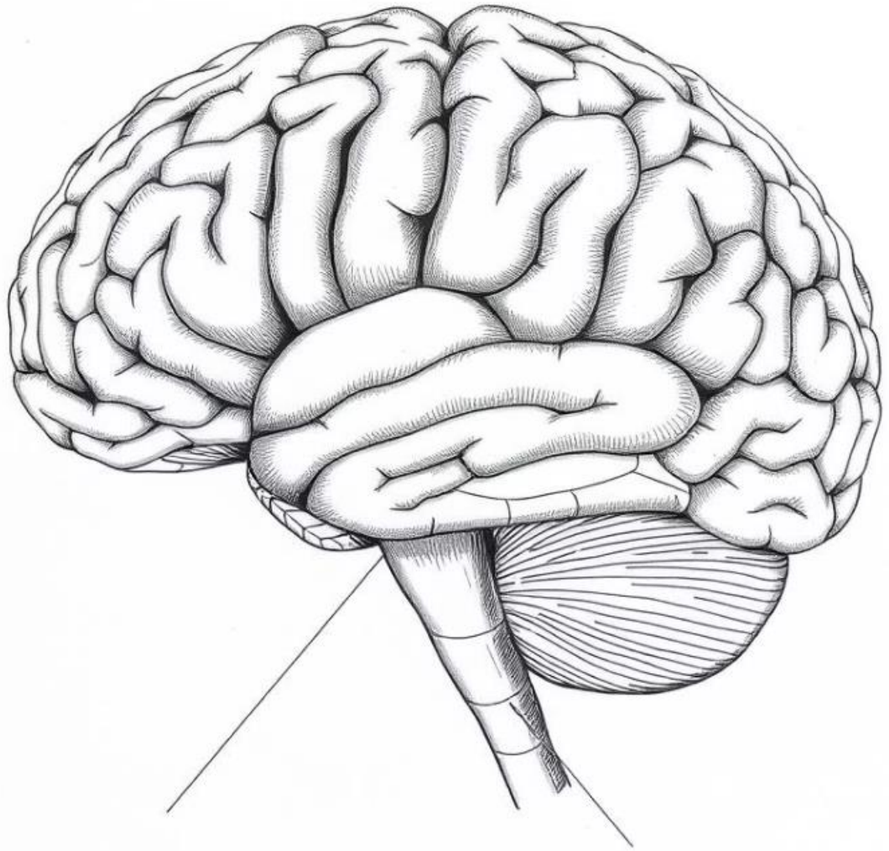
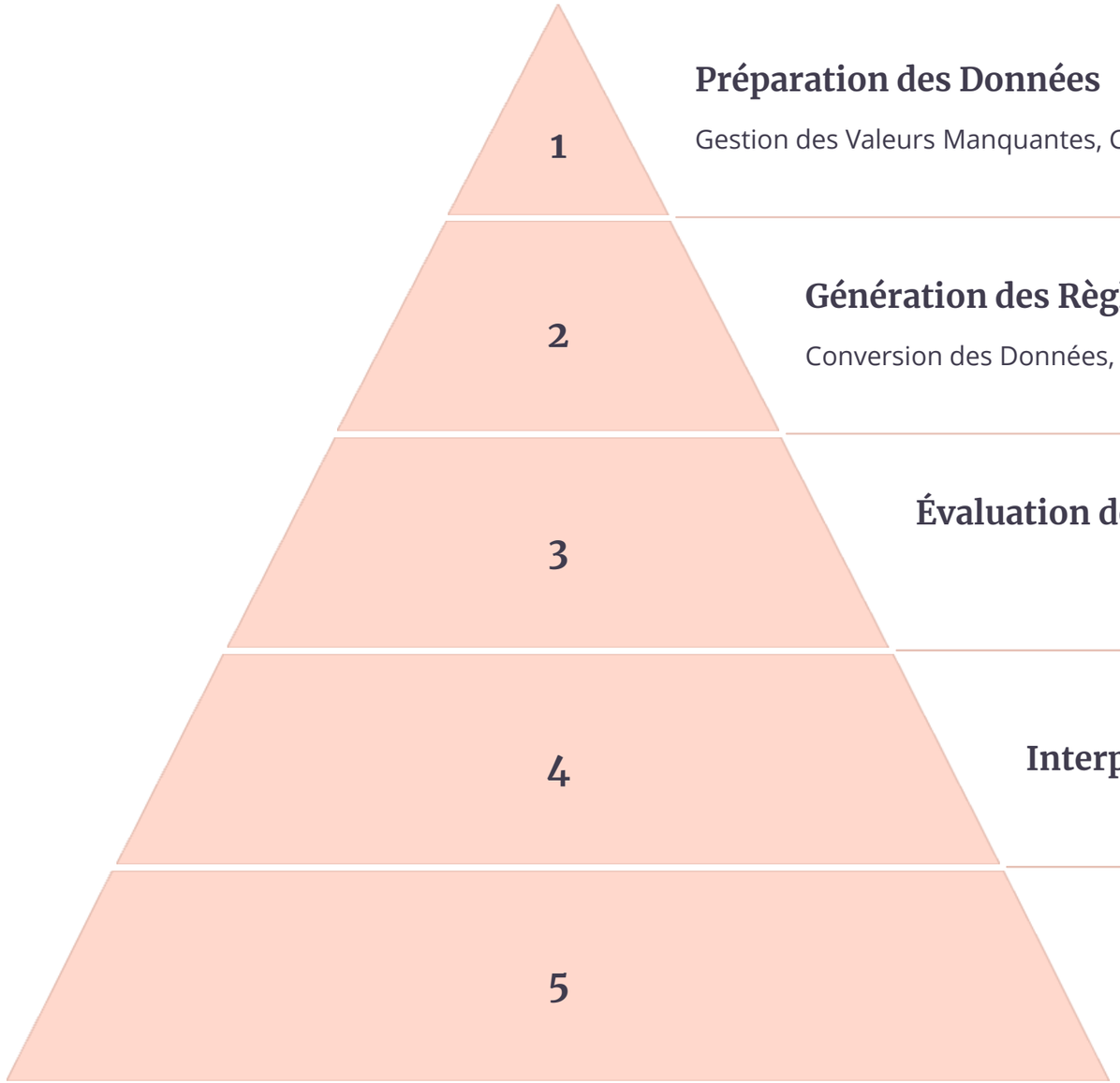


Génération et Analyse des Règles d'Association

PREDICTION DES RISQUES D'AVC

Plan du projet



Préparation des Données

Gestion des Valeurs Manquantes, Conversion des Variables Catégorielles et Normalisation des Variables Numériques

Génération des Règles d'Association

Conversion des Données, Génération des Itemsets Fréquents et Génération des Règles

Évaluation des Règles d'Association

Interprétation des résultats

Interface Utilisateur

Introduction

Selon l'Organisation mondiale de la Santé (OMS), l'AVC est la deuxième cause de décès dans le monde, responsable d'environ 11 % du total des décès.

L'ensemble de données qu'on va utiliser (**Stroke Prediction Dataset**) permet de prédire le risque d'AVC chez un patient, en fonction de paramètres tels que le sexe, l'âge, les différentes pathologies et le tabagisme. Chaque ligne de données fournit des informations pertinentes sur le patient.

Préparation des Données

Gestion des Valeurs Manquantes

Variables Numériques (age, avg_glucose_level, bmi**) :**

- Utilisation de la stratégie d'imputation par la moyenne

- Cette approche préserve la distribution des données numériques

Variables Catégorielles (gender, hypertension, heart_disease, ever_married, work_type, Residence_type, smoking_status**):**

- Utilisation de la stratégie d'imputation par le mode (valeur la plus fréquente)

- Cette méthode est plus appropriée pour les données catégorielles car elle maintient la cohérence des catégories

Conversion des Variables Catégorielles

- Variables traitées : `gender`, `ever_married`, `work_type`, `Residence_type`, `smoking_status`

Méthode : One-hot encoding avec préfixe

Justification : Simplification pour l'analyse des règles d'association

Normalisation des Variables Numériques

- Variables Traitées :
- `age`, `avg_glucose_level`, `bmi`

Méthode : Standardisation (StandardScaler)

Processus :

- Centrage des données (moyenne = 0)

- Réduction (écart-type = 1)

Justification : Permet de comparer des variables sur des échelles différentes

Génération des Règles d'Association

Conversion des Données

- Utilisation de `TransactionEncoder` pour transformer les données en format binaire

- Création d'un DataFrame binaire où chaque colonne représente une variable

- Cette étape est cruciale car l'algorithme Apriori nécessite des données binaires

Génération des Itemsets Fréquents

- Application de l'algorithme Apriori avec le seuil de support minimum

- Utilisation de `use_colnames=True` pour conserver les noms des variables

- Cette étape identifie les combinaisons de variables qui apparaissent fréquemment dans le dataset

Génération des Règles

- Application de la fonction `association_rules` avec le seuil de confiance minimum

- Utilisation de la métrique "**confidence**" pour évaluer la fiabilité des règles

- Les règles générées représentent des relations significatives entre les variables

Évaluation des Règles d'Association

Les mesures sélectionnées dans cette analyse **lift**, **confidence**, **conviction** et **information gain** ont été retenues en raison de leur pertinence démontrée dans l'article "**On selecting interestingness measures for association rules"**, basé sur une évaluation multicritère.

Mesure	Utilité principale	Interprétation des valeurs
Lift	Mesure l'indépendance entre les items	> 1 : corrélation positive (apparition conjointe plus fréquente que prévu) < 1 : corrélation négative (moins fréquente que prévu) = 1 : indépendance
Confiance (Confidence)	Évalue la fiabilité de la règle (proba que Y apparaisse si X est là)	Valeur proche de 1 : règle souvent vraieMais peut être trompeuse si Y est très fréquent
Conviction	Détecte les dépendances logiques	= 1 : indépendance 1 : corrélation positive Valeur élevée : forte dépendance logique
Gain d'information	Met en valeur l'aspect surprenant ou informatif d'une règle	Plus la valeur est élevée, plus la règle apporte de l'information nouvelle ou inattendue

Ces quatre mesures présentent un bon équilibre entre **interprétabilité**, **stabilité** et **capacité à distinguer les règles réellement intéressantes**, ce qui les rend particulièrement adaptées à l'analyse des règles d'association.

Évaluation des Règles d'Association

Les mesures **certainty factor** et **jaccard** ne figurent pas dans l’article, mais ont été ajoutées dans notre analyse pour enrichir l’évaluation des règles d’association.

L’**indice de Jaccard** permet de mesurer la similarité entre l’antécédent et le conséquent, ce qui est utile pour identifier les règles basées sur une forte cooccurrence relative.

La mesure de **certitude** (certainty) prend en compte les faux positifs et permet d’évaluer la confiance d’une règle tout en corrigeant les biais liés aux déséquilibres de classe.

Ces deux mesures offrent ainsi des perspectives complémentaires, notamment dans des contextes où la précision ou la compacité des règles est recherchée.

Interprétation des résultats

Les seuils utilisés:

Seuil de Confiance (min_confidence = 0.3)

Ce seuil de 30% est modéré et approprié car : Il permet de capturer des associations qui ne sont pas évidentes mais qui ont une signification statistique

Il évite les règles trop faibles (qui pourraient être dues au hasard) Il reste suffisamment bas pour détecter des relations complexes entre facteurs de risqué.

Dans le domaine médical, une confiance de 30% peut être significative, surtout si le support est faible.

Seuil de Support (min_support = 0.001)

Ce seuil très bas (0,1 %) est adapté au domaine médical, car il permet de détecter des associations rares mais importantes.
En effet :

- Certaines maladies ou conditions médicales rares doivent être prises en compte.
- Des combinaisons spécifiques de facteurs de risque, bien que peu fréquentes, peuvent avoir une grande valeur clinique.
  Si on choisissait un seuil plus élevé, on risquerait de passer à côté de ces associations rares mais pourtant significatives.
  Dans le domaine médical, même des patterns peu fréquents peuvent avoir une réelle importance.

Interprétation des résultats

Les résultats:

```
(venv) C:\Users\kmahd\OneDrive\Bureau\mini_projet_BI>c:/Users/kmahd/OneDrive/Bureau/mini_projet_BI/venv/Scripts/python.exe c:/Users/kmahd/OneDrive/Bureau/mini_p
rojet_BI/utils.py

(venv) C:\Users\kmahd\OneDrive\Bureau\mini_projet_BI>c:/Users/kmahd/OneDrive/Bureau/mini_projet_BI/venv/Scripts/python.exe c:/Users/kmahd/OneDrive/Bureau/mini_p
rojet_BI/stroke_analysis.py
Chargement et prétraitement des données...
Gestion des valeurs manquantes...
Conversion des variables catégorielles...
Normalisation des variables numériques...

Résumé du prétraitement:
Nombre total de variables: 22
Nombre de variables catégorielles encodées: 5
Nombre de variables numériques normalisées: 3
Nombre de variables binaires: 2

Génération des règles d'association avec:
Support minimum: 0.001
Confiance minimum: 0.3
```

Le nombre total de variables (22) correspond au nombre de colonnes du DataFrame après prétraitement. Ce nombre est supérieur au nombre de colonnes du fichier CSV d'origine, car les variables catégorielles ont été transformées en plusieurs colonnes via l'encodage (one-hot encoding), et certaines variables numériques ou binaires ont pu être transformées ou normalisées.

Interprétation des résultats

Meilleures 5 règles par Lift :

```
Meilleures règles par Lift (Indépendance):

Règle: ['o', 'd', 'a'] => ['r', 'm', 'e', '_', 'i']
Support: 0.001
Confidence: 1.000
Lift: 730.000
Conviction: inf
Leverage: 0.001
Jaccard: 0.857
Certainty Factor: 1.000
Information Gain: 0.011

Règle: ['n', '_', 's'] => ['u']
Support: 0.001
Confidence: 0.857
Lift: 730.000
Conviction: 6.992
Leverage: 0.001
Jaccard: 0.857
Certainty Factor: 0.857
Information Gain: 0.011

Règle: ['g', 'o'] => ['n', 'i', 'm', 'a']
Support: 0.001
Confidence: 1.000
Lift: 730.000
Conviction: inf
Leverage: 0.001
Jaccard: 0.857
Certainty Factor: 1.000
```

```
Information Gain: 0.011

Règle: ['i', 'r', 'm'] => ['a', 'e', '_', 'o', 'd']
Support: 0.001
Confidence: 0.857
Lift: 730.000
Conviction: 6.992
Leverage: 0.001
Jaccard: 0.857
Certainty Factor: 0.857
Information Gain: 0.011

Règle: ['o', '_', 'd', 'a'] => ['i', 'r', 'm', 'e']
Support: 0.001
Confidence: 1.000
Lift: 730.000
Conviction: inf
Leverage: 0.001
Jaccard: 0.857
Certainty Factor: 1.000
Information Gain: 0.011
```

Interprétation des résultats

Meilleures règles par Confiance (Fiabilité) :

```
Meilleures règles par Confiance (Fiabilité):

Règle: ['o', 'd', 'a'] => ['r', 'm', 'e', '_', 'i']
Support: 0.001
Confidence: 1.000
Lift: 730.000
Conviction: inf
Leverage: 0.001
Jaccard: 0.857
Certainty Factor: 1.000
Information Gain: 0.011

Règle: ['v'] => ['_']
Support: 0.001
Confidence: 1.000
Lift: 255.500
Conviction: inf
Leverage: 0.001
Jaccard: 0.350
Certainty Factor: 1.000
Information Gain: 0.011

Règle: ['u'] => ['_']
Support: 0.001
Confidence: 1.000
Lift: 255.500
Conviction: inf
Leverage: 0.001
Jaccard: 0.300
Certainty Factor: 1.000
```

```
Information Gain: 0.009

Règle: ['o', 'i', 'd', 'a'] => ['_', 'r', 'm', 'e']
Support: 0.001
Confidence: 1.000
Lift: 567.778
Conviction: inf
Leverage: 0.001
Jaccard: 0.667
Certainty Factor: 1.000
Information Gain: 0.011

Règle: ['o', '_', 'd', 'a'] => ['i', 'r', 'm', 'e']
Support: 0.001
Confidence: 1.000
Lift: 730.000
Conviction: inf
Leverage: 0.001
Jaccard: 0.857
Certainty Factor: 1.000
Information Gain: 0.011
```

Interprétation des résultats

Meilleures règles par Conviction (Dépendance):

```
Meilleures règles par Conviction (Dépendance):

Règle: ['o', 'd', 'a'] => ['r', 'm', 'e', '_', 'i']
Support: 0.001
Confidence: 1.000
Lift: 730.000
Conviction: inf
Leverage: 0.001
Jaccard: 0.857
Certainty Factor: 1.000
Information Gain: 0.011

Règle: ['v'] => ['_']
Support: 0.001
Confidence: 1.000
Lift: 255.500
Conviction: inf
Leverage: 0.001
Jaccard: 0.350
Certainty Factor: 1.000
Information Gain: 0.011

Règle: ['u'] => ['_']
Support: 0.001
Confidence: 1.000
Lift: 255.500
Conviction: inf
Leverage: 0.001
Jaccard: 0.300
```

```
Certainty Factor: 1.000
Information Gain: 0.009

Règle: ['o', 'i', 'd', 'a'] => ['_', 'r', 'm', 'e']
Support: 0.001
Confidence: 1.000
Lift: 567.778
Conviction: inf
Leverage: 0.001
Jaccard: 0.667
Certainty Factor: 1.000
Information Gain: 0.011

Règle: ['o', '_', 'd', 'a'] => ['i', 'r', 'm', 'e']
Support: 0.001
Confidence: 1.000
Lift: 730.000
Conviction: inf
Leverage: 0.001
Jaccard: 0.857
Certainty Factor: 1.000
Information Gain: 0.011
```

Interprétation des résultats

Meilleures règles par Jaccard (Similarité):

```
Meilleures règles par Jaccard (Similarité):

Règle: ['r'] => ['e']
Support: 0.004
Confidence: 1.000
Lift: 243.333
Conviction: inf
Leverage: 0.004
Jaccard: 0.952
Certainty Factor: 1.000
Information Gain: 0.031

Règle: ['e'] => ['r']
Support: 0.004
Confidence: 0.952
Lift: 243.333
Conviction: 20.918
Leverage: 0.004
Jaccard: 0.952
Certainty Factor: 0.952
Information Gain: 0.031

Règle: ['k'] => ['o', 't']
Support: 0.002
Confidence: 1.000
Lift: 464.545
Conviction: inf
Leverage: 0.002
Jaccard: 0.909
Certainty Factor: 1.000
Information Gain: 0.017

Règle: ['o', 't'] => ['k']
Support: 0.002
Confidence: 0.909
Lift: 464.545
Conviction: 10.978
Leverage: 0.002
Jaccard: 0.909
Certainty Factor: 0.909
Information Gain: 0.017

Règle: ['k'] => ['o', '_', 't']
Support: 0.002
Confidence: 0.900
Lift: 511.000
Conviction: 9.982
Leverage: 0.002
Jaccard: 0.900
Certainty Factor: 0.900
Information Gain: 0.016
```

Interprétation des résultats

Meilleures règles par Certainty Factor (Certitude):

```
Meilleures règles par Certainty Factor (Certitude):

Règle: ['o', 'd', 'a'] => ['r', 'm', 'e', '_', 'i']
Support: 0.001
Confidence: 1.000
Lift: 730.000
Conviction: inf
Leverage: 0.001
Jaccard: 0.857
Certainty Factor: 1.000
Information Gain: 0.011

Règle: ['v'] => ['_']
Support: 0.001
Confidence: 1.000
Lift: 255.500
Conviction: inf
Leverage: 0.001
Jaccard: 0.350
Certainty Factor: 1.000
Information Gain: 0.011

Règle: ['u'] => ['_']
Support: 0.001
Confidence: 1.000
Lift: 255.500
Conviction: inf
Leverage: 0.001
Jaccard: 0.300
Certainty Factor: 1.000
```

```
Information Gain: 0.009

Règle: ['o', 'i', 'd', 'a'] => ['_', 'r', 'm', 'e']
Support: 0.001
Confidence: 1.000
Lift: 567.778
Conviction: inf
Leverage: 0.001
Jaccard: 0.667
Certainty Factor: 1.000
Information Gain: 0.011

Règle: ['o', '_', 'd', 'a'] => ['i', 'r', 'm', 'e']
Support: 0.001
Confidence: 1.000
Lift: 730.000
Conviction: inf
Leverage: 0.001
Jaccard: 0.857
Certainty Factor: 1.000
Information Gain: 0.011
```

Interprétation des résultats

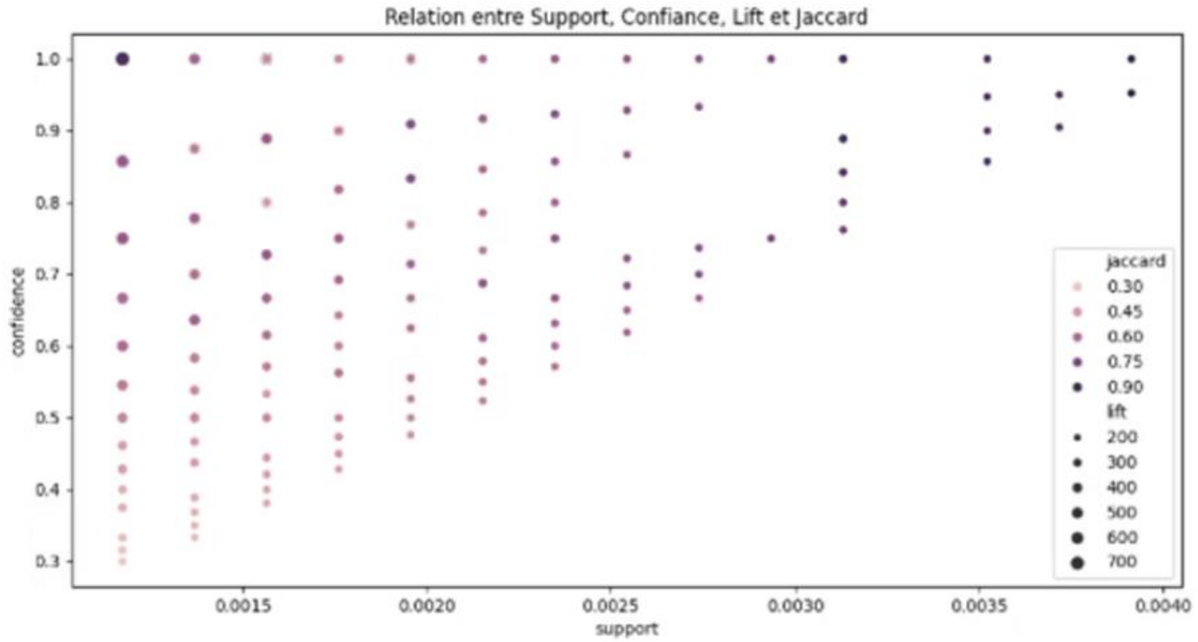
Meilleures règles par Information Gain (Information):

```
Meilleures règles par Information Gain (Information):

Règle: ['r'] => ['e']
Support: 0.004
Confidence: 1.000
Lift: 243.333
Conviction: inf
Leverage: 0.004
Jaccard: 0.952
Certainty Factor: 1.000
Information Gain: 0.031

Règle: ['e'] => ['r']
Support: 0.004
Confidence: 0.952
Lift: 243.333
Conviction: 20.918
Leverage: 0.004
Jaccard: 0.952
Certainty Factor: 0.952
Information Gain: 0.031

Règle: ['_'] => ['e']
Support: 0.004
Confidence: 0.950
Lift: 231.167
Conviction: 19.918
Leverage: 0.004
Jaccard: 0.864
Certainty Factor: 0.950
```

```
Information Gain: 0.029

Règle: ['e'] => ['_']
Support: 0.004
Confidence: 0.905
Lift: 231.167
Conviction: 10.459
Leverage: 0.004
Jaccard: 0.864
Certainty Factor: 0.904
Information Gain: 0.029

Règle: ['r', '_'] => ['e']
Support: 0.004
Confidence: 1.000
Lift: 243.333
Conviction: inf
Leverage: 0.004
Jaccard: 0.857
Certainty Factor: 1.000
Information Gain: 0.028

(venv) C:\Users\kmahd\OneDrive\Bureau\mini_projet_BI>
```

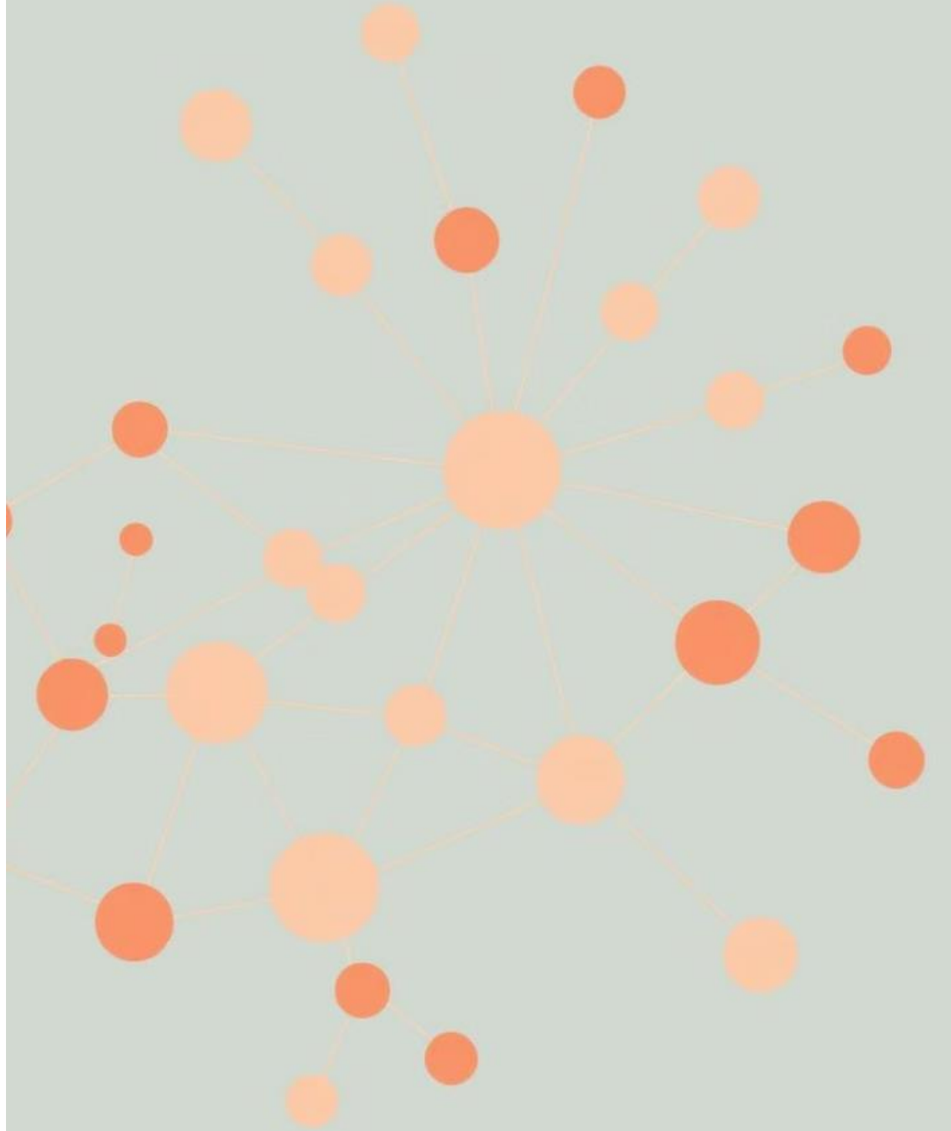
Interprétation des résultats

Relation entre Support, Confiance, Lift et Jaccard :



La visualisation montre la relation entre le support, la confiance, le lift et le coefficient de Jaccard, ce qui permet de voir comment ces différentes métriques sont corrélées.

Interface Utilisateur



Merci !