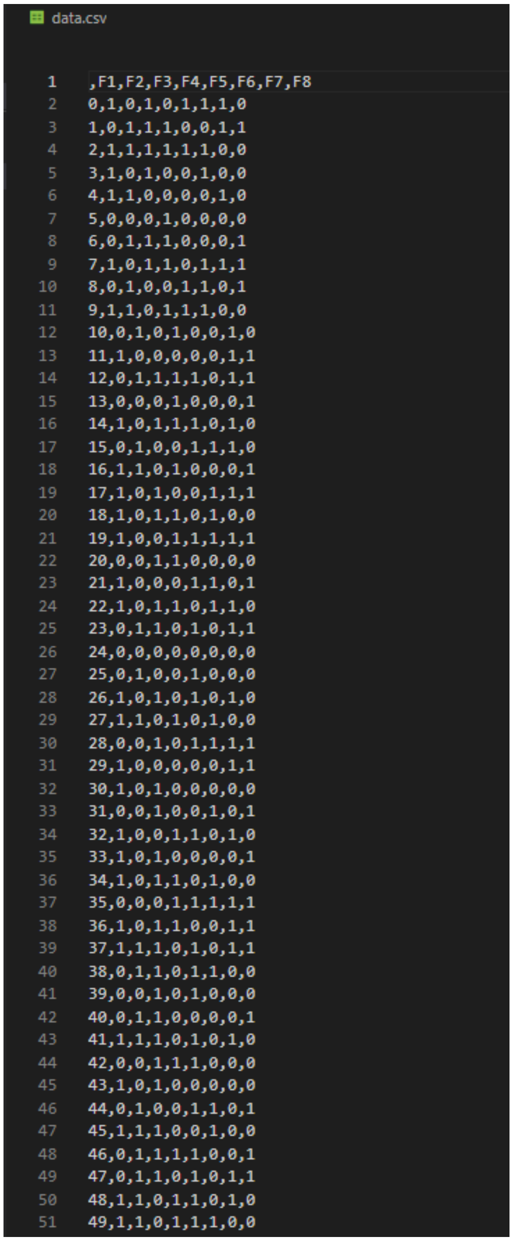
IT414 Assignment-2

Name: Durga Supriya HL
Roll no.: 201IT121

Generate the dataset: First, the data set is generated randomly and stored in a CSV file so that each time the training dataset doesn't change.



```
,F1,F2,F3,F4,F5,F6,F7,F8
0,1,0,1,0,1,1,1,0
1,0,1,1,1,0,0,1,1
2,1,1,1,1,1,1,0,0
3,1,0,1,0,0,1,0,0
4,1,1,0,0,0,0,1,0
5,0,0,0,1,0,0,0,0
6,0,1,1,1,0,0,0,1
7,1,0,1,1,0,1,1,1
8,0,1,0,0,1,1,0,1
9,1,1,0,1,1,1,0,0
10,0,1,0,1,0,0,1,0
11,1,0,0,0,0,0,1,1
12,0,1,1,1,1,0,1,1
13,0,0,0,1,0,0,0,1
14,1,0,1,1,1,0,1,0
15,0,1,0,0,1,1,1,0
16,1,1,0,1,0,0,0,1
17,1,0,1,0,0,1,1,1
18,1,0,1,1,0,1,0,0
19,1,0,0,1,1,1,1,1
20,0,0,1,1,0,0,0,0
21,1,0,0,0,1,1,0,1
22,1,0,1,1,0,1,1,0
23,0,1,1,0,1,0,1,1
24,0,0,0,0,0,0,0,0
25,0,1,0,0,1,0,0,0
26,1,0,1,0,1,0,1,0
27,1,1,0,1,0,1,0,0
28,0,0,1,0,1,1,1,1
29,1,0,0,0,0,0,1,1
30,1,0,1,0,0,0,0,0
31,0,0,1,0,0,1,0,1
32,1,0,0,1,1,0,1,0
33,1,0,1,0,0,0,0,1
34,1,0,1,1,0,1,0,0
35,0,0,0,1,1,1,1,1
36,1,0,1,1,0,0,1,1
37,1,1,1,0,1,0,1,1
38,0,1,1,0,1,1,0,0
39,0,0,1,0,1,0,0,0
40,0,1,1,0,0,0,0,1
41,1,1,1,0,1,0,1,0
42,0,0,1,1,1,0,0,0
43,1,0,1,0,0,0,0,0
44,0,1,0,0,1,1,0,1
45,1,1,1,0,0,1,0,0
46,0,1,1,1,1,0,0,1
47,0,1,1,0,1,0,1,1
48,1,1,0,1,1,0,1,0
49,1,1,0,1,1,1,0,0
```

Preprocess the dataset: The second code block preprocesses the dataset by splitting it into training and testing sets and standardizing the features.

Implement the decision tree algorithm: The third code block contains the implementation of the decision tree algorithm. The algorithm is implemented as a class called DecisionTreeClassifier, which has methods for fitting the model to the training data and making predictions on new data. The algorithm uses a recursive process to build the decision tree by selecting the best feature to split on at each node based on the Gini impurity score.

Train the model: The fourth code block creates an instance of the DecisionTreeClassifier class and trains the model using the training data.

Test the model: The fifth code block uses the trained model to make predictions on the test data set that is given.

Decision tree algorithm

Choosing the best feature: At each node of the decision tree, the algorithm calculates the Gini impurity score for each feature in the dataset and selects the feature with the lowest score as the best feature to split on.

Splitting the node: Once the best feature has been chosen, the algorithm splits the node into child nodes based on the possible values of the selected feature.

Repeating the process: The algorithm recursively repeats the process of choosing the best feature and splitting the node until it reaches the maximum depth or all instances at a node belong to the same class.

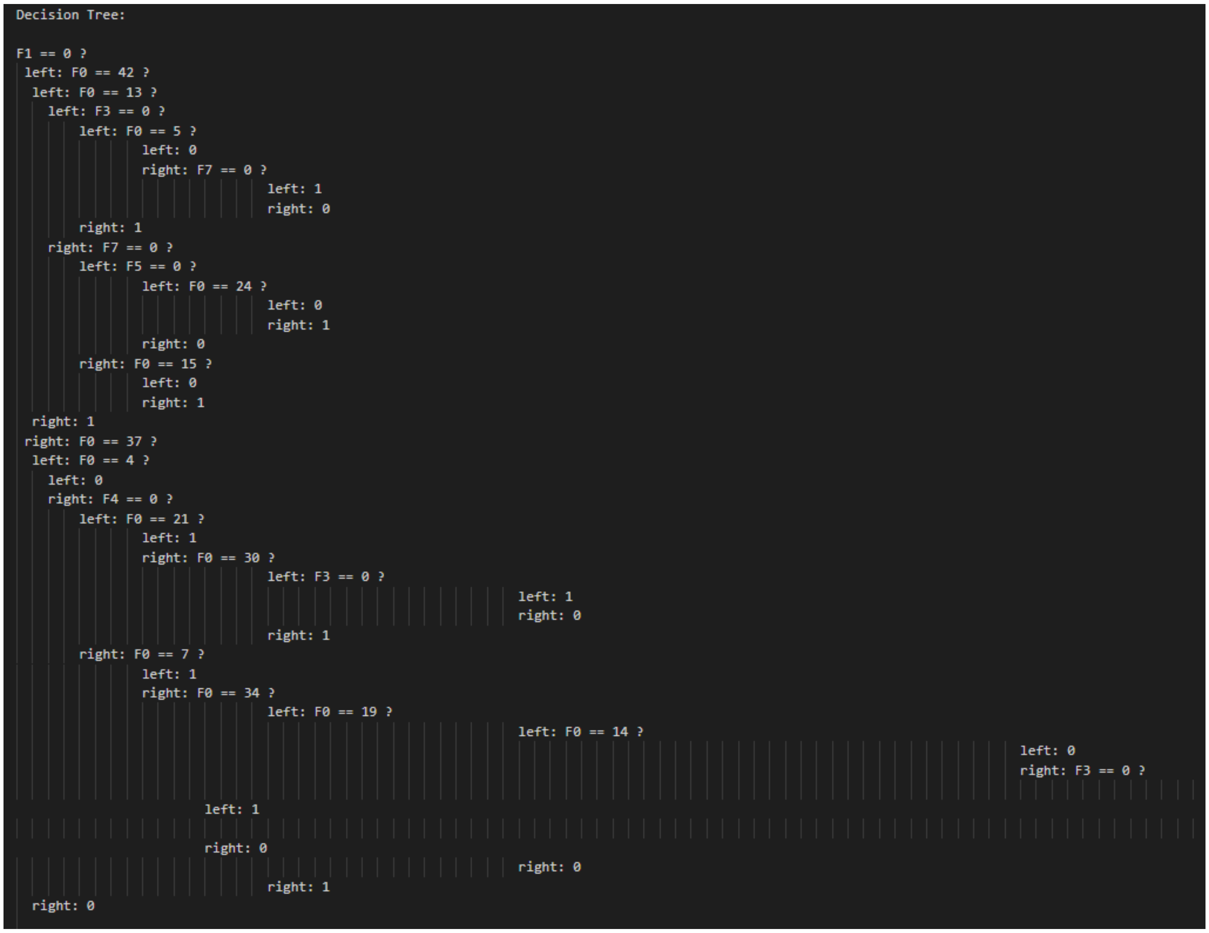
Entropy:

Entropy is a measure of the impurity of a set of instances. The entropy of a set S with respect to a binary classification problem is defined as:

Entropy = -(p(0) * log(P(0)) + p(1) * log(P(1)))

where p0 is the proportion of instances in S that belong to class 1, and p1 is the proportion of instances that belong to class 2 (i.e., p0 + p1 = 1. The entropy is 0 when all instances belong to the same class (i.e., S is pure), and it is maximal when the set is evenly split between the two classes.

In the decision tree algorithm, the entropy is calculated for each subset of instances at a node, and the weighted average entropy of the child nodes is subtracted from the entropy of the parent node to obtain the information gain.

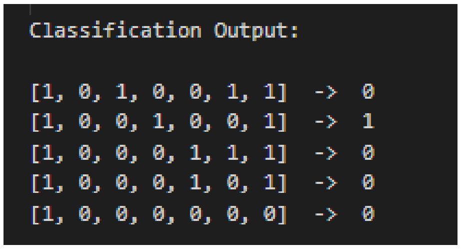


```
Decision Tree:

F1 == 0 ?
 left: F0 == 42 ?
  left: F0 == 13 ?
    left: F3 == 0 ?
        left: F0 == 5 ?
                left: 0
                right: F7 == 0 ?
                                left: 1
                                right: 0
        right: 1
    right: F7 == 0 ?
        left: F5 == 0 ?
                left: F0 == 24 ?
                                left: 0
                                right: 1
                right: 0
        right: F0 == 15 ?
                left: 0
                right: 1
  right: 1
 right: F0 == 37 ?
  left: F0 == 4 ?
    left: 0
    right: F4 == 0 ?
        left: F0 == 21 ?
                left: 1
                right: F0 == 30 ?
                                left: F3 == 0 ?
                                                                left: 1
                                                                right: 0
                                right: 1
        right: F0 == 7 ?
                left: 1
                right: F0 == 34 ?
                                left: F0 == 19 ?
                                                                left: F0 == 14 ?
                                                                                                                                left: 0
                                                                                                                                right: F3 == 0 ?
                        left: 1
                        right: 0
                                                                right: 0
                                right: 1
  right: 0
```

Making predictions: To make a prediction on a new instance, the algorithm traverses the decision tree by following the path of nodes that correspond to the feature values of the instance until it reaches a leaf node. The class label of the leaf node is then returned as the predicted class label for the instance.

The predictions for the given test dataset are below:



```
Classification Output:

[1, 0, 1, 0, 0, 1, 1]  ->  0
[1, 0, 0, 1, 0, 0, 1]  ->  1
[1, 0, 0, 0, 1, 1, 1]  ->  0
[1, 0, 0, 0, 1, 0, 1]  ->  0
[1, 0, 0, 0, 0, 0, 0]  ->  0
```