

Projet de TP – Machine Learning

Licence Flex

Objectif du projet

L'objectif de ce projet est d'appliquer des techniques de *machine learning* sur des jeux de données réels à travers une étude de cas complète.

Le projet s'étalera sur les trois dernières séances de TP (TP8, TP9 et TP10). Lors de la séance TP8, vous formerez des groupes de 3 étudiants et choisirez le sujet que vous étudierez. Les séances TP8 et TP9 seront dédiées à l'analyse, à la modélisation et aux échanges avec l'enseignant. La séance TP10 sera consacrée à la présentation de vos travaux.

Choix du sujet

Vous pouvez choisir l'un des sujets ci-dessous ou en proposer un autre (par exemple depuis la plateforme Kaggle) sous réserve de validation par votre enseignant. Selon votre sujet, vous suivrez les consignes associés aux sujets 1 ou 2.

Sujet 1 : Régression

Objectif : Prédire la concentration d'ozone dans l'air à partir du jeu de données `ozone.csv`.

Modèles à tester :

- k -plus proches voisins,
- Régression Lasso,
- Régression Ridge,
- Au choix...

Métrique d'évaluation : RMSE (*Root Mean Squared Error*).

Sujet 2 : Classification

Objectif : Prédire la survie d'un passager (`Survived`) à partir du jeu de données `titanic.csv`.

Modèles à tester :

- k -plus proches voisins,
- Support Vector Machine (SVM),
- Régression logistique,

- Au choix...

Métrique d'évaluation : AUC (*Area Under the ROC Curve*).

Démarche recommandée

Pour tous les sujets, vous suivrez une démarche structurée inspirée des TP précédents :

1. Analyse exploratoire et nettoyage des données.
2. Construction et entraînement des modèles.
3. Optimisation des hyperparamètres.
4. Évaluation des performances.
5. Présentation visuelle et interprétation des résultats (tableaux et graphiques soignés).

La mise en œuvre de techniques de *feature engineering* est fortement recommandée pour améliorer les performances de vos modèles.

Évaluation

1. Jupyter Notebook

Chaque groupe devra remettre un notebook complet contenant :

- Une introduction présentant la problématique et le jeu de données.
- Une section de préparation et nettoyage des données.
- L'implémentation des modèles et l'optimisation des hyperparamètres.
- Une analyse comparative des résultats obtenus.

Le notebook doit être clair, documenté, commenté, propre et illustré par des graphiques.

2. Présentation orale

Lors de la séance TP10, chaque groupe présentera son travail :

- Durée : 10 minutes de présentation + 5 minutes de questions.
- Le respect du temps de présentation sera pris en compte dans l'évaluation.

Bon courage ! Soyez curieux, rigoureux et créatifs.