**Classification Assignment**

**1.** **Problem Statement**

Predict the Chronic Kidney Disease (CKD) on the several parameters.

**2.** **Dataset Information**

Dataset contains 27 columns and 400 rows of data values including header

**3.** **Data Preprocessing**

In the given dataset, there are twelve columns contains string values. Since Python AI models cannot directly handle categorical data, so those column’s values will be converted into meaningful numerical data.

4. R&D on Model Selection

1. Logistics Regression

The Mean test score value obtained in the Logistics Regression model is **0.96**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Sl No | mean\_fit\_time | std\_fit\_time | mean\_score\_time | std\_score\_time | param\_penalty | param\_solver | params | split0\_test\_score | split1\_test\_score | split2\_test\_score | split3\_test\_score | split4\_test\_score | mean\_test\_score | std\_test\_score | rank\_test\_score |
| 0 | 0.013201 | 0.001621 | 0.002909 | 0.000971 | l2 | lbfgs | {'penalty': 'l2', 'solver': 'lbfgs'} | 0.851852 | 0.942166 | 0.925146 | 0.886792 | 0.961826 | 0.913557 | 0.039513 | 4 |
| 1 | 0.005597 | 0.002373 | 0.001653 | 0.000728 | l2 | liblinear | {'penalty': 'l2', 'solver': 'liblinear'} | 0.962963 | 0.981014 | 0.962573 | 0.943651 | 0.961826 | 0.962405 | 0.011821 | 3 |
| 2 | 0.033414 | 0.004342 | 0.002543 | 0.002102 | l2 | newton-cg | {'penalty': 'l2', 'solver': 'newton-cg'} | 0.981569 | 0.981014 | 0.962573 | 0.962264 | 0.961826 | 0.969849 | 0.009347 | 1 |
| 3 | 0.004749 | 0.003465 | 0.001214 | 0.000188 | l2 | newton-cholesky | {'penalty': 'l2', 'solver': 'newton-cholesky'} | 0.981569 | 0.981014 | 0.962573 | 0.962264 | 0.961826 | 0.969849 | 0.009347 | 1 |
| 4 | 0.006982 | 0.001705 | 0.001139 | 0.00002 | l2 | sag | {'penalty': 'l2', 'solver': 'sag'} | 0.486532 | 0.50141 | 0.477841 | 0.477841 | 0.477841 | 0.484293 | 0.009196 | 5 |
| 5 | 0.006413 | 0.001674 | 0.001185 | 0.000137 | l2 | saga | {'penalty': 'l2', 'solver': 'saga'} | 0.486532 | 0.50141 | 0.477841 | 0.477841 | 0.477841 | 0.484293 | 0.009196 | 5 |

1. K-Neighbors Classifier

The Mean test score value obtained in the Logistics Regression model is **0.75**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Sl No** | **mean\_fit\_time** | **std\_fit\_time** | **mean\_score\_time** | **std\_score\_time** | **param\_algorithm** | **param\_weights** | **params** | **split0\_test\_score** | **split1\_test\_score** | **split2\_test\_score** | **split3\_test\_score** | **split4\_test\_score** | **mean\_test\_score** | **std\_test\_score** | **rank\_test\_score** |
| **0** | 0.003734 | 0.001524 | 0.004948 | 0.001698 | auto | uniform | {'algorithm': 'auto', 'weights': 'uniform'} | 0.744423 | 0.685491 | 0.664016 | 0.702707 | 0.677527 | 0.694833 | 0.027776 | 5 |
| **1** | 0.00317 | 0.000541 | 0.004565 | 0.002366 | auto | distance | {'algorithm': 'auto', 'weights': 'distance'} | 0.706269 | 0.685491 | 0.701348 | 0.758432 | 0.738011 | 0.71791 | 0.026483 | 1 |
| **2** | 0.00223 | 0.001132 | 0.003722 | 0.001195 | ball\_tree | uniform | {'algorithm': 'ball\_tree', 'weights': 'uniform'} | 0.744423 | 0.685491 | 0.664016 | 0.702707 | 0.677527 | 0.694833 | 0.027776 | 5 |
| **3** | 0.002457 | 0.000773 | 0.002859 | 0.000892 | ball\_tree | distance | {'algorithm': 'ball\_tree', 'weights': 'distance'} | 0.706269 | 0.685491 | 0.701348 | 0.758432 | 0.738011 | 0.71791 | 0.026483 | 1 |
| **4** | 0.001203 | 0.000063 | 0.002294 | 0.000052 | kd\_tree | uniform | {'algorithm': 'kd\_tree', 'weights': 'uniform'} | 0.744423 | 0.685491 | 0.664016 | 0.702707 | 0.677527 | 0.694833 | 0.027776 | 5 |
| **5** | 0.001372 | 0.000474 | 0.001644 | 0.000289 | kd\_tree | distance | {'algorithm': 'kd\_tree', 'weights': 'distance'} | 0.706269 | 0.685491 | 0.701348 | 0.758432 | 0.738011 | 0.71791 | 0.026483 | 1 |
| **6** | 0.001488 | 0.000376 | 0.002854 | 0.000568 | brute | uniform | {'algorithm': 'brute', 'weights': 'uniform'} | 0.744423 | 0.685491 | 0.664016 | 0.702707 | 0.677527 | 0.694833 | 0.027776 | 5 |
| **7** | 0.000952 | 0.000097 | 0.001455 | 0.000047 | brute | distance | {'algorithm': 'brute', 'weights': 'distance'} | 0.706269 | 0.685491 | 0.701348 | 0.758432 | 0.738011 | 0.71791 | 0.026483 | 1 |

1. Naive Bayes – GaussianNB

The GaussianNB model is **0.98**

A number of numbers in a row

AI-generated content may be incorrect.

1. Naive Bayes – MultinomialNB

A screenshot of a graph

AI-generated content may be incorrect.The MultinomialNB model is **0.83**

1. Naive Bayes – ComplementNB

A close-up of numbers

AI-generated content may be incorrect.The ComplementNB model is **0.95**

1. Random Forest

The Mean test score value obtained in the Logistics Regression model is **0.98**

A table of numbers and letters

AI-generated content may be incorrect.*('class\_weight': 'balanced', 'criterion': 'gini', 'max\_features': 'log2')*

1. Random Forest

The Mean test score value obtained in the Logistics Regression model is **0.95**

*A white sheet with black and green text

AI-generated content may be incorrect.('criterion': 'gini', 'max\_features': 'sqrt', 'splitter': 'random')*

In the above experiment, based on the accuracy score, the **Random Forest** (*'class\_weight': 'balanced', 'criterion': 'gini', 'max\_features': 'log2'*) model appears to be the best-performing model.