Cardiovascular Disease
for Classification

SDAIA Academy
أكاديمية سدايا

Abstract:

Cardiovascular disease is the leading cause of death throughout the United States, we explore several machine learning approaches to detect the presence of cardiovascular disease using only standard health information. we will also investigate which data features are most indicative of disease.

Design:

Our objective is to design a classifier to identify the presence of cardiovascular disease.

Data:

Our dataset has standard health information and information on the presence/absence of cardiovascular disease for over 70,000 patients records (34,979 presenting with cardiovascular disease and 35,021 not presenting with cardiovascular disease) and contains 11 features (4 demographic, 4 examination, and 3 social history).

Algorithms:

Logistic Regression.k-nearest neighbors, Decision tree, ExtraTrees, Random forest, gaussian naive bayes, Bernoulli naive bayes.

Tools:

The Tool That we need is Python and Jupyter notebook to execute the code. Pandas packages to manipulate data, Pandas_profiling , Statsmodels Pickle NumPy Sqlalchemy and visualisation library (such as Seaborn and Matplotlib), Sklearn ,Metrics Library to see the model performance on data.

Communication:

Score of 7 models for train and test datasets

ROC curve

LogisticRegression (AUC = 0.78)
KNeighborsClassifier (AUC = 0.73)
DecisionTreeClassifier (AUC = 0.61)
RandomForestClassifier (AUC = 0.74)
ExtraTreesClassifier (AUC = 0.71)
GaussianNB (AUC = 0.77)
BernoulliNB (AUC = 0.76)

Comparison of Scores of the Models

Accuracy
F1
Precision
Recall
ROC Score

Done by Ghadah Alharbi – Rahaf Alyousef