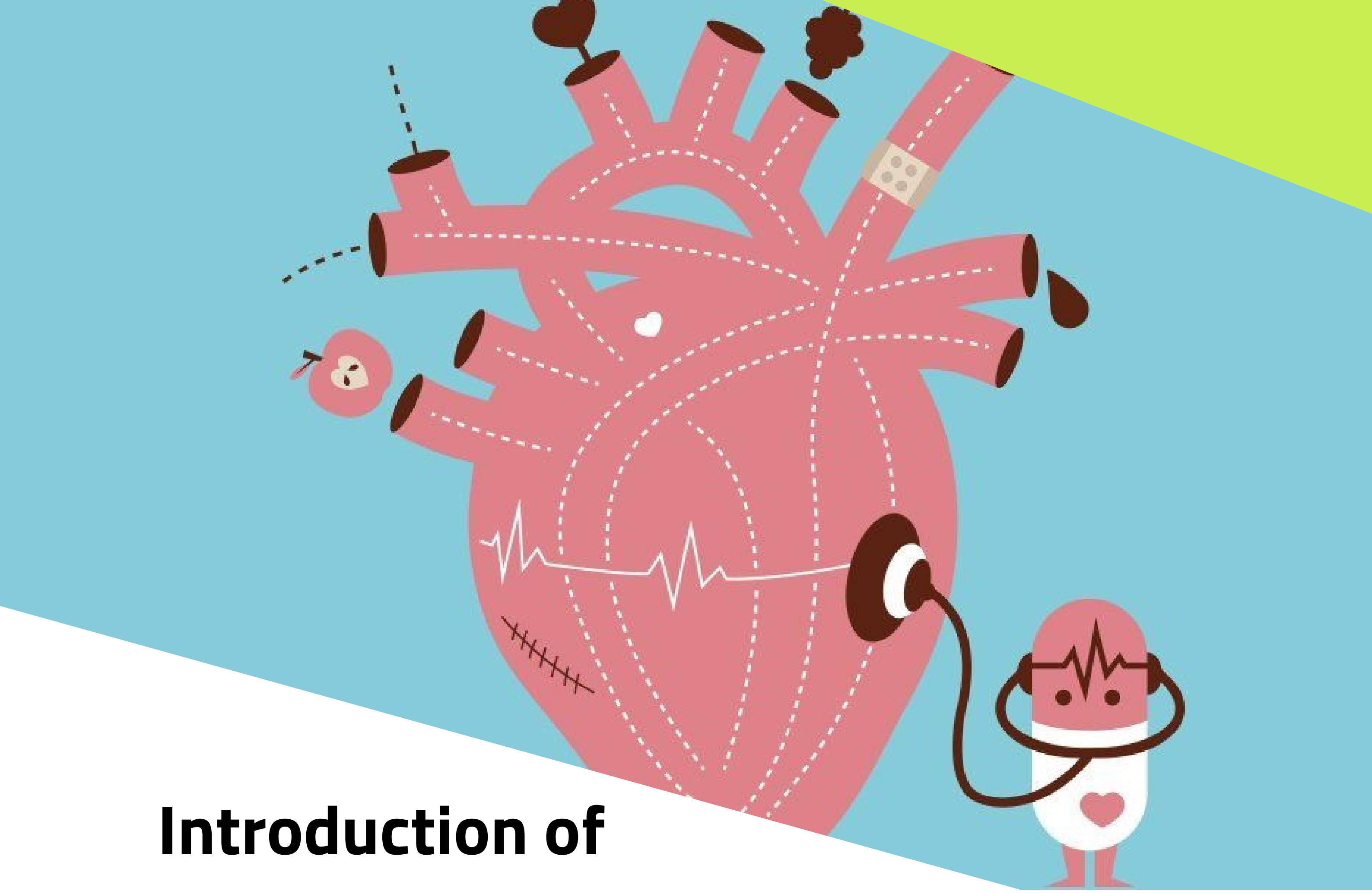
DATA SCIENCE

CARDIOVASCULAR DISEASE REPORT

May 2021

Prepared by
Sarthak Chandel



Introduction of problem

Heart diseases or CVDs are the number one cause of death globally with more than 18 million deaths every year . CVDs are concertedly contributed by hypertension , diabetes , overweight and unhealthy lifestyles .

This report trains the model on SVM and KNN algorithms with the dataset to predict the probability of a person suffering from a heart condition.

Dataset Description

The dataset consists of 70,000 records of patients data,
11 features + 1 target.
There are 3 types of input features:

4	4	3
OBJECTIVE	EXAMINATION	SUBJECTIVE

1. Age | Objective Feature | int (days)
2. Height | Objective Feature | int (cm) |
3. Weight | Objective Feature | float (kg) |
4. Gender | Objective Feature | categorical code |
5. Systolic blood pressure | Examination Feature | int |
6. Diastolic blood pressure | Examination Feature | int |
7. Cholesterol | Examination Feature | 1: normal, 2: above normal, 3: well above normal |
8. Glucose | Examination Feature | 1: normal, 2: above normal, 3: well above normal |
9. Smoking | Subjective Feature | binary |
10. Alcohol intake | Subjective Feature | binary |
11. Physical activity | Subjective Feature | binary |
12. Presence or absence of cardiovascular disease | Target Variable| binary |

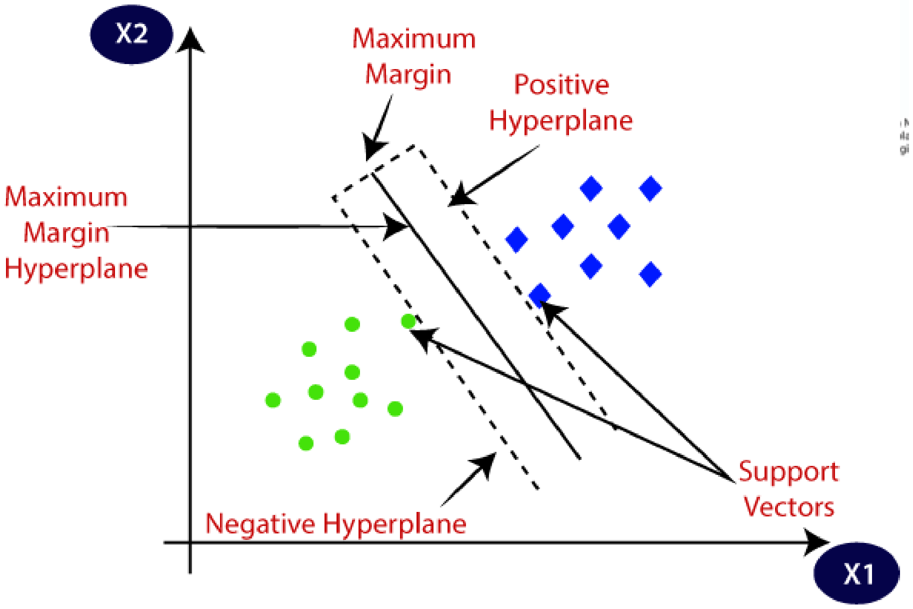
Algorithms Used

Support Vector Machine

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification.

The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine. Consider the below diagram in which there are two different categories that are classified using a decision boundary or hyperplane:



The distance between the hyperplane and the nearest data point from either set is known as the margin. The goal is to choose a hyperplane with the greatest possible margin between the hyperplane and any point within the training set, giving a greater chance of new data being classified correctly.

Algorithms Used

K- Nearest Neighbors

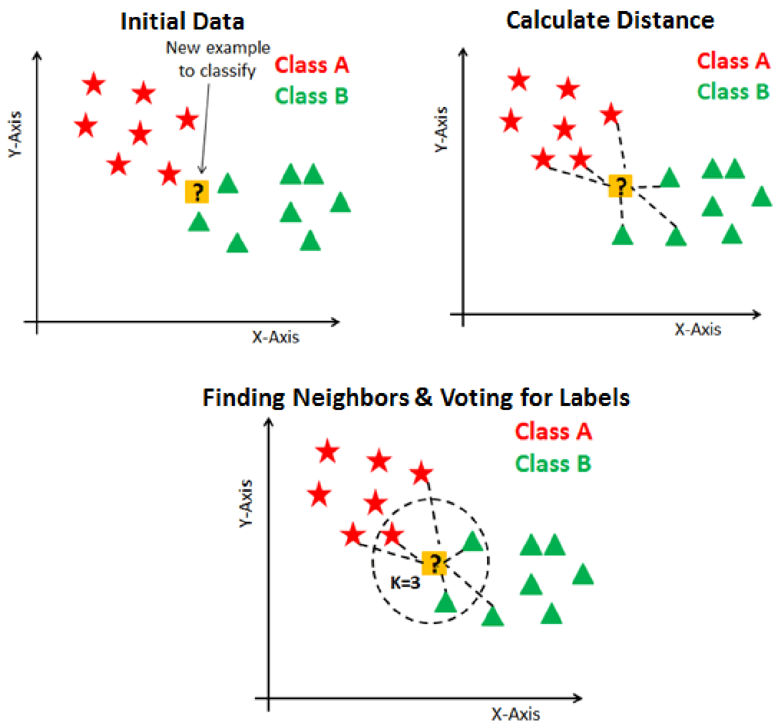
The K-nearest neighbors (KNN) algorithm is a supervised learning algorithm used to solve both regression and classification problems.

K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most like the available categories.

The input consists of the k closest training example in the data set.

To make a prediction, the KNN algorithm will use the entire dataset. For an observation that is not part of the dataset and is the actual value we want to predict, the algorithm will look for the K instances of the dataset closest to our observation.

Then for these K neighbors, the algorithm will use their output to calculate the variable y of the observation that we want to predict.

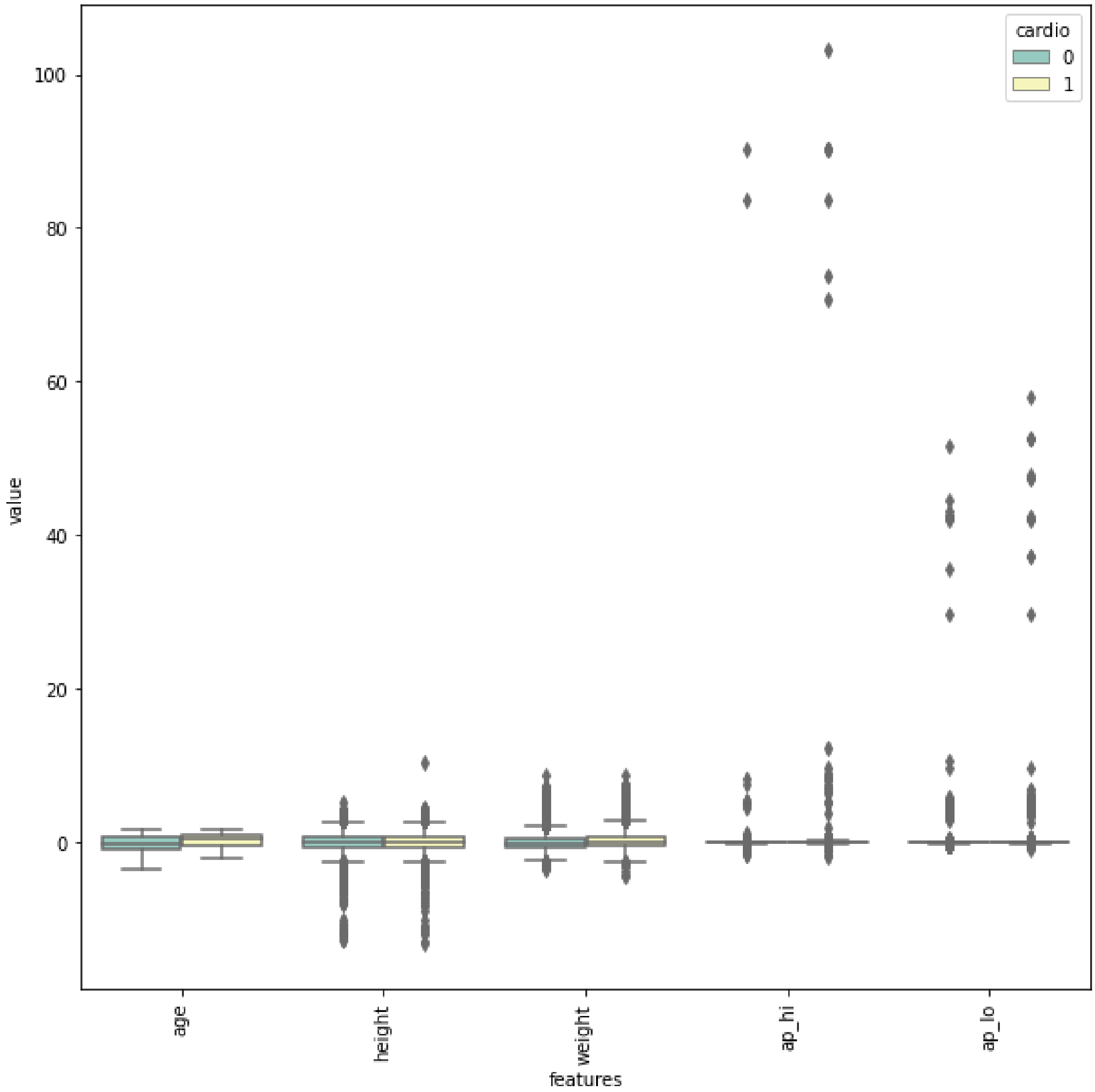


A common formula used in this algorithm to find the distance.

$$\text{Euclidean distance} = \sqrt{\sum_{i=1}^{n} \left(x_i - y_i\right)^2}$$

Data Visualization

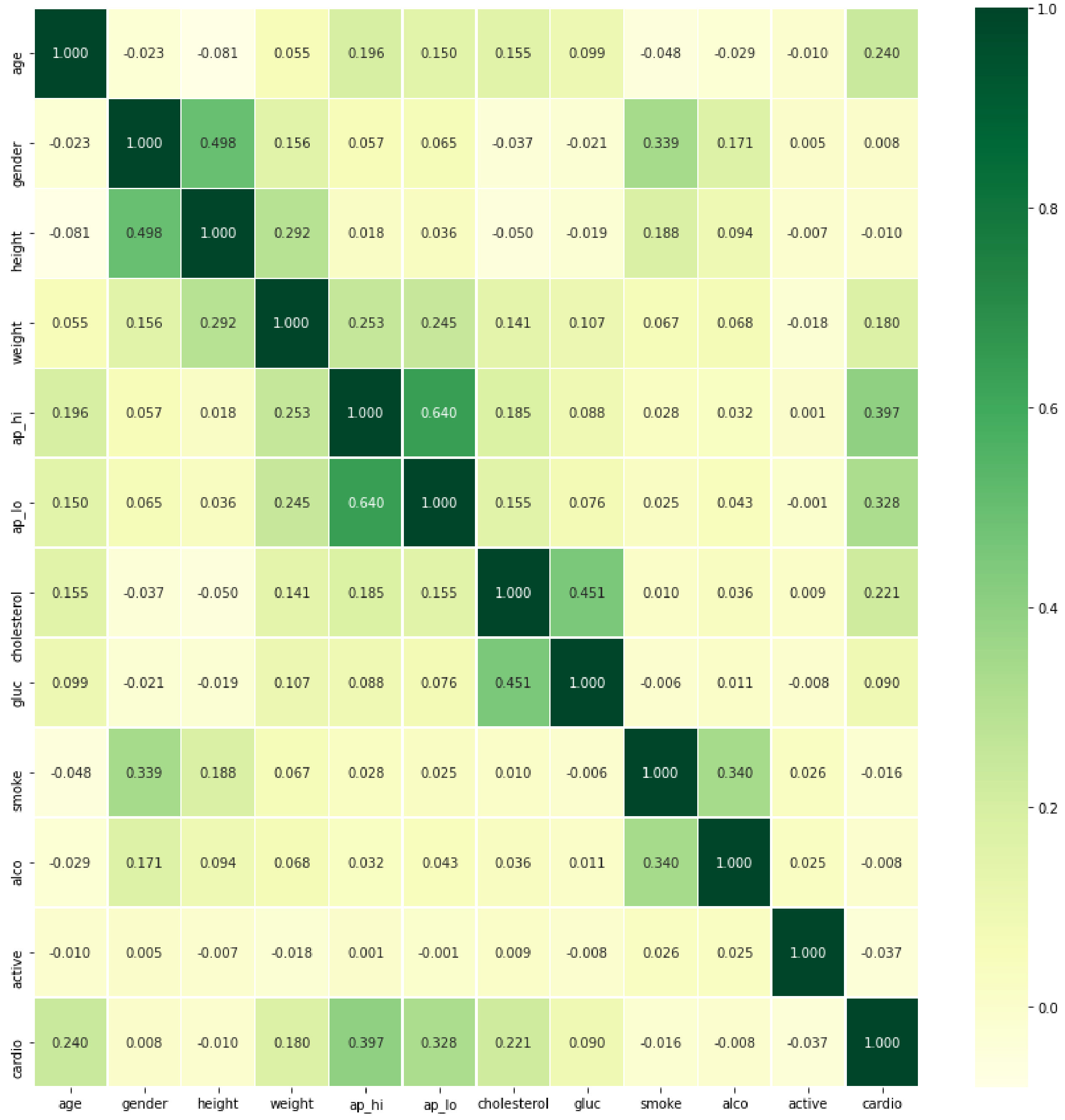
Box Plot



From this box plot we can spot the outliers. With extreme abnormalities in the ap_hi and ap_lo or the blood pressure columns.

Data Visualization

Heatmap

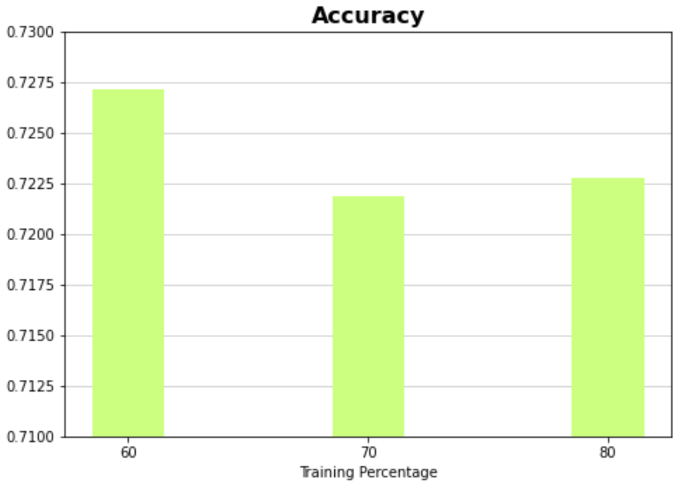


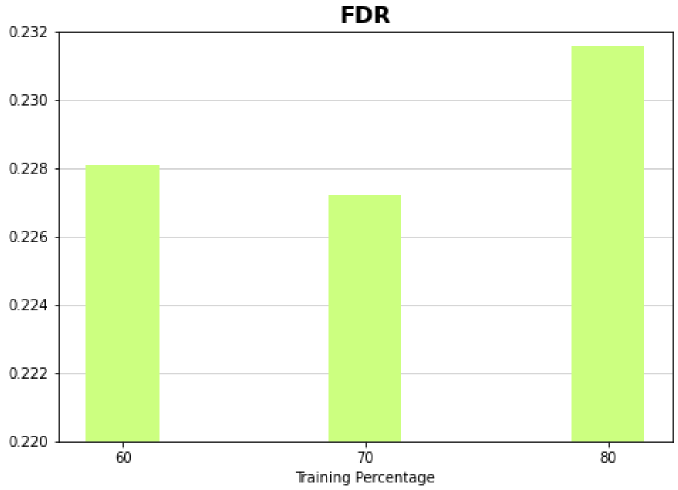
	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio
age	1.000	-0.023	-0.081	0.055	0.196	0.150	0.155	0.099	-0.048	-0.029	-0.010	0.240
gender	-0.023	1.000	0.498	0.156	0.057	0.065	-0.037	-0.021	0.339	0.171	0.005	0.008
height	-0.081	0.498	1.000	0.292	0.018	0.036	-0.050	-0.019	0.188	0.094	-0.007	-0.010
weight	0.055	0.156	0.292	1.000	0.253	0.245	0.141	0.107	0.067	0.068	-0.018	0.180
ap_hi	0.196	0.057	0.018	0.253	1.000	0.640	0.185	0.088	0.028	0.032	0.001	0.397
ap_lo	0.150	0.065	0.036	0.245	0.640	1.000	0.155	0.076	0.025	0.043	-0.001	0.328
cholesterol	0.155	-0.037	-0.050	0.141	0.185	0.155	1.000	0.451	0.010	0.036	0.009	0.221
gluc	0.099	-0.021	-0.019	0.107	0.088	0.076	0.451	1.000	-0.006	0.011	-0.008	0.090
smoke	-0.048	0.339	0.188	0.067	0.028	0.025	0.010	-0.006	1.000	0.340	0.026	-0.016
alco	-0.029	0.171	0.094	0.068	0.032	0.043	0.036	0.011	0.340	1.000	0.025	-0.008
active	-0.010	0.005	-0.007	-0.018	0.001	-0.001	0.009	-0.008	0.026	0.025	1.000	-0.037
cardio	0.240	0.008	-0.010	0.180	0.397	0.328	0.221	0.090	-0.016	-0.008	-0.037	1.000

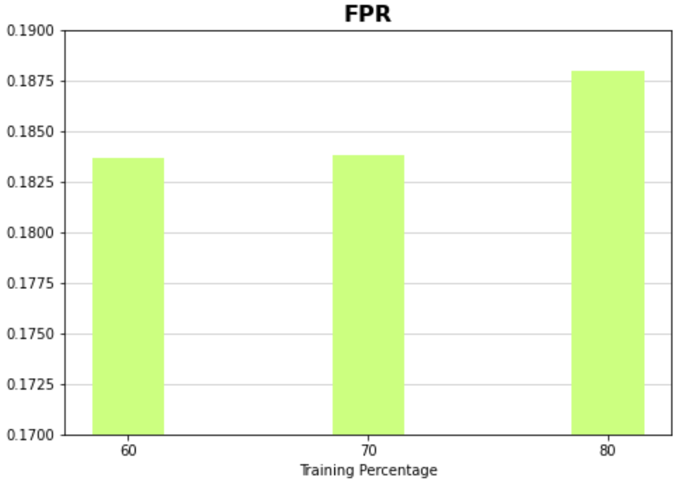
This heatmap shows us the strong correlation between cholesterol, blood pressure, age, glycogen, cholesterol levels and cardiovascular diseases

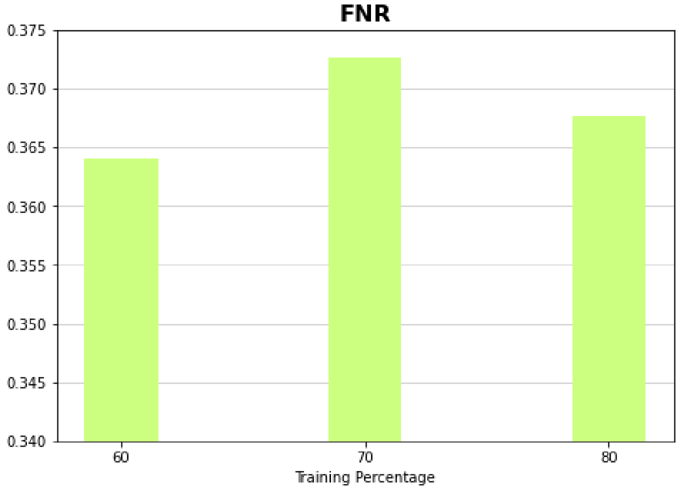
Results

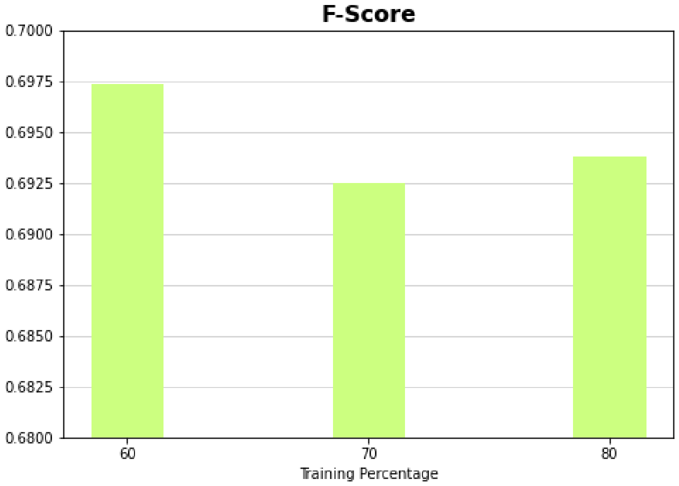
SVM

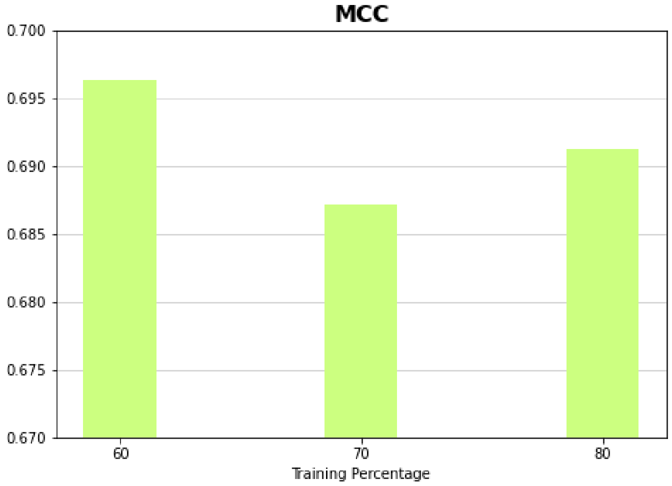

Accuracy
0.7300
0.7275
0.7250
0.7225
0.7200
0.7175
0.7150
0.7125
0.7100
60
70
80
Training Percentage


FDR
0.232
0.230
0.228
0.226
0.224
0.222
0.220
60
70
80
Training Percentage


FPR
0.1900
0.1875
0.1850
0.1825
0.1800
0.1775
0.1750
0.1725
0.1700
60
70
80
Training Percentage

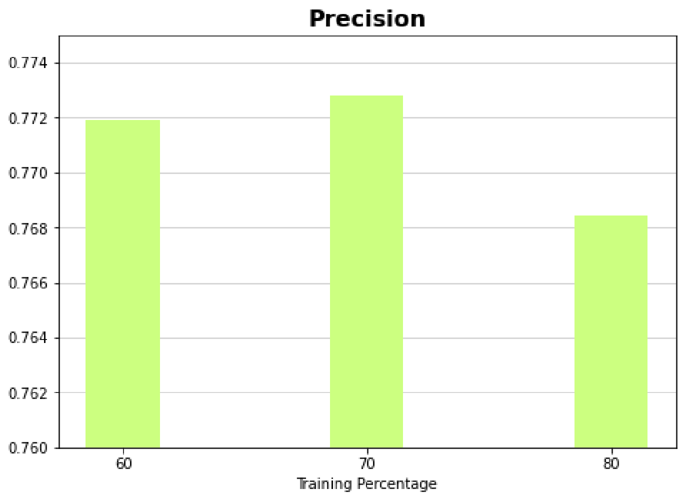

FNR
0.375
0.370
0.365
0.360
0.355
0.350
0.345
0.340
60
70
80
Training Percentage

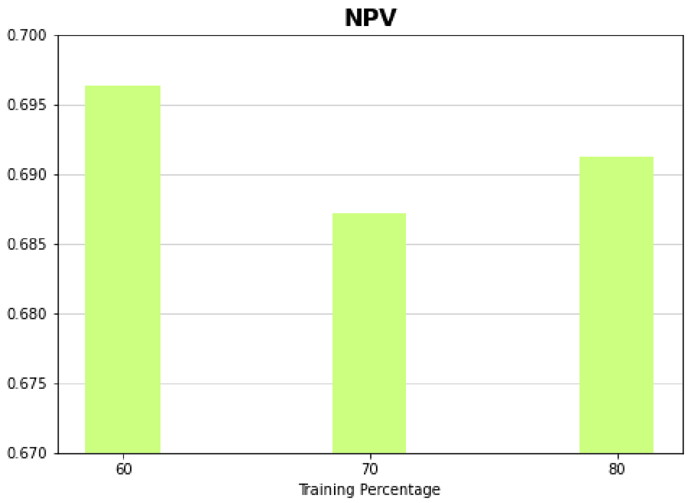

F-Score
0.7000
0.6975
0.6950
0.6925
0.6900
0.6875
0.6850
0.6825
0.6800
60
70
80
Training Percentage

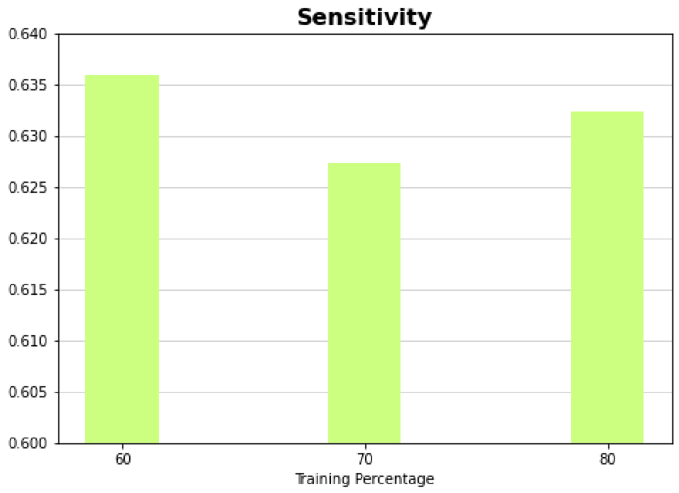

MCC
0.700
0.695
0.690
0.685
0.680
0.675
0.670
60
70
80
Training Percentage

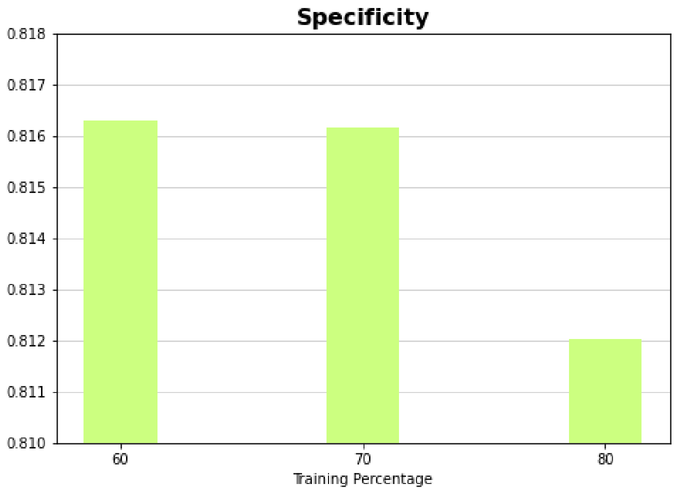
Results

SVM


Precision
0.774
0.772
0.770
0.768
0.766
0.764
0.762
0.760
60
70
80
Training Percentage

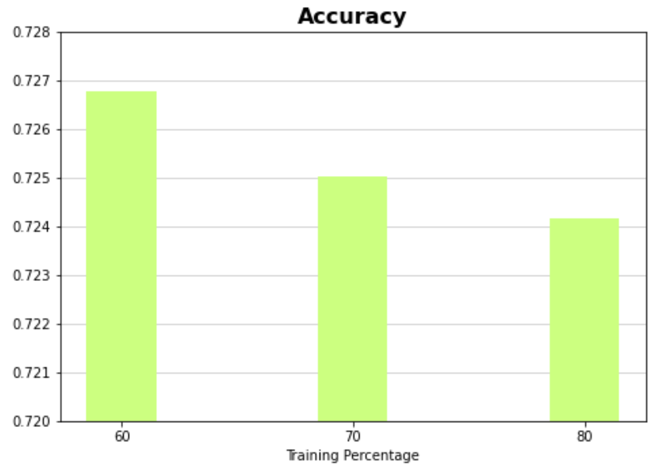

NPV
0.700
0.695
0.690
0.685
0.680
0.675
0.670
60
70
80
Training Percentage

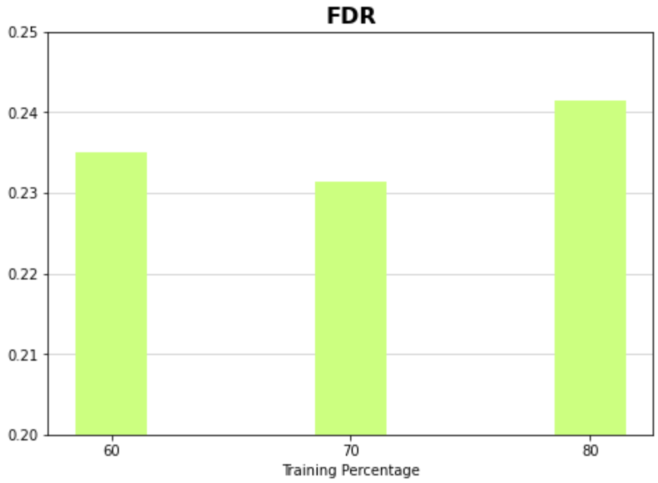

Sensitivity
0.640
0.635
0.630
0.625
0.620
0.615
0.610
0.605
0.600
60
70
80
Training Percentage

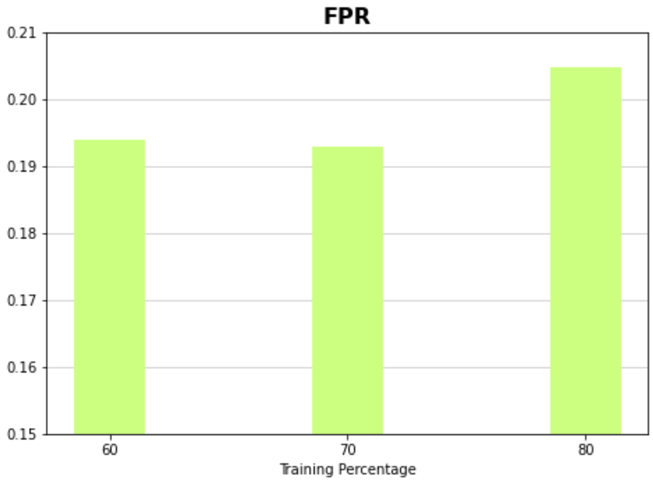

Specificity
0.818
0.817
0.816
0.815
0.814
0.813
0.812
0.811
0.810
60
70
80
Training Percentage

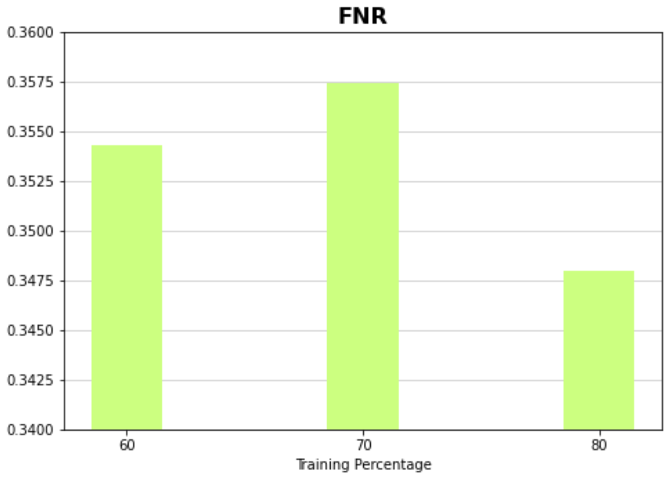
Results

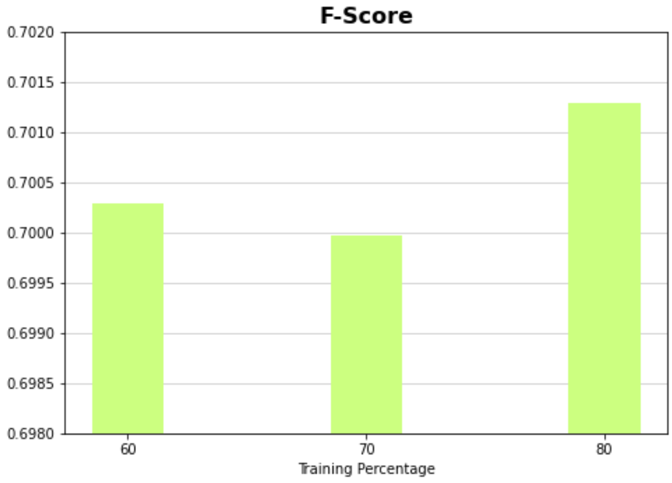
KNN

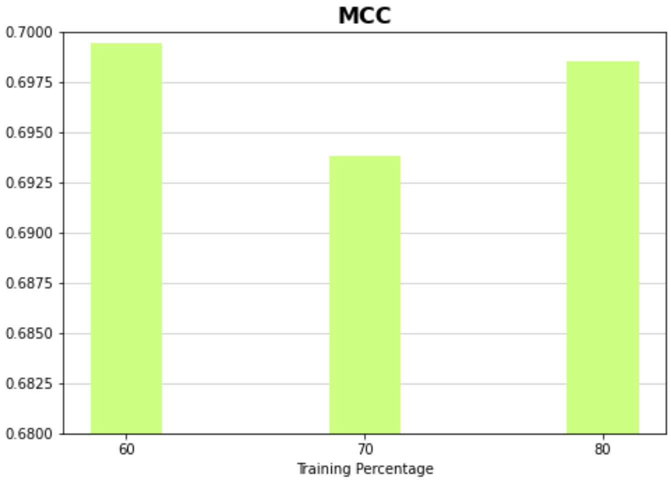

Accuracy
0.728
0.727
0.726
0.725
0.724
0.723
0.722
0.721
0.720
60
70
80
Training Percentage


FDR
0.25
0.24
0.23
0.22
0.21
0.20
60
70
80
Training Percentage


FPR
0.21
0.20
0.19
0.18
0.17
0.16
0.15
60
70
80
Training Percentage

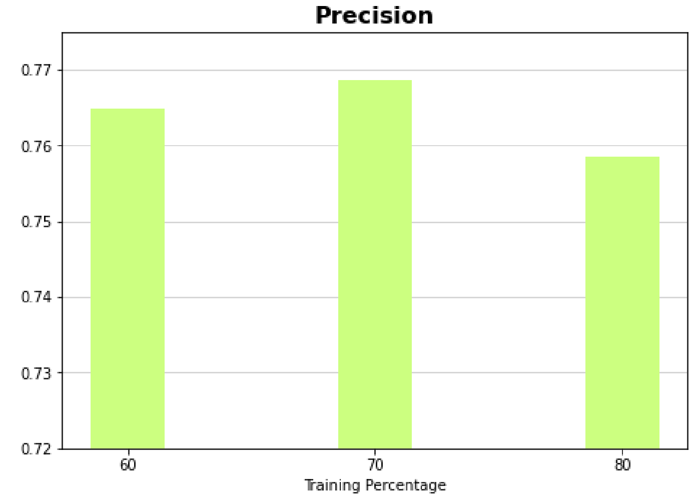

FNR
0.3600
0.3575
0.3550
0.3525
0.3500
0.3475
0.3450
0.3425
0.3400
60
70
80
Training Percentage

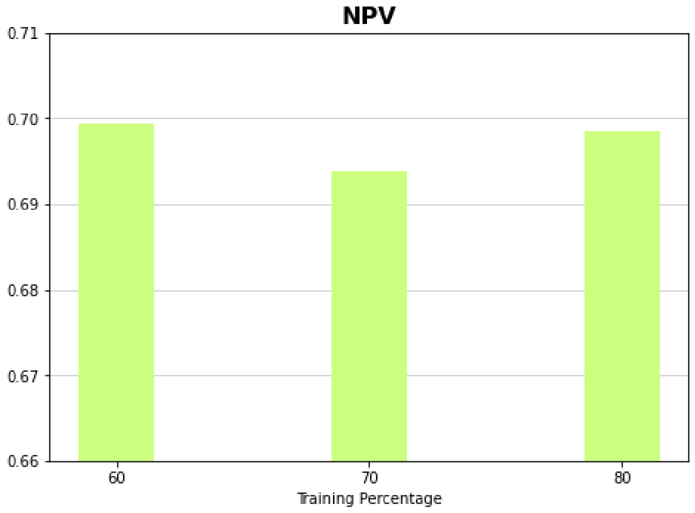

F-Score
0.7020
0.7015
0.7010
0.7005
0.7000
0.6995
0.6990
0.6985
0.6980
60
70
80
Training Percentage

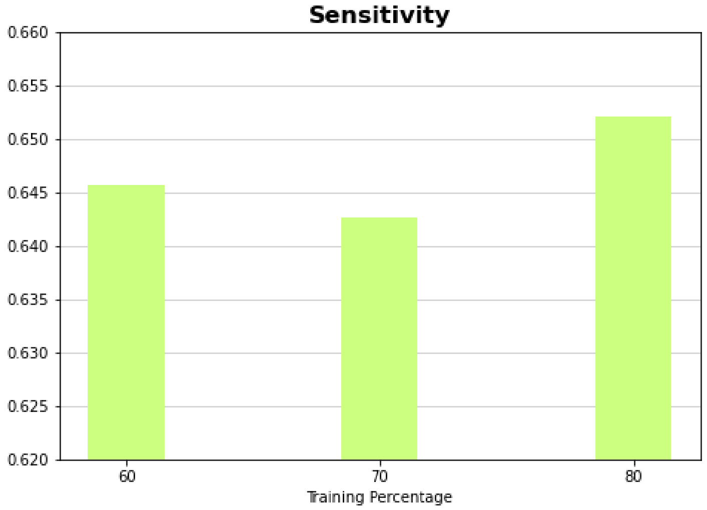

MCC
0.7000
0.6975
0.6950
0.6925
0.6900
0.6875
0.6850
0.6825
0.6800
60
70
80
Training Percentage

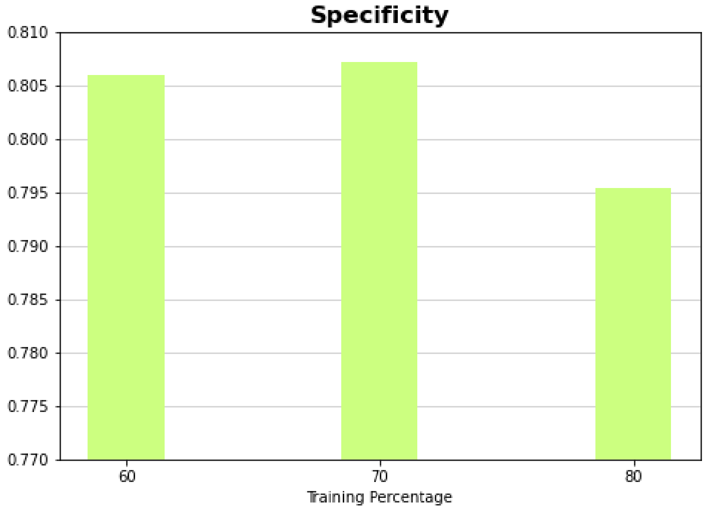
Results

KNN


Precision
0.77
0.76
0.75
0.74
0.73
0.72
60
70
80
Training Percentage


NPV
0.71
0.70
0.69
0.68
0.67
0.66
60
70
80
Training Percentage


Sensitivity
0.660
0.655
0.650
0.645
0.640
0.635
0.630
0.625
0.620
60
70
80
Training Percentage


Specificity
0.810
0.805
0.800
0.795
0.790
0.785
0.780
0.775
0.770
60
70
80
Training Percentage

Result Analysis

Confusion Matrix

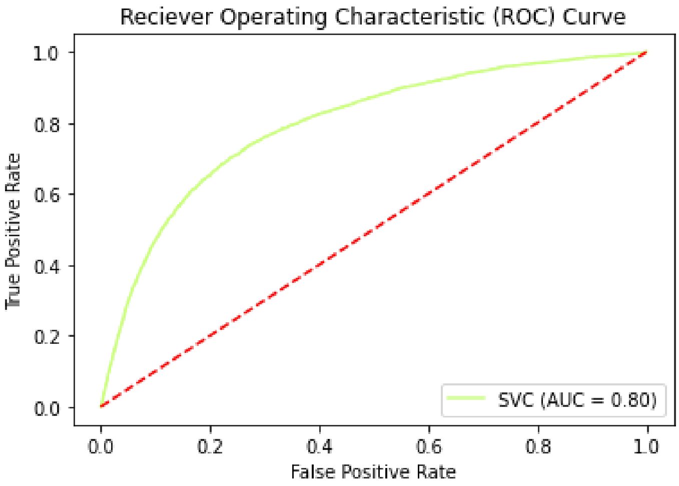
90:10 Split	Predicted Negative	Predicted Positive
Actual Negative	2855	628
Actual Positive	1265	2153

SVM

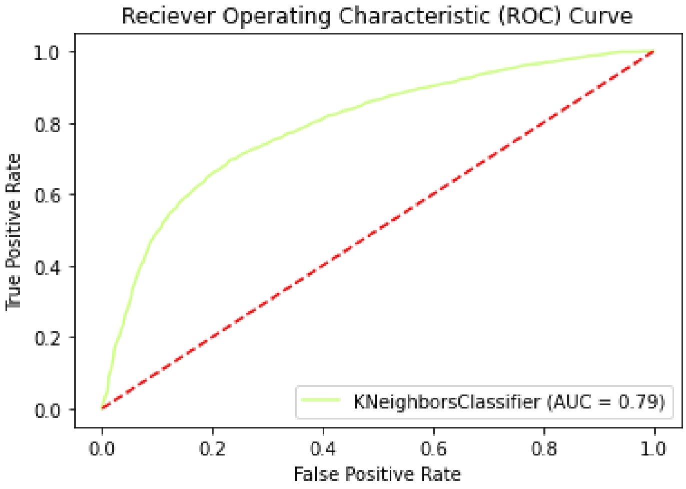
90:10 Split	Predicted Negative	Predicted Positive
Actual Negative	2767	681
Actual Positive	1189	2264

KNN

ROC



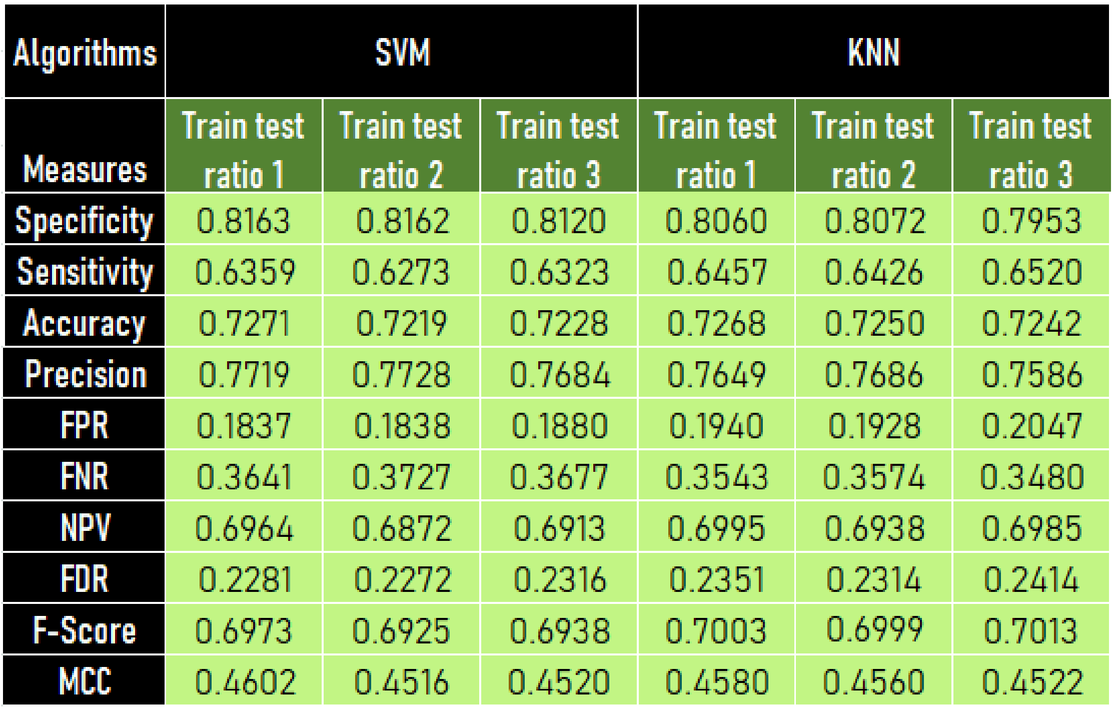
SVM



KNN

Result Analysis

Table for evaluation parameter



Algorithms	SVM			KNN		
Measures	Train test ratio 1	Train test ratio 2	Train test ratio 3	Train test ratio 1	Train test ratio 2	Train test ratio 3
Specificity	0.8163	0.8162	0.8120	0.8060	0.8072	0.7953
Sensitivity	0.6359	0.6273	0.6323	0.6457	0.6426	0.6520
Accuracy	0.7271	0.7219	0.7228	0.7268	0.7250	0.7242
Precision	0.7719	0.7728	0.7684	0.7649	0.7686	0.7586
FPR	0.1837	0.1838	0.1880	0.1940	0.1928	0.2047
FNR	0.3641	0.3727	0.3677	0.3543	0.3574	0.3480
NPV	0.6964	0.6872	0.6913	0.6995	0.6938	0.6985
FDR	0.2281	0.2272	0.2316	0.2351	0.2314	0.2414
F-Score	0.6973	0.6925	0.6938	0.7003	0.6999	0.7013
MCC	0.4602	0.4516	0.4520	0.4580	0.4560	0.4522

SVM and KNN both seem to have similar values across all measures. In this case, using KNN would be beneficial as it has a lower FNR rate and false negatives in this case could be life threatening.

Having a slightly higher FPR is not a major issue in this study.