

# Synthetic DATA

Performance de SDV dans la génération de données

Canva

# SDV : Metrics

- MultiTableMetric
- Multi Table Quality Report
- Multi Table BayesianNetwork Metrics
- CardinalityStatisticSimilarity Metrics
- Multi Table Detection Metrics



# MultiTableMetric

Les métriques trouvées dans ce dossier fonctionnent sur des ensembles de données multi-tables, passés comme deux dicts python contenant des tables comme *pandas.DataFrame*.

**Parent-Child Detection metrics:** Métriques qui dé-normalisent chaque relation parent-enfant dans le dataset, puis exécutent une métrique de détection de single table sur les tables générées.

- LogisticParentChildDetection
- SVCParentChildDetection

**Multi Single Table Metrics:** Métrique qui exécute une métrique de single table sur chaque table du dataset et renvoie ensuite le score moyen obtenu par celle-ci.

- CSTest
- KSComplement
- LogisticDetection
- SVCDetection
- BNLikelihood
- BNLogLikelihood

# Multi Table Quality Report

Le Quality Report de SDMetrics évalue dans quelle mesure vos données synthétiques capturent les propriétés mathématiques de vos données réelles.

C'est ce qu'on appelle la fidélité des données synthétiques. Le rapport exécute certaines métriques pour mesurer ces propriétés et résume les résultats.

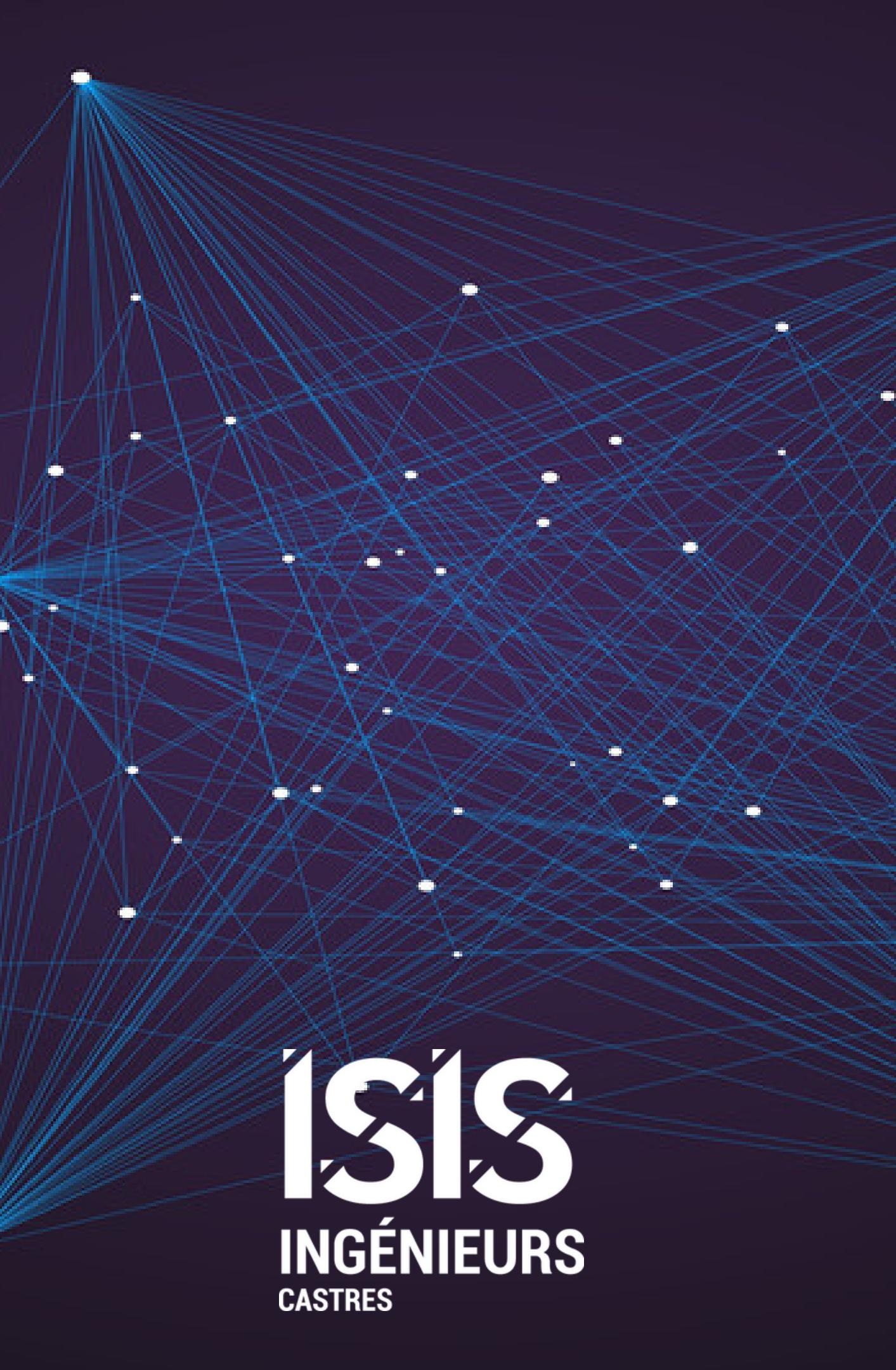
Une fois le rapport généré, vous pouvez obtenir plus de détails pour expliquer les résultats, visualiser les scores et enregistrer le rapport pour le partager.

*Canva*

Les données synthétiques rendent-elles compte de la 'shape' de chaque colonne ?

La 'shape' d'une colonne décrit sa distribution globale. Plus le score est élevé, plus les distributions des données réelles et synthétiques sont similaires.





**ISIS**  
**INGÉNIEURS**  
CASTRES

# Multi Table BayesianNetwork Metrics

MultiSingleTableMetric est basé sur le SingleTable BNLikelihood.

Cette métrique adapte un BayesianNetwork aux données réelles, puis évalue la probabilité que les données synthétiques appartiennent à la même distribution.

La sortie est la probabilité moyenne sur toutes les lignes synthétiques.

Canva

# CardinalityStatistic

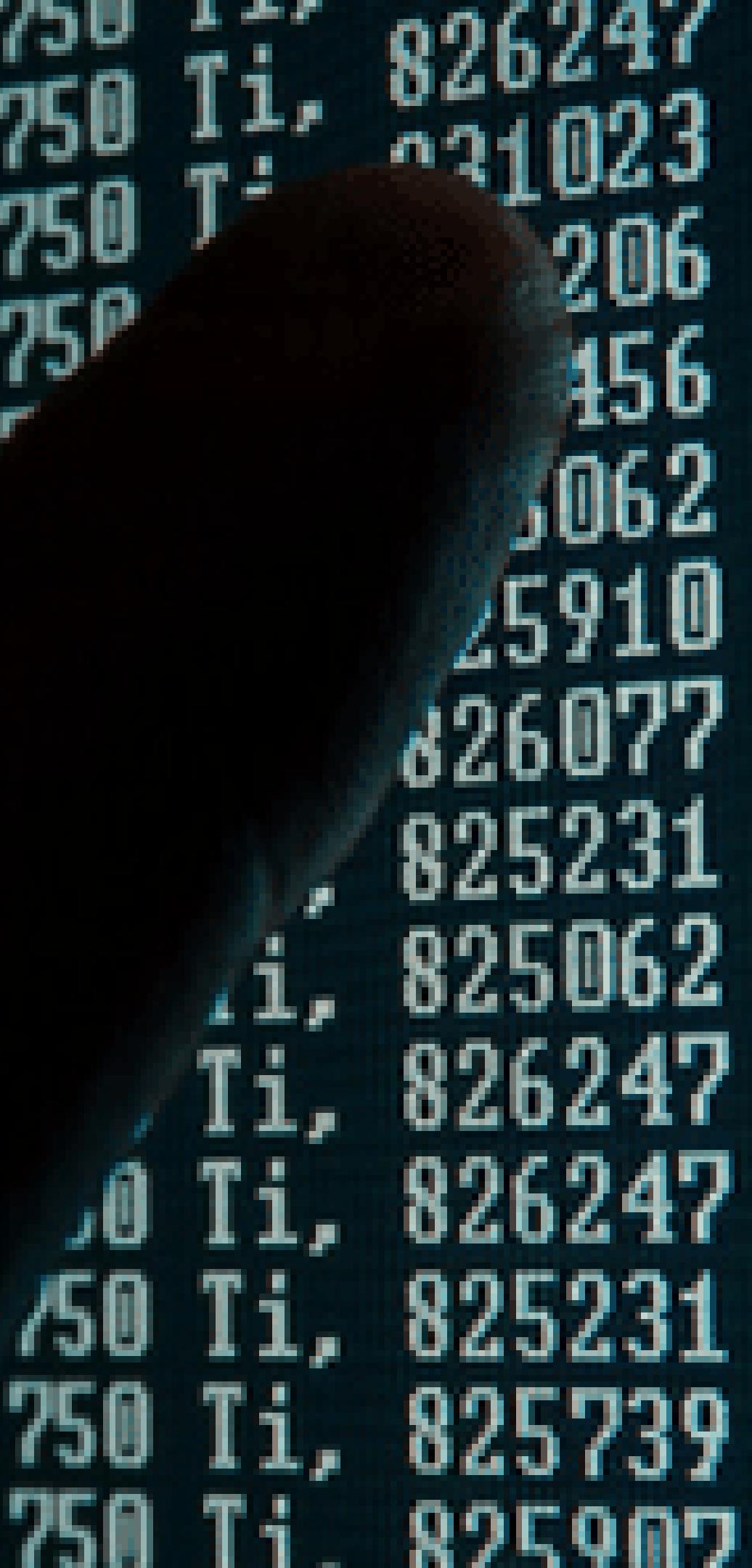
## Similarity metric

Si vous avez des multi tables et connectées, cette métrique mesure si la cardinalité de la table parent est la même entre les ensembles de données réels et synthétiques. La cardinalité est définie comme le nombre de lignes enfants pour chaque parent.

Cette méthode est destinée à être utilisée sur les colonnes d'identification (clés primaires et étrangères). Les ID des clés primaires doivent être uniques alors que les ID des clés étrangères peuvent se répéter.

Canva

- (meilleur) 1,0 : Les valeurs de cardinalité sont les mêmes dans les données réelles et synthétiques.
- (pire) 0,0 : Les valeurs de cardinalité sont totalement différentes.



# Multi Table Detection Metrics

Classe de base pour la détection par apprentissage automatique basée sur des métriques sur plusieurs tableaux.

Ces métriques construisent un classificateur d'apprentissage automatique qui apprend à distinguer les données synthétiques des données réelles, qui sont ensuite évaluées par validation croisée.

Canva

Le résultat de la métrique est égal à un moins le score moyen ROC AUC obtenu.



**ISIS**  
**INGÉNIEURS**  
CASTRES

# Merci pour votre écoute

Synthetic Data - PTUT

**ISIS**  
INGÉNIEURS  
CASTRES