White Paper

Cardiovascular Disease Factor Analysis

DSC 680

Prof. Amirfarrokh Iranitalab

Carlos Cano

Business Problem:

From the moment life begins and each of our hearts take their first beats it will continue to do so until we die. Several billion-dollar industries exist from Healthcare to Fitness the world over, it is from this that we must focus on what factors contribute to cardiac disease. By understanding the factors and the magnitudes they play on our cardiac health we can all understand the contributing nature through the use of data science.

Background/History:

Trash in, trash out is a saying taken from data science. It expresses the causality of how we process data. This adage has been expressed many times over by fitness junkies who thrive and seek better ways of understanding their bodies. In aspect this too is akin to the relationship and sometimes self-preserved nature we undertake in maintaining our own health as we age.

The goal of the data is to express the effect of how these factors can lead to cardiovascular disease which ultimately could lead to a dire situation in which a patient may be at risk for a cardiovascular event of a heart attack.

Data Explanation (Data Prep):

Data processing for this EDA consisted of importation, exploration, modification, and implementation. The variables within were viewed several were chosen as their provided an intuitive explanation for the risk of cardiovascular disease that can found in patients across the world over.

Methods:

The methods focused on this research will focus on both a linear regression-based analysis and a KNN classification approach that will be used express the efficacy of how our model is able to predict instances of cardiovascular disease in patients.

Analysis:

A brief graphical analysis highlights the variable of Cholesterol within the patients in the dataset, this can be highlighted by Figure 1 and 2, which both express the relationship from a macro prospective relative to the grand populous and by a gender component. This data showcases the variations found with the data and serves as a clear indicator of how much this variable relates to cardiovascular disease. By undertaking this EDA, we can begin to assess how the data can be transform in such a way to express the original problem of understanding how the variables within the data can reasonable predict the presence of Cardiovascular disease within a patient. With a simple linear regression of the model, it becomes evident that the accuracy of 52% is not sufficient enough to justify the application of these metric to predict reasonably the presence of cardiovascular disease. By imploring the use of a more advance data science technique such as KNN with Hyperparameter Modeling with can see a more robust exploration and explanation of the data's predictable efforts as highlighted with an accuracy of 91.5%.

Conclusion:

The efficacy of this EDA highlights the importance of applying the right tool for the job, while the use of a Linear Relationship Model is typically the go to for how measure the relationship and efficacy of models in data science, it is not the end all by any means. The application of a KNN Classifier with a hyperparameters was of a more statistically significant solution for the efficacy in this study and exploration. As such it is of a conclusive nature that the Model in this circumstance has enough bearing that it can be used to accurately predict the presence of cardiovascular disease in patients.

Assumptions:

The main assumptions here are that the is whole in its nature and has not been tampered with, as the source data did not contain any Nans, it is high suspect in its original format. Given this fact, the data may have been preprocessed in some fashion undisclosed.

Limitations:

The data doesn't highlight enough external variables such as diet and lifestyle factors that may have contributed to each variable functional output. In several iterations of this study account for the removal of some or most variables the results varied sometimes by up to 15% in both Models. A tertiary logistical regression model served as guidance in variable selection. So, although a 91.5% accuracy was garnered the limitations found were in the ability of the model to benefit the most from key removal of certain variables.

Challenges:

Hyper parameterization of the model most challenging part of this EDA, understand and applying this technique which the model while a working theory had original exist was proved in valuable in its application. A separate challenge that existed was datasets curation which was aided through the use of logistical regression to calculate the weight of each variable effect on the model.

Future Users/Additional Applications:

Future users may find more refined techniques which could be explored to future curate and refine the model as it exists. The bearing of this on the EDA can one day be further refined by more advance techniques. This modeling application can be used to further examine other disease and medical factors that can be found in, such as diabetes, kidney disease, or possibly a way to predict future instances of strokes in patient before they even happen.

Recommendations:

Continued exploration and refinement technique should be explored as this area of research contain many significances in quite possibly all of our lives. By further studying the variables and their measured weights on our overall cardiovascular disease predictability we can all better understand the role each of them plays. This data, possibly studied over time can be used to create more robust examinations into how we view our cardiovascular health. In the exploration of the graphical data the explored variable of Cholesterol which typically comes to mind when associated cardiovascular disease to a single data point, highlights many of the singular effects. Though this approach has many fallacies it can be a steppingstone for many to expand how we view this disease in a more robust and possibly complete picture.

Implementation Plan:

The execution of the initial plan has several diversions which couldn't account for the variations in this the focus of narrowing down the data was spawned. Further explorations of medical journals and research therein would serve as specific guidance of factors that would contribute to cardiovascular disease and overall cardiac health. It was in the spirit of that understanding that this plan was laid out, and it is in that spirit that if further research into this area is done that, we much continue to strive to make greater leaps and bounds to further refine this and future models.

Ethical Assessment:

An initial ethical dilemma of this EDA would be the tunnel vision approach set in by focusing on any one single variable over another. It was for this purpose the imploration of a logistical regression was used. As such the manipulation of data serves as a remind of how an incomplete or manipulated dataset can be used to construe the data expressly coerce a specific agenda.

Graphical Representations

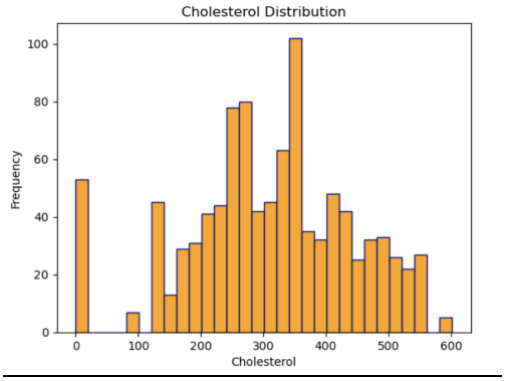

Figure 1.

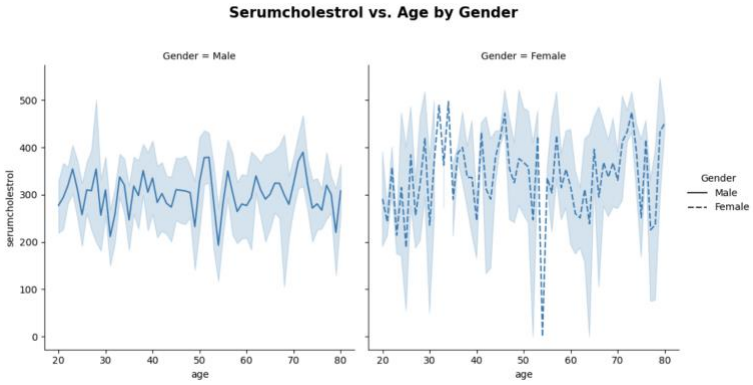

Figure 2.

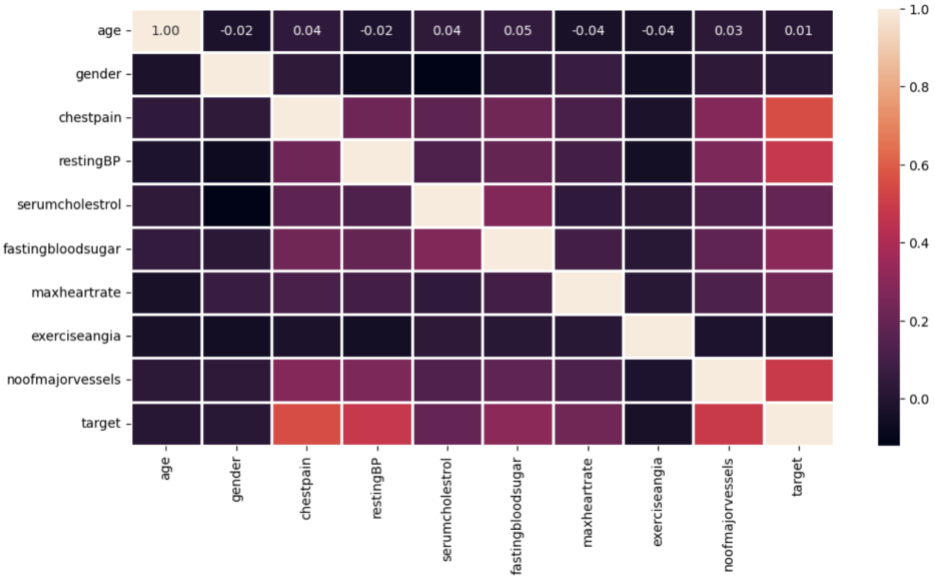

Figure 3.

References

Doppala, B. P. (2021, April 16). *Cardiovascular_Disease_Dataset*. Mendeley Data. https://data.mendeley.com/datasets/dzz48mvjht/1