CIS 635 Project Progress Report - Predicting Cardiovascular Disease

Team members: Alyssa Adamczak, Leah Bishop, Laxmi Sowjanya Doddi, & Jonathan Kivuva

Dataset: [Cardiovascular Disease dataset - kaggle.com]

Project Progress Overview

This project aims to build a model that predicts the presence or absence of cardiovascular disease (CVD) by assessing which features are significant risk factors for CVD. The dataset utilized by this project - *Cardiovascular Disease* - includes data from 70,000 patient records, eleven features, and the target feature of the presence or absence of CVD. The dataset is linked above.

This project is organized into four main sections: data collection and preprocessing, exploratory data analysis, modeling, and modeling evaluation. We are currently finishing up the exploratory data analysis section. We have completed all steps in the data collection and preprocessing section, including importing the dataset, handling missing values, and handling duplicates. For the exploratory data analysis portion of this project, we have computed descriptive statistics and visually inspected the data using histograms and Q-Q plots. We identified outliers 1.5 times above or below the interquartile range (IQR), however, we are still discussing how to handle these data points. After determining future steps, we will continue with our modeling and modeling evaluation steps.

Overall, we believe we are making good progress on this project. We look forward to building our model and assessing its performance on predicting the presence or absence of CVD.

Challenges

While we have been making good progress on this project, we did face a minor issue while working through the data preprocessing and exploratory data analysis sections. When formulating our implementation plan, we anticipated that removing missing values and duplicates from our dataset could pose a challenge, especially considering the dataset's large size. However, after performing descriptive statistics, we found an issue with the 'age' feature. Descriptive statistics for 'age' gave seemingly unrealistic results (ex. mean value of 19468.87) and inspection of the raw data revealed values that also appeared unrealistic (ex. age of 18393). After referencing the data source, kaggle.com, we realized that 'age' was given in units of days. To make our analysis more logical and coherent, we converted all values of 'age' from units of days to years. While this was only a minor issue, we are confident in our ability to handle any future challenges we may encounter as we move forward with this project.

Collaboration

This project has highlighted the importance of collaboration and communication among team members. So far, our team has relied heavily on tools like Google Docs, Google Colab, and Slack to facilitate communication and idea sharing. We have found Google Docs to be a particularly useful platform while working on our project proposal – and even this progress report – because everyone has access to the most up-to-date document with editing capabilities. Using the "comments" feature of Google Docs has helped facilitate the brainstorming of ideas and refinement of solutions. We have utilized Google Colab in much the same way for the coding portion of this project. Due to differences in schedules, it is difficult to coordinate meeting times and because of this, we have only been able to meet once in person as a full group. Despite this, we have found Slack to be an effective way to stay in communication with one another on the status of our project.

At this point in the project, we have recognized that there is room for improvement in our collaborative efforts. While each member brings their own valuable skill sets and insights to this project, there have been some challenges with consistently sharing the workload thus far. Moving forward, we plan to better delegate tasks so that each member has a more clearly defined role within the project. This will not only help to achieve the end goal of a successful project, but it will also help foster accountability and provide each member with an opportunity to contribute work they are proud of. To support this, we plan to hold regular check-ins via Slack and keep open communication to help one another as needed. We are optimistic that these adjustments will create a more productive and unified approach for the remainder of the project.

Next Steps

Our next steps involve implementing supervised machine learning algorithms to enhance the model's accuracy in predicting CVD. Specifically, we will begin by constructing a decision tree and a regression tree to explore the relationships between features and the target variable, the presence or absence of CVD. Decision trees are advantageous as they can capture non-linear relationships and provide a visual representation of feature importance, which may be useful in identifying significant CVD risk factors.

To evaluate our model's effectiveness, we will generate a confusion matrix to assess its predictive accuracy. The confusion matrix will allow us to calculate key performance metrics such as accuracy, recall, precision, and F1-score. These metrics will provide insights into our model's predictive capability and reveal any areas for improvement, guiding further refinement of our approach to ensure robust performance in predicting the presence or absence of CVD.