PatientPal: The Interactive Heart Disease Prediction Tool

Ryan Stopczynski

Business Analytics Capstone

ISM 4300: Managing Information Resources

Dr. Markum Reed

April 17, 2024

**Introduction:**

In today's fast-paced healthcare landscape, accessing timely medical assistance is often challenging, leading to delays in detecting and addressing serious health issues. This project acknowledges the critical need for proactive healthcare solutions that empower medical professionals to swiftly diagnose patients, especially those at risk of medical emergencies like heart disease, during doctor visits.

**Methodology:**

**Data Strategy:**

Utilized medical datasets for anonymized patient data, health surveys that collect data on risk factors, and research literature. For the main medial dataset, data cleaning, normalization and standardization, and feature engineering processes was used.

**Analytical Techniques:**

Utilized machine learning algorithms such as: logistic regression, support vector machines, decision tree, random forest, and k-nearest neighbor to identify the relationship and importance of features. Python offers a rich ecosystem of libraries and tools for data manipulation, analysis, and modeling, of which I used numpy, panda, matplotlib, sklearn, and seaborn. Tableau was essential for building an interactive and dynamic user interface for making predictions based on the important features found through the machine learning algorithms in python.

**Project Management:**

This project was managed using an Agile methodology, phase one is data collection and preprocessing which took two weeks, phase two is model development and training which six weeks, phase three is Tableau implementation and integration which took two weeks, phase four and phase five is the completion and final presentation which took the remaining two weeks.

Changes:

Throughout the course of the project PatientPal underwent sweeping changes. The original idea was to create an interactive AI-driven chatbot that a patient at home could use to predict how at-risk they are based on the feature importance established within the Python code, and to feature a SQL driven database for patients to sign up for clinician visits. However, due to a plethora of both internal and external factors, the direction of the project had to shift into its current iteration.

First, the scope of creating an AI-driven chatbot was too ambitious given my fundamental programming skills, time, and available resources. I tried to create the chatbot using HTML, CSS, and JavaScript, however I had never worked with any of these languages prior to this course. So, when trying to program a chatbot I ran into countless errors that would take considerable time and effort each time to fix. Also, due to the complexity of AI I had to integrate open-AI into my chatbot which worked, but it is an outsourced product that I lack full control over which introduced a slew of problems. The breaking point is when I tried to implement my Python code into my chatbot, and the errors started to pile up which made it clear that I lacked the fundamental skills and time required to finish this project by the due date should I continue down this path. Therefore, I switched to Tableau which would accomplish my desired task, allow for my full control, exclude the fluff like begin able to schedule appointments which is unrelated to the task, allow me to work with a tool that I have extensive knowledge of and am comfortable with, and ensure that the project is finished by the due date.

Next, an ethical dilemma arose as I began working on the first iteration of my project. As previously stated, the original idea was to create an interactive AI-driven chatbot that a patient at home could use to predict how at-risk they are. Putting the control of diagnosing a serious health condition such as heart disease in the hands of the patient shows a clear lack of fundamental understanding of the medical field and the responsibilities of health care providers. First, if you are concerned enough to use a chatbot to see how at-risk you are for heart disease then you should seek professional medical advice and or treatment immediately. Next, it is financially beneficial for the medical industry to have patients come in to get assessed for heart disease rather than a patient indulging in self-diagnosis and or self-treatment. Finally, medical professionals have an ethical responsibility to protect the individual and having a chatbot that relies on a patient imputing their own data creates a serious risk to that patient. Therefore, another reason why the project was switched from a chatbot to be used by the general population, to a dashboard predictor that should only be used by medical professionals to aid in the in-person clinical diagnosing of heart disease.

Finally, after extensive contemplation over the idea, I came to the realization that my project would be more effective in a Tableau visualization as opposed to a chatbot. With Tableau visualization I can easily showcase all the factors in the form of bar charts with corresponding drop-down menus to select specified data points. A medical professional can then record the patients thalach, trestbps, exang, cp, ca, and age directly into the interactive dashboard and in real-time the graphs will change to display how many individuals with the same characteristics have been diagnosed with heart disease. The medical professional should not solely rely on this predictor; however, it will serve as an accurate prediction based on real patient data. This is more interactive, more intuitive, and overall, a more effective model than having to manually enter the patients data into a chatbot and then instead of seeing an actual visualization, only seeing a percentage.

Findings:

Throughout the course of this project a variety of findings were made. First, I utilized five different machine learning algorithms, which included: logistic regression, support vector machines, decision tree, random forest, and k-nearest neighbor. Of the five, k-nearest neighbor had the highest accuracy and was therefore used when finding the feature importance. Next, using k-nearest neighbor I identified the feature importance. There are fourteen features in all, however I used the most important six when creating my Tableau prediction dashboard, which includes: thalach, trestbps, exang, cp, ca, and age. Finally, through using my prediction dashboard I have found what characteristics a person diagnosed with heart disease displays and therefore can predict the risk of a patient based on their personal characteristics exhibited.

Conclusions:

This project addresses the urgent need for proactive healthcare solutions by leveraging machine learning algorithms to predict the risk of heart disease in patients during medical consultations. The methodology involved data collection from medical datasets, surveys, and research literature, followed by data preprocessing and feature engineering. Analytical techniques such as logistic regression, support vector machines, decision trees, random forests, and k-nearest neighbors were applied to identify feature importance. Python libraries like numpy, pandas, matplotlib, scikit-learn, and seaborn facilitated data analysis and modeling, while Tableau was instrumental in creating an interactive dashboard for predictions.

The project underwent significant changes, transitioning from an ambitious AI-driven chatbot to a more feasible Tableau visualization due to technical constraints and ethical considerations. The final approach involves a predictive dashboard tailored for medical professionals, highlighting key features such as thalach, trestbps, exang, cp, ca, and age in diagnosing heart disease. K-nearest neighbors emerged as the most accurate algorithm, leading to insightful findings about the characteristics associated with heart disease and enabling risk prediction based on individual patient traits.

In conclusion, this project showcases the effectiveness of machine learning in healthcare decision-making and underscores the importance of ethical considerations and technical feasibility in project development. The predictive dashboard offers a valuable tool for medical professionals to enhance clinical diagnosis and improve patient outcomes in the context of heart disease management.

References:

Found on 2/7/2024: <https://chat.openai.com/>

Note: OpeanAI Chatbot

Found on 2/7/2024: <https://www.kaggle.com/datasets/priyanka841/heart-disease-prediction-uci/data>

Note: Heart Disease Prediction UCI dataset

Found on 2/7/2024: <https://www.projectpro.io/article/heart-disease-prediction-using-machine-learning-project/615>

Note: Tutorial on using Python to create a heart disease prediction model

Found on 2/8/2024: <https://towardsdatascience.com/heart-disease-uci-diagnosis-prediction-b1943ee835a7>

Note: Secondary Tutorial on using Python to create a heart disease prediction model

Found on 3/12/2024: <https://www.geeksforgeeks.org/machine-learning/>

Note: Description and Tutorial of Machine Learning in python

Found on 3/12/2024: <https://www.geeksforgeeks.org/getting-started-with-classification/>

Note: Description and Tutorial of for Getting Started with Classification

Found on 3/13/2024: <https://www.geeksforgeeks.org/introduction-to-seaborn-python/>

Note: Description and Tutorial of Seaborn

Found on 3/13/2024: <https://www.atlassian.com/data/notebook/how-to-save-a-plot-to-a-file-using-matplotlib>

Note: Tutorial on How to Save a Plot Using Matplotlib

Found on 3/14/2024: <https://www.geeksforgeeks.org/understanding-logistic-regression/>

Note: Description and Tutorial of Logistic Regression in Machine Learning

Found on 3/16/2024: <https://www.geeksforgeeks.org/decision-tree/>

Note: Description and Tutorial of Decision Tree Classifier

Found on 3/18/2024: <https://www.geeksforgeeks.org/random-forest-algorithm-in-machine-learning/>

Note: Description and Tutorial of Random Forest Algorithm in Machine Learning

Found on 3/20/2024: <https://www.geeksforgeeks.org/support-vector-machine-algorithm/>

Note: Description and Tutorial of Support Vector Machine (SVM) Algorithm

Found on 3/23/2024: <https://www.geeksforgeeks.org/k-nearest-neighbours/>

Note: Description and Tutorial of K-Nearest Neighbor (KNN) Algorithm

Found on 3/24/2024: <https://machinelearningmastery.com/calculate-feature-importance-with-python/>

Note: Tutorial on Calculating Feature Importance in Python

Found on 3/24/2024: <https://builtin.com/data-science/feature-importance>

Note: Secondary Tutorial on Calculating Feature Importance in Python

Found on 3/28/2024: <https://www.geeksforgeeks.org/tableau-tutorial/>

Note: Tutorial on Tableau

Found on 4/3/2024: <https://help.tableau.com/current/pro/desktop/en-us/calculations_bins.htm>

Note: Tutorial on Creating Bins in Tableau

Found on 4/6/2024: <https://www.datacamp.com/tutorial/calculalting-fields-tableau-tutorial>

Note: Tutorial on Creating Calculated Fields in Tableau

Found on 4/11/2024: <https://stackoverflow.com/questions/31684375/automatically-create-file-requirements-txt>

Note: Tutorial on Creating an Automatic requirements.txt File

Found on: 4/13/2024: <https://www.cdc.gov/heartdisease/facts.htm#:~:text=Heart%20disease%20is%20the%20leading,groups%20in%20the%20United%20States.&text=One%20person%20dies%20every%2033,United%20States%20from%20cardiovascular%20disease.&text=About%20695%2C000%20people%20in%20the,1%20in%20every%205%20deaths>.

Note: Statistics on Heart Disease in the United States