Classification 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

Machine Learning - Supervised – Classification

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

Rows - 399

Columns - 25

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

Yes, converting string column (rbc, pc, pcc, ba, htn, dm, cad, appet, pe, ane, classification) to numbers using **One hot encoding method and Standard scaler method** for standardising the 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.

These three Classification Algorithms gives the best result for the final model:

- **Logistic Regression**
- **Random Forest**
- **Support Vector Machine**

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

Decision Tree Classification:

```
The 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
   macro avg       0.97      0.98      0.98       133
weighted avg       0.98      0.98      0.98       133


The Confusion Matrix:
 [[51  0]
 [ 3 79]]
```

Roc_auc_score is: 0.9817073170731707

K-Nearest Neighbour Classification:

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

           0       0.86      1.00      0.93        51
           1       1.00      0.90      0.95        82

    accuracy                           0.94       133
   macro avg       0.93      0.95      0.94       133
weighted avg       0.95      0.94      0.94       133


The Confusion Matrix:
 [[51  0]
 [ 8 74]]
```

Roc_auc_score is: 0.9992826398852224

Logistic Regression Classification:

```
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


The Confusion Matrix:
 [[51  0]
 [ 1 81]]              Roc_auc_score is: 1.0
```

Naive Bayes Classification:

```
The 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
   macro avg       0.97      0.98      0.98       133
weighted avg       0.98      0.98      0.98       133


The Confusion Matrix:
 [[51  0]
 [ 3 79]]              Roc_auc_score is: 1.0
```

Random Forest Classification:

```
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


The Confusion Matrix:
 [[51  0]
 [ 1 81]]              Roc_auc_score is: 1.0
```

Support Vector Machine Classification:

```
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


The Confusion Matrix:
 [[51  0]
 [ 1 81]]              Roc_auc_score is: 1.0
```

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

The models created in **Random Forest Algorithm, Logistic Regression Algorithm and Support Vector Machine Algorithm** seems to be the best model as the accuracy in confusion matrix is nearer to 1 (For best model, accuracy should be nearer to 1) comparing to other algorithms, with less error and ***Receiver operating characteristic curve is also perfect*** for all three algorithms.