Chronic Kidney Disease (CKD) Prediction Model Documentation

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.Problem Statement Identification

Title: Chronic Kidney Disease Predictor

Dataset ➽ Numerical Dataset

Supervised Learning ➽ Input and Output data are clearly defined.

Classification ➽ Target variable ➽ Categorical

2.Basic Information About the Dataset

- **Total Number of Rows:** 399
- **Total Number of Columns:** 25

3. Data Preprocessing

For this project, the following preprocessing steps were implemented:

- **One-Hot Encoding:** convert categorical string data into numerical data (nominal data)
- **Standard Scaling:** feature scaling technique to standardize the feature of dataset, ensuring that they have a mean of zero and a standard deviation of one.

4.Model Development and Evaluation Metrics

Multiple machine learning algorithms were applied to develop the predictive model. The performance of each model was evaluated using the **f1 score** and the **ROC_AUC score**.

The model with the best performance based on these metrics was selected as the final model.

Model Performance Summary:

Classification Algorithm	f1 Score	ROC_AUC Score
Logistic Regression	0.99	1
SVM	0.98	1
Random Forest	0.98	1
Decision Tree	0.97	0.97
K Nearest Neighbors	0.96	1
Gaussian NB	0.98	1
Complement NB	0.83	0.93
Categorical NB	1	1
Bernoulli NB	0.98	1
Multinomial NB	0.83	0.93

5.Research Findings

All the research and performance metrics for each algorithm were thoroughly documented. Below is a summary of the results obtained:

- **Logistic Regression:** ROC-AUC Score = 1.00, F1 Score = 0.99
- **SVM:** ROC-AUC Score = 1.00, F1 Score = 0.98
- **Random Forest:** ROC-AUC Score = 1.00, F1 Score = 0.98
- **Decision Tree:** ROC-AUC Score = 0.97, F1 Score = 0.97
- **KNN:** ROC-AUC Score = 1.00, F1 Score = 0.96
- **Gaussian NB:** ROC-AUC Score = 1.00, F1 Score = 0.98
- **Complement NB:** ROC-AUC Score = 0.93, F1 Score = 0.83
- **Bernoulli NB:** ROC-AUC Score = 1.00, F1 Score = 0.98
- **Multinomial NB:** ROC-AUC Score = 0.93, F1 Score = 0.83
- **Categorical NB:** ROC-AUC Score = 1, F1 Score = 1 (without train & test set)

6.Final Model Selection

The **Logistic Regression** model was chosen as the final model because it exhibited the highest f1 score and ROC-AUC score among all the models evaluated. These metrics indicate that the Logistic Regression model provides the most accurate and reliable predictions for CKD.

GitHub Link for Source Code: GitHub