**Classification Assignment**

**Step 1:** Identify the Problem Statement

The objective is to develop a predictive model to identify whether a patient has Chronic Kidney Disease (CKD) based on several parameters (such as age, blood pressure, blood sugar levels, and other medical indicators).

Type of Problem: Binary Classification

Target variable: class (CKD or not CKD)

Goal: Predict if a patient has CKD (1) or does not have CKD (0).

**Step 2:** Load and Explore the Dataset

A screenshot of a computer

AI-generated content may be incorrect.

**Total Rows:** The number of patient records = 399

**Total Columns:** The features available for prediction (including the target label) = 28

**Step 3:** Pre-processing the Data

**Encoding Categorical Variables**

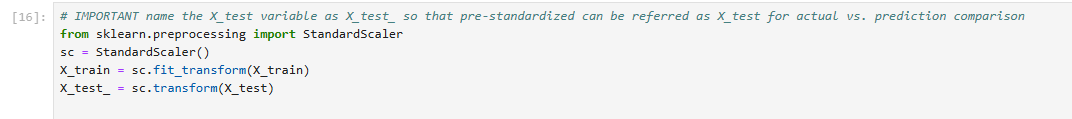
* Convert categorical data to numerical using **One-Hot Encoding**.

A screenshot of a computer

AI-generated content may be incorrect.

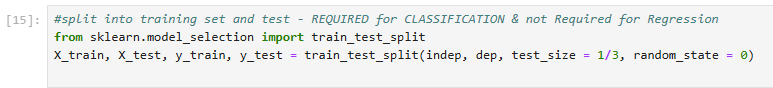
**Feature Scaling**

* Use **StandardScaler** to scale the numerical features.



**Step 4:** Split the Dataset and Train Multiple Models

Split the dataset into training and testing sets



Train Multiple Models

1. Logistic Regression

A screenshot of a computer

AI-generated content may be incorrect.

1. Support Vector Machine (SVM)

A screenshot of a computer program

AI-generated content may be incorrect.

1. Decision Tree

A screenshot of a computer program

AI-generated content may be incorrect.

1. Random Forest

A screenshot of a computer program

AI-generated content may be incorrect.

**Step 5:** Evaluate and Document the Results

**Model Performance Comparison Table**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-Score** | **AUC Score** |
| Logistic Regression | 0.99 | 1.00 | 0.99 | 0.99 | 0.95 |
| SVM | 0.99 | 1.00 | 0.99 | 0.99 | 0.46 |
| Decision Tree | 0.97 | 1.00 | 0.95 | 0.97 | 0.48 |
| Random Forest | 0.98 | 0.99 | 0.99 | 0.99 | 0.99 |

**Step 6:** Select the Best Model / Justification

**Best Model: Random Forest**

It achieves **strong overall metrics**:

* Accuracy: **0.98;** Precision: **0.99;** Recall: **0.99;** F1-Score: **0.99**

It has the **highest AUC score (0.99)**, indicating it has the best ability to distinguish between positive and negative classes.

Even though Logistic Regression has slightly higher accuracy, **Random Forest's superior AUC** makes it the best choice overall, as AUC is a more reliable metric for assessing model performance in classification tasks.

This makes it the **most robust and reliable model** for predicting Chronic Kidney Disease (CKD) in this scenario.