Projets d'Intelligence Artificielle – Instructions & Guide Étudiant

Module Fondement de l'IA, 2 LIG - ISGBizert, 2025-2026,

enseigné par Dr.Alaa Bessadok

2LIG - G1 et G2: Date de remise 11 decembre 2025

2LIG - G3: Date de remise 16 decembre 2025

Ce document présente les instructions pour des projets d'IA. Chaque projet contient une courte description, l'objectif, ainsi que les étapes à suivre pour réaliser un notebook clair et bien structuré.

Liste des étudiants concernés par le projet:

2 LIG - G1	2 LIG - G2	2 LIG - G3	2 LIG - G3
Chayma Brahim Boubaker	Ranim Fraj Hkiri Arij	Doua Nasri Ranim Manoubia Nasri Nour Guesmi	Malek Sassi Ines Zoghlemi Sinda Moussaoui Mariem Tayachi

Dans le cadre de ce projet, votre travail sera évalué à la fois sur la qualité de votre notebook et sur votre compréhension du cours que nous avons étudié ensemble. L'objectif n'est pas seulement d'exécuter du code, mais de démontrer que vous maîtrisez les étapes essentielles d'un processus de l'IA: analyse, préparation des données, modélisation et interprétation des résultats.

L'évaluation se fera selon deux composantes:

1) Évaluation du projet technique (60%):

Présentation Exécutive - 20%: 7 min de presentation

Notebook Technique - 40%: Cette partie correspond au notebook que vous allez rendre. Elle sera notée selon les critères suivants:

1. Structure et organisation du notebook
2. Nettoyage et analyse du dataset
3. Implémentation correcte de l'algorithme demandé
4. Visualisations et interprétation des résultats

5. Clarté et qualité des explications
6. Si l'étudiant ajoute du contenu supplémentaire pertinent, cela sera comptabilisé comme un bonus dans la note.

2) Questions sur le cours (40%):

Je pourrai vous poser quelques questions rapides liées aux notions expliquées en classe. Ces questions permettent de vérifier votre compréhension personnelle du travail réalisé et des concepts fondamentaux enseignés.

L'objectif global est de vous aider à **développer une vraie compétence pratique et théorique** en intelligence artificielle.

Projet de Régression Linéaire avec Validation Croisée - USA Housing Dataset

Objectif

Prédire le prix des maisons aux États-Unis à partir de plusieurs caractéristiques en utilisant la régression linéaire et comparer différentes méthodes de validation croisée, avec une attention particulière à la détection de l'overfitting et de l'underfitting.

Lien dataset: https://www.kaggle.com/datasets/vedavyasv/usa-housing

Étapes à Suivre

1. Importation et Inspection des Données

- Importer le fichier du dataset
- Inspecter le dataset (nombre de lignes, colonnes, types de données)
- Vérifier la présence de valeurs dupliquées
- Afficher les statistiques descriptives (moyenne, écart-type, min/max)

2. Nettoyage et Préparation des Données

- Vérifier et traiter les valeurs manquantes
- Identifier la variable cible et les variables indépendantes
- Analyser la colinéarité entre variables indépendantes

3. Analyse Exploratoire et Visualisation

- Visualiser les relations entre variables (matrice de scatter plots)
- Calculer et interpréter la matrice de corrélation avec heatmap
- Analyser la distribution des variables (histogrammes, boxplots)

4. Préparation des Données pour la Modélisation

- Séparer les variables indépendantes (X) et la variable cible (y)
- Normaliser/standardiser les données (StandardScaler)
- Feature engineering - création de nouvelles variables (si pertinent)

5. Évaluation des Modèles avec Analyse de Overfitting/Underfitting

A. Split Train-Test Simple avec Analyse de Biais/Variance

- Diviser le dataset (80%/20%) avec `random_state=42`
- Entraîner un modèle de régression linéaire
- Calculer les scores sur train et test :
    - Différence Train-Test = indicateur d'overfitting
    - R^2_train vs R^2_test
    - MSE_train vs MSE_test

B. Validation Croisée Standard (3, 5 et 10 folds) avec Analyse par Fold

- Implémenter validation croisée avec 3, 5 et 10 folds
- Pour chaque fold, analyser:
    - Performance sur train du fold
    - Performance sur validation du fold
    - Écart Train-Val pour chaque fold
- Calculer indicateurs d'overfitting/underfitting:
    - Gap moyen = moyenne(R^2_train) - moyenne(R^2_val)
    - Variabilité des écarts entre folds
- Visualisation des écarts Train-Val par fold (bar plot)
- Heatmap des performances par fold

C. Leave-One-Out Cross Validation (LOOCV) avec Analyse de Stabilité

- Implémenter LOOCV
- Analyser la distribution des erreurs
    - Si erreurs très variables → possible overfitting
    - Si erreurs systématiques → possible underfitting

- Calculer:
    - Biais: moyenne des erreurs
    - Variance: variance des prédictions
    - Bias-Variance Tradeoff = biais² + variance
- Plot de l'évolution du modèle au cours de LOOCV

D. Repeated K-Fold Cross Validation avec Analyse de Robustesse

- Implémenter Repeated K-Fold (5 folds, 10 répétitions)
- Analyser la stabilité du biais et de la variance
- Calculer pour chaque répétition:
    - Écart Train-Val moyen
    - Variance des performances
- Boxplot des écarts Train-Val par répétition

6. Analyse Comparative des Méthodes de Validation

Tableau Comparatif Complet des Performance sur Test Set

1. Créer un tableau synthétique comparant toutes les méthodes de validation testées
2. Pour chaque méthode, calculer et reporter trois métriques sur l'ensemble de test uniquement:
    - R^2 (Coefficient de Détermination)
    - RMSE (Root Mean Square Error)
    - MAE (Mean Absolute Error)
3. Visualisations Complémentaires
    - Bar Plot Comparatif :
        - Un graphique à barres pour comparer les R^2
        - Un graphique à barres pour comparer les RMSE
        - Inclure les barres d'erreur (écart-type)

Conclusions à Tirer:

1. Méthode recommandée pour ce dataset?
    - Justifier le choix (performance + stabilité + temps)
    - Ex: "Le 5-Fold CV offre le meilleur équilibre entre performance (R^2=0.85) et stabilité (σ=0.02)"
2. Limitations identifiées?
    - Ex: "Le modèle a des difficultés avec les maisons > $1M (outliers)"
    - Ex: "La performance varie selon les régions (analyse géographique à faire)"

Livrables Demandés

1. Code complet avec commentaires

Fichier: `Projet_Linear_Regression_CV_Housing_NomPrenom.ipynb`

- Notebook exécutable de bout en bout
- Code bien structuré et commenté
- Visualisations claires et professionnelles

2. Présentation Exécutive (5 slides)

- Slide 1: Problème business et objectifs
- Slide 2: Approche méthodologique
- Slide 3: Résultats principaux
- Slide 4: Insights business
- Slide 5: Recommendations et next steps