**Curneu**

**Problem Statement-2 Documentation**

**Problem statement-2**

Try to understand the dataset of Social\_Network\_Ads.csv and try to find the best suitable ML algorithm and write the code in python for algorithm from scratch and try to achieve the below output plot

**Dataset:**

The dataset we have the User ID, Gender – of user ,Age– of user ,Salary– of user and the data if Purchase made by a user.

**Data cleaning:**

**Data cleaning** is the process of preparing **data** for analysis by removing or modifying **data** that is incorrect, incomplete, irrelevant, duplicated, or improperly formatted. This **data** is usually not necessary or helpful when it comes to analysing **data** because it may hinder the process or provide inaccurate results.

* Check if there is any null values in the dataset

**Data Splitting:**

The basic idea is to divide the dataset T into two subsets – one subset is used for training while the other subset is left out and the performance of the final model is evaluated on it. The main purpose of cross-validation is to achieve a stable and confident estimate of the model performance

* We split the model into 75% for training and 25% for test

**Data scaling :**

Feature **scaling** (also known as **data** normalization) is the method used to standardize the range of features of **data**. Since, the range of values of **data** may vary widely, it becomes a necessary step in **data** preprocessing while using machine learning algorithms

* We have scaled the data because we have considered Age and Estimated salary to make the purchase prediction.There is a big difference in the reange of these features.

**Model Selection**

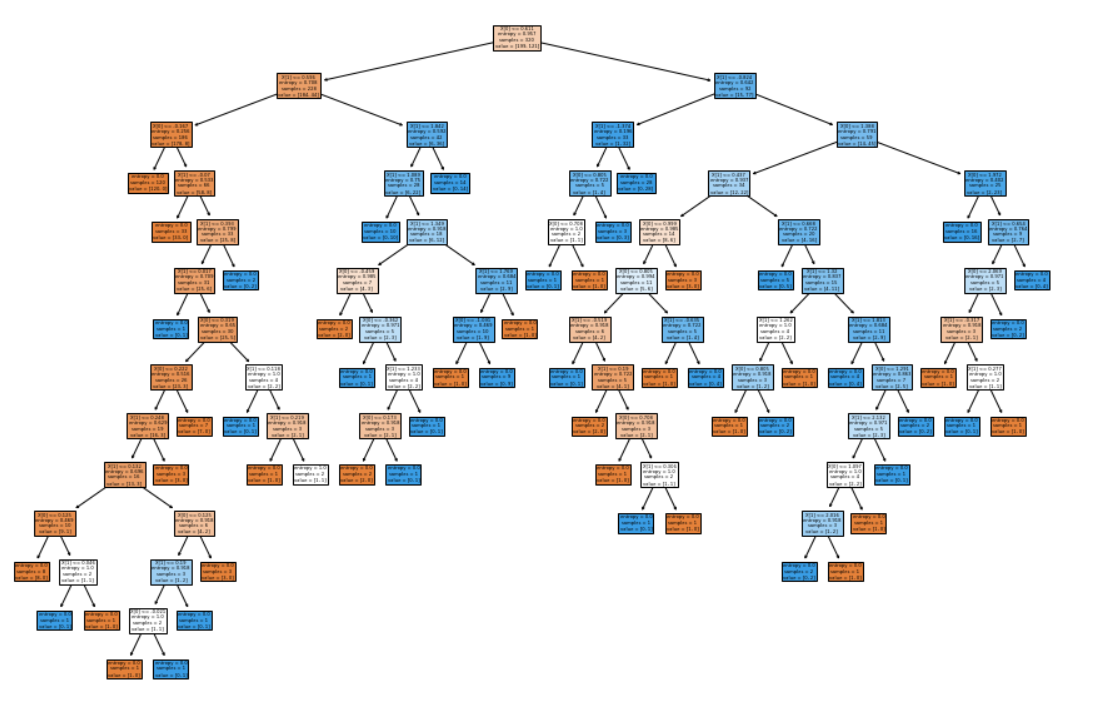
**Decision Tree Classifier:**

**Decision trees** use multiple **algorithms** to decide to split a node into two or more sub-nodes. The creation of sub-nodes increases the homogeneity of resultant sub-nodes.The **decision tree** splits the nodes on all available variables and then selects the split which results in most homogeneous sub-nodes.

**Model Evaluation:**

**Plotting Decision Tree:**

* We can see that our tree has many nodes. So Splitting has taken place many times. The problem with this is that we may have good accuracy on our training Dataset because the model learns by heart the values. We we try to predict the result for unseen data (test) data there is possibility of our accuracy reducing.



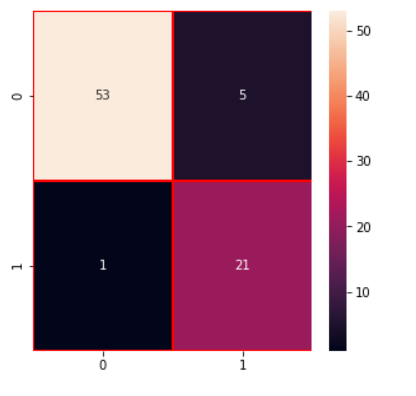
**Accuracy(r2 score):**

**R2** corresponds to the squared correlation between the observed outcome values and the predicted values by the model. The Higher the R-squared, the better the model

* We can see that we have 100% accuracy for our training test and around 92.5% accuracy for our test set.This shows that our training model set is overfitting.

### Confusion Matrix :

A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known. The confusion matrix itself is relatively simple to understand, but the related terminology can be confusing.

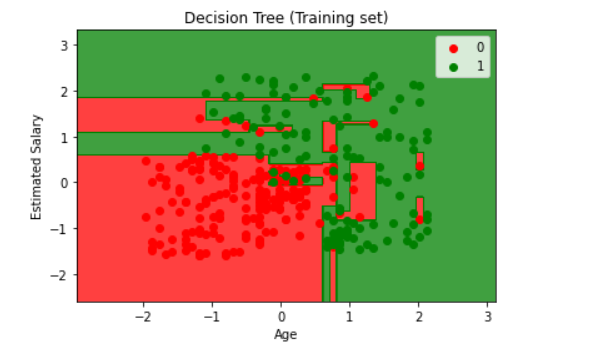


### Classification Report:

A **Classification report** is used to measure the quality of predictions from a **classification** algorithm. How many predictions are True and how many are False. More specifically, True Positives, False Positives, True negatives and False Negatives are used to predict the metrics of a **classification report.**

* We have quite good value of Accuracy and F1 Score.

**Visualising the Training Set:**



**Model From Scratch:**

* Create two classes one for node and other for the decision tree.
* Inside the decision tree class define functions for model fitting, prediction, tree working using information gain and tree traversal methods
* Pass the train and test inputs to the model.
* It will display the predicted results and the accuracy.