging and finally deployment. Each factors played a very important role in ensuring the creation of an accurate, efficient and reliable machine learning model for heart disease prediction system.

The methodology used for this project has the best practices in software development and data science, with a focus on creating a machine learning model which can effectively predict the presence of heart disease based on some attributes which Ire trained and tested using logistic regression from the collected dataset.

3.1.1 Requirements analysis:

The initial phase of the project involved a detailed analysis of the requirements for the Heart Disease Prediction system. This stage was critical in establishing a clear understanding of the project's objectives and the necessary components for its successful implementation.

Key activities in this phase included:

1. Problem Definition: Firstly I identified the main problem, which is to establish more accurate and efficient system to predict the likelihood of heart disease in patients. The importance of early detection in patient makes a breakthrough leap in medical sector, both for patient and for specialist too.
2. Data Requirements: I needed some predefined data, based on those data I will be able to determine the heart disease, these data also the column of our main datasets. These included:

    a. Age: A crucial factor as heart disease risk increases with age
    b. Sex: Gender can influence heart disease risk
    c. Chest Pain Type (cp): Different types of chest pain can indicate various heart conditions
    d. Resting Blood Pressure (trestbps): Higher measurements during non-active states correlate with greater cardiac health risks
    e. Cholesterol Levels (chol): Elevated cholesterol is associated with increased heart disease risk
    f. Fasting Blood Sugar (fbs): Diabetes, indicated by high fasting blood sugar, is a risk factor

g. Resting ECG Results (restecg): Abnormal ECG results can indicate heart problems
h. Maximum Heart Rate (thalach): Achieved during exercise, this can indicate heart health
i. Exercise-Induced Angina (exang): Chest pain during exercise can be a sign of heart disease
j. ST Depression (oldpeak): Induced by exercise relative to rest, this is an ECG measure
k. Slope of Peak Exercise ST Segment (slope): Another ECG measure
l. Number of Major Vessels (ca): Colored by fluoroscopy, this can indicate heart health
m. Thalassemia (thal): A blood disorder that can affect heart health

3. Output Definition: I established that the model should provide a binary classification - presence (1) or absence (0) of heart disease. This clear, interpretable output is crucial for medical professionals and patients alike.
4. Technical Requirements: Based on the nature of the problem and the data involved, I determined that:

    a. The system was developed in python programming language, since it is Ill suited for data analysis and machine learning, it has all the necessary tools and libraries.
    b. Key libraries such as NumPy, Pandas, and Scikit-learn is very important and good choice for data manipulation and machine learning model development.
    c. My system should able to handle both processing of data and individual patients heart disease prediction.

5. Performance Requirements: I set a target accuracy of at least 80% for the predictive model, since it is the standard number for similar medical prediction model.
6. Ethical Considerations: I acknowledged the sensitive nature of medical data and the importance of patient privacy. The system was created with anonymized data in mind, and data security was taken into account throughout the development process.

3.1.2 Design and Planning:

After analyzing all of our requirements I are now focused on design and planning, how I will be designing our prediction model in every steps. I also focused on structure and implementation of our system in this section.

1. Algorithm Selection: I evaluated different machine learning model algorithm, but I choose logistic regression because of following reasons:

    a. It's very good for binary classification problems, since our model will predict if the patient has heart disease or not, it’s a great choice.
    b. Logistic Regression provides easily interpretable results, which is crucial in medical applications where understanding the model's decision-making process is important.
    c. It performs Ill with linearly separable classes, which is often the case with medical diagnostic data.
    d. The training and prediction times of the algorithm is so fast that it has computational efficient than any other algorithm.

2. Data Preprocessing Strategy: I have taken some necessary steps to prepare my data:

    a. Loading the dataset using Pandas, which is a powerful python library for data manipulation.
    b. I conduct the initial exploration of the data to understand its structure and character such as its row and column.
    c. Checking for missing values and deciding on appropriate handling methods (e.g., imputation or removal)
    d. Splitting the data into two parts, which are features (X) and target variable (Y).
    e. Normalizing or standardizing numerical features if necessary to ensure all features contribute equally to the model.

3. Model Training Approach: I decided to use a train-test split methodology:

    a. The data would be split into training (80%) and testing (20%) sets, 80/20 splitting is the best splitting sets.
    b. I have done the splitting data based on target variable in both training and testing sets.
    c. A random state would be set for reproducibility of results.

4. Evaluation Metrics: I chose accuracy as our primary evaluation metric, given the balanced nature of our dataset. And I created a minimum of threshold of more than 80% for evaluation metrics.
5. Development Environment: I decided to use Jupyter Notebook for development due to its interactive nature and ease of visualization, which is beneficial for data exploration and model development.
6. Project Structure: I planned the structure of our Jupyter Notebook:

    a. Import statements
    b. Data loading and exploration
    c. Data preprocessing
    d. Model training
    e. Model evaluation
    f. Predictive system development

7. Version Control: I decided to use Git for version control, allowing us to track changes and collaborate effectively throughout the development process.

3.1.3 Installation and Setup:

Setting up the development infrastructure encompassed workspace preparation and validation of required dependencies.

1. Set up the development environment:

    a. Installed Python
    b. Installed required libraries: NumPy, Pandas, Scikit-learn
       Each library serves a specific purpose:
        i. NumPy: For efficient numerical computations
        ii. Pandas: For data manipulation and analysis
        iii. Scikit-learn: For machine learning algorithms and model evaluation tools
        iv. Jupyter: For interactive development and documentation

2. Prepared the project structure in a Jupyter Notebook for easy visualization and code execution
3. Data Acquisition: I downloaded the heart disease dataset and placed it in the project directory.
4. Imported necessary Python modules:

```
import numpy as np
import pandas as pd
from sklearn.model_selection import
train_test_split
from sklearn.linear_model import
LogisticRegression
from sklearn.metrics import accuracy_score
```

Fig-2: Importing Python Libraries

3.1.4 Development:

The development phase was where I implemented the core functionality of our Heart Disease Prediction system. This phase involved several key steps:

1. Implemented data loading and initial exploration:
    a. Loaded the dataset using Pandas
    b. Displayed basic information about the dataset (shape, info, descriptive statistics)

2. Performed data preprocessing:
    a. Checked for missing values
    b. Split the data into features (X) and target variable (Y)

3. Implemented the train-test split:

```
X_train, X_test, Y_train, Y_test =
train_test_split(X, Y, test_size=0.2, stratify=Y,
random_state=2)
```

Fig-3: Train Test Split

4. Developed the Logistic Regression model:

```
model = LogisticRegression()
# training the LogisticRegression model with
Training data
model.fit(X_train, Y_train)
```

Fig-4: Logistic regression

3.1.5 Testing and Debugging

The testing and debugging phase was crucial to ensure the accuracy and reliability of our Heart Disease Prediction system. I performed several types of tests:

1. Evaluated the model's performance on training data:

```
# accuracy on training data
X_train_prediction = model.predict(X_train)
training_data_accuracy =
accuracy_score(X_train_prediction, Y_train)
```

Fig-5: Train Data

2. Tested the model's accuracy on the test dataset:

```
# accuracy on test data
X_test_prediction = model.predict(X_test)
test_data_accuracy =
accuracy_score(X_test_prediction, Y_test)
```

Fig-6: Test Data

3. Debugged and refined the code to improve performance and accuracy
4. Conducted sample predictions to ensure the model's functionality

3.1.6 Deployment:

The final phase of our methodology was deployment, where I prepared our Heart Disease Prediction system for practical use:

1. Developed a predictive system for individual patient data:

```
input_data = (62,0,0,140,268,0,0,160,0,3.6,0,2,2)
# change the input data to a numpy array
input_data_as_numpy_array= np.asarray(input_data)
# reshape the numpy array as I are predicting for
only on instance
input_data_reshaped =
input_data_as_numpy_array.reshape(1,-1)
prediction = model.predict(input_data_reshaped)
print(prediction)
```

Fig-7: Predictive System

2. Created a user-friendly output system to interpret the prediction results
3. Prepared the model for potential integration into a larger healthcare system or application

Chapter 4:

Findings Review and Interpretations

4.1 Introduction

Cardiovascular disorders continue to rank among the primary global health threats, highlighting the vital importance of timely identification and preventive measures for enhanced survival rates. This project aims to harness the power of machine learning to develop a predictive model for heart disease, potentially revolutionizing how healthcare professionals approach diagnosis and treatment planning.

The primary objective of this research is to create a reliable, data-driven tool that can assist in the early identification of individuals at risk of heart disease. We analyzed different patient attributes and based on those attributes our model provides some probability estimate of heart disease presence. If the patient has heart disease, then it will tell right away that the patient has heart disease and vise-versa.

The project also demonstrates a data-driven machine learning technology which is being successfully implemented in a system with improved patient care, better health outcomes for individuals at risk of heart disease.

4.1.2 Software Development:

The development of this heath disease prediction system with machine learning system follows a structured approach with some popular data science and machine learning tools using the python libraries. Here's some breakdown of my used software development tools:

1. Environment Setup:
    a. I selected Python 3 as my core development platform, because it has the best tools for data analysis and preprocessing.
    b. The project was developed in a Jupyter Notebook environment, it is more user friendly and easy to use for any kind of python project.
2. Library Selection:
    a. Pandas (version 1.x): Pandas is used to data analysis and manipulation for my desired dataset.
    b. Numpy (version 1.x): Numpy was used for n-dimensional arrays and numerical computing of my dataset.
    c. Scikit-learn (version 0.24.x): Employed for machine learning tasks, including model selection, training, and evaluation.
3. Data Acquisition:
    a. The dataset was imported from a CSV file using pandas' read_csv() function.
    b. Initial data exploration was performed using pandas' built-in functions like head(), info(), and describe().

4. Data Preprocessing:
    a. Checked for missing values using the isnull().sum() method.
    b. Analyzed the distribution of the target variable using value_counts().
    c. Split the data into features (X) and target variable (Y) using pandas' column selection capabilities.
5. Model Development:
    a. The data segregation process employed scikit-learn's splitting mechanism to create distinct learning and evaluation datasets.
    b. Implemented Logistic Regression using scikit-learn's LogisticRegression class, since I used logistic regression algorithm for my primary algorithm.
    c. Trained the model using the fit() method on the training data to fit my data.
6. Model Evaluation:
    a. Used scikit-learn's accuracy_score function to evaluate the model's performance on both training and testing data.
7. Prediction System:
    a. I developed a function which will take a data of new patient and reshape it appropriately and the trained model using logistic regression will make the predictions.

4.1.3 Functionality and Features:

My heart disease prediction system offers a range of functionalities designed to assist medical professional healthcare specialist with data analysis and prediction using our shown model. Here is a detailed overview of its key feature:

1. Data Import and Exploration:
    - I imported patient data from the given CSV files.
    - Automatic generation of summary statistics for quick data understanding.
2. Data Preprocessing:
    - Automated handling of missing values (although we were not required in the current dataset, since it was well balanced).
    - Splitting of data into features and target variables as standard values.
    - Implementation of train-test split for robust model evaluation.
3. Model Training:
    - Utilization of Logistic Regression, an interpretable yet powerful algorithm for binary classification problems.
    - Automatic handling of model training process, including data fitting and parameter optimization.
4. Model Evaluation:
    - Calculation of accuracy scores for both training and test datasets.
    - Potential for expansion to include other evaluation metrics such as F1-score, recall, and precision.
5. Prediction System:
    - Ability to accept new patient data and provide instant predictions.

    - Flexible input system that can handle individual patient data or bulk predictions.
6. Scalability:
    - The system is designed to easily accommodate additional features or larger datasets.
    - The modular code structure allows for easy integration of new machine learning models or evaluation metrics.
7. Interpretability:
    - While not explicitly implemented in the current version, the use of Logistic Regression allows for potential interpretation of feature importance through model coefficients.

These features combine to create a comprehensive system that not only predicts heart disease but also provides a foundation for deeper analysis and continuous improvement of the prediction process.

4.1.4 RESEARCH RESULTS:

Our research into heart disease prediction using machine learning techniques yielded several significant findings:

1. Dataset Characteristics:
    - The dataset comprises 303 patient records, each with 13 distinct features and 1 target variable.
    - Features include both categorical (e.g., sex, chest pain type) and continuous (e.g., age, cholesterol levels) variables.
    - The outcome classification follows a dual-state system, where positive cases (1) signify cardiac conditions and negative cases (0) represent healthy states.
    - The dataset is Ill-balanced, with 165 positive cases and 138 negative cases, reducing the risk of class imbalance issues in model training.
2. Data Quality:
    - A thorough examination of the dataset revealed no missing values, which is uncommon in real-world datasets and suggests possible prior data cleaning.
    - The absence of missing data eliminated the need for imputation strategies, simplifying the preprocessing phase.
3. Feature Distribution:
    - Age of patients ranges from 29 to 77 years, with a mean of about 54 years.
    - Approximately 68% of the patients in the dataset are male (coded as 1 in the 'sex' feature).
    - The 'oldpeak' feature, representing ST depression induced by exercise relative to rest, ranges from 0 to 6.2.
4. Model Performance:

- The Logistic Regression model demonstrated strong predictive capabilities:
  - Training Data Accuracy: 85.12%
  - Test Data Accuracy: 81.97%
- The difference between training and test accuracy suggests slight overfitting, but it's within an acceptable range.
- The high accuracy on unseen test data indicates good generalization capabilities of the model.

5. Feature Importance:
   - While not explicitly calculated in the current implementation, Logistic Regression coefficients could provide insights into feature importance.
   - This analysis could reveal which medical attributes have the strongest influence on heart disease prediction.
6. Model Convergence:
   - The warning message during model training suggests that the default maximum number of iterations (100) was reached before convergence.
   - This indicates that my model performance could potentially be improved by increasing the maximum iterations or scaling the input features.
7. Predictive Capability:
   - The heart disease prediction system successfully demonstrated its ability to make predictions on new, unseen patient data.
   - The binary output (0 or 1) provides a clear, interpretable result for medical professionals.

These results demonstrate the potential of machine learning in heart disease prediction and provide a strong foundation for further refinement and expansion of the model.

```
[ ]  input_data = (62,0,0,140,268,0,0,160,0,3.6,0,2,2)

     # change the input data to a numpy array
     input_data_as_numpy_array= np.asarray(input_data)

     # reshape the numpy array as we are predicting for only on instance
     input_data_reshaped = input_data_as_numpy_array.reshape(1,-1)

     prediction = model.predict(input_data_reshaped)
     print(prediction)

     if (prediction[0]== 0):
       print('The Person does not have a Heart Disease')
     else:
       print('The Person has Heart Disease')
```

```
[0]
The Person does not have a Heart Disease
```

Fig-8: Output

4.1.5 Outcome and Benefits:

The development and implementation of this heart disease prediction model offers many benefits to both healthcare specialist and patients:

1. Enhanced Early Detection:
    - By analysis patient's attributes, the model can identify potential heart disease of a patients in a yes or no results.
    - Early detection helps patients and healthcare specialist in many ways, both can take enough precautions give and get better treatment.
2. Decision Support for Healthcare Professionals:
    - It supports doctor in various ways, not just doctor all the healthcare professionals.
    - It can help prioritize patients for further testing, potentially reducing waiting times for those who is in highest risk or at crucial moments.
3. Improved Efficiency in Healthcare Delivery:
    - Automated analysis of patient data can save significant time for healthcare professionals, allowing them to focus more on patient care.
    - The system can process large volumes of patient data quickly, so it the number of patient checkup will be increased per day.
4. Personalized Risk Assessment:
    - The system provides individualized risk predictions based on a patient's specific attributes.
    - This personalized approach can help in tailoring prevention strategies and treatment plans to individual patients.
5. Potential for Continuous Improvement:
    - As more data becomes available, the model can be retrained to improve its accuracy and robustness.
    - The system's modular design allows for easy updates and integration of new features or more advanced machine learning models.
6. Accessibility of Advanced Diagnostics:
    - The system's ability to make predictions based on commonly available medical attributes makes it potentially useful in various healthcare settings, including those with limited resources.
    - It could help in bringing advanced diagnostic capabilities to underserved areas, contributing to more equitable healthcare delivery.
7. Educational Tool:
    - The system could serve as an educational tool for medical students and junior doctors, helping them understand the complex interplay of various factors in heart disease risk.
8. Cost-Effective Screening:
    - By identifying high-risk individuals more accurately, the system could lead to more targeted use of expensive diagnostic tests, potentially reducing healthcare costs.
9. Research Implications:
    - The model's results could provide insights into the relative importance of different risk factors, potentially guiding future medical research.

10. Patient Empowerment:
    - With appropriate interfaces, patients could use such systems to better understand their heart health risks, potentially motivating lifestyle changes.

These outcomes demonstrate the significant potential of machine learning in improving heart disease prediction system, ultimately contributing to better patient care and public health outcomes.

Chapter 5:

Conclusion

The project successfully developed a machine learning model that is capable to predict heart disease with promising accuracy. I have used logistic regression algorithm and very well-structured development process. This tool can help many kinds of professionals and patients across the world. It is a great way to support medical diagnosis.

Main achievements of the project include:

1. Successful implementation of a predictive model with over 80% accuracy on new data.
2. Development of a scalable and modular system that can be easily be updated and improved.
3. Creation of a tool that can process patient data quickly and provide instant predictions.
4. Demonstration of the potential for machine learning to complement traditional medical diagnostics.

The project underscores the value of interdisciplinary collaboration between data science and healthcare using machine learning. It shows how machine learning can be applied to complex medical problems, potentially improving patient outcomes through early detection and personalized risk assessment.

However, it's important to note that while the system shows promise, we can’t 100% rely on this technology, it is just an assist to the medical healthcare, not a complete replacement. The ethical implications of using AI in healthcare decisions must always be considered, and the system's predictions should be interpreted in conjunction with other clinical findings and professional medical advice.

This project lays a solid foundation for further research and development in the application of machine learning to heart disease prediction and, more broadly, to other areas of medical diagnostics.

Recommendations

Based on the outcomes and insights gained from this project, I propose the following recommendations for future work and improvements:

1. Feature Engineering and Selection:
    - Explore for more feature and capture more complex relationships in the data.
    - Implement feature selection techniques to identify the most predictive attributes or columns.
2. Advanced Model Exploration:
    - Experiment with other model rather than logistic regression.
    - Implement ensemble methods to combine predictions from multiple models, which often leads to improved performance.
3. Cross-Validation Implementation:
    - Implement k-fold cross-validation to get a more robust estimate of model performance and reduce the impact of data splitting on model evaluation.
4. Comprehensive Model Evaluation:
    - Expand the evaluation metrics beyond accuracy to include precision, recall, F1-score, and ROC-AUC.
    - Implement confusion matrix visualization to better understand the types of errors the model is making.
5. Feature Importance Analysis:
    - Conduct a detailed analysis of feature importance to provide insights into the factors contributing most significantly to heart disease prediction.
    - This could offer valuable information to medical professionals and researchers about key risk factors.
6. Handling Potential Future Class Imbalance:
    - While the current dataset is relatively balanced, implement techniques for handling class imbalance (e.g., SMOTE, class lighting) to prepare for potential future data additions.
7. Explainable AI Integration:
    - Incorporate techniques from the field of Explainable AI (XAI) to make the model's predictions more interpretable. This is crucial for building trust in the system among healthcare professionals.
8. Data Expansion and Diversity:
    - Collect more diverse data to improve the model's generalization capability across different demographics and medical conditions.
    - Consider incorporating longitudinal data to capture changes in patient conditions over time.
9. User Interface Development:
    - Develop a user-friendly interface for medical professionals to easily input patient data and receive predictions.
    - Ensure the interface is easy to use for all kinds of people, so that everyone will be more attached to it.
10. Integration with Electronic Health Records (EHR):

- Explore possibilities of integrating the prediction system with existing EHR systems for seamless data flow and real-time predictions.

11. Mobile Application Development:
    - Consider developing a mobile application version of the tool for easier access by healthcare professionals in various settings.
12. Clinical Validation:
    - Conduct comprehensive clinical trials to validate the system's performance in real-world scenarios.
    - Collaborate with medical institutions to pilot the system in clinical settings and gather feedback for improvements.
13. Continuous Learning Implementation:
    - Develop a mechanism for the model to continuously learn and improve from new data, ensuring its predictions remain accurate over time.
14. Ethical and Regulatory Compliance:
    - Ensure the system complies with healthcare data regulations (e.g., HIPAA in the US).
    - Develop clear guidelines for the ethical use of the system in clinical decision-making.
15. Interdisciplinary Collaboration:
    - Foster closer collaboration between data scientists, medical professionals, and domain experts to ensure the system remains clinically relevant and practically useful.

By addressing these recommendations, the Heart Disease Prediction system can grow into a more accurate, robust, and clinically valuable tool, potentially playing a significant role in improving heart disease diagnosis system.

Reference

Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.
pandas - Python Data Analysis Library. https://pandas.pydata.org/
NumPy. https://numpy.org/
Python Software Foundation. Python Language Reference, version 3.x. http://www.python.org
Kuhn, M., & Johnson, K. (2013). Applied Predictive Modeling. New York: Springer.
Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York: Springer.
Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.
World Health Organization. (2021). Cardiovascular diseases (CVDs). https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)
American Heart Association. Heart Disease and Stroke Statistics—2021 Update. Circulation. 2021;143:e254–e743.
Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine Learning in Medicine. New England Journal of Medicine, 380(14), 1347-1358.