EVALUATION OF FEATURE SELECTION METHODS

DATA EXPLORATION AND VISUALIZATION – EXERCISE 1

Marta Szuwarska

June 2025

GOAL

Evaluation of feature selection methods:

- Mutual Information Importance (with known number of relevant features)
- Boruta Algorithm
- LASSO Regularization

and dimensionality reduction techniques:

- PCA
- Multidimensional scaling
- tSNE

PIPELINE

1. Generate dataset configurations -> list of configurations
2. For each configuration:
    1. Generate data -> X, y, list of relevant features, list of all features
    2. For each feature selection method:
        - Select features -> List of selected features
    3. Evaluate feature selection methods with metrics
    4. For each dimensionality reduction technique:
        - Transform matrix X -> Projected X
    5. For each classifier (Random Forest or SVM):
        - For each transformation method (no transformation, feature selection, relevant features, dimensionality reduction):
            - Split dataset into train and test datasets -> X_train, X_test, y_train, y_test
            - Train model and predict -> model, predictions
            - Calculate metrics -> accuracy, precision, recall, F1 score
3. Collect and aggregate (calculate average) metrics across all runs

DATA GENERATION

INFORMATIVE FEATURES

Generated by make_classification(), directly contribute to determining the class label.

IRRELEVANT FEATURES

Random features sampled independently from normal distribution N(0, 1).

REDUNDANT FEATURES

Generated by make_classification(), linear combinations of informative features.

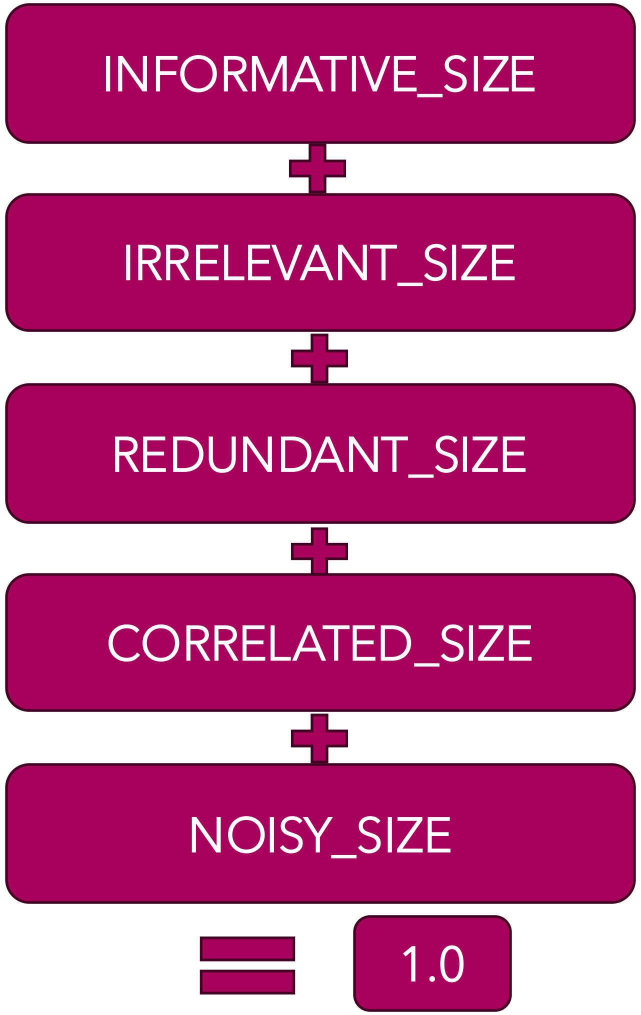
CORRELATED FEATURES

Manually created features as copies of informative features with added noise ~N(0, 0.1).

NOISY FEATURES

Manually created features as copies of informative features with added strong noise ~N(0, 1).

DATASET CONFIGURATION


INFORMATIVE_SIZE

IRRELEVANT_SIZE

REDUNDANT_SIZE

CORRELATED_SIZE

NOISY_SIZE

1.0

Dataset configuration consists of the number of samples, number of features and proportions of categories of features that sum up to 1.

Those proportions were drawn from Dirichlet distribution Dir(1, 1, 1, 1, 1), while making sure there are always at least two informative features.

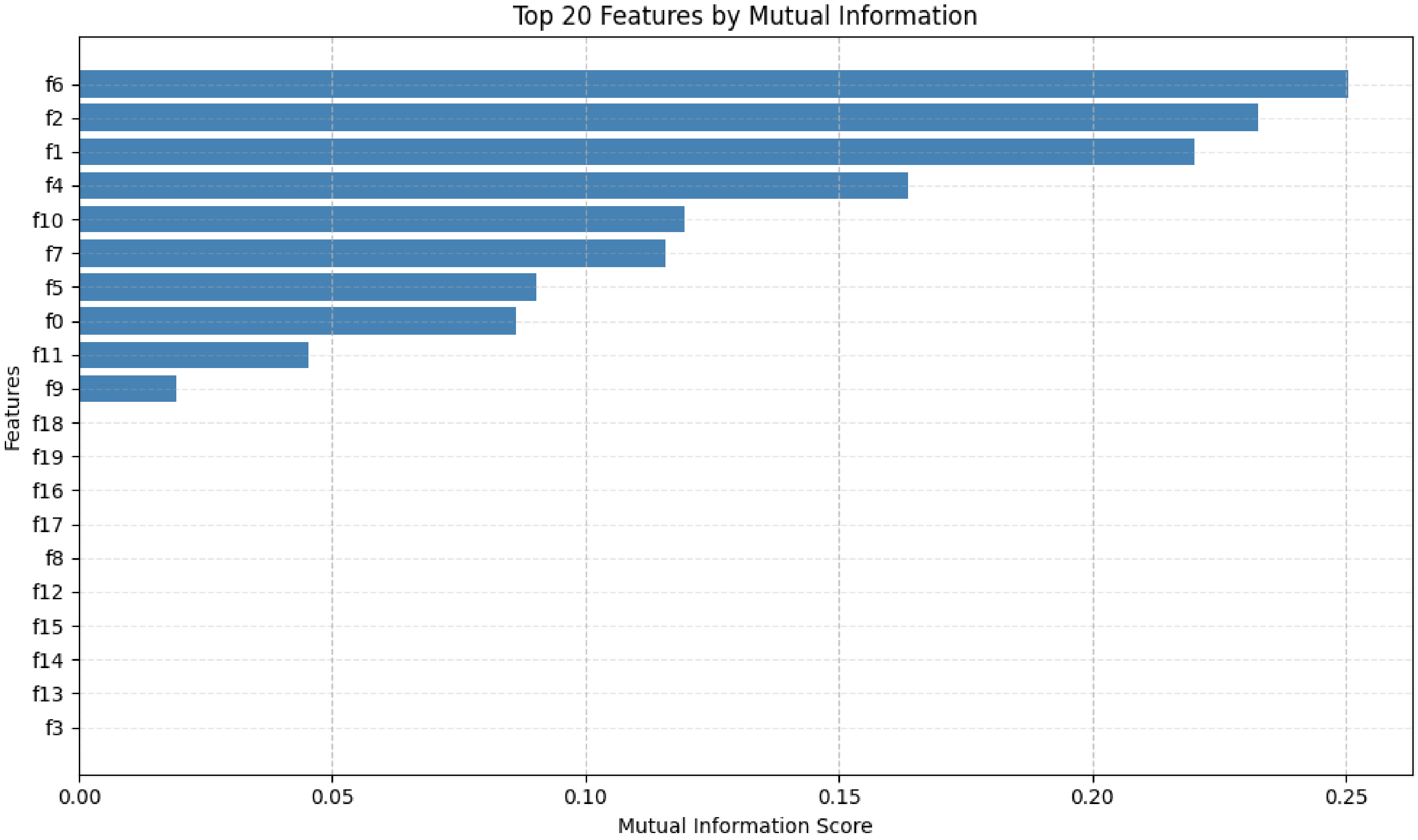
EXAMPLE DATASET

Parameter	Value
n_samples	1000
n_features	20
iInformative_size	0.3
irrelevant_size	0.4
redundant_size	0.1
correlated_size	0.1
noisy_size	0.1

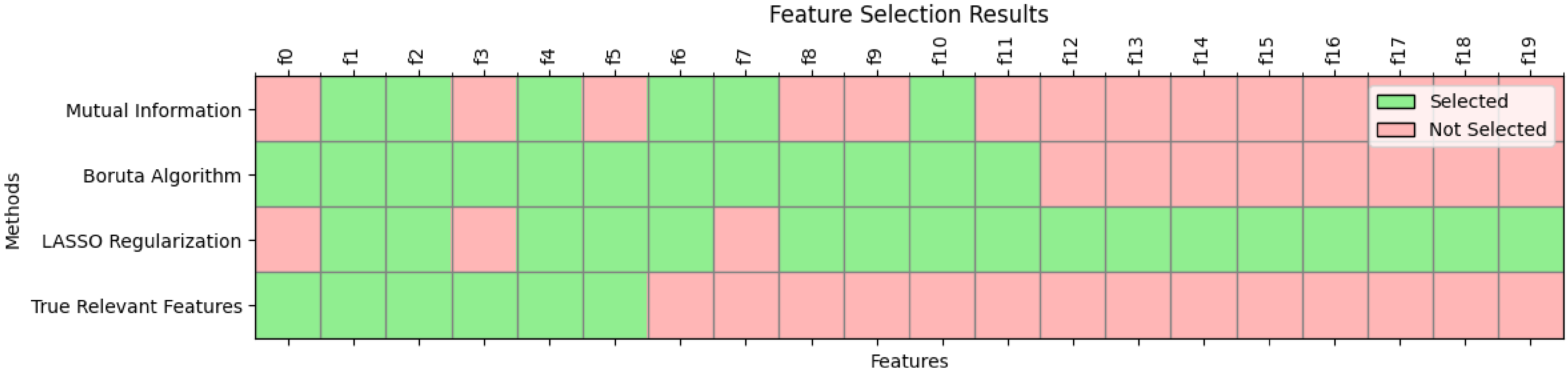
Relevant features:

f0, f1, f2, f3, f4, f5

EXAMPLE DATASET



EXAMPLE DATASET



Feature Selection Results

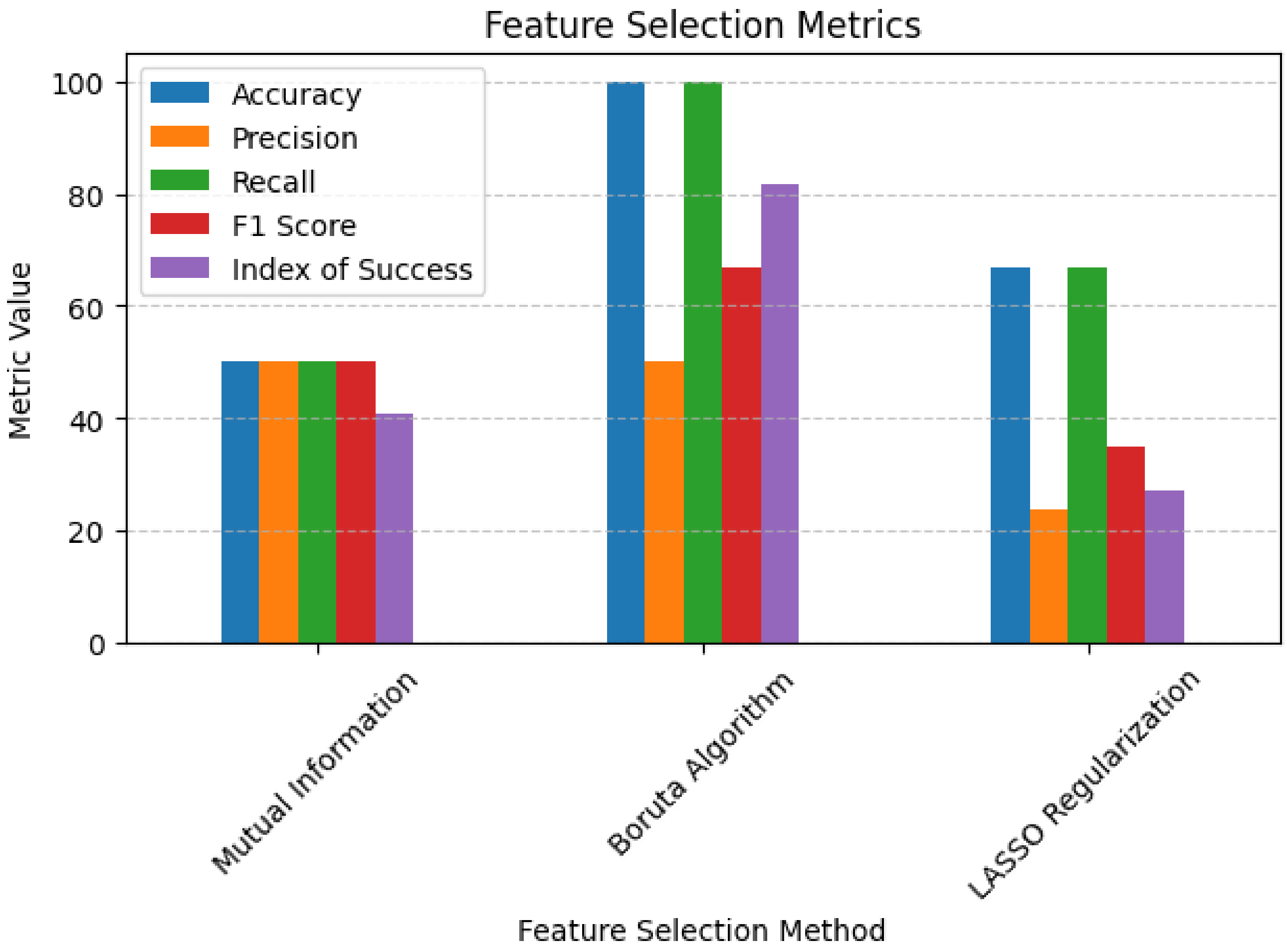
Methods	f0	f1	f2	f3	f4	f5	f6	f7	f8	f9	f10	f11	f12	f13	f14	f15	f16	f17	f18	f19
Mutual Information	🟥	🟩	🟩	🟥	🟩	🟥	🟩	🟩	🟥	🟥	🟩	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥
Boruta Algorithm	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥
LASSO Regularization	🟥	🟩	🟩	🟥	🟩	🟩	🟩	🟥	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
True Relevant Features	🟩	🟩	🟩	🟩	🟩	🟩	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥

Features

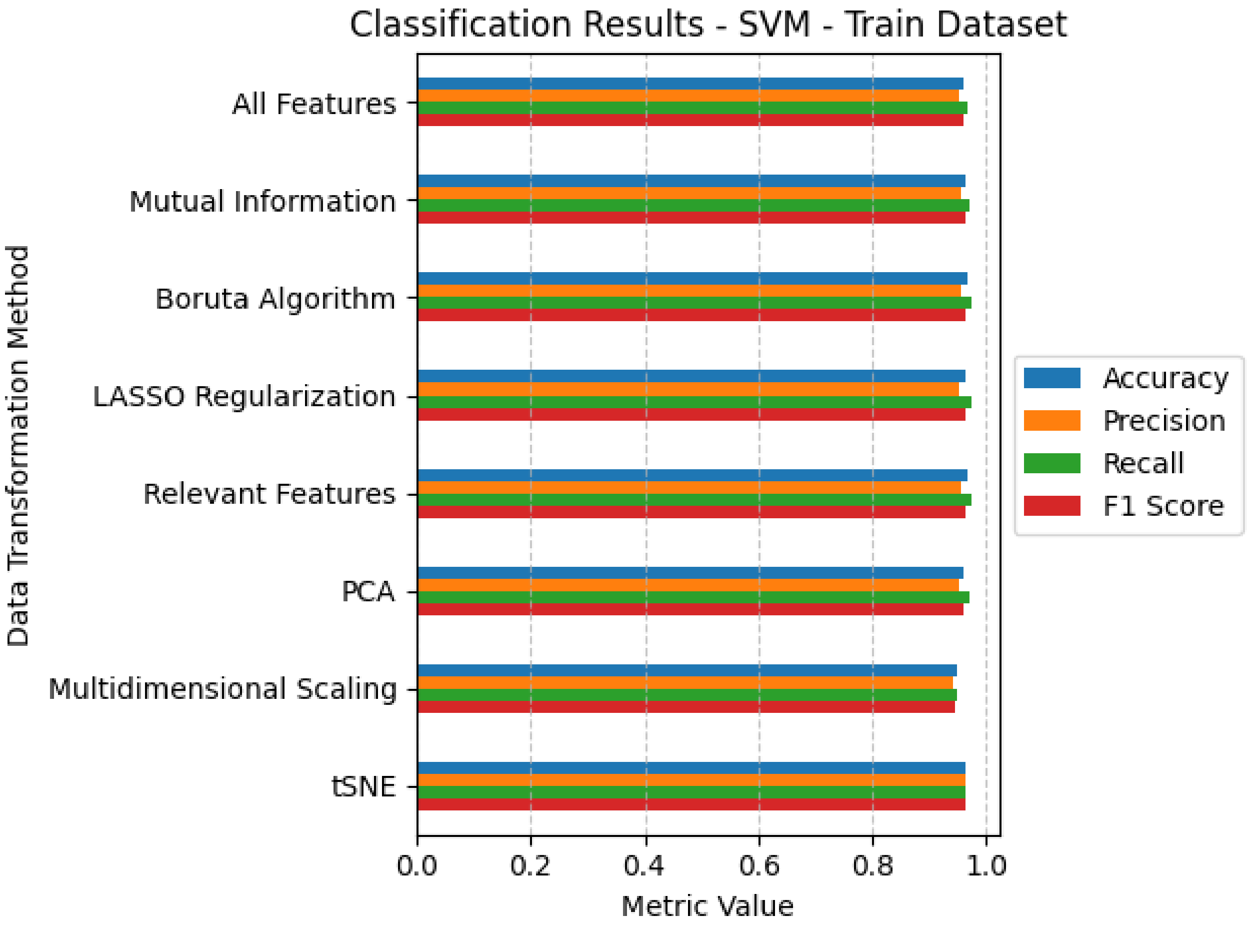
🟩 Selected

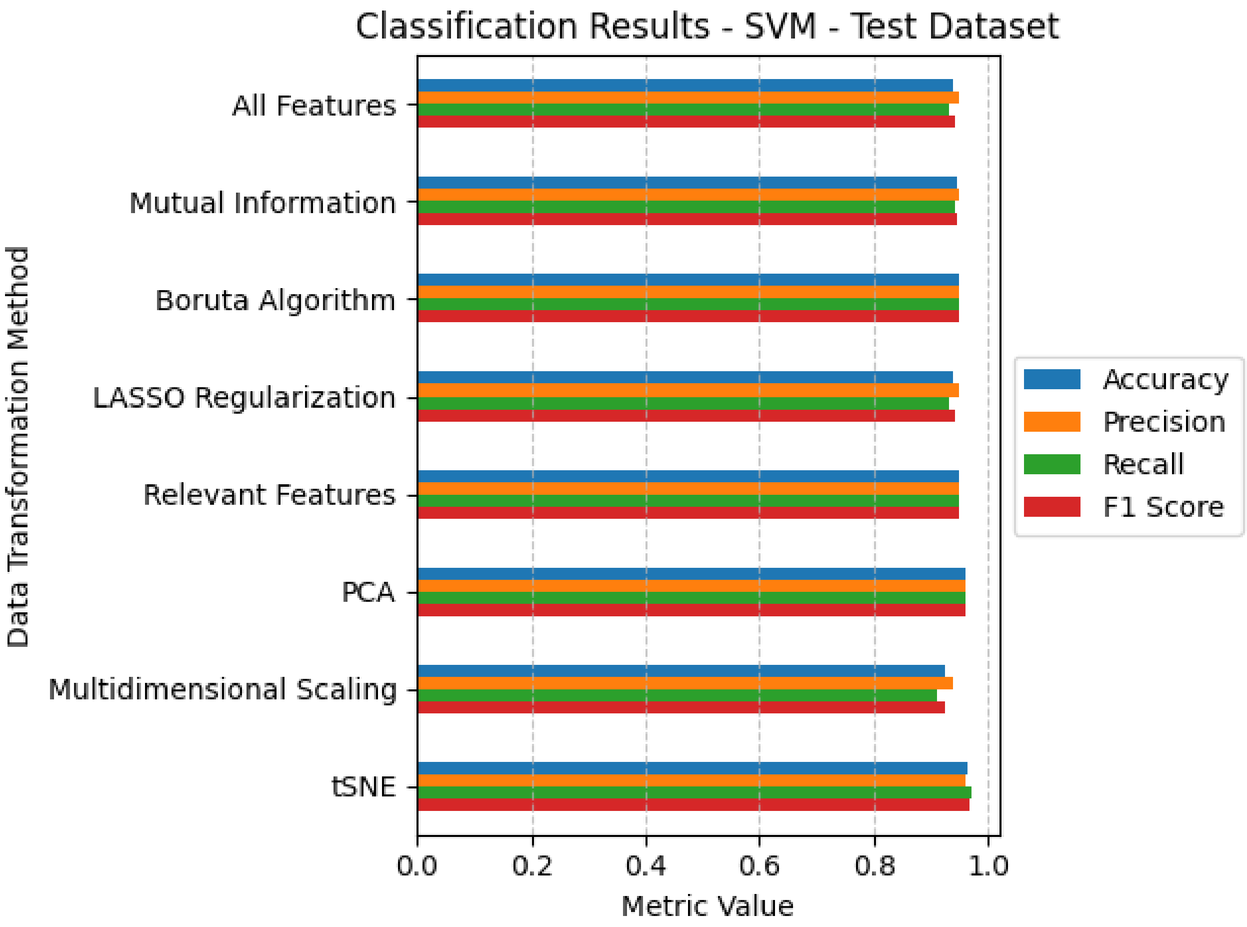
🟥 Not Selected

EXAMPLE DATASET

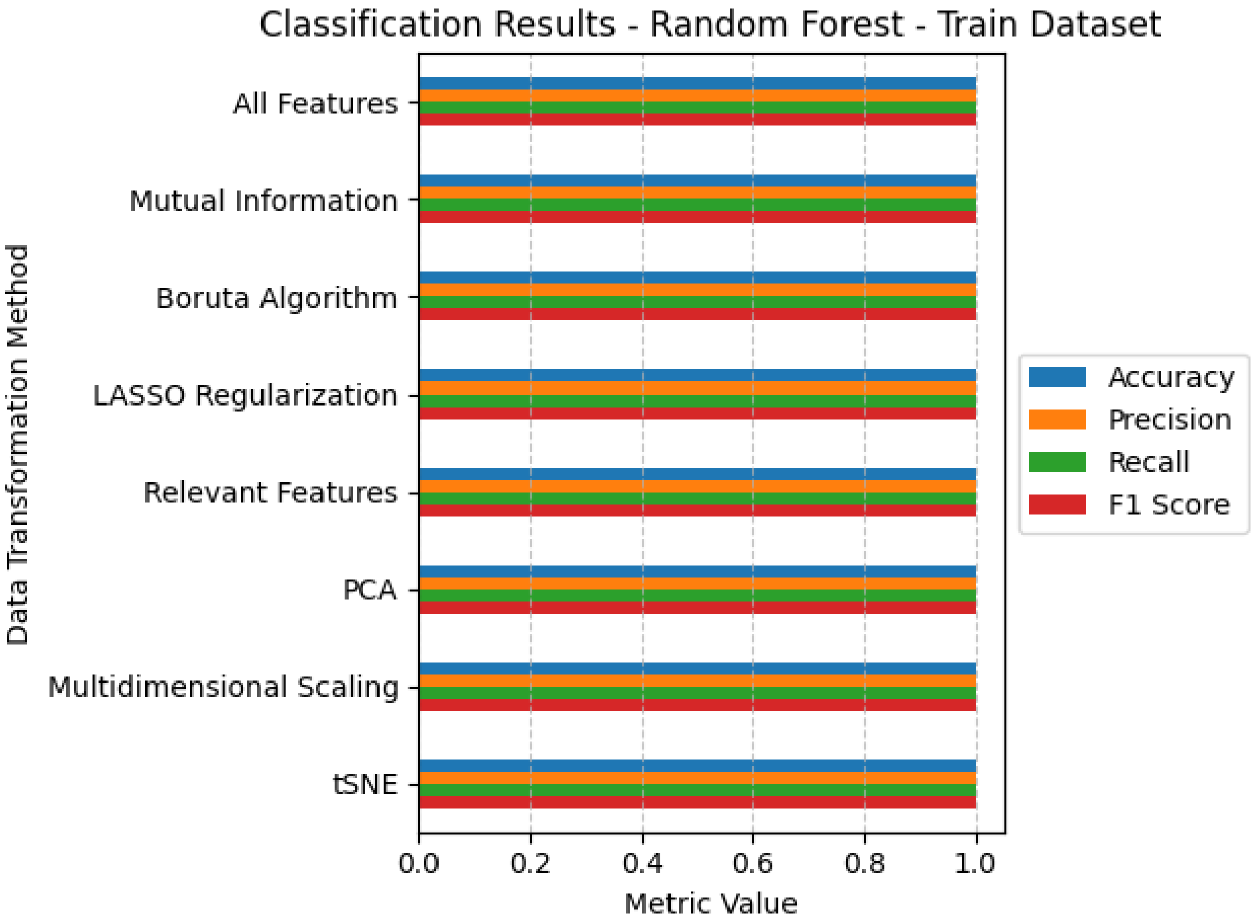


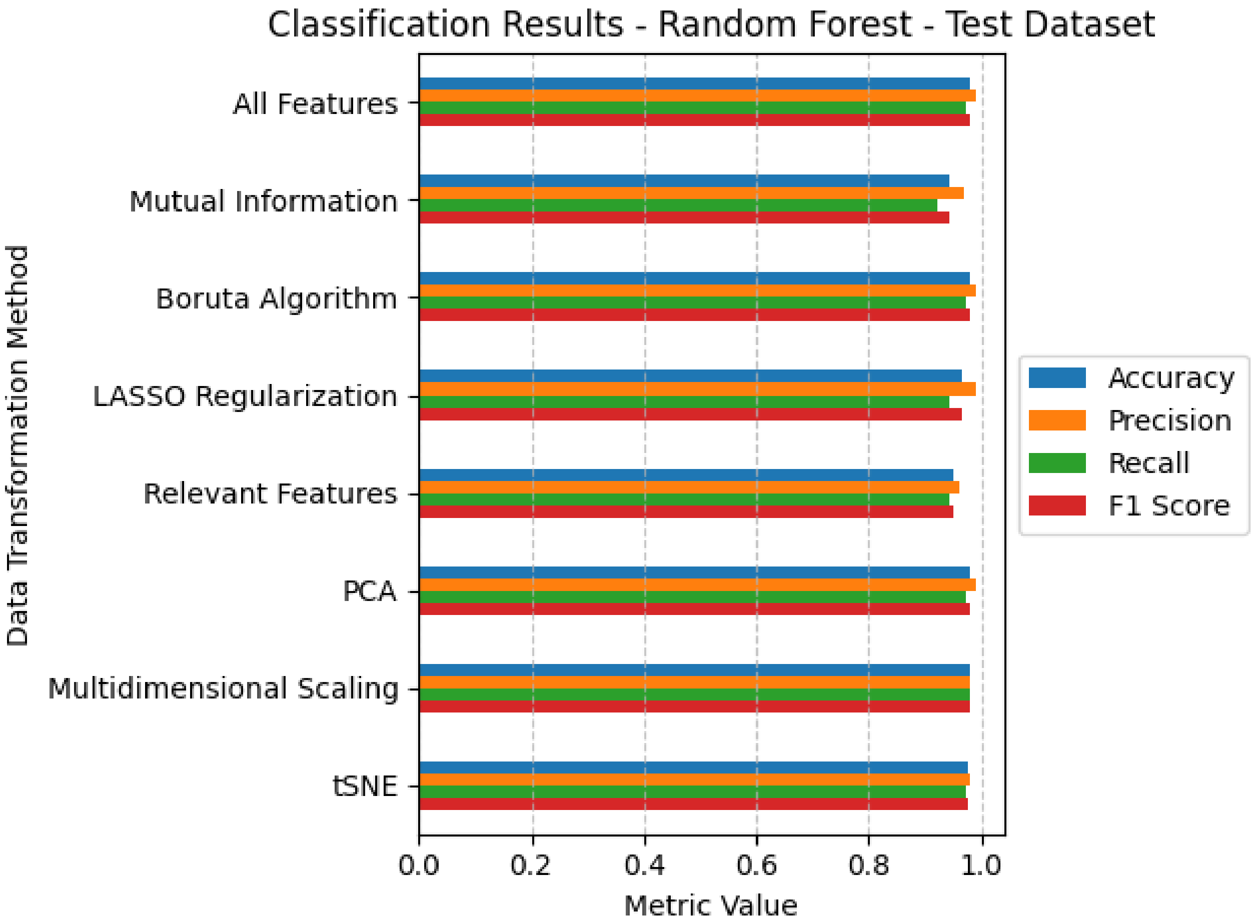
EXAMPLE DATASET





EXAMPLE DATASET

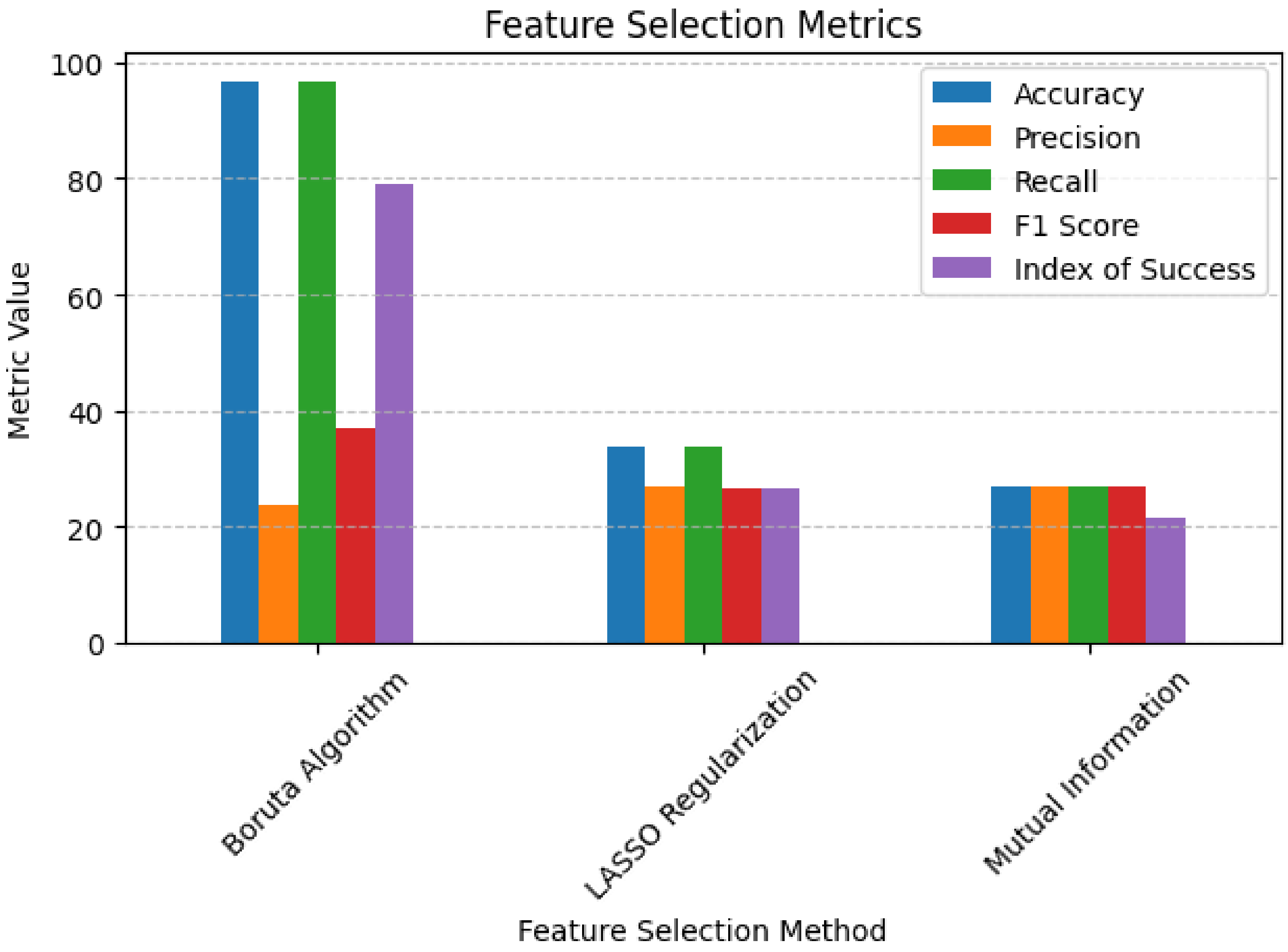




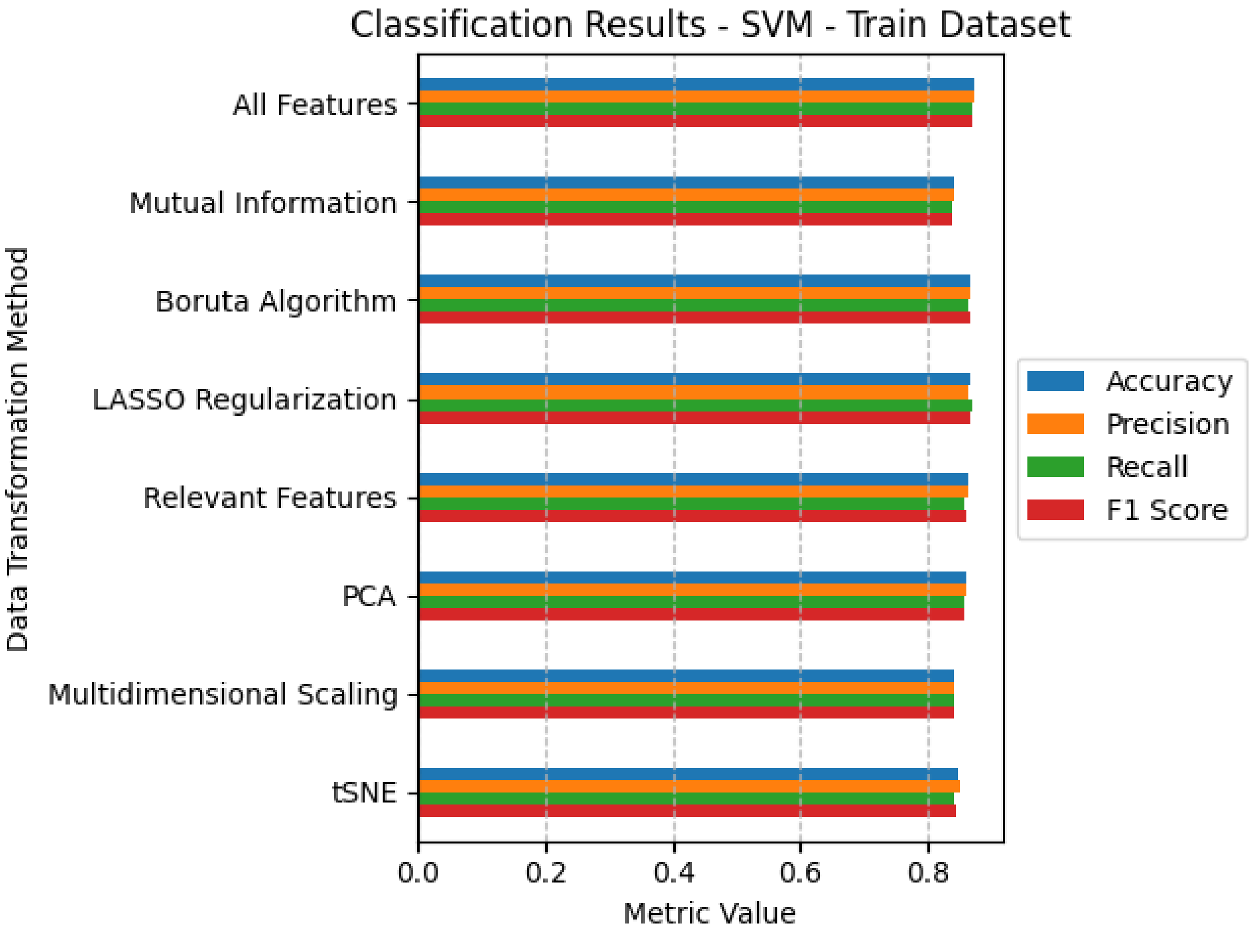
TESTED CONFIGURATIONS

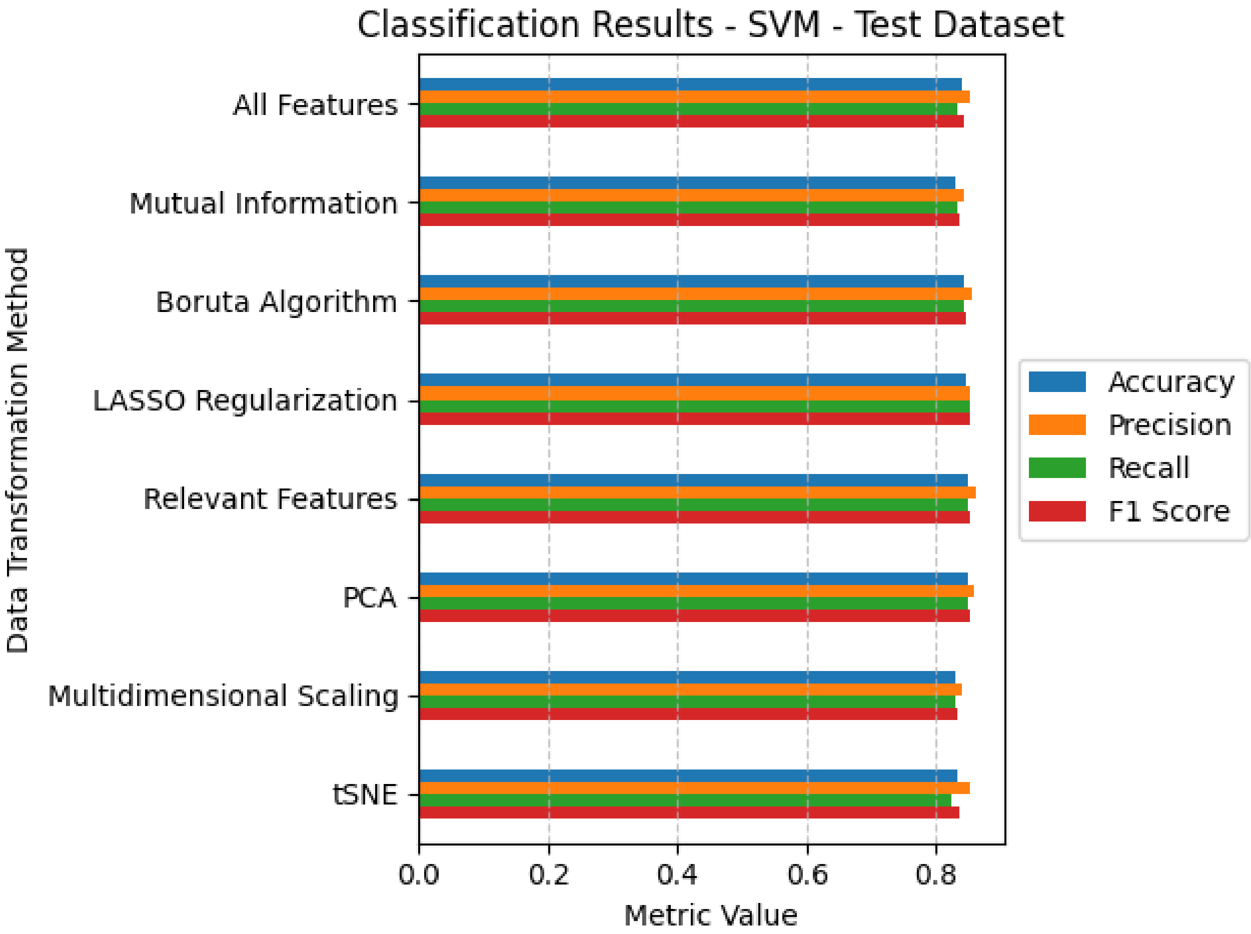
Number of samples	Number of features	Number of datasets
1000	20	10
1000	50	10
1000	100	10
500	50	10
3000	100	10

AVERAGED RESULTS

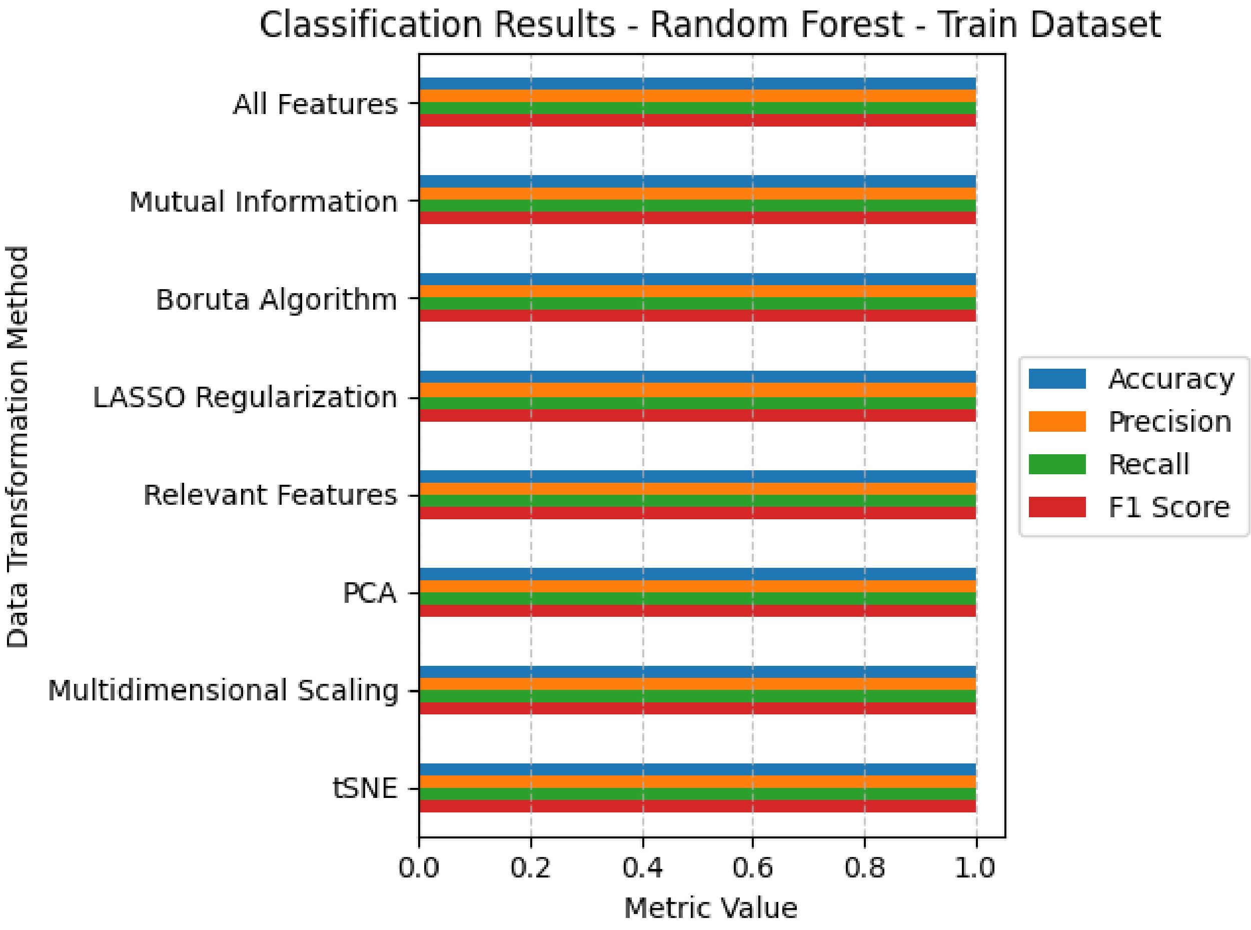


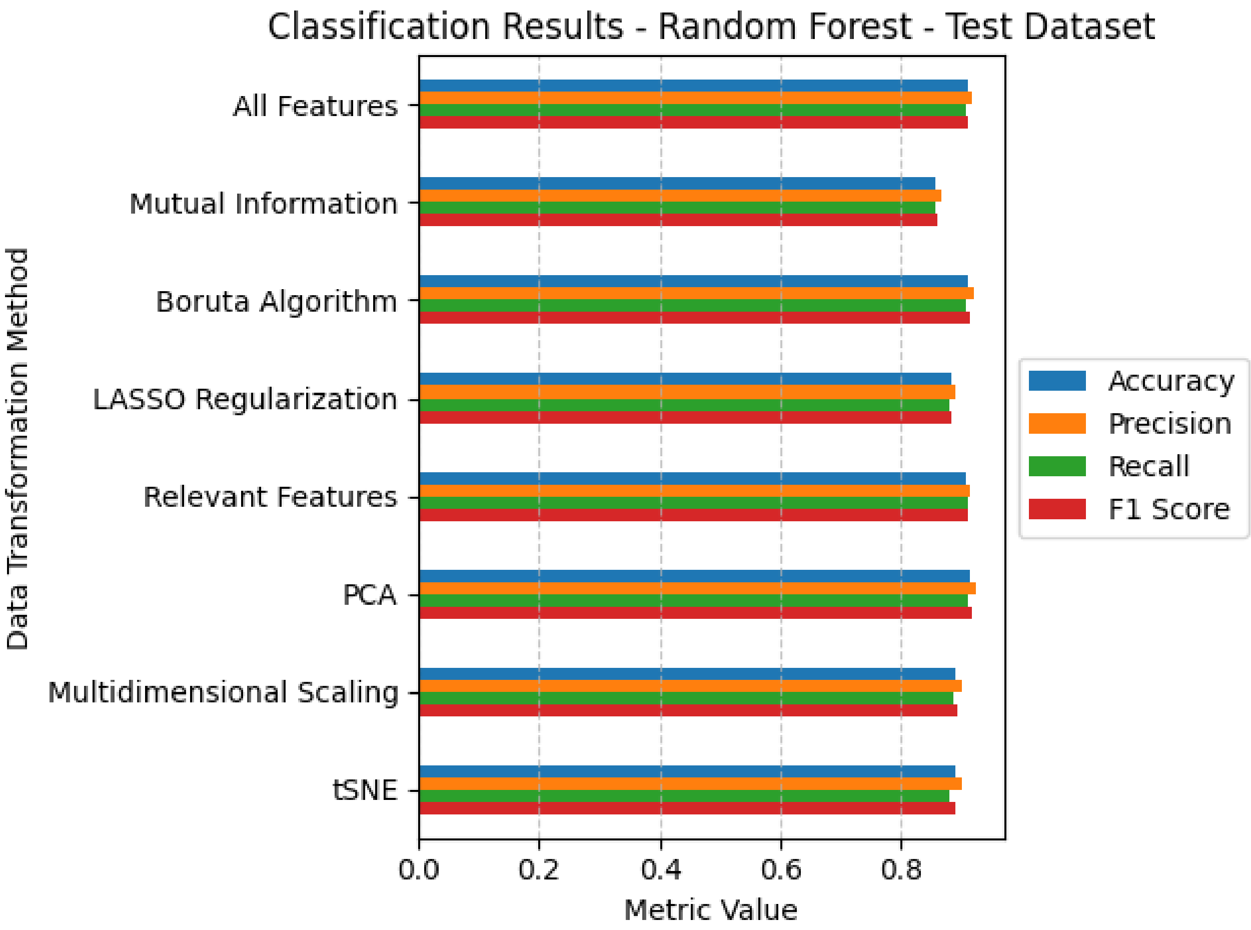
AVERAGED RESULTS





AVERAGED RESULTS





EVALUATION OF DATA CLUSTERING ALGORITHMS

DATA EXPLORATION AND VISUALIZATION – EXERCISE 2

Marta Szuwarska

June 2025

GOAL

Evaluation of data clustering algorithms:

- K-means
- Genie ($g \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$)
- Agglomerative Hierarchical Clustering (with single, average, complete or Ward linkage)
- DBSCAN ($\varepsilon \in \{0.01, 0.05, 0.1, 0.5, 1.0, 1.5\}$, minimum number of samples $\in \{2, 3, 5, 10\}$)

PIPELINE

1. Read datasets -> X, y_trues, k_values
2. For each dataset:
    I. For each k (number of clusters):
        - Fit and predict clusters with KMeans -> labels
        - For each gini threshold:
            - Fit and predict clusters with Genie -> labels
        - For each linkage:
            - Fit and predict clusters with Agglomerative Hierarchical Clustering -> labels

    II. For each **ε** :
        - For each min_samples value:
            - Fit and predict with DBSCAN -> labels

    III. Evaluate all predicted labels -> adjusted rand index, normalized mutual information, fowlkes mallows score, adjusted assymetric accuracy, silhouette score
3. Collect and aggregate (calculate average) metrics across all runs or selected runs

PARTITION EVALUATION

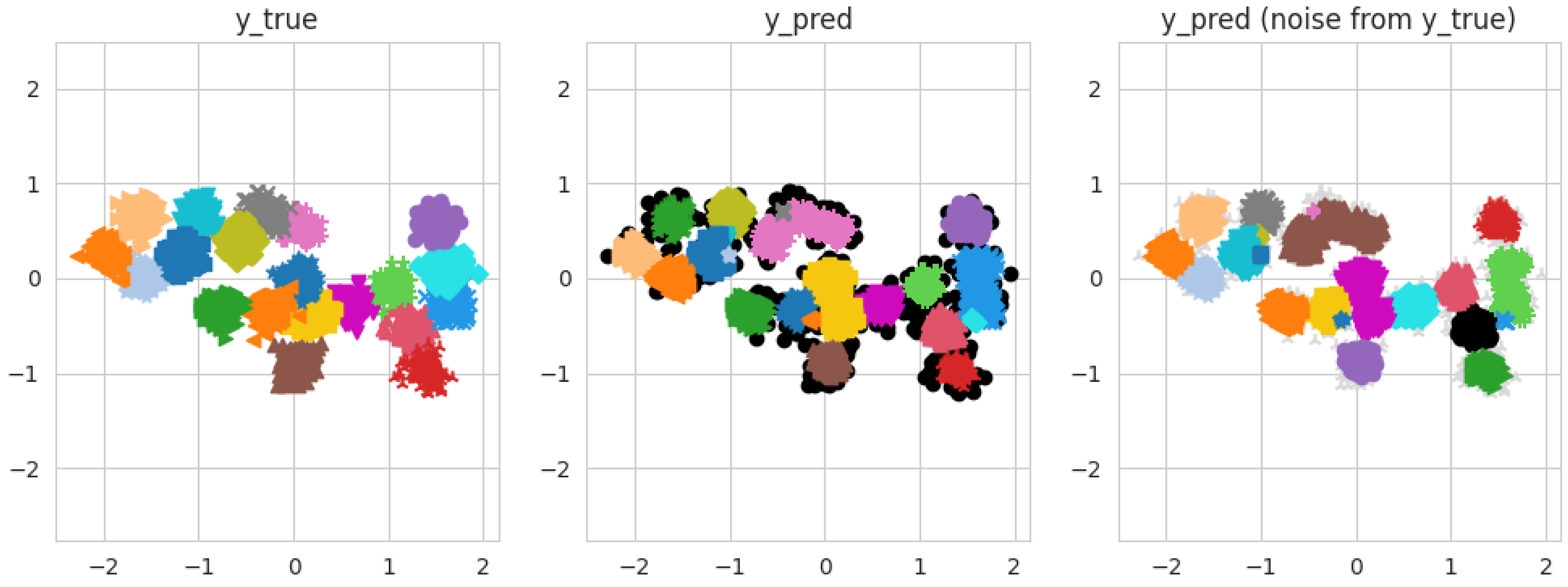
Different labels handling

Since in true labels 0 is reserved for noise, all predicted labels are shifted to 1 (keeping –1 as noise if it occurs).

Noise handling

For evaluation with metrics, all noise points (y_true == 0 or y_pred == -1) are excluded from calculations.

EXAMPLE DATASET - A1

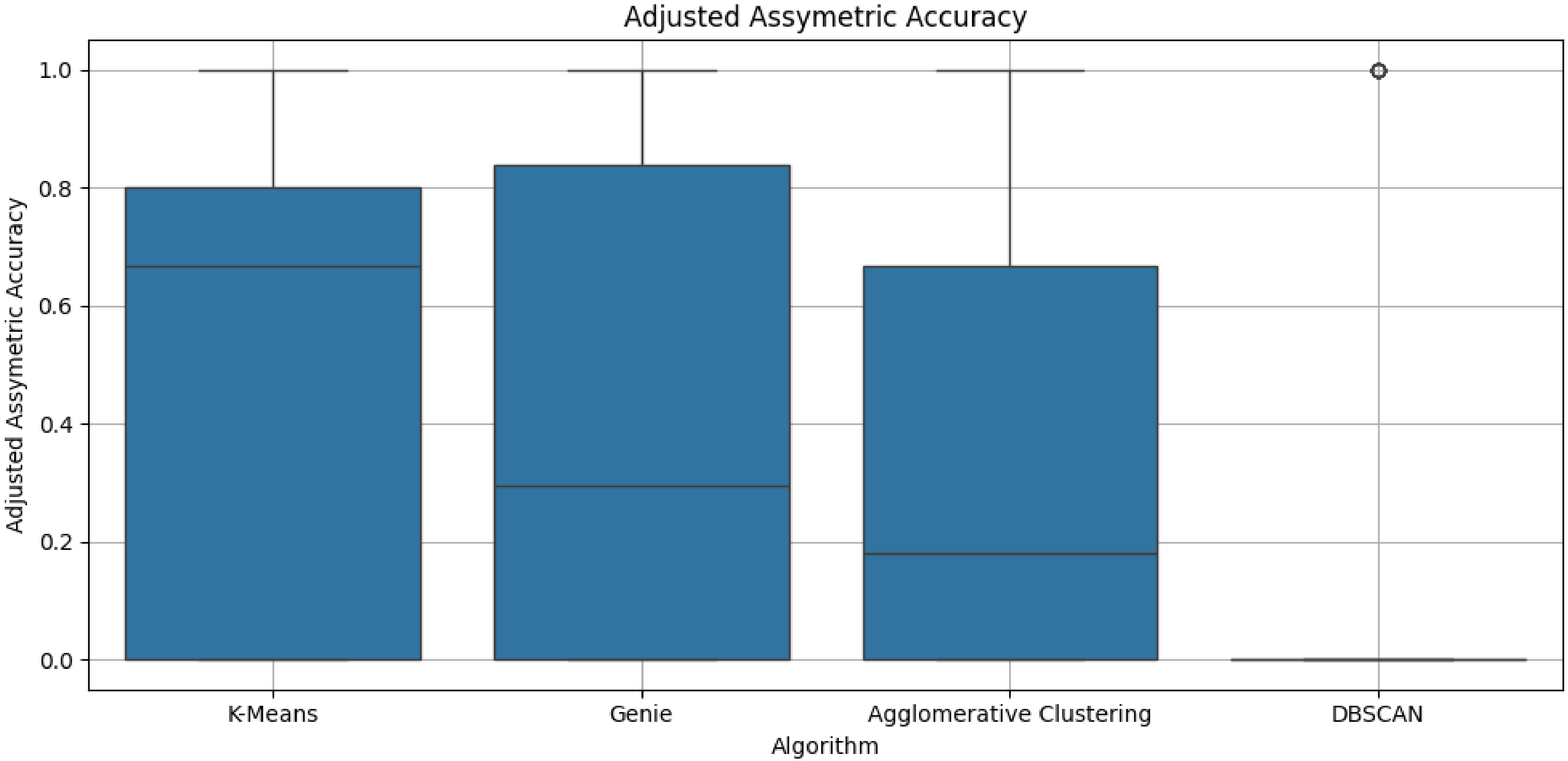


DBSCAN

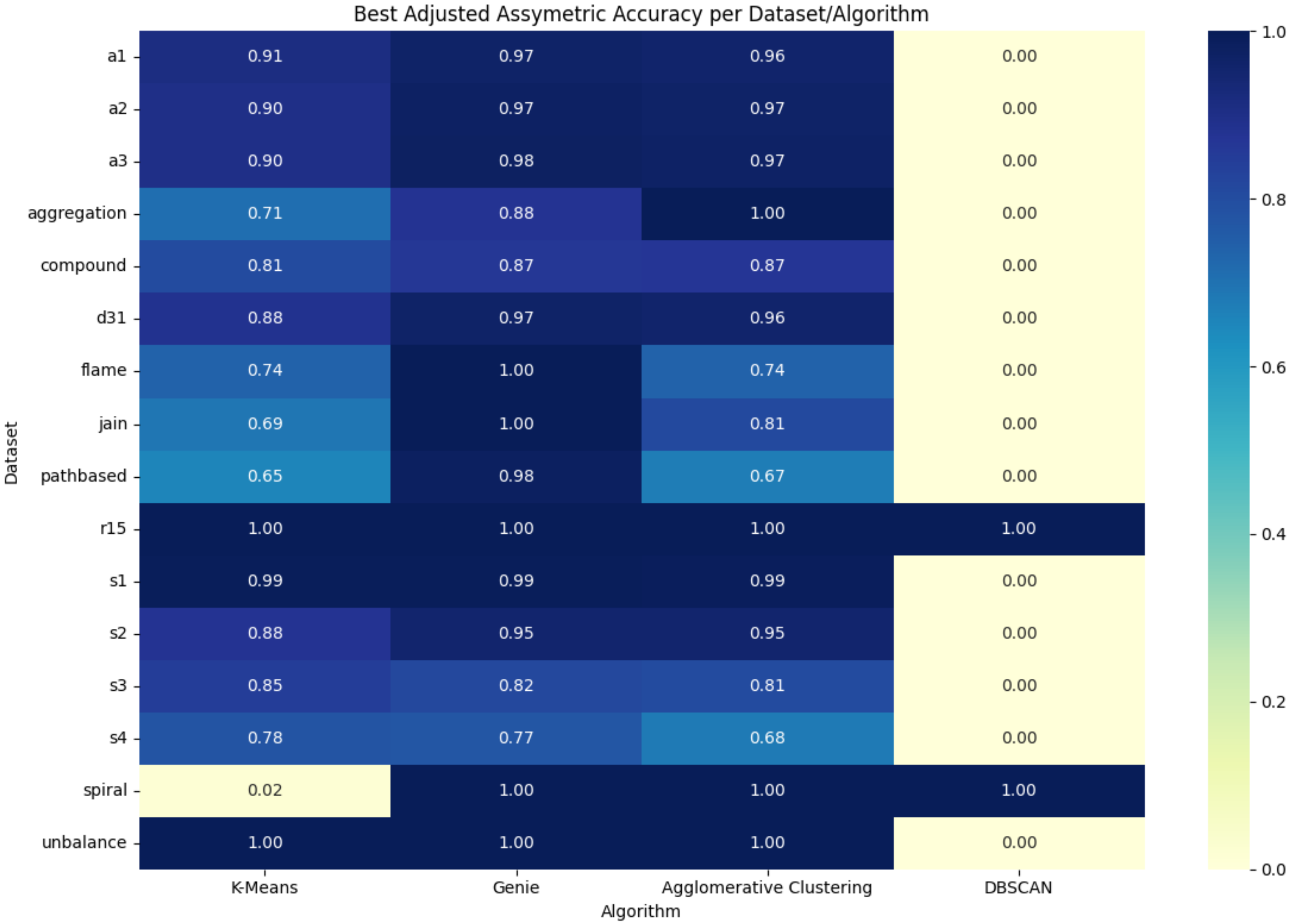
ε = 0.05

min_samples = 5

AGGREGATED RESULTS



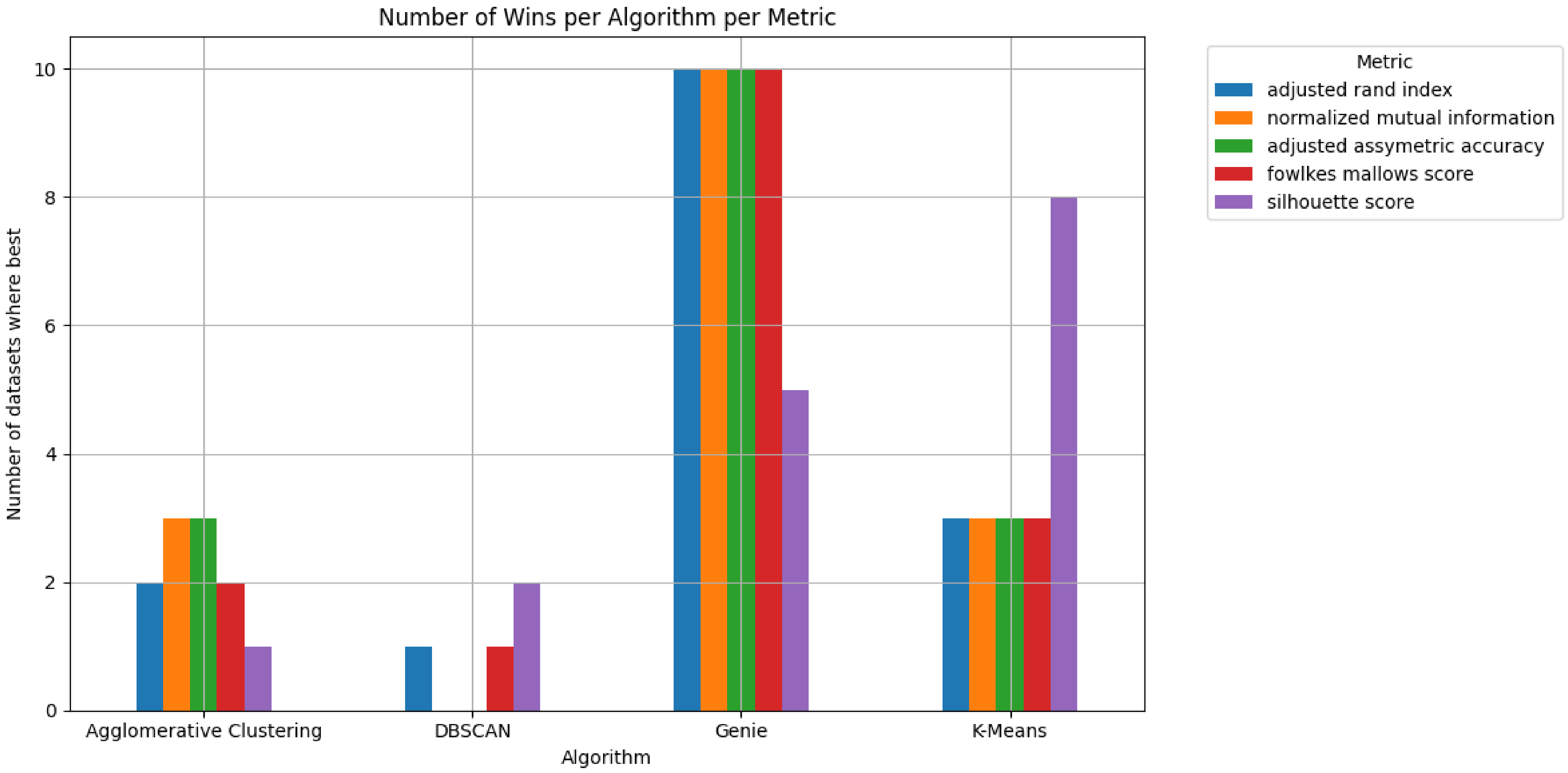
AGGREGATED RESULTS



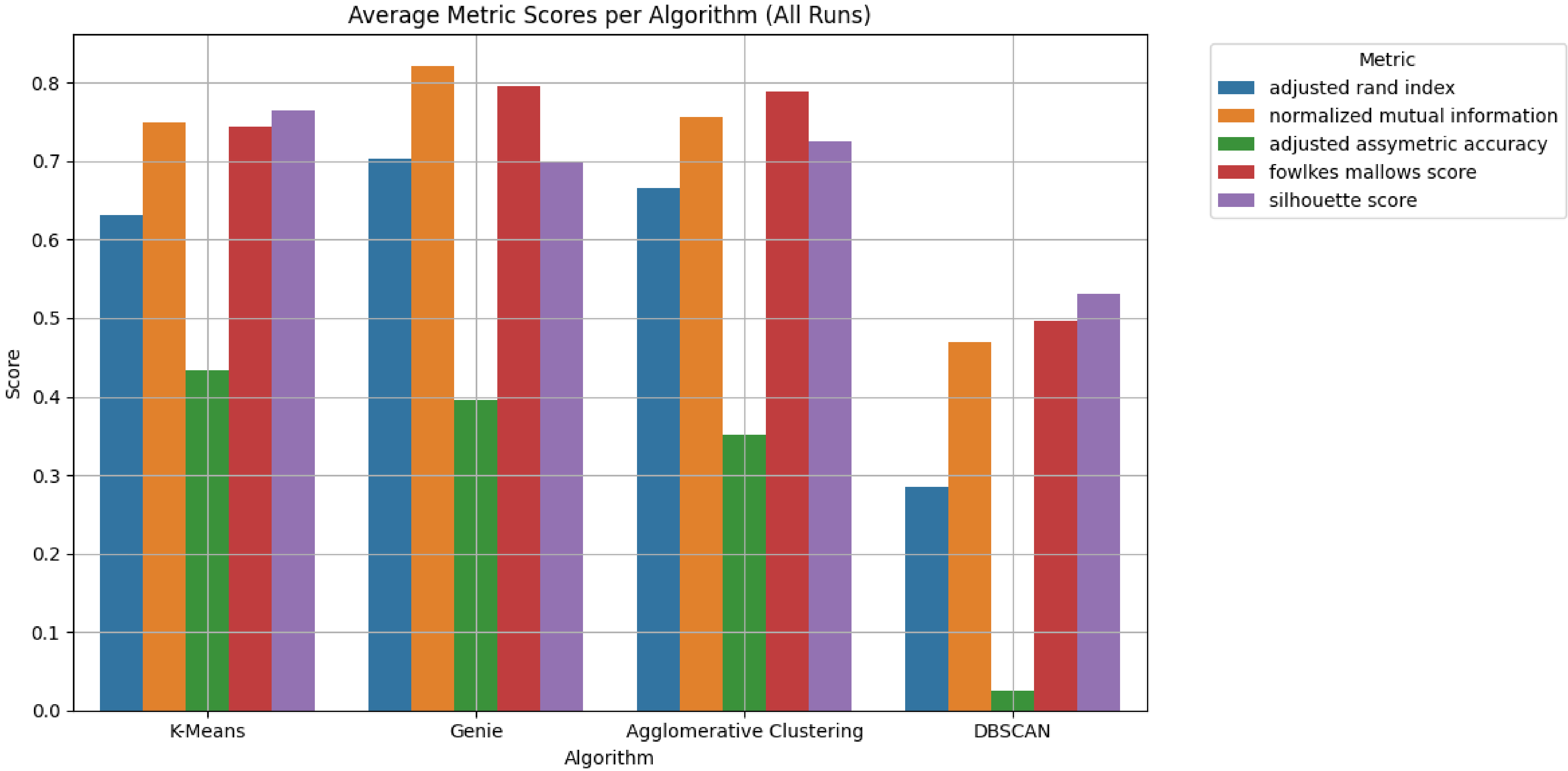
Best Adjusted Assymetric Accuracy per Dataset/Algorithm

Dataset	K-Means	Genie	Agglomerative Clustering	DBSCAN
a1	0.91	0.97	0.96	0.00
a2	0.90	0.97	0.97	0.00
a3	0.90	0.98	0.97	0.00
aggregation	0.71	0.88	1.00	0.00
compound	0.81	0.87	0.87	0.00
d31	0.88	0.97	0.96	0.00
flame	0.74	1.00	0.74	0.00
jain	0.69	1.00	0.81	0.00
pathbased	0.65	0.98	0.67	0.00
r15	1.00	1.00	1.00	1.00
s1	0.99	0.99	0.99	0.00
s2	0.88	0.95	0.95	0.00
s3	0.85	0.82	0.81	0.00
s4	0.78	0.77	0.68	0.00
spiral	0.02	1.00	1.00	1.00
unbalance	1.00	1.00	1.00	0.00

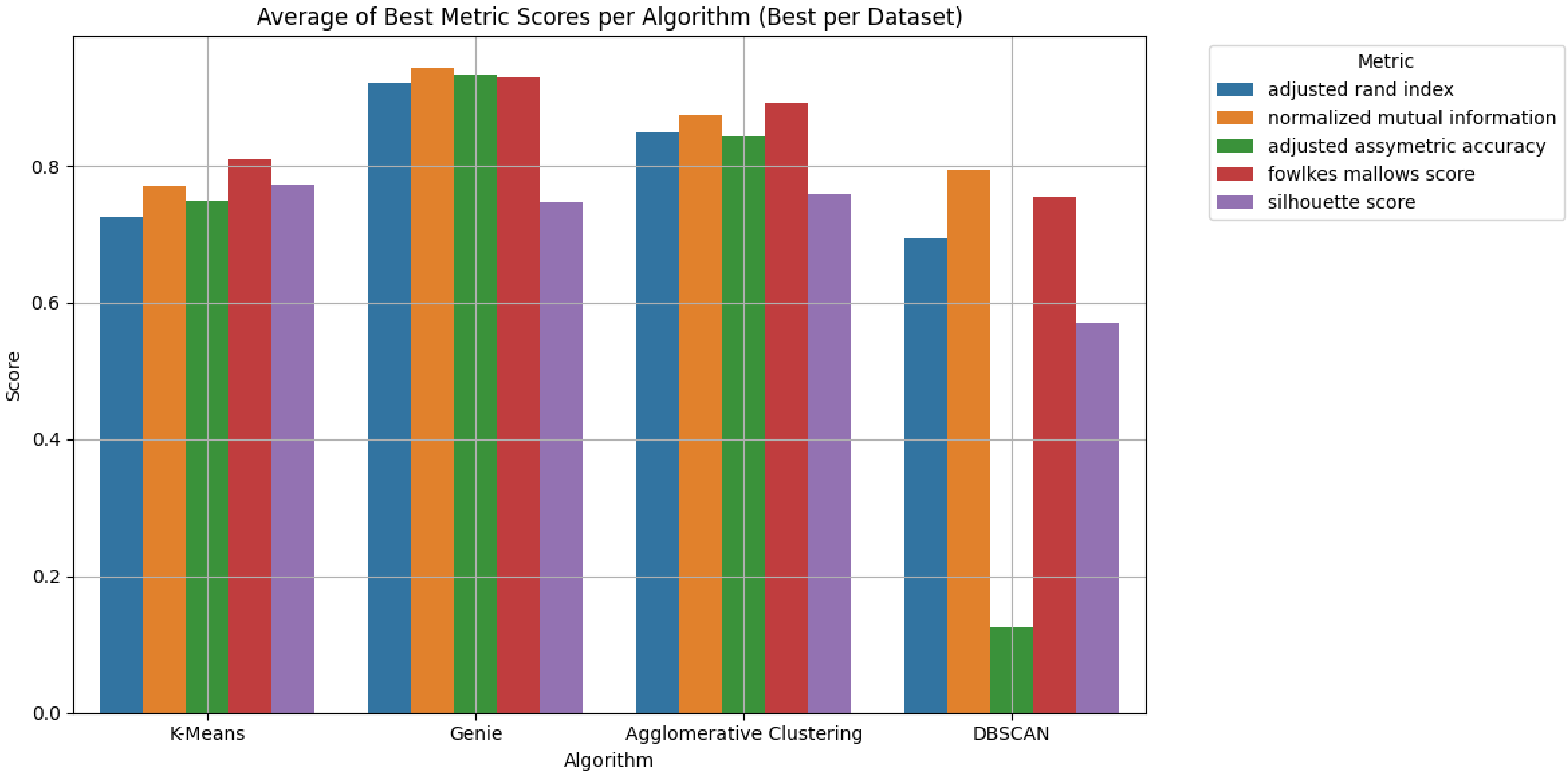
AGGREGATED RESULTS



AGGREGATED RESULTS



AGGREGATED RESULTS



THANK YOU FOR YOUR ATTENTION