Série de révision

1.Quelle est la principale différence entre l'apprentissage supervisé et l'apprentissage non supervisé ?

a) L'apprentissage supervisé utilise des labels tandis que l'apprentissage non supervisé apprend à partir de données non étiquetées

b) L'apprentissage supervisé ne nécessite pas de données d'entraînement

c) L'apprentissage non supervisé a toujours une sortie binaire

d) L'apprentissage supervisé ne peut être utilisé que pour la classification

2.Quelle affirmation est correcte concernant le sur-apprentissage (overfitting) ?

a) Il se produit lorsque le modèle généralise bien sur les nouvelles données

b) Il signifie que le modèle ne s'adapte pas bien aux données d'entraînement

c) Il se produit lorsque le modèle est trop complexe et s'ajuste trop aux données d'entraînement

d) Il est souhaitable car il réduit l'erreur d'apprentissage

3.Quelle technique peut être utilisée pour réduire l'overfitting ?

a) Ajouter plus de couches dans le modèle

b) Réduire le volume des données d'entraînement

c) Augmenter le taux d'apprentissage du modèle

d) Utiliser la régularisation L1 ou L2

4.Dans quel domaine l'apprentissage non supervisé est-il couramment utilisé ?

a) La reconnaissance faciale

b) Le clustering de clients en marketing

c) La prédiction de prix de logements

d) La classification d'images médicales

5.Quelle affirmation décrit le mieux l'underfitting (sous-ajustement) dans un modèle de Machine Learning ?

a) Le modèle s'adapte trop bien aux données d'entraînement et ne généralise pas bien aux nouvelles données

b) Le modèle est trop simple et ne capture pas correctement les tendances des données d'entraînement

c) Le modèle atteint une précision de 100% sur les données d'entraînement

d) L'underfitting se produit uniquement lorsque l'on utilise trop de données

6.Un scientifique des données entraîne un modèle sur un jeu de données et veut s'assurer que son modèle généralise bien. Il utilise un ensemble d'entraînement (train set), un ensemble de validation (validation set) et un ensemble de test (test set). Quelle stratégie suivante est la plus adaptée pour obtenir une évaluation fiable du modèle ?

a) Utiliser l'ensemble d'entraînement pour ajuster les paramètres du modèle, l'ensemble de validation pour ajuster les hyperparamètres, et l'ensemble de test uniquement pour évaluer la performance finale sans aucun ajustement.
b) Utiliser l'ensemble d'entraînement pour ajuster les hyperparamètres et les paramètres du modèle, et l'ensemble de test pour sélectionner la meilleure version du modèle.
c) Mélanger les ensembles de validation et de test pour augmenter la quantité de données d'entraînement et maximiser la performance du modèle.
d) Utiliser le test set pour valider et ajuster les hyperparamètres du modèle, puis entraîner à nouveau sur tout le dataset pour maximiser la précision.

7.Quelle est la méthode la plus appropriée pour gérer les valeurs manquantes dans un jeu de données ?
a) Supprimer toutes les lignes contenant des valeurs manquantes
b) Remplacer les valeurs manquantes par la moyenne, médiane ou la valeur la plus fréquente
c) Remplir les valeurs manquantes avec des zéros
d) Ignorer complètement les valeurs manquantes

8.Quelle transformation est souvent nécessaire pour les algorithmes basés sur la distance, comme KNN ?
a) Standardisation des données
b) Réduction du nombre de colonnes
c) Ajout de bruit aux données
d) Suppression des valeurs aberrantes

9.Quelle méthode permet de réduire la dimensionnalité d'un jeu de données ?
a) Augmentation des échantillons d'entraînement
b) Principal Component Analysis (PCA)
c) Encodage des données catégoriques
d) Algorithme des K-means

10.Pourquoi est-il important d'effectuer une normalisation des données avant d'entraîner un modèle ?
a) Pour réduire la taille du dataset
b) Pour permettre aux modèles de convergence plus rapidement et d'éviter des biais liés aux échelles des variables
c) Pour introduire des valeurs aberrantes qui rendent le modèle plus robuste
d) Pour rendre le modèle plus complexe

11.Quelle méthode d'encodage est la plus adaptée pour une variable catégorielle ordinale ?
a) One-Hot Encoding
b) Label Encoding
c) Encodage fréquentiel
d) Binary Encoding

12.Un scientifique des données souhaite réduire la dimensionnalité d'un jeu de données en sélectionnant uniquement les caractéristiques les plus pertinentes. Il hésite entre les méthodes basées sur les filtres, les wrappers et les méthodes intégrées (embedded methods).

Quelle affirmation est correcte concernant ces approches ?

a) Les méthodes basées sur les filtres utilisent directement les performances du modèle pour sélectionner les meilleures caractéristiques.
b) Les méthodes basées sur les wrappers testent différentes combinaisons de caractéristiques en évaluant leur impact sur la performance du modèle.
c) Les méthodes intégrées (embedded methods) sont toujours plus performantes que les méthodes wrappers.

13. Un modèle de Machine Learning contient des caractéristiques fortement corrélées entre elles. Quelle est la conséquence principale de cette redondance et quelle méthode est la plus appropriée pour résoudre ce problème ?

a) La redondance n'a aucun impact, le modèle apprend mieux avec plus de variables.
b) La présence de caractéristiques redondantes peut entraîner un sur-apprentissage (overfitting) et augmenter le temps de calcul du modèle. Une solution efficace est d'utiliser la régression L1 (Lasso) pour forcer à zéro les coefficients des variables inutiles.
c) Les caractéristiques redondantes améliorent toujours la précision du modèle car elles contiennent plus d'informations sur les données.
d) La seule façon de gérer les caractéristiques redondantes est de supprimer manuellement les variables les plus corrélées.

14.Quelle métrique est la plus appropriée pour un problème de classification où les classes sont déséquilibrées ?
a) Accuracy
b) Precision-Recall Curve
c) Mean Squared Error (MSE)
d) Coefficient de corrélation

15.Que représente l'AUC-ROC ?
a) Une mesure de la précision du modèle sur l'ensemble d'entraînement
b) La courbe qui évalue la séparation entre les classes en fonction du seuil de

classification
c) Une technique de régularisation pour éviter l'overfitting
d) Un algorithme de classification

16.Dans une matrice de confusion, qu'est-ce qu'un faux négatif (FN) ?
a) Une prédiction correcte d'une classe négative
b) Une prédiction correcte d'une classe positive
c) Un cas où l'algorithme prédit négatif alors que la vraie classe est positive
d) Un cas où l'algorithme prédit positif alors que la vraie classe est négative

17.Quelle métrique combine précision et rappel dans un seul score ?
a) L'accuracy
b) Le score F1
c) La courbe ROC
d) Le nombre total de prédictions correctes

18.Un data scientist entraîne un modèle de classification binaire sur un dataset où 95% des données appartiennent à la classe négative et 5% à la classe positive. Après évaluation, il obtient une accuracy de 95%.

Que peut-on en conclure sur la performance du modèle ?

a) Le modèle est très performant car il atteint une accuracy élevée.
b) Le modèle est biaisé en faveur de la classe majoritaire et l'accuracy seule n'est pas une bonne métrique dans ce cas.
c) Une accuracy élevée signifie que le modèle a également une précision et un rappel élevés.
d) Pour améliorer la précision, il suffit d'augmenter la quantité de données dans la classe majoritaire.

19.Un modèle de classification binaire a une AUC (Area Under the Curve) de 0.75. Que peut-on en déduire ?

a) Le modèle est parfait et classe toutes les instances correctement.
b) Le modèle a une performance meilleure qu'un modèle aléatoire mais n'est pas optimal.
c) L'AUC de 0.75 signifie que le modèle a un taux de vrais positifs (TPR) de 75% et un taux de faux positifs (FPR) de 25%.
d) Une AUC de 0.75 indique que le modèle prédit correctement 75% du temps sur l'ensemble d'entraînement.

20.Pourquoi est-il important de normaliser les données avant d'utiliser KNN ?
a) Pour améliorer la vitesse d'exécution
b) Parce que KNN est basé sur des distances et est sensible aux différences d'échelle entre les variables

c) Parce que KNN ne fonctionne qu'avec des données normalisées
d) Pour éviter l'overfitting

21.Quelle distance est couramment utilisée pour KNN ?
a) Distance Euclidienne
b) Cosinus de Similarité
c) Distance de Manhattan
d) Distance de Hamming

22.Quel est l'impact de choisir un K trop grand dans KNN ?
a) Risque de sous-ajustement (underfitting)
b) Risque de sur-ajustement (overfitting)
c) Augmentation de la complexité du modèle
d) Augmentation du biais du modèle

23.Un data scientist applique une validation croisée k-fold avec k=10 pour entraîner un modèle de classification. Pourquoi cette méthode est-elle préférable à une simple division train/test ?

a) Parce qu'elle réduit le risque de sur-apprentissage en testant le modèle sur plusieurs sous-ensembles de données.
b) Parce qu'elle garantit que le modèle obtient toujours un score plus élevé qu'avec un simple train/test split.
c) Parce qu'elle augmente automatiquement la taille du dataset en créant de nouvelles données à chaque fold.
d) Parce qu'elle est plus rapide que l'entraînement sur un seul split, car elle entraîne moins de modèles.

24.Un ingénieur en machine learning veut choisir la meilleure valeur de K pour KNN en utilisant une variante de la méthode du coude (Elbow Method). Il trace la courbe de l'erreur moyenne en fonction de K et observe le comportement suivant :

- **Pour K=1, l'erreur est très faible.**
- **À mesure que K augmente, l'erreur commence à augmenter lentement.**
- **À partir de K=10, l'erreur stagne et ne diminue plus significativement.**

Que peut-on en conclure ?

a) K=1 est le meilleur choix car l'erreur est minimale, garantissant une bonne performance.
b) Le coude de la courbe (K ≈ 10) est le meilleur choix, car il représente le point où l'augmentation de K n'améliore plus la généralisation.
c) K=20 est préférable car un K plus grand améliore toujours la robustesse du modèle.

d) La méthode Elbow ne peut pas être appliquée à KNN, elle est uniquement valable pour K-Means.

25.Un data scientist veut optimiser le nombre de voisins (K) dans un modèle K-Nearest Neighbors (KNN). Il décide d'utiliser Grid Search avec validation croisée pour trouver la meilleure valeur de K parmi les choix suivants :

$$K \in \{1,3,5,7,9,11\}$$

Quelle est la raison principale pour laquelle il utilise Grid Search avec validation croisée au lieu de tester un seul train/test split ?

a) Grid Search teste automatiquement toutes les valeurs de K et sélectionne celle qui minimise l'erreur généralisée, offrant ainsi un modèle plus robuste.
b) Grid Search entraîne KNN plus rapidement, car il évite d'entraîner le modèle plusieurs fois.
c) Grid Search empêche l'overfitting, car il ajuste automatiquement K pour donner la meilleure performance sur toutes les nouvelles données.
d) Grid Search est inutile pour KNN, car cet algorithme ne dépend pas de paramètres à optimiser.

26.L'algorithme K-Nearest Neighbors (KNN) est souvent qualifié d'algorithme paresseux ("lazy algorithm") en machine learning. Quelle est la raison principale de cette classification ?

a) Parce qu'il ne nécessite pas de phase d'entraînement, il stocke simplement les données et effectue les calculs uniquement lors de la prédiction.
b) Parce qu'il utilise des heuristiques pour accélérer la recherche des voisins les plus proches.
c) Parce qu'il entraîne un modèle complexe avant d'effectuer des prédictions, ce qui ralentit son exécution.
d) Parce qu'il fonctionne uniquement avec des données de petite taille et ne peut pas être utilisé sur de grands datasets.

27. Quel est le principal objectif de l'algorithme SVM ?
a) Maximiser la marge entre les classes
b) Minimiser l'erreur quadratique moyenne
c) Générer des arbres de décision
d) Réduire le nombre de variables

28.Quel concept permet à SVM de gérer les données non linéairement séparables ?
a) Le paramètre C
b) La transformation polynomiale

c) Le Kernel Trick
d) La normalisation

29.Un scientifique des données entraîne un SVM (Support Vector Machine) sur un jeu de données. Il hésite entre une séparation avec marge dure (Hard Margin) et une séparation avec marge souple (Soft Margin).

Dans quel cas l'utilisation d'un Soft Margin est préférable ?

a) Lorsque les données sont parfaitement séparables, pour éviter d'avoir des points mal classés.
b) Lorsque les données contiennent du bruit ou des valeurs aberrantes, car une marge souple permet d'éviter un sur-apprentissage.
c) Lorsque l'on veut minimiser le temps d'entraînement du modèle, car Soft Margin converge plus rapidement.
d) Lorsque l'on veut que le modèle ait une marge maximale, car un Soft Margin offre toujours une plus grande distance entre les classes.

30.Un SVM linéaire ne peut pas séparer un dataset où les classes ne sont pas séparables dans l'espace d'origine. L'ingénieur décide d'utiliser le Kernel Trick pour transformer les données dans un espace de dimension supérieure.

Que fait exactement le Kernel Trick ?

a) Il ajoute plus de caractéristiques (features) au dataset pour améliorer la classification.
b) Il projette les données dans un espace de plus grande dimension sans calculer explicitement les nouvelles coordonnées.
c) Il supprime les données non séparables et entraîne le modèle uniquement sur les points bien classés.
d) Il applique une fonction de coût pour minimiser la perte sur les données mal classées.

31.Un data scientist utilise un SVM pour classifier des données non linéairement séparables. Il doit choisir une fonction de noyau (kernel function) pour améliorer la séparabilité.

Quelle fonction de noyau est généralement la plus utilisée pour les données très complexes ?

a) Noyau linéaire
b) Noyau polynomial
c) Noyau gaussien (RBF - Radial Basis Function)
d) Noyau sigmoïde

32.Un ingénieur en machine learning doit entraîner un SVM multi-classe et hésite entre les approches One-vs-All et One-vs-One.

Quelle est la principale différence entre ces deux méthodes ?

a) One-vs-All entraîne un seul classificateur SVM, tandis que One-vs-One entraîne un réseau de neurones avant d'utiliser SVM.
b) One-vs-All entraîne un classificateur pour chaque classe contre toutes les autres, tandis que One-vs-One entraîne un classificateur pour chaque paire de classes.
c) One-vs-All est plus précis que One-vs-One dans tous les cas car il utilise un seul modèle pour éviter les erreurs de prédiction.
d) One-vs-One est toujours préférable car il nécessite moins de calculs et de mémoire que One-vs-All.

33.Lequel des éléments suivants est un avantage clé du SVM en machine learning ?

a) Il fonctionne bien avec des données de haute dimension et permet la classification non linéaire grâce au "kernel trick".
b) Il est plus rapide que les arbres de décision et les réseaux de neurones pour de grands datasets.
c) Il est conçu principalement pour les problèmes multi-classes et gère naturellement plus de trois classes sans adaptation.
d) Il n'est pas affecté par les noyaux mal choisis et fonctionne toujours avec une bonne précision.

34.Pourquoi SVM n'est-il pas toujours le meilleur choix pour un problème de classification ?

a) Il est moins performant sur les jeux de données bruités et volumineux.
b) Il est toujours plus lent qu'un réseau de neurones, quelle que soit la taille du dataset.
c) Il n'a pas besoin de régularisation car il ne souffre jamais de sur-apprentissage (overfitting).
d) Il ne peut pas être utilisé avec des noyaux autres que le noyau linéaire.

35.Quel est le principe de base d'un arbre de décision ?
a) Utiliser une régression linéaire pour classifier les données
b) Construire un modèle prédictif sous forme d'arbre en utilisant des règles de décision
c) Associer chaque donnée d'entrée à une classe grâce à une table de correspondance
d) Séparer les données en groupes de même taille avant d'effectuer une classification

36.Quelle mesure est utilisée pour choisir le meilleur attribut à chaque division de l'arbre ?

a) Moyenne arithmétique des valeurs
b) Nombre d'instances dans chaque classe
c) Indice de Gini ou Entropie
d) Variance des données

37.Quelle est la principale différence entre les algorithmes CART et ID3 pour les arbres de décision ?

a) CART utilise l'indice de Gini tandis qu'ID3 utilise l'entropie
b) ID3 est basé sur la régression tandis que CART est basé sur la classification
c) CART ne fonctionne que pour des problèmes de classification binaire
d) ID3 est plus rapide que CART pour tous les types de jeux de données

38.Qu'est-ce qu'un nœud pur dans un arbre de décision ?

a) Un nœud qui contient des instances de plusieurs classes
b) Un nœud qui contient uniquement des instances d'une seule classe
c) Un nœud dont la taille est supérieure à un certain seuil
d) Un nœud où l'indice de Gini est maximal

39.Quel est le principal inconvénient des arbres de décision ?

a) Ils ne fonctionnent pas avec des données numériques
b) Ils nécessitent des données parfaitement équilibrées pour être efficaces
c) Ils sont sensibles au surapprentissage (overfitting)
d) Ils ne peuvent pas être utilisés pour des problèmes de classification

40.Dans un Arbre de Décision, un data scientist remarque que l'ajout de nombreuses caractéristiques à son dataset ne semble pas toujours améliorer les performances du modèle.

Pourquoi l'ajout d'un grand nombre de caractéristiques peut être problématique dans un Arbre de Décision ?

a) Parce que l'arbre devient plus profond et complexe, ce qui augmente le risque de sur-apprentissage (overfitting).
b) Parce qu'un arbre de décision ne peut pas gérer plus de trois caractéristiques à la fois.
c) Parce que l'algorithme de sélection de caractéristiques choisit toujours les plus pertinentes, donc ajouter plus de variables ne change rien.
d) Parce que les arbres de décision fonctionnent uniquement avec un nombre fixe de caractéristiques et ignorent les autres.

41. Quelle est la première étape du processus CRISP-DM ?

a) Préparation des données
b) Modélisation
c) Compréhension du métier
d) Évaluation

42.Dans le cadre de TDSP, quel rôle est principalement responsable de la gestion de projet et de la coordination d'équipe ?

a) Data Engineer
b) Data Scientist
c) Project Manager
d) Business Analyst

43.Quel est l'objectif principal de la phase "Compréhension des données" dans le processus CRISP-DM ?

a) Créer un modèle prédictif
b) Nettoyer les données
c) Analyser et explorer les données pour comprendre leur structure et leur qualité
d) Déployer le modèle en production

44.Quelle est la principale différence entre TDSP et CRISP-DM ?

a) TDSP est spécifique à Microsoft, tandis que CRISP-DM est un standard ouvert
b) TDSP se concentre davantage sur la phase de prétraitement des données
c) CRISP-DM est orienté vers les projets de business intelligence uniquement
d) Il n'y a aucune différence entre les deux

45.Quel rôle joue la phase "Évaluation" dans CRISP-DM ?

a) Déterminer les étapes à suivre pour le déploiement du modèle
b) Mesurer les performances du modèle et son adéquation avec les objectifs
c) Nettoyer et prétraiter les données
d) Préparer les données pour la modélisation