


# Plan global projet final jedha

👤 Créé par	 marine deldicque
🕒 Heure de création	@June 18, 2025 8:08 PM
🏷️ Étiquettes	

Bien sûr ! Voici la **version enrichie et complète** du plan en 3 blocs pour le projet **SmartInvest**, en intégrant **les données spatiales (scraping/API)**, **les features enrichies**, **MLflow**, et **Streamlit** – le tout **sans perdre la problématique centrale** :

🧠 Problématique : "Comment savoir sur quel logement investir à Paris ?"

## Plan de projet SmartInvest – Organisation en 3 blocs




### Bloc A – Modélisation & Tracking ML (MLflow)

 **Responsable** : Marine (ou membre 1)

#### Objectifs :

- Créer et améliorer un modèle de prédiction du **prix au m<sup>2</sup>**
- Tracker toutes les expérimentations avec **MLflow**
- Identifier les **facteurs les plus influents** pour la prise de décision d'investissement

#### Tâches :


- Tester plusieurs modèles :
  -  Random Forest (avec GridSearch)
  -  Régression Linéaire / Ridge / Lasso
  -  XGBoost ou LightGBM

- Suivi des scores ( $R^2$ , MAE) via **MLflow**
- Sauvegarde du meilleur modèle `.pkl`
- Affichage de l'importance des variables ( `feature_importances_` )

## Dossiers :

```
notebooks/02_model_experiments.ipynb
mlflow/
models/random_forest_model.pkl
```

## ◆ Bloc B – Feature Engineering & Enrichissement Spatial/Temporel

 **Responsable** : Collaborateur 2

### Objectifs :




- Enrichir le dataset avec des données **locales, temporelles et géographiques**
- Fournir des **features différenciantes** pour l'investissement immobilier

### Tâches :

#### Données internes :

- Extraction de `année` , `mois` , `arrondissement`
- Calcul du `prix_m2_moyen_100m` avec **KDTree**
- Détection de biens **sous le prix du quartier**

#### Données externes à intégrer :

- **Latitude/Longitude** par géocodage ( `geopy` , `Nominatim` )
- Scraping ou API pour :
  -  Distance à la station de métro la plus proche
  -  Distance à l'école la plus proche
  -  Comptage des commerces à 300 m

-  Espaces verts à proximité (via Open Data Paris)

## Création de nouvelles colonnes :

Feature	Description
<code>prix_m2_moyen_100m</code>	Moyenne locale dans un rayon de 100 m
<code>distance_metro_m</code>	Distance en mètres à la station de métro la plus proche
<code>nb_commerces_300m</code>	Nombre de commerces dans 300 m
<code>sous_eval</code>	Booléen si le bien est sous le prix médian local

## Dossiers :

notebooks/01\_feature\_engineering.ipynb  
data/enriched\_features.csv


# ◆ Bloc C – Application interactive (Streamlit) & Visualisation

 **Responsable** : Collaborateur 3

## Objectifs :

- Créer une **application Streamlit** pour aider l'utilisateur à savoir **si un bien vaut le coup**
- Proposer des **recommandations** visuelles & chiffrées

## Contenu de l'application :

- Formulaire utilisateur (surface, code postal, pièces, etc.)
- Prédiction du prix/m<sup>2</sup> avec modèle `.pkl`
- Comparaison avec :
  - Moyenne du quartier
  - Moyenne 100 m autour
  - Évolution sur 5 ans (graphique)
- Résumé "  Opportunité ou non ?" basé sur les écarts
- Carte interactive possible (bonus)

## Dossiers :

streamlit\_app/app.py  
plots/

## Arborescence du projet (résumé)

```
SmartInvest/
├── README.md
├── notebooks/
│   ├── 01_feature_engineering.ipynb
│   └── 02_model_experiments.ipynb
├── streamlit_app/
│   └── app.py
├── data/
│   ├── dvf_paris_clean.csv
│   └── enriched_features.csv
├── models/
│   └── random_forest_model.pkl
├── mlflow/
├── requirements.txt
└── .gitignore
```

## Suivi collaboratif

Bloc	Branche GitHub
A – Modélisation + MLflow	ml-dev
B – Feature engineering + enrichissements	features-dev
C – Streamlit App + visualisations	app-dev

Souhaites-tu maintenant :

- Un kanban Notion ou Trello prêt à copier avec toutes ces tâches ?
- Un fichier `.py` ou `.ipynb` pour récupérer les distances avec métro/écoles via OpenStreetMap ou géopy ?

- Un exemple de `app.py` avec formulaire + prédiction ?