Group 19
Smoker Status Analysis using Bio-Signals

Zhiyang Zhang, Unay Shah,
Pin-Ying Wu, Zoom Chow

Outline

- Task Description
- Dataset Introduction
- Data Analysis and Visualization
- Top-influential Bio-factors
- Conclusions
- Reference
- Q & A

Task Description

- Smoking ranks as a primary factor in preventable diseases and fatalities globally, with negative impacts across multiple health aspects.
- Despite evidence-based treatments, smoking cessation rates remain low, partly due to perceived inefficacy and time constraints found in physician counseling.
- There's a critical need for physicians to effectively identify smokers who are more likely to quit. The proposed solution involves mathematical analysis of datasets to pinpoint influential bio-signals for smoker identification.

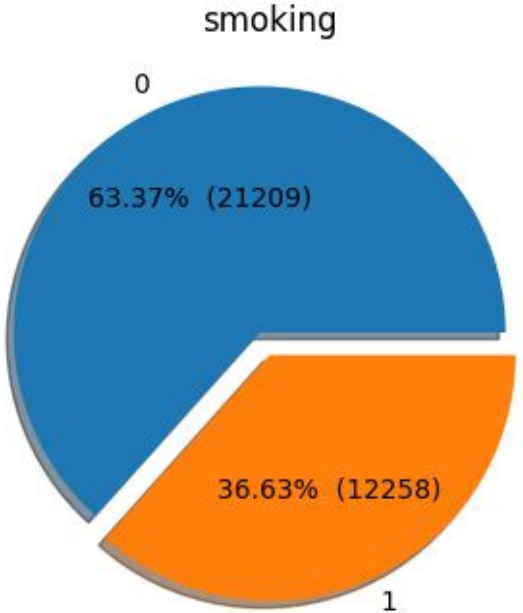
Dataset Introduction

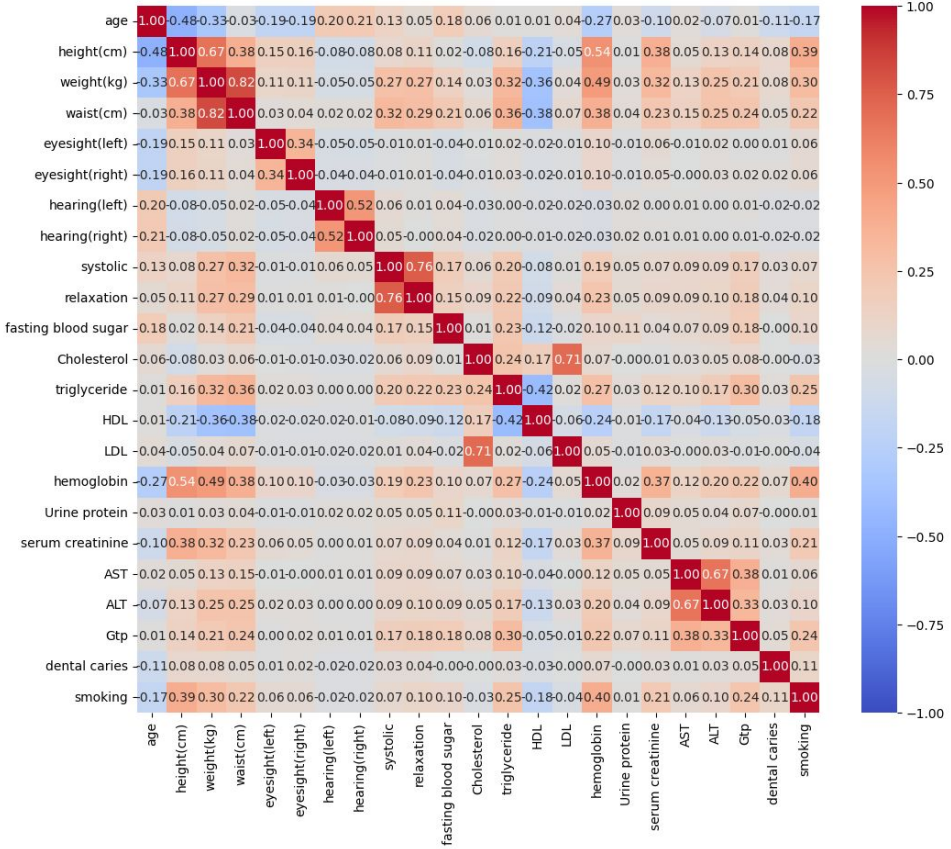
- ML Olympiad Dataset [1] contains training and testing CSV files with 23 bio-signal features per patient.
  - Numerical features: age, height, weight, waist, eyesight (left & right), systolic, relaxation, fasting blood sugar, cholesterol, triglyceride, HDL, LDL, hemoglobin, serum creatinine, AST, ALT, Gtp
  - Categorical features: hearing (left & right), urine protein, dental caries
  - Target prediction: smoking status
- We focus on the data analysis and visualization, so we only use the training data.

Data Analysis and Visualization

Relationship between Different Features

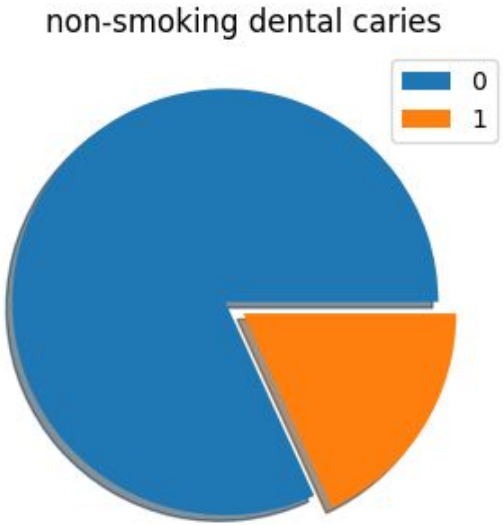
- Many possible features, some more distinctive than others

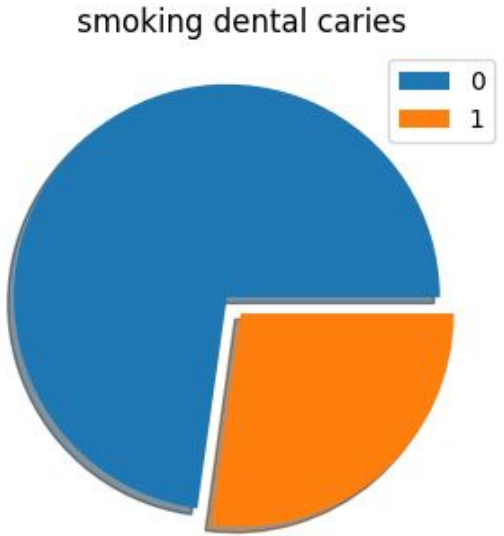




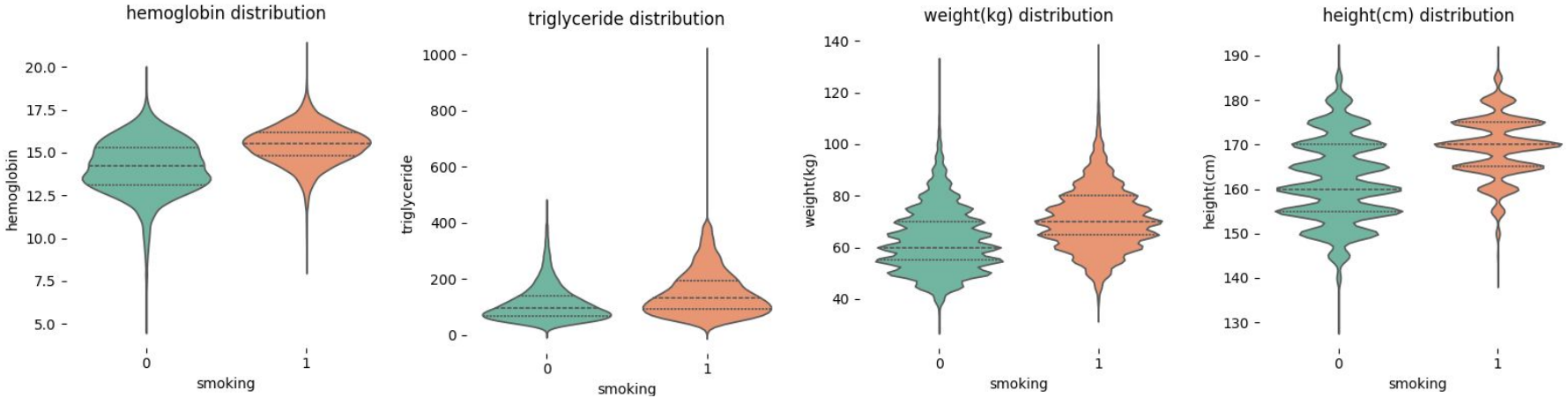
Pie Chart of Categorical Features

- Dental caries are more commonly known as tooth decay/cavities
- Less dental caries in non-smokers than those who smoke



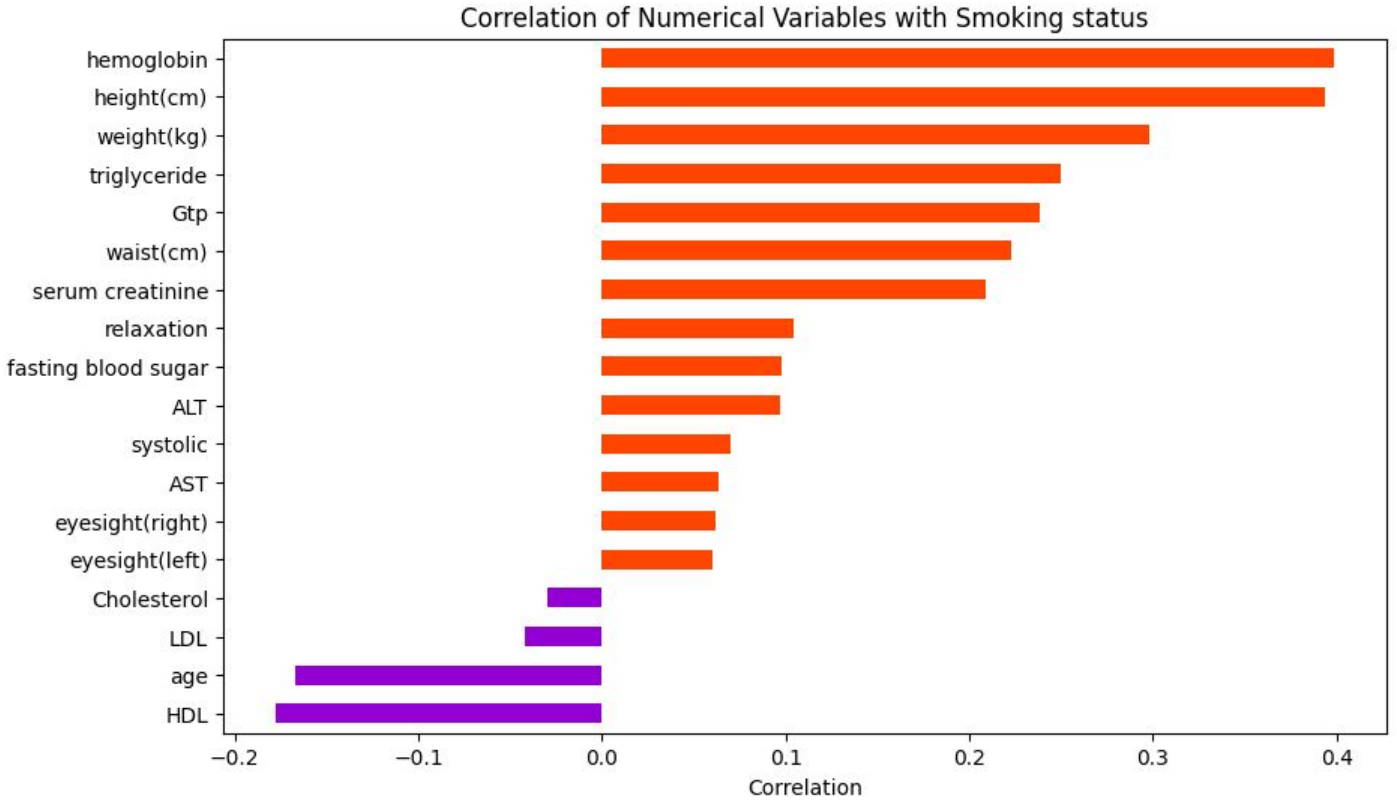


Violin Chart of Numerical Features

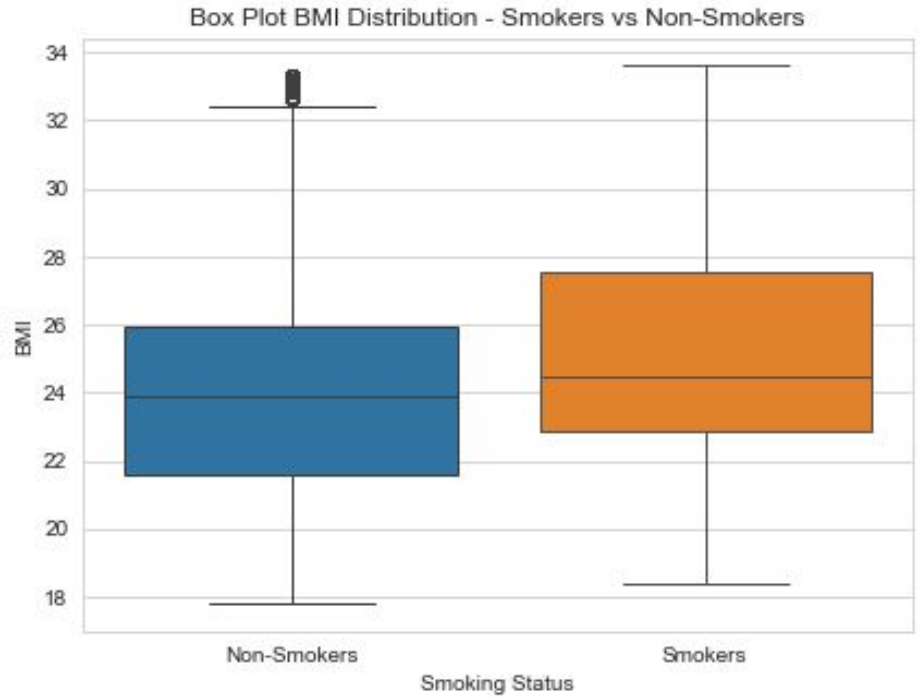


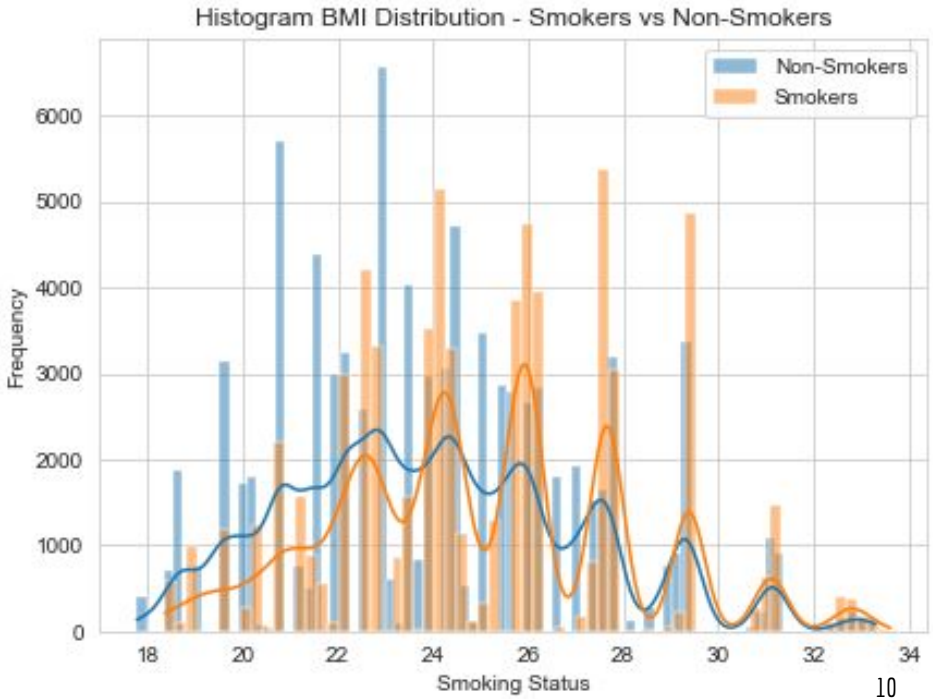
- Larger hemoglobin and triglyceride distribution in those who smoke
- Larger weight/height in those who smoke could suggest higher proportion of male smokers

Correlation of Numerical Variables with Smoking



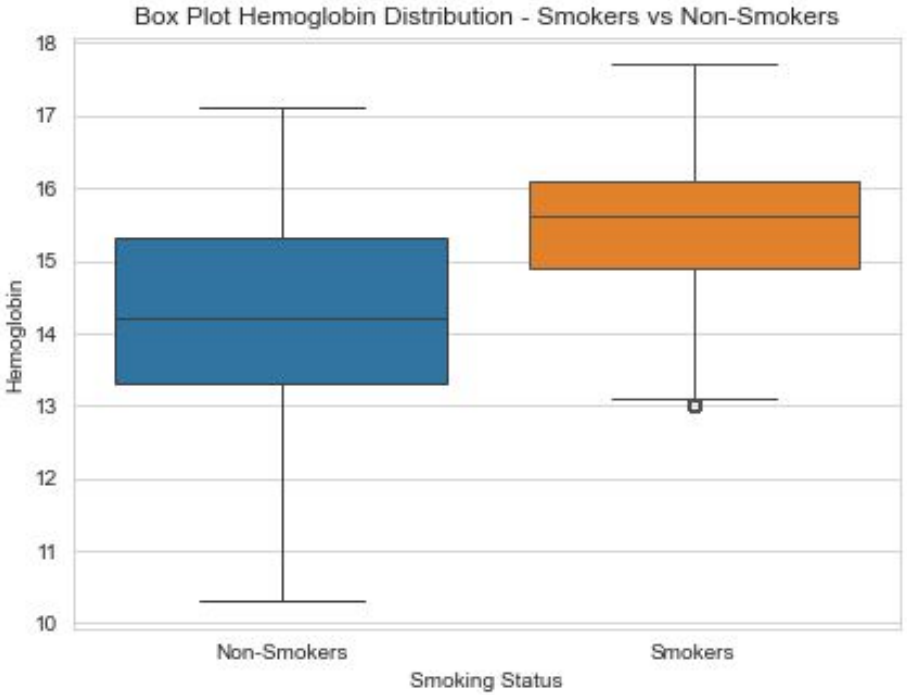
BMI Analysis

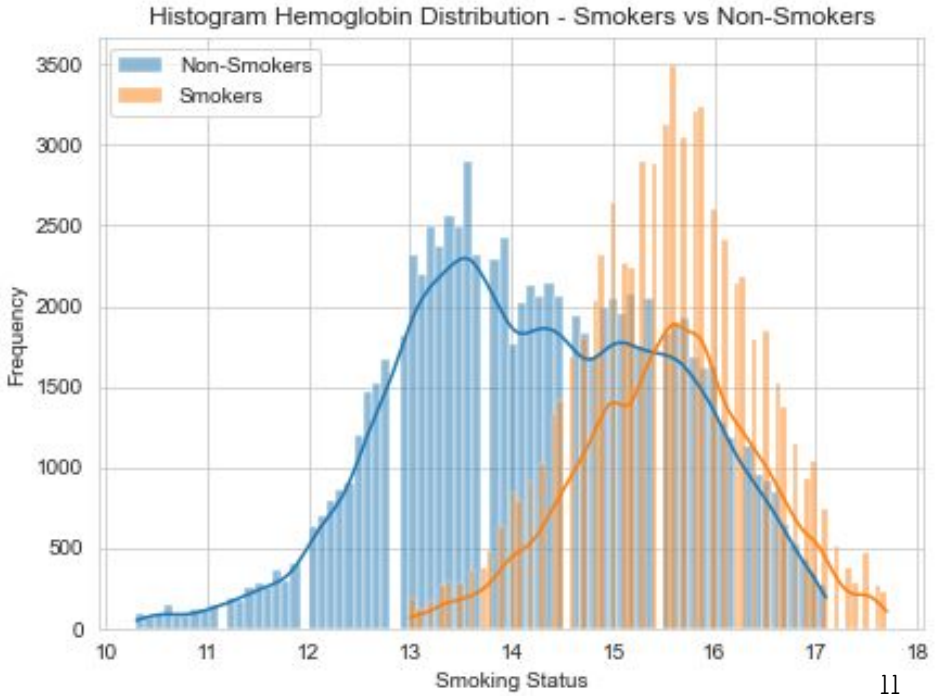




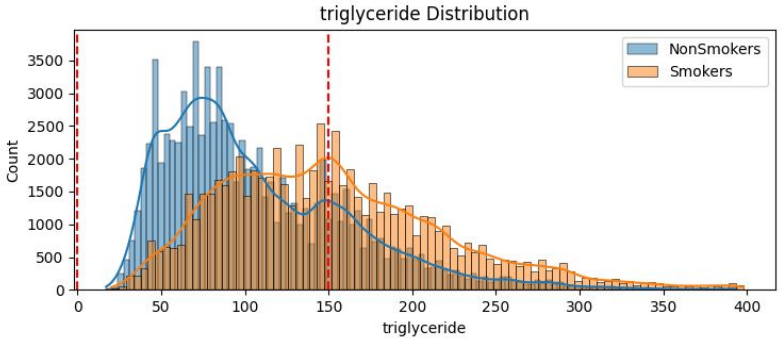
Blood Analysis

- Hemoglobin of smokers tends to spike [2]

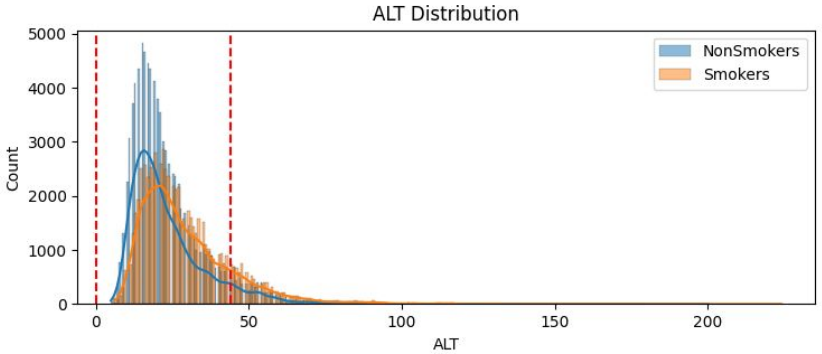




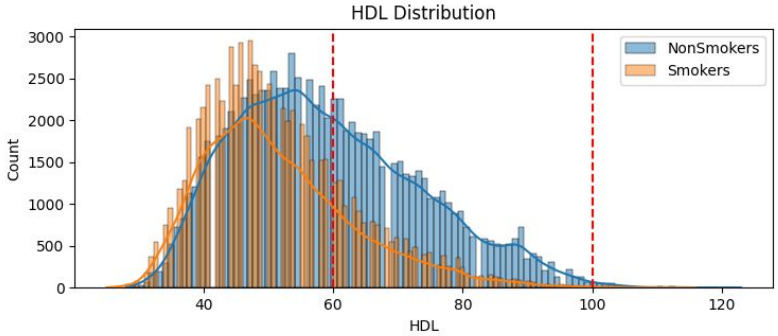
Health Features



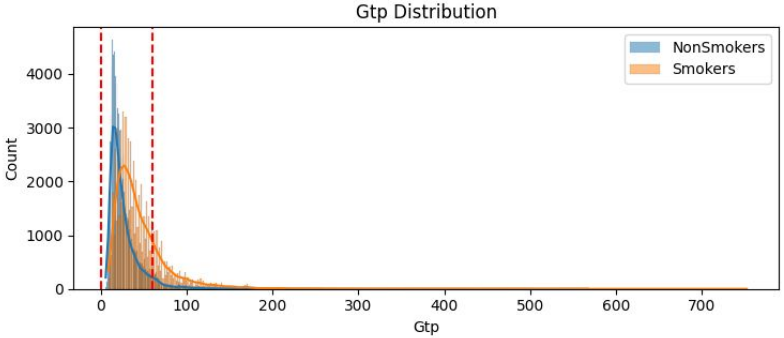
Much higher Triglyceride Levels



Much higher ALT Levels



Much lower HDL Levels



Much higher Guanosine-5'-triphosphate (GTP) Levels

Top-influential Bio-factors

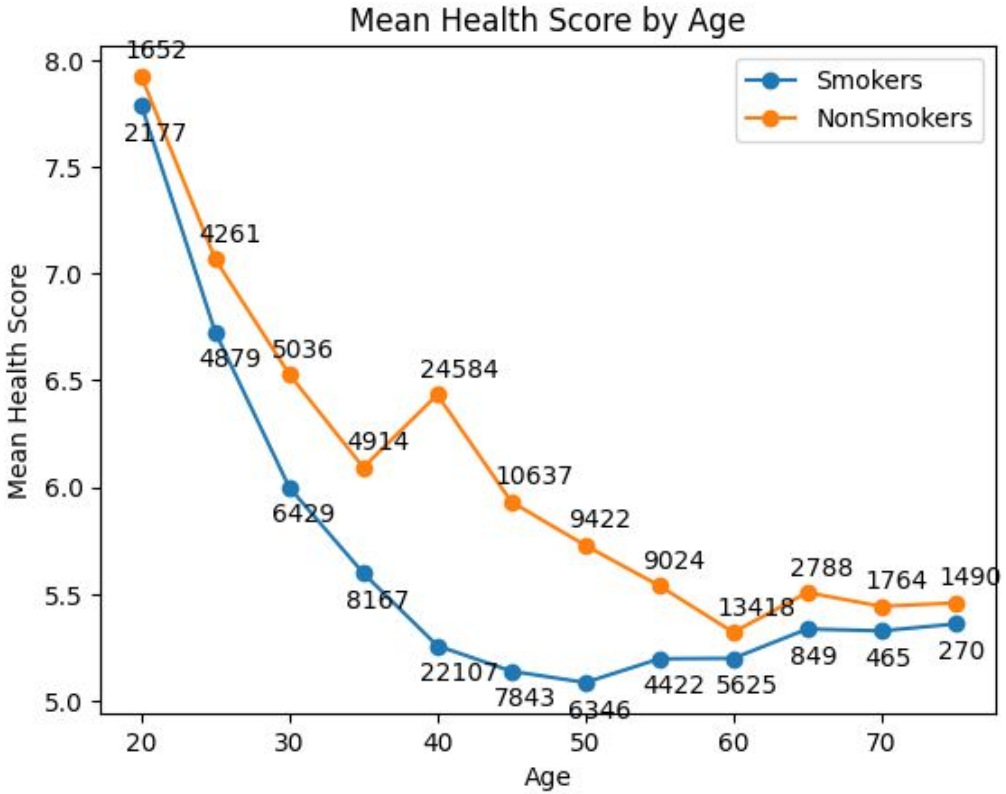
- The most distinctive and helpful bio-factors to aid physicians include:
  - Hemoglobin levels
  - Triglyceride levels
  - BMI
- Study shows that obese people are at higher risk of smoking. [4]
- However, there is an inverse relationship between BMI and smoking [3]
  - Smokers with higher BMI consume more cigarettes per day and might be more nicotine-dependent than lean smokers.

⇒ There is a bi-directional relationship between BMI and smoking.

Calculating a Health Score

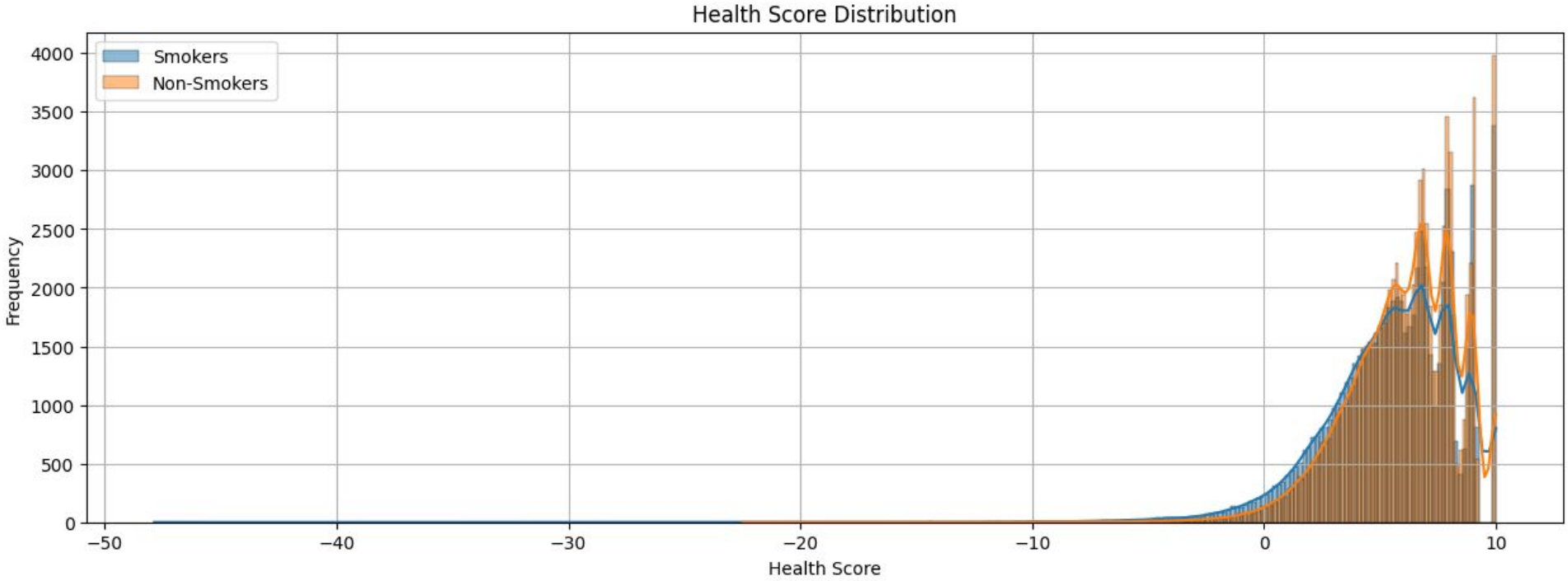
A health score is calculated using the following features and ranges:

- Hemoglobin: 13.5 - 17.5
- Serum creatinine: 0.5 - 1.2
- AST: 0 - 40,
- ALT: 0 - 44,
- Gtp: 0 - 60,
- BMI: 18.5 - 24.9
- Fasting blood sugar: 70 - 100
- LDL: 0 - 100
- Total cholesterol: 0 - 200
- Triglyceride: 0 - 150
- Systolic: 90 - 120
- Relaxation: 60 - 80



Number indicates count of samples used to calculate mean

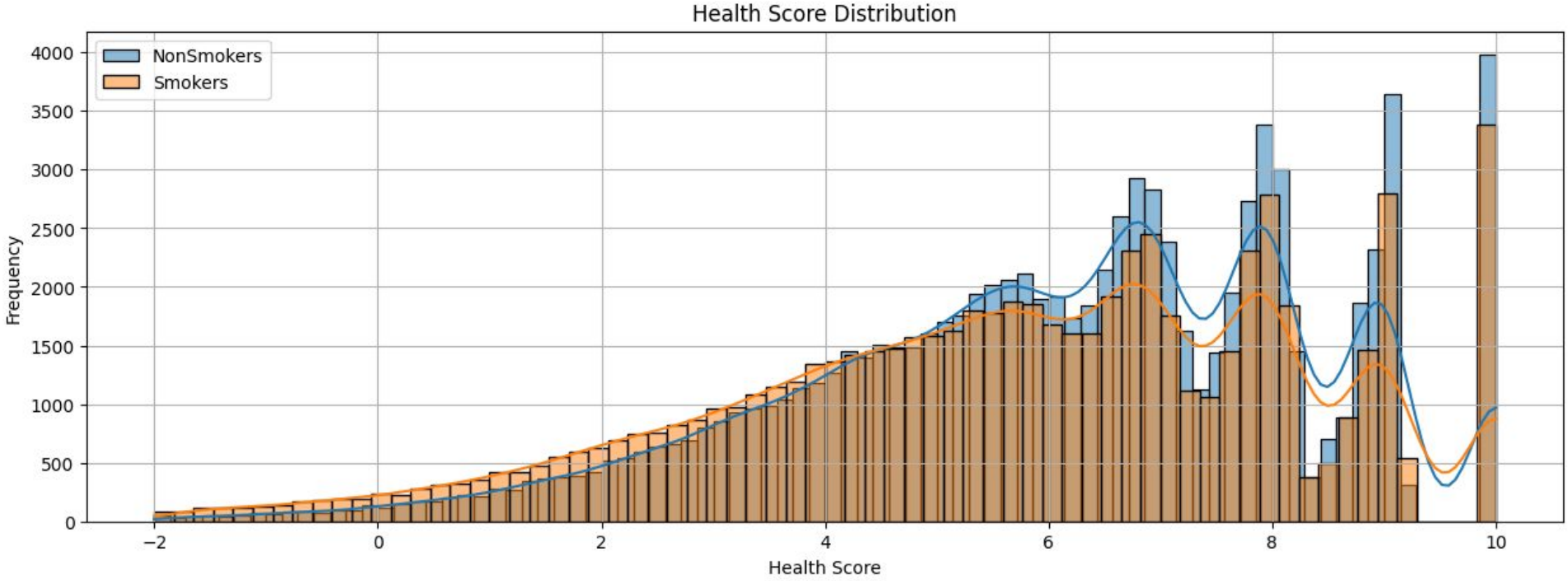
Health Score for Entire Dataset



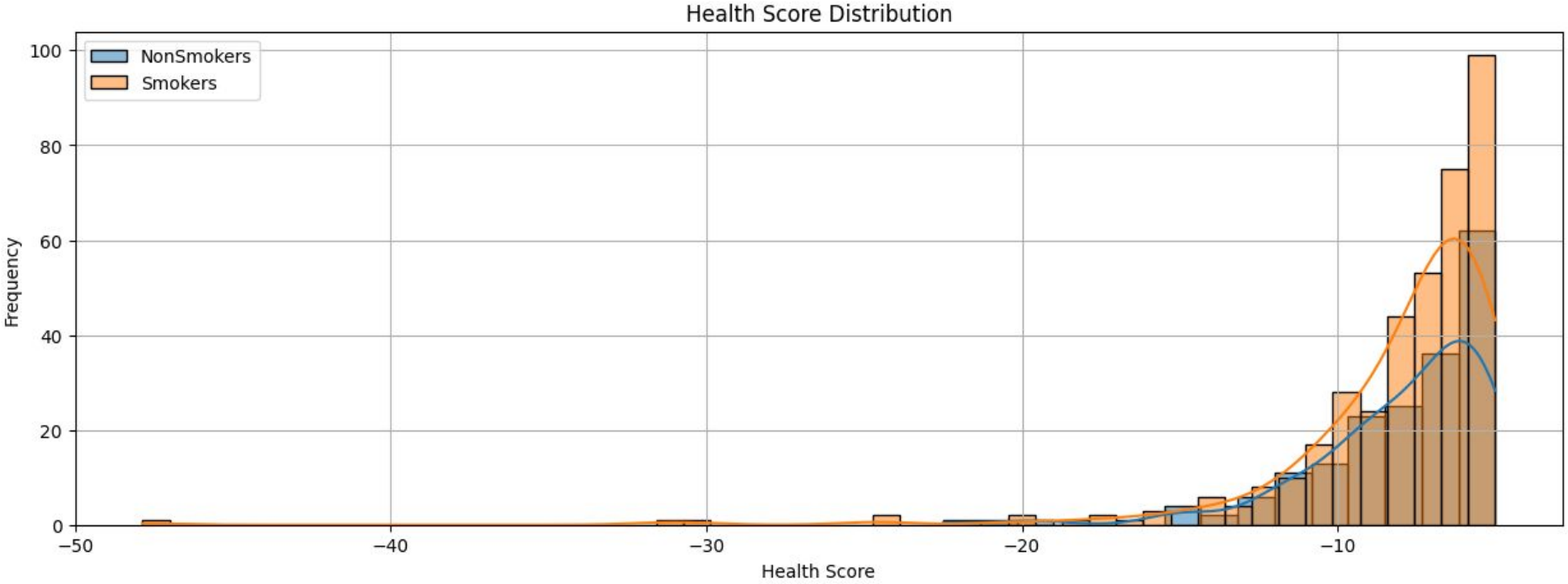
Smokers: Mean: 5.512, Min: -47.866

Non-Smokers: Mean: 6.015, Min: -22.487

Health Score for Entire Dataset



Health Score for Entire Dataset



Conclusion

- From the health score distribution plot, it is evident that there is a larger number of smokers in the lower range, and their average health score is lower compared to non-smokers. Additionally, many smokers have a health score less than -2, indicating significant harm to health due to smoking.
- Health score reduces as age increases, but the health score of smokers is consistently lower than that of non-smokers.

Reference

- [1] ML Olympiad Dataset (https://www.kaggle.com/competitions/ml-olympiad-smoking/data)
- [2] Tobacco smoking causes secondary polycythemia and a mild leukocytosis among heavy smokers in Taif City in Saudi Arabia. [Alkhedaide, A. Q. (2020). In Saudi Journal of Biological Sciences, 27(1), 407-411.]
- [3] Obese Smokers as a Potential Subpopulation of Risk in Tobacco Reduction Policy. [Rupprecht LE, Donny EC, Sved AF.]. In Yale J Biol Med. 2015 Sep 3;88(3):289-94]
- [4] New evidence on link between obesity and smoking behaviour from genetic data: obese people at higher risk of smoking. (https://www.bristol.ac.uk/news/2018/may/obesitysmoking.html)
- [5] Alanine Aminotransferase (https://emedicine.medscape.com/article/2087247-overview)
- [6] Triglycerides: Why do they matter? (https://www.mayoclinic.org/diseases-conditions/high-blood-cholesterol/in-depth/triglycerides/art-20048186)
- [7] Lowering GTP Level Increases Survival of Amino Acid Starvation but Slows Growth Rate for Bacillus subtilis Cells Lacking (p)ppGpp [Bittner AN, Kriel A, Wang JD 2014 Jun;196(11):2067-76. doi: 10.1128/JB.01471-14]

Thanks for listening!
Any questions?