

Projet Guidé Débutant : Prédire le type de cancer à partir des facteurs de risque

Contexte du projet

Tu es **data scientist** à l'hôpital de Pointe-Noire.

Les médecins du service "Prévention des cancers" veulent comprendre **pourquoi certains patients développent un certain type de cancer** (sein, poumon, colon, etc.).

Ils te demandent d'entraîner une **intelligence artificielle** qui apprend à reconnaître les **facteurs de risque** les plus liés à chaque **type de cancer**.

Le but n'est **pas de remplacer le médecin**, mais d'aider à repérer plus vite **les profils à surveiller**.

Objectif

Créer un modèle capable de **prédire le type de cancer le plus probable** selon les habitudes de vie, les antécédents et l'environnement du patient.

Exemples :

- Un fumeur exposé à la pollution aura plus de risque de cancer du **poumon**.
- Une femme avec une mutation **BRCA** aura plus de risque de cancer du **sein**.

Le jeu de données

Chaque ligne représente **un patient**, et chaque colonne correspond à **un facteur de risque** ou une **caractéristique**.

Variables importantes (simplifiées)

Nom	Description	Exemple	Interprétation
Age	Âge du patient	65	Les cancers sont plus fréquents à partir d'un certain âge
Gender	0 = Femme, 1 = Homme	1	Certains cancers sont spécifiques au sexe
Smoking	Niveau de tabagisme (0 = pas du tout, 10 = beaucoup)	8	Plus la valeur est haute, plus le patient fume
Alcohol_Use	Niveau de consommation d'alcool (0-10)	3	Facteur aggravant pour certains cancers
Obesity	Niveau d'obésité (0-10)	7	10 = obésité sévère
Family_History	Antécédents familiaux (0/1)	1	Oui = un parent a eu un cancer
Air_Pollution	Niveau de pollution dans l'environnement	8	Pollution forte = risque plus élevé
Occupational_Hazards	Risques liés au travail (chimie, radiations...)	5	Plus la valeur est élevée, plus le risque augmente
BRCA_Mutation	Mutation génétique associée au cancer du sein	1	Présente = facteur important
Cancer_Type	🎯 Type de cancer (notre "cible à prédire")	"Lung"	C'est ce que notre modèle doit deviner

Les colonnes **Overall_Risk_Score** et **Risk_Level** ne seront **pas utilisées** car elles ont été calculées à partir des autres colonnes.

🧩 Étape 1 — Comprendre le problème

Pourquoi ?

Avant de faire du code, il faut **savoir ce qu'on veut prédire**.

👉 Ici, on veut deviner **le type de cancer** en fonction du profil du patient.

Type de problème :

C'est une **classification multiclasse** (plusieurs catégories possibles).

Ce qu'on veut faire :

- Entrée du modèle : âge, tabac, pollution, etc.
- Sortie du modèle : "probabilité d'avoir un cancer du sein / poumon / colon / peau / prostate".

🧹 Étape 2 — Nettoyage des données

Pourquoi ?

Les données réelles contiennent parfois :


- des lignes incomplètes,
- des erreurs de saisie,
- des colonnes inutiles.

Actions à faire :

1. Supprimer les colonnes inutiles : Patient_ID, Overall_Risk_Score, Risk_Level
2. Vérifier les valeurs manquantes
3. Vérifier les doublons
4. Vérifier les types de données (par ex. âge doit être un nombre)

Exemple de code :

python

 Copier

```
import pandas as pd

df = pd.read_csv("cancer-risk-factors.csv")

# 1. Supprimer les colonnes inutiles
df = df.drop(columns=["Patient_ID", "Overall_Risk_Score", "Risk_Level"])

# 2. Vérifier les valeurs manquantes
print(df.isna().sum())

# 3. Supprimer les lignes incomplètes
df = df.dropna()

# 4. Vérifier les doublons
print("Doublons :", df.duplicated().sum())
```

Étape 3 — EDA (Analyse exploratoire des données)

Pourquoi ?

Avant de construire un modèle, il faut **comprendre ce que les données racontent**.

✅ Actions à faire pour t'entraîner :

- Regarder la répartition des types de cancer

python

 Copier

```
df["Cancer_Type"].value_counts().plot(kind="bar", title="Nombre de cas par type de cancer")
```

- Comparer les moyennes de certains facteurs selon le cancer
Exemple : est-ce que les fumeurs ont plus de cancers du poumon ?

python

 Copier

```
df.groupby("Cancer_Type")["Smoking"].mean()
```

- Voir la relation entre âge et type de cancer

python

 Copier

```
import seaborn as sns
sns.boxplot(x="Cancer_Type", y="Age", data=df)
```

- Chercher les variables qui varient ensemble (corrélation)

python

 Copier

```
import matplotlib.pyplot as plt
corr = df.corr(numeric_only=True)
plt.figure(figsize=(10,6))
sns.heatmap(corr, cmap="coolwarm", annot=False)
plt.show()
```

- **Repérer les valeurs extrêmes (outliers)**

Par exemple, un âge de 5 ans ou 200 ans n'est pas logique.

Étape 4 — Préparation pour le modèle

Pourquoi ?

Les algorithmes ne comprennent **que les nombres**.

Il faut donc :

- transformer les mots en chiffres,
- et s'assurer que toutes les valeurs sont dans des plages correctes.

Étapes à faire :

1. Transformer Cancer_Type (le texte) en chiffres.
2. Normaliser les valeurs si besoin (pour que tout soit à la même échelle).
3. Vérifier que la cible (y) est bien séparée des autres colonnes (X).

Code :

python

 Copier

```
from sklearn.preprocessing import LabelEncoder

encoder = LabelEncoder()
df["Cancer_Type"] = encoder.fit_transform(df["Cancer_Type"])

X = df.drop(columns=["Cancer_Type"])
y = df["Cancer_Type"]
```

Étape 5 — Séparation des données (entraînement et test)

Pourquoi ?

C'est comme un examen :

on **apprend** sur une partie des données (train),

et on **teste** ce qu'on a appris sur une autre (test).

Code :

python

 Copier

```
from sklearn.model_selection import train_test_split  
  
X_train, X_test, y_train, y_test = train_test_split(  
    X, y, test_size=0.2, random_state=42, stratify=y  
)
```



Étape 6 — Choisir le modèle



But de cette étape :

Essayer plusieurs modèles pour trouver celui qui prédit le mieux **le type de cancer**.



Les modèles utiles dans ce genre de projet :

Modèle	Description simple	Avantages	Inconvénients
Logistic Regression	Essaye de tracer une frontière entre les classes	Simple, rapide, facile à expliquer	Ne marche pas bien si les données ne sont pas "linéaires"
Decision Tree	Arbre de décision (ex. : "Si fumeur>7 → cancer du poumon")	Très intuitif et visuel	Peut vite surapprendre (overfitting)
Random Forest	Ensemble de plusieurs arbres de décision	Précis, robuste, bon choix par défaut	Moins facile à expliquer
XGBoost / LightGBM	Arbres plus "intelligents", apprennent vite	Très performants	Plus techniques à régler
KNN (K plus proches voisins)	Compare un nouveau patient aux plus semblables	Facile à comprendre	Lent avec beaucoup de données
Naive Bayes	Basé sur des probabilités	Simple, efficace si les variables sont indépendantes	Moins précis si les variables sont corrélées

🏆 Recommandation pour débuter :

Commence avec **Random Forest**.

C'est un bon équilibre entre **facilité** et **performance**.

Code :

```
python
from sklearn.ensemble import RandomForestClassifier

model = RandomForestClassifier(random_state=42)
model.fit(X_train, y_train)
```

 Copier

Étape 7 — Évaluer ton modèle

Pourquoi ?

Pour savoir si ton modèle “pense bien” ou “se trompe souvent”.

Mesures principales :

- **Accuracy (précision globale)** : % de bonnes réponses.
- **Classification report** : détail par type de cancer (utile en santé).

Code :

```
python

from sklearn.metrics import accuracy_score, classification_report

y_pred = model.predict(X_test)

print("Taux de réussite :", accuracy_score(y_test, y_pred))
print("\nRapport détaillé :\n", classification_report(y_test, y_pred))
```

 Copier

🧩 Étape 8 — Interprétation des résultats

Pourquoi ?

Un modèle utile doit être **compris** par les médecins.
On doit savoir **quelles variables** influencent le plus les prédictions.

Code simple :

```
python

import numpy as np

importances = model.feature_importances_
indices = np.argsort(importances)[::-1]

for i in indices[:10]:
    print(f"{X.columns[i]} : {importances[i]:.3f}")
```

 Copier

🩺 Étape 9 — Conclusion du projet

Résumé :

- Le modèle prédit le **type de cancer** à partir des **habitudes et antécédents**.

- On peut maintenant identifier les **facteurs les plus dangereux** selon le type.
- Le modèle peut être intégré dans un outil pour les médecins.

Exemple d'interprétation :

- **Tabagisme + pollution** = fort lien avec le **cancer du poumon**.
- **BRCA Mutation** = lié au **cancer du sein**.
- **Obésité + faible activité physique** = lien avec **colon** et **prostate**.



Pour aller plus loin (facultatif)

Tu peux :

- tester d'autres modèles (ex : XGBoost),
- visualiser la matrice de confusion,
- créer un **petit tableau de bord** (avec Power BI, Streamlit ou Dash),
- ou même une **API** qui donne le type de cancer probable à partir d'un formulaire patient.