ANALYSE ET VISUALISATION DES DONNÉES

Alexandra-Maria DOBRESCU

| Cours 10 |

Travailler avec les données existantes

AVD | Cours 10 | AMD

- TECHNIQUES VARIÉES
- MÉTHODES DE PRÉDICTION
- MÉTHODES DE DESCRIPTION
- CONSEILS TRAVAUX PRATIQUES

Techniques variées

LES MÉTHODES DE PRÉDICTION ET LES MÉTHODES DE DESCRIPTION SONT DEUX APPROCHES DISTINCTES UTILISÉES DANS DIFFÉRENTS DOMAINES, NOTAMMENT LA SCIENCE, L'ANALYSE DES DONNÉES ET LA PRISE DE DÉCISION.

Méthodes de prédiction

- ✓ **Objectif:** Les méthodes de prédiction visent *à prévoir* ou *à estimer* les résultats ou les tendances futurs sur la base de modèles de données existants.

- ✓ **Processus:** Ces méthodes impliquent souvent l'utilisation:
  - de modèles statistiques,
  - d'algorithmes d'apprentissage automatique,
  - d'autres techniques informatiques,

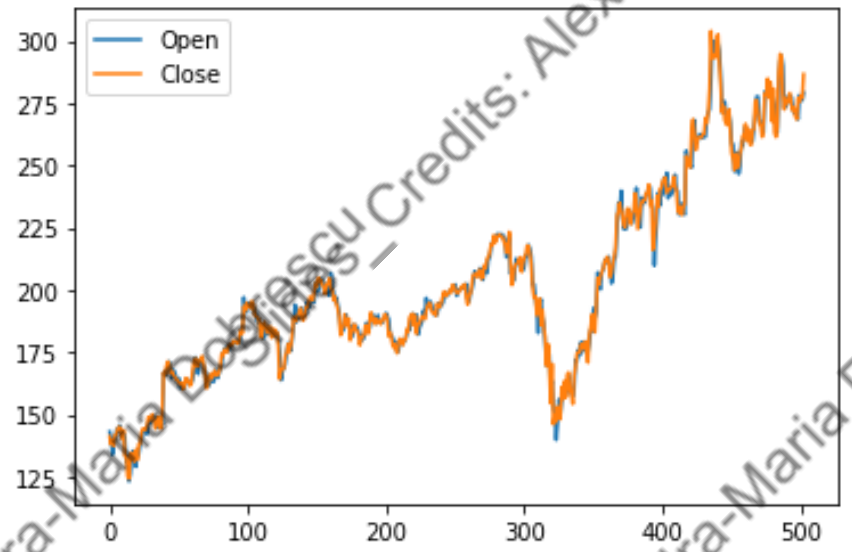
  pour identifier les relations et les modèles au sein d'un ensemble de données.

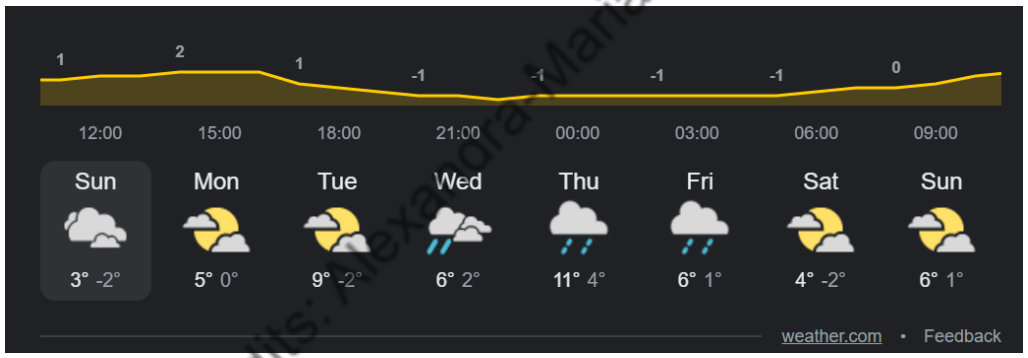
  L'objectif est de créer un modèle capable de faire des prédictions précises sur de nouvelles données inédites.

Méthodes de prédiction

- ✓ **Exemple:** Dans l'apprentissage automatique, un modèle prédictif peut être formé sur des données historiques pour prédire des valeurs futures:
  - telles que les prix des actions,
  - les conditions météorologiques,
  - le comportement des clients.

Source[3]





Source[4]

Techniques variées

LES MÉTHODES DE PRÉDICTION ET LES MÉTHODES DE DESCRIPTION SONT DEUX APPROCHES DISTINCTES UTILISÉES DANS DIFFÉRENTS DOMAINES, NOTAMMENT LA SCIENCE, L'ANALYSE DES DONNÉES ET LA PRISE DE DÉCISION.

Méthodes de description

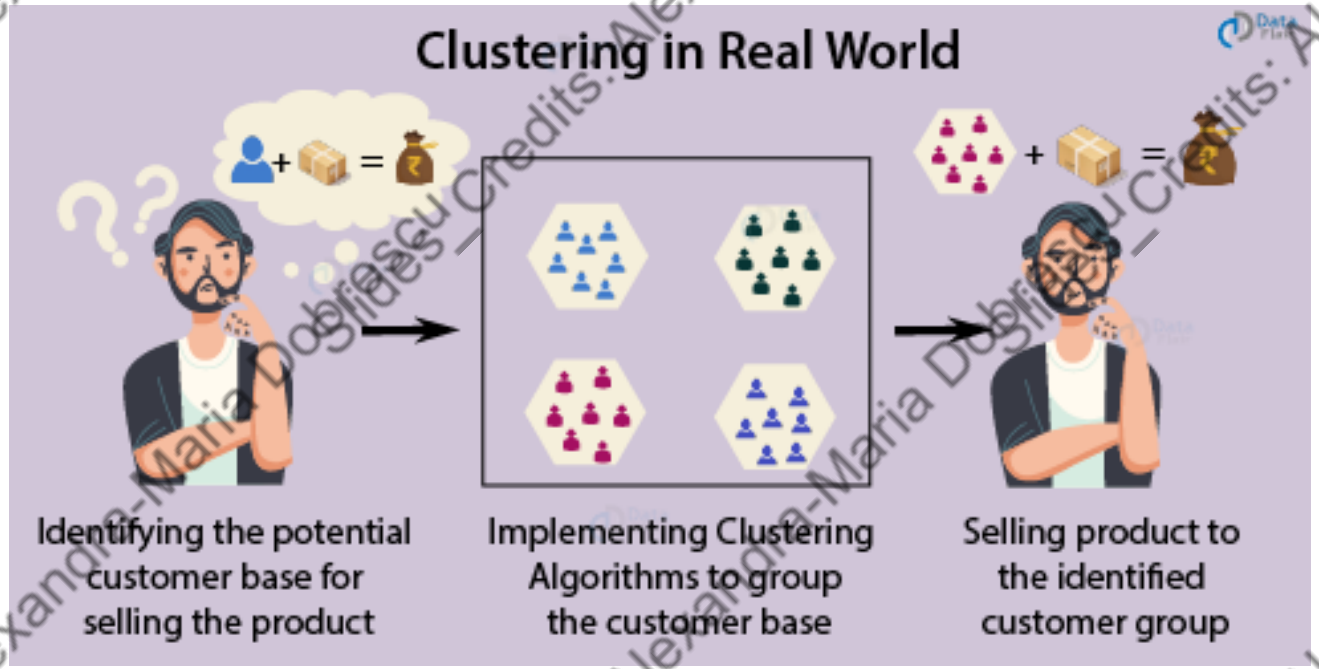
- ✓ **Objectif:** Les méthodes de description se concentrent sur *la compréhension* et *l'explication* de la structure et des caractéristiques sous-jacentes d'un système ou d'un phénomène.

- ✓ **Processus:** Ces méthodes impliquent souvent l'analyse de données afin d'identifier des modèles, des tendances et des relations.

Toutefois, l'accent est mis sur la description et l'interprétation des données plutôt que sur la prédiction des résultats futurs.

Méthodes de prédiction

- ✓ **Exemples:**
  - Dans les statistiques descriptives, les analystes peuvent utiliser des mesures telles que la moyenne, la médiane et l'écart-type pour résumer et décrire les tendances centrales et la variabilité au sein d'un ensemble de données.
  - Les algorithmes de clustering permettent de trouver des groupes d'objets similaires dans un ensemble de données (appelés clusters) et d'éventuels objets isolés, éloignés de tout cluster, appelés valeurs aberrantes.



Comparaison

Focus

Les **méthodes de prédiction** se concentrent sur *les résultats futurs* et visent à faire des *prévisions précises*.
Les **méthodes de description** se concentrent sur *la compréhension* et *l'explication* de l'état présent ou passé d'un système.

Utilisation des données

Les **méthodes de prédiction** s'appuient fortement sur des données historiques pour former des modèles permettant de faire des prédictions futures.
Les **méthodes de description** analysent les données historiques pour mieux comprendre les caractéristiques et la structure des données elles-mêmes.

Comparaison

Application

Les **méthodes de prédiction** sont souvent utilisées dans des scénarios où l'objectif est de prendre des décisions éclairées sur l'avenir en se basant sur les données disponibles.
Les **méthodes de description** sont généralement employées lorsque l'objectif principal est de comprendre les propriétés sous-jacentes des données ou du système.

Flexibilité

Les **méthodes de prédiction** peuvent donner la priorité à la précision du modèle et peuvent être moins concernées par l'interprétabilité du modèle.
Les **méthodes de description** mettent souvent l'accent sur l'interprétabilité des résultats afin de fournir des informations et une compréhension claire des données.

Comparaison

Méthodes de description ⟵ les processus d'analyse des données ⟶ la prise de décision ⟶ **Méthodes de prédiction**

Dans la pratique, les méthodes de prédiction et de description ont toutes deux leur place dans les processus d'analyse des données et de prise de décision, et le choix entre les deux dépend des objectifs et des exigences spécifiques de la tâche à accomplir.

Dans certains cas, une combinaison des deux approches peut être utilisée pour parvenir à une compréhension plus complète d'un système ou d'un phénomène.

Méthodes de description

Problème concret

Méthodes de prédiction

Prenons l'exemple de la gestion et de l'amélioration des résultats pour les patients atteints d'une maladie chronique, telle que le diabète.

- ✓ **Objectif:** Prédire la probabilité de complications ou d'exacerbations futures chez les patients diabétiques.
- ✓ **Approche:** Utiliser des algorithmes d'apprentissage automatique pour analyser les données historiques des patients, notamment les taux de glycémie, l'observance du traitement, les facteurs liés au mode de vie et les complications antérieures.

Former un modèle prédictif pour prévoir la probabilité d'événements futurs, tels que les hospitalisations ou les fluctuations graves de la glycémie.

Méthodes de description

Problème concret

Méthodes de prédiction

Combinaison

- ✓ **Intégration des informations:** Combinez le modèle prédictif avec des informations descriptives pour obtenir une compréhension globale.

Exemple:

- Si le modèle prédictif indique un risque élevé de complications pour un sous-groupe de patients, les analyses descriptives peuvent aider à expliquer les facteurs contributifs, tels que les modes de vie courants ou la non-observance des médicaments au sein de ce sous-groupe.
- *D'autres?*

Méthodes de description

Problème concret

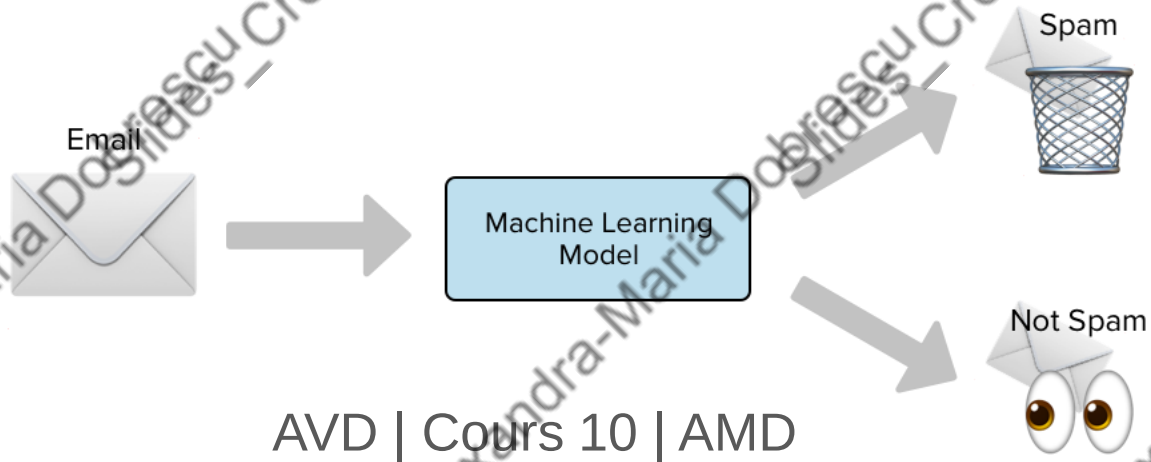
Méthodes de prédiction

Combinaison

- ✓ **Interventions personnalisées**
  - On peut utiliser l'analyse prédictive pour adapter les interventions aux patients à haut risque en fonction de la probabilité d'événements futurs.
  - Parallèlement, les résultats descriptifs peuvent être exploités pour comprendre les caractéristiques uniques de différents groupes de patients, ce qui permet d'élaborer des plans de soins personnalisés qui répondent à des besoins et à des défis spécifiques
- ✓ **Amélioration continue**
  - Utiliser une boucle de rétroaction dans laquelle les connaissances acquises par les méthodes de prédiction et de description s'enrichissent mutuellement.
  - Ajustez et affinez les modèles prédictifs sur la base des analyses descriptives en cours, en veillant à ce que les modèles restent pertinents et reflètent l'évolution de la population de patients.

Classification

- ✓ **Définition**
  La classification est une tâche d'apprentissage supervisé qui consiste à attribuer des étiquettes ou des catégories prédéfinies aux données d'entrée en fonction de leurs caractéristiques.
- ✓ **Example**
  Étant donné un ensemble de données d'e-mails, un algorithme de classification pourrait être entraîné à prédire si chaque e-mail est un « spam » ou un « non-spam ».
- ✓ **Résultat**
  La sortie d'un modèle de classification est une étiquette ou une catégorie de **classe discrète.**



Classification - Concepts clés

Classes/étiquettes	Données pour le *Training*	Limite de décision
Les catégories ou classes distinctes que l'algorithme vise à attribuer aux données d'entrée.	Données étiquetées utilisées pour former le modèle, où chaque exemple possède des caractéristiques d'entrée et des étiquettes de classe correspondantes.	Séparation entre différentes classes dans l'espace des caractéristiques, déterminée par le modèle au cours de la formation.

Classification – Mesures d'évaluation

Précision	Précision et *Recall*	Score F1
La proportion d'instances correctement classées.	Mesures de la capacité d'un modèle à identifier correctement les instances positives.	Moyenne harmonique de la précision et du rappel.

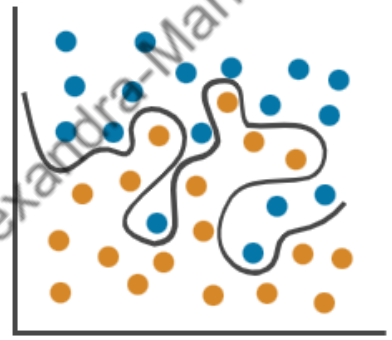
Régression

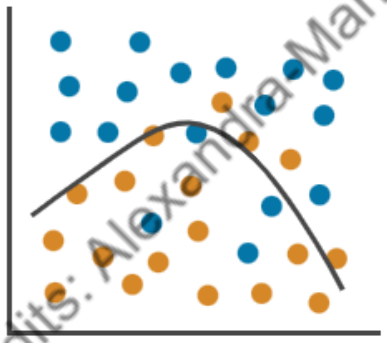
- ✓ **Définition**
  La régression est une tâche d'apprentissage supervisé qui consiste à prédire une sortie continue ou une valeur numérique sur la base de caractéristiques d'entrée.
- ✓ **Example**
  Prédire le prix d'une maison (une valeur continue) en fonction de caractéristiques telles que la surface, le nombre de chambres et l'emplacement.
- ✓ **Résultat**
  La sortie d'un modèle de régression est **une valeur numérique continue**.

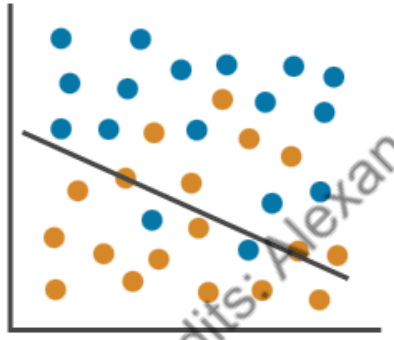
Régression - Concepts clés

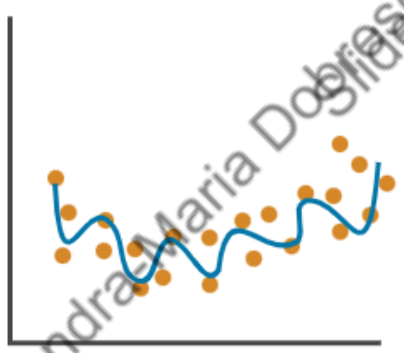
Variable dépendante	Variables indépendantes	Ligne de régression
La variable à prédire ou à expliquer, généralement représentée sur l'axe des y.	Les caractéristiques d'entrée qui influencent la variable dépendante, généralement représentées sur l'axe des x.	La ligne qui s'ajuste le mieux aux points de données, en minimisant la différence entre les valeurs prédites et les valeurs réelles.

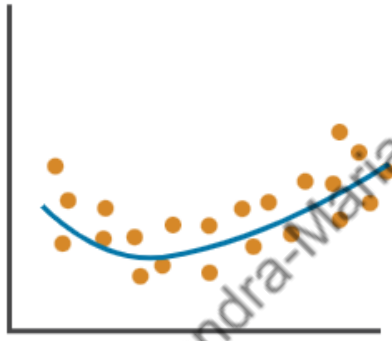
Ligne de décision

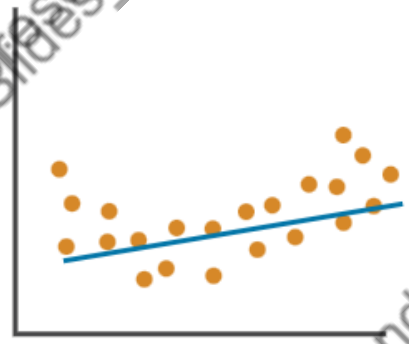






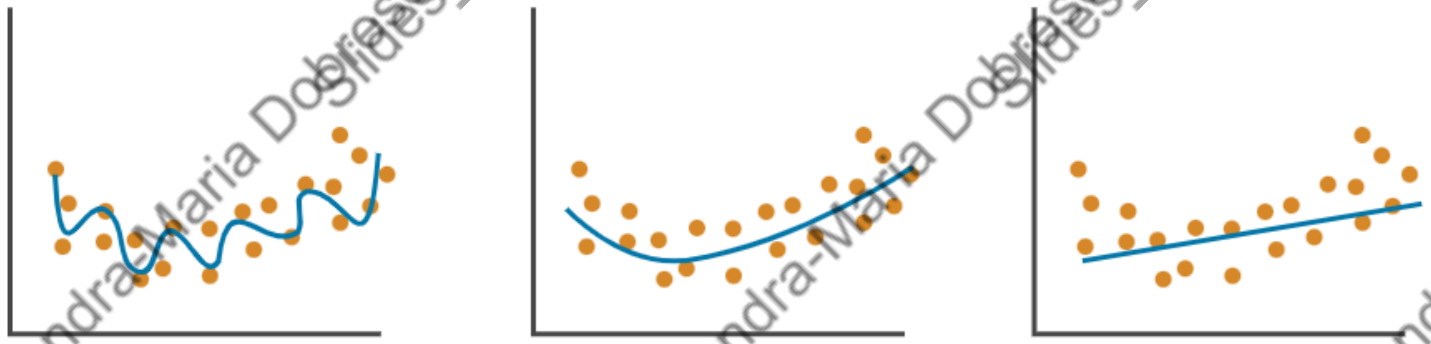






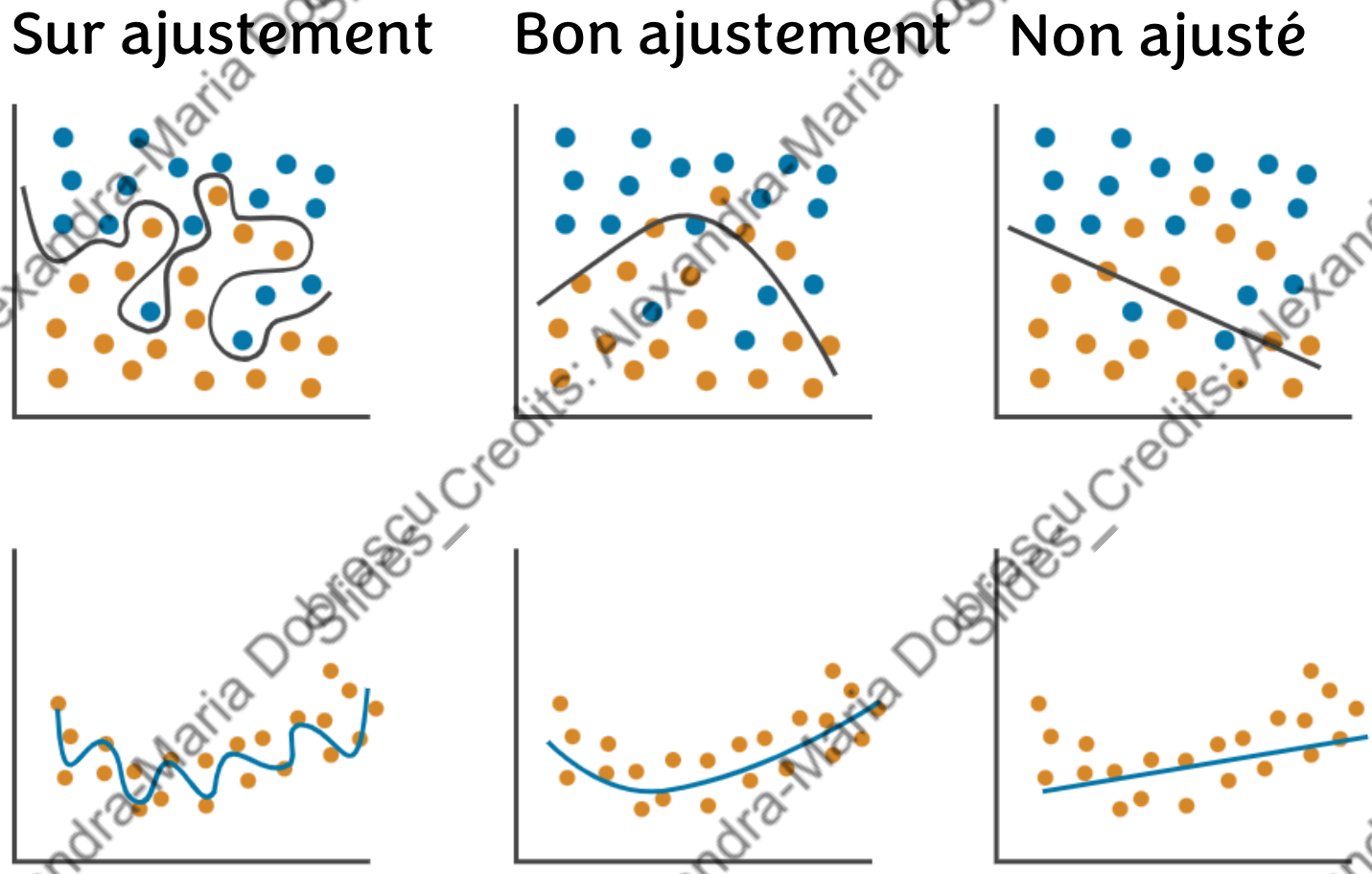
Ligne de décision

Classification



Régression

Ligne de décision



Régression - Mesures d'évaluation

Mean Squared Error (MSE)	Root Mean Squared Error (RMSE)	R au carré (R^2)
La moyenne des différences au carré entre les valeurs prédites et les valeurs réelles.	La racine carrée de l'MSE, qui fournit une mesure dans les unités d'origine de la variable dépendante.	Indique la proportion de la variance de la variable dépendante qui est prévisible à partir des variables indépendantes.

Méthodes de prédiction – Algorithmes

	Forêt aléatoire	Machines de renforcement du gradient (par exemple, XGBoost)
Type	Apprentissage d'ensemble	Apprentissage d'ensemble
Utilisation	Random Forest est un algorithme polyvalent utilisé pour la classification et la régression. Il construit plusieurs arbres de décision et combine leurs prédictions, ce qui permet d'obtenir des résultats robustes et précis tout en limitant le surajustement.	Le Gradient Boosting construit un ensemble d'apprenants faibles de manière séquentielle, chacun corrigeant les erreurs des précédents. XGBoost, une implémentation populaire, est largement utilisée pour la modélisation prédictive en raison de ses performances élevées et de son évolutivité.

Méthodes de prédiction – Algorithmes

	Réseaux de mémoire à courte terme (LSTM)	Machines à vecteurs de support (SVM) pour la classification
Type	Réseau neuronal récurrent (RNN)	Apprentissage supervisé
Utilisation	Les LSTM sont efficaces pour les tâches de prédiction de séquences, en particulier pour les prévisions de séries temporelles. Ils peuvent saisir les dépendances sur de longues séquences, ce qui les rend aptes à prédire les valeurs futures sur la base de modèles historiques.	Les SVM sont puissants pour les tâches de classification binaire et multiclasse. Ils fonctionnent bien dans les espaces à haute dimension et peuvent gérer des limites de décision complexes, ce qui les rend appropriés pour une variété de problèmes de prédiction.

Méthodes de description - Algorithmes

	Analyse en composantes principales (PCA)	Regroupement K-Means
Type	Réduction de la dimensionnalité	Apprentissage non-supervisé
Utilisation	L'PCA est utilisée pour transformer des données de haute dimension en une forme de dimension inférieure, afin de capturer les informations les plus importantes. Elle est souvent utilisée pour comprendre la structure sous-jacente des données et identifier des modèles.	K-Means est un algorithme de clustering qui regroupe les points de données similaires en clusters. Il est fréquemment utilisé pour l'analyse descriptive afin d'identifier les regroupements naturels au sein d'un ensemble de données, ce qui permet de découvrir des modèles ou de segmenter les données.

Méthodes de description - Algorithmes

	Regroupement hiérarchique	T-SNE
Type	Apprentissage non-supervisé	Réduction de la dimensionnalité
Utilisation	Cet algorithme construit une hiérarchie de clusters en les fusionnant ou en les divisant successivement en fonction de leur similarité. Il est utile pour visualiser les relations et les similitudes dans les données à différents niveaux de granularité	t-SNE est une autre technique de réduction de la dimensionnalité qui se concentre sur la visualisation de données à haute dimension en deux ou trois dimensions. Elle est souvent utilisée pour explorer la structure et les relations inhérentes aux points de données.

Conseils travaux pratiques

Démonstration et discussion en direct

Bibliographie

[1] Subramaniam, A. (2020). What Is Big Data Analytics| Big Data Analytics Tools and Trends| Edureka.
[2] Bishop, C. M., & Nasrabadi, N. M. (2006). *Pattern recognition and machine learning* (Vol. 4, No. 4, p. 738). New York: springer.
[3] https://frinkiac.com/
[4] https://towardsdatascience.com/predictive-analytics-predicting-consumer-behavior-with-data-analytics-8ca51abb8dc2