# Research & Prediction of Tuberculosis cases in a country.

By Ibrahim Khan

## Context & Introduction :

**Tuberculosis (TB) is caused by bacteria (Mycobacterium tuberculosis) that most often affect the lungs. Tuberculosis is curable and preventable.**

**TB is spread from person to person through the air. When people with lung TB cough, sneeze or spit, they propel the TB germs into the air. A person needs to inhale only a few of these germs to become infected.**

**About one-quarter of the world's population has latent TB, which means people have been infected by TB bacteria but are not (yet) ill with the disease and cannot transmit the disease.**

**TB occurs in every part of the world. In 2018, the largest number of new TB cases occurred in the South-East Asian region, with 44% of new cases, followed by the African region, with 24% of new cases and the Western Pacific with 18%.**

**In 2018, 87% of new TB cases occurred in the 30 high TB burden countries. Eight countries accounted for two thirds of the new TB cases: India, China, Indonesia, Philippines, Pakistan, Nigeria, Bangladesh and South Africa.**

## Problem Statement :

**With the number of TB incidences for every country in 2018 & the data of factors affecting TB in a country such as it's Human Development Index (HDI), it's expenditure on healthcare, it's GDP per capita, the PM 2.5 air pollution exposure to its citizens, the population density we wish to figure out the following :**

- **Which risk factor (alcohol, tobacco, diabetes) contributes most for TB occurrence?**
- **Which age groups are most prone to TB?**
- **Which countries have the highest ratio of cases in adults vs adolescents for every risk factor?**
- **Correlation between every attribute as listed above to the country's TB incidences?**
- **How accurately can we predict from machine learning the no. of cases in a country with respect to its attributes?**

## Criteria of Success :

**The machine learning model is able to predict the number of cases for at least the 8 major countries within the 95% confidence interval range.**

## Scope of Solution space :

**Finding correlations between the various factors & the cases might help us in figuring out about other factors which we might have not considered yet.**

**Initially we will predict the cases irrespective of the risk factor but if the model demands data of specific risk factors we will incorporate that.**

**We will consider the 'best' column which is the mean value for the number of cases in the normal distribution for every country.**

## Constraints :

**Since there are several third world countries that have poor economic status & healthcare facilities available. But TB incidences are low in them which might create a problem resulting in inaccurate predictions. Hence one eye has to be kept on the weightage given to every factor.**

## Clients:

**Institutions / Researchers / Students or whichever party finds the study about TB & the factors affecting it useful.**

## Data Sources:

- **TB incidence data by country (2018) -** https://www.who.int/tb/country/data/download/en/
- **Healthcare expenditure data per capita in USD - https://data.worldbank.org/indicator/SH.XPD.CHEX.PC.CD?most_recent_year_desc=true**
- **Population density (per square km)- https://www.gapminder.org/data/**
- **Current Healthcare expenditure (% of GDP) - http://hdr.undp.org/en/data#**
- **Human Development Index (HDI) - http://hdr.undp.org/en/data#**
- **PM2.5 air pollution (mean annual exposure) - https://www.gapminder.org/data/**
- **Total Population (in millions) - http://hdr.undp.org/en/data#**
- **GDP per capita (USD) - https://www.gapminder.org/data/**