Prédiction du type de cancer à partir des facteurs de risque

Gloria SAFOUESS

Introduction

Ce projet a pour objectif de concevoir un modèle de machine learning capable de prédire le type de cancer le plus probable chez un patient à partir de ses facteurs de risque.

L'enjeu principal est de comprendre l'influence des habitudes de vie, des antécédents familiaux et de l'environnement sur la survenue de différents cancers.

Ce travail s'inscrit dans un contexte hospitalier simulé : le service de prévention des cancers de Pointe-Noire souhaite disposer d'un outil d'aide à la décision fondé sur les données.

Description du jeu de données

Le jeu de données utilisé comprend 2000 patients et 21 variables.
Chaque ligne correspond à un patient, et chaque colonne représente un facteur de risque.

Les cinq types de cancer à prédire sont : **Lung, Breast, Colon, Prostate** et **Skin**.

Quelques exemples de variables :

- Age, Gender (homme/femme),
- Smoking, Alcohol_Use, Obesity,
- Family_History, Air_Pollution, BRCA_Mutation.

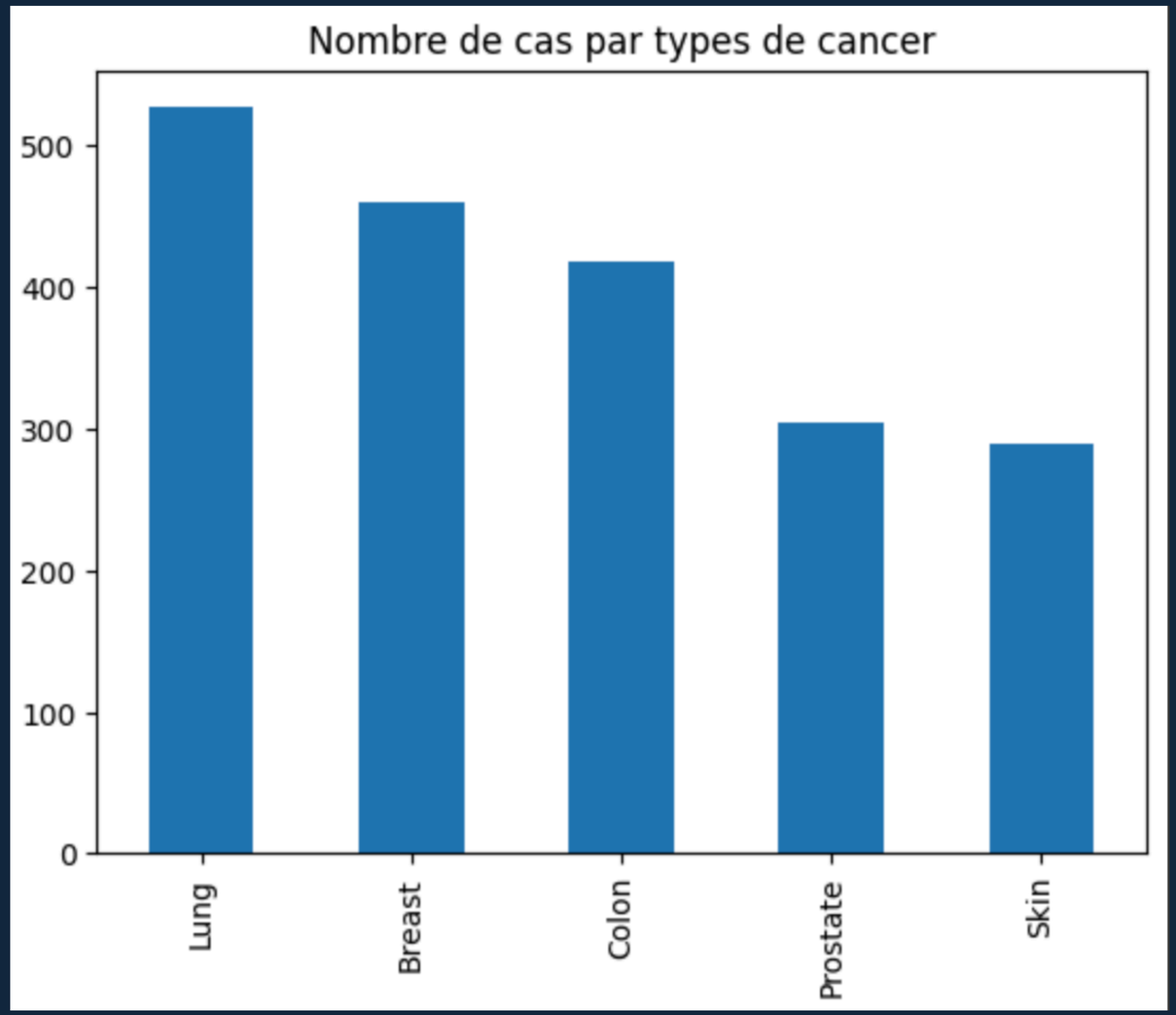
Les colonnes inutiles comme Patient_ID ou Overall_Risk_Score ont été retirées pour éviter les biais.

Nettoyage et Préparation

- Suppression des colonnes inutiles
- Suppression des valeurs manquantes et des doublons
- Encodage de la variable cible Cancer_Type
- Normalisation de certaines variables numériques

Répartition des types de cancer

- **Cancer du poumon:** près de 550 cas
- **Cancer du sein:** près de 480 cas
- **Cancer du colon:** près de 420 cas
- **Cancer de la prostate et de la peau**: près de 300 cas



Relation entre l’age et les types de cancer

- **Prostate :**

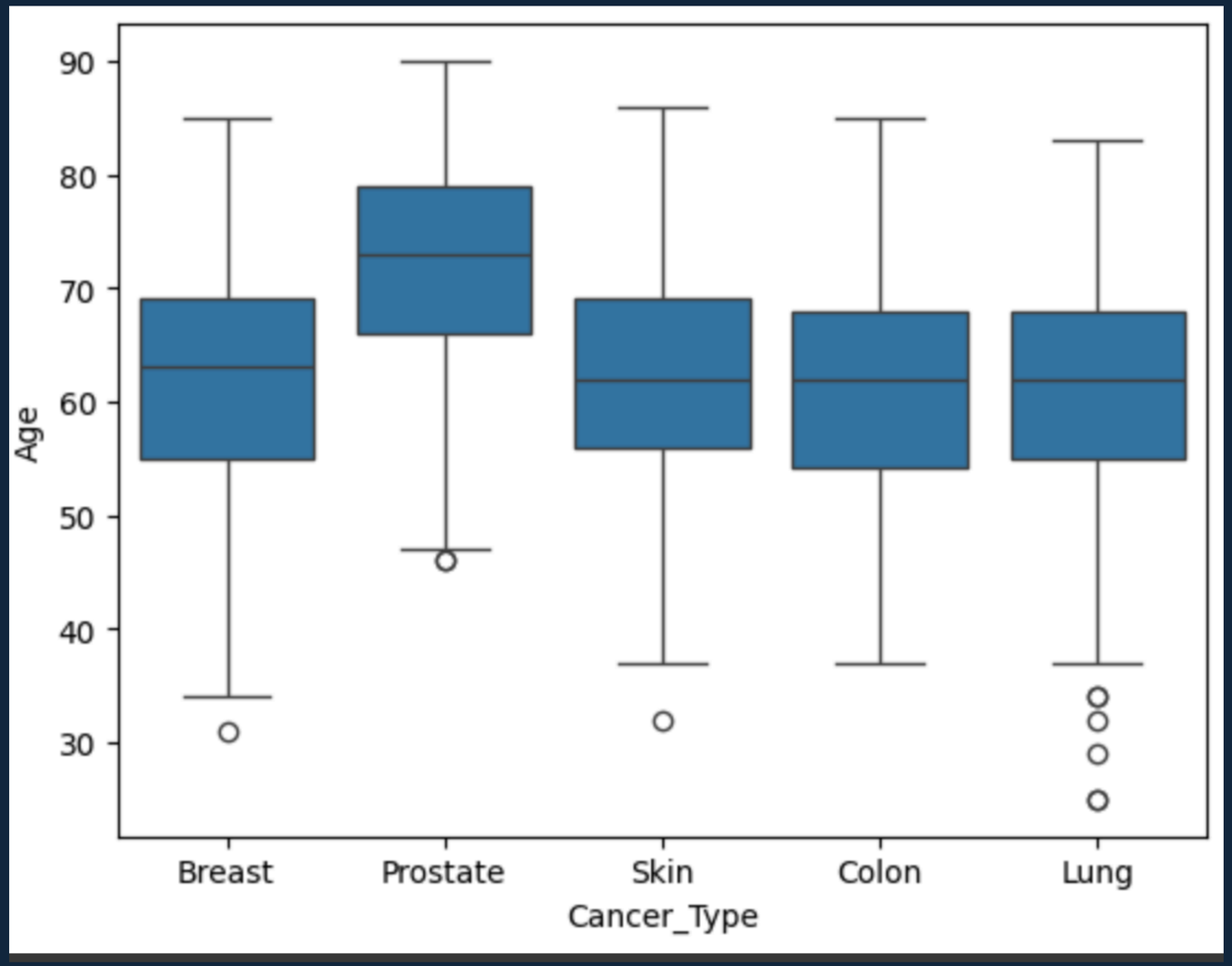
Médiane la plus élevée (~70–75 ans).

- **Breast (sein) :**

Médiane autour de 60–65 ans.
Quelques cas plus jeunes (~30 ans)

- **Skin, Colon et Lung :**

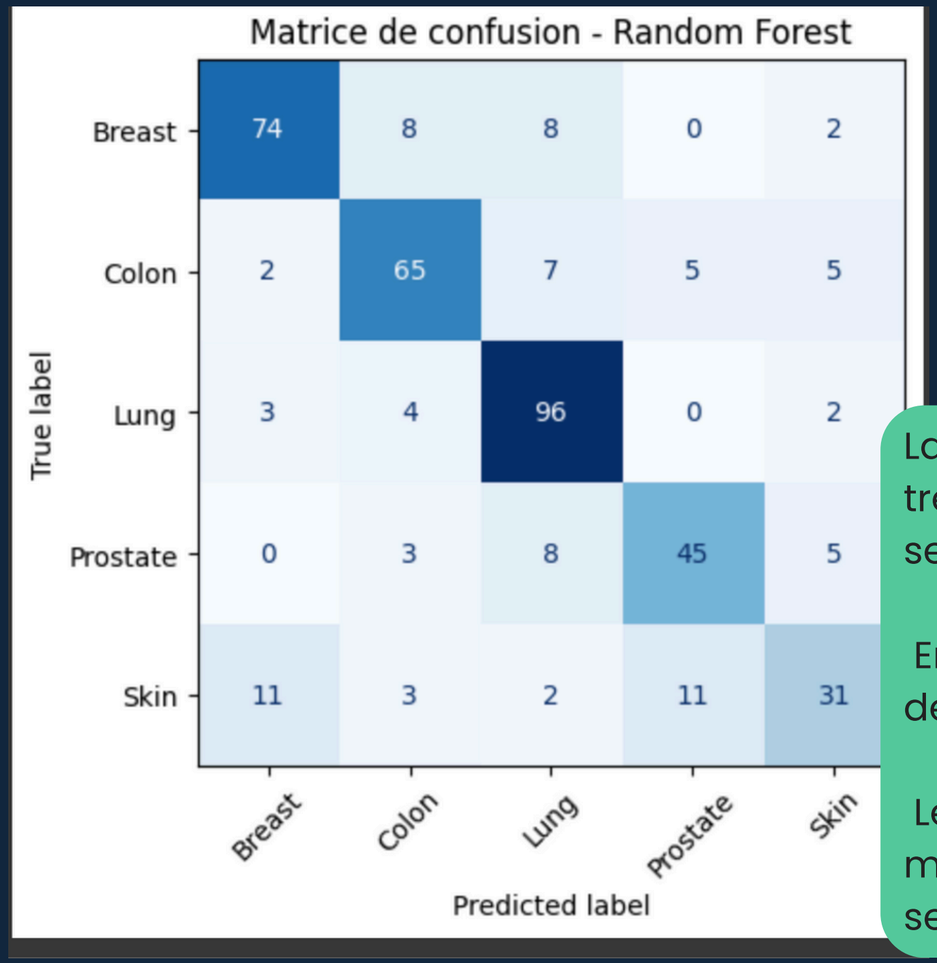
Médianes similaires (~60–65 ans).
Le cancer du poumon montre plusieurs outliers à des âges plus jeunes.



Modélisation et évaluation

RANDOM FOREST

- Split : 80 % entraînement / 20 % test (stratifié).
- Modèle : RandomForestClassifier (200 arbres, random_state=42).
- Motivations : robuste, bon compromis performance / interprétabilité.

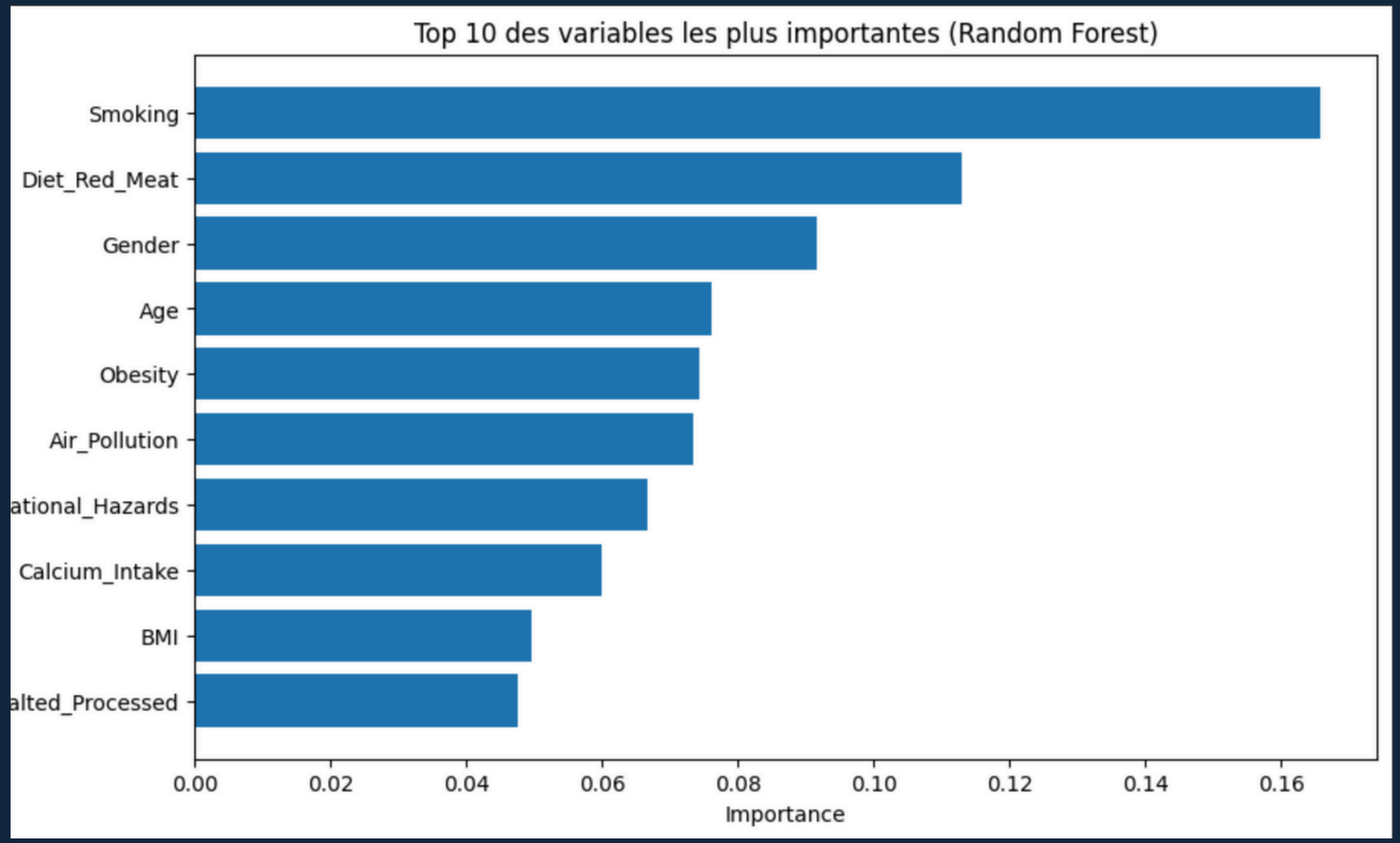


Matrice de confusion

La matrice de confusion montre que le modèle reconnaît très bien les cancers du poumon (96/105 bien classés) et du sein (74/92).

En revanche, il confond parfois le cancer du colon et celui de la prostate, qui présentent des profils similaires.

Le cancer de la peau reste difficile à identifier : près de la moitié des cas sont mal classés, souvent confondus avec le sein ou la prostate.



Insights clés:

- **Smoking** = facteur 1 du modèle
- **Diet_Red_Meat & Age/Obesity** contribuent fortement
- **Air_Pollution** pèse mais moins que prévu (effet partagé avec Smoking)
- **Attention :** importance ≠ causalité ; variables corrélées se partagent le poids

Conclusion

Le modèle met en évidence des facteurs de risque cohérents (tabac, pollution, âge, obésité, BRCA).
Utilité : aide à la prévention et à la priorisation des profils à surveiller.
Limites : confusions Colon/Prostate ; données synthétiques ; à valider cliniquement.