Capstone: Predicting Mortality by Heart Disease

Conner Murowchick

Table of contents

**Prompt A**

Letter of Transmittal 4

Project Recommendation 6

Problem Summary 6

Application Benefits 6

Application Description 7

Data Description 7

Objective and Hypothesis 7

Methodology 8

Funding Requirements 9

Stakeholders Impact 9

Data Precautions 9

Developer Expertise 9

**Prompt B**

Project Proposal 10

Problem Statement 10

Customer Summary 10

Existing System Analysis 11

Data 11

Project Methodology 11

Project Outcomes 12

Implementation Plan 13

Evaluation Plan 13

Resources and Costs 14

Timeline and Milestones 14

**Prompt C**

Application Files 16

**Prompt D**

Post-implementation Report 17

Project Purpose 17

Datasets 17

Data Product Code 18

Hypothesis Verification 19

Effective visualizations and reporting 20

Accuracy analysis 20

Application Testing 21

**Appendices**

Installation Guide 22

User Guide 22

Summation of Learning Experience 25

References 26

**Prompt A**

**LETTER OF TRANSMITTAL**

March 4th, 2022

John Workman, CEO

Healthcare Solutions Inc.

1555 Developer Road

Seattle, Washington

Dear Mr. Workman:  
Heart disease is the leading cause of death in the United States. As such a means of predicting whether an individual is expected to die of heart disease quickly and accurately would be exceedingly useful to health care professionals. With this in mind, I recommend a software solution that would take in client information, process it, and then make an accurate prediction as to if they are expected to die of heart disease.

Having this software solution would allow your company to sell this product to hospitals across the U.S. This product will provide those hospitals with a tool substantially faster than any human reviewer at determining the likelihood of mortality by heart disease for a given individual. This would allow hospitals the ability to rapidly assess the risk of heart disease-related death for a patient and set up treatment and prevention plans immediately.

The objective of this project is to create a user-friendly program that will take in an individual’s information, and intuitively display whether or not they are predicted to die of heart disease. The total funding requirement for this project is $25,000. This will be primarily for the payment of the developer, as the software and developing environment are open-source and free. The proposed developer for this solution is me, Conner Murowchick. I have extensive experience in developing projects of this machine learning nature and have been successful in creating successful software solutions for this company in the past. Please let me know if you have any concerns.

Sincerely,

Conner Murowchick

**Project Recommendation**

**Problem Summary**

Heart disease is the leading cause of death in the United States. This being the case, hospitals need a way of quickly and accurately predicting if an individual is expected to die of heart disease or not. The solution to this problem is an easy-to-use product that will quickly and accurately predict if an individual is expected to die of heart disease. This project will include a user-friendly dashboard with several visuals describing the data and the result that the neural network predicts. A neural network is a series of algorithms that seeks to recognize the underlying patterns and relationships in a set of data in a way that mimics the way the human brain operates. A neural network will be the backbone of how this project operates.

**Application Benefits**

This software solution will give hospitals a tool that will be substantially faster than any human reviewer at determining the likelihood of mortality by heart disease for a given individual. This project will also provide health care staff with a second opinion that will be invaluable in providing patient care. With the global pandemic still a relevant issue putting pressure on the health care system, this software solution would take some of the pressure off of healthcare workers by giving them a second opinion or a reassurance of their own belief when determining the state of an individual’s heart health.

**Application Description**

This software project will be comprised of an interactable dashboard that will display information about the data being used. The dashboard will show descriptive data indicating important properties of the data as well as predictive data which will be the output of the neural network that will predict if a patient is expected to die of heart disease. This dashboard will include a form for users to input unique data and receive a prediction from then neural network regarding death expectancy. This application lines up with the business priorities of both Healthcare Solutions Inc. and the hospitals it will eventually be sold to by Healthcare Solutions Inc. because it provides a quick and accurate forecast of a client’s heart disease health, as well as back up to a doctor’s claim, which is important for any hospital.

**Data Description**

The data that will be used is from Kaggle.com and was curated by university researchers. The data for this application is a CSV file containing 12 attributes of information that the neural network will use to find underlying patterns in the data. The data that will be used to train and test the neural network will need to be split up into training and testing sets for the neural network. The data will also need to be scaled with a data scaler for the neural network to process it correctly and deal with outliers and anomalies.

**Objective and Hypothesis**

The objective of this project and application are to provide a valuable tool that can be sold to hospitals that will be able to predict if patients are expected to die of heart disease. If I create this machine learning-based program, then I will be able to address the organizational need of identifying patients at risk of heart disease with over 70% accuracy.

**Methodology**

I will be using a CRISP-DM methodology to implement this project. CRISP-DM stands for Cross Industry Standard Process for Data Mining. The six steps are business understanding, data understanding, data preparation, modeling, evaluation, and deployment. Starting with business understanding, the business (hospitals), needs an easy-to-use machine learning-based program that can detect with a high level of accuracy if patients are likely to die of heart disease. Next, with data understanding, the data has 304 data entries with 14 attributes in each entry and the data is clean. The data is currently in a CSV file, so for data preparation, I will import that file into my python project and use pandas (a data manipulation software) to manipulate and organize it. For modeling, I will be using a neural network. I will split the data into a training set (80% of the data) to train the neural network, and a testing set (20% of the data) to test the results. For evaluation, I will check what percent of the testing set the neural network correctly guesses to evaluate the program. I will also check that the dashboard correctly displays the data visualizations. For deployment, I will develop and document a plan for deploying the model, that will be used throughout development as a reference for progress. Then I will conduct a project retrospective to see what was done well and what could be done better. This methodology is a good fit as it ties in well with data manipulation and deep learning.

**Funding Requirements**

The funding requirements for this project will be $25,000 for the payment of the developer. The software and developing environment are open-source and free so those will not be a factor in the funding requirement.

**Stakeholders Impact**

The key stakeholders in this project are my boss Steve Workman, as well as the CEO of Healthcare Solutions Inc. John Workman. The completion of this project will cement our company as a cornerstone of cutting-edge technology in the hospital industry and will enable this company to be the first thing that comes to mind when one requires hospital-related tech.

**Data Precautions**

This data is from real patients, but no sensitive data is present. This is because all identifying information has been removed from the data set, so there is no risk of violating HIPAA. This is also a public dataset from Kaggle.com, which ensures that there is no sensitive data present.

**Developer Expertise**

The software developer who will be constructing this program is, Conner Murowchick. I have done similar projects like this one in the past for the company, and they have all ended successfully. I have 5 years of industry experience working with neural networks and their applications, so I am qualified to take on this project.

**Prompt B**

**Project Proposal**

**Problem Statement**

Heart disease is the leading cause of death in the United States. Because of this, hospitals need a way of quickly and accurately predicting if an individual is expected to die of heart disease or not. Hospitals will need accurate analysis and predictive representation of that data if they are going to be able to act on it. Doctors and other healthcare staff would benefit greatly from having a second opinion, or simply a reassurance that they are making the right choice regarding patient care.

**Customer Summary**

The customer of this project will be the hospitals, and ultimately the hospital staff that will be using this software. My proposed solution to this problem is to use a machine learning-based neural network to provide accurate analysis and a dashboard to provide the predictive representation and data visualization that these hospitals will need. This application will be used by doctors and other medical staff on their hospital desktops, the software will be easy to use and intuitive and only require basic computer skills to operate.

**Existing System Analysis**

Currently, the customer for Healthcare Solutions, All American Hospital, uses clinicians’ judgment to determine if patients are expected to die of heart disease. The doctor at hand reviews the patient data and comes to a conclusion based on their experience.

After the project, the software will be installed on the All-American Hospital internal network. Python.3.9.10 will be installed with the following libraries: pandas, TensorFlow, sklearn, NumPy, beautifultable, and matplotlib.

The final, intended result of this project is to provide a quick and accurate second opinion for doctors looking to diagnose patients. If successful, this project will improve the outcomes of all involved stakeholders.

**Data**

The data chosen to train and validate this neural network application is a CSV file containing 304 data entries with 12 attributes for each entry. As the data is already in CSV file format, the data is clean and will need further processing to use. To deal with data anomalies and outliers, the data will be scaled using a data scaler known as “PowerTransformer” that will be using the “Yeo-Johnson” method. This will optimally group the data for the neural network to process it accurately.

**Project Methodology**

I will be using a CRISP-DM methodology to implement this project. CRISP-DM stands for Cross Industry Standard Process for Data Mining. The six steps are business understanding, data understanding, data preparation, modeling, evaluation, and deployment. Starting with the business understanding, the business (hospitals), needs an easy-to-use machine learning-based program that will detect with a high level of accuracy if patients are likely to die of heart disease. Next, with data understanding, the data has 304 data entries with 14 attributes in each entry and the data is clean. The data is currently in a CSV file, so for data preparation, I will import it into Jupyter Notebooks and use pandas to manipulate and organize it. For modeling, I will be using a neural network. I will split the data into a training set (80% of the data) to train the neural network, and a testing set (20% of the data) to test the results. For evaluation, I will check what percent of the testing set the neural network correctly guessed after being trained to evaluate the program. I will also check that the dashboard correctly displays the data visualizations. For deployment, I will develop and document a plan for deploying the model, that will be used throughout development as a reference for progress. I will start with setting up the neural network to work with the data, then I will work on making the dashboard functional and visually pleasing. Lastly, I will conduct a project retrospective to see what was done well and what could be done better.

**Project Outcomes**

The project deliverables associated with the design of this application will be a schedule, plans for testing, and mock-ups of the dashboard. The product deliverables associated with the development of the application will be a functional dashboard, maintenance tools, and useful documentation. For documentation purposes, a user guide and an installation guide will be provided, which will include instructions on installing and operating the interface of the project. Additionally, a schedule with projected and actual milestone completion dates will be created.

**Implementation Plan**

This project will be implemented with the previously discussed CRISP-DM methodology. I will start with business understanding, data understanding, and data preparation, which will give me a clear goal and usable data. Then I will work on modeling, which will produce a functional neural network that will be able to take any new data with the same 11 attributes for any patient, and predict if they are expected to die of heart disease. Next up will be, an evaluation where I will check what percent of the testing set the neural network currently guessed after being trained by the program. I will also check to see if the dashboard is functional and accurate in this phase. Lastly, in the deployment phase, I will thoroughly document the code.

**Evaluation Plan**

For evaluating the success of this project during development, each section of the code will be checked for meeting the key performance indicators required of it before moving on. For the neural network, this will be running in a reasonable amount of time and achieving an accuracy of at least 70%. For the form that will take in unique user data, this will be achieving the same accuracy as the neural network and validate the user input for each attribute. For the data visualizations, success will be defined by accurate and clean representations of the data.

For evaluating the success of this project upon completion, I will check to see if the neural network can predict with an accuracy of higher than 70% if a person is predicted to die of heart disease. I will also check to see if the data visualizations in the dashboard are all functional, and the dashboard is easy to use. If the accuracy is lower than 70%, for the sake of quality assurance, I will rework the neural network until it has reached this industry standard.

**Resources and Costs**

For a programming environment, I will be using Jupyter Notebooks (as a compiler and coding environment), NumPy (as a third-party library), and pandas (for data manipulation). All of this software is open-source and free. There are no additional environmental costs associated with this application. There will be no startup, first-time, or licensing costs. The projected time and cost for labor to complete the application is 128 hours and 25,000 dollars.

**Timeline and Milestones**

Below is a table representing the timeline and milestones of this project:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Mile  stone | Pre-  requisites | Activity | Resource Assigned | Hours | Start | End |
| 1 | - | Requirements Approval | Project Manager | 8 | 3/1/22 | 3/2/22 |
| 2 | 1 | Architecture Design | Software Engineer | 8 | 3/3/22 | 3/4/22 |
| 3 | 1 | Development Environment  Configuration | Software Engineer | 8 | 3/5/22 | 3/6/22 |
| 4 | 1 | Data curation | Software Engineer | 8 | 3/7/22 | 3/8/22 |
| 5 | 2 | Interface Mockup | Software Engineer | 8 | 3/9/22 | 3/10/22 |
| 6 | 2,3,4 | Neural Network Development | Software Engineer | 40 | 3/11/22 | 3/20/22 |
| 7 | 1 | Jupyter Notebooks Markdown formatting | Software Engineer | 8 | 3/21/22 | 3/21/22 |
| 8 | 4 | Descriptive Data Table | Software Engineer | 8 | 3/22/22 | 3/23/22 |
| 9 | 4 | Data Distribution Histograms | Software Engineer | 8 | 3/24/22 | 3/25/22 |
| 10 | 2,3,4,6 | Neural Network Testing Form | Software Engineer | 8 | 3/26/22 | 3/27/22 |
| 11 | 10 | Documentation | Software Engineer | 8 | 3/28/22 | 3/29/22 |
| 12 | 11 | Final Project Delivery | Software Engineer | 8 | 3/30/22 | 3/31/22 |

**Prompt C**

**Application Files**

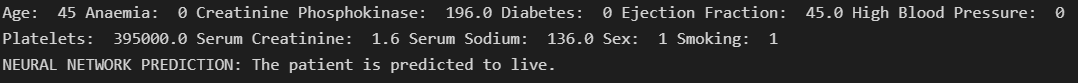
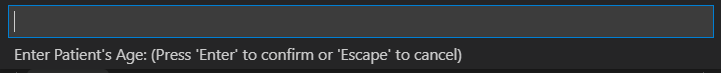
\heartdisease\_task.ipynb Jupyter notebook containing all code and markdown describing the code for the project.

\heart\_failure\_clinical\_records\_dataset.csv The CSV file containing the data to be used by the neural network.

**Prompt D**

**Post-Implementation Report**

**Project purpose**

This project will provide decision-making support to hospital staff by using a database of information relating to past patients to predict if a current patient will die of heart disease. By comparing which patient information attributes are the most important in determining death by heart disease, the application uses these as a basis for predicting on unseen data whether or not death by heart disease is expected. By giving hospital staff a second opinion or reassurance of their thoughts, these hospital staff can be reassured that they are doing their utmost to care for the patients. This can be accomplished by filling out the form in the application with the unique data for the client they are treating, to get a prediction on whether or not they are expected to die of heart disease. 

**Datasets**

The data used in this task was scrubbed for sensitive information before its use. Thus, the data was used freely without massive security concerns. The format of the data was a CSV file of 12 attributes of patient data with 298 rows. Here is an example of the unprocessed data in CSV file format for one patient. 

Access to the use of this data set will be provided in the form of the CSV file that will be included with this submission.

**Data product code**

For processing the data from the CSV file, I used pandas to read the CSV file, then split the data between two datasets, X and Y, which were the attributes and the expected result respectively. I then used a function from sklearn called ‘train\_test\_split’ that split the data into training and tests sets. Next, I found I had to scale the data, because one of the attributes, platelets, was consistently over a hundred thousand while the rest of the data was below or around 100, and this was wildly throwing off the accuracy of the neural network. I tried several data scalers, StandardScaler, RobustScaler, MinMaxScaler, QuantileTransformer, and found the PowerTransformer with the method ‘Yeo-Johnson’ yielded the best results. Before scaling the data with PowerTransformer the neural network accuracy was around 55- 60% which isn’t much better than guessing randomly. But after scaling the data the accuracy was risen to consistently between 70 and 80%, which is a much better result.

For the descriptive data table, I calculated the mean, standard deviation, min, max, and percentiles of the data separately as this was simpler, then simply put them in a table format using beautifultable and made the formatting look nice by adjusting some values. For the data distribution histograms, I simply split the data in the CSV file into 11 different pandas’ data frames and made a distribution histogram with the data from each of them.

The neural network uses a TensorFlow Keras Sequential model. The input layer has 11 neurons, there are three hidden layers with 320, 448, 384 neurons respectively, and the output layer has 1 neuron. The activation functions for the three hidden layers are ReLu as this was the most effective when testing and the output layers activation function is sigmoid because sigmoid always returns a value between 0 and 1, and that is the desired output. This number of layers may seem deranged and random at first glance, but the number of neurons and layers was intentional. I used a tool called KerasTuner, which runs the neural network with many different layers, neurons, and hyperparameter configurations, and the configuration I am using was the best result returned by KerasTuner. A neural network was chosen as the analytic method for this project because binary classification via a neural network fits the organizational need perfectly. This neural network was trained on 70% of the data and tested on the remaining 30% for accuracy. Given the data analysis method of a neural network, a data distribution histogram, as well as a descriptive data table proved useful in understanding how the neural network recognized the patterns involved when predicting whether or not a patient is predicted to die of heart disease.

**Hypothesis verification**

The design of this application was built around the hypothesis that if I create this machine learning program, then I will be able to address the organizational need of predicting patient death by heart disease with over 70% accuracy. Using this dataset, the hypothesis was confirmed. The neural network was able to correctly predict patient death with over 70% accuracy when shown new data. Improvements to the hyperparameters of the neural networks could show significant improvements to this accuracy if desired.

**Effective Visualizations and Reporting**

To help support decision-making for doctors, the visualization of primary importance is the form that can take in unique data and make a prediction. From the Jupyter Notebook, the user can quickly fill out a form with a patient’s information and get a prediction from the neural network to aid in their work.

From the same Jupyter Notebook histograms of distributions of each attribute of the data, as well as a descriptive table are available, which can be used as a reference to expected levels for a patient’s information. The form, the histograms, and the descriptive table are all clearly labeled and simple to navigate.

**Accuracy Analysis**

For the accuracy metric of this application, it was determined that success would be measured on whether or not the neural network was capable of achieving over 70% accuracy in identifying on unseen data if a patient was expected to die of heart disease.





While the accuracy of the neural network changes each time it’s run, as is the nature of stochastic gradient descent learning, the accuracy is consistently above 70%. Based upon this predetermined metric of success, this application succeeded in performing at acceptable levels.

**Application Testing**

Each module of this project was thoroughly tested for expected output before moving on to the next module. For the neural network code, this involved running numerous tests with varying hyperparameters such as learning rate, as well as with differently shaped neural network architectures with varying layers and numbers of neurons in those layers. Adjusting these bells and whistles allowed for an increase in accuracy and computation time for the neural network. For the form for inputting unique data, this included ensuring that input was validated, and not allowing the user to input data of the wrong type. Also, for the form, testing included ensuring that the neural network was loaded and used correctly and as expected. For the data visualizations, testing involved verifying that the output was as expected, and presented neatly.

**Appendices**

**Installation Guide**

Prerequisites:

Visual Studio Code

Jupyter Notebooks

Python 3.9.10 with supporting libraries: pandas, TensorFlow, sklearn, beautifultable, and matplotlib.

1) Install the prerequisite applications.

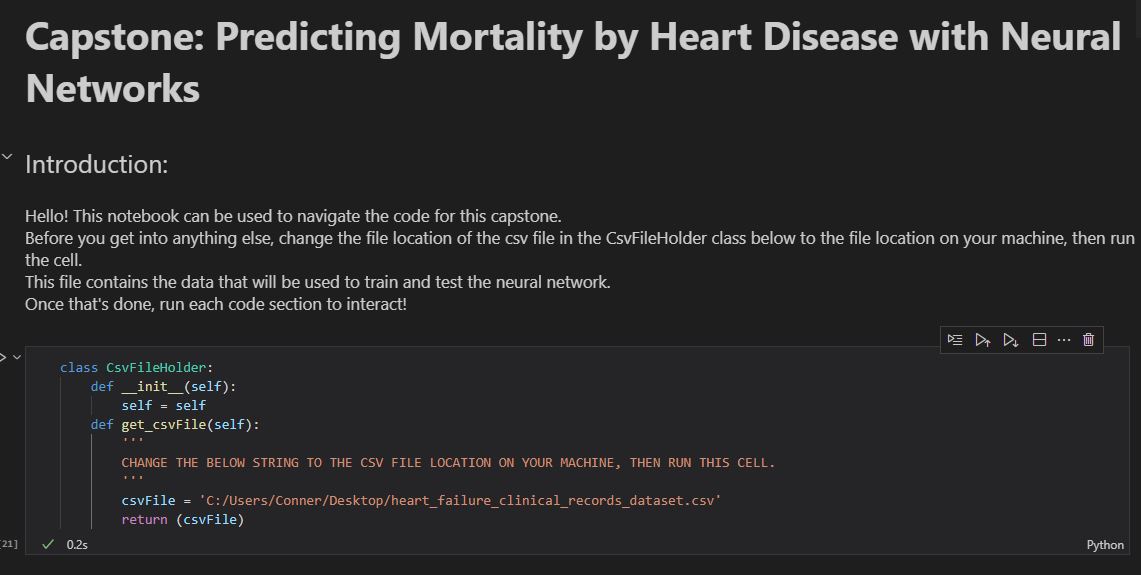
2) Download the heartdisease\_task.ipynb file and the heart\_failure\_clinical\_records\_dataset.csv file.

3) Open the heartdisease\_task.ipynb file with Visual Studio Code

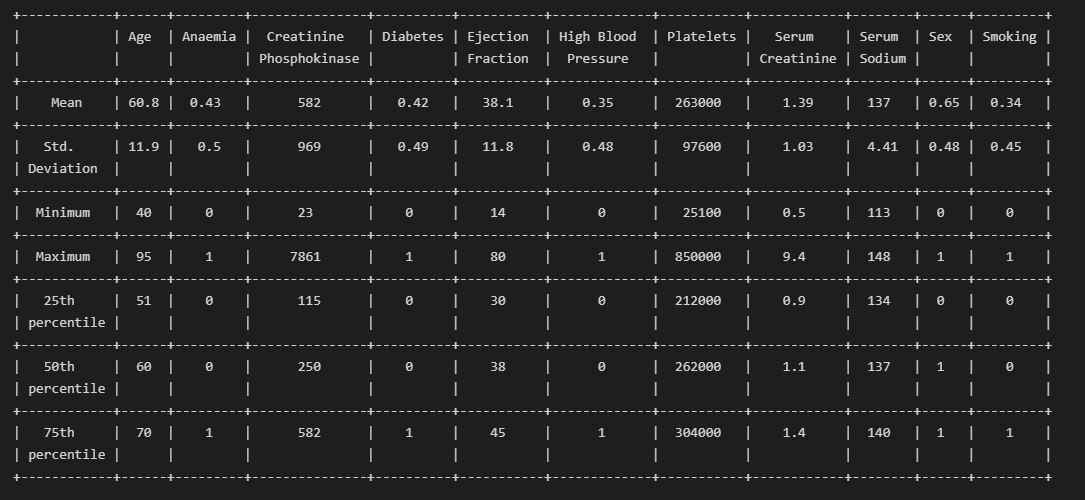
4) Follow the directions in the introduction of the Jupyter Notebook file to correctly configure the CSV file.

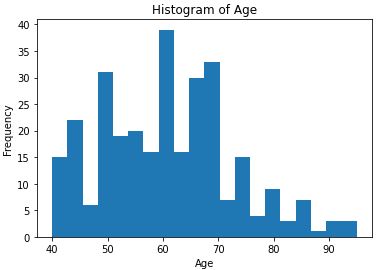
**User Guide**

To start predicting with this neural network, first, follow the instructions in the introduction, and change the “csvFile” variable in the “CsvFileHolder” class to the CSV file location on your machine, then run the cell. This class is located at the top of the notebook, right beneath the introduction section.

****

Once this is done, run the neural network code cell, to save a model of the neural network and get the accuracy with which the neural network can correctly predict if a patient is expected to die of heart disease.  Then, to test the neural network with unique data, run the next cell, and input the unique data in the prompt box at the top of the application. There are several appropriate sample data pieces provided if none are on hand. 

This will return the neural network’s prediction based on your unique data entry. Also included in this Jupyter Notebook is a table of the patient data that will train the neural network, 

As well as distribution histograms for all attributes of the data, that can be accessed by running their cells. 

**Summation of Learning Experience**

When working on this assignment, I was confronted with my issues with settling on a project to commit to and complete, and my troubles with vague directions. My educational experiences up to this point and a history of heart disease in my family encouraged me to choose a task that would potentially help hospital staff to make decisions regarding patient care. Learning project management skills allowed to me follow a project methodology well and avoid common pitfalls. Learning how to build multiple projects from scratch, gave me the experience and the confidence to complete this project. When researching the ins and outs of what it means to build a binary classification neural network, I came across many articles and books that I used to better understand the nature of the problem I faced. Through, learning about neural networks and completing this project, I found there is so much more out there for me to learn, as the complexity of machine learning is astounding. By working on this project, I realized that machine learning is a fascinating field and I plan to pursue a career in that direction if possible.

**References**

This space was intentionally left blank