CLASSIFICATION ALGORITHM-ASSIGNMENT

Problem Statement or Requirement:

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.

The Client has provided the dataset of the same.

1.) **Identify your problem statement:**

Stage 1: Machine Learning (Dataset in csv file)

Stage 2: Supervised Learning (Requirement is clear that client's wants to predict the Chronic Kidney Disease and while checking input and output, they clearly provide the information).

Stage 3: Classification. (because output of this dataset is categorical form).

2.) **Tell basic info about the dataset (Total number of rows, columns):**

Total number of rows: 399 rows.

Total number of columns: 25 columns.

It has 25 inputs as (Age, bp, sg, al, su, rbc, pc, pcc, ba, bgr, bu, sc, sod, pot, hrmo, pcv, wc, rc, htn, dm, cad, appet, pe, ane) and 1 output(classification).

3.) **Mention the pre-processing method if you're doing any (like converting string to number – nominal data):**

• Yes, we have to do pre-processing for this dataset because we have many categorical columns in this dataset.

4.) **Develop a good model with good evaluation metric. You can use any machine learning algorithm; you can create many models. Finally, you have to come up with final model.**

This data set is created the very good model in all algorithms.

Here we have same scores for three algorithms, they are:

- SVM
- Random Forest
- Logistic Regression

	A	B	C	D	E	F	G	H	I	J
1	**Evaluation metrics**	**SVM**	**Decision Tree**	**Random Forest**	**Logistic Regression**	**KNN**	**MultinomialNB**	**BernoulliNB**	**ComplementNB**	**GaussianNB**
2	what is the overall performance? (Accuracy)	0.99	0.96	0.99	0.99	0.96	0.98	0.98	0.98	0.98
3	What is the percentage of correctly classification of class 0 or not purchased to the total input of class 0 or not purchased in the test set (Recall)?	1	0.98	1	1	1	1	1	1	1
4	What is the percentage of correctly classification of class 1 or purchased to the total input of class 1 or purchased in the test set (Recall)?	0.99	0.95	0.99	0.99	0.94	0.98	0.96	0.98	0.96
5	what is the percentage of correct classification of class 0 or not purchased to sum of correctly and wrongly classified a class 0 or not-purchased in the test set (precision)?	0.98	0.93	0.98	0.98	0.91	0.96	0.94	0.96	0.94
6	what is the percentage of correct classification of class 1 or purchased to sum of correctly and wrongly classified a class 1 or purchased in the test set (precision)?	1	0.99	1	1	1	1	1	1	1
7	What is the overall performance of class 0 or not purchased (F1- Measure)?	0.99	0.95	0.99	0.99	0.95	0.98	0.97	0.98	0.97
8	What is the overall performance of class 0 or purchased (F1- Measure)?	0.99	0.97	0.99	0.99	0.97	0.99	0.98	0.99	0.98
9	what is the average performance of Recall?	0.99	0.97	0.99	0.99	0.97	0.99	0.98	0.99	0.98
10	what is the average performance of precision?	0.99	0.96	0.99	0.99	0.96	0.98	0.97	0.98	0.97
11	what is the average performance of F1-Measure?	0.99	0.96	0.99	0.99	0.96	0.98	0.98	0.98	0.98
12	What is the sum of product of proportion rate of each class (Recall)?	0.99	0.96	0.99	0.99	0.96	0.98	0.98	0.98	0.98
13	What is the sum of product of proportion rate of each class (precision)?	0.99	0.96	0.99	0.99	0.97	0.99	0.98	0.99	0.98
14	What is the sum of product of proportion rate of each class (F1-Measure)?	0.99	0.96	0.99	0.99	0.96	0.99	0.98	0.99	0.98

5.) All the research values of each algorithm should be documented. (You can make tabulation or screenshot of the results.)

Support Vector Machine (SVM):

```
The report:
              precision    recall  f1-score   support

           0       0.98      1.00      0.99        51
           1       1.00      0.99      0.99        82

    accuracy                           0.99       133
   macro avg       0.99      0.99      0.99       133
weighted avg       0.99      0.99      0.99       133
```

Receiver Operating Characteristic - Area Under the Curve (roc_auc):

1.0

Decision Tree:

The report:

	precision	recall	f1-score	support
0	0.93	0.98	0.95	51
1	0.99	0.95	0.97	82
accuracy			0.96	133
macro avg	0.96	0.97	0.96	133
weighted avg	0.96	0.96	0.96	133

Receiver Operating Characteristic - Area Under the Curve (roc_auc): 0.965805834528933

Random Forest:

The report:

	precision	recall	f1-score	support
0	0.98	1.00	0.99	51
1	1.00	0.99	0.99	82
accuracy			0.99	133
macro avg	0.99	0.99	0.99	133
weighted avg	0.99	0.99	0.99	133

Receiver Operating Characteristic - Area Under the Curve (roc_auc): 0.99976087996174

Logistic Regression:

The report:

	precision	recall	f1-score	support
0	0.98	1.00	0.99	51
1	1.00	0.99	0.99	82
accuracy			0.99	133
macro avg	0.99	0.99	0.99	133
weighted avg	0.99	0.99	0.99	133

Receiver Operating Characteristic - Area Under the Curve (roc_auc): 1.0

K-Nearest Neighbour:

The report:

	precision	recall	f1-score	support
0	0.91	1.00	0.95	51
1	1.00	0.94	0.97	82
accuracy			0.96	133
macro avg	0.96	0.97	0.96	133
weighted avg	0.97	0.96	0.96	133

Receiver Operating Characteristic - Area Under the Curve (roc_auc): 0.9995217599234816

Navie Bayes:

MultinominalNB:

	precision	recall	f1-score	support
0	0.96	1.00	0.98	51
1	1.00	0.98	0.99	82
accuracy			0.98	133
macro avg	0.98	0.99	0.98	133
weighted avg	0.99	0.98	0.99	133

Receiver Operating Characteristic - Area Under the Curve (roc_auc): 1.0

BernoulliNB:

	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
macro avg	0.97	0.98	0.98	133
weighted avg	0.98	0.98	0.98	133

Receiver Operating Characteristic - Area Under the Curve (roc_auc): 0.996652319464371

ComplementNB:

	precision	recall	f1-score	support
0	0.96	1.00	0.98	51
1	1.00	0.98	0.99	82
accuracy			0.98	133
macro avg	0.98	0.99	0.98	133
weighted avg	0.99	0.98	0.99	133

Receiver Operating Characteristic - Area Under the Curve (roc_auc):
1.0

GaussianNB:

	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
macro avg	0.97	0.98	0.98	133
weighted avg	0.98	0.98	0.98	133

Receiver Operating Characteristic - Area Under the Curve (roc_auc):
1.0

6.) Mention your final model, justify why u have chosen the same.

- As per our research **SVM** and **Logistic Regression** both are best model.
- Based on evaluation metric report(Accuracy:0.99) and roc_auc(1.0) we finalize these two alogorthim.