Cardiac disorder prediction using Statistical Machine Learning - Final Submission

Samridh Gupta(2022441)
Rachit Arora(2022384)

May 13, 2024

Contents

Chapter 1

Introduction and methodology

1.1 Overview

- Heart disease is one of the major causes of death in modern society
  - Therefore it is extremely important to have predictive models
  - In this project, we will use popular statistical machine learning techniques on multiple public datasets
    - To accurately classify those people who have heart diseases and those who don't
    - Using techniques like Bagging, Random Forests, QDA, etc
  - Using the results that we obtain,
    - We will draw conclusions as to which attributes majorly contribute to heart disease
    - And we will compare our results to popular academic papers in the area of heart disease, cardiology and heart attacks.
  - We will develop a small application that predicts accurately about the heart disease and heart attack risk of an individual.

1.2 Datasets used

- Kaggle: Heart Attack Risk Prediction dataset https://www.kaggle.com/datasets/iamsouravbanerjee/heart-attack-prediction-dataset
- Kaggle: Heart Disease Health Indicators dataset https://www.kaggle.com/datasets/alexteboul/heart-disease-health-indicators-dataset

1.3 Techniques used

- Our problem is a binary classification problem. Thus, we used the following techniques on each dataset:
  - **Decision Trees :** Decision trees are a popular supervised learning algorithm. They work by recursively splitting the dataset into subsets based on the feature that provides the best split according to the gini index of the splits.
  - **Random Forests :** Random forests are a learning method that constructs multiple decision trees during training and outputs the majority of the classes . Each tree in the forest is trained on a random subset of the features at each split.
  - **Adaboost :** Adaboost is a learning method that combines multiple decision trees to create a strong learner. It works by training a series of decision trees on weighted versions of the training data. In each iteration, the weights of misclassified instances are increased, due to which decision trees focus more on the miclassified points. The final prediction is a weighted sum of the predictions from all weak learners.
  - **LDA :** Linear Discriminant Analysis is a classification algorithm used for classification. It works by modeling the distribution of the input features for each class and then estimating the probability of each class given the input features. It assumes all classes have the same covariance matrix.
  - **QDA :** Quadratic Discriminant Analysis is like the LDA except that there is a different covariance matrix for each classes. This allows for a wider range of relationships between different data elements
  - **Bagging :** Bagging is a learning technique that builds multiple models using different subsets of the training data (some are repeated in order to make size of each bag same as the training set. Each model is trained independently using the entire feature set. It reduces variance and improves accuracy.We applied to try to increase overall accuracy.
  - **Logistic Regression :** Logistic Regression is a linear classification algorithm used for binary classification tasks. It models the probability that an instance belongs to a particular class using the sigmoid function.Logistic regression is just a single layer neural network.

Chapter 2

Dataset descriptions

2.1 Dataset 1

- This dataset consists of 25 attributes (24 if Systolic and Diastolic BP are in the same category)
  - Including but not limited to: age, sex, cholesterol, blood pressure, heart rate, diabetes, family history, smoking, obesity, alcohol consumption, exercise, BMI, sleep, income
  - Also includes possible geolocation data like continent and country
- It has about 8700 training samples
- We separated a random 1700 sample set as a val set for testing (0.2 fraction of the original dataset)
- A single binary attribute "Heart Attack Risk" is used for classification.

2.2 Dataset 2

- This dataset consists of 21 attributes
  - Including but not limited to: age, sex, high BP, high cholesterol, cigarettes per day, diabetes type, heavy alcohol consumption, income, etc.
  - Most attributes are discrete instead of continuous, making end user input simpler
- It has about 253k training samples We separated a random 50k sample set as a val set for testing (0.2 fraction of the original dataset)
- A single binary attribute "HeartDiseaseOrAttack", 1 if yes, 0 if no.

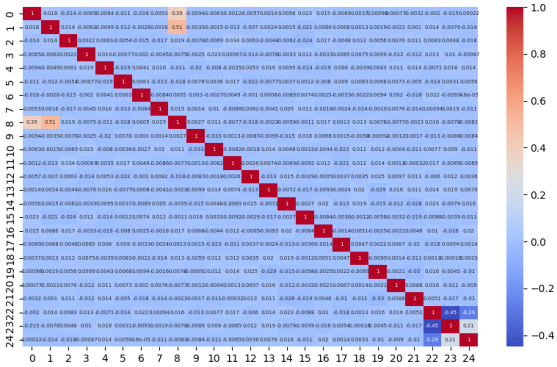

Figure 2.1: Correlation HeatMap for dataset-1

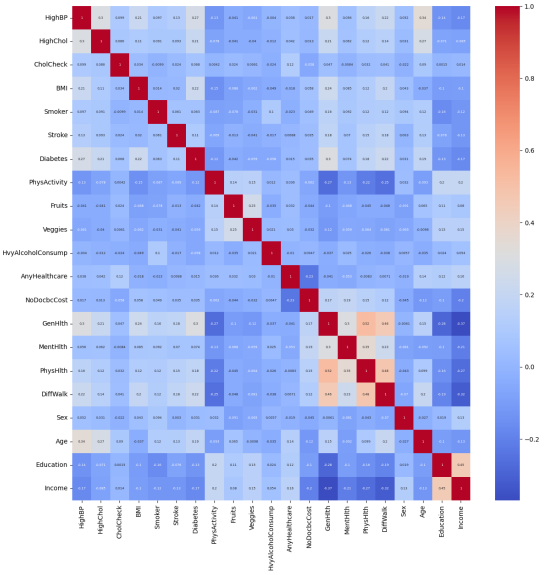

Figure 2.2: Correlation HeatMap for dataset-2

2.3 Final choice of dataset and model

- Since dataset 2 produced consistently better results than dataset 1, we are choosing dataset 2 as our final training set.
- Decision trees and logistic regression gave nearly identical accuracy in dataset 2.
  - We chose decision trees because it semantically makes more sense in a binary classification problem and its results can be easily explained to others.
  - It can also be easily visualized as we can see below.

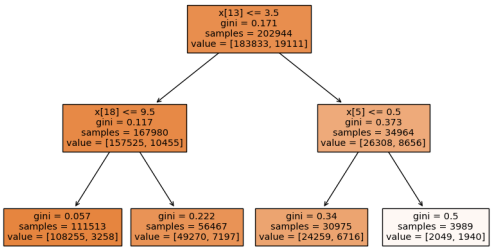

Figure 2.3: Decision tree for dataset-2

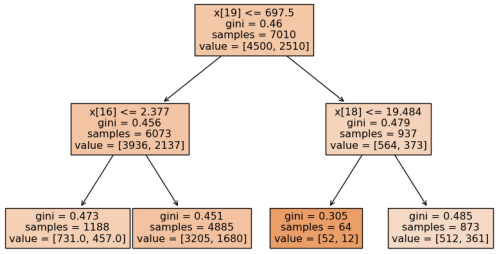

Figure 2.4: Decision tree for dataset-1

Chapter 3

Comparison to scientific literature

3.1 Factors related to physical health

- Our models have consistently shown that physical health factors like obesity (measured objectively using the BMI scale) and diet-related factors like cholesterol and blood sugar levels
- 2 of the most popular studies in cardiovascular health also agree with our conclusions
  - "Ischemic Heart Disease in Women: A Focus on Risk Factors" published in the United States' National Library of Medicine by Dr. Mehta et al in 2014, gives very remarkable results for women specifically.
    - For example, "a diabetic woman has 4-6 times more chance of developing narrow arteries". Our model showed that type 2 diabetes patients have a large risk of having heart diseases. Similarities also follow for other physical traits.
  - "The contribution of major depression to the global burden of ischemic heart disease: a comparative risk assessment" by Charlson et al is an objective meta-analysis of heart attack studies
    - This study also agrees with our conclusion of BMI and cholesterol levels being significant factors.

3.2 Factors related to substance abuse

- Alcohol consumption and smoking frequency is also a major factor, as determined by our model.

- 2 of the most popular studies in cardiovascular health also agree with our conclusions
  - Dr. Mehta's study (mentioned in physical health) of IHDs in women says:
    - "Alcohol intake has a J-shaped relationship with IHD risk in generally healthy populations, and healthy women are recommended to have no more than one drink per day on average. Excessive alcohol consumption is associated with hyperlipidemia, hypertension, vasoconstriction, hypercoagulability, and a lower ventricular fibrillation threshold"
    - This observation aligns with our view because we observed that most healthy people are not affected by heavy alcohol consumption, but those with other types of unhealthy factors suffer from it significantly.
  - "The contribution of major depression to the global burden of ischemic heart disease: a comparative risk assessment" by Charlson et al is an objective meta-analysis of heart attack studies
    - This study also agrees with our conclusion of smoking and alcohol being major factors.

Chapter 4

Results and Conclusion

4.1 Results

For Dataset 1:

Model	Accuracy
LDA	64.11%
QDA	61.20%
Decision Tree	64.46%
Random Forest	63.03%
AdaBoost Accuracy	64.11%
Bagging Accuracy	63.14%
Logistic Regression	64.11%

For Dataset 2:

Model	Accuracy
LDA	90.29%
QDA	83.68%
Decision Tree	90.78%
Random Forest	90.32%
AdaBoost Accuracy	90.73%
Bagging Accuracy	89.88%
Logistic Regression	90.81%

4.2 Dataset

The dataset used for this project can be found on Kaggle:
Kaggle: Heart Attack Risk Prediction Dataset
Kaggle: Heart Disease Health Indicators Dataset

4.3 Team Members

- Project Manager: Rachit Arora ,Samridh Gupta
- Software Engineer: Samridh Gupta
- Software Engineer: Rachit Arora

4.4 Concluding notes

- It is remarkable that we can make strong observations about popular issues like heart attacks using just statistical machine learning techniques.
  - And using academic literature which does not rely on statistics, we can reassure our accuracy and precision.
- Our study also presented general principles that one must keep in mind to prevent risk of heart disease:
  - Like following a good diet, avoiding alcohol and smoking, and consistently getting physical activity.
  - Even though this is common knowledge in the modern day, it is helpful to know that machines also agree
- Finally, we developed a small application for a person to check their own risk.

4.5 Bibliography

- Study by Dr. Mehta et al: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4336825/
- Study by Dr. Charlson et al: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4222499/