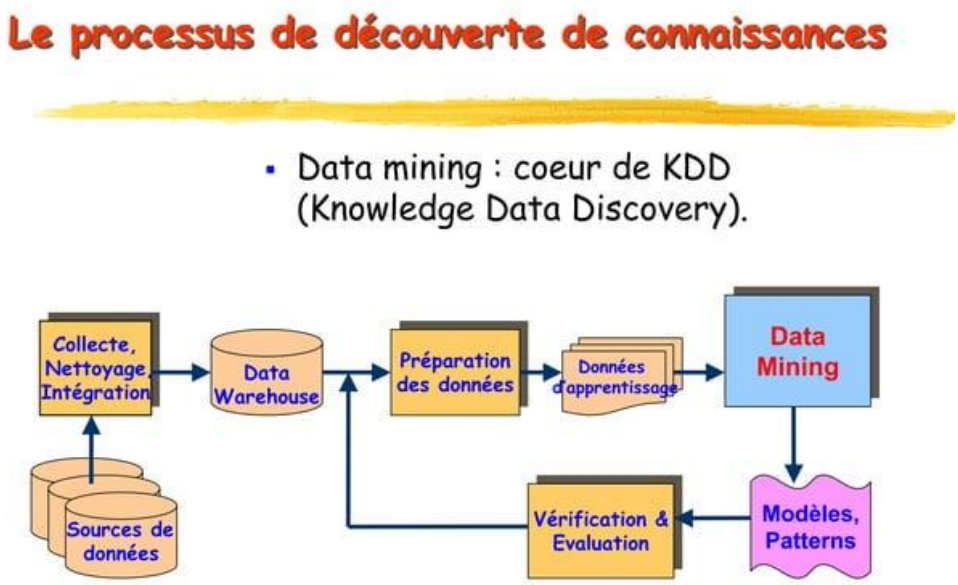
PLAN DE L'EXPOSE

INTRODUCTION

Dans un monde où les données sont de plus en plus abondantes et complexes il est devenu essentiel pour les organisations de trouver des moyens efficaces pour les analyser et les exploiter c'est dans ce contexte que le datamining est apparu comme un concept clé pour extraire des connaissances et des modèles à partir des grandes quantités de données. Le datamining également appelé fouille de donnée est un processus qui consiste à appliquer des techniques statistiques et informatiques pour découvrir des relations, des tendances et des modèles dans des ensembles de données. Ses techniques permettent aux organisations de prendre des décisions éclairées, d'améliorer leur efficacité et de créer de nouvelles opportunités de croissance. Dans cet exposé, nous alors vous présentez les processus du datamining, leurs méthodes et les différents outils du datamining. Enfin fin nous allons discuter des défis et des limites du datamining, ainsi que des perspectives futures pour ses techniques.

I. PROCESSUS DU DATA MINING

Le datamining est l'un des composants essentiels de la technologie Big Data et de technique d'analyse de données volumineuses. Il s'agit de logiciels en parti des outils analytiques pour l'analyse des données. A cet effet, le datamining ou encore l'exploration des données est un processus complexe qui implique plusieurs étapes pour l'extraction des informations utiles à partir de grande quantité de données. Nous avons identifié 10 principales étapes qui constituent ce dernier à savoir :



- **Définition des objectifs**

Ici il s'agit d'identifier clairement les objectifs de l'analyse et les questions auxquelles on souhaite répondre.

- **Collecte de données**

Elle revient à rassembler les données pertinentes à partir de différentes sources (Bases de données internes, fichiers, API, données publique).

- **Préparation de données**

C'est transformer les données afin de les rendre exploitables. Cela inclut la gestion des valeurs manquantes, la suppression des doublons, la normalisation des données et de formatage.

- **Exploration des données**

Analyser les données à l'aide d'étude statistique et de visualisation pour comprendre leur structure, leurs tendances et leurs anomalies.

- **Sélection des techniques**

Choisir la méthode approprier en fonction des objectifs définis. Cela peut inclure des techniques de classification, régression et clustering, association …etc.

- **Modélisation**

Appliquer les techniques choisis pour construire des modelés qui peuvent prédire ou classifier les données.

- **Evaluation du modelé**

Tester le modelé sur un ensemble de données distinct pour évaluer sa performance.

- **Interprétation des résultats**

Analyser le résultat obtenu et tirer des conclusions significatives par rapport aux objectifs initiaux.

- **Déploiement**

Mettre en œuvre le modelé dans un environnement opérationnel afin qu'il puisse être utilise pour prendre des décisions basées sur les données.

- **Suivi et maintenance**

Surveiller la performance du modèle au fil du temps et effectuer des mises à jour ou des ajustements si nécessaire.

II. METHODES DU DATAMINING

Intuitivement, nous pourrions penser que « l'exploration » de données fait référence à l'extraction de nouvelles données, mais ce n'est pas le cas. Le datamining consiste plutôt à extrapoler des modèles et des connaissances à partir des données que nous avons déjà recueillies. En s'appuyant sur des techniques et des technologies à l'intersection de la gestion des bases de données, des statistiques et du machine learning, les spécialistes du datamining ont consacré leurs carrières à mieux comprendre comment traiter et tirer des conclusions de grandes quantités d'informations. Mais quelles sont les techniques qu'ils utilisent pour y parvenir ?

L'exploration de données via le datamining est très efficace, pour autant qu'elle s'appuie sur une ou plusieurs de ces techniques :

1. Classification

La classification est une méthode qui consiste à prédire la catégorie à laquelle appartient un nouvel exemple, basé sur un ensemble d'exemples précédemment classés.

a. TECHNIQUES UTILISER

- Arbres de décision : Modèles qui prédisent la valeur d'une variable cible en apprenant des règles de décision à partir des attributs d'entrée.
- Forêts aléatoires : Ensemble d'arbres de décision qui améliorent la précision en réduisant le risque de sur apprentissage.
- Régression logistique : Utilisée pour des problèmes de classification binaire, prédit la probabilité d'appartenance à une classe.

b. APPLICATIONS

Cette méthode peut être utile dans le cadre de :

- Détection de fraude
- Diagnostic médical
- -Classification d'e-mails (spam ou non spam)

2. REGRESSION

La régression est une méthode utilisée pour prédire une valeur numérique continue. Elle établit une relation entre une variable dépendante et une ou plusieurs variables indépendantes.

a. LES TECHNIQUES UTILISEES

- Régression linéaire : Modèle simple qui prédit une variable continue en se basant sur une relation linéaire.
- Régression polynomiale : Extension de la régression linéaire qui peut capturer des relations non linéaires.

b. APPLICATIONS

On l'utilise généralement pour :

- Prévision des ventes
- Analyse des tendances économiques.

3. CLUSTERING

Le clustering est une technique qui regroupe des données similaires en clusters (ou groupes) sans avoir besoin de connaître à l'avance les catégories.

a. TECHNIQUES UTILISEES

Les méthodes de clustering incluent :

- K-means : Algorithme qui partitionne les données en K groupes en minimisant la variance intra-groupe.
- Algorithme de regroupement hiérarchique : Crée une hiérarchie de clusters en fusionnant ou en divisant les groupes.

b. APPLICATIONS

- Segmentation de la clientèle
- Analyse de marché.

4. ASSOCIATION

Les règles d'association sont utilisées pour découvrir des relations intéressantes entre les variables dans de grandes bases de données. L'exemple classique est l'analyse du panier d'achat, qui cherche à identifier quels produits sont souvent achetés ensemble.

a. TECHNIQUES UTILISEES

- Algorithme Apriori : Utilisé pour générer des règles d'association en identifiant les ensembles d'items fréquents.
- Algorithme FP-Growth : Une méthode plus efficace que l'Apriori pour extraire les règles d'association.

b. APPLICATIONS

- Recommandations de produits
- Marketing ciblé.

5. DETECTION D'ANOMALIES

La détection d'anomalies identifie des points de données qui diffèrent significativement du reste des données. Cela est crucial pour des applications telles que la détection de fraudes ou la surveillance de la santé des systèmes.

a. TECHNIQUES UTILISEES

- Méthodes statistiques : Utilisent des modèles statistiques pour identifier des valeurs qui se situent en dehors des normes attendues.

- Apprentissage automatique : Algorithmes tels que les réseaux de neurones ou les SVM (machines à vecteurs de support) pour détecter des anomalies.

b. APPLICATIONS

- Sécurité informatique
- Finance
- Santé

En fonction des objectifs et des types de données, les méthodes de classification, régression, clustering, association et détection d'anomalies peuvent être appliquées de manière complémentaire pour obtenir des résultats significatifs. L'utilisation judicieuse de ces techniques peut aider les entreprises à prendre des décisions éclairées et à optimiser leurs opérations.

III. OUTILS DU DATAMINING

Les principaux outils du datamining sont :

- R : Langage de programmation pour l'analyse de données et le datamining.
- Python : Langage de programmation polyvalent pour le datamining, notamment avec les bibliothèques Pandas, NumPy et Scikit-learn.
- SQL : Langage de requête pour la gestion de bases de données relationnelles.
- Tableau : Outil de visualisation de données pour créer des tableaux de bord interactifs.
- Power BI : Outil de business intelligence pour créer des rapports et des tableaux de bord interactifs.

❖ Outils d'exploration de données

- Pandas : Bibliothèque Python pour la manipulation et l'analyse de données.
- NumPy : Bibliothèque Python pour les calculs numériques.
- Matplotlib : Bibliothèque Python pour la visualisation de données.
- Seaborn : Bibliothèque Python pour la visualisation de données.
- Plotly : Bibliothèque Python pour la visualisation de données interactives.

❖ Outils de modélisation prédictive

- Scikit-learn : Bibliothèque Python pour l'apprentissage automatique.
- TensorFlow : Bibliothèque open-source pour l'apprentissage automatique.
- PyTorch : Bibliothèque open-source pour l'apprentissage automatique.
- Keras : Bibliothèque open-source pour l'apprentissage automatique.
- XGBoost: Bibliothèque pour l'apprentissage automatique.

❖ **Outils de visualisation de données**

- D3.js : Bibliothèque JavaScript pour la visualisation de données.
- Chart.js : Bibliothèque JavaScript pour la visualisation de données.
- Highcharts : Bibliothèque JavaScript pour la visualisation de données.
- QlikView : Outil de business intelligence pour la visualisation de données.
- SAS : Outil de business intelligence pour la visualisation de données.

❖ **Outils de gestion de données**

- MySQL : Système de gestion de base de données relationnelle.
- PostgreSQL : Système de gestion de base de données relationnelle.
- MongoDB : Système de gestion de base de données NoSQL.
- Hadoop : Framework pour la gestion de données massives.
- Spark : Framework pour la gestion de données massives.

❖ **Outils de préparation de données**

- Trifacta : Outil pour la préparation de données.
- Dataiku : Outil pour la préparation de données.
- Talend : Outil pour l'intégration de données.
- Informatica : Outil pour l'intégration de données.
- Microsoft Power Query : Outil pour la préparation de données.

❖ **Outils de minage de données**

- Weka : Outil pour le datamining et l'apprentissage automatique.
- Orange : Outil pour le datamining et la visualisation de données.
- RapidMiner : Outil pour le datamining et l'apprentissage automatique.
- KNIME : Outil pour le datamining et l'apprentissage automatique.
- SAS Enterprise Miner : Outil pour le datamining et l'apprentissage automatique.

CONCLUSION

En conclusion le datamining, est une technique puissante qui permet aux organisations de découvrir des connaissances être des modèles à partir de grandes quantités de données. Ses techniques ont des applications dans de nombreux domaines tels que la finance la sante, le marketing et la logistique. Grace au datamining, les organisations peuvent prendre des décisions éclairées, améliorer leur efficacité et créer de nouvelles opportunités de croissance. Cependant, il est important de noter que le datamining nécessite une approche méthodique et une bonne compréhension des techniques statistiques et informations.

TABLE DES MATIERES