**Chronic Kidney Disease (CKD) Predictive Model**

**SUMMARY**

**Objective:**Build a reliable model to predict CKD presence from routine clinical parameters.

**Best Model:**LogisticRegression with a preprocessing pipeline (imputation, scaling, one-hot encoding).

**PROBLEM STATEMENT & DATASET:**

**Problem:**Predict CKD (Yes/No) to aid early detection and resource allocation.

**Dataset:**399 rows, 28 columns, target = classification.

**Numeric features Categorical features**

**16 11**

**Cleaning Dataset:**

1. Yes/No Variants.
2. **Preprocessing Pipeline:** numeric and standardization; categorical → one-hot encoding.
3. **Modeling:** trained Logistic Regression, SVC (RBF), and Random Forest on stratified train/test split.
4. **Evaluation:** computed Accuracy, Precision, Recall, F1, ROC AUC on held-out test set; inspected confusion matrix and ROC curve.
5. **Selection:** chose the simplest model with Best Performance and best clinical interpretability.

**RESULTS:**

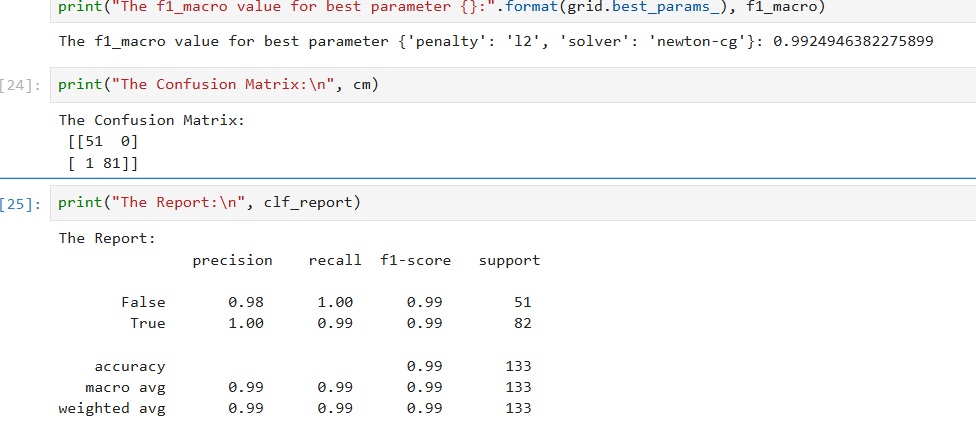
**Model Comparison**

| **model** | **accuracy** | **precision** | **recall** | **f1** | **ROC\_AUC** |
| --- | --- | --- | --- | --- | --- |
| LogisticRegression | 0.99 | 1.0 | 0.98 | 0.9899 | 1.0000 |
| SVC | 0.99 | 1.0 | 0.98 | 0.9899 | 1.0000 |
| RandomForest | 0.9875 | 1.0 | 0.98 | 0.9899 | 0.9993 |

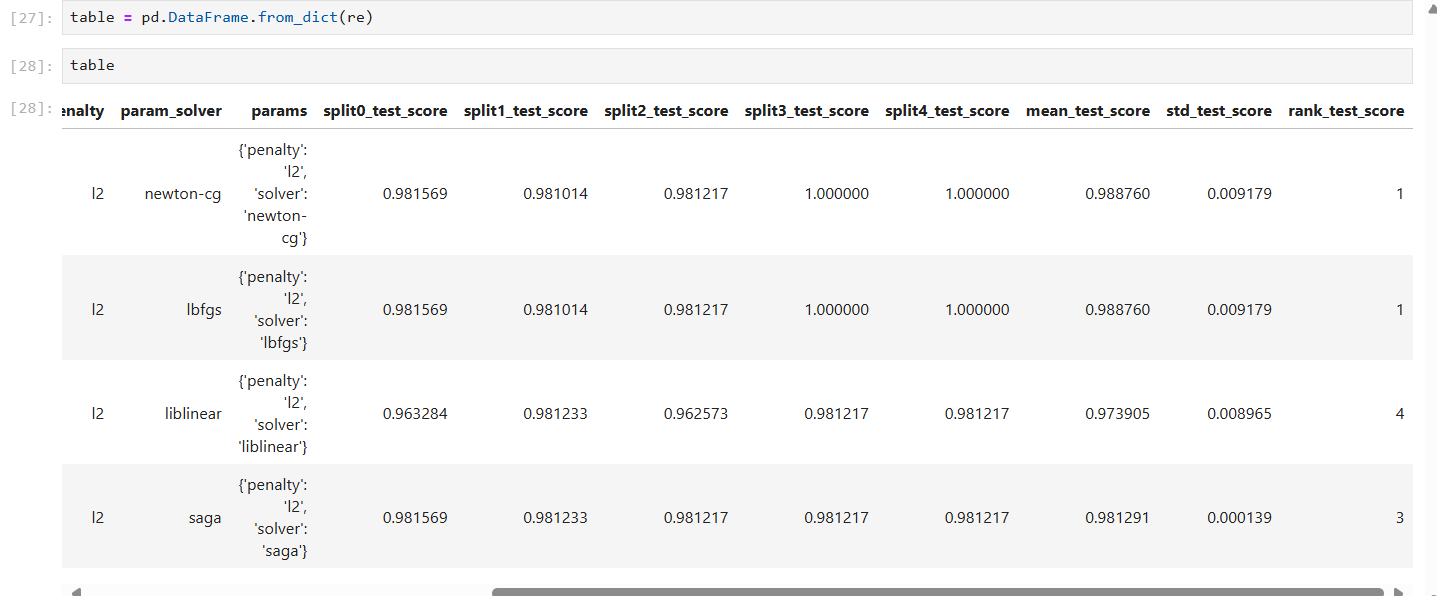
**CONFUSION MATRIX:**

|  | **Pred 0** | **Pred 1** |
| --- | --- | --- |
| **Actual 0** | 51 | 0 |
| **Actual 1** | 1 | 81 |

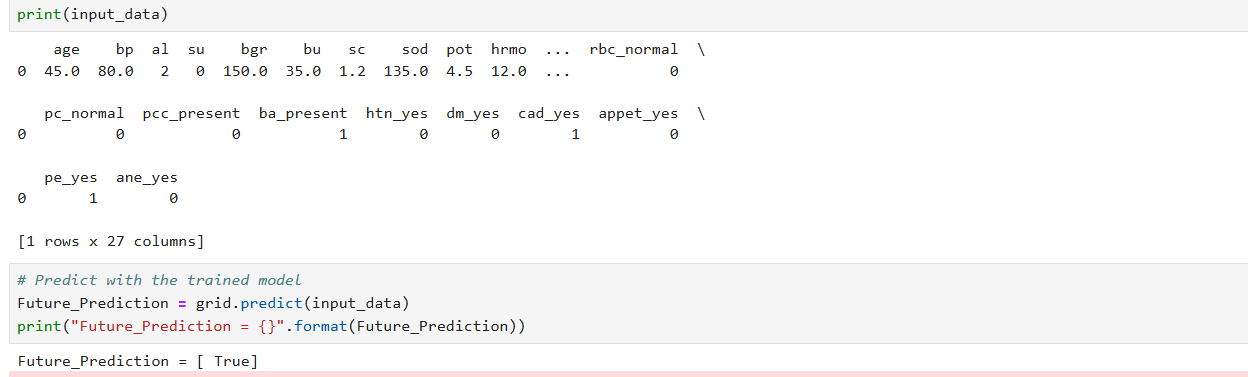
**FINAL REPORT [Logistic Regression]:**

****

**Grid Table [Logistic Regression]:**

****

**FUTURE PREDICTIONS: Future\_Prediction = [1]**

****

**Why Choosing Logistic Regression:**

We chose **Logistic Regression** because the **CKD prediction is a binary classification** problem. Logistic Regression is interpretable, **provides probabilistic outputs, handles both numerical and categorical data efficiently**, and works well with relatively small medical datasets. Most importantly, it allows clinicians to understand the impact of each medical parameter on the disease outcome, which is essential in healthcare applications.

Unlike complex models (RandomForest, Neural Networks), logistic regression is **less prone to overfitting** on small datasets.

**SAVED MODEL:**

****

**DEPLOYMENT & HANDOVER:**

* **Model:** finalized\_model\_CKD\_LogisticRegression\_classification.sav

**Algorithms:**

* **Metrics:** model comparison, test report, confusion matrix
* **Figure:** ROC curve