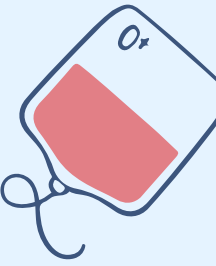
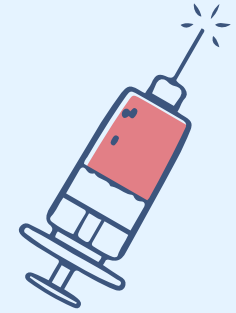


Cardio Disease

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



Project Workflow

01

EDA

Chiacking duplicate
Converting age from
days to years

02

**Feature
Engineering**

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

03

Classification



age	1	-0.024	-0.084	0.052	0.021	0.018	0.16	0.099	-0.05	-0.032	-0.0094	0.23
gender	-0.024	1	0.5	0.15	0.0052	0.014	-0.045	-0.027	0.34	0.17	0.0093	0.0014
height	-0.084	0.5	1	0.29	0.0056	0.0063	-0.05	-0.018	0.19	0.096	-0.0064	-0.0093
weight	0.052	0.15	0.29	1	0.029	0.042	0.13	0.1	0.061	0.062	-0.011	0.17
ap_hi	0.021	0.0052	0.0056	0.029	1	0.016	0.022	0.011	-0.0018	0.00067	0.00084	0.054
ap_lo	0.018	0.014	0.0063	0.042	0.016	1	0.022	0.009	0.0037	0.0094	0.0063	0.065
cholesterol	0.16	-0.045	-0.05	0.13	0.022	0.022	1	0.44	0.0025	0.03	0.019	0.21
gluc	0.099	-0.027	-0.018	0.1	0.011	0.009	0.44	1	-0.011	0.0066	0.00058	0.08
smoke	-0.05	0.34	0.19	0.061	-0.0018	0.0037	0.0025	-0.011	1	0.34	0.031	-0.024
alco	-0.032	0.17	0.096	0.062	0.00067	0.0094	0.03	0.0066	0.34	1	0.029	-0.014
active	-0.0094	0.0093	-0.0064	-0.011	0.00084	0.0063	0.019	0.00058	0.031	0.029	1	-0.027
cardio	0.23	0.0014	-0.0093	0.17	0.054	0.065	0.21	0.08	-0.024	-0.014	-0.027	1
	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio

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

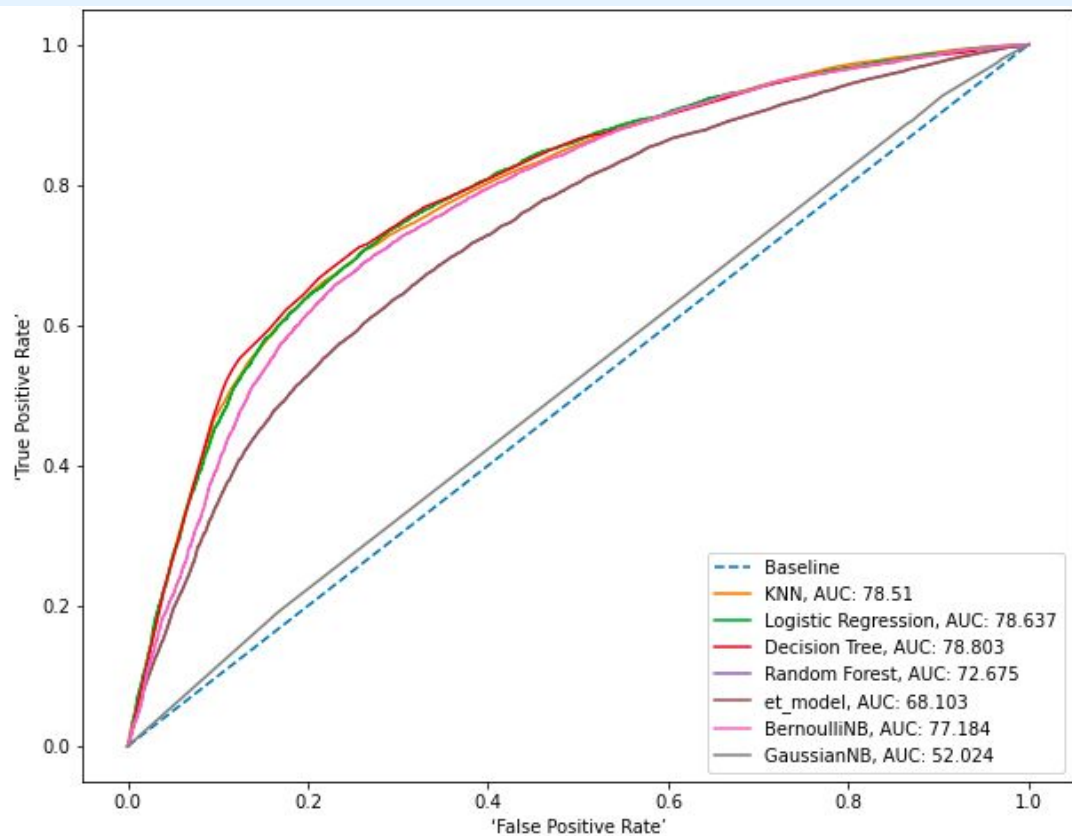


Best Model

Decision Tree

Accuracy	F score	Precision	Recall
0.726	0.726	0.743	0.711

Roc curve



Thanks!

