**SANCHEZ INNOCENCIA D**

**MSC DECISION AND COMPUTING SCIENCES**

**PROBLEM STATEMENT 1: - MANDATORY**

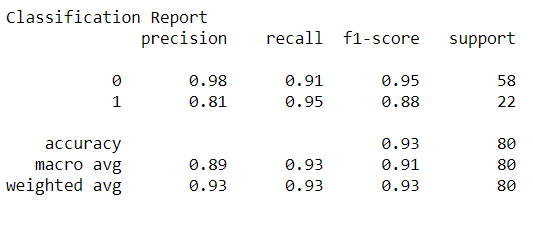
Try to understand the dataset of Social\_Network\_Ads.csv and try to find the best suitable ML algorithm and write the code in python for algorithm from scratch and try to achieve the below output plot.

**ML ALGORITHM:** DECISION TREE CLASSIFIER

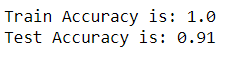
The data has been first imported and then pre-processed. And then the data has been scaled because we have considered Age and Estimated salary to make the purchase prediction. There is a big difference in the range of these features. And then the model is built. The tree has many nodes. So, Splitting 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. Then we have predicted the result for unseen data (test) data there is possibility of our accuracy reducing. Then the accuracy was found, where we have 100% accuracy for our training set and around 92.5% accuracy for our test set. This shows that our training model set is overfitting. However, the model is overfitted by capturing irrelevant patterns that does not have any/less data points. We can prune the data. And through the Classification Report, we have a quite good value of Accuracy and F1 Score. We can see that the decision tree classifier is creating optimal splits to predict the outcome based on the variables Age and Estimated Salary. The Decision Tree Classifier Model is been fitted. And also the Decision Tree is been plotted.

**OUTPUT:**

**CLASSIFICATION REPORT**

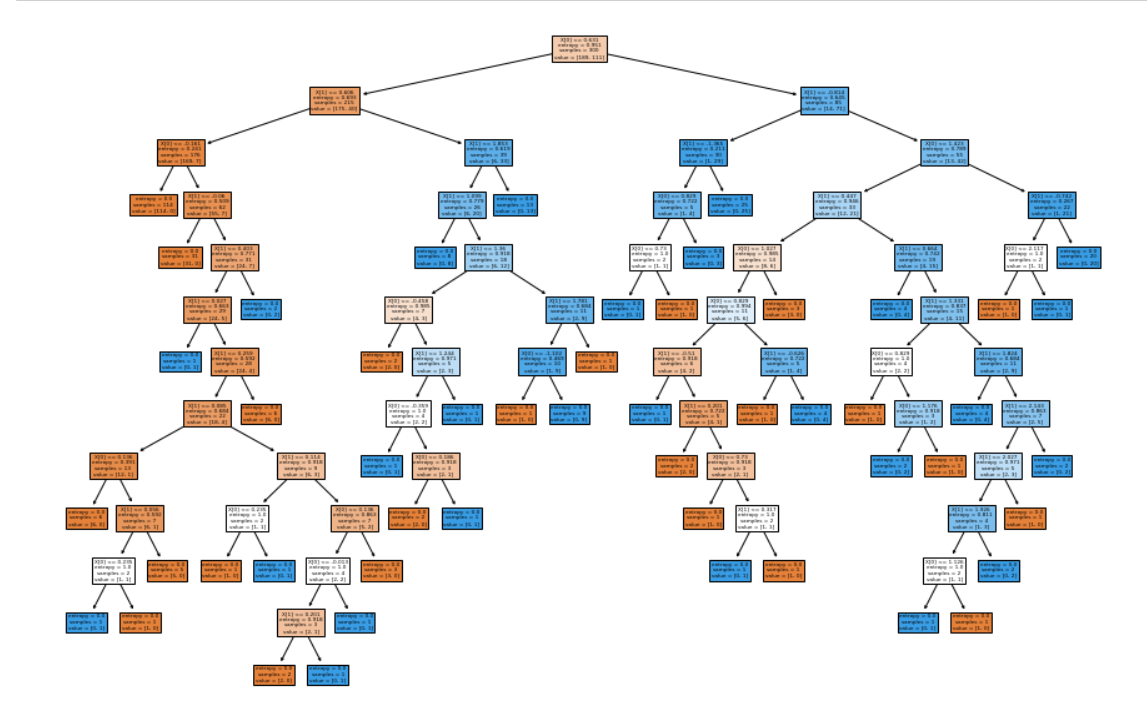


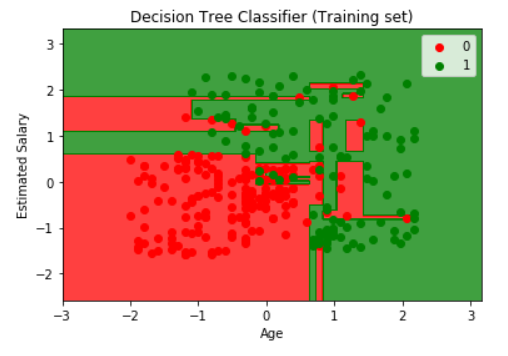
**ACCURACY SCORE**



**DECISION TREE**

The Decision Tree has been plotted



**DECISION TREE CLASSIFIER (TRAINING SET)**

**DECISION TREE CLASSIFIER (TESTING SET)**

