Thanks to H. Ozgur Cagdas on the Slack channel, I discovered GridSearchCV, a tool which takes a dictionary of parameter values, builds all possible models, and returns the model with the best result. My understanding is that we can only use the initial 16 records to build our model as the last 8 are the records we have to predict.

These are the steps I took:

1. Clean the data. This time, I divided the age feature into 3 groups and did not binarize the first- and third-party claims– I just made sure that they were integers. I removed No Claims and Experienced Driver as these were not included in the data we had to predict Good Driver for.
2. Build the model using GridSearchCV and the following parameter ranges: 'min\_samples\_split':range(2,5), 'max\_depth':range(2,6), 'min\_samples\_leaf': range(1,4). I also set cv=4 which means that the X\_train group would be split into 4 sub-groups for the cross validation.
3. Printed out the scores achieved by the best model, and the parameters it used.
4. Predicted the results of Good Driver for the 8 records we were given.

**Timeline

Description automatically generated**

Results on the training data:

[[0 0]

[0 0]

[0 0]

[1 0]

[0 0]

[0 0]

[1 0]

[1 1]

[0 0]

[0 0]

[0 0]

[0 0]

[1 1]

[1 1]

[0 0]

[1 1]]

[[10 2]

[ 0 4]]

Accuracy Score:

0.875

precision recall f1-score support

0 1.00 0.83 0.91 12

1 0.67 1.00 0.80 4

accuracy 0.88 16

macro avg 0.83 0.92 0.85 16

weighted avg 0.92 0.88 0.88 16

Best parameter setting:

GridSearchCV classifier score 0.875

{'max\_depth': 4, 'min\_samples\_leaf': 1, 'min\_samples\_split': 3}

Predicted results for the 8 records:

[0 1 0 0 1 1 0 0]