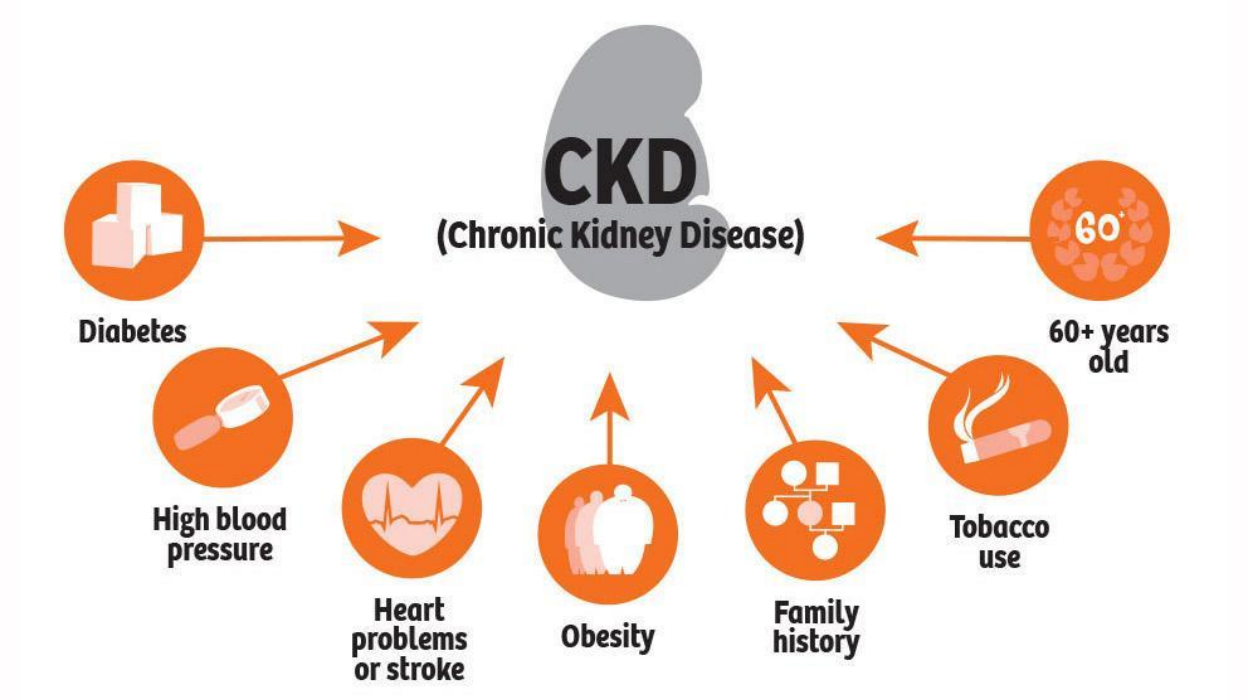
Chronic Kidney Decease



1. Problem statement: Prediction of chronic kidney decease from given dataset.

   (Machine learning-supervised learning-classification)

2. Dataset details: Number of Columns-25 & Number of Rows-399.

3. Data preprocessing performed for column name – rbc, pc, pcc, ba.

4. Machine learning algorithms for classification using random forest produce good results.

```
[[45  0]
 [ 2 73]]
              precision    recall  f1-score   support

           0       0.96      1.00      0.98        45
           1       1.00      0.97      0.99        75

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

0.9834018801410106
0.9997037037037038
```

5. Other classification algorithm research values:

   a. **Decision Tree:**

```
Fitting 3 folds for each of 12 candidates, totalling 36 fits
[[44  1]
 [ 3 72]]
              precision    recall  f1-score   support

           0       0.94      0.98      0.96        45
           1       0.99      0.96      0.97        75

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

0.9668037602820211
0.9688888888888889
```

   b. **Random Forest:**

```
[[45  0]
 [ 2 73]]
              precision    recall  f1-score   support

           0       0.96      1.00      0.98        45
           1       1.00      0.97      0.99        75

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

0.9834018801410106
0.9997037037037038
```

   c. **Logistic Regression:**

```
Fitting 3 folds for each of 4 candidates, totalling 12 fits
[[43  2]
 [ 0 75]]
              precision    recall  f1-score   support

           0       1.00      0.96      0.98        45
           1       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

0.9832535885167464
0.9973333333333334
```

d. Naïve Bayes:

```
Fitting 3 folds for each of 1 candidates, totalling 3 fits
[[45  0]
 [ 2 73]]
              precision    recall  f1-score   support

           0       0.96      1.00      0.98        45
           1       1.00      0.97      0.99        75

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

0.9834018801410106
1.0
```

e. K Nearest neighbors:

```
Fitting 3 folds for each of 4 candidates, totalling 12 fits
[[41  4]
 [22 53]]
              precision    recall  f1-score   support

           0       0.65      0.91      0.76        45
           1       0.93      0.71      0.80        75

    accuracy                           0.78       120
   macro avg       0.79      0.81      0.78       120
weighted avg       0.83      0.78      0.79       120

0.7866161616161614
0.8558518518518519
```

f. Support vector machine:

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

           0       0.98      0.98      0.98        45
           1       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

0.9833333333333335
0.9985185185185185
```

6. Final Model analysis:

The trained model's development indicated that the Random Forest algorithm delivered the best performance across all parameters.

- It predicts a 98% total model performance.
- Type I error in the confusion matrix is 0, and Type II error is 2. Here, type I error has a lower value compared to type II error, which suggests that predicting chronic renal disease is accurate.
- A slight increase or decrease in recall and precision values. Only F1 score is taken into consideration for the best outcome.
- The ROC value for this problem is 0.9997, or almost 1, indicating that the random forest algorithm is effective.