Comparative study of ML model to predict heart disease

Asalah Zarban

Abstract

The goal of this project is to use logistic regression models to predict heart disease in order to help patients make a decision on lifestyle changes to reduce complications and save the life of high-risk patients. Also Comparative accuracy of ML model.

Design

The dataset is publicly available on the Kaggle website, and it is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. The classification goal is to predict whether the patient has 10-year risk of future coronary heart disease (CHD). Based on several kinds of features as age, sex male, …, Ten Year CHD, we predict the data using various ML models algorithm.

Data

The dataset contains more than 4,000 observations with 15 features for each, and we divided it into 5 categorical to be clearer.

Algorithms

P-value approach

I take different machine learning algorithm as Logistic Regression, Naive Bayes, Random Forest Classifier, Extreme Gradient Boost, K-Nearest Neighbour, Decision Tree, Support Vector Machine and try find algorithm which predict accurately.

Tools

Pandas, numpy, seaborn for loading dataset

sklearn.metrics, sklearn.linear_model, sklearn.naive_bayes, sklearn.ensemble, xgboost, sklearn.neighbors, sklearn.tree, sklearn.svm for modeling dataset

matplotlib.pyplot for visualisation

sklearn.model_selection, statsmodels.api, scipy.stats, collections, statsmodels.tools for splitting dataset

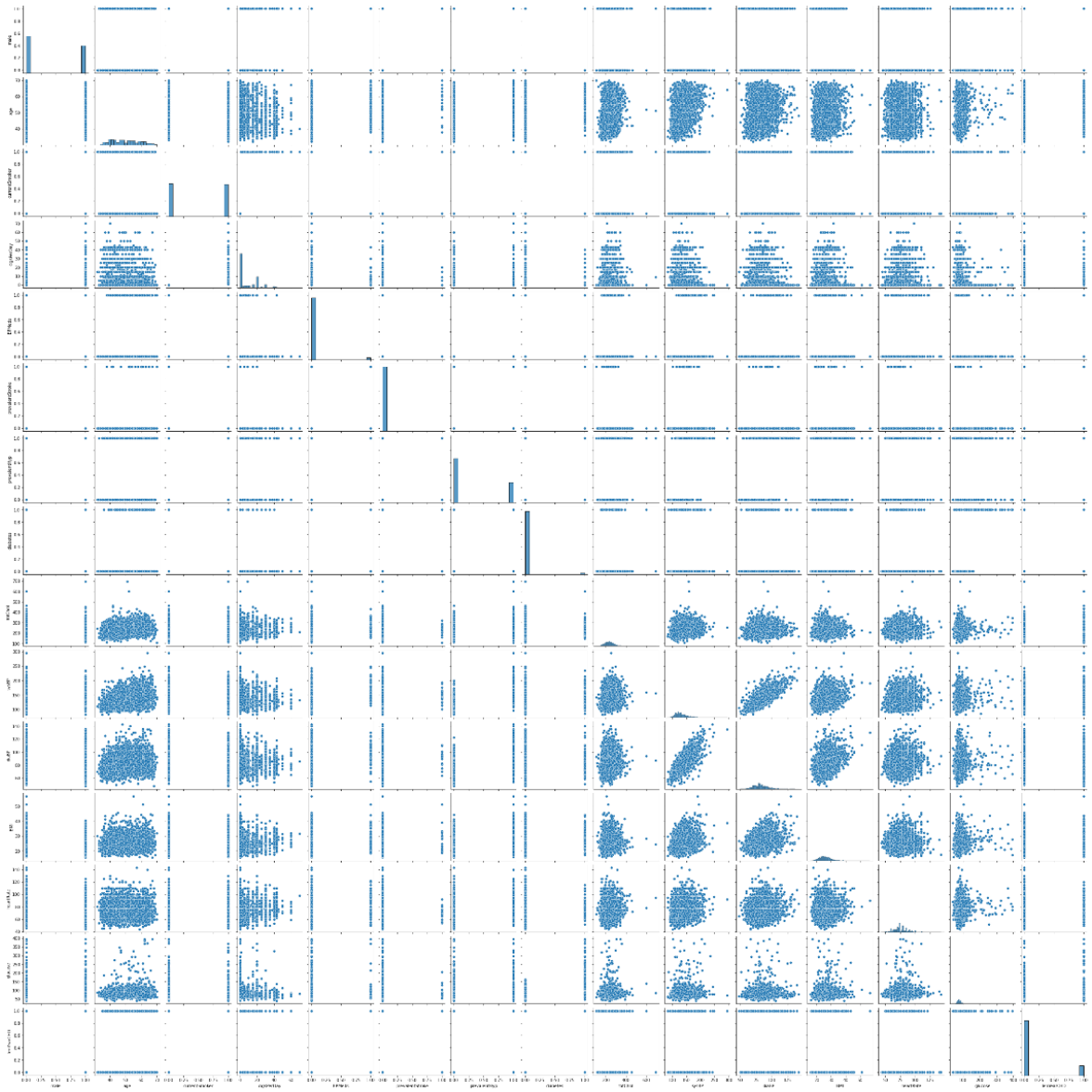
sklearn.preprocessing for preprocessing dataset

mlxtend.classifier for ensembling dataset

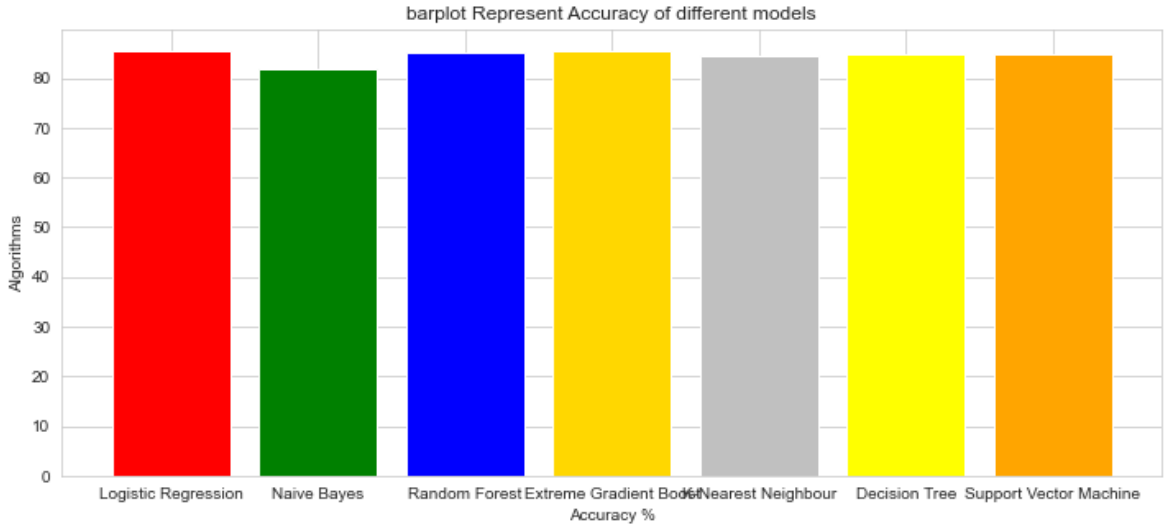
Communication

These are some of the visuals presented in project.

Exploratory analysis for our features by pairplot:



Bar plot Represent Accuracy of different models:



Bar plot Represent feature importance: