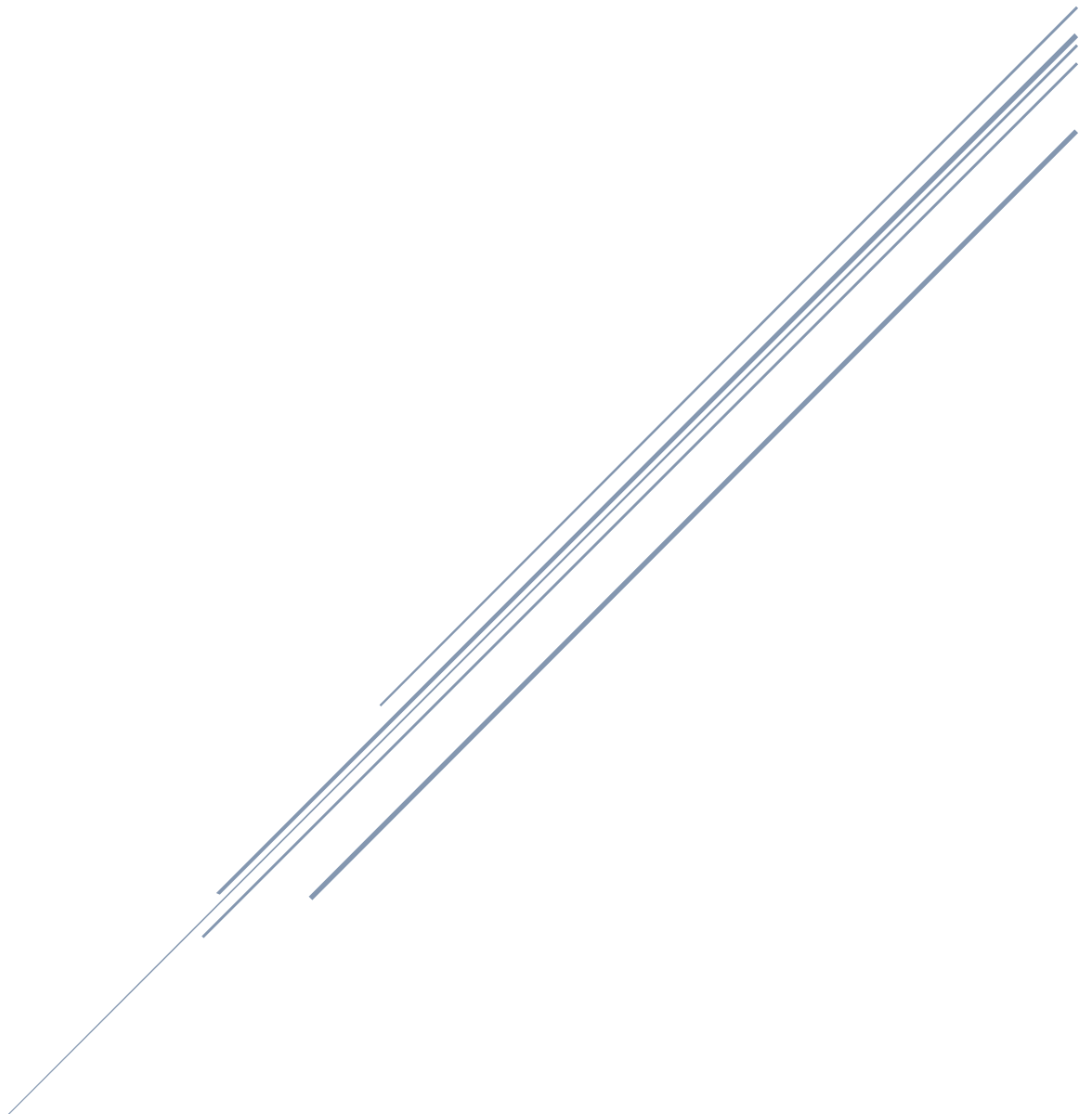


HEART DISEASE PREDICTION USING MACHINE LEARNING

John Randall McKahan II – Student ID:



Western Governors University

C964 – Computer Science Capstone

Table of Contents

Section A – Page 2

Letter of Transmittal – Page 2

Project Recommendation – Page 3

Section B – Page 9

Problem Statement – Page 9

Customer Summary – Page 10

Existing System Analysis - Page 11

Data – Page 11

Project Methodology – Page 12

Project Outcomes – Page 13

Implementation Plan – Page 13

Evaluation Plan – Page 15

Resources and Costs – Page 16

Timeline and Milestones – Page 17

Section C – Page 18

Data Methods – Page 18

Datasets – Page 19

Analytics – Page 19

Data Cleaning – Page 19

Data Visualization – Page 20

Real-Time Queries – Page 20

Adaptive Element – Page 21

Outcome Accuracy – Page 21

Security Measures – Page 22

Product Health Monitoring – Page 22

Section D – Page 23

Project Purpose – Page 23

Datasets – Page 23

Data Product Code – Page 24

Hypothesis Verification – Page 24

Visualization and Reporting – Page 25

Accuracy Analysis – Page 26

Application Testing – Page 26

Application Files – Page 26

User's Guide – Page 27

Summation of Learning – Page 34

Section E – Page 35

Sources – Page 35

October 29, 2020

Dr. John H. Watson, Director

Greathealth Hospital 221b

Baker Street.

London, WV 25126

Subject: Letter of Transmittal – Heart Disease Application

Dear Dr. Watson,

It has come to my attention that many of your colleagues are struggling to keep up with an increasing number of cardiac patients. This has become problematic, as Greathealth Hospital is currently operating with sufficient staffing, and the recent addition of supplementary cardiac doctors has not proved beneficial in reducing the workload of your colleagues. Even with ample testing, your doctors are struggling to find the time to properly assess the presence of heart disease in patients.

I would like to take this time to recommend investing in a machine learning application that can take patient data and predict, much like a doctor, whether a patient has heart disease. This machine learning application would be a standalone application, no internet required, and capable of running on most any modern computer. This application would be easy to use, making data entry a breeze, and would allow your medical professionals to see disease prediction in real time. Say a doctor wants to see if a patient's heart disease risk will increase with age? This application would be able to do it!

Imagine your doctors being able to spend more time with patients, and less time calculating and second guessing themselves over their disease predictions. This is an attainable goal, and one that would allow Greathealth Hospital to continue delivering the best in personalized healthcare. The amount of time saved on calculations alone would increase the efficiency Greathealth Hospital.

The objective of this project is to create an easy-to-use standalone software solution, with a high degree of accuracy – at least 80%, that allows medical professionals to input patient data, and predicts the percent chance of heart disease.

In order to complete this project, funding in the amount of \$20,000.00 will need to be secured. This amount, though steep, is a one-time payment amount in which Greathealth Hospital will reap benefits from for years to come.

I will be the developer of this proposed solution, I have a year of experience in developing software solutions, and I am able to provide samples of my previous work upon request.

I look forward to your correspondence,

Respectfully,

Project Recommendation

Problem Summary

Currently, Greathealth Hospital doctors are overwhelmed with their current cardiac patient workloads. This has reduced the amount of time that medical professionals are able to spend providing the personalized healthcare that Greathealth Hospital is known for. Within the scope of this project is the creation a standalone application which can run on most any modern Windows 10 based computer. The application will display graphs based upon heart disease data obtained by other medical institutions.

Users, intended to be medical professionals, will be able to easily input patient medical data, and obtain heart disease prediction percentages, with an accuracy of at least 80%. Higher accuracies may be achieved but are not guaranteed. The proposed software will not feature the ability to save patient information, or input sensitive information such as names, birthdates, and social security numbers. This is designed to uphold HIPAA compliance, and to prevent the theft of sensitive data. The application will have the capability of being hosted on a local network or over the internet, however, the client will assume all security risks and responsibility if they choose to host the application in this manner as the proposed application is intended to be used without a network connection.

Application Benefits

The proposed application will benefit Greathealth Hospital medical professionals by allowing them to spend more time with patients and less time calculating data and making predictions. Additionally, this application will benefit the average doctor's accuracy as well. A

2019 study by the Human Diagnosis Project showed that medical specialists had a diagnostic accuracy percent of only 66.3% (Michael L. Barnett, 2019). As the proposed software solution aims to provide a minimum accuracy of 80%, doctors would be benefitting not only by saving time on calculations, but by increasing their accuracy with the added benefit of the software acting as a second opinion.

Application Description

The proposed application will be developed in Python 3.83, using Anaconda3 and Streamlit 0.69.2. The application will be hosted locally on any of Greathealth Hospital's Windows 10 computer systems. As mentioned previously, this application does have the ability to be hosted over a local network, or the internet, but it is designed to be used on a machine without using a network connection. The proposed application will feature an easy-to-use graphical interface providing slider bars for doctors to select values of specific patient data, such as blood pressure, and buttons for selecting data such as the presence of chest pain in the patient. The dashboard of the application will feature an area that displays the patient data that is currently being selected from the sliders and buttons on the sidebar for the medical professional to confirm the data selected. Within the dashboard will be an area featuring an output saying whether the application predicts heart disease within the patient. Below the application's prediction will be an area displaying the risk percentage of heart disease. Furthermore, an area which displays the accuracy percentage of the machine learning model will be displayed in order to confirm the 80% or higher accuracy requirement. Lastly, the application will contain an area with three labeled graphs based upon previous data for the review of medical professionals.

Data Description

The data used within the proposed application will be a public dataset available at the following link <https://www.kaggle.com/volodymyrgavrysh/heart-disease>. The linked dataset was obtained by the Hungarian Institute of Cardiology, University Hospitals of Zurich and Basel Switzerland, the V.A. Medical Center of Long Beach, and the Cleveland Clinic Foundation. The data contains 304 records of patient data, saved in a comma separated value format, pertinent only to the presence or lack of heart disease, and does not contain sensitive information such as names or social security numbers. All data is within integer format and is composed of the following.

- Age
- Sex (Male = 1, Female = 0)
- Chest Pain (Asymptomatic = 0, Typical = 1, Non-Anginal = 2, Non-Typical = 3)
- Cholesterol (Blood serum in units of mg/dL)
- Fasting Blood Sugar (Blood sugar levels above 120 mg/dL, 0 = False, 1 = True)
- Resting ECG (Resting Electrocardiogram results. Normal = 0, Level 1 = 1, Level 2 = 2)
- Exercise Induced Angina (0 = False, 1 = True)
- Maximum Heart Rate (In beats per minute)
- ST Depression (ST segment Depression on ECG induced by exercise compared to resting)
- Slope (Slope of the peak of the ST segment of ECG induced by exercise)
- Ca (Number of major vessels colored by fluoroscopy)
- Thallium Stress Test Results (1 = Normal, 2 = Fixed Defect, 3 = Reversible Defect)
- Target (Presence of heart disease in patient, 0 = false, 1 = true)

The target, which is the presence of heart disease, is the dependent variable in this case with the independent variables of the patient's health data, such as maximum heart rate, contributing to the outcome of the heart disease diagnosis. The data to be used in the proposed software solution is limited in size, as there are only 304 data entries. Additionally, the patient data provided is only a snapshot of one's overall health data, there are many additional factors that could be used such as a family history of heart disease, weight, and height. Some anomalies are noted within the data, such as patients with excessively high blood pressure yet no heart disease.

Objectives and Hypothesis

The goal of the proposed software solution is to create a machine-learning heart disease prediction application with an easy-to-use graphical interface that medical professionals can use to assist in reducing their workload and enhancing their accuracy in heart disease prediction. The proposed project's objectives would include the ability to view previous data from the proposed data file in the form of scatter plot charts allowing doctors to see medical trends that may be related to the risk, or lack thereof, of heart disease.

The proposed project's hypothesis is that a heart disease prediction application can be created with an accuracy rating of at least 80%, which would make the application competitive with a medical specialist. Additionally, if successful, this project may pave the way for future research on disease prediction and the creation of highly accurate disease prediction tools.

Methodology

The proposed project plans to be developed using the waterfall methodology. As the proposed project is a relatively small project with well-defined requirements, the waterfall

methodology will best suit the development of the software. The following steps of the waterfall method will be used to produce and deliver the proposed solution.

1. **Requirements:** The requirements of the proposed software will be well defined in a meeting with the appropriate stakeholders and representatives from Greathealth Hospital.
2. **Design:** With the requirements obtained from the initial meeting, an overall design of the proposed software will be produced, which will include the expected features of the solution, as well as the programming languages used.
3. **Implementation:** Using the documents provided within the design stage, a functional software will be produced.
4. **Verification and Testing:** The proposed software will be verified against the initial requirements as well as tested to ensure as many bugs as possible are caught before the initial launch.
5. **Launch:** The software solution is delivered to the client, in this case Greathealth Hospital, and upon acceptance approval, installed on their hospital computer systems.

Funding Requirements

The funding requirements to produce and deliver the proposed software solution amounts to **\$20,000.00**. This amount is based upon the current scope and requirements as laid out previously. Additional features may be added to the proposed project at a negotiable cost should the client decide so. The proposed project will be produced using Python 3.83, Anaconda3, and Streamlit 0.69.2, which are all free to use. The developer of this project is incurring all environmental costs as they will be working from their home office. Additionally, costs are kept

lower as the personnel on this project consists of solely the developer. Greathealth Hospital currently has the computer infrastructure to run the proposed software and requires no additional hardware at this time.

Stakeholders Impact

The successful completion and delivery of this software solution is designed to add value to all stakeholders of Greathealth Hospital. Medical specialists will see value in time saved on “number-crunching” as well as by assisting them in providing the best medical care possible by enhancing their heart disease prediction accuracy. Patients, another stakeholder of Greathealth Hospital, will notice an increased level of personalized healthcare as Greathealth Hospital doctors will be able to allocate more time to working with patients on an individual level. Board members of Greathealth Hospital will see an increase in patient satisfaction ratings, patient retention, and increasing Greathealth Hospital’s reputation of being the leader in personalized healthcare.

Data Precautions

The dataset which the proposed software solution will be based upon contains no sensitive or protected information. All health-related information within the dataset cannot be traced back to any individual, and therefore is not protected by HIPAA. If full birthdates were to have been used, phone numbers, account numbers, medical record numbers, Social Security numbers, email addresses, facial photographic information, or any information which may be traced back to an individual were used then the dataset would have contained information protected by HIPAA. As the proposed dataset does not contain PHI (Protected Health Information), it is free from protection under HIPAA.

Developer's Expertise

The developer of the proposed software solution has one year of software development experience and is graduating from Western Governors University with a Bachelor of Science in computer science in November 2020. The combined software development experience with the rigorous and hands-on education of Western Governors University makes this developer the ideal candidate to develop the proposed solution properly, and within a timely manner. **Section B**

Problem Statement

Greathealth Hospital is currently overwhelmed with the number of cardiac patients they are receiving. Due to the increased workload, physicians and other medical professionals at Greathealth Hospital are unable to spend as much time with patients, and thus reducing the level of personalized healthcare that Greathealth Hospital is known for. The proposed solution aims to reduce the workload by calculating heart disease risk and assist medical professionals in making proper assessments of each patient's heart health. Medical professionals will have more time to work with patients individually and will spend less time performing calculations by hand. Additionally, the proposed software solution will have a high level of accuracy which will act as a virtual "second opinion" increasing the confidence of physicians in their predictions. The proposed software will include a graphical user interface featuring scatter plots based upon previously obtained data, easy to use sliders and buttons to input patient data, and a prediction as to whether a patient suffers from heart disease based upon the inputted data.

Customer Summary

The intended audience of the proposed software solution consists of medical professionals, specifically physicians and all associated medical assistants. Medical data obtained from patients will be easy to input into the proposed software solution, and from there, easy to read heart disease predictions will be displayed in a graphical user interface. This proposed software will greatly benefit those medical professionals who had previously spent hours crunching numbers to make determinations of the heart health of their patients. Additionally, research physicians may find the included scatter plots beneficial in determining links between certain health statistics and the presence of heart disease.

The proposed software is intended to be used within a medical office environment, specifically, on the computers of physicians and associated medical assistants. The reason a medical environment is advised is that the proposed software is to be developed with the intention of assisting medical professionals in real-time, and to provide patients with their results as quickly as possible.

Basic computer skills such as clicking and dragging and using the keyboard to input data will be required to properly operate and use the proposed software. While there aren't any specialized skills or training needed to obtain the output of the software, whether or not a patient has heart disease, it is important to note that the proposed software solution is designed to be used as a tool by medical professionals. The proposed software will be able to make a determination as to the health of the patient's heart based upon inputted medical data, but it is up

to a trained medical professional to make the final determination based upon their knowledge and experience.

Existing System Analysis

There is not an existing computer-based system for the analysis of heart disease risk in patients at Greathealth Hospital. Currently, physicians gather patient health information, such as blood pressure and heart rate, and use their knowledge and training to decide as to whether a patient is at risk for heart disease. The creation of the proposed software solution will provide a completely new tool, and act as a sort of virtual assistant or second opinion for healthcare workers at Greathealth Hospital.

Data

The proposed software will be using data obtained from a publicly available dataset, <https://www.kaggle.com/volodymyrgavrysh/heart-disease> This dataset contains information from various medical institutions including Hungarian Institute of Cardiology, University Hospitals of Zurich and Basel Switzerland, the V.A. Medical Center of Long Beach, and the Cleveland Clinic Foundation. There are 304 records within the dataset, which includes critical information such as the age of the patient, maximum heart rate, presence of chest pain, among other vital information. The information contained within the dataset is sufficient for determining the presence of heart disease in a patient, however it does not contain complete medical records. As the dataset is already cleaned and normalized, and every record contains a full set of values, no additional work is needed on the dataset in order to use it for the proposed software.

Project Methodology

The proposed project will be developed using the waterfall methodology. Waterfall methodology is suitable for this project as there are well defined requirements, and it is a relatively small project. Unlike agile development, in the waterfall methodology, the requirements of the proposed solution cannot be easily changed. The waterfall methodology, with its rigidity will allow for a faster development time, as well as providing Greathealth Hospital with an accurate cost for development as there will be no hidden costs.

Firstly, within the waterfall methodology requirements and documentation will be discussed and obtained with Greathealth Hospital. Currently, the developer has proposed the desired functionality of the product as well as a candidate dataset to be used. During this phase, Greathealth Hospital and the developer may negotiate the usage of alternative datasets or the inclusion or exclusion of certain features. Secondly, in the design stage of development, the overall design of the software solution will be developed, including the data models. Thirdly, using the documentation produced by the design phase, a functional software solution will be produced in the implementation stage. Fourthly, the produced software will be subjected to various testing methods to ensure a complete and functional product.

Each unit of the software solution will be subjected to white-box testing, that will test the internal workings of the units. Afterwards, units will be combined with other previously tested units and subjected to integration testing. Once the complete software solution is produced, it will then be subjected to system testing. Lastly, the produced program will be tested for acceptance, assuring that it meets the requirements with those of Greathealth Hospital.

Project Outcomes

There will be both project deliverables, as well as product deliverables associated with the development of the proposed software solution. The following lists are provided to break down what deliverables will be associated with the development of the product.

Project Deliverables

- 1. Letter of Transmittal**
- 2. Project scope**
- 3. Project Proposal**
- 4. Software Testing Plans**
- 5. Project Development Schedule**
- 6. Proposed Software Mockup**

Product Deliverables

- 1. A standalone heart disease prediction application**
- 2. The source code of the produced application**
- 3. User installation manual**

Implementation Plan

Once the software solution is created it will need to be implemented in order to be used by Greathealth Hospital. Outlined below is a plan which details how the proposed solution will be implemented once completed.

Strategy for Implementation

Greathealth Hospital currently has computers capable of running the completed software solution, and the ones to be used to host the proposed software have been previously identified. Anaconda3 and Streamlit 0.69.2, or higher versions if applicable at the time of installation, will be installed on the candidate computer. The developer will install the software solution on the chosen device of Greathealth Hospital, but an installation guide will be provided in the case that Greathealth Hospital wishes to install the developed software on additional devices.

Phases of Rollout

The proposed software will be developed using the waterfall methodology, and a complete software solution will be provided when development has finished. This means that the software will be rolled out in completion.

Details for Levels of Testing and Final Distribution

Unit testing, integration testing, system testing, and lastly acceptance testing will be performed on the proposed software solution after it has been developed. Once testing is completed, and the software functions properly and is accepted by Greathealth Hospital, the final distribution will be made available.

Dependencies and Milestones

Dependencies and milestones of the proposed software solution are provided further in this document, under the section header of “Timeline and Milestones”.

Deliverables

Project and product deliverables outlined within the section titled “Project Outcomes” will be provided to Greathealth Hospital at the completion of the project.

User Testing

Before the completion of the software, future users from Greathealth Hospital will be invited to beta test the application. Users from Greathealth Hospital may note desired changes or potentially discover bugs not previously caught during previous testing. Changes to the proposed software are not included in the scope of this project but may be added at an additional cost should the client request them.

Evaluation Plan

Once the product has been developed it will be evaluated to ensure it meets the requirements and needs of Greathealth Hospital. Software testing will be completed covering unit, integration, system, and afterwards acceptance testing. The accuracy rating of the software will be evaluated, using test data, to ensure an accuracy of 80% or higher as proposed to the client. The accuracy will be based upon the application successfully determining the presence or absence of heart disease given the sample patient data from the proposed dataset.

After the product has been tested for accuracy, and successfully meets the accuracy requirements of the client, it will then be evaluated for acceptance by users from Greathealth Hospital. Users will be asked to give feedback on their experience to ensure a high degree of user satisfaction with the product. Additionally, the proposed software solution does not interfere with

any regulatory requirements, such as HIPAA, and therefore does not contain any personally identifiable patient information.

Resources and Costs

Programming Environment

Hardware

1. Windows 10 Computer, 16GB DDR4 DRAM, Intel Core i7-7700 CPU, 1 TB HDD
2. 32" 1920x1080p Computer Monitor
3. Laser Mouse and Mousepad
4. Full Computer Keyboard

Software

1. Windows 10 Operating System
2. Anaconda 3
3. Python 3.83
4. Streamlit 0.69.2

Most of the software used to create the proposed software solution, Anaconda 3, Python 3.83, and Streamlit 0.69.2 are open source and are free to use. The developer is not charging a fee for the usage of their existing computer setup.

Environment Costs

Open-source software and programming languages are to be used in the creation of the proposed software. Therefore, there are no environmental costs associated with the production of the software. The proposed software is designed to be locally hosted on an existing machine

located at Greathealth Hospital. As no new hardware or proprietary software is required in the production and hosting of the proposed application, environmental costs are not factored in the development price.

Human Resource Requirements

The proposed project is estimated to take a total of 140 hours to complete. The developer is choosing to charge a flat rate of **\$125.00** per development hour, totaling **\$17,500.00**. This includes all hours spend gathering resources, developing, and testing the software. In addition, a consulting fee of **\$500.00** will be charged. Upon completion of the project the developer will install the completed software on a single machine at Greathealth Hospital. The installation fee is **\$2,000.00** and is to cover all associated travel costs as well as the developer's time to install the software. The total cost to complete this project is **\$20,000.00**.

Timeline and Milestones

Phase	Dependencies	Milestone	Start Date	End Date	Resources
Requirements	Requirements to be obtained from stakeholders	Requirement documentation produced	10/10/2020	10/15/2020	Stakeholders
Design	Requirements	Design documentation produced	10/16/2020	10/18/2020	Software Developer
Implementation	Design	Software solution is created	10/19/2020	10/24/2020	Software Developer

Verification and Testing	Implementation	Software solution is tested and meets all functional requirements	10/25/2020	10/30/2020	Software Developer
Launch	Verification and Testing	Software solution is shipped, and installed on Greathealth Hospital's computer system	11/1/2020	11/10/2020	Software Developer

Section C

Data Methods

Descriptive Method

K-means clustering was used in the creation of the software solution's visualization of previous data. Three different scatter plots were created using K-means clustering allowing the users to visualize heart disease trends across a variety of factors. The K-means clustering has an accuracy of 81.8%.

Non-Descriptive Method

Logistic regression was used in the creation of the software solution's heart disease prediction tool. As one of the goals of this project was to create a software solution that could determine the presence or absence of heart disease, logistic regression

was a suitable candidate for the non-description method as it works well with binary outcomes. The logistic regression model has an accuracy of 85.48%.

Datasets

The dataset used in the creation of the software solution is sourced from publicly available information and can be downloaded at the following web address.

<https://www.kaggle.com/volodymyrgavrysh/heart-disease>. The dataset contains 304 patient records, and is free from any personally identifiable health information, and therefore does not fall under the protection of HIPAA. As the dataset is in a CSV format, it was easy to use in the creation of the software by using Python 3.83's built in "CSV.reader".

Analytics

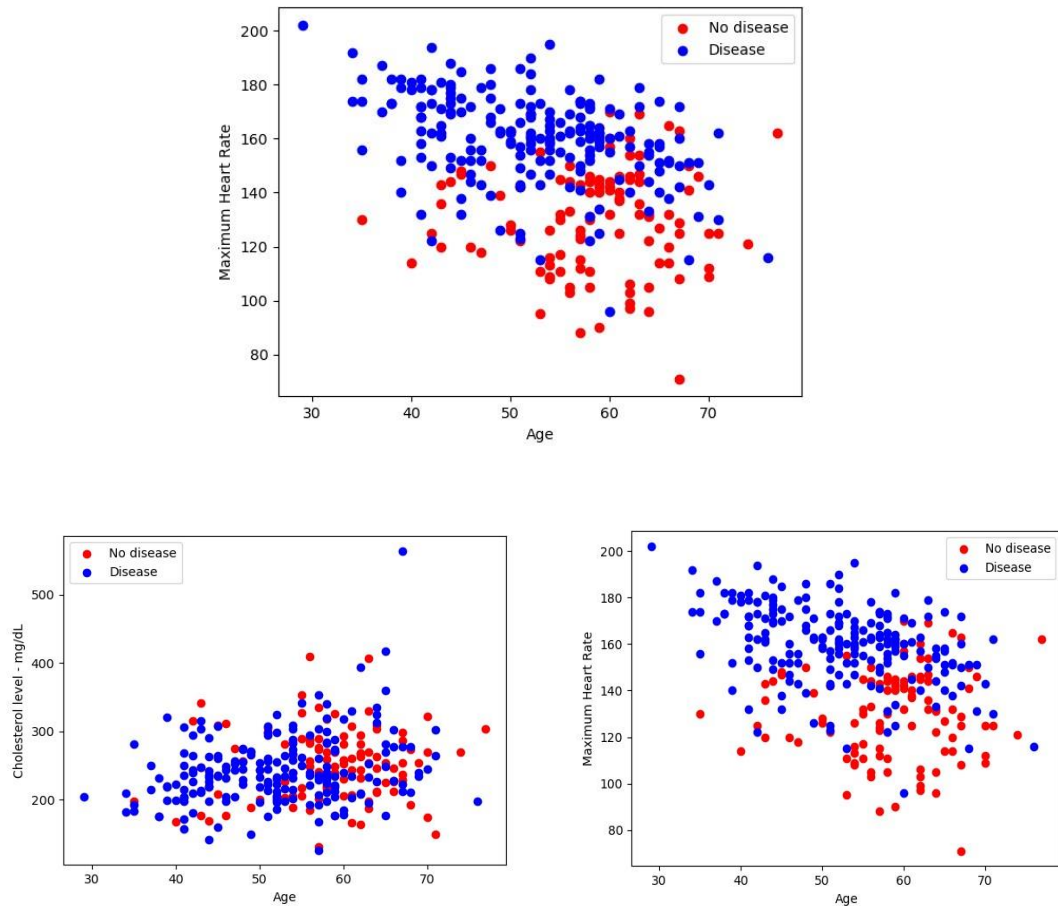
The software can determine the risk of heart disease given patient data with an accuracy of 85.48%. Medical professionals can use this tool for decision making when assessing a patient's heart health. Additionally, the software provides three scatter plots, allowing for past data to be visualized for determining trends in heart disease risk.

Data Cleaning

The data contained within the provided dataset did not require cleaning, and all records were complete and free of null values.

Data Visualization

Three scatter plots are provided within the software for medical professionals to review and determine trends in heart disease risk. A sample is provided below.



Real-Time Queries

The response time of the application as the user interacts with patient health parameters is near instantaneous. This allows for seemingly instant results as medical professionals insert the patient data.

Adaptive Element

Using logistic regression, the application can take patient data and determining the presence or absence of heart disease based upon the inputted data. The accuracy of the logistic regression model is displayed on the GUI and is 85.48%.

Outcome Accuracy

The accuracy of the logistic regression model is generated every time the software is run and is displayed on the GUI for users to see. The confidence percentage, under prediction probability rating, is displayed for users to see the confidence of the logistic regression model's prediction. Additionally, users can see the patient data which has been inputted from parameters located on the sidebar of the application in order to confirm they have inputted the data correctly. A partial screenshot is provided below to demonstrate the accuracy features of the software.

Patient Data

	Age	Sex	Chest Pain	Resting BP	Cholesterol	FBS	Rest ECG	Thalach	Ex
0	18	1	3	100	524	1	2	201	

Heart Disease Prediction

Predicted Heart Disease From Patient Data

Prediction Probability Rating

0 : No Heart Disease

1 : Heart Disease

	0	1
0	0.0284	0.9716

Logistic Regression Model Accuracy Rating

The model has an accuracy of 85.48%

Security Measures

The software features a login screen and requests a password upon starting the application. A user is not able to access any of the features, such as the scatter plots and heart disease prediction tool, without inputting the correct password. The current password of the software is “nightowl”.

Product Health Monitoring

The health of the application is monitored every time a prediction is made by the logistic regression model. The accuracy of the logistic regression model is displayed on the GUI and is also outputted to the command line interface used to initially run the software. This allows users to have confidence in the health and accuracy of the application.

Dashboard

A screenshot is provided below of the user-friendly dashboard of the software solution.

Please select data for heart disease prediction

Age
18 100

Sex
☒ Male
☐ Female

Chest Pain
☒ Typical
☐ Asymptomatic
☐ Nonanginal
☐ Nontypical

Resting Blood Pressure
80 200

Cholesterol
120 500

Fasting Blood Sugar Above 120 m/dl
☒ Yes
☐ No

Resting ECG Results
☒ Normal
☐ Level 1
☐ Level 2

Maximum heart rate
70 210

Exercise Induced Angina
☒ Yes

Heart Disease Prediction Application



Patient Data

Age	Sex	Chest Pain	Resting BP	Cholesterol	FBS	Rest. ECG	Thalach	Ex
18	1	3	120	200	1	0	104	

Heart Disease Prediction

Predicted Heart Disease From Patient Data

Prediction Probability Rating

Section D

Project Purpose

The purpose of this project was to provide a software solution that would increase the speed of medical professionals in making heart disease predictions in patients. The client was assured the software solution would have a prediction accuracy of at least 80%. This requirement was satisfied by the K-means clustering accuracy of 81.8% and the logistic regression model's accuracy of 85.48%. The software was successful in reducing the workload of medical professionals at Greathealth Hospital, as well as acting as a virtual second opinion when physicians were unsure on the heart health of a patient. The software solution proved easy-to-use, with a graphical user interface, and was accepted and installed on Greathealth Hospital's Cardiac Wing Computer.

Datasets

A copy of the dataset used is included within the software solution and can be downloaded at the following link. <https://www.kaggle.com/volodymyrgavrysh/heart-disease>. The dataset did not require any cleaning and contained no null values.

Data Product Code

The code of the data product has been provided to Greathealth Hospital. Within the folder "Heartproject", "KMeansPlotting.py" was used to generate the KMeans clustering based scatter plots. Using Pandas, Sklearn, numpy, and Matplotlib, the raw CSV data from the provided dataset was able to be immediately used for scaling and clustering.

The main functionality of the program is provided within the file “main.py”. This file uses Streamlit, Pandas, and Sklearn, and PIL. Using these libraries, the raw CSV data was able to be used for logistic regression modeling. PIL was used to handle displaying images, such as the main logo, and scatter plots. Streamlit was used in the creation of the GUI of the software. Pandas was used in handling the CSV data and dataframes. Lastly, Sklearn was used in the creation of the logistic regression model.

Hypothesis Verification

It was hypothesized that the workflow of medical professionals at Greathealth Hospital would improve and the workload would decrease with the creation of a heart disease prediction tool. In turn, it was hypothesized that medical professionals would be able to spend more time giving patients more personalized healthcare, and thus increasing patient satisfaction ratings of Greathealth Hospital.

Since the implementation of the data product, medical professionals at Greathealth Hospital have reported 80% less time spent predicting heart disease, and on average, spend two hours less per day on calculations. Patients have also reported higher satisfaction ratings with Greathealth Hospital since the introduction of the data product. Additionally, physicians have been able to use the software as a virtual second opinion, which has increased their confidence in predicting heart disease from an internal survey at Greathealth Hospital. Given the outcome of this product, the hypothesis has been verified.

Effective Visualization and Reporting

The produced data product effectively visualized heart disease prediction based upon new data. The easy-to-use interface provided users with sliders and buttons for inputting patient data, a table to provide users with the ability to double check that they have input their data correctly, and prediction results in real-time based upon the inputted data. A section of the GUI, in plain text, provided the users with a clear “yes or no” in determining the heart health of a patient. Additionally, a statistical section was provided, including percentages of confidence in the prediction.

Scatter plots based upon previously obtained datapoints were generated and allowed the user to clearly view correlations between certain risk factors. This allowed medical professionals to get a good look at the data from a first glance, without needing to focus on raw statistical data. As an interesting note, the maximum heart rate of a patient was shown to be indicative of the presence of heart disease in patients, regardless of age. The ability to visualize data allows users to see trends that they may not have noticed from raw data alone.

Accuracy Analysis

The goal of the project was to create a data product with an accuracy of 80% or higher. The logistic regression model, as well as the K-means clustering were both found to be above 80% in accuracy. Greathealth Hospital was pleased to discover the logistic regression model, which is used in predicting heart disease from inputted data, had an accuracy of 85.48%.

Application Testing

The application was subjected to unit testing to ensure that both the logistic regression model, and the K-Means clustering were performing as intended. Integration testing was performed after the units were combined to ensure functionality when put together. System testing was performed once the entire software solution was produced. Testing was sufficient and ensured the product would find most major bugs prior to launch. Acceptance testing was performed with users from Greathealth Hospital, and upon their approval, the application was cleared for deployment.

Application Files

The following application files are found within the “Heartproject” folder, which has been provided to Greathealth Hospital.

AgeVSHeartRate.png – This is a PNG image of the scatter plot of age vs heart rate.

CholesterolAgeScatter.png – This is a PNG image of the scatter plot of cholesterol vs age.

Heart.csv – This is the CSV file containing the data obtained from

<https://www.kaggle.com/volodymyrgavrysh/heart-disease>

Heartpredictiontoolimage.png – This is the application’s main logo in PNG format.

KMeansPlotting.py – This is a python file which was used to generate **AgeVSHeartRate.png**,

CholesterolAgeScatter.png, and **RestBPvsThalach.png**. This file is dependent on **Heart.csv**.

Main.py – This is a python file which contains the main application. This file is dependent on

Heart.csv, **AgeVSHeartRate.png**, **CholesterolAgeScatter.png**,

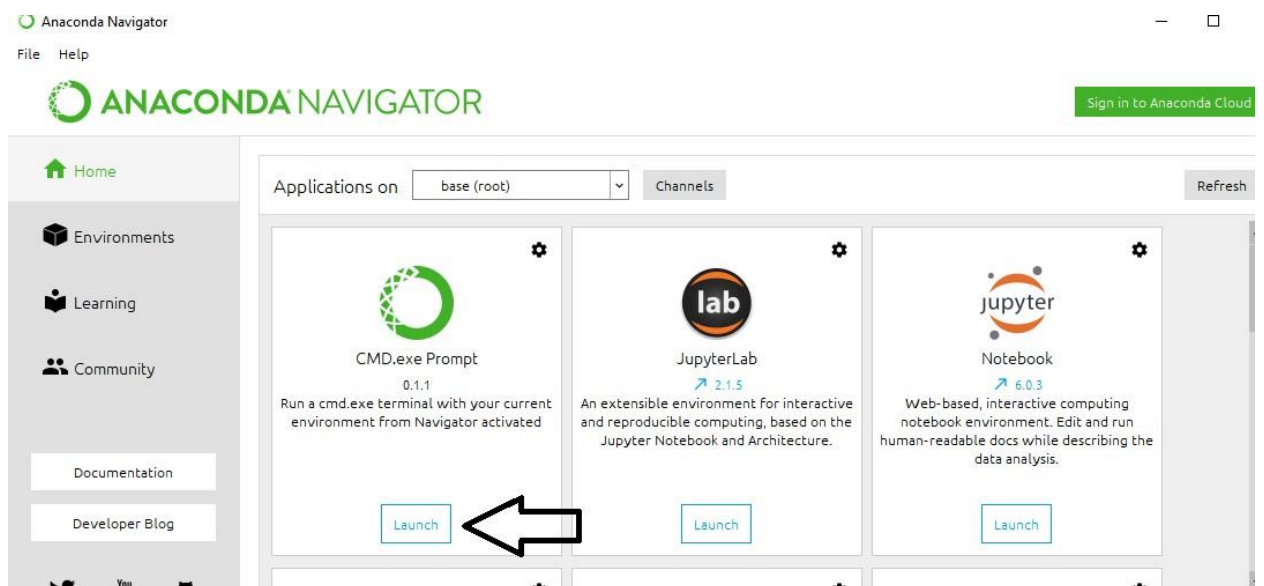
Heartpredictiontoolimage.png, and **RestBPvsThalach.png**

RestBPvsThalach.png – This is a PNG image of the scatter plot of resting blood pressure vs maximum heart rate.

User's Guide

Installation and User's Guide

1. Download Anaconda3 from the following website
<https://www.anaconda.com/products/individual>
2. Upon completing the installation of Anaconda3, open “Anaconda Navigator” and navigator to the “Home” tab
3. Click on the “Launch” button, located under “CMD.exe Prompt” as shown in the picture below.



4. Type in the command prompt “pip install streamlit” without quotations, then, press the enter key.



A screenshot of a Windows command prompt window. The title bar reads "Select C:\Windows\system32\cmd.exe". The window content shows the following text: "Microsoft Windows [Version 10.0.19041.572] (c) 2020 Microsoft Corporation. All rights reserved. (base) C:\Users\ [redacted] > pip install streamlit_". A large white arrow points upwards towards the command prompt.

5. Upon successful installation of Streamlit, ensure the folder titled “Heartproject” downloaded. C:\Users\Test\PycharmProjects\Heartproject\main.py is used within this example, however yours will be unique to your machine. In order to determine the address, a screenshot is provided.




Copy, or write down the address in which your “Heartproject” is located and add “\main.py” to the end of the address, without quotations.

6. Within the command prompt, type in “streamlit run your-unique-address” without quotation marks, and replacing “your-unique-address” with the unique address obtained from step 5. After you have typed this in, hit the enter key.
7. Upon hitting enter, your default web browser should pop up containing the application. If in the case the web browser does not automatically pop up, copy the local link provided within the command prompt and navigate to it using your desired web browser.

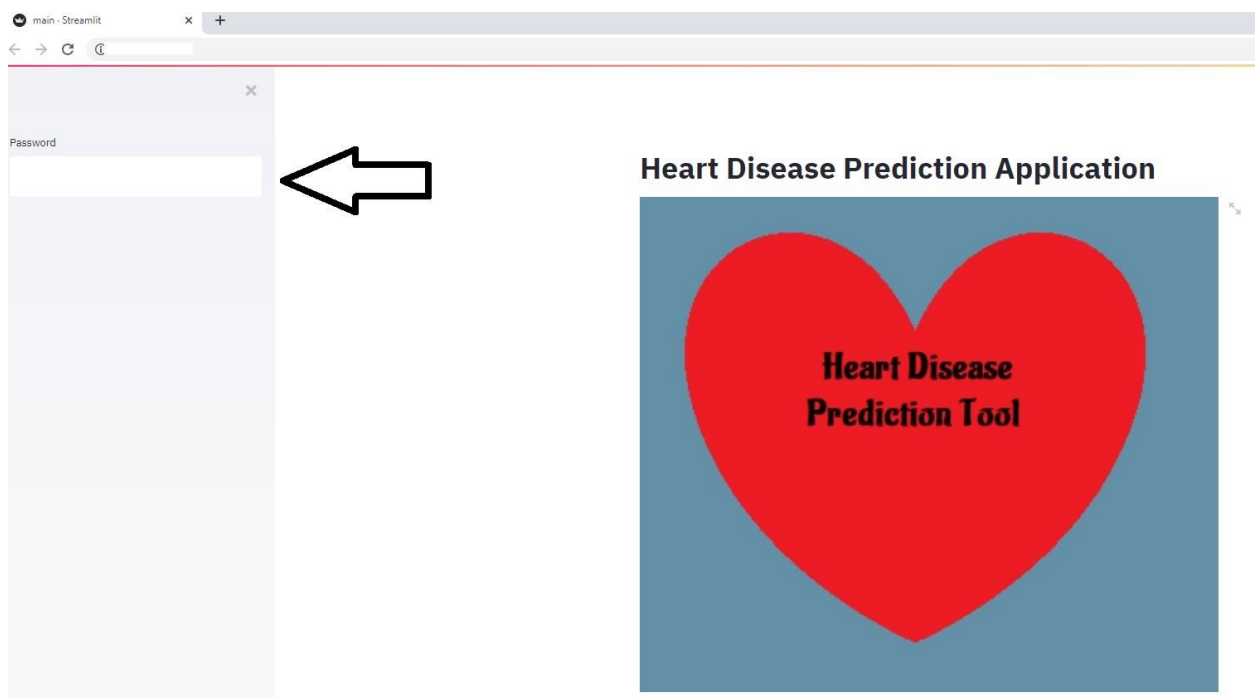
```
Microsoft Windows [Version 10.0.19041.572]
(c) 2020 Microsoft Corporation. All rights reserved.

(base) C:\Users\██████>streamlit run c:\users\██████\pycharmprojects\heartproject\main.py

You can now view your Streamlit app in your browser.
Local URL: http://██████
Network URL: http://██████
```



8. In the password box, type in “nightowl” without quotation marks and press enter.



9. You are now logged in to the application! Use the sidebar to input patient data (1), and the “Patient Data” (2) table to ensure you have input the correct patient data.

The screenshot displays the 'Heart Disease Prediction Application' interface. On the left is a sidebar titled 'Please select data for heart disease prediction'. It contains several input fields: a password field with 'nightowl' entered, an age slider set to 18, a sex selection with 'Male' chosen, a chest pain selection with 'Typical' chosen, a resting blood pressure slider set to 120, a cholesterol slider set to 200, a fasting blood sugar selection with 'Yes' chosen, and a resting ECG results section. A large white arrow labeled '1' points from the sidebar to the main application area. The main area features a large red heart graphic with the text 'Heart Disease Prediction Tool' inside. Below the heart is a table titled 'Patient Data' with the following data:

	Age	Sex	Chest Pain	Resting BP	Cholesterol	FBS	Rest ECG	Thalach	Ex
0	18	1	3	120	200	1	0	100	

A large white arrow labeled '2' points from the 'Patient Data' table back to the sidebar.

10. Scroll down on the application to view additional features. Heart Disease Prediction (1) will predict the presence or absence of heart disease based upon the inputted data. Prediction Probability Rating (2) will display a percentage in decimal format of the presence or absence of heart disease. Logistic Regression Model Accuracy (3) displays the model's accuracy rating.

Heart Disease Prediction

Predicted Heart Disease From Patient Data

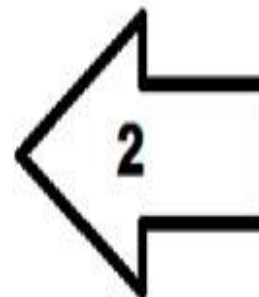


Prediction Probability Rating

0 : No Heart Disease

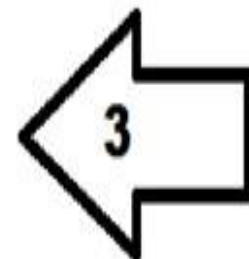
1 : Heart Disease

	0	1
0	0.3883	0.6117



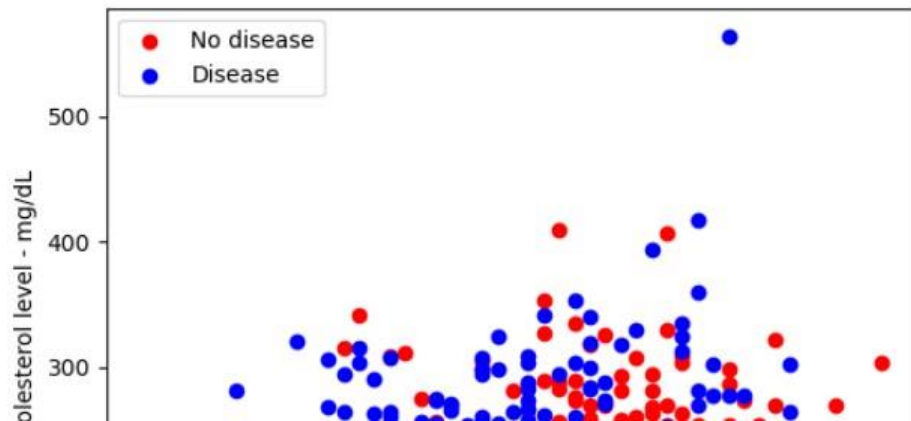
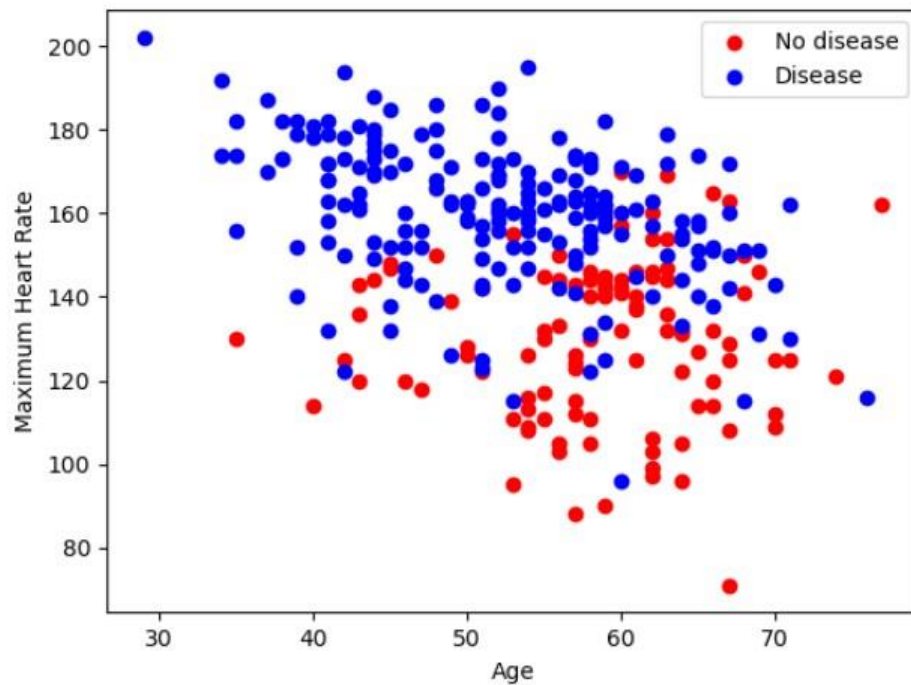
Logistic Regression Model Accuracy Rating

The model has an accuracy of 85.48%



11. Scroll down on the application to view the three scatter plots that are provided for reviewing data trends.

Scatter Plots From Previous Data Using K Means Clustering



12. Lastly, this tool is not intended to replace the advice of a medical professional. Please use the tool responsibly.

Summation of Learning Experience

Prior to enrolling as a student with Western Governors University I had a minuscule amount of programming experience. After completing coursework which gave me the chance to develop hands on skills with Java, C++, and in this case Python, it gave me both the confidence and experience I needed to start this project.

I reached out to my course instructor, Dr. Charlie Paddock, who worked with me and guided me in the right direction from the start and into the development of this project. I realized I still had a lot to learn if I wanted to complete this project, such as learning how to use “Streamlit”. Through a combination of reading manuals and watching videos, I was able to develop the data product.

The completion of this project was possible due to the skills I had gained and honed throughout my time as an undergraduate student at Western Governors University. My mentor, Professor Mary Rousseau has been incredibly supportive on my academic journey, and through their mentorship I feel I have grown both academically and personally.

According to the CDC, heart disease is the leading cause of death for most people within the United States (Heart Disease Facts, 2020). This project allowed me to see just how computer science, and machine learning, can be used to benefit humanity. This contributed to my concept of life-long learning, showing me just how versatile and helpful the skills I have gained as an undergraduate student can be. I hope to continue my academic journey and learn new skills that can one day be used to help humanity.

Section E

Sources

Michael L. Barnett, M. (2019, March 01). Diagnostic Accuracy of Collective Intelligence of Multiple Physicians vs Individual Physicians. Retrieved November 10, 2020, from <https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2726709>

Heart Disease Facts. (2020, September 08). Retrieved November 10, 2020, from <https://www.cdc.gov/heartdisease/facts.htm>