Smoke Signals: Predict Smokers by Vital Signs

Saketh Ragirolla 2021092
Saumil Lakra 2021097
Ishwar Babu 2021532
Rahul Oberoi 2021555

1. Motivation

Smoking is a leading cause of cancer, stroke, heart attack, and lung disease and it can also escalate the risk of depression and anxiety. Timely and accurate diagnosis has to be done to prevent such serious health consequences. Through this project, we aim to develop ML models to detect a smoker through his/her body signs such as weight, cholesterol, and hemoglobin levels. We are using the body signal of smoking kaggle dataset which consists of 55,692 rows and 27 columns.

2. Related Work

[1] Consequences of smoking for body weight, body fat distribution, and insulin resistance: This study finds the effects of smoking on a person's health. It helps in identifying important features such as body weight during and after cessation of smoking. It shall act as a means to generate features for our dataset and provide insights into how the health factors change due to smoking.
[2] Predicting Smoking Status Using Machine Learning Algorithms and Statistical Analysis: In this study, five machine learning algorithms were used to predict smoking status based on blood tests and vital reading data. The research also conducted a one-way analysis of variance to highlight differences in blood test results between smokers and non-smokers. Among the algorithms, Logistic Regression outperformed the others with precision, recall, F-measure, and accuracy of approximately **83%**.
[3] Artificial neural network machine learning prediction of the smoking behavior and health risks perception of Indonesian health professionals: This study makes predictions using Artificial Neural Networks and obtains an accuracy of **81%**. It works on a smaller dataset of 240 people and considers features such as diabetes, place of work, any respiratory disease, etc. We shall expand the research scope of this study and use this as a baseline for our models.

3. Timeline

Week 1-2:	Data Cleaning
Week 3:	Data Pre-Processing
Week 4:	Feature Extraction and Analysis
Week 5:	EDA & Data Visualization
Week 6-8:	Deploying ML models like Logistic Regression etc.
Week 9:	Improving the performance of models
Week 10:	Hyperparameter Tuning
Week 11-12:	Reporting the scores

4. Individual Tasks

Week	Tasks	Team
1 - 2	Data Cleaning	Saketh, Saumil
3	Data Pre-Processing	Ishwar, Rahul
4	Feature Extraction and Analysis	Saketh, Ishwar
5	Exploratory Data Analysis & Data Visualization	Saumil, Rahul
6-8	Deploying ML models like Support Vector Classifier, Logistic Regression, etc.	All members
9-10	Improving the performance of models and Hyperparameter Tuning	All members
11-12	Reporting the scores	All members

5. Final Outcome

This project aims to develop a Machine Learning model to accurately predict whether a given individual is a smoker or not given his/her vital biological signals. By applying classical algorithms, the project seeks to uncover patterns in smoker classification tasks, creating a model that is both accurate and easily interpretable, with the potential for continuous improvement to achieve even better prediction outcomes.

References

[1] Arnaud Chiolero, David Faeh, Fred Paccaud, and Jacques Cornuz. Consequences of smoking for body weight, body fat distribution, and insulin resistance. *The American Journal of Clinical Nutrition*, 87(4):801–809, Apr 2008.

[2] Charles Frank, Asmail Habach, Raed Seetan, and Abdullah Wahbeh. Predicting smoking status using machine learning algorithms and statistical analysis. *Advances in Science, Technology and Engineering Systems Journal*, 3:184–189, 2018.

[3] D Nuryunarsih, O Okatiranti, and L Herawati. Artificial neural network machine learning prediction of the smoking behavior and health risks perception of indonesian health professionals. *Environmental Analysis Health and Toxicology*, 38(1):e2023003–0, Mar 2023. PMID: 37100398; PMCID: PMC10195675.