Lab AI

Patrick AOUN | 201710226

# Why is it that every time you rerun your complete code, you get different results?

Every time we train\_test\_split our data, we get random sample.

# We saw in the Loans problem that we need to replace the purpose column by numerical values. Explain why we did not have to do the same thing in the Kyphosis problem for the Kyphosis column

It’s depends on the correlation. The purpose column affects the prediction results for loans while it doesn't for Kyphosis so translating it to numerical value will let the program take  this column into consideration and be able to identify the relation between this column and  the predicted one

# Table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | accuracy | precision | recall | F1-score |
| DT | 0.74 | 0.19 | 0.24 | 0.21 |
| 1 | 0.74 | 0.16 | 0.18 | 0.16 |
| 10 | 0.81 | 0.26 | 0.04 | 0.05 |
| 100 | 0.82 | 0.44 | 0.02 | 0.03 |
| 1000 | 0.82 | 0.41 | 0.01 | 0.02 |
| 10000 | 0.84 | 0.45 | 0.02 | 0.04 |

# From the results you obtained in part 3, you may see that the  predictions can/might be somehow improved. ‘Feature engineering’ can  be a good way to do it. Conduct some research about the “feature  engineering” process and explain how it could improve your model

Feature engineering, the second step in the [machine learning pipeline](https://towardsdatascience.com/how-to-create-value-with-machine-learning-eb09585b332e), takes in the [label times from the first step](https://towardsdatascience.com/prediction-engineering-how-to-set-up-your-machine-learning-problem-b3b8f622683b) — prediction engineering — and a raw dataset that needs to be refined. Feature engineering means building features for each label while filtering the data used for the feature based on the label’s cutoff time to make valid features. These features and labels are then passed to modeling where they will be used for training a machine learning algorithm.

The RF algorithm works by making vertical and horizontal  splits of the data it creates half spaces of the form. If your data lives on some oblique  plane, then the splitting algorithm has to work harder to capture these diagonal lines,  which could justify a rotation of the data.