République Tunisienne
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique Université de Sousse

Institut Supérieur des Sciences Appliquées et de Technologie de Sousse

DEPARTEMENT INFORMATIQUE

RAPPORT DU PROJET BASES DE DONNEES AVANCEES

Prédiction du risque d'atteinte par le cancer du poumon à partir des données reliées au style de vie

Elaboré par :

Ahmed Khmis , Mohamed Chachia, Aymen Hammami

Groupe :

FIA2-GL-01

Enseignante :

Mme. Ammar Emna

Année Universitaire : **2021 / 2022**

Table des matières

1 Résumé :

1. Problématique :

Le cancer de poumons, l'un des types de cancer le plus répandues au monde, est de plus en plus en propagation.

En effet, nos données personnelles ainsi que quelques habitudes quotidiennes, souvent négligés peuvent être le signe d'alarme à l'augmentation de risque pour le cancer.

Donc notre projet a comme rôle d'analyser, étudier et mettre l'accent sur ses facteurs pour sensibiliser au cancer d'une manière orienté données.

2. La méthode de prédiction appliquée :

Pour faire une étude du prédiction du cancer, nous allons recours à

Intelligence Artificielle (AI) [1].

En fait, L'intelligence artificielle consiste à mettre en œuvre un certain nombre de techniques visant à permettre aux machines d'imiter une forme d'intelligence réelle. L'IA se retrouve implémentée dans un nombre grandissant de domaines d'application.

Parmi les résultats obtenus par cette analyse, nous trouvons :

La consommation combinée du l'alcool et du tabac augmente exponentiellement le risque d'atteindre par le cancer du poumon.

Il existe une possibilité substantielle que l'on puisse encore avoir un cancer du poumon sans fumer ni consommer d'alcool.

3. Conclusion principale :

L'augmentation de nombre de cas d'atteint pour le cancer fait appel à revoir ses habitudes quotidiennes, de plus il faut faire un bilan de santé chaque année pour tout le monde pour le prévoir le plus tôt possible (on peut être atteint ayant un style de vie sain).

2 Introduction :

1. Définition du domaine :

Dans ce projet, nous nous concentrons dans le domaine du santé, et plus précisément, la biostatistique.

Nous allons étudier les données d'un data set de personnes atteints et non atteints de cancer et nous allons observer leur données généraux ainsi que leur habitudes de styles de vie.

2. Les lacunes du domaine :

Les causes directes de cancer ne sont pas encore connues jusqu'a présent .Les études ont montré les facteurs à risques ,mais n'ont pas bien mis en évidence une corrélation ou un potentiel risque augmenté suite à la combination de ces facteurs.

De plus,on a un manque de données sur les nouveaux cas des pays sous developpés ce qui peut réduire la certitude des résultats(différents styles de vie).

les questions à poser sont donc:

-Quel rapport entre les différents facteurs?

-Les cas atteints par le cancer sans facteurs à risque,quel pourcentage ?

-Quel est la limite de cette prédiction ?

3. Objectif du projet :

Parmi les objectifs à atteindre dans ce projet, on trouve :

- Déterminer les données (facteurs) qui ont un rapport avec l'atteint du cancer
- Mettre en œuvre la corelation entre ces différents facteurs

3 Analyse comparative :

1. Comparer l'outil utilisé avec d'autres outils du BI :

Dans cette partie nous allons comparer notre étude effectuée avec deux outils d'analyse utilisés à l'informatique décisionnelle ou Business Intelligence (BI) qui sont « Jupyter » et « Power BI ».

Caractéristiques	Cette Etude	Jupyter	Talend
Outil Gratuit	X	X	X
Analyse Personnalisé	X		X
Rapport		X	
Génère un tableau de bord	X	X	X
Génère un rapport		X	
Taux d'exactitude	X	X	
Processus Lourd			X
Utilisation possède une formation			X
Prérequis en Python	X	X	
Lourd en mémoire		X	X

2. Comparer la prédiction avec une autre prédiction similaire :

La prédiction similaire est un programme fait mieux que détecter le cancer, il calcule le risque que celui-ci survienne dans les cinq ans par une analyse fine des structures tissulaires.

effectue par **Amin Emad, Morag Park** 2 chercheurs à **Genome Québec** en Canada [4].

Caractéristiques	Cette Etude	L'étude similaire
Méthode	Machine Learning	Machine Learning
Algorithme	Algorithmes python (XGB Classifier), Random Forest	Tree Decision
Outil de test	Algorithmes de l'erreur relative et absolue	LandSat7 et LandSat8 images
Zone	Allmagne - France	Canada
Type du patient	309 patients du different aspect	Plus de 10.000 patient
Nature des données	Informations générales relié à la personne	Information qui concernes les symptomes

4 Méthodologie : Processus de et techniques de collecte et d'analyse des données :

4.1 Les ressources utilisées :

Pour étudier les deux phénomènes de plantation et de déforestation, nous avons utilisé 4 datasets qui contiennent un ensemble d'informations pertinentes comme le montre ci-dessous :

- Data-Initial.csv : [6] Ce dataset contient 13 champs principaux qui sont des données personnel avec des habitudes de vie. Voici quelques exemples :
  - ✓ Gender
  - ✓ Age
  - ✓ Smoking
  - ✓ Yellow_Fingers
  - ✓ Anxiety
  - ✓ Peer Pressure
  - ✓ Chronic Disease
  - ✓ Alcohol Consuming

Ce datasets possèdent 309 lignes qui sont titré par les indacteurs suivants : (M = Male , F=Female ,1=NO , 2 = YES)

4.2 La chaine décisionnelle :

L'informatique décisionnelle possède une chaine décisionnelle importante pour que les résultats obtenues soit précises. Cette chaine contient cinq phases de bases.

4.2.1 Planification :

Cette phase consiste déterminer l'ensemble de taches à faire tout au long du processus. Dans notre cas, les principales taches à faire sont :

- Préciser le domaine de prédiction.
- Collecter les données. Cette tâche est la plus difficile du processus.
- Extraire les données de l'ensemble des datasets.
- Choix des centres d'intérêt et des axes d'analyse.
- Transformer les données vers un format homogène.
- Charger les données dans **l'entrepôt de données (ED)** ou le **Data WareHouse (DW).** (C'est une base de données homogène qui aide la prise de décision).
- Modéliser les données selon un modèle bien défini (modèle en étoile dans notre cas).
- Analyser les données : développer un ensemble d'algorithme d'analyse des données.
- Générer un tableau de bord.
- Déterminer des rapports à partir du tableau de bord.
- Prendre des décisions.

4.2.2 Nettoyage et normalisation des données :

Cette phase consiste à mettre les données sous formes standardisé(M , F => 1 , 0) et à supprimer les champs non nécessaire

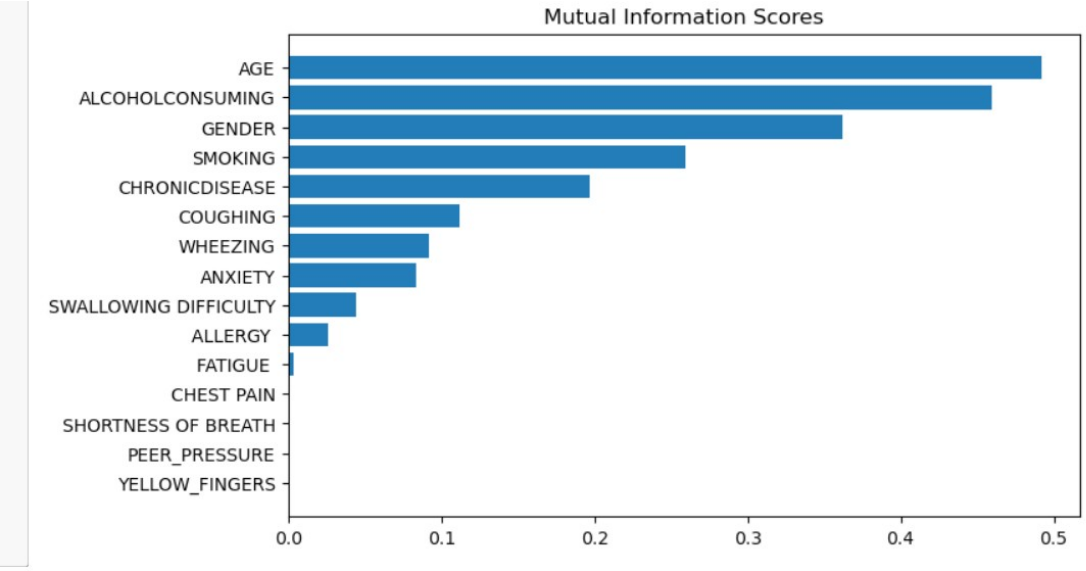
```
# X.head()
key_rev = {'YES' : 1, 'NO' : 0}

df = df.replace(key_rev)
df.head(5)
```

```
#Changing Value for Gender Column Male : 1, Female : 0

df['GENDER'] = df['GENDER'].replace({'M' : 1, 'F' : 0})
df['GENDER'].value_counts()
```

4.2.3 Balance de données

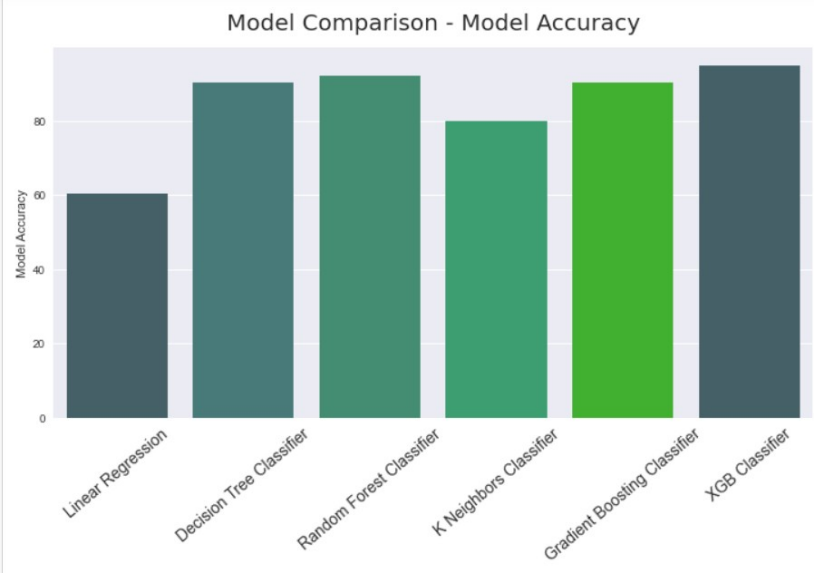


Nous avons choisit les facteurs qui ont dépasser une valeur de 0.2 (le milieu) au dessous du milieu les facteurs n'ont plus une relation avec le phénoméne étudié

4.2.4 Segmentation des données

Répartition des données :25 % des données dédiées au test et 75 % pour l'apprentisage du model

4.2.5 Test avec plusieur algorithmes



les différents model du traitement appliqué ont des résultats inégales selon les algorithmes utilisées et précisemment le model XGB Classifier que nous avons utilisé dans notre application.

```
1  XGBClassifierScore = xgb.score(X_test,y_test)
2  print("Précision obtenue par XGB Classifier model:",XGBClassifierScore*100)
```

```
Accuracy obtained by XGB Classifier model: 94.07407407407408
```

Le taux de précision et l'efficacité du model selon l'algorithme "XGB Classifier" est 94.95 %.

```
1  KNeighborsClassifierScore = knn.score(X_test, y_test)
2  print("Précision obtenue par K Neighbors Classifier model:",KNeighborsClassifierScore*100)
```

```
Précision obtenue par K Neighbors Classifier model: 80.0
```

```
5  lr = LinearRegression()
6  lr.fit(X_train, y_train)
7  LinearRegressionScore = lr.score(X_test,y_test)
8  print("Précision obtenue par Linear Regression model:",LinearRegressionScore*100)
```

```
Précision obtenue par Linear Regression model: 60.40411179294546
```

4.2.6 Analyse

```
DecisionTreeClassifierScore = dtc.score(X_test,y_test)
print("Précision obtenue par Decision Tree Classifier model:",DecisionTreeClassifierScore*100)
```

```
Précision obtenue par Decision Tree Classifier model: 90.29629629629629
```

La phase d'analyse est la quatrième phase du processus effectué dans

```
Entrée [43]:
RandomForestClassifierScore = rfc.score(X_test, y_test)
print("Précision obtenue par Random Forest Classifier model:",RandomForestClassifierScore*100)
```

```
Précision obtenue par Random Forest Classifier model: 92.29629629629629
```

l'informatique décisionnelle. En fait, cette étape se focalise à la définition d'un ensemble d'algorithmes python de traçage des courbes liées aux données

```
GradientBoostingClassifierScore = gb.score(X_test,y_test)
print("Précision obtenue par Gradient Boosting Classifier model:",GradientBoostingClassifierScore*100)
```

```
Précision obtenue par Gradient Boosting Classifier model: 90.29629629629629
```

collectées.

Parmi ces courbes nous avons choisis :

- ✓ Diagramme de Venn : visualisation des données sous formes du cercle ayant des intersections (point commun)
- ✓ Histogramme: Clarifier les données binaires .

Nous avons défini aussi dans cette étape des algorithmes qui permet de déterminer le taux d'exactitude ou le degré de précision des données utilisées dans la base décisionnelle.

Parmi les algorithmes nous avons choisis :

- ✓ Taux d'exactitude pour chaque Personne.
- ✓ Taux d'exactitude du monde en intégrant le nombre maximum d'évaluations pour chaque candidat.

4.2.7 Restitution :

La restitution est la dernière étape du processus. Dans cette étape, nous avons générer un ensemble de rapports et de tableaux de bord. Ces derniers nous permettent de prendre une décision finale selon les conditions et les circonstances.

4.3 Schématisation du processus :

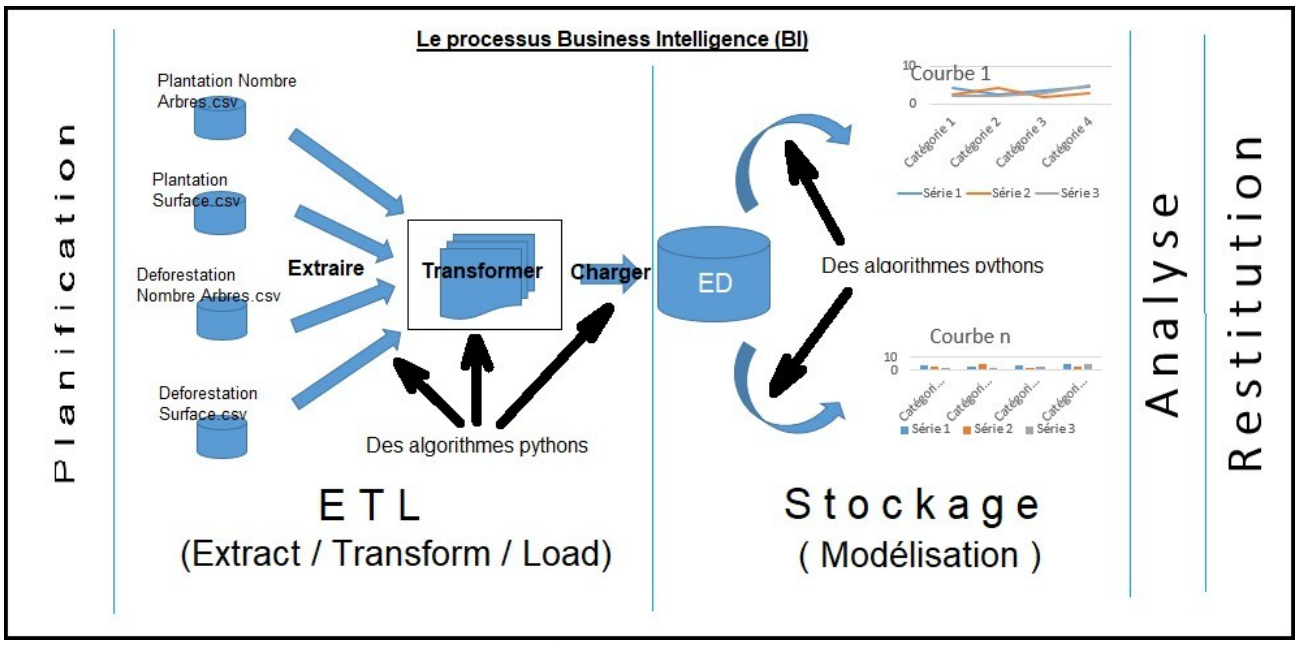


Figure 3: Schématisation du processus

5 Résultats

5.1 Histogramme montrant l'atteint au cancer en fonction de l'age

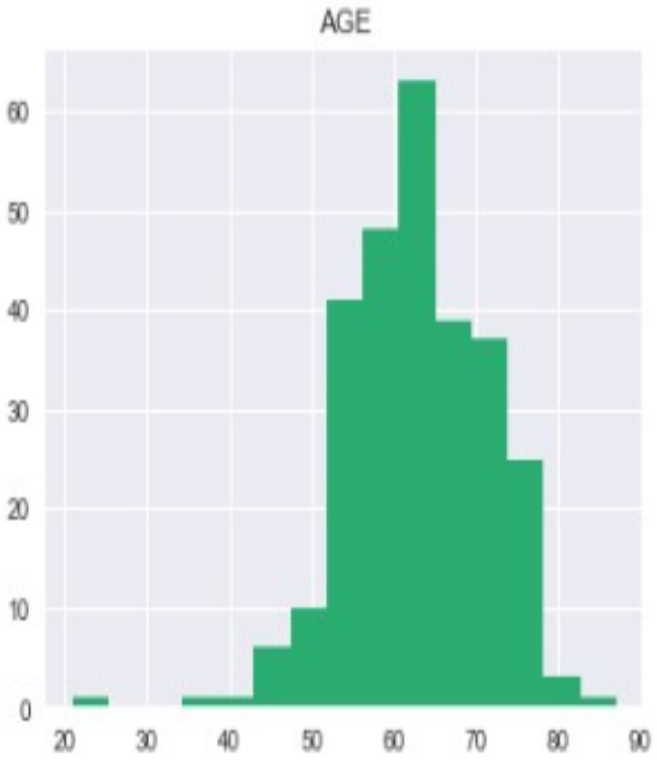


Figure 4: Histogramme selon tranche d'âge

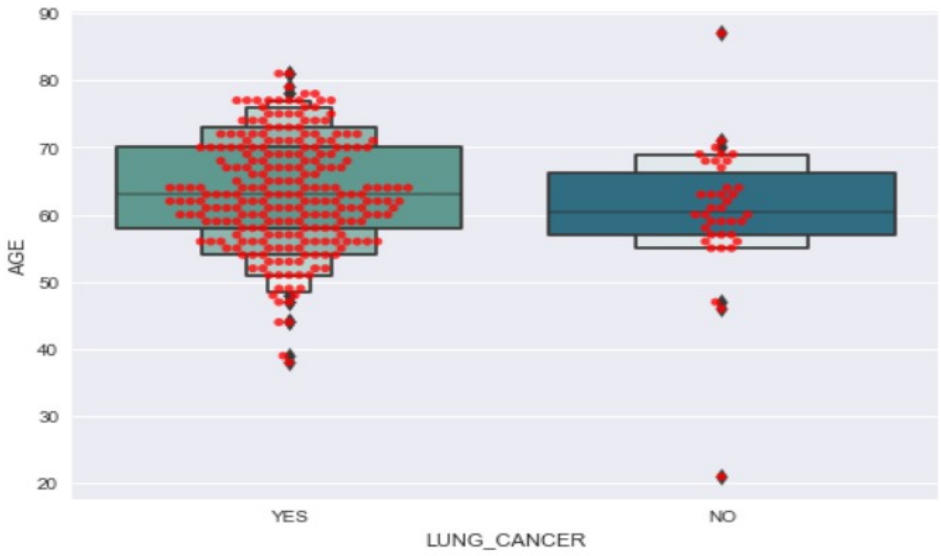


Figure 5: Diagramme l'infection selon tranche d'âge

5.2 Diagramme de Venn d'intersection des habitudes à risque avec l'atteint au cancer :

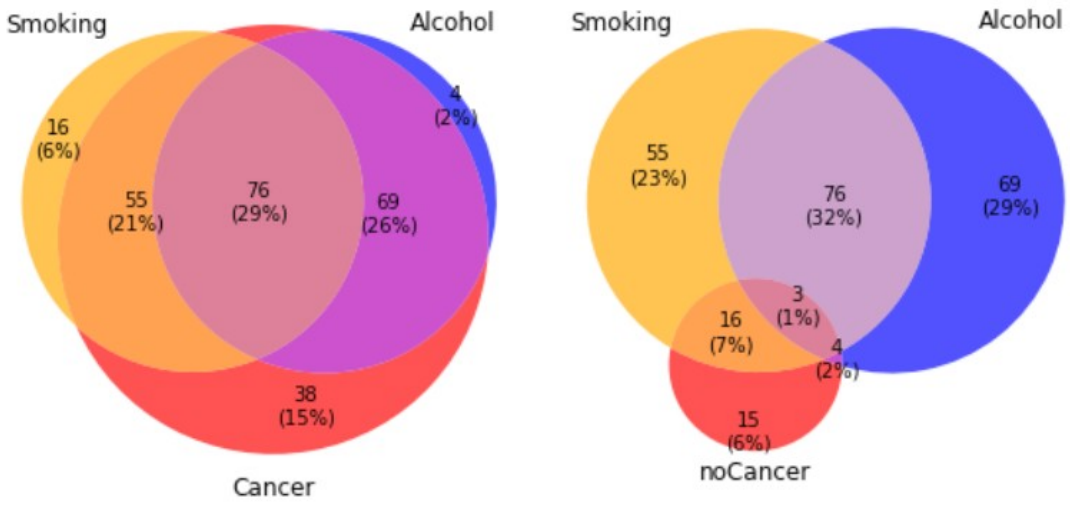


Figure 6: Diagramme de venn d'habitude des patients

5.3 Mise en œuvre du risque obtenue lors du combinaison des facteurs

Figure 7: Combinaison des facteurs et la probabilité

6 Discussion

6.1 Discussion des résultats :

- D'après la figure 4, nous constatons que :
  - ✓ La possibilité d'étre infecté par le cancer de plus on plus propable lorsque l'age augment précisemment entre [50 , 79].

- D'après la figure 5, nous constatons que :
  - ✓ Dans la population âgée, cette pathologie doit être considérée différemment qu'au jeune âge. En effet, les caractéristiques, la prise en charge, et surtout l'évolution de la maladie sont bien distinctes chez le patient âgé.
- D'après la figure 6, nous constatons que :

Pour mieux visualiser les résultats on a choisit d'utiliser le diagramme de Venn pour mieux mettre en évidence la correlation entre les différents facteurs à risque.

Analyse :

En effet le premier diagramme montre :
Violet : 26% alcool seulement + atteint au cancer
Violet claire: 29% alcool + tabac + atteint au cancer
Orangé : 21% tabac seulement + atteint au cancer
Jaune: 6% tabac seulement + non atteint au cancer
Bleu : 2% alcool seulement + non atteint au cancer
Rouge :15% ni alcool ni tabac + non atteint au cancer

Intérprétation :

La plus grande population de personne atteint au cancer sont ceux qui consomme de l'alcool et fume au même temps (29 %).
Si on ne fume pas et on ne cosomme pas de l'alcool on a une probabilité minoritaire de 15% (0.15) " de ratrapper le cancer "

Conclusion :

la consommation **combiné** du l'alcool et du tabac **double la possibilité** d'être atteint par le cancer.

6.2 Evaluation :

En se basant principalement et uniquement sur ce dataset:

- L'anxieté et les facteurs sociales associés (peer pressure) sont loin d'étre responsable de l'évolution et l'atteinte par le cancer de poumons.
- Selon le dataset ,l'alcool SEULE augmente plus le risque d'atteinte par le cancer que le tabagisme SEUL
- La consommation combiné d'alcool et de tabac prend la majeure parti des patients de cancer de poumons(plus que la moitié 55%).Combiné,ce sont des pratiques à très haut risque
- L'age joue un role plus élevé que toutes les autres facteurs,à partir des quarantaines,il faut bien revoir ses habitudes quotidiennes
- L'incidence du cancer augmente régulièrement au cours de la vie. Le cancer chez les 65 ans et plus représentent ainsi 62,4 % des cancers estimés tous âges confondus en 2021
- Chez l'homme, le cancer du poumon dont l'incidence estimée est la plus élevée après 65 ans que les femmes (20,214 nouveaux cas estimés en 2022 par rapport 9,328 cas chez les femmes)

6.3 Limites, recommandations :

Bien que nous ayons pu extraire des statistiques importantes du plusieur dataset, nous avons eu du mal à obtenir un jeu de données avec de nombreux attributs et capable de couvrir le nombre maximum d'axes dont dépend l'infection du cancer du poumon

En revanche, il serait très utile de récupérer des résultats d'analyse liés au pays d'un individu plutôt qu'un géante dataset qui ne fait pas de différence en termes de pays d'origine des individus, notamment au profit du secteur national de la santé .

7 Conclusion

Les choses à retenir de cette étude sur ce dataset:

- En termes de données personnels,l'age est le premier facteur.On ne commence pas à considèrer sérieusement le risque d'atteinte par le cancer que après l'age de quarentaine(l'age moyen étant 65 ans)
- En terme d'habitude unique(non combiné) et contrairement à ce qui est répandu ,l'alcool est le premier facteur à risque et non pas le tabagisme .

  Combiné les deux,il occupe la moitié des personnes affecté
- Tout ces conclusions sont des probabilités et non pas des causes directes ,la seule manière de confirmation est le bilan médical

8 Références

[1] https://www.genomequebec.com/239-projet/developpement-de-modeles-d-intelligence-artificielle-pour-predire-la-reponse-aux-combinaisons-de-medicaments-chez-les-patients-atteints-d-un-cancer-ayant-un-mauvais-pronostic/

[2] https://kaduceo.com/prediction-de-deces-precoces/

[3] https://www.em-consulte.com/article/956142/article/prediction-du-risque-cumule-de-deuxieme-cancer-pri

[4] https://www.revmed.ch/revue-medicale-suisse/2011/revue-medicale-suisse-296/cancer-et-vieillissement-une-evidence-epidemiologique

[5] https://beei.org/index.php/EEI/article/view/2532/1923

[6] https://cancer.ca/fr/cancer-information/cancer-types/lung/risks