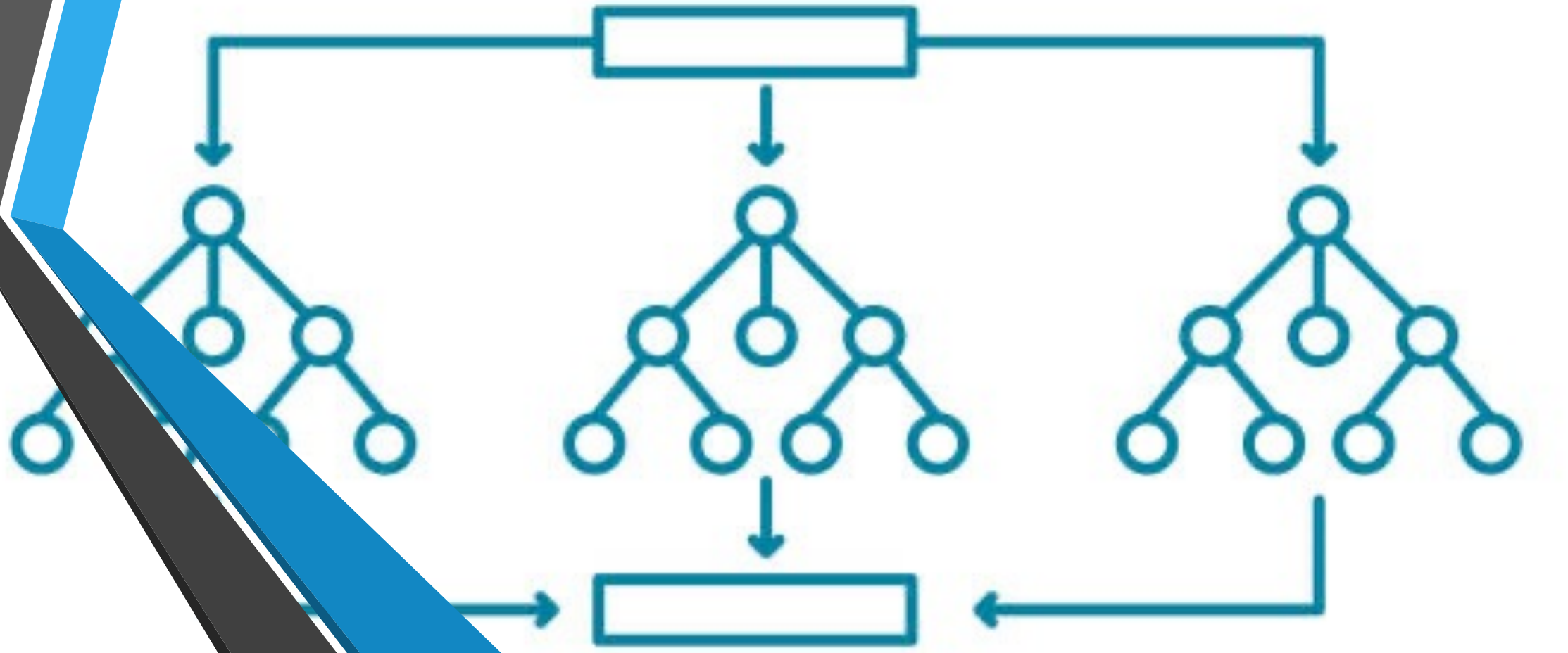
Decision Tree



- Raj Sarode

Data Pre-processing

```
loan_data.isna().sum()
```

```
Sex                0
Age                0
Time_at_address    0
Res_status         0
Occupation         0
Job_status         0
Time_employed      0
Time_bank          0
Liab_ref           0
Acc_ref            0
Home_Expn          0
Balance            0
Decision           0
dtype: int64
```

1. Checking for missing values & null values.
2. Encoding the categorical variables using OneHotencoding and LabelEncoding.
3. Scaled the data and found that accuracy was reduced.

	Age	Time_at_address	Time_employed	Time_bank	Home_Expn	Balance	Sex_F	Sex_M	Res_status_owner	Res_status_rent	...	Job_status_military
0	50.750000	0.585	0	0	145	0	0	1	1	0	...	0
1	19.670000	10.000	0	0	140	0	0	1	0	1	...	0
2	52.830002	15.000	5	14	0	2200	1	0	1	0	...	0
3	22.670000	2.540	2	0	0	0	0	1	0	1	...	0
4	29.250000	13.000	0	0	228	0	0	1	1	0	...	0
...	...	...	...	...	...	...	...	...	...	...	...	...
424	34.169998	2.750	2	0	232	200	0	1	1	0	...	0
425	22.250000	1.250	3	0	280	0	1	0	0	1	...	0
426	23.330000	1.500	1	0	422	200	0	1	1	0	...	0
427	21.000000	4.790	2	1	80	300	0	1	0	1	...	0
428	27.750000	1.290	0	0	140	0	0	1	1	0	...	0

Entropy Method VS Gini Method

1. The gini impurity measures the frequency at which any element of the dataset will be mislabelled when it is randomly labelled.

2. The minimum value of the Gini Index is 0. This happens when the node is **pure**, this means that all the contained elements in the node are of one unique class. Therefore, this node will not be split again.

3. Thus, the optimum split is chosen by the features with less Gini Index. Moreover, it gets the maximum value when the probability of the two classes are the same.

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2$$

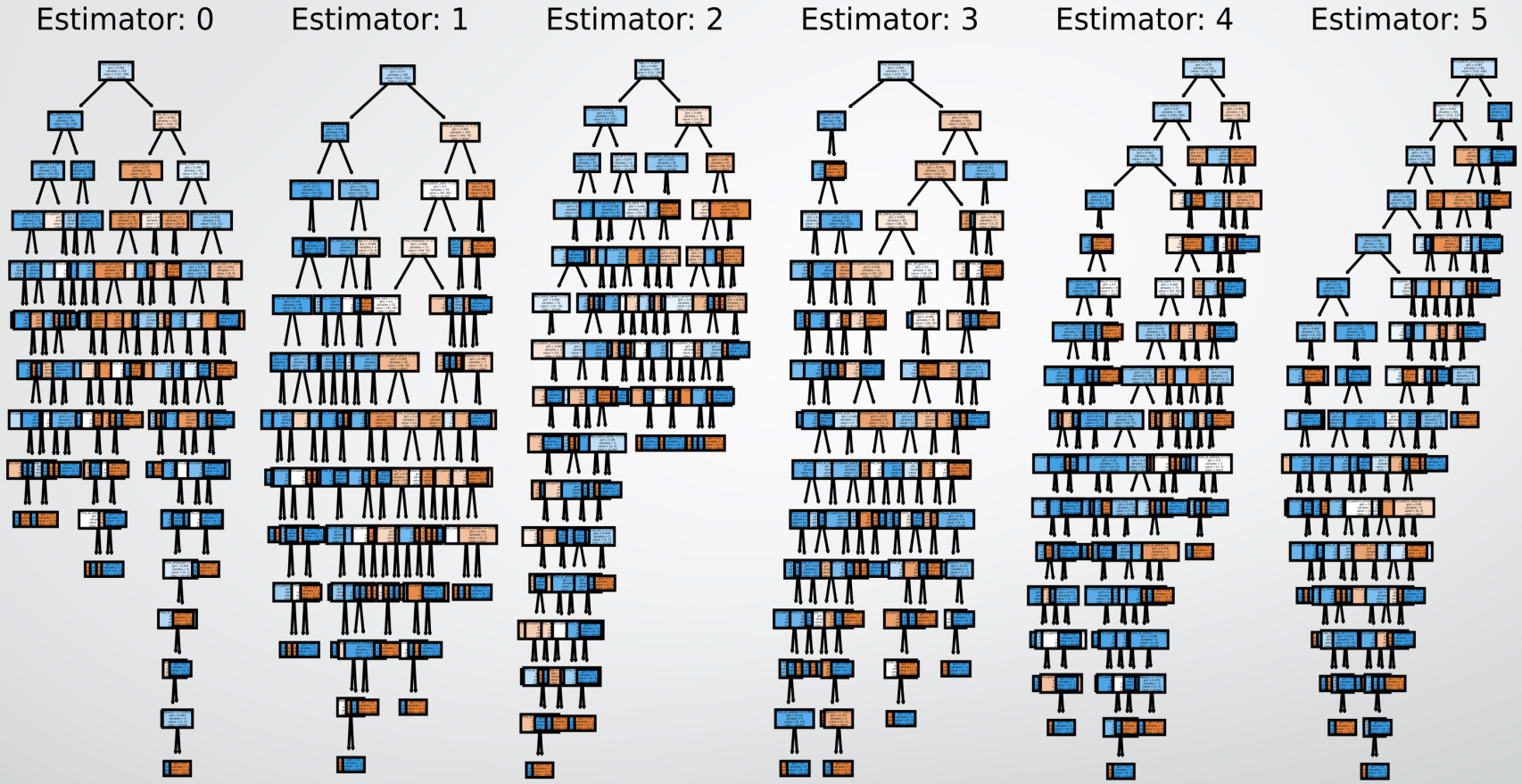
1. Entropy is defined as the randomness or measuring the disorder of the information being processed in Machine Learning.

2. Further, in other words, we can say that entropy is the machine learning metric that measures the unpredictability or impurity in the system.

3. When entropy becomes 0, then the dataset has no impurity. Datasets with 0 impurities are not useful for learning. Further, if the entropy is 1, then this kind of dataset is good for learning.

$$Entropy = \sum_{i=1}^{C} -p_i * \log_2(p_i)$$

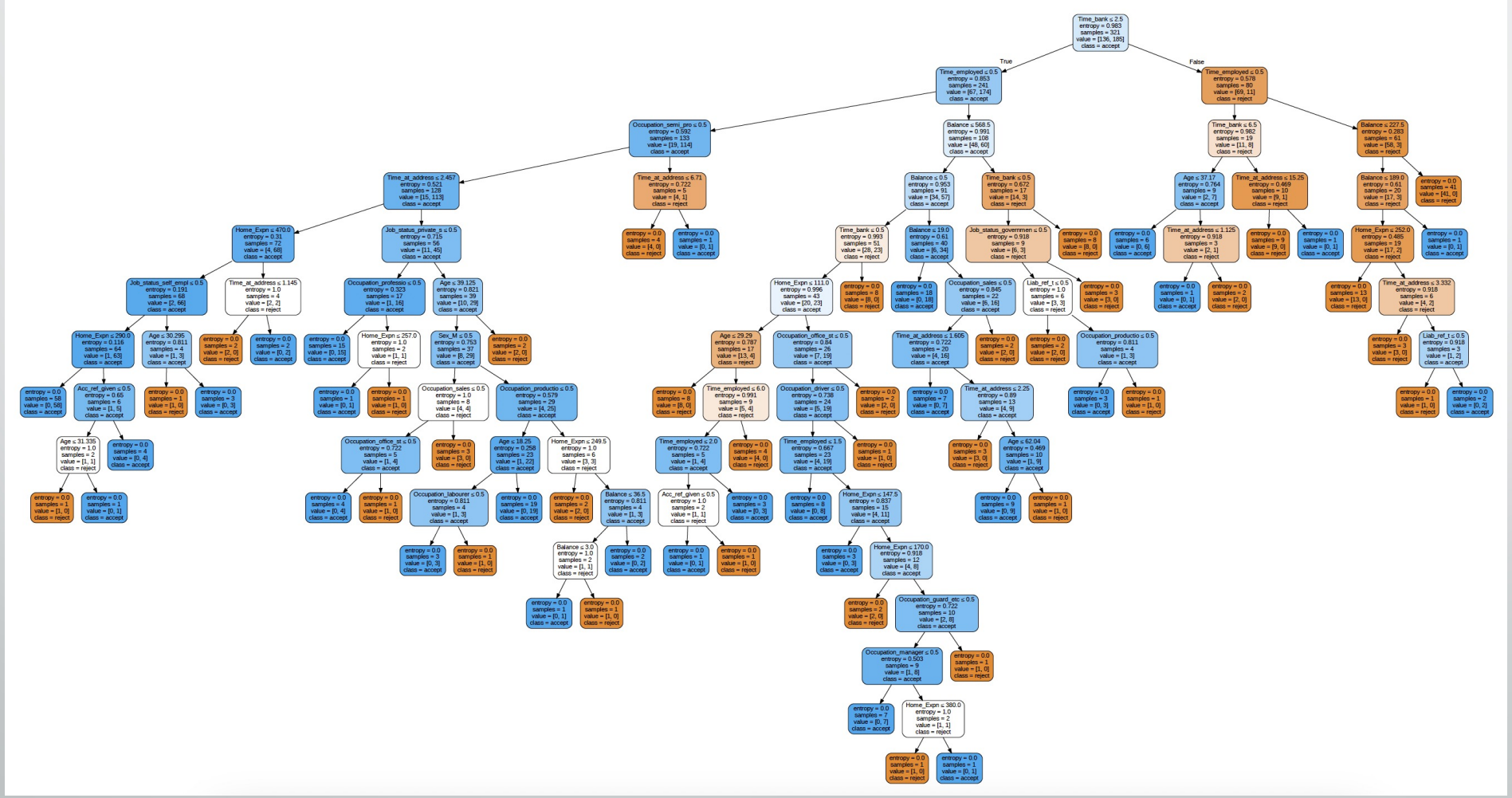
Random Forest



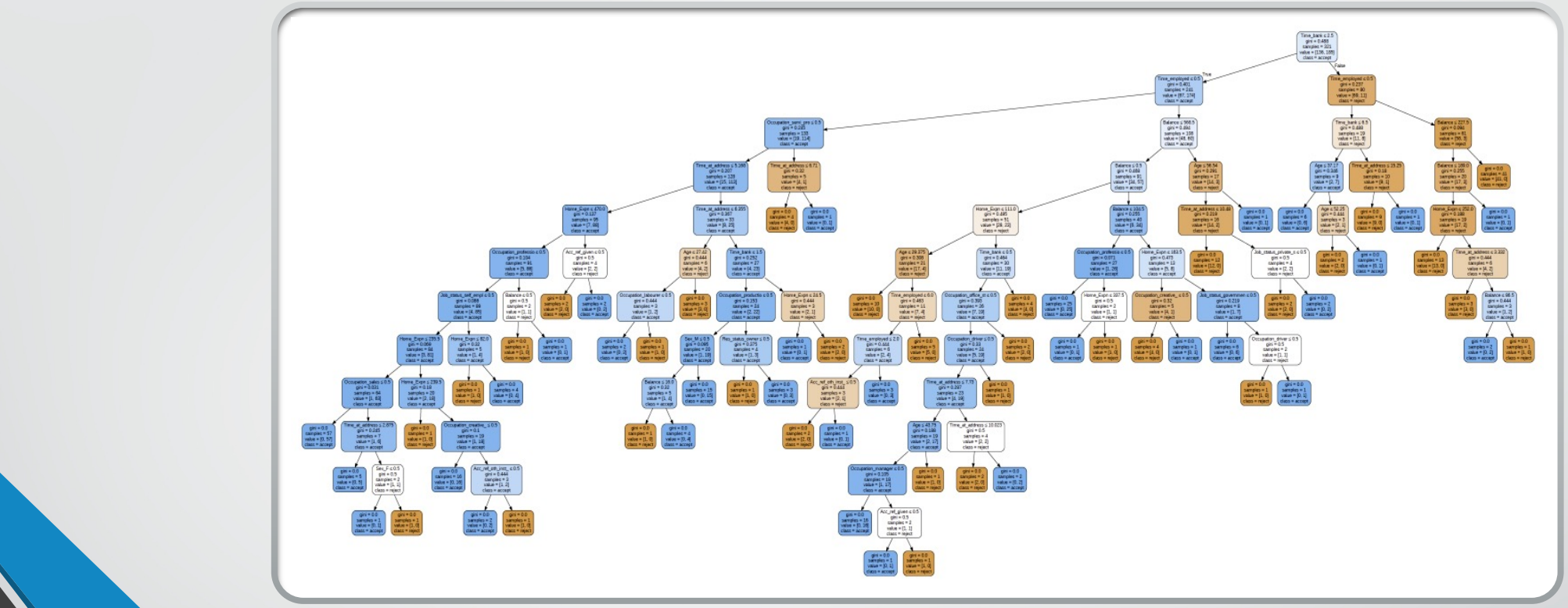
It can perform both regression and classification tasks. A random forest produces good predictions that can be understood easily. It can handle large datasets efficiently. The random forest algorithm provides a higher level of accuracy in predicting outcomes over the decision tree algorithm. Works well with non-linear data. Lower risk of overfitting. Runs efficiently on a large dataset. Better accuracy than other classification algorithms.

RESULTS

Decision Tree by Entropy Method



RESULTS



Decision Tree by Gini Method