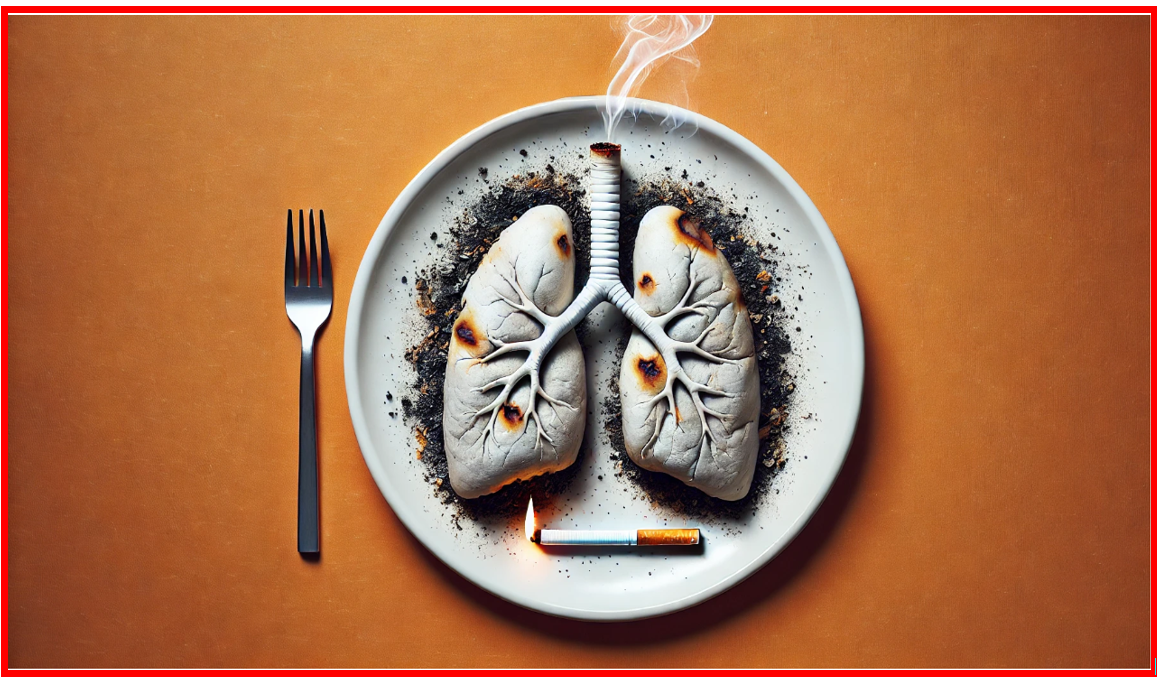
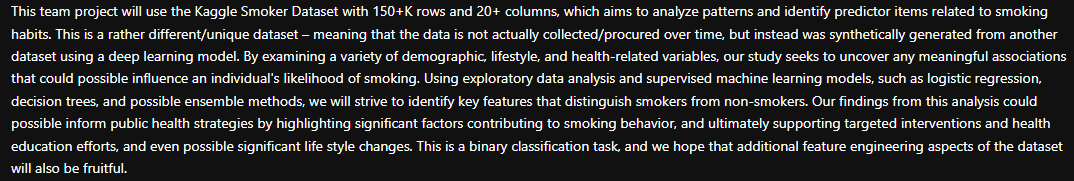
Team 6: Tyler Earps, Ryan Smith, Basil Mullings, Ean Vandergraaf

## Team6 Kaggle’s Smoker Status Prediction - BioSignals

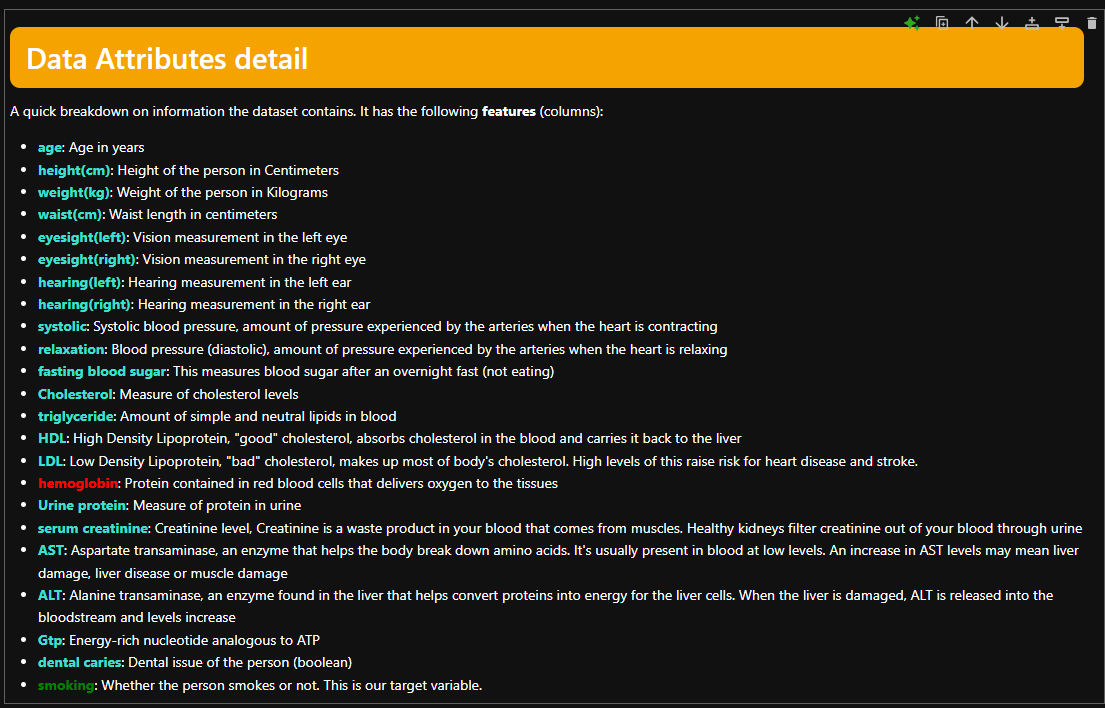


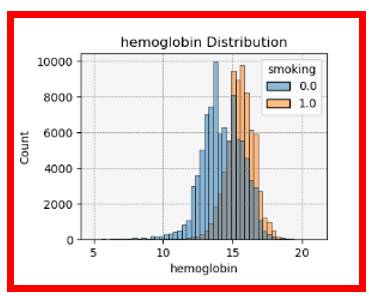
## Project Introduction:



1. **Dataset Overview**: The Kaggle Smoker Dataset contains 150K+ rows and 20+ columns, synthetically generated through a deep learning model.
2. **Objective**: Analyze patterns and identify predictors related to smoking habits using demographic, lifestyle, and health data.
3. **Analysis Approach**: We Employed exploratory data analysis (EDA) and supervised machine learning models, including logistic regression, decision trees, random forest and Gradient boosted trees
4. **Binary Classification**: Our goal is to classify individuals as smokers or non-smokers, identifying key features influencing smoking behavior.
5. **Potential Impact**: Findings could inform public health strategies, targeting interventions, health education, and lifestyle changes.

## Explain an example of the data (with numbers/visuals, etc):





Among the many features of the dataset – hemoglobin stood out as the one that could closely indicate if an individual is a smoker or not.

As showing from the graph above, once its level has surpassed a certain value ( > 15), then those folks that are smokers typically have elevated values. **However, this doesn’t necessarily mean that everyone with and elevated value is indeed a smoker**.

Hemoglobin is a protein found in our red blood cells, and plays a crucial role in transporting oxygen from the lungs to the rest of the body and also returning carbon dioxide from the tissues back to the lungs.

Hemoglobin levels can typically be higher in smokers due to our body's response to reduced oxygen levels resulting from smoking. Reasons:

### 1. **Carbon Monoxide Exposure**:

* **Carbon monoxide (CO)** is a toxic gas present in cigarette smoke. When inhaled, carbon monoxide binds to hemoglobin in the blood, forming. This reduces the ability of hemoglobin to carry oxygen.
* To compensate for the reduced oxygen-carrying capacity of hemoglobin, the body increases the **production of red blood cells**, leading to **higher hemoglobin levels**. This is a compensatory mechanism aimed at ensuring sufficient oxygen delivery to tissues despite the reduced oxygen availability.

### 2. **Hypoxia (Reduced Oxygen Supply)**:

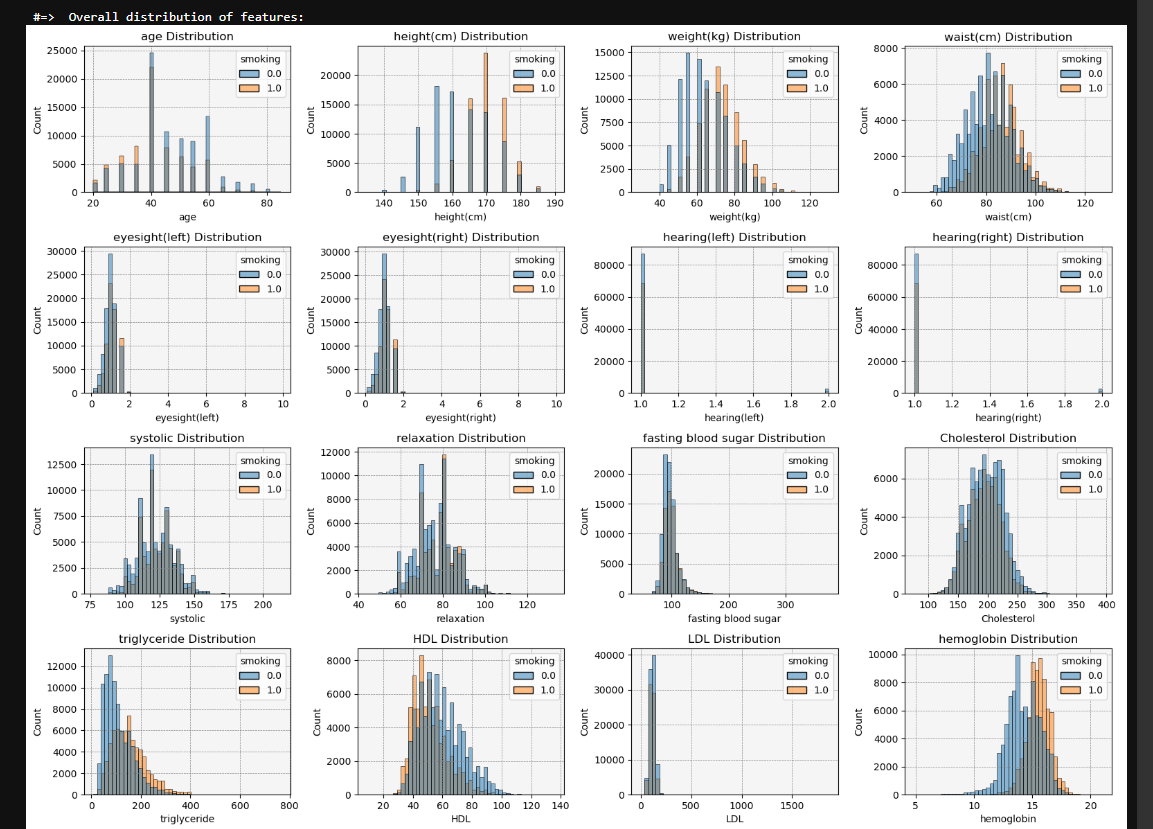
* Smoking causes **chronic low-grade hypoxia**, or a lower-than-normal oxygen level in the bloodstream, because of the reduced efficiency of oxygen transport due to carbon monoxide and other toxic chemicals.
* In response, the body stimulates the **bone marrow** to produce more red blood cells in an effort to improve oxygen delivery. This increase in red blood cell production raises the levels of hemoglobin, the protein responsible for carrying oxygen in the blood.

### 3. **Chronic Smoking and Adaptation**:

* Smokers may experience **long-term adaptations** to deal with the ongoing exposure to reduced oxygen levels. Over time, their body may produce more hemoglobin as a regular adaptive response, even though this can be harmful in the long run. Elevated hemoglobin levels are often observed in chronic smokers and those living in environments with poor oxygen levels.

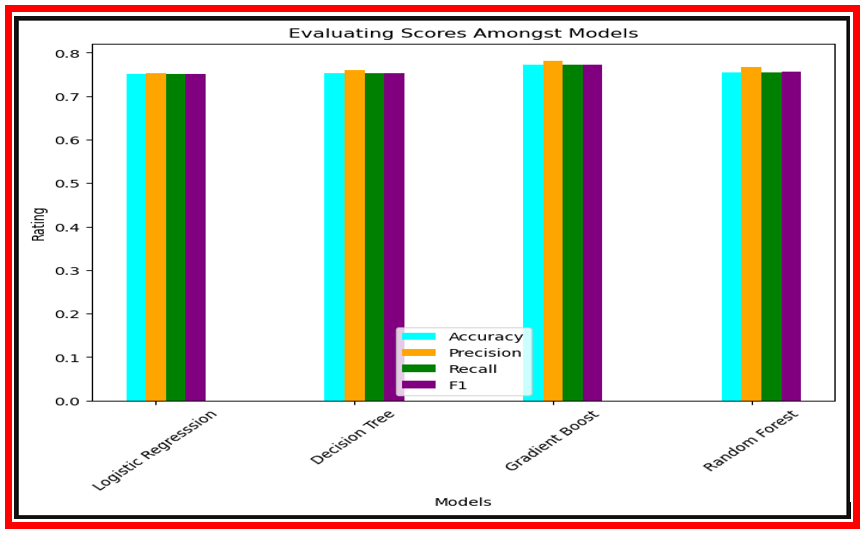
### 4. **Polycythemia**:

* In some chronic smokers, the increase in red blood cell production can lead to a condition called **polycythemia**, where the number of red blood cells becomes abnormally high. This condition is associated with higher hemoglobin levels, which can make the blood thicker and increase the risk of clotting and other cardiovascular issues.



## Best ML results:

1. We tried a few supervised models – including Logistic Regression, Decision Tree, Random Forest, Gradient-Boosted Trees
2. While none of the models performance was awful, they were all very close
3. The Gradient-Boosted Trees performed the best though with an accuracy score of approx. 78%. This was probably due to a few reasons:
   1. It is considered somewhat of an ensemble method that builds sequential trees that tries to correct errors from the previous tree, thus leading to overall better performance and accuracy
   2. It seems to be able to capture non-linear relationships in the dataset
   3. Adaptability to the tuning parameters like the number of trees and depth



## Conclusion/Takeaway:

1. Overall, a very interesting project that show cases the potential of machine learning
2. Using ML for predicting smoker status probably already has real life implementations and is actively being used
   1. For example, this could be used by health insurance providers in some capacity to assess the insured’s smoker status
3. We should be cautious that even if the model predicts a subject as a smoker, this may not necessarily be the case, hence perhaps other measures should be in place as a “sanity” check – to do continuous validation of the ML results.
4. For example, there are other scientific & biologic tests that can be used to determine smoking status:
   1. Carbon Monoxide measurement
   2. Exhaled Nitric Oxide presence
   3. Saliva and hair samples
   4. Chest x-rays
   5. Lung pulmonary tests for lung capacity and air flow
   6. Liver tests

## 