Cardio Disease

Shahad Alowais - Amena Algroon

introduction

Cardiovascular diseases are the leading cause of death worldwide. In this “Cardio Disease” dataset, We'll use it to build classification models and try to analyze and gather the insights of a dataset and predict the possibility of a person having Cardiovascular disease based on various parameters specified in this dataset.

Data Set

The cardiovascular disease dataset is found on Kaggle.

The data consists of 70,000 patient records and 13 features.

Tools


python


jupyter


seaborn


NumPy


matplotlib


pandas


scikit
learn

Project Workflow

EDA

Chiacking duplicate
Converting age from days to years

Feature Engineering

Drop columns (weight - height).
Add new column (bmi).
Remove outliers

Classification

EDA

Cardio disease is more relative with AGE, WEIGHT and CHOLESTEROL

Baseline

	Score train	Score Test
LogisticRegression	0.716	0.719
Polynomial(degree=2)	0.708	0.710

Model Scores

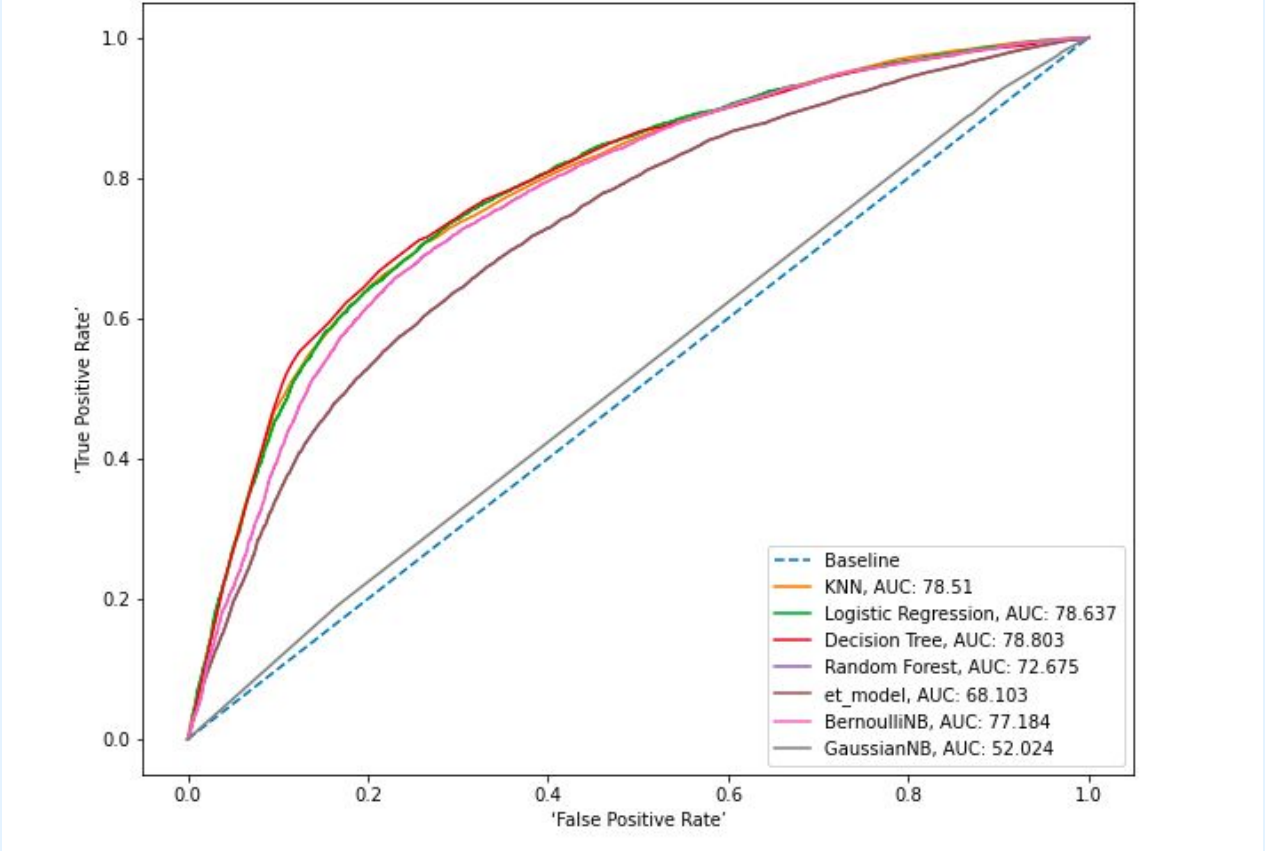
Classifier	Accuracy	F score	Precision	Recall
Logistic Regression	0.720	0.714	0.747	0.684
K-nearest neighbors	0.720	0.710	0.754	0.671
Decision Tree	0.726	0.726	0.743	0.711
Random Forest	0.669	0.681	0.671	0.691
Extra Trees	0.647	0.661	0.650	0.671
Bernoulli Naive Bayes	0.519	0.653	0.517	0.887
Gaussian Naive Bayes	0.708	0.688	0.757	0.631

Best Model

Decision Tree

Accuracy	F score	Precision	Recall
0.726	0.726	0.743	0.711

Roc curve


1.0
0.8
0.6
0.4
0.2
0.0
'True Positive Rate'
0.0
0.2
0.4
0.6
0.8
1.0
'False Positive Rate'
Baseline
KNN, AUC: 78.51
Logistic Regression, AUC: 78.637
Decision Tree, AUC: 78.803
Random Forest, AUC: 72.675
et_model, AUC: 68.103
BernoulliNB, AUC: 77.184
GaussianNB, AUC: 52.024

Thanks!