RÉPUBLIQUE DU CAMEROUN
REPUBLIC OF CAMEROON
Peace – Work – Fatherland
UNIVERSITÉ DE DSCHANG
UNIVERSITY OF DSCHANG
Scholae Thesaurus Dschangensis Ibi Cordum
P 96, Dschang (Cameroun) -Tél./Fax
(237) 233 45 13 811
Website: http://www.univ-dschang.org

FACULTE DES SCIENCES
FACULTY OF SCIENCE
Département de Mathématiques et Informatique
Department of Mathematics and Computer Science
BP 67, Dschang (Cameroun)
E-mail : udsrectorat@univ-dschang.org

Master 1 : Intelligence Artificielle

IAR 418 : FOUILLE DE DONNEE

Chapitre 14: Pré-traitement et Post-traitement.

(Université de dschang)

9 avril 2025

Rédigé Par :

Noms et prénoms	Matricules
KENFACK FONGANG Victor Cyntiche *	CM-UDS-21SCI0555
MFENTAM Mohammed Salam	CM-UDS-21SCI0941
KOUOKAM TALLA Eugene Asaph	CM-UDS-21SCI0021
CHEGUEP Marcelle Fadhy	CM-UDS-21SCI0709

Sous la supervision de : **Pr KENGNE TCHENDJI Vianney.**

Année académique : **2024/2025**

Table des matières

Table des figures

1 Introduction

De nos jours nous collectons de plus en plus de grandes masses de données et nous récoltons plus que nous pouvons en traiter ; ainsi, les capacités de stockage sont en progression quasi exponentielles. Les données accumulées sont souvent brutes et loin d'être de bonnes qualités ; elles contiennent des valeurs manquantes, du bruit, et surtout des informations redondantes. La présence des valeurs manquantes et la redondance de l'information au niveau des dimensions sont des inconvénients bien souvent insurmontables pour la plupart des algorithmes de data mining. Intuitivement, l'information pertinente est noyée dans de nombreux attributs et d'entrées que son extraction n'est possible que si les données originales sont nettoyées et pré-traitées. Pour remédier à ce problème et aussi au problème de très grandes tailles des données, nous proposons une nouvelle technique de pré-traitement de données qui vise à éliminer les valeurs manquantes, les bruits, et les attributs redondants et aussi de réduire leur taille en générant des échantillons représentatifs et de qualités. Le pré-traitement et le post-traitement sont des tâches cruciales dans la découverte de connaissances dans les bases de données KDD [1]. Le KDD correspond à des processus avancés de gestion des données qui, au cours des dernières années, sont devenus très intéressants pour les chercheurs dans le domaine de l'ingénierie. Les étapes des processus KDD sont les suivantes :

1. compréhension du domaine,
2. collecte des données,
3. pré-traitement des données,
4. exploration des données,
5. et le post-traitement des connaissances dérivées

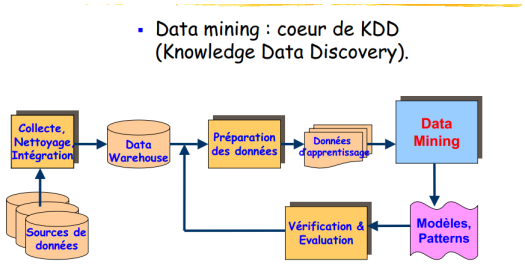


FIGURE 1 – KDD

1. KDD : Knowledge Discovery in Databases

Les données doivent être traitées par un algorithme d'acquisition de connaissances. Le problème majeur est dont que ces données sont généralement bruyantes, incomplètes et souvent incohérentes.

Ainsi, De nombreuses étapes doivent être réalisées avant l'analyse des données proprement dite. C'est la raison pour laquelle des procédures de prétraitement sont en constante évolution.

De même, un résultat provenant d'un algorithme d'apprentissage automatique en utilisant des données non traitées, tel qu'un arbre de décision, un ensemble de règles de décision ou un réseau neuronal artificiel etc., peut s'avérer être inapproprié du point de vue des clients ou pour un usage commercial. Par conséquent, une description conceptuelle (modèle, base de connaissances) produite par un tel processus inductif doit généralement être traitée a posteriori. D'où Les procédures de post-traitement. Elles comprennent généralement des routines variées d'élagage, de filtrage des règles ou même d'intégration des connaissances

Les tâches de pré et post-traitement dans l'exploration de données appliquées à un problème réel soulèvent donc l'importance de la préparation des données en raison de trois aspects :

- les données du monde réel sont très difficiles à traiter car elles sont impures ;
- les systèmes d'exploration à haute performance nécessitent des données de qualité ;
- et des données de qualité permettent d'obtenir et produisent des résultats de haute qualité.

Dans la suite, nous présentons des techniques traditionnelles et bien connues de pré-traitement et de post-traitement dans le but de montrer son importance dans l'exploration de données (DM [2]), et de renforcer le besoin d'interprétabilité des résultats lorsqu'il s'agit de données réelles. L'interprétation des résultats lors du traitement d'un ensemble de données réelles.

2. DM : Data Mining

2 Pré-traitement

2.1 Qu'est-ce que le pré-traitement des données ?

Le pré-traitement des données consiste à évaluer, filtrer, manipuler et encoder les données afin qu'un algorithme d'apprentissage automatique puisse les comprendre et exploiter les résultats. L'objectif principal du pré-traitement des données est d'éliminer les problèmes tels que les valeurs manquantes, d'améliorer la qualité des données et de les rendre utiles à l'analyse de données. il comprend les étapes que nous devons suivre pour transformer ou encoder les données afin qu'elles puissent être facilement analysées par la machine.

L'objectif principal pour qu'un modèle soit précis et exact dans ses prédictions est que l'algorithme soit capable d'interpréter facilement les caractéristiques des données.

Cette tâche prend généralement beaucoup de temps (entre 70% et 80%), surtout lorsque de nombreuses agrégations sont nécessaires sur les données. Les opérations effectuées dans le cadre d'un processus de pré-traitement peuvent être réduites à deux grandes familles de techniques :

- **Les techniques de détection (DT [3])** pour détecter les imperfections dans les ensembles de données
- **et les techniques de transformation (TT [4])** orientées vers l'obtention de données plus précises.

D'une part, Les DT comprennent la détection des valeurs aberrantes, la détection des données manquantes, détection des observations influentes, l'évaluation de la normalité, de la linéarité et de l'indépendance.

D'autre part, Les techniques de transformation (TT) visant à obtenir des ensembles de données plus faciles à gérer et comprend le traitement des valeurs aberrantes, l'amputation des données manquantes, les techniques de réduction de dimension ou de projection des données, les techniques de dérivation de nouveaux attributs, le filtrage et le ré-échantillonnage.

3. DT : Detection Technique
4. TT : Transformation Technique

En outre, les statistiques de nettoyage des données et les techniques de visualisation jouent également un rôle important dans le pré-traitement des données.

2.2 Pourquoi le pré-traitement des données ?

Les algorithmes basés sur les données sont des équations statistiques qui opèrent sur les valeurs d'une base de données. Comme le dit l'adage : « Si des données erronées entrent, des données erronées sortent. » La réussite de votre projet de données dépend des données d'entrée que vous alimentez.

Étant donné que de nombreuses personnes, processus métier et applications produisent, traitent et stockent fréquemment des données réelles, le processus est voué à devenir chaotique. Ce phénomène est généralement dû à des erreurs manuelles, des événements imprévus, des défaillances technologiques ou à plusieurs autres facteurs. Les algorithmes ne peuvent pas ingérer de données incomplètes ou bruitées, car ils ne sont généralement pas conçus pour gérer les valeurs manquantes. Or, le bruit perturbe la structure réelle de l'échantillon.

C'est pourquoi le pré-traitement des données est nécessaire pour presque tous les types d'analyse de données, de science des données et de développement d'IA afin de produire des résultats fiables, précis et résilients pour les applications d'entreprise.

De plus, La majorité des ensembles de données du monde réel pour l'apprentissage automatique sont très susceptibles d'être manquants, incohérents et bruyants en raison de leur origine hétérogène.

L'application d'algorithmes d'exploration de données à ces données bruitées ne donnerait pas de résultats de qualité, car ils ne parviendraient pas à identifier efficacement les tendances. Le pré-traitement des données est donc important pour améliorer la qualité globale des données.

Les valeurs en double ou manquantes peuvent donner une vue incorrecte des statistiques globales des données.

Les valeurs aberrantes et les points de données incohérents ont souvent tendance à perturber l'apprentissage global du modèle, conduisant à de fausses prédictions.

Les décisions de qualité doivent être fondées sur des données de qualité. Le pré-traitement des données est essentiel pour obtenir ces données de qualité, sans quoi le scénario serait un véritable échec.

2.3 Importance la préparation des données

Les algorithmes de data mining sont plus performants lorsque les données sont présentées de manière à simplifier la résolution d'un problème. Le traitement, la transformation et la réduction des données, la sélection et la mise à l'échelle des caractéristiques sont autant d'exemples d'approches de pré-traitement des données utilisées par les équipes pour réorganiser les données brutes dans un format adapté à certains algorithmes. Cela peut réduire considérablement la puissance de traitement et le temps nécessaires à l'entraînement d'un nouveau système d'apprentissage automatique ou d'IA, ou à la réalisation d'une inférence.

Il améliore la qualité des données : Le pré-traitement des données est la voie rapide vers l'amélioration de la qualité des données, car bon nombre de ses étapes reflètent les activités que vous trouverez dans tout processus de gestion de la qualité des données , telles que le nettoyage des données, le profilage des données, l'intégration des données, etc.

Il gère les données manquantes :

Il normalise et met à l'échelle les données : Plusieurs raisons peuvent expliquer l'absence de valeurs dans une collection de données (certains champs de données). Les praticiens des données doivent déterminer s'il est préférable de rejeter les enregistrements manquants, de les ignorer ou de les compléter avec une valeur estimée.

Il normalise et met à l'échelle les données : Les variables dépendantes et indépendantes évoluent à des échelles distinctes, ou bien l'une évolue linéairement tandis que l'autre évolue de manière exponentielle. Le salaire, par exemple, peut être exprimé en chiffres multiples, tandis que l'âge est exprimé en chiffres à deux chiffres. La normalisation et la mise à l'échelle permettent de modifier les données de manière à permettre aux ordinateurs d'établir un lien significatif entre ces variables.

Il élimine les doublons d'enregistrements : Lorsque deux enregistrements semblent se répéter, un algorithme doit déterminer si la même

mesure a été saisie deux fois ou si les données reflètent des occurrences distinctes. Dans de rares cas, un enregistrement peut présenter des divergences mineures dues à un champ erroné. Des techniques de recherche, de suppression ou de connexion des doublons permettent de résoudre automatiquement ces problèmes de qualité des données.

Il gère les valeurs aberrantes : Les spécialistes des données doivent parfois fusionner plusieurs sources de données pour construire un nouveau modèle. L'analyse en composantes principales, par exemple, est une technique importante pour réduire le nombre de dimensions dans l'ensemble de données d'entraînement et produire une représentation plus efficace.

contribue à améliorer les performances du modèle : Le prétraitement implique souvent le développement de nouvelles fonctionnalités ou la modification de fonctionnalités existantes afin de mieux cerner le problème sous-jacent et d'améliorer les performances du modèle. Cela peut inclure le codage de variables de catégorie, le développement de termes d'interaction et la récupération de données pertinentes à partir de textes ou d'horodatages.

2.4 étapes de pré-traitement des données

2.4.1 Acquérir l'ensemble de données

Naturellement, la collecte de données constitue la première étape de tout projet d'apprentissage automatique et la première des étapes de prétraitement des données. La collecte de données peut sembler simple, mais c'est loin d'être le cas. La plupart des entreprises finissent par conserver leurs données en silos et les répartir entre plusieurs services, équipes et solutions numériques. Par exemple, l'équipe marketing peut avoir accès à un système CRM, mais celui-ci peut fonctionner indépendamment de la solution d'analyse web. Combiner tous les flux de données dans un stockage consolidé s'avère complexe.

2.4.2 Charger les données

une fois les donnes collecté, comme à notre habitude, on va charger nos données.

2.4.3 Importer des bibliothèques

Il est ensuite temps d'importer les bibliothèques nécessaires à votre projet d'apprentissage automatique. Une bibliothèque est un ensemble de fonctions

qu'un algorithme peut appeler et utiliser. Vous pouvez rationaliser les procédures de pré-traitement des données grâce à des outils et des frameworks qui simplifient l'organisation et l'exécution du processus. Sans certaines bibliothèques, le codage et l'optimisation de solutions simples peuvent prendre des heures.

2.4.4 Importer des ensembles de données

L'étape clé suivante consiste à charger les données qui seront utilisées. Il s'agit de l'étape de pré-traitement la plus critique. De nombreuses entreprises commencent par stocker leurs données dans des entrepôts, qui nécessitent un ETL [5]. Le problème avec cette méthode est qu'on ne sait jamais quelles données seront utiles à un projet. Par conséquent, les entrepôts sont couramment utilisés pour accéder aux données via des interfaces de business intelligence afin d'observer les indicateurs que nous savons devoir surveiller. Les lacs de données sont utilisés pour les données structurées et non structurées, notamment les photos, les vidéos, les enregistrements vocaux et les fichiers PDF. Cependant, même structurées, les données ne sont pas transformées avant leur stockage. Vous chargez les données telles quelles, puis vous décidez comment les utiliser et les modifier ultérieurement.

2.4.5 Vérifiez les valeurs manquantes

Évaluez les données et recherchez les valeurs manquantes. Ces dernières peuvent perturber les tendances réelles et entraîner des pertes de données supplémentaires lorsque des lignes et des colonnes entières sont supprimées en raison de quelques cellules manquantes dans l'ensemble de données. Si vous en découvrez, vous pouvez choisir entre deux méthodes pour résoudre ce problème :

* Supprimez la ligne entière contenant une valeur manquante. Cependant, supprimer une ligne entière augmente le risque de perte de données critiques. Cette stratégie est avantageuse si l'ensemble de données est volumineux.
* Estimez la valeur en utilisant la moyenne, la médiane ou le mode.

2.4.6 Encoder les données

Les données non numériques sont incompréhensibles pour les modules d'apprentissage automatique. Pour éviter tout problème ultérieur, il est conseillé de les organiser numériquement. La solution consiste à convertir toutes les

5. ETL : Extraction, Transformation and Loading

valeurs textuelles en valeurs numériques. L'encodage est une étape essentielle pour pouvoir faire du data mining. En effet, il permet de transformer les données texte en chiffres que la machine peut interpréter ou que la machine peut comprendre. Il existe différents types d'encodage et nous allons dès maintenant aborder les plus connus :

* **One-Hot Encoding :**Le One-Hot Encoding est à la fois la méthode la plus connu, la plus simple à réaliser, qui consite à créer un dictionnaire avec tous les mots qui apparaissent dans nos phrases nettoyées.Ce dictionnaire est en fait un tableau où chaque colonne représente un mot et chaque ligne représente une phrase. Si tel mot apparaît dans tel phrase, on met une valeur de 1 dans l'élément du tableau, sinon on met une valeur de 0. On aura donc un tableau composé uniquement de 0 et de 1. Le seul inconvénient du One-Hot Encoding c'est que l'on perd la hiérarchie, l'ordre des mots. Cela nous fait donc perdre le contexte, le sens de la phrase et en théorie cela devrait appauvrir les résultats de notre modèle.
* **Word embeddings (Encodage Hiérarchique) :**Ici on utilise un autre type d'encodage : l'encodage hiérarchique. Contrairement au One-Hot Encoding, vous vous en doutez, on garde la hiérarchie, l'ordre des mots et donc le sens de la phrase.

2.4.7 Mise à l'échelle

La mise à l'échelle est inutile pour les algorithmes non basés sur la distance (comme l'arbre de décision). Les modèles basés sur la distance, en revanche, nécessitent que toutes les caractéristiques soient mises à l'échelle.

2.4.8 Diviser l'ensemble de données en ensembles d'entraînement, d'évaluation et de validation

Il s'agit de la dernière étape du pré-traitement des données. Il est temps de diviser votre jeu de données en ensembles d'entraînement, d'évaluation et de validation. L'ensemble d'entraînement correspond aux données que vous utiliserez pour entraîner votre modèle de machine learning. L'ensemble d'évaluation analysera les données et le modèle, tandis que l'ensemble de validation le validera.

2.5 Exemples et techniques de pré-traitement des données

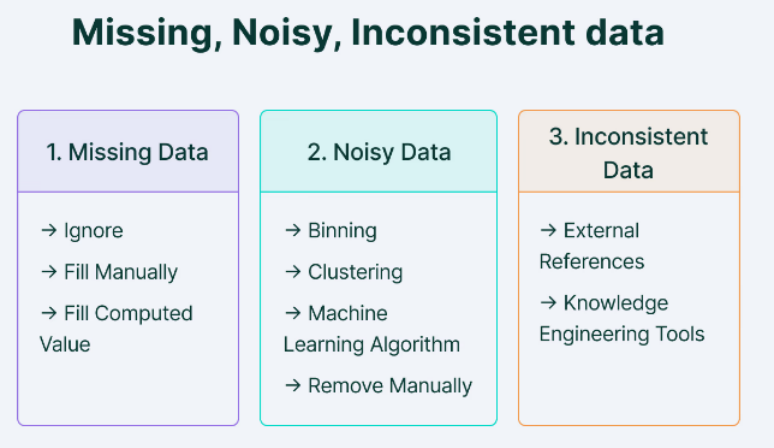
2.5.1 Nettoyage des données

Une fois que les données sont chargées il faut les nettoyer. L'objectif ici est d'identifier la solution la plus simple pour corriger les problèmes de qualité, tels que la suppression des données incorrectes, le remplissage des données manquantes ou la garantie que les données brutes sont appropriées pour l'ingénierie des fonctionnalités, en lissant les données bruyantes, en résolvant les incohérences et en supprimant les valeurs aberrantes.

1. **Valeurs manquantes :**
   Voici quelques façons de résoudre ce problème :
   * Ignorez ces tuples Cette méthode doit être envisagée lorsque l'ensemble de données est volumineux et que de nombreuses valeurs manquantes sont présentes dans un tuple.
   * Remplissez les valeurs manquantes Il existe de nombreuses méthodes pour y parvenir, comme le remplissage manuel des valeurs, la prédiction des valeurs manquantes à l'aide de la méthode de régression ou de méthodes numériques comme la moyenne des attributs, le smoothing(lissage) etc.
2. **Données bruyantes :**
   Il s'agit d'éliminer une erreur aléatoire ou une variance d'une variable mesurée. Cela peut être réalisé à l'aide des techniques suivantes :
   * Binning : Il s'agit d'une technique qui s'applique aux valeurs de données triées afin de lisser le bruit qu'elles contiennent. Les données sont divisées en compartiments de taille égale, chaque compartiment étant traité indépendamment. Toutes les données d'un segment peuvent être remplacées par leur moyenne, leur médiane ou leurs valeurs limites.
   * Régression : Cette technique d'exploration de données est généralement utilisée à des fins de prédiction. Elle permet de lisser le bruit en intégrant tous les points de données dans une fonction de régression. L'équation de régression linéaire est utilisée s'il n'y a qu'un seul attribut indépendant ; sinon, des équations polynomiales sont utilisées.
   * Regroupement : Création de groupes/clusters à partir de données ayant des valeurs similaires. Les valeurs non incluses dans le cluster peuvent être traitées comme des données bruitées et supprimées.
3. **Suppression des valeurs aberrantes :** Les techniques de clustering regroupent des points de données similaires. Les tuples situés en dehors

du cluster sont des valeurs aberrantes/incohérentes.

Garder en tête cependant que pour certains type de problèmes il peut être intéressant de préserver certains types de caractères.
Par exemple : pour analyser si un email est un spam ou non, on peut imaginer que les ' !' sont un bon indicateur et donc ne pas les enlever lors du nettoyage.



2.5.2 Transformation des données

L'une des étapes les plus importantes de la phase de préparation est la transformation des données , qui consiste à les convertir d'un format à un autre. Certains algorithmes nécessitent une modification des données d'entrée . Une fois le nettoyage des données effectué, nous devons consolider les données de qualité sous des formes alternatives en modifiant la valeur, la structure ou le format des données à l'aide des stratégies de transformation des données mentionnées ci-dessous.

* **Réduction des données :** Les collections de données brutes contiennent souvent des données en double résultant de diverses méthodes de définition d'événements, ainsi que du matériel qui ne fonctionne tout simplement pas pour votre architecture d'apprentissage automatique ou la portée de votre projet. Les techniques de réduction des données, telles que l'analyse en composantes principales, sont utilisées pour convertir les données brutes dans un format simplifié adapté à certains cas d'utilisation.

  1. **Agrégation de cubes de données :**Il s'agit d'une méthode de réduction des données, dans laquelle les données collectées sont exprimées sous une forme résumée.
  2. **Réduction de dimension :**Les techniques de réduction de dimension permettent d'extraire des caractéristiques. La dimension

d'un jeu de données désigne les attributs ou les caractéristiques individuelles des données. Cette technique vise à réduire le nombre de caractéristiques redondantes prises en compte dans les algorithmes d'apprentissage automatique. La réduction de dimension peut être réalisée à l'aide de techniques telles que l'analyse en composantes principales.

3. **Compression des données :**L'utilisation de technologies de codage permet de réduire considérablement la taille des données. Cependant, la compression des données peut être avec ou sans perte. Si les données originales peuvent être récupérées après reconstruction à partir des données compressées, on parle de réduction sans perte ; dans le cas contraire, on parle de réduction avec perte.
4. **Discrétisation :** La discrétisation des données permet de diviser les attributs continus en données avec intervalles. Cette méthode est utilisée car les caractéristiques continues ont tendance à avoir une plus faible probabilité de corrélation avec la variable cible. Par conséquent, l'interprétation des résultats peut être plus complexe. Après discrétisation d'une variable, les groupes correspondant à la cible peuvent être interprétés. Par exemple, l'âge d'un attribut peut être discrétisé en tranches telles que : moins de 18 ans, 18-44 ans, 44-60 ans et plus de 60 ans.
5. **Réduction de la numérotation :**Réduction de la numérotation Les données peuvent être représentées sous forme de modèle ou d'équation, comme un modèle de régression. Cela permettrait d'éviter de stocker d'énormes ensembles de données plutôt qu'un modèle.
6. **Sélection de sous-ensembles d'attributs :** Il est essentiel de sélectionner les attributs avec précision. Dans le cas contraire, des données de grande dimension pourraient être générées, difficiles à entraîner en raison de problèmes de sous-apprentissage ou de surapprentissage. Seuls les attributs apportant une valeur ajoutée à l'entraînement du modèle doivent être pris en compte, les autres pouvant être ignorés.
7. **Normalisation :**Il s'agit de la technique de transformation de données la plus répandue. Les attributs numériques sont agrandis ou réduits pour s'adapter à une plage spécifiée. Dans cette approche, nous limitons notre attribut de données à un conteneur spécifique afin de développer une corrélation entre différents points de données. La normalisation peut s'effectuer de plusieurs manières, présentées ici :

— Normalisation min-max
— Normalisation du score Z
— Normalisation de l'échelle décimale

8. **Généralisation :**Les données de bas niveau ou granulaires sont converties en informations de haut niveau grâce à des hiérarchies de concepts. Nous pouvons transformer les données primitives de l'adresse, comme la ville, en informations de niveau supérieur, comme le pays.
9. **Agrégation :**Il s'agit d'une méthode de stockage et de présentation de données sous forme synthétique. Par exemple, pour les ventes, les données peuvent être agrégées et transformées pour être affichées au format mensuel ou annuel.
10. **Intégration des données :**L'intégration des données est l'une des étapes de pré-traitement des données utilisées pour fusionner les données présentes dans plusieurs sources dans un seul magasin de données plus grand comme un entrepôt de données.

L'intégration des données est particulièrement nécessaire lorsqu'il s'agit de résoudre un scénario concret comme la détection de nodules à partir d'images de tomodensitométrie. La seule solution consiste à intégrer les images de plusieurs nœuds médicaux pour constituer une base de données plus vaste.

Nous pourrions rencontrer certains problèmes lors de l'adoption de l'intégration des données comme l'une des étapes de prétraitement des données :

— Intégration de schéma et correspondance d'objets : les données peuvent être présentes dans différents formats et attributs qui peuvent entraîner des difficultés dans l'intégration des données.
— Suppression des attributs redondants de toutes les sources de données.
— Détection et résolution des conflits de valeurs de données : La taille de l'ensemble de données dans un entrepôt de données peut être trop importante pour être gérée par les algorithmes d'analyse et d'exploration de données. Une solution possible consiste à obtenir une représentation réduite de l'ensemble de données, beaucoup plus petite en volume mais produisant la même qualité de résultats analytiques

2.5.3 Ingénierie des fonctionnalités

La stratégie d'ingénierie des caractéristiques vise à produire de meilleures caractéristiques pour votre jeu de données, ce qui optimisera les performances du modèle. Nous utilisons principalement les connaissances du domaine pour produire ces caractéristiques, que nous générons manuellement à partir de caractéristiques existantes après leur avoir appliqué une transformation.

Voici quelques exemples simples pour vous aider à comprendre cela : Imaginez que vos données contiennent une caractéristique de couleur de cheveux avec des valeurs de brun, noir ou inconnue. Dans ce cas, vous pouvez ajouter une nouvelle colonne intitulée « a une couleur » et attribuer la valeur 1 si la couleur est présente et 0 si la valeur est inconnue.

Un autre exemple est la dé-construction d'une caractéristique date/heure, qui fournit des informations importantes, mais est difficile à exploiter dans un modèle dans son format d'origine. Ainsi, si vous pensez que votre problème implique des dépendances temporelles et que vous découvrez un lien entre la date/heure et la variable de sortie, consacrez du temps à essayer de transformer cette colonne date/heure en une caractéristique plus intelligible pour votre modèle, telle que « période de la journée », « jour de la semaine », etc.

2.5.4 Données déséquilibrées

L'un des problèmes les plus courants que vous pouvez rencontrer lorsque vous travaillez avec la catégorisation de données du monde réel est que les classes sont déséquilibrées (l'une contient plus d'échantillons que l'autre), ce qui entraîne un biais important pour le modèle. Imaginez que vous souhaitiez prédire si une transaction est frauduleuse. D'après vos données d'entraînement, 95% de votre ensemble de données est constitué d'enregistrements de transactions légitimes, tandis que seulement 5% sont des transactions frauduleuses. Sur cette base, votre modèle prédira probablement la classe majoritaire, identifiant les transactions frauduleuses comme d'habitude. Pour résoudre cette faiblesse dans l'ensemble de données, vous pouvez utiliser trois techniques :

* Sur-échantillonnage – Le sur-échantillonnage est une technique qui consiste à enrichir un ensemble de données avec des données générées par la classe minoritaire. La technique de sur-échantillonnage synthétique des minorités (SMOTE) est la méthode la plus couramment

utilisée ; elle sélectionne un échantillon aléatoire de la classe minoritaire

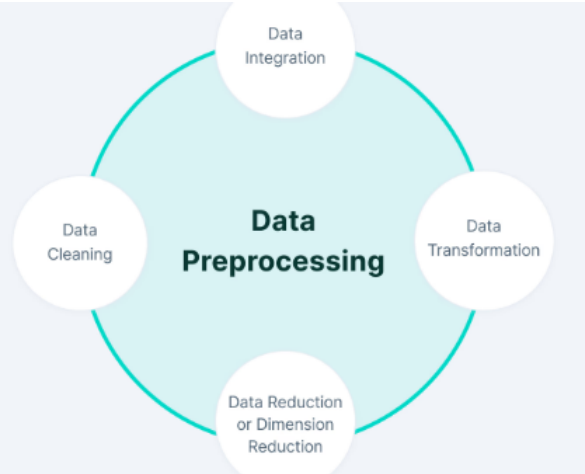
* Sous-échantillonnage – Le sous-échantillonnage consiste à réduire un ensemble de données et à éliminer les données identiques de la classe majoritaire. Les deux principaux algorithmes utilisés dans cette méthode sont TomekLinks , qui élimine les observations basées sur le plus proche voisin, et Edited Nearest Neighbors (ENN).
* Sur-échantillonnage hybride – La stratégie hybride intègre à la fois le sur-échantillonnage et le sous-échantillonnage dans votre jeu de données. L'une des méthodes utilisées est SMOTEENN , qui utilise l'algorithme SMOTE pour le sur-échantillonnage minoritaire et l'algorithme ENN pour le sous-échantillonnage majoritaire.

2.5.5 Données d'échantillonnage

Plus vous disposez de données, plus le modèle est précis. Cependant, certains algorithmes d'apprentissage automatique peuvent avoir du mal à gérer une grande quantité de données, ce qui peut entraîner des problèmes tels que la saturation de la mémoire, une augmentation des calculs nécessaires à la mise à jour des paramètres du modèle, etc.

2.6 Meilleures pratiques de pré-traitement des données

1. **Enrichissement des données :** À cette étape, les spécialistes des données utilisent diverses bibliothèques d'ingénierie de caractéristiques pour apporter les modifications nécessaires. Le résultat final devrait être un ensemble de données organisé de manière à trouver le meilleur équilibre entre le temps d'apprentissage d'un nouveau modèle et les besoins de calcul avec par exemple.
2. **Validation des données :** La validation des données commence par leur séparation en deux ensembles. Le premier sert à entraîner un algorithme de machine learning ou de deep learning. Le second sert de données de test et permet d'évaluer l'exactitude et la robustesse du modèle final. Cette deuxième étape permet d'identifier les problèmes liés aux hypothèses utilisées pour le nettoyage des données et l'ingénierie des caractéristiques. Si l'équipe est satisfaite des résultats, elle peut confier le pré-traitement à un ingénieur de données, qui choisira comment le mettre à l'échelle pour la production. Dans le cas contraire, les spécialistes des données peuvent revenir en arrière et ajuster leurs procédures de nettoyage des données et d'ingénierie des fonctionnalités.



3 Post-traitement

3.1 Qu'est-ce que le post-traitement ?

Le post-traitement est défini comme le traitement des résultats d'un modèle après son exécution. Il est utile pour appliquer des contraintes d'équité sans nécessairement modifier le modèle.
Par exemple, on pourrait post-traiter un classificateur binaire en définissant un seuil de classification pour garantir que l'égalité des chances pour un attribut donné est maintenue en vérifiant si les taux de vrais positifs sont les mêmes quel que soit l'attribut en question.
Le post-traitement des données est une étape essentielle en fouille de données qui vise à exploiter efficacement les résultats obtenus après l'application des algorithmes de data mining. Il permet d'affiner ou évaluer la performance des modèles et identifier les limites, de visualiser et d'interpréter les résultats de manière compréhensible et enfin d'exploiter les connaissances extraites (résultats) pour garantir des décisions pertinentes. Cette étape suit généralement trois phases principales : l'évaluation des modèles, la visualisation des résultats et l'exploitation des analyses pour la prise de décision.

3.2 Importance du post-traitement

La découverte de connaissances dans les bases de données (KDD) est devenue une discipline très attractive, tant pour la recherche que pour l'industrie, ces dernières années. Son objectif est d'extraire des « morceaux » de connaissances de bases de données généralement très volumineuses. Elle décrit une séquence robuste de procédures à exécuter pour obtenir des résultats

raisonnables et compréhensibles.

Selon notre compréhension, la découverte de connaissances désigne le processus global de détermination de connaissances utiles à partir de bases de données, c'est-à-dire l'extraction de connaissances de haut niveau à partir de données de bas niveau dans le contexte de grandes bases de données. La découverte de connaissances peut être considérée comme une activité multidisciplinaire car elle exploite plusieurs disciplines de recherche en intelligence artificielle telles que l'apprentissage automatique, la reconnaissance de formes, les systèmes experts, l'acquisition de connaissances, ainsi que des disciplines mathématiques telles que les statistiques, la théorie de l'information et le traitement des incertitudes.

Les composants de post-traitement en KDD peuvent être classés dans les groupes suivants :

— filtrage des connaissances ;
— interprétation et explication ;
— évaluation ;
— et intégration des connaissances.

Dans le cas d'algorithmes d'apprentissage automatique tels que les arbres ou les règles de décision formées à partir de données bruitées, les résultats sont générés à partir de peu de données d'apprentissage. Ceci est dû au fait que les algorithmes d'induction tentent de subdiviser l'ensemble des données d'apprentissage. Pour surmonter ce problème, les arbres de décision ou les règles doivent être réduits, soit par élagage (arbres de décision), soit par troncature (règles de décision). Après avoir obtenu de nouvelles connaissances, celles-ci peuvent être soit implémentées dans un système expert, soit utilisées par un utilisateur final.

Dans ce dernier cas, les résultats de la connaissance doivent être documentés pour permettre l'interprétation par l'utilisateur final. Une autre possibilité est d'afficher les connaissances et de les transformer en une forme compréhensible pour l'utilisateur final. Nous pouvons également vérifier que les nouvelles connaissances ne sont pas en conflit avec les connaissances induites précédemment.

1. **Filtrage des connaissances :** Troncature des règles et élagage a posteriori Si les données d'apprentissage sont bruyantes, l'algorithme inductif génère des feuilles d'un arbre de décision ou des règles de décision qui couvrent un très petit nombre d'objets d'apprentissage.

Cela est dû au fait que l'algorithme inductif (d'apprentissage) tente de diviser les sous-ensembles d'objets d'apprentissage en sous-ensembles encore plus petits qui seraient véritablement cohérents.

Pour surmonter ce problème, un arbre ou un ensemble de règles de décision doit être réduit, soit par élagage (arbres de décision), soit par troncature. Post-pruning (arbres de décision) ou par troncation (règles de décision)

2. **Interprétation et explication :** Maintenant, nous pouvons utiliser les connaissances acquises directement pour la prédiction ou dans un système expert comme une base de connaissances. en tant que base de connaissances, Si le processus de découverte des connaissances est pour un utilisateur final, nous documentons généralement les résultats obtenus. Une autre possibilité consiste à visualiser les connaissances ou à les transformer en une forme compréhensible pour l'utilisateur final. En outre, nous pouvons également vérifier que les nouvelles connaissances ne sont pas en conflit avec les connaissances induites précédemment. Dans cette étape, nous pouvons également résumer les règles et les combiner avec une connaissance spécifique au domaine fournie pour la tâche donnée.
3. **Évaluation :** Après qu'un système d'apprentissage a induit des hypothèses conceptuelles (modèles) à partir de l'ensemble d'apprentissage, leur évaluation (ou test) doit avoir lieu. Il existe plusieurs critères largement utilisés à cette fin de classification, la compréhensibilité, la complexité de calcul, etc.
4.) **Intégration des connaissances :** Les systèmes traditionnels de prise de décision traditionnels dépendent d'une technique, d'une stratégie ou d'un modèle unique. Les nouveaux systèmes sophistiqués d'aide à la décision, combinent ou affinent les résultats obtenus à partir de plusieurs modèles, produits généralement par différentes méthodes. Ce processus augmente la précision et les chances de succès.

3.3 Quelques méthodes de post-traitement

3.3.1 Évaluation des modèles et interprétation des résultats

1. Un modèle de classification est un modèle qui prédit des catégories il en existe deux types d'algorithmes :
   — **Régression logistique simple :** qui est un algorithme d'apprentissage permettant de prédire une variable catégorielle en fonction d'une et une seule variable d'entrée.

— **Régression logistique multiple :** qui est un algorithme permettant de prédire une variable catégorielle en fonction de plusieurs variables d'entrées.

Pour mesurer la performance de ce modèle, on utilise les métriques de classification suivantes :

(a) **Matrice de confusion :** qui est une matrice qui croise les valeurs réelles (CP et CN) avec les valeurs de prédictions ceci en montrant la répartition des vraies et fausses prédictions (PP et PN) entre les classes et en indiquant les faux positifs (FP) et faux négatifs (FN). Elle est définie comme suit :

-	PP	PN
CP	VP	FN
CN	FP	VN

(b) **Accuracy (taux de précision globale) :** il permet de donner l'exactitude sur les classements ceci en mesurant le pourcentage des bons classements du modèle. Il est donné par :

$$Accuracy = \frac{VP + VN}{VP + VN + FP + FN}$$

(c) **Précision, recall et f1-score :**

— **Précision :** mesure la proportion ou le pourcentage de prédictions correctes parmi celles qui ont été classées comme positives par le modèle.

$$Prcision = \frac{VP}{VP + FP}$$

— **Rappel (recall) :** qui est la proportion des prédictions positives classées dans le modèle parmi les prédictions positives effectuées par le modèle.

$$Rappel|recall = \frac{VP}{VP + FN}$$

— **F1-score :** qui est l'équilibre entre précision et rappel, permet de calculer la moyenne harmonique entre la précision et le rappel.

$$\frac{2}{F1-Score} = \frac{1}{rappel} + \frac{1}{precision}$$

$$F1-score = \frac{2RappelPrecision}{Rappel + Precision}$$

(d) **Courbe ROC (Receiver Operating Characteristic) & AUC (Area Under the Curve) :** les courbes ROC qui sont utilisées pour évaluer les modèles de classification binaire et montrées le compromis entre le taux de vrais positifs et le taux de faux positifs en faisant varier le seuil de décision du modèle , et l'AUC qui mesure la surface sous la courbe roc : un AUC proche de 1 signifie un bon modèle, tandis qu'un AUC proche de 0.5 signifie un modèle aléatoire.

2. **Évaluation des modèles de régression**
Les modèles de régression prédisent une valeur dépendante ou continue, il existe deux algorithmes de ce modèle :
   - δ **Régression linéaire simple :** qui vise à prédire une variable cible (ou dépendante) à partir d'une variable d'entrée. Elle est donnée par l'équation suivante : $Y = aX + b$
   - δ **régression linéaire multiple :** qui permet de prédire une variable cible à partir de plusieurs variables indépendantes, elle est donnée par : $Y = a_0 + a_1x_1 + a_2x_2 + \cdots + a_nx_n = \sum_{i=1}^{n} a_ix_i$
   Sa performance est mesurée en utilisant les métriques de régression suivantes :
     - ϕ **MSE (Mean Squared Error) :** qui est l'erreur quadratique moyenne plus elle est faible, plus le modèle est précis. Elle est donnée par : $\sum(y-\bar{y})^2$
     - ϕ **MAE (Mean Absolute Error) :** qui est l'erreur absolue moyenne et indique l'erreur moyenne en valeurs absolues. Elle est donnée par : $MAE = \sum \mid y-\bar{y}\mid$
   - δ **Le coefficient de détermination (RMSE ou R^2) :** elle mesure la proportion de variance expliquée par le modèle. Plus il est proche de 1, mieux le modèle explique les données.

3. **Évaluation des modèles de clustering** Les modèles de clustering sont des modèles qui regroupent les données sans labels prédéfinis. Comme algorithme de ce modèle nous avons : le k-means qui est un algorithme permettant de regrouper les données en k-groupes (appelés clusters) en fonction d'une mesure de similarité eu de dissimilarité. Sa performance est mesurée par l'utilisation de métriques suivantes :
   — **Indice de silhouette ou silhouet score :** qui est une métrique d'apprentissage non supervisé permettant de superviser à quel point un individu est bien classe dans son cluster par rapport aux autres clusters. Il permet de déterminer le niveau d'homogénéité entre clusters, sa valeur est toujours comprise entre [-1,1].
   Interprétations :

     Si l'indice de silhouette est proche de 1 alors les individus sont bien classés par leurs clusters
     (a) Si elle est proche de 0 alors l'individu est en frontière entre plusieurs clusters
     (b) Si elle est négative alors l'individu est mal classé.
   — **SSE (Sum of Squared Errors) & Distorsion Intra-Cluster :** elle mesure la compacité des clusters c'est-à-dire a quel point les points d'un même cluster sont rapprochés autour du centroïde. Elle est donnée par : $\sum_{k=1}^{n} \sum_{x_i \in K} \|x_i - \mu_k\|^2$
   C_k : un cluster,
   X_i : les points appartenant au cluster C_k,
   μ_k : le centroide du cluster,
   $\|x_i - \mu_k\|^2$ est la distance quadratique entre chaque point et son centroide.
   Interprétations :

     (a) Si les points sont proches du centre du cluster alors on a une faible distorsion intra-cluster et par conséquent un meilleur clustering.
     (b) Si les points sont éloignés du centre du cluster alors on a une forte distorsion intra-cluster et par conséquent une mauvaise qualité du clustering.
   — **Davies-Bouldin Index :** Mesure la séparation entre les clusters.
   — **Calinski-Harabasz Index :** Évalue la densité des clusters par rapport à la dispersion. Ces deux méthodes sont possible avec des outils comme Scikit-learn (Python) et clusterCrit (R).

4. **Évaluation de modèles par arbres de décisions**
   Ils sont représentés sous forme d'arborescence dans laquelle les nœuds feuilles représentent les classes à prédire et un nœud interne et une arête sortante une règle de décision. On évalue sa performance par :
   (a) IDs Entropie $C_{4.5}$
   (b) CART Indice de Gini
5. **Evaluation de modèle par SVM (Support Vector Machin) ou machine de support de vecteur**
   Il permet de trouver l'hyperplan optimal de réparation d'un ensemble de points dans un ensemble de caractéristiques.
6. etc.

3.3.2 Visualisation des résultats

La visualisation des résultats est essentielle car les résultats des analyses peuvent être complexes et difficiles à interpréter sous forme brute (tableaux, valeurs numériques, métriques, etc. Elle consiste à représenter graphiquement les données et les résultats des algorithmes de data mining afin de mieux :

- ϕ Faciliter l'interprétation des données c'est-à-dire comprendre la structure et les relations entre les variables.
- ϕ Évaluer la qualité des modèles c'est-à-dire identifier les erreurs, biais et performances des algorithmes.
- ϕ communiquer les résultats efficacement : permettre aux décideurs de prendre de décisions basées sur les analyses.
- ϕ etc.

1. **Types de visualisation selon le type d'analyse**
   La visualisation des données permet d'explorer, d'analyser et de présenter l'information de manière claire et intuitive. Voici les principaux types de visualisation, classés selon leur usage dans le tableau ci-dessous :

(a) **Visualisation pour la distribution des données**

i. **Histogramme** : qui montre la fréquence des valeurs d'une variable.

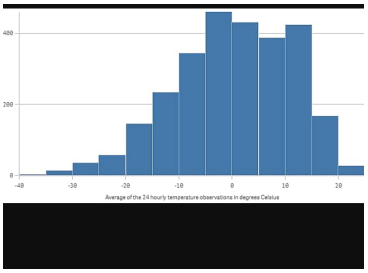


FIGURE 2 – histogramme

ii. **Boîte à moustaches (Boxplot)** : qui détecte les outliers (ou valeurs aberrantes celles qui diffèrent significativement des autres observations du jeu de données).

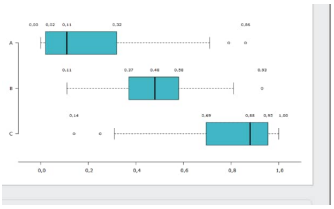


FIGURE 3 – Boxplot

iii. **Nuage de points (scatter plot) :** qui visualise la relation entre deux variables

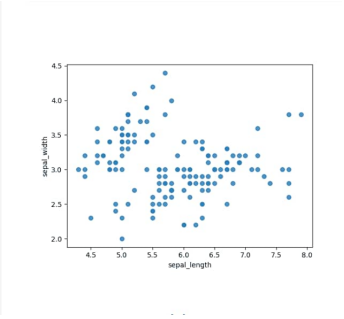


FIGURE 4 – scatter plot

(b) **Visualisation des résultats pour la classification**

i. **Matrice de confusion :** qui montre les erreurs de classification (faux positifs, faux négatifs).

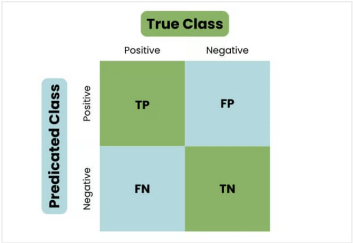


FIGURE 5 – Matrice de confusion

ii. **Courbe ROC & AUC :** qui évalue la capacité du modèle à distinguer les classes.

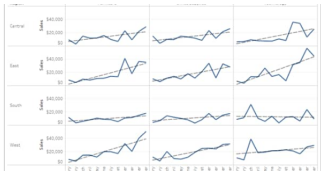


FIGURE 6 – Courbe ROC & AUC

iii. **Diagramme en barres :** pour montrer la répartition des prédictions par classe.

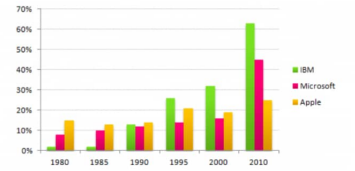


FIGURE 7 – Diagramme en barres

(c) **Visualisation pour la régression**

i. **Courbe de tendance :** qui montre la relation entre la variable cible et les prédicteurs.

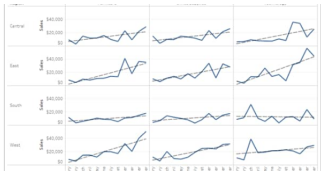


FIGURE 8 – Courbe de tendance

ii. **Résidus vs valeurs prédites** : qui identifie si le modèle a un biais ou une variance excessive.

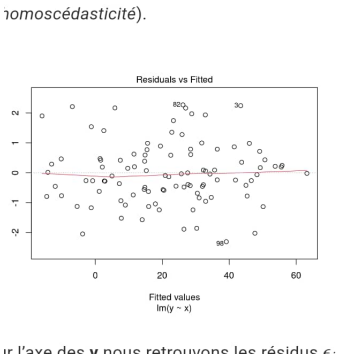


FIGURE 9 – Résidus vs valeurs prédites

(d) **Visualisation pour le clustering**

i. **Nuage de points coloré par cluster :** permettant de voir si les groupes sont bien distincts.

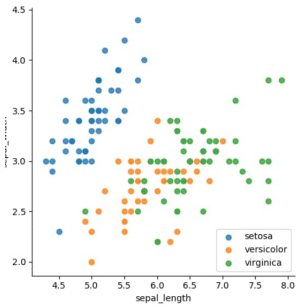


FIGURE 10 – Nuage de points coloré par cluster

ii. **Indice de silhouette** : qui mesure la qualité du clustering.

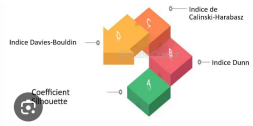


FIGURE 11 – Indice de silhouette

(e) **Visualisation pour les corrélations et relations entre variables**

i. **Heatmap (carte de chaleur)** : qui montre les corrélations entre variables.

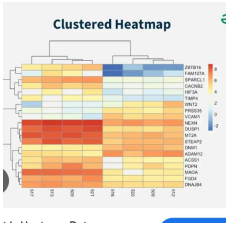


FIGURE 12 – Heatmap (carte de chaleur)

ii. **Graphes de réseau** : qui sont utile pour analyser les relations entre éléments

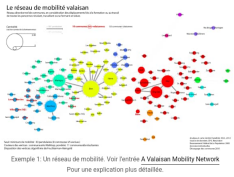


FIGURE 13 – Graphes de réseau

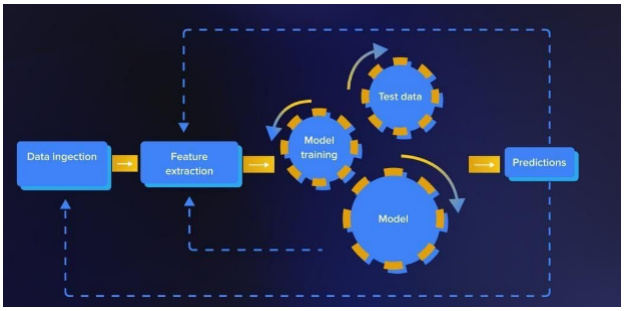


FIGURE 14 – récapitulatif

4 Conclusion

Le pré-traitement [1] des données est essentiel aux premières phases de l'extraction de données. Dans le domaine de l'IA [6], le pré-traitement améliore la qualité des données [2] en les nettoyant, en les transformant et en les formatant afin d'accroître la précision d'un nouveau modèle tout en minimisant la quantité de calculs nécessaires tandis que le post-traitement [5] consiste à raffiner, au monitoring et à optimiser notre modèle pour une amélioration de performance.efficaces. Ainsi, le pré-traitement et le post-traitement ne sont pas de simples étapes techniques : ils constituent des leviers stratégiques pour exploiter pleinement le potentiel des données et répondre aux défis du monde numérique. sans un bon post-traitement, même les modèles les plus avancés peuvent produire des résultats difficiles à exploiter et à interpréter, réduisant ainsi leur impact dans la prise de décision. De ce fait, avec l'ère du numérique où les décisions basées sur les données [3] influencent des domaines aussi

6. IA : Intelligence artificielle

variés que la finance, la santé, la sécurité et l'industrie, maîtriser ces deux étapes est un impératif. L'enjeu est donc de développer des méthodes et des outils performants permettant d'optimiser chaque phase du cycle d'analyse, afin de garantir des décisions éclairées et efficaces. Ainsi, le pré-traitement et le post-traitement ne sont pas de simples étapes techniques : ils constituent des leviers stratégiques pour exploiter[4] pleinement le potentiel des données et répondre aux défis du monde numérique.

Références

[1] Grignola ANTOINE. *La Couche de Prétraitement : Fondamentaux en Data Science*. 2025 DataBird. 2025. URL : https://www.data-bird.co/blog/couche-de-pre-traitement.

[2] Grignola ANTOINE. *Tout savoir sur le traitement des données*. 2025 Talend. 2005 - 2025. URL : https://www.talend.com/fr/resources/what-is-data-processing/.

[3] Talend DATA FABRIC. *Liste et explication des 16 techniques de forage de données*. 2005 - 2025 Talend, Inc., 2025. URL : https://www.talend.com/fr/resources/data-mining-techniques/.

[4] Studocu STUDOCU. *Liste et explication des 16 techniques de forage de données*. 2025 StudeerSnel B.V, 2025. URL : https://www.studocu.com/row/document/universite-djillali-liabes-de-sidi-bel-abbes/data-mining/m1-data-mining-exam/42265090.

[5] RELX TM. *Postprocessing*. 2025 serlier. 2025. URL : https://www.data-bird.co/blog/couche-de-pre-traitement.

5 Fiche de TD

Exercice 1

Pourquoi et comment traiter les valeurs manquantes. Expliquer en utilisant des exemples les méthodes traitées au cours et TP.
Pour diverses raisons (oublis, erreurs, non disponibilité, non communiqué,...) des valeurs sont manquantes (absentes), Il parait judicieux de les remplacer par des valeurs estimées pour permettre aux algorithmes d'apprentissage automatique de mieux fonctionner et d'améliorer leurs performances. Sans oublier que certains algorithmes ne fonctionnent pas si le dataset contient des valeurs manquantes.

Réponses :

1. Supprimer tout simplement l'instance contenant des valeurs manquantes
2. Considérer la valeur manquante comme une valeur à part
3. Remplacer par la valeur (moyenne, médiane, mode) des valeurs des attributs de la même classe
4. Remplacer par la valeur la plus fréquente.
5. Faire une prédiction de la valeur manquante en construisant un modèle à l'aide d'un algorithme d'apprentissage KNN ou autres.

Exercice 2

Supposez que l'on veuille développer un modèle de prédiction du succès d'un smartphone. Expliquer en détail comment on procédera dans l'approche data mining. Quel est l'intérêt d'une telle approche ? Proposer un bon résumé de fichier arff pour, selon vous, résoudre ce problème.

Réponses : Créer une base de données en faisant une recherche sur le net ou ailleurs. Puisque la classe est succès, la base contiendra un max d'exemples positifs et négatifs. Chaque smartphone sur le marché devient un exemple. Chaque caractéristique deviendra une valeur d'attribut possible. L'ensemble des caractéristique sera maximum défini et fixé dès le départ.
Intérêt : Seulement à partir d'exemples (positifs et négatifs), des algorithmes dédiés sont capable la connaissance descriptive du concept « smartphone à succès ». Aucun autre apport n'est nécessaire. D'où l'intérêt de cette science

Exercice 3

Donner le pourquoi et le principe des méthodes ensemblistes. Comparer les algorithmes bagging et boosting (vus au cours) étape par étape (apprentissage et classification), donner les similarités et les différences. Expliquez les paramètres de chacun. Quel est le meilleur et pourquoi ?
Réponses :

Pourquoi et principe : Faire Travailler ensemble plusieurs modèles de performances moyennes d'une même application développés sur un même dataset modifiés à chaque fois, ou avec différentes algorithmes d'apprentissage automatique. Dans l'espoir de voir l'ensemble prédire, classifier, clusteriser avec plus d'efficacité que le modèle unique.

BAGGING	BOOSTING
1. Un seul dataset initial 2. Générer n dataset exactement dès le départ (parallèlisable) 3. Appliquer le même algorithme sur tout les datasets (parallélisable) 4. Prédit la classe en utilisant tout les modeles crées. 5. La classe a plus fréquente en cas de classification. La moyenne des réponses en cas de régression	1. Un seul dataset initial 2. N ou moins datasets seront générés les uns après les autres par repondération des instances mal classées par les modèles précédents. 3. Appliquer le même algorithme de manière sequentielle. 4. Calculer l'erreur sur le dataset pondéré. 5. S'arrêter si l'erreur est nulle ou sup »rieur à 0.5 6. Reponderer les instances en utilisant cette erreur avant d'appliquer de nouveau l'algorithme. 7. Normaliser les poids. 8. Prédit la classe en utilisant tout les modeles crées.

Les paramètres sont l'algorithme et le nombre de modèles à combiner. Le meilleur et pourquoi : Chaque méta algorithme fonctionnera de manière optimale, si des conditions précises sont vérifiées. Pour le bagging, par exemple, on aura besoin d'un algorithme instable, sans oublier que des instances seront repetés ans le dataset construit, alors qu'ils ne le sont pas dans le dataset original. Pour le boosting, par exemple, les instances seront pondérés de manière particulière, d'où la nécessité pour l'algorithme de les gérer.

Exercice 4 :

Selon la kappa statistic, lequel des deux modèles de classification est le plus performant :

		Classe Prédite		
		a	b	c
Classe réelle	a	45	11	9
	b	13	15	6
	c	10	3	32

TABLE 1 – Matrice de confusion Modèle 1

		Classe Prédite		
		a	b	c
Classe réelle	a	47	8	10
	b	5	16	13
	c	5	5	35

TABLE 2 – Matrice de confusion Modèle 2

réponse : $KappaM1 = 0.43$
$KappaM2 = 0.50$
Le meilleur modèle est M2

Exercice 5 :

Selon la MAE (mean absolut error), lequel des deux modèles de régression est le plus performant :

Valeurs réelles A_i	12.3	6.5	10.7	7.9	11.5	13.7	12.6	9.4
M1 Valeurs prédites P_i	11.7	7.8	10.9	9	10.8	13.2	11.6	9.3
M2 Valeurs prédites P_i	13	6.2	11.2	8.5	12.1	12.9	12.5	9.2

TABLE 3 – données récoltées

réponse : $MAE1 = 5.5/8 = 0.6875$
$MAE2 = 3.8/8 = 0.475$
Le meilleur modèle est M2