Cardiovascular Disease

manar almohaimeed

##Design

This project is one of the T5 Data Science BootCamp requirements. Data provided by Kaggle has been used in this project. We will analyze the data to find the relation between the features and use machine learning algorithms that would enable us to predict if the patient has cardiovascular disease or not.

##Data

The dataset is provided in .csv format. It contains of 70,000 patient record with 13 features for each patients doing cardiovascular disease examination. All of the dataset values were collected at the moment of medical examination.

There are 3 types of input features:Objective: factual information, Examination: results of medical examination, and Subjective: information given by the patient.

The features list:

- Age | Objective Feature | age | int (days)
- Height | Objective Feature | height | int (cm)|
- Weight | Objective Feature | weight | float (kg)|
- Gender | Objective Feature | gender | categorical code| 1:women, 2:men
- Systolic blood pressure | Examination Feature | ap_hi | int|
- Diastolic blood pressure | Examination Feature | ap_lo | int|
- Cholesterol | Examination Feature | cholesterol | 1: normal, 2: above normal, 3: well above normal|
- Glucose | Examination Feature | gluc | 1: normal, 2: above normal, 3: well above normal|
- Smoking | Subjective Feature | smoke | binary|
- Alcohol intake | Subjective Feature | alco | binary|
- Physical activity | Subjective Feature | active | binary|
- Presence or absence of cardiovascular disease | Target Variable | cardio | binary |

##Algorithms

Feature Engineering

- Cleaning the data by converting the Hight from CM to M, age from days to year , and change the data type to intger.
- Calculating the BMI by apply this formula (wight / height^2)

1. Finding the **most age** affecting the occurrence of cardiovascular disease?
2. Find if **smoke** affecting the occurrence of cardiovascular disease?

- The gender that most suffer from cardiovascular diseases

Models

Nai, and random forest were used to classify the tweets. The naive bayes get the higher accuracy with `Blobtext` labeling.

Model Evaluation and Selection

Logistic Regression, and random forest were used to molding the data . The random forest get the higher accuracy . The models were trained on a 25/75 test vs. train. The official metric was the accuracy of the model, where the model tested on the accuracy. The result of used model:

- Accuracy:71%

##Tools

-Numpy and Pandas for data manipulation

-Scikit-learn for modeling

-Matplotlib and Seaborn for plotting

-LogisticRegression and RandomForestClassifier for modling

##Communication

The slides are provided here, besides details are provided at the readme of the project.