Rapport de Projet : Analyse des Facteurs de Risque du Cancer Oral

1. Introduction

Problématique :

Le cancer oral est une pathologie souvent liée à des comportements comme le tabagisme et la consommation d'alcool. L'objectif de ce projet est d'analyser un jeu de données pour identifier les facteurs de risque associés à cette maladie, en se concentrant principalement sur le tabagisme et la consommation d'alcool. Nous avons utilisé des méthodes statistiques classiques, telles que la corrélation et les tests Chi-Carré, pour explorer ces relations.

Objectifs :

- Identifier les variables les plus corrélées à la présence de cancer oral.
- Tester les associations statistiques entre des facteurs comme le tabagisme et l'alcool et la présence de cancer oral.

2. Analyse des Données

2.1. Présentation du Jeu de Données :

Le jeu de données utilisé contient plusieurs variables concernant des comportements à risque, comme le tabagisme et la consommation d'alcool, ainsi que des informations démographiques et médicales. La variable cible est **Oral Cancer (Diagnosis)**, qui indique si le cancer oral a été diagnostiqué.

Les principales variables incluent :

- **Tabagisme** (Tobacco Use)
- **Consommation d'alcool** (Alcohol Consumption)
- **Infection par le virus HPV** (HPV Infection)
- **Hygiène buccale** (Poor Oral Hygiene)

2.2. Exploration des Données :

Une analyse des corrélations a été effectuée pour identifier les relations entre les différentes variables et la variable cible. Les résultats obtenus, notamment à partir de la **matrice de corrélation**, ont montré des **corrélations élevées entre certaines variables**, mais les comportements de tabagisme et d'alcool ne semblent pas être des facteurs significatifs dans la prédiction du cancer oral, ce qui est incohérent avec la littérature scientifique.

Matrice de Corrélation des Variables :

La matrice de corrélation montre les relations entre les variables. Voici la matrice affichée lors de l'analyse :

Les variables comme la **taille de la tumeur** et le **stade du cancer** sont fortement corrélées avec le diagnostic de cancer oral, ce qui est attendu. Cependant, **tabagisme** et **consommation d'alcool** montrent des corrélations faibles ou nulles avec la variable cible, ce qui est surprenant.

Corrélation avec le Cancer Oral :

Un graphique supplémentaire a été produit pour examiner spécifiquement les variables corrélées avec la présence de cancer oral. Ce graphique indique que des facteurs comme la **taille de la tumeur** et le **stade du cancer** sont très fortement corrélés avec le diagnostic de cancer oral, mais **le tabagisme** et **l'alcool** apparaissent comme faiblement corrélés.

3. Méthodologie

3.1. Modèle Statistique :

Nous avons utilisé des méthodes statistiques classiques pour analyser les relations entre les variables et le cancer oral :

- **Matrice de corrélation** pour explorer les relations linéaires entre toutes les variables.
- **Test du Chi-Carré** pour tester l'association entre les variables catégorielles (tabagisme, alcool) et la présence du cancer oral.

3.2. Hypothèses :

- **Hypothèse 1 :** Il existe une association significative entre le tabagisme et le cancer oral.
- **Hypothèse 2 :** Il existe une association significative entre la consommation d'alcool et le cancer oral.

4. Résultats

4.1. Résultats des Tests Chi-Carré :

Les tests Chi-Carré ont été appliqués pour tester l'association entre **tabagisme**, **alcool** et la présence de cancer oral. Les résultats sont présentés ci-dessous :

Test Chi-Carré pour le tabagisme :

- **p-value pour le tabagisme : 0.5864**
- Cette p-value étant supérieure à 0.05, nous **ne rejetons pas l'hypothèse nulle**. Il n'y a donc pas d'association significative entre le tabagisme et le cancer oral dans ce jeu de données.

Test Chi-Carré pour la consommation d'alcool :

- **p-value pour l'alcool : 0.6457**
- La p-value étant également supérieure à 0.05, nous **ne rejetons pas l'hypothèse nulle**. Aucune association significative n'est observée entre la consommation d'alcool et la présence de cancer oral.

4.2. Interprétation des Résultats

Les résultats des tests Chi-Carré indiquent que **ni le tabagisme, ni la consommation d'alcool ne sont significativement associés au cancer oral** dans ce jeu de données. Cependant, cela semble incohérent avec le sens commun qui démontre généralement une forte corrélation entre ces comportements et le cancer oral.

5. Analyse Critique

5.1. Forces :

- Les visualisations et la matrice de corrélation offrent une bonne vue d'ensemble des relations entre les différentes variables.
- Le test du Chi-Carré est une méthode classique et appropriée pour tester l'association entre des variables catégorielles.

5.2. Limites et Incohérence des Résultats :

Les résultats obtenus ne semblent pas cohérents avec les attentes basées sur la recherche scientifique. Le **tabagisme** et **la consommation d'alcool** devraient logiquement montrer des associations significatives avec le cancer oral, mais cela n'a pas été le cas. Cette incohérence pourrait être due à plusieurs facteurs :

1. **Problèmes avec le jeu de données :** Le fichier CSV pourrait être mal structuré, contenir des données manquantes ou incorrectement étiquetées, ce qui pourrait fausser les résultats des analyses.
2. **Erreurs dans le pré-traitement des données :** La transformation des données catégorielles en variables numériques via un **label encoding** pourrait avoir introduit des biais, surtout si certaines catégories sont mal représentées ou inégales.

5.3. Pistes d'Amélioration :

- Il est fortement recommandé de **vérifier et nettoyer le jeu de données** pour détecter d'éventuelles erreurs ou anomalies dans les valeurs du CSV.
- Une approche de **régression logistique** ou d'autres modèles plus sophistiqués pourrait être envisagée pour mieux capturer les relations entre les variables.
- Il serait également pertinent d'intégrer **plus de données** sur d'autres facteurs de risque non présents dans ce jeu de données.

6. Conclusion

Cette étude révèle que **ni le tabagisme, ni la consommation d'alcool** ne sont associés de manière significative au cancer oral dans ce jeu de données, ce qui est incohérent avec les attentes basées sur la littérature scientifique. Cette incohérence pourrait être due à des problèmes dans la structuration des données, ce qui suggère la

nécessité de vérifier et de nettoyer le fichier CSV. Un travail de pré-traitement plus approfondi et l'utilisation de modèles plus complexes pourraient améliorer la qualité des résultats et des prédictions futures.