AIT-664-DL3– Represent, Process & Visualize Applied Information Technology

PROJECT MILESTONE 1

UNVEILING CARDIOVASCULAR HEALTH PATTERNS USING STATISTICAL AND PREDICTIVE ANALYTICS

GROUP- 1

ABHISHEK ANUMALLA
AAKASH BOENAL
PAVAN TEJAVATH
SHASHANK YELAGANDULA

Prof. EBRIMA CEESAY

TABLE OF CONTENTS

INTRODUCTION

Heart disease remains a formidable health concern worldwide, exerting a profound impact on both individuals and communities. As a leading cause of mortality globally, heart-related fatalities surpass even cancer-related deaths in prevalence. Addressing this pervasive issue necessitates a comprehensive understanding of the multifaceted factors contributing to heart disease and its diverse manifestations. By unraveling intricate patterns and relationships embedded within extensive datasets, our study endeavors to shed light on the underlying risk factors and predictors associated with heart disease. Through meticulous analysis, we aim to discern the predictive power of various demographic, clinical, and lifestyle variables in assessing an individual's susceptibility to heart disease. Furthermore, our research aspires to extend beyond mere observation, delving into actionable insights that can inform the development of targeted strategies for early detection, prevention, and management of heart disease. By enhancing our understanding of the complex interplay between risk factors and disease outcomes, we aspire to catalyze advancements in cardiovascular healthcare, ultimately fostering improved health outcomes for both individuals and communities worldwide.

WHY THE PROBLEM IS IMPORTANT

Understanding the significance of the problem is crucial as heart disease stands as a pervasive threat to public health, claiming countless lives globally each year. It not only ranks as the leading cause of mortality but also imposes a substantial burden on individuals, families, and healthcare systems worldwide. Heart-related deaths continue to have a severe impact on people, families, and healthcare systems throughout the world despite tremendous advances in medical research. Developing successful prevention and management methods requires an understanding of the complex variables causing heart disease and its different symptoms. We seek to understand the complex links and patterns that underlie cardiovascular health by analyzing a large dataset that includes clinical and demographic characteristics. Our goal is to uncover important heart disease risk factors and predictions using rigorous data analysis and modeling methodologies.

This study is important in ways that go beyond scholarly interest. By providing guidance for evidence-based initiatives and policies targeted at reducing the burden of heart disease, it has the potential to have a direct impact on public health outcomes. We want to provide useful tools for early diagnosis and tailored risk assessment by utilizing machine learning and predictive analytics. This will enable people and healthcare providers to take on this difficult health challenge together. Furthermore, the fact that the goals of our project are in line with more public health objectives highlights the significance of this work and its potential to improve general well-being. We hope to make a significant contribution to the development of cardiovascular health research and practice by utilizing cutting-edge technologies and data-driven insights. Our ultimate goal is to create a healthier future for people all over the world.

ABOUT THE DATASET

In this project, we have acquired the dataset from Huggingface. It contains 70000 records in total with 12 features and a target variable ‘Cardio’. This research aims to analyze a dataset comprising the following features: age (recorded in days), height (in centimeters), weight (in kilograms), and gender (categorical code). Examination features encompass systolic and diastolic blood pressure measurements (ap_hi and ap_lo, respectively), as well as cholesterol and glucose levels, categorized into normal, above normal, and well above normal. Subjective features involve smoking status, alcohol intake, and physical activity engagement, represented as binary variables. The target variable indicates the presence (1) or absence (0) of cardiovascular disease in patients. These comprehensive features offer a multidimensional view of cardiovascular health, enabling detailed analyses and insights into potential risk factors and their impact on disease prevalence. [1]

LITERATURE REVIEW

Prediction of heart disease at early stage using data mining and big data analytics

The literature review in the paper focuses on the utilization of data mining (DM) models and techniques for forecasting heart disease (HD) based on patient datasets. It emphasizes how crucial data mining is for drawing insightful conclusions from massive volumes of medical data, assisting in the early diagnosis and prevention of heart disease. The review covers a range of DM approaches, including support vector machines, neural networks, naïve bayes, decision trees, genetic algorithms, K-NN, and clustering algorithms. These methods are used to create prediction models for heart disease with the goal of enhancing patient outcomes and diagnostic precision. The review also highlights the importance of big data analytics in managing extensive and intricate medical datasets, making it easier to extract useful information for the prediction of heart disease. Overall, the review of the literature highlights how DM and big data analytics can be combined to improve predictive modeling and develop strategies for managing heart disease. [2]

Analysis of heart disease using statistical techniques

This paper centers on heart disease, emphasizing its diverse manifestations and related risk factors, including age, gender, obesity, smoking history, and particular symptoms like dyspnea and chest pain. Binary logistic regression is used in the research methodology to analyze data and determine associations between these risk factors and the chance of developing heart disease. The findings show that depression, obesity, smoking history, and chest pain are important indicators of heart disease. Understanding the connection between these variables and the presence of cardiovascular disease is possible thanks to the logistic regression model. The model's suitability for predicting heart disease risk is confirmed by statistical tests used to evaluate the model's goodness of fit. The significance of logistic regression in medical research is emphasized in the conclusion, especially when it comes to identifying and controlling cardiovascular risk factors to avoid unfavorable outcomes. [3]

Cardiovascular disease analysis using supervised and unsupervised data mining techniques

The paper addresses the critical issue of cardiovascular diseases (CVDs), which are a leading cause of global mortality. Because CVDs have a major impact on public health, the study highlights the significance of early detection and treatment. A dataset with 14 characteristics linked to the diagnosis of heart disease is analyzed using a variety of data mining techniques, such as decision trees, support vector machines, Bayesian networks, and k-nearest neighbors. The methodology consists of preparing the dataset, applying various data mining algorithms, and segmenting the data using Simple K-Means clustering. Findings show that the support vector machines approach performed best in terms of precision (97.70%) and recall (97.70%), indicating that it is a useful diagnostic tool for cardiovascular disorders. The study highlights the potential of data mining techniques in healthcare decision-making and advances the development of diagnostic tools for CVDs. [4]

Machine Learning prediction in cardiovascular diseases: a meta-analysis

A machine learning (ML) method was presented by Krittanawong et al. (2020) to predict mortality following cardiac arrest by analyzing electrocardiogram (ECG) parameters. By using machine learning algorithms to evaluate ECG data and forecast patient outcomes the research advances customized treatment in critical care environments. The study looks closely at ECG features such waveform patterns and abnormal signals in an effort to find predictive indicators linked to post-arrest death. The application of machine learning techniques indicates that prediction accuracy can be improved beyond conventional risk assessment methodologies. The importance of combining clinical data with computational methods to enhance patient care and decision-making in cardiac emergencies is highlighted by this work. In the end, the results help doctors by providing insightful information on how to use technology-driven methods for risk prediction and stratification in cardiac arrest scenarios. [5]

Primary Prevention of Cardiovascular Disease: A special report from AHACC
The American Heart Association and American College of Cardiology released a special study in 2019 that Lloyd-Jones et al. published. The paper focused on the use of risk assessment tools to guide decision-making in the primary prevention of atherosclerotic cardiovascular disease (ASCVD). The study emphasizes how crucial it is to precisely determine each person's cardiovascular risk profile to successfully direct tailored preventative measures. The paper offers thorough guidance on the selection and use of risk assessment techniques such as the American College of Cardiology/American Heart Association Pooled Cohort Equations and other validated risk scores by integrating the most recent research and expert opinions. These instruments incorporate a range of clinical, lifestyle, and demographic variables to calculate a person's chance of developing ASCVD events within a given time period. [6]

PROPOSED APPROACH

The proposed approach includes data collection, cleaning, modeling, exploratory data analysis (EDA), visualization, development of a prediction tool (tentative). The proposed approach entails a systematic methodology for leveraging machine learning techniques to analyze a comprehensive dataset on cardiovascular health. Beginning with data collection from reliable sources, rigorous data cleaning and preprocessing steps will be undertaken to ensure data quality and consistency. Exploratory data analysis (EDA) will follow, involving statistical analysis and visualization techniques to uncover patterns, correlations, and trends within the dataset. Feature selection and engineering will then be employed to identify relevant variables and enhance model performance. Subsequently, various machine learning algorithms will be applied to develop predictive models for heart disease risk assessment, with careful evaluation and validation to ensure robustness and generalization. Interpretation of model results will provide insights into the factors influencing heart disease risk. Finally, stakeholder engagement and communication will facilitate the dissemination of research findings and the development of evidence-based interventions and policies aimed at improving cardiovascular health outcomes. We are planning to develop a **prediction tool with a potential web user interface** to predict the heart disease of an individual based on their values.

PROPOSED METHODS

The target variable showing the presence or absence of cardiovascular illness is accompanied by 12 features in each of the 70,000 patient data records that make up the dataset. The dataset provides a comprehensive view of the patients' health profiles by incorporating many input feature types, such as Objective, Examination, and Subjective data. We intend to implement the data cleaning process by first checking for the duplicates, null values, discrepancies in the data set and omitting them. For the exploratory data analysis, we are using Python(Datacamp) which would assist us in gaining valuable information with respect to the dataset. We intend to implement various machine learning models and find out which model is going to be the most accurate one. We will be implementing the models(Logistic Regression, SVM, Random Forest, Decision tree, Gradient Boosting Classifiers etc.) in Python. Furthermore, we intend to take our research to the next level by planning to develop a potential heart disease prediction tool which will provide the user an interface to predict if an individual has heart disease or not based on his vitals/data. The development of prediction tool is tentative and is contingent upon the workload throughout the semester.

PRELIMINARY RESULTS

Through correlation analysis, we observed significant associations between various demographic, clinical, and lifestyle variables and the presence of cardiovascular disease (CVD). Firstly, we tried to visualize the distribution of each numerical variable in the dataset stratified by the presence or absence of CVD represented by the cardio variable. Notably, age and cholesterol demonstrated a moderate positive correlation with CVD, suggesting an increased risk with advancing age. Furthermore, we analyzed and compared the ap_hi and ap_lo trends for people with and without cardiovascular, there by gaining more insights. indicating their potential role as predictors of cardiovascular risk.

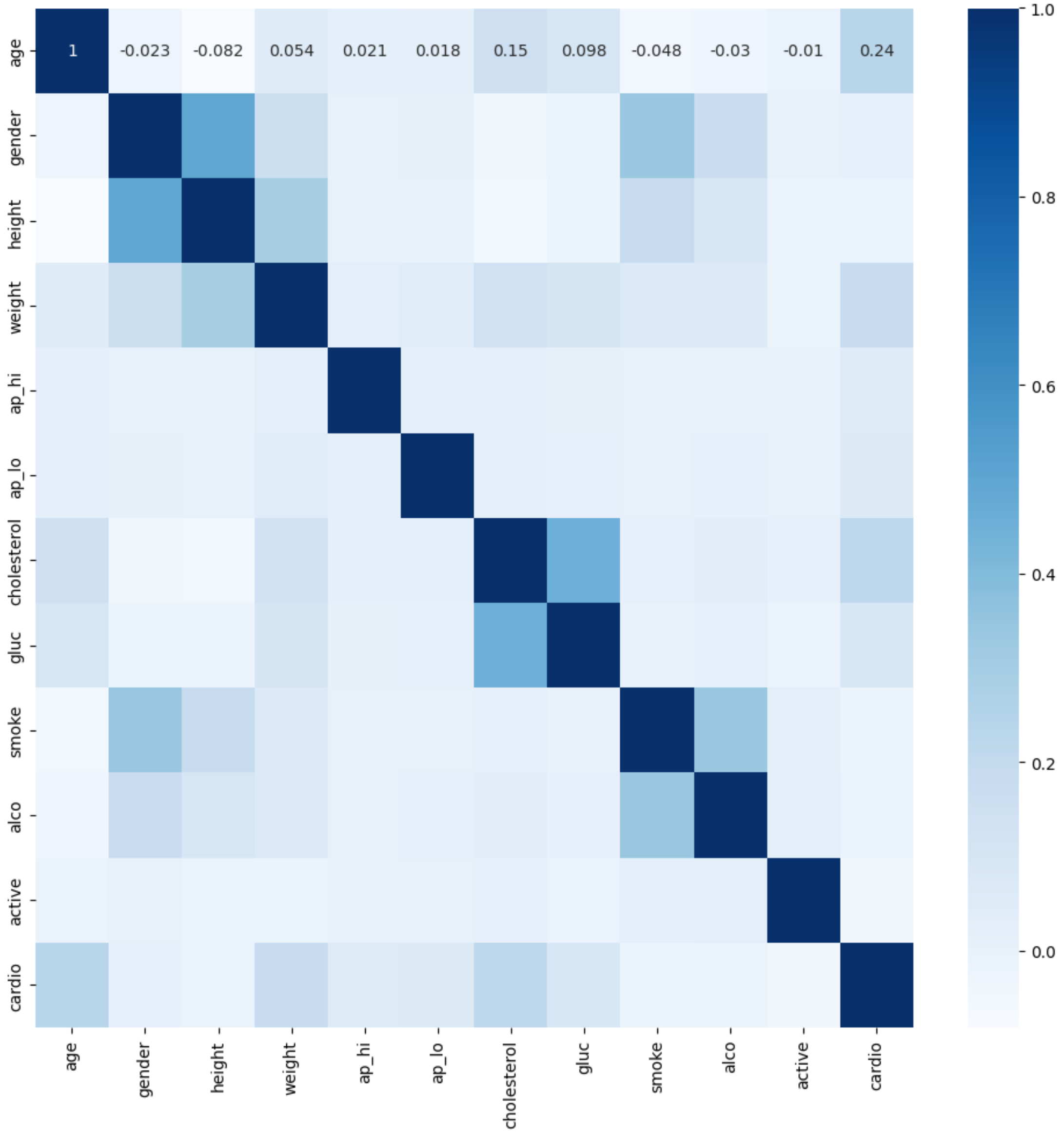

Fig 1: Correlation analysis to visualize the relationships between multiple attributes

Visualizations complemented our correlation findings, providing a comprehensive understanding of the distribution and relationships among key variables. Additionally, histograms and kernel density plots highlighted the scaling distribution of numerical variables, offering insights into the variability and spread of age, height, weight, and blood pressure measurements across the dataset.

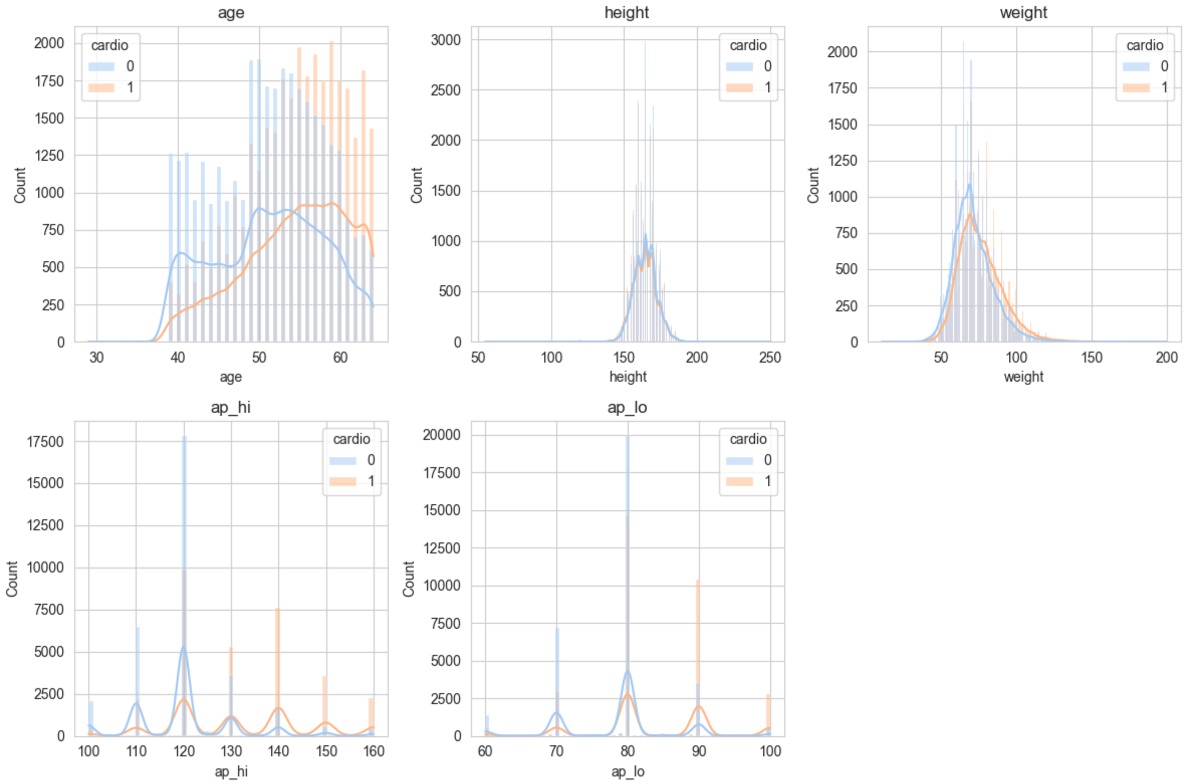

Fig 2: Histograms and kernel density plots to visualize the scaling distribution of numerical variables

The below plots illustrated the age, cholesterol effects and also the distribution of systolic and diastolic blood pressure measurements, showcasing differences between individuals with and without CVD.

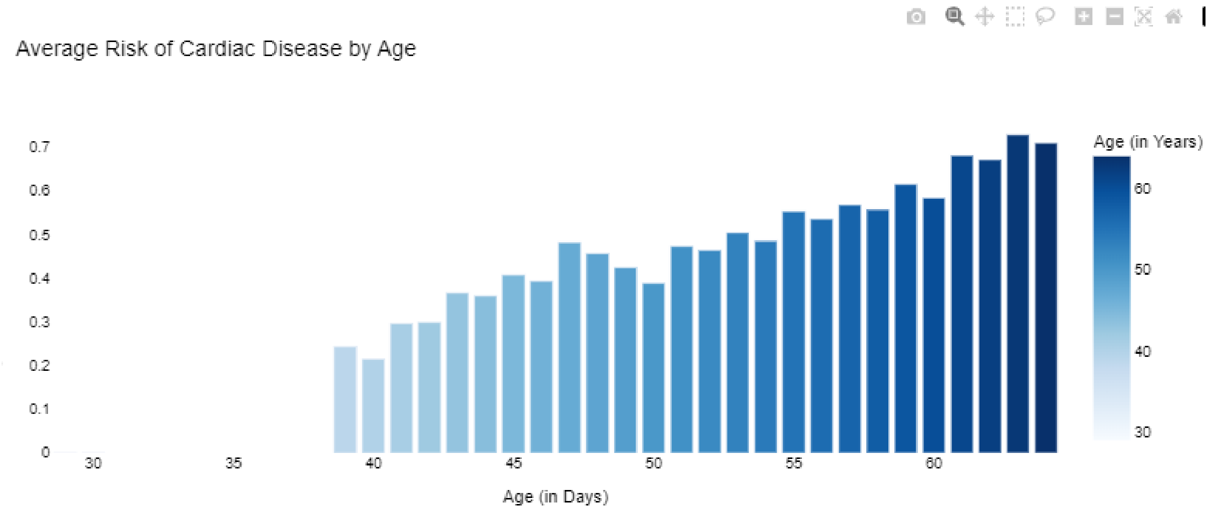

Fig 3: Bar plots to visualize average risk of Cardiac disease by Age

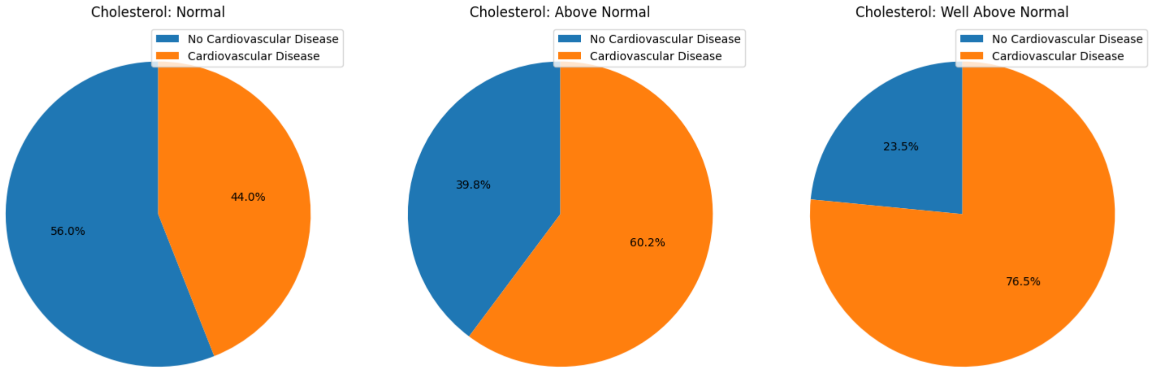

Fig 4: Pie charts to visualize the risk of Cardiac disease by Cholesterol levels

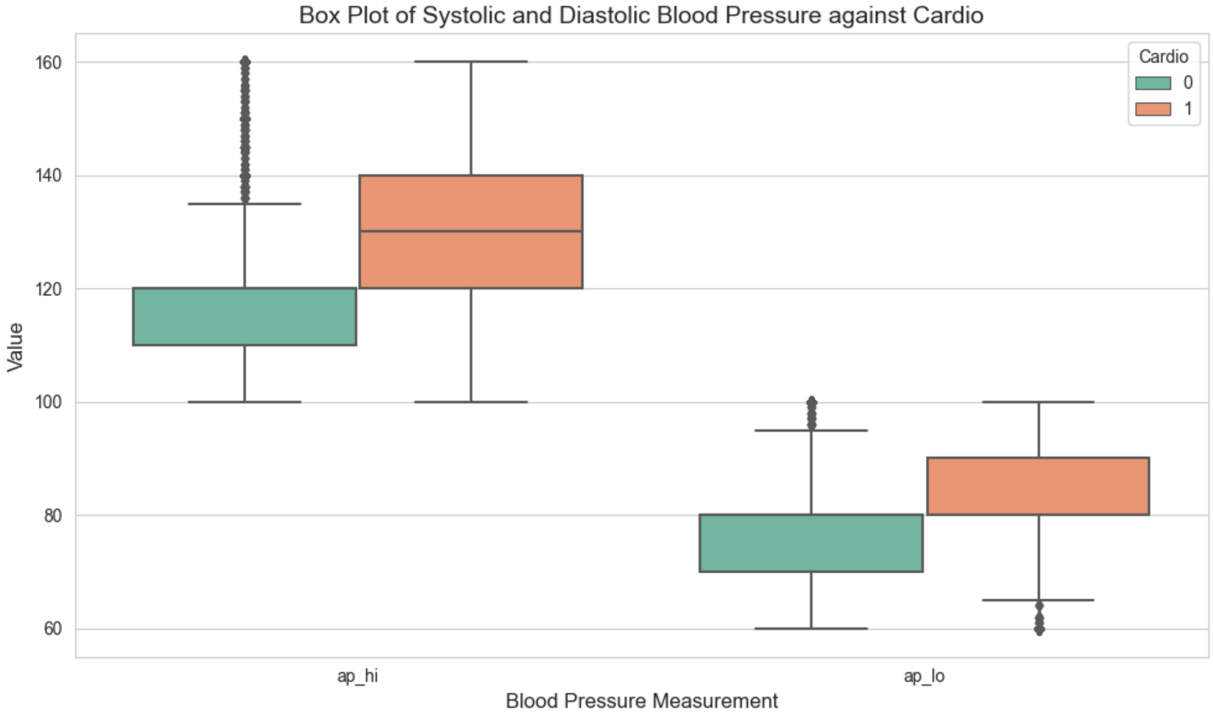

Fig 5: Box plot to visualize the Systolic and Diastolic Blood Pressure against Cardio

These preliminary results underscore the complexity of cardiovascular health and underscore the importance of considering multiple factors in assessing disease risk. Moving forward, further analysis and modeling efforts will be conducted to elucidate the predictive power of these variables and develop robust strategies for early detection, prevention, and management of cardiovascular disease. Through continued exploration and interpretation of the dataset, we aim to enhance our understanding of cardiovascular health and contribute to improved health outcomes for individuals and communities.

PROJECT LINKS

Dataset Link: https://huggingface.co/datasets/AlexCambell/HeartFailureDataset [1]
Project Website: https://mason.gmu.edu/~aanumall/

PROJECT TIMELINE

We are still working out the exact details, but we intend to complete the project in a time period of 10-12 weeks. Below is the tentative schedule for the project:

TASK NAME	TENTATIVE TIME	STATUS
Project Initiation Phase	Week 1	Done
Data Collection	Week 2-3	Done
Project Proposal	Week 4	Done
Data Cleaning & Preparation Phase	Week 5-6	Done
Project Milestone 1	Week 7-8	Done
Visualizations	Week 9	Done
Model Development Phase	Week 10	
Project Milestone 2	Week 11	
Model Evaluation and Interpretation Phase	Week 12	
Heart Disease Prediction Tool Development	Week 13	
Reporting and Presentation Phase	Week 14	
Final Report	Week 15	
Final Project Submission	Week 16	

REFERENCES

[1] S. S. Salma Banu, Prediction of heart disease at early stage using data mining and big data analytics: A survey, IEEE, 2017. https://ieeexplore.ieee.org/document/7955226

[2] R. G. Priyadarshini, Analysis of heart disease using statistical techniques, IOPScience, 2021. https://iopscience.iop.org/article/10.1088/1742-6596/1770/1/012105/meta

[3] A. P. J. Fabio Mendoza, Cardiovascular Disease Analysis Using Supervised and, JSW-Journal of Software, 2016. https://www.jsoftware.us/index.php?m=content&c=index&a=show&catid=178&id=2727

[4] Alex, Heart Failure Dataset, Huggingface. https://huggingface.co/datasets/AlexCambell/HeartFailureDataset

[5] Krittanawong, Machine learning prediction in cardiovascular diseases: a meta-analysis, Scientific Reports, 2020. https://doi.org/10.1038/s41598-020-72685-1

[6] D. Jones, Special Report on CVD by AHAAC, Science Direct, 2018. https://www.sciencedirect.com/science/article/pii/S0735109718390363?via%3Dihub