Classification Assignment

1. Problem Statement:

Develop a predictive machine learning model that identifies Chronic Kidney Disease (CKD) in patients using clinical parameters provided by hospital management. The goal is to assist medical staff in early diagnosis and effective management of CKD.

2. Dataset Overview:

- 399 rows × 28 columns

3. Encoding Categorical Data:

Converted columns from string to numbers using one-hot encoding.

Handling Missing Data:

Verified there are no missing values; ready for modeling.

4. Model Development and Evaluation:

Experimented with the following machine learning algorithms using Python's sklearn and libraries.

5. The research values:

- **Logistic Regression - GRID**

```
[17]: print("The confusion Matrix:\n",cm)

The confusion Matrix:
 [[51  0]
 [ 1 81]]

[18]: print("The report:\n",clf_report)

The report:
              precision    recall  f1-score   support

       False       0.98      1.00      0.99        51
        True       1.00      0.99      0.99        82

    accuracy                           0.99       133
   macro avg       0.99      0.99      0.99       133
weighted avg       0.99      0.99      0.99       133
```

- **Decision Tree-GRID:**

```
[16]: print("The confusion Matrix:\n",cm)

The confusion Matrix:
 [[50  1]
 [ 1 81]]

[17]: print("The report:\n",clf_report)

The report:
              precision    recall  f1-score   support

       False       0.98      0.98      0.98        51
        True       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
```

- **Random Forest-GRID:**

```
print("The confusion Matrix:\n",cm)

The confusion Matrix:
 [[ 0 51]
 [ 0 82]]

print("The report:\n",clf_report)

The report:
              precision    recall  f1-score   support

       False       0.00      0.00      0.00        51
        True       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
```

- **Support Vector Machine (SVM)-GRID**

```
print("The confusion Matrix:\n",cm)

The confusion Matrix:
 [[51  0]
 [ 2 80]]

print("The report:\n",clf_report)

The report:
              precision    recall  f1-score   support

           0       0.96      1.00      0.98        51
           1       1.00      0.98      0.99        82

    accuracy                           0.98       133
   macro avg       0.98      0.99      0.98       133
weighted avg       0.99      0.98      0.99       133
```

6. Final Model Selection:

Logistic Regression -GRID – Classifier hasthe highest overall accuracy (99%) and excellent precision, recall, and F1 scores balanced perfectly across classes.

Model for Deployment:

- The model represented is the best suited for deployment due to its highest accuracy, balanced precision/recall, and near-perfect F1 scores.
- This model ensures minimal false positives/negatives and is robust across both classes.

Report 4 is considered the best model because it demonstrates near-perfect performance across all important classification metrics, indicating that it reliably distinguishes between classes with minimal errors. Here's why:

1. Extremely High Precision (False: 0.98, True: 1.00)

- Precision measures the accuracy of positive predictions.
- A precision of 1.00 for the True class means no false positives for that class, i.e., every sample predicted as True is actually True.
- For the False class, 0.98 precision means very few incorrect positive predictions.

2. Almost Perfect Recall (False: 1.00, True: 0.99)

- Recall measures the model's ability to identify all actual positives.
- Recall of 1.00 for the False class means the model identifies all false class samples correctly (no false negatives).
- Recall of 0.99 for the True class shows it almost perfectly detects true positives, missing very few.

3. High F1-Score for Both Classes (0.99)

- The F1-score balances precision and recall.
- Scores of 0.99 mean the model has excellent balance, minimizing both false positives and false negatives.

4. High Overall Accuracy (0.99)

- The model correctly predicts 99% of all samples, demonstrating very strong generalization.

5. Balance Across Classes

- The macro and weighted averages (~0.99) confirm the model performs well on both classes equally, avoiding bias toward one class.
- Balanced performance is crucial especially if both classes are important.

Summary

These metrics collectively show the model:

- Minimizes both types of classification errors.
- Detects and classifies samples from both classes with great reliability.
- Is well-generalized and not overfitted to just one class.

Due to such strong, consistent, and balanced results, model is the best choice for deployment to ensure accuracy and reliability in real-world applications.