Heart Disease Detection

Mikayla McCormack[1], Brennon Fitzpatrick[2], Jayden Leuciuc[3], Devanshi Patel[4], and Chase Woodfill[5]
The University of Tennessee - Knoxville, CS445 Fall 2024

***Abstract*— This project creates a website to predict heart disease risk using user-provided health data. Leveraging a UCI dataset and predictive algorithms, the site offers an accessible tool for early detection. The project will be completed in eight weeks with defined team roles in design, data handling, and implementation.**

I. INTRODUCTION

A. Objectives

Our project aims to develop a website that predicts heart disease risk based on user-provided health data. Users will input specific values such as age, sex, chest pain type, resting blood pressure, cholesterol levels, fasting blood sugar, resting electrocardiographic results, maximum heart rate, and other key indicators. These inputs will be processed to predict the likelihood of heart disease, represented by a "target" value indicating the presence (1) or absence (0) of heart disease. The website will leverage anonymized patient data to ensure privacy while offering an accessible, user-friendly tool for early detection and prevention efforts.

B. Motivations

Heart disease remains one of the leading causes of death worldwide, with early detection playing a crucial role in prevention and effective treatment. Over 80 percent of heart diseases are preventable, including cardiovascular disease and strokes ([1], [2]). However, access to healthcare services for early diagnosis can be limited due to geographical, economic, and social barriers. Our project is motivated by the desire to provide individuals with an accessible, easy-to-use tool for assessing their potential risk of heart disease. By leveraging predictive algorithms and anonymized health data, we aim to empower users with insights into their cardiovascular health, potentially prompting timely medical consultation and lifestyle changes that can reduce the risk of severe outcomes.

II. DATA DISCUSSION

A. Data Collection

We plan on using an open-source dataset available from the UCI Machine Learning Repository [3]. This dataset contains 303 samples with 14 features per patient, with the main objective being to predict whether or not a patient has heart disease. These features include medical and demographic information, such as age, sex, cholesterol levels, and other health-related parameters, making it suitable for building a heart disease detection model. Table 1 contains the attribute content we'll be collecting from the user.

TABLE I
FEATURES AND THEIR DESCRIPTIONS FOR THE HEART DISEASE DATASET.

Order	Feature	Description	Feature Range Value
1	Age	Age in years	29 to 77
2	Sex	Gender	1 = Male 0 = Female
3	Cp	Chest Pain Type	0 = Typical angina 1 = Atypical angina 2 = Non-anginal pain 3 = Asymptomatic
4	Trestbps	Resting blood pressure	94 to 200 mm Hg
5	Chol	Serum Cholesterol (mg/dL)	125 to 564
6	Fbs	Fasting blood sugar (>120mg/dL)	1 = True 0 = False
7	Restecg	Resting electrocardiographic results	0 = Normal 1 = ST-T wave abnormality 2 = Left ventricular hypertrophy by Estes Criteria
8	Thalac	Maximum Heart Rate	71 to 202
9	Exang	Exercise-induced angina	1 = Yes 0 = No
10	OldPeak	Stress test depression	0 to 6.2
11	Slope	Slope of ST segment	0 = upsloping 1 = flat 2 = downsloping
12	Ca	Number of major vessels	0 to 3 (colored by fluoroscopy)
13	Thal	Thallium Heart Rate	0 = Normal 1 = Fixed defect 2 = Reversible defect
14	Target	Diagnosis of heart disease	0 = No disease 1 = Disease

B. 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.

C. Data Cleaning

A crucial first step in preparing the data is performing thorough data cleaning to ensure high-quality inputs for our model. We will:

- Handle Missing Values: We will inspect the data for any missing values, using mean or median imputation techniques to fill them in when appropriate. For rows

with significant missing data, we will consider removing them to prevent skewing the model's performance.

- Outlier Detection: To address outliers, we will visualize the dataset using tools such as boxplots. This will help us identify extreme values that might distort the model. Techniques such as Winsorization (limiting extreme values) will be applied to manage outliers effectively, preventing them from impacting model accuracy.
- Feature Encoding: Since some features are categorical, such as gender or chest pain type, we will apply feature scaling techniques. Using a One-Hot Encoder, we will convert categorical labels into numerical values, making them suitable for machine learning models.

D. Data Exploration

We will conduct extensive exploratory data analysis (EDA) to understand the relationships between features and identify any patterns that could improve the model's performance:

- Correlation Analysis: By using a correlation heatmap, we will visualize the relationships between different features, helping to detect redundant or highly correlated variables. Identifying such relationships can guide us in feature selection, ensuring we only use the most relevant inputs for the model.
- Dataset Splitting: The dataset will be split into training and testing sets to evaluate model performance. This approach helps prevent overfitting and ensures that our model generalizes well to new, unseen data.

The data exploration phase will be vital in optimizing our model for predicting heart disease with higher accuracy.

III. PROJECT PLAN

A. Timeline

Our project will follow a structure of one-week sprints, with weekly meetings to review progress, address challenges, and plan upcoming tasks. While this schedule is not set in stone, it provides a flexible framework that helps maintain organization and keeps the team aligned. The sprint approach allows us to break the project into manageable chunks, making certain that each aspect is addressed efficiently while also allowing for adjustments as needed throughout the development process. Table II provides a weekly breakdown of tasks.

TABLE II
PROJECT TIMELINE AND TASKS

Week	Tasks
Week 1 (ending 10/11)	Start on design and website. 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 the design. Have the basic website UI done. Continue to work on data handling and website integration.
Week 3 (ending 10/25)	Have the dataset and website communicate.
Week 4 (ending 11/1)	Finish functionality between the dataset and website. Ensure dataset handling is complete.
Week 5 (ending 11/8)	Ensure the website returns the expected information from the dataset.
Week 6/7 (ending 11/15)	Polish the website, fixing any design and functionality issues.
Week 8 (ending 11/22)	Test all website functionality, ensuring everything works properly. Prepare for presentation to the class.

B. Team Member Roles

In our project, the responsibilities are divided among the team to ensure efficient collaboration and progress. Devanshi and Tully will be primarily focused on managing the data and constructing the predictive models. Their role involves handling the input health values and ensuring the accuracy of the data processing, which is critical for generating reliable heart disease risk assessments. By focusing on the development of the predictive algorithm and data handling, they will ensure that the backend of the project is robust and capable of delivering precise health risk predictions.

Jayden is tasked with setting up the website infrastructure and managing the overall implementation. His responsibilities include ensuring that the platform is fully functional, secure, and user-friendly. He will focus on establishing a seamless connection between the frontend and backend systems, as well as integrating the predictive models into the website. Jayden's work is crucial for creating a stable and accessible environment where users can input their health data and receive risk assessments efficiently.

Mikayla and Chase are responsible for the project's design and documentation components. They will work on crafting an intuitive, visually appealing website layout that enhances user experience and simplifies navigation. In addition, they will document the project's development process in detail, including design decisions, technical challenges, and solutions. This documentation will not only provide clarity for the current project but also ensure that future updates or scalability efforts can be implemented smoothly by maintaining thorough records of the team's work.

IV. EXPECTED OUTCOMES

We expect to develop a functional website that allows users to input their health data and receive an immediate risk assessment for heart disease. The prediction will be based on well-established medical indicators, providing users with a "target" value that signifies the likelihood of heart disease. Additionally, the tool will prioritize user privacy by ensuring that all data is processed anonymously. Our platform is anticipated to contribute to preventive healthcare efforts by raising awareness and encouraging proactive measures. In the long term, we hope to reduce the burden of heart disease through accessible digital tools that promote early detection and health consciousness.

V. CONCLUSIONS

In conclusion, addressing heart disease through accessible and preventative tools is crucial, as early detection can significantly reduce the risk of severe health outcomes. Many

individuals face barriers to healthcare, and providing an easy-to-use platform allows them to assess their risk and take proactive steps towards better health. By empowering users with this knowledge, we hope to contribute to the prevention of heart disease, ultimately improving lives and reducing the global burden of cardiovascular diseases.

REFERENCES

[1] Centers for Disease Control and Prevention, "Heart Disease Facts," Heart Disease, Apr. 29, 2024. https://www.cdc.gov/heart-disease/data-research/facts-stats/index.html

[2] World Heart Federation, "CVD Prevention," World Heart Federation, 2023. https://world-heart-federation.org/what-we-do/prevention/

[3] D. Lapp, "Heart Disease Dataset," www.kaggle.com, 2019. https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset