Assignment - Classification

Identify your problem statement

The objective is to develop a machine learning model that predicts the likelihood of Chronic Kidney Disease (CKD) based on various clinical and laboratory input features such as age, blood pressure, albumin levels, blood glucose, serum creatinine, hemoglobin, and other relevant medical attributes. This is a supervised learning classification problem, and the project can be titled "Chronic Kidney Disease Prediction.

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

The dataset consists of **400 rows and 25 columns**.The input features include various **clinical and laboratory parameters** such as age, blood pressure, blood glucose random (bgr), blood urea (bu), serum creatinine (sc), sodium (sod), potassium (pot), hemoglobin (hemo), packed cell volume (pcv), white blood cell count (wc), red blood cell count (rc), along with categorical attributes such as hypertension, diabetes mellitus, anemia, pus cell clumps, bacteria presence, and others.
The **target variable** is **classification of CKD** (1 = Chronic Kidney Disease, 0 = Not CKD).

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

The dataset contains several categorical variables such as **red blood cell (normal/abnormal)**, **pus cell (normal/abnormal)**, **pus cell clumps (present/not present)**, **bacteria (present/not present)**, **hypertension (yes/no)**, **diabetes mellitus (yes/no)**, **coronary artery disease (yes/no)**, **appetite (good/poor)**, **pedal edema (yes/no)**, and **anemia (yes/no)**.

Since these variables are **categorical without any sequential order**, I applied **nominal encoding** methods. Specifically, I used **one-hot encoding (via** pd.get_dummies()**)** to convert these categorical values into numerical format so they can be processed by machine learning algorithms.

Algorithms with Report:

Logistic Regression:

	precision	recall	f1-score	support
0	0.71	1.00	0.83	51
1	1.00	0.74	0.85	82
accuracy			0.84	133
macro avg	0.85	0.87	0.84	133
weighted avg	0.89	0.84	0.84	133

Class 0 (No CKD): Perfect recall (1.0), but lower precision (0.71).
Class 1 (CKD): Perfect precision (1.0), but recall is lower (0.74), meaning some CKD cases are missed.
Overall accuracy: 84%.

SVM:

	precision	recall	f1-score	support
0	0.00	0.00	0.00	51
1	0.62	1.00	0.76	82
accuracy			0.62	133
macro avg	0.31	0.50	0.38	133
weighted avg	0.38	0.62	0.47	133

Class 0 (No CKD): Precision = 0.00, Recall = 0.00 → The model failed to identify any non-CKD cases (all misclassified).
Class 1 (CKD): Precision = 0.62, Recall = 1.00 → It caught all CKD cases, but with moderate precision (some false positives).
Accuracy = 62% → Much lower than before (84%).
Macro avg (0.38 F1) → Overall poor balance across both classes.
Weighted avg (0.47 F1) → Performance heavily biased toward CKD class.

Decision Tree:

	precision	recall	f1-score	support
0	0.00	0.00	0.00	51
1	0.62	1.00	0.76	82
accuracy			0.62	133
macro avg	0.31	0.50	0.38	133
weighted avg	0.38	0.62	0.47	133

Class 0 (No CKD): Precision = 0.00, Recall = 0.00 → It failed to detect any non-CKD patients.
Class 1 (CKD): Precision = 0.62, Recall = 1.00 → It detects all CKD patients, but only 62% of those predictions are actually correct.
Accuracy = 62% → Not reliable, since it ignores one whole class.
Macro/Weighted avg F1-scores are very low → confirms poor balance across classes.

Random Forest:

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

- **Class 0 (No CKD):** Precision = 0.98, Recall = 0.98 → Hardly misses any non-CKD cases.
- **Class 1 (CKD):** Precision = 0.99, Recall = 0.99 → Almost perfect detection of CKD patients.
- **Accuracy = 98%** → Very high overall performance.
- **Macro & Weighted avg F1 = 0.98** → Balanced performance across both classes.

Conclusion:

We experimented with multiple algorithms to predict Chronic Kidney Disease (CKD). **Logistic Regression** achieved an accuracy of **84%**, showing good balance between CKD and non-CKD cases. However, **Decision Tree** and **SVM** both performed poorly with an accuracy of only **62%**, as they were biased towards predicting only CKD cases and failed to classify non-CKD patients effectively. On the other hand, **Random Forest** delivered the best performance with an outstanding **98% accuracy**, along with excellent precision, recall, and F1-scores across both classes, making it the most reliable model for CKD prediction.