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

**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**

**Supervised Learning:** Input and Output data are clearly defined.

**Classification**: **Target variable: Categorical Data**

**2.Basic Information About the Dataset:**

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

**3. Data Preprocessing:**

* **One-Hot Encoding:** convert categorical string data into numerical data (nominal data)
* **Standard Scaling**: (to define a range between e.g. -2 to 2) 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**.

**F1 Score:**

* **Harmonic mean of Precision and Recall**
* Useful when need a **balance between Precision and Recall**, especially for **imbalanced classes** (e.g., more negatives than positives)
* **Formula**:

F1 = (2 \* Precision \* Recall) / (Precision + Recall)

**ROC AUC Score:**

* **ROC –** Receiver Operating Characteristic.
* **AUC –** Area Under the Curve
* **ROC AUC Score –** measures the model’s ability to distinguish between classes and scores ranges from **0 to 1.**
  + **1.0 =** perfect classifier (high score – very good model)
  + **0.5 =** random guess (threshold value)
  + **< 0.5 =** worse than random

**5.Research Findings:**

|  |  |  |  |
| --- | --- | --- | --- |
| **NO** | **Classification Algorithms** | **F1 Score** | **ROC\_AUC Score** |
| 1 | Logistic Regression | 0.99 | 1 |
| 2 | Support Vector Machine | 0.98 | 1 |
| 3 | Random Forest | 0.97 | 1 |
| 4 | Decision Tree | 0.97 | 0.97 |
| 5 | K Nearest Neighbors | 0.95 | 1 |
| 6 | Gaussian NB | 0.98 | 1 |
| 7 | Complement NB | 0.83 | 0.93 |
| 8 | Categorical NB (no train & test set) | 1 | 1 |
| 9 | Bernoulli NB | 0.98 | 0.99 |
| 10 | Multinomial NB | 0.870.9 | 0.95 |

**6.Final Model Selection:**

* The **Logistic Regression model** was chosen as the **final model** because it exhibited the **highest** **f1 score (0.99) and ROC-AUC Score (1)** 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**