**Livret Technique – Big Data Challenge**

**Projet :** Optimisation de la performance énergétique des bâtiments  
**Réalisé par :** Alaa Mazouz, Gilles Adrien KOUEBOU DJONKO, Franck Russell FONGANG, Abrar BOUSLAHI  
**Encadrants :** M. Abdelhadi MIFDAL, M. Arnaud BOUJUT

**1. Contexte du projet**

La transition énergétique en France s’appuie sur la réduction des émissions de gaz à effet de serre, notamment dans le secteur du bâtiment. Ce projet vise à améliorer l’analyse du Diagnostic de Performance Énergétique (DPE) en le comparant aux consommations électriques réelles, pour fiabiliser les prévisions et appuyer les décisions de rénovation.

**2. Objectifs du projet**

**Objectifs métiers :**

* Quantifier l’impact réel des rénovations énergétiques sur la consommation électrique.
* Valider les estimations fournies dans les DPE par des données réelles.

**Objectifs data science :**

* Modéliser la consommation réelle à partir des caractéristiques des logements (DSO2).
* Visualiser les écarts entre estimation et réalité à l’aide de tableaux de bord (DSO5).
* Segmenter les logements selon leurs caractéristiques énergétiques (DSO4).

**3. Architecture du projet**

Le projet suit une architecture typique d’analyse Big Data :

* Collecte de données depuis l’open data (ADEME, ENEDIS, BDNB, BAN)
* Prétraitement et normalisation des données (Python, Jupyter)
* Modélisation prédictive (régression supervisée, clustering)
* Visualisation des résultats (Power BI)
* Suivi de projet (JIRA)

**Une image contenant texte, diagramme, capture d’écran, Police

Le contenu généré par l’IA peut être incorrect.4. Données utilisées**

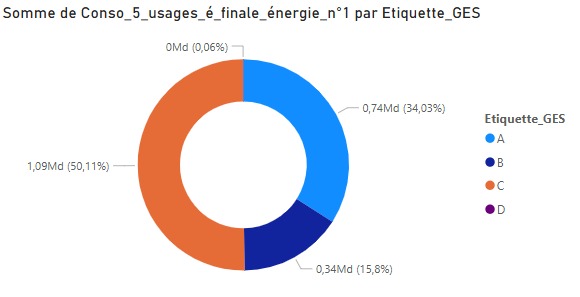
Les principales sources utilisées sont :

* DPE logements neufs (ADEME)
* DPE bâtiments tertiaires (ADEME)
* Données de consommation électrique annuelle (ENEDIS)
* BAN : Base Adresse Nationale (normalisation des adresses)

Les données sont stockées dans l’environnement Data.gov.fr

**5. Méthodologie**

1. Nettoyage et exploration des données (technique)
2. Implémentation de modèles prédictifs avec Azure ML
3. Visualisation avec Power BI

**

**6. Résultats obtenus**

* Métrique des modèles prédictifs : XGBoostRegressor
* Analyse des données avec Power BI
* Application web de prédiction de la consommation d’éclairage avec Streamlit (inachevée)

**7. Guide utilisateur**

**✅ Prérequis techniques** Pour exécuter le projet localement, il vous faut :

* Python ≥ 3.8
* Azure Machine Learning
* Jupyter Notebook (ou Google Colab)
* Power BI Desktop (version récente)
* Bibliothèques Python :

pip install pandas numpy matplotlib seaborn scikit-learn

**📁 Structure du projet** Le dossier contient plusieurs parties organisées comme suit :

* Azure/ → Captures du déploiement dans Azure ML Studio
* Codes/ → Notebooks d’analyse et préparation des données
* modèle\_Optimal\_complet/ → Modèle prédictif local avec environnement
* Modèle\_prediction\_azure/ → Script et modèle pour déploiement Azure
* README.md → Documentation

**🚀 Étapes d’exécution**

1. Ouvrir Jupyter Notebook (ou Google Colab)
2. Lancer Abrar\_dataPrep.ipynb
   * Objectif : fusion, nettoyage, normalisation
   * Résultat : df\_clean.csv
3. Lancer Data Exploration.ipynb
   * Objectif : analyse, modélisation, segmentation
   * Résultat : resultats\_modeles.csv, heatmap.png, etc.
4. (Optionnel) Lancer Franck\_dataPrep.ipynb (variante)

**☁️ Déploiement du modèle sur Azure** Les dossiers « Modèle\_prediction\_azure » et « Azure » contiennent :

* Le script de scoring scoring\_file\_v\_2\_0\_0.py
* L’environnement Conda
* Le modèle model.pkl
* Des captures d’écran du pipeline dans Azure ML Studio

**📊 Visualisation avec Power BI**

1. Ouvrir Power BI Desktop
2. Importer les CSV (df\_clean.csv, resultats\_modeles.csv, ...)
3. Ouvrir dashboard.pbix si fourni, ou créer vos propres visualisations :
   * Comparatif consommation estimée vs réelle
   * Segmentation selon le DPE
   * Heatmaps de corrélation

**📌 Fichiers générés attendus**

* df\_clean.csv, resultats\_modeles.csv, heatmap.png, etc.

**💡 Conseils d’exécution**

* Assurez-vous que tous les CSV sont dans un même dossier
* Sur Google Colab : importer manuellement les CSV dans Power BI
* Data Exploration.ipynb contient des visualisations automatiques

**8. Problèmes rencontrés**

* Normalisation des adresses (API BAN)
* Croisement de données hétérogènes (DPE + ENEDIS)
* Données manquantes ou peu fiables (filtrage nécessaire)
* Limites de représentativité (agrégation à ≥10 logements)

**9. Conclusion et perspectives**

Ce projet a permis d’approfondir l’analyse entre DPE et consommation réelle d’électricité.  
Les modèles prédictifs sont performants et les visualisations facilitent la prise de décision.  
**Perspectives :** déploiement en temps réel, intégration avec des API open data, application web finalisée.