

CLASSIFICATION ASSIGNMENT

Requirement or Problem statement:

A requirement from the Hospital, Management asked us to create a predictive model which will predict the Chronic Kidney Disease (CKD) based on the several parameters.

1. Identification of statement

- In this case, i am clearly identify the problem state
- I am going to predict the Chronic Kidney Disease (CKD) based on the several parameters given by the client.
- As a data scientist, i am going create good model and predict the Chronic Kidney Disease (CKD).

2. Basics information about the dataset

INPUT DATA					OUTPUT DATA
Age	Bp	...	pe	ane	classification
2	76.34	...	yes	no	yes
3	76.34	...	poor	no	no
4	76.34	...	poor	no	no

- In the dataset, i have 399 row and 25 columns

3. Pre-Processing method

➤ In the dataset, 12 categories are categorical data. So i need to change the data as a meaningful number.

➤ In this case, i am going to use nominal data to expanse the column and change the values to meaningful number.

➤ In sg category- In this case, yes or no response are considered nominal data. yes or no is not express any form of rank or order.

➤ So i am going to use nominal data to change the smoke category for meaning data (sting to number)

4. Confusion matrix and Classification Report comparison

1.LOGISTIC REGRESSION.

1.Logistic Regression cm = [39 12]

[6 76]

2.Logistic regression clf report

	precision	recall	F1 score	support
0	0.87	0.76	0.81	51
1	0.86	0.93	0.89	82
accuracy			0.86	133
Macro avg	0.86	0.85	0.85	133
Weighted avg	0.86	0.80	0.86	133

3.Logistic Regression roc score=0.95

4.Logistic regression's best mode is= 'penalty':'l2',
'solver': 'newton-cg'

2. KNN(K-NEAREST NEIGHBOUR)

1. KNN cm= [40 11]

[30 52]

2.KNN Classification report

	precision	recall	F1 score	support
0	0.57	0.78	0.66	51
1	0.83	0.63	0.72	82
accuracy			0.69	133
Marco avg	0.70	0.71	0.69	133
Weighted avg	0.73	0.69	0.70	133

3.NAVIES BAYES

1.Navie bayes cm= [51 0]

[3 79]

2.Navie bayes clf report

	precision	recall	F1 score	support
0	0.94	1.00	0.97	51
1	1.00	0.96	0.98	82
accuracy			0.98	133
Marco avg	0.97	0.98	0.98	133
Weighted avg	0.98	0.98	0.98	133

4.DECISION TREE

1.Decision Tree cm [0 51]
[0 82]

2.Decision Tree classification report

	precision	recall	F1 score	support
0	0.00	0.00	0.00	51
1	0.62	1.00	0.76	82
accuracy			0.62	133
Macro avg	0.31	0.50	0.38	133
Weighted avg	0.38	0.62	0.77	133

3.Decision Tree Roc score = 0.5

4.Decision Tree best model is 'criterion': 'gini',
'max_features': 'sqrt',
'splitter': 'random'}:

5.RANDOM FOREST

1.Random forest cm= [0 51]
[0 82]

2.Random Forest classification report

	precision	recall	F1 score	support
0	0.00	0.00	0.00	51
1	0.62	1.00	0.76	82
accuracy			0.62	133
Marco avg	0.31	0.50	0.38	133
Weighted avg	0.38	0.62	0.47	133

3.Ranodm forest roc score 0.98

4.Random Forest best model is 'criterion': 'entropy',
'max_features': 'sqrt',
'n_estimators': 100}

6.SVM(SUPPORT VECTOR MACHINE)

1.Svm cm= [51 0]
[82 0]

2.Svm classification report

	precision	recall	F1 score	support
0	0.38	1.00	0.55	51
1	0.00	0.00	0.00	82

accuracy			0.38	133
Marco avg	0.19	0.50	0.28	133
Weighted avg	0.15	0.38	0.21	133

3.Svm best model is = 'C': 10,
gamma': 'auto',
'kernel': 'poly'}

6. FINAL MODEL

My final model is **NAVIES BAYES(GaussianNB)**. Because comparing all the clf report values with respective algorithms, NAVIES BAYES values are good. So finally select the NAVIES BAYES algorithm.

