Predicting Heart Attacks with Machine Learning: Integrating Wearables for Personalized Prevention Strategies

[redacted]

Introduction:

Heart disease remains one of the leading causes of death worldwide, imposing a substantial burden on individuals, healthcare systems, and societies. Despite advances in treatment options, the most effective strategy lies in preventive measures that mitigate risk factors and promote heart-healthy lifestyles. Early identification of individuals at elevated risk is crucial for implementing timely interventions and empowering them to take proactive steps towards improving their cardiovascular health.

With the widespread adoption of wearable devices and mobile health applications, we have an unprecedented opportunity to integrate advanced risk prediction models with personalized, real-time monitoring and coaching tools. By leveraging the rich data from national health surveys and harnessing the power of machine learning algorithms, we can develop robust models that accurately assess an individual's heart attack risk based on their health indicators, lifestyle factors, medical history, and demographics.

This study aims to develop cutting-edge machine learning models using the CDC National Health Interview Survey dataset, which includes responses from over 400,000 adults across the United States. Through exploratory data analysis, we identify key risk factors and their relationships with heart attack prevalence. Feature selection techniques are employed to determine the most predictive variables from the 40 variables available. We then evaluate the performance of several machine learning algorithms including logistic regression, random forests, and gradient boosting models in predicting heart attack risk. Model accuracy is assessed using metrics such as precision, recall, accuracy, and confusion matrix. The most accurate model is further analyzed to generate insights into significant predictors and their impact on cardiovascular health outcomes.

However, our ambition goes beyond merely predicting risk – we strive to empower individuals with actionable insights and practical tools to facilitate sustainable lifestyle modifications. By seamlessly integrating our risk prediction models with mobile applications and wearable devices, we can provide personalized recommendations tailored to each individual's risk profile, enabling real-time monitoring of health metrics, and guiding them through evidence-based interventions for nutrition, physical activity, stress management, and behavior change.

The convergence of data-driven risk modeling, digital health technologies, and behavior change principles presents a unique opportunity to democratize heart attack prevention at an unprecedented scale. By empowering individuals with personalized risk assessments, contextualized guidance, and interactive coaching, we can foster a proactive approach to heart health management and drive positive lifestyle transformations across diverse populations.

Motivation:

As a cardiologist's kid deeply committed to improving cardiovascular health outcomes, I am driven by the immense potential of this interdisciplinary approach to address a pressing public health challenge. Witnessing firsthand the devastating impact of heart attacks on individuals and their loved ones has fueled my determination to contribute to more effective prevention strategies.

By leveraging the power of data science, machine learning, and digital health technologies, we can transcend the limitations of traditional risk assessment methods and provide individuals with personalized, accessible, and engaging tools to take control of their heart health. The ability to continuously monitor risk factors, receive real-time feedback, and adapt interventions based on individual progress holds the promise of fostering long-lasting behavior change and ultimately reducing the burden of heart attack.

Furthermore, the integration of wearable devices and mobile applications aligns with the growing trend of consumer-centric healthcare, empowering individuals to actively participate in their own well-being. By making heart attack prevention accessible and engaging, we can inspire a cultural shift towards proactive health management and cultivate a deeper understanding of the modifiable risk factors within our control.

Ultimately, this project represents a significant step towards realizing the vision of precision healthcare, where personalized interventions are tailored to each individual's unique circumstances, preferences, and needs. By combining the power of data science, digital technologies, and evidence-based strategies, we can pave the way for a future where heart attack is no longer an inevitable fate but a preventable condition through informed and sustained lifestyle choices.

Previous work:

Because of the nature of the Kaggle heart disease dataset, many prediction systems have been developed. However, [1] Most prior studies relied on small, curated datasets or focused on specific subpopulations (2020 CDC dataset with 18 features). In contrast, our study utilizes the large-scale, comprehensive, and recently released 2022 CDC National Health Interview Survey dataset with over 400,000 adult respondents across the United States [4]. [2] Unlike previous works that used cleansed datasets, our approach tackles the complexities of an uncleaned, raw dataset requiring robust data preprocessing, missing value imputation, and feature engineering techniques.[3] While existing research has primarily concentrated on developing accurate predictive models, our work proposes the novel integration of these models with wearable devices and mobile applications for personalized heart attack prevention.Thus, bridging the gap between risk assessment and practical implementation of preventive measures.[4] In contrast to previous studies that primarily employed traditional models like logistic regression, decision trees, or random forests, our work utilizes the Hybrid Random Forest Linear Model (HRFLM). The HRFLM is a cutting-edge ensemble model that combines the strengths of random forests and linear models, potentially capturing complex non-linear relationships and interactions among predictors more effectively. HRFLM demonstrates high efficacy when provided with numerous features, and exhibits performance on par with logistic regression even when working with a smaller feature set.

Approach

In this study, we employed a comprehensive and rigorous approach to develop accurate machine-learning models for predicting heart attack risk using the 2022 CDC National Health Interview Survey dataset. The process involved several key stages, including data preprocessing, exploratory data analysis, feature engineering, model development, and evaluation.

A. Data Preprocessing - Data Cleaning

Before proceeding with model development, we addressed potential data quality issues within the dataset. Our initial inspection identified missing values, inconsistent data formats, and potential outliers, which we remedied through the following strategies:

- Missing Value Imputation: To ensure the completeness of the dataset, missing values were imputed using appropriate techniques. For numerical variables, mean imputation was employed, while mode imputation was applied for categorical variables.
- Feature Selection: To prevent overfitting of the model and enhance the cleanliness of the dataset, we judiciously removed certain features. Columns with high missing values were omitted, along with those deemed irrelevant to the analysis.
- Outlier Detection and Treatment: Potential outliers were identified using the Interquartile Range (IQR) method to establish lower and upper limits. Subsequently, outliers were addressed by replacing them with the respective limit values, taking into consideration the nature and impact of these outliers on the analysis.

B. Feature Engineering - One Hot Encoding

Before diving into model development, we addressed the challenge of handling categorical variables within the dataset. Recognizing the need for numerical representations, we explored the application of one-hot encoding as a preprocessing strategy.

- We identified columns containing categorical variables, such as 'Sex', 'GeneralHealth', 'RaceEthnicityCategory', and others that were not suitable for direct use in machine learning algorithms.
- Leveraging Pandas' get_dummies() function, we performed one-hot encoding on the identified categorical columns. This process transformed categorical variables into binary representations,

creating new binary columns for each category within a feature.
- We opted to drop the first binary column for each categorical feature to mitigate multicollinearity issues.
- We validated the effectiveness of the one-hot encoding process by inspecting the structure and content of the encoded data frame.
- The transformation ensured compatibility with machine learning algorithms, enhancing the dataset's readiness for model training and analysis.

C. Data Visualization - Analyzing Feature Relationships with the Target Variable.

We meticulously examined the dataset to explore the relationship between its features and the target variable through visualization. Our initial inspection identified a diverse range of features encompassing both numerical and categorical data. To gain deeper insights, we employed various graphical representations tailored to the nature of the data.

Numerical Data Analysis:

- Boxplots: I utilized boxplots to visualize the distribution of numerical features concerning the target variable. These plots provided valuable insights into the central tendency, spread, and presence of outliers within the data.
- Histograms: Histograms were instrumental in providing a detailed view of the distribution of numerical features. By assessing the frequency and density of data points across different ranges, I gained a better understanding of their distribution patterns.

Categorical Data Analysis:

- Pie Charts: For categorical features, I employed pie charts to illustrate the proportion of each category within the dataset. These charts offered a clear visual representation of the distribution of categorical variables.
- Bar Graphs: Bar graphs were utilized to compare different categories within categorical features, allowing me to visualize the proportion or count of each category effectively.
- Heatmaps: To explore the correlation between some categorical features and the target variable, I utilized heatmaps. These visualizations provided a color-coded representation of the strength and direction of the correlation, aiding in feature selection and model interpretation.

D. Dataframes and Pandas - Integral Tools for Data Cleaning, EDA, and Visualization

Throughout the entire data cleaning, exploratory data analysis (EDA), and visualization processes, we relied heavily on DataFrames and the Pandas library. These powerful tools enabled me to efficiently manage, manipulate, and analyze the dataset, facilitating various tasks such as:

Data Cleaning:

- With DataFrames, we identified and addressed potential data quality issues, including missing values, inconsistent data formats, and outliers.
- Utilizing Pandas functions, we performed missing value imputation, feature selection, and outlier detection and treatment, ensuring the dataset's integrity and reliability.

Exploratory Data Analysis (EDA):

- DataFrames provided a structured framework for exploring the dataset's characteristics, distributions, and relationships between variables.
- We employed Pandas functions to calculate descriptive statistics, visualize data distributions, and analyze correlations, gaining valuable insights into the dataset's underlying patterns and trends.

Visualization:

- Leveraging Pandas' integration with visualization libraries such as Matplotlib and Seaborn, we created insightful plots and charts to visualize relationships between features and the target variable.
- DataFrames facilitated the aggregation and manipulation of data, enabling the calculation of proportions and the creation of informative visualizations such as pie charts, bar graphs, and heatmaps.

E. Regularization - Selecting Lasso Regularization for Logistic Regression

In our quest to optimize the logistic regression model, we explored three regularization techniques: Lasso, Ridge, and Elastic Net. Here's a concise summary of the process and the rationale behind our decision:

Evaluation of Regularization Techniques:

- We systematically applied Lasso, Ridge, and Elastic Net regularization to logistic regression models.
- Through rigorous evaluation, including cross-validation, we compared their performance in terms of accuracy scores and model generalization.

Impact of Regularization on Logistic Regression:

- Given the inherent simplicity of logistic regression compared to more complex models like decision trees or hierarchical linear models (HRFLM), we anticipated minimal variation in performance between Lasso and Ridge regularization.
- Elastic Net regularization exhibited slightly inferior performance compared to Lasso and Ridge, likely due to its additional complexity.

Decision and Justification:

- Despite similar accuracy scores between Lasso and Ridge, Lasso regularization showed slightly better performance during cross-validation.
- Considering the simplicity and interpretability of logistic regression, we concluded that the marginal improvements offered by Lasso regularization justify its selection over Ridge.
- Therefore, we decided to proceed with Lasso regularization for training the logistic regression model, as it strikes a balance between model simplicity and performance enhancement.

F. Logistic Regression Performance

In our analysis, logistic regression emerged as a standout performer among other algorithms for our given dataset. This success can likely be attributed to the characteristics of our dataset, which may have been well-suited for logistic regression. Notably, logistic regression tends to perform well when dealing with binary classification tasks and datasets with a relatively smaller number of features.

To enhance the logistic regression model's performance and address potential overfitting, we employed cross-validation to systematically evaluate different regularization techniques. Through this process, Lasso regularization stood out as particularly effective for our dataset. Lasso regularization, known for its ability to induce sparsity in the model by shrinking coefficients towards zero, helped mitigate overfitting while retaining important features for classification.

Furthermore, logistic regression served as a valuable tool in our analysis not only for its predictive performance but also for its capability to detect and mitigate overfitting. By leveraging techniques such as cross-validation and regularization, we were able to ensure the robustness and generalization ability of the logistic regression model, making it a reliable choice for our classification task.

Additionally, we utilized the liblinear solver to optimize the logistic regression model's performance. The solver employs a Coordinate Descent (CD) algorithm, which solves optimization problems by successively performing approximate minimization along coordinate directions or coordinate hyperplanes. This approach is similar to gradient descent in linear regression, facilitating efficient optimization of the logistic regression model's parameters and enhancing its effectiveness in handling our dataset [2].

G. Classification with Decision Tree and Regularization

Given the nature of the task as a classification problem, the decision tree model emerges as a suitable choice among other algorithms due to its inherent interpretability and ability to handle both numerical and categorical data effectively.

The implementation of a Decision Tree Classifier with regularization techniques addresses the inherent challenge of overfitting in decision tree models. Overfitting occurs when a model learns the training data too precisely, leading to poor generalization on unseen data. By employing regularization, we aim to control the complexity of the decision tree model, thus enhancing its ability to generalize to new data.

Initially, we instantiate a base Decision Tree Classifier with a random state to ensure reproducibility. Subsequently, we define a grid of hyperparameters, including max_depth, min_samples_split, min_samples_leaf, and max_features, which govern the structure and complexity of the decision tree.

Utilizing Grid Search Cross Validation (GridSearchCV), we conduct an exhaustive search through the hyperparameter grid. This involves evaluating each combination of hyperparameters using 5-fold cross-validation to identify the combination that maximizes the model's accuracy.

Upon determining the optimal hyperparameter combination, we re-train the model on the entire training dataset using these settings. We then evaluate the trained model's performance on the test dataset, utilizing metrics such as accuracy and a classification report.

Overall, this systematic approach enables us to fine-tune the hyperparameters of the decision tree classifier, thereby improving its ability to generalize and perform well on unseen data. Additionally, we conducted experiments with a plain decision tree lacking hyperparameters or regularization. While this approach yielded acceptable results, the grid search method demonstrated superior performance, albeit requiring higher computational resources and time due to its exhaustive search strategy.

H. Hybrid random forest linear model-HRFLM

In our analysis, we explored hybrid random forest linear models (HRFLM) in addition to Logistic Regression. This comparative analysis was conducted to determine the most suitable model for our dataset, considering the number of features and the complexity of the problem.

We conducted the comparison between Logistic Regression and HRFLM in our analysis: First, we trained a Random Forest Classifier on the training dataset and generated predictions on the test dataset. Then, we calculated the residuals by subtracting the Random Forest predictions from the actual target values and trained a Linear Regression model on these residuals. Next, we combined the predictions from both models to form predictions for the HRFLM. Finally, we evaluated the performance of the HRFLM by converting the continuous predictions into binary predictions using a threshold of 0.5 and calculated relevant evaluation metrics such as accuracy score and classification report. This approach allowed us to compare the performance of the HRFLM with other models and gain insights into the efficacy of different modeling techniques for our dataset.

HRFLM performed on par with the Logistic Regression model, particularly due to the limited number of features in our dataset. However, according to recent research findings, HRFLM tends to excel when dealing with datasets containing a large number of features. In such scenarios, this model has been observed to outperform other models such as Xgboost, LightGBM, and traditional random forest [1].

Given our dataset's characteristics, where the number of features was not extensive, the performance of HRFLM was comparable to Logistic Regression. However, it's important to note that in scenarios with more features, HRFLM may exhibit superior performance and could potentially outperform other advanced models. Therefore, the choice of model should be tailored to the specific characteristics of the dataset and the objectives of the analysis.

Data Analysis

Based on the data analysis some of the following trends were noticed:

- Age emerges as a significant factor influencing heart attack risk, underscoring the importance of age-related interventions and screenings for early detection and management of heart attack.

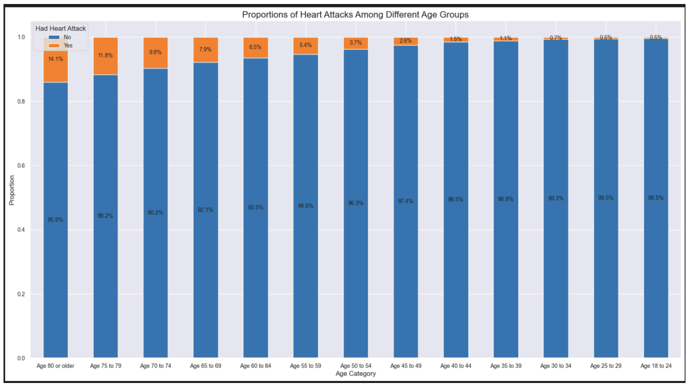

Fig 1. The proportion of Heart Attacks Among Different Age Groups

- The data suggests a notable increase in the occurrence of heart attacks among individuals aged 80 and above. Examining the proportion table reinforces this trend, indicating a higher likelihood of experiencing a heart attack with advancing age.

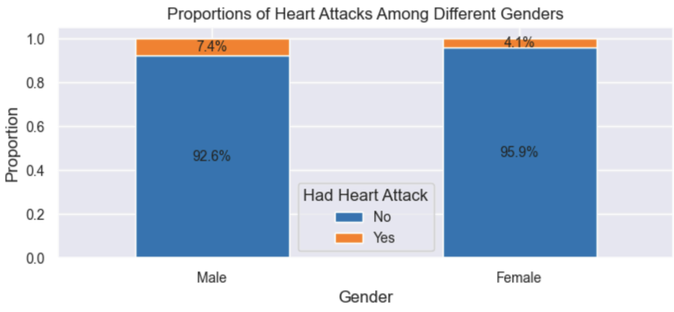

Fig 2. The proportion of Heart Attacks Among Different Gender

- Based on the above figure, the data reveals a clear disparity, with a higher prevalence of heart attacks observed among males in contrast to females.

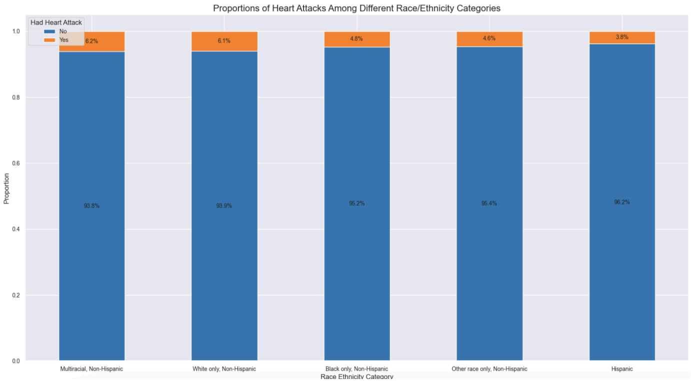

Fig 3. The proportion of Heart Attacks Among Race Category

- Also from the figure in the given sample population, multiracial individuals have the highest proportion of heart attacks (6.2%), followed by white individuals (6%), black individuals (4.8%), and other race individuals (4.6%). Hispanic individuals have the lowest proportion of heart attacks (3.8%).

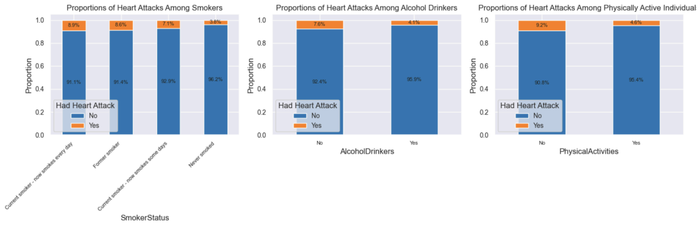

Fig 4. The proportion of Heart Attacks Among Smokers, Alcohol Drinkers, and Physically Active individuals.

From Figure 4, the following conclusions can be made:

- Former smokers and current daily smokers exhibit a significantly higher proportion of heart attacks.
- Individuals who consume alcohol have a higher likelihood of experiencing heart attacks compared to non-drinkers.
- Physically inactive individuals are at a heightened risk of heart attacks.

Similarly based on the EDA, the following analyses were made;

- Some chronic conditions exhibit weaker associations with cardiovascular health disparities.
- Individuals previously diagnosed with chronic conditions such as COPD (Chronic obstructive pulmonary disease), diabetes, obesity, and arthritis diseases are notably predisposed to heart diseases, with approximately 12%, 17%, 18%, and 30% respectively having co-existing heart conditions.
- Other conditions like kidney disease (7%), asthma (9%), and skin cancer (7%) also pose some risk of developing heart disease.
- Considering heart patients with pre-existing chronic conditions overall, nearly one-third had arthritis, approximately 17-18% were diabetic, and another 18% were obese.
- Almost 50% of the people suffered Angina pain before a heart attack.
- Significant people had a stroke before a heart attack.
- The correlation is evident: individuals with poorer overall health are at a higher risk of experiencing a heart attack.

- As individuals age, they tend to report an increase in the number of days with poor physical health over the past 30 days. Conversely, there is a trend of experiencing fewer days with poor mental health as age increases. Furthermore, older individuals, particularly those aged 65 and above, generally sleep slightly more hours on average.
- Individuals who had a depressive disorder are more likely to have had a heart attack compared to those who did not have a depressive disorder. About 7.03% of individuals with a depressive disorder had a heart attack, while only 5.32% of those without a depressive disorder had a heart attack.
- 15% of people who have difficulty walking have had a heart attack.
- 13% of blind and deaf people are heart patients.
- 9% of people who have difficulty concentrating had a heart attack.
- Sensory impairment appears to exert a notable influence on heart attack prevalence, with a percentage higher than the 6% observed in the total sample population.
- Individuals with a BMI greater than 30, indicating obesity, demonstrated a higher likelihood of having a heart attack based on the analysis of the dataset.

Evaluations

A. Performance Metrics

For each model developed from the datasets, we utilized various performance metrics to assess their predictive power and effectiveness:

• Logistic Regression: In evaluating the logistic regression model for heart attack prediction, we employed metrics such as confusion matrix, accuracy, precision, and recall. These metrics provided insights into the model's ability to correctly predict heart attack, considering the imbalanced nature of the dataset. The logistic regression model achieved an accuracy score of 94.54%. The classification report revealed high precision (0.95) and recall (0.99) for class 0 (non-heart disease), indicating strong predictive performance. However, precision and recall for class 1 (heart disease) were comparatively lower (0.57 and 0.24, respectively), resulting in a lower F1-score of 0.33 for class 1.

• Decision Tree Classifier with Hyperparameters and Regularization: Similarly, the decision tree classifier with hyperparameters and regularization was assessed using metrics such as confusion matrix, accuracy, precision, and recall. These metrics allowed us to evaluate the model's performance in comparison to logistic regression and identify any differences in predictive power. Additionally, we examined the impact of hyperparameters and regularization techniques on the decision tree's performance. The decision tree model with optimized hyperparameters attained an accuracy score of 94.28%. Although the model demonstrated improved precision and recall for class 1 compared to the simple decision tree, with precision of 0.52 and recall of 0.17, the F1-score remained relatively low at 0.26 for class 1.

• Regular Decision Tree: We also evaluated the performance of the regular decision tree model using the same set of metrics. Despite its simplicity and lack of hyperparameter tuning or regularization, this baseline model served as a point of comparison to assess the effectiveness of more complex approaches. The baseline decision tree model yielded an accuracy score of 90.68%. While the precision and recall for class 1 were higher compared to the decision tree with hyperparameters, with precision of 0.25 and recall of 0.29, the overall performance of the model was lower, as reflected by the lower accuracy score and F1-score for class 1.

• Hybrid Random Forest Linear Model(HRFLM): In evaluating the HRFLM for heart attack prediction, we obtained a detailed classification report showcasing precision, recall, and F1-score for both classes. The report indicated that while the model demonstrated high precision for class 0 (non-heart attack), its recall for class 1 (heart attack) was relatively lower, resulting in a lower F1-score for class 1. Additionally, the accuracy score provided an overall measure of the model's predictive performance, indicating a high accuracy of 94.45%. Furthermore, we visualized the model's performance using a confusion matrix, which highlighted the distribution of correct and incorrect classifications across the two classes.

Conclusion

Based on the exploratory data analysis (EDA) conducted on the dataset, several key insights have emerged regarding the prevalence and risk factors associated with heart disease:

- Through the analysis, it became evident that certain variables significantly influence the likelihood of heart disease. These include obesity, angina pain, depressive disorder, smoking, arthritis, diabetes, and lack of physical activity. These findings provide valuable insights into the factors that contribute to heart disease and can inform targeted interventions and preventive measures.
- Almost 70% of individuals in the dataset exhibit at least one of the major risk factors for heart attacks. This highlights the significant prevalence

of heart disease risk factors within the population and underscores the importance of addressing these factors to prevent heart attacks.

- Approximately 45% of individuals possess one of the three major risk factors for heart attack, further emphasizing the widespread presence of cardiovascular risk factors in the population.
- Furthermore, the correlation between these risk factors and existing research from reputable sources, such as the Better Health Channel, reinforces the validity of the findings. Age emerges as a significant factor, aligning with established knowledge about the relationship between age and cardiovascular health [3].

Overall, the findings from this analysis not only contribute to our understanding of heart disease risk factors but also have practical implications for improving preventive healthcare interventions and promoting better cardiovascular health outcomes. Through the integration of data-driven approaches into wearable technologies, we can revolutionize personalized health monitoring and empower individuals to take proactive steps towards better heart health.

Resources

[1] M. B. N. Nafouanti, J. Li, E. E. Nyakilla, G. C. Mwakipunda, and A. Mulashani, “A novel hybrid random forest linear model approach for forecasting groundwater fluoride contamination,” Environmental science and pollution research international, https://pubmed.ncbi.nlm.nih.gov/36800089/ (accessed Apr. 28, 2024).

[2] A. R, “Scikit-Learn Solvers explained,” Medium, https://medium.com/@arnavr/scikit-learn-solvers-explained-780a17bc322d#:~:text=It’s%20a%20linear%20classification%20that,coordinate%20directions%20or%20coordinate%20hyperplanes. (accessed Apr. 28, 2024).

[3] “Heart disease - risk factors,” Better Health Channel, https://www.betterhealth.vic.gov.au/health/conditionsandtreatments/heart-disease-risk-factors#risk-factors (accessed Apr. 28, 2024).

[4] K. Pytlak, “Indicators of heart disease (2022 update),” Kaggle, https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease/data (accessed Apr. 28, 2024).