合成資料評估報告

Alex Chen

2025-02-04

A. 合成欄位

資料集名稱：penguins.csv

列數：333

欄位數：7

其中各資料型態之欄位數目如下：

Data Types str	count u32
”Float64”	2
”Int64”	2
”String”	3

B. 合成方法

合成模型：

```
# ENTER YOUR METHOD HERE
SDV - GaussianCopulaSynthesizer
```

合成參數：

```
# ENTER YOUR PARAMETERS HERE
{
    'enforce_min_max_values': True,
    'enforce_rounding': True,
    'locales': ['en_US'],
    'numerical_distributions': {},
    'default_distribution': 'beta'
}
```

前處理方法：

```
# ENTER YOUR PARAMETERS HERE
{
    'Species': UniformEncoder(),
    'Island': UniformEncoder(),
    'Sex': UniformEncoder(),
    'Culmen Length (mm)': FloatFormatter(learn_rounding_scheme=True,
     enforce_min_max_values=True),
    'Culmen Depth (mm)': FloatFormatter(learn_rounding_scheme=True,
    enforce_min_max_values=True),
    'Flipper Length (mm)': FloatFormatter(learn_rounding_scheme=True,
     enforce_min_max_values=True),
    'Body Mass (g)': FloatFormatter(learn_rounding_scheme=True,
    enforce_min_max_values=True)
}
```

合成條件設定：

```
# ENTER YOUR PARAMETERS HERE
[
    {
        'constraint_class': 'FixedIncrements',
        'constraint_parameters': {
            'column_name': 'Body Mass (g)',
            'increment_value': 5
        }
    }
]
```

備註：

```
# ENTER YOUR COMMENTS HERE
無
```

C. 合成表現

資料抽樣方法：

```
# ENTER YOUR METHOD HERE
普通抽樣
```

合成筆數：333

合成欄位數：7

其中各資料型態之欄位數目如下：

Data Types str	count u32
”Float64”	2
”Int64”	2
”String”	3

C-1. 合成資料診斷

分數：1.0

診斷分數介於 0.0 至 1.0 之間。此分數旨在確認合成資料的結構與原始資料相似。合成資料診斷分為兩項目，詳見細項。

資料有效性 (Data Validity)

分數：1.0

	Column	Metric	Score
0	Species	CategoryAdherence	1.0
1	Island	CategoryAdherence	1.0
2	Sex	CategoryAdherence	1.0
3	Culmen Length (mm)	BoundaryAdherence	1.0
4	Culmen Depth (mm)	BoundaryAdherence	1.0
5	Flipper Length (mm)	BoundaryAdherence	1.0
6	Body Mass (g)	BoundaryAdherence	1.0

判讀方法

資料有效性 (Data Validity) 的數值介於 0.0 至 1.0 之間，此數值應接近 1.0。若未達 1.0，可能代表以下情形：

1. 合成資料主鍵（若有）不唯一或有空值
2. 連續型資料數值超過原始範圍
3. 類別型資料類別數與原始資料不同

可根據以上原則檢查未達 1.0 之欄位，並應視具體資料情境決定是否接受。

資料結構 (Data Structure)

分數：1.0

	Metric	Score
0	TableStructure	1.0

判讀方法

資料結構 (Data Structure) 的數值介於 0.0 至 1.0 之間，此數值應接近 1.0。若未達 1.0，代表欄位數目或欄位名稱與原始資料不同，需要進行調整。

C-2. 保真度

分數：0.85

保真度分數介於 0.0 至 1.0 之間，此分數旨在確認合成資料的品質，數值越大代表資料型態/趨勢越相似。CAPE 建議保真度分數 0.75 以上為可接受。保真度分為兩項目，詳見細項。

欄位型態 (Column Shapes)

分數：0.92

	Column	Metric	Score
0	Species	TVComplement	0.969970
1	Island	TVComplement	0.942943
2	Sex	TVComplement	0.990991
3	Culmen Length (mm)	KSComplement	0.897898
4	Culmen Depth (mm)	KSComplement	0.921922
5	Flipper Length (mm)	KSComplement	0.840841
6	Body Mass (g)	KSComplement	0.894895

判讀方法

欄位型態 (Column Shapes) 的數值介於 0.0 至 1.0 之間，此數值越高越好。此項目計算每個欄位在原始資料與合成資料之間的統計相似度，即各欄位的邊際分配 (Marginal Distribution) 之相似度。欄位型態未達 0.75 者，代表合成資料有部分欄位跟原始資料分佈不一致，可觀察是否低分欄位皆有相同特徵，並針對該批欄位做特徵工程，例如連續型變項可考慮尺度調整、轉換；類別型變項可考慮概化（調整 bins 數）。

欄位對趨勢 (Column Pair Trends)

分數：0.78

	Column 1	Column 2	Metric	Score
0	Species	Island	ContingencySimilarity	0.630631
1	Species	Sex	ContingencySimilarity	0.969970
2	Species	Culmen Length (mm)	ContingencySimilarity	0.630631
3	Species	Culmen Depth (mm)	ContingencySimilarity	0.747748
4	Species	Flipper Length (mm)	ContingencySimilarity	0.537538
5	Species	Body Mass (g)	ContingencySimilarity	0.678679
6	Island	Sex	ContingencySimilarity	0.936937
7	Island	Culmen Length (mm)	ContingencySimilarity	0.714715
8	Island	Culmen Depth (mm)	ContingencySimilarity	0.648649
9	Island	Flipper Length (mm)	ContingencySimilarity	0.537538
10	Island	Body Mass (g)	ContingencySimilarity	0.651652
11	Sex	Culmen Length (mm)	ContingencySimilarity	0.762763
12	Sex	Culmen Depth (mm)	ContingencySimilarity	0.732733
13	Sex	Flipper Length (mm)	ContingencySimilarity	0.651652
14	Sex	Body Mass (g)	ContingencySimilarity	0.735736
15	Culmen Length (mm)	Culmen Depth (mm)	CorrelationSimilarity	0.989867
16	Culmen Length (mm)	Flipper Length (mm)	CorrelationSimilarity	0.969011
17	Culmen Length (mm)	Body Mass (g)	CorrelationSimilarity	0.994332
18	Culmen Depth (mm)	Flipper Length (mm)	CorrelationSimilarity	0.948690
19	Culmen Depth (mm)	Body Mass (g)	CorrelationSimilarity	0.943426
20	Flipper Length (mm)	Body Mass (g)	CorrelationSimilarity	0.952075

判讀方法

欄位對趨勢 (Column Pair Trends) 的數值介於 0.0 至 1.0 之間，此數值越高越好。此項目計算兩欄位相關性在原始資料與合成資料間的相似度。欄位對趨勢較低，代表合成資料的部分欄位組合與原始資料趨勢不一致，很可能是跨數值與類別的欄位對趨勢不佳，可嘗試做特徵工程改善。以實務上來說，欄位對趨勢通常較難達到高分數，需視需求進行不同門檻的要求。

其他保真度改善建議

除了上述原因造成保真度偏低之外，也有可能是由於原始資料有內隱的強制性資料邏輯，而生成出偏移的結果，導致保真度降低，此時可做約束條件限制產出。

C-3. 保護力

指認性 (Singling-Out)：

```
ENTER YOUR VALUE HERE
```

使用參數：

```
ENTER YOUR PARAMETERS HERE
```

判讀方法

指認性 (Singling-Out) 風險的數值介於 0.0 至 1.0 之間，此數值越低越好。此項目代表合成資料中只有特定一筆資料有獨一無二組合的風險，不必然可以進行再辨認（即使知道這個資訊，也不代表可以知道這個人是誰）。若此指標高於 0.09，代表隱私風險較高，合成資料中有多筆跟原始資料高度相同的紀錄，可利用約束條件來加強資料邏輯，剔除極端資料組合，減少指認性風險。

連結性 (Linkability)：

```
ENTER YOUR VALUE HERE
```

使用參數：

```
ENTER YOUR PARAMETERS HERE
```

判讀方法

連結性 (Linkability) 風險的數值介於 0.0 至 1.0 之間，此數值越低越好。此項目代表判斷兩筆（或以上）資料屬於同個人或同個團體的風險，意即當攻擊者同時有兩個與合成資料欄位重複的資料集，但缺乏將兩份資料連結起來的線索時，合成資料可以成為此線索的風險高低。若此指標高於 0.09，代表隱私風險較高，需重新評估合成資料中所指定的欄位分組方式，再決定該如何減少連結性風險。

推論性 (Inference)：

```
ENTER YOUR VALUE HERE
```

使用參數：

```
ENTER YOUR PARAMETERS HERE
```

> **判讀方法**
>
> 推論性 (Inference) 風險的數值介於 0.0 至 1.0 之間，此數值越低越好。此項目代表攻擊者可以猜測（推論）出資料中未知變數的值，例如攻擊者知道一筆資料的部分資訊，可藉由合成資料推論此筆資料的秘密資訊。若此指標高於 0.09，代表隱私風險較高，合成資料中秘密資訊欄位很容易被猜到，應先檢查秘密資訊欄位是否與輔助資訊欄位有高度邏輯依賴關係，再行處置。需注意的是，實務上由於此方法計算方式與機器學習高度重疊，因此結果僅供參考。

C-4. 實用性

下游任務：

```
ENTER YOUR TASK HERE
分類/聚類/迴歸
```

原始資料訓練模型：

```
ENTER YOUR MODEL HERE
```

原始資料訓練參數：

```
ENTER YOUR PARAMETERS HERE
```

合成資料訓練模型：

```
ENTER YOUR MODEL HERE
```

合成資料訓練參數：

```
ENTER YOUR PARAMETERS HERE
```

其他訓練參數：

```
ENTER YOUR PARAMETERS HERE
e.g., train_test_split, cross_validate
```

驗證指標：

```
ENTER YOUR METRICS HERE
```

原始訓練資料對原始測試資料驗證分數：

```
ENTER YOUR SCORES HERE
```

合成訓練資料對原始測試資料驗證分數：

```
ENTER YOUR SCORES HERE
```

> **判讀方法**
>
> 實用性旨在比較原始資料與合成資料在原始資料「測試資料集」上的表現，合成資料上的表現應越高越好，至少在原始資料與合成資料的表現差異不應太大。若此情況發生，應從特徵工程、資料前處理、合成模型、機器學習等角度切入。請與 CAPE 團隊討論可能的可行方案。

> **分類任務的指標判讀**
>
> CAPE 建議關注以下關鍵指標在測試集上的表現即可：
>
> - ROC-AUC：大部分分類情境使用，0.8 以上可接受
> - MCC：適合處理不平衡資料集，0.5 以上可接受

D. CAPE 意見

```
ENTER YOUR COMMENTS HERE
```