IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

FISE - A2

UE BC : A Journey To A Data Scientist

Modeling & Performance of algorithms

Groupe n°4

Sommaire

I. Introduction et problématique

La pollution de l'air, du sol et de l'eau en France représente un défi majeur pour la santé publique, avec des impacts graves sur diverses maladies, notamment certains types de cancers. Ce projet a pour objectif d'analyser les liens entre les niveaux de pollution et la prévalence de deux types de cancer : le cancer broncho-pulmonaire et le cancer colorectal . L'objectif est de fournir aux agences régionales de santé (ARS) des outils permettant d'anticiper les taux de ces maladies de ces cancers afin de mieux adapter leurs ressources médicales et aussi de les aider à expliquer l'apparition de ces maladies .

En traduisant cette problématique en une approche de data science, l'objectif est de mieux comprendre les interactions entre les polluants et la santé, et de développer des modèles prédictifs capables d'estimer la prévalence de ces cancers à partir des données de pollution des années précédentes. Les résultats attendus incluent l'identification des polluants les plus fortement associés à des problèmes de santé spécifiques.

De plus, ce projet fournira des prédictions sur la prévalence des cancers dans différentes régions, permettant ainsi d'orienter les actions sanitaires de manière proactive. Ces analyses visent à renforcer l'efficacité des efforts de réduction des risques liés à la pollution et à protéger la santé publique.

Problématique

Comment la pollution de l'air et de l'eau influence-t-elle l'incidence des cancers broncho-pulmonaire et colorectal dans les régions les plus touchées ? L'objectif est d'aider les agences régionales à mieux comprendre ces liens et à ajuster leurs stratégies de sensibilisation et d'allocation de ressources médicales en fonction des spécificités de chaque département .

Problématique Data Science

Pour répondre à cette problématique, l'approche data science se concentre sur l'analyse des corrélations entre les indicateurs de pollution et la prévalence des maladies et d' essayer de construire un modèle de prédiction des prévalences des différentes pathologies en se basant sur les quantités des polluants . Les objectifs analytiques incluent :

1. **Prédictions :** Construire des modèles prédictifs pour estimer la prévalence des maladies à partir des données sur la pollution.
2. **Explications :** Pouvoir expliquer les prévalences obtenues à partir des quantités des polluants les plus importants.

Résultats attendus

Les résultats de cette analyse permettront de :

- Générer des **modèles prédictifs** fiables pour estimer la prévalence des maladies
- Identifier les **polluants critiques** liés à la prévalence de cancers

II. Récapitulatif des étapes passées

1. Définition de la problématique:

- Objectifs: Analyser les liens entre les polluants et les cancers broncho-pulmonaire et colorectal pour prédire leur prévalence et en expliquer les causes.

2. Préparation des données

- **2 Datasets :**
  - Pathologies: effectif de patients par pathologie, sexe, classe d'âge et territoire.
  - Registre Français des émissions polluantes
- Nettoyage des données : gestion des valeurs manquantes.

3. Étude de corrélation :

Afin de mieux comprendre la relation entre les quantités de polluants émises et la prévalence des cancers broncho-pulmonaire et colorectal, une étude de corrélation a été réalisée entre les

Les résultats obtenus suggèrent qu'il n'existe pas de lien direct ou significatif entre les quantités de polluants étudiés et la prévalence des cancers analysés.

D'où l'explication de la prévalence ne peut être totalement expliquée par les quantités émises de chaque polluants .

4. Fusion des données: Croisement des datasets par département, année.

5. Identification des méthodes: Modèles prédictifs pour estimer la prévalence des maladies et essayer de les expliquer.

6. Variables d'intérêt :

- **Variable Cible :** La prévalence des maladie
- **Variables d'intérêt** : les noms des départements, l'année d'émission des polluants, l'année de pathologie, la quantité des polluants émises.

III. Choix des modèles

Dans le but de prédire la prévalence des cancers broncho-pulmonaire et colorectal, une étape de sélection de modèles a été réalisée pour chaque type de cancer. Cette démarche a consisté à évaluer divers algorithmes d'apprentissage automatique de régression afin d'identifier celui offrant la meilleure précision prédictive tout en étant le plus explicable. Les performances des modèles ont été comparées à l'aide de métriques telles que l'erreur quadratique moyenne (MSE) et l'erreur absolue moyenne (MAE). Cette analyse a permis de sélectionner le modèle le plus adapté pour prédire efficacement la prévalence de ces cancers en fonction des quantités de polluants émises, tout en favorisant une meilleure interprétabilité des résultats.

L'objectif est de prédire la prévalence du cancer bronchopulmonaire pour une année donnée à partir des quantités de polluants émises 8 ans avant. Nous avons testé plusieurs modèles d'apprentissage supervisé pour accomplir cette tâche.
Étant donné que notre problème consiste à prédire des valeurs continues, nous avons choisi d'utiliser des modèles de régression. Plusieurs modèles ont été envisagées :

1. **Decision Tree et Random Forest** : Captent des interactions non linéaires, avec Random Forest réduisant les risques de surapprentissage.
2. **Boosting Methods (XGBoost, AdaBoost, Gradient Boosting)** : Méthodes puissantes pour modéliser des relations complexes via une approche itérative.
3. **SVR (Support Vector Regressor)** : Modèle efficace pour des relations complexes en minimisant l'erreur et les écarts.
4. **MLP Regressor (Multi-Layer Perceptron)** : Réseau de neurones capable de modéliser des relations très complexes grâce à ses couches cachées.
5. **KNeighbors Regressor** : Prédictions basées sur la proximité locale des données, adaptées aux relations régionales.

Préparation des données

La préparation des données a été réalisée en plusieurs étapes clés:

1. **Fusion des données de pollution et de pathologies**: Pour chaque année de pathologie, seules les quantités de polluants émises 8 ans plus tôt ont été considérées . Par exemple, pour analyser les données de l'année 2019, nous utilisons uniquement les quantités de polluants émises en 2011. Par la suite, les données relatives aux "cancers bronchopulmonaires" et au "cancer colorectal" ont été extraites, puis regroupées par département pour calculer la prévalence moyenne des cas.
2. **Mapping des départements**:Nous avons attribué à chaque département son code, afin d'uniformiser les représentations des départements, qui différait entre les deux datasets.
3. **Calcul des émissions de polluants par département**: Pour chaque polluant, les émissions par département ont été agrégées, et les données ont été fusionnées avec

celles de la prévalence des pathologies(Par exemple, les quantités émises de polluants en 2014 ont été agrégées avec les prévalences en 2022). Les départements sans données sur un polluant donné ont été remplis avec une valeur nulle.

4. **Consolidation des données**: Pour chaque type de cancer, les données finales ont été combinées dans un seul dataframe où chaque ligne correspond à un département avec ses informations de prévalence et de pollution pour une année donnée. La taille du dataset final est donc 707 lignes et 133 colonnes. Pour garantir une évaluation fiable du modèle, les données ont été divisées en deux ensembles:

- ***Ensemble d'entraînement et de validation :*** composé de **606 lignes**, correspondant aux données des années **2016 à 2021**.
- ***Ensemble de Test :*** composé de **101 lignes**, correspondant à prévalence de l'année **2022**.

➔ Cette approche permet d'entraîner le modèle sur les six premières années, puis d'utiliser les données de l'année 2022 pour prédire la prévalence moyenne des pathologies. Les prédictions obtenues sont ensuite comparées aux valeurs réelles afin d'évaluer les performances du modèle.

5. **Séparation des données et préparation pour le modèle:** Pour l'entraînement et la validation des modèles, une validation croisée K-fold a été utilisée (Avec K = 5), permettant de tester les modèles sur plusieurs sous-ensembles des données pour évaluer leur performance de manière plus robuste étant donnée que notre dataset finale n'est pas très large.

   De plus deux approches ont été testées :

- ***Avec PCA****:* Une réduction de la dimensionnalité a été appliquée pour réduire la complexité des données, mais cette approche n'a pas montré de meilleures performances.
- ***Sans PCA****:* En l'absence de réduction de dimensionnalité, les modèles ont montré des résultats meilleurs, probablement en raison de la conservation des caractéristiques originales des données.

De plus, un Scaling des données a été effectué pour normaliser les valeurs, en particulier pour ceux sensibles à l'échelle des données comme le Support Vector Regressor et les réseaux de neurones.

Cette préparation a permis d'obtenir un ensemble de données propre et prêt à être utilisé pour entraîner et tester les différents modèles de régression.

Évaluation des modèles

Pour évaluer les performances des modèles, plusieurs métriques ont été utilisées :

- **MSE (Erreure Quadratique Moyenne)** : Permet de mesurer l'écart moyen entre les prévisions et les valeurs réelles. Un faible MSE indique que les prédictions sont proches des valeurs observées.

- **MAE (Erreure Absolute Moyenne)** : Permet de mesurer la moyenne de la différence entre la valeur réelle et la valeur prédites. Un faible MAE indique que les prédictions sont proches des valeurs observées.

Tous les modèles ont été testé avec les mêmes “k-fold”. De plus, la comparaison a été faite avec plusieurs valeurs du k-fold (Nous avons testé les performances pour k=5, k= 10 et k=15).

1. Cancer Bronchopulmonaire:

Le tableau ci-dessous montre les performances des différents modèles pour k=5 :

Modèle\Metriques	MSE	MAE
Random Forest Regressor	0.0043 ± 0.0020	0.0425 ± 0.0075
Gradient Boosting	0.0048 ± 0.0022	0.0429 ± 0.0080
Decision Tree	0.0049 ± 0.0027	0.0400 ± 0.0104
Support Vector Regressor	0.0050 ± 0.0023	0.0430 ± 0.0084
AdaBoost	0.0051 ± 0.0019	0.0468 ± 0.0085
XGBoost	0.0054 ± 0.0023	0.0470 ± 0.0089
MLP Regressor	0.0569 ± 0.0106	0.1464 ± 0.0134
K neighbors Regressor	0.0061 ± 0.0026	0.0491 ± 0.0077

Figure n°1: Comparaison des performances des modèles pour le cancer bronchopulmonaire

Le tableau montre que les modèles Random Forest et Gradient Boosting obtiennent les meilleures performances en termes de MSE et MAE, suivis de près par le Decision Tree, qui offre une performance légèrement inférieure mais reste compétitive.
Nb: **Pour chaque modèle on a fait un Grid Search pour que la comparaison soit plus ou moins équitable.**

2. Cancer colorectal :

Modèle\Metriques	MSE	MAE
Random Forest Regressor	0.0201 ± 0.0067	0.0915 ± 0.0081
Gradient Boosting	0.0211 ± 0.0075	0.0951 ± 0.0086
Decision Tree	0.0267 ± 0.0066	0.0994 ± 0.0079
Support Vector Regressor	0.0245 ± 0.0108	0.0980 ± 0.0120
AdaBoost	0.0250 ± 0.0046	0.1076 ± 0.0083
XGBoost	0.0240 ± 0.0084:	0.1061 ± 0.0089
MLP Regressor	0.0887 ± 0.0157:	0.1946 ± 0.0118
Kneighbors Regressor	0.0283 ± 0.0086	0.1162 ± 0.0087

Figure n°2: Comparaison des performances des modèles pour le cancer Colorectal

Finalement, le **Decision Tree Regressor a été choisi pour les deux cancers** car, bien que légèrement moins performant, il offre une meilleure capacité d'explication grâce à sa structure intuitive et compréhensible.

Cela permet de voir comment le modèle prend ses décisions, et quels sont les paramètres les plus utilisés dans la tâche de régression de l'arbre de décision ce qui est important pour l'explication

IV. Hyperparamètres et Évaluation du modèle

Hyperparamètres du modèle

Grid Search:

Afin d'optimiser les performances du modèle, un Grid Search a été effectué pour sélectionner les meilleurs hyperparamètres.

1. Configuration :
Les hyperparamètres explorés et leurs plages respectives sont les suivants :

- max_depth : [5, 6, 8, 9]
- min_samples_split : [2, 5, 10]
- min_samples_leaf : [1, 2, 4]

2. Méthodologie :
Un K-fold a été appliqué (K=5) pour évaluer les performances des combinaisons d'hyperparamètres. La métrique utilisée pour évaluer la qualité du modèle est l'erreur quadratique moyenne négative, afin de minimiser les erreurs de prédiction.

3. Résultats :
Le modèle avec les meilleurs hyperparamètres a été sélectionné automatiquement. Voici les paramètres optimaux identifiés :

- max_depth : 6
- min_samples_split : 4
- min_samples_leaf : 2

Les performances du modèle optimisé sont les suivantes:

- **MSE:** 0.01931589245746108
- **MAE** : 0.089751239644382

Post Pruning
On a utilisé le post-pruning pour simplifier l'arbre de décision tout en maintenant de bonnes performances prédictives. Cela permet d'éviter le surapprentissage, en supprimant les branches peu significatives, et améliore ainsi la capacité du modèle à généraliser sur de nouvelles données. De plus, un arbre plus simple est plus lisible et explicable, facilitant la communication des résultats aux parties prenantes comme les ARS.

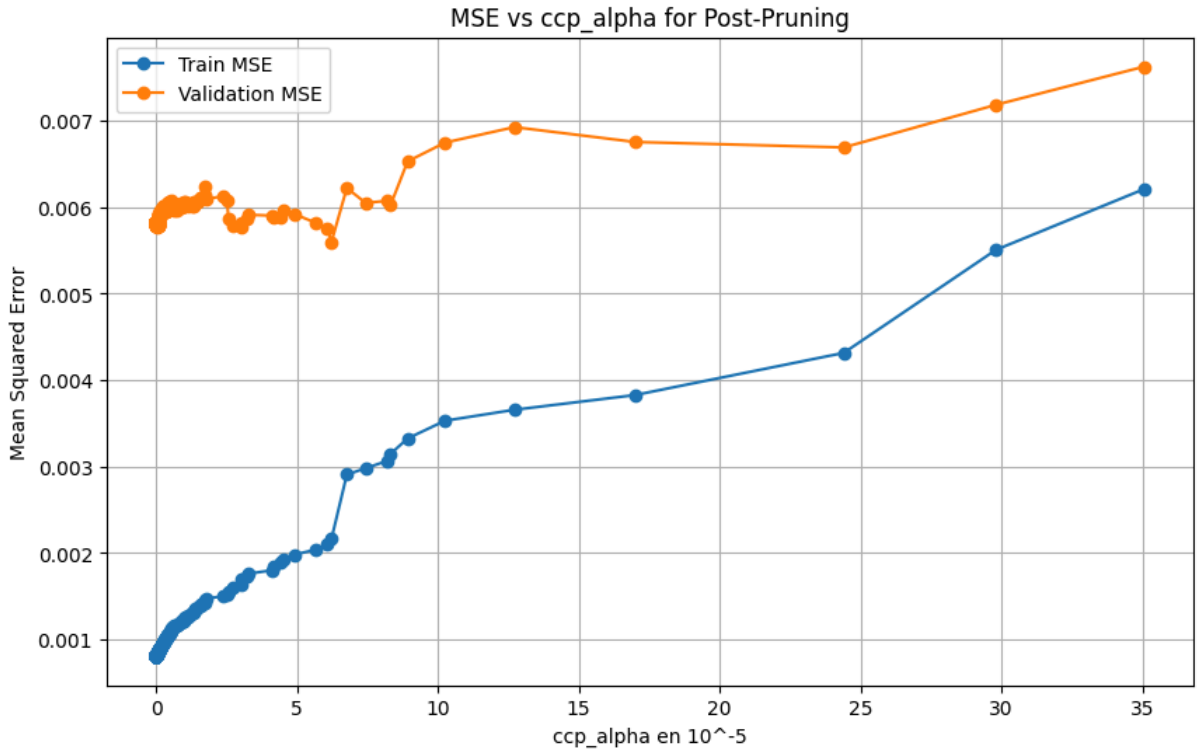

Figure n°3: Relation entre le degré de Post-pruning et MSE

Commentaire: La courbe montre l'évolution de l'erreur quadratique moyenne (MSE) en fonction du paramètre ccp alpha, qui contrôle la complexité de l'arbre après le post-pruning :

- Zone gauche : L'arbre est très complexe, avec de nombreuses branches, ce qui entraîne un surapprentissage. Cela se traduit par un MSE faible sur les données d'entraînement, mais un MSE plus élevé sur les données de validation.
- Zone droite: L'arbre est trop simplifié (sous-apprentissage), ce qui augmente à la fois le MSE d'entraînement et de validation.
- La valeur optimale de ccp_alpha est 5×10^{-5}, ce qui est relativement faible. Cela reflète un bon compromis entre la complexité de l'arbre et ses performances : un ccp_alpha faible indique que l'arbre conserve davantage de branches, évitant un excès de simplification, tout en limitant le surapprentissage et en maintenant une bonne capacité de généralisation, comme le montre le MSE minimal sur les données de validation.

Comparaison des performances vis-à-vis le post pruning

Profondeur de l'arbre	Avec Post pruning	Sans Post pruning
3	MSE: 0.0060 ± 0.0008 MAE: 0.0505 ± 0.0040	MSE: 0.0060 ± 0.0008 MAE: 0.0505 ± 0.0041
5	MSE: 0.0057 ± 0.0016 MAE: 0.0477 ± 0.0065	MSE: 0.0056 ± 0.0015 MAE: 0.0469 ± 0.0058
7	MSE: 0.0059 ± 0.0019 MAE: 0.0475 ± 0.0075	MSE: 0.0057 ± 0.0017 MAE: 0.0452 ± 0.0060
10	MSE: 0.0053 ± 0.0018 MAE: 0.0435 ± 0.0059	MSE: 0.0052 ± 0.0016 MAE: 0.0410 ± 0.0048

Figure n°4: Performances avec et sans Post pruning et différentes profondeurs

Commentaire: Le tableau montre que les performances sont légèrement meilleures sans post-pruning, surtout pour des arbres profonds. Cependant la différence est très faible (de l'ordre de 0.01 en valeur absolue ce qui revient à 14 personnes pour une population de 139 000 personnes comme Brest). D'ailleurs le post-pruning stabilise les résultats et prévient le surapprentissage, tout en simplifiant le modèle. Cela permet un bon équilibre entre précision et interprétabilité.

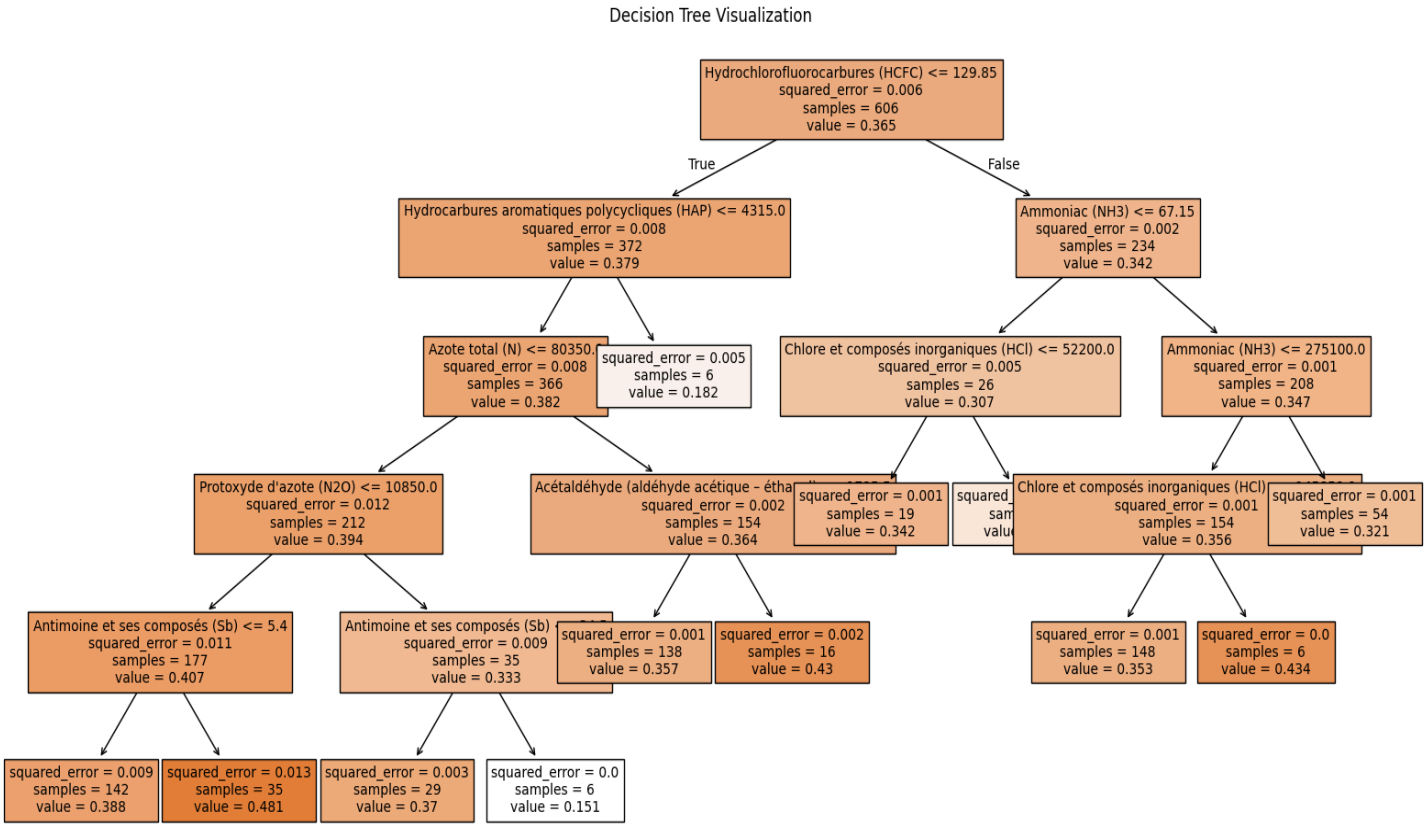

Figure n°5: Visualisation de l'arbre de décision pour le Cancer Broncho-pulmonaire (Sans normalisation des données)

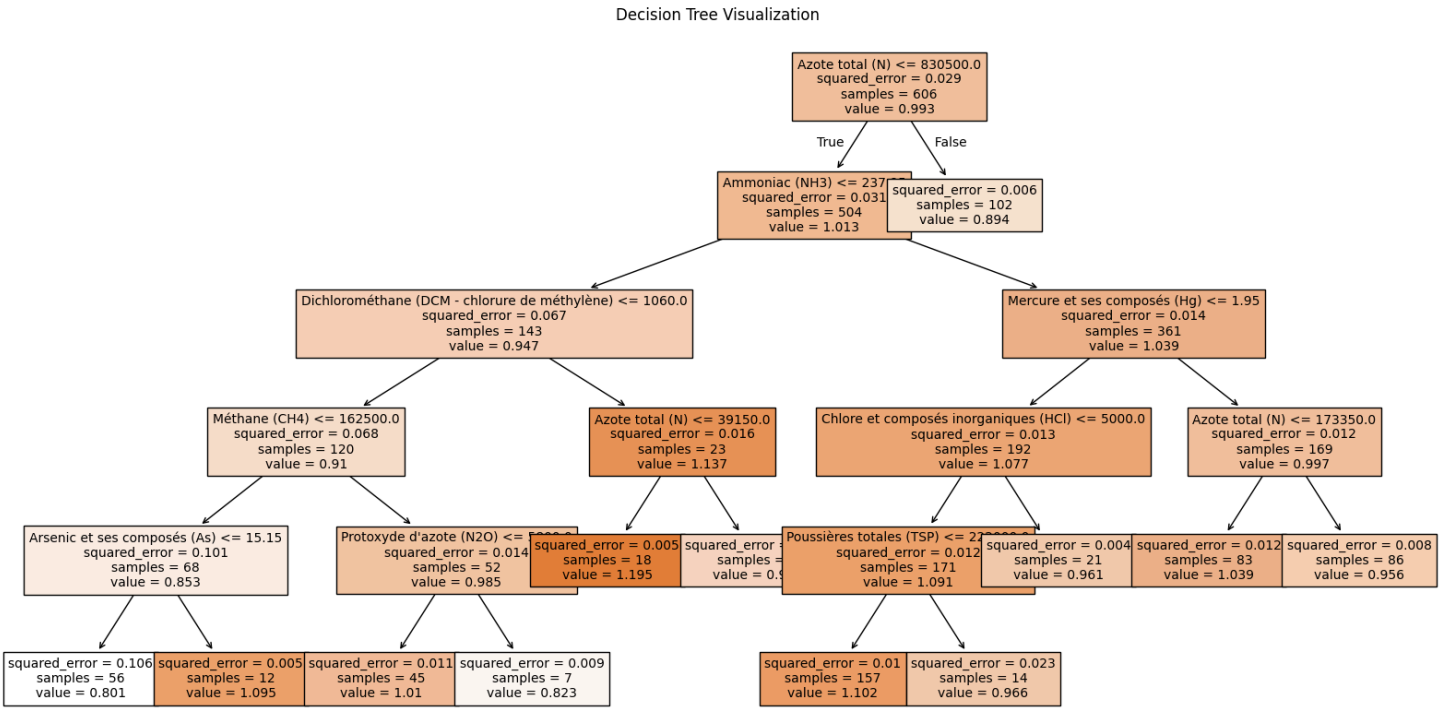

Figure n°6: Visualisation de l'arbre de décision pour le Cancer Colorectal (Sans normalisation des données)

Commentaire : Ce graphique représente un arbre de décision pour prédire la prévalence en fonction des seuils sur les quantités de divers polluants. À chaque nœud, une variable est utilisée pour diviser les données selon un seuil, avec les erreurs quadratiques et le nombre d'échantillons affichés. La prédiction finale est donnée dans les feuilles terminales. Pour prendre une décision, on suit les branches en fonction des valeurs des polluants mesurés, jusqu'à arriver à une feuille qui donne la prédiction optimale.

Nb: Pour faire la visualisation on a ignoré la normalisation des données (qui a un impact très faible sur l'erreur) afin d'avoir des valeurs de quantité de polluants significatives.

➔ **Finalement le modèle choisi est :** Decision Tree Regressor avec 5 comme profondeur max et avec le post-pruning

Explication des résultats

- **Pour la figure *n°4:***

L'interprétation de l'arbre de décision révèle des liens significatifs entre les polluants atmosphériques et la prévalence des cancers broncho-pulmonaires. Les polluants identifiés comme clés, à savoir les Hydrochlorofluorocarbures (HCFC), les Hydrocarbures Aromatiques Polycycliques (HAP) et l'Ammoniac (NH3), sont majoritairement associés au milieu "AIR", soulignant le rôle prédominant de la pollution atmosphérique dans l'apparition de ces cancers. Cette corrélation est biologiquement plausible, car la respiration humaine est directement influencée par la qualité de l'air. L'arbre de décision reflète cette réalité en plaçant ces polluants dans les premiers niveaux de division, ce qui indique leur importance pour expliquer la prévalence. Par ailleurs, ces résultats orientent les efforts de prévention : pour réduire l'incidence des cancers broncho-pulmonaires, il serait stratégique de cibler prioritairement les sources d'émissions de ces polluants dans l'air. Une analyse complémentaire pourrait explorer les impacts des autres milieux (eau, sol) pour confirmer leur rôle éventuel dans des contextes spécifiques.

- **Pour la figure *n°5:***

L'interprétation de l'arbre de décision montre que le cancer colorectal est principalement associé à des polluants provenant de l'eau, notamment l'Azote total, le cadmium et ses composés, ainsi que le Dichlorométhane (DCM - chlorure de méthylène). Ces polluants proviennent principalement de sources agricoles (engrais, pesticides), de déversements industriels (métallurgie, solvants) et de rejets urbains insuffisamment traités, contaminant les eaux de surface et souterraines.

Ce lien est cohérent, car la santé du côlon et du rectum dépend de ce que nous consommons, notamment la qualité de l'eau. Une exposition prolongée à ces polluants favorise

l'inflammation chronique et les mutations génétiques, augmentant le risque de cancer colorectal. Ces résultats soulignent l'importance de renforcer la surveillance des eaux, de limiter les rejets polluants, et d'orienter les politiques sanitaires vers une meilleure prévention et sensibilisation des populations à risque.

Evaluation du modèle

Prédictions de la prévalence pour l'année 2022:

Afin de tester le modèle, nous avons prédit la prévalence pour l'année 2022 en utilisant les quantités de polluants en 2014. La prédiction donne les résultats suivant :

	MSE	MAE
Cancer Bronchopulmonaire	0.006	0.0494
Cancer Colorectal	0.0197	0.0953

Figure n°6 : Résultats des tests du modèle

Commentaire: Les résultats montrent que le modèle prédit avec précision la prévalence des cancers bronchopulmonaires avec des erreurs faibles. Pour les cancers colorectaux, l'erreur est légèrement plus élevée mais reste acceptable. Globalement, le modèle est performant, bien que les prédictions soient légèrement moins précises pour le cancer colorectal

V. Contributions du Modèle et de son Interprétation pour les ARS : Actions Ciblées

En France, le coût moyen annuel de traitement d'un patient atteint de cancer est de 14 600 €, mais peut atteindre 72 000 € pour des traitements spécifiques comme l'immunothérapie (source : *Que Choisir*), ce qui montre qu'une mauvaise gestion des ressources peut engendrer des surcoûts importants, surtout pour les cas diagnostiqués tardivement. Grâce à notre modèle, une réduction de 10 % , par exemple, des erreurs d'allocation des ressources, pour 1 000 patients, permettrait une économie annuelle de 1,46 million d'euros, en optimisant l'utilisation des lits, des équipements et du personnel médical, ce qui aiderait les ARS à mieux gérer les ressources médicales disponibles.

Par ailleurs, selon l'Organisation Mondiale de la Santé, un diagnostic précoce peut réduire le coût des traitements de 40 %, passant de 72 000 € à 43 200 € pour les cas les plus coûteux. Si notre modèle permet de détecter précocement 20 % des cas supplémentaires parmi 1 000 patients, le gain total s'élèverait à 5,76 millions d'euros par an. En combinant la diminution des erreurs d'allocation et l'impact du diagnostic précoce, notre modèle pourrait ainsi générer des économies totales estimées à 7,22 millions d'euros par an pour 1 000 patients, tout en permettant aux ARS d'améliorer la qualité des soins et les chances de survie.

D'autre part, nous avons pu détecter les polluants les plus liés à chaque type de cancer, ce qui permet à l'ARS de collaborer avec d'autres organisations pour prendre des actions directes visant à contrôler les quantités de ces polluants, notamment dans les rejets industriels et les pratiques agricoles.

Citons l'exemple de la collaboration entre l'ARS Auvergne-Rhône-Alpes et la DREAL en 2022, qui a porté sur la surveillance des substances perfluoroalkylées (PFAS) dans les eaux au sud de Lyon. Cette initiative a permis d'intégrer les PFAS dans la liste des substances contrôlées lors des inspections des rejets aqueux des sites industriels, renforçant ainsi la surveillance environnementale et protégeant la santé publique en limitant l'exposition à ces polluants. Une telle collaboration illustre comment des actions conjointes peuvent répondre efficacement aux problématiques environnementales et sanitaires identifiées par notre modèle, tout en orientant les politiques vers une prévention ciblée et des mesures correctives adaptées.

VI. Conclusion

Pour accomplir la tâche de prédiction et fournir des explications interprétables, nous avons suivi un processus structuré en plusieurs étapes. Tout d'abord, nous avons créé le dataset final en fusionnant deux bases de données, l'une relative aux pathologies et l'autre à la pollution, en les croisant par année et département. Ce dataset a ensuite été divisé en un ensemble d'entraînement et de validation (pour les années 2016-2021, validé par K-fold) et un ensemble de test correspondant à l'année 2022.

Plusieurs modèles de régression ont été évalués, et nous avons retenu les arbres de décision en raison de leur simplicité d'interprétation et de leur bonne performance dans notre contexte. Le modèle a été optimisé vis-à-vis l’explicabilité grâce à un post-pruning et une limitation de la profondeur tout en gardant de bonnes performances. Finalement, le modèle final a permis de prédire efficacement les prévalences, notamment pour l'année 2022, et d'identifier les polluants les plus associés à chaque type de cancer. Ces résultats seront immédiatement exploitables par les ARS pour optimiser la gestion des ressources médicales et renforcer les contrôles sur les rejets industriels et les pratiques agricoles grâce à leurs collaborations existantes.

Utilisation de l’IA générative :

- Correction des fautes d'orthographe
- Résumé des paragraphes et reformulation
- Correction des erreurs de code