

University of Tennessee, Knoxville

COSC 445

12/09/24

Heart Disease Detection Final Report

Mikayla McCormack, Devanshi Patel, Chase Woodfill, Tully Fitzpatrick, Jayden Leuciuc

I. Objective

Our project is dedicated to advancing preventive healthcare by focusing on raising awareness and promoting proactive measures to combat heart disease. By empowering individuals with knowledge about risk factors and symptoms, we aim to bridge the gap between awareness and action. Our digital tools are designed to provide accessible, user-friendly resources that enable users to monitor their health and recognize early warning signs. Through targeted education and engagement, we strive to make heart health a priority in everyday life, fostering a culture of prevention and treatment.

Looking ahead, our vision is to reduce the global burden of heart disease by leveraging the power of technology to encourage early detection and sustained health consciousness. We aim to make preventive care not just more accessible but also more effective. Through these efforts, we aspire to create a lasting impact on public health, reducing both the incidence and severity of heart disease and contributing to a healthier, more informed society.

II. Dataset Breakdown

We used an open-source dataset from UCI Machine Learning Repository, also available on Kaggle. This dataset had 303 patient records, consisting of 909 entries in total and 14 different features to categorize if the person is likely to have a heart disease. Here's the table to enlist all 14 features:

Order	Feature	Description	Feature Range Value
1	Age	Age is Years	29 to 77
2	Sex	Gender	Value 1 = Male Value 0 = Female
3	Cp	Chest Pain Type	Value 0: typical angina Value 1: atypical angina Value 2: non-anginal pain Value 3: asymptotic
4	Trestbps	Resting blood pressure (in mm Hg on admision to hospital)	94 to 200
5	Chol	Serum Cholestrol in mg/dL	125 to 564
6	Fbs	Fasting blood sugar > 120mg/dL	Value 1 = true Value 0 = false
7	Restecg	Resting electrocardiographic results	Value 0: Normal Value 1: having ST-T wave Abnormality (T wave inversions and or ST elevation or depression of > 0.05 mV) Value 2: showing probable or define left ventricular hypertrophy by Estes Criteria
8	Thalac	Maximum Heart Rate achieved	71 to 202
9	Exang	Exercised- induced angina	Value 1 = yes Value 0 = no
10	OldPeak	Stress test depression	0 to 6.2

		induced by exercise relative to rest	
11	Slope	The slope of peak exercise ST segment	Value 0: upsloping Value 1: flat Value 2: downsloping
12	Ca	Number of major vessels	Number of major vessels(0-3) Colored by fluroscopy
13	Thal	Thallium Heart Rate	Value 0: Normal Value 1: fixed defect Value 2: reversible defect
14	Target	Diagnosis of heart disease	Value 0 : no disease Value 1: disease

A. Data Storage

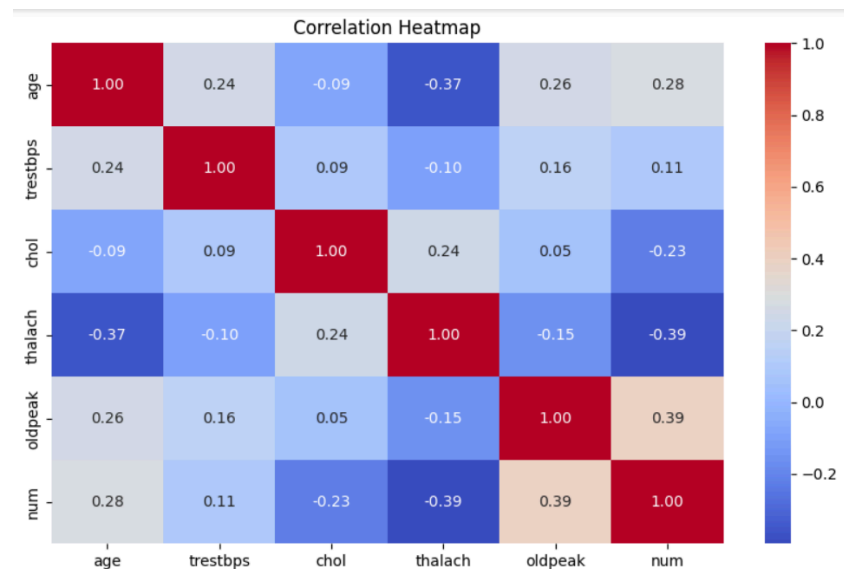
Given the dataset's relatively small size, it will be easy to manage using common data-handling tools. We plan to store the dataset in CSV format, as it is lightweight, widely supported, and compatible with most data preprocessing libraries, particularly in Python. We will use the Pandas library to efficiently load and manipulate the data, ensuring it remains accessible for analysis and model training.

B. Data Processing and Exploration

The dataset consisted of a combination of numerical and categorical variables. To standardize the data, all categorical variables were transformed into numeric representations using one-hot encoding. Additionally, missing values were addressed by removing any rows containing incomplete data, ensuring that only complete records were retained for analysis. A heatmap was plotted to analyze correlations between features, helping to identify relationships and potential redundancies in the dataset. The target variable, which originally ranged from 0 (no disease) to 4 (varying levels of disease severity), was simplified into a binary classification:

- **0**: No Heart Disease
- **1, 2, 3, 4**: Presence of Heart Disease

We plotted the Target Variable Distribution graph and the target labels as can be seen from the plot are fairly distributed. The dataset is balanced, this ensures that the model doesn't favor one class over another.



(Optional) HeatMap Description: The heatmap revealed several notable correlations between features and the target variable (**num**):

- **Oldpeak ($r = 0.39$)**: This feature exhibited the strongest positive correlation with the target, indicating that higher values of ST depression are associated with an increased likelihood of heart disease.
- **Thalach ($r = -0.39$)**: The strongest negative correlation was observed with maximum heart rate, suggesting that higher heart rate levels are linked to a lower likelihood of heart disease.
- **Age ($r = 0.28$)**: A moderate positive correlation with heart disease was identified, implying that the probability of heart disease tends to increase with age.

Our initial approach was to drop the columns that had the most missing values (i.e Number of major vessels (ca), slope, and Thallium Heart Rate (thal)). However, it didn't work well with accuracy. Hence, we changed our approach to keep all 14 features for training to achieve better results.

C. Data Integration

The data was standardized to ensure compatibility with the Machine Learning Model.

D. Feature Scaling

Since our features have different ranges (e.g., age in years vs. chol in mg/dL), **StandardScaler** from sklearn was used for normalization. Training Data was scaled using fit_transform to compute mean and variance. While, the testing was scaled using transform to ensure uniform scaling.

Train-Test Split: We performed an 80-20 split for our training dataset. Random seed was set at 42 for reproducibility.

III. Model Description

For training, we used two different machine learning models in addition to Support Vector Machine(SVM). These models were chosen to explore different approaches to classification and ensure the best performance for heart disease prediction.

The models used:

- **Logistic Regression:** A linear model suitable for binary classification, leveraging the relationship between input features and the log odds of the target variable.
- **Random Forest Classifier:** An ensemble learning method that builds multiple decision trees and aggregates their predictions for improved accuracy and reduced overfitting.
- **Support Vector Machine (SVM):** A kernel-based model that finds the optimal hyperplane for separating the classes.

For each of this model, pipeline was set up to perform one-hot encoding and it was later fitted using the training dataset. Each model was tested, and its performance was assessed using accuracy as an evaluation metric.

A. Results

Model	Accuracy Score
Logistic Regression	79%
SVM	85%
Random Forest Classifier	81%

From the results above, SVM outperformed Logistic Regression and Random Forest Classifier. This can be because of the fact that SVM has a better ability to handle high-dimensional data. Additionally, the model was saved in a serialized format (.pkl) for future use, enabling easy deployment and testing with new input data.

IV. Approach

To develop the heart disease prediction website, we used an open-source dataset from the UCI Machine Learning Repository, which contained 14 features and 303 patient records. The dataset was first cleaned by addressing missing values and reducing the target variable to binary labels for ease of interpretation. Using StandardScaler, we standardized numerical features to ensure they were on a similar scale, which is crucial for improving model performance. Correlation analysis through a heatmap helped us identify relationships between variables, informing feature selection and engineering decisions. Categorical variables like chest pain type were one-hot encoded to allow machine learning models to process them effectively. For model training, we opted to test multiple algorithms, including Random Forest Classifier, Logistic Regression, and Support Vector Machines (SVM), to determine which provided the best balance of accuracy and generalization. We split the dataset into an 80-20 train-validation set to validate model performance rigorously.

The integration of the machine learning model with the website involved hosting the trained model on a Flask server, which acted as an API for the React-based frontend. Bootstrap and CSS were used for the frontend design, ensuring a clean and responsive interface. Flask endpoints were designed to handle requests efficiently, enabling the user to input medical data

and receive predictions in real time. This connection required careful handling of data formats and debugging issues like CORS restrictions to ensure smooth functionality.

In addition to the prediction feature, we prioritized making the front-end of the website informative and user-friendly. We included resources, statistics, and clear details about the purpose of the site to enhance user engagement and understanding. By presenting users with information on heart disease, its risk factors, and the importance of early detection, we aimed to create a tool that not only predicts risk but also educates and empowers users. The inclusion of this information was a deliberate decision to ensure the website serves as a comprehensive health resource, reinforcing the site's purpose as a supportive tool for heart disease awareness and prevention.

Heart Disease Detection Tool

Welcome to our Heart Health Tool – Empowering you to take charge of your health.

As dedicated students committed to making a difference, we developed this tool to help empower individuals with knowledge about their heart health. Our mission is to provide easy access to insights that can encourage proactive choices for a healthier future.


Our Goal

As part of our studies, we developed this tool to help you quickly assess your heart disease risk. Our goal is to empower you with insights for informed health choices. If you may be at risk, we encourage consulting a medical professional. Early awareness is key to prevention, and we're here to support your journey to a healthier future.

Heart Statistics

Heart disease is a serious health concern, affecting more than 1 in 10 adults in the U.S. It's the leading cause of death, accounting for 1 in 4 deaths, with over 700,000 lives lost in 2022 alone. Advanced imaging shows that nearly 50% of people tested have early signs of coronary artery disease, yet 3 in 5 of these cases go undiagnosed.

Happy + Healthy



Please fill out this field. Using information from a medical assessment provided by your doctor, simply answer the questions, and we'll assess your risk of heart disease based on our pre-trained model. Together, we can take proactive steps towards preventing heart disease and fostering healthier lives for all.

Age:

Sex:

Chest Pain Type:

Resting Blood Pressure:

Serum Cholesterol mg/dL:

Fasting Blood Sugar > 120 mg/dL:

Resting Electrocardiographic Results:

Maximum Heart Rate Achieved:

Exercise-Induced Angina:

Oldpeak (ST Depression) Induced by Exercise Relative to R:

Slope of Peak Exercise ST Segment:

Number of Major Vessels (0-3) Colored by Fluoroscopy:

Thalassemia:

Results for Your Heart Health Assessment

Please find your heart disease risk assessment results below.

Based on our model, you are predicted to be:

High Risk

What do your results mean?

You are predicted to be at **High Risk**. Please note that this assessment is not a medical diagnosis. We strongly encourage you to consult with a healthcare professional to discuss your results and take appropriate action to address your heart health.

Happy + Healthy



Ultimately, our approach combined robust preprocessing, thorough exploration of algorithms, and thoughtful design of the web application, resulting in a user-friendly tool for predicting heart disease while highlighting the strengths of data-driven healthcare solutions.

V. Results

We successfully developed a web-based platform designed to empower users with insights into their heart health. By allowing users to input relevant health data, such as age, blood pressure, and cholesterol levels, the platform calculates their predicted risk of heart disease. This user-friendly tool leverages advanced machine learning models to analyze the input and provide actionable insights, helping individuals understand their risk profile and make informed decisions to improve their heart health. The platform's seamless interface ensures accessibility, making it an effective resource for preventive healthcare.

At the core of our project is a Support Vector Machine model, which achieved an impressive accuracy of 85% in predicting heart disease risk. This high-performance model was trained and validated on a robust dataset, ensuring reliability and precision in its predictions. By combining the predictive power of the Support Vector Model with an intuitive user experience, our platform sets a strong foundation for scalable digital healthcare solutions.

VI. Primary Issues Encountered

Developing a website to predict heart disease using a machine learning model involved several significant challenges. The first hurdle was identifying the appropriate tech stack for the project. This required balancing the strengths of different tools to ensure compatibility and efficiency. Flask was chosen as the API platform for the ML model due to its simplicity and flexibility, while Bootstrap and React were selected to handle the frontend for their responsiveness and dynamic capabilities. However, ensuring seamless integration between these

technologies was complex, requiring extensive testing to establish smooth communication between the server and client sides. Additionally, selecting the right feature set for training the ML model proved critical. The team had to identify and preprocess relevant medical data to maximize prediction accuracy while avoiding overfitting or underfitting, which involved experimenting with various combinations of features and hyperparameters.

Once the foundational setup was in place, the next challenge was connecting the machine learning model hosted on the Flask server to the React-based web frontend. This integration demanded precise API endpoint design and careful handling of data exchanges between the server and frontend to ensure that predictions were delivered in real time without significant latency. Additionally, ensuring the usability and visual appeal of the website through Bootstrap required additional iterations to make the application intuitive for non-technical users like medical professionals. Despite these challenges, the project ultimately provided valuable insights into the complexities of integrating machine learning models into user-facing applications, highlighting the importance of collaborative problem-solving and iterative development.

VII. Future Work

This heart disease tool has the potential to be expanded into a comprehensive platform capable of identifying a wide range of health risk factors. By providing similar tools, the platform could provide insights into various medical conditions, enhancing its utility and application across healthcare domains, making the data more useful.

Given that the current data utilized in the tool dates back to 1988, obtaining updated and more comprehensive datasets would significantly improve its relevance and accuracy. Modern data would better reflect current health trends, demographics, and risk profiles, leading to more reliable analyses and predictions, especially in the obesity epidemic many countries are facing.

Compared to 1988, advancements in the medical field would give the tool more access to specific information, bettering accuracy.

Additionally, incorporating more specific and detailed input data would up the tool's accuracy. By integrating nuanced variables and specific health metrics, the platform could deliver more precise risk assessments and recommendations for users and healthcare providers. The downside to this is the platform would become more complex.

VIII. Timeline and Responsibilities

The team was able to reach all milestones with success. The following timeline was followed to ensure progress:

Week 1 ending 10/11:

- Start on design and website. But, ideally finish the design.
- Research implementation of data handling.
- Have a plan for implementation between the dataset and website.

Week 2 ending 10/18:

- Complete design.
- Have basic website UI done.
- Continue to work on data handling/ website integration.

Week 3 ending 10/25:

- Have dataset and website communicate

Week 4 ending 11/1

- Finish functionality between dataset and website.

- Have dataset handling done.

Week 5 ending 11/8:

- Have the website return expected information as a result of the dataset.

Week 6/7 ending 11/15:

- Polish website, fix any design/ functionality issues.

Week 8 ending 11/22

- Test all website functionality, make sure it's working properly.
- Prepare for presentation to class.

Devanshi	Working on data handling / predictions
Jayden	Help create website and offer recommendations for tech stack
Mikayla	Design the website and create integration between frontend and flask server
Tully	Design and create flask middleware to connect frontend to model
Chase	Help with design and documentation