Quiz sur l'Analyse Exploratoire des Données (EDA) en Apprentissage Machine

1. Comprendre la distribution des données

Question :

Pourquoi est-il important d'analyser la distribution des données avant d'entraîner un modèle d'apprentissage machine ?

Réponse : Cela permet d'identifier d'éventuelles asymétries, tendances ou valeurs extrêmes qui pourraient impacter la performance du modèle.

2. Détecter les valeurs manquantes et aberrantes

Question :

Quelle méthode peut être utilisée pour détecter les valeurs aberrantes dans un jeu de données ?

Réponse : La méthode des quartiles (IQR - Interquartile Range), le Z-score, ainsi que des techniques avancées comme Isolation Forest et LOF (Local Outlier Factor).

3. Étudier les relations entre les variables

Question :

Quelle est l'utilité d'une matrice de corrélation dans l'analyse exploratoire des données ?

Réponse : Elle permet d'identifier les relations entre différentes variables et de détecter les corrélations fortes qui pourraient influencer la modélisation.

4. Détection des biais et déséquilibres des classes

Question :

Comment peut-on équilibrer un jeu de données fortement déséquilibré ?

Réponse : On peut utiliser l'oversampling (ex : SMOTE) pour augmenter les classes minoritaires, l'undersampling pour réduire la taille des classes majoritaires, ou encore ajuster les poids des classes dans l'algorithme d'apprentissage.

5. Analyse des distributions conditionnelles

Question :

Quelle technique peut être utilisée pour comparer la distribution d'une variable continue en fonction d'une variable catégorielle ?

Réponse : On peut utiliser les boxplots, les violin plots, ou encore les histogrammes conditionnels.

6. Identification de groupes sous-représentés

Question :

Pourquoi est-il important d'identifier les groupes sous-représentés dans un jeu de données ?

Réponse : Cela permet d'éviter que le modèle soit biaisé en faveur des classes majoritaires et assure une meilleure généralisation des prédictions.

7. Guider la préparation des données

Question :

Quelle est la différence entre la normalisation et la standardisation ?

Réponse : La normalisation (MinMaxScaler) met les valeurs dans une plage spécifique (généralement entre 0 et 1), tandis que la standardisation (StandardScaler) centre les données autour de la moyenne avec un écart-type de 1.

8. Importance de la qualité des données

Question :

Pourquoi la qualité des données est-elle cruciale dans le développement d'une application d'IA ?

Réponse : Des données de mauvaise qualité peuvent introduire du bruit, biaiser les modèles et réduire leur performance et leur capacité à généraliser à de nouvelles données.

PARTIE QCM

1. Comprendre la distribution des données

Question :

Quelle méthode permet d'analyser la répartition des valeurs d'une variable ?

A) Matrice de corrélation
B) Histogramme
C) SMOTE
D) Isolation Forest

Réponse : B) Histogramme

2. Détecter les valeurs manquantes et aberrantes

Question :

Quelle technique n'est pas utilisée pour gérer les valeurs manquantes ?

A) Remplacement par la moyenne
B) Suppression des lignes concernées
C) PCA
D) KNN Imputer

Réponse : C) PCA

3. Étudier les relations entre les variables

Question :

Quel outil est couramment utilisé pour visualiser les relations entre plusieurs variables numériques ?

A) Boxplot
B) Heatmap de corrélation
C) Diagramme en barres
D) One-hot encoding

Réponse : B) Heatmap de corrélation

4. Détection des biais et déséquilibres des classes

Question :

Quelle approche permet de rééquilibrer un jeu de données fortement déséquilibré ?

A) Oversampling avec SMOTE
B) Réduction de la dimensionnalité
C) Utilisation des statistiques descriptives
D) Clustering K-Means

Réponse : A) Oversampling avec SMOTE

5. Analyse des distributions conditionnelles

Question :

Quel graphique est le plus adapté pour comparer une variable numérique selon une catégorie ?

A) Scatter plot
B) Boxplot
C) Heatmap

D) Pair plot

Réponse : B) Boxplot

6. Identification de groupes sous-représentés

Question :

Pourquoi faut-il détecter les catégories sous-représentées dans un jeu de données ?

A) Pour accélérer l'entraînement du modèle
B) Pour éviter un biais du modèle envers les classes majoritaires
C) Pour réduire la taille du dataset
D) Pour rendre les calculs plus complexes

Réponse : B) Pour éviter un biais du modèle envers les classes majoritaires

7. Guider la préparation des données

Question :

Quel est l'objectif principal de la standardisation des données ?

A) Transformer les variables catégorielles en numériques
B) Mettre toutes les valeurs entre 0 et 1
C) Centrer les données autour de la moyenne avec un écart-type de 1
D) Supprimer les valeurs aberrantes

Réponse : C) Centrer les données autour de la moyenne avec un écart-type de 1

8. Importance de la qualité des données

Question :

Quel impact une mauvaise qualité des données peut-elle avoir sur un modèle d'apprentissage machine ?

A) Améliorer la précision
B) Accélérer l’entraînement
C) Introduire du bruit et biaiser les prédictions
D) Réduire le nombre de variables

Réponse : C) Introduire du bruit et biaiser les prédictions