EFFİCİENT TELEMARKETİNG

Enes YILDIRIM

20211603001

Demİr SARRAÇ

20231603050

BASİS OF OUR DATA

	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	y
3571	26	admin.	married	secondary	no	2469	no	no	cellular	16	jul	136	8	-1	0	unknown	no
3366	52	unknown	married	primary	no	247	no	no	cellular	29	jul	268	6	-1	0	unknown	no
2722	53	blue-collar	married	secondary	no	25	no	no	cellular	22	aug	528	2	-1	0	unknown	yes
1916	51	entrepreneur	married	tertiary	no	3921	yes	no	cellular	5	may	168	1	-1	0	unknown	no
2923	39	admin.	married	secondary	no	260	yes	no	cellular	17	apr	146	1	281	1	failure	no

Age: Numeric, continious

Job: Categorical, should be learnt how many unique job are there and using skleran.OneHotEncoder we should make this column numeric 1

Marital: Categorical

Education: Categorical but also ordinal. Education can be unknown,secondary,primary, or tertiary. To make it more reasonable, we'll replace "unknown" with NaN and then fill them with mode. And then, using OrdinalEncoder we make this column ordinal. 2

Default: Categorical,

Balance: Numeric, continious. Since there are many rich people, we may need to normalize this feature.

Housing: Categorical

Loan: Categorical

Contact: Categorical, communication type

Day: Numeric, day of the month of last contact. We may drop tihs column, but first, we should see whether is there a correlation between day and y or not

Month: Categorical, we may drop this column.

Duratin: Numeric, duration of the last contact in seconds. We may want to drop this column too, there might be a data leakage, which makes our model working exceptionally. 3

Campaign: Numeric

Pdays: Numeric, days since the client was las contacted (1 means never contacted before). We may need to create a feature called isContacted to make#### 1 values more meaningful.

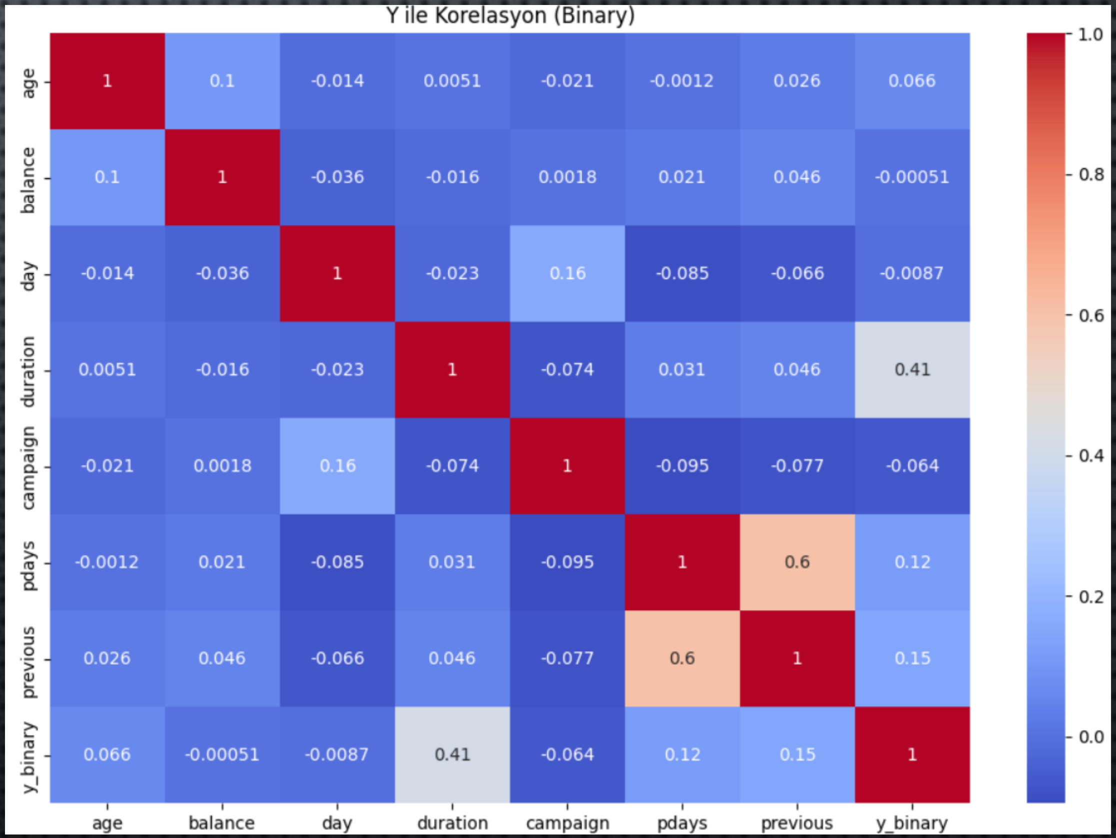
Previous: Numeric, number of previous contacts before this campaign.

Poutcome: Categorical, outcome of the previous marketing campaign. This can be unknown, failure, success. We may replace unknown with NaN an then fill this blanks with the mode.

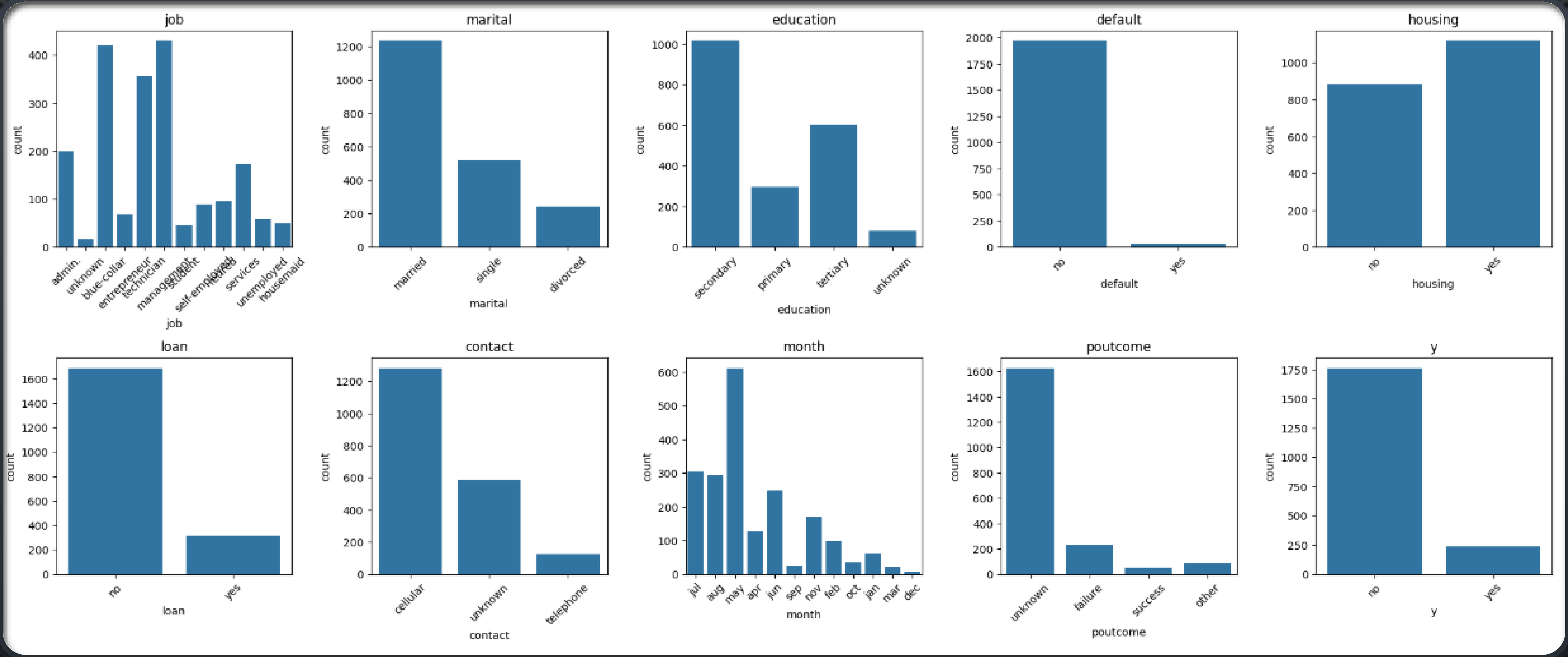
y: Categorical, has the client subscribed to a term deposit? The feature we're trying to predict.

PROBLEM DEFİNİTİON

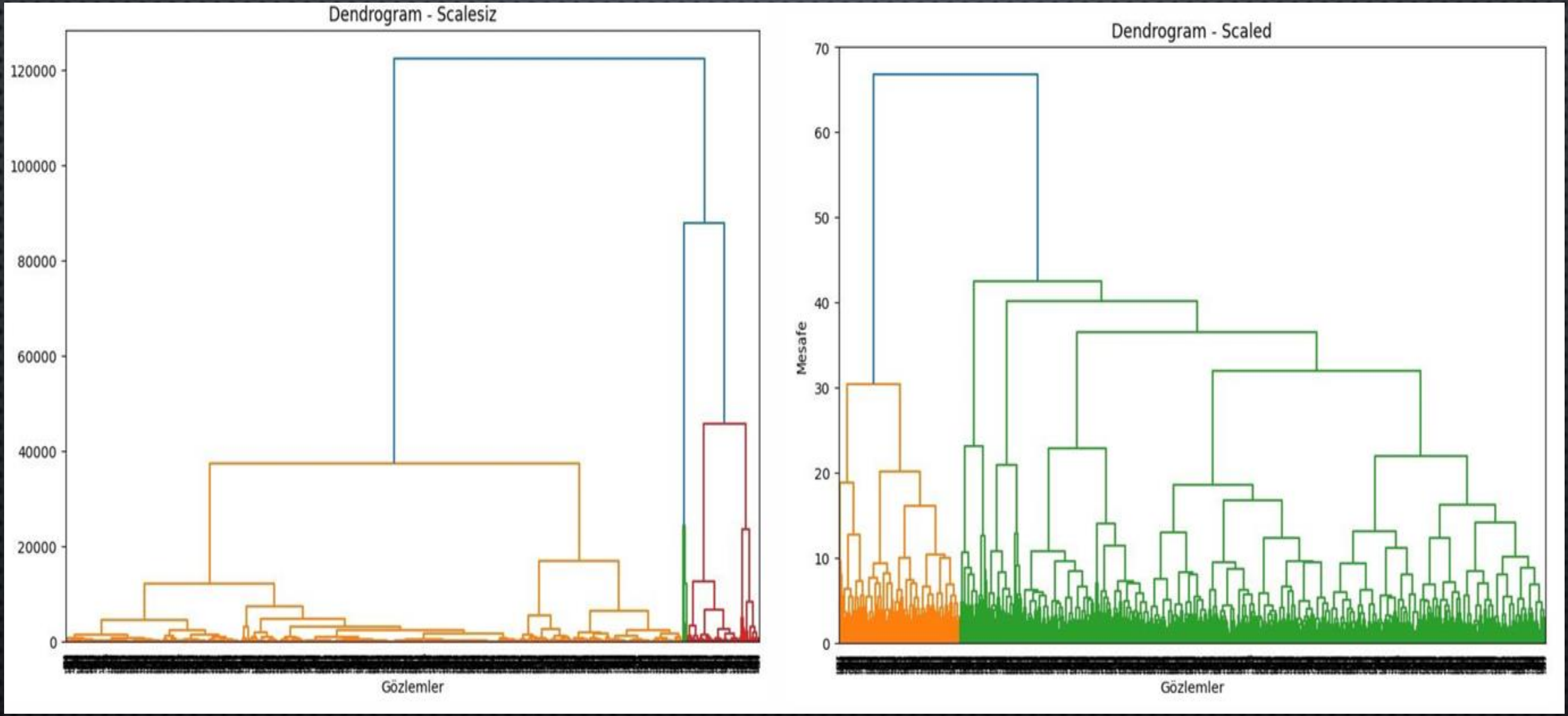
At that poİnt, we realİzed that the duratİon column had an overly strong İnfluence on the model's accuracy. Therefore, we decİded to drop İt İn order to achİeve more effİcİent results.



SO WE AND UP WITH THESE DATAS

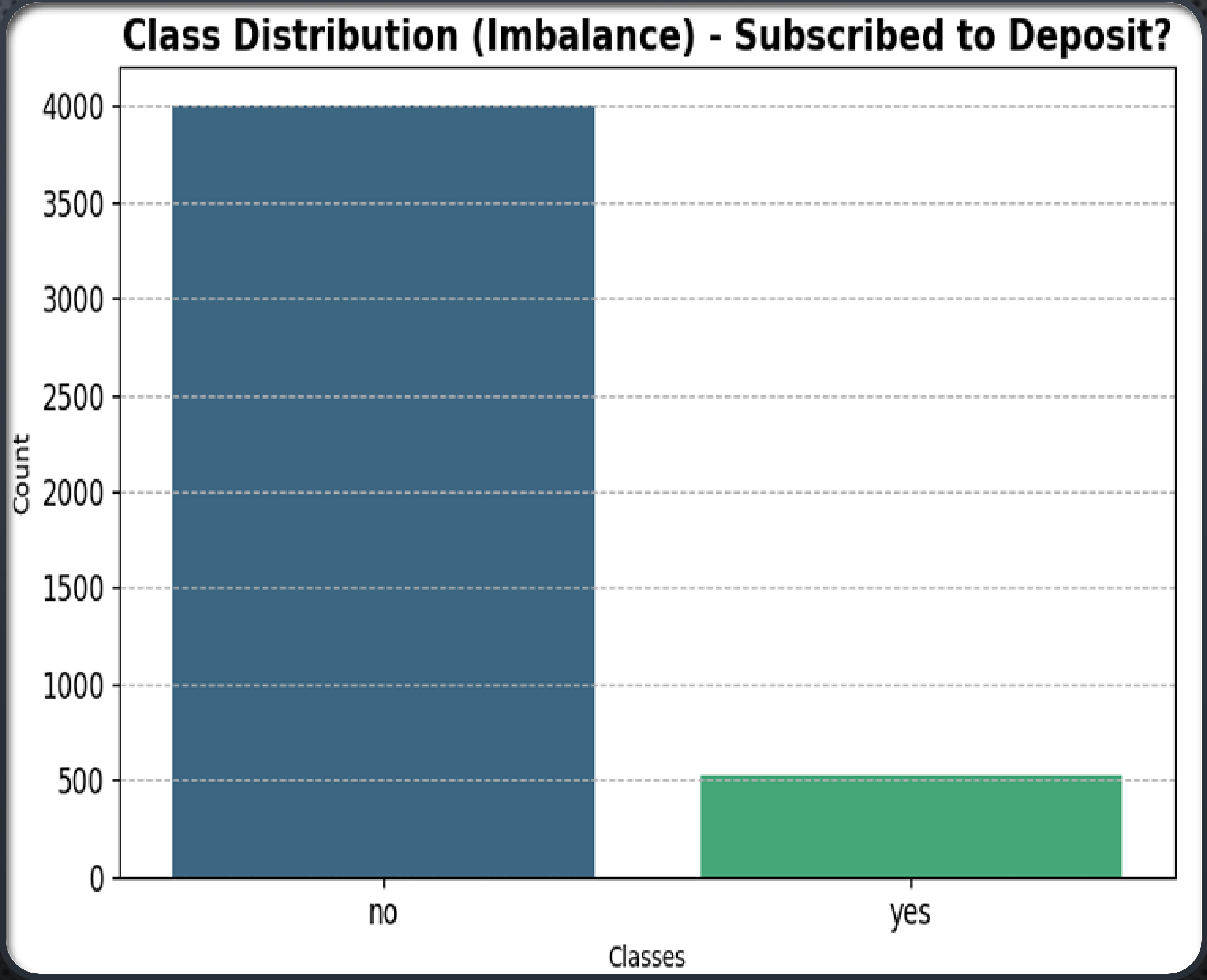


DENDROGRAMS

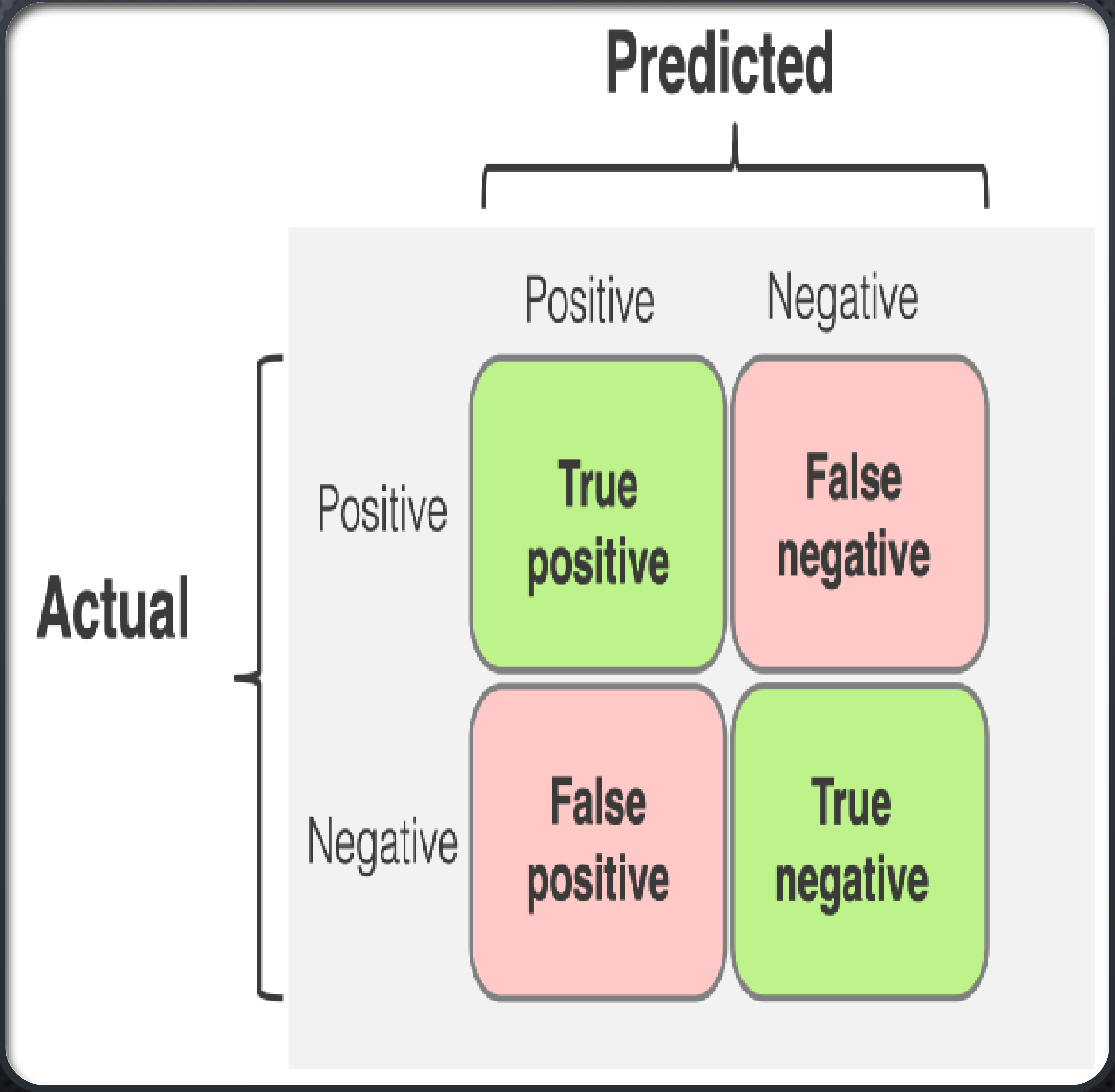


IMBALANCE DATA SET

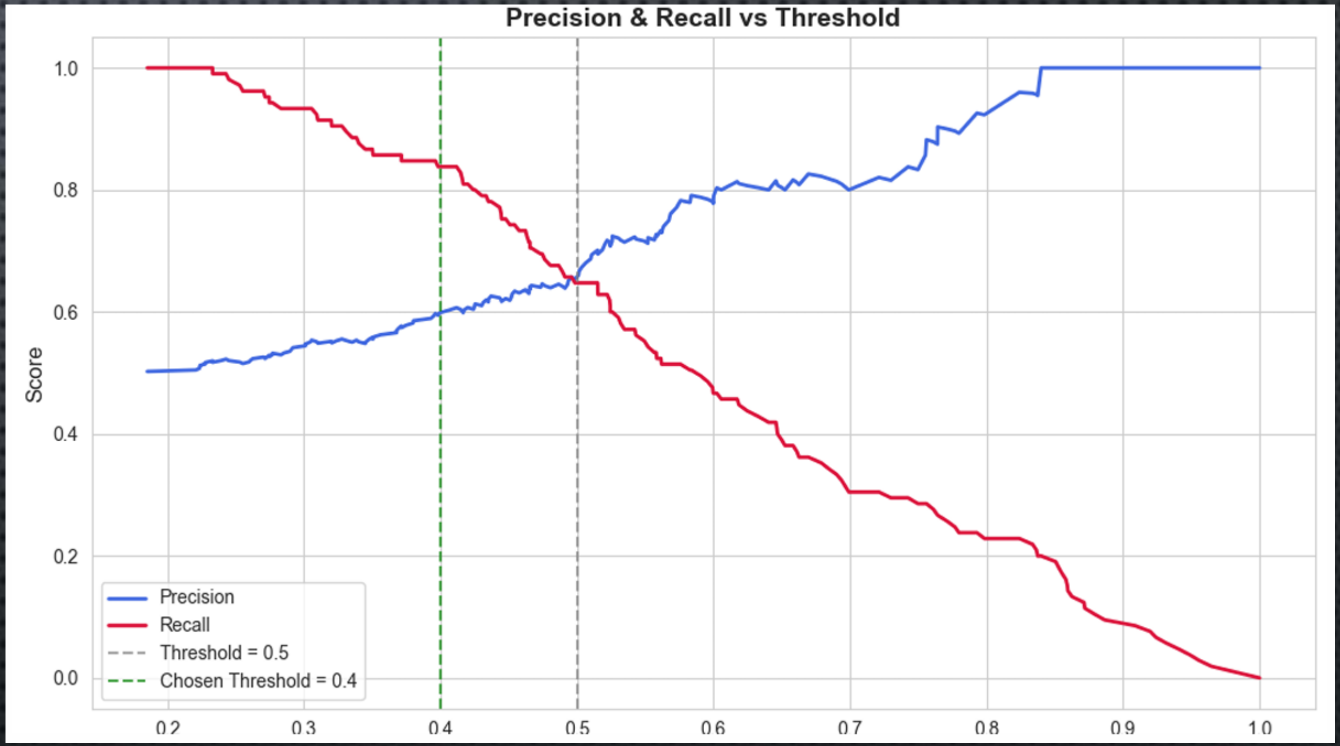
Our dataset is imbalanced, with more 'no' (4000) than 'yes' (521) labels. We'll address this by downsampling the majority class and later compare results using SMOTE.



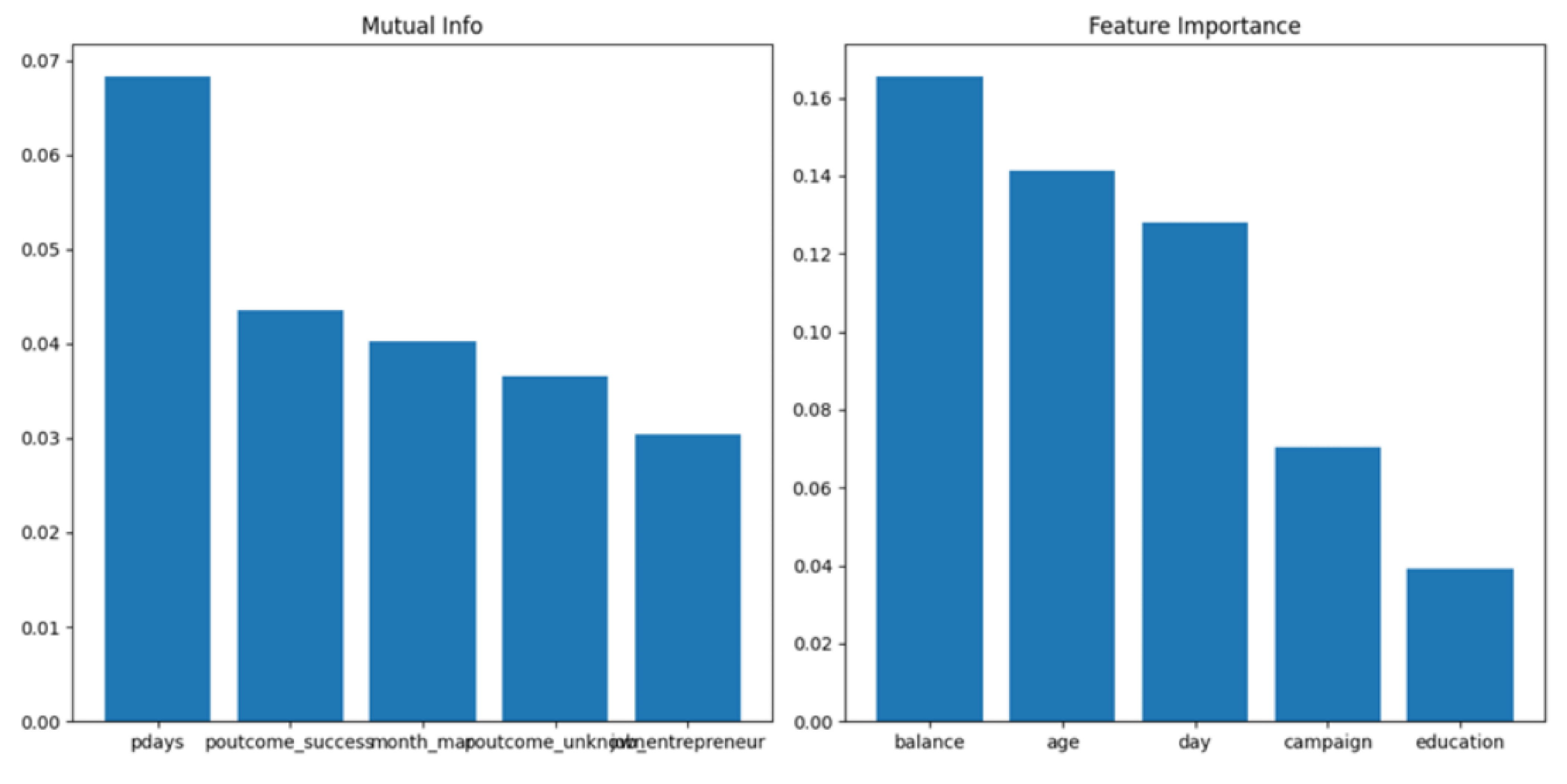
We aim to identify clients likely to subscribe in advance, given our limited human resources. To achieve this, we focus on maximizing recall, which helps reduce false negative, missing actual subscribers. This approach improves efficiency and prevents potential revenue loss.



Determining the optimal trade-off point depends on understanding the impact of false positives and false negatives. However, conducting a comprehensive cost-sensitive evaluation falls beyond the current scope of this study.

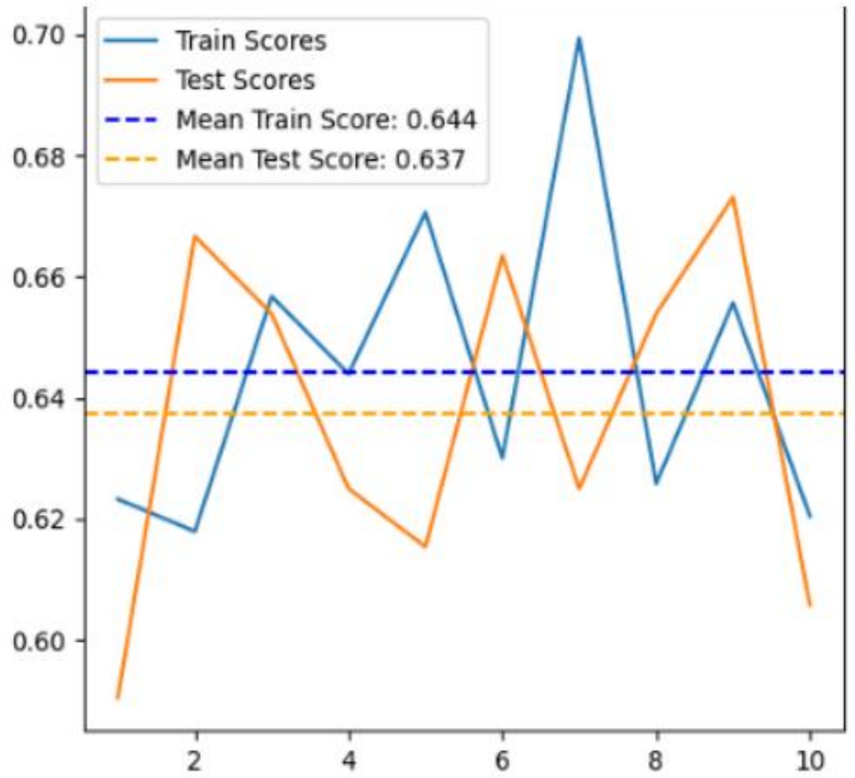


FEATURE RELEVANCE: MUTUAL INFORMATİON VS. MODEL IMPORTANCE

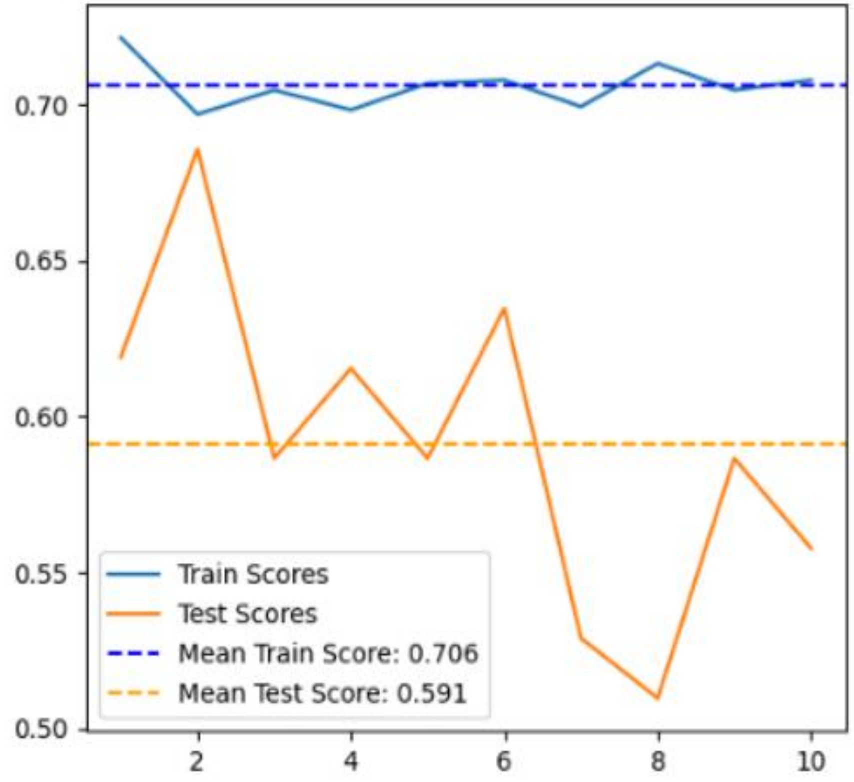


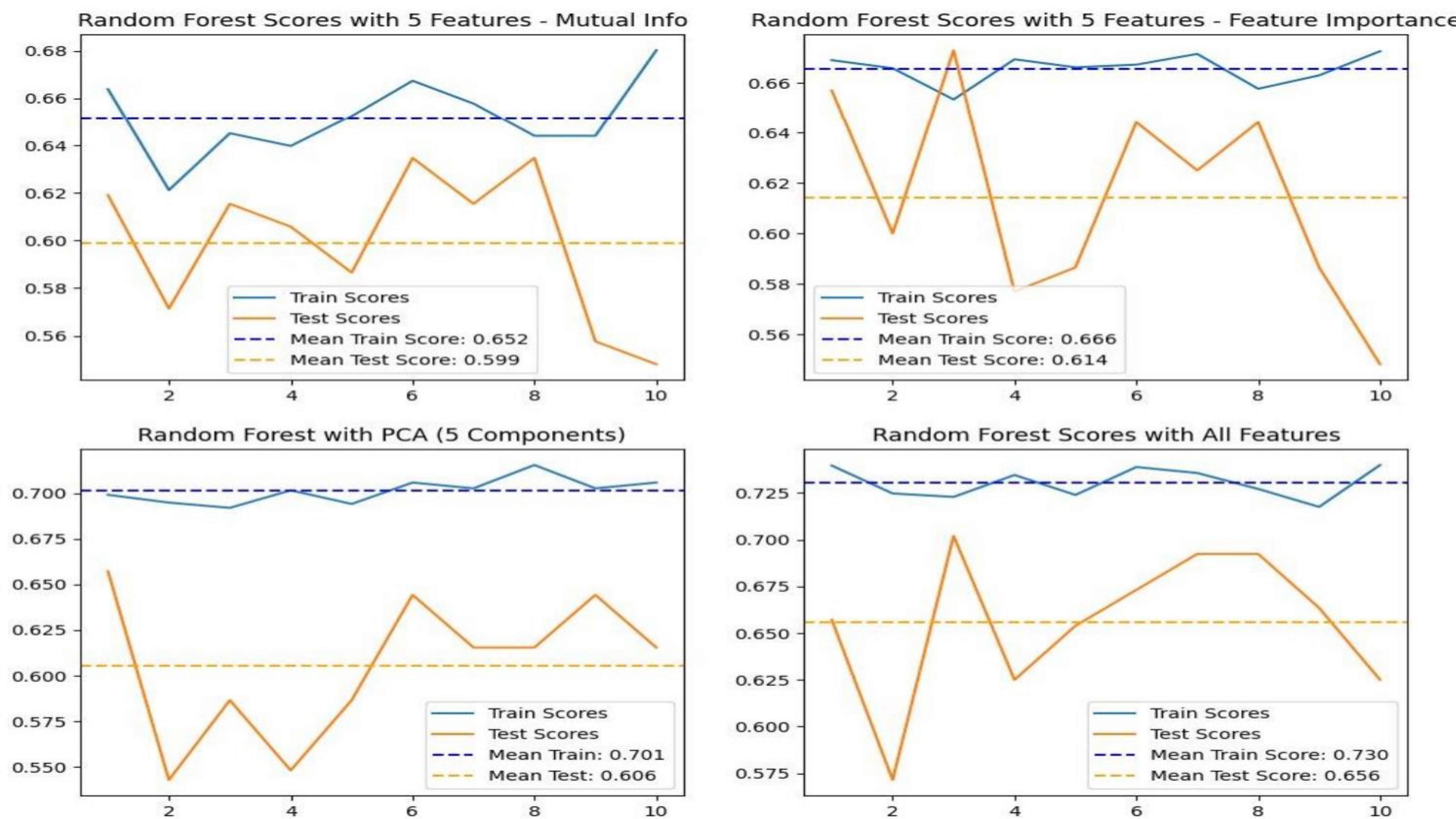
RANDOM FOREST PERFORMANCE COMPARİSON: MUTUAL INFO VS. FEATURE IMPORTANCE

Random Forest Scores with 5 features – Mutual Info



Random Forest Scores with 5 features – feature Importance




Random Forest Scores with 5 Features - Mutual Info
Train Scores
Test Scores
Mean Train Score: 0.652
Mean Test Score: 0.599
0.68
0.66
0.64
0.62
0.60
0.58
0.56
2
4
6
8
10
Random Forest Scores with 5 Features - Feature Importance
Train Scores
Test Scores
Mean Train Score: 0.666
Mean Test Score: 0.614
0.66
0.64
0.62
0.60
0.58
0.56
2
4
6
8
10
Random Forest with PCA (5 Components)
Train Scores
Test Scores
Mean Train: 0.701
Mean Test: 0.606
0.700
0.675
0.650
0.625
0.600
0.575
0.550
2
4
6
8
10
Random Forest Scores with All Features
Train Scores
Test Scores
Mean Train Score: 0.730
Mean Test Score: 0.656
0.725
0.700
0.675
0.650
0.625
0.600
0.575
2
4
6
8
10