Introduction au Data Mining

DNJOMOU YONMBA WILFRIED LOIC CM-UDS-24SCI0999
KENFACK LEONEL CM-UDS-17SCI0998
SAGUEU WAKAM DILANE CM-UDS-24SCI1040

Mars 2024

Plan

Plan

Introduction

Le **Data Mining** permet d'extraire des informations exploitables à partir de vastes ensembles de **données**.
Son développement est motivé par l'explosion des données et l'essor des capacités de calcul et il trouve des applications dans la santé, la finance, le marketing, la cybersécurité et l'agriculture.
Durant notre présentation, nous aborderons les principes, techniques, outils et défis du Data Mining

C'est quoi le data mining

Le **Data Mining**, ou exploration de données ou encore forage de données, est le processus d'analyse de grands volumes de données du big data pour découvrir des modèles et des tendances cachés. Il utilise des techniques sophistiquées issues de la statistique, et de l'intelligence artificielle pour analyser en profondeur des données sous différents angles afin de tirer des informations utiles. Ces informations peuvent ensuite être utilisées par les entreprises pour augmenter un chiffre d'affaires ou pour réduire des coûts.

C'est quoi une donnée

Dans le contexte du data mining, **une donnée** représente une unité d'information brute collectée à partir de diverses sources. Ces données peuvent être de différentes natures, telles que des chiffres, du texte, des images ou des enregistrements audio.
En data mining, les données sont classées en trois catégories principales : structurées, semi-structurées et non structurées.

Données Structurées

Les données structurées sont organisées selon un format prédéfini, généralement sous forme de tableaux avec des lignes et des colonnes. Chaque champ possède un type de données spécifique, facilitant le stockage, la recherche et l'analyse.

Exemples :

- Bases de données relationnelles contenant des informations clients (nom, adresse, numéro de téléphone).
- Feuilles de calcul avec des données financières.

Données non structurées

Les données non structurées n'ont pas de format ou de modèle prédéfini, ce qui les rend plus complexes à collecter, traiter et analyser. Elles sont caractérisées par une absence de structure formelle rendant l'analyse plus complexe.
Exemples :

- Documents texte tels que des rapports ou des articles.
- Contenus multimédias comme des images, vidéos et fichiers audio.
- Publications sur les réseaux sociaux ou messages instantanés.

Données semi-structurées

Les données semi-structurées ne suivent pas un schéma strict, mais possèdent des balises ou des marqueurs pour séparer les éléments de données. Elles combinent des aspects des données structurées et non structurées, offrant une certaine organisation sans la rigidité des bases de données relationnelles

Exemples :

- Fichiers XML ou JSON utilisés pour échanger des données entre systèmes.
- Emails contenant des métadonnées (expéditeur, destinataire) et du contenu libre.

Nature des variables

Une **variable** peut contenir deux type de données :
Une données qualitative exprime une qualité, c'est-à-dire un statut unique et est de nature discrète : les valeurs peuvent être listées et répétées pour plusieurs enregistrements. On distingue les variables qualitative ordinale et nominale. Une variable quantitative contient des valeurs numériques - qui peuvent être mesurées ou agrégées (que l'on peut quantifier). Nous retrouvons comme exemple la taille, le poids, le revenu, l'âge, la température,... Une variable quantitative est également dissociée en deux sous-catégories : les variables quantitative proportionnelle et par intervalle.

Enjeux du data mining

- **Prédire et anticiper les tendances :** une entreprise qui commercialise des produits peut utiliser le Data mining pour prédire les produits qui seront les plus populaires lors de certaines saisons ou événements.
- **Optimisation des processus internes :** une entreprise de fabrication peut utiliser l'exploration de données pour identifier les goulots d'étranglement dans sa chaîne d'approvisionnement et les résoudre pour améliorer l'efficacité globale.
- **Segmentation de la clientèle et personnalisation :** permet aux entreprises de diviser leur base de clients en segments homogènes en fonction de diverses caractéristiques telles que le comportement d'achat, les préférences ou la démographie

Enjeux du data mining

- **Détection de fraudes et de risques :** En analysant les schémas de comportement et les anomalies dans les données, le Data mining peut aider les entreprises à détecter les activités frauduleuses ou à haut risque.
- **Mesure de l'efficacité des campagnes marketing :** Les entreprises évaluent l'impact réel de leurs campagnes marketing. Cela permet d'ajuster les stratégies marketing pour des résultats optimaux.
- **Développement de nouveaux produits et services :** Le Data Mining offre d'identifier les besoins non satisfaits des clients en analysant leurs préférences et leurs comportements. Cette connaissance précieuse peut orienter le développement de nouveaux produits et services adaptés au marché.

Relation entre Big data et data mining

Le Big Data concerne le stockage et le traitement de grands ensembles de données, tandis que le Data Mining vise à en extraire des modèles intéressants.
Le Data Mining fait partie du Big Data, mais toutes les tâches du Big Data ne relèvent pas du Data Mining (ex : indexation). De même, certaines analyses de Data Mining peuvent être réalisées sur de petits fichiers sans nécessiter d'infrastructures massives.

Relation entre data mining et statistique

Le **Data Mining** et la **statistique** sont complémentaires pour analyser et extraire des informations pertinentes à partir de données. La statistique fournit des méthodes rigoureuses pour identifier des tendances et modéliser des relations, garantissant la validité des résultats. Des techniques comme la régression, l'analyse factorielle et la classification sont essentielles pour structurer et interpréter les données dans divers domaines comme la finance, la santé ou le marketing.

Relation entre data mining et machine learning

Le **Data Mining** consiste à analyser de grands ensembles de données pour identifier des modèles cachés et extraire des informations utiles, souvent à l'aide de techniques statistiques.
Le **Machine Learning**, quant à lui, est une branche de l'intelligence artificielle qui permet aux systèmes d'apprendre à partir des données et de faire des prédictions sans programmation explicite.
Bien que distincts, ces deux domaines sont complémentaires : le Data Mining permet d'extraire des connaissances exploitables, tandis que le Machine Learning utilise ces informations pour entraîner des modèles capables d'automatiser des tâches et d'améliorer la prise de décision.

Relation entre data mining et Business Intelligence

La Business Intelligence (BI) transforme les données brutes en informations exploitables pour améliorer la prise de décision. Elle repose sur l'analyse descriptive à travers des tableaux de bord et des rapports, offrant une vue d'ensemble des performances passées et actuelles.
En revanche, le Data Mining identifie des modèles cachés pour prédire des tendances futures. Intégré à la BI, il enrichit les analyses en fournissant des insights prédictifs, aidant ainsi les entreprises à anticiper et optimiser leurs décisions stratégiques.

Le processus de découverte de connaissances (KDD)

La découverte de connaissances dans les bases de données (KDD : Knowledge Discovery in Databases), désigne l'extraction non triviale d'informations implicites, jusqu'alors inconnues et potentiellement utiles, à partir de données stockées dans des bases de données.

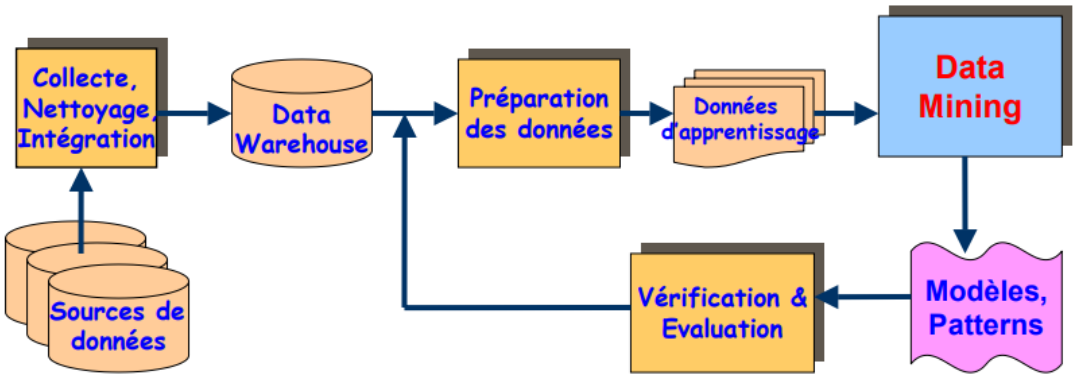


Figure – Data mining : coeur de KDD.

Le processus de découverte de connaissances (KDD)

Le processus de découverte de connaissances dans les bases de données comprend plusieurs étapes, allant de la collecte de données brutes à la production de nouvelles connaissances. Ce processus itératif comprend les étapes suivantes.

1. **Nettoyage des données**
2. **Intégration des données**
3. **Sélection des données**
4. **Transformation des données**
5. **Data Mining**
6. **Évaluation des modèles**
7. **Représentation des connaissances**

la méthodologie CRISP-DM

CRISP-DM, qui signifie Cross-Industry Standard Process for Data Mining, est une méthode mise à l'épreuve sur le terrain permettant d'orienter les travaux d'exploration de données.

- En tant que **méthodologie**, CRISP-DM comprend des descriptions des phases typiques d'un projet et des tâches comprises dans chaque phase, et une explication des relations entre ces tâches.
- En tant que **modèle de processus**, CRISP-DM offre un aperçu du cycle de vie de l'exploration de données.

la méthodologie CRISP-DM

1. **La compréhension du besoin client**
2. **La compréhension des données**
3. **La préparation des données**
4. **La modélisation ou modeling**
5. **L'évaluation**
6. **Le déploiement**

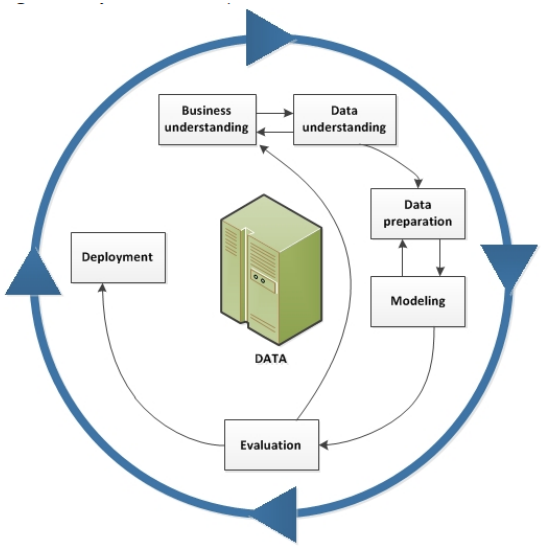


Figure – Le cycle de vie de l'exploration des données.

Nettoyage et prétraitement des données

Avant d'appliquer des techniques de Data Mining, il est crucial de nettoyer et de préparer les données. Cela implique **le traitement des valeurs manquantes, la normalisation des données, la suppression des doublons et la transformation des données non structurées** en un format exploitable.

Apprentissage Supervisé

Définition Elle repose sur des ensembles de données étiquetées,

c'est-à-dire que chaque donnée d'entrée est associée à une sortie connue (comme une catégorie ou une valeur), permettant à l'algorithme d'apprendre à prédire ces sorties pour de nouvelles données.

- À partir des données $\{(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}, i = 1, \ldots, N\}$, où x_i représente les caractéristiques d'un exemple (comme des symptômes) et y_i son étiquette (comme un diagnostic), estimer les dépendances entre $\mathcal{X}$ et $\mathcal{Y}$.
- On parle d'apprentissage **supervisé** car les y_i permettent de guider le processus d'estimation.

Apprentissage Supervisé

Exemple

- Prédire le risque d'infarctus à partir de données sur l'alimentation et l'âge d'un patient. Ici, x_i correspond à d attributs concernant le régime d'un patient, et y_i à sa catégorie (risque ou pas de risque).
- Détecter des fraudes bancaires en analysant des transactions étiquetées comme frauduleuses ou non.

Techniques

L'apprentissage supervisé peut être divisé en deux types de problèmes : la **classification** (ex. machines à vecteurs de support, arbres de décision, réseaux de neurones) et la **régression** (ex. régression linéaire, régression logistique).

Apprentissage Non Supervisé

Définition

L'apprentissage **non supervisé** utilise des algorithmes pour analyser des données non étiquetées et identifier des modèles ou regroupements sans aucune indication préalable sur les résultats attendus. Ces algorithmes découvrent des structures cachées dans les données sans intervention humaine, d'où le terme ≪ non supervisé ≫.

À partir des données $\{x_i \in \mathcal{X}, i = 1, \ldots, N\}$, décrire l'organisation des données, extraire des sous-ensembles homogènes ou explorer des structures cachées.

Apprentissage Non Supervisé

Exemple

- Regrouper les clients d'un supermarché selon leurs habitudes d'achat. Ici, x_i représente un individu (adresse, âge, habitudes de courses, etc.).
- Détecter des anomalies dans des données industrielles, comme des défauts de fabrication.
- Applications : segmentation de marchés, catégorisation de documents, compression d'images, etc.

Techniques

Les techniques incluent le **clustering** (ex. k-means, clustering hiérarchique), l'**association** (ex. Apriori pour les règles d'association) et la **réduction de dimensionnalité** (ex. analyse en composantes principales - PCA).

Types d'apprentissage - Semi-Supervisé

Définition

L'apprentissage **semi-supervisé** vise à exploiter un petit ensemble de données étiquetées $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ et un grand ensemble de données non étiquetées $\{x_{n+1}, \ldots, x_N\}$, notamment lorsque l'étiquetage est coûteux ou chronophage. L'objectif est similaire à celui de l'apprentissage supervisé, mais en tirant parti des données non étiquetées pour améliorer la performance.

Utiliser les données étiquetées pour guider l'apprentissage tout en exploitant les données non étiquetées pour mieux comprendre la structure des données.

Classification

Principe général

- Objectif : apprendre un modèle à partir de données étiquetées pour prédire la classe d'objets inconnus.
- Fonctionnement : le modèle est entraîné sur un ensemble de données où chaque exemple est associé à une classe, puis il est utilisé pour classer de nouvelles données.

La classification utilise un ensemble D de données, chaque donnée étant un vecteur d'attributs $x = \langle x_1, x_2, \ldots, x_m, y \rangle$ avec y comme attribut de classe.
L'objectif est d'entraîner un algorithme de classification A sur S pour approximer une fonction $f(x) = y$. La fonction approximée Cl est le classifieur.
L'évaluation de Cl se fait sur un ensemble de test T indépendant

Classification

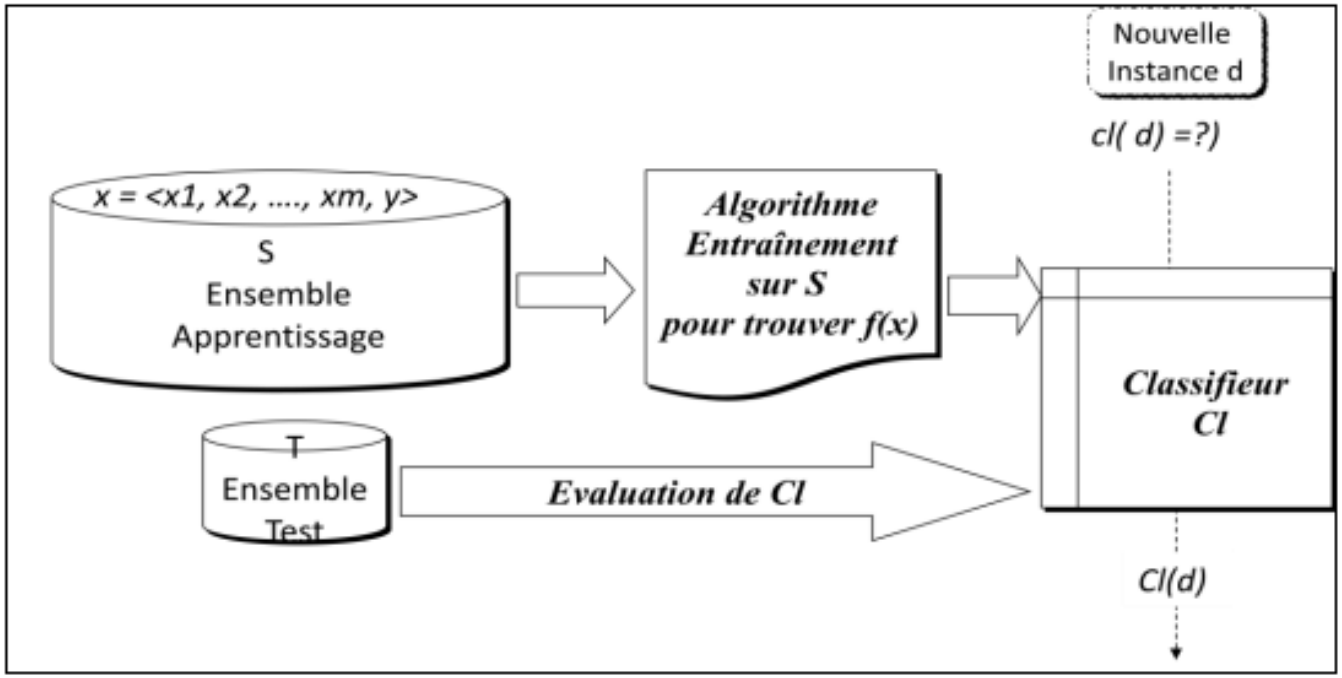


Figure – Schéma général de la tâche de classification

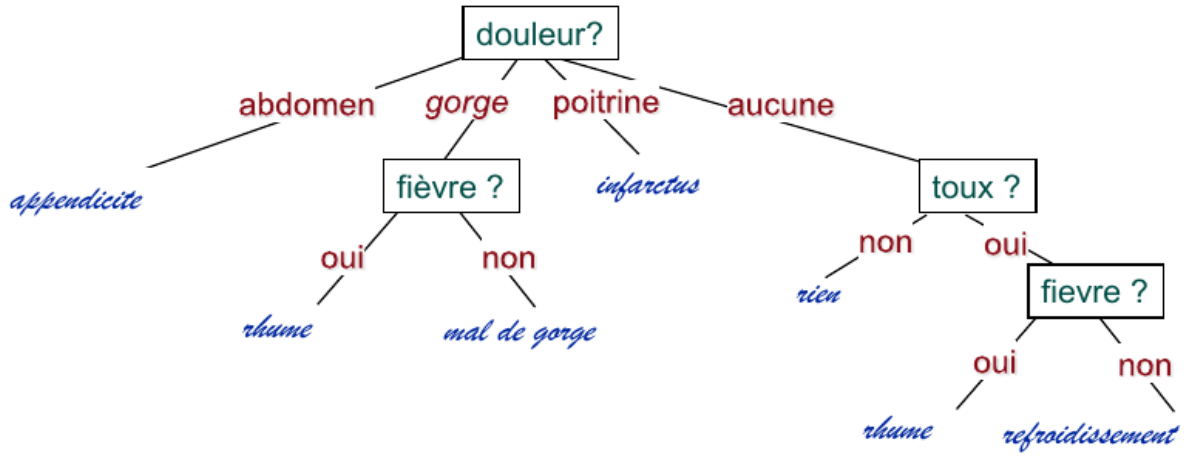
Classification

Exemples

- Détection de spam : classer les emails comme « spam » ou « non spam ».
- Diagnostic médical : prédire si un patient a une maladie spécifique en fonction de ses symptômes.

Techniques

- **Arbres de Décision**



Classification

- **Machines à Vecteurs de Support**
  La machine à vecteurs de support (SVM, pour "Support Vector Machine" en anglais) vise à trouver l'hyperplan optimal qui sépare au mieux les différentes classes de données dans l'espace des caractéristiques. Cet hyperplan est choisi de manière à maximiser la marge entre les deux classes les plus proches de l'hyperplan, ces données étant appelées vecteurs de support.

  Exemples d'application :
  Reconnaissance de visages : Les SVM peuvent être utilisés pour identifier les personnes à partir de caractéristiques extraites de leurs visages, ce qui est utile dans les systèmes de sécurité et de surveillance.

Classification

- **Réseaux de Neurones**
  Les réseaux de neurones sont des modèles d'apprentissage automatique inspirés du cerveau humain, composés de couches de neurones interconnectés pour traiter des données et apprendre des patterns complexes.

  Exemples d'application : Reconnaissance d'images : Les réseaux de neurones convolutifs (CNN) sont largement utilisés pour classer des images dans des catégories, telles que la reconnaissance faciale ou l'identification d'objets.

Clustering (Segmentation)

Le **clustering** (ou regroupement) vise à regrouper des objets similaires en clusters, sans utiliser de classes prédéfinies. Il est utilisé pour découvrir des structures ou des motifs cachés dans des données non étiquetées.

Principe général

- Objectif : identifier des groupes naturels dans les données, où les objets d'un même cluster sont plus similaires entre eux que ceux des autres clusters.
- Fonctionnement : les algorithmes de clustering analysent les similitudes entre les objets et les regroupent en conséquence.

Clustering (Segmentation)

Le problème de segmentation optimale en k groupes est NP-complet. Il faut donc chercher des algorithmes calculant une bonne décomposition, sans espérer être sûr de trouver la meilleure. On distingue essentiellement deux types de méthodes ≪non hiérarchiques≫ ou ≪hiérarchques≫. L'outils de base dans cette tâche est la mesure de similarité entre les données.

Techniques

- Approche hiérarchique (Exemple : ≪k-means≫)
- Approche non hiérarchique

Clustering (Segmentation)

Exemples

- Segmentation de marché : regrouper les clients en fonction de leurs comportements d'achat.
- Analyse de données sociales : identifier des communautés dans les réseaux sociaux.
- Applications : compression d'images, recommandation de contenu.

Régression

La **régression** est une tâche du data mining qui vise à prédire une valeur numérique continue en fonction d'une ou plusieurs variables prédictives. Contrairement à la classification, qui prédit des étiquettes de classe catégoriques, la régression modélise des fonctions à valeurs continues. Elle est particulièrement utile lorsque les variables prédictives sont également continues.

- Objectif : apprendre un modèle à partir de données étiquetées pour prédire une valeur numérique pour de nouvelles données.
- Fonctionnement : le modèle est entraîné sur un ensemble de données où chaque exemple est associé à une valeur cible continue, puis il est utilisé pour prédire cette valeur pour de nouvelles instances.

Régression

La régression cherche à modéliser la relation entre une variable dépendante y (la variable à prédire) et une ou plusieurs variables indépendantes $x_1, x_2, \ldots, x_n$ (les prédicteurs). Le modèle de régression le plus simple est la régression linéaire, où la relation est supposée linéaire.

Techniques

- **Régression linéaire** : modélise la relation entre les variables par une ligne droite. Pour un seul prédicteur, elle prend la forme $y = w_0 + w_1 x$

Régression

Exemple : Prédiction du salaire : estimer le salaire d'un employé en fonction de son expérience professionnelle.

x years experience	*y salary (in $1000s)*
3	30
8	57
9	64
13	72
3	36
6	43
11	59
21	90
1	20
16	83

Figure – Table Data

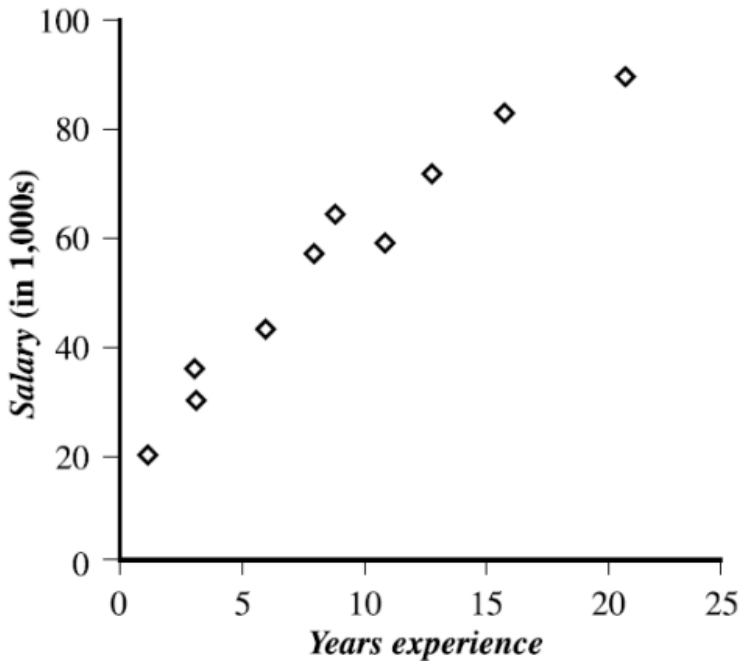


Figure – 2-D data on a scatter plot.

Régression

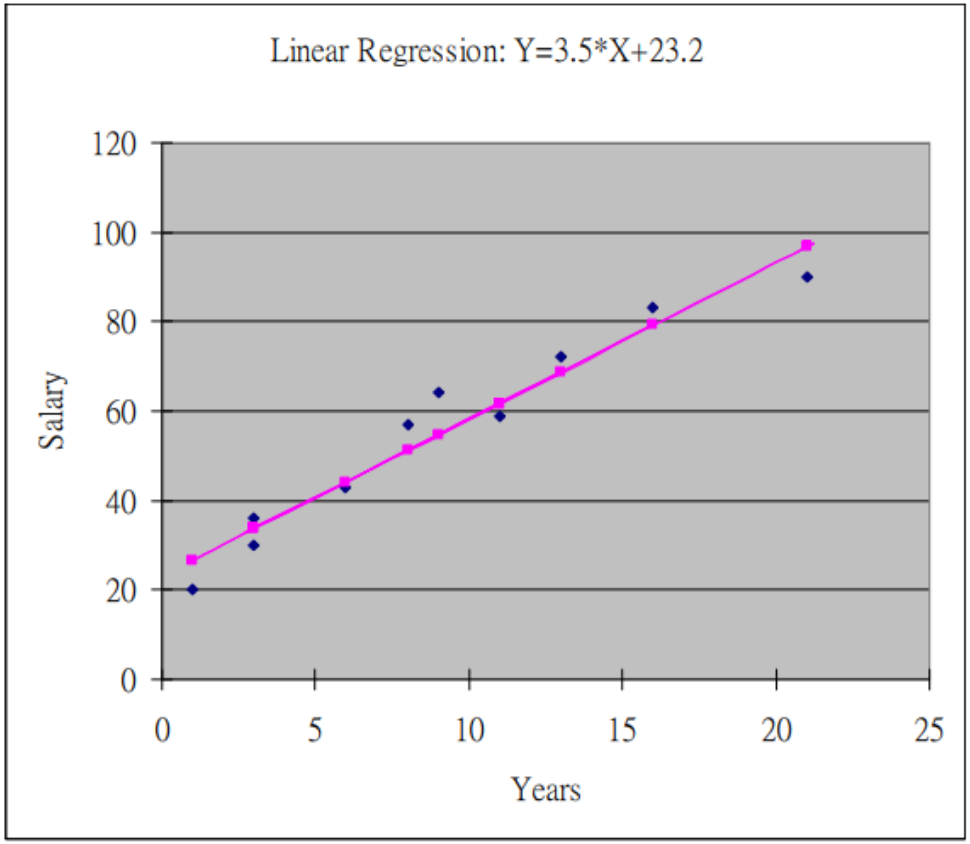


Figure – Exemple de régression linéaire

Régression

- **Régression linéaire multiple** : La régression linéaire multiple implique l'utilisation de plusieurs variables prédictives avec des données de la forme :

$$(\mathbf{X}_1, y_1),\ (\mathbf{X}_2, y_2),\ \ldots,\ (\mathbf{X}_p, y_p)$$

où :

  - $\mathbf{X}_i \in \mathbb{R}^n$: vecteur de données de dimension n
  - y_i : étiquette de classe associée

c'extension de la régression linéaire à plusieurs prédicteurs, de la forme $y = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_n x_n$.

- **Régression non linéaire** : utilisée lorsque la relation entre les variables n'est pas linéaire. Par exemple, la régression polynomiale : $y = w_0 + w_1 x + w_2 x^2 + w_3 x^3$. Elle peut capturer des relations plus complexes mais nécessite plus de données et peut être sujette au surajustement.

Régression

- **Modèles linéaires généralisés** : incluent la régression logistique, qui modélise la probabilité d'un événement, et la régression de Poisson, pour des données de comptage. Ils étendent la régression linéaire à des distributions non normales.

Extraction de règles d'association

L'**extraction de règles d'association** est une tâche fondamentale en fouille de données (data mining), visant à identifier des relations implicites et significatives entre les attributs dans de grandes bases de données. cette technique est particulièrement reconnue pour son

application dans l'analyse du panier de marché. Elle permet, par exemple, de dégager des règles telles que « 70% des clients achetant du lait et du thé achètent aussi du pain », utiles pour optimiser l'agencement des rayons, planifier des promotions ou gérer les stocks dans un objectif d'amélioration des profits.

Extraction de règles d'association

Problématique

Le cadre classique de la fouille des règles d'association peut être décrit ainsi : soit $A = \{x_1, x_2, \ldots, x_m\}$ un ensemble de m attributs (ou items), et $E = \{e_1, e_2, \ldots, e_n\}$ un ensemble de n transactions, chaque transaction e_i étant un sous-ensemble de A. Chaque transaction est associée à un identifiant unique (TID, Transaction IDentifier). Un motif est un sous-ensemble de A, et une règle d'association est une implication $X \rightarrow Y$ (où $Y \neq \emptyset$), indiquant que les attributs de X tendent à apparaître avec ceux de Y.

Languages de programmation

1. **R** : C'est Un langage de programmation statistique avec de nombreuses bibliothèques pour la fouille de données. Un des atouts de R est la facilité avec laquelle des graphiques bien conçus, de qualité digne de publication, peuvent être produits, contenant des symboles mathématiques et des formules si besoin est
2. **Python** : C'est un *langage de programmation* polyvalent largement utilisé dans le domaine de la Data Science et du Data Mining. Il offre une vaste gamme de bibliothèques et de frameworks tels que Pandas, NumPy, scikit-learn, TensorFlow, etc., qui permettent d'effectuer diverses tâches d'analyse de données, de modélisation et de visualisation.

Outils logiciels

1. **RapidMiner** : Propose une interface conviviale pour la préparation des données, la modélisation, l'évaluation et le déploiement des modèles. Il offre une grande variété d'algorithmes de Data Mining et de fonctionnalités de traitement des données
2. **WEKA** (Waikato Environment for Knowledge Analysis) : logiciel open source qui prend en charge toutes les principales tâches d'exploration de données. Cependant, il excelle en particulier dans la classification.
3. **Orange** : logicielle open-source qui propose une interface graphique intuitive qui permet aux utilisateurs de créer et d'exécuter des workflows de traitement de données sans écrire de code

Outils logiciels

1. **KNIME** : plateforme open-source qui permet aux utilisateurs de créer des workflows visuels pour l'analyse des données et le déploiement de modèles.
2. **SAS Enterprise Miner** : offre une large gamme d'algorithmes de modélisation avancés, une interface conviviale pour la création de workflows et des fonctionnalités de déploiement de modèles.

Défis et limites du data mining

Un biais est une distorsion systématique dans l'analyse des données qui peut fausser les résultats et entraîner des décisions erronées.
Il peut apparaître à différentes étapes, de la collecte à l'interprétation des données.
Pourquoi est-ce un problème ?

- Influence négative sur la qualité des modèles prédictifs.
- Peut mener à des décisions commerciales inappropriées.
- Dans certains cas, risque de discrimination involontaire dans des domaines comme le recrutement ou le crédit.

Types de Biais Courants

- **Biais de Sélection :** Se produit lorsque l'échantillon de données n'est pas représentatif de la population cible.
  Exemple : Un algorithme bancaire entraîné uniquement sur des clients ayant déjà un crédit risque d'exclure les nouveaux emprunteurs.
- **Biais de Confirmation :** Tendance à favoriser les données qui confirment une hypothèse préexistante, en ignorant celles qui la contredisent. Les conséquences possible sont l'interprétation biaisé des résultats, mauvaise décisions basées sur des suppositions erronées
  Exemple : Une entreprise analyse uniquement les avis positifs pour mesurer la satisfaction client, ignorant les retours négatifs.

Types de Biais Courants

- **Biais de Survivance :** Se produit lorsque seules les données des cas réussis sont prises en compte, en excluant celles ayant échoué. Exemple : Étudier uniquement les startups à succès pour identifier des facteurs de réussite, sans analyser celles qui ont échoué.
- **Biais de Mesure :** Résulte d'erreurs dans la collecte ou la mesure des données, affectant la précision des analyses. Elle peut être causé par des instruments de mesure défectueux ou une incohérence dans les méthodes collecte

Stratégies pour Surmonter les Biais

Évaluation et Nettoyage des Données : Vérifier la qualité des données avant l'analyse.

- Supprimer les valeurs aberrantes et incohérentes.
- Traiter les valeurs manquantes pour éviter les distorsions.

Standardisation et Normalisation : Assurer l'uniformité des données pour éviter les erreurs d'interprétation.

- Convertir toutes les devises en une unité unique avant une analyse financière.
- Unifier les formats de date et heure dans un dataset international.