École nationale de la Statistique et de l'Analyse économique-Pierre Ndiaye

Machine learning 1

Prédiction de l'émission du CO_2 dans la ville de Seattle

Rédigé par :

Moussa DIEME

Ben Idriss SOMA

Papa Ahmadou NIANG

Prosper LAWA FOUMSOU

Tamsir NDONG

Élèves en ISE 2

Sous la supervision de :

Mme Mously DIAW

Enseignante à l'ENSAE

Décembre 2025

Introduction

Contexte et justification

Le changement climatique constitue l'un des principaux défis contemporains, et les émissions de dioxyde de carbone (CO_2) issues des activités humaines en sont un facteur déterminant. Dans les zones urbaines, les bâtiments résidentiels, commerciaux et industriels représentent une part importante de la consommation énergétique et des émissions de gaz à effet de serre (GES). La ville de Seattle, située dans l'État de Washington aux États-Unis, s'est engagée à lutter contre le changement climatique en réduisant ses émissions de GES et en atteignant la neutralité carbone d'ici 2050. Pour suivre ses progrès et orienter ses politiques publiques, Seattle produit régulièrement un inventaire des émissions de GES (Greenhouse Gas Inventory), mis à jour tous les deux ans, couvrant les principaux secteurs : transports, bâtiments et déchets.

Le secteur du bâtiment est responsable de près d'un tiers des émissions mondiales de CO_2. À Seattle, il représente environ 40 % des émissions locales, toutes catégories confondues, juste derrière le secteur des transports. Afin de favoriser l'analyse et la transparence, la ville met à disposition des données ouvertes issues du Seattle Building Energy Benchmarking, qui décrivent les caractéristiques des bâtiments (surface, usage, année de construction), leur consommation énergétique et leurs émissions de CO_2.

Dans ce contexte, le présent projet vise à exploiter les techniques de machine learning pour prédire les émissions de CO_2 des bâtiments non destinés à l'habitation de Seattle à partir de leurs caractéristiques. Un tel modèle pourra servir d'outil d'aide à la décision pour les autorités publiques et les urbanistes, en identifiant les facteurs clés d'émissions et en orientant les politiques de réduction carbone.

Problématique

Les méthodes traditionnelles d'évaluation des émissions de CO_2 reposent sur des données mesurées, telles que les relevés de compteurs énergétiques ou les factures de consommation. Bien que précises, ces informations ne sont disponibles qu'après la période de consommation et ne permettent pas d'anticiper les émissions futures d'un bâtiment, notamment lors de projets de construction ou de rénovation.

Par ailleurs, la réalisation d'audits énergétiques exhaustifs est coûteuse et difficilement généralisable à l'ensemble du parc immobilier. Dans ce contexte, la capacité à prédire les émissions de CO_2 à partir de caractéristiques structurelles observables constitue un enjeu majeur pour l'aide à la décision publique. Il apparaît donc logique de se poser les questions suivantes : peut-on prédire avec précision les émissions de CO_2 des bâtiments non résidentiels de la ville de Seattle à partir de données déclaratives et structurelles ? Peut-on se passer de certaines variables coûteuses à calculer, comme le **ENERGY STAR Score** ?

Objectifs du projet

L'objectif est de développer un modèle de Machine Learning capable de prédire les émissions de CO_2 des bâtiments non destinés à l'habitation. Pour répondre aux enjeux stratégiques susmentionnés, l'équipe technique se fixe quatre objectifs opérationnels précis, mesurables et temporellement définis.

> **Modélisation Prédictive** : le premier objectif consiste à construire un modèle robuste capable d'estimer les émissions de gaz à effet de serre (totales ou par unité de surface) en se basant sur les caractéristiques structurelles et déclaratives du bâtiment, sans avoir nécessairement accès aux relevés de compteurs détaillés au moment de la prédiction. Cela permettrait d'identifier les bâtiments énergivores nécessitant une intervention prioritaire dans le cadre du **Building Emissions Performance Standard** (BEPS).

Analyse de l'Energy Star Score : le second objectif porte spécifiquement sur l'analyse de l'Energy Star Score. Une évaluation rigoureuse de l'impact de cet indicateur sur la précision prédictive sera menée en comparant systématiquement les performances de modèles incluant versus excluant cette variable. L'objectif est de quantifier précisément le compromis entre simplicité opérationnelle (modèle sans Energy Star Score) et gain prédictif potentiel, permettant ainsi une décision éclairée sur l'architecture finale du système.

Interprétabilité des Résultats : le troisième objectif consiste à identifier et hiérarchiser les variables explicatives qui influencent le plus la consommation via des méthodes d'explicabilité globale et locale, afin de fournir des leviers d'action clairs aux décideurs.

Industrialisation du Pipeline : le quatrième objectif vise l'industrialisation complète du pipeline de machine learning. Au-delà de la phase expérimentale, le projet doit aboutir à une architecture logicielle reproductible, versionnable et maintenable. Cela implique l'automatisation de l'ensemble de la chaîne de traitement (ingestion des données, preprocessing, entraînement, évaluation, déploiement) selon les principes MLOps, garantissant que le système puisse être mis à jour automatiquement lors de l'arrivée de nouvelles données ou de l'évolution des exigences métier.

Dans la suite de ce document, nous présenterons en premier lieu la revue de littérature justifiant le choix des principales variables explicatives de notre variable cible, en second lieu, nous présenterons les données utilisées dans le cadre de notre travail puis en dernier lieu nous aborderons la méthodologie adoptée.

Revue de la littérature

La littérature consacrée à la performance énergétique et environnementale des bâtiments met en évidence que les émissions de dioxyde de carbone (CO_2) résultent d'un ensemble de facteurs structurels, fonctionnels, énergétiques, technologiques et contextuels. En premier lieu, les variables morphologiques et structurelles apparaissent comme des déterminants majeurs des émissions. La surface du bâtiment (Gross Floor Area) est systématiquement identifiée comme le prédicteur le plus robuste et le plus stable des consommations énergétiques et des émissions associées, en raison de son lien direct avec les besoins en chauffage, climatisation et éclairage (Perez-Lombard, Ortiz et Pout, 2008 ; IEA, 2019). De nombreuses études empiriques montrent que la relation entre surface et émissions est quasi proportionnelle, bien que des effets de non-linéarité puissent apparaître pour les bâtiments de très grande taille (Hong, Taylor-Lange, D'Oca, Yan et Corgnati, 2015). Aussi, l'année de construction et, le cas échéant, l'année de rénovation, jouent également un rôle significatif : les bâtiments récents tendent à présenter de meilleures performances énergétiques en raison de normes de construction plus strictes et de technologies plus efficaces (Filippín, 2000 ; IPCC, 2022).

Les variables fonctionnelles liées à l'usage du bâtiment constituent un second groupe fondamental de prédicteurs. La littérature souligne que le type d'usage (bureaux, établissements scolaires, hôpitaux, commerces, entrepôts) influence fortement le niveau d'émissions, indépendamment de la taille du bâtiment (Santamouris et al., 2010). À surface équivalente, les bâtiments à usage intensif, tels que les hôpitaux ou les centres de données, présentent des émissions nettement supérieures à celles des bâtiments administratifs ou logistiques, en raison d'horaires d'occupation prolongés, d'équipements spécifiques et de besoins énergétiques continus (Chung, Rhee et Im, 2011). Ces résultats justifient l'intégration de variables catégorielles décrivant l'usage principal et secondaire dans les modèles de prédiction.

Par ailleurs, les variables énergétiques constituent le lien le plus direct entre les caractéristiques du bâtiment et les émissions de CO_2. La consommation d'énergie, qu'il s'agisse d'électricité, de gaz naturel ou d'autres combustibles, est un déterminant central, les émissions étant proportionnelles à la quantité d'énergie consommée et au facteur d'émission associé à chaque vecteur énergétique (IEA, 2021). Plusieurs études insistent sur l'importance de distinguer les différentes sources d'énergie, notamment dans les régions où le mix électrique est relativement décarboné, ce qui renforce le rôle explicatif de la consommation de combustibles fossiles comme le gaz naturel (Gurney et al., 2012). La littérature recommande également l'utilisation de variables dérivées, telles que l'intensité énergétique (Energy Use Intensity, EUI), définie comme le ratio entre la consommation totale d'énergie et la surface du bâtiment, afin de neutraliser l'effet de taille et de faciliter la comparaison entre bâtiments hétérogènes (Chung, 2011 ; Hong et al., 2015).

À un niveau plus agrégé, des indicateurs synthétiques de performance énergétique, tels que le ENERGY STAR Score, sont largement utilisés dans les démarches de benchmarking énergétique, en particulier aux États-Unis (U.S. EPA, 2018). Toutefois, plusieurs travaux montrent que ces scores apportent une information marginale limitée dans les modèles prédictifs une fois les variables de consommation, de surface et d'usage incluses, en raison de leur construction basée sur ces mêmes informations et des problèmes de colinéarité qui en résultent (Kontokosta, 2012 ; Robinson et al., 2017). À Seattle, l'électricité est majoritairement d'origine hydroélectrique, ce qui entraîne un facteur d'émission de CO_2 très faible. Ainsi, une consommation électrique élevée n'implique pas nécessairement des émissions importantes, contrairement aux sources énergétiques fossiles comme le gaz naturel.

Enfin, la littérature souligne le rôle des variables technologiques et contextuelles, telles que les systèmes de chauffage, ventilation et climatisation (CVC), la qualité de l'enveloppe thermique, ainsi que les conditions climatiques locales, dans l'explication des émissions de CO_2 des bâtiments (Santamouris, 2015 ; IPCC, 2022). Bien que ces

facteurs soient parfois observés de manière imparfaite dans les bases de données administratives, ils contribuent à expliquer les différences résiduelles de performance énergétique entre bâtiments comparables. L'ensemble de ces résultats justifie très souvent le recours à des modèles de machine learning non linéaires, capables de capturer les interactions complexes entre la taille, l'usage, l'énergie et la technologie, pour la prédiction des émissions de CO_2 des bâtiments non résidentiels.

Champ de l'étude et données

Champ de l'étude

L'étude porte principalement sur les bâtiments non résidentiels de la ville de Seattle (bureaux, établissements scolaires, hôpitaux, bâtiments commerciaux et industriels). Ces bâtiments présentent une forte hétérogénéité de consommation énergétique et constituent une cible prioritaire des politiques publiques de décarbonation.

Source des données

Les données utilisées proviennent du Seattle Building Energy Benchmarking Dataset pour l'année 2016, mis à disposition en open data par la ville de Seattle. Les données utilisées dans cette étude proviennent des déclarations administratives associées aux permis d'exploitation commerciale des bâtiments non résidentiels. Ces données, fournies annuellement par les propriétaires ou exploitants, décrivent les caractéristiques physiques, l'usage et les consommations énergétiques des bâtiments. Bien qu'elles ne reposent pas sur des mesures en temps réel, elles constituent une source d'information standardisée et largement utilisée par les collectivités pour le suivi de la performance énergétique et environnementale du parc immobilier.

Variable cible

La variable cible du modèle est le niveau total annuel d'émissions de gaz à effet de serre, mesuré en tonnes de CO_2 équivalent (**TotalGHGEmissions**). Cette variable synthétise l'impact environnemental direct et indirect des consommations énergétiques du bâtiment, en tenant compte des facteurs d'émission associés aux différentes sources d'énergie utilisées. Le choix de cette variable se justifie par sa centralité dans les objectifs de neutralité carbone poursuivis par les collectivités locales et par sa disponibilité homogène dans la base de données étudiée.

La variable **GHGEmissionsIntensity** bien que portant sur les émissions de CO_2 introduit un niveau de complexité supplémentaire car il combine les émissions totales avec d'autres variables (ex. surface, consommation d'énergie). Cela peut créer une corrélation forte avec les variables explicatives, rendant le modèle plus instable. Dans ce contexte, le choix de cette variable comme variable ciblée serait mal adapté.

Variables explicatives

Conformément à la revue de littérature, les variables explicatives sont regroupées en plusieurs catégories. Les variables morphologiques et structurelles incluent notamment la surface totale du bâtiment (Gross Floor Area), le nombre d'étages et l'année de construction ou de rénovation, qui capturent l'effet de taille et les évolutions technologiques du bâti. Les variables fonctionnelles décrivent l'usage principal et secondaire du bâtiment (bureaux, commerces, établissements de santé, etc.), reflétant les différences d'intensité d'occupation et de besoins énergétiques. Les variables énergétiques correspondent aux consommations annuelles d'électricité et de gaz naturel, ainsi qu'à des indicateurs dérivés tels que l'intensité énergétique (Energy Use Intensity, EUI). Enfin, un indicateur synthétique de performance énergétique, le ENERGY STAR Score, est intégré de manière optionnelle afin d'évaluer son apport marginal à la performance prédictive des modèles.

Méthodologie

Présentation générale de la démarche

La méthodologie adoptée repose sur une démarche structurée, progressive et reproductible, couvrant l'ensemble du cycle de vie d'un projet de data science : depuis le traitement des données brutes jusqu'au déploiement du modèle sous forme d'API et de tableau de bord interactif. Afin de garantir la clarté, la traçabilité et une collaboration efficace au sein de l'équipe, le travail est organisé en plusieurs notebooks distincts, chacun correspondant à une étape précise de l'analyse.

Traitement et préparation des données

Chargement et exploration initiale des données

Les données brutes sont importées dans les notebooks initiaux afin d'obtenir une première vue d'ensemble du dataset. Cette étape permet d'identifier la structure générale des données, les types de variables (numériques, catégorielles, booléennes) ainsi que les éventuelles incohérences telles que les valeurs manquantes, les doublons ou les formats incorrects. Cette exploration initiale constitue une étape essentielle pour orienter les choix méthodologiques ultérieurs.

Nettoyage des données

Un processus de nettoyage rigoureux est mis en œuvre afin d'assurer la fiabilité des analyses. Il comprend notamment : la correction ou la suppression des valeurs aberrantes évidentes, l'harmonisation des noms de variables, la gestion des valeurs manquantes selon la nature des variables (suppression, imputation ou conservation), la conversion des variables vers des types de données appropriés (entiers, flottants, catégories). Ce travail permet d'obtenir un jeu de données cohérent et exploitable pour les analyses exploratoires.

Préparation et enrichissement des variables

À cette étape, certaines variables dérivées sont créées afin d'enrichir l'information disponible, par exemple l'âge du bâtiment à partir de son année de construction. Les variables redondantes ou non informatives sont identifiées et supprimées. Le jeu de données final issu de cette phase sert de base commune pour l'ensemble des analyses exploratoires.

Analyse exploratoire des données (EDA)

L'analyse exploratoire des données est menée de manière progressive, en allant de l'analyse univariée vers des analyses plus complexes, afin de mieux comprendre la structure des données avant toute modélisation.

Analyse univariée

L'analyse univariée consiste à étudier chaque variable individuellement afin de comprendre son comportement statistique. Les méthodes utilisées incluent le calcul de statistiques descriptives (moyenne, médiane, écart-type, quartiles), la visualisation des distributions à l'aide d'histogrammes et de boxplots, l'analyse de la normalité à l'aide de QQ-plots et de tests statistiques, ainsi que l'étude de la proportion de valeurs manquantes. Cette étape permet d'identifier les variables fortement asymétriques, la présence de valeurs extrêmes et les transformations nécessaires avant la modélisation, notamment pour la variable cible des émissions de CO_2.

Analyse bivariée

L'analyse bivariée vise à étudier les relations entre deux variables afin d'identifier les dépendances et les corrélations significatives. Elle repose sur des visualisations telles que les scatterplots, les boxplots conditionnels et les matrices de corrélation. Des comparaisons par groupes sont également réalisées selon le type de bâtiment, l'usage principal ou le quartier. Cette analyse permet de mettre en évidence des patterns explicatifs et d'identifier les variables les plus pertinentes pour la prédiction des

émissions.

Analyse multivariée et feature engineering

L'analyse multivariée a pour objectif d'explorer les interactions complexes entre plusieurs variables et de préparer un jeu de données optimisé pour la modélisation. Elle inclut la sélection de variables, ainsi que la création de nouvelles variables plus informatives. Cette phase vise à améliorer à la fois la performance prédictive des modèles et leur interprétabilité.

Analyse spatiale et temporelle

Une analyse spatiale et temporelle est menée afin d'intégrer les dimensions géographiques et temporelles dans l'étude des émissions de CO_2. Cette étape permet d'analyser les différences entre quartiers, d'identifier d'éventuelles disparités territoriales et d'étudier l'évolution des performances énergétiques dans le temps.

Modélisation prédictive

Définition des variables cibles

Deux stratégies de modélisation sont explorées dans le projet : la prédiction des émissions totales de gaz à effet de serre, la prédiction du score de performance énergétique ENERGY STAR. Ces deux approches permettent de comparer des objectifs de modélisation complémentaires.

Modèles de Machine Learning testés

Plusieurs familles de modèles sont évaluées, notamment des modèles linéaires, des modèles basés sur les arbres de décision et des méthodes d'ensemble. Chaque modèle est entraîné sur les mêmes données afin de permettre une comparaison équitable.

Évaluation des performances

Les performances des modèles sont évaluées à l'aide d'une séparation des données en

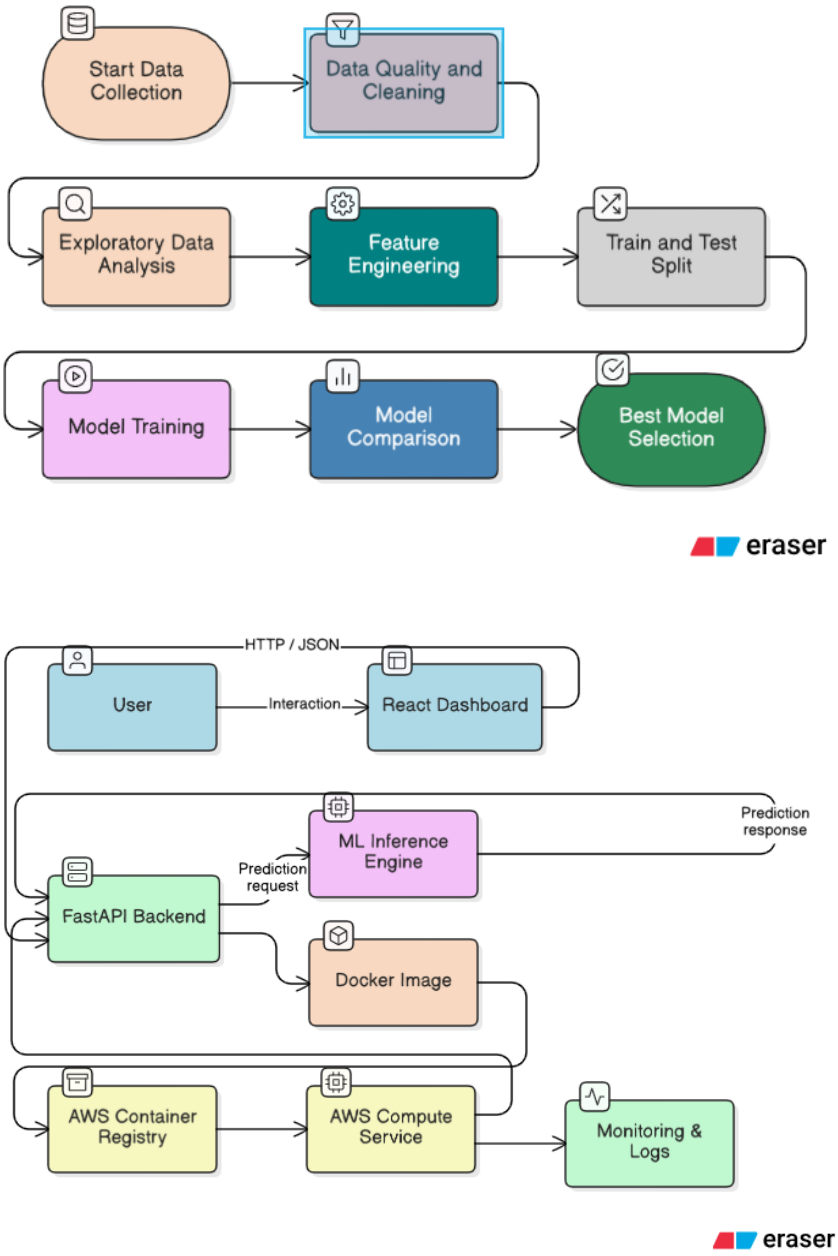
ensembles d'entraînement et de test, complétée par des procédures de validation croisée. Les indicateurs utilisés incluent notamment la RMSE, la MAE et le coefficient de détermination R^2. Le modèle final est sélectionné sur la base d'un compromis entre performance prédictive et interprétabilité.

ML Ops : déploiement et valorisation

La dernière phase du projet vise à rendre le modèle exploitable en dehors des notebooks. Elle comprend le développement d'une API de prédiction, la création d'un tableau de bord interactif et la mise en place d'un pipeline reproductible. Cette étape permet de transformer le projet académique en une solution opérationnelle.

La méthodologie adoptée dans ce projet repose sur une progression logique et rigoureuse, une séparation claire des responsabilités entre les notebooks et une cohérence forte entre l'analyse des données, la modélisation et le déploiement. Cette approche garantit la robustesse des résultats, la reproductibilité du travail et une collaboration efficace au sein de l'équipe projet.

Figure 1 : Diagramme illustrant le projet



Source : Travaux des auteurs

Conclusion

Ce projet s'inscrit dans une démarche analytique appliquée aux enjeux environnementaux urbains. En mobilisant des techniques de machine learning sur des données de benchmarking énergétique, il vise à évaluer la capacité de modèles prédictifs à anticiper les émissions de CO_2 des bâtiments et à éclairer les politiques publiques de transition énergétique de la ville de Seattle.

Table des matières