Prédiction de maladies pulmonaires liées à la consummation de chicha

Par : DIANE Kallo Mory

Formation : Master en Systèmes d'Information

Professeur : M. TRAORE

Résumé

Ce projet de recherche a pour objectif de prédire le risque de développement de maladies pulmonaires chez les consommateurs de chicha en s'appuyant sur des techniques de Machine Learning. Pour cela, un jeu de données synthétique a été conçu, intégrant plusieurs variables explicatives pertinentes : l'âge et le sexe des individus, la durée et la fréquence de consommation de chicha, le nombre de cigarettes consommées parallèlement, l'exposition passive à la fumée, ainsi que la pratique ou non d'activités sportives.

Plusieurs modèles d'apprentissage supervisé ont été expérimentés afin de comparer leur performance prédictive : la régression logistique, l'arbre de décision et la Random Forest. L'évaluation de ces modèles a été réalisée à l'aide de métriques reconnues en classification, notamment la précision globale (accuracy), la matrice de confusion permettant d'analyser les erreurs de prédiction, et les courbes ROC/AUC mesurant la capacité de distinction entre personnes à risque et non à risque.

Les résultats obtenus mettent en évidence que [insérer résumé des performances, par exemple : la Random Forest présente la meilleure précision et une AUC élevée, tandis que la régression logistique fournit une bonne interprétabilité des variables]. L'analyse des facteurs les plus influents suggère que [insérer ici : par exemple la fréquence de consommation et l'exposition passive sont parmi les variables déterminantes].

En définitive, cette étude illustre le potentiel du Machine Learning comme outil d'aide à la décision dans le domaine de la santé publique. Les modèles développés pourraient contribuer à renforcer les actions de prévention et à favoriser un dépistage plus précoce des risques pulmonaires liés à la consommation de chicha, en particulier chez les jeunes adultes.

1. Introduction

La consommation de chicha, également appelée narguilé, est devenue une pratique courante, notamment chez les jeunes adultes. Bien qu'elle soit souvent perçue comme moins nocive que la cigarette, de nombreuses études scientifiques ont démontré qu'elle entraîne des risques significatifs pour la santé, en particulier pour le système respiratoire. L'inhalation de fumée de chicha expose les poumons à des substances toxiques telles que le monoxyde de carbone, les métaux lourds et diverses particules fines, pouvant favoriser l'apparition de maladies pulmonaires chroniques.

Dans ce contexte, l'identification précoce des individus à risque constitue un enjeu majeur de santé publique. Détecter les profils vulnérables avant l'apparition des premiers symptômes permettrait non seulement d'améliorer la prévention, mais aussi de mettre en place des stratégies de dépistage ciblées et d'encourager des changements de comportements.

L'objectif principal de ce projet est donc de développer un modèle prédictif basé sur des données liées aux habitudes de consommation et au mode de vie. Ce modèle vise à classifier les patients en deux catégories : **« sain »** ou **« à risque de maladie pulmonaire »**. Pour ce faire, plusieurs variables sont prises en compte, telles que la fréquence et la durée de consommation de chicha, l'âge, le sexe, le tabagisme parallèle, l'exposition passive à la fumée ainsi que la pratique sportive. L'approche retenue repose sur des méthodes de Machine Learning, permettant d'analyser ces facteurs et d'estimer la probabilité de développer des complications respiratoires.

2. Méthodologie

2.1 Description des données

Dataset : [500 observations, 7 variables explicatives + 1 variable cible].

Variables :

- Age (années)
- Sexe (0 = femme, 1 = homme)
- DuréeConsommation (années)
- FrequenceParSemaine

- NbCigarettesParJour
- ExpositionPassive (0 = non, 1 = oui)
- Sport (0 = non, 1 = oui)
- MaladiePulmonaire (0 = sain, 1 = malade)

2.2 Exploration des données

- Distribution de la cible : [Insérer figure countplot]
- Histogrammes des variables : [Insérer figure histograms]
- Corrélations : [Insérer heatmap]

2.3 Prétraitement

- Standardisation des variables continues.
- Split train/test 80/20 stratifié.
- Pas de valeurs manquantes dans le dataset synthétique.

3. Modélisation

3.1 Modèles utilisés

- Régression logistique (Logistic Regression)
- Arbre de décision (Decision Tree, max_depth=5)
- Random Forest (pour comparaison)

3.2 Évaluation

- Accuracy, matrice de confusion, classification report.
- Courbes ROC et AUC.
- Validation croisée stratifiée (Stratified K-Fold 5).

4. Résultats

4.1 Performance des modèles

Modèle	Accuracy	AUC
Logistic Regression	[ ]	[ ]
Decision Tree	[ ]	[ ]
Random Forest	[ ]	[ ]

- Confusion matrices : [Insérer figures]

- Courbes ROC : [Insérer figure ROC]

4.2 Importance des variables

- Decision Tree : [Insérer figure barplot importance features]
- Logistic Regression : [Insérer tableau coefficients]

5. Discussion

- L'arbre de décision permet d'identifier les variables les plus influentes (DuréeConsommation, FréquenceParSemaine, ExpositionPassive).
- La régression logistique fournit des coefficients interprétables.
- Limites : dataset synthétique, simplification des habitudes, absence de certaines variables environnementales ou génétiques.
- Perspectives : utiliser un dataset réel, tester d'autres modèles, deep learning.

6. Conclusion

- Le projet a permis de construire et comparer plusieurs modèles de prédiction des maladies pulmonaires liées à la chicha.
- Les modèles peuvent servir de support pour le dépistage préventif et la sensibilisation aux risques de consommation.
- Les résultats sont prometteurs mais doivent être validés sur des données réelles.

7. Annexes / Code

Inclure le Jupyter Notebook complet avec toutes les cellules et figures.
Exemple de prédiction sur nouveaux patients :

```python
nouveaux = pd.DataFrame([
    [30,1,5,4,0,0,1],
    [45,0,10,8,5,1,0],
    [22,1,1,1,0,0,1]
], columns=X.columns)
nouveaux_scaled = scaler.transform(nouveaux)
print(lr.predict(nouveaux_scaled))

print(lr.predict_proba(nouveaux_scaled)[:,1])
```