Imperial College London

# An exposome approach: Predicting the incidence of Asthma Attacks using multiple environmental exposures and lifestyle factors.

Malo Dirou

Project Presentation, May 4 2021

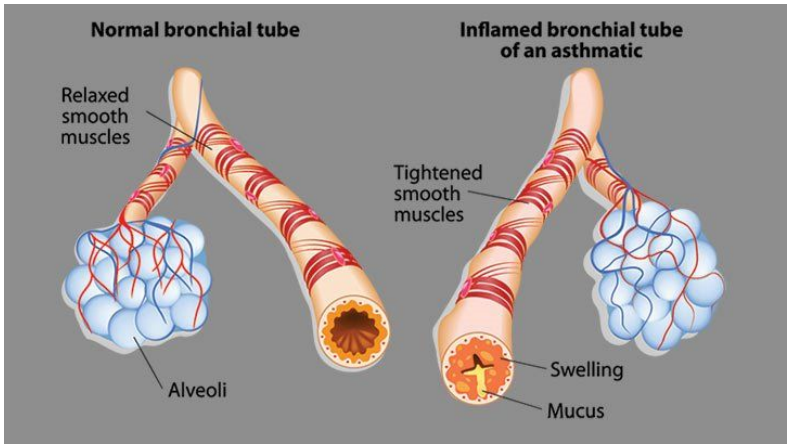
Imperial College London

# Table of Contents

Imperial College London

# Table of Contents





# Information about the disease

According to WHO,

- Asthma is a major non-communicable disease.
- Asthma is a long-term multi-factorial disease. Asthma Attacks are the results of a combination of genetic, physiological, environmental and behavioural factors.
- Asthma has a complex etiology: variation in frequency and severity within and between individuals.
- Asthma is a common chronic disease in children
- Most asthma related-deaths occurs in adults, essentially in low and lower-middle income countries

# A Few Numbers

- Estimated that 339 Million people suffer from asthma worldwide.
- 417,918 deaths due to asthma at the global level in 2016.
- 24.8 Million DALYS attributable to asthma in 2016.
- Recall 1 DALY = 1 Lost year of healthy life (1B + 1C)

# Aim of the research & Dataset

**Today's aim:**

- Complementing genome studies with an exposome study.
- Specifically, we will try to identify a large range of environmental factors and behaviors that are suspected to be triggers of asthma attacks.

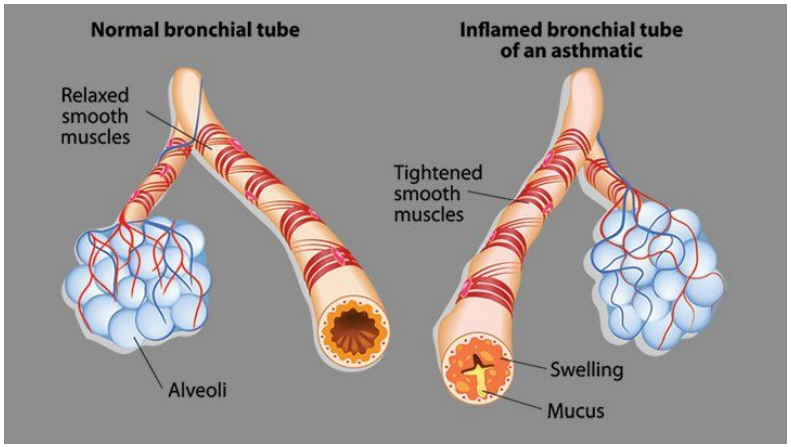
**Dataset: NHANES**

- The National Health and Nutrition Examination Survey is a program of studies designed to assess the health and nutritional status of adults and children in the United States.
- How is NHANES unique? It combines interviews and physical examination.

Imperial College London

# Table of Contents





# Data processing Steps (1)

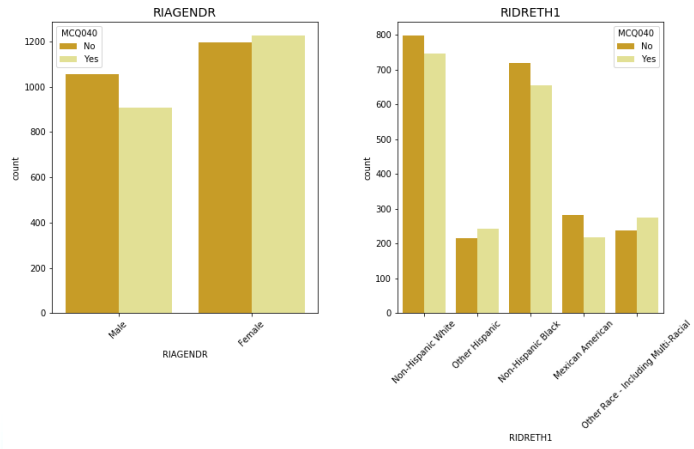
- Combination of 24 datasets selected from NHANES publicly available data.
- The 24 datasets contain questionnaire information, demographics, examination data and laboratory data.
- Each of the 24 datasets contain 2 years of information for participants.
- I appended 10 years of information (2009-2018) to end up with 2134 unique cases.
- Cases were defined as individuals who had asthma attacks in the past year.

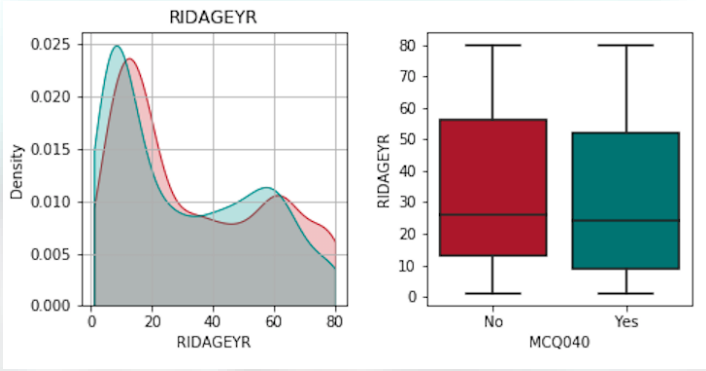
# Data processing Steps (2)

**Selection of controls:**

- Matching using matchit function in R to create case and control groups balanced on age, sex and ethnicity.
- Used a ratio of 4 controls per case.
- Removal of variables with NAs $>$ 50% of all observations
- Removal of variables with same meaning
- Ends up with a dataframe of 10,670 observations and 105 variables

# Data Exploration




Gender, ethnicity and age differences before matching among individuals who suffered from asthma attacks in the past year and the healthy individuals

## Data Exploration

- Gender, Age and Ethnicity differences after matching
- Age distribution is identical among cases and controls
- Balanced ratio among cases and controls for gender and ethnicity

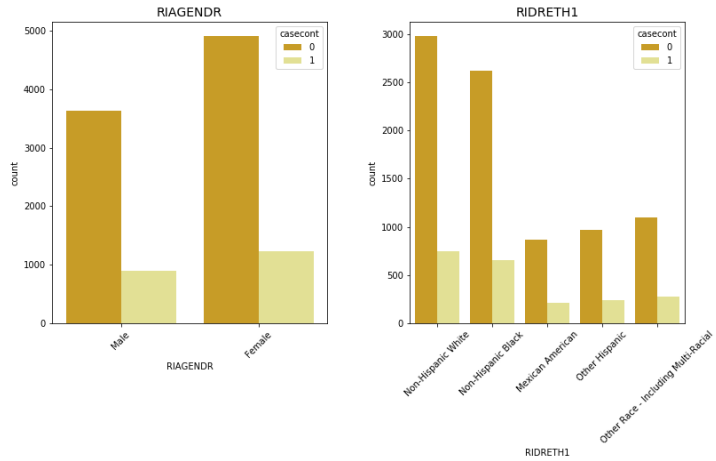



Figure:

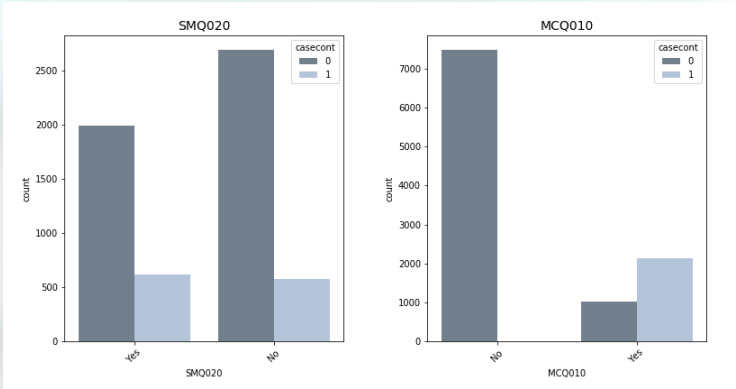



Figure:

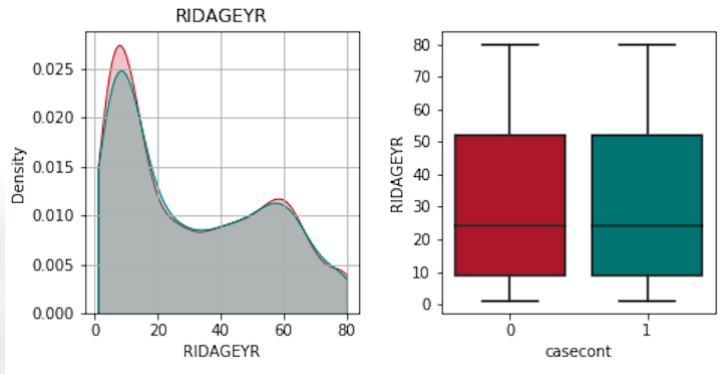



Figure:

# Algorithms choices

**Unsupervised Learning Technique: KMeans**

- Mean imputation of missing values for numerical variables
- Standardization of numerical values
- One-hot encoding of categorical variables
- Dimensionality transformation of the entire dataset using PCA (2PCS used)

**Supervised Learning Techniques: SVM**

- Same handling of missing values as KMeans - No PCA

**Distributed Random Forest and Gradient Boost**

- Only differences with KMeans and SVM data: No PCA. Use of most frequent imputation for categorical variables instead of one-hot encoding.

# Models Parametrization (1)

**KMEANS**

- Use of Within Cluster Sum of Squares (WCSS).
- Elbow method associated with gradual decrease in WCSS to find optimal number of clusters k.

**SVM**

- AUC GridSearch on linear kernel with C = [1,10,30,50].
- On RBF kernel with C = [1,10,30,50], gamma = [0.1, 0.2, 0.3, 0.4, 0.5].
- On polynomial kernel with C = [1,10,30,50], degree = [2,3,4], gamma = [0.01,0.02,0.03,0.04,0.05,0.06,0.07,0.08,0.09].

# Models Parametrization (2)

**GB**

- AUC GridSearch with learning rate = [0.01, 0.1], maximum depth per tree = [3, 5, 9], adding some variations on row and columns sample rates.

**RF**

- AUC GridSearch with max depth per tree = [3,4,5,6,7,8,9] and minimum number of observations per leaf node = [5,7,10,15,20], adding some variations on row sample rate too.

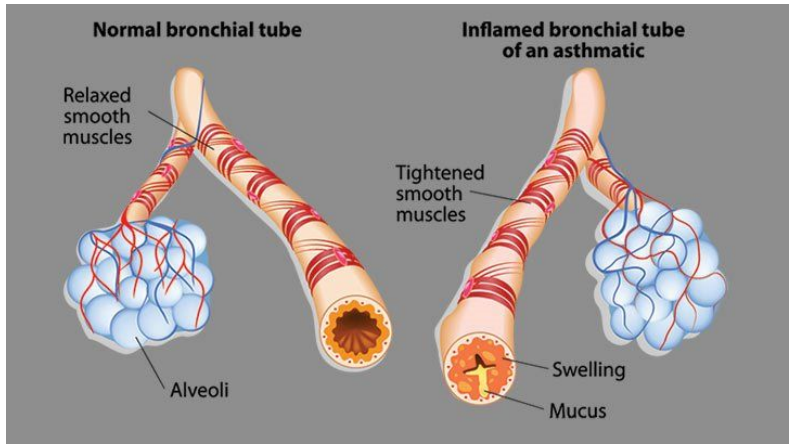
**Common uses among SVM, RF, GB**

- number of trees = 10000, stopping rounds = 5, stopping tolerance = 1e-4, stopping metric = "AUC", score tree interval = 10 for trees.
- Use of 5-fold cross validation for all models

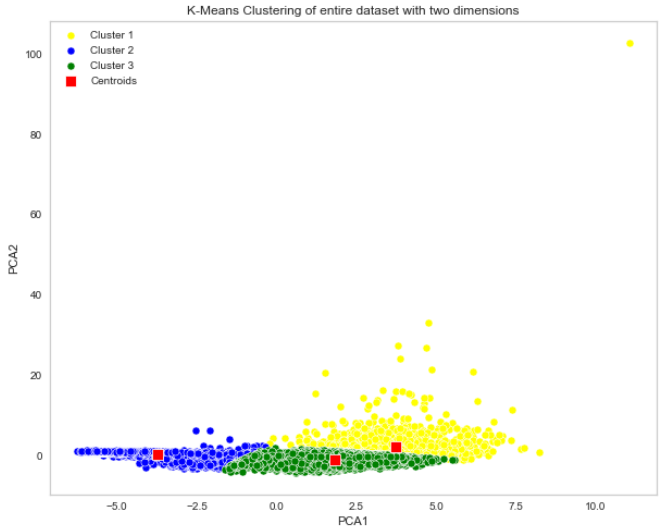
Imperial College London

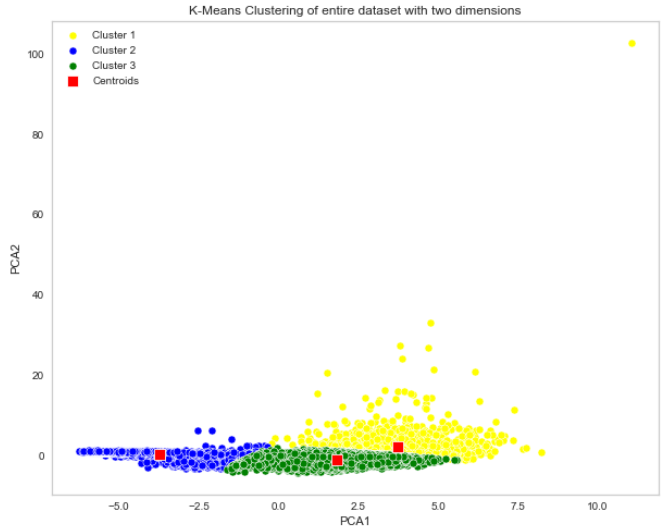
# Table of Contents





# KMEANS Visualization and Goodness of fit

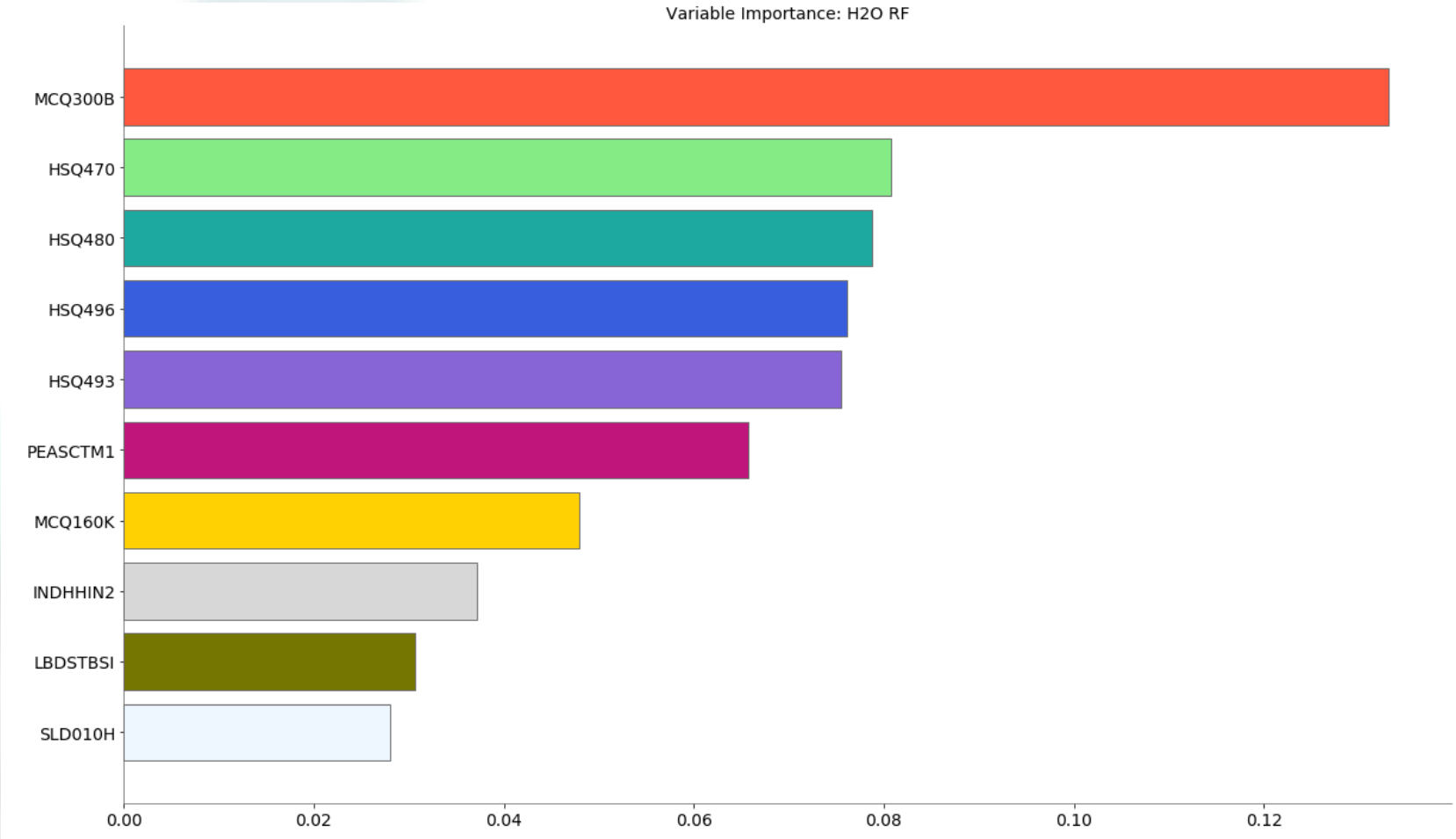



- Clusters are quite compact

Imperial College London

# Random Forest Variable Importance Plot

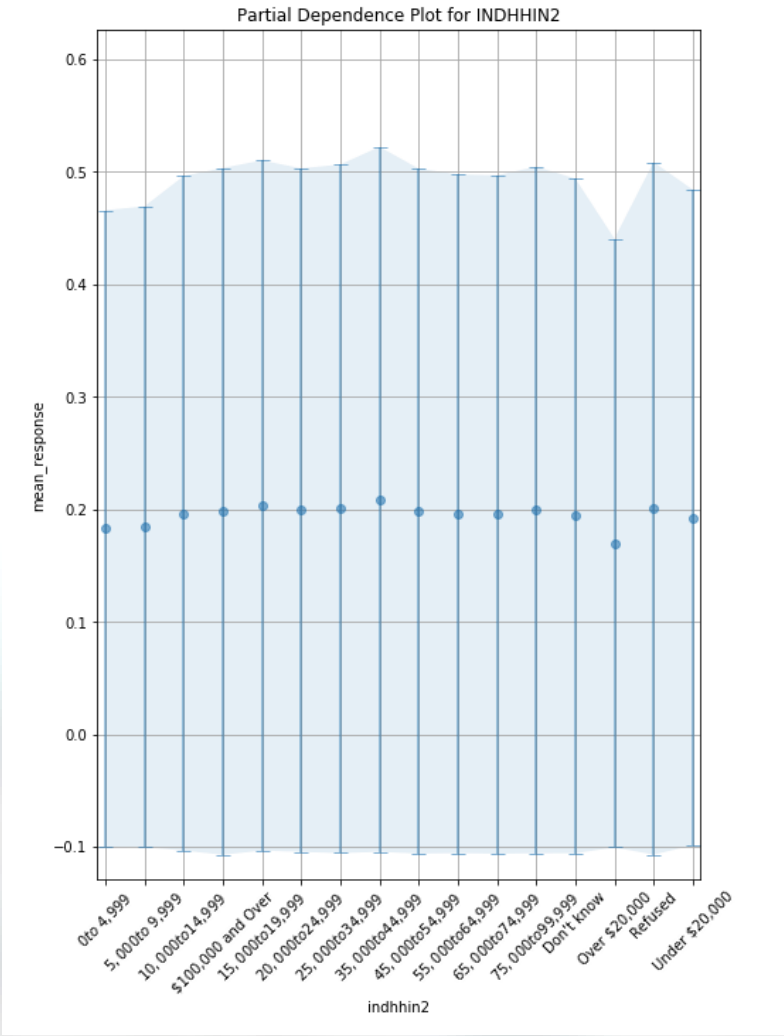


# Variable Importance Plot Meaning

By order of importance:

- Close relative had asthma,
- no. of days physical health was not good
- no. of days mental health was not good
- How many days feel anxious
- Pain make it hard for usual activities
- Blood Pressure
- Ever told you had chronic bronchitis
- Annual household income
- Total Bilirubin (umol/L)
- How much sleep do you get (hours)?

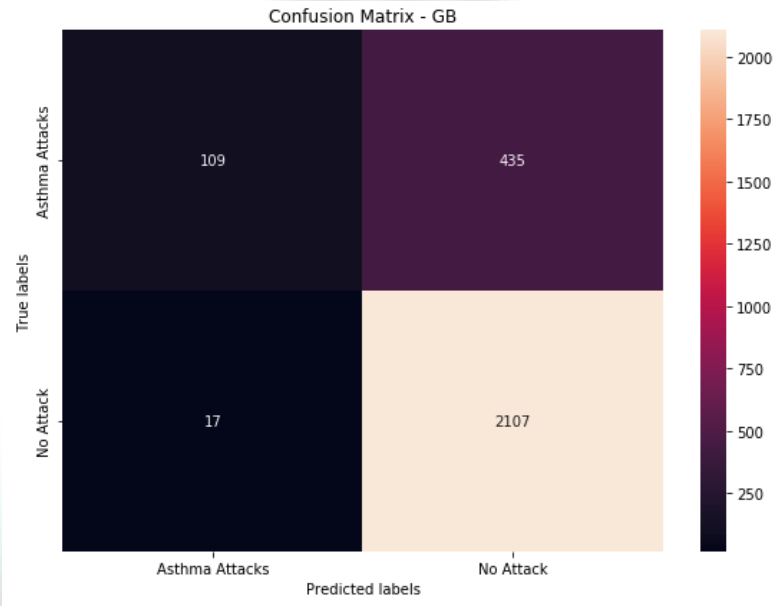
# Partial Dependence plot example

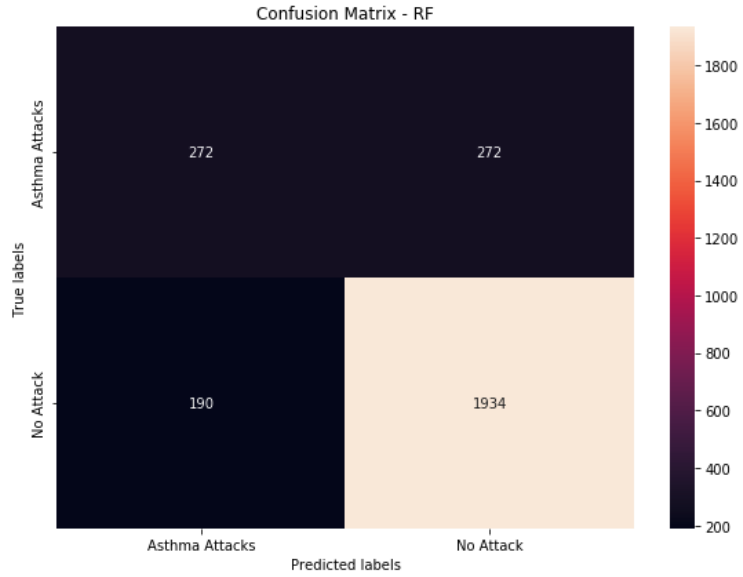



- The PDP shows the dependence between asthma attacks and annual household income.
- Plotting an individual feature against the outcome is interesting to tackle black-box effects

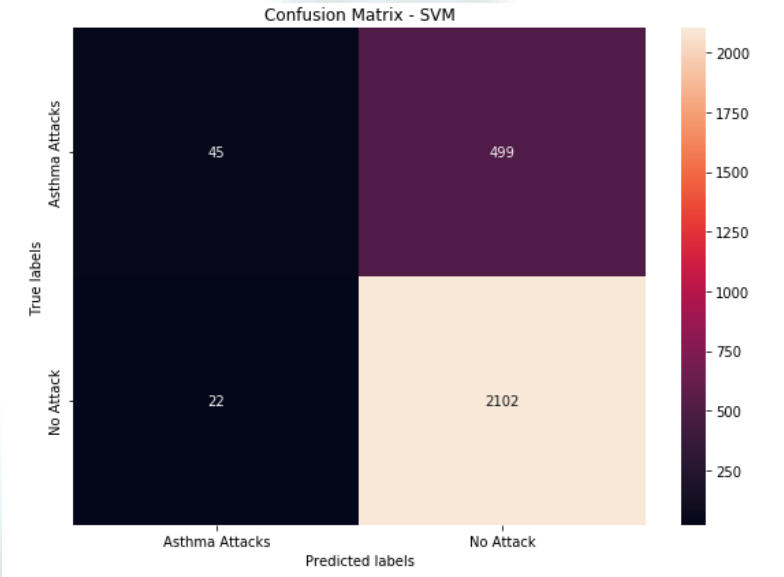
Imperial College London

# Confusion Matrices for Gradient Boost and Random Forest



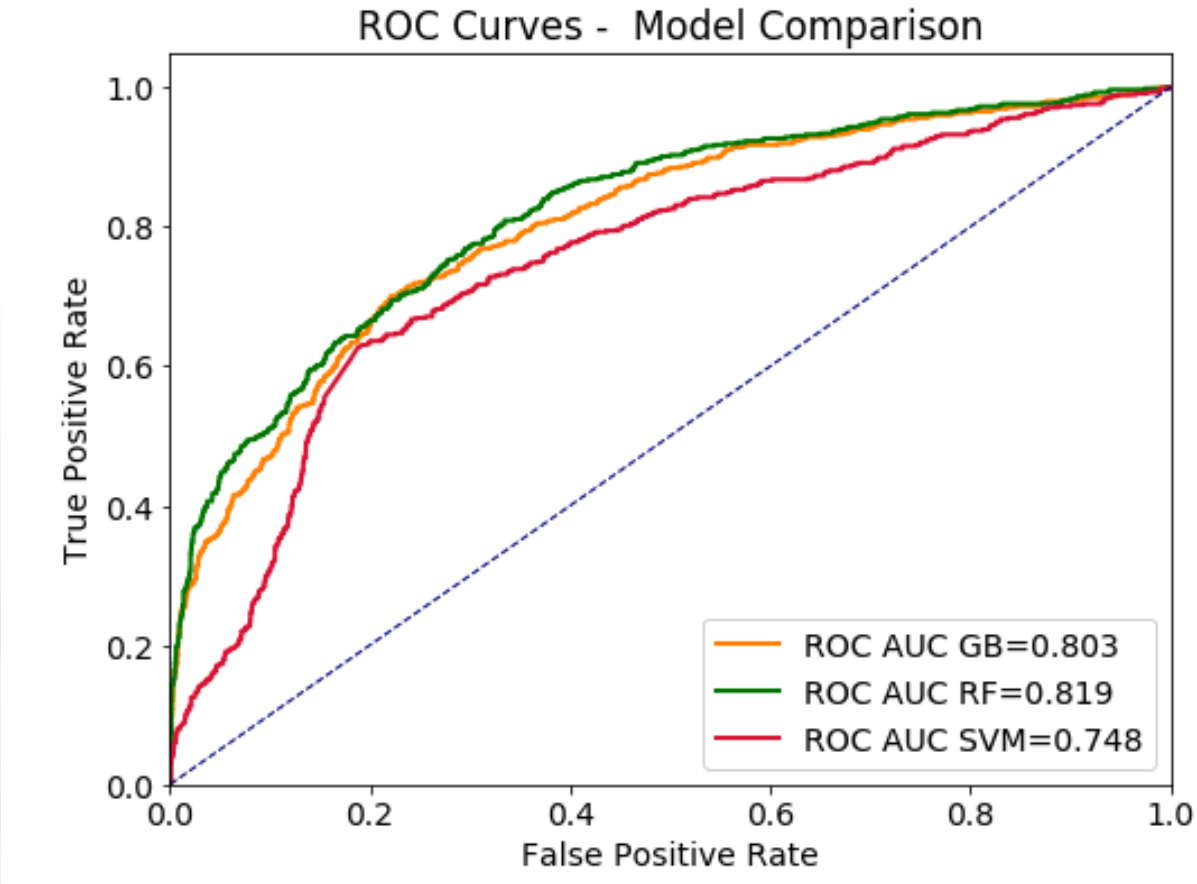
# Confusion Matrix for SVM and score metrics for all models




| | accuracy | f1 | precision | recall |
|---|---|---|---|---|
| **Random Forest** | 0.826837 | 0.540755 | 0.588745 | 0.5 |
| **Gradient Boost** | 0.830585 | 0.325373 | 0.865079 | 0.200368 |
| **SVM** | 0.804723 | 0.1473 | 0.671642 | 0.0827206 |

- Gradient Boost has a slightly higher accuracy than RF and SVM
- But Random forest has the highest recall

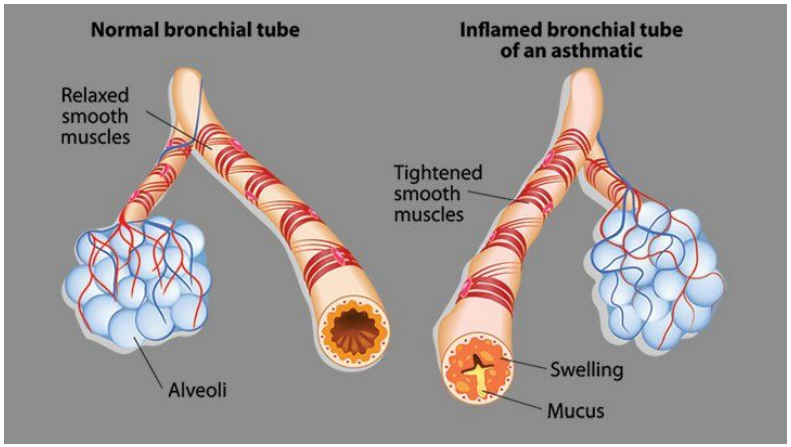
# AUC-ROC curves for RF, GB and SVM models



Imperial College London

# Table of Contents





# Limitations

- Laboratory and Questionnaire data have too many missing values.
- Self-reporting bias in data from surveys.
- Implementation of some population bias with the nature of the nested case-control study developed.
- Imbalanced available information in variables. Imputation techniques remain specific to the developer.
- GridSearch analyses can always be improved.
- Monitoring environmental exposure is complex and highly subjective.
- What is highest priority and target ? Reducing disability or premature deaths arsising from asthma attacks ?
- Nationally representative data might not generalize well globally.

# Actionable plans

- Clustering results could be further developed towards semi-supervised learning
- Restrict study population to healthy controls to prevent population bias
- In-depth investigation of co-occurring factors that might be indicators of future asthma attacks
- Improving access to cost-effective interventions in lower-income countries
- Surveillance to map the magnitude of asthma worldwide
- Primary prevention to reduce the level of exposure to common risk factors

# Table of Contents