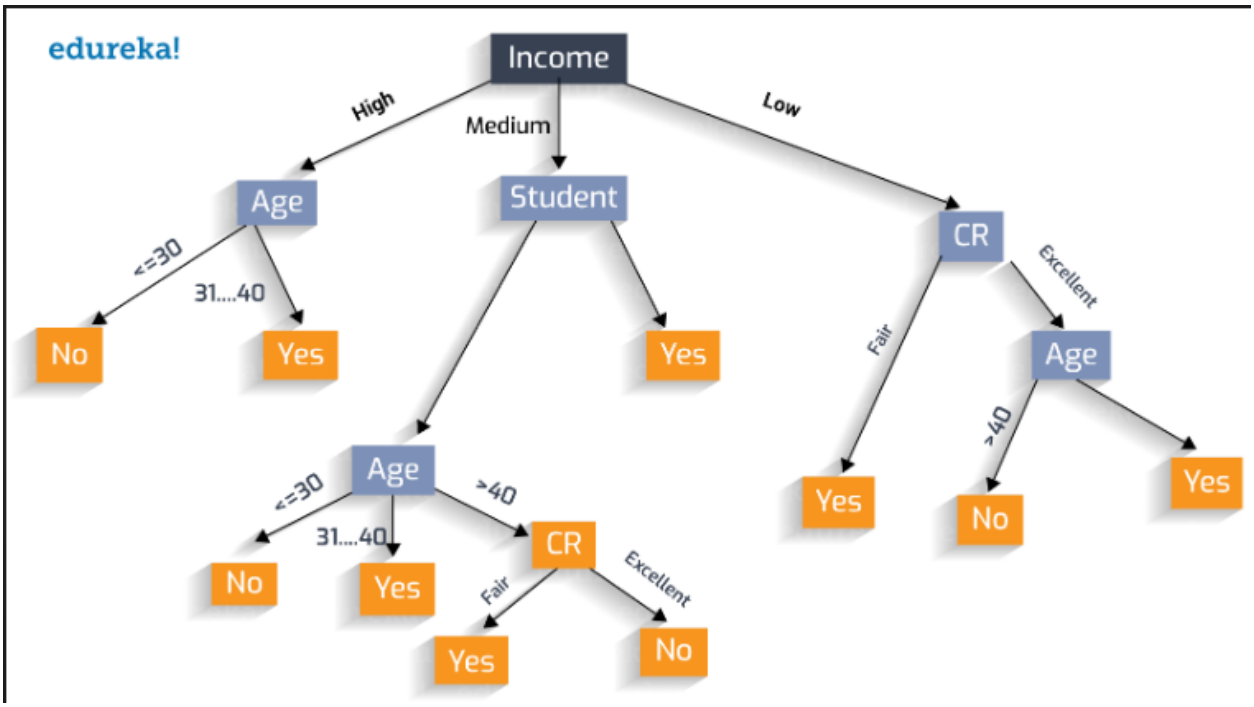
machine learning - decision tree classifiers

this algorithm can be called as “White Box” algorithm . it is easy to train and they do not depend on probability assumptions . they have high accuracy over high dimensional data .

example of a decision tree :



decision trees have three types of nodes:

- Root node
- Decision node
- Leaf node

decision trees split into groups of binary decision . the iteration flows on the basis of the decision made by the previous nodes .

the iteration flows through the end of the leaf node and will return the predicted classification.

the algorithm uses different way to measure the data sets . one of the way is to measure the **GINI impurity .**

the **GINI impurity** decides the optimal split from a root node that makes effective decision in a decision tree . it is the likelihood of incorrect classification in a new instance of a random variable if that instance is randomly classified to distribution of class variables in a dataset.

best example to understand what a **GINI impurity is :**

- **Perfectly sorted box (Gini impurity = 0):** All oranges are on one side, all apples on the other. Easy to pick the right fruit!
- **Totally mixed box (Gini impurity = high):** Oranges and apples are all jumbled together. Picking the right fruit is a guessing game!

Note : lower the impurity , cleaner the seperation and easy to make predictions.

making a decision tree classifier in python:

in this code , we are using the basic decisiontreeclassifier

```
import pandas as pd
from sklearn.tree import DecisionTreeClassifier
from sklearn import tree
from sklearn.model_selection import train_test_split
from sklearn import metrics
```

from sklearn.preprocessing import LabelEncoder
import numpy as np
import matplotlib.pyplot as plt # Install using `pip install graphviz`

Read the CSV data

col_names = ['pregnant', 'glucose', 'bp', 'skin', 'insulin', 'bmi', 'pedigree', 'age', 'label']
pima = pd.read_csv("diabetes.csv", header=None, names=col_names)

Handle potential missing values (optional)

pima.replace('?', np.nan, inplace=True) # Replace '?' with NaN
pima.dropna(inplace=True) # Drop rows with missing values (if necessary)

Convert categorical features to numerical using LabelEncoder

categorical_features = ['pregnant', 'glucose', 'bp', 'skin', 'insulin', 'bmi', 'pedigree', 'age', 'label']
le = LabelEncoder()
for col in categorical_features:
pima[col] = le.fit_transform(pima[col])

Separate features (X) and target variable (y)

X = pima.drop('label', axis=1)
y = pima['label']

No need for scaling numerical features with DecisionTreeClassifier

Split data into training and testing sets

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

Train the decision tree classifier

clf = DecisionTreeClassifier()
clf = clf.fit(X_train, y_train)

Make predictions

y_pred = clf.predict(X_test)

Evaluate accuracy

print("Accuracy:", metrics.accuracy_score(y_test, y_pred))

text_representation = tree.export_text(clf)
print(text_representation)

the output is given as :

```
Accuracy: 0.6428571428571429
|--- feature_1 <= 99.00
|   |--- feature_1 <= 55.50
|   |   |--- feature_5 <= 69.50
|   |   |   |--- feature_5 <= 59.00
|   |   |   |   |--- feature_1 <= 33.50
|   |   |   |   |   |--- feature_4 <= 123.00
|   |   |   |   |   |   |--- class: 0
|   |   |   |   |   |--- feature_4 >  123.00
|   |   |   |   |   |   |--- feature_4 <= 130.50
|   |   |   |   |   |   |   |--- class: 1
|   |   |   |   |   |   |--- feature_4 >  130.50
|   |   |   |   |   |   |   |--- class: 0
|   |   |   |   |--- feature_1 >  33.50
|   |   |   |   |   |--- feature_2 <= 19.50
|   |   |   |   |   |   |--- class: 1
|   |   |   |   |   |--- feature_2 >  19.50
```

```
|   |   |   |   |   |   |--- feature_1 <= 36.00
|   |   |   |   |   |   |   |--- feature_3 <= 12.00
|   |   |   |   |   |   |   |   |--- class: 1
|   |   |   |   |   |   |   |--- feature_3 >  12.00
|   |   |   |   |   |   |   |   |--- class: 0
|   |   |   |   |   |   |--- feature_1 >  36.00
|   |   |   |   |   |   |   |--- class: 0
|   |   |   |--- feature_5 >  59.00
|   |   |   |   |--- feature_0 <= 6.50
|   |   |   |   |   |--- class: 0
|   |   |   |   |--- feature_0 >  6.50
|   |   |   |   |   |--- feature_3 <= 7.50
|   |   |   |   |   |   |--- feature_5 <= 65.00
|   |   |   |   |   |   |   |--- class: 0
|   |   |   |   |   |   |--- feature_5 >  65.00
|   |   |   |   |   |   |   |--- class: 1
|   |   |   |   |   |--- feature_3 >  7.50
|   |   |   |   |   |   |--- class: 1
|   |   |--- feature_5 >  69.50
|   |   |   |--- feature_7 <= 7.50
|   |   |   |   |--- feature_2 <= 0.50
|   |   |   |   |   |--- class: 1
|   |   |   |   |--- feature_2 >  0.50
|   |   |   |   |   |--- feature_5 <= 223.00
|   |   |   |   |   |   |--- feature_6 <= 291.00
|   |   |   |   |   |   |   |--- feature_1 <= 39.50
|   |   |   |   |   |   |   |   |--- feature_4 <= 0.50
|   |   |   |   |   |   |   |   |   |--- feature_2 <= 19.00
|   |   |   |   |   |   |   |   |   |   |--- class: 1
|   |   |   |   |   |   |   |   |   |--- feature_2 >  19.00
|   |   |   |   |   |   |   |   |   |   |--- feature_3 <= 30.50
|   |   |   |   |   |   |   |   |   |   |   |--- truncated branch of depth 5
|   |   |   |   |   |   |   |   |   |--- feature_3 >  30.50
|   |   |   |   |   |   |   |   |   |   |--- class: 1
|   |   |   |   |   |   |   |   |--- feature_4 >  0.50
|   |   |   |   |   |   |   |   |   |--- class: 0
|   |   |   |   |   |   |   |--- feature_1 >  39.50
|   |   |   |   |   |   |   |   |--- feature_1 <= 46.00
|   |   |   |   |   |   |   |   |   |--- feature_4 <= 9.00
|   |   |   |   |   |   |   |   |   |   |--- class: 0
|   |   |   |   |   |   |   |   |   |--- feature_4 >  9.00
|   |   |   |   |   |   |   |   |   |   |--- class: 1
|   |   |   |   |   |   |   |   |--- feature_1 >  46.00
|   |   |   |   |   |   |   |   |   |--- class: 0
|   |   |   |   |   |   |--- feature_6 >  291.00
|   |   |   |   |   |   |   |--- feature_6 <= 323.00
|   |   |   |   |   |   |   |   |--- feature_6 <= 302.50
|   |   |   |   |   |   |   |   |   |--- feature_1 <= 11.50
|   |   |   |   |   |   |   |   |   |   |--- class: 1
|   |   |   |   |   |   |   |   |--- feature_1 >  11.50
|   |   |   |   |   |   |   |   |   |--- class: 0
|   |   |   |   |   |   |   |--- feature_6 >  302.50
```

```
|   |   |   |   |   |   |   |   |   |--- class: 1
|   |   |   |   |   |   |   |   |--- feature_6 >  323.00
|   |   |   |   |   |   |   |   |   |--- feature_3 <= 1.00
|   |   |   |   |   |   |   |   |   |   |--- feature_7 <= 2.00
|   |   |   |   |   |   |   |   |   |   |   |--- class: 1
|   |   |   |   |   |   |   |   |   |   |--- feature_7 >  2.00
|   |   |   |   |   |   |   |   |   |   |   |--- class: 0
|   |   |   |   |   |   |   |   |   |--- feature_3 >  1.00
|   |   |   |   |   |   |   |   |   |   |--- feature_2 <= 27.50
|   |   |   |   |   |   |   |   |   |   |   |--- feature_7 <= 2.50
|   |   |   |   |   |   |   |   |   |   |   |   |--- class: 0
|   |   |   |   |   |   |   |   |   |   |   |--- feature_7 >  2.50
|   |   |   |   |   |   |   |   |   |   |   |   |--- truncated branch of depth 2
|   |   |   |   |   |   |   |   |   |   |--- feature_2 >  27.50
|   |   |   |   |   |   |   |   |   |   |   |--- class: 0
|   |   |   |   |   |   |--- feature_5 >  223.00
|   |   |   |   |   |   |   |--- feature_2 <= 3.50
|   |   |   |   |   |   |   |   |--- class: 0
|   |   |   |   |   |   |   |--- feature_2 >  3.50
|   |   |   |   |   |   |   |   |--- class: 1
|   |   |   |--- feature_7 >  7.50
|   |   |   |   |--- feature_6 <= 300.50
|   |   |   |   |   |--- feature_1 <= 15.50
|   |   |   |   |   |   |--- feature_2 <= 5.50
|   |   |   |   |   |   |   |--- class: 1
|   |   |   |   |   |   |--- feature_2 >  5.50
|   |   |   |   |   |   |   |--- feature_3 <= 1.50
|   |   |   |   |   |   |   |   |--- feature_1 <= 10.00
|   |   |   |   |   |   |   |   |   |--- feature_0 <= 14.50
|   |   |   |   |   |   |   |   |   |   |--- feature_0 <= 0.50
|   |   |   |   |   |   |   |   |   |   |   |--- class: 0
|   |   |   |   |   |   |   |   |   |   |--- feature_0 >  0.50
|   |   |   |   |   |   |   |   |   |   |   |--- class: 1
|   |   |   |   |   |   |   |   |   |--- feature_0 >  14.50
|   |   |   |   |   |   |   |   |   |   |--- class: 0
|   |   |   |   |   |   |   |   |--- feature_1 >  10.00
|   |   |   |   |   |   |   |   |   |--- class: 0
|   |   |   |   |   |   |   |--- feature_3 >  1.50
|   |   |   |   |   |   |   |   |--- class: 0
|   |   |   |   |   |--- feature_1 >  15.50
|   |   |   |   |   |   |--- feature_0 <= 1.50
|   |   |   |   |   |   |   |--- class: 1
|   |   |   |   |   |   |--- feature_0 >  1.50
|   |   |   |   |   |   |   |--- feature_4 <= 96.00
|   |   |   |   |   |   |   |   |--- feature_4 <= 13.50
|   |   |   |   |   |   |   |   |   |--- feature_2 <= 36.50
|   |   |   |   |   |   |   |   |   |   |--- feature_0 <= 10.50
|   |   |   |   |   |   |   |   |   |   |   |--- truncated branch of depth 6
|   |   |   |   |   |   |   |   |   |--- feature_0 >  10.50
|   |   |   |   |   |   |   |   |   |   |--- truncated branch of depth 4
|   |   |   |   |   |   |   |   |--- feature_2 >  36.50
|   |   |   |   |   |   |   |   |   |--- feature_3 <= 33.50
```

```
|   |   |   |   |   |   |   |   |   |   |   |--- class: 0
|   |   |   |   |   |   |   |   |   |   |--- feature_3 >  33.50
|   |   |   |   |   |   |   |   |   |   |   |--- class: 1
|   |   |   |   |   |   |   |   |   |--- feature_4 >  13.50
|   |   |   |   |   |   |   |   |   |--- feature_7 <= 24.50
|   |   |   |   |   |   |   |   |   |   |--- class: 0
|   |   |   |   |   |   |   |   |   |--- feature_7 >  24.50
|   |   |   |   |   |   |   |   |   |   |--- class: 1
|   |   |   |   |   |   |   |   |--- feature_4 >  96.00
|   |   |   |   |   |   |   |   |   |--- feature_1 <= 28.00
|   |   |   |   |   |   |   |   |   |--- class: 0
|   |   |   |   |   |   |   |   |--- feature_1 >  28.00
|   |   |   |   |   |   |   |   |   |--- class: 1
|   |   |   |   |--- feature_6 >  300.50
|   |   |   |   |   |--- feature_1 <= 53.50
|   |   |   |   |   |   |--- feature_3 <= 4.50
|   |   |   |   |   |   |   |--- feature_7 <= 9.00
|   |   |   |   |   |   |   |   |--- class: 0
|   |   |   |   |   |   |   |--- feature_7 >  9.00
|   |   |   |   |   |   |   |   |--- feature_5 <= 123.00
|   |   |   |   |   |   |   |   |   |--- class: 1
|   |   |   |   |   |   |   |   |--- feature_5 >  123.00
|   |   |   |   |   |   |   |   |   |--- feature_6 <= 370.00
|   |   |   |   |   |   |   |   |   |   |--- class: 1
|   |   |   |   |   |   |   |   |   |--- feature_6 >  370.00
|   |   |   |   |   |   |   |   |   |   |--- feature_6 <= 458.00
|   |   |   |   |   |   |   |   |   |   |   |--- class: 0
|   |   |   |   |   |   |   |   |   |   |--- feature_6 >  458.00
|   |   |   |   |   |   |   |   |   |   |   |--- truncated branch of depth 2
|   |   |   |   |   |   |--- feature_3 >  4.50
|   |   |   |   |   |   |   |--- feature_7 <= 38.50
|   |   |   |   |   |   |   |   |--- feature_3 <= 43.00
|   |   |   |   |   |   |   |   |   |--- feature_3 <= 35.00
|   |   |   |   |   |   |   |   |   |   |--- feature_2 <= 22.50
|   |   |   |   |   |   |   |   |   |   |   |--- truncated branch of depth 2
|   |   |   |   |   |   |   |   |   |   |--- feature_2 >  22.50
|   |   |   |   |   |   |   |   |   |   |   |--- class: 1
|   |   |   |   |   |   |   |   |   |--- feature_3 >  35.00
|   |   |   |   |   |   |   |   |   |   |--- feature_5 <= 123.50
|   |   |   |   |   |   |   |   |   |   |   |--- class: 0
|   |   |   |   |   |   |   |   |   |   |--- feature_5 >  123.50
|   |   |   |   |   |   |   |   |   |   |   |--- truncated branch of depth 3
|   |   |   |   |   |   |   |   |   |--- feature_3 >  43.00
|   |   |   |   |   |   |   |   |   |--- class: 0
|   |   |   |   |   |   |   |   |--- feature_7 >  38.50
|   |   |   |   |   |   |   |   |   |--- class: 0
|   |   |   |   |   |--- feature_1 >  53.50
|   |   |   |   |   |   |--- class: 0
|   |--- feature_1 >  55.50
|   |   |--- feature_5 <= 28.00
|   |   |   |--- class: 0
|   |   |--- feature_5 >  28.00
```

```
|   |   |   |--- feature_4 <= 140.50
|   |   |   |   |--- feature_5 <= 86.00
|   |   |   |   |   |--- feature_5 <= 52.50
|   |   |   |   |   |   |--- class: 1
|   |   |   |   |   |--- feature_5 >  52.50
|   |   |   |   |   |   |--- feature_4 <= 41.50
|   |   |   |   |   |   |   |--- class: 0
|   |   |   |   |   |   |--- feature_4 >  41.50
|   |   |   |   |   |   |   |--- class: 1
|   |   |   |   |--- feature_5 >  86.00
|   |   |   |   |   |--- feature_2 <= 42.50
|   |   |   |   |   |   |--- feature_6 <= 29.00
|   |   |   |   |   |   |   |--- feature_6 <= 13.00
|   |   |   |   |   |   |   |   |--- class: 1
|   |   |   |   |   |   |   |--- feature_6 >  13.00
|   |   |   |   |   |   |   |   |--- class: 0
|   |   |   |   |   |   |--- feature_6 >  29.00
|   |   |   |   |   |   |   |--- feature_2 <= 28.50
|   |   |   |   |   |   |   |   |--- feature_5 <= 124.50
|   |   |   |   |   |   |   |   |   |--- feature_2 <= 23.00
|   |   |   |   |   |   |   |   |   |   |--- class: 1
|   |   |   |   |   |   |   |   |   |--- feature_2 >  23.00
|   |   |   |   |   |   |   |   |   |   |--- feature_0 <= 14.00
|   |   |   |   |   |   |   |   |   |   |   |--- class: 0
|   |   |   |   |   |   |   |   |   |   |--- feature_0 >  14.00
|   |   |   |   |   |   |   |   |   |   |   |--- class: 1
|   |   |   |   |   |   |   |   |--- feature_5 >  124.50
|   |   |   |   |   |   |   |   |   |--- class: 1
|   |   |   |   |   |   |   |--- feature_2 >  28.50
|   |   |   |   |   |   |   |   |--- feature_7 <= 36.50
|   |   |   |   |   |   |   |   |   |--- class: 1
|   |   |   |   |   |   |   |   |--- feature_7 >  36.50
|   |   |   |   |   |   |   |   |   |--- feature_4 <= 40.50
|   |   |   |   |   |   |   |   |   |   |--- class: 1
|   |   |   |   |   |   |   |   |   |--- feature_4 >  40.50
|   |   |   |   |   |   |   |   |   |   |--- class: 0
|   |   |   |   |   |--- feature_2 >  42.50
|   |   |   |   |   |   |--- class: 0
|   |   |   |--- feature_4 >  140.50
|   |   |   |   |--- feature_0 <= 9.50
|   |   |   |   |   |--- feature_7 <= 4.00
|   |   |   |   |   |   |--- class: 0
|   |   |   |   |   |--- feature_7 >  4.00
|   |   |   |   |   |   |--- class: 1
|   |   |   |   |--- feature_0 >  9.50
|   |   |   |   |   |--- class: 0
|--- feature_1 >  99.00
|   |--- feature_5 <= 238.00
|   |   |--- feature_7 <= 21.50
|   |   |   |--- feature_5 <= 160.50
|   |   |   |   |--- feature_6 <= 485.50
|   |   |   |   |   |--- feature_6 <= 49.50
```

```
|   |   |   |   |   |   |--- feature_6 <= 44.50
|   |   |   |   |   |   |   |--- class: 0
|   |   |   |   |   |   |--- feature_6 >  44.50
|   |   |   |   |   |   |   |--- class: 1
|   |   |   |   |   |--- feature_6 >  49.50
|   |   |   |   |   |   |--- class: 0
|   |   |   |   |--- feature_6 >  485.50
|   |   |   |   |   |--- feature_0 <= 10.50
|   |   |   |   |   |   |--- class: 0
|   |   |   |   |   |--- feature_0 >  10.50
|   |   |   |   |   |   |--- class: 1
|   |   |   |--- feature_5 >  160.50
|   |   |   |   |--- feature_5 <= 162.00
|   |   |   |   |   |--- class: 1
|   |   |   |   |--- feature_5 >  162.00
|   |   |   |   |   |--- feature_4 <= 173.00
|   |   |   |   |   |   |--- feature_2 <= 25.50
|   |   |   |   |   |   |   |--- feature_3 <= 26.50
|   |   |   |   |   |   |   |   |--- class: 1
|   |   |   |   |   |   |   |--- feature_3 >  26.50
|   |   |   |   |   |   |   |   |--- class: 0
|   |   |   |   |   |   |--- feature_2 >  25.50
|   |   |   |   |   |   |   |--- class: 0
|   |   |   |   |   |--- feature_4 >  173.00
|   |   |   |   |   |   |--- class: 1
|   |   |--- feature_7 >  21.50
|   |   |   |--- feature_3 <= 15.00
|   |   |   |   |--- class: 0
|   |   |   |--- feature_3 >  15.00
|   |   |   |   |--- feature_1 <= 125.00
|   |   |   |   |   |--- class: 0
|   |   |   |   |--- feature_1 >  125.00
|   |   |   |   |   |--- class: 1
|   |--- feature_5 >  238.00
|   |   |--- feature_4 <= 185.50
|   |   |   |--- class: 1
|   |   |--- feature_4 >  185.50
|   |   |   |--- class: 2
```

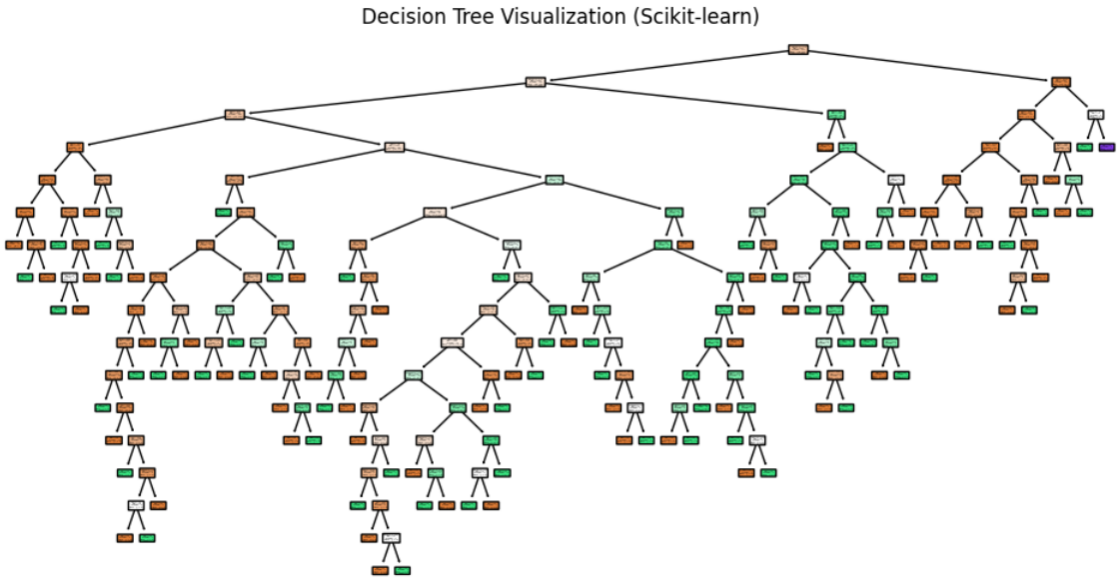
this is the representation of the tree but in a text format .

if you want to view it as a pictorial representation .

then replace the text_representation code with:

plt.figure(figsize=(12, 6)) # Adjust figure size as needed
plot_tree(clf, rounded=True, filled=True, feature_names=X.columns,

```
class_names=le.inverse_transform(y))
plt.title("Decision Tree Visualization (Scikit-learn)")
plt.show()
```



the font is tiny as the tree has more than 4 nodes .

hence, this is how you create a decision tree.