Mini-Projet DATA832

Mise en place d'un workflow d'apprentissage supervisé avec scikit-learn

1 Objectifs pédagogiques

Ce projet a pour objectif d'amener les étudiants à mettre en pratique les concepts fondamentaux de l'apprentissage supervisé, en appliquant différentes méthodologies sur un problème de classification. Ils devront :

- Comprendre et appliquer un workflow rigoureux en machine learning.
- Comparer l'efficacité de plusieurs algorithmes de classification : kNN, arbres de décision et réseaux de neurones.
- Mettre en œuvre une validation croisée avec k-fold.
- Optimiser les hyperparamètres via une recherche par grille.
- Évaluer les performances des modèles avec des métriques adaptées.
- Rédiger un rapport structuré et analytique des résultats obtenus.

2 Description des défis

Les étudiants pourront choisir l'un des défis suivants. **Les données sont téléchargeables sur Moodle. Le lien est uniquement dédié au descriptif des datasets.**

2.1 Défi 1 : Classification des Pokémon légendaires

Contexte : Les Pokémon sont des créatures aux caractéristiques variées, certaines étant classées comme "légendaires" en raison de leur rareté et de leur puissance. L'objectif est d'entraîner un modèle permettant de prédire si un Pokémon est légendaire ou non à partir de ses statistiques.

Description des données : Le jeu de données comprend des informations sur 800 Pokémon, incluant des caractéristiques comme les points de vie (HP), l'attaque, la défense, la vitesse, ainsi que des attributs catégoriels (type, génération, etc.).

Pistes à explorer :

- Sélection des meilleures caractéristiques pour la classification.
- Comparaison des performances des modèles (arbres de décision, kNN, réseaux de neurones).
- Impact de la normalisation des données sur les résultats.

Lien du jeu de données : `https://www.kaggle.com/abcsds/pokemon`

2.2 Défi 2 : Reconnaissance de chiffres manuscrits (MNIST)

Contexte : La reconnaissance de chiffres manuscrits est une application classique de l'apprentissage automatique, largement utilisée dans la numérisation de documents et la lecture automatique de codes postaux.

Description des données : Le dataset MNIST contient 70 000 images de chiffres manuscrits (28x28 pixels, niveaux de gris) étiquetées de 0 à 9.

Pistes à explorer :

- Utilisation de PCA ou t-SNE pour la réduction de dimension.
- Comparaison des performances entre méthodes basées sur des distances (kNN) et des approches plus complexes (réseaux de neurones).

Lien du jeu de données : Voir TD partie supervisée.

2.3 Défi 3 : Classification de genres musicaux

Contexte : La reconnaissance automatique de genres musicaux est une tâche difficile, impliquant des caractéristiques temporelles et spectrales extraites des fichiers audio.

Description des données : Le dataset GTZAN contient 1 000 morceaux de 10 genres musicaux différents (rock, jazz, blues, reggae, etc.), avec des fichiers audio au format WAV.

Pistes à explorer :

- Extraction de caractéristiques spectrales (MFCCs, chroma, spectrogrammes).

— Utilisation de modèles adaptés aux données temporelles (ex. réseaux de neurones).
— Analyse des erreurs de classification et des genres les plus difficiles à différencier.

Lien du jeu de données : https://www.kaggle.com/andradaolteanu/gtzan-dataset-music-genre-classification

2.4 Défi 4 : Reconnaissance d'images de chiens et de chats

Contexte : La distinction entre chiens et chats est un problème classique de classification d'images en vision par ordinateur.

Description des données : Le jeu de données contient 25 000 images étiquetées de chiens et de chats.

Pistes à explorer :

— Prétraitement des images (redimensionnement, normalisation).
— Exploration de méthodes classiques (SIFT + SVM) et basées sur les réseaux de neurones.
— Impact de la taille de l'ensemble d'entraînement sur les performances.

Lien du jeu de données : https://www.kaggle.com/c/dogs-vs-cats

3 Grille d'évaluation

Table 1 – Grille d'évaluation du projet (Total : 20 points + 1 bonus)

Catégorie	Niveau 1	Niveau 2	Niveau 3
Difficulté du challenge (3 pts)	1 (Pokémon, MNIST)	2 (Musique)	3 (Chien vs Chat)
Rigueur méthodologique (3 pts)	1	2	3
Sélection des features (4 pts)	1-2	3	4
Tuning des hyperparamètres (3 pts)	1	2	3
Qualité du rapport (5 pts)	1-2	3-4	5
Utilisation d'outils IA (2 pts)	1	2	-
Bonus LaTeX (1 pt)	1 point si rapport en LaTeX		

4 Date limite de rendu

Les projets devront être soumis avant le **7 avril 2025**. Toute soumission après cette date entraînera une pénalité.