Activité pratique 2: Techniques de classification binaire

Ce travail pratique a pour objectif de **mettre en œuvre l'ensemble du processus de classification supervisée** appliqué à un cas concret de **prédiction de la solvabilité des clients bancaires**.

1. Importez les bibliothèques nécessaires pour l'analyse et la modélisation.
2. Chargez le fichier `scoring.sav` dans un DataFrame.
3. Affichez les 5 premières lignes et les dimensions de la base de données.
4. Listez toutes les variables et leurs types.
5. Décrivez statistiquement la base (minimum, maximum, moyenne, effectifs, etc.).

Questions :

- Combien d'observations et de variables contient la base ?
- Quelles sont les principales variables numériques et catégorielles ?
- Que remarquez-vous à propos de la variable cible `Statut1` ?

6. Étudiez la distribution de `Statut1` (effectifs, pourcentages).
7. Recodez la variable `Statut1` pour qu'elle prenne des valeurs numériques (ex. : 0 et 1).
8. Vérifiez le résultat du recodage.

Questions :

- La base de données est-elle équilibrée ou déséquilibrée ?
- Quelles conséquences ce déséquilibre peut-il avoir sur la performance du modèle ?

9. Identifiez les valeurs manquantes et représentez-les graphiquement.
10. Décidez d'une méthode de traitement des valeurs manquantes (suppression ou imputation).
11. Calculez et affichez la matrice de corrélation.

Questions :

- Quelles variables sont les plus corrélées avec la variable `Statut1` ?
- Quelles paires de variables présentent une corrélation forte (risque de redondance) ?

12. Séparez la variable cible y et les variables explicatives X.
13. Divisez les données en deux sous-ensembles : apprentissage (80 %) et test (20 %).

Questions :

- Quelles sont les dimensions de chaque sous-ensemble ?
- Pourquoi faut-il réserver un jeu de test ?

Activité pratique 2: Techniques de classification binaire

14. Entraînez un modèle de régression logistique avec les paramètres par défaut.
15. Affichez les coefficients du modèle et l'interception.
16. Faites des prédictions sur le jeu de test.

Questions :

- Quelle est la précision (accuracy) du modèle sur les données test ?
- Que signifient les coefficients positifs et négatifs dans ce contexte ?

17. Calculez la matrice de confusion et les indicateurs de performance : **précision, rappel, F1-score**.
18. Tracez la courbe ROC et calculez l'AUC.

Questions :

- Interpréter la matrice de confusion obtenue
- Quelle est la valeur de l'AUC et comment l'interpréter ?
- Quelle métrique vous semble la plus pertinente ici ?

19. Entraînez les trois modèles de régression logistique suivants :
    1. Lasso (L1)
    2. Ridge (L2)
    3. Elastic Net (L1 + L2)
20. Comparez leurs performances sur le jeu test.

Questions :

- Quelles différences observez-vous entre les trois modèles ?
- Quel solveur est le plus adapté à chaque type de régularisation ?
- Comment la régularisation influence-t-elle les coefficients ?

21. Mettez en œuvre une validation croisée à 5 plis (K=5) pour optimiser les hyperparamètres du modèle **KNN et évaluer sa performance**:
    - Définissez une grille de valeurs pour les hyperparamètres `n_neighbors` et `metric`.
    - Lancez une recherche exhaustive par **GridSearchCV** (CV=5, scoring='accuracy').
    - Affichez le meilleur modèle et ses hyperparamètres optimaux.
    - Utiliser la validation KFold classique et puis **StratifiedKFold.** Pourquoi la stratification est-elle utile pour ce jeu de données ?
    - Quels sont les meilleurs hyperparamètres ?Quelle est la performance moyenne associée ?
    - Comment évolue l'accuracy quand le nombre de voisins augmente

Activité pratique 2: Techniques de classification binaire

22. Implémentez une validation croisée imbriquée avec :
    1. KFold externe (5 plis)
    2. KFold interne (3 plis)
23. Vous allez maintenant utiliser votre **modèle final de classification** pour prédire la **solvabilité de ce nouveau client** :

Variable	Valeur
Age	**41**
Marital	**2** (marié)
Expenses	**65.0**
Income	**160.0**
Amount	**1200.0**
Price	**1350.0**

- Quelle est la **classe prédite** pour ce nouveau client ?
- Quelle est la **probabilité estimée** de non-solvabilité ?
- D'après ce résultat, **accepteriez-vous ou refuseriez-vous** la demande de crédit ?
- Que se passe-t-il si vous augmentez son revenu à 120 000 DH ? Observez et expliquez la variation de la probabilité.

Livrables attendus

- **Notebook complet** avec code, visualisations et réponses argumentées.
- **Note synthétique** (1 page) présentant vos conclusions et recommandations.