**Machine Learning Classification Assignment**

**Problem Statement or Customer Requirement:**

A client’s requirement is, to create a predictive model which will predict the Chronic Kidney Disease (CKD) based on the several input parameters, The Client has provided the dataset of the same.

**Identifying problem statement**

Stage -1: AI Domain selected here is – Machine Learning

Stage- 2: Learning selection is – Supervised Learning

Stage -3: Based on patient’s input health details, we need to predict whether they are affected by Chronic Kidney Disease or NOT.

Hence, we decided to create Machine Learning – Supervised – Classification model.

**Basic information about the dataset**

Total number of rows – 399 rows

Total number of inputs – 27

Input columns are –

'age', 'bp', 'al', 'su', 'bgr', 'bu', 'sc', 'sod', 'pot', 'hrmo', 'pcv','wc', 'rc', 'sg\_b', 'sg\_c', 'sg\_d', 'sg\_e', 'rbc\_normal', 'pc\_normal','pcc\_present', 'ba\_present', 'htn\_yes', 'dm\_yes', 'cad\_yes','appet\_yes', 'pe\_yes', 'ane\_yes'.

**Final output that needs to be predicted here is – CKD exist or not.**

**Dataset** **Pre**-**processing** **method**

Pre-processing method is – One hot encoding.

Pre-processing technique – StandardScaler from scikit-learn

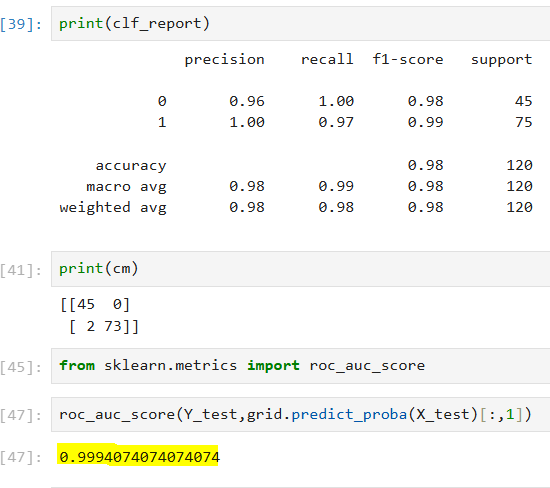
Inputs that needs to be converted from string to number (classification data to Nominal data) are -

sg, rbc, pc, pcc, ba, htn, dm, cad, appet, pe and ane.

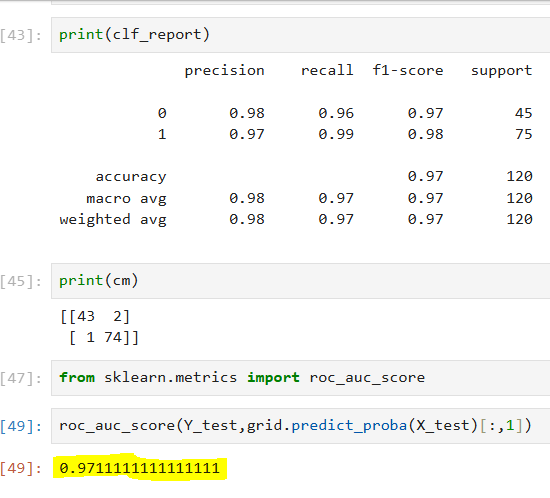
**ROC\_AUC score values of Machine Learning Classification Models.**

|  |  |
| --- | --- |
| **Classification Model** | **ROC\_AUC score values** |
| Support Vector Machine | **0.9994074074074074** |
| Decision Tree | **0.9711111111111111** |
| Random Forest | **0.9997037037037036** |
| Logistic Regression | **1.0** |

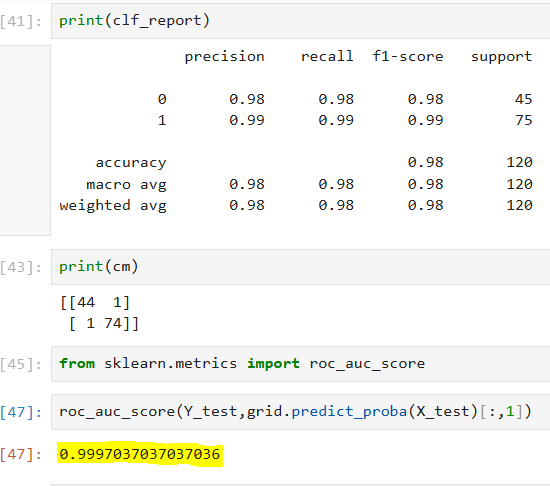
**Support Vector Machine research values:**



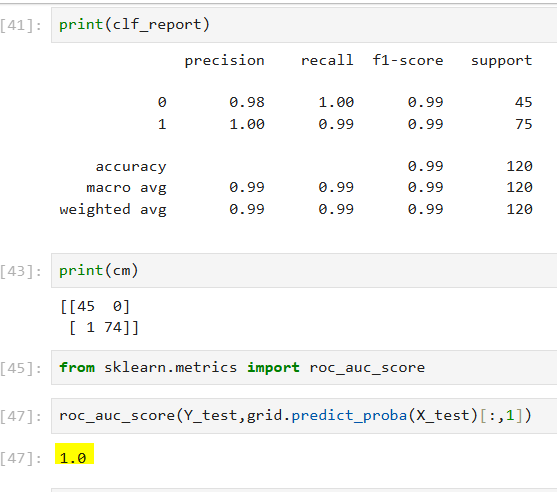
**Decision Tree research values:**



**Random Forest research values:**



**Logistic Regression research values:**



**Conclusion:**

The model we created using Logistic Regression algorithm is the final machine Learning classification model that has been selected and saved as the best model in order to predict CKD based on the patient’s input medical specifications.

**Justification why I have chosen the same is:**

Among other research values of various models created with hyper parameters, the ROC\_AUC score value we obtained which is 🡪 **1.0,** by using the algorithm “Logistic Regression” with tuning parameters applied below,

Solver – 'newton-cg' and 'lbfgs’.