La préparation et la qualification des données

Introduction :

Préparation des données

- Le nettoyage et la préparation des données avec pour finalité le besoin d'avoir des données propres et représentatives de la réalité
- Des informations exactes et cohérentes pour les applications BI et BA

Qualification des données

70 % des données dont disposent les entreprises ne sont jamais utilisées, faute d'avoir été qualifiées (vérifiées, complétées, rendues exploitables). Ne pas les enrichir, c'est générer une perte sèche, et limiter son ROI

Objectif : améliorer le suivi, augmenter la valeur des données

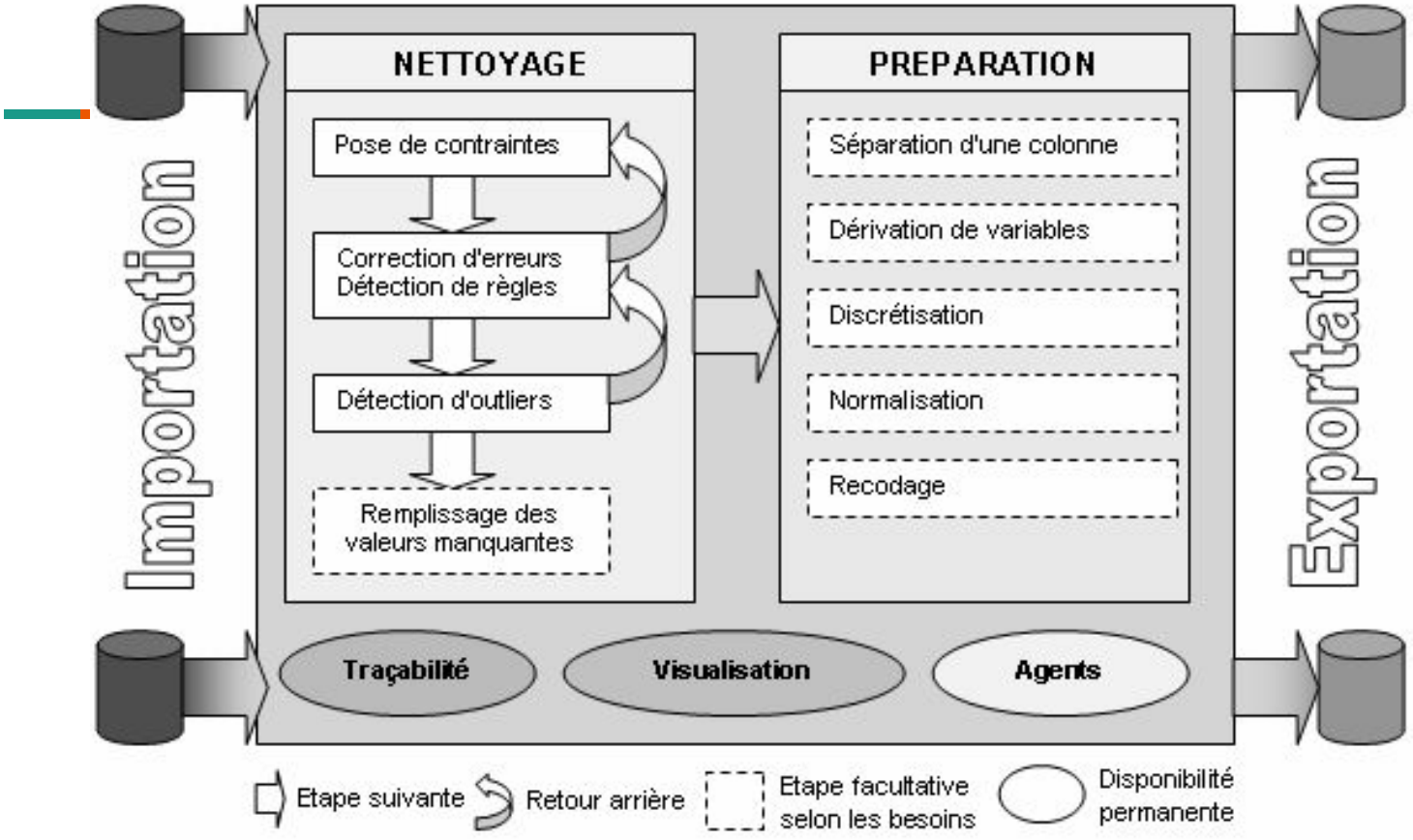
- Qualifier les bases clientes existantes
- Acquérir de nouveaux prospects pour les rentrer en base

Pourquoi prétraiter et nettoyer les données ?

Qualité des données & modèles prédictifs performant

Anomalies ou des valeurs incorrectes qui compromettent la qualité du jeu de données

- **Caractère incomplet :** des valeurs ou des attributs sont manquants
- **Bruit :** les données contiennent des enregistrements erronés ou des aberrations
- **Incohérence :** les données contiennent des enregistrements en conflit ou des contradictions


Importation
NETTOYAGE
Pose de contraintes
Correction d'erreurs
Détection de règles
Détection d'outliers
Remplissage des valeurs manquantes
PREPARATION
Séparation d'une colonne
Dérivation de variables
Discrétisation
Normalisation
Recodage
Exportation
Traçabilité
Visualisation
Agents
Etape suivante
Retour arrière
Etape facultative selon les besoins
Disponibilité permanente

Normalisation des données

Méthode de prétraitement des données qui permet de réduire la complexité des modèles

Objectif : modifier les valeurs des colonnes numériques du jeu de données pour utiliser une échelle commune, sans que les différences de plages de valeurs ne soient faussées et sans perte d'informations. Certains algorithmes ont également besoin d'une normalisation pour modéliser correctement les données.

Exemple :

Le jeu de données d'entrée contient une colonne avec des valeurs allant de 0 à 1 et une autre colonne avec des valeurs allant de 10 000 à 100 000. La grande différence d'*échelle* des nombres peut poser problème lorsque vous essayez de combiner les valeurs sous forme de fonctions lors de la modélisation.

La normalisation permet d'échapper à ce type de problème en créant de nouvelles valeurs qui conservent la même distribution générale et les mêmes ratios que les données sources tout en appliquant la même échelle aux valeurs des différentes colonnes numériques utilisées dans le modèle.

Quels sont les critères d’intégrité des données

Il est possible d’évaluer la qualité globale des données en vérifiant les éléments suivants :

- Le nombre d' enregistrements.
- Le nombre de valeurs manquantes.
- Données bien formées. (format TSV ou CSV, vérifiez que les séparateurs de colonne et de ligne séparent correctement les colonnes et les lignes.)
- Enregistrements de données incohérents. (Par exemple, si les données sont des notes moyennes d’étudiant, vérifiez qu’elles sont comprises dans la plage désignée.)

Lorsque vous détectez des problèmes dans les données, des étapes de traitement s’imposent : nettoyage des valeurs manquantes, traitement de texte pour supprimer et/ou remplacer des caractères incorporés susceptibles de perturber l’alignement des données etc etc.

Les principales opérations effectuées lors du prétraitement des données ?

- Nettoyage des données. (compléter les valeurs manquantes, détecter et supprimer les données non conformes)
- Transformation des données. (normaliser les données pour réduire le volume.)
- Réduction des données. (échantillonner les enregistrements de données pour faciliter la manipulation des données.)
- Nettoyage du texte. (notamment supprimer les caractères incorporés pouvant perturber l'alignement des données.)

Conclusion

Actuellement il y a beaucoup de tâches de préparation des données avant qu'elle soit utilisable pour l'apprentissage automatique amélioré, nous avons pu en voir quelques-une primordiale mais il en reste d'autres bien sûrs comme : gérer les valeurs manquantes(suppression ou le remplacement par une valeur factice), réduire les données(pour faciliter la manipulation ou encore ne garder qu'une partie des enregistrements qui serait quand même représentatif.), nettoyer les données textuelles(Certains champs de texte peuvent contenir des caractères qui perturbent l’alignement des données ou une manipulation inappropriée pendant l’écriture ou la lecture de textes peuvent provoquer une perte d’informations.)