Descriptive Analysis and Classification Modeling for Cardiovascular Disease

Salma AlArfaj

Abstract

The goal of this project is to use a very large dataset to apply machine learning algorithms, exploratory data analysis, data cleaning, feature engineering and statistical processes. The effect of each provided feature on the target was studied by answering questions each time about the possible relation between them. We created a model that will predict if a person is likely to have cardiovascular disease or not. We run an evaluation metrics in four fitted models to compare scores. We voted for the highest score model (i.e., the random forest model).

Design

The dataset cardiovascular disease (CVDs) is provided by Kaggle. We used it to find the presence or absence of cardiovascular disease. Studying such data classifying statuses accurately via machine learning models and descriptive analysis would enable us to know which common factors usually accompany cardiovascular diseases. Features like age, body mass index, gender, blood pressure, cholesterol, glucose, smoking and alcohol intake, physical activity and obesity were studied.

Data

The dataset consists of 70,000 records of patients' data, with eleven features and the target variable "Cardio" with two labels: presence or absence of cardiovascular disease. It contains three types of input features: objective or factual information, examination or results of medical examination and subjective information given by the patient were all collected at the moment of medical examination. Grouping features into more general categories such as age-groups, and obesity were undertaken for the sake of descriptive analysis.

Algorithms

Feature Engineering

1. Transferring the age into years for a better interpretation
2. Converting categorical features to binary dummy variables and defining functions to generate features based on features we have.
3. Creating masks and graphs to highlight outliers and inspect unusual values and boundaries during EDA
4. Dropping columns that does not provide any additional information to the problem and exclude impossible values such as negative systolic blood pressure and zeros in diastolic blood pressure.

Models

K-nearest neighbors, logistic regression, decision tree, and random forest classifiers were used. Based on the findings from the classification report, we have our winner model to be the fitted random forest model. Although, here we have one of the best scores for this dataset in the literature, that has been explored, there is a place for improvement. For the sake of this project, we will end here. However, randomized search on hyper parameters via RandomizedSearchCV could be beneficial for better choice of the models' parameters. In addition, implementation of neural network is highly suggested to be further examined as part of the next steps.

Model Evaluation and Selection

The entire training dataset of 68,675 records was split into 80/20 train vs. test.

Final random forest scores: (n_estimators=1000, max_depth=10 , min_samples_leaf=15, min_samples_split=4)

- Accuracy 0.733
- F1 0.746 no cardio, 0.717 no cardio
- precision 0.717 no cardio, 0.752 no cardio
- recall 0.779 no cardio, 0.686 no cardio

Tools

- Numpy and Pandas for data manipulation
- Scikit-learn for modeling
- Matplotlib and Seaborn for plotting

Sources:

- Reference blood pressure categories: heart.org (https://www.heart.org/-/media/files/health-topics/high-blood-pressure/hbp-rainbow-chart-english.pdf)
- Reference obesity classes: NIH , source:(https://www.euro.who.int/en/health-topics/disease-prevention/nutrition/a-healthy-lifestyle/body-mass-index-bmi