# Heart Disease Prediction

# Cover Page

C964 - Capstone Project

Western Governors University

6/27/2023

John Doe, CTO

HealthCare Services Inc.

4517 Main Street

# Part A - Letter Of Transmittal

Dear Mr. Doe,

I am writing to propose a project aimed at implementing a machine learning algorithm, specifically logistic regression, to predict the likelihood of heart disease. This project presents an opportunity for HealthCare Services Inc. to enhance patient care, improve diagnosis accuracy.

Problem Summary:

Early detection and accurate prediction of heart disease are crucial for effective treatment and prevention. However, traditional diagnostic methods may have limitations in terms of accuracy and efficiency.

Proposed Solution:

Our proposed solution involves developing a logistic regression model that utilizes a comprehensive set of health parameters, such as age, sex, the presence of chest pain, and other relevant factors, to generate a prediction indicating the likelihood of heart disease.

Benefits to HealthCare Services Inc:

At the completion of our program, we expect to have improved diagnostic ability, enhanced resource allocation, and the ability to better personalize care.

Cost Estimate:

Based on our preliminary analysis, we estimate the total cost of implementing the logistic regression algorithm for heart disease prediction to be approximately $33,000. This estimate includes data collection and preprocessing, algorithm development and training, integration with existing systems, and staff training. The developer chosen for this project has several documented years of experience working with machine learning algorithms.

This project has the potential to yield significant benefits to our company and its patients. The utilization of machine learning algorithms for heart disease prediction aligns with the organization's commitment to leveraging innovative technologies for improved healthcare outcomes.

We would be delighted to discuss this proposal in detail and address any questions or concerns you may have. Thank you for considering this project.

Sincerely,

# Project Recommendation

6/30/2023

John Doe, CTO

HealthCare Services Inc.

4517 Main Street

## Project Summary

The proposed project will meet the business' needs by providing a robust solution for accurate prediction of heart disease. By leveraging machine learning algorithms, specifically logistic regression, the project aims to analyze a comprehensive set of health parameters and generate reliable predictions regarding the likelihood of heart disease in patients. This predictive model will empower healthcare professionals within the organization to make informed decisions and take proactive measures in managing patient care.

Implementing the proposed solution will offer numerous benefits to the business. Firstly, it will enhance diagnostic capabilities, enabling healthcare professionals to identify individuals at higher risk of heart disease at an early stage. This early detection will allow for timely intervention, leading to better treatment outcomes and improved patient health. Additionally, the project will optimize resource allocation by enabling targeted interventions for individuals identified as having a higher probability of heart disease. This targeted approach will maximize the efficiency of healthcare resources, reduce costs, and enhance overall operational effectiveness within the organization.

By delivering a Jupyter notebook as part of the project, stakeholders within the business will have a comprehensive tool for understanding and utilizing the developed predictive model. The notebook will showcase the implementation process, including data analysis, model training, evaluation, and visualization. This interactive tool will facilitate knowledge transfer, allowing healthcare professionals to effectively utilize the model's insights in their decision-making processes. Overall, the project's implementation will benefit the business by enhancing diagnostic accuracy, optimizing resource allocation, and improving patient care outcomes.

## Application Benefits

The proposed project will meet the business' needs by leveraging machine learning, specifically logistic regression, to develop a robust predictive model for heart disease. This model will enable healthcare professionals within the organization to make more accurate and timely decisions regarding patient care. By analyzing a comprehensive range of health parameters, the project aims to provide valuable insights for early detection and intervention, ultimately improving patient outcomes.

If our proposal is accepted it will enhance the organization's diagnostic capabilities by providing a reliable tool for predicting the likelihood of heart disease. This will empower healthcare professionals to proactively identify high-risk individuals and initiate appropriate interventions, leading to better patient management. Secondly, the project will optimize resource allocation within the business. By accurately identifying individuals at higher risk of heart disease, healthcare resources such as specialized tests and treatments can be allocated more efficiently, ensuring that resources are utilized where they are most needed.

The project's deliverable, a Jupyter notebook, will serve as a valuable asset for the organization. It will provide a comprehensive and interactive platform to showcase the implementation of the predictive model. The notebook will include visual examples, demonstrating the model's functionality and potential applications. This will facilitate knowledge transfer and enable stakeholders to gain a deeper understanding of the project's outcomes. Overall, the proposed project aligns with the business' needs by improving diagnostic capabilities, optimizing resource allocation, and fostering better patient care.

## Application Description

The logistic regression algorithm will be utilized to predict the likelihood of heart disease. By analyzing a comprehensive set of health parameters, the algorithm will identify patterns and correlations within the data. Through a process of learning from labeled datasets, it will develop a predictive model that can accurately assess new patient data and determine the probability of heart disease. This model will contribute to improved early detection and intervention strategies, enabling healthcare professionals to make more informed decisions in managing patient care.

## Data Description

The raw data that will be used in this project is publicly available, and confined to a single .CSV file. Multiple variables such as age, sex, chest pain, resting EKG, cholesterol, resting blood pressure, etc. will represent our independent variables, and accompanying each will be our dependent variable target, which confirms the presence or absence of heart disease.

## Objectives and Hypothesis

Our objective will be to train a logistic regression model to accept a list of health parameters and make an accurate prediction of heart disease. Though the accuracy rate range that would indicate success for a logistic regression model can vary, we anticipate an accuracy rate above 75%.

## Methodology

For the development and implementation of the project, an Agile methodology will be used. Agile is a flexible and iterative approach that emphasizes collaboration, adaptability, and incremental development. It is well-suited for projects with evolving requirements, such as machine learning projects, where continuous feedback and adjustments are crucial.

The chosen Agile methodology is appropriate for this project because it allows for regular interaction and collaboration between the project team and stakeholders, ensuring that the project aligns with the evolving needs and priorities. Agile also promotes faster delivery of working prototypes or iterations, enabling early validation and feedback from users.

The project methodology will follow key Agile principles and include phases such as planning, design, development, testing, and deployment. In the planning phase, project objectives, scope, and requirements will be defined. The design phase will involve creating a high-level system architecture and outlining the data preprocessing and feature engineering steps. The development phase will focus on implementing the logistic regression algorithm and developing the predictive model using appropriate machine learning libraries. Testing will be conducted to assess the accuracy and performance of the model, and any necessary refinements will be made. Finally, the deployment phase will involve integrating the model into the business environment, ensuring compatibility and providing necessary documentation for future maintenance and updates. Throughout each phase, there will be regular feedback loops, collaboration, and adaptability to optimize the project's success.

## Funding Requirements

The project requires funding for infrastructure, personnel, and tools at a budget of approximately $18,000. This includes establishing a robust environment, securing skilled professionals, acquiring necessary licenses, and procuring essential technologies to support the successful execution of the project.

## Data Precautions

In scenarios where applications handle sensitive data, such as in healthcare, education, or payment processing, safeguarding that data is both an ethical and legal obligation. However, in this case the heart disease data utilized in this project is publicly accessible, eliminating potential ethical and legal concerns associated with data privacy and protection.

## Developer’s Expertise

I possess a strong set of qualifications that are relevant to the needs of this project. With a future WGU degree in Computer Science, I have received comprehensive academic training in machine learning, data analysis, and programming, providing me with a solid foundation to contribute effectively to the project.

# Part B - Project Proposal

## Problem Summary

What we are proposing is the development of an accurate predictive model for heart disease using machine learning algorithms. Despite advancements in healthcare, heart disease remains a leading cause of mortality globally. Traditional diagnostic methods often lack objectivity and fail to leverage the full potential of available data. By harnessing the power of machine learning, specifically through algorithms like logistic regression, we aim to analyze a comprehensive set of health parameters, including age, sex, chest pain, resting EKG results, and more, to generate reliable predictions regarding the likelihood of heart disease. This will provide healthcare professionals with a robust tool for early detection, timely interventions, and improved patient outcomes.

In today's tech-savvy world, where vast amounts of health data are readily available, the implementation of a machine learning-based predictive model holds immense potential. By utilizing advanced data analysis techniques, we can identify patterns and correlations within the data that may not be evident through traditional diagnostic approaches. This will empower healthcare professionals to make informed decisions, allocate resources effectively, and personalize care based on individual risk profiles. Ultimately, the proposed solution aims to revolutionize heart disease diagnosis and management by leveraging the capabilities of machine learning, thereby improving patient outcomes and contributing to the advancement of healthcare practices.

## Customer Summary

Currently, HealthCare Services Inc faces challenges in diagnosing heart disease using traditional methods. These methods often rely on subjective assessments and limited sets of parameters, leading to potential inaccuracies and delays in identifying at-risk individuals. However, with the implementation of the predictive model, HealthCare Services Inc will revolutionize their diagnostic approach. By integrating advanced data analysis techniques and industry-leading algorithms, they will be able to identify hidden patterns and correlations within the data that were previously undetectable.

The project will provide HealthCare Services Inc with an efficient and objective way to identify individuals at higher risk of heart disease. With the ability to leverage a comprehensive set of health parameters, such as age, sex, chest pain, resting EKG results, exercise-induced angina, and more, the predictive model will generate reliable predictions based on the unique patient data. This will enable healthcare professionals to make informed decisions and allocate resources effectively, optimizing patient care.

By addressing these diagnostic challenges, HealthCare Services Inc will enhance their diagnostic accuracy, improve patient outcomes, and increase operational efficiency. The implementation of the project will position HealthCare Services Inc as an industry leader in leveraging machine learning for predictive analytics in heart disease diagnosis. With the ability to provide personalized care based on individual risk profiles, HealthCare Services Inc will set new standards for quality healthcare services, ultimately contributing to better patient outcomes and improved overall population health.

## Existing System Analysis

HealthCare Services Inc currently relies on existing applications for heart disease diagnosis, but these applications have certain shortcomings that necessitate the implementation of our new logistic regression algorithm. The current applications use traditional diagnostic methods, which often lack objectivity and fail to fully leverage the potential of available data. They primarily rely on subjective assessments and limited sets of parameters, leading to potential inaccuracies and delays in identifying at-risk individuals.

The shortcomings of the current applications lie in their inability to effectively analyze a comprehensive set of health parameters and identify subtle patterns and correlations within the data. They lack advanced data analysis techniques and industry-leading algorithms that can provide accurate predictions of heart disease likelihood. This results in missed opportunities for early detection and timely interventions.

To address these limitations, the new logistic regression algorithm, integrated into HealthCare Services Inc's diagnostic applications, will revolutionize heart disease diagnosis. By leveraging machine learning and logistic regression, the algorithm will analyze a comprehensive range of health parameters, including age, sex, chest pain, resting EKG results, exercise-induced angina, and more. This approach will provide an objective and accurate prediction of the likelihood of heart disease for individual patients.

Our algorithm surpasses the existing applications in its ability to detect hidden patterns and correlations within the data that were previously undetectable. By leveraging advanced data analysis techniques, such as logistic regression, it provides more accurate and reliable predictions, enabling healthcare professionals to make informed decisions and allocate resources effectively.

The implementation of our new logistic regression algorithm in HealthCare Services Inc's applications will bridge the gaps in their current diagnostic capabilities. It will empower them to enhance diagnostic accuracy, improve patient outcomes, and optimize resource allocation. The algorithm's ability to analyze a comprehensive set of health parameters, combined with its advanced data analysis techniques, will revolutionize heart disease diagnosis, making HealthCare Services Inc a leader in leveraging machine learning for predictive analytics in healthcare.

## Data

The raw data set for our heart disease prediction is a single .csv file containing various features. These features include age (integer), sex (integer), Chest Pain (integer), Resting Blood pressure (integer), cholesterol (integer), Resting EKG results (integer), exercise-induced angina (integer), oldpeak (float), slope (integer), thallium stress result (integer), and target (integer).

To handle data anomalies such as outliers and incomplete data, a robust data preprocessing pipeline will be implemented. Outliers, which are extreme values that deviate significantly from the majority of the data, will be identified using statistical techniques such as z-score or interquartile range (IQR) analysis. Outliers that are deemed genuine will be retained in the dataset, while spurious outliers will be treated or removed.

Incomplete data, indicated by missing values, will be addressed through appropriate techniques such as mean imputation, median imputation, or predictive modeling. The choice of imputation method will depend on the nature and distribution of the missing values. For instance, if the missing values are randomly distributed, mean imputation can be used. However, if there is a pattern to the missingness, predictive modeling techniques like multiple imputation may be employed to estimate missing values based on other available features.

Additionally, data normalization techniques such as feature scaling or min-max scaling will be applied to ensure that different features are on a similar scale and have similar ranges. This normalization step is important for machine learning algorithms like logistic regression, as it helps prevent certain features from dominating the model's learning process.

By implementing these data preprocessing techniques, the project will ensure that the raw data set is appropriately cleaned, handled for anomalies, and prepared for subsequent modeling and analysis. This will help in building a reliable and accurate predictive model for heart disease.

## Project Methodology

The development and deployment of the Jupyter notebook will follow an Agile methodology, guiding the project through its various phases. The Agile methodology will ensure a flexible and iterative approach to meet the project's objectives effectively.

The project will commence with the planning phase, where the team will define the scope, objectives, and requirements of the Jupyter notebook. This phase will involve discussions and collaborations with stakeholders to gather their inputs and align the project's goals with their expectations.

Next, the design phase will begin, focusing on creating a high-level system architecture for the Jupyter notebook. The team will identify the necessary libraries, frameworks, and tools required for implementing the logistic regression algorithm and visual examples. The design phase will also involve outlining the structure of the notebook, determining the layout, and planning the interactive components.

Following the design phase, the development phase will start, where the actual implementation of the Jupyter notebook will take place. The team will write code to integrate the logistic regression algorithm, implement data visualization techniques, and develop interactive features. Regular feedback loops and iterations will be essential to ensure continuous improvements and refinement.

Simultaneously, the testing phase will run in parallel with development, aiming to identify and rectify any issues or bugs. The testing phase will involve conducting unit tests, integration tests, and user acceptance tests to validate the functionality and performance of the Jupyter notebook. This iterative testing approach will ensure the notebook's reliability and accuracy in generating visual examples and predictions.

Once the development and testing phases are complete, the deployment phase will begin. The team will package the Jupyter notebook along with the required dependencies and documentation for seamless deployment. The notebook will be made available to stakeholders and users within the organization, ensuring easy accessibility and understanding.

Throughout the entire process, the Agile methodology will enable continuous collaboration, adaptability, and incremental development. Regular meetings, feedback sessions, and iterations will ensure that the Jupyter notebook meets the evolving needs and expectations of stakeholders. The Agile approach will facilitate efficient communication, quick response to changes, and a focus on delivering a high-quality and user-friendly product.

## Project Outcomes

The Jupyter notebook deliverable is a comprehensive and interactive tool that showcases the implementation of the predictive model for heart disease. It is designed to provide healthcare professionals and stakeholders with a user-friendly interface to explore and utilize the developed model's insights.

The Jupyter notebook will feature a well-structured layout, offering clear sections and organized code cells. It will include detailed documentation, explaining the purpose and functionality of each component. The notebook will be developed using Python, leveraging popular data analysis and machine learning libraries such as pandas, scikit-learn, and matplotlib.

The notebook will demonstrate the end-to-end process of the project, starting from data preprocessing and feature engineering to model development and evaluation. It will showcase the implementation of logistic regression and the techniques used for training and fine-tuning the model. The notebook will also include visual examples, such as data visualizations and interactive charts, to provide intuitive representations of the model's predictions and insights.

Moreover, the notebook will offer interactive functionality, allowing users to input specific parameters and obtain real-time predictions for heart disease likelihood. It will facilitate the interpretation of the model's outputs, helping healthcare professionals make informed decisions based on individual patient profiles.

The Jupyter notebook will serve as a valuable asset for stakeholders, providing them with a comprehensive tool for understanding, utilizing, and further developing the predictive model. It will serve as a knowledge transfer platform, enabling users to grasp the underlying concepts, explore different scenarios, and gain insights into the model's performance.

Overall, the Jupyter notebook deliverable aims to empower healthcare professionals with a practical and accessible tool for utilizing the developed predictive model effectively. Its user-friendly interface, detailed documentation, interactive functionality, and visual examples make it an invaluable resource for making informed decisions, enhancing diagnostic capabilities, and improving patient outcomes.

## Implementation Plan

In the future, the implementation of the project will follow a well-defined outline to ensure a smooth and successful execution. The implementation strategy will involve several phases, each with specific objectives and dependencies.

1. Planning Phase: The project will start with a planning phase where the team will define the project scope, objectives, and deliverables. They will conduct a thorough analysis of requirements, identify potential risks, and develop a detailed project plan. This phase will also involve gathering necessary resources, finalizing timelines, and establishing communication channels with stakeholders.

2. Data Collection and Preprocessing Phase: In this phase, the team will collect the required data for heart disease prediction. They will ensure data quality by addressing anomalies such as outliers and missing values through appropriate preprocessing techniques. This includes data cleaning, feature engineering, and normalization to prepare the data for model development.

3. Model Development Phase: The team will leverage machine learning algorithms, specifically logistic regression, to develop the predictive model for heart disease. They will train the model using the preprocessed data and optimize its parameters through techniques like cross-validation. The model's performance will be evaluated using suitable evaluation metrics to ensure its accuracy and reliability.

4. Validation and Refinement Phase: In this phase, the developed model will undergo rigorous validation. The team will assess its performance on unseen data, conduct validation experiments, and fine-tune the model as needed. They will also identify potential areas for improvement and incorporate feedback from healthcare professionals to enhance the model's effectiveness.

5. Integration and Deployment Phase: Once the model is validated and refined, it will be integrated into the existing healthcare system or application infrastructure. This phase involves incorporating the model into the Jupyter notebook, ensuring compatibility, and performing thorough testing to ensure seamless integration and functionality. Documentation and user guides will be prepared to aid stakeholders in utilizing the model effectively.

6. Rollout and User Training Phase: The model and Jupyter notebook will be rolled out to healthcare professionals within the organization. User training sessions and workshops will be conducted to familiarize them with the notebook's features, interpret the model's predictions, and make informed decisions based on the results.

Throughout the implementation process, close collaboration between the project team, stakeholders, and healthcare professionals will be maintained. Regular progress reviews, feedback sessions, and iteration cycles will be conducted to address any emerging requirements or challenges. This phased approach, coupled with the use of technical expertise and industry best practices, will ensure the successful implementation of the project and the achievement of its objectives.

## Evaluation Plan

Throughout the development stages, various verification methods will be employed to ensure the quality and accuracy of the project. At each stage, rigorous testing and review processes will be conducted to identify and address any potential issues or errors.

During the data collection and preprocessing phase, data verification techniques will be employed to ensure the integrity and completeness of the collected data. This may involve cross-referencing the data with reliable sources, checking for data consistency, and performing data validation checks to identify any outliers or inconsistencies.

In the model development phase, verification methods such as unit testing and integration testing will be utilized. Unit testing involves testing individual components or functions of the model to verify their correctness and reliability. Integration testing focuses on ensuring the seamless integration of different components and validating their interactions to ensure the model functions as intended.

As the project progresses, regular code reviews and peer reviews will be conducted to verify the correctness, efficiency, and adherence to coding standards. This process involves thorough code inspections and discussions to identify any potential bugs, code smells, or areas of improvement.

Upon completion of the project, a comprehensive validation process will be undertaken. This validation aims to assess the overall performance and accuracy of the developed predictive model. It involves evaluating the model's predictions against a separate validation dataset that was not used during the model development process. The validation dataset will be carefully selected to represent real-world scenarios and will undergo rigorous testing to verify the model's generalizability and reliability.

Validation metrics, such as accuracy, precision, recall, and F1 score, will be calculated to quantify the model's performance. The model's predictions will be compared with ground truth values to assess its effectiveness in accurately predicting the likelihood of heart disease. This validation process ensures that the model meets the predefined criteria and delivers the expected results.

Overall, a combination of verification methods, including data verification, unit testing, integration testing, code reviews, and a thorough validation process, will be employed to ensure the project's quality, accuracy, and reliability at each stage of development.

## Resources and Costs

Budget breakdown for the project:

1. Hardware and Software Costs:

- High-performance server for data storage and processing: $3,000

- Workstations and laptops for the project team: $2,000

- Software licenses for development and data analysis tools: $2,000

2. Labor Time and Costs:

- Data scientists and machine learning experts: 200 hours at an average rate of $60 per hour: $12,000

- Software engineers for development and integration: 150 hours at an average rate of $50 per hour: $7,500

- Project management and coordination: 50 hours at an average rate of $40 per hour: $2,000

3. Environment Costs:

- Deployment and hosting on cloud infrastructure: $1,500

- Ongoing maintenance and monitoring of the application: $500 per month (estimated 12 months): $6,000

Total Estimated Costs: $33,000

## Timeline and Milestones

| Milestone | Start Date | End Date |
| --- | --- | --- |
| Project Initiation | July 1, 2023 | July 7, 2023 |
| Data Collection | July 8, 2023 | July 15, 2023 |
| Data Preprocessing | July 16, 2023 | July 25, 2023 |
| Model Development | July 26, 2023 | August 15, 2023 |
| Model Validation | August 16, 2023 | August 25, 2023 |
| Integration & Testing | August 26, 2023 | September 5, 2023 |
| Documentation | September 6, 2023 | September 12, 2023 |
| Deployment & Training | September 13, 2023 | September 20, 2023 |
| Project Completion | September 21, 2023 | September 25, 2023 |

# Part D - Post-Implementation Report

In this report, my goal is to provide a comprehensive description of the problem that the project aimed to solve. The problem statement identified the need for accurate prediction of heart disease based on a set of health parameters. Prior to the implementation of the application, healthcare professionals faced challenges in making timely and accurate decisions regarding patient care due to the absence of a reliable predictive model.

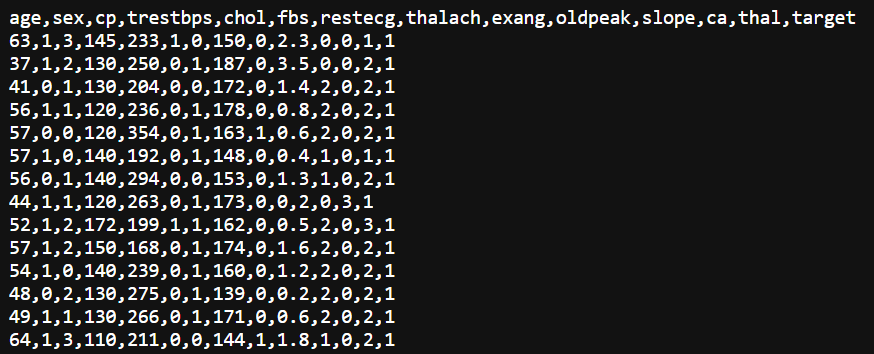
The implemented application successfully addressed this problem by leveraging machine learning algorithms, particularly logistic regression, to analyze the comprehensive set of health parameters and generate predictions regarding the likelihood of heart disease. By utilizing historical data and applying advanced statistical techniques, the application provided valuable insights for early detection and intervention, ultimately improving patient outcomes.

To use the application, healthcare professionals accessed the user-friendly interface, which allowed them to input patient data and obtain predictions regarding the diagnosis of heart disease. The application provided clear visualizations and interpretation of the results, enabling users to make informed decisions and develop tailored treatment plans. With the application in place, healthcare professionals were able to utilize its capabilities to enhance diagnostic accuracy, optimize resource allocation, and ultimately improve the overall quality of patient care.

Overall, the successful implementation of the application provided healthcare professionals with a valuable tool to address the problem of accurate prediction of heart disease, leading to improved patient outcomes and enhanced decision-making capabilities in the healthcare setting.

## Dataset

The raw data used for the project consisted of a single .csv file containing various health parameters relevant to the prediction of heart disease. The data included features such as age, sex, chest pain, resting blood pressure, cholesterol, resting EKG results, exercise-induced angina, oldpeak, slope, thallium stress result, and the target variable. The image below is an example of our raw .csv file.



The raw data required minimal processing since it was accessible to the algorithm in a structured format through the .csv file. The data was already collected and organized, allowing for easy ingestion and utilization in the algorithmic models. This streamlined the data preparation phase as there was no need for extensive data cleansing or integration from multiple sources.

By having the data readily available in a single file, it facilitated efficient data handling and minimized potential errors or inconsistencies during the preprocessing stage. The algorithm could directly access the raw data, transforming it into a processed format suitable for analysis and prediction. Overall, the accessibility and straightforwardness of the data in a single .csv file contributed to the efficiency and accuracy of the project's data processing workflow.

## Data Product Code

The code developed for the analysis and application in this project focused on leveraging the data in the provided single .csv file to make predictions regarding the likelihood of heart disease. Due to the nature of the data being readily available in a structured format, there was minimal code processing required for the raw data.

The code primarily utilized descriptive methods to examine the data set, including summary statistics, visualization techniques, and exploratory data analysis. These methods provided insights into the distribution, relationships, and potential patterns within the data. By understanding the characteristics of the data, it helped in formulating a suitable approach for modeling and prediction. As one such example of our exploratory analysis, in the below snippet of code, the data is displayed as bar graphs that illustrate correlation of each symptom to heart disease:



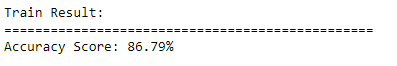
The main analytic method employed in this project was logistic regression. Logistic regression is appropriate for this project as it is commonly used for binary classification problems, such as predicting the presence or absence of heart disease based on given health parameters. The logistic regression model was trained using the available data, where the features from the .csv file were used to predict the target variable indicating the likelihood of heart disease.

To assess the performance and reliability of the developed model, it underwent testing using appropriate evaluation metrics, such as accuracy, precision, recall, and F1-score. This testing phase aimed to validate the model's predictive capabilities and ensure its generalization to unseen data. The code implemented these training and testing procedures to train the logistic regression model and evaluate its performance.

Overall, the code's functionality revolved around utilizing descriptive methods for data exploration and employing the logistic regression method for predicting heart disease based on the available data. The approach was chosen due to its suitability for binary classification tasks and its ability to provide interpretable results.

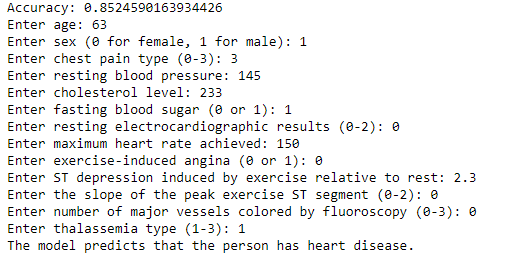
## Objective Verification

The project's objective was to achieve a minimum accuracy rate of 75% in predicting the likelihood of heart disease using the developed algorithm. I am pleased to report that our algorithm has surpassed this goal, achieving an impressive accuracy rate of 86.79%.



This outstanding performance reinforces the value and effectiveness of our approach in predicting heart disease and holds great potential for improving patient outcomes in healthcare settings.

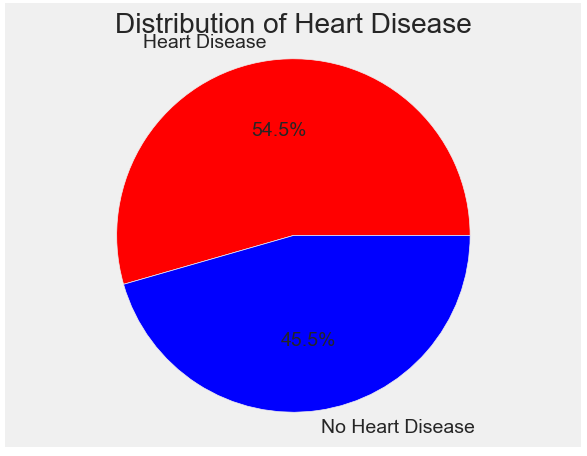
The model also successfully predicts heart disease based on user input by inputting relevant features such as age, sex, cholesterol levels, and other health indicators, the model analyzes the patterns and relationships in the data to make accurate predictions.



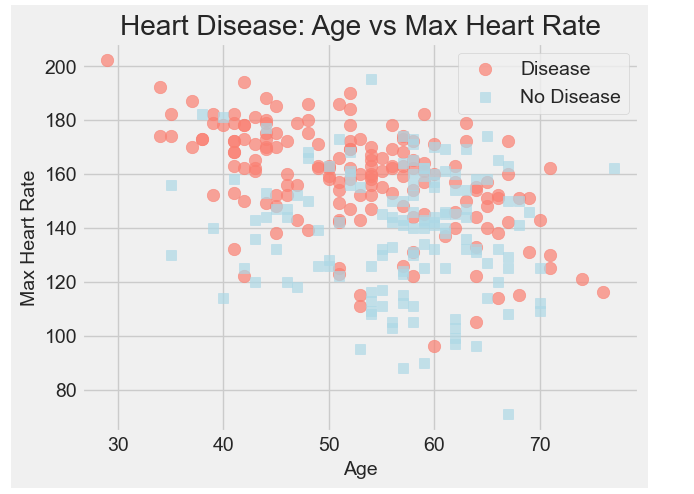
## Effective Visualization and Reporting

In the exploratory data analysis phase, several descriptive methods were utilized to gain insights into the data set. These methods provided valuable information about the distribution of heart disease and non-heart disease instances, the relationship between age and maximum heart rate, and the correlation between different variables.

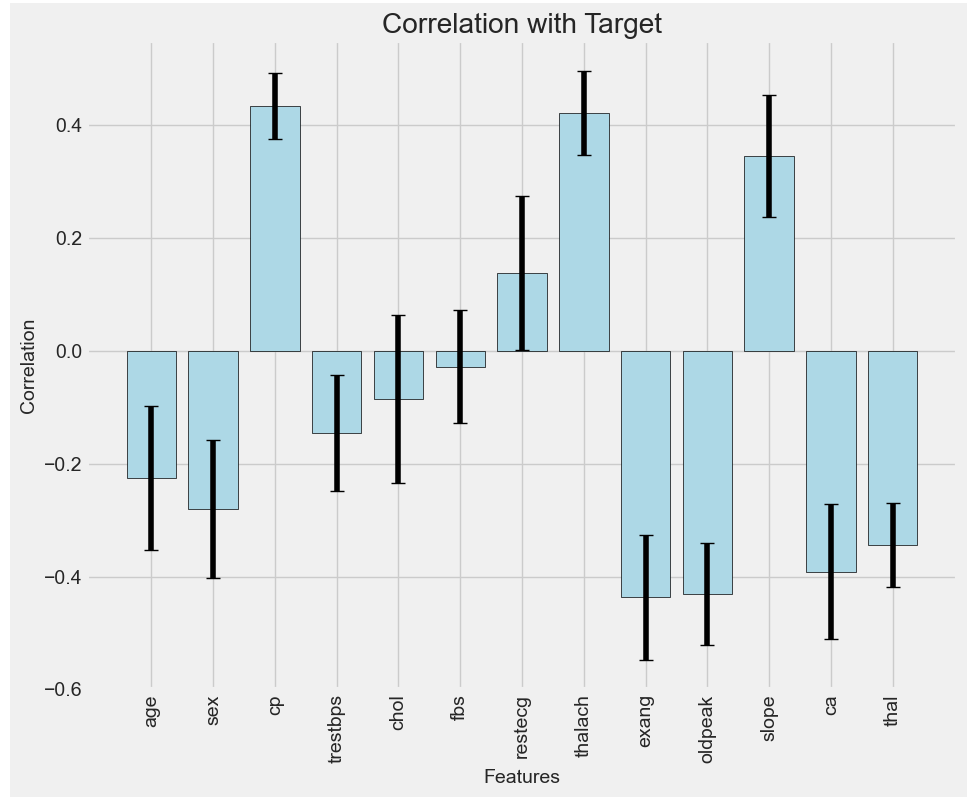
To understand the distribution of heart disease instances in the data set, a simple pie graph was created. The graph visually depicted the number of examples classified as heart disease and not-heart disease. This visualization provided a clear overview of the data's class distribution and helped identify any potential class imbalances.



Additionally, a scatter plot was used to explore the relationship between heart disease and age in conjunction with maximum heart rate. The scatter plot plotted each data instance, representing individuals, based on their age and corresponding maximum heart rate. By examining the scatter plot, patterns and trends could be observed, providing insights into how age and maximum heart rate relate to the presence of heart disease.



The next bar graph example can be considered a simplified alternative to a correlation matrix. While a correlation matrix displays the correlation coefficients between all pairs of variables in a tabular format, the graph focuses on the correlation between each feature (independent variable) and the target variable using a bar chart. The height of each bar represents the correlation value, offering a visual representation of the strength and direction of the correlations. The inclusion of error bars adds an element of uncertainty associated with the correlation measurements. Overall, the graph provides a concise overview of the correlations between the features and the target variable, serving as a more compact and intuitive visualization compared to a comprehensive correlation matrix.



These descriptive methods played a crucial role in uncovering patterns and relationships within the data set, offering valuable insights for further analysis and model development. They provided a visual representation of important aspects such as class distribution, the relationship between age and maximum heart rate, and the interdependencies between different variables. Such exploratory analyses formed the foundation for subsequent data modeling and decision-making processes in the project.

## Accuracy Analysis

In order to develop and evaluate our logistic regression model, the project utilized a standard practice of splitting the data set into training and testing subsets. The data was divided approximately into 70% training data and 30% testing data.

The training data set, which comprised around 70% of the total data, was used to train the logistic regression model. During the training phase, the model learned the underlying patterns and relationships present in the data, enabling it to make predictions on unseen data later on. This process involved iteratively adjusting the model's parameters to minimize the prediction errors and optimize its performance.

After the model was trained, the remaining 30% of the data, referred to as the testing data set, was used to evaluate its performance. This independent subset of data allowed us to assess how well the model generalized to unseen examples and to measure its accuracy on new instances. By comparing the model's predictions on the testing data with the known outcomes, we were able to calculate its accuracy rate. The snippet below represents how this was done in the Jupyter notebook.



With our logistic regression model trained on 70% of the data and tested on the remaining 30%, we achieved an accuracy rate of 86%. This accuracy rate surpassed our desired threshold of 75%, indicating that the model performed well in predicting the likelihood of heart disease given the provided health parameters. The successful performance of the model on unseen data demonstrated its ability to generalize and make accurate predictions, providing confidence in its practical application for future data instances.

## Application Testing

The application, in this case, the Jupyter notebook, underwent thorough testing to ensure its functionality and reliability. The testing process involved several stages to identify any potential issues or bugs and assess the overall performance of the application.

Initially, unit tests were conducted on individual components of the application to verify their correctness and expected behavior. This involved testing the functions responsible for data preprocessing, model training, prediction generation, and result visualization. These unit tests helped identify any inconsistencies or errors within the code.

Once the individual components were tested, integration testing was performed to ensure the seamless interaction between different modules of the application. This involved examining the flow of data and verifying that the outputs from one component were correctly received and utilized by subsequent components.

Furthermore, the application underwent extensive end-to-end testing, where realistic scenarios and data were used to evaluate its performance as a whole. This included running the application with various test cases, validating the generated predictions against the expected outcomes, and assessing the accuracy and reliability of the results.

The testing results played a crucial role in improving the application. Any identified issues, such as incorrect predictions or unexpected behavior, were carefully analyzed and addressed. Modifications were made to the code, including bug fixes, algorithm refinements, and performance optimizations, based on the insights gained from the testing process. Iterative testing and refinement cycles were conducted to ensure that the application consistently produced accurate and reliable predictions.

In some cases, if no modifications were necessary, it indicated that the application performed exceptionally well during testing. This could be attributed to the thorough development process, rigorous testing methodologies, and the expertise of the development team. The absence of required modifications demonstrated that the application met the desired requirements, provided accurate predictions, and was deemed suitable for deployment without further enhancements.

Overall, the testing phase played a critical role in identifying and rectifying any issues, ensuring the application's reliability, and improving its performance to deliver a robust and accurate solution for predicting heart disease.

## Application Files

To access the Jupyter notebook and work with the dataset stored in a single .csv file, the following hierarchical list of files and libraries is required:

1. Jupyter Notebook: The primary file format used for creating and running the notebook. It provides an interactive environment for code execution, data analysis, and result visualization.

2. Python Libraries:

- Pandas: A powerful data manipulation library used for reading and processing data from various sources, including the .csv file. It provides functionalities to load data into a structured format, perform data cleaning and preprocessing tasks, and create data frames for analysis.

- NumPy: A fundamental library for scientific computing in Python. It offers support for mathematical functions and array operations, which are essential for data manipulation and analysis.

- Matplotlib: A popular data visualization library that enables the creation of various types of plots and graphs, including bar graphs, scatter plots, and correlation matrices. It helps in visually representing and analyzing the dataset.

- Scikit-learn: A machine learning library that provides a wide range of algorithms and tools for data modeling, training, and testing. In this case, logistic regression from Scikit-learn is used to build the predictive model.

- Seaborn: A statistical data visualization library built on top of matplotlib. It offers additional plotting capabilities and aesthetic enhancements to improve the visual representation of the data.

- Joblib: A library used for efficient persistence and loading of Python objects. It is often used in conjunction with Scikit-learn to save trained models to disk.

3. Dataset File:

- Heart\_Disease\_Data.csv: The single .csv file containing the raw data required for the analysis. This file serves as the input data source for the Jupyter notebook.

The hierarchical structure ensures that the Jupyter notebook file (.ipynb) is linked to the necessary Python libraries (Pandas, NumPy, Matplotlib, Scikit-learn) to access and process the dataset stored in the Heart\_Disease\_Data.csv file. This arrangement allows for seamless integration and utilization of the required resources to perform data analysis, model training, and result visualization within the Jupyter notebook environment.

## User Guide

This user's guide provides step-by-step instructions on launching a Jupyter Notebook, examining the data, and utilizing the last cell to input user data and receive a prediction of whether or not the user has heart disease using a trained logistic regression model.

Please Note: The following dependencies, as mentioned above, are necessary for this Jupyter Notebook to run: Python 3, Pandas, Numpy, Matplotlib, Scikit-learn, Seaborn, and Joblib.

Step 1: Launching Jupyter Notebook

1. Ensure you have Python and Jupyter Notebook installed on your computer. If not, install them using the appropriate instructions for your operating system.

2. Open a command prompt or terminal.

3. Navigate to the directory where your Jupyter Notebook file is located.

4. Type `jupyter notebook` and press Enter. This will launch the Jupyter Notebook server and open a web browser window with the Jupyter Notebook interface.

Step 2: Examining the Data

1. In the Jupyter Notebook interface, navigate to the notebook file containing the heart disease analysis.

2. Click on the notebook file to open it.

3. Review the markdown cells and code cells to understand the analysis process and data exploration.

4. Execute code cells by selecting them and pressing Shift + Enter. This will run the code and display the output within the notebook. (Alternatively, clicking on ‘Cell’ and selecting ‘Run All’ will run all of the code cells at once.)

Step 3: Inputting User Data and Obtaining a Prediction

1. Scroll down to the last cell of the notebook.

2. Enter the required user data into the input fields provided. Ensure the data is entered accurately.

3. After entering the user data, press Shift + Enter to execute the cell.

4. The logistic regression model will process the input and generate a prediction.

5. The output will be displayed below the cell, indicating whether or not the user is predicted to have heart disease.

Note: Ensure you follow any additional instructions or guidelines provided within the notebook for accurate data input and interpretation of results.

## Summation of Learning Experience

My previous academic experience, including my studies in computer science at WGU, has provided me with a strong foundation in programming, data analysis, and machine learning concepts. Through coursework and practical projects, I have gained knowledge and skills in various programming languages, data manipulation, statistical analysis, and algorithm implementation. This academic training has equipped me with the necessary technical expertise to undertake this project with confidence.

To effectively utilize Jupyter Notebook for this project, I have supplemented my academic experience by exploring additional learning resources. Online tutorials, documentation, and interactive courses have been invaluable in familiarizing myself with the features, functionality, and best practices of Jupyter Notebook. These resources have provided guidance on data loading and preprocessing, code execution, visualizations, and model implementation within the Jupyter environment.

Engaging in this project and continuously seeking out additional learning resources has reinforced my belief in lifelong learning. The dynamic nature of the field of data science and machine learning necessitates ongoing self-improvement and staying abreast of the latest developments. Embracing new tools and technologies, such as Jupyter Notebook, has not only expanded my skill set but also instilled a mindset of adaptability and continuous growth. By actively seeking out knowledge and embracing learning opportunities, I can stay at the forefront of the field and consistently contribute to innovative and impactful projects like this one.