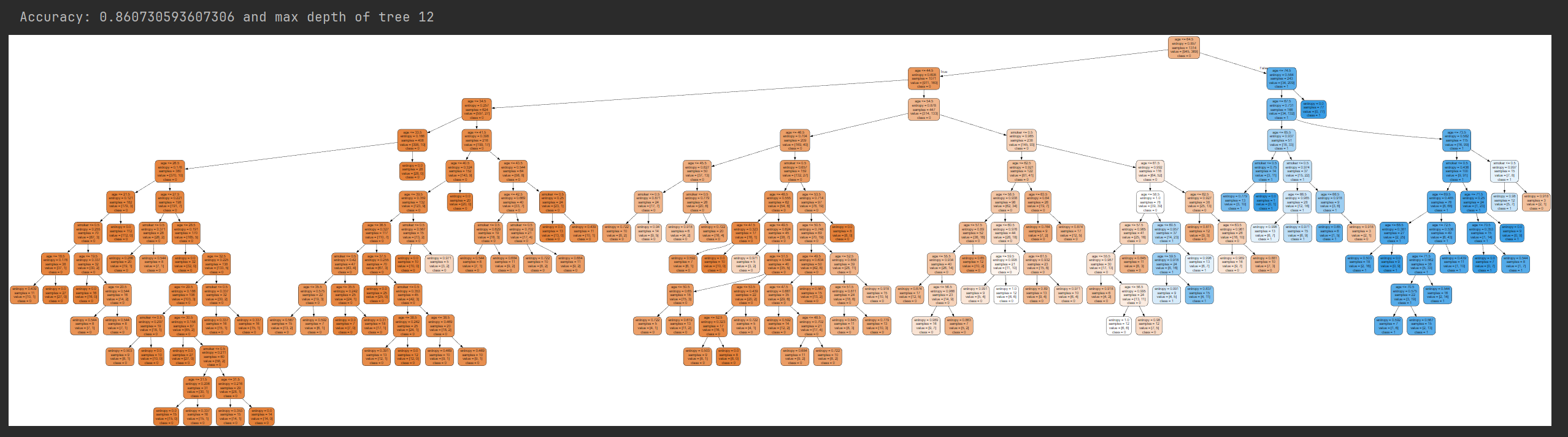
* Original data set

The model has used default parameters with criterion='entropy', which creates max depth tree of 12.



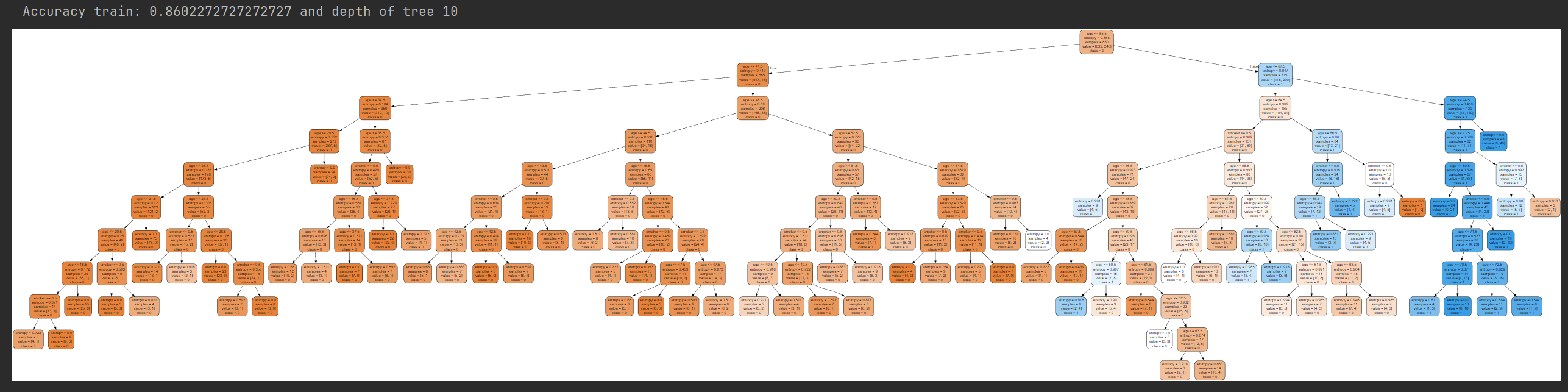
For the following task we are going to use models from 1 to 13 levels of depth and going to compare how the Accuracy of the model is changing by increasing the depth of the tree

* Split dataset

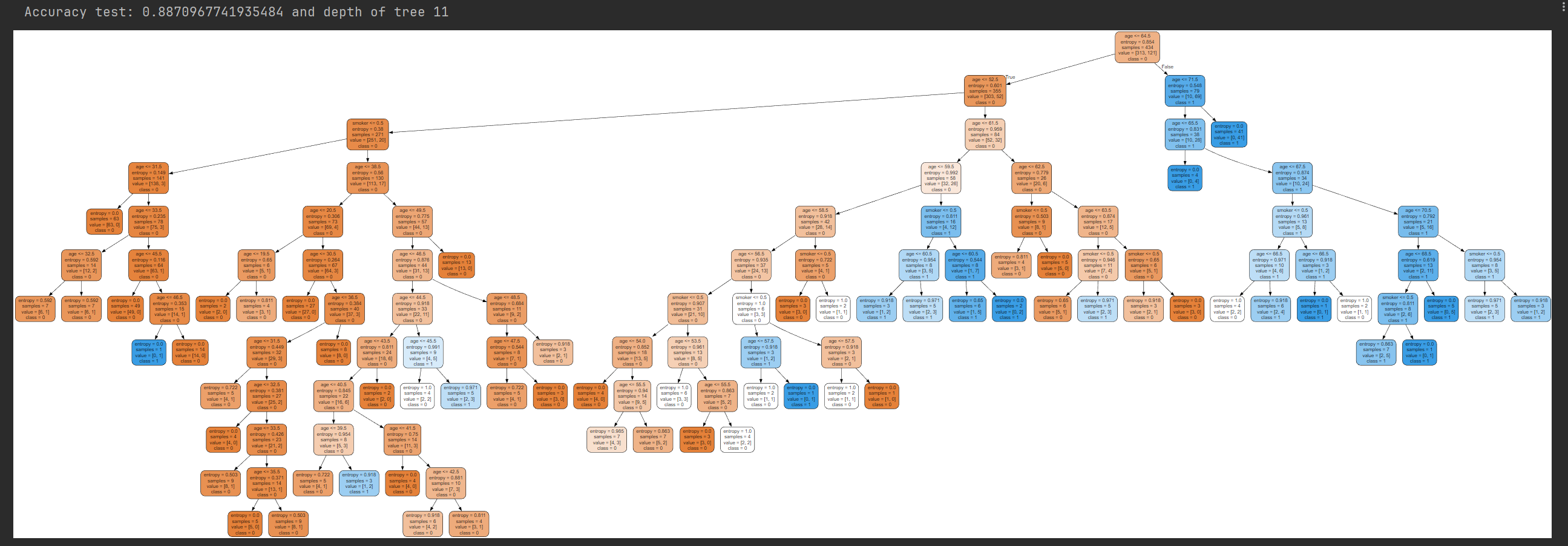
Data has been split on training and test data set using sklearn module

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.33, random\_state=42)

Train Decision tree model



Test Decision tree model





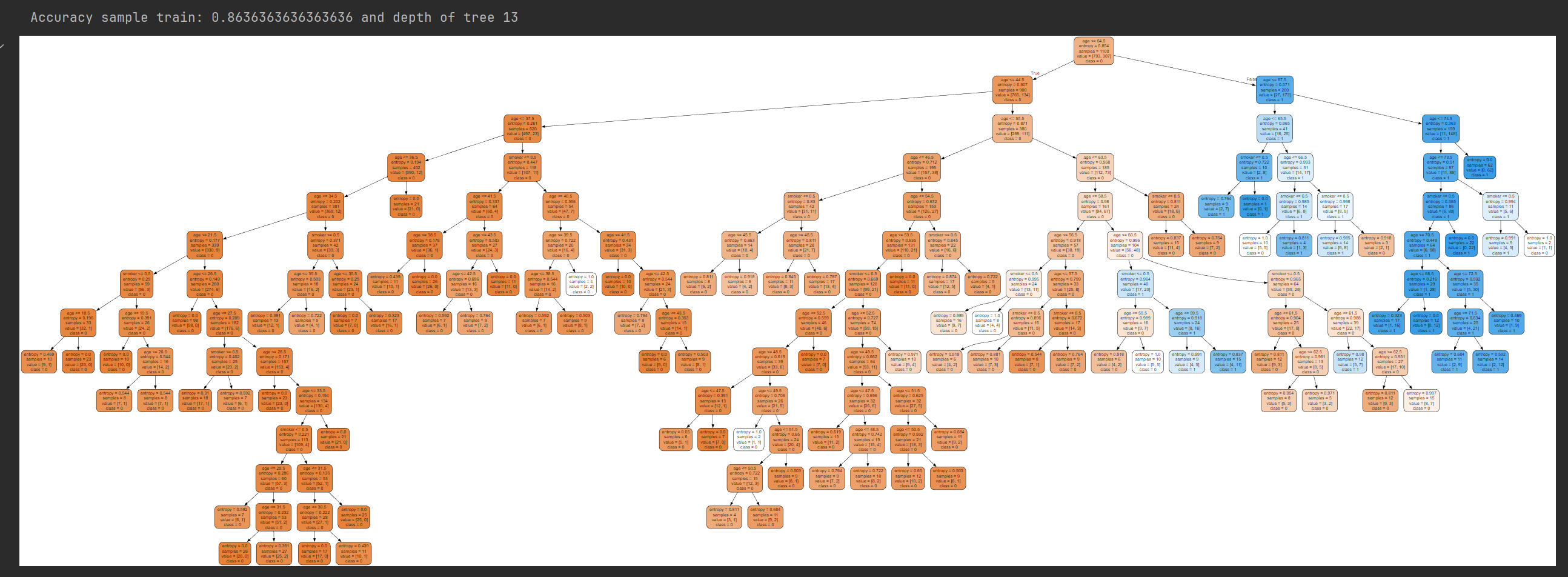
From the above plot we can see the test dataset has the best accuracy of the model at depth of tree of 9 and above

* Split sample dataset

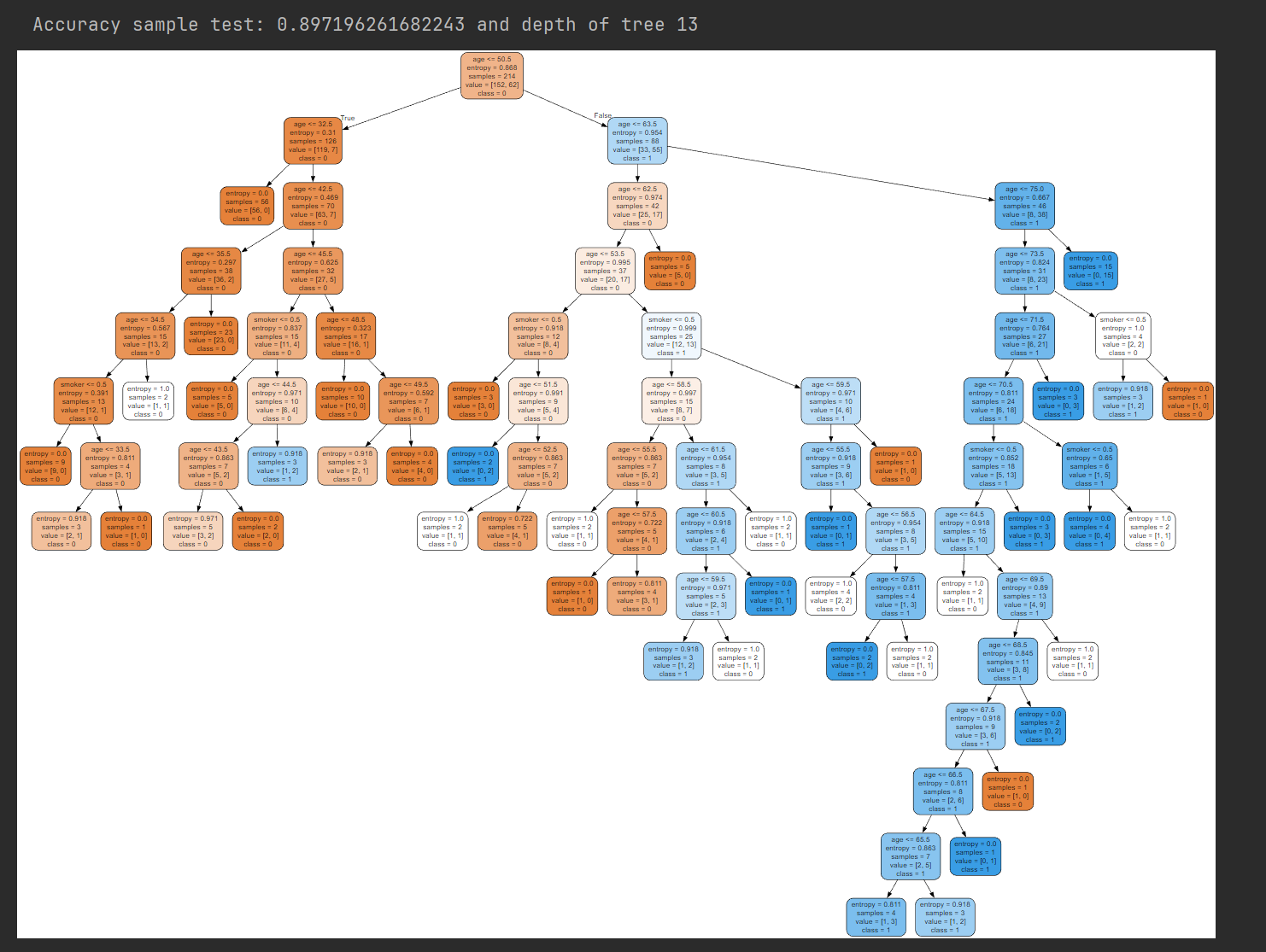
Data has been split on random sample of 1100 out of 1300 using numpy library

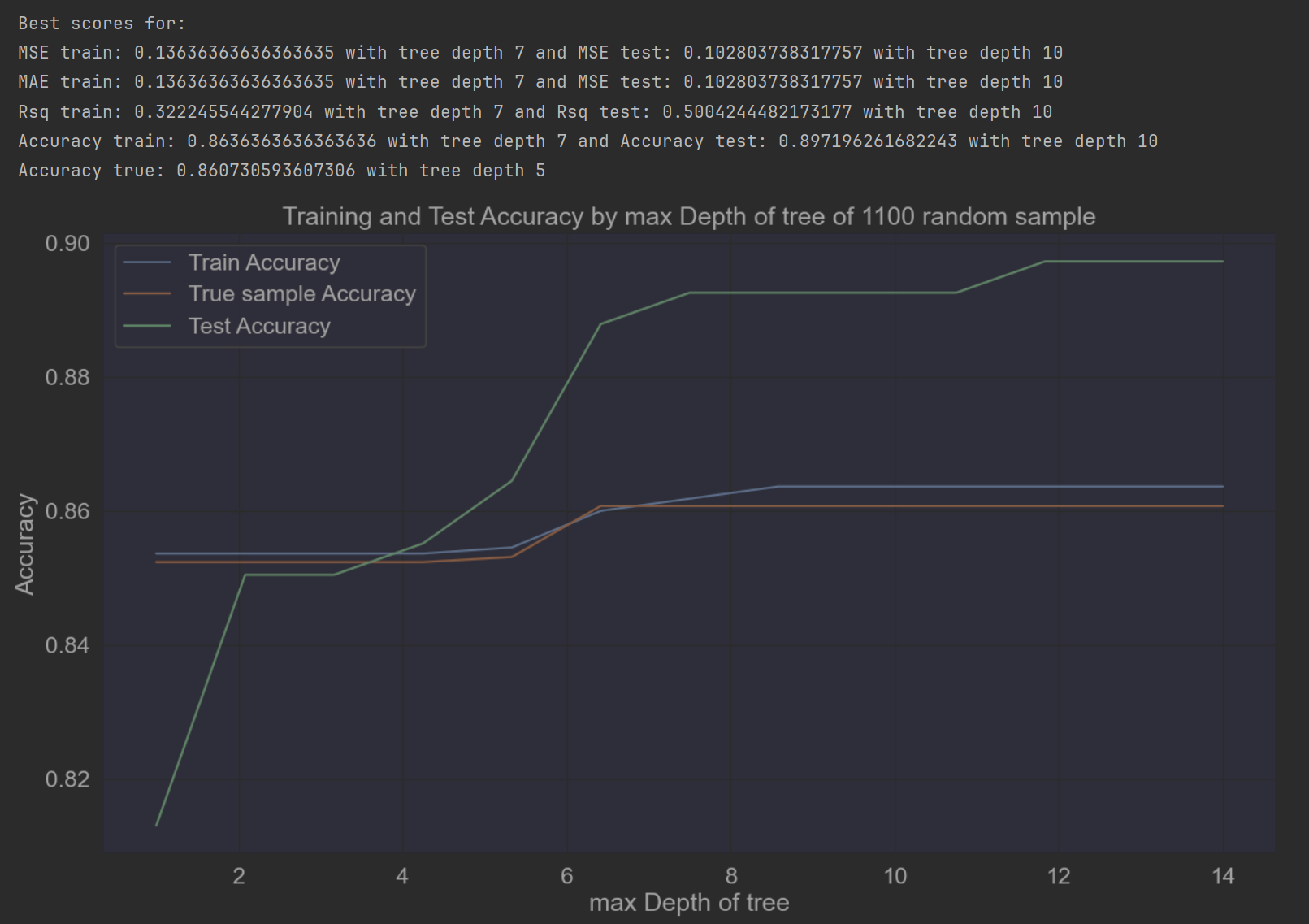
train\_data, test\_data = np.split(data.sample(frac=1), [1100])

Train Decision tree model



Test Decision tree model





From the above plot we notice that the Accuracy of the test sample dataset has higher value than training and the original true dataset.

Conclusion is the training of the sample has the best Accuracy of the model and the least MSE and MAE metrics

From both splits we can say that the model based on the training dataset has got the highest Accuracy value at depth of tree 7. As the model based on true dataset has the best Accuracy value at depth level of 5, we may say the model will overfit between depth level 5-7.

You are right I noticed that convergence at depth level 13, 14, 15, but after splitting the dataset the tree will reshape and it might go further up.