Synthetic Data Error Rate Analysis

Overview

We compared synthetic samples against real asthma patient measurements to see how many "fake" points mistakenly clung to the majority class—using both Euclidean and Hassanat distance metrics on a 2-D toy dataset (x1, x2).

1. Data Loading

- Loaded Excel workbook Small Dataset Scatter Plots-01222023.xlsx.
- Parsed the **Original** sheet for real points and all method sheets (ROS, SMOTE, GAMMA, GSMOTE, SDDSMOTE, ADKNN) for synthetic points.

2. Real Data Extraction

- From **Original**:
  - **Majority**: columns Majority (x1) & Unnamed: 1 (x2)
  - **Minority**: columns Minority (x1) & Unnamed: 3 (x2)
- Combined into X_real (shape Nreal×2N_{real}\times 2) with labels y_real (0 = majority, 1 = minority).

3. Synthetic Data Extraction

- In each method sheet, all columns after the first four form (x1,x2) pairs.
- Stacked them into X_synth (shape Nsynth×2N_{synth}\times 2) for that method.

4. Nearest-Neighbor Classification

We implement two distance metrics:

4.1 Euclidean Distance

```
nn = NearestNeighbors(n_neighbors=1, metric='euclidean')

nn.fit(X_real)

_, idx = nn.kneighbors(X_synth)

nearest_labels = y_real[idx.flatten()]
```

4.2 Hassanat Distance (Eq. 1 in the paper)

For scalars p,qp,q:

dH(p,q)=1−1+min(p,q)1+max(p,q) d_H(p,q) = 1 - \frac{1 + \min(p,q)}{1 + \max(p,q)}

Vectorized over two features:

```
def hassanat_dist(p, q):
    mn, mx = np.minimum(p, q), np.maximum(p, q)
    return 1 - (1 + mn) / (1 + mx)
```

```
dist = np.zeros((S, R))  # S synth, R real

for j in range(2):

    p = X_synth[:, j][:, None]

    q = X_real[:, j][None, :]

    dist += hassanat_dist(p, q)

idx = np.argmin(dist, axis=1)

nearest_labels = y_real[idx]
```

5. Error Rate Calculation

- **CM** = number of synthetic points whose NN label = majority (0).
- **SS** = total synthetic points.
- **ErrorRate** = CM / SS.

6. Visualization

Separate scatter plots per method with:

- Majority → black hollow squares
- Minority → solid blue circles
- Synthetic → red triangles

7. Results & Outputs

- **Excel workbook** (per_point_error_rates_by_method.xlsx): one sheet per method listing (x1, x2), nearest label, SS, CM, ErrorRate for both Euclidean and Hassanat metrics.
- **Summary table** of ErrorRate per method:

Method	EuclidError	HassanatError
ROS	0.0000	0.0000
SMOTE	0.0083	0.0330
GAMMA	0.0583	0.0583

GSMOTE	0.0500	0.0500
SDDSMOTE	0.0000	0.0000
ADKNN	0.0000	0.0000