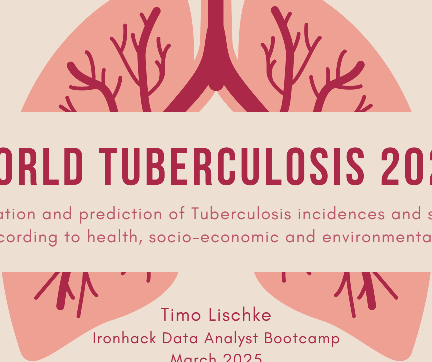


WORLD TUBERCULOSIS 2023

Correlation and prediction of Tuberculosis incidences and severity level according to health, socio-economic and environmental factors

Timo Lischke
Ironhack Data Analyst Bootcamp

March 2025

CONTENT

1. introduction and TB burden data set overview
2. additional data acquistion for data set enrichment
3. data cleaning and wrangling
4. exploratory data analysis
5. feature engineering: correlation of features with TB severity
6. data set preparation for modelling: splitting, normalization, balancing?
7. implementation of different supervised machine learning models to predict number of TB incidences for 2023 according added data, optional: same for TB severity level (groups of incidences)
8. hyperparameter tuning?
9. confusion matrices, other visualizations (tableau)
10. conclusion (key findings), challenges, outlook

WHO TUBERCULOSIS REPORT 2024

Tuberculosis (TB) = contagious lung infection caused by
Mycobacterium tuberculosis (MTB) bacteria.
TB was the world´s leading infectious disease killer in 2023.
Worldwide 1.25 million people died due to TB in 2023.
Worldwide 10.8 million people fell ill with TB in 2023.
Development of incidence rates per country over time were reported.
Still no effective prevention (vaccine) available.
Bacillus Calmette-Guérin (BCG) vaccine statistics reported.
Only suboptimal treatment options available.
https://www.who.int/teams/global-tuberculosis-programme/data

DATA ENRICHMENT OF TUBERCULOSIS REPORT

Objective: Correlate TB incidences in 2023 with further disease-related information (treatment resistance & BCG vaccination rate), other health indicators (smoking rates), socio-economic (population density, poverty index) and environmental (air pollution) circumstances.

Task: data acqusition and enrichment

- Data on treatment resistance & BCG vaccination rate downloaded from WHO.
- Include air pollution data (average annual fine particulate matter <2.5 μm diameter in μg/m³) per country for 2023 obtained from IQAIR (https://www.iqair.com/us/world-most-polluted-countries) or for 2019 from WHO (https://www.who.int/data/gho/data/themes/air-pollution/who-air-quality-database)
- Include multidimensional poverty index (MPI) data per country for 2023 obtained from UNDP (United Nations Development Programme) Human Development Report (HDR) (https://hdr.undp.org/content/2023-global-multidimensional-poverty-index-mpi)
- Include population density (https://database.earth/population/density/2023)
- Include smoking rates per country for 2022 (https://worldpopulationreview.com/country-rankings/smoking-rates-by-country)

EXPLORATORY DATA ANALYSIS

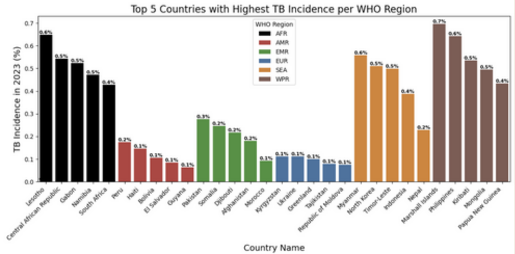
Objective: Display top 5 countries per world region according to: TB incidences, treatment resistance, BCG vaccination rate, population density, poverty index, smoking rates, air pollution

Three-letter code of WHO regions:

- AFR → African Region
- AMR → Region of the Americas
- EMR → Eastern Mediterranean Region
- EUR → European Region
- SEA → South-East Asia Region
- WPR → Western Pacific Region

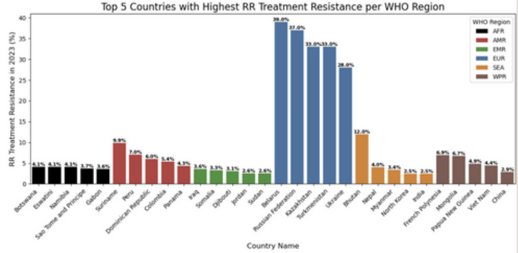
EXPLORATORY DATA ANALYSIS

Top 5 countries per world region according to: TB incidences



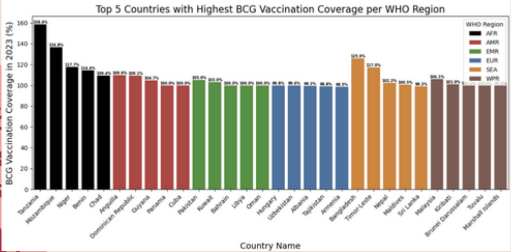
EXPLORATORY DATA ANALYSIS

Top 5 countries per world region according to: treatment resistance



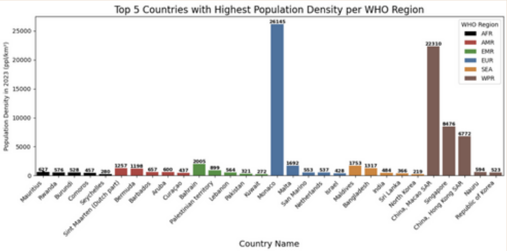
EXPLORATORY DATA ANALYSIS

Top 5 countries per world region according to: BCG vaccination rate



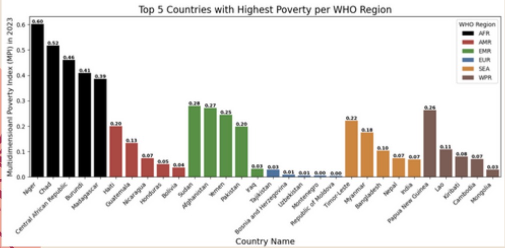
EXPLORATORY DATA ANALYSIS

Top 5 countries per world region according to: population density



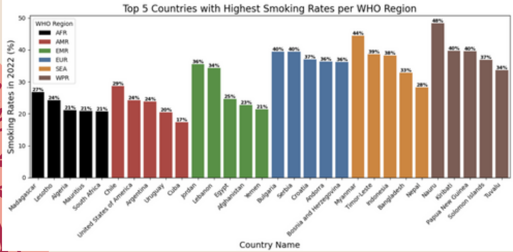
EXPLORATORY DATA ANALYSIS

Top 5 countries per world region according to: poverty index



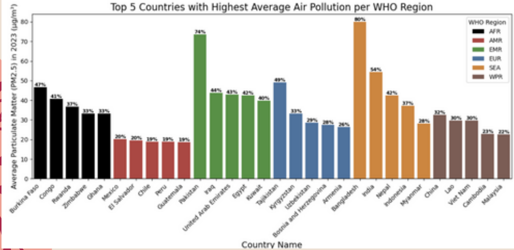
EXPLORATORY DATA ANALYSIS

Top 5 countries per world region according to: smoking rates



EXPLORATORY DATA ANALYSIS

Top 5 countries per world region according to: air pollution

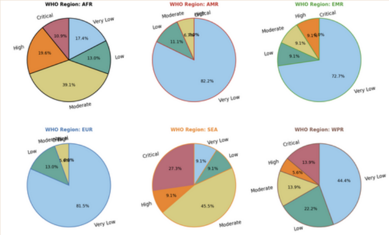


TUBERCULOSIS INCIDENCE & SEVERITY LEVEL

Distribution of TB severity level in the 6 world regions (target)

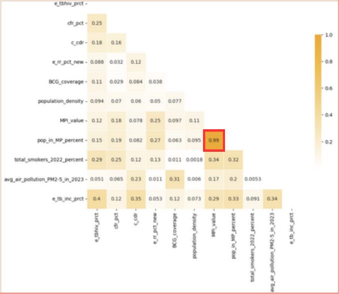
Based on SD intervals: 0.1 (Mean), 0.244 (Mean to Mean + 1 SD), & 0.388 (Mean + 1 SD to Mean + 2 SD)

Levels of TB severity: Very Low ⩽0.05, Low ⩽ 0.1, Moderate ⩽ 0.244, High ⩽ 0.388, Critical > 0.388



CORRELATION OF FEATURES WITH TB INCIDENCE

3 features from TB data set,
7 enriched features,
--> drop % pop in pov



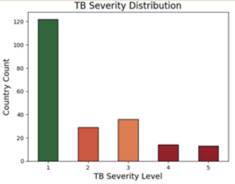
	e_tbhiv_prct	cfr_pct	c_cdr	e_rr_pct_new	BCG_coverage	population_density	MPI_value	pop_in_MP_percent	total_smokers_2022_percent	avg_air_pollution_PM2-5_in_2023	e_tb_inc_prct
e_tbhiv_prct											
cfr_pct	0.25										
c_cdr	0.18	0.16									
e_rr_pct_new	0.088	0.032	0.12								
BCG_coverage	0.11	0.029	0.084	0.038							
population_density	0.094	0.07	0.06	0.05	0.077						
MPI_value	0.12	0.18	0.078	0.25	0.097	0.11					
pop_in_MP_percent	0.15	0.19	0.082	0.27	0.063	0.095	0.99				
total_smokers_2022_percent	0.29	0.25	0.12	0.13	0.011	0.0018	0.34	0.32			
avg_air_pollution_PM2-5_in_2023	0.051	0.065	0.23	0.011	0.31	0.006	0.17	0.2	0.0053		
e_tb_inc_prct	0.4	0.12	0.35	0.053	0.12	0.073	0.29	0.33	0.091	0.34	

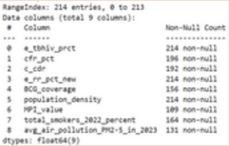
DATA PREPARATION FOR PREDICTION MODELLING

total 214 countries: train & test splitting = 171 (80%) & 43 (20%)

normalization: Min/Max scaling, many NaN values in data,

SMOTE target balancing, class weights





```
RangeIndex: 214 entries, 0 to 213
Data columns (total 9 columns):
 #   Column                             Non-Null Count
---  ------                             --------------
 0   e_tbhiv_prct                       214 non-null
 1   cfr_pct                            196 non-null
 2   c_cdr                              192 non-null
 3   e_rr_pct_new                       214 non-null
 4   BCG_coverage                       156 non-null
 5   population_density                 214 non-null
 6   MPI_value                          109 non-null
 7   total_smokers_2022_percent         164 non-null
 8   avg_air_pollution_PM2-5_in_2023    131 non-null
dtypes: float64(9)
```

PREDICTION OF TB SEVERITY USING ML MODELS

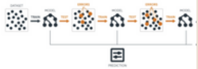
Implement supervised ML models to predict TB severity levels

Ensemble prediction model testing:

- HistGradientBoostingClassifier
- RandomForestClassifier (DecTree + RandPatch)

Model modifications/improvements:

- impute missing NaN using KNN
- target parameter balancing using SMOTE or Class Weight balancing on RFC model
- hyperparameter tuning for HGBC model



RANDOM FOREST

EVALUATE PREDICTION MODEL´S PERFORMANCE

Evaluation according to prediction precision, recall, F1-score, accuracy

model	precision	recall	F1-score	accuracy
HistGradBoost	0.59	0.56	0.54	0.56
Random Forest	0.48	0.63	0.54	0.63
RFC + CW	0.47	0.65	0.54	0.65
RFC + SMOTE	0.58	0.49	0.47	0.49
RFC(CW)+SMOTE	0.38	0.40	0.38	0.40
HGBC_CV				

XXXXXXXXXXXXXXXXXX

XXXXXXXXXXXXXXXX

XXXXX

- XXXXX

XXXXXXXXXXXXXXXXXX

XXXXXXXXXXXXXXXX

XXXXX

- XXXXX

XXXXXX

XXXXX

XXXXX

- XXXXX