**CLASSIFICATION ASSIGNMENT USING GRID WITH DATA PREPROCESSING**

**Problem statement**

The client is from a health sector and they want help in classifying the people based on the results from their health check reports.

**Dataset**

The dataset contains 400 rows and 25 columns of data.

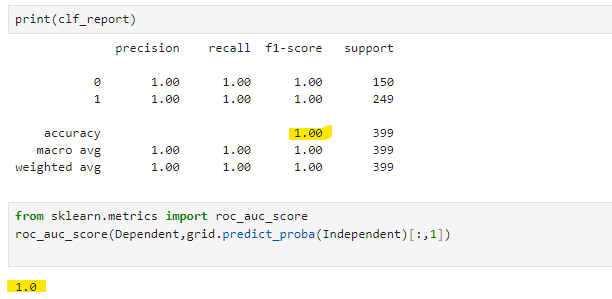
**Data pre-processing**

The data is a mix of categorical and numerical value. Here the categorical data is converted into numerical data using get dummies function from pandas.

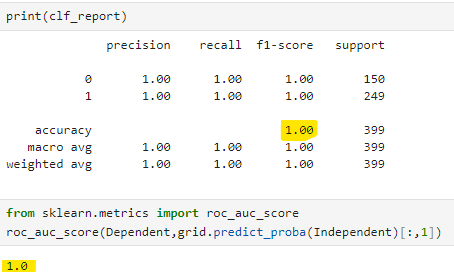
**Creating a good model**

The goal is to create a model with good accuracy and roc\_auc\_curve using machine learning algorithm. Here the output is categorical so, “classifier” is used. It’s a supervised learning as the input and outputs is clearly defined.

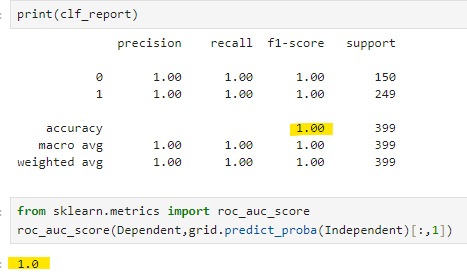
**1.SUPPORT VECTOR MACHINE**



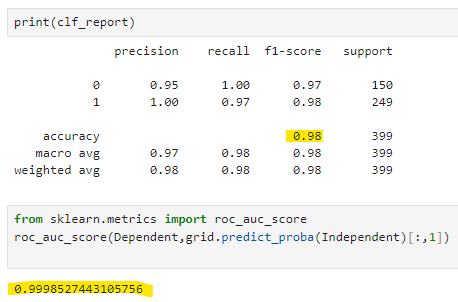
**2.DECISION TREE**

****

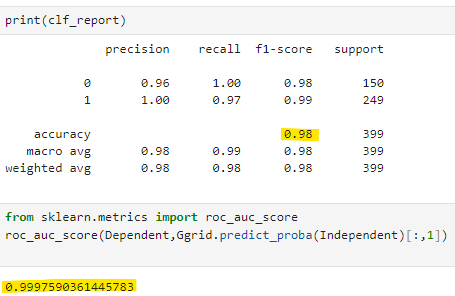
**3.Random Forest**

****

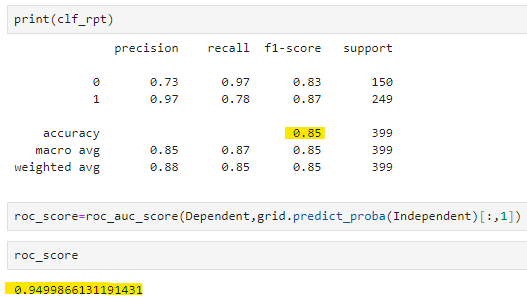
**4.KNN**

****

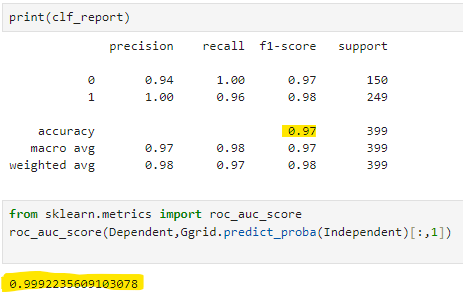
**5.GausianNB**

****

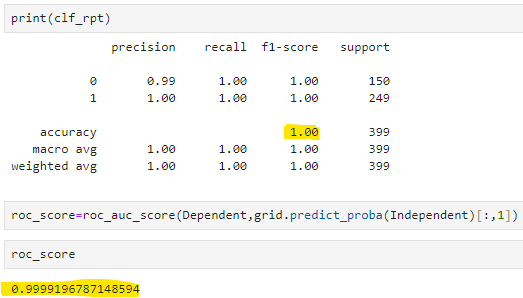
**6.MultinomialNB**

****

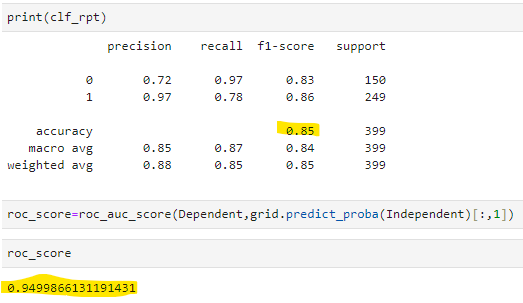
**7.BernoulliNB**

****

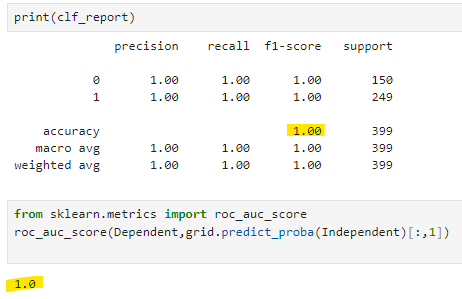
**8.CategoricalNB**

****

**9.ComplementNB**

****

**10.Logistic Regression**

****

**Final Model**

**The accuracy and roc\_score are 0.1 for the following algorithms.**

* **Logistic Regression**
* **Random Forest**
* **Support Vector Machine**
* **Decision Tree**

**Among this best model selecting the “Decision Tree” algorithm for Deployment**

The final model is **Machine learning>>Classification>>Decision Tree**

Justification: While using the Decision Tree algorithm, we get the accuracy and roc\_score as **1.0**

**The best parameters for Decision Tree are {'criterion': 'entropy', 'max\_features': 'log2',**

**'splitter': 'random'}**