Heart Disease Analysis

DS 5110: Introduction to Data Management and Processing

Maalolan Bharaniraj, Ngoc Khanh Vy Le, Madhuri Krishnamurthy, Subhankar Shah

2024-04-21

1. Summary

Heart disease is a leading cause of death in the United States. According to the CDC, every 33 seconds, there will be one person in the US who dies due to cardiovascular disease. And every year around 800,000 people have heart attacks, of which 1 in 5 cases will be silent heart attacks. In addition, heart disease causes around $200 billion each year just from 2018 to 2019. This prompts our project's focus on leveraging data to tackle this critical health issue. Our project aims to predict the likelihood of heart disease, heart attack, and the mortality rate of heart disease based on different factors. By doing so, we can lay the groundwork for interventions and create plans and strategies to predict the risk of heart patients in the future.

We gathered data from different sources - Kaggle, UCI, and CDC, which have different information including demographic information, crucial indicators such as cholesterol levels, resting blood sugar, chest pain type and geography information. The data from Kaggle gave us information about the lifestyle and the medical history of patients who did and did not experience heart attacks including their lifestyle like exercise, sleep, smoking, underlying health, etc. The UCI data collected cardiac data from thousands of candidates who did and did not have heart disease. Apart from these datasets, we also used data from the CDC which contained stroke and heart mortality based on region, sex, and race.

Following the data science workflows, we imported and cleaned the dataset to ensure their suitability for analysis. We performed exploratory data analysis on the data to find the trends and primary contributors to heart disease. Reflecting on the inferences, we applied feature engineering techniques to simplify and help enhance the model performances. We splitted our data into training/testing and then built and trained our models - linear regression, logistic regression, neural network, and decision tree to predict the risk of heart disease based on the common factors we found in EDA. Furthermore, we calculated ROC curves, sensitivity, specificity, and F-1 scores in order to evaluate our models' accuracy. Finally, we observed and culminated the insights from the results from models and drew conclusions regarding the efficacy of our predictive models.

2. Methods

2.1 Data Processing

We imported and cleaned our datasets to prepare and ensure that our datasets were standardized, cleaned, and ready for further analysis which will help us easily identify and find meaningful insights into heart disease and related factors. Each dataset was cleaned differently but missing values were handled similarly to ensure consistency and clarity. Missing values in numeric columns are replaced with mean values while missing categorical values are replaced with mode values.

a. Kaggle : The creators mainly cleaned the dataset but we removed unnecessary columns like *PhysicalHealthDays*, and *MentalHealthDays* as their information was already expanded and detailed in other columns.

b. UCI : The datasets from UCI had different files that contained the data from different locations but they all conducted the same data. The initial dataset does not have columns' names so we renamed the columns of the dataset based on the documents provided by the source All numerical columns such as gender, chest pain type, fasting blood sugar, etc., are transformed into meaningful categories, and the values were also based on the documents.

c. CDC : The dataset had several similar columns that expressed the details of locations like lat, long, or descriptions and abbreviations of other columns. Those columns were removed since we just wanted to analyze the relationship between the locations and heart disease. Additionally, a *Region* column was added to categorize locations (Northeast, Midwest, South, and West) based on their states, this was useful for us to identify the mortality rates based on the locations. Lastly, we renamed *Data_Value* column to *Data_Value_Per_100000_Population* to make it clearly show where the value was from.

2.2 Feature Engineering

Various important transformations were used in feature engineering for heart disease analysis to understand the effects of different factors on heart health. Age binning was applied to distinguish between age groups and understand how varying age ranges impact the risk of heart disease. BMI binning was used to classify people according to their BMI and determine how it affected their risk of heart disease. To understand how sleep patterns affect heart health, sleep category binning was used. In order to determine the connection between vaccination and health status, a binary variable was developed using feature combination to indicate whether a person is vaccinated and whether they have any pre-existing conditions. The analysis of the contributions of these factors to heart disease was determined by the application of one-hot encoding, which transformed categorical variables into binary form. When combined, these methods improve the precision of the analysis of heart disease which leads to a more sophisticated comprehension of its causes.

2.3 Exploratory Data Analysis (EDA)

We performed EDA to help us better understand the data and identify the patterns, trends, and relationships between factors. We started by analyzing the relationship between age, sex, and heart disease to determine which demographic group is the most commonly affected. The graph in *Figure 1* confirms that most people having heart disease peak in old age, while those without heart disease are evenly distributed across ages. In addition, besides ages, we wanted to know how the region and the practices, environment, and habits that come with the region affect the probability of getting heart attacks. We plot bar graphs that display the total cases of heart disease per 100,000 population across different regions.

Figure 1: Age Distribution of Heart Disease Prevalence

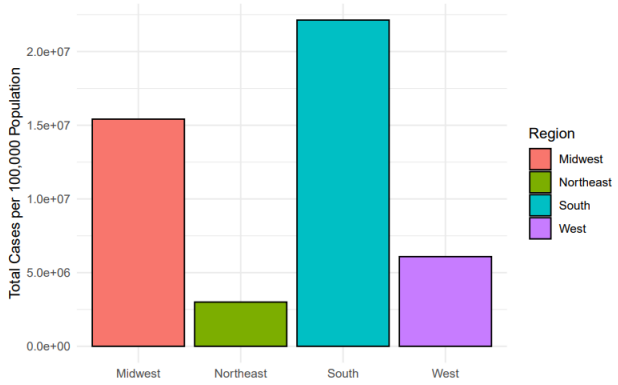
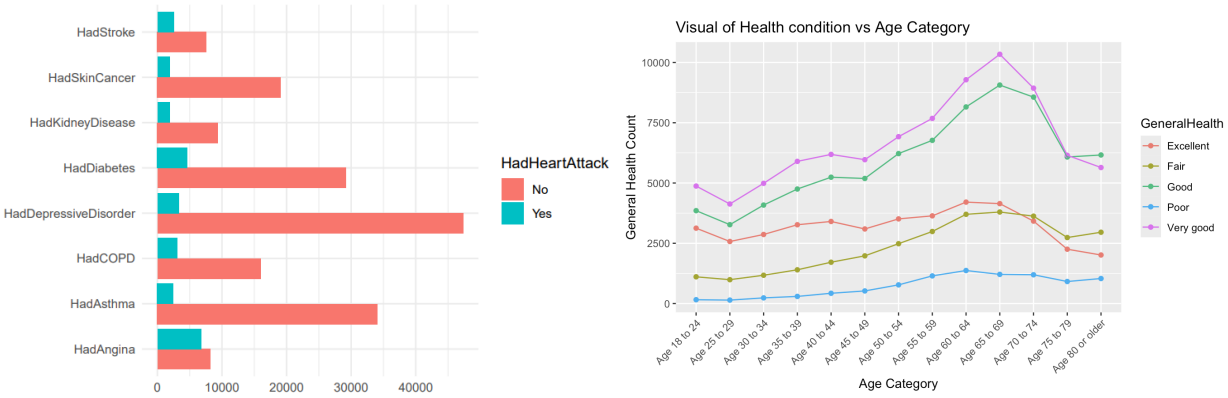
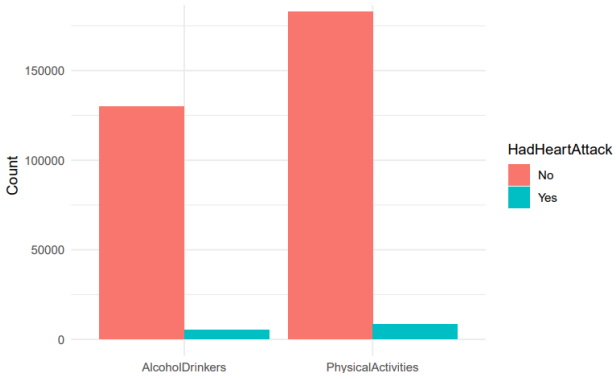


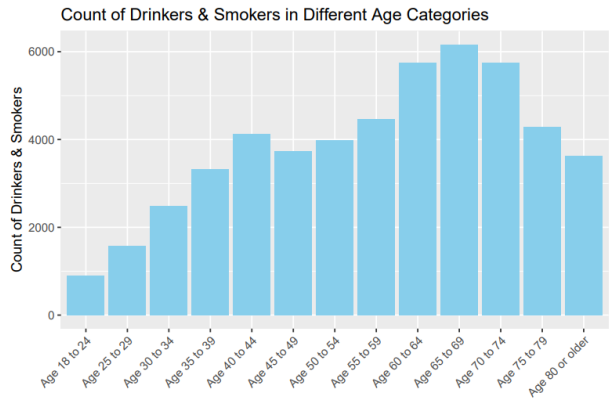
Figure 2: Heart Disease Mortality Cases in Different Regions

Figure 3 presents the medical history distribution of patients who did and did not experience heart attacks. We can conclude that various underlying health conditions are associated with heart attacks. The largest number of patients who had heart attacks were with Diabetes, Cancer, COPD, Asthma, and Depressive Disorder, indicating these are the most common conditions contributing to the risk of having heart attacks. The visualization in Figure 4 illustrates that individuals with the highest number of good, very good, and excellent general health fall under the age category of 65 to 69, which is surprising given they are among one of the most prone to heart disease age categories. This goes to show that symptoms, and external factors like diet, habits etc play a major role in the probability of getting a heart attack.



Furthermore, because Figure 4 shows that most people in group 65-74 have excellent health but *Figure 1* demonstrates that most people having heart disease peak in old age, we plotted boxplot (*Figure 5*) to observer the lifestyle and to identify whether or not lifestyle also contributes to the risk of heart attack. The plot shows that most people who experienced heart attacks did not usually exercise, but surprisingly, the majority of patients who drank alcohol did not have heart attacks. To analyze further to see which age, drinking, and smoking affect the patients experiencing heart attacks, we plotted the histogram graph shown in *Figure 6*. The graph visualizes the count of drinkers and smokers across different age categories, highlighting higher prevalence among older age groups. This gives us another solid factor as to why individuals in the older age category are more likely to get heart attacks.





2.4 Data Preparation for Model Input

All the datasets were split into training and testing data - 80/20 since it is the best way to access the ability of the models into the new data. We used the training data to fit the model and testing to evaluate the accuracy of the model.

2.5 Feature Selection

We want to find the relationship between heart disease mortality rates and other factors in the dataset like regions, races, and sex. Since the data values for the mortality rate - *Data_Value_Per_100000_Population* - a continuous values, we plan to use linear regression in this dataset. We find the best predictors for the models based on BIC values (*Table 1*). We chose the model which has the least BIC, which is *Region + Sex + Age*

Table 1: BIC result for Linear Regression model

Models	BIC
Region	103.6074
Region + Sex	98.89635
Region + Sex + Age	91.56468

Additionally, EDA shows that all health factors contribute to the chances of having heart disease; therefore, for the models for logistic regression, we chose all the attributes from the datasets (general health, medical history, lifestyle, geographic and demographic information)

2.6 Models

a. Linear Regression

With the selected predictors mentioned in *Section 2.5*, we fit the training data into the model and predicted the relationship between the heart disease mortality rate with demographic and geographic information. In addition, we chose *MSE - mean squared error* - to evaluate the model since *MSE* as its conceptually straightforward and is one of the common ways to evaluate the predictive models. The results and metrics of this model are discussed in the Result section (Section 3).

b. Logistic Regression

We applied logistic regression models to predict the risk of having heart disease using datasets from Kaggle and UCI, which are called *Model 1* and *Model 2*, respectively, since dependent values in both of the datasets are encoded into binary. *Model 1* has *have_heart_disease* as the dependent variable, while *Model 2* uses *HadHeartAttack* as a dependent variable. We used different health factors as predictors in *Model 1* but in *Model 2*, we also included lifestyle factors.

In addition to evaluating the accuracy of the model, we calculated the area under ROC curves, sensitivity, specificity, and F-1 scores. Each of these metrics gives different insights. The sensitivity to measure where missing out on true positives has severe consequences, while specificity helps to measure false positives. The ROC curves show the performance of a classification model at all classification thresholds and F-1 helps us not overlook the importance of either precision or recall. The results and metrics of this model are discussed in the Result section *(Section 3)*.

c. Decision Tree

We used Decision Tree to predict and support the logistic regression model for the prediction of heart attack risk since EDA overall does not show us which are the important factors that cause heart attack, and the Decision Tree model can handle the complex and non-linear relationship between features and offers us excellent interpretability. The results and metrics of this model are discussed in the Result section *(Section 3)*.

d.Neural Networks

We wanted to know how the neural network model performed when we trained the model with the same parameters. The neural network model was trained to predict if the person will have a heart attack in the future when given the various variables. The results and metrics of this model are discussed in the Result section *(Section 3)*.

3. Results

Based on the analysis from our EDAs, we can conclude that demography and underlying health are the primary factors contributing to the chances of having heart disease and heart attack. In addition to age, sex, and medical history, smoking and alcohol consumption in older age and not doing exercise regularly also contribute to the probability of having heart attacks. Furthermore, our analysis shows that there is a link between the demographic and geographic factors in the mortality rates of heart disease, as *Figure 2* illustrates that the South region stands out with a significantly higher number of mortality cases.

To confirm the hypothesis we conclude from EDAs, we then created and compared 5 different models - 2 logistic regression, 1 linear regression, 1 neural network, and 1 decision tree to predict the risk of having heart disease, heart attack, and mortality rate which are based on different factors found from EDAs.

3.1 Logistics regression model

As for the logistic regression models, we conducted two different models - *Model 1 and Model 2*, respectively, based on two datasets UCI and Kaggle to predict the chances of having a heart attack and heart disease (*Table 2*). Model 1 performs well in specificity and F1- score with a decent accuracy, 76.77%. F1-score - above 80% shows us that there is a good balance between precision and recall indicating Model 1 is fairly reliable to use to predict the risk of heart disease with just based on different health factors.

Table 2: Accuracy, Sensitivity, Specificity, F1, and AUC scores of logistic regression models

	Accuracy	Sensitivity	Specificity	F1-score	AUC
Model 1	0.7677	0.6429	0.8596	0.8099174	0.7799

	Accuracy	Sensitivity	Specificity	F1-score	AUC
Model 2	09488	0.9896	0.2423	0.3404812	0.9027

In the linear regression model, as mentioned in *section 2.5*, we chose *Region, Sex, and Age* as predictors for the model and fit in the training data. However, we got the *MSE* values of 8384.09 which is too high, and R-square values of 0.2539, which is too low (*Table 3*). It suggests that the model lacks predictive strength, more data and variables may be needed to have a better model.

3.2 Decision Tree Model

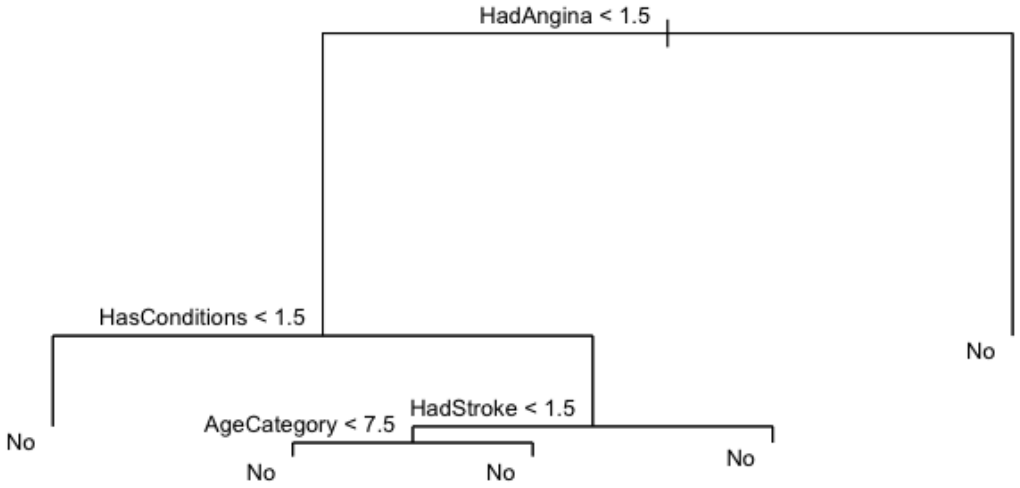


Figure 3: Decision Tree Model tree structure

The root node splits on the *HadAngina* variable. Left branch - *HadAngina* < 1.5 and right branch - *HadAngina* > 1.5. We further see that the left branch then further splits on the *HasConditions* variable, with those having *HasConditions* < 1.5 being classified as "No" (no heart attack) and those with *HasConditions* > 1.5 splitting again on the *HadStroke* variable. The right branch with *HadAngina* > 1.5 is classified as "No" (no heart attack). The tree has 5 terminal nodes, representing the final predictions.

The variables used in the tree construction are *HadAngina*, *HasConditions*, *HadStroke*, and *AgeCategory*. This indicates these variables are the most important predictors in the model.

3.3 Neural Network Model

We wanted to know how the neural network model performed when we trained the model with the same parameters. The neural network model was trained to predict if the person will have a heart attack in the future when given the various metrics mentioned above.

The neural network model accurately predicts if the person will suffer from a heart attack or not. The accuracy of the model is 94.53%

4. Discussions

5. Statement of Contributions

	Maalolan Bharaniraj	Ngoc Khanh Vy Le	Madhuri Krishnamurthy	Subhankar Shah
Data Processing		✓	✓	
DExploratory Data Analysis	✓	✓		
Feature Engineering			✓	
Data Preparation + Feature Selection		✓		
Linear Regression		✓		
Logistic Regression		✓		✓
Decision Tree + Neural Networks				✓
Model Evaluation + Insights		✓		✓
Project Report	✓	✓	✓	✓

6. References

1. Dataset:

- Heart Disease Mortality Data Among US Adults (35+) by State/Territory and County – 2016-2018 https://catalog.data.gov/dataset/heart-disease-mortality-data-among-us-adults-35-by-state-territory-and-county-2016-2018-c0d58
- Indicators of Heart Disease https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease
- UCI - Heart Disease https://archive.ics.uci.edu/dataset/45/heart+disease

2. Paper references

- Detrano, Robert C. et al. "International application of a new probability algorithm for the diagnosis of coronary artery disease." The American journal of cardiology 64 5 (1989): 304-10.
- "Heart Disease." Centers for Disease Control and Prevention, U. S. Department of Health and Human Services, 19 Jan. 2021, https://www.cdc.gov/heartdisease/index.htm

7. Appendix

Screenshots of the three datasets used in the project :

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	have_heart_disease	age_group	bps_scaled[, 1]	cholesterol_group
1	67	2	1	160	286	2	1	108	2	1.5	2	3	2	yes	Elderly	0.800	High
2	67	2	1	120	229	2	1	Other	2	2.6	2	2	3	yes	Elderly	0.600	Borderline High
3	37	2	3	130	250	2	2	Other	1	3.5	1	0	2	no	Young	0.650	High
4	41	1	2	130	204	2	1	172	1	1.4	3	0	2	no	Middle-aged	0.650	Borderline High
5	56	2	2	120	236	2	2	178	1	0.8	3	0	2	no	Middle-aged	0.600	Borderline High
6	62	1	1	140	268	2	1	160	1	3.6	1	2	2	yes	Elderly	0.700	High
7	57	1	1	120	354	2	2	163	2	0.6	3	0	2	no	Middle-aged	0.600	High
8	63	2	1	130	254	2	1	147	1	1.4	2	1	3	yes	Elderly	0.650	High
9	53	2	1	140	203	1	1	155	2	3.1	1	0	3	yes	Middle-aged	0.700	Borderline High
10	57	2	1	140	192	2	2	148	1	0.4	2	0	1	no	Middle-aged	0.700	Normal
11	56	1	2	140	294	2	1	Other	1	1.3	2	0	2	no	Middle-aged	0.700	High
12	56	2	3	130	256	1	1	142	2	0.6	2	1	1	yes	Middle-aged	0.650	High
13	44	2	2	120	263	2	2	173	1	0.0	3	0	3	no	Middle-aged	0.600	High

Figure 4: UCI dataset

	Year	LocationAbbr	LocationDesc	GeographicLevel	DataSource	Class	Topic	Data_Value	Data_Value_Unit	Data_Value_Type
1	2017	AK	Aleutians East	County	NVSS	Cardiovascular Diseases	Heart Disease Mortality	172.9	per 100,000 population	Age-adjusted, Spatially Smoothed, 3-year Average
2	2017	AK	Aleutians West	County	NVSS	Cardiovascular Diseases	Heart Disease Mortality	172.2	per 100,000 population	Age-adjusted, Spatially Smoothed, 3-year Average
3	2017	AK	Anchorage	County	NVSS	Cardiovascular Diseases	Heart Disease Mortality	243.3	per 100,000 population	Age-adjusted, Spatially Smoothed, 3-year Average
4	2017	AK	Bethel	County	NVSS	Cardiovascular Diseases	Heart Disease Mortality	337.1	per 100,000 population	Age-adjusted, Spatially Smoothed, 3-year Average
5	2017	AK	Bristol Bay	County	NVSS	Cardiovascular Diseases	Heart Disease Mortality	*NA*	per 100,000 population	Age-adjusted, Spatially Smoothed, 3-year Average
6	2017	AK	Denali	County	NVSS	Cardiovascular Diseases	Heart Disease Mortality	240.9	per 100,000 population	Age-adjusted, Spatially Smoothed, 3-year Average
7	2017	AK	Dillingham	County	NVSS	Cardiovascular Diseases	Heart Disease Mortality	366.5	per 100,000 population	Age-adjusted, Spatially Smoothed, 3-year Average
8	2017	AK	Fairbanks North Star	County	NVSS	Cardiovascular Diseases	Heart Disease Mortality	276.7	per 100,000 population	Age-adjusted, Spatially Smoothed, 3-year Average
9	2017	AK	Haines	County	NVSS	Cardiovascular Diseases	Heart Disease Mortality	261.0	per 100,000 population	Age-adjusted, Spatially Smoothed, 3-year Average
10	2017	AK	Hoonah-Angoon	County	NVSS	Cardiovascular Diseases	Heart Disease Mortality	272.0	per 100,000 population	Age-adjusted, Spatially Smoothed, 3-year Average
11	2017	AK	Juneau	County	NVSS	Cardiovascular Diseases	Heart Disease Mortality	266.8	per 100,000 population	Age-adjusted, Spatially Smoothed, 3-year Average
12	2017	AK	Kenai Peninsula	County	NVSS	Cardiovascular Diseases	Heart Disease Mortality	284.2	per 100,000 population	Age-adjusted, Spatially Smoothed, 3-year Average

Figure 5: CDC dataset

	State	Sex	GeneralHealth	PhysicalActivities	SleepHours	HadHeartAttack	HadAngina	HadStroke	HadAsthma	HadSkinCancer	HadCOPD	HadDepressiveDisorder
1	Alabama	1	5	2	9	No	1	1	No	1	1	1
2	Alabama	2	5	2	6	No	1	1	No	1	1	1
3	Alabama	2	5	1	8	No	1	1	No	1	1	1
4	Alabama	1	2	2	9	No	1	1	No	2	1	2
5	Alabama	1	3	2	5	No	1	1	No	1	1	1
6	Alabama	2	3	2	7	No	1	1	No	1	1	1
7	Alabama	1	3	2	8	No	1	2	No	1	1	1
8	Alabama	2	2	2	8	Yes	2	1	No	2	1	1
9	Alabama	2	3	1	6	No	1	1	No	1	1	1
10	Alabama	1	5	2	7	No	1	1	Yes	2	1	1
11	Alabama	2	5	2	8	No	1	1	No	1	1	1
12	Alabama	1	3	2	5	No	1	1	No	1	1	1

Figure 6: Kaggle dataset