# Smoker Prediction

Using Machine Learning to make predictions on smokers from various health inputs

## Collaborators


Karl Winterling - https://github.com/kwinterling
Meri Medel - https://github.com/memusic1523
Kathryn Bottenberg https://github.com/KathrynBottenberg
Zev Bennett - https://github.com/zevbennett
Kellen Finn - https://github.com/kelmo974
Jyotsna Muppasani - https://github.com/JyotsnaMuppasani

# Why Smoker Predictions?

Smoking is Common

- Throughout the world, 20% of adults smoke tobacco products. (Our World in Data, 2013)
- Almost 14% of U.S. adults aged 18 years old or older were cigarette smokers in 2018. (CDC, 2021)
- About 16% of male adults and just over 12% of female adults were smokers in 2018 in the U.S. (CDC, 2021)

Smoking is Unhealthy

- Smokers are two to four times more likely to visit the emergency room than nonsmokers. (*Wolters Kluwer*, 2015)
- In the United States, between 80% and 90% of lung cancer deaths are linked to cigarette smoking. (CDC, 2020)
- Around 32% of coronary heart disease deaths in the United States are linked to smoking. (U.S. Surgeon General, 2014)
- Around 79% of chronic obstructive pulmonary disease (COPD) cases in the United States are linked to smoking. (U.S. Surgeon General, 2014)

Questions

1. Can we predict a smoker based on certain variables?
2. What are the top variables which attribute to a smoker vs non smoker?
3. What health problems may be a factor from smoking?

# Datasets Details

- train_df
  - 38,984 rows
  - 70.0 % of the dataset
  - No null values
- test_df
  - 16,708 rows
  - 30.0 % of the dataset
  - No null values

## 23 variables

- Age
- height(cm)
- weight(kg)
- waist(cm)
- eyesight(left)
- eyesight(right)
- hearing(left)
- hearing(right)
- Systolic
- relaxation
- fasting blood sugar
- Cholesterol
- Triglyceride
- High-Density Lipoprotein (HDL)
- Low-Density Lipoprotein (LDL)
- hemoglobin
- Urine protein
- serum creatinine
- Aspartate Transferase Blood Test (AST)
- Alanine Transaminase Blood Test (ALT)
- Gtp
- dental caries
- **Smoking * not in test_df**

# PROJECT ROADMAP

**STEP 1:**
Finding Dataset, ETL,

**STEP 2:**

**STEP 3:**
Machine Learning

**STEP 4:**
Visualizations: Tableau, Plotly, MatPlotLib

**Completed**

**Upcoming**

# Resources

Smoker Status Prediction | Kaggle- creator Gaurav Dutta