Modeling Heart Attack Risk

Alice Ding

DATA621

Fall 2023

Abstract

Myocardial infarctions, or heart attacks, have grown increasingly problematic within the eyes of global and public health. The purpose of this paper and corresponding analysis is to see if there is a way to predict whether a patient is at risk of having a heart attack given certain attributes ranging from age, cholesterol levels, blood pressure, smoking habits, exercise patterns, income, and more. By using predictive analytics, specifically binomial regression and decision trees, the goal of this is to determine whether a combination of certain attributes or the presence of others can reliably predict the likelihood of a heart attack.

Key Words: heart attack, cardiovascular disease, cardiovascular risk factors, prediction of heart attacks

Introduction

Heart attacks occur when blood flow to part of the heart muscle is blocked, often due to the formation of a blood clot or other health issues within the body. As myocardial infarctions are a leading cause of mortality worldwide, this paper serves as a way to pinpoint the various factors that contribute to heightened risk for these deadly attacks and to help inform those that they may be at risk.

Etiologically, heart attacks can be caused by a combination of genetic predispositions and modifiable risk factors such as lifestyle choices and environmental influences. Some examples of contributors can include but are not limited to hypertension, cholesterol levels, diabetes, smoking habits, activity levels, and more. Understanding how these various factors can and do interact is crucial for developing prevention strategies and intervention measures.

Given how prevalent heart attacks are, it is imperative that we explore this risk and derive its causes and drivers. By shedding light on the web of factors contributing to this health issue, we can empower individuals, healthcare professionals, policy makers, and more with the knowledge to implement proactive and preventative measures, thus mitigating risk and helping decrease the prevalence of this issue slowly as medicine progresses.

Literature Review

Upon research within this topic, several articles regarding cardiovascular disease risk emerged. In a recent journal published by Adhikary, Barman, Ranjan, & Stone (2022), they noted that different countries had a higher prevalence of CVD (cardiovascular diseases) due to the coexistence of multiple risk factors that were present in certain areas compared to others. Adhikary, Barman, Ranjan, Stone (2022) also mention that, “potential risk factors for coronary artery diseases are hypertension, obesity, and physical inactivity.” Adhikary, Barman, Ranjan, & Stone (2022) have done some extensive research regarding the topic and their analyses compares all age groups and genders from across the global population, however their limitations include focusing only on English-language papers and the available ones they chose from are mainly from Asia, Europe, Africa, and the United States. Given the ultimate goal would be to help remove heart attack risk from around the world, the scope of their work is limited. This is somewhat similar to the work done in this project though as the dataset being analyzed is also very biased towards Asia and Europe, meaning neither pieces of work encompass the entire world’s population or are equipped to predict based on any and all individuals.

Meanwhile, it has also been noted elsewhere that those with a total cholesterol level of < 180 mg per deciliter, < 120 mm Hg systolic and 80 mm Hg diastolic blood pressure, and those with nonsmoking and nondiabetic statuses have substantially lower risks of death from cardiovascular disease through the age of 80 years than those with two or more major risk factors (Berry, Dyer, Cai, Garside, Ning, Thomas, Greenland, Van Horn, Tracy, & Lloyd-Jones (2012)).This analysis however was only done for black and white individuals at the ages of 45, 55, 65, and 75, thus missing a large portion of the global population. The fact that they measured risk though at different ages across the same people's lives is a good observation -- the same people with different health issues across the span of one's life is an helpful aspect to examine and can help predict the actual effect of certain health issues depending on age and other factors. This is not similar to the work done in this paper / project as there is no limitation on race and this is a snapshot of data taken from one point in time rather than over the scope of individuals' lives.

Hypertension, or high blood pressure (BP), in particular has been known to be a major cause of cardiovascular disease (CVD). Interestingly, Fuchs and Whelton (2019) propose that as a society, we have been artificially inflating what is deemed as a normal BP and thus underrepresenting the impact of what high BP on the likelihood of CVD. This hypothesis is strong in the fact that it does its best to meet Occam's razor promise (has the least amount of assumptions), however it still is hard to say how much hypertension contributes to heart attack risk as other aspects such as sodium, physical inactivity, lipid abnormalities, smoking habits, and more are also large factors. The analysis done by these folks is different from the one discussed in this paper as it focuses just on one variable that can lead to heart attacks rather than a myriad of them, although it does address that CVD is impacted by other factors; overall, it does not quantify how impactful each one is though unlike this one which has the ability to do so. This analysis though does not address the fact that society seemingly has been inflating BP.

Methodology

Given a dataset with almost 9,000 individuals of all genders and ages 18 to 90, the presence of specific health and activity-related variables was used to predict whether each user at the time of the data collection was at risk of a heart attack or not. Various socioeconomic fields were included in our dataset as well to see if external factors beyond just health fields contribute to this risk as well (Appendix A).

Basic data exploration was conducted by graphing and showing distributions and correlations for all of the fields individually and in relation to the target variable (Appendices B and C). Upon discovering if there were nulls and outliers, data preparation occurred to address those instances if they existed and normalization for numerical fields was applied as well.

For model creation and selection, two sets of models were created after splitting the training data: one using binomial logistic regression and the other using a decision tree. These models were refined and adjusted depending on initial outputs to try to create the most accurate versions of themselves. The ultimate decision came down to overall performance with the test dataset to determine which model performed the best under those specific conditions.

Experimentation and Results

Beginning with data exploration, overall distributions and correlations were first charted. It was noted that there were no outliers in any of the numerical variables which was an interesting observation and helpful in terms of data preparation later down the line as nothing had to be imputed or removed. There was not much of a normal distribution for any of the variables though which led to slight issues in normalization during the data preparation phase. Correlations also were not particularly strong amongst variables or with the target variable itself, something that was interesting and helpful as no variables had to be removed as a result..

At a high-level, it should be noted that a majority of our sample size came from Asia and Europe, leading to a majority of our participants being located in the Northern Hemisphere. Other disproportionate categorical variable observations were a large proportion of our population being non-smoking, diabetic, and male. Whether this is proportional to the current state of the entire world's population, this is unknown. While this is not something that could be adjusted, there was also an observation that our target variable's distribution was not equal so up-sampling was done in order to remedy that and account for class bias.

As mentioned previously, normalization posed an issue. Several methods were attempted such as log, scaling, and square-rooting, however none of these assisted in adjusting the distribution. Scaling was eventually chosen as the methodology and those transformed fields were used instead of the ones presented in the original dataset as it was slightly better than the non-transformed values, however the lack of normal data was not ideal for our modeling step.

Utilizing R's generalized linear models (glm) function, binomial regression was chosen as the first algorithm. This was picked as the problem statement had two different outcomes: 0 for not at risk for a heart attack, 1 for at risk. Two models were created with this methodology and compared to see which had the greatest success in predicting whether each person was at risk of a heart attack, the first using all variables the dataset had to offer and the second using only those that were statistically significant.

The second model was created using rpart's decision tree modeling. The first iteration of this model used all available fields, similar to how the first model was approached, but the tree that was created utilized only one variable: country (Appendix D). The second iteration of this model removed that field to see what other factors could be used and ultimately, the one that was created utilized income, BMI, and cholesterol (Appendix E).

Once these four models were created, confusion matrices were formed using the test dataset and then compared using five metrics: accuracy, precision, specificity, recall, and F1 (Appendix F). While the second iteration of the decision tree model performed the best in accuracy, precision, recall, and F1, its specificity was extremely low; recall on the other hand was very high, insinuating that this model was over capturing positives. Implication wise, this would mean that people would be deemed as at risk for a heart attack when in reality, they're not. This could then lead to preventative measures being implemented and could potentially be very costly and time-consuming. In order to avoid this and inducing stress on those given false positives, a choice was made to select the second iteration of our binomial regression model as it performed second best and thus was the leading choice.

Discussion and Conclusions

While the models created for this project were pruned down to the best fitting one, the reality is that accuracy is still only 50% overall for the final decision; this means that flipping a coin has the same results as running through the selected model. In the future, perhaps it would be best to create models based on data from certain regions of the world given the reality that this is a global issue and not every human is the same. Race or country of origin can heavily contribute, as discussed by Adhikary, Barman, Ranjan, and Stone (2022). In order to address this issue wholeheartedly, creating a model for every country and appropriately categorizing that way could lead to more accurate models rather than one just generated with a huge breadth of individuals to choose from.

Additionally and in a similar vein, with the size of the dataset, it's hard to say that this work can be directly applied to the greater global population, especially with a 50/50 outcome. Future iterations of this work could be to get more data from various countries and regions throughout the world, enough to create well-fitted models, and then also track these individuals over the course of their lives rather than just being a snapshot in time, similar to Berry, Dyer, Cai, Garside, Ning, Thomas, Greenland, Van Horn, Tracy, & Lloyd-Jones (2012).

References

Adhikary, D., Barman, S., Ranjan, R., & Stone, H. (2022, October 10). A systematic review of major cardiovascular risk factors: A growing global health concern. Cureus. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9644238/

Berry, J. D., Dyer, A., Cai, X., Garside, D. B., Ning, H., Thomas, A., Greenland, P., Van Horn, L., Tracy, R. P., & Lloyd-Jones, D. M. (2012, January 26). Lifetime risks of cardiovascular disease. New England Journal of Medicine. https://www.nejm.org/doi/full/10.1056/NEJMoa1012848

Fuchs, F. D., &amp; Whelton, P. K. (2019, December 23). High blood pressure and cardiovascular disease | hypertension. AHA Journals. https://www.ahajournals.org/doi/10.1161/HYPERTENSIONAHA.119.14240

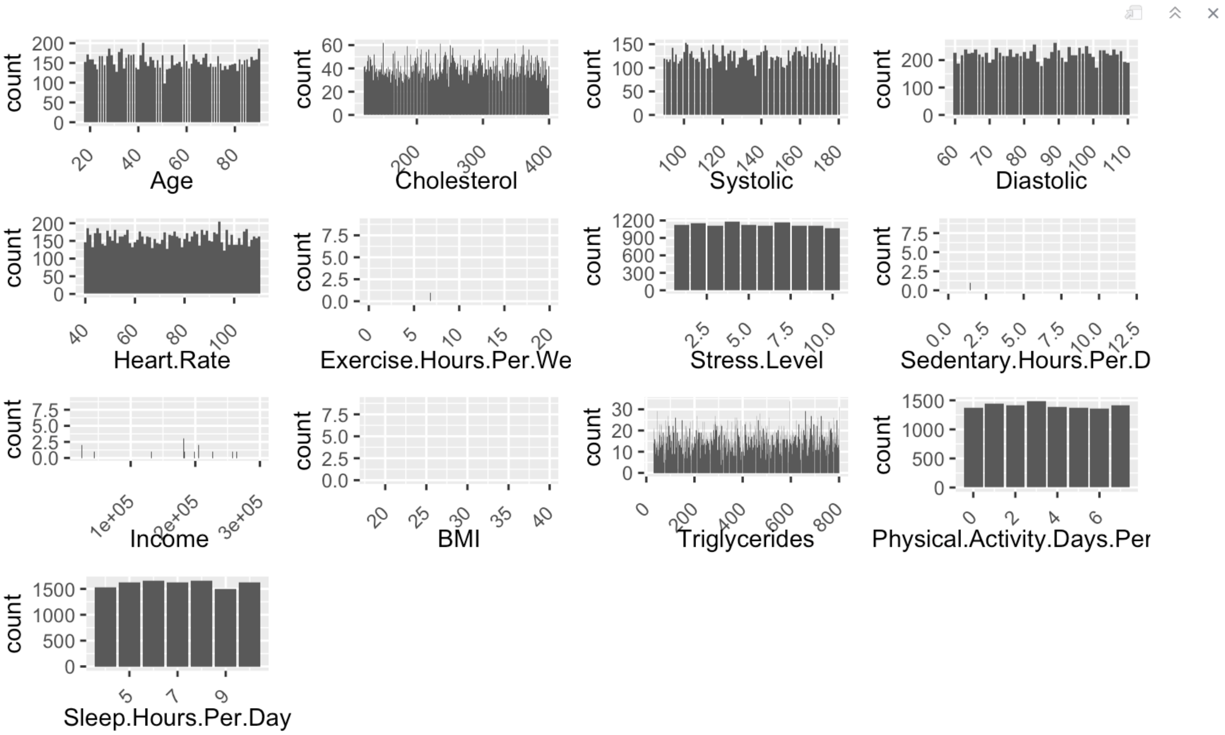
Appendices

Appendix A: Dataset Fields

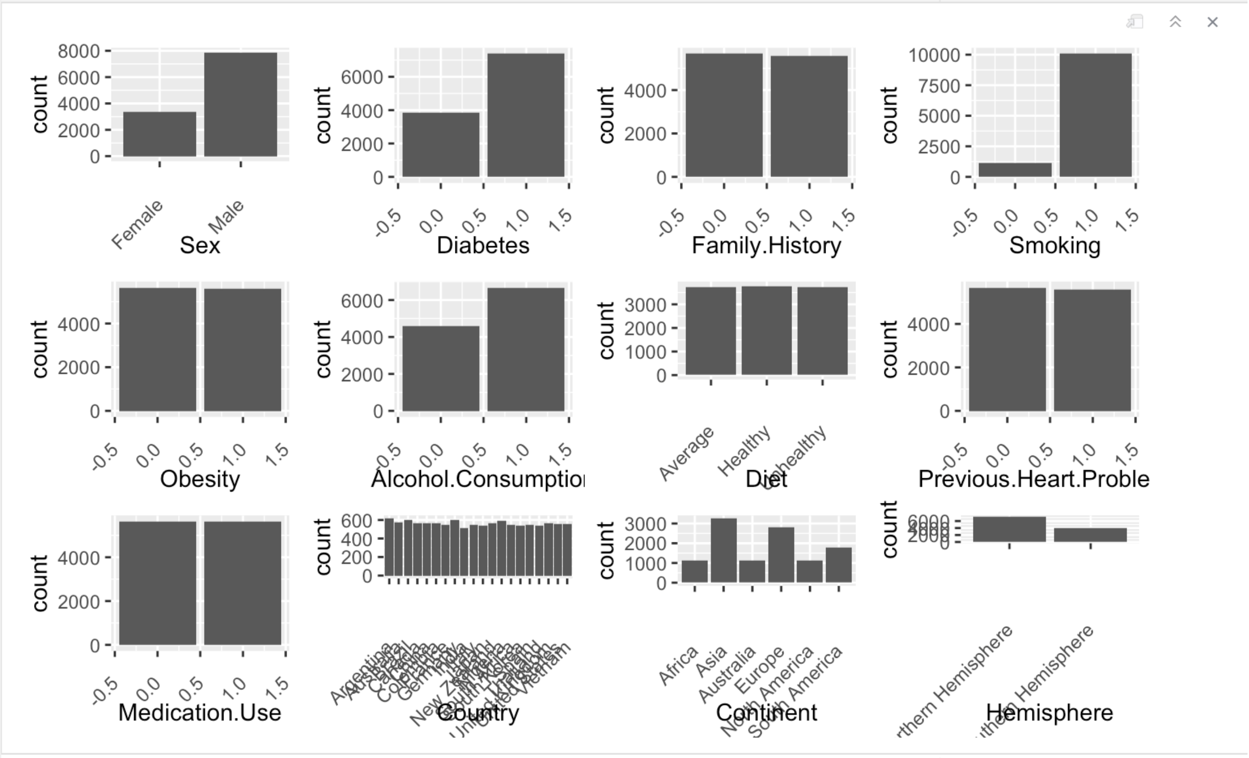
Patient ID	Unique identifier for each patient
Age	Age of the patient
Sex	Gender of the patient (Male/Female)
Cholesterol	Cholesterol levels of the patient
Blood Pressure	Blood pressure of the patient (systolic/diastolic)
Heart Rate	Heart rate of the patient
Diabetes	Whether the patient has diabetes (Yes/No)
Family History	Family history of heart
Smoking	Smoking status of the patient (1: Smoker, 0: Non
Obesity	Obesity status of the patient (1: Obese, 0: Not obese)
Alcohol Consumption	Whether the patient regularly consumes alcohol (1: Yes, 0: No)
Exercise Hours Per Week	Number of exercise hours per week
Diet	Dietary habits of the patient (Healthy/Average/Unhealthy)
Previous Heart Problems	Previous heart problems of the patient (1: Yes, 0: No)
Medication Use	Medication usage by the patient (1: Yes, 0: No)

Stress Level	Stress level reported by the patient (1
Sedentary Hours Per Day	Hours of sedentary activity per day
Income	Income level of the patient
BMI	Body Mass Index (BMI) of the patient
Triglycerides	Triglyceride levels of the patient
Physical Activity Days Per Week	Days of physical activity per week
Sleep Hours Per Day	Hours of sleep per day
Country	Country of the patient
Continent	Continent where the patient resides
Hemisphere	Hemisphere where the patient resides
Heart Attack Risk	Presence of heart attack risk (1: Yes, 0: No) // TARGET VARIABLE

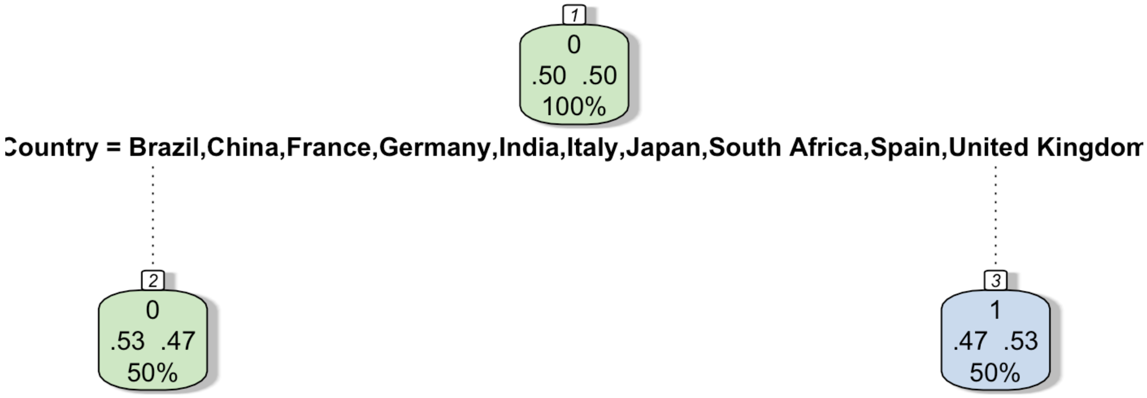
Appendix B: Numeric Field Distributions



Appendix C: Categorical Field Distributions

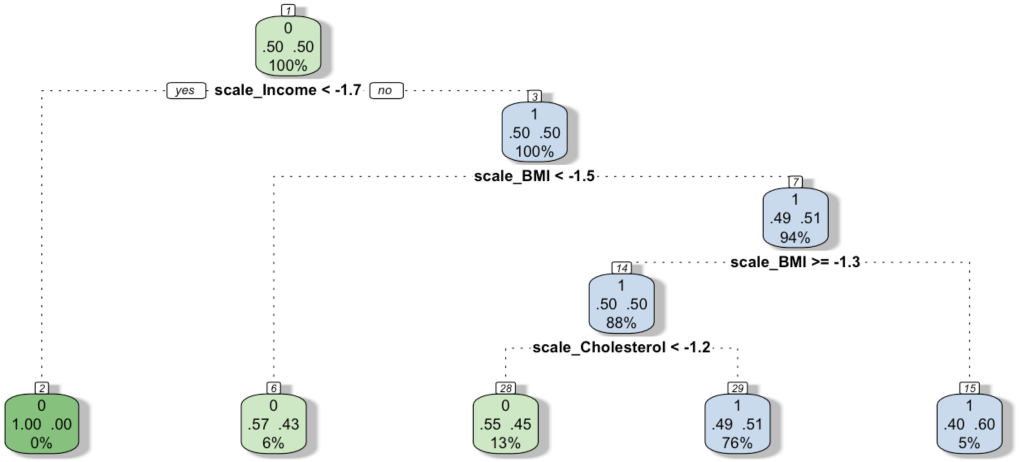


Appendix D: Decision Tree Model 1


1
0
.50 .50
100%
Country = Brazil,China,France,Germany,India,Italy,Japan,South Africa,Spain,United Kingdom
2
0
.53 .47
50%
3
1
.47 .53
50%

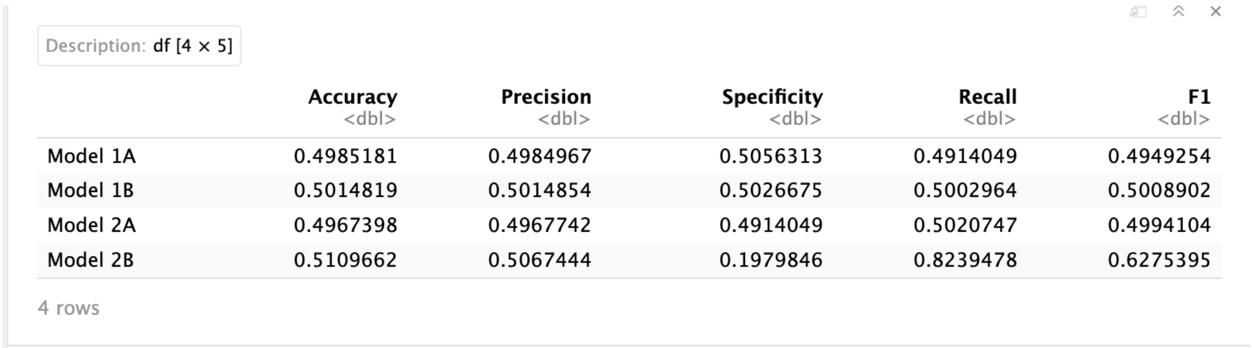
Rattle 2023-Dec-14 17:01:12 ading

Appendix E: Decision Tree Model 2



Rattle 2023-Dec-14 17:01:20 ading

Appendix F: Score Comparisons



Description: df [4 × 5]

	Accuracy	Precision	Specificity	Recall	F1
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
Model 1A	0.4985181	0.4984967	0.5056313	0.4914049	0.4949254
Model 1B	0.5014819	0.5014854	0.5026675	0.5002964	0.5008902
Model 2A	0.4967398	0.4967742	0.4914049	0.5020747	0.4994104
Model 2B	0.5109662	0.5067444	0.1979846	0.8239478	0.6275395

4 rows