Predictive Analysis of Cardiovascular Disease Using Python

Ty Johnson

tjohnson@bellarmine.edu

16 January 2024

Executive Summary

Cardiovascular disease is the leading cause of death worldwide and in our country. This project will dive into the group of diseases and seek out to understand the main factors that contribute to them. Using a relatively large dataset of patients, several predictive analytic models will be created in an attempt to assess whether or not an individual will have a cardiovascular disease. Each of these models will utilize a different separate predictive analysis technique in order to differentiate between which model will perform best for the type of information dealt with in the project. In addition to model building, an exploratory data analysis will be undertaken that aims to answer several key questions related to the factors that cause cardiovascular disease. The project will utilize the Python programming language and several of the libraries included in the language, in addition to the data visualization program, Tableau. Can cardiovascular disease be accurately predicted with demographic, lifestyle, and biometrical information? The models created in this project seek to answer this question.

Project Idea

The project will aim to investigate the various factors that contribute to cardiovascular disease and develop mulitple predictive models that will be able to assess whether or not an individual will be diagnosed with a disease that falls into the category. The ultimate goal of the project and models will be used to predict a binary outcome: 0 - no heart disease, and 1 - diagnosis of cardiovascular disease. There will be various questions in the exploratory data analysis that the project will attempt to answer. How does alcohol consumption and smoking play a role in the cause? How effective is being physically active as a preventative measure? What is the correlation between obesity and cardiovascular disease? These are only a few examples of questions that the project will attempt to answer, and it will experiment to see how changes in the underlying factors may change the diagnosis.

Background

Cardiovascular disease is the leading cause of death in the United States and globally. In 2021, it accounted for 1 in every 5 deaths. This rate is higher worldwide, as 32% of global deaths fell into the category of cardiovascular diseases. Cardiovascular disease repersents the general group of disorders for the heart and blood vessels. The most common type is cornorary heart disease, which has to do with the blood vessels that supply the heart muscle. Cardiovascular disease is amongst the most researched medical fields, with the main causes of it being understood. There are many resources out there to see whether or not an individual is at risk, such as reading articles and info pages of medical institutions around the world. This project will aim to serve as another resource that could

be utilized for understanding the risk that an individual faces. The dataset that will be used includes 70,000 entries with information related to cardiovascular disease. The information that is tracked in the dataset can be separated into three main categories. First, there are the lifestyle factors that play a role in heart disease: Is the person a smoker? Do they consume alcohol? Are they physically active? Second, there are demographic factors such as age and gender. Third, there are biometrical indicators, such as height, weight, blood pressure levels, cholestorol, and glucose. The dataset also tracks if the person suffers from a cardiovascular heart disease, which will be the dependent variable of the predictive analysis.

Modeling

There will be a few models used in the project. Since it is attempting to predict a binary outcome, a logistic regression model will be developed first in order to serve as a baseline for the models that will come later. There are three different main models that may be used. The first model could be k-Nearest neighbor. It is easy to implement, making it the perfect follow up model to logistic regression. Since the dataset is large, it should allow the data to be easily balanced while keeping many predictions available for the model to be built on. The results are supposed to be easy to interpret as well, making it a good one to utilize. The second model could be a decision tree. The dataset uses both numeric and categorical data, which can be handled by a decision tree. Furthermore, the results of the decision tree model should be easy to interpret. It may allow us to understand the most likely path that can lead to someone developing cardiovascular disease. It can handle nonlinear relationships, which is the case in this project. The third model will be a random forest model. This model should build on the decision tree model and add improved accuracy. Random forest models emphasize each feature of the model, which may be beneficial since one of the goals in the project is to understand how changes in the underlying factors may produce a different diagnosis. This type of model is able to handle larger and more complex datasets, though this dataset may not be complex, it is large. If there are strong outliers present in the data, this model will be able to withstand them as well. The implementation of each model could be a good way to understand which type of model performs best for datasets that are structured similar to the one used in this project.

Tools

The project will use different tools for the various stages of the project. Python will be the only programming language used. The simplicity of the language and support from various libraries make it the perfect language for this predictive analysis. R shouldn’t be needed since there will not be any heavy statistical analysis.

VSCode will be used in conjunction with the Jupyter Notebooks extension. There are a lot of extensions and features available in the editor that will make the programming experience easier. For data cleaning and preprocessing, the pandas library will be used. The dataset is in a CSV file, so pandas will be great for making manipulations to the dataset. For model building, the scikit-learn library will be used so the models will not have to be built from scratch. There will be a few different tools used in the visualization process. Matplotlib and seaborn will be used to make some visualizations within the notebook. Tableau will be used to make a dashboard. The reason that a dashboard will be to put multiple graphs together into one visual that anyone can look at and easily understand the data.

Conclusion

In conclusion, numerous demographic, lifestyle, and biometrical factors contribute to an individual developing cardiovascular disease. Using these factors, the project will aim to predict if an individual has a cardiovascular disease. The project will seek to understand questions related to the factors that cause the group of diseases as well, and visualize them by means of graphs. It is important to understand cardiovascular diseases and the factors that contribute to them as it accounts for 32% of deaths globally and around 20% of deaths in the United States each year. Several different models will be created with the Python programming language and libraries. By leveraging machine learning models such as k-Nearest means, decision trees, and random forest, this project aims to become a reliable prediction tool. Overall, this project should highlight the importance of data-driven approaches in understanding and combatting major health issues around the world.

References

Cardiovascular diseases (CVDs) (who.int)

Heart Disease | cdc.gov

Risk Factors for Cardiovascular Heart Disease (kaggle.com)

K-Nearest Neighbor(KNN) Algorithm - GeeksforGeeks

Python | Decision tree implementation - GeeksforGeeks

A Simple Introduction to Random Forests (statology.org)