INSTITUT SUPERIEUR POLYTECHNIQUE
DE MADAGASCAR

Machine Learning
TP 1

Thème : Initiation au Data Science

Livrables

Pour chaque partie, un fichier notebook contenant codes, graphiques, résultats et interprétations.

Partie A – Clustering

Dataset : Mall Customer Segmentation Data
Objectif : Segmenter les clients du centre commercial en groupes homogènes.
Plan :

1. Chargement : utiliser le fichier `Mall_Customers.csv` fourni.
2. Nettoyage et prétraitement : Traitez les valeurs manquantes le cas échéant. Standardisez les variables quantitatives (âge, revenu annuel, score de dépense).
3. EDA univariée et multivariée : Histogrammes et boxplots de chaque variable. Matrice de corrélation et scatter-plot matrix.
4. Feature engineering / sélection : Supprimez les features non-informatives ou redondantes. (Optionnel) Créez des ratios ou catégorisez des variables (ex. classes d'âge).
5. Réduction de dimension : Appliquez une PCA pour visualiser les données en 2D.
6. Clustering : K-means (k de 2 à 6), Évaluation et stabilité : Calcul des indices (silhouette et elbow)
7. Interprétation des clusters : Analyse des centroïdes ou profils, visualisation en 2D (PCA) colorée par cluster.
8. Conclusions : Synthèse de la pertinence des clusters et suggestions d'améliorations.

Partie B – Régression

Dataset : California Housing
Objectif : Prédire le prix médian des maisons (variable continue).
Plan :

1. Chargement : Partir du notebook `Regression - California Housing.ipynb`
2. Nettoyage et prétraitement : Imputez/supprimez les valeurs manquantes. Vérifiez unités et typages.
3. EDA : Statistiques descriptives, histogrammes, boxplots, scatter-plots prix vs variables clés, matrice de corrélation.
4. Split *train/validation/test* : 60% train, 20% validation, 20% test. Le *validation set* sera utilisé pour choisir le modèle approprié et le *test set* sera utilisé pour l'évaluation finale.
5. Feature engineering / sélection : Normalisation (StandardScaler), encodage catégoriel …
6. Baseline et modélisation initiale : Régression linéaire, rapporter MSE, MAE, R^2 sur le set de *validation*.
7. Régularisation : *Ridge* : pénalisation L2 pour la multicolinéarité et *Lasso* : pénalisation L1 pour la sélection de features.
8. Évaluation finale : Application du meilleur modèle sur le *test* set, présentation de MSE, MAE, R^2 et intervalles de confiance.
9. Conclusions : Interprétation de l'importance des variables et pistes d'amélioration.

Partie C – Classification

Dataset : Pima Indians Diabetes
Objectif : Prédire la présence de diabète (étiquette binaire).
Plan :

1. Chargement : utiliser le fichier `diabetes.csv` fourni.
2. Nettoyage et prétraitement : Gérez les zéros/missing dans les variables cliniques, standardisez ou normalisez les mesures.
3. EDA et déséquilibre : Barplots du nombre de positifs/négatifs, statistiques descriptives par classe.
4. Split *train/validation/test* : 60% train, 20% validation, 20% test. Le *validation set* sera utilisé pour choisir le modèle approprié et le *test set* sera utilisé pour l'évaluation finale.
5. Feature engineering / sélection : Encodage, normalisation/standardisation, …
6. Gestion du déséquilibre
7. Modélisation initiale : Logistic Regression. Evaluation Accuracy, Precision, Recall, F1 sur *validation*.
8. Évaluation finale : Matrice de confusion, analyses sur le *test* set.
9. Explainability : Importance des features, analyse des erreurs (faux positifs/faux négatifs, f1-score).
10. Conclusions : Synthèse de la performance et recommandations pour le déploiement/collecte de données.

Instructions

1. Les parties A, B et C sont indépendantes.
2. Travail à faire par groupe de trois étudiants. Les parties sont indépendantes. Vous devez inscrire votre groupe dans le google form : https://forms.gle/ofuX8SmiQNix9kyr9. La répartition de groupe est permanente pour toute l'Année Universitaire. Le formulaire sera disponible jusqu'au vendredi 20 juin 2025 à 23 heures 59.

3. Placer vos fichiers dans le repository GIT renseigné dans le formulaire précedent. Vous ne devez plus toucher à votre travail après mardi 24 juin 2025 à 23 heures 59.
4. Les Datasets CSV et le fichier `Regression - California Housing.ipynb` sont disponibles dans le dossier drive : https://drive.google.com/drive/folders/1f29qMWqIsOBmcgwlnUf9JUYoOOlWAYvp?usp=drive_link