IND5022 Data Analytics for Smart Manufacturing

Team Project Part II
Analysis and Prediction of the Prevalence of Heart Disease Based on Multi-Features

Propose To:
Prof. Guo Lei

Proposed By:

Du Zhouyang	A0254065H
Mo Xixi	A
Tian Shuo	A0254073J
Wu Jinqian	A0250914J
Yang Leying	A0254381E
Zhang Wentao	A0232358E

Table of Contents:

1. Background

According to the World Health Organization, heart disease is the leading cause of death for both men and women worldwide, accounting for an estimated 32 percent of deaths, making it even more important to focus on and prioritize heart health. In 2017, heart disease was listed as the underlying cause of over 860,000 deaths in the United States. Each year, approximately 659,000 people in the United States die as a result of heart disease. This means that heart disease kills one out of every four people in the United States.

The increasing number of heart disease deaths caught the attention and interest of our team. Our team is committed to using data analysis and mechanistic learning to solve real-world problems. We decided to use data analysis and machine learning to analyze and solve potential problems inside the data of heart disease and work to contribute to global health by studying and attempting to reduce the likelihood of heart disease.

2. Data introduction:

Our dataset comes from Kaggle, collected by the US Centers for Disease Control and

Prevention (CDC) through telephone survey. There are nearly 300 variables in the original dataset. After processing, these 300 variables are reduced to about 20 variables, which is the dataset used by our team now. There are more than three hundred thousand respondents and eighteen variables in this dataset. The descriptions of variables are in the table 1.

Variables	Description
HeartDisease (boolean)	Respondents that have ever reported having coronary heart disease (CHD) or myocardial infarction (MI).
BMI (decimals)	Body Mass Index (BMI)
Smoking (boolean)	Have you smoked at least 100 cigarettes in your entire life?
AlcoholDrinking (boolean)	Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week).
Stroke (boolean)	Whether you had a stroke?
PhysicalHealth (decimal from 0 to 30)	How many days during the past 30 was your physical health not good, including physical illness and injury?
MentalHealth (decimal from 0 to 30)	How many days during the past 30 days was your mental health not good?
DiffWalking (boolean)	Do you have serious difficulty walking or climbing stairs?
Sex (string)	Gender
AgeCategory (string)	Fourteen-level age category
Race (string)	Including White, Hispanic, Black, Asian and Other.
Diabetic (boolean)	Whether you had diabetes?
PhysicalActivity (boolean)	Adults who reported doing physical activity or exercise during the past 30 days other than their regular job.
GenHealth (string)	Your health in general is excellent, very good, good, fair, and poor.
SleepTime (decimal from 0 to 24)	On average, how many hours of sleep do you get in a 24-hour period?
Asthma (boolean)	Whether you had asthma?
KidneyDisease (boolean)	Whether you had kidney disease not including kidney stones, bladder infection or incontinence?
SkinCancer (boolean)	Whether you had skin cancer?

Table 2.1. The descriptions of variables

3. Business questions:

3.1. According to the above characteristics, establish a prediction model to predict whether the patient has heart disease. (prediction)

3.1.1. Flow Chart

The diagram below shows the process of our project, which we will discuss in detail in part 4 and 5.

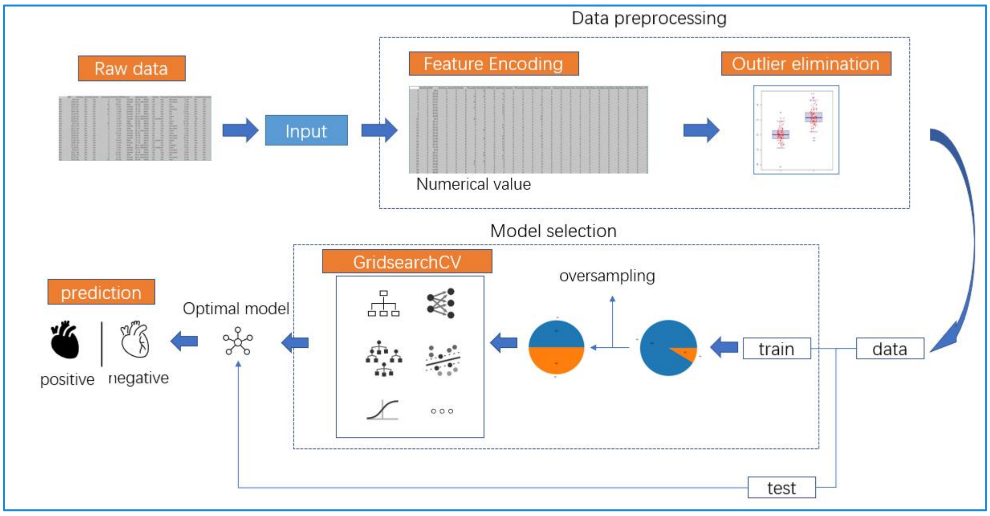

Figure 3.1 Flow Chart

3.1.2. Data Preprocessing

On the one hand, data preprocessing is to improve the quality of data, on the other hand, it is also to adapt to the software or methods of data analysis. The data preprocessing process includes the following steps.

(1) Data cleaning

In the process of data cleaning, we need to check if there are some missing values and outliers, and duplicates in the dataset and remove them. In this data, only BMI and SleepTime have outliers. We use the box chart method to filter the outliers. The box chart can be programmed to set a standard for identifying outliers, that is, the values greater than or less than the upper and lower bounds set by the box chart are identified as outliers.

$$\text{Upper Limit} = Q3 + 1.5 \times IQR \tag{1}$$

$$\text{Lower Limit} = Q1 - 1.5 \times IQR \tag{2}$$

$$IQR = Q3 - Q1 \tag{3}$$

Sort the data from small to large. The median is 50% quantile, which is Q2. The 75%

quantile or the third quarter quantile is Q3, and the 25% quantile, or the quarter quantile is Q1. Upper Limit (1) is the maximum value in non-outliers and Lower Limit (2) is the minimum value in non-outliers. Values beyond this range are outliers.

(2) Data transformation

There are some text data in the original data, such as Yes, no, male and female. We need to convert the text data into numbers for analysis. There are three kinds of features of our dataset that need to be transformed. The first kind of feature is binary data, including HeartDisease, Smoking, AlcoholDrinking, Stroke, DiffWalking, PhysicalActivity, Asthma, KidneyDisease, SkinCancer and Dicabetic. These features are transformed into 1 and 0, where 1 represents Yes and 0 represents No. The second kind of features are Sex and Race. They are categorical variables. We use one-hot encoding to transform this kind of feature, that is, to split a feature into multiple features. For example, Sex will be split into male and female these two features. The third kind of features is AgeCategory and GenHealth. We use Label Encoder to transform them.

(3) Dealing with imbalance problems

The data suffers from serious imbalance problem. We decide to use RandomOverSampler to generate synthetic samples and solve the problem. Aftering dividing the data into train set and test set, we use RandomOverSampler to generate synthetic samples of the data in train set.

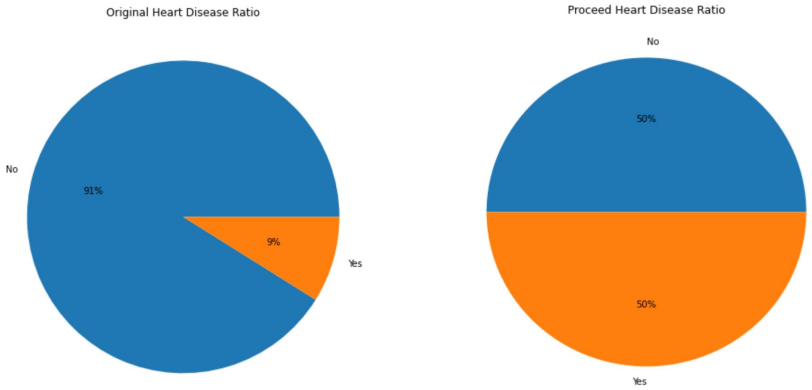

Figure 3.2. the original data label ratio and proceed data label ratio

(4) Multicollinearity check

Since we used models based on linear regression, we checked variance inflation factors of all features and find potential multicollinearity problem may exist in our data. Then, we removed features 'Sex_Female', 'Race_White', 'Race_Others' to effectively solve the multicollinearity problem. The vif value figure is shown in Figure 3.3.

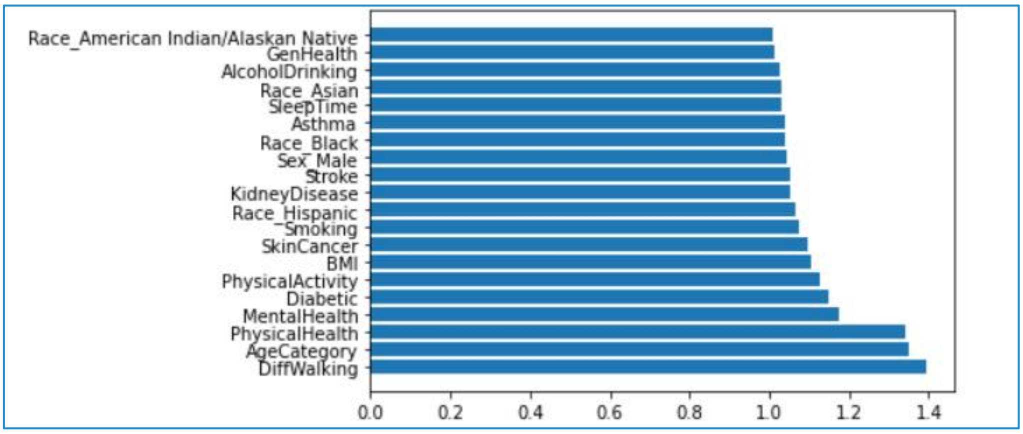

Figure 3.3. the vif of features after removing those contains duplicate information

3.1.3. Optimal Model Selection

After balancing the positive and negative samples of the train set, we selected seven classification models: Adaboost, Xgboost, MLP classifier, Logistic Regression, Linear SVC, Decision Tree, Random Forest. We use GridsearchCV from Sklearn to look for the optimal super parameters of some models. Grid search tries every specified parameter combination through loop traversal and uses cross-validation to verify these parameters. The best parameter combination is our final settings.

For evaluation metrics, we choose accuracy, recall, and AUC value as key evaluation metrics. The accuracy represents the ratio of how many samples are predicted correctly, and the recall means the proportion of how many true positive samples are predicted correctly overall positive samples. When we use the True Positive Rate, TPR (same as Recall) as the vertical axis and the False Positive Ratio, FPR which tells us the proportion of the negative class got incorrectly as the horizontal axis, we can draw a ROC curve from the different combination of TPR and FPR. The Area Under the Curve, AUC can reflect the ability of the classifier which is always between 0 and 1. The higher value means the classifier has better performance.

In this project, we mainly focus on the recall metric of heart disease. The recall denotes the proportion of how many heart disease patients are classified correctly. Because we hope our classifier has more reliability for the real patient.

In all, the evaluation result is shown in Fig below, since we choose recall as the most important reference, Adaboost is chosen as our final model.

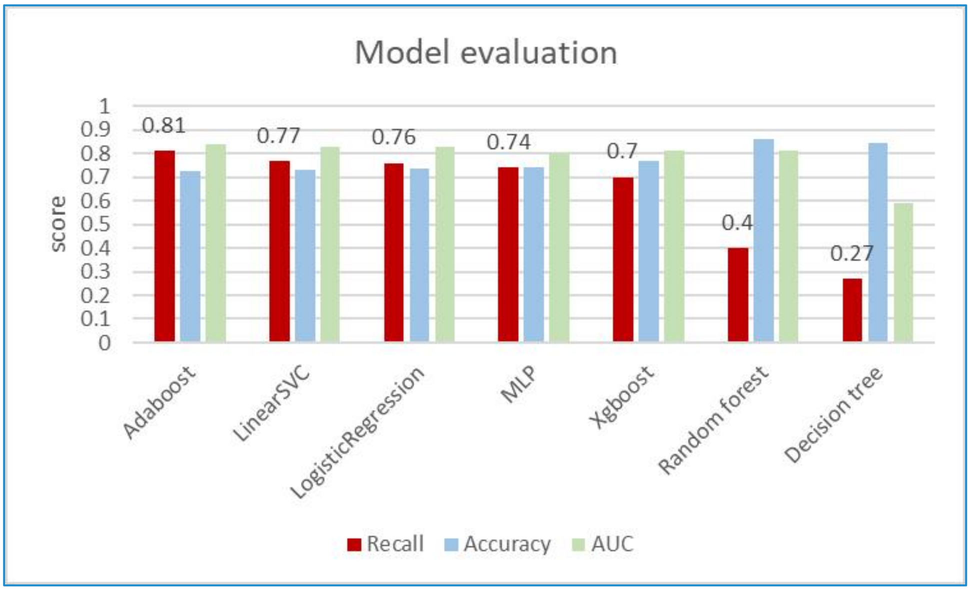

Figure 3.3. Recall, Accuracy score and AUC of seven models

3.1.4. Model Analysis (Limitation)

The most special characteristic of the original data is the high imbalanced ratio of negative and positive samples. If we train the model without any data preprocessing, the recall of positive samples is even lower than 0.1, meaning that most of the positive samples are classified as having no heart disease. The oversampling helps us create more than 200000 new samples in the training set, in exchange for the accuracy value to raise the recall score. Here is our model result.

	Precision	Recall	Support samples
Non heart disease (negative)	0.97	0.71	54807
Heart disease (positive)	0.22	0.81	5334

Table 3.2. Model result of Adaboost

As you can see, the precision of positive samples is very low for our model, meaning that many of the positive samples predicted are actually negative. However, precision is not the priority of our consideration, because compared with classifying healthy people as having heart disease, the consequence of classifying heart disease patients as healthy people is much worse.

Based on the result, we believe our model is useful in the reality. If we dignose a negative sample as result, it is convinced because of the 0.97 high precision probability. On the other hand, if we find a positive predicted result, it means there will be about 20 percent chance for the tested person to have heart disease, then we will suggest he or she to look for professional advice. Our model has 0.81 recall rate which means most of patient can be diagnosed correctly.

3.2. How can people reduce the risk of developing heart disease?

3.2.1. Analysis of external risk factors

In this section, we analyze how the risk factors may be related to having heart disease, based on our chosen optimal model. The figure below shows the ranking of feature importance given by our optimal model.

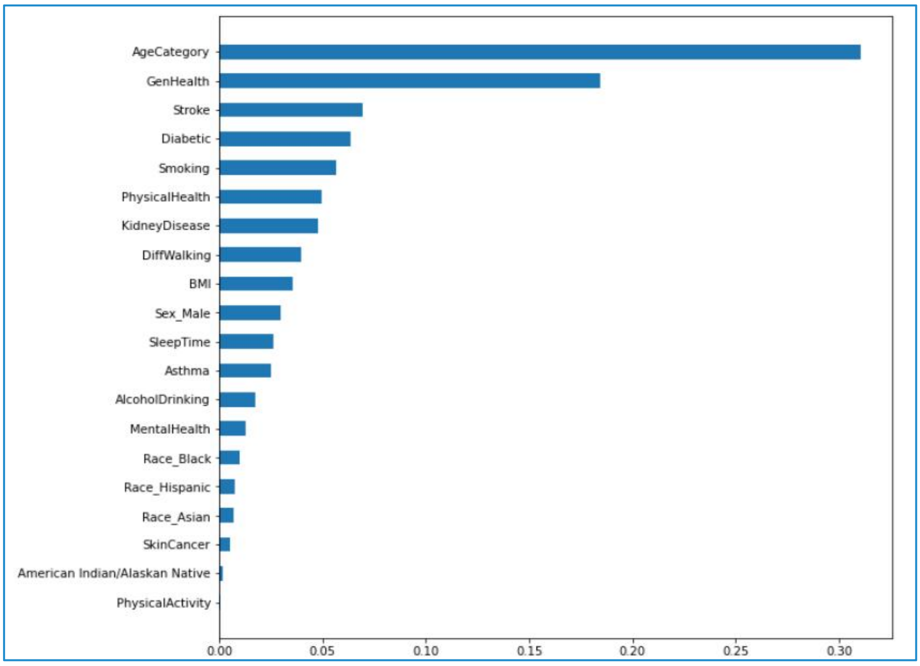

Figure 3.4. Ranking of feature importance given by Adaboost

As is shown, the most related 5 external risk foctors given by our model are ‘Stroke’, ‘Diabetic’, ‘Smoking’, ‘KidneyDisease’, and ‘DiffWalking’. To give a deeper insight, We analyze if an individual has more than 1 risk factor, what is the likelihood that he has heart disease.

We group our external risk factors according to the relationship between different medical symptoms or diseases, for example, stroke and DiffWalking are grouped together because many stroke patients may have difficulties in walking. The result of the statistics is shown in the figure below.

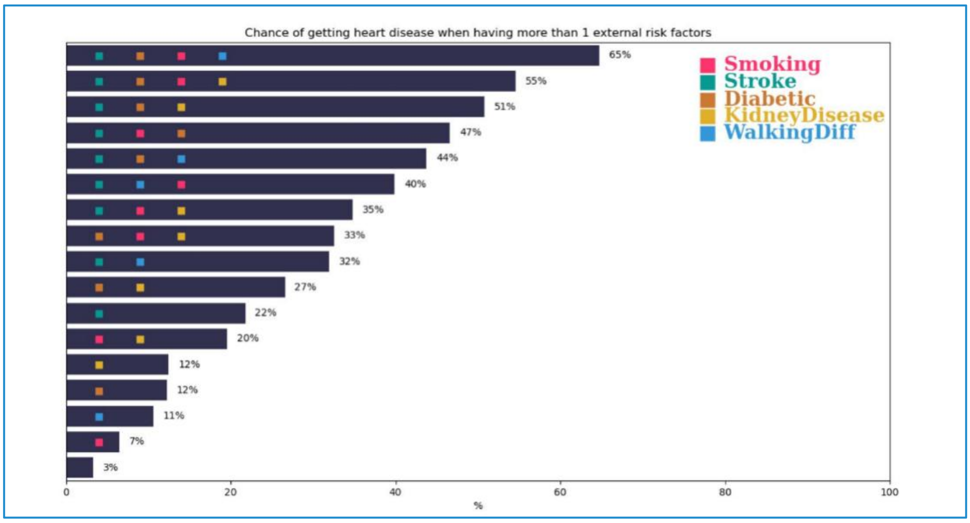

Figure 3.5. Top 5 external risk factors

Our model takes these as the 5 most external risk factors for getting heartdisease. When people have none of these risk factors, the probability of having heart disease is only 3%. Moreover, the visualization shows stroke to be the most effective external factor for getting heart disease, which is consistent with the result from our Adaboost classification model.

3.2.2. Analysis of Human Controllable Factors

Based on the given features, we apply one of the dimension reductions, which is feature selection. We manually selected 5 features out of 23 features based on whether the feature can be controllable by human behaviors, which are Alcohol Drinking, BMI, Sleep Time, Smoking and Physical Activity.

Following that, we processed the data further. The BMI, for example, was divided into five groups, which can be referred to as following chart.

BMI	Status
BMI<18.5	Underweight
18.5≤BMI＜23.9	Normal
23.9≤BMI＜27	Overweight

27≤BMI＜32	Fat
32≤BMI	Overfat

Table 3.3. BMI groups

Subsequently, we divided the sleep duration into three groups, which can be seen in the following chart.

Sleep time	**Status**
Sleep time<6	Sleep deprivation
6≤Sleep time＜9	Normal sleep
9≤ Sleep time	Excessive Sleep

Table 3.4. Sleep time groups

Overall, we discovered that people of normal weight were the least likely to have heart disease, while people who were malnourished were the second least likely. The remaining section indicates that the greater the weight, the greater the risk of heart disease. When sleep duration was considered, people with normal sleep duration were less likely to have heart disease. Nonsmokers and people who exercise regularly are less likely to have heart disease, according to the other two characteristics. As a result, we advise people to try to maintain a healthy weight, get enough sleep, avoid smoking, and exercise regularly. Detailed graphics can be found in Appendix A.

The data on alcohol consumption that followed was surprising. This is because the study discovered that heavy drinkers are less likely to have heart disease, which defies logic. As a result, we considered other factors, such as sample data bias. Nondrinkers accounted for 93% of the total number of lookalikes, while drinkers accounted for 7% of the total number of drinkers, indicating a significant difference between the two lookalike data sets. As a result, the effect of whether or not to drink alcohol is not informative for this analysis, so we eliminated it.

4. Conclusion

In this project, we used Kaggle's data set to conduct heart disease related research. We mainly explored two problems. The first question is according to the above characteristics, establish a prediction model to predict whether the patient has heart disease. This is a prediction problems. Through data preprocessing and model selecting, we choose Adaboost as our prediction model because it has the highest

Recall. The Recall of Adaboost is 0.81. Then we use the output of Adaboost to solve the second question, How can people reduce the risk of developing heart disease? In this problem, we mainly use analysis of external risk factors and human controllable factors to solve it. Through our analysis, we recommend that people try to maintain a healthy weight, get enough sleep, avoid smoking, and exercise regularly. But strangely, the study found that alcoholics are less likely to suffer from heart disease, and we believe that this problem may be caused by the deviation of the data.

We believe that our project will assist more people in predicting and preventing heart disease. We will later develop a heart disease prediction website based on our prediction model and provide advice on heart disease prevention based on the user's specific situation, thereby assisting more people in lowering their risk of heart disease and detecting heart disease early so that they can actively treat it and reduce their risk of death from heart disease.

Appendix A

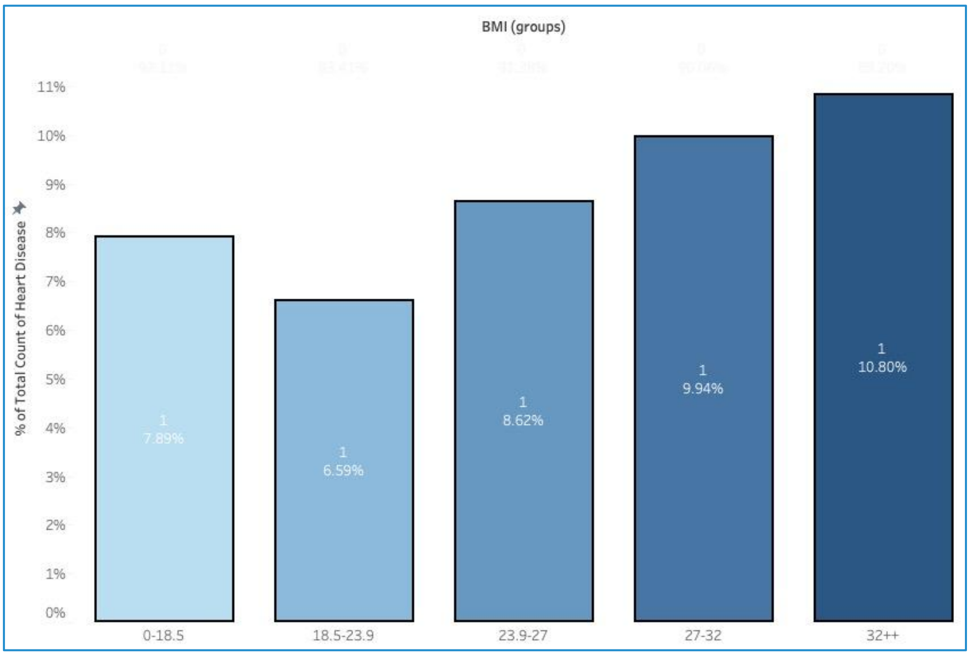

Figure 1. BMI

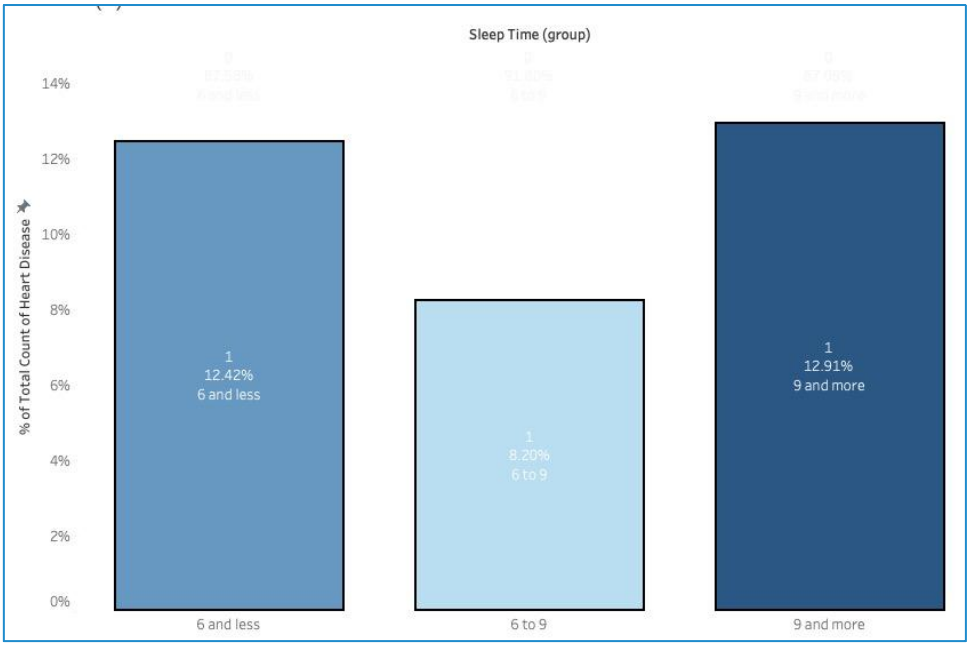

Figure 2. Sleep time

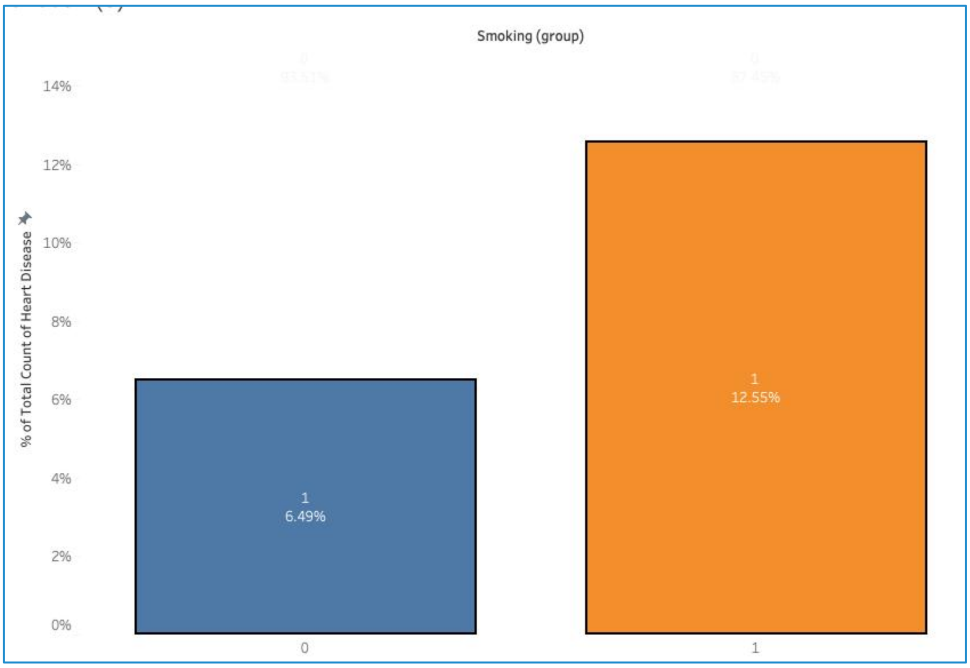

Figure 3. Smoking

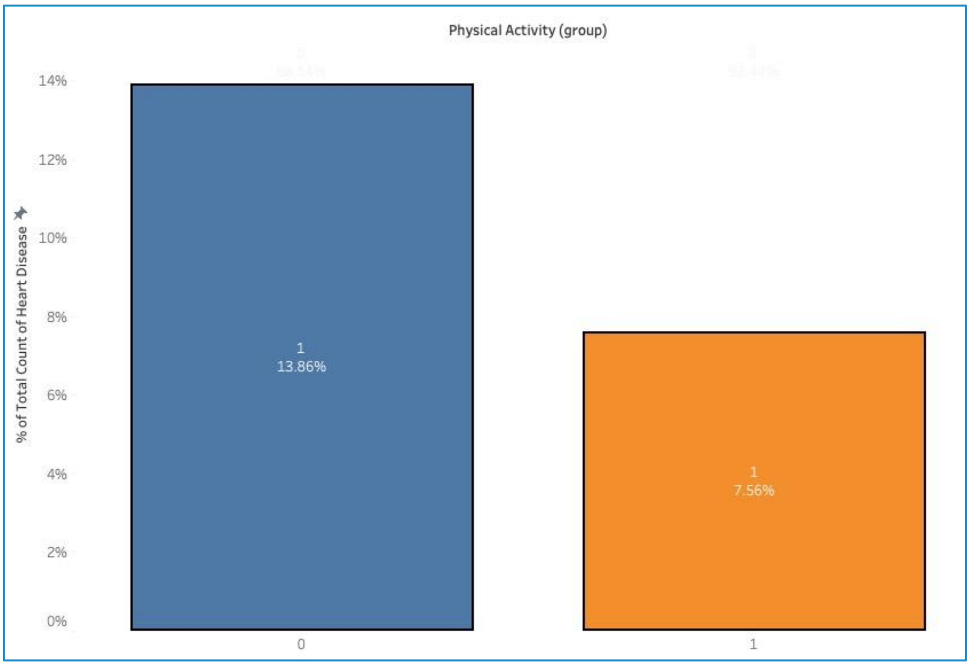

Figure 1. Physical Activity

References

[1]kaggle dataset:

https://www.kaggle.com/code/ahmedmohsen2002/heart-disease-eda-business-questions

[2] https://www.cdc.gov/heartdisease/facts.htm