Classification Assignment

Problem Statement: **Date – 9.9.25**

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:

Client input – Dataset containing Several parameters that contains data about whether the patient has chronic kidney disease or not depending on the values of those parameters.

Solution – Build a model that would clearly predict if the patient has chronic kidney disease or not based on the input parameters.

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

Dataset – CKD.csv

No of Rows – 399

No of Columns – 25

```
]:  classification_yes
    1    249
    0    150
    Name: count, dtype: int64
```

Imbalanced Dataset

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

pandas.get_dummies() --- transforms categorical data into binary (0/1) "dummy" columns—one per unique value—making it easy for machine learning models to process.

 # Changing categorical data to Numerical data

dataset = pd.get_dummies(dataset, drop_first = True, dtype = int)

drop_first=True in pandas.get_dummies() removes the first dummy column (out of k) to prevent multicollinearity, by representing k categories using only k–1 binary feature.

Model Selection:

S.no	Model	Accuracy without grid search	Best Model with Grid Search	F1_Score	ROC AUC Score	Output prediction for sample i/p
1	Support Vector Machine	0.62	Coding running for too much time and not generating model			
2	Decision Tree Classifier	0.92 (output was predicted wrong with sample I/P)	{'criterion': 'gini', 'max_features': 'sqrt', 'splitter': 'random'}	0.9420122887864824	0.9444444444444445	Yes
3	Random Forest Classifier	0.97	{'criterion': 'entropy', 'max_features': 'sqrt', 'n_estimators': 100}	0.9916474440062505	1.0	Yes
4	Logistic Regression Classifier	0.93	{'penalty': 'l2', 'solver': 'liblinear'}	0.9916474440062505	0.9991111111111111	Yes
5	K Nearest Neighbor Classifier	0.71	{'algorithm': 'auto', 'n_neighbors': 5, 'weights': 'distance'}	0.76953125	0.8542222222222222	Yes

6	Gaussian Naïve Bayes	0.98	{'var_smoothing': 1e-09}	0.9834018801410106	1.0	Yes
7	Multinomial Naïve Bayes	--	{'alpha': 0.001, 'fit_prior': True}	0.9014285714285715	0.9525925925925927	Yes
8	Bernoulli Naïve Bayes	0.93	{'alpha': 0.001, 'binarize': 0.0, 'fit_prior': True}	0.9751481237656352	0.9967407407407407	Yes

Observations:

1. The accuracy value increases if grid search is done.

2. Decision Tree classifier predicted wrong for a sample I/P data when traditional method was followed but it predicted correct with improved accuracy after Grid search was executed

3. Best value for both F1_Score and ROC AUC from the above table for the given CKD dataset is:

Model - **Random Forest Classifier**

F1_Score = 0.9916474440062505

ROC AUC Value = 1.0
Best Parameters = {'criterion': 'entropy', 'max_features': 'sqrt', 'n_estimators': 100}