In [1]:
```
from pycaret.classification import *
```

In [2]:
```
import pandas as pd
import numpy as np
```

In [4]:
```
df=pd.read_csv('mst2-data.csv')
df.head()
```

Out[4]:

	f0	f1	f2	f3	f4	f5	f6	f7	f8	f9	f10	f11	f12	f13	f14	f15	f16	f17	f18	f19
0	0	0	1	1	1	1	1	0	1	0	4.60	3.30	6.30	8.48	5.44	8.83	5.64	0.87	7.66	2.84
1	1	1	1	1	0	0	1	0	0	1	9.51	9.52	7.11	6.35	4.90	2.11	7.98	4.57	1.17	5.65
2	0	1	1	0	1	0	0	0	1	0	1.62	8.32	7.05	2.68	7.00	3.79	0.48	2.93	9.62	1.89
3	1	0	0	1	0	0	0	0	1	0	10.45	4.29	4.22	8.59	3.03	2.06	4.16	1.54	5.50	3.10
4	0	1	0	1	1	0	0	1	0	1	4.23	7.06	2.57	6.13	9.85	4.01	1.48	6.84	3.44	8.53

In [5]:

```
clf = setup(data = df, target = 'f0', session_id=123)
```

	Description	Value
0	session_id	123
1	Target	f0
2	Target Type	Binary
3	Label Encoded	0: 0, 1: 1
4	Original Data	(2999, 26)
5	Missing Values	True
6	Numeric Features	16
7	Categorical Features	9
8	Ordinal Features	False
9	High Cardinality Features	False
10	High Cardinality Method	None
11	Transformed Train Set	(2099, 25)
12	Transformed Test Set	(900, 25)
13	Shuffle Train-Test	True
14	Stratify Train-Test	False
15	Fold Generator	StratifiedKFold
16	Fold Number	10
17	CPU Jobs	-1
18	Use GPU	False
19	Log Experiment	False
20	Experiment Name	clf-default-name
21	USI	6dff
22	Imputation Type	simple
23	Iterative Imputation Iteration	None
24	Numeric Imputer	mean
25	Iterative Imputation Numeric Model	None
26	Categorical Imputer	constant
27	Iterative Imputation Categorical Model	None
28	Unknown Categoricals Handling	least_frequent
29	Normalize	False
30	Normalize Method	None
31	Transformation	False
32	Transformation Method	None
33	PCA	False

	Description	Value
34	PCA Method	None
35	PCA Components	None
36	Ignore Low Variance	False
37	Combine Rare Levels	False
38	Rare Level Threshold	None
39	Numeric Binning	False
40	Remove Outliers	False
41	Outliers Threshold	None
42	Remove Multicollinearity	False
43	Multicollinearity Threshold	None
44	Clustering	False
45	Clustering Iteration	None
46	Polynomial Features	False
47	Polynomial Degree	None
48	Trignometry Features	False
49	Polynomial Threshold	None
50	Group Features	False
51	Feature Selection	False
52	Features Selection Threshold	None
53	Feature Interaction	False
54	Feature Ratio	False
55	Interaction Threshold	None
56	Fix Imbalance	False
57	Fix Imbalance Method	SMOTE

In [6]:
```
df.replace('', np.NaN)
df.dropna(inplace = True)
```

In [8]:
```
df=df.drop_duplicates()
```

Since more than one models are showing equal accuracy and other metrics, TT (sec) is being compared to select best model.

Ans1

In [11]:
```
clf = setup(data = df, target = 'f0', session_id=123, data_split_shuffle=False)
```

	Description	Value
0	session_id	123
1	Target	f0
2	Target Type	Binary
3	Label Encoded	0: 0, 1: 1
4	Original Data	(1996, 26)
5	Missing Values	False
6	Numeric Features	16
7	Categorical Features	9
8	Ordinal Features	False
9	High Cardinality Features	False
10	High Cardinality Method	None
11	Transformed Train Set	(1397, 25)
12	Transformed Test Set	(599, 25)
13	Shuffle Train-Test	False
14	Stratify Train-Test	False
15	Fold Generator	StratifiedKFold
16	Fold Number	10
17	CPU Jobs	-1
18	Use GPU	False
19	Log Experiment	False
20	Experiment Name	clf-default-name
21	USI	a099
22	Imputation Type	simple
23	Iterative Imputation Iteration	None
24	Numeric Imputer	mean
25	Iterative Imputation Numeric Model	None
26	Categorical Imputer	constant
27	Iterative Imputation Categorical Model	None
28	Unknown Categoricals Handling	least_frequent
29	Normalize	False
30	Normalize Method	None
31	Transformation	False
32	Transformation Method	None
33	PCA	False

	Description	Value
34	PCA Method	None
35	PCA Components	None
36	Ignore Low Variance	False
37	Combine Rare Levels	False
38	Rare Level Threshold	None
39	Numeric Binning	False
40	Remove Outliers	False
41	Outliers Threshold	None
42	Remove Multicollinearity	False
43	Multicollinearity Threshold	None
44	Clustering	False
45	Clustering Iteration	None
46	Polynomial Features	False
47	Polynomial Degree	None
48	Trignometry Features	False
49	Polynomial Threshold	None
50	Group Features	False
51	Feature Selection	False
52	Features Selection Threshold	None
53	Feature Interaction	False
54	Feature Ratio	False
55	Interaction Threshold	None
56	Fix Imbalance	False
57	Fix Imbalance Method	SMOTE

```
In [12]: best_model = compare_models(fold = 12)
```

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
dt	Decision Tree Classifier	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.0283
rf	Random Forest Classifier	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.2475
ada	Ada Boost Classifier	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.0242
gbc	Gradient Boosting Classifier	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.1508
xgboost	Extreme Gradient Boosting	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.1125
lightgbm	Light Gradient Boosting Machine	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.0592
catboost	CatBoost Classifier	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	5.5333
lr	Logistic Regression	0.9907	0.9996	0.9935	0.9896	0.9915	0.9812	0.9813	0.0492
et	Extra Trees Classifier	0.9843	0.9992	0.9909	0.9808	0.9858	0.9681	0.9683	0.2267
nb	Naive Bayes	0.9814	0.9989	0.9817	0.9849	0.9831	0.9624	0.9629	0.0200
ridge	Ridge Classifier	0.9749	0.0000	0.9673	0.9869	0.9769	0.9495	0.9500	0.0317
lda	Linear Discriminant Analysis	0.9749	0.9983	0.9673	0.9869	0.9769	0.9495	0.9500	0.0333
qda	Quadratic Discriminant Analysis	0.9592	0.9934	0.9661	0.9602	0.9628	0.9176	0.9183	0.0300
svm	SVM - Linear Kernel	0.9513	0.0000	0.9557	0.9593	0.9548	0.9020	0.9072	0.0267
knn	K Neighbors Classifier	0.8719	0.9387	0.8903	0.8787	0.8840	0.7409	0.7417	0.0642

Ans2

In [13]:
```
clf = setup(data = df, target = 'f0', session_id=123, data_split_shuffle=False, r
```

	Description	Value
0	session_id	123
1	Target	f0
2	Target Type	Binary
3	Label Encoded	0: 0, 1: 1
4	Original Data	(1996, 26)
5	Missing Values	False
6	Numeric Features	16
7	Categorical Features	9
8	Ordinal Features	False
9	High Cardinality Features	False
10	High Cardinality Method	None
11	Transformed Train Set	(1397, 25)
12	Transformed Test Set	(599, 25)
13	Shuffle Train-Test	False
14	Stratify Train-Test	False
15	Fold Generator	StratifiedKFold
16	Fold Number	10
17	CPU Jobs	-1
18	Use GPU	False
19	Log Experiment	False
20	Experiment Name	clf-default-name
21	USI	d12a
22	Imputation Type	simple
23	Iterative Imputation Iteration	None
24	Numeric Imputer	mean
25	Iterative Imputation Numeric Model	None
26	Categorical Imputer	constant
27	Iterative Imputation Categorical Model	None
28	Unknown Categoricals Handling	least_frequent
29	Normalize	True
30	Normalize Method	zscore
31	Transformation	False
32	Transformation Method	None
33	PCA	False

	Description	Value
34	PCA Method	None
35	PCA Components	None
36	Ignore Low Variance	False
37	Combine Rare Levels	False
38	Rare Level Threshold	None
39	Numeric Binning	False
40	Remove Outliers	False
41	Outliers Threshold	None
42	Remove Multicollinearity	False
43	Multicollinearity Threshold	None
44	Clustering	False
45	Clustering Iteration	None
46	Polynomial Features	False
47	Polynomial Degree	None
48	Trignometry Features	False
49	Polynomial Threshold	None
50	Group Features	False
51	Feature Selection	False
52	Features Selection Threshold	None
53	Feature Interaction	False
54	Feature Ratio	False
55	Interaction Threshold	None
56	Fix Imbalance	False
57	Fix Imbalance Method	SMOTE

```
In [14]: best_model = compare_models(fold = 12)
```

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
dt	Decision Tree Classifier	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.0308
rf	Random Forest Classifier	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.2383
ada	Ada Boost Classifier	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.0317
gbc	Gradient Boosting Classifier	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.1708
xgboost	Extreme Gradient Boosting	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.1242
catboost	CatBoost Classifier	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	5.4117
lightgbm	Light Gradient Boosting Machine	0.9964	1.0000	0.9974	0.9962	0.9967	0.9928	0.9929	0.0808
et	Extra Trees Classifier	0.9893	0.9996	0.9935	0.9873	0.9903	0.9783	0.9786	0.2258
lr	Logistic Regression	0.9857	0.9992	0.9883	0.9857	0.9870	0.9711	0.9712	0.0308
svm	SVM - Linear Kernel	0.9850	0.0000	0.9869	0.9858	0.9863	0.9697	0.9699	0.0308
nb	Naive Bayes	0.9814	0.9989	0.9817	0.9849	0.9831	0.9624	0.9629	0.0217
ridge	Ridge Classifier	0.9749	0.0000	0.9673	0.9869	0.9769	0.9495	0.9500	0.0192
lda	Linear Discriminant Analysis	0.9749	0.9983	0.9673	0.9869	0.9769	0.9495	0.9500	0.0342
qda	Quadratic Discriminant Analysis	0.9592	0.9934	0.9661	0.9602	0.9628	0.9176	0.9183	0.0233
knn	K Neighbors Classifier	0.8661	0.9267	0.8785	0.8785	0.8778	0.7298	0.7311	0.0683

Ans3

In [15]:
```
clf = setup(data=df, target='f0',session_id=123, normalize=True, normalize_methoc
```

	Description	Value
0	session_id	123
1	Target	f0
2	Target Type	Binary
3	Label Encoded	0: 0, 1: 1
4	Original Data	(1996, 26)
5	Missing Values	False
6	Numeric Features	16
7	Categorical Features	9
8	Ordinal Features	False
9	High Cardinality Features	False
10	High Cardinality Method	None
11	Transformed Train Set	(1397, 25)
12	Transformed Test Set	(599, 25)
13	Shuffle Train-Test	False
14	Stratify Train-Test	False
15	Fold Generator	StratifiedKFold
16	Fold Number	10
17	CPU Jobs	-1
18	Use GPU	False
19	Log Experiment	False
20	Experiment Name	clf-default-name
21	USI	4b5f
22	Imputation Type	simple
23	Iterative Imputation Iteration	None
24	Numeric Imputer	mean
25	Iterative Imputation Numeric Model	None
26	Categorical Imputer	constant
27	Iterative Imputation Categorical Model	None
28	Unknown Categoricals Handling	least_frequent
29	Normalize	True
30	Normalize Method	zscore
31	Transformation	True
32	Transformation Method	yeo-johnson

	Description	Value
33	PCA	False
34	PCA Method	None
35	PCA Components	None
36	Ignore Low Variance	False
37	Combine Rare Levels	False
38	Rare Level Threshold	None
39	Numeric Binning	False
40	Remove Outliers	False
41	Outliers Threshold	None
42	Remove Multicollinearity	False
43	Multicollinearity Threshold	None
44	Clustering	False
45	Clustering Iteration	None
46	Polynomial Features	False
47	Polynomial Degree	None
48	Trignometry Features	False
49	Polynomial Threshold	None
50	Group Features	False
51	Feature Selection	False
52	Features Selection Threshold	None
53	Feature Interaction	False
54	Feature Ratio	False
55	Interaction Threshold	None
56	Fix Imbalance	False
57	Fix Imbalance Method	SMOTE

```
In [16]: best_model = compare_models(fold = 12)
```

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
dt	Decision Tree Classifier	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.0308
rf	Random Forest Classifier	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.2208
ada	Ada Boost Classifier	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.0300
gbc	Gradient Boosting Classifier	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.1750
xgboost	Extreme Gradient Boosting	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.1158
catboost	CatBoost Classifier	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	5.3917
lightgbm	Light Gradient Boosting Machine	0.9964	1.0000	0.9974	0.9962	0.9967	0.9928	0.9929	0.0858
lr	Logistic Regression	0.9857	0.9993	0.9883	0.9857	0.9870	0.9711	0.9712	0.0300
et	Extra Trees Classifier	0.9857	0.9992	0.9909	0.9833	0.9870	0.9711	0.9713	0.2233
nb	Naive Bayes	0.9814	0.9989	0.9817	0.9849	0.9831	0.9624	0.9629	0.0225
svm	SVM - Linear Kernel	0.9778	0.0000	0.9778	0.9819	0.9797	0.9552	0.9556	0.0250
ridge	Ridge Classifier	0.9764	0.0000	0.9713	0.9856	0.9783	0.9524	0.9527	0.0308
lda	Linear Discriminant Analysis	0.9757	0.9983	0.9700	0.9856	0.9776	0.9509	0.9513	0.0350
qda	Quadratic Discriminant Analysis	0.9585	0.9935	0.9648	0.9602	0.9622	0.9162	0.9168	0.0258
knn	K Neighbors Classifier	0.8654	0.9276	0.8811	0.8759	0.8776	0.7281	0.7298	0.0542

Ans4

In [18]:
```
clf = setup(data=df, target='f0',session_id=123, normalize=True, normalize_metho
```

	Description	Value
0	session_id	123
1	Target	f0
2	Target Type	Binary
3	Label Encoded	0: 0, 1: 1
4	Original Data	(1996, 26)
5	Missing Values	False
6	Numeric Features	16
7	Categorical Features	9
8	Ordinal Features	False
9	High Cardinality Features	False
10	High Cardinality Method	None
11	Transformed Train Set	(1257, 25)
12	Transformed Test Set	(599, 25)
13	Shuffle Train-Test	False
14	Stratify Train-Test	False
15	Fold Generator	StratifiedKFold
16	Fold Number	10
17	CPU Jobs	-1
18	Use GPU	False
19	Log Experiment	False
20	Experiment Name	clf-default-name
21	USI	04aa
22	Imputation Type	simple
23	Iterative Imputation Iteration	None
24	Numeric Imputer	mean
25	Iterative Imputation Numeric Model	None
26	Categorical Imputer	constant
27	Iterative Imputation Categorical Model	None
28	Unknown Categoricals Handling	least_frequent
29	Normalize	True
30	Normalize Method	zscore
31	Transformation	False
32	Transformation Method	None

	Description	Value
33	PCA	False
34	PCA Method	None
35	PCA Components	None
36	Ignore Low Variance	False
37	Combine Rare Levels	False
38	Rare Level Threshold	None
39	Numeric Binning	False
40	Remove Outliers	True
41	Outliers Threshold	0.100000
42	Remove Multicollinearity	False
43	Multicollinearity Threshold	None
44	Clustering	False
45	Clustering Iteration	None
46	Polynomial Features	False
47	Polynomial Degree	None
48	Trignometry Features	False
49	Polynomial Threshold	None
50	Group Features	False
51	Feature Selection	False
52	Features Selection Threshold	None
53	Feature Interaction	False
54	Feature Ratio	False
55	Interaction Threshold	None
56	Fix Imbalance	False
57	Fix Imbalance Method	SMOTE

```
In [19]: best_model = compare_models(fold = 12)
```

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
dt	Decision Tree Classifier	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.0308
rf	Random Forest Classifier	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.2225
ada	Ada Boost Classifier	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.0242
gbc	Gradient Boosting Classifier	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.1525
xgboost	Extreme Gradient Boosting	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.1092
catboost	CatBoost Classifier	0.9984	1.0000	0.9985	0.9986	0.9985	0.9968	0.9968	5.1392
lightgbm	Light Gradient Boosting Machine	0.9976	1.0000	1.0000	0.9957	0.9978	0.9952	0.9952	0.0742
lr	Logistic Regression	0.9841	0.9989	0.9883	0.9827	0.9855	0.9679	0.9680	0.0308
svm	SVM - Linear Kernel	0.9833	0.0000	0.9810	0.9885	0.9846	0.9664	0.9668	0.0258
et	Extra Trees Classifier	0.9817	0.9991	0.9868	0.9800	0.9833	0.9631	0.9634	0.2150
nb	Naive Bayes	0.9793	0.9983	0.9810	0.9817	0.9812	0.9583	0.9587	0.0258
ridge	Ridge Classifier	0.9714	0.0000	0.9679	0.9796	0.9736	0.9423	0.9426	0.0233
lda	Linear Discriminant Analysis	0.9714	0.9980	0.9679	0.9796	0.9736	0.9423	0.9426	0.0317
qda	Quadratic Discriminant Analysis	0.9586	0.9930	0.9650	0.9597	0.9622	0.9165	0.9169	0.0292
knn	K Neighbors Classifier	0.8631	0.9208	0.8715	0.8777	0.8737	0.7243	0.7259	0.0542

Ans5

In [20]:
```
clf = setup(data=df, target='f0',session_id=123, pca = True, pca_method='linear',
```

	Description	Value
0	session_id	123
1	Target	f0
2	Target Type	Binary
3	Label Encoded	0: 0, 1: 1
4	Original Data	(1996, 26)
5	Missing Values	False
6	Numeric Features	16
7	Categorical Features	9
8	Ordinal Features	False
9	High Cardinality Features	False
10	High Cardinality Method	None
11	Transformed Train Set	(1187, 16)
12	Transformed Test Set	(599, 16)
13	Shuffle Train-Test	False
14	Stratify Train-Test	False
15	Fold Generator	StratifiedKFold
16	Fold Number	10
17	CPU Jobs	-1
18	Use GPU	False
19	Log Experiment	False
20	Experiment Name	clf-default-name
21	USI	1bae
22	Imputation Type	simple
23	Iterative Imputation Iteration	None
24	Numeric Imputer	mean
25	Iterative Imputation Numeric Model	None
26	Categorical Imputer	constant
27	Iterative Imputation Categorical Model	None
28	Unknown Categoricals Handling	least_frequent
29	Normalize	False
30	Normalize Method	None
31	Transformation	False
32	Transformation Method	None

	Description	Value
33	PCA	True
34	PCA Method	linear
35	PCA Components	0.990000
36	Ignore Low Variance	False
37	Combine Rare Levels	False
38	Rare Level Threshold	None
39	Numeric Binning	False
40	Remove Outliers	True
41	Outliers Threshold	0.150000
42	Remove Multicollinearity	False
43	Multicollinearity Threshold	None
44	Clustering	False
45	Clustering Iteration	None
46	Polynomial Features	False
47	Polynomial Degree	None
48	Trignometry Features	False
49	Polynomial Threshold	None
50	Group Features	False
51	Feature Selection	False
52	Features Selection Threshold	None
53	Feature Interaction	False
54	Feature Ratio	False
55	Interaction Threshold	None
56	Fix Imbalance	False
57	Fix Imbalance Method	SMOTE

```
In [21]: best_model = compare_models(fold = 12)
```

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
lr	Logistic Regression	0.9865	0.9995	0.9922	0.9833	0.9877	0.9728	0.9730	0.0292
svm	SVM - Linear Kernel	0.9823	0.0000	0.9799	0.9877	0.9836	0.9644	0.9647	0.0225
ridge	Ridge Classifier	0.9688	0.0000	0.9550	0.9873	0.9708	0.9374	0.9382	0.0258
lda	Linear Discriminant Analysis	0.9688	0.9978	0.9550	0.9873	0.9708	0.9374	0.9382	0.0242
qda	Quadratic Discriminant Analysis	0.9579	0.9953	0.9737	0.9509	0.9617	0.9149	0.9162	0.0208
catboost	CatBoost Classifier	0.9486	0.9908	0.9596	0.9473	0.9530	0.8963	0.8974	4.8167
et	Extra Trees Classifier	0.9452	0.9906	0.9627	0.9391	0.9502	0.8893	0.8910	0.2300
nb	Naive Bayes	0.9435	0.9917	0.9658	0.9334	0.9490	0.8858	0.8874	0.0258
xgboost	Extreme Gradient Boosting	0.9317	0.9847	0.9457	0.9302	0.9376	0.8623	0.8631	0.2383
lightgbm	Light Gradient Boosting Machine	0.9292	0.9850	0.9396	0.9313	0.9351	0.8572	0.8579	0.1142
rf	Random Forest Classifier	0.9182	0.9794	0.9301	0.9204	0.9250	0.8351	0.8357	0.2550
ada	Ada Boost Classifier	0.9141	0.9743	0.9209	0.9225	0.9207	0.8268	0.8289	0.0917
gbc	Gradient Boosting Classifier	0.9141	0.9768	0.9410	0.9051	0.9225	0.8261	0.8275	0.2200
knn	K Neighbors Classifier	0.8643	0.9325	0.8868	0.8675	0.8765	0.7261	0.7273	0.0550
dt	Decision Tree Classifier	0.7919	0.7906	0.8045	0.8138	0.8079	0.5806	0.5827	0.0267

```
In [ ]:
```