Classification Assignment

Dataset : CKD.csv

1.Problem Statement :

To create a predictive model which will predict if the patient will have chronic kidney disease based on several other health conditions.

2.Basic Info about dataset: 399 rows × 25 columns

3.Preprocessing Technique: As the value of the data are categorical and does not have any sense of order , we use One hot encoding.

4. Model creation

a.Logistic Regression Classifier:

```
[[43  2]
 [ 0 75]]
              precision    recall  f1-score   support

       False       1.00      0.96      0.98        45
        True       0.97      1.00      0.99        75

    accuracy                           0.98       120
   macro avg       0.99      0.98      0.98       120
weighted avg       0.98      0.98      0.98       120

The roc curve of the model for best parameter {'penalty': 'l2', 'solver': 'liblinear'}: 0.9976296296296296
The accuracy of the model for best parameter {'penalty': 'l2', 'solver': 'liblinear'}: 0.9833333333333333
```

b.Decision Tree Classifier:

```
[[45  0]
 [ 4 71]]
              precision    recall  f1-score   support

       False       0.92      1.00      0.96        45
        True       1.00      0.95      0.97        75

    accuracy                           0.97       120
   macro avg       0.96      0.97      0.97       120
weighted avg       0.97      0.97      0.97       120

The roc curve of the model for best parameter {'criterion': 'entropy', 'max_depth': 10, 'min_samples_split': 5, 'splitter': 'random'}: 0.9860740740740741
The accuracy of the model for best parameter {'criterion': 'entropy', 'max_depth': 10, 'min_samples_split': 5, 'splitter': 'random'}: 0.9666666666666667
```

c.Random Forest Classifier:

```
[[44  1]
 [ 1 74]]
              precision    recall  f1-score   support

       False       0.98      0.98      0.98        45
        True       0.99      0.99      0.99        75

    accuracy                           0.98       120
   macro avg       0.98      0.98      0.98       120
weighted avg       0.98      0.98      0.98       120

The roc curve of the model for best parameter {'criterion': 'gini', 'max_depth': 3, 'min_samples_split': 2, 'n_estimators': 100}: 0.9994074074074074
The accuracy of the model for best parameter {'criterion': 'gini', 'max_depth': 3, 'min_samples_split': 2, 'n_estimators': 100}: 0.9833333333333333
```

d.Support Vector Machine:

```
[[44  1]
 [ 4 71]]
              precision    recall  f1-score   support

       False       0.92      0.98      0.95        45
        True       0.99      0.95      0.97        75

    accuracy                           0.96       120
   macro avg       0.95      0.96      0.96       120
weighted avg       0.96      0.96      0.96       120

The accuracy of the model for best parameter {'C': 10, 'gamma': 'scale', 'kernel': 'linear'}: 0.9583333333333334
```

5.Final Model:

As the dataset is imbalanced, we can consider the value of roc_curve to find the best model.

Random Forest Classifier has highest roc curve with parameters criterion:gini, max_depth:3,min_samples_split:2,n_estimators:100 with 0.99940.

6.Questions:

1.How well is the model performing?

Ans : Accuracy: The model performs with the accuracy of 99%.

2.How much has model predicted correct classification to the total input of Chronic dieases yes value?

Ans: Recall: 99%

3.How much has the model predicted correct classification to the total input of chronic disease yes and no value?

Ans: Precision:99%

4.F1 score : 0.99, Macro Average=0.98, Weighted Average=0.98