Cardiovascular Disease Prediction

Shahd Omar Aljabarti

ABSTRACT

The cardiovascular diseases lead the cause of death across the world as the statement of World Health Organization. In 2019, 32% of the global death were due to the heart attacks and strokes. [1] This proposed project intended to predict the heart disease presence based on combination of several features and that could help in prevent life threats.

PROBLEM STATEMENT

The cardiovascular disease prediction is an interesting problem in the data science field. This project aims to explore the different features that might increase or affect the chance of cardiovascular disease occurrence and then use the extracted knowledge to predict whether a person has a cardiovascular disease or not.

DATA DESCRIPTION

Source of the dataset is Kaggle website where the dataset was extracted in excel format then converted to csv format. It has 70,000 records and 12 features. For the features, it has an age column that measured in days, gender which has two values 1 for female and 2 for male, height in cm and weight in kg, systolic blood pressure, diastolic blood pressure, cholesterol levels (1: normal, 2: above normal, 3: well above normal), glucose levels (1: normal, 2: above normal, 3: well above normal), alcohol intake, physical activity which has binary values 1 for active and 0 for inactive, smoking , and the last feature is the cardio disease presence which has 1 for heart disease presence and 0 for no presence.

DATA CLEANING

By using python packages (numpy, pandas, matplitlib) I started with renaming some columns to understandable names. I filled the null values of the height, systolic blood pressure and diastolic blood pressure columns with their mean values. I added new column named ‘age_in_years’ and filled it with the age in years by applying a function to the age in days column. Pandas.corr() function was used to discover the correlated features.

I also created new column named “Blood_pressure” which contains the levels of the blood pressure reading (systolic blood pressures/ diastolic blood pressure) whether it’s normal, at risk or high blood pressure based on the Center of Disease Control and Prevention categorization [2].

I created new column named BMI (Body Mass Index) as well, which is computed using heights and weight. After cleaning, the dataset ends up with 69,856 rows and 10 features.

ALGORITHMS

Logistic regression and random forest classifiers (with n_estimators = 100) were used before choosing logistic regression as the model with better performance.

MODEL EVALUATION

For testing and evaluation, I strive for high recall and high accuracy in prediction, I used recall, precision, and cross validation to estimate how accurately each model will perform. The following table is a comparison between logistic regression and random forest and the results show that logistic regression perform better.

	Logistic Regression	Random Forest
Recall	67.06%	66.51%
Precision	73.25%	66.6%
Cross Validation	71.72%	66.08%
ROC AUC	77.91%	59.24%

TOOLS

I used Jupyter Notebook to create the machine learning models and applied the following packages: NumPy and Pandas for data manipulation, Scikit-learn for modeling, Matplotlib and Seaborn for plotting.

COMMUNICATION

The following figures show interesting distribution and correlations of the cardio_disease_presence, figure 1 shows the counts of age observations in cardio_disease_presence categories, and figure 2 shows pairwise correlation of all columns.

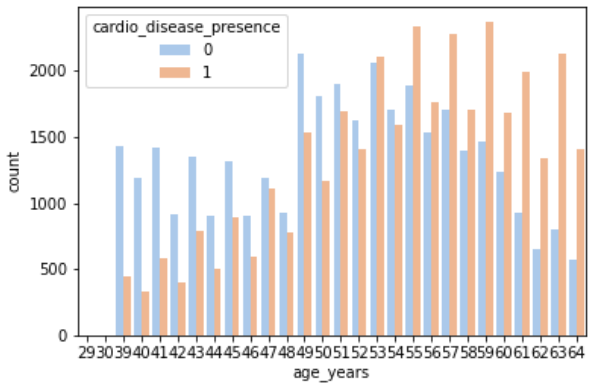

Figure 2: count plot for age-cardio_disease_presence

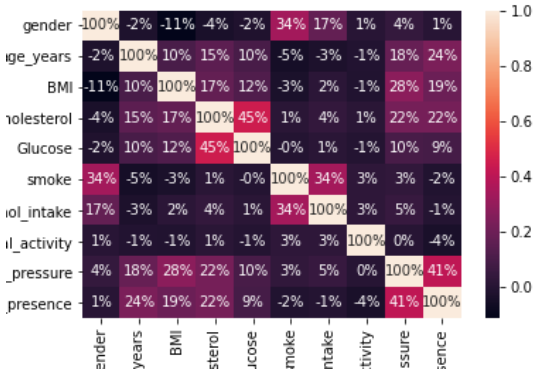

Figure 1: features correlations

Reference:

[1] 'Cardiovascular diseases (CVDs)'. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds) (accessed Oct. 03, 2021).

[2] CDC, 'High Blood Pressure Symptoms, Causes, and Problems | cdc.gov', *Centers for Disease Control and Prevention*, May 18, 2021. https://www.cdc.gov/bloodpressure/about.htm (accessed Oct. 04, 2021).