Data mining project:

**Project Name: Indian Liver Patient**

**Data description:**

**Attribute Information:**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Age | Gender | TB | DB | Alkphos | Sgpt | Sgot | TP | ALB | A/G | label |

1. Age: age of the patient
2. Gender: gender of the patient
3. TB Total Bilirubin
4. DB Direct Bilirubin
5. Alkphos: Alkaline Phosphatase
6. Sgpt Alamine Aminotransferase
7. Sgot Aspartate Aminotransferase
8. TP Total Proteins
9. ALB Albumin
10. A/G Ratio Albumin and Globulin Ratio
11. Selector field used to split the data into two sets (labeled by the experts)

This data set contains 416 liver patient records and 167 non liver patient records. The data set was collected from north east of Andhra Pradesh, India. Selector is a class label used to divide into groups (liver patient or not). This data set contains 441 male patient records and 142 female patient records.

**Problem definition:**

This data set contains 10 variables that are age, gender, total bilirubin, direct bilirubin, total protein, albumin, A/G ratio, SGPT, SGOT and Alk Phos. From these variables, I can determine whether this person is Indian liver disease or not and determine the proportion of the disease in men and women

We represented male as 1 and female as 0.

And we represented Patient as 1 and non-patient as 0.

**Data mining algorithms used and why**

1: **Decision Tree:**

Decision tree is the most powerful and popular tool for classification and prediction. A Decision tree is a flowchart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label.

**2: KNN**

* KNN is a very simple algorithm used to solve classification problems. KNN stands for K-Nearest Neighbors. K is the number of neighbors in KNN. Let’s find out some advantages and disadvantages of KNN algorithm
* Since the KNN algorithm requires no training before making predictions, new data can be added seamlessly which will not impact the accuracy of the algorithm.

**Visualization techniques:**

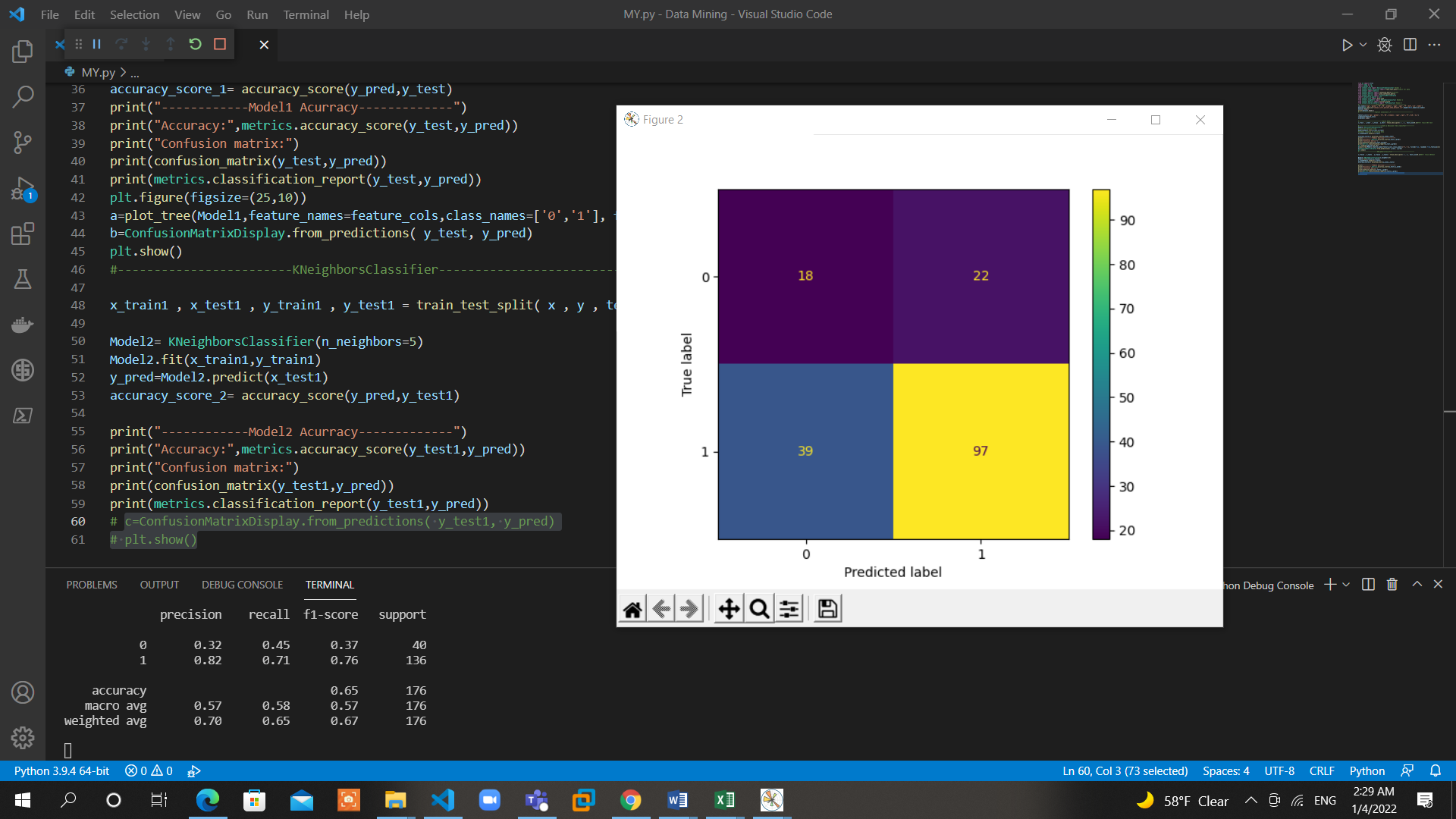
1: confusionMatrixDisplay (plot\_confusionMatrix ())

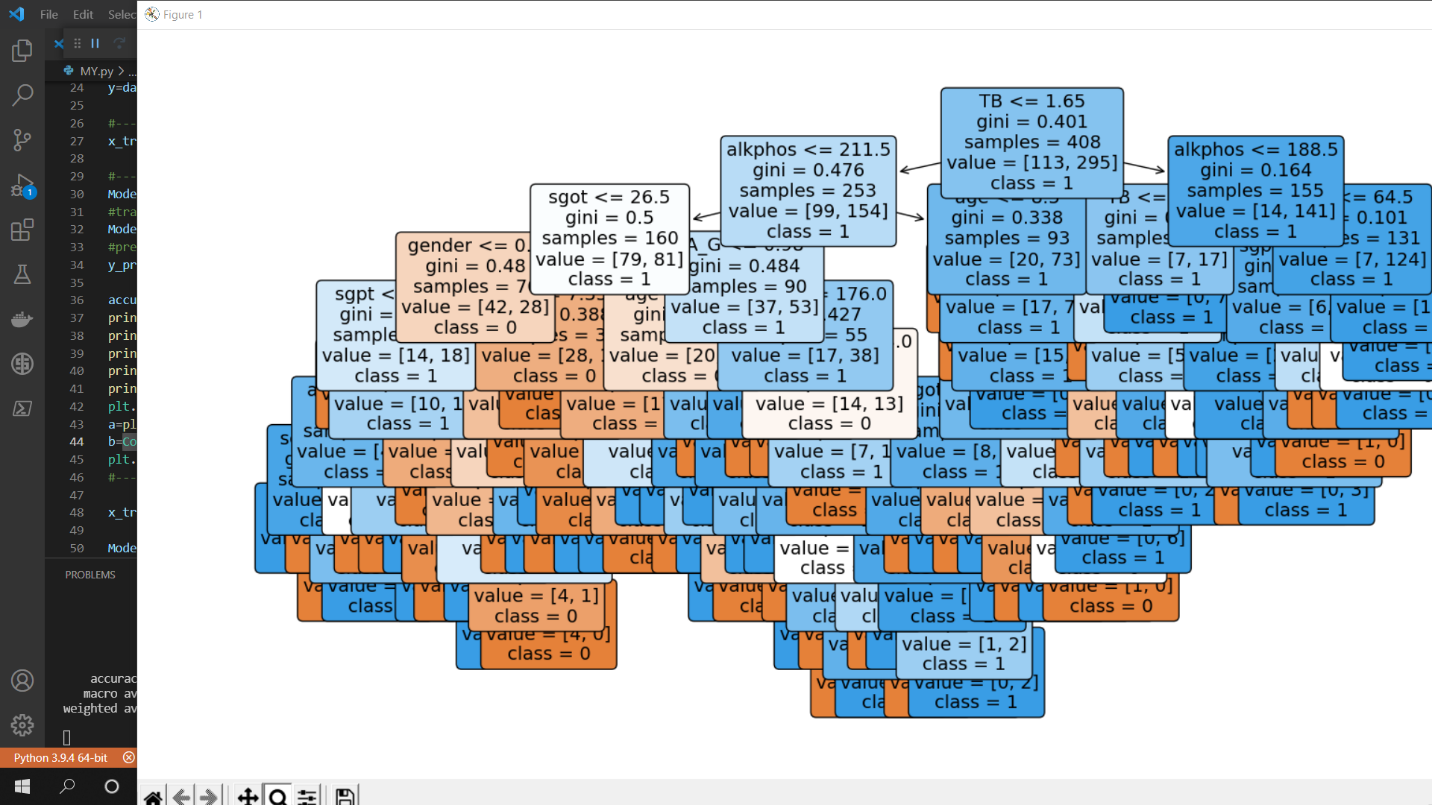
2: plot tree ()

3: show ()

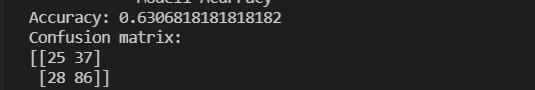
**Evaluation of models**

**Decision Tree**

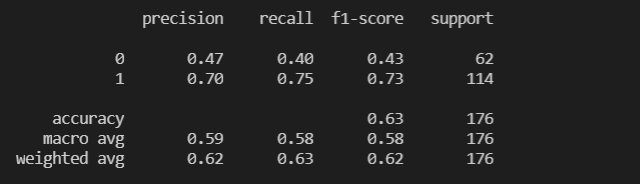
**plot\_confusionMatrix**

  **Plot tree**

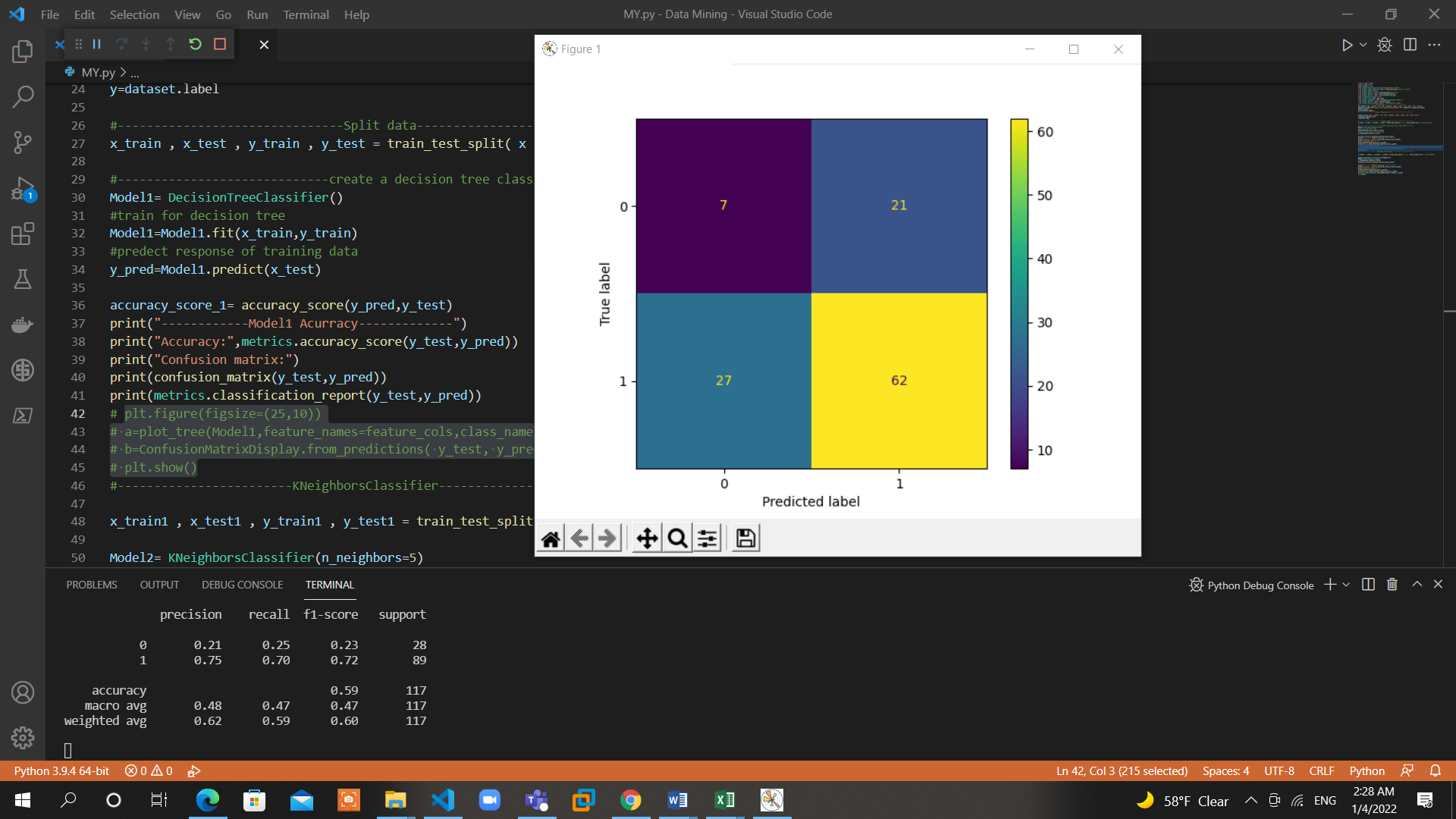
Accuracy and confusion matrix

****

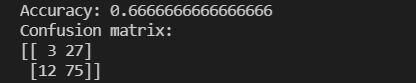
**Report**

****

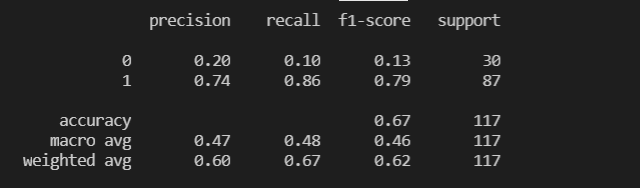
KNN

**plot\_confusionMatrix ()**

Accuracy and confusion matrix



Report



**The Best Model in 2 Model is**

****

**Because it has the best Accuracy**