Decision Tree

- Decision tree is one of the most popular machine learning algorithms used all along.
- Decision trees are used for both classification and regression problems.

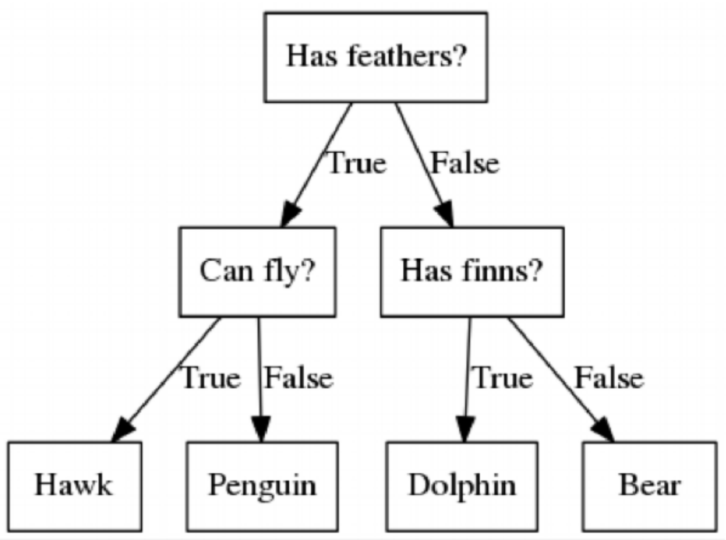
1. Why Decision trees?

We have couple of other algorithms there, so why do we have to choose Decision trees??

- Decision tress often mimic the human level thinking so its so simple to understand the data and make some good interpretations.
- Decision trees actually make you see the logic for the data to interpret(not like black box algorithms like SVM,NN,etc..)
- A decision tree follows a set of if-else conditions to visualize the data and classify it according to the conditions.

For example :

animal_tree

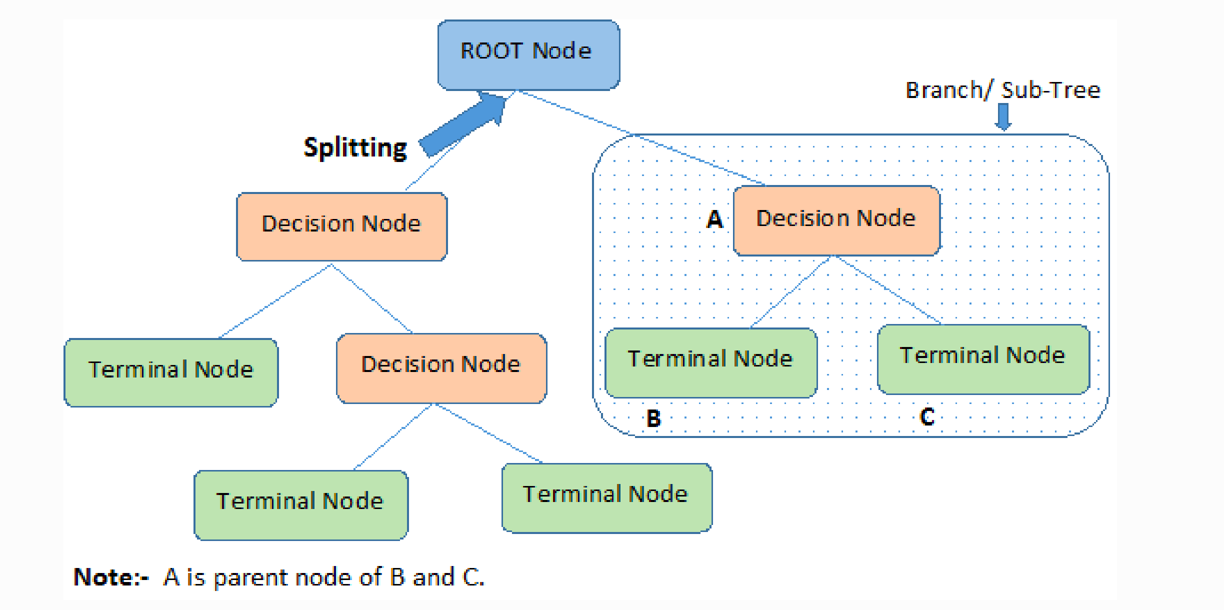


- A decision tree is a tree where each node represents a feature(attribute), each link(branch) represents a decision(rule) and each leaf represents an outcome(categorical or continues value).

Important terminology In DT

- Root Node: This attribute is used for dividing the data into two or more sets. The feature attribute in this node is selected based on Attribute Selection Techniques.
- Branch or Sub-Tree: A part of the entire decision tree is called branch or sub-tree.
- Splitting: Dividing a node into two or more sub-nodes based on if-else conditions.
- Decision Node: After splitting the sub-nodes into further sub-nodes, then it is called as the decision node.

- Leaf or Terminal Node: This is the end of the decision tree where it cannot be split into further sub-nodes.
- Pruning: Removing a sub-node from the tree is called pruning.



Types of Decision Trees

- Types of decision tree is based on the type of target variable we have. It can be of two types:

1. Categorical Variable Decision Tree: Decision Tree which has categorical target variable then it called as categorical variable decision tree. Example:- In above scenario of student problem, where the target variable was “Student will play cricket or not” i.e. YES or NO.

2. Continuous Variable Decision Tree: Decision Tree has continuous target variable then it is called as Continuous Variable Decision Tree.

Three Important Term In DT

- Entropy(H)
- Information Gain(IG)
- Gini Impurity

Entropy

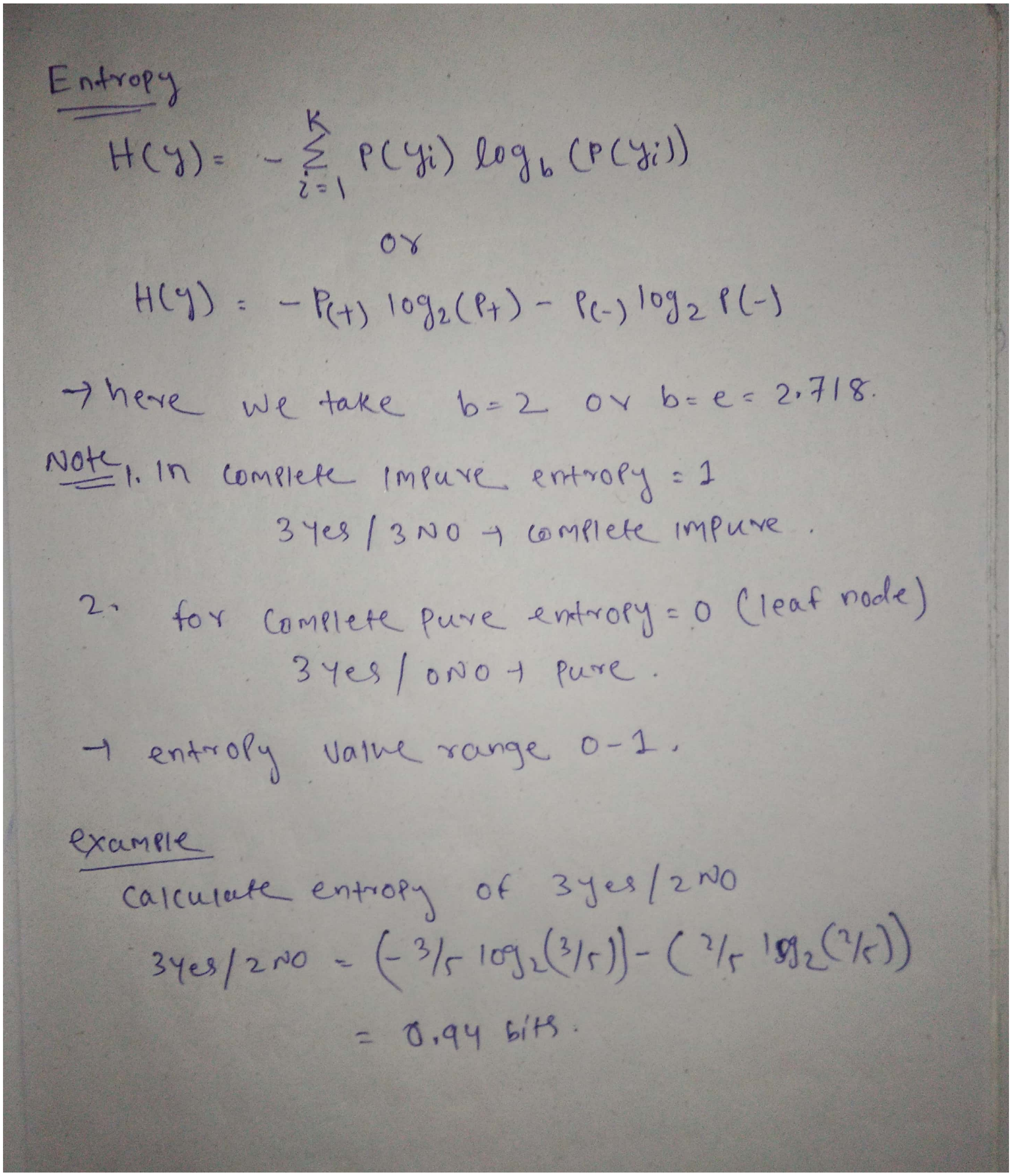
-> It is select the best feature/attribute in DT for Splitting Purpose

-> It helps to measure purity of split

-> Pure split is either get yes or no but not both.

-> Eg. 3yes/2No - it is not pure But 3Yes/0No - it is pure

-> we keep split whenever pure split not come.



Entropy

$$H(y) = -\sum_{i=1}^{K} P(y_i) \log_b (P(y_i))$$

or

$$H(y) = -P_{(+)} \log_2(P_+) - P_{(-)} \log_2 P(-)$$

→ here we take $b = 2$ or $b = e = 2.718$.

Note 1. In complete impure entropy = 1

3 yes / 3 NO → complete impure.

2. for Complete pure entropy = 0 (leaf node)

3 yes / 0 NO → pure.

→ entropy value range 0-1.

example

Calculate entropy of 3 yes / 2 NO

$$3\text{yes}/2\text{NO} = \left(-\tfrac{3}{5}\log_2\left(\tfrac{3}{5}\right)\right) - \left(\tfrac{2}{5}\log_2\left(\tfrac{2}{5}\right)\right)$$

$$= 0.94 \text{ bits}.$$

2. Information Gain(IG)

-> Information gain is calculated by comparing the entropy of the dataset before and after a transformation.

-> If there are many entropy then we compute average of all entropy and compare to find which is best this done by information gain

-> Information gain is the reduction in entropy or surprise by transforming a dataset and is often used in training decision trees.

$$GAIN_{split} = Entropy(p) - \left(\sum_{i=1}^{k} \frac{n_i}{n} Entropy(i) \right)$$

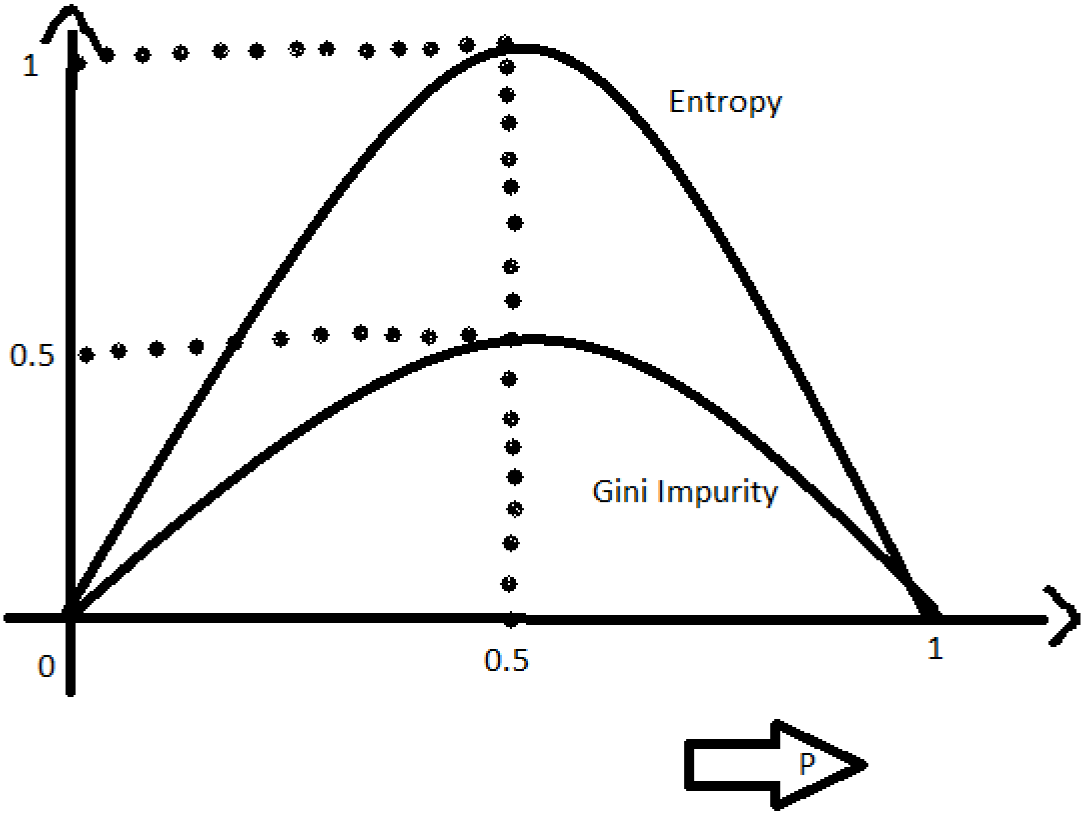
3. Gini Impurity

-> Both Entropy and Gini Impurity Calculate the purity of split in DT.

-> But Gini Impurity is better than Entropy. So we use Gini Impurity to find purity of split.

-> Gini Impurity is a measurement of the likelihood of an incorrect classification of a new instance of a random variable, if that new instance were randomly classified according to the distribution of class labels from the data set.

Lets plot the Entropy Vs Gini impurity



$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

$$Gini(E) = 1 - \sum_{j=1}^{c} p_j^2$$

-> Entropy Max Value is 1 but Gini Impurity Max Value is 0.5.

-> As Gini Impurity has no logarithmic function so it is more efficient than Entropy.

Tree Pruning

Tree pruning is the method of trimming down a full tree (obtained through the above process) to reduce the complexity and variance in the data. Just as we regularised linear regression, we can also regularise the decision tree model by adding a new term.

Post-pruning

Post-pruning, also known as backward pruning, is the process where the decision tree is generated first and then the non-significant branches are removed. Cross-validation set of data is used to check the effect of pruning and tests whether expanding a node will make an improvement or not. If any improvement is there then we continue by expanding that node else if there is reduction in accuracy then the node not be expanded and should be converted in a leaf node.

Pre-pruning

Pre-pruning, also known as forward pruning, stops the non-significant branches from generating. It uses a condition to decide when should it terminate splitting of some of the branches prematurely as the tree is generated.

How does the Decision Tree algorithm Work?

- In a decision tree, for predicting the class of the given dataset, the algorithm starts from the root node of the tree. This algorithm compares the values of root attribute with the record (real dataset) attribute and, based on the comparison, follows the branch and jumps to the next node.
- For the next node, the algorithm again compares the attribute value with the other sub-nodes and move further. It continues the process until it reaches the leaf node of the tree. The complete process can be better understood using the below algorithm:

Step-1: Begin the tree with the root node, says S, which contains the complete dataset.

Step-2: Find the best attribute in the dataset using Attribute Selection Measure (ASM). Here it use Entropy or Gini Impurity.

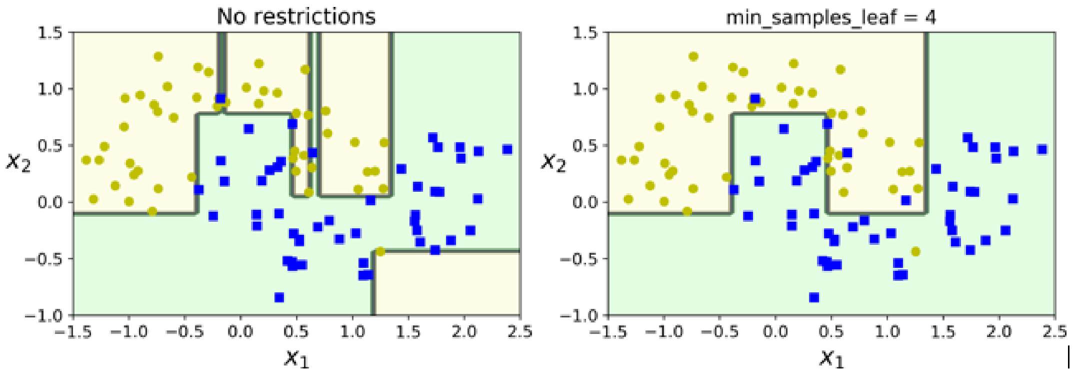
Step-3: Divide the S into subsets that contains possible values for the best attributes.

Step-4: Generate the decision tree node, which contains the best attribute.

Step-5: Recursively make new decision trees using the subsets of the dataset created in step -3. Continue this process until a stage is reached where you cannot further classify the nodes and called the final node as a leaf node.

See An Example Of Above Algorithm

- Here YOu see the overview of DT, On the left, the Decision Tree is trained with the default hyperparameters (i.e., no restrictions), and on the right the Decision Tree is trained with min_sam ples_leaf=4. It is quite obvious that the model on the left is overfitting, and the model on the right will probably generalize better.



Lets Implement With Python

Importing the libraries

In [1]:
```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
```

Importing the dataset

In [2]:
```
dataset = pd.read_csv('Social_Network_Ads.csv')
X = dataset.iloc[:, [2, 3]].values
y = dataset.iloc[:, -1].values
```

Splitting the dataset into the Training set and Test set

In [3]:
```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.25, rando
```

Feature Scaling

In [4]:
```
from sklearn.preprocessing import StandardScaler
sc = StandardScaler()
X_train = sc.fit_transform(X_train)
X_test = sc.transform(X_test)
```

Training the Decision Tree Classification model on the Training set

In [5]:

```
from sklearn.tree import DecisionTreeClassifier
classifier = DecisionTreeClassifier(criterion = 'entropy', random_state = 0)
classifier.fit(X_train, y_train)
```

Out[5]:

```
DecisionTreeClassifier(ccp_alpha=0.0, class_weight=None, criterion='entropy',
                       max_depth=None, max_features=None, max_leaf_nodes=None,
                       min_impurity_decrease=0.0, min_impurity_split=None,
                       min_samples_leaf=1, min_samples_split=2,
                       min_weight_fraction_leaf=0.0, presort='deprecated',
                       random_state=0, splitter='best')
```

Predicting the Test set results

In [6]:

```
y_pred = classifier.predict(X_test)
```

Making the Confusion Matrix

In [7]:

```
from sklearn.metrics import confusion_matrix
cm = confusion_matrix(y_test, y_pred)
print(cm)
```

```
[[62  6]
 [ 3 29]]
```

In [15]:

```
dataset.head()
```

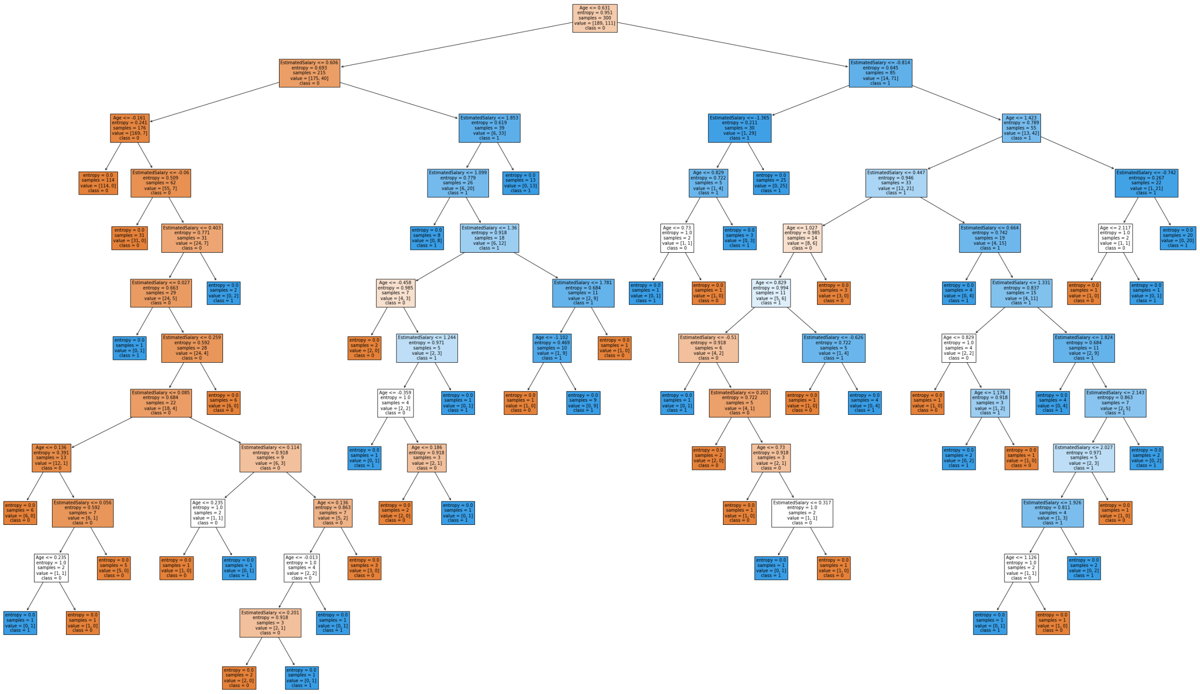
Out[15]:

	User ID	Gender	Age	EstimatedSalary	Purchased
0	15624510	Male	19	19000	0
1	15810944	Male	35	20000	0
2	15668575	Female	26	43000	0
3	15603246	Female	27	57000	0
4	15804002	Male	19	76000	0

In [10]:

```
from sklearn import tree
```

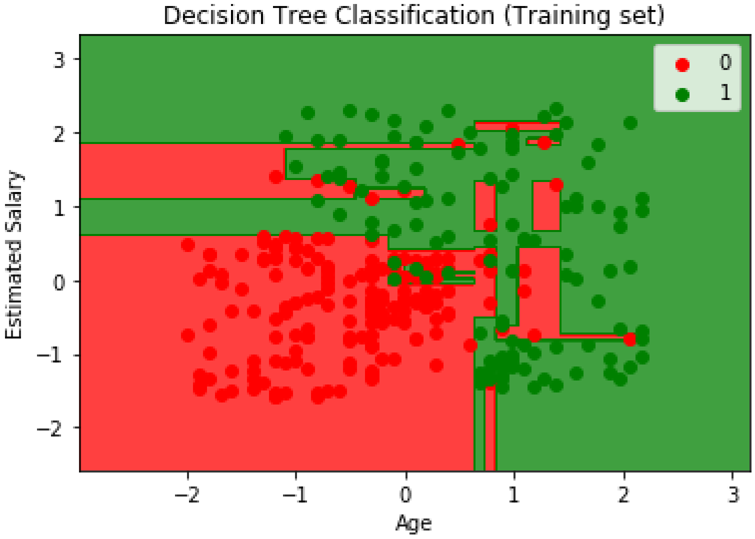
In [20]:
```
fn=['Age','EstimatedSalary']
cn=['0','1']
fig, axes = plt.subplots(nrows = 1,ncols = 1,figsize = (50,30))
tree.plot_tree(classifier,
               feature_names = fn,
               class_names=cn,
               filled = True);
fig.savefig('imagename.png')
```


Age <= 0.631 entropy = 0.951 samples = 300 value = [189, 111] class = 0
EstimatedSalary <= 0.606 entropy = 0.693 samples = 215 value = [175, 40] class = 0
EstimatedSalary <= -0.814 entropy = 0.645 samples = 85 value = [14, 71] class = 1
Age <= -0.161 entropy = 0.241 samples = 176 value = [169, 7] class = 0
EstimatedSalary <= 1.853 entropy = 0.619 samples = 39 value = [6, 33] class = 1
EstimatedSalary <= -1.365 entropy = 0.211 samples = 30 value = [1, 29] class = 1
Age <= 1.423 entropy = 0.789 samples = 55 value = [13, 42] class = 1
entropy = 0.0 samples = 114 value = [114, 0] class = 0
EstimatedSalary <= -0.06 entropy = 0.509 samples = 62 value = [55, 7] class = 0
EstimatedSalary <= 1.099 entropy = 0.779 samples = 26 value = [6, 20] class = 1
entropy = 0.0 samples = 13 value = [0, 13] class = 1
Age <= 0.829 entropy = 0.722 samples = 5 value = [1, 4] class = 1
entropy = 0.0 samples = 25 value = [0, 25] class = 1
EstimatedSalary <= 0.447 entropy = 0.946 samples = 33 value = [12, 21] class = 1
EstimatedSalary <= -0.742 entropy = 0.267 samples = 22 value = [1, 21] class = 1
entropy = 0.0 samples = 31 value = [31, 0] class = 0
EstimatedSalary <= 0.403 entropy = 0.771 samples = 31 value = [24, 7] class = 0
entropy = 0.0 samples = 8 value = [0, 8] class = 1
EstimatedSalary <= 1.36 entropy = 0.918 samples = 18 value = [6, 12] class = 1
Age <= 0.73 entropy = 1.0 samples = 2 value = [1, 1] class = 0
entropy = 0.0 samples = 3 value = [0, 3] class = 1
Age <= 1.027 entropy = 0.985 samples = 14 value = [8, 6] class = 0
EstimatedSalary <= 0.664 entropy = 0.742 samples = 19 value = [4, 15] class = 1
Age <= 2.117 entropy = 1.0 samples = 2 value = [1, 1] class = 0
entropy = 0.0 samples = 20 value = [0, 20] class = 1
EstimatedSalary <= 0.027 entropy = 0.663 samples = 29 value = [24, 5] class = 0
entropy = 0.0 samples = 2 value = [0, 2] class = 1
Age <= -0.458 entropy = 0.985 samples = 7 value = [4, 3] class = 0
EstimatedSalary <= 1.781 entropy = 0.684 samples = 11 value = [2, 9] class = 1
entropy = 0.0 samples = 1 value = [0, 1] class = 1
entropy = 0.0 samples = 1 value = [1, 0] class = 0
Age <= 0.829 entropy = 0.994 samples = 11 value = [5, 6] class = 1
entropy = 0.0 samples = 3 value = [3, 0] class = 0
entropy = 0.0 samples = 4 value = [0, 4] class = 1
EstimatedSalary <= 1.331 entropy = 0.837 samples = 15 value = [4, 11] class = 1
entropy = 0.0 samples = 1 value = [1, 0] class = 0
entropy = 0.0 samples = 1 value = [0, 1] class = 1
entropy = 0.0 samples = 1 value = [0, 1] class = 1
EstimatedSalary <= 0.259 entropy = 0.592 samples = 28 value = [24, 4] class = 0
entropy = 0.0 samples = 2 value = [2, 0] class = 0
EstimatedSalary <= 1.244 entropy = 0.971 samples = 5 value = [2, 3] class = 1
Age <= -1.102 entropy = 0.469 samples = 10 value = [1, 9] class = 1
entropy = 0.0 samples = 1 value = [1, 0] class = 0
EstimatedSalary <= -0.51 entropy = 0.918 samples = 6 value = [4, 2] class = 0
EstimatedSalary <= -0.626 entropy = 0.722 samples = 5 value = [1, 4] class = 1
Age <= 0.829 entropy = 1.0 samples = 4 value = [2, 2] class = 0
EstimatedSalary <= 1.824 entropy = 0.684 samples = 11 value = [2, 9] class = 1
EstimatedSalary <= 0.085 entropy = 0.684 samples = 22 value = [18, 4] class = 0
entropy = 0.0 samples = 6 value = [6, 0] class = 0
Age <= -0.359 entropy = 1.0 samples = 4 value = [2, 2] class = 0
entropy = 0.0 samples = 1 value = [0, 1] class = 1
entropy = 0.0 samples = 1 value = [1, 0] class = 0
entropy = 0.0 samples = 9 value = [0, 9] class = 1
entropy = 0.0 samples = 1 value = [0, 1] class = 1
EstimatedSalary <= 0.201 entropy = 0.722 samples = 5 value = [4, 1] class = 0
entropy = 0.0 samples = 1 value = [1, 0] class = 0
entropy = 0.0 samples = 4 value = [0, 4] class = 1
entropy = 0.0 samples = 1 value = [1, 0] class = 0
Age <= 1.176 entropy = 0.918 samples = 3 value = [1, 2] class = 1
entropy = 0.0 samples = 4 value = [0, 4] class = 1
EstimatedSalary <= 2.143 entropy = 0.863 samples = 7 value = [2, 5] class = 1
Age <= 0.136 entropy = 0.391 samples = 13 value = [12, 1] class = 0
EstimatedSalary <= 0.114 entropy = 0.918 samples = 9 value = [6, 3] class = 0
entropy = 0.0 samples = 1 value = [0, 1] class = 1
Age <= 0.186 entropy = 0.918 samples = 3 value = [2, 1] class = 0
entropy = 0.0 samples = 2 value = [2, 0] class = 0
Age <= 0.73 entropy = 0.918 samples = 3 value = [2, 1] class = 0
entropy = 0.0 samples = 2 value = [0, 2] class = 1
entropy = 0.0 samples = 1 value = [1, 0] class = 0
EstimatedSalary <= 2.027 entropy = 0.971 samples = 5 value = [2, 3] class = 1
entropy = 0.0 samples = 2 value = [0, 2] class = 1
entropy = 0.0 samples = 6 value = [6, 0] class = 0
EstimatedSalary <= 0.056 entropy = 0.592 samples = 7 value = [6, 1] class = 0
Age <= 0.235 entropy = 1.0 samples = 2 value = [1, 1] class = 0
Age <= 0.136 entropy = 0.863 samples = 7 value = [5, 2] class = 0
entropy = 0.0 samples = 2 value = [2, 0] class = 0
entropy = 0.0 samples = 1 value = [0, 1] class = 1
entropy = 0.0 samples = 1 value = [1, 0] class = 0
EstimatedSalary <= 0.317 entropy = 1.0 samples = 2 value = [1, 1] class = 0
EstimatedSalary <= 1.926 entropy = 0.811 samples = 4 value = [1, 3] class = 1
entropy = 0.0 samples = 1 value = [1, 0] class = 0
Age <= 0.235 entropy = 1.0 samples = 2 value = [1, 1] class = 0
entropy = 0.0 samples = 5 value = [5, 0] class = 0
entropy = 0.0 samples = 1 value = [1, 0] class = 0
entropy = 0.0 samples = 1 value = [0, 1] class = 1
Age <= -0.013 entropy = 1.0 samples = 4 value = [2, 2] class = 0
entropy = 0.0 samples = 3 value = [3, 0] class = 0
entropy = 0.0 samples = 1 value = [0, 1] class = 1
entropy = 0.0 samples = 1 value = [1, 0] class = 0
Age <= 1.126 entropy = 1.0 samples = 2 value = [1, 1] class = 0
entropy = 0.0 samples = 2 value = [0, 2] class = 1
entropy = 0.0 samples = 1 value = [0, 1] class = 1
entropy = 0.0 samples = 1 value = [1, 0] class = 0
EstimatedSalary <= 0.201 entropy = 0.918 samples = 3 value = [2, 1] class = 0
entropy = 0.0 samples = 1 value = [0, 1] class = 1
entropy = 0.0 samples = 1 value = [0, 1] class = 1
entropy = 0.0 samples = 1 value = [1, 0] class = 0
entropy = 0.0 samples = 2 value = [2, 0] class = 0
entropy = 0.0 samples = 1 value = [0, 1] class = 1

Visualising the Training set results

In [8]:
```
from matplotlib.colors import ListedColormap
X_set, y_set = X_train, y_train
X1, X2 = np.meshgrid(np.arange(start = X_set[:, 0].min() - 1, stop = X_set[:, 0].
                     np.arange(start = X_set[:, 1].min() - 1, stop = X_set[:, 1].
plt.contourf(X1, X2, classifier.predict(np.array([X1.ravel(), X2.ravel()]).T).res
             alpha = 0.75, cmap = ListedColormap(('red', 'green')))
plt.xlim(X1.min(), X1.max())
plt.ylim(X2.min(), X2.max())
for i, j in enumerate(np.unique(y_set)):
    plt.scatter(X_set[y_set == j, 0], X_set[y_set == j, 1],
                c = ListedColormap(('red', 'green'))(i), label = j)
plt.title('Decision Tree Classification (Training set)')
plt.xlabel('Age')
plt.ylabel('Estimated Salary')
plt.legend()
plt.show()
```

```
'c' argument looks like a single numeric RGB or RGBA sequence, which should be
avoided as value-mapping will have precedence in case its length matches with
'x' & 'y'.  Please use a 2-D array with a single row if you really want to spec
ify the same RGB or RGBA value for all points.
'c' argument looks like a single numeric RGB or RGBA sequence, which should be
avoided as value-mapping will have precedence in case its length matches with
'x' & 'y'.  Please use a 2-D array with a single row if you really want to spec
ify the same RGB or RGBA value for all points.
```



- The above output is completely different from the rest classification models. It has both vertical

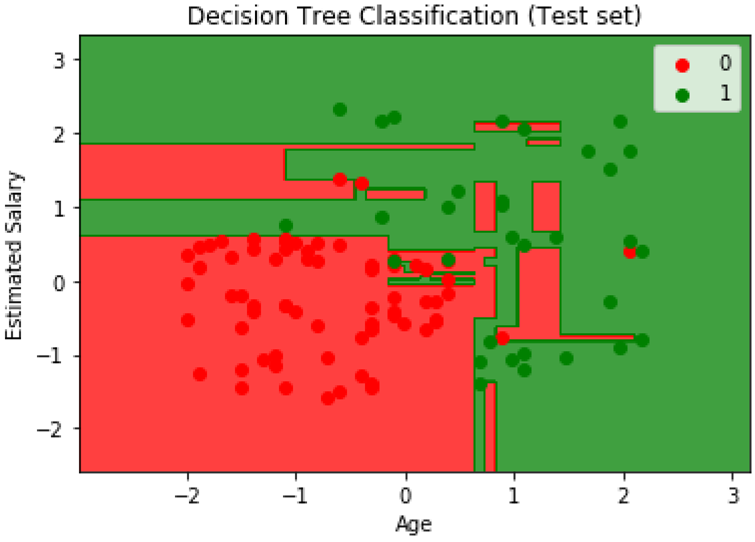
- The above output is completely different from the rest classification models. It has both vertical and horizontal lines that are splitting the dataset according to the age and estimated salary variable.
- As we can see, the tree is trying to capture each dataset, which is the case of overfitting.

Visualising the Test set results

In [9]:

```
from matplotlib.colors import ListedColormap
X_set, y_set = X_test, y_test
X1, X2 = np.meshgrid(np.arange(start = X_set[:, 0].min() - 1, stop = X_set[:, 0].
                     np.arange(start = X_set[:, 1].min() - 1, stop = X_set[:, 1].
plt.contourf(X1, X2, classifier.predict(np.array([X1.ravel(), X2.ravel()]).T).res
             alpha = 0.75, cmap = ListedColormap(('red', 'green')))
plt.xlim(X1.min(), X1.max())
plt.ylim(X2.min(), X2.max())
for i, j in enumerate(np.unique(y_set)):
    plt.scatter(X_set[y_set == j, 0], X_set[y_set == j, 1],
                c = ListedColormap(('red', 'green'))(i), label = j)
plt.title('Decision Tree Classification (Test set)')
plt.xlabel('Age')
plt.ylabel('Estimated Salary')
plt.legend()
plt.show()
```

```
'c' argument looks like a single numeric RGB or RGBA sequence, which should be
avoided as value-mapping will have precedence in case its length matches with
'x' & 'y'.  Please use a 2-D array with a single row if you really want to spec
ify the same RGB or RGBA value for all points.
'c' argument looks like a single numeric RGB or RGBA sequence, which should be
avoided as value-mapping will have precedence in case its length matches with
'x' & 'y'.  Please use a 2-D array with a single row if you really want to spec
ify the same RGB or RGBA value for all points.
```



As we can see in the above image that there are some green data points within the red region and vice versa. So, these are the incorrect predictions.