**Project Proposal**

Group 84

Venkata Subrahmanya Ankit Vennelakanti

Arijit Mukherjee

vennelakanti.v@northeastern.edu

mukherjee.ari@northeastern.edu

**Percentage of Effort Contributed by Student 1: 50**

**Percentage of Effort Contributed by Student 2: 50**

**Signature of Student 1: Ankith Vennelakanti**

**Signature of Student 2: Arijit Mukherjee**

**Submission Date: 01/27/2023**

**Problem Setting:**

The problem of smoking is a significant public health concern with far-reaching negative effects on individuals and society. Not only does smoking harm nearly every organ of the body and increase the risk of numerous diseases, such as lung cancer and heart disease, but it also reduces the life expectancy of smokers. Despite this, smoking remains a prevalent habit and the number of deaths caused by smoking is projected to reach 10 million by 2030, according to a World Health Organization report. Furthermore, evidence-based treatment options for smoking cessation, such as counseling, have been proposed and promoted but be ineffective and time-consuming for many physicians.

**Problem Definition:**

This analysis intends to identify the best classification and predictors to identify how various attributes point toward a smoker which eventually leads to other numerous diseases. We look for information from attributes such as systolic (measures the pressure in your arteries when your heart beats), hemoglobin measurement in blood, glutamic oxaloacetic transaminase type (an enzyme found in the liver, heart, and other tissues) which measures liver and heart damage along with few more attributes.

**Data Sources:**

https://www.kaggle.com/datasets/gauravduttakiit/smoker-status-prediction

**Data Description:** There are two datasets available for addressing the problem, one is a Train dataset, and the other is a test dataset. The training dataset consists of 38984 instances and 23 attributes, whereas the test dataset consists of 16709 instances and 22 attributes. Data presents us with the physical and health parameters of each subject as mentioned below. In the test dataset, column “Smoking” is not present as this parameter will be predicted with the trained model

age (in the multiples of 5)   
height(cm)  
weight(kg)  
waist(cm): Waist circumference length  
eyesight(left)  
eyesight(right)  
hearing(left)  
hearing(right)  
systolic: Blood pressure  
relaxation: Blood pressure  
fasting blood sugar  
Cholesterol: total  
triglyceride  
HDL: cholesterol type  
LDL: cholesterol type  
hemoglobin  
Urine protein  
serum creatinine  
AST : glutamic oxaloacetic transaminase type  
ALT : glutamic oxaloacetic transaminase type  
Gtp: γ-GTP  
dental caries  
smoking