Heart disease prediction using different machine learning techniques

Abstract

Heart disease is considered as the most common disease among people, this project aims to find the most reasons for it by using different data visualizations, in addition to building different models to predict whether the person has a heart disease/heart attack, or not, the models will be build using a different classifications algorithm.

Design

The data has been got from Kaggle, with different data explorer and machine learning algorithms, the following questions will be answered hopefully:

- What are the most features effects in heart diseases?
- How is the model trusted?
- Which gender is likely to have a high chance of heart disease?
- Does diabetes have a relationship with heart diseases?

Data

The dataset contains 253680 instances with 22 features with the target, the dataset suffers from imbalanced data.

Link: Heart Disease Health Indicators Dataset | Kaggle

Algorithms

Feature selection:

Select the most relevant features to build a model.

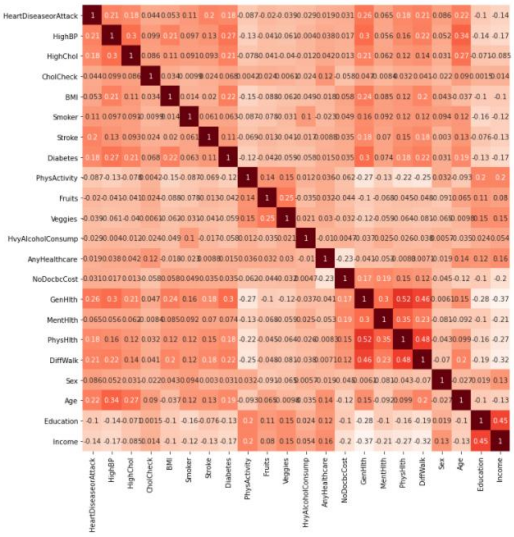
modelling:

Building six different models with different number of features based on correlation , with two types of algorithms: Decision Tree ,and Linear Regression .

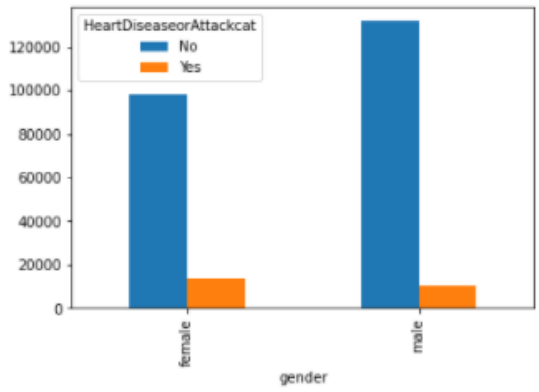
Tools:

- Seaborn
- Pandas
- Skit-learn
- Matplotlib

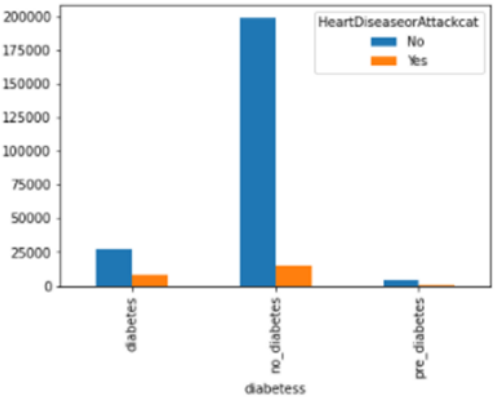
Answered questions:



The most effecting features in heart diseases are: CholCheck,Age, DiffWalk, HighBP,Stroke ,HighChol ,Diabetes,PhysHlth ,Smoker ,sex . Descending order



Women are likely to have a high chance of heart disease



Diabetes has a weak with heart diseases

Experimental Results

Measure/Algorithm	LR N=21	LR N=10	LR N=5	DT N=21	DT N=10	DT N=5
Accuracy	100%	91%	91%	100%	90%	83%
precision	100%	73%	71%	100%	65%	50%
recall	100%	55%	53%	100%	54%	50%