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 (web scraping)
3. data cleaning and wrangling
4. exploratory data analysis
5. feature engineering: correlation of features with TB severity
6. data model preparation: splitting, normalization, balance check
7. supML model implementation: predict TB severity according added data
8. HGBC: hyperparameter tuning, RFC: CW + NaN imputing/SMOTE-balancing
9. model evaluation: classification, feature importance, confusion matrices
10. unsupervised ML models: Kmeans & hierarchical clustering
11. further visualizations (tableau)
12. 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 /severity level 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)

DATA CLEANING AND WRANGLING

Preprocessing of data for modelling:

- all column names conclusive & lower case
- some columns dropped
- all values numerical
- still many null values, replaced with NaN
- no duplicated values
- country iso2 and iso3 codes introduced

-> Extensive cleaning needed before table merging

EXPLORATORY DATA ANALYSIS

Top 5 countries per world region displayed according to:

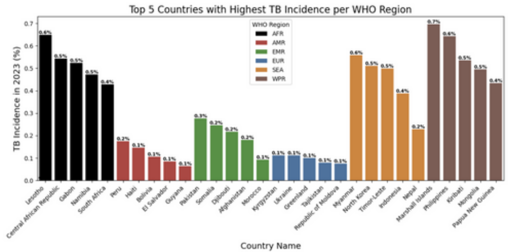
- TB incidences
- treatment resistance
- BCG vaccination rate
- population density
- poverty index
- smoking rates
- air pollution
- TB severity

Three-letter code of the 6 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

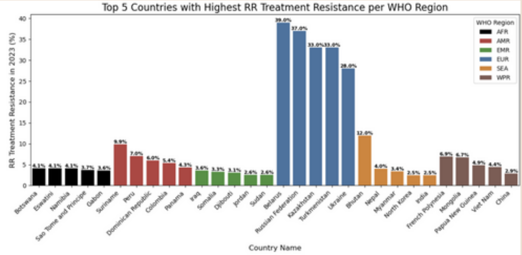
EXPLORATORY DATA ANALYSIS

Top 5 countries per world region according to: TB incidences



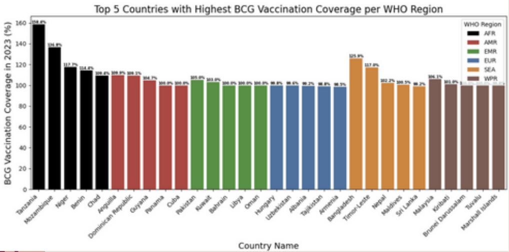
EXPLORATORY DATA ANALYSIS

Top 5 countries per world region according to: treatment resistance



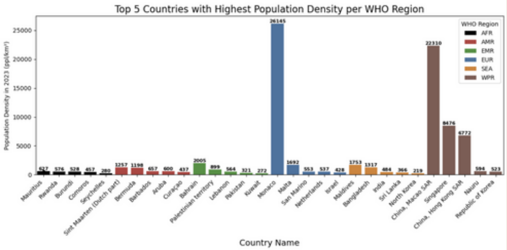
EXPLORATORY DATA ANALYSIS

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



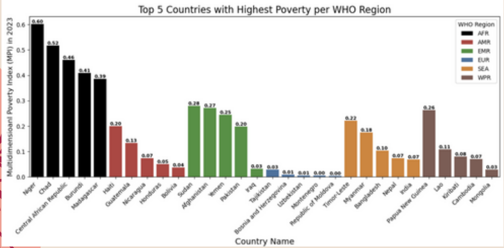
EXPLORATORY DATA ANALYSIS

Top 5 countries per world region according to: population density



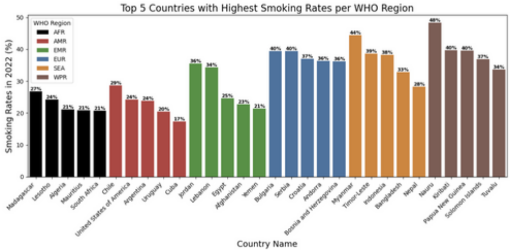
EXPLORATORY DATA ANALYSIS

Top 5 countries per world region according to: poverty index



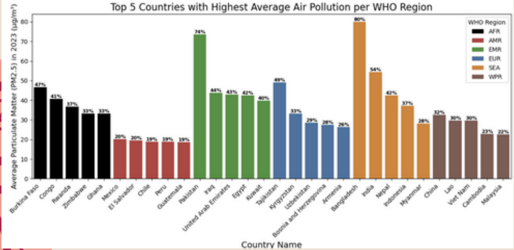
EXPLORATORY DATA ANALYSIS

Top 5 countries per world region according to: smoking rates



EXPLORATORY DATA ANALYSIS

Top 5 countries per world region according to: air pollution

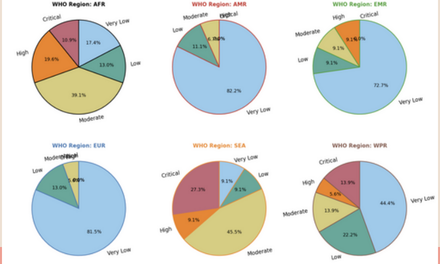


TUBERCULOSIS INCIDENCES: SEVERITY LEVEL

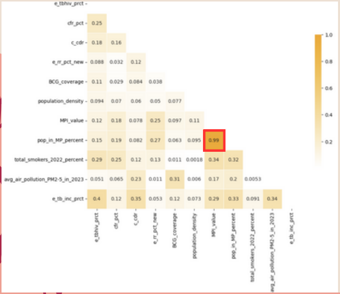
Distribution of TB severity level in the 6 world regions (target)

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



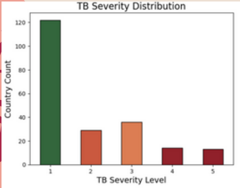
FEATURE ENGINEERING AND SELECTION



Correlation of features with TB incidence:
3 features from TB burden data set,
7 enriched features -> drop % pop in pov
-> 9 features for model

DATA PREPARATION FOR PREDICTION MODELLING

total 214 countries: train & test split = 171 (80%) & 43 (20%)
normalization: Min/Max scaling, tuning: hyperparameters
balancing: impute NaN -> SMOTE, class weighting




```
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 optimization:

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

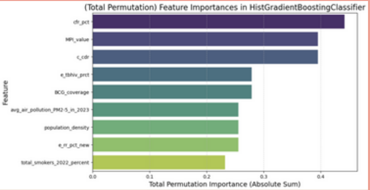
EVALUATE PREDICTION MODEL´S PERFORMANCE

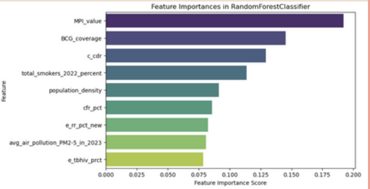
Evaluation according to prediction precision, recall, F1-score, accuracy

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_HT	0.47	0.65	0.54	0.65

EVALUATION

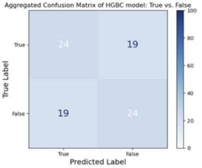
FEATURE IMPORTANCE

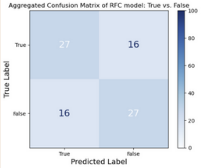


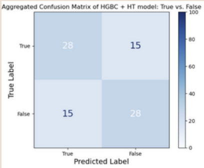


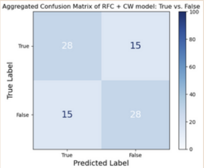
EVALUATION

CONFUSION MATRICES









PREDICTION OF TB SEVERITY USING ML MODELS

Implement **unsupervised ML models** to predict TB severity levels

Clustering prediction model testing:

- no train/test split, use whole data for clustering
- impute missing NaN using KNN, Min/Max norm.
- KMeans clustering (not useful)
  - Adjusted Rand Index (ARI): -0.0361 (similarity)
  - Normalized Mutual Information (NMI): 0.0378 (dep.)
- create dendrogram to determine cluster no.
- Hierarchical clustering (not useful either)
  - Adjusted Rand Index (ARI): 0.1552 (similarity)
  - Normalized Mutual Information (NMI): 0.1043 (dep.)

VISUALIZATIONS (TABLEAU)

XXXXXXXXXXXXXXXX

XXXXX

- XXXXX

KEY FINDINGS AND INSIGHTS

XXXXXXXXXXXXXXXX

XXXXX

- XXXXX

REAL WORLD APPLICATION AND IMPACT

XXXXXXXXXXXXXXXX

XXXXX

- XXXXX

CHALLENGES AND LEARNINGS

XXXXXXXXXXXXXXXX

XXXXX

- XXXXX

FUTURE WORK AND IMPROVEMENTS

Improving data set:

- try to fill in missing values for 2023
- try to find other data for enrichment with higher correlation to target
- include data of years before 2023

Improving supervised ML models:

- run time corrleated predictions

Improving unsupervised clustering models:

- Perform PCA before running unsupervised clustering models to reduce noise and redundant features.
- Experiment with different values of K for Kmeans (e.g., using the Elbow Method or Silhouette Score).
- Consider other clustering algorithms like DBSCAN.

THANK YOU !

Correlation and prediction of Tuberculosis incidences and severity level according to health, socio-economic and environmental factors (based on WHO Tuberculosis report 2024

Timo Lischke
Ironhack Data Analyst Bootcamp
March 2025