Devoir Surveillé

Semestre : 1 ■ 2 □
Session : Principale ■ Rattrapage □

Module: Machine Learning Appliqué
Classes: 4 ERP-BI
Enseignantes : Equipe ML Appliqué
Nombre de pages : 4 Documents : NON Autorisé Internet : NON Autorisé
Date : 31/10/2024 Heure : 11h00 Durée :1h

Merci de répondre sur la feuille des réponses

1. Laquelle de ces affirmations concernant le CRISP-DM est fausse ?
a. Le CRISP-DM est un processus linéaire et séquentiel.
b. Le CRISP-DM est un processus itératif.
c. Le CRISP-DM est adaptable à différents types de projets de data science.
d. Le CRISP-DM est un standard largement utilisé dans l'industrie.

2. Que signifie la technique d'imputation des données ?
a. Supprimer les lignes manquantes
b. Remplacer les valeurs manquantes par des valeurs estimées
c. Convertir les données catégorielles en numériques
d. Augmenter les données par des techniques de synthèse

3. Pourquoi est-il important de nettoyer les données avant de les utiliser dans un modèle de machine learning ?
a. Pour améliorer la précision du modèle
b. Pour réduire le temps de calcul
c. Pour rendre les données plus compréhensibles
d. Toutes ces réponses

4. Quelle est la différence entre la standardisation et la normalisation ?
a. La standardisation met les données à l'échelle, tandis que la normalisation les contraint entre 0 et 1.
b. La normalisation met les données à l'échelle, tandis que la standardisation les contraint entre 0 et 1.
c. Il n'y a pas de différence.
d. La standardisation est utilisée pour les données catégorielles, tandis que la normalisation est utilisée pour les données numériques.

5. Quel est l'un des principaux avantages de la réduction de dimensionnalité ?
a. Augmenter la taille des données
b. Améliorer la visualisation et la compréhension des données
c. Éliminer toutes les valeurs manquantes
d. Augmenter la complexité du modèle.

6. Quel est le risque de ne pas effectuer de nettoyage de données ?
a. Augmentation de la complexité
b. Réduction de la performance du modèle
c. Amélioration de la précision
d. Augmentation de la taille du dataset.

7. Quel est le principal inconvénient de l'élimination de caractéristiques peu pertinentes ?
a. Risque de perdre des informations importantes
b. Augmentation du temps d'exécution
c. Diminution de la précision
d. Aucun inconvénient

8. Vous avez un dataset avec une colonne "Prix" contenant des valeurs aberrantes (par exemple, un prix de 100 000 DT pour un produit qui coûte normalement 10 DT). Quelle technique de nettoyage est la plus adaptée pour gérer ces valeurs ?
a. Suppression des lignes
b. Imputation par la moyenne
c. Détection et remplacement par des valeurs plausibles
d. Standardisation

9. À quelle étape du CRISP-DM effectue-t-on généralement l'analyse exploratoire des données ?
a. Compréhension du besoin métier
b. Compréhension des données
c. Préparation des données
d. Modélisation

10. Quelle est l'étape la plus coûteuse en termes de calcul dans l'algorithme KNN ?
a. Le training

b. Le test
c. Déploiement
d. Evaluation

11. Que se passe-t-il si la valeur de K est trop petite ?

a. Il peut surapprendre et être trop sensible au bruit
b. Le modèle peut surapprendre et être trop sensible au bruit
c. Le modèle ignorera les points de données voisins
d. Le modèle ne fonctionnera pas

12. KNN peut-il être utilisé pour des problèmes de régression ?

a. Oui, en prenant la moyenne des voisins
b. Non, KNN est exclusivement utilisé pour la classification
c. Oui, en utilisant un arbre de régression
d. Non, KNN ne peut pas traiter les données continues

13. Quelle est l'étape la plus coûteuse en termes de calcul dans l'algorithme KNN ?

a. La recherche du K optimal
b. Le calcul des distances entre les points
c. La construction du modèle
d. La sélection des caractéristiques

14. Quelles sont les données utilisées dans KNN pendant la phase de test ?

a. Seulement les K plus proches voisins
b. L'ensemble complet des données.
c. Les données les plus récentes
d. Les données filtrées par le modèle

15. Quel est le principal inconvénient de KNN pour des ensembles de données volumineux ?

a. Il ne peut pas traiter des caractéristiques multiples
b. Il consomme beaucoup de temps et de mémoire pour calculer les distances
c. Il a une complexité de calcul faible
d. Il ignore les classes minoritaires

16. Dans quel cas l'algorithme KNN n'est-il pas recommandé ?

a. Lorsque les données sont très bruitées.
b. Lorsque les données sont en grande dimension avec de nombreuses caractéristiques.
c. Lorsque les données sont réparties de manière uniforme.
d. Lorsque la taille du jeu de données est petite et équilibrée.

17. Est-ce que l'augmentation de K améliore la qualité de prédiction ?

a. L'augmentation améliore toujours la qualité de la prédiction
b. Oui Si le dataset est trop bruité
c. Oui Si le dataset est balancé
d. L'augmentation n'influe pas toujours la qualité de la prédiction

18. Que représente la "marge" dans le contexte SVM ?

a. La distance entre les vecteurs supports
b. La distance entre les classes
c. La quantité d'erreurs de classification
d. La vitesse d'exécution de l'algorithme

19. Que se passe-t-il si la marge d'un SVM est très étroite ?

a. Le modèle est moins robuste et risque de surajuster les données d'entraînement
b. Le modèle sera toujours précis sur les nouvelles données
c. Le modèle ne sera pas affecté par le bruit dans les données
d. Le modèle aura une complexité computationnelle plus faible

20. Quelle méthode est utilisée pour gérer les données non linéaires avec SVM ?

a. Normalisation des données
b. Transformation des données
c. Utilisation de fonctions noyau
d. Utilisation de modèles d'ensemble

21. Quel est l'inconvénient d'utiliser un noyau linéaire pour des problèmes non linéaires ?

a. Le noyau linéaire est plus complexe à calculer.
b. Il risque de ne pas trouver une bonne séparation des classes.
c. Il génère trop de vecteurs de support.
d. Il nécessite plus de données pour fonctionner correctement.

22. Comment l'algorithme SVM maximise-t-il la marge entre les classes ?

a. En minimisant l'erreur quadratique moyenne.
b. En utilisant un algorithme de la descente de gradient.
c. en maximisant le coefficient de pondération de la fonction
d. En maximisant la somme des distances entre tous les points et l'hyperplan.

23. Quel type de noyau peut être considéré comme une généralisation du noyau linéaire ?

a. Noyau polynomial
b. Noyau gaussien (RBF)
c. Noyau sigmoid
d. Noyau de Laplace

24. Que mesure la métrique AUC-ROC dans les modèles de classification binaire ?

a. La précision d'un modèle sur l'ensemble des classes.
b. La capacité d'un modèle à distinguer entre les classes positives et négatives.
c. Le nombre moyen d'erreurs commises par le modèle.
d. La dispersion des données autour de la prédiction moyenne.

25. Dans un SVM avec noyau, quelle est la principale conséquence de l'augmentation de la dimensionnalité des données ?

a. Cela réduit le risque de surajustement.
b. Cela augmente le temps d'entraînement et peut rendre le modèle plus complexe.
c. Cela n'a aucun impact sur la performance du modèle.
d. Cela simplifie le modèle et améliore la vitesse d'exécution.

26. Quel est un inconvénient majeur des arbres de décision ?

a. Ils sont difficiles à interpréter
b. Ils sont sensibles aux données bruitées
c. Ils ne gèrent pas bien les données catégorielles
d. Ils ne peuvent pas être utilisés pour la régression

27. Quelle est une des principales limites des arbres de décision lorsqu'ils sont utilisés sur des données de très haute dimensionnalité ?

a. Ils sont inefficaces pour traiter des variables catégorielles
b. Ils ont tendance à sous-apprendre les données
c. Ils peuvent devenir instables et très sensibles aux petites variations dans les données
d. Ils ne peuvent pas gérer les données manquantes

28. Lors de la construction d'un arbre de décision, quel est le principal objectif de la fonction de coût (comme l'indice de Gini ou l'entropie) ?

a. Maximiser la taille des nœuds
b. Minimiser l'hétérogénéité des nœuds
c. Augmenter le nombre de feuilles
d. Minimiser la hauteur de l'arbre

29. Comment un arbre de décision gère-t-il les variables catégorielles ?

a. En les convertissant en variables continues
b. En les divisant en sous-groupes possibles à chaque nœud
c. En les remplaçant par la moyenne de leurs occurrences
d. En les ignorant dans la construction de l'arbre

30. Quel est l'inconvénient principal de la méthode des arbres de décision lorsqu'elle est utilisée seule, sans technique d'ensemble comme le bagging ou le boosting ?

a. Elle nécessite un long temps de calcul
b. Elle est sujette à l'overfitting
c. Elle ne peut pas gérer des jeux de données larges
d. Elle ne supporte pas les données manquantes

31. Si l'arbre de décision montre des performances très différentes à chaque exécution, quelle en pourrait être la cause ?

a. La taille du jeu de données est trop importante
b. Le modèle est stable et robuste, donc cela ne devrait pas arriver
c. L'arbre est sensible à la variabilité dans l'échantillonnage des données
d. Il y a un trop grand nombre de variables catégorielles dans le modèle

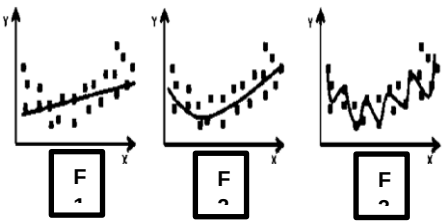
32. Quelle est le taux de bonne classification ?

a. 88.16%
b. 13.33 %
c. 88.23 %
d. 90.8 %

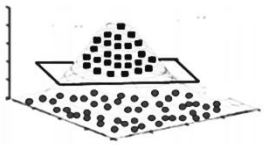
	English Speaker	Spanish Speaker
English Speaker	86	12
Spanish Speaker	10	79

33. Quelle figure représente un surapprentissage (overfitting) ?

a. Figure F1
b. Figure F2
c. Figure F3
d. Aucune réponse



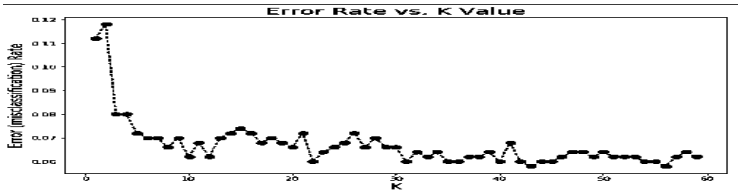
34. Que représente la figure ci-dessus ?



a. Les données d'origine dans leur espace 2D
b. les données transformées dans un espace de dimension supérieure grâce à une fonction noyau
c. Une séparation linéaire du noyau
d. Les données transformées dans un espace de dimension supérieure grâce à une séparation noyau

35. Quelle est la valeur de K qui donne la meilleure prédiction ?

a. 5
b. 10 et 12
c. 43 et 56
d. 8



Mise en situation 1 : Une entreprise de e-commerce souhaite améliorer la prédiction du comportement d'achat de ses utilisateurs en ligne. L'équipe de data science décide de construire un modèle basé sur un arbre de décision pour prédire si un utilisateur effectuera un achat ou non, en se basant sur plusieurs facteurs comme : **Temps passé sur le site** (en minutes), **Nombre de pages visitées, Source du trafic** (direct, via une publicité, ou via une recherche organique), **Historique d'achat** (si l'utilisateur a déjà acheté auparavant), et **Montant dépensé lors des visites précédentes.**

Après avoir entraîné le modèle, l'équipe observe que l'arbre de décision a tendance à sur-apprendre le jeu de données d'entraînement. L'équipe cherche des solutions pour éviter cela et pour améliorer les performances sur les données de test.

36. Quel serait un bon indicateur que l'arbre de décision est en train de sur-apprendre les données d'entraînement ?

a. La performance sur le jeu d'entraînement est excellente, mais sur le jeu de test, elle est faible
b. La performance sur le jeu d'entraînement et le jeu de test est similaire
c. L'arbre de décision ne parvient pas à faire de bonnes prédictions sur le jeu d'entraînement
d. Le modèle prend beaucoup de temps à s'entraîner

37. Comment la variable "Source du trafic" est-elle probablement traitée dans l'arbre de décision ?

a. Elle est transformée en une variable continue
b. Elle est divisée en catégories distinctes pour chaque source
c. Elle est ignorée car elle est catégorielle
d. Elle est normalisée avant d'être utilisée

38. Pour éviter que le Temps passé sur le site n'ait un poids disproportionné dans l'arbre de décision, quelle technique pourrait être utilisée ?

a. Supprimer la variable " Temps passé sur le site "
b. Appliquer une technique de normalisation sur les variables continues
c. Limiter le nombre de divisions basées sur le Temps passé sur le site
d. Diviser la variable " Temps passé sur le site " en catégories

Mise en situation 2 : Un hôpital souhaite prédire si un patient est susceptible de développer des complications post-opératoires après une chirurgie. Pour cela, ils utilisent un modèle d'arbre de décision basé sur les données de santé des patients. Les variables prises en compte incluent : **Âge du patient, Durée de la chirurgie (en heures), Type de chirurgie (orthopédique, cardiaque, généraliste, etc.), Antécédents médicaux (présence ou absence de maladies chroniques),** et **Niveau d'activité physique (faible, moyen, élevé).**

Après avoir construit un premier modèle, l'équipe médicale constate que l'arbre de décision accorde beaucoup trop d'importance à l'âge des patients et néglige d'autres variables comme les antécédents médicaux ou la durée de la chirurgie.

39. Quel problème peut survenir si l'arbre de décision accorde une trop grande importance à une variable (comme l'âge) ?

a. Cela signifie que l'âge est la seule variable importante
b. L'arbre peut sur-apprendre en se basant principalement sur cette variable, négligeant des informations cruciales des autres variables
c. L'arbre va avoir une précision élevée sur les données de test
d. Cela n'affecte pas la performance du modèle

40. Si les antécédents médicaux sont fortement corrélés avec l'âge du patient, que pourrait-il se passer si cette variable est supprimée du modèle ?

a. Eviter l'overfitting
b. Le modèle aurait une meilleure performance car il éviterait la colinéarité
c. Le modèle deviendrait plus complexe
d. La suppression d'une variable corrélée n'affecterait pas la performance du modèle

Bon Travail