Decision Tree – Clinical Dataset

Problem Statement – Predict claim status based on clinical drug details

Data View

	List ID	PGO Generic Product ID	PNO Current Drug Status	PNO Qty Maximum	PNO Days Supply Minimum	PNO Qty Minimum	PNO Days Supply Maximum	PNO Period Qty Days	PNO Period Qty Maximum	RSTNDCLST	SMART_PA_SCH	Status
0	ACEACSF	07000070000120	N	0.0	0	0.0	0	0	0.0	NaN	NaN	P
1	ACEACSF	07000070002520	N	0.0	0	0.0	0	0	0.0	ACFPG00089	ACFPG00089	R
2	ACEACSF	12109025000320	R	0.0	0	0.0	0	0	0.0	NaN	NaN	R
3	ACEACSF	12109902260320	R	0.0	0	0.0	0	0	0.0	NaN	NaN	R
4	ACEACSF	12109903270320	R	0.0	0	0.0	0	0	0.0	NaN	NaN	R

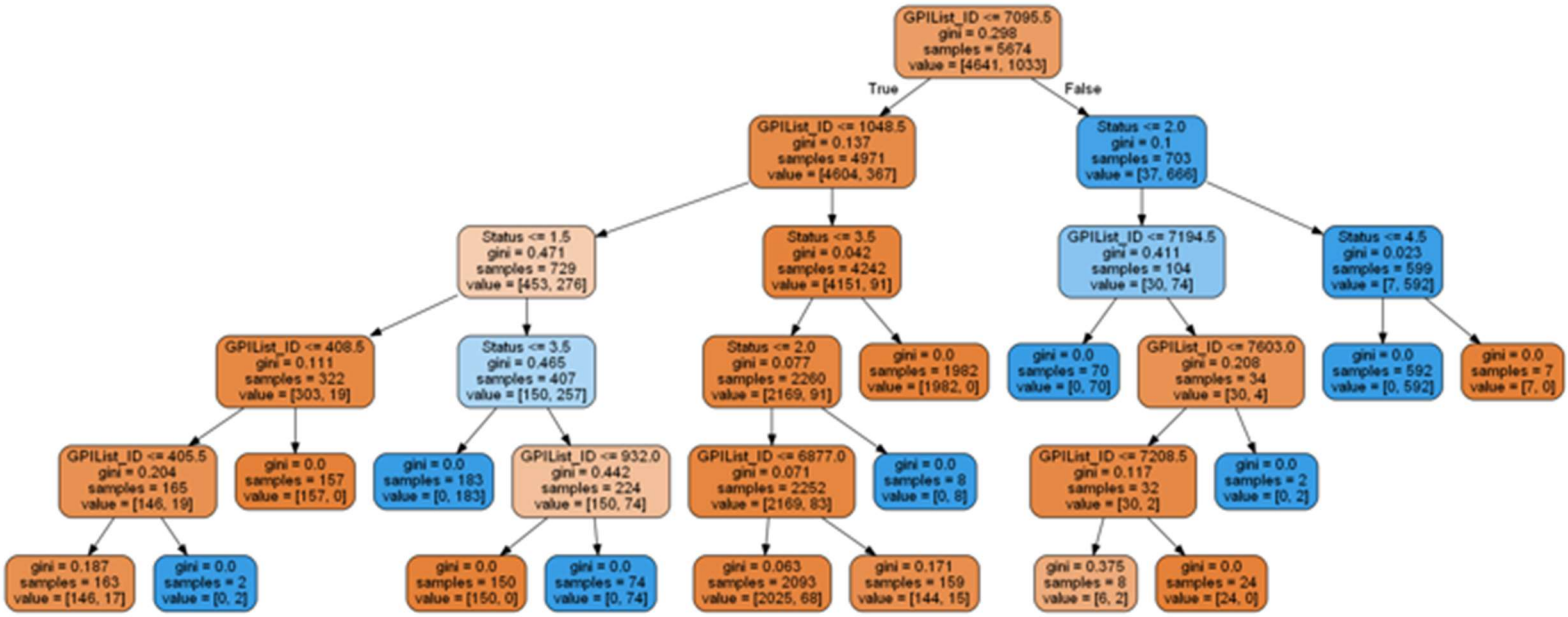
Observations:

1) There are categorical and numerical values
2) Categorical variables need encoding
3) Missing values needs treatment

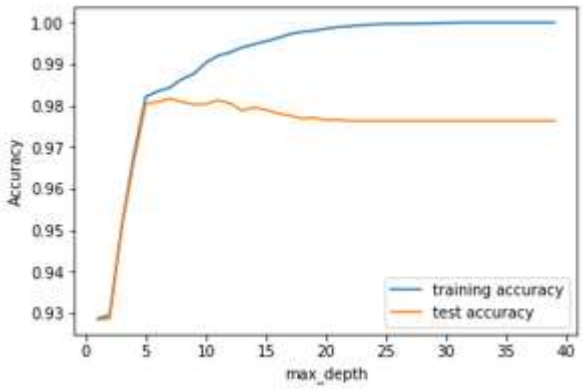
Data View after Exploratory Data Analysis

	PNO Qty Maximum	PNO Days Supply Minimum	PNO Qty Minimum	PNO Days Supply Maximum	PNO Period Qty Days	PNO Period Qty Maximum	PNO Current Drug Status	Status	GPIList_ID
0	0.0	0	0.0	0	0	0.0	1	0	63
1	0.0	0	0.0	0	0	0.0	1	1	64
2	0.0	0	0.0	0	0	0.0	3	1	65
3	0.0	0	0.0	0	0	0.0	3	1	66
4	0.0	0	0.0	0	0	0.0	3	1	67

First Decision tree with default hyperparameters and no-cross-validation tuning.

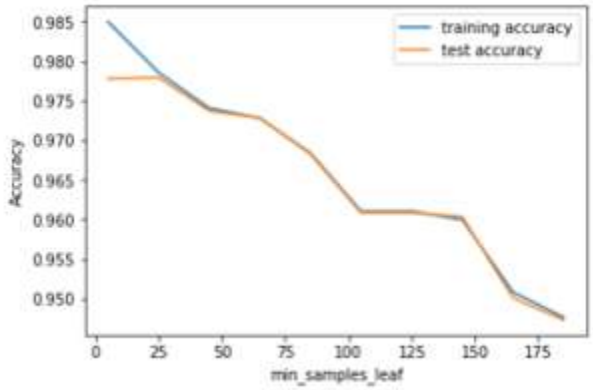


Tuning max_depth



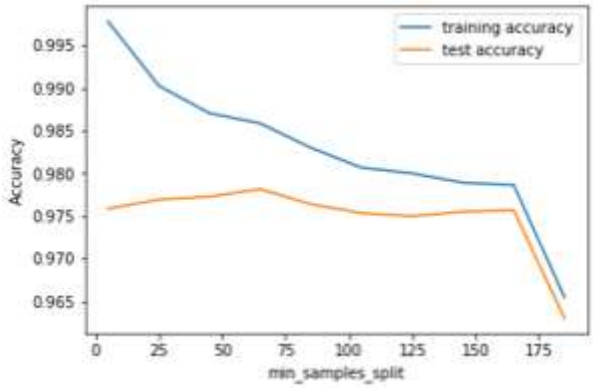
You can see that as we increase the value of max_depth, both training and test score increase till about max-depth = 5, after which the test score gradually reduces. Note that the scores are average accuracies across the 5-folds.

Tuning min_samples_leaf



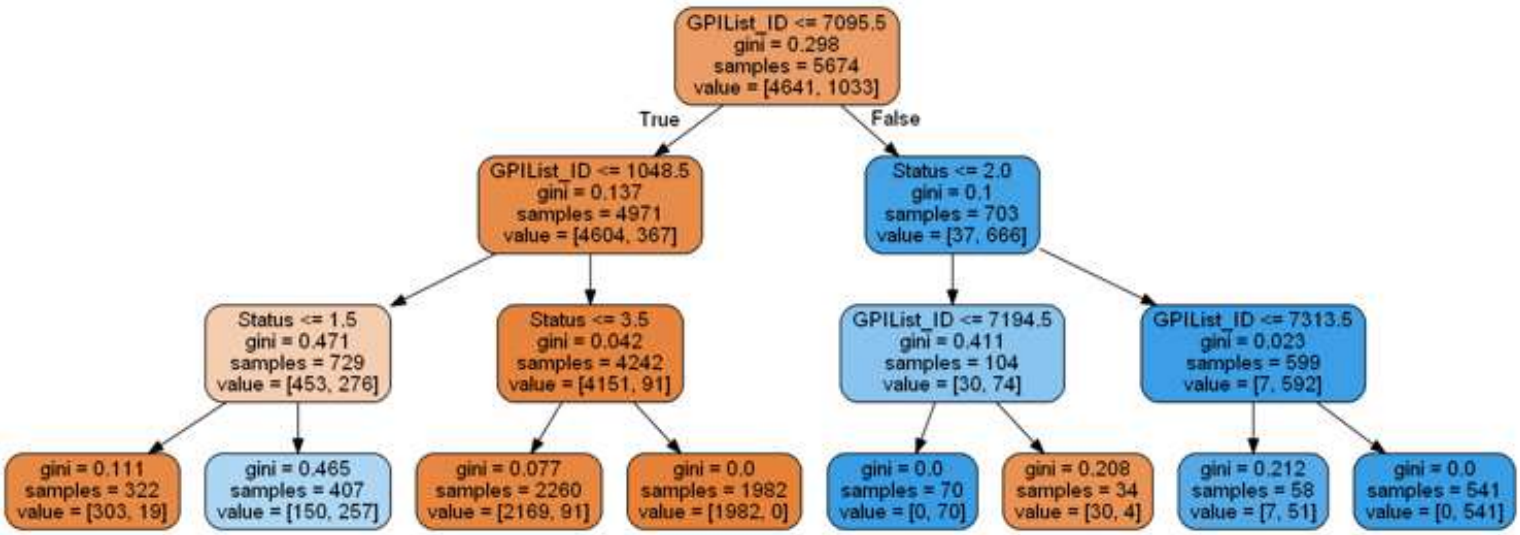
You can see that at low values of min_samples_leaf, the tree gets a bit overfitted. At values > 25, however, the model becomes more stable and the training and test accuracy start to converge.

Tuning min_samples_split



This shows that as you increase the min_samples_split, the tree overfits lesser since the model is less complex.

Decision tree with hyperparameters and cross-validation tuning. Also have reduced the max_depth to 3 to reduce the model complexity



Classification metrics

	precision	recall	f1-score	support
0	0.97	0.97	0.97	1968
1	0.89	0.88	0.88	465
micro avg	0.96	0.96	0.96	2433
macro avg	0.93	0.93	0.93	2433
weighted avg	0.96	0.96	0.96	2433

Confusion metrics

```
[[1915   53]
 [  54  411]]
```