TEST #1 — Machine Learning for Cyber security

Ce test de TP en machine learning couvre la régression linéaire et polynomiale, la régression logistique, et la classification par les arbres de décision, nous choisirons un dataset polyvalent qui peut être utilisé pour à la fois des problèmes de régression et de classification. Le dataset "California Housing" est un excellent candidat pour cela. Il contient des données sur le logement en Californie, avec des caractéristiques comme la médiane du revenu, l'âge moyen d'une maison, le nombre de pièces, le nombre de chambres, la population, le nombre de ménages, et la médiane des prix des maisons. Pour les besoins de ce TP, nous pourrions ajouter une variable binaire artificielle pour la classification, par exemple, classer les maisons comme "chères" ou "abordables" basées sur un seuil de la médiane des prix des maisons.

Partie 1: Régression Linéaire et Polynomiale

Objectifs:

- Comprendre et appliquer les concepts de la régression linéaire simple et multiple.
- Explorer l'extension de la régression linéaire aux modèles polynomiaux pour capturer les relations non linéaires entre les caractéristiques et la cible.

Questions:

1. Prétraitement des données:
    - Charger le dataset et effectuer un nettoyage initial (traitement des valeurs manquantes, normalisation des caractéristiques, etc.).
    - Sélectionner une caractéristique pertinente (par exemple, la médiane du revenu) et la médiane des prix des maisons comme cible pour la régression linéaire simple.
2. Régression linéaire simple:
    - Effectuer une régression linéaire simple avec la caractéristique sélectionnée. Évaluer les performances du modèle en utilisant le R^2 et RMSE (Root Mean Square Error).
3. Régression linéaire multiple:
    - Ajouter plus de caractéristiques pour créer un modèle de régression linéaire multiple. Comparer les performances avec le modèle de régression linéaire simple.
4. Régression polynomiale:
    - Transformer les caractéristiques pour créer un modèle de régression polynomiale (degré 2 ou 3). Évaluer et comparer les performances avec les modèles linéaires.

Partie 2: Régression Logistique

Objectifs:

- Appliquer la régression logistique pour un problème de classification binaire.

Questions:

1. Préparation des données pour la classification:
    - Créer une nouvelle variable cible binaire indiquant si le prix d'une maison est "élevé" ou "bas" basé sur un seuil prédéfini.
    - Sélectionner des caractéristiques pertinentes pour la tâche de classification.
2. Modélisation et évaluation:
    - Construire un modèle de régression logistique. Utiliser la validation croisée pour évaluer la précision, le rappel, et le score F1 du modèle.

3. Interprétation et visualisation du modèle:
    - Visualiser la courbe ROC et calculer l'aire sous la courbe (AUC).

Partie 3: Classification par Arbres de Décision

Objectifs:

- Comprendre et implémenter un modèle de classification basé sur les arbres de décision.

Questions:

1. Construction de l'arbre de décision:
   - Utiliser les mêmes données de la **Partie 2** pour entraîner un arbre de décision. Expérimenter avec différents paramètres tels que la profondeur maximale de l'arbre et le nombre minimum d'échantillons requis pour être une feuille.
2. Évaluation du modèle:
   - Comparer les performances de l'arbre de décision avec le modèle de régression logistique de la **Partie 2** en utilisant les mêmes métriques d'évaluation.
3. Interprétation du modèle:
   - Visualiser l'arbre de décision et discuter de l'importance des différentes caractéristiques utilisées pour la classification.
   - Analyser les différentes branches de l'arbre et identifier les règles de décision les plus importantes.
4. Interprétation et comparaison du modèle:
   - Evaluer les performances du modèle d'arbre de décision sur l'ensemble de test.
   - Comparer les performances du modèle d'arbre de décision avec celles du modèle de régression logistique.

Dataset Suggéré:

- Dataset "California Housing" disponible dans plusieurs bibliothèques de machine learning comme scikit-learn. Ce dataset offre une bonne variété de caractéristiques pour explorer à la fois des problèmes de régression et de classification.
  - Vous pouvez le charger directement à partir d'anaconda :

```
from sklearn.datasets import fetch_california_housing
import numpy as np

# Charger le dataset California Housing
data = fetch_california_housing()
```

  - Vous pouvez également le charger directement à partir du fichier csv « *california_housing.csv*»:

```
import pandas as pd

# Load the dataset
file_path = 'california_housing.csv'
data = pd.read_csv(file_path)

# Display the first 10 records
data.head(10)
```

Ce TP offre une couverture complète des techniques de base en machine learning, permettant aux étudiants de pratiquer le prétraitement des données, la modélisation, et l'évaluation des modèles dans différents contextes.