**Machine Learning -Classification Assignment – CKD**

Dataset: CKD.csv

**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.) Identify your problem statement

2.) Tell basic info about the dataset (Total number of rows, columns)

3.) Mention the pre-processing method if you’re doing any (like converting string to number – nominal data)

4.) Develop a good model with good evaluation metric. You can use any machine learning algorithm; you can create many models. Finally, you have to come up with final model.

5.) All the research values of each algorithm should be documented. (You can make tabulation or screenshot of the results.)

6.) Mention your final model, justify why u have chosen the same.

**Problem Statement:**

* A requirement from the Hospital, Management asked us to create a predictive model which will predict the Chronic Kidney Disease (CKD) based on the parameters.

**Requirement:**

* To predict the Chronic Kidney Disease from given dataset CKD.csv based on the input and output column variables.

**Problem Identification:**

**The 3 stages of identifying the problem:**

**Stage 1 – Domain Selection**

* Here the dataset contains numbers. So we for the domain **Machine Learning.**

**Stage 2 – Learning Selection**

* In this dataset, both inputs and output variables are present and the requirement is clear. So that it comes under **Supervised** **Learning**.

**Stage 3 – Classification or Regression**

* The Prediction Value “Classification” Column

**yes** – denotes that the patient will have the possibility of Chronic Kidney Disease.

**no** - denotes that the patient will not have the possibility of Chronic Kidney Disease.

Therefore, it would be a Classification data. So that, it comes under **Classification**.

**Classification**

**Machine Learning**

**Supervised Learning**

**Basic information about the dataset:**

|  |  |
| --- | --- |
| Dataset Name | **CKD.csv** |
| Total Number of Rows | **399** |
| Total Number of Columns | **28** |
| Input / Independent Variables | **age**  **bp**  **sg**  **al**  **su**  **rbc**  **pc**  **pcc**  **ba**  **bgr**  **bu**  **sc**  **sod**  **pot**  **hrmo**  **pcv**  **wc**  **rc**  **htn**  **dm**  **cad**  **appet**  **pe**  **ane** |
| Output / Dependent Variable | **charges** |

**Preprocessing Method:**

* Here, there are the11 columns ‘rbc’, ’pc’, ‘pcc’, ’ba’,’ htn’,’dm’,’ cad’,’ appet’,’ pe’,

’ane’ and ’classification’ is containing string values. So that, we convert the string value into numerical value by using “**One Hot Encoding Method**”.

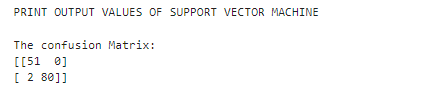
* Best model would be selected by using **GridSearchCV** method. For best practice, **StandardScaler** is used to standardize the given dataset before the training models, to best performance of the model.
* Also, **MinMaxScaler** - used to **normalize numerical data** by scaling values to a fixed range, typically **between 0 and 1** to avoid negative values. Because MultinomialNB, CategoricalNB does not handle negative values.

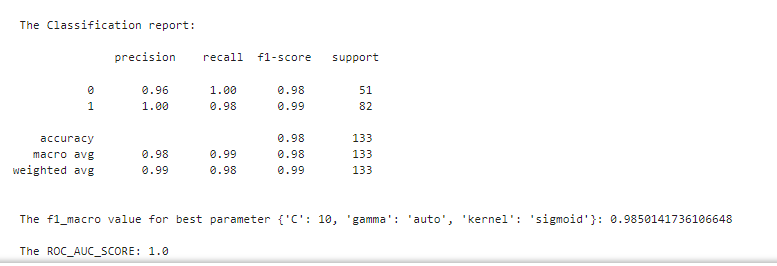
**MODELS**

To create and select best model for the following dataset using Classification method by finding Confusion Matrix and classification Report.

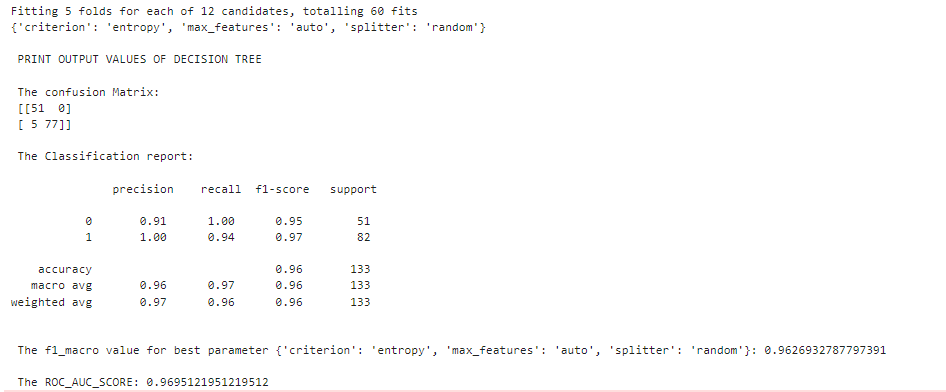
**Dataset**: CKD.csv

1. **Support Vector Machine Classification**

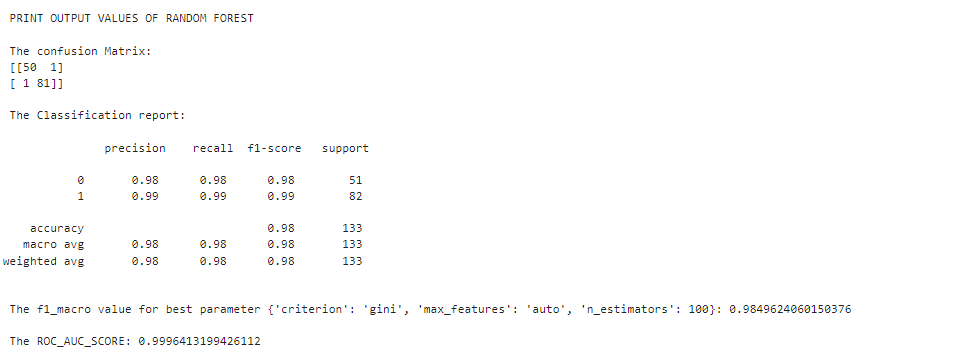
****

****

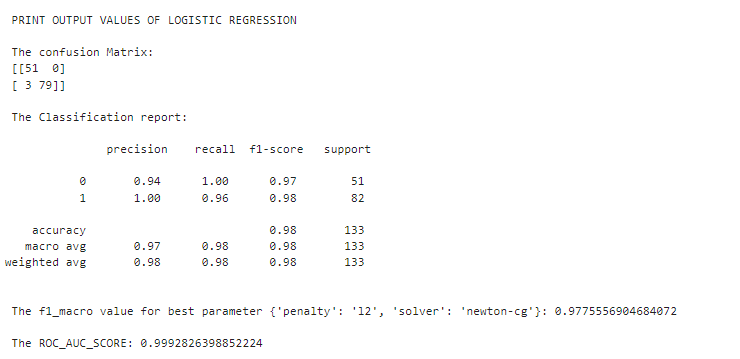
1. **Decision Tree Classification**

****

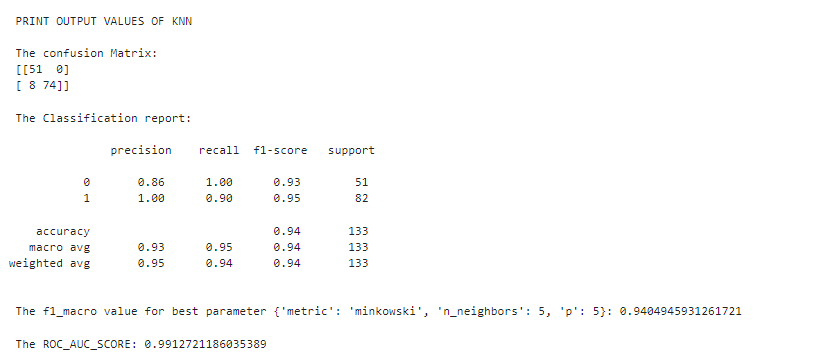
1. **Random Forest Classification**

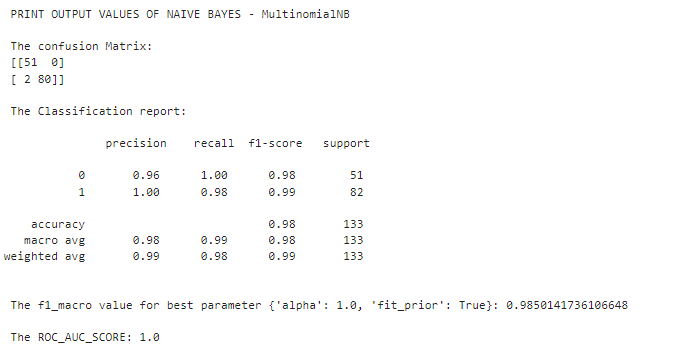
****

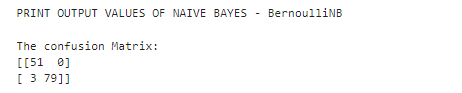
1. **Logistic Regression Classification**

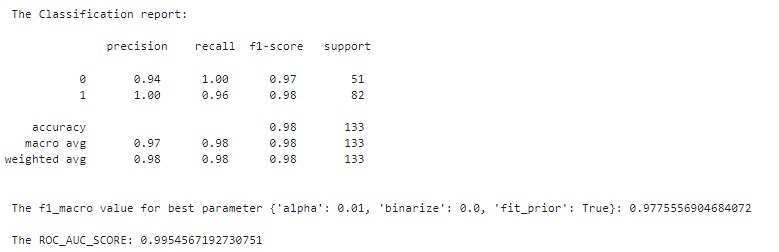
****

1. **KNN Classification**

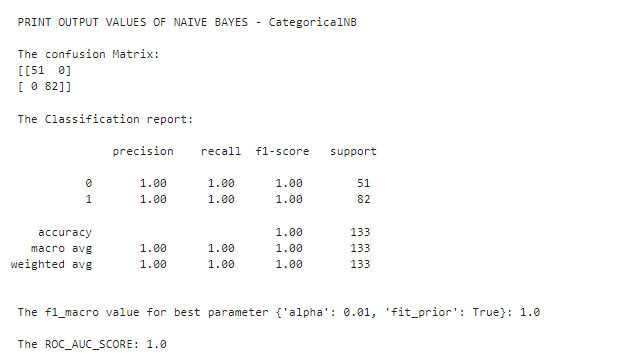
****

1. **Naive Bayes Classification**
2. **MultinomialNB**
3. **BernoulliNB**

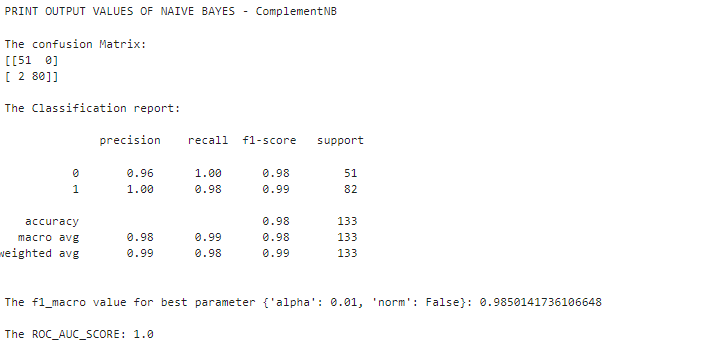
****

****

1. **CategoricalNB**

****

1. **ComplementNB**



**Conclusion: Best Model Selection**

Based on the results from the **SVM** , **Naive Bayes – MultinomialNB and CategoricalNB classifiers** achieved:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **F1-Score (0)** | **F1-Score (1)** | **ROC-AUC Score** |
| SVM | 0.98 | 0.98 | 0.99 | 1 |
| MultinomialNB | 0.98 | 0.98 | 0.99 | 1 |
| CategoricalNB | 1 | 1 | 1 | 1 |

* Though, CategoricalNB gives better (100%) results, it suits for the categorical dataset.
* Comparing SVM , Naive Bayes – MultinomialNB (both gives same results):
* ROC-AUC Score is 1.0 for all models, indicating perfect classification ability.
* Speed & Efficiency - Naive Bayes (MultinomialNB/CategoricalNB) is better for large datasets.
* Complex Decision Boundaries- SVM is preferable for more complex patterns.

Therefore, we have numerical datas with smaller dataset. So that, **Support Vector Machine is preferable as the Best Model** for this dataset.