12.9.23

MTA School of Computer Science
Project Number 231006

ADVISORS

Prof. Adi Shraibman
Dr. Dorit Shweiki
Dr. Yonatan Bilu

STUDENSTS

Dan Naftaly
Tamir Matok

Medical Data Analysis

lung Cancer Research

Table of Contents

Research Question

While smoking is the leading cause of lung cancer,
many non-smokers also develop the disease.
Other possible causes include exposure to air pollution, radon, asbestos, and genetic factors.

We want to rank and investigate the different risk factors and understand how they contribute to the development of lung cancer.

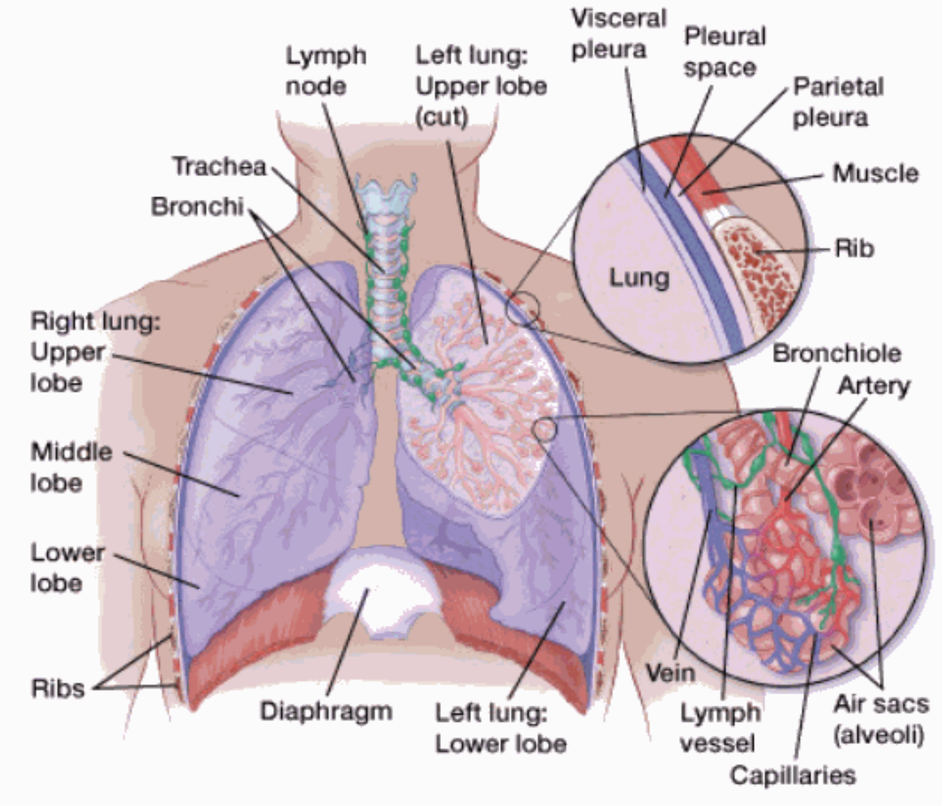
Motivation

- Personal connection as smokers adds a layer of personal significance to this research, driving us to unravel the complexities surrounding lung cancer risks.

- Need to understand the diverse factors contributing to the development of lung cancer.

- Uncovering insights into potential contributors such as air pollution, radon, asbestos, and genetic predisposition, with the ultimate goal of enhancing prevention and treatment strategies for this devastating disease.

Selected types of lung cancer



C34	**Malignant neoplasm of bronchus and lung**
C34.0	**Main bronchus** Carina Hilus (of lung)
C34.1	**Upper lobe, bronchus or lung**
C34.2	**Middle lobe, bronchus or lung**
C34.3	**Lower lobe, bronchus or lung**
C34.8	**Overlapping lesion of bronchus and lung** [See note 5 at the beginning of this chapter]
C34.9	**Bronchus or lung, unspecified**

Data Fields

Behavioral

- Smoking status
- Number of cigarettes previously smoked daily
- Chest pain during physical activity
- Doctor diagnosed asbestosis
- Doctor diagnosed COPD
- Cough on most days

Physiological

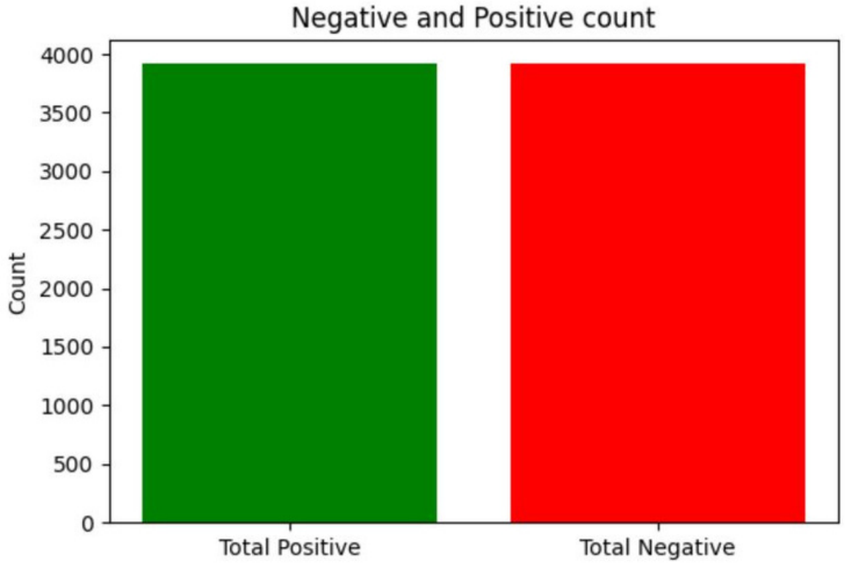
- **Sex**
- **Year of birth**
- BMI
- Albumin
- Platelet count
- Pulse rate
- C-recative protein
- Lymphocyte count
- ALP
- White blood cell count
- Forced Vital Capacity
- FEV1
- Maximum heart rate during fitness test

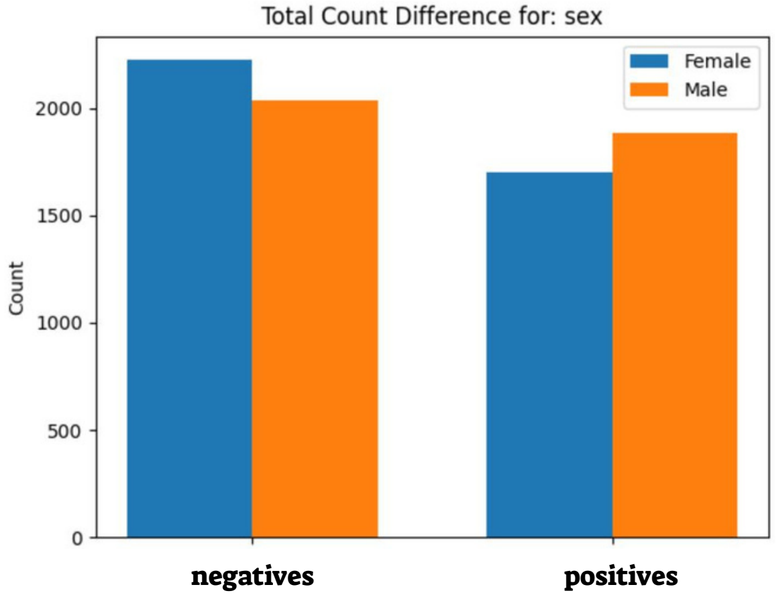
Demographic

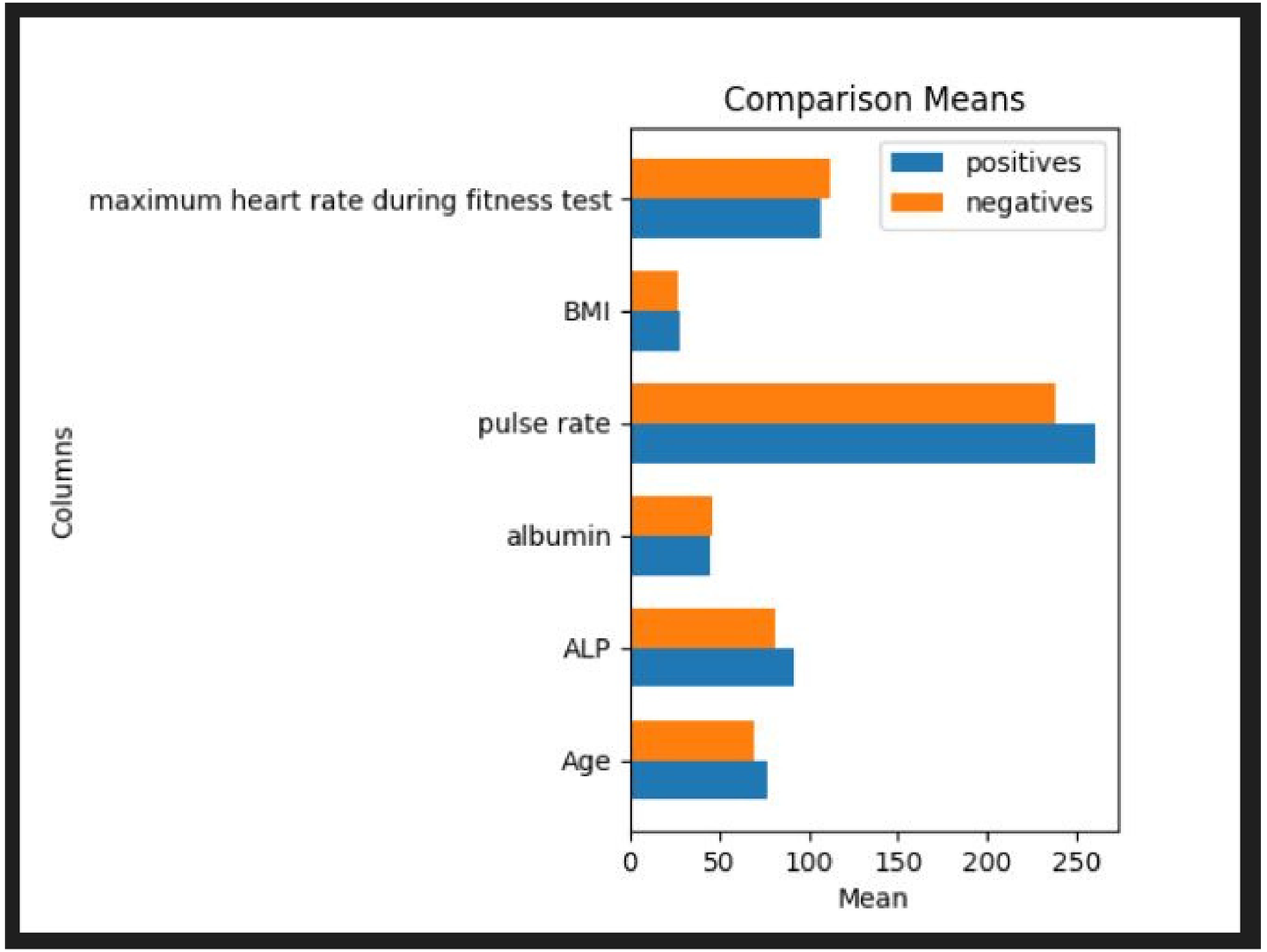
- Ethnic background

Enviormental

- Close to major road
- Nitrogen dioxide air pollution
- Particulate matter air pollution 2.5-10um


Negative and Positive count
Count
4000
3500
3000
2500
2000
1500
1000
500
0
Total Positive
Total Negative

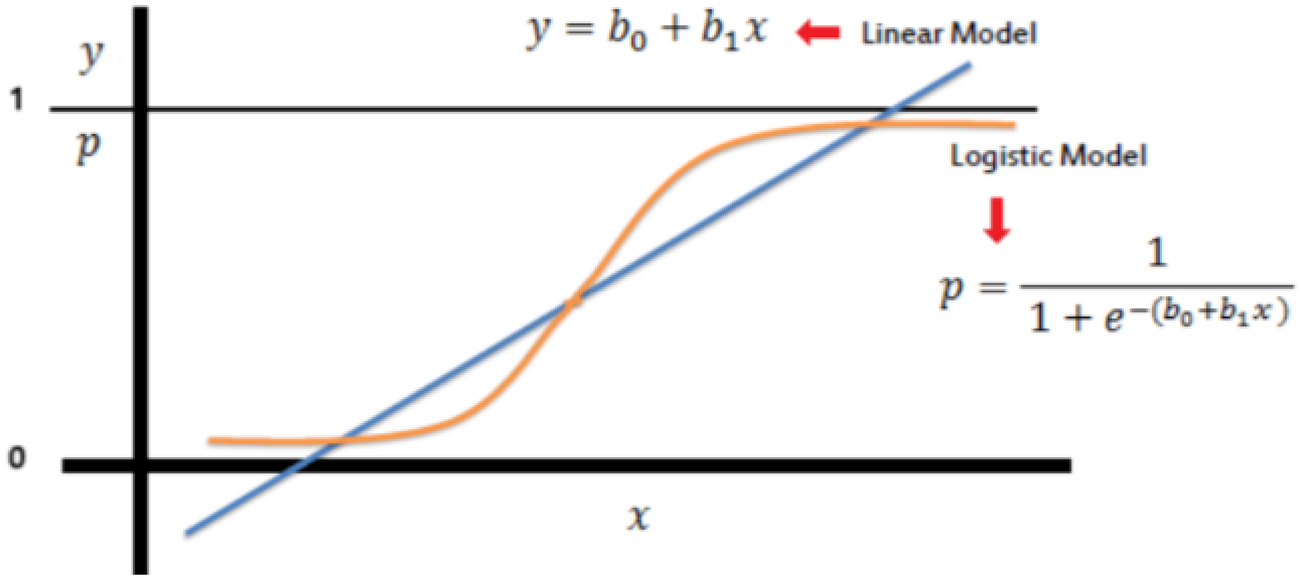

Total Count Difference for: sex
Female
Male
Count
2000
1500
1000
500
0
negatives
positives


Comparison Means
positives
negatives
maximum heart rate during fitness test
BMI
pulse rate
albumin
ALP
Age
Columns
0
50
100
150
200
250
Mean

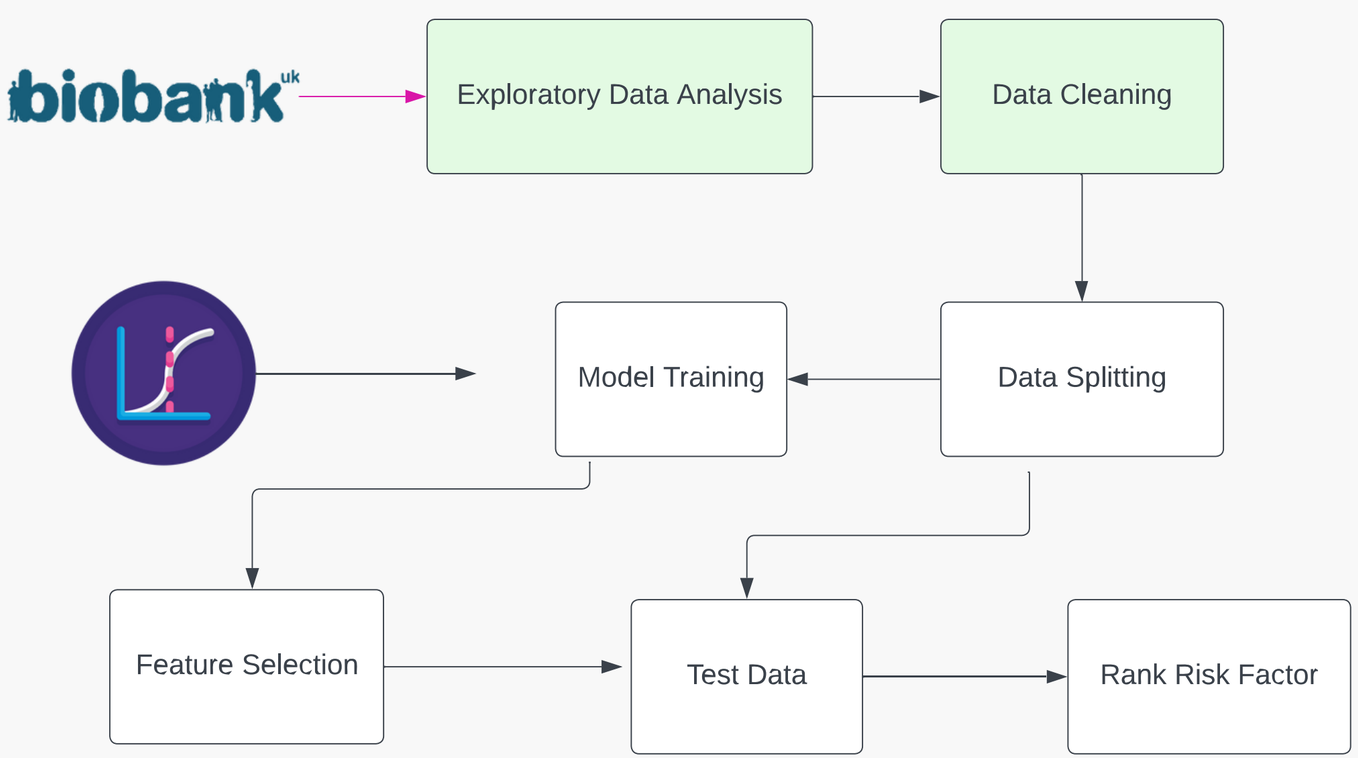
Model selection

Logistic Regression

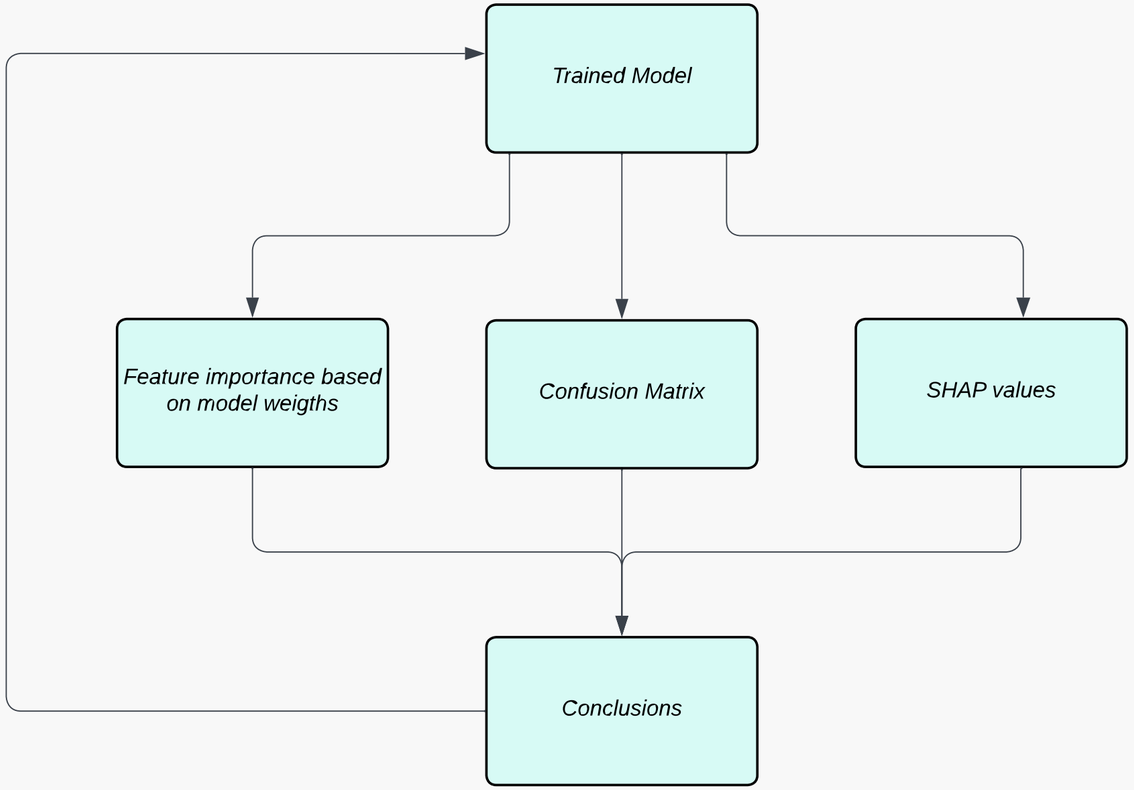
Logistic regression is a supervised learning algorithm that makes use of logistic functions to predict the probability of a binary outcome.

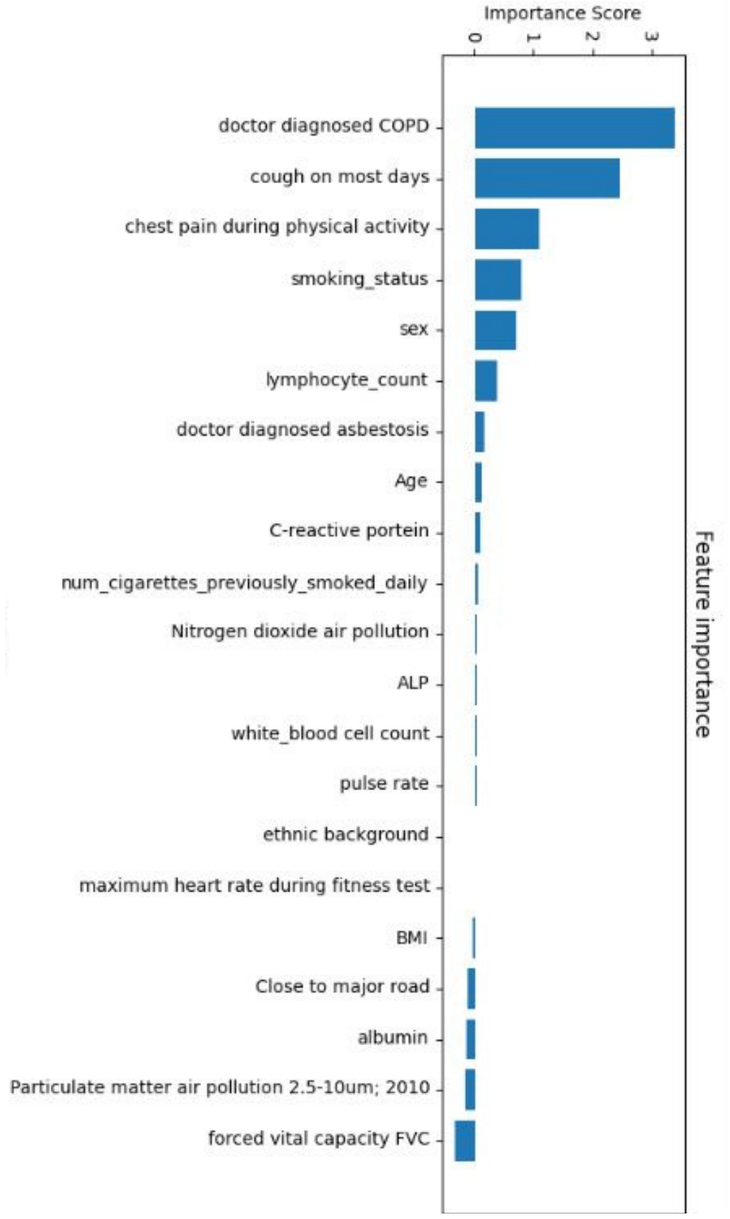


High Level Architecture and Design

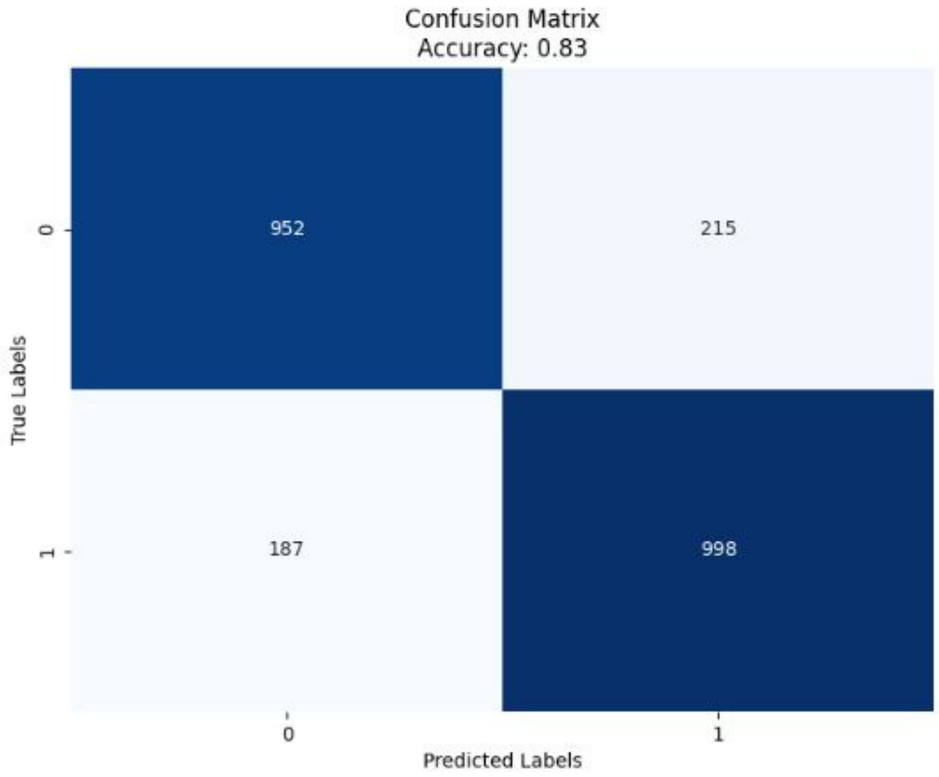


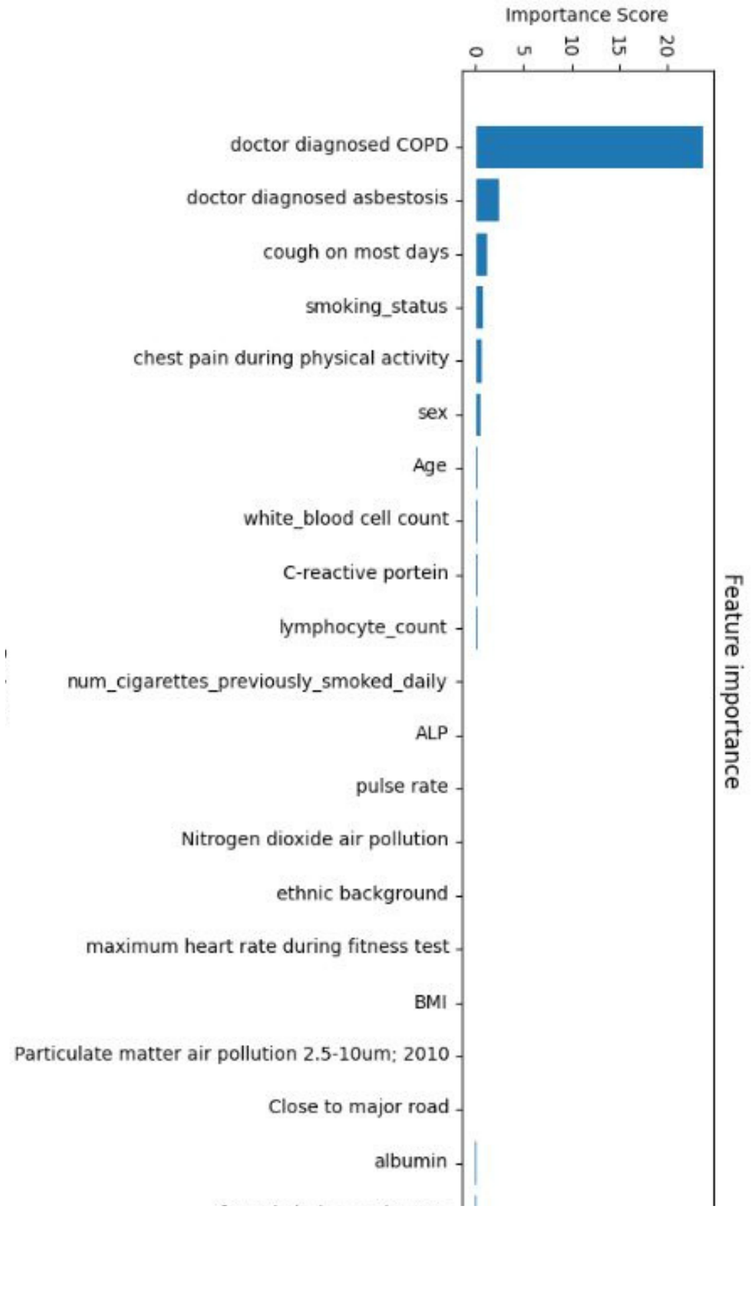
Model evaluation process



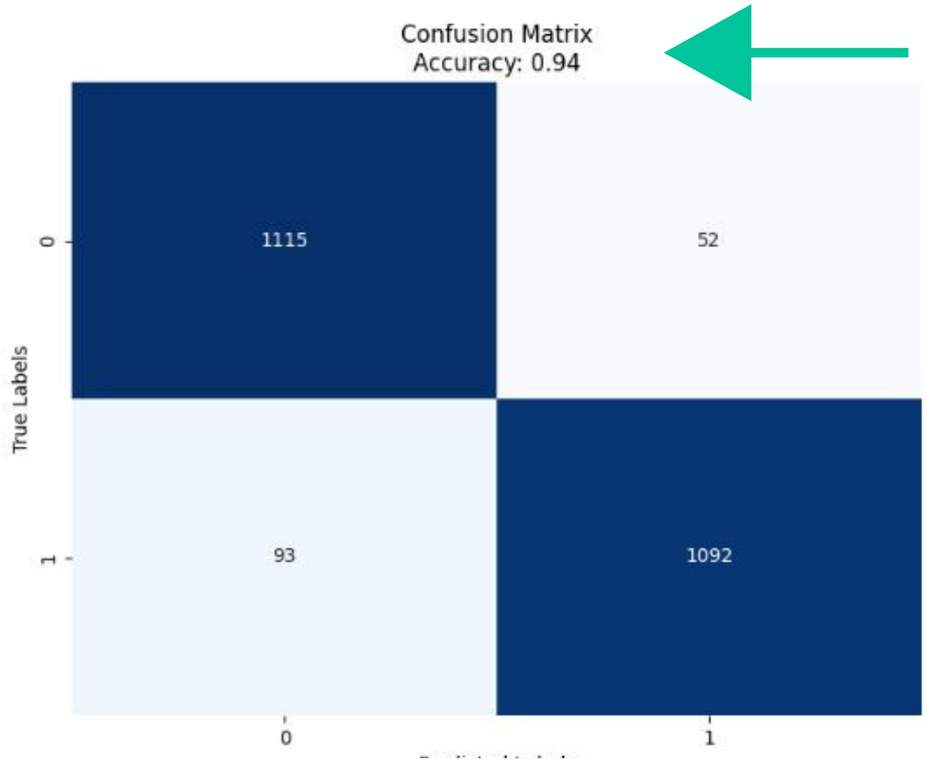


- **At first glance - model results are satisfying based on the confusion matrix.**
- **Several columns seem to have a significant impact on the model results, which raises concerns about the authenticity of our results.**

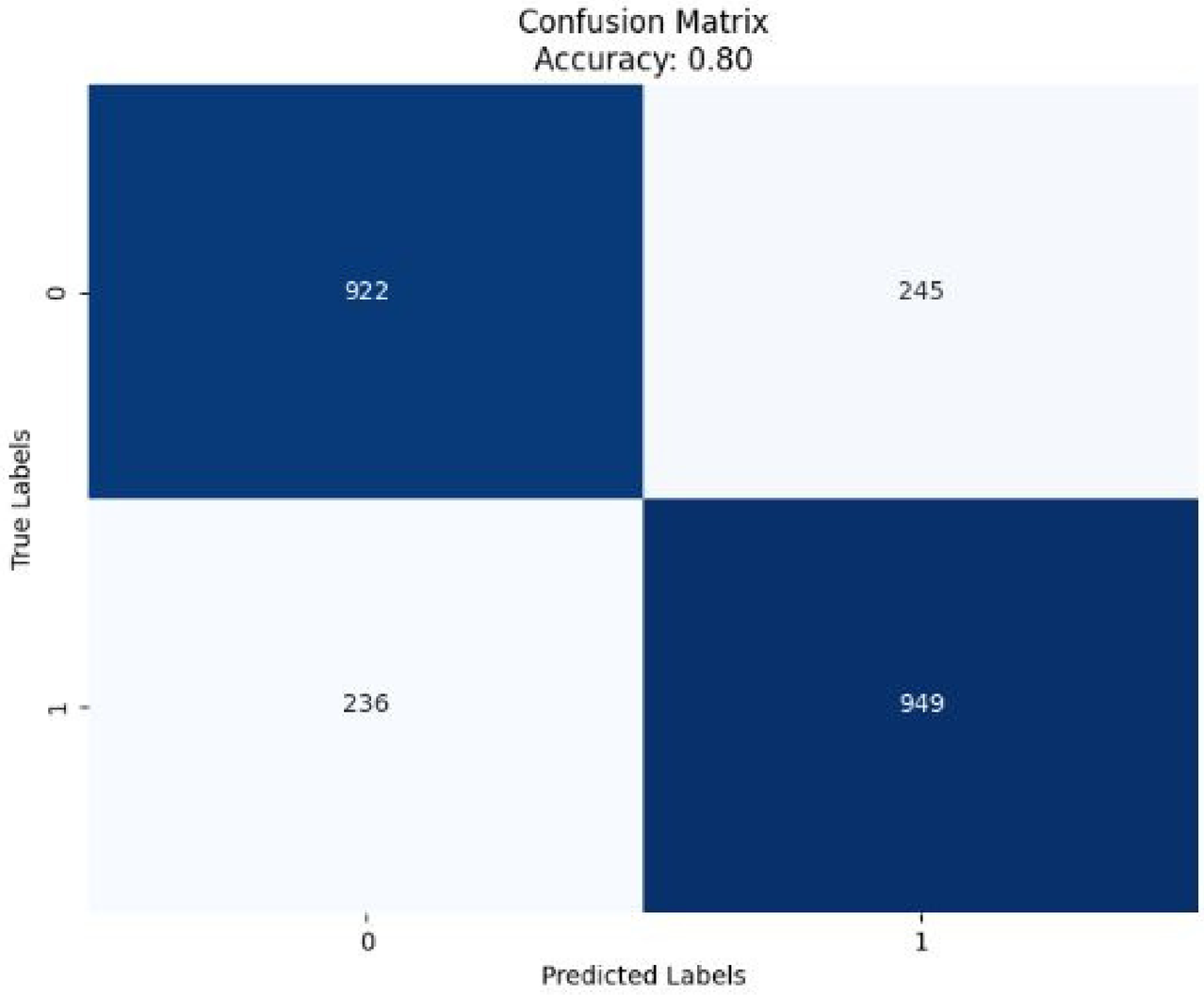


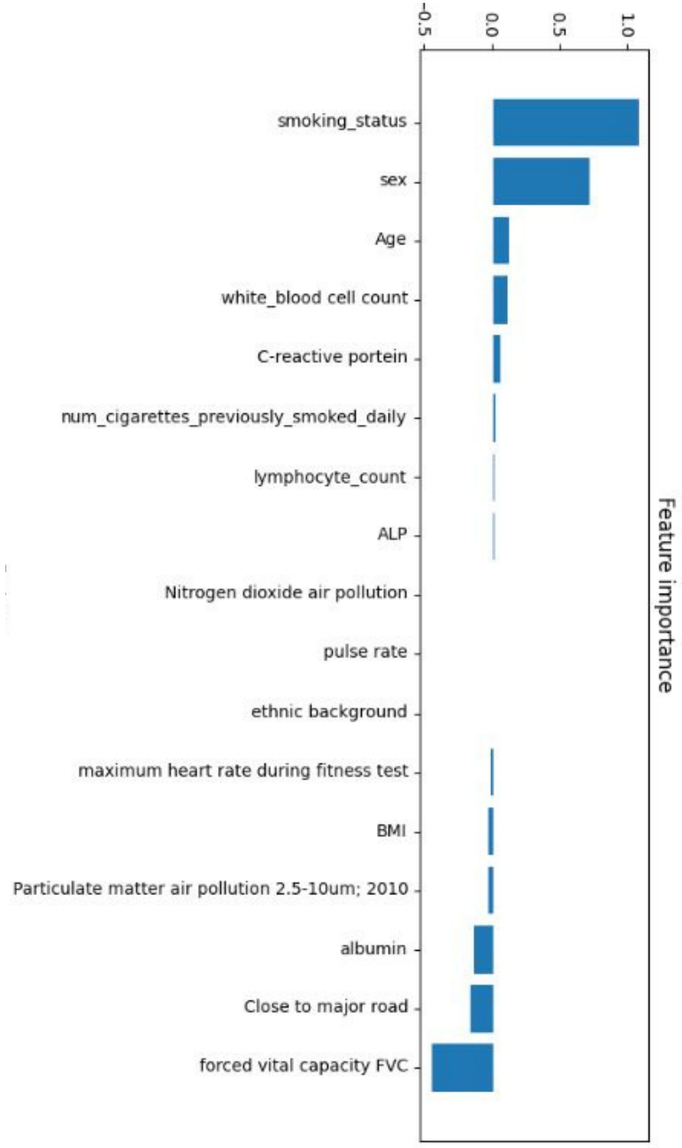


- **By incorporating L1 regularization, we can identify and select the most important features related to lung cancer risk.**
- **Among smokers, COPD is an important risk factor for developing lung cancer, and predates lung cancer in up to 70–80% of cases (National Institutes of Health)**

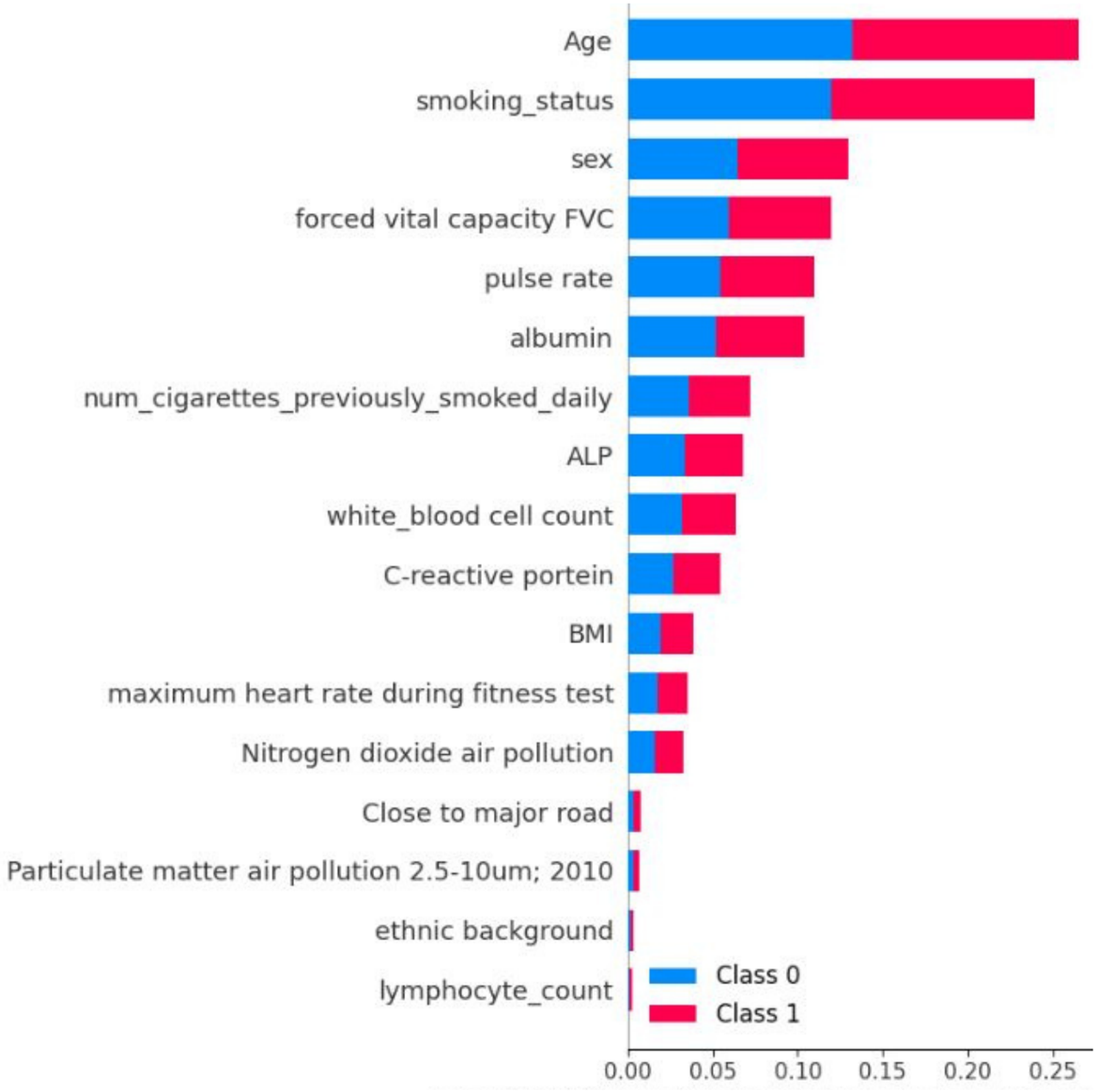


Final Model Results





SHAP values


Age
smoking_status
sex
forced vital capacity FVC
pulse rate
albumin
num_cigarettes_previously_smoked_daily
ALP
white_blood cell count
C-reactive portein
BMI
maximum heart rate during fitness test
Nitrogen dioxide air pollution
Close to major road
Particulate matter air pollution 2.5-10um; 2010
ethnic background
lymphocyte_count
Class 0
Class 1
0.00
0.05
0.10
0.15
0.20
0.25

Summary

- The model has provided valuable insights into the factors influencing this deadly disease.
- With age and smoking status emerging as the foremost contributors to lung cancer risk, the findings reaffirm the importance of addressing these well-known risk factors.
- Exploration of additional variables, such as ethnic background and air pollution, revealed their comparatively lower significance in the context of lung cancer.

This research not only enhances our understanding of the disease but also underscores the need for continued efforts in smoking cessation and age-appropriate screenings to combat lung cancer effectively.