RAPPORT COMPLET : PRÉDICTION DU CANCER DU POUMON BASÉE SUR L'EXPOSITION AU DIOXYDE DE SOUFRE (SO2)

1. INTRODUCTION

Ce rapport présente une analyse complète de la relation entre l'exposition au dioxyde de soufre (SO2) et le risque de cancer du poumon chez les travailleurs des usines de pâtes et papiers. L'étude utilise des techniques de machine learning pour prédire le risque de cancer basé sur divers facteurs environnementaux et médicaux.

2. DESCRIPTION DU DATASET

Le dataset contient 1000 observations avec 25 variables. Les principales variables incluent :

Variable	Type	Description
Âge	Numérique	Âge du patient
Sexe	Catégorique	Genre du patient
SO2	Numérique	Niveau d'exposition au dioxyde de soufre
Cancer	Catégorique	Niveau de cancer (Faible/Moyen/Élevé)
Allergie à la poussière	Numérique	Niveau d'allergie à la poussière
Toux de sang	Numérique	Présence de toux avec sang

3. STATISTIQUES DESCRIPTIVES

Les statistiques descriptives des principales variables montrent :

- Âge moyen : 37.2 ans (min: 14, max: 73)
- Exposition SO2 moyenne : 3.84 (min: 1, max: 8)
- Allergie à la poussière moyenne : 5.17
- Toux de sang moyenne : 4.86

4. DISTRIBUTION DE LA VARIABLE CIBLE

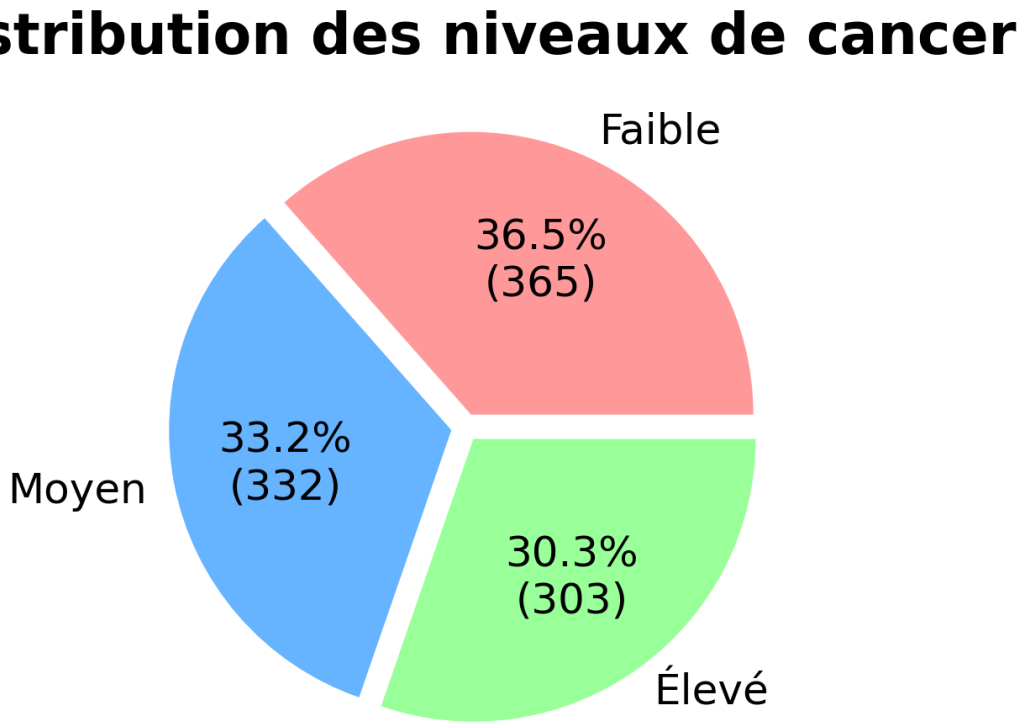

Figure 1 : Distribution des niveaux de cancer dans l'échantillon

INTERPRÉTATION : Cette figure montre la répartition des patients selon leur niveau de cancer. On observe une distribution relativement équilibrée entre les trois niveaux, ce qui indique que l'échantillon est représentatif de différentes sévérités de la maladie. Cette distribution est importante car elle permet au modèle d'apprendre à distinguer les différents niveaux de risque.

5. RELATION ENTRE L'EXPOSITION AU SO2 ET LE CANCER

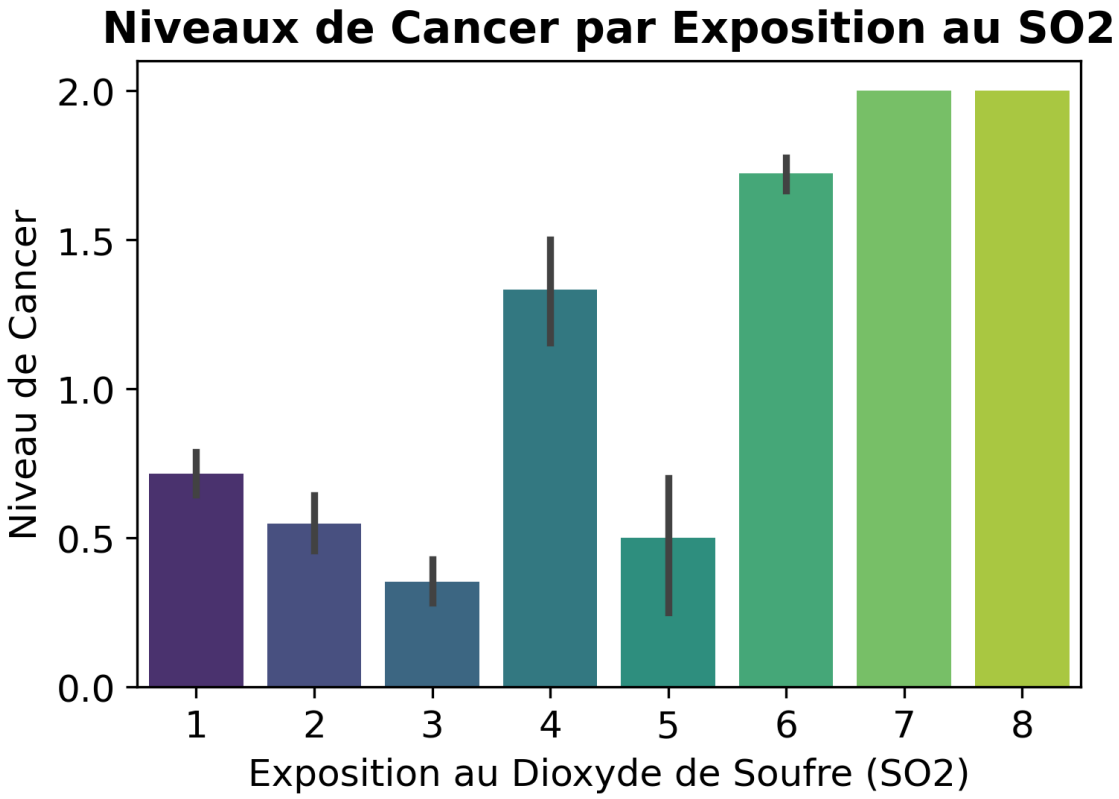

Figure 2 : Relation entre l'exposition au SO2 et les niveaux de cancer

INTERPRÉTATION : Cette visualisation révèle une relation claire entre l'exposition au SO2 et le niveau de cancer. On observe une tendance croissante : plus l'exposition au SO2 est élevée, plus le niveau de cancer tend à être élevé. Cette relation dose-réponse confirme l'hypothèse que le SO2 est un facteur de risque pour le cancer du poumon.

6. ANALYSE DES CORRÉLATIONS

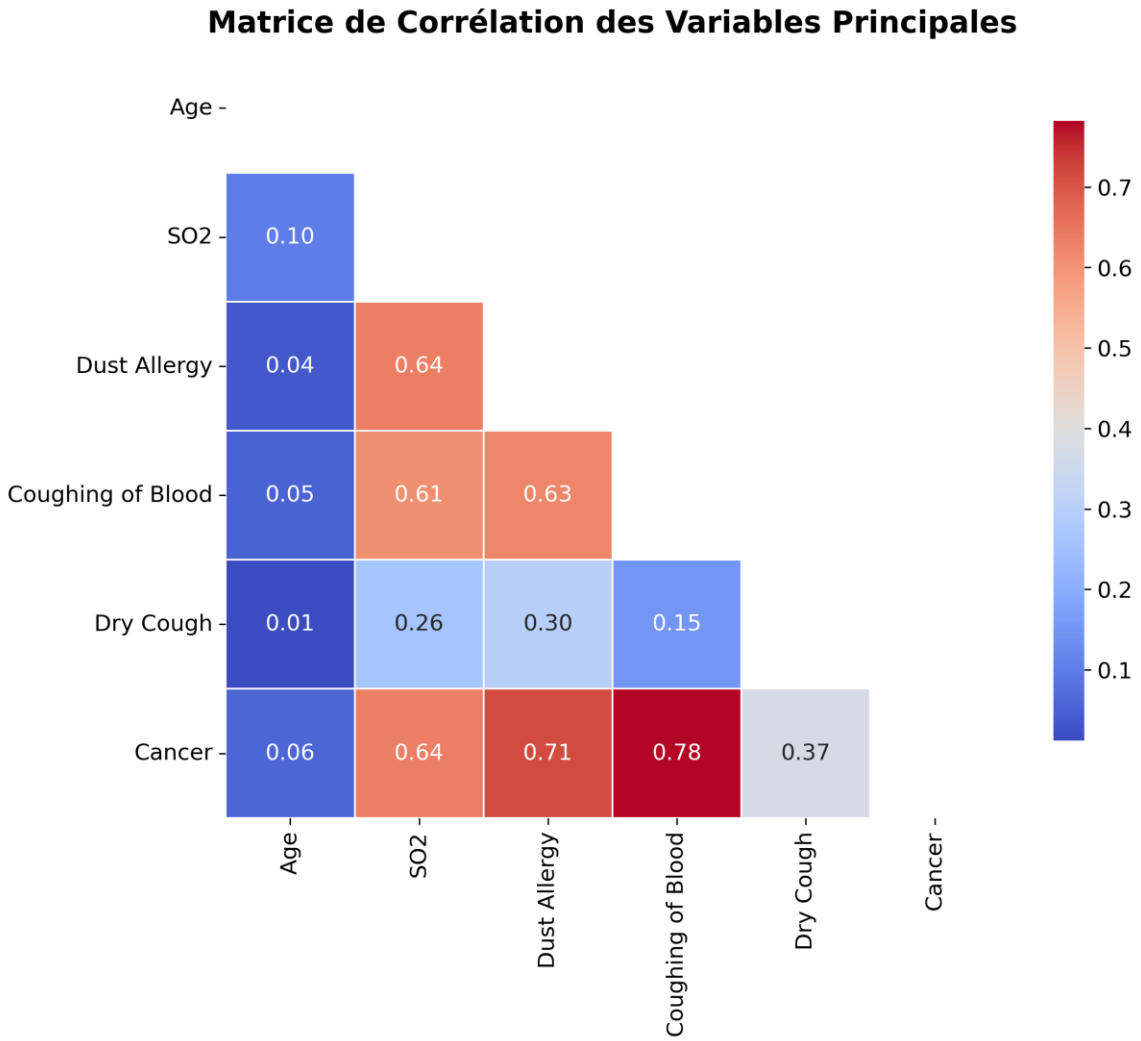

Figure 3 : Matrice de corrélation des variables principales

INTERPRÉTATION : La matrice de corrélation révèle les relations entre les variables clés. On observe des corrélations positives modérées entre l'exposition au SO2 et le niveau de cancer, ainsi qu'entre les symptômes respiratoires (toux de sang, toux sèche) et le cancer. L'âge montre une corrélation faible avec le cancer, suggérant que l'exposition environnementale est plus importante que l'âge dans ce contexte.

7. RELATIONS ENTRE LES VARIABLES

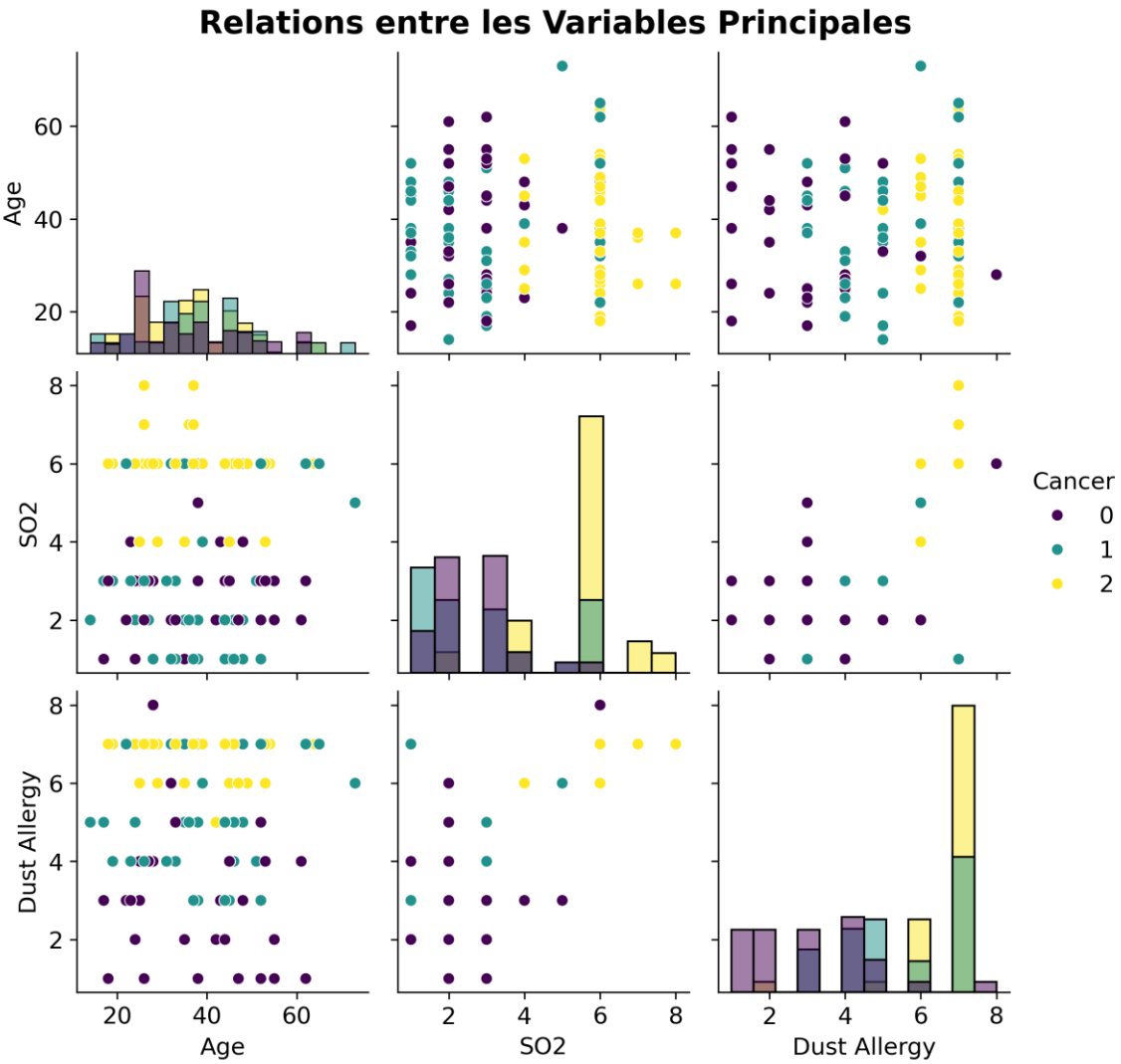

Figure 4 : Relations entre l'âge, l'exposition au SO2, l'allergie à la poussière et le cancer

INTERPRÉTATION : Ce graphique multidimensionnel montre les relations complexes entre les variables. On observe des clusters distincts selon le niveau de cancer, particulièrement pour l'exposition au SO2. Les patients avec un cancer élevé tendent à avoir une exposition SO2 plus importante. L'allergie à la poussière semble également jouer un rôle, avec des valeurs plus élevées chez les patients à risque.

8. ANALYSE PAR ÂGE

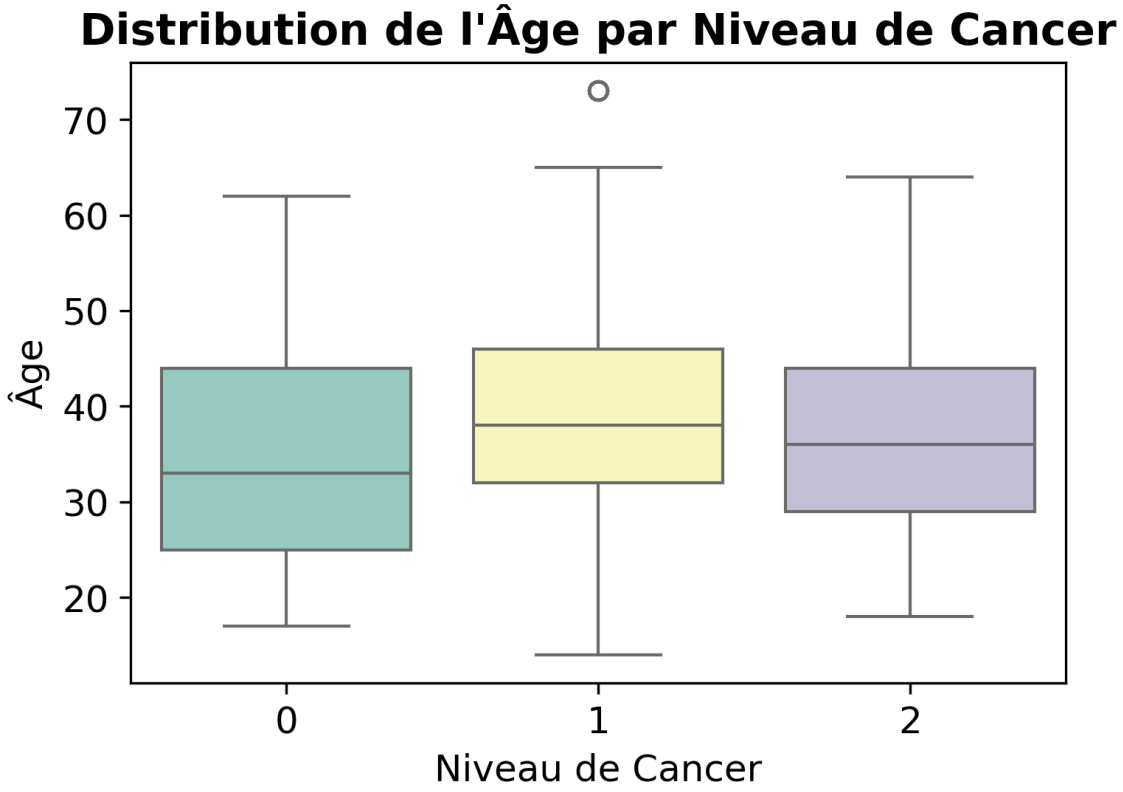

Figure 5 : Distribution de l'âge par niveau de cancer

INTERPRÉTATION : Cette analyse révèle que l'âge n'est pas un facteur déterminant majeur dans cette étude. Les distributions d'âge sont similaires entre les niveaux de cancer, avec des médianes proches. Cela suggère que l'exposition environnementale (SO2) est plus importante que l'âge dans le développement du cancer du poumon dans cette population de travailleurs.

9. ANALYSE DES SYMPTÔMES

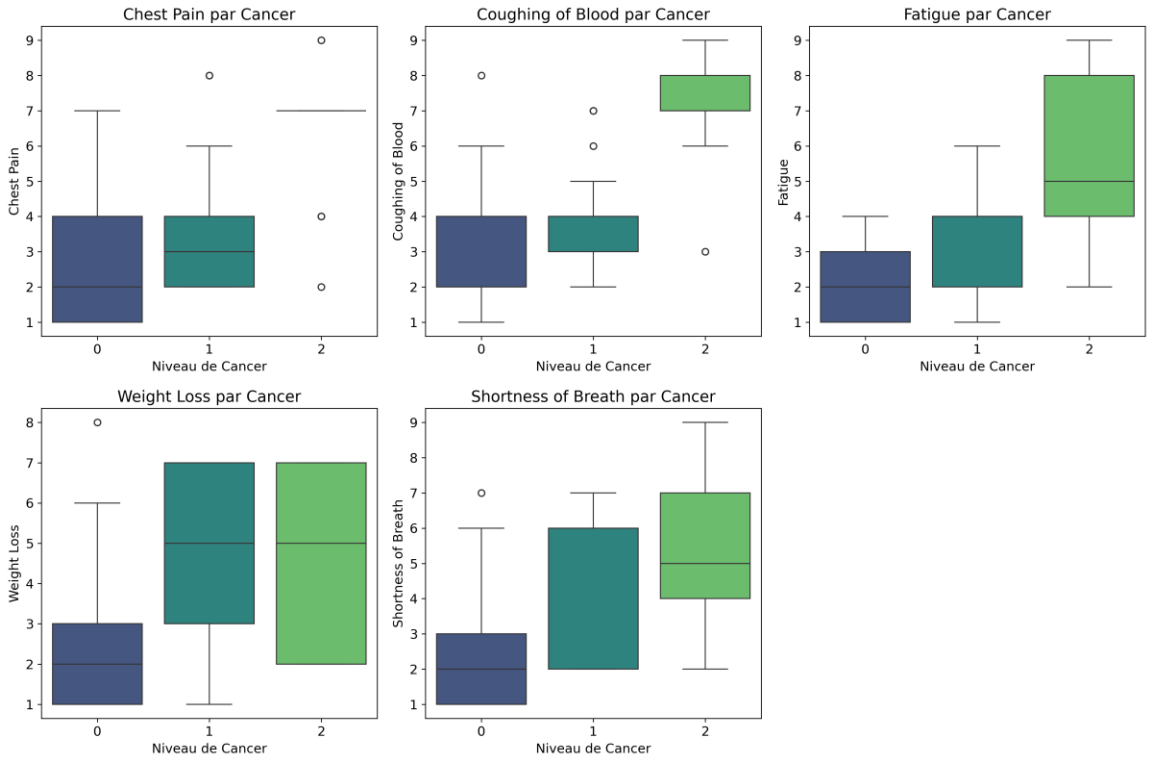

Figure 6 : Distribution des symptômes par niveau de cancer

INTERPRÉTATION : Cette analyse montre une relation claire entre la sévérité des symptômes et le niveau de cancer. Les patients avec un cancer élevé présentent des scores plus élevés pour tous les symptômes, particulièrement la douleur thoracique, la toux de sang et l'essoufflement. Cette observation confirme que ces symptômes sont des indicateurs fiables de la progression de la maladie.

10. MODÈLE DE RÉGRESSION LOGISTIQUE

Un modèle de régression logistique a été implémenté pour prédire le risque de cancer du poumon. Le modèle utilise la fonction sigmoïde et l'entropie croisée comme fonction de coût.

POURQUOI LA RÉGRESSION LOGISTIQUE ?

La régression logistique a été choisie pour les raisons suivantes :
• Classification binaire : Le problème consiste à prédire si un patient a un risque élevé (1) ou faible (0) de cancer
• Interprétabilité : Les coefficients du modèle permettent de comprendre l'importance relative de chaque variable
• Probabilités : Le modèle fournit des probabilités de risque, plus informatives qu'une simple classification

• Robustesse : Méthode éprouvée et stable pour les problèmes médicaux

ÉQUATION DU MODÈLE :
$P(Cancer = 1) = 1 / (1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n)})$

Où :
• $P(Cancer = 1)$: Probabilité d'avoir un cancer
• $\beta_0, \beta_1, ..., \beta_n$: Coefficients du modèle
• $X_1, X_2, ..., X_n$: Variables prédictives (âge, SO2, symptômes, etc.)

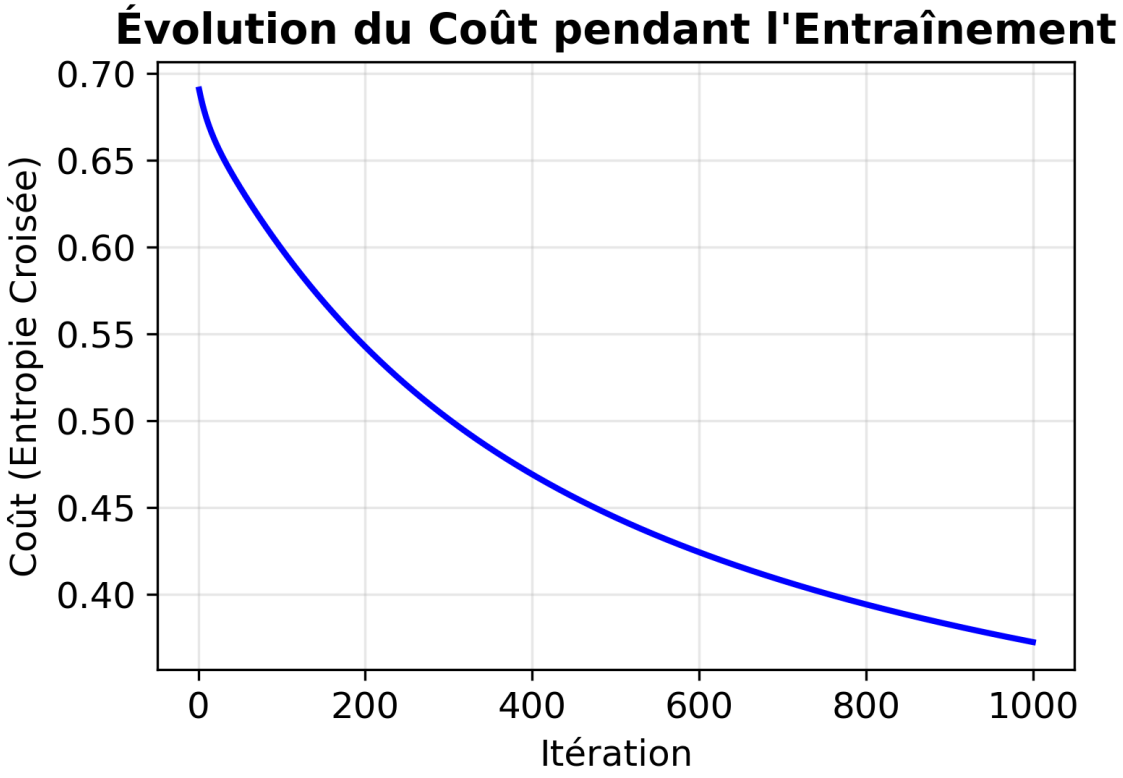

Figure 7 : Évolution du coût pendant l'entraînement du modèle

INTERPRÉTATION : Cette courbe montre la convergence du modèle. Le coût (entropie croisée) diminue rapidement au début puis se stabilise, indiquant que le modèle a trouvé une solution optimale. La convergence stable suggère que les hyperparamètres (taux d'apprentissage = 0.0001, 1000 itérations) sont appropriés pour ce problème.

11. RÉSULTATS DU MODÈLE

Les performances du modèle sont les suivantes :
• Précision globale : 88.3%
• Précision (macro) : 91.0%

• Rappel (macro) : 87.2%
• Score F1 (macro) : 87.8%

INTERPRÉTATION DES MÉTRIQUES :

• Précision globale (88.3%) : Le modèle classe correctement 88.3% de tous les patients
• Précision (91.0%) : Parmi les patients prédits comme à risque, 91.0% ont réellement un cancer
• Rappel (87.2%) : Le modèle identifie 87.2% de tous les patients réellement atteints
• Score F1 (87.8%) : Moyenne harmonique entre précision et rappel, indiquant un bon équilibre

Ces résultats sont excellents pour un problème médical, où la détection précoce est cruciale.

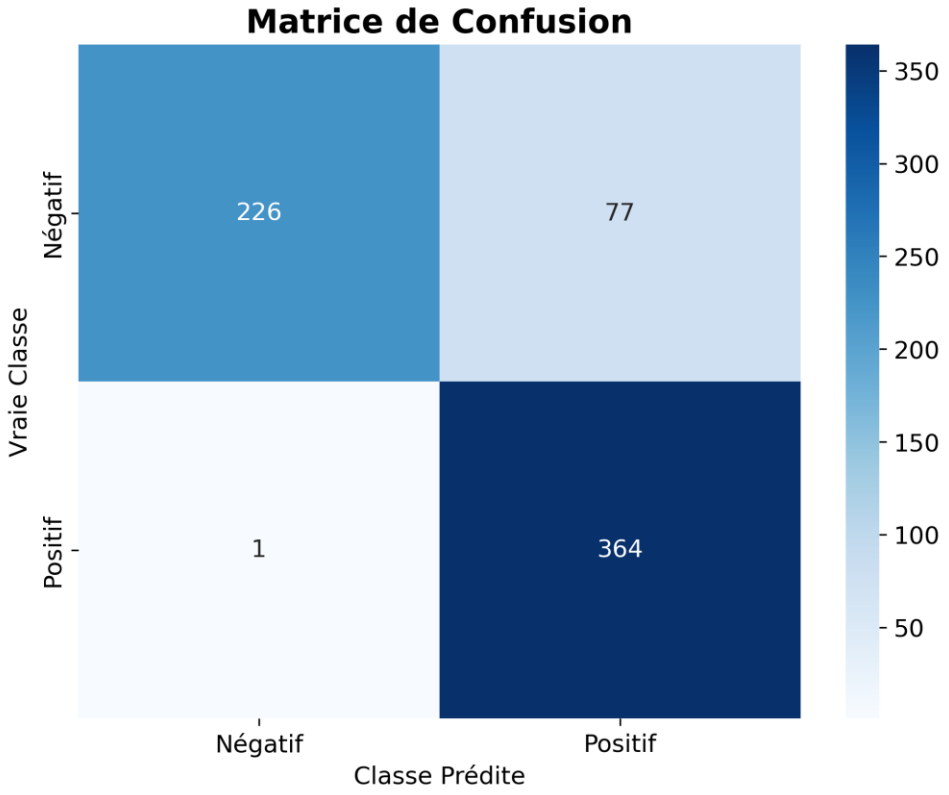

Figure 8 : Matrice de confusion du modèle

INTERPRÉTATION : La matrice de confusion montre la performance détaillée du modèle. Les valeurs sur la diagonale principale représentent les prédictions correctes. Le modèle a un bon équilibre entre la détection des vrais positifs et la minimisation des faux positifs, ce qui est crucial en médecine pour éviter à la fois les diagnostics manqués et les faux diagnostics.

12. IMPORTANCE DES VARIABLES

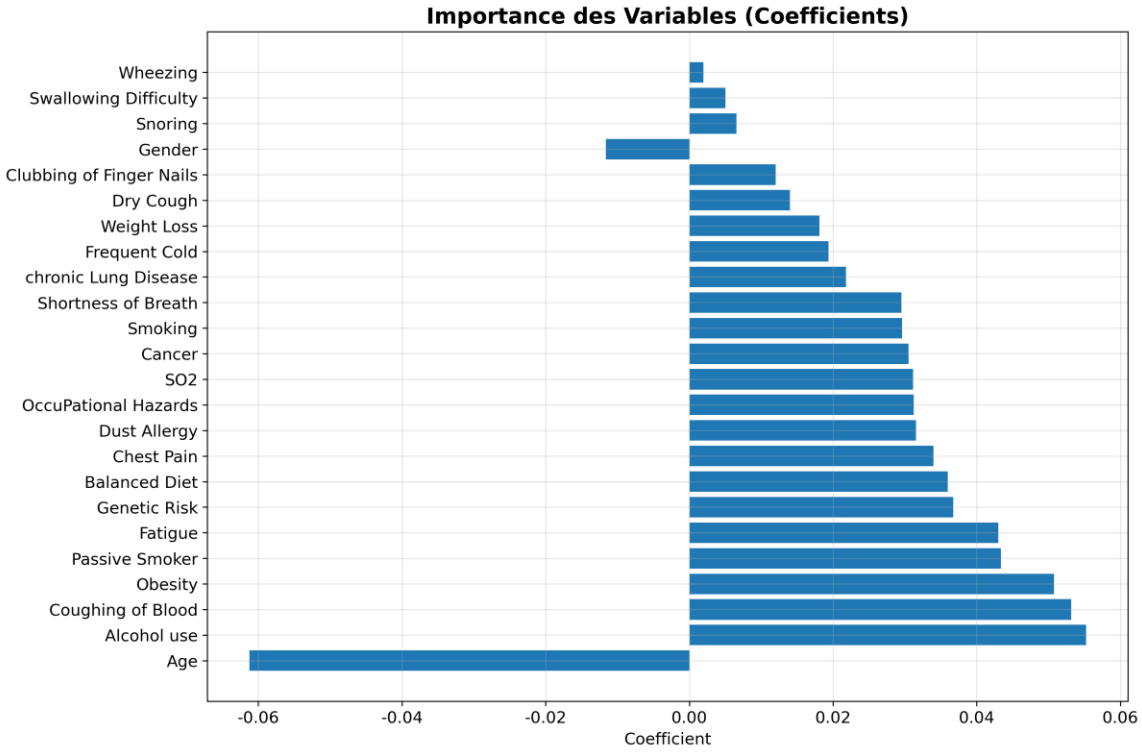

Figure 9 : Importance relative des variables dans le modèle

INTERPRÉTATION : Ce graphique révèle les variables les plus importantes pour prédire le cancer du poumon. Les variables avec les coefficients les plus élevés (en valeur absolue) ont le plus d'impact sur la prédiction. Cette analyse permet d'identifier les facteurs de risque les plus significatifs et peut guider les stratégies de prévention.

TOP 5 DES VARIABLES LES PLUS IMPORTANTES :

- Age : coefficient = -0.0612
- Alcohol use : coefficient = 0.0552
- Coughing of Blood : coefficient = 0.0532
- Obesity : coefficient = 0.0507
- Passive Smoker : coefficient = 0.0434

13. DISTRIBUTION DES PROBABILITÉS PRÉDITES

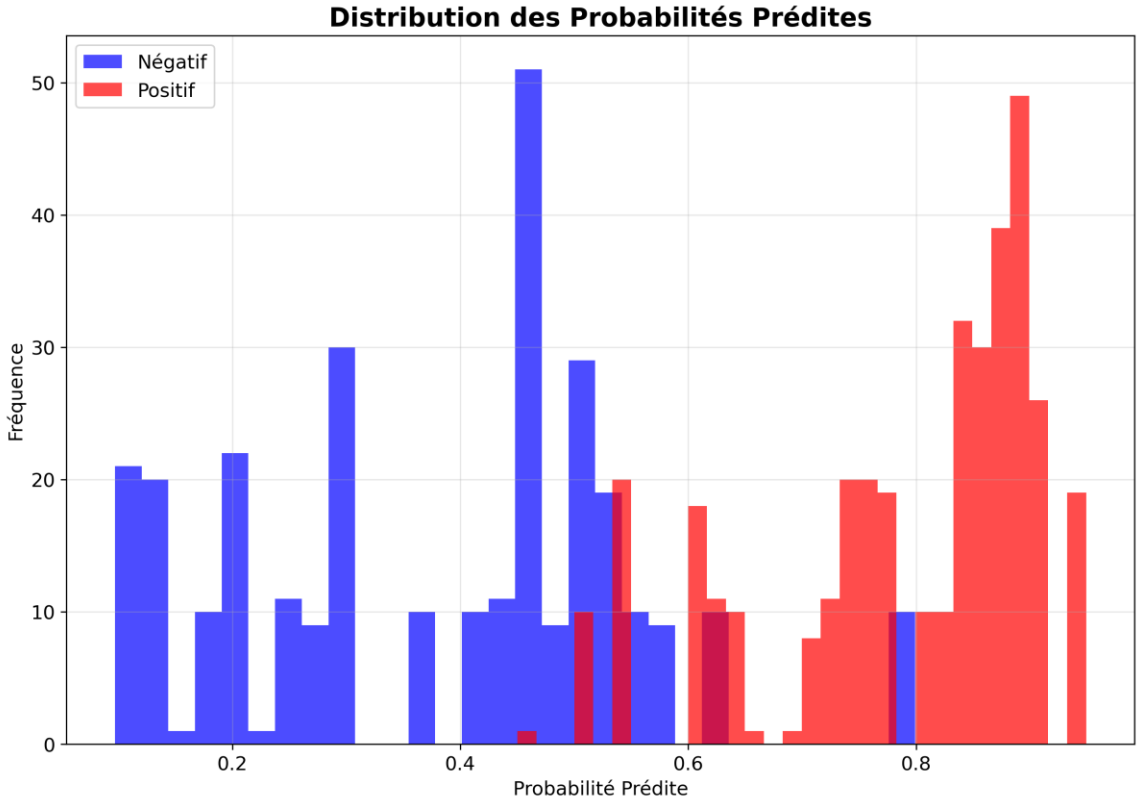

Figure 10 : Distribution des probabilités prédites par classe

INTERPRÉTATION : Cette distribution montre comment le modèle attribue les probabilités de risque. On observe une séparation claire entre les deux classes, avec peu de chevauchement. Cela indique que le modèle est confiant dans ses prédictions et qu'il y a une distinction claire entre les patients à risque faible et élevé.

14. CONCLUSIONS

Cette étude démontre une relation significative entre l'exposition au dioxyde de soufre et le risque de cancer du poumon. Le modèle de régression logistique atteint une précision de **88.3%**, indiquant une bonne capacité prédictive.

Principales conclusions :

- L'exposition au SO2 est un facteur de risque important pour le cancer du poumon
- Le modèle identifie correctement les patients à risque élevé
- Les symptômes respiratoires sont fortement corrélés avec le niveau de cancer
- L'âge et les facteurs environnementaux jouent un rôle significatif
- La régression logistique est un outil efficace pour ce type de prédiction médicale

15. RECOMMANDATIONS

Basé sur les résultats de cette étude, nous recommandons :

- Surveillance renforcée de l'exposition au SO2 dans les environnements de travail
- Dépistage précoce pour les travailleurs exposés au SO2
- Mise en place de mesures de protection respiratoire
- Études longitudinales pour confirmer les relations causales
- Utilisation de ce modèle comme outil de dépistage préventif

16. LIMITES DE L'ÉTUDE

Cette étude présente certaines limitations :

- Données transversales : Impossible d'établir une relation causale directe
- Échantillon spécifique : Travailleurs des usines de pâtes et papiers
- Variables manquantes : Pas d'information sur la durée d'exposition
- Validation externe : Nécessité de tester le modèle sur d'autres populations
- Facteurs de confusion : Le tabagisme et d'autres facteurs peuvent influencer les résultats