



# Openclassrooms: Projet 4

## Anticipez les besoins en consommation électrique de bâtiments

Denis MUTOMBO TSHITUKA Avril 2025 Parcours «Data scientist»

# Sommaire



## Introduction et méthodologie

Contexte, objectifs et approche générale du projet



## Analyse et modélisation

Exploration des données, modèles de consommation et d'émissions



## ENERGY STAR Score

Analyse de la pertinence de cet indicateur



## Résultats et recommandations

Performances, facteurs déterminants et conseils pour Seattle

Notre présentation suit une progression logique, depuis l'introduction du projet jusqu'aux recommandations finales. Nous examinerons en détail chaque étape de notre analyse et les implications pratiques pour la ville de Seattle.



# Introduction et Objectifs

## Contexte

La ville de Seattle s'est fixé un objectif ambitieux de neutralité carbone d'ici 2050. Les bâtiments non résidentiels constituent une source majeure d'émissions urbaines, mais l'obtention de relevés énergétiques précis reste coûteuse et complexe à grande échelle.

## Objectifs du projet

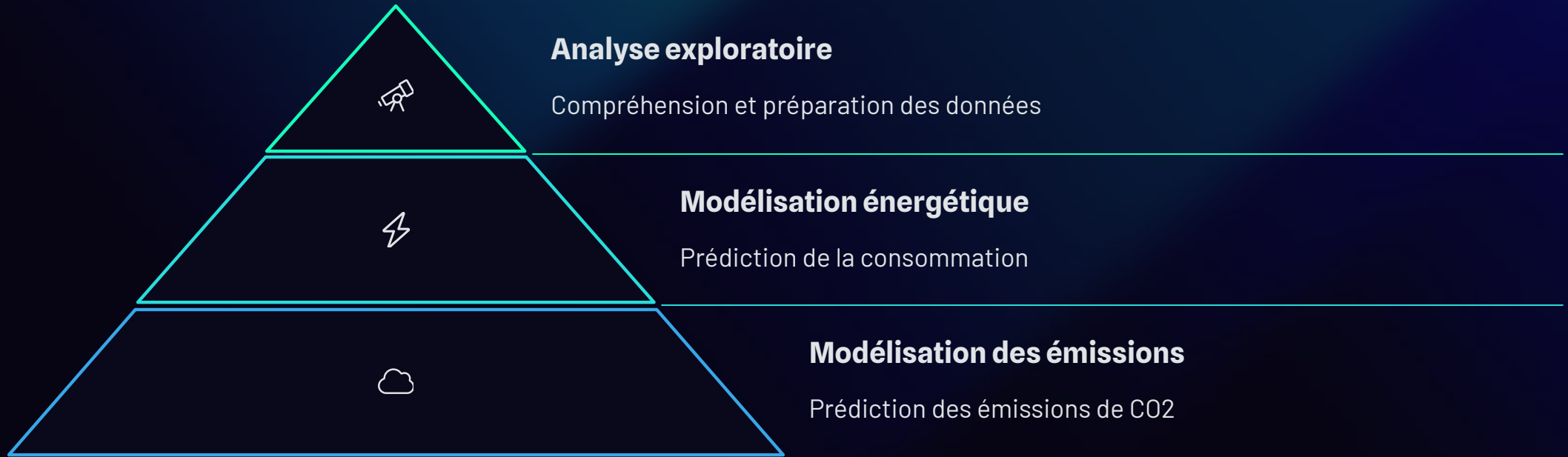
Développer des modèles prédictifs fiables pour estimer la consommation énergétique des bâtiments sans mesures directes. Calculer les émissions de CO<sub>2</sub> associées à partir de données déclaratives facilement accessibles.

## Applications visées

Évaluer l'efficacité de l'ENERGY STAR Score comme indicateur prédictif et identifier les facteurs les plus influents pour orienter efficacement les politiques publiques de réduction des émissions carbone.

Ce projet s'inscrit dans une démarche d'optimisation des ressources de la ville, en permettant de cibler les efforts de mesure et d'intervention sur les bâtiments présentant le plus fort potentiel d'amélioration énergétique.

# Méthodologie et Données



Notre approche méthodologique s'articule autour de trois phases complémentaires, formant une pyramide d'analyse où chaque niveau s'appuie sur les précédents. Nous avons utilisé les relevés énergétiques de Seattle pour les années 2015-2016.

Les données comprennent diverses variables déclaratives (taille, usage, année de construction), des variables énergétiques (consommation par source, intensité) et l'ENERGY STAR Score comme indice de performance standardisé. Cette méthodologie progressive nous permet de construire des modèles robustes tout en comprenant les relations sous-jacentes.



# 1. Analyse Exploratoire

## Prétraitement des données

Le jeu de données initial présentait des défis importants, notamment une forte proportion de valeurs manquantes dépassant 50% dans certaines colonnes. Nous avons procédé à un nettoyage rigoureux, incluant la conversion des types de données et la création de variables dérivées pertinentes comme l'âge des bâtiments.

## Observations principales

Notre analyse a révélé une distribution fortement asymétrique des consommations et émissions, nécessitant des transformations logarithmiques. La taille des bâtiments s'est avérée fortement corrélée à leur consommation totale, mais avec des variations significatives selon le type d'usage. L'âge du bâtiment montre une influence modérée sur la performance énergétique.

## Visualisations réalisées

Nous avons développé des visualisations complexes pour comprendre les distributions des variables cibles. Les matrices de corrélation ont permis d'identifier les relations entre variables, complétées par des analyses segmentées par type de bâtiment pour capturer les spécificités sectorielles.

## 2. Modélisation de la Consommation



### Préparation des données

Variable cible: Intensité énergétique (Site EUI)



### Construction de pipelines

Traitement numérique et catégoriel



### Évaluation d'algorithmes

5 modèles de régression comparés



### Optimisation

Hyperparamètres affinés

Notre approche de modélisation a ciblé l'intensité énergétique comme variable à prédire. Les algorithmes de Random Forest et Gradient Boosting ont démontré les meilleures performances avec un coefficient de détermination ( $R^2$ ) d'environ 0,85, surpassant significativement la régression linéaire ( $R^2 \sim 0,70$ ).

L'analyse d'importance des variables a révélé que le type d'usage du bâtiment, sa surface totale, le nombre d'étages et la source d'énergie principale constituent les facteurs les plus déterminants de la consommation énergétique.

# 3. Modélisation des Émissions de CO2



## Spécificités méthodologiques

Pour la modélisation des émissions de gaz à effet de serre, nous avons appliqué une transformation logarithmique à la variable cible pour normaliser sa distribution fortement asymétrique. Cette approche a permis d'améliorer significativement la qualité des prédictions.



## Performance des modèles

L'algorithme de Gradient Boosting s'est distingué comme le plus performant, atteignant un coefficient de détermination exceptionnel d'environ 0,90. L'optimisation via GridSearchCV a permis un gain supplémentaire de 3 à 5% en performance prédictive.



## Facteurs d'influence identifiés

Notre analyse a révélé que le mix énergétique utilisé par le bâtiment constitue un facteur déterminant des émissions, conjointement avec la surface totale qui reste le facteur dominant. Le type de bâtiment influence fortement l'intensité des émissions.

La validation croisée a confirmé la robustesse de nos modèles, assurant leur fiabilité pour des applications pratiques à l'échelle de la ville.

# Analyse de l'ENERGY STAR Score

## Test comme prédicteur unique

Performance:  $R^2 \sim 0.60$  pour consommation,  
 $\sim 0.55$  pour CO2

## Intérêt communicationnel

Standardisé et reconnu par le public



## Comparaison avec modèles complets

Écart significatif de  $\sim 0.25$  points de  $R^2$

## Utilité comme filtre

Indicateur de premier niveau efficace

Notre analyse spécifique de l'ENERGY STAR Score révèle une corrélation négative significative avec la consommation énergétique ( $-0,75$ ), confirmant sa pertinence comme indicateur de performance. Cependant, utilisé seul, il présente un écart de performance d'environ 0,25 points de  $R^2$  par rapport aux modèles complets.

Nous recommandons donc d'utiliser ce score comme filtre initial pour identifier rapidement les bâtiments nécessitant une analyse approfondie, tout en le complétant par d'autres variables pour des prédictions plus précises.



# Facteurs Déterminants Identifiés

## Pour la consommation énergétique

1. Type d'usage du bâtiment (bureaux, hôpitaux, data centers)
2. Surface totale du bâtiment
3. Année de construction/rénovation
4. Nombre d'étages
5. ENERGY STAR Score

## Pour les émissions de CO2

1. Surface totale du bâtiment
2. Sources d'énergie utilisées
3. Type d'usage du bâtiment
4. ENERGY STAR Score
5. Localisation géographique

Notre analyse d'importance des variables révèle des différences notables entre les facteurs influençant la consommation énergétique et ceux déterminant les émissions de CO2. Pour la consommation, l'usage du bâtiment apparaît comme le facteur principal, suivi de près par la surface totale.

En revanche, pour les émissions, la surface reste dominante mais le mix énergétique prend une importance considérable. Cette distinction est cruciale pour élaborer des stratégies différenciées selon que l'on cible la réduction de consommation ou la diminution des émissions carbone.

# Recommandations pour Seattle



## Applications des modèles

Nous recommandons à la ville de Seattle d'utiliser nos modèles prédictifs pour prioriser les inspections des bâtiments à haut potentiel d'économies. Un système de notation énergétique basé sur ces prédictions permettrait également de cartographier les émissions estimées de l'ensemble du parc immobilier.



## Politiques publiques

Les incitations à la rénovation devraient cibler en priorité les types de bâtiments identifiés comme les plus émetteurs. Les standards d'efficacité gagneraient à être adaptés par catégorie de bâtiment, avec un accent particulier sur la promotion des transitions vers des sources d'énergie moins carbonées.



## Collecte de données

Nous suggérons de concentrer les efforts de collecte sur les variables identifiées comme les plus prédictives, tout en standardisant les protocoles de mesure. Un suivi temporel permettrait d'évaluer efficacement l'impact des mesures mises en place.

Ces recommandations visent à optimiser l'allocation des ressources limitées de la ville pour maximiser l'impact des politiques énergétiques sur la réduction des émissions de CO<sub>2</sub>.

# Perspectives et Améliorations

## Limites actuelles

Notre étude présente certaines limites à considérer. Les données sont restreintes à la période 2015-2016, ce qui ne permet pas d'analyser les évolutions temporelles. Nous notons également l'absence de certaines variables potentiellement influentes comme les données d'occupation réelle.



## Prochaines étapes

Nous recommandons de tester la généralisation de nos modèles sur d'autres villes aux caractéristiques similaires. Le développement d'un outil web de simulation permettrait aux propriétaires d'estimer l'impact de différentes interventions sur leur performance énergétique.



## Pistes d'amélioration

L'intégration de données météorologiques (degrés-jours) permettrait d'affiner les modèles en tenant compte des variations climatiques. Les méthodes d'apprentissage profond pourraient capturer des relations plus complexes, tandis qu'un modèle combiné consommation/émissions offrirait une approche plus intégrée.

Un suivi longitudinal serait également précieux pour évaluer l'évolution des performances énergétiques et l'efficacité des politiques mises en œuvre par la ville de Seattle.

# Conclusion

**85%**

## **Précision consommation**

Performance du modèle optimal

**90%**

## **Précision émissions**

Fiabilité des prédictions CO2

**2050**

## **Objectif neutralité**

Horizon pour Seattle

Nos travaux ont permis de développer des modèles prédictifs hautement performants pour estimer à la fois la consommation énergétique et les émissions de CO2 des bâtiments non résidentiels de Seattle. L'ENERGY STAR Score s'est confirmé comme un indicateur utile mais insuffisant s'il est utilisé seul.

L'identification précise des facteurs clés permet désormais de cibler efficacement les actions de réduction des émissions. L'impact attendu inclut une optimisation des campagnes de mesure et d'audit, des politiques publiques mieux ciblées, et une contribution significative à l'ambitieux objectif de neutralité carbone pour 2050.

Cette étude démontre la puissance de la data science au service de la transition écologique urbaine, avec une approche adaptable à d'autres villes et contextes.