Cardiovascular Disease Classification Report

Abstract:

This project is to build classification models and try to analyze and collect insights from the data set and predict the likelihood of a person developing a cardiovascular disease based on different criteria defined in this data set.

Data:

The data that will be used in this project is downloaded from kaggle.com (https://www.kaggle.com/sulianova/cardiovascular-disease-dataset). The data is Cardiovascular Disease data.The dataset consists of over 70000 observations with 13 features.

Algorithms:

- Exploratory Data Analysis was done to the dataset.
- Building multiple models and finding out the well-suited one for this specific dataset.

Cleaning:

- Checked null values
- Drop duplicate values / Converting age from days to years.

Feature Engineering:

New feature: we have height and weight, we calculated **BMI**.

Model Building:

Around 7 models were tried and played with to get the best model that goes hand in hand with the dataset. After performing simple train and validation on the models one was chosen for further investigation. Models trained was:
Logistic regression (Baseline)/ K-NN
/ Random forest/ Extra Trees /Decision trees/ Bernoulli Naive Bayes / Gaussian Naive Bayes

The Best Model:
Decision trees.

Tools:

- Numpy - Pandas - Matplotlib - Seaborn - Sklearn