**Anticipez les besoins en consommation de bâtiments**

**Enoncé :**

Vous travaillez pour la ville de Seattle.

Pour atteindre son objectif de ville neutre en émissions de carbone en 2050, votre équipe s’intéresse de près à la consommation et aux émissions des bâtiments non destinés à l’habitation.

Des relevés minutieux ont été effectués par les agents de la ville en 2016.

Cependant, ces relevés sont coûteux à obtenir, et à partir de ceux déjà réalisés, vous voulez tenter de prédire **les émissions de CO2** et **la consommation totale d’énergie** des bâtiments non destinés à l’habitation pour lesquels elles n’ont pas encore été mesurées.

Votre prédiction se basera sur les données structurelles des bâtiments (taille et usage des bâtiments, date de construction, situation géographique, ...)

Vous cherchez également à évaluer l’intérêt de l’"ENERGY STAR Score" pour la prédiction d’émissions (expliqué sur ce site en anglais : www.energystar.gov), qui est fastidieux à calculer avec l’approche utilisée actuellement par votre équipe.

Vous l'intégrerez dans la modélisation et jugerez de son intérêt.

Vous sortez tout juste d’une réunion de brief avec votre équipe. Voici un récapitulatif de votre mission :

> Réaliser une courte analyse exploratoire.

> Tester différents modèles de prédiction afin de répondre au mieux à la problématique.

Avant de quitter la salle de brief, Douglas, le ‘project lead’, vous donne quelques pistes et erreurs à éviter :

*L’objectif est de te passer des relevés de consommation annuels futurs (attention à la fuite de données). Nous ferons de toute façon pour tout nouveau bâtiment un premier relevé de référence la première année, donc rien ne t'interdit d’en déduire des variables structurelles aux bâtiments, par exemple la nature et proportions des sources d’énergie utilisées...*

*Fais bien attention au traitement des différentes variables, à la fois pour trouver de nouvelles informations (peut-on déduire des choses intéressantes d’une simple adresse ?) et optimiser les performances en appliquant des transformations simples aux variables (normalisation, passage au log, etc.).*

*Mets en place une évaluation rigoureuse des performances, et optimise les hyperparamètres et le choix d’algorithmes de ML à l’aide d’une validation croisée. Tu testeras au minimum 4 algorithmes de familles différentes (par exemple : ElasticNet, SVM, GradientBoosting, RandomForest).*

**Livrables :**

Un notebook de l'analyse exploratoire mis au propre et annoté.

Un notebook pour chaque prédiction (émissions de CO2 et consommation totale d’énergie) des différents tests de modèles mis au propre, dans lequel vous identifierez clairement le modèle final choisi.

Un support de présentation pour la soutenance (entre 15 et 25 slides).

**Référentiel d'évaluation :**

Compétences évaluées :

> Créer les variables pertinentes d'un modèle d'apprentissage supervisé ou non-supervisé

> Sélectionner, entraîner et évaluer des modèles d’apprentissage supervisé

Sélectionner, transformer et créer les variables pertinentes d'un modèle d'apprentissage supervisé ou non supervisé (feature engineering) afin de réaliser un apprentissage optimal.

CE1 Vous avez identifié les variables catégorielles (qualitatives).

CE2 Les variables catégorielles identifiées ont été transformées en fonction du besoin (par exemple via OneHotEncoder ou TargetEncoder).

CE3 Vous avez créé de nouvelles variables à partir de variables existantes.

CE4 Vous avez réalisé des transformations mathématiques lorsque c'est requis pour transformer les distributions de variables.

CE5 Vous avez normalisé les variables lorsque c'est requis.

Sélectionner, entraîner et évaluer des modèles d’apprentissage supervisé, afin de choisir le modèle le plus performant pour la problématique métier.

CE1 Vous avez choisi la ou les variables cibles pertinentes. Cela signifie dans le cadre de ce projet :

> Vous avez choisi la variable cible pour la consommation totale d’énergie,

> Vous avez choisi la variable cible pour la consommation de CO2.

CE2 Vous avez vérifié qu'il n’y a pas de problème de 'data leakage' (c'est-à-dire, des variables trop corrélées à la variable cible et inconnues a priori dans les données en entrée du modèle).

CE3 Vous avez choisi des métriques adaptées pour évaluer la performance d'un algorithme (par exemple : R2, RMSE, MAE en régression.)

CE4 Vous avez exploré d'autres indicateurs de performance que le score pour comprendre les résultats (par exemple : coefficients des variables en fonction de la pénalisation, visualisation des erreurs en fonction des variables du modèle, temps de calcul...).

CE5 Vous avez séparé les données en 'train/test' pour les évaluer de façon pertinente et détecter l'overfitting ('surajustement').

CE6 Vous avez mis en place une validation croisée (via GridsearchCV, RandomizedSearchCV ou équivalent) afin de réaliser une Cross Validation sur le set de 'train' et d’optimiser les hyperparamètres.

CE7 Vous avez testé au minimum 4 algorithmes de familles différentes (par exemple : ElasticNet, SVM, GradientBoosting, RandomForest). Vous avez présenté une synthèse comparative de l'ensemble des résultats. Vous avez justifié le choix final de l'algorithme et des hyperparamètres.

CE8 Vous avez réalisé l’analyse de l’importance des variables (feature importance) globales sur l’ensemble du jeu de données et variables locales sur chaque individu du jeu de données.