Chronic Kidney Disease (CKD) Prediction

Problem Statement:

Chronic Kidney Disease (CKD) is a significant health concern characterized by a steady increase in cases. Since the human body can only survive without functioning kidneys for an average of 18 days, there is a high demand for kidney transplants and dialysis treatments. Early prediction of CKD is crucial for effective management. Machine learning techniques offer effective tools for predicting CKD status based on clinical data.

Machine Learning → Supervised Learning→ Classification(categorical value)

Total number of rows, columns:

399 rows × 25 columns

24 columns as input parameters

01 columns as output parameters

data types: float64(10), int64(4), object(11)

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

Ordinal data refers to categorical data with a specific order or ranking associated with its values. When dealing with ordinal data in machine learning, label encoding algorithms are commonly used to convert string representations of categories into numerical representations while preserving their ordinal relationships.

Choosing Best Model

Algorithms (Classifier)	Confusion matrix	Classification report and	Roc score
Support Vector Machine	[[51 0] [0 82]]	precision recall f1-score support 0 1.00 1.00 1.00 51 1 1.00 1.00 1.00 82 accuracy 1.00 133 macro avg 1.00 1.00 1.00 133 weighted avg 1.00 1.00 1.00 133 **The f1_macro value for best parameter {'C': 10, 'gamma': 'scale', 'kernel': 'poly'}: 1.0**	**1.0**
Decision Tree	[[47 4] [0 82]]	The report: precision recall f1-score support 0 1.00 0.92 0.96 51 1 0.95 1.00 0.98 82 accuracy 0.97 133 macro avg 0.98 0.96 0.97 133 weighted avg 0.97 0.97 0.97 133 **The f1_macro value for best parameter {'criterion': 'entropy', 'max_features': None, 'splitter': 'random'}: 0.9696690706357731**	**0.9607843 13725490**
Random Forest	**[[51 0] [0 82]]**	precision recall f1-score support 0 1.00 1.00 1.00 51 1 1.00 1.00 1.00 82 accuracy 1.00 133 macro avg 1.00 1.00 1.00 133 weighted avg 1.00 1.00 1.00 133 **The f1_macro value for best parameter {'criterion': 'log_loss', 'max_features': 'sqrt', 'n_estimators': 100}: 1.0**	**0.8640602 58249641**

Logistics Regression Classifier	[[51 0] [3 79]]	precision recall f1-score support 0 0.94 1.00 0.97 51 1 1.00 0.96 0.98 82 accuracy 0.98 133 macro avg 0.97 0.98 0.98 133 weighted avg 0.98 0.98 0.98 133 **The f1_macro value for best parameter {'penalty': 'l2', 'solver': 'newton-cg'}: 0.9775556904684072**	**0.999043519846963**
KNN	[[51 0] [5 77]]	precision recall f1-score support 0 0.91 1.00 0.95 51 1 1.00 0.94 0.97 82 accuracy 0.96 133 macro avg 0.96 0.97 0.96 133 weighted avg 0.97 0.96 0.96 133 **The f1_macro value for best parameter {'algorithm': 'auto', 'metric': 'chebyshev', 'n_neighbors': 5, 'weights': 'distance'}: 0.9626932787797391**	**0.5**
Navies bayes [Gaussian Naive Bayes]	[[51 0] [6 76]]	precision recall f1-score support 0 0.89 1.00 0.94 51 1 1.00 0.93 0.96 82 accuracy 0.95 133 macro avg 0.95 0.96 0.95 133 weighted avg 0.96 0.95 0.96 133	**1.0**
(MultinomialNB)	[[50 1] [23 59]]	precision recall f1-score support 0 0.68 0.98 0.81 51 1 0.98 0.72 0.83 82 accuracy 0.82 133 macro avg 0.83 0.85 0.82 133 weighted avg 0.87 0.82 0.82 133	**0.901721664275466 4**

(BernoulliNB)	[[51 0] [10 72]]	precision recall f1-score support 0 0.84 1.00 0.91 51 1 1.00 0.88 0.94 82 accuracy 0.92 133 macro avg 0.92 0.94 0.92 133 weighted avg 0.94 0.92 0.93 133	**0.9774031563845051**
(ComplementNB)	[[50 1] [23 59]]	precision recall f1-score support 0 0.68 0.98 0.81 51 1 0.98 0.72 0.83 82 accuracy 0.82 133 macro avg 0.83 0.85 0.82 133 weighted avg 0.87 0.82 0.82 133	**0.9017216642754664**

Mention your final model, justify why u have chosen the same:

Final Model:

Support Vector Machine

The f1_macro value for best parameter {'C': 10, 'gamma': 'scale', 'kernel': 'poly'}: 1.0

Roc score=1.0

Accuracy 100%

[OR]

Random Forest

The f1_macro value for best parameter {'criterion': 'log_loss', 'max_features': 'sqrt', 'n_estimators': 100}: 1.0

Roc score=0.864060258249641

Accuracy 100%

This study proposes a systematic approach that includes data preprocessing and selecting relevant attributes. Among the 9 machine learning methods tested**, the Support Vector Machine and random forest classifier demonstrate the highest accuracy.** Furthermore, the research emphasizes the importance of incorporating domain knowledge during data collection and analysis to enhance the reliability of machine learning models for predicting CKD status.