Cardiovascular Analysis

Likhithasree Kommineni
Kyle Kane
Anto Lourdu Xavier Raj Arockia Selvarathinam

Kamrul Hasan
22 April 2024

Acknowledgements

We would like to express my sincere gratitude to all those who have contributed to the completion of this project.

First and foremost, We deeply thankful to my professor Kamrul Hasan for their guidance, support, and valuable insights throughout the duration of this project. Their expertise and encouragement have been invaluable in shaping the direction of the research and overcoming various challenges.

We extend my appreciation to my colleagues and peers for their collaboration, feedback, and constructive discussions. Their input has enriched the quality of the work and fostered a stimulating academic environment.

Last but not least, We are deeply indebted to my family and friends for their unwavering support, understanding, and encouragement throughout this journey. Their love and encouragement have been my source of strength and motivation.

In conclusion, We are grateful to everyone who has played a part, no matter how big or small, in the completion of this project. Your support and assistance have been invaluable, and we are truly thankful for the opportunity to undertake this endeavour.

Introduction

Heart disease is one of the leading causes of death across the world. The general cause is related to atherosclerosis which is the buildup plaque in blood vessels. This leads to a decrease in bloodflow throughout the body which can cause heart attacks. Overall, heart disease is a relatively well understood topic and as such multiple other factors can be attributed to an increased risk of heart disease. Some of these include

blood pressure, age, smoking, activity level, weight, and alcohol intake. While these are important to understand further information should be gathered as to which cofactors have the most influence and which of these are most linked to actual death due to heart disease.

Methods

At the beginning of this project, we obtained 2 separate datasets which looked at cofactors and heart disease. However, due to the difficulty merging them we decided to pick a separate data set that would work better for our analysis. This dataset was sourced from Kaggle and contained 253,680 patients. These patients were measured for a variety of variables including high blood pressure, high cholesterol, cholesterol check, BMI, smoker, stroke, diabetes, physical activity, fruits, veggies, alcohol consumption, healthcare, no doctor, mental health, physical health, sex, age, education and income.

Data Extraction and mining

The datasets were initially merged using a standard merge technique however since each column represented a unique patient, combining them with this method would incorrectly assign values to these unique patients. The next attempt was simply concatenating the data creating a significant number of null values. From there the null values were filled in. With columns that had few missing values the mean or median replaced the value. With the columns where there were many missing values a KNN imputer was used to find the nearest neighbor to the missing values. This method technically worked but practically did not. This is because there was a significant difference in the size of the two datasets. Out of the roughly 70,300 patients if values were imputed for 70,000 of them it would lead to significant variability to the overall value of the combining the datasets. Based on this information, the datasets were determined to be analyzed separately.

After combining the datasets and facing difficulty we ultimately decided to use the larger dataset to build our models and visualise our data.

Modeling

Logistic Regression Model:

This model is used for classification between 2 data factors.

K-Nearest Neighbor

This method is a non-parametric, supervised learning classifier. It uses distance to make predictions about an individual data point.

Decision Tree

This model has a hierarchical nature that uses a tree-like model of decisions and chance outcomes.

Random Forest

The random forest model uses random subsets of data to create multiple decision trees. It is a supervised machine learning algorithm and can be used for both classification and for regression. The reason for selecting random forest as a model was because of its advantages as a model. This model can incorporate a wide variety of datatypes and can also handle outlier and missing values without issue.

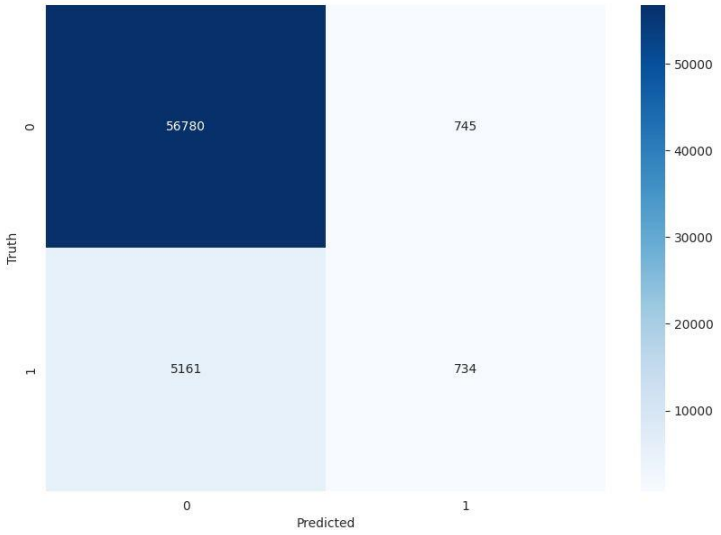
Neural Network

A neural network is an artificial intelligence method that is modeled after the human brain. The human brain includes nodes that are connected to each other, based on how often a particular network is used and it becomes strengthened. Similarly in the neural network model there are nodes which are layered. These layers include a input layer, hidden layers (can be one layer or multiple layers), and an output layer. These layers are linked similarly to neurons and weights can be applied to these links based on the strength of the connection. This is useful for detecting nonlinear relationships between variables as well as detecting all possible interactions between variables. These advantages make this model applicable for the datasets we have and what we intended to do with them.

Results from models

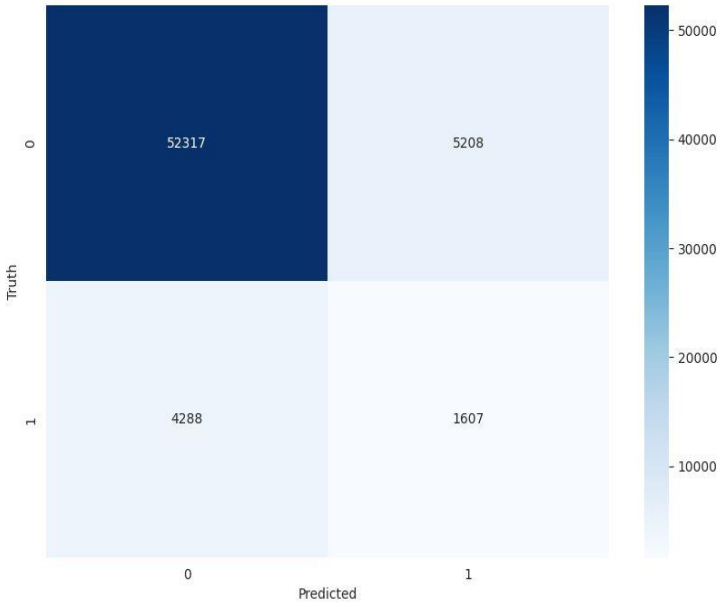
Logistic Regression Model

The results of the logistic regression model showed an accuracy of 0.91, precision of 0.88, recall of 0.91 and finally an F1 score of 0.88. Below is an image of the Confusion Matrix Heatmap:



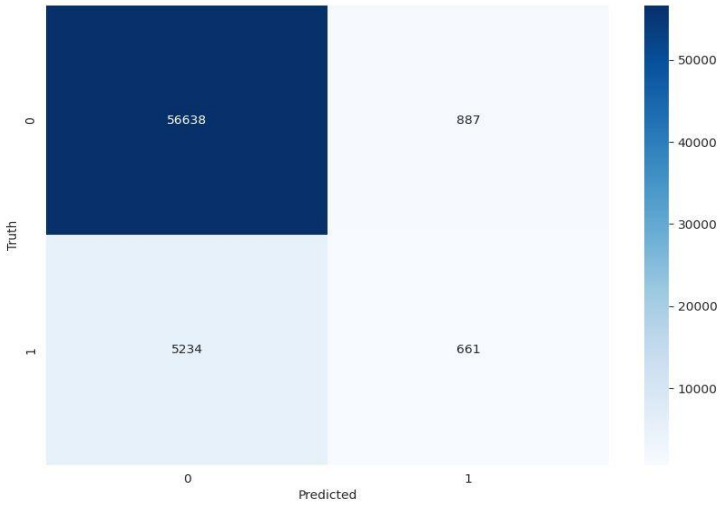
K-nearest neighbor

The results of the KNN model showed an accuracy of 0.895, precision of 0.86, recall of 0.895 and finally an F1 score of 0.88. Below is an image of theConfusion Matrix Heatmap:



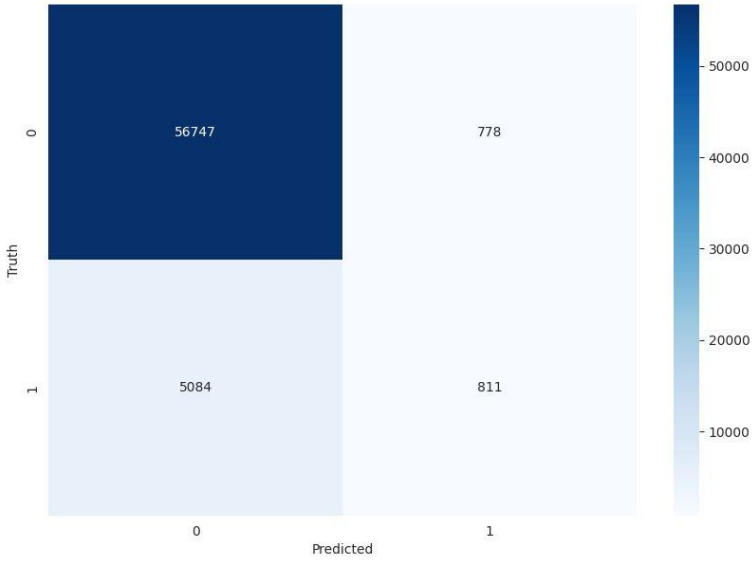
Decision Tree

The results of the decision tree model showed an accuracy of 0.85, precision of 0.86, recall of 0.85, and finally an F1 score of 0.86. Below is an image of the Confusion Matrix Heatmap:



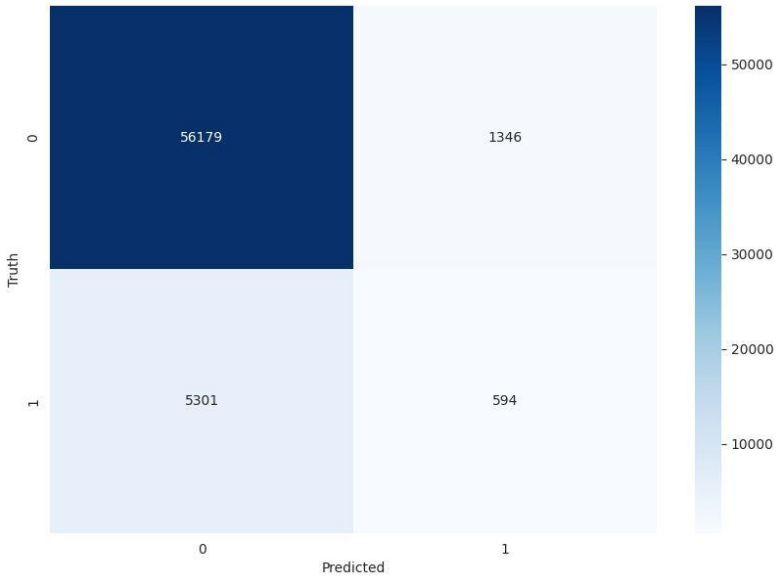
Random Forest

The results of the random forest model showed an accuracy of 0.90, precision of 0.86, recall of 0.90, and finally an F1 score of 0.88. Below is an image of the Confusion Matrix Heatmap:



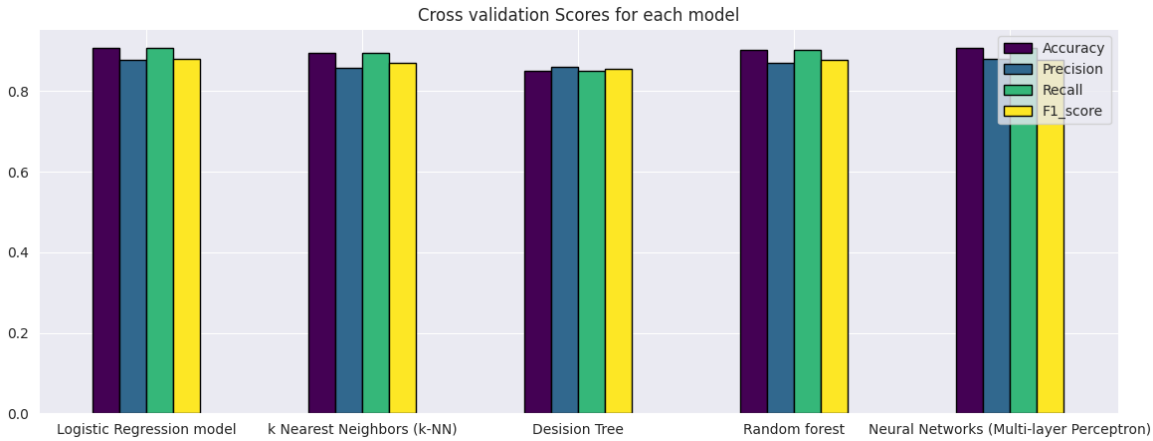
Neural Networks

The results of the neural network model showed an accuracy of 0.90, precision of 0.88, recall of 0.91, and finally an F1 score of 0.88. Below is an image of the Confusion Matrix Heatmap:



Results Table

	Accuracy	Precision	Recall	F1_score	Training time total
Logistic Regression model	0.906875	0.877602	0.906875	0.880711	4.572931
k Nearest Neighbors (k-NN)	0.895191	0.857300	0.895191	0.870479	62.281650
Desision Tree	0.851403	0.860722	0.851403	0.855905	1.531201
Random forest	0.903390	0.869399	0.903390	0.876579	22.671538
Neural Networks (Multi-layer Perceptron)	0.908199	0.878855	0.908199	0.876538	205.082504



Results Graph

Based on the results of the various models, all 5 models were successful at modeling the relationship between the given variables and cardiovascular events. The logistic regression model showed the highest accuracy out of the 5 models. The logistic regression model and neural network both had the highest precision and recall. The decision tree was the only model that had a slightly lower F1 score. Logistic regression and neural network had the most success modeling the dataset.

Discussion

With heart disease being one of the leading causes of death throughout the world successful modeling the causes is incredibly important. Based on the results the variables age, gender, smoking, systolic and diastolic blood pressure, alcohol consumption, activity level, weight, and blood glucose levels play a significant role in the progression of atherosclerosis and heart disease. These results also indicate a level of variability on whether heart disease will progress from person to person as well as potential indication for more variables to be incorporated as well. Potential variables that could be included in future research could include looking at if there is any genetic components that control progression of plaque buildup within the blood vessels which ultimately leads to cardiovascular events. Another potential variable could be how diet or cholesterol impact the progression of atherosclerosis. If all other things are equal perhaps consumption of more cholesterol may impact cardiovascular disease. In conclusion while this was a successful attempt at modeling heart disease risk factors, more research can still be done to continue improving the model and are deeper understanding of what impacts atherosclerosis and heart disease.