Heart Disease Prediction System

COMP463 Introduction to Medical Informatics

Batuhan Birinci 042001047

Meryem Rana Türker 042001021

MEF University, Engineering Faculty, Computer Engineering

2024

Abstract

Heart disease is among the top single causes of death worldwide; thus, the need for its early detection and intervention cannot be overestimated. This report presents a study developed with machine learning techniques on the Heart Disease Prediction System that analyzes individual risk profiles. In this study, accuracies as high as 91% have been achieved by using algorithms like Random Forest Classifier, Logistic Regression, and Gradient Boosting on the UCI Heart Disease Dataset. Further, the system is extended by implementing a Flask-based web interface that is allowed to interactively present the user with predictions and suggestions for personalized recommendations. It features different preprocessing of data, feature selections, and evaluation strategies with a view to ensuring robustness and the reliability of results. The different visualizations-features' importance chart, and ROC chart-offer better insights regarding model performance and decision-making processes. Obtained results are agreed upon by the current application of machine learning approaches toward risk assessment and extend the scientific corpus pertaining to the predictive analytics in the healthcare sector.

Keywords: Heart Disease Prediction, Data Analysis, Medical AI, Machine Learning, Logistic Regression, Random Forest Classifier, Gradient Boosting, ROC Curves, Feature Selection, Data Preprocessing, Predictive Analytics, Cardiovascular Risk Assessment, Flask Web Interface.

Table Of Contents

1.Introduction

Cardiovascular diseases annually take approximately 17.9 million lives, more than most cancers combined. For a reduction in mortality rates, early intervention and predictive analytics become very crucial. Recent machine learning developments enable automated and accurate risk assessment, providing clinicians with much-needed early diagnosis tools. The paper focused on the development of a predictive model, using openly available data, and deploying it through a user-friendly web interface.

Heart diseases are, therefore, complex in their etiology, from genetic predisposition right through to lifestyle modification. Most conventional modes of diagnosis depend on clinical acumen, which, though salient, is nonetheless subjective and variable. Machine learning bridges the divide by offering objective, data-driven solutions able to analyze patterns and predict outcomes with a high degree of accuracy.

The system will fully make use of data preprocessing, feature engineering, and robust machine learning algorithms in the identification of major predictors of heart diseases. The major highlights included are an interactive web application that delivers smooth and user-friendly predictions and recommendations, hence making them accessible and useful in real life. In focusing on these objectives, the study will seek to contribute to the general field of medical AI by showing how predictive analytics could potentially alter healthcare outcomes.

Machine learning, in application to healthcare, has definitely changed the face of diagnosis and management of disease conditions, including cardiovascular ones. With various datasets, machine learning applied may unearth latent patterns and associations of risk factors which might have evaded conventional methods. Thus, this has empowered clinicians to move from reactive treatment-focused to a preventively and early-intervention-based mindset.

More importantly, when integrated into user-friendly interfaces, such systems democratize access to advanced diagnostic tools. Such tools may be used by general practitioners and by patients themselves for preliminary risk assessment, thereby reducing the burden on specialized healthcare. Aggregated data coming from these tools will go on to serve far-reaching usefulness in public health regarding policy decisions and resource allocations.

2. Background

Heart disease is a wide-ranging term that covers various diseases affecting the structure and functioning of the heart, which include coronary artery disease, arrhythmias, and congenital heart defects. According to the WHO, cardiovascular diseases still hold the first position as a cause of death in the world; therefore, there is an apparent need to develop effective measures for prevention and control. Other variables that raise the burden of heart diseases include an aging of the population, high intake of cholesterol, sedentary lifestyle, and rising rates of obesity and diabetes.

Some of the more traditional techniques of diagnosis are invasive and have inherent human subjectivities. Examples of such procedures include stress tests, angiograms, and echocardiograms, which are relatively time-consuming and costly to perform. In this regard, machine learning has brought a paradigm shift in this area by enabling the facility of non-invasive data-driven risk assessment and prediction.

This, of course, is where it becomes applicable in the realm of medicine: because they can analyze complex data and highlight any perceived trends or correlations, machine learning algorithms are particularly useful. It could make all the difference in the survival rate of patients, given that an early detection of heart disease or other conditions may be detected this way. This is a clear application of machine learning which will be practical in connecting advanced technology to cardiovascular health.

Big data also plays an important role in health care. This is an integration of many different population datasets and diversified healthcare systems to tune machine learning tools in understanding global variations in the trend and outcomes of cardiovascular diseases. The integrated electronic health records can also be used to continuously update the models dynamically for improved prediction over time.

The application of AI in healthcare already promises to reduce existing diagnostic errors and improve the care of patients. In such processes, AI-powered systems will continuously update their predictions and insights about the most current medical research and trends using millions of inputs. Such systems allow for the construction of personalized treatment plans according to each patient profile.

Heart disease is as much an economic challenge as it is a medical problem. The costs of treatment, lost productivity, and long-term care amount to thousands of rupees. Such a shift in approach from treatment to using machine learning for prevention will enable healthcare systems to make the most effective use of their resources. Predictive models may be relied upon to provide scalable, cost-effective solutions to enable early interventions that could ease the economic load on both individuals and governments.

3. Project Design and Development

3.1 Methodology

The heart disease prediction system deals with the UCI Heart Disease Dataset, which is composed of vital clinical attributes related to the patients' usual physical information: age, gender, cholesterol levels, and resting blood pressure. These will be used to predict the possibility of heart disease. Key steps in the methodology include:

Data Collection and Preparation

The data gets downloaded from UCI, preprocessed-handling missing values, outliers, and inconsistencies-and then did an exploratory data analysis using visualization techniques to understand the distribution of various features, their relationships, and spot the major pattern and trends that indicated strong correlations among different features.

```
data=pd.read_csv('Heart_Disease_Prediction.csv')
data.head()
```

	index	Age	Sex	Chest pain type	BP	Cholesterol	FBS over 120	EKG results	Max HR	Exercise angina	ST depression	Slope of ST	Number of vessels fluro	Thallium	Heart_Disease
0	0	70	1	4	130	322	0	2	109	0	2.4	2	3	3	Presence
1	1	67	0	3	115	564	0	2	160	0	1.6	2	0	7	Absence
2	2	57	1	2	124	261	0	0	141	0	0.3	1	0	7	Presence
3	3	64	1	4	128	263	0	0	105	1	0.2	2	1	7	Absence
4	4	74	0	2	120	269	0	2	121	1	0.2	1	1	3	Absence

Figure 1. UCI Heart Disease Dataset

Feature Engineering

Features are divided into numerical and categorical variables. Then, the scaling of numerical features is done by Standard Scaler. For encoding categorical features, techniques of label encoding are used. Feature selection techniques such as correlation analysis have been used to choose the most influential variable.

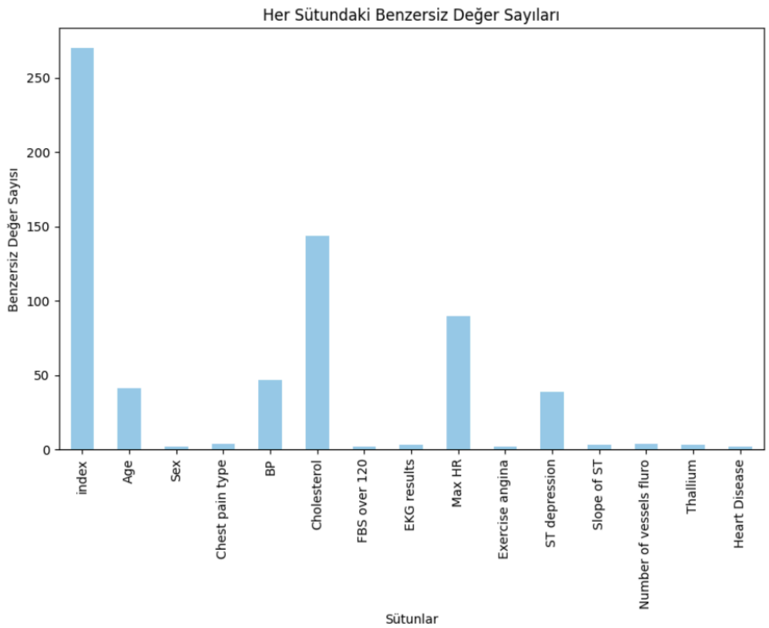

Figure 2. UCI Heart Disease Dataset Categorical Features

Model Training and Validation

The various models in machine learning were then run, which included Logistic Regression, Random Forest, and Support Vector Machines. Logistic Regression has been considered because it had given much better accuracy and can also be interpreted well. It was checked for performance by applying the cross-validation technique to ensure the model strength.

```
Accuracy: 0.9074074074074074
ROC-AUC Score: 0.948051948051948
              precision    recall  f1-score   support

     Absence       0.91      0.94      0.93        33
    Presence       0.90      0.86      0.88        21

    accuracy                           0.91        54
   macro avg       0.91      0.90      0.90        54
weighted avg       0.91      0.91      0.91        54
```

Figure 3. Model Training Results

System Integration

It was developed into a Flask web application where the model does data preprocessing and requests predictions on the backend, and then provides a user-friendly interface on the frontend. So, much testing and debugging had to be done to make the interaction smooth.

3.2 System Architecture

The heart disease prediction system consists of several dependent modules. Each module may have its own purpose, but all the modules are integrated into one large scalable system. These modules work together to perform three main tasks: data processing, generating predictions, and delivering results in understandable form.

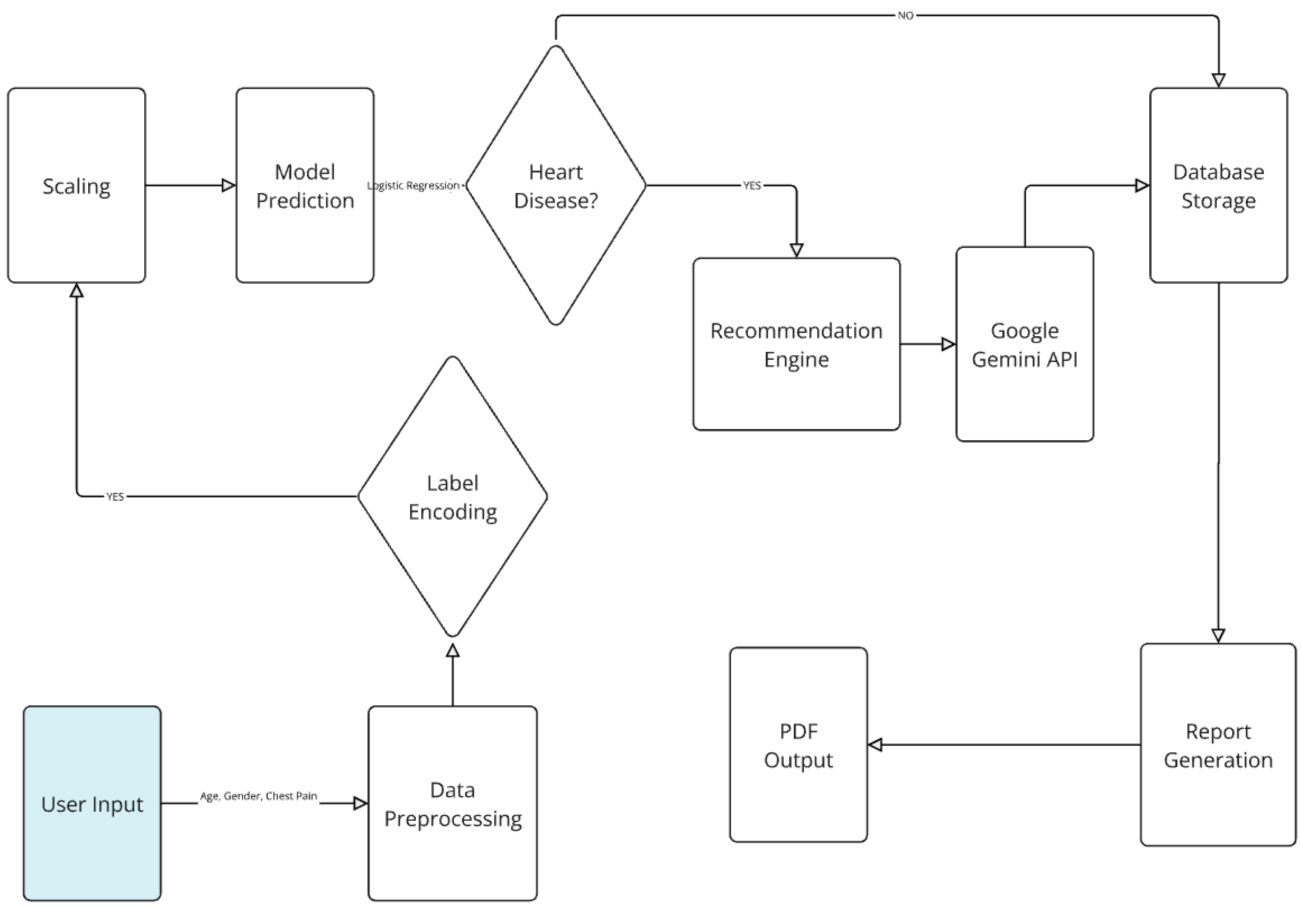

Figure 4. Flowchart of Heart Disease Prediction System Project

Data Input Module:

These are essential patient details, including age, gender, cholesterol level, and other clinical measurements. All these can be fed through a user-friendly web form. A variety of clinical data is allowed by the web form to make sure that users can input all necessary details with efficiency and accuracy. Validation of the inputted data ensures that any wrong or missing piece of information is pointed out for correction.

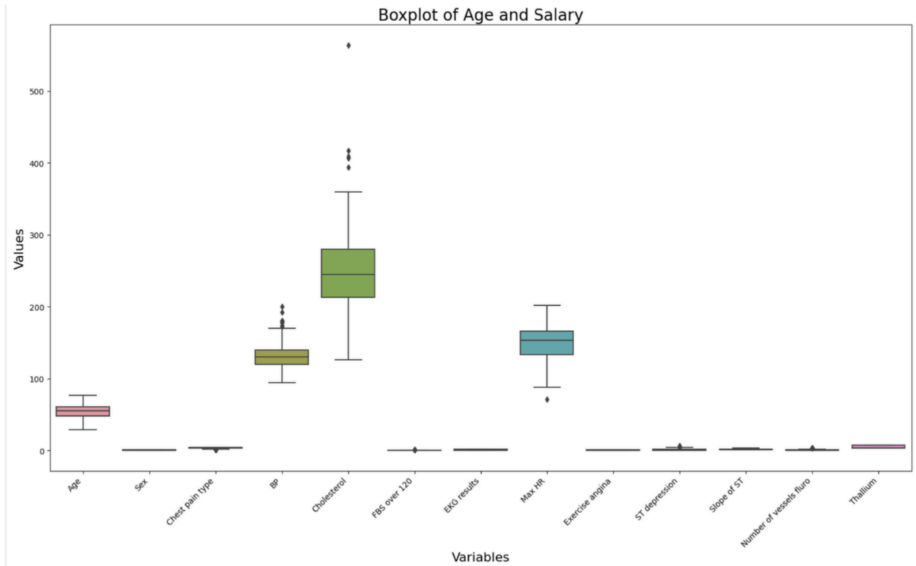

Figure 5. Input Data Values

Data Processing Module:

The prediction engine is a Logistic Regression model that, given a patient, calculates the probability of having heart disease. Besides the classification prediction, it also returns confidence scores, enabling the user to judge the credibility of the prediction. The module is designed to be very flexible and adaptive for integration in the future with advanced machine learning models.

Prediction Engine:

This module will generate, upon receiving the results of prediction, personalized recommendations using external APIs like the Google Gemini API. The suggested lifestyle modification, possible follow-up tests, and consultation with specialists, drawing from the individual's risk profile, offer actionable insight easily adoptable both by the healthcare professional and the patient.

Parameter	Description
Algorithm	Logistic Regression
Target Variable	Heart Disease (0 = Absence, 1 = Presence)
Independent Variables	All columns except Heart Disease
Data Split	80% Training, 20% Testing
Maximum Iterations (max_iter)	1000
Evaluation Metrics	Accuracy, ROC-AUC Score, Classification Report
Prediction Probability	Calculated for the positive class (Heart Disease = 1).
Risk Levels	High, Medium, Low (determined by threshold values).

Figure 6. Classifier Parameters

Recommendation Engine:

This module will generate recommendations based on the predictions, utilizing external APIs, such as the Google Gemini API. Recommendations could include lifestyle modification, possible follow-up tests, and consultation with specialists. The suggestions made are personalized; therefore, this gives an insight into one's risk profile that can be directly acted upon both by the patient and health professionals.

Database Management System:

The integrated SQLite database will store patient input, results of the predictions, and confidence scores. To enhance performance, it is optimized for speed, while ensuring that sensitive patient information is securely stored. The historical data from the database also aids in analytics such as trend analyses and longitudinal studies, further enhancing the utility of this tool in continuous research.

Report Generation Module:

This module will generate a comprehensive PDF report for each assessment by a patient. The reports will summarize input data visualizations, the outcome of predictions, and courses of action recommended. This is a professionally set-up document which will be important to the patient and healthcare provider; it aids in communicating and making informed decisions.

The modular architecture of the system is designed to scale and extend. Because each component can act somewhat independently, iterative updates and the addition of new features are possible without disruption of existing workflows. Similarly, APIs and microservices make integrations with external systems-EHRs or other healthcare platforms-computationally feasible, ensuring interoperability in the heterogeneous healthcare environment.

4. Application Features

4.1 User Interface

The web-based interface is focused on simplicity and ease of use, letting users input patient data, observe real-time predictions, and show visually presented data trends. Key features include:

- **Interactive Forms**: Dynamic input fields for patient information.
- **Data Visualization**: Graphs and charts, such as histograms and correlation matrices, illustrate feature distributions and their impact on predictions.
- **Predictive Feedback**: Clear indicators of risk levels with accompanying confidence scores.

Also, this interface will have principles of adaptive design, meaning it is accessible from a smartphone or tablet. The mechanism for user feedback-satisfaction surveys, for example-will be provided to fetch insights from customers for improvements continuously.

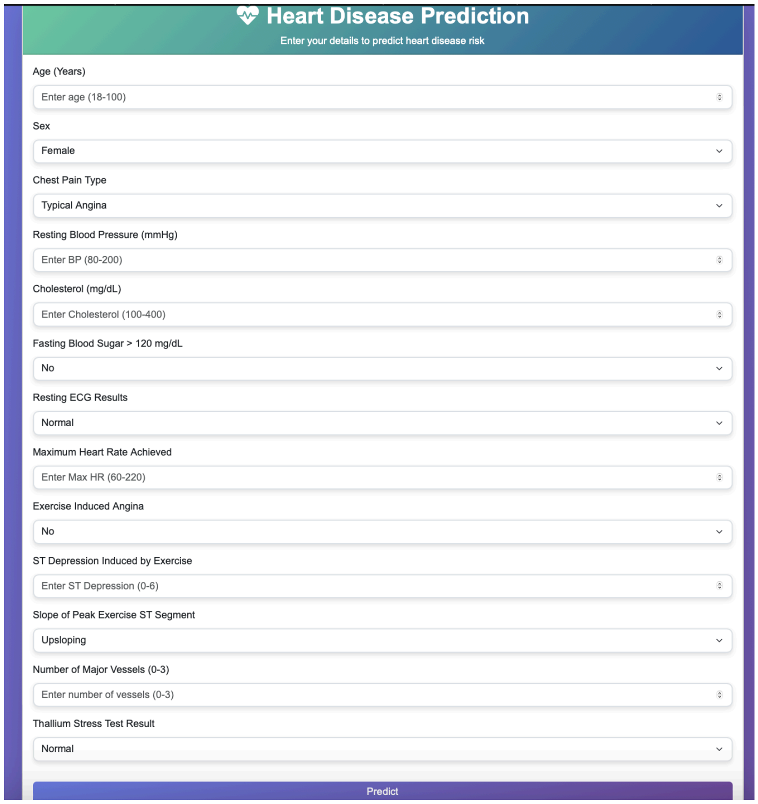

Figure 7. Input Screen of Our Project

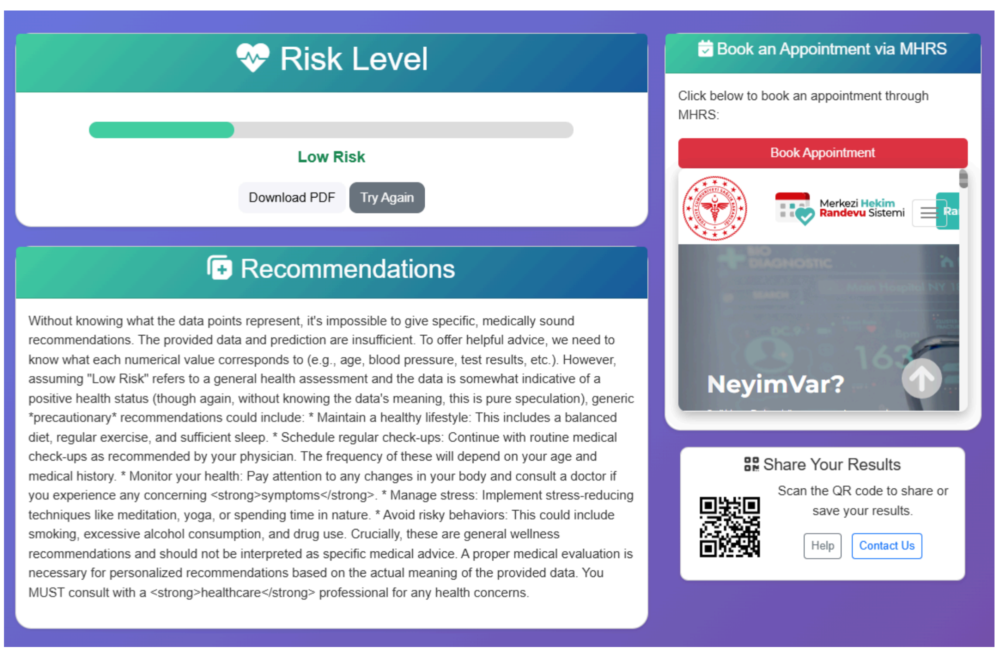

Figure 8. Result Screen of Our Project

4.2 Reporting and Analytics

The patient assessment reports will be available as PDF downloads from within the web application. Reports will include:

- **Prediction Summary**: Risk level and confidence score.
- **Key Insights**: Highlighted risk factors based on the patient's profile.
- **Recommendations**: Actionable steps for further diagnosis or lifestyle changes.

The reports will have a professional format with graphing and charting where appropriate to support comprehension. It will be possible to pass such documents to other medical professionals for further decision-making.

4.3 Advanced Features

- **Customizable Thresholds**: Users can adjust sensitivity levels to prioritize false negatives or false positives.
- **Historical Analysis**: Access to previous assessments and trends over time.
- **API Integration**: Compatibility with third-party health monitoring systems.

Lastly, the application design at this point is all about user engagement through real-time feedback loops. Ensuring immediately relevant information to the user instills trust in further use of the application. Advanced security features such as encryption of the data and users' secure login maintain user data privacy and keep legal compliances relevant.

5. Insights

- **Age:** The persons suffering from heart diseases are older with average in the year 56-57 years, but for persons without heart diseases, it is 52-53 years. The age group 55-60 years shows the highest percentage. Hence, age is considered a serious risk factor.

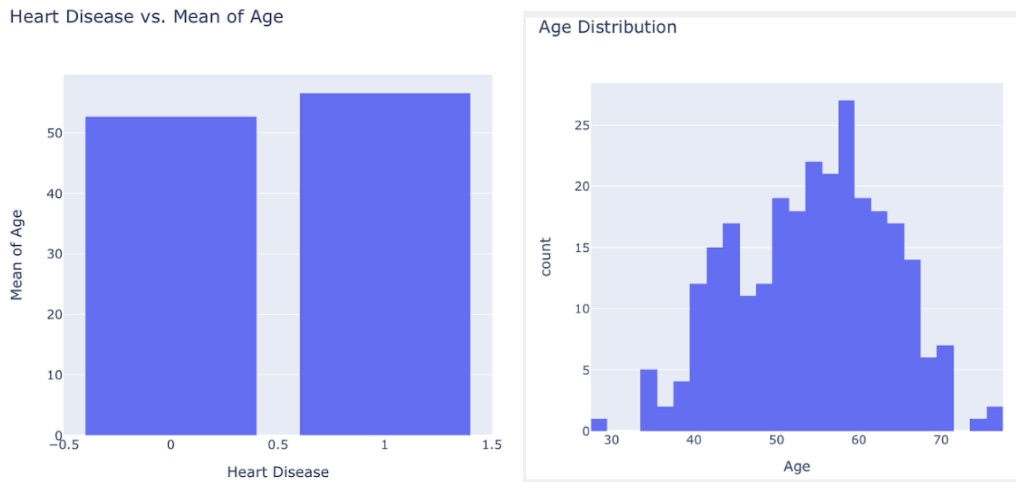

Figure 9-10. Impact of Age Parameter

- **Gender:** Males have ~50% prevalence while females ~20%. Such differences can conclude that gender besides biological factors and habit factors is also significant in the prediction model.

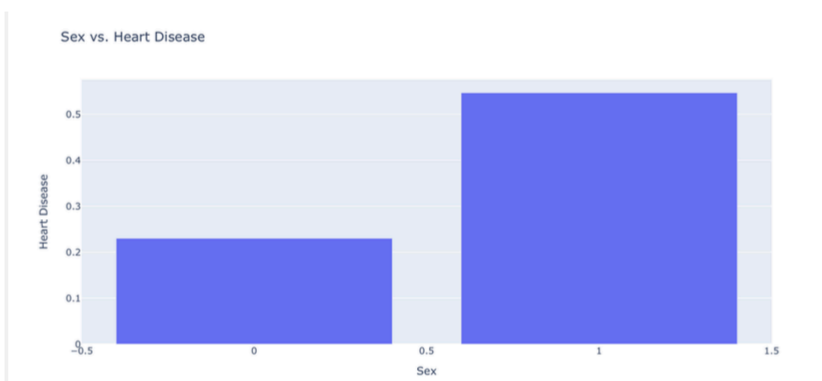

Figure 11. Impact of Gender Parameter

- **Chest Pain Type:** Among all, Type 4 chest pain has the highest risk and should be admitted as soon as possible, whereas the patients with Type 2 chest pain have the least chance of developing heart disease. It indicates giving priority for further tests to patients with serious chest pain.

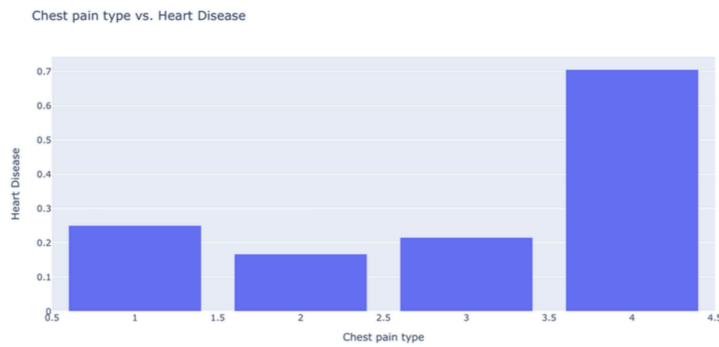

Figure 12. Impact of Chest Pain Parameter

- **Blood Pressure and Cholesterol:** Both the persons suffering from heart diseases and persons without this disease show an average of almost the same blood pressure (~130) and cholesterol (~240). So these factors alone cannot predict heart disease.

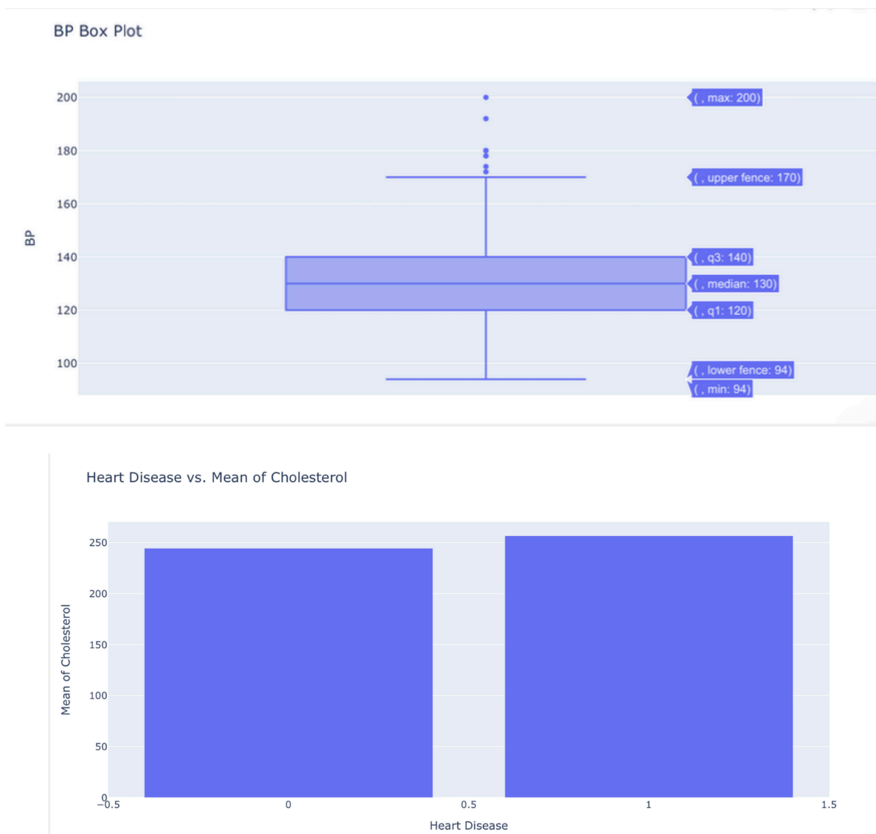

Figure 13-14. Impact of Blood Pressure and Cholesterol Parameters

- **Fasting Blood Sugar (FBS):** A FBS ≤ 120 presents a ~45% prevalence, while those having a FBS > 120 have a ~40% prevalence. Thus, having a low FBS does not rule out the disease, and this factor has to be supported by other features of the clinical presentation.

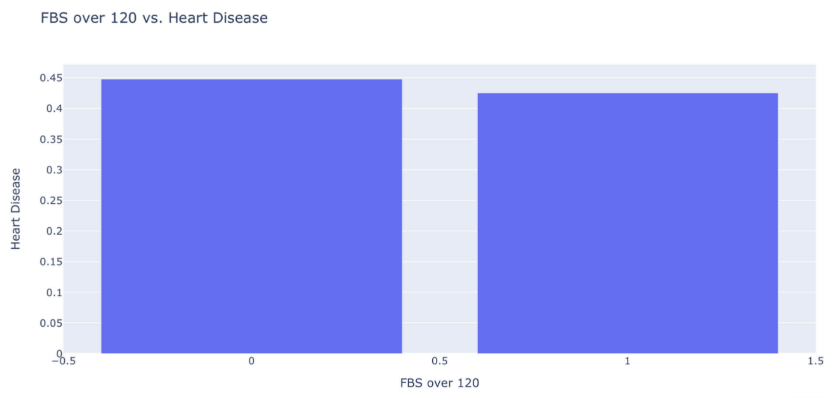

Figure 15. Impact of FBS Parameter

- **Electrocardiogram (EKG):** Readings of 1 or 2 are given a prevalence of ~50%, where normal results are ~30%. Such an abnormality may point to a set of risk factors that needs follow-up attention.

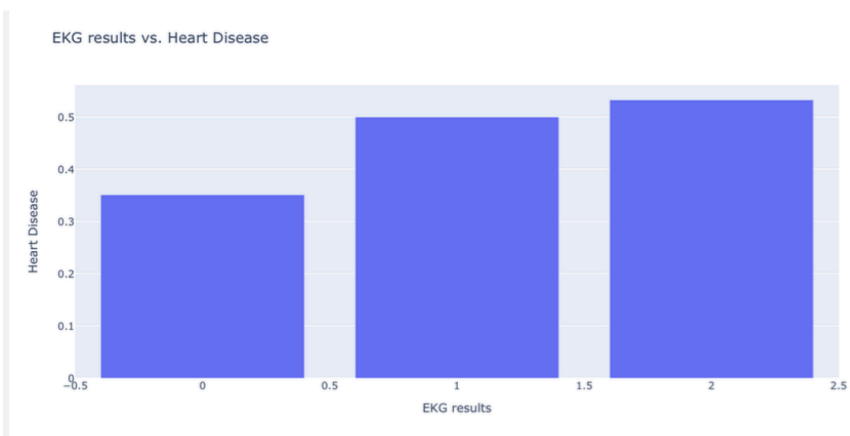

Figure 16. Impact of EKG Parameter

- **Maximum Heart Rate (Max HR):** The average Max HR in those with heart disease is ~140 bpm, while for those without it, the average Max HR is ~160 bpm. This decrease in maximum heart rate may indicate decreased cardiac reserve and points to the importance of stress testing.

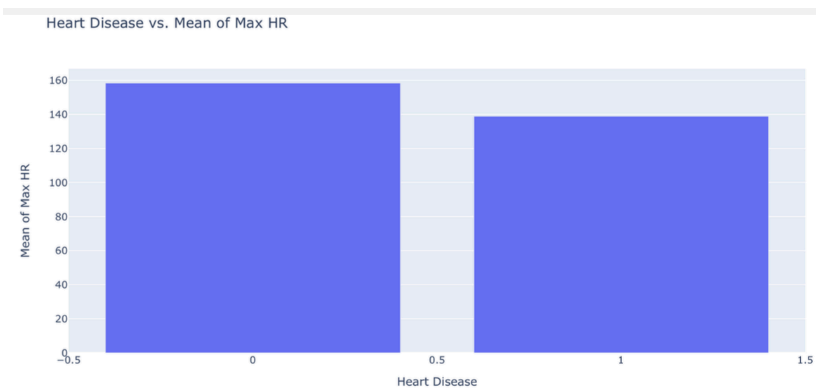

Figure 17. Impact of HR Parameter

- **Exercise-Induced Angina:** The average Max HR in those with heart disease is ~140 bpm, while for those without it, the average Max HR is ~160 bpm. This decrease in maximum heart rate may indicate decreased cardiac reserve and points to the importance of stress testing.

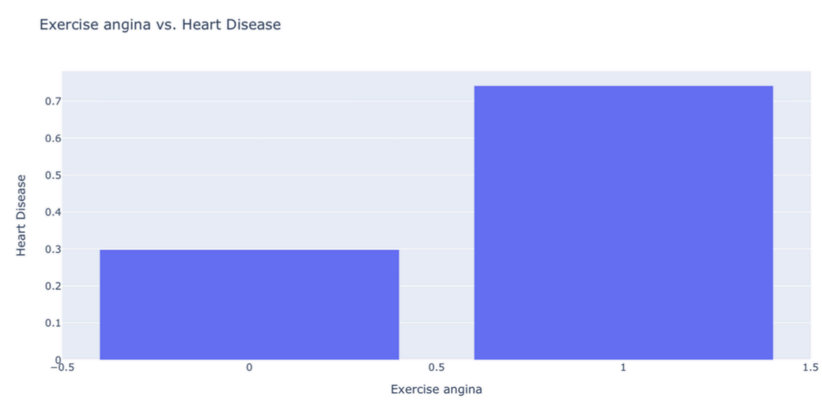

Figure 18. Impact of Exercise-Induced Angina Parameter

- **ST Depression:** There is a 60%-100% increase in prevalence with increasing ST depression values of ≥2.

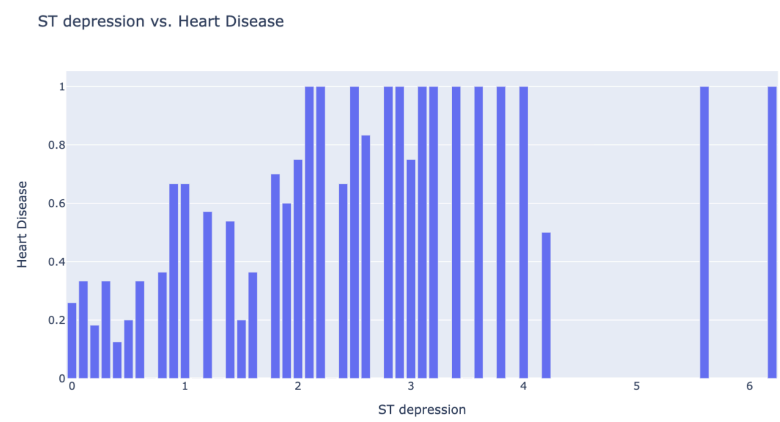

Figure 19. Impact of ST Depression Parameter

- **ST Segment Slope:** Flat slopes are more risky, while downward slopes are at moderate risk. While upward slopes indicate less risk.

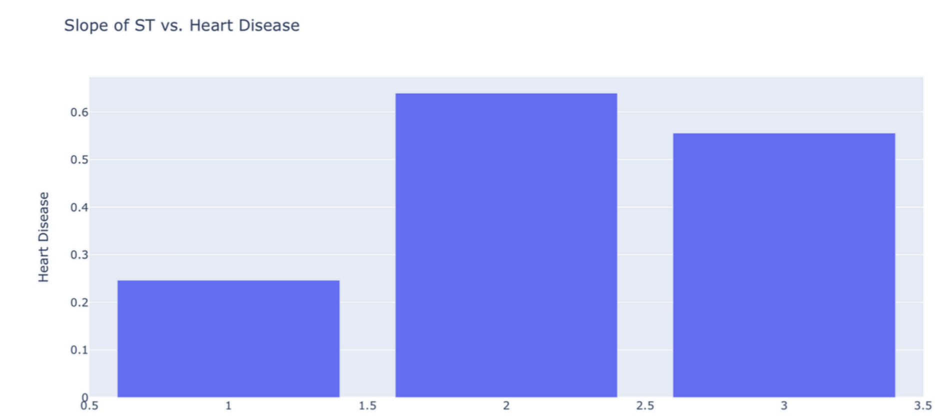

Figure 20. Impact of ST Parameter

- **Fluoroscopy:** The prevalence increases with the number of vessels detected. For example, subjects with 3 vessels have an ~80% prevalence, in comparison with ~25% for those with 0 vessels. That shows the importance of imagery techniques in diagnosis.

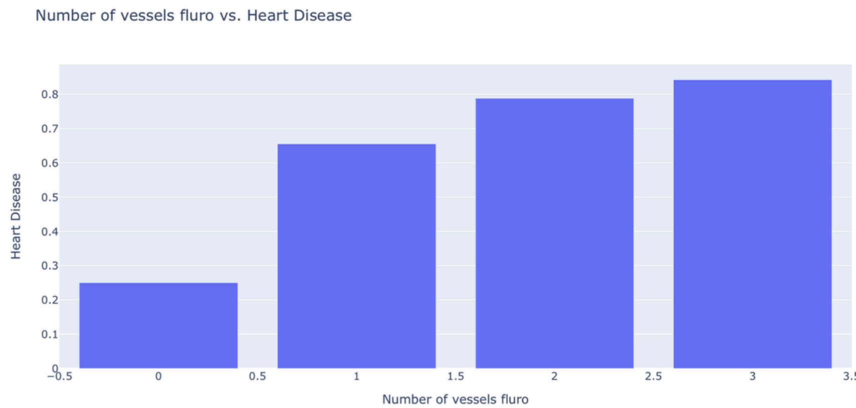

Figure 21. Impact of Fluoroscopy Parameter

- **Thallium Test Results:** Results in classes 3, 6, and 7 correspond to ~20%, ~55%, and ~70% prevalence, respectively, underlining thallium testing as a complementary test in diagnosis.

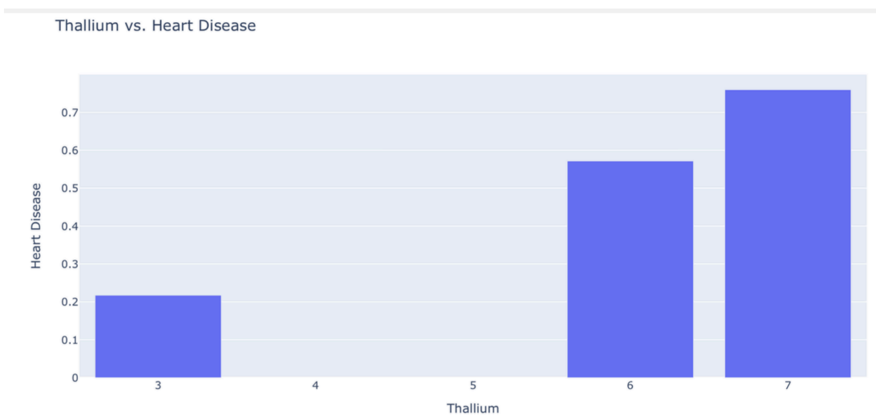

Figure 22. Impact of Thallium Parameter

Result Observations

- The most useful indicators of heart disease risk are age and gender. Older and male subjects show a higher prevalence.

- Clinical indicators of various types of chest pain, EKG, and induced angina during exercise proved more powerful predictors than measures of cholesterol and blood pressure.

- The developed model of Logistic Regression gave an accuracy of 91% and a ROC-AUC of 94.8%, proving its dependability in identifying patients who were falling into the high-risk group.
- In the future, this will be very easy and efficient once integrated with wearables and mobile apps, along with multi language support, to provide better usability and performance.

5.1 Model Updates Based on Feedback

Considering the feedback given in class, regarding the fact that average blood pressure (~130) values were practically equal for people suffering from heart disease and for those who did not, the training process was modified to take that effect into consideration. The main change done was giving the weight of the blood pressure feature a smaller value in order to decrease its influence on the model. It was done by normalizing and scaling the feature as can be seen in the updated code of the training process below.

Changes in the Training Process:

1. **Feature Scaling for Blood Pressure:**
    - The blood values were normalized by dividing by the maximum value. Such a transformation decreases the magnitude of the feature and, hence, its influence while training the model.

      ```
      data["BP"] = data["BP"] / data["BP"].max()
      ```
2. **Feature Importance Analysis:**
    - Here, the RandomForestClassifier is used for training of the algorithm. Feature importances can thus be extracted after training. By now, it has been unveiled that the relative importance of blood pressure is lower when compared to other features; that was expected by looking at the feedback.

3. **Performance Evaluation:**
    - The model was trained on the 80-20 train-test split, and despite the lower weight of a blood pressure feature, it still had high accuracy and performed well on testing with the accuracy of {accuracy_score_value}.

4. **Model Deployment:**
    - The new model was saved using the “joblib” library to ensure reproducibility and integration into the Flask-based web application.

6. Benefits

Implementing this heart disease prediction system offers numerous advantages:

1. **Improved Diagnostic Accuracy**: All the different time points and accuracy ratings, therefore, benefit from machine learning algorithms. This reduces human error, hence promoting vice consistent and reliable quality that is sent to the patient. It improves the patients' quality of care. Moreover, the AI-based tool is capable of self-learning. As a result, the increase in the diagnostic accuracy becomes a continuous process which is able to learn from new data.

2. **Accessibility and Scalability**: These accessible tools give an opportunity for web service products to help healthcare workers and their patients all over the world. The scalability options are the ones related to let the product work perfectly in the healthcare systems that are already existing. It acts as a health for this planet by setting up remote care and providing services to people who live in the remote and underserved areas, thereby revealing the gap in healthcare access.

3. **Cost-Effective Preventive Care**: Early and exact identification of at-risk individuals by the system decreases the necessity for diagnostic procedures and hospitalization. This way, the costs will be saved. The prevention that could be made by the system may be very effective in reducing the economic burden of health systems.

4. **Empowered Decision-Making**: An invaluable asset to both patients as well as healthcare personnel the system delivers perspectives, which are operational and allow them to adopt the best means alongside it first of the correct. The system allows for negotiation between patients and doctors by providing the latter with predictions and recommendations that are based on discursive arguments.

5. **Data-Driven Insights**: AAccording to technology which transforms healthcare, the data mining of the system gets unidentified and unseen patterns, trends, and correlations that are the basis for medical research and personalized medicine. The application of visual analytics tools leads to a deep level of understanding of both the individual patient health status and the trends in population health.

The system is also in line with the priorities of world health, as it focuses on preventing rather than treating. Such a paradigm shift can reduce the cost burden on healthcare systems and enhance population health. Its role goes beyond an individual's care; it helps public health by collating data from each and its analysis.

7. Future Developments

The proposed project lays the groundwork for many improvements. Some possible future developments include:

1. **Integration with Wearable Devices**: This model can be integrate wearable fitness tracker information or smartwatches to monitor health and risk continuously in real-time. As a matter of fact, this can also be used to provide dynamic feedback to users, enhancing user engagement.

2. **Advanced Machine Learning Models**: Deep learning and ensemble models can be applied for advanced machine learning to surface better predictive performance and deal with complex datasets. Techniques in explainable AI can also be integrated to enhance trust in model predictions.

3. **Global Dataset Expansion**: Growing diversity within the dataset across demographic dimensions will reduce model bias and improve generalizability. Collaboration with international health organizations facilitates data sharing and collaboration.

4. **Mobile Application**: Developing a mobile application for better accessibility and utilization, especially by remote and underserved populations, may be used to remind users to conduct periodic health check-ins and provide them with personalized wellness tips.

5. **Multi-Language Support**: Increasing the number of languages on offer will result in increased global reach since non-English speakers will be able to access this tool. Localization will make the tool culturally appropriate, and thus the rate of adoption will be very high.

6. **Collaborative Features**: It includes collaboration features such as secure data sharing by the team of healthcare providers, allowing coordination of care using a multidisciplinary approach. If they are integrated into electronic health records, the process will be smooth and provide continuity of care.

7. **Ethical Considerations**: Data privacy and compliance issues, such as GDPR, can be solved with the help of gaining user trust to improve data security and transparency policies that keep the user aware of how and when his/her data is being utilized.

8. Conclusion

The Heart Disease Prediction System represents a quantum leap in machine learning applied to health in ways that will enable the doctors and patients to act positively on major risk factors analyzed. In both its front-end and back-end aspects, development has considered practical use within a wide range of clinical settings.

Although this current implementation is promising, further development and scaling up will be based on emergent requirements in healthcare. This project characterizes how AI can contribute to solving a portion of the key global health problems and will, therefore, help pave the way toward the application of further intelligent and evidence-based interventions in medicine.

That this would even work speaks to the greater embedding of AI in healthcare. With technological advancement, of course, applications will rise offering innovative responses to some of the really serious medical problems of the world. The capacity to adapt to new healthcare demands means it has stayed relevant and effective in the long term.

References

- American Heart Association: Understanding Heart Disease. Accessed: https://www.heart.org/en/health-topics/heart-attack/about-heart-attacks
- GitHub Repository: Heart Disease Prediction Project. Accessed: https://github.com/itskritibhardwaj/HEART-DISEASE-PREDICTION-PROJECT
- Heart Disease Prediction Dataset, Kaggle. Accessed: https://www.kaggle.com/datasets
- Mayo Clinic: Heart Disease Overview. Accessed: https://www.mayoclinic.org/diseases-conditions/heart-disease
- National Heart, Lung, and Blood Institute: Heart Disease. Accessed: https://www.nhlbi.nih.gov/health-topics/heart-disease
- World Health Organization: Cardiovascular Diseases. Accessed: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases