Classification Assignment-Solution

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

Step1: Domain selection: Machine learning

Step 2: Learning Selection: Supervised Learning

Step 3: Classification

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

The dataset initially consists of 399 rows and 25 columns.

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

After applying pd.get_dummies() with drop_first=True, the dataset now has 399 rows and 28 columns. It is like converting categorical data into numeric data.

```
[5]: dataset.shape
[5]: (399, 25)
[6]: dataset=pd.get_dummies(dataset,dtype=int,drop_first=True)
[10]: dataset.shape
[10]: (399, 28)
```

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.

Between Random Forest, SVM, and Logistic Regression (after standardization), all achieve **99% accuracy**.
However, since the selection criterion based on confusion matrix which I have taken, is that **Type I error should be lower than Type II**,
so the **final model should be either SVM or Logistic Regression (Standardized)**.

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

Algorithm	Precision (Class 0)	Recall (Class 0)	F1-Score (Class 0)	Precision (Class 1)	Recall (Class 1)	F1-Score (Class 1)	Accuracy	Macro Avg (F1)	Weighted Avg (F1)
Logistic Regression (without Standardization)	0.89	0.92	0.90	0.95	0.93	0.94	**0.92**	0.92	0.93
Logistic Regression (with Standardization)	0.98	1.00	0.99	1.00	0.99	0.99	**0.99**	0.99	0.99
Multinomial Naive Bayes	0.68	0.98	0.81	0.98	0.72	0.83	**0.82**	0.82	0.82
Bernoulli Naive Bayes	0.86	1.00	0.93	1.00	0.90	0.95	**0.94**	0.94	0.94
Complement Naive Bayes	0.68	0.98	0.81	0.98	0.72	0.83	**0.82**	0.82	0.82
Decision Tree Classifier	0.91	0.96	0.93	0.97	0.94	0.96	**0.95**	0.94	0.95
K-Nearest Neighbors (KNN)	0.57	0.78	0.66	0.83	0.63	0.72	**0.69**	0.69	0.70
Random Forest Classifier	1.00	0.98	0.99	0.99	1.00	0.99	**0.99**	0.99	0.99
Support Vector Machine (SVM)	0.98	1.00	0.99	1.00	0.99	0.99	**0.99**	0.99	0.99

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

SVM was chosen as the final model because it achieved excellent classification performance with balanced Type I and Type II errors. Since the dataset is of small-to-medium scale, SVM is well-suited due to its ability to create optimal decision boundaries and generalize effectively. It also performed equally well as Logistic Regression after

standardization but provides better robustness and margin-based separation between classes, making it a more reliable choice for this dataset.