Project Proposal for Data Mining

Title Data: Heart disease prediction.

Group member: Anibal Ruiz

**Intro**

For this project I have chosen data about heart disease prediction because in the last months I’ve been trying to change my lifestyle and be healthier, and having good habits are probably the most important things we need to improve to accomplishes that change. The data is proposed to help in understanding the relationship between some attributes and having a heart disease. According to the CDC, heart disease is one of the leading causes of death for people of most races in the US (African Americans, American Indians and Alaska Natives, and white people). About half of all Americans (47%) have at least 1 of 3 key risk factors for heart disease: high blood pressure, high cholesterol, and smoking. Other key indicator include diabetic status, obesity (high BMI), not getting enough physical activity or drinking too much alcohol. Detecting and preventing the factors that have the greatest impact on heart disease is very important in healthcare. Computational developments, in turn, allow the application of machine learning methods to detect "patterns" from the data that can predict a patient's condition.

First at all, my plan is to visualize the data. I would like to see graphs to discover things like noise. My prediction of the data is that if you are a smoker and have bad habits like drink alcohol or not sleeping enough time, you’ll probably suffer a heart disease.

**Description**

The original dataset of nearly 300 variables was reduced to just about 20 variables. In addition to classical EDA, this dataset can be used to apply a range of machine learning methods, most notably classifier models (logistic regression, SVM, random forest, etc.). You should treat the variable "HeartDisease" as a binary ("Yes" - respondent had heart disease; "No" - respondent had no heart disease). But note that classes are not balanced, so the classic model application approach is not advisable. Fixing the weights/under sampling should yield significantly betters results.

The data has 17 attributes and the predicted class: heart disease: yes or no. the explanation of the attributes is:

1. Heart Disease: Respondents that have ever reported having coronary heart disease (CHD) or myocardial infarction (MI).
2. BMI: Body Mass Index (BMI).
3. Smoking: Have you smoked at least 100 cigarettes in your entire life? ( The answer Yes or No ).
4. Alcohol Drinking: Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week
5. Stroke: (Ever told) (you had) a stroke?
6. Physical Health: Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days was your physical health not good? (0-30 days).
7. Mental Health: Thinking about your mental health, for how many days during the past 30 days was your mental health not good? (0-30 days).
8. DiffWalking: Do you have serious difficulty walking or climbing stairs?
9. Sex: Are you male or female?
10. Age Category: Fourteen-level age category.
11. Race: Imputed race/ethnicity value.
12. Diabetic: (Ever told) (you had) diabetes?
13. Physical Activity: Adults who reported doing physical activity or exercise during the past 30 days other than their regular job.
14. GenHealth: Would you say that in general your health is...
15. SleepTime: On average, how many hours of sleep do you get in a 24-hour period?
16. Asthma: (Ever told) (you had) asthma?
17. Kidney Disease: Not including kidney stones, bladder infection or incontinence, were you ever told you had kidney disease?
18. Skin Cancer: (Ever told) (you had) skin cancer?

Size of data: 18 columns and 319795 entries.

If there are missing values, I would probably use the most common value of that specific attribute. I will do **Classification tasks** using probably a **decision tree and K-Nearest Neighbors (KNN)**. Naïve baiyes and ANN

I think smoking and mental health will be important factors to decide whether you have a heart disease. Using python with panda, NumPy, seaborn to draw graphs and plot diagrams will be enough to process my data.

Training and testing data:

Typically, when you separate a data ser into a training and testing data set, most of the data is used for training, and a smaller portion of the data is used for testing. The simplest way would be to split the training sets into 2/3 of the data and the rest 1/3 for the testing data.

Parameter Estimation:

One approach to overcoming this bias is to nest the hyperparameter optimization procedure under the model selection procedure. This is callednested cross-validation and is the preferred way to evaluate and compare tuned machine learning models. The k-fold cross-validation procedure is used to estimate the performance of machine learning models when making predictions on data not used during training. (Maximum Likelihood)

Timeline:

For the second part, I intend to visualize my data and see if it needs some cleaning or noise removal, and then data preprocessing.

For the last part, I’ll submit my conclusion, description and the final implementation with the presentation to explain what I did.

Conclusion:

I explored a lot of data sets for my project, and I think I found the right one for my exploration and implementation. It is something that, for example, in The United States a lot of people are suffering. And I would like to discover the main causes of heart problems.

Cite:

https://www.kaggle.com/code/ahmedklabi/heart-disease-pred/notebook

2020 annual CDC survey data of 400k adults related to their health status.