**FINAL PROJECT**

**GLOBAL TUBERCULOSIS TRENDS ANALYSIS AND PREDICTIONS**

**KRUTIKA DESHMUKH**

**HAP 780**

**PROF DR. SANJA AVRAMOVIC**

### **Abstract:**

Tuberculosis (TB) remains a significant public health challenge, particularly in low- and middle-income countries. This project focuses on analyzing global trends in TB incidence, exploring the relationship between TB and key risk factors such as HIV prevalence, and predicting future trends through data analysis. The primary objective is to generate actionable insights that can inform public health interventions, optimize resource allocation, and guide policy decisions aimed at reducing the global burden of TB.

### **Objective:**

The objective of this project is to conduct an in-depth analysis of global trends in tuberculosis (TB) incidence and mortality, with an emphasis on identifying correlations between TB and key risk factors such as HIV prevalence. By leveraging advanced data mining methodologies and machine learning techniques, including Logistic Regression, Random Forest, Clustering, and Naive Bayes, the project seeks to uncover patterns in TB data across various developed and underdeveloped countries. The ultimate goal is to predict future TB outbreaks, provide actionable insights for public health interventions, optimize resource allocation, and inform policy decisions, particularly in regions most affected by the disease.

### **Data:**

The dataset utilized in this study was sourced from Kaggle and includes comprehensive data on tuberculosis, such as age-specific TB deaths, incidence rates for all forms of tuberculosis, case detection rates, treatment success rates for all forms of TB, MDR, and XDR TB, as well as HIV incidence rates related to tuberculosis across various developed and underdeveloped countries.

### **Method:**

The study began with preprocessing the dataset using SQL, followed by further preprocessing in Weka. Classification algorithms, including Logistic Regression and Random Tree, were employed to build predictive models. These models were evaluated using various metrics such as precision, accuracy, recall, and F1-score, with the ROC area value used to compare the overall performance of the models.

### **Results:**

The results indicate that the Logistic Regression model achieved the highest accuracy, at 99.3%, compared to the Random Tree algorithm. Although clustering was employed to gain additional insights, it did not significantly enhance the model's performance.

### **Keywords:** Tuberculosis, incidence rate, classification algorithms, Weka, predictive modeling

**Introduction:** TB is the leading cause of death among people living with HIV and contributes significantly to antimicrobial resistance, posing a major challenge to global health. The disease predominantly affects individuals in low- and middle-income countries.

Globally, approximately 25% of the population is estimated to be infected with TB bacteria, although the majority will not develop the disease. However, those who are immunocompromised, such as individuals with HIV, malnutrition, diabetes, or those who use tobacco, face a higher risk of developing active TB. Understanding these trends and predicting future outbreaks is crucial for effective public health planning

MDR : Multi Drug Resistant TB

XDR : Extensively Drug Resistant

**STEPS IN THE DATA ANALSYIS:**

1. **Data Preprocessing using SQL**: Preparing the dataset by cleaning, organizing, and structuring the data for analysis.
2. **Descriptive Statistics and Initial Observations**: Summarizing the data to identify key patterns and trends at the outset of the analysis.
3. **Regional Analysis of Tuberculosis Incidence**: Examining TB incidence across different regions to identify high-risk areas.
4. **Age Group Analysis for Mortality**: Analysing mortality rates by age group to understand which populations are most vulnerable to TB.
5. **Comparative and Correlational Analysis**: Identifying correlations and comparing different factors, such as incidence rates across regions and age groups, treatment success rates, and the impact of HIV on tuberculosis outcomes.
6. **Predictive Modelling and Classification using WEKA**: Developing predictive models to forecast future trends or classify data points, such as identifying countries at risk of high tuberculosis incidence.

**DATASET DESCRIPTION:**

For Analysis I am using Five tables they are [dbo].[ incidence-of-tuberculosis-sdgs], [dbo].[tuberculosis-deaths-by-age], dbo].[tuberculosis-case-detection-rate] , [dbo].[4- tuberculosis-treatment-success-rate-by-type], [Estimated\_HIV\_in\_incident\_tuberculosis]. Below are the variables from the tables-

1. **Entity**: This column likely represents the country or region for which the tuberculosis data is recorded (e.g., "Afghanistan"). There are total of 203 countries in the dataset out of which I have selected the countries which are common in all the tables.
2. **Code**: This column contains the three-letter country code (e.g., "AFG" for Afghanistan), representing the corresponding entity.
3. **Year**: This column records the year in which the data was collected or reported (e.g., 2000, 2001). In the Data this column contains the records from the Year 1990 to 2019
4. **Estimated\_incidence\_of\_all\_forms\_of\_tuberculosis**: This column indicates the estimated incidence rate of all forms of tuberculosis within a specific entity for a particular year.
5. **Indicator\_Treatment\_success\_rate\_new\_TB\_cases**: This column shows the success rate of treatment for new TB cases within the entity for a given year.
6. **Indicator\_Treatment\_success\_rate\_for\_patients\_treated\_for\_MDR\_TB**: This column represents the treatment success rate specifically for patients treated for Multi-Drug Resistant Tuberculosis (MDR-TB).
7. **Estimated\_HIV\_in\_incident\_tuberculosis**: This column provides an estimate of HIV incidence among tuberculosis cases within the specified entity for that year.
8. **Case\_detection\_rate\_all\_forms**: This column indicates the detection rate of all forms of tuberculosis cases within a specific entity for a particular year.
9. **Deaths\_Tuberculosis\_Sex\_Both\_Age\_70\_years\_Number**: This column provides the number of deaths due to tuberculosis in individuals aged 70 years and above, irrespective of sex, within the specified entity and year.
10. **Deaths\_Tuberculosis\_Sex\_Both\_Age\_50\_69\_years\_Number**: This column provides the number of deaths due to tuberculosis in individuals aged between 50 and69 years, irrespective of sex, within the specified entity and year.
11.  **Deaths\_Tuberculosis\_Sex\_Both\_Age\_15\_49\_years\_Number**: This column provides the number of deaths due to tuberculosis in individuals aged between 15 and 49 years, irrespective of sex, within the specified entity and year.
12.  **Deaths\_Tuberculosis\_Sex\_Both\_Age\_5\_14\_years\_Number**: This column indicates the number of deaths due to tuberculosis in individuals aged between 5 and 14 years, irrespective of sex, within the specified entity and year.
13.  **Deaths\_Tuberculosis\_Sex\_Both\_Age\_Under\_5\_Number**: This column represents the number of deaths due to tuberculosis in children under 5 years of age, irrespective of sex, within the specified entity and year.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

### **DATA PREPROCESSING**

### Before diving into analysis, we need to thoroughly explore and clean the data, ensuring it is ready for meaningful analysis.

Let’s find the similar countries in all the Four tables so that we can Join them.

-- Countries in 'incidence-of-tuberculosis-sdgs' not in 'tuberculosis-deaths-by-age'

SELECT [Entity] AS CountryName FROM [dbo].[ incidence-of-tuberculosis-sdgs]

EXCEPT

SELECT [Entity] FROM [dbo].[tuberculosis-deaths-by-age];

-- Countries in 'tuberculosis-deaths-by-age' not in 'incidence-of-tuberculosis-sdgs'

SELECT [Entity] AS CountryName FROM [dbo].[tuberculosis-deaths-by-age]

EXCEPT

SELECT [Entity] FROM [dbo].[ incidence-of-tuberculosis-sdgs]

A screenshot of a computer

Description automatically generated

SELECT [Entity] AS CountryName FROM [dbo].[tuberculosis-case-detection-rate]

EXCEPT

SELECT [Entity] FROM [dbo].[4- tuberculosis-treatment-success-rate-by-type]

SELECT [Entity] AS CountryName FROM [dbo].[4- tuberculosis-treatment-success-rate-by-type]

EXCEPT

SELECT [Entity] FROM [dbo].[tuberculosis-case-detection-rate]

A screenshot of a computer

Description automatically generated

In the above screenshots you can find the similar Countries present in 4 tables they are [dbo].[ incidence-of-tuberculosis-sdgs], [dbo].[tuberculosis-deaths-by-age], [dbo].[tuberculosis-case-detection-rate], [dbo].[4- tuberculosis-treatment-success-rate-by-type],

In next step is joining all the four tables to a final table which is dbo.finalTB on the column Entity nothing but name of the country.

------------- Joining two tables-------

SELECT a.[Estimated\_incidence\_of\_all\_forms\_of\_tuberculosis],b.\* into #newtable

FROM [dbo].[ incidence-of-tuberculosis-sdgs] AS a

JOIN [dbo].[tuberculosis-deaths-by-age] AS b

ON a.[Entity] = b.[Entity]

ORDER BY a.[Entity]

-------------203 same countries-----

------139380-----

A screenshot of a computer

Description automatically generated

SELECT a.\*,b.[Indicator\_Treatment\_success\_rate\_new\_TB\_cases],b.[Indicator\_Treatment\_success\_rate\_for\_patients\_treated\_for\_MDR\_TB],b.[Indicator\_Treatment\_success\_rate\_XDR\_TB\_cases]

into #secondtable

FROM #newtable AS a

INNER JOIN [dbo].[4- tuberculosis-treatment-success-rate-by-type] AS b

ON a.[Entity] = b.[Entity]

ORDER BY a.[Entity]

----------------2911890-------------

A screenshot of a computer

Description automatically generated

SELECT a.\*,b.[Case\_detection\_rate\_all\_forms]

into dbo.finalTB

FROM #secondtable AS a

INNER JOIN [dbo].[tuberculosis-case-detection-rate] AS b

ON a.[Entity] = b.[Entity]

ORDER BY a.[Entity]

-------65251170------

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

After Joining the tables I removed all the Null values from the Final Table.

delete from [dbo].[finalTB] where [Indicator\_Treatment\_success\_rate\_new\_TB\_cases] is Null

delete from [dbo].[finalTB] where [Indicator\_Treatment\_success\_rate\_for\_patients\_treated\_for\_MDR\_TB] is Null

delete from [dbo].[finalTB] where [Indicator\_Treatment\_success\_rate\_XDR\_TB\_cases] is Null

delete from [dbo].[finalTB] where [Case\_detection\_rate\_all\_forms] is Null

A screenshot of a computer

Description automatically generated

In the Next step select all the distinct countries from the final table into new table which is dbo.TB

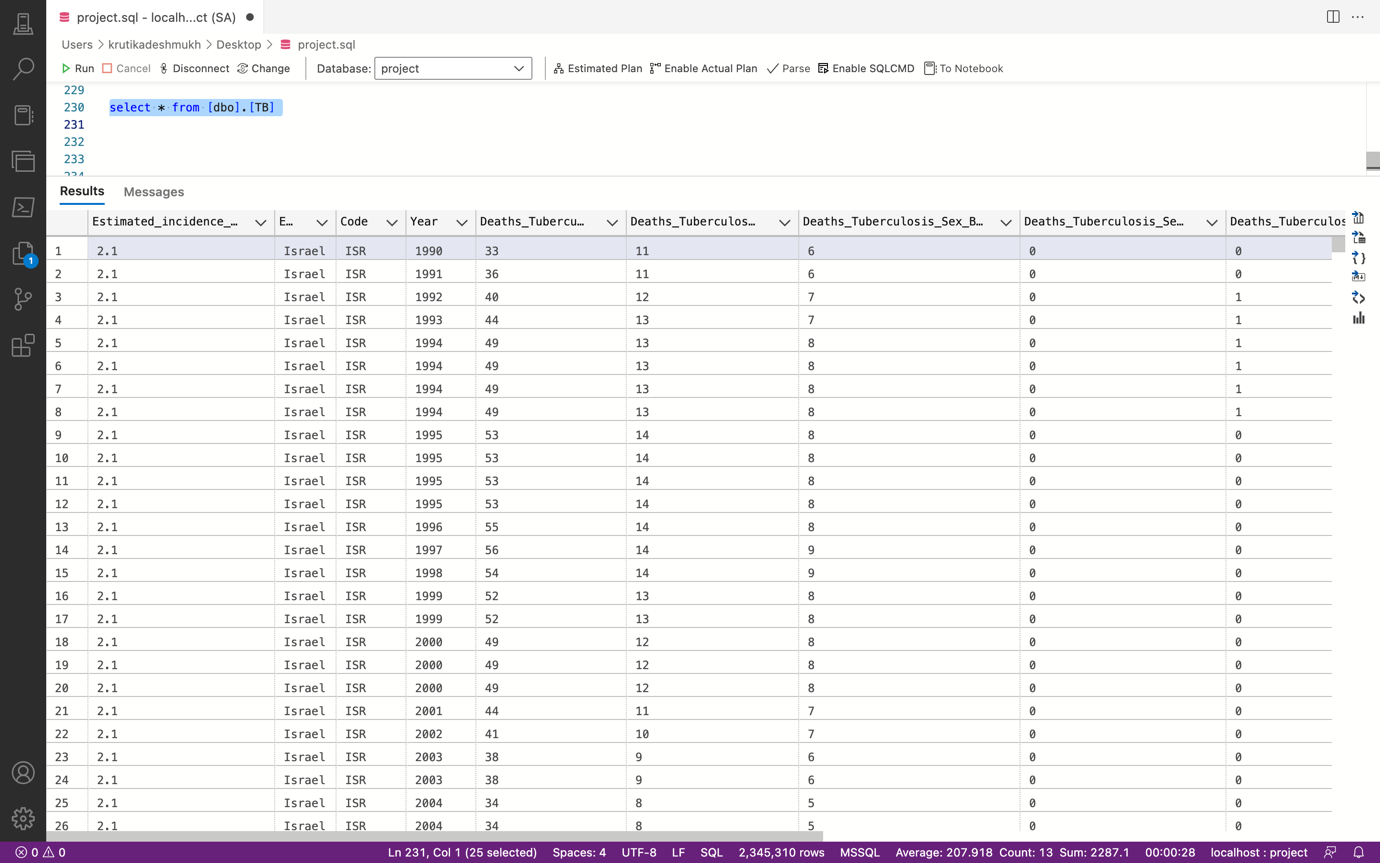
select distinct \* into dbo.TB from [dbo].[finalTB]

A screenshot of a computer

Description automatically generated

**Final Table:**

select \* from [dbo].[TB]



1. **Descriptive Statistics and Initial Observations**

**INCIDENCE RATE:**

The incidence rate typically refers to the number of new cases of a disease or condition reported during a specific time period, per unit of population. It's usually expressed as the number of cases per 100,000 people per year. This rate helps public health officials and researchers understand how widely a disease is spreading within a community, making it a crucial measure in epidemiology for tracking the spread of diseases, including infectious diseases like tuberculosis (TB), and chronic conditions like diabetes or cancer.

--------1.2 Descriptive Statistics and Initial Observations-----------

-- Summary statistics for tuberculosis incidence

SELECT

MIN([Estimated\_incidence\_of\_all\_forms\_of\_tuberculosis]) AS min\_incidence,

MAX([Estimated\_incidence\_of\_all\_forms\_of\_tuberculosis]) AS max\_incidence,

AVG([Estimated\_incidence\_of\_all\_forms\_of\_tuberculosis]) AS avg\_incidence,

STDEV([Estimated\_incidence\_of\_all\_forms\_of\_tuberculosis]) AS stddev\_incidence

FROM [dbo].[TB]

--------min\_incidence : 2.1 ,max\_incidence : 1590 , avg\_incidence : 263.0318866162682 , stddev\_incidence : 273.00749104379094

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

--------min\_incidence : 2.1 ,max\_incidence : 1590 , avg\_incidence : 263.0318866162682 , stddev\_incidence : 273.00749104379094

* **Minimum Incidence (2.1)**

Indicates effective TB control and possibly strong healthcare systems in certain regions or periods.

* **Maximum Incidence (1590)**

Highlights regions or times with severe TB challenges, likely due to factors like poor healthcare infrastructure or high

HIV co-infection rates.

* **Average Incidence (263.03)**

Represents the general TB burden across studied areas, useful for assessing global TB control efforts.

* **Standard Deviation (273.01)**

Shows substantial variability in TB incidence, suggesting a need for region-specific health interventions.

### **3)Regional Analysis of Tuberculosis Incidence:**

--------Analyze the trends of TB incidence rates over the years to understand how TB is evolving globally or in specific countries

SELECT [Year], [Entity], AVG([Estimated\_incidence\_of\_all\_forms\_of\_tuberculosis]) AS avg\_incidence\_rate

FROM [dbo].[finalTB]

GROUP BY [Year],[Entity]

ORDER BY year, avg\_incidence\_rate DESC

--------Eswatini -- 971.5652173913044

A screenshot of a computer

Description automatically generated

* **High Incidence Rates**: The countries listed show relatively high TB incidence rates in 1990, particularly in southern Africa.
* **Eswatini (formerly Swaziland)** and **South Africa** have the highest rates, nearing or exceeding 965 cases per 100,000 population. This highlights a significant health burden and suggests that TB was a major public health issue in these countries at the time.
* **Regional Impact**: The data indicates a regional trend in southern Africa (Eswatini, South Africa, Lesotho, Namibia) where TB incidence rates are particularly high. This region may have faced specific socio-economic or healthcare-related challenges contributing to these high rates.

**Insights**: From this analysis, you can identify which regions have higher or lower average incidence rates and how these have changed over time.

#### **4)Age Group Analysis for Mortality**

SELECT

year,

CONVERT(FLOAT, Deaths\_Age\_70\_plus) / CONVERT(FLOAT, Total\_Deaths) \* 100 AS Pct\_Deaths\_Age\_70\_plus,

CONVERT(FLOAT, Deaths\_Age\_50\_to\_69) / CONVERT(FLOAT, Total\_Deaths) \* 100 AS Pct\_Deaths\_Age\_50\_to\_69,

CONVERT(FLOAT, Deaths\_Age\_15\_to\_49) / CONVERT(FLOAT, Total\_Deaths) \* 100 AS Pct\_Deaths\_Age\_15\_to\_49,

CONVERT(FLOAT, Deaths\_Age\_5\_to\_14) / CONVERT(FLOAT, Total\_Deaths) \* 100 AS Pct\_Deaths\_Age\_5\_to\_14,

CONVERT(FLOAT, Deaths\_Age\_Under\_5) / CONVERT(FLOAT, Total\_Deaths) \* 100 AS Pct\_Deaths\_Age\_Under\_5

FROM

(SELECT

year,

SUM(CONVERT(BIGINT, [Deaths\_Tuberculosis\_Sex\_Both\_Age\_70\_years\_Number])) AS Deaths\_Age\_70\_plus,

SUM(CONVERT(BIGINT, [Deaths\_Tuberculosis\_Sex\_Both\_Age\_50\_69\_years\_Number])) AS Deaths\_Age\_50\_to\_69,

SUM(CONVERT(BIGINT, [Deaths\_Tuberculosis\_Sex\_Both\_Age\_15\_49\_years\_Number])) AS Deaths\_Age\_15\_to\_49,

SUM(CONVERT(BIGINT, [Deaths\_Tuberculosis\_Sex\_Both\_Age\_5\_14\_years\_Number])) AS Deaths\_Age\_5\_to\_14,

SUM(CONVERT(BIGINT, [Deaths\_Tuberculosis\_Sex\_Both\_Age\_Under\_5\_Number])) AS Deaths\_Age\_Under\_5,

SUM(CONVERT(BIGINT, [Deaths\_Tuberculosis\_Sex\_Both\_Age\_70\_years\_Number])

+ CONVERT(BIGINT, [Deaths\_Tuberculosis\_Sex\_Both\_Age\_50\_69\_years\_Number])

+ CONVERT(BIGINT, [Deaths\_Tuberculosis\_Sex\_Both\_Age\_15\_49\_years\_Number])

+ CONVERT(BIGINT, [Deaths\_Tuberculosis\_Sex\_Both\_Age\_5\_14\_years\_Number])

+ CONVERT(BIGINT, [Deaths\_Tuberculosis\_Sex\_Both\_Age\_Under\_5\_Number])) AS Total\_Deaths

FROM

[dbo].[TB]

GROUP BY

year

) AS SubQuery

ORDER BY

year DESC

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

* **Least Affected Age Groups:**

**Under 5 Years:** Rarely exceeds 5%.

**Ages 5 to 14:** Low impact, percentages around 1% to 5%.

* **Most Affected Age Groups:**

**Ages 50 to 69:** Often above 30%, indicating high vulnerability.

**Ages 70+:** Frequently in the high 20s to 30s percent range.

* These insights suggest targeted health interventions might be particularly necessary for the older age groups, while the resilience of younger populations should be further studied to understand protective factors or reporting discrepancies.

**Insights**: This analysis reveals which age groups are most vulnerable to tuberculosis. It can highlight the need for targeted interventions for specific age groups.

### **5)Comparative and Correlational Analysis**

### ***Correlation Between HIV Prevalence and Tuberculosis Incidence***

----Comparative and Correlational Analysis------

-----Correlation Between HIV Prevalence and Tuberculosis Incidence-----

--------- Correlation between HIV prevalence and tuberculosis incidence

SELECT distinct

t1.[Entity],

t1.[Year],

t1.[Estimated\_incidence\_of\_all\_forms\_of\_tuberculosis],

[Estimated\_HIV\_in\_incident\_tuberculosis],

([Estimated\_HIV\_in\_incident\_tuberculosis] \* [Estimated\_incidence\_of\_all\_forms\_of\_tuberculosis]) / (AVG([Estimated\_HIV\_in\_incident\_tuberculosis]) OVER() \* AVG([Estimated\_incidence\_of\_all\_forms\_of\_tuberculosis]) OVER()) AS correlation\_factor

FROM [dbo].[TB] t1

JOIN [dbo].[5- tuberculosis-patients-with-hiv-share] t2 ON t1.[Entity] = t2.[Entity] AND t1.year = t2.year

ORDER BY correlation\_factor DESC;

A screenshot of a computer

Description automatically generated

* · **Persistent TB Burden**: Eswatini consistently shows high TB incidence rates, indicating a severe ongoing challenge.
* · **Significant HIV Prevalence**: HIV prevalence among TB patients remains high, reinforcing the link between TB and HIV infections.
* · **Strong Correlation**: Correlation factors above 25 indicate a robust positive association between HIV prevalence and TB incidence, emphasizing the impact of HIV on TB dynamics.
* · **Stable Yearly Trends**: Despite yearly variations in data, the correlation between HIV and TB remains strong, highlighting their interdependence.
* · **Public Health Strategy**: The data supports integrated public health approaches to effectively manage both HIV and TB in Eswatini.

**Insights**: High correlation factors suggest that HIV prevalence strongly influences tuberculosis incidence in certain regions. This insight can be used to advocate for integrated healthcare strategies targeting both conditions.

#### **6)Treatment Success Rate Comparison**

------------ Avg\_success rate------

SELECT [Entity], AVG([Indicator\_Treatment\_success\_rate\_for\_patients\_treated\_for\_MDR\_TB]) AS avg\_success\_rate\_mdr, AVG([Indicator\_Treatment\_success\_rate\_XDR\_TB\_cases]) AS avg\_success\_rate\_xdr, AVG([Indicator\_Treatment\_success\_rate\_new\_TB\_cases])

AS avg\_success\_rate\_newcases FROM [dbo].[TB]

GROUP BY [Entity]

ORDER BY avg\_success\_rate\_mdr DESC, avg\_success\_rate\_xdr DESC,avg\_success\_rate\_newcases DESC ;

A screenshot of a computer

Description automatically generated

1. **High MDR TB Treatment Success**: Benin and Sweden excel in MDR TB treatment, with Benin achieving a 100% success rate, indicating effective management strategies.
2. **Variable XDR TB Success Rates**: Significant variation exists in XDR TB treatment success, with Sweden at 60% and Haiti at 0%, reflecting disparities in treatment capabilities.
3. **Effective New TB Case Management**: Countries like Benin lead with high success rates (88%) for new TB cases, showcasing strong initial treatment protocols.
4. **Challenges with Resistant TB**: Countries such as Oman and Myanmar display lower success rates for resistant TB, underscoring the need for enhanced management approaches.
5. **Call for Targeted Interventions**: The data highlights the need for specific health interventions to improve outcomes, particularly in countries with lower success rates for resistant TB.

* These insights can guide improvements in global TB management, particularly for resistant strains.

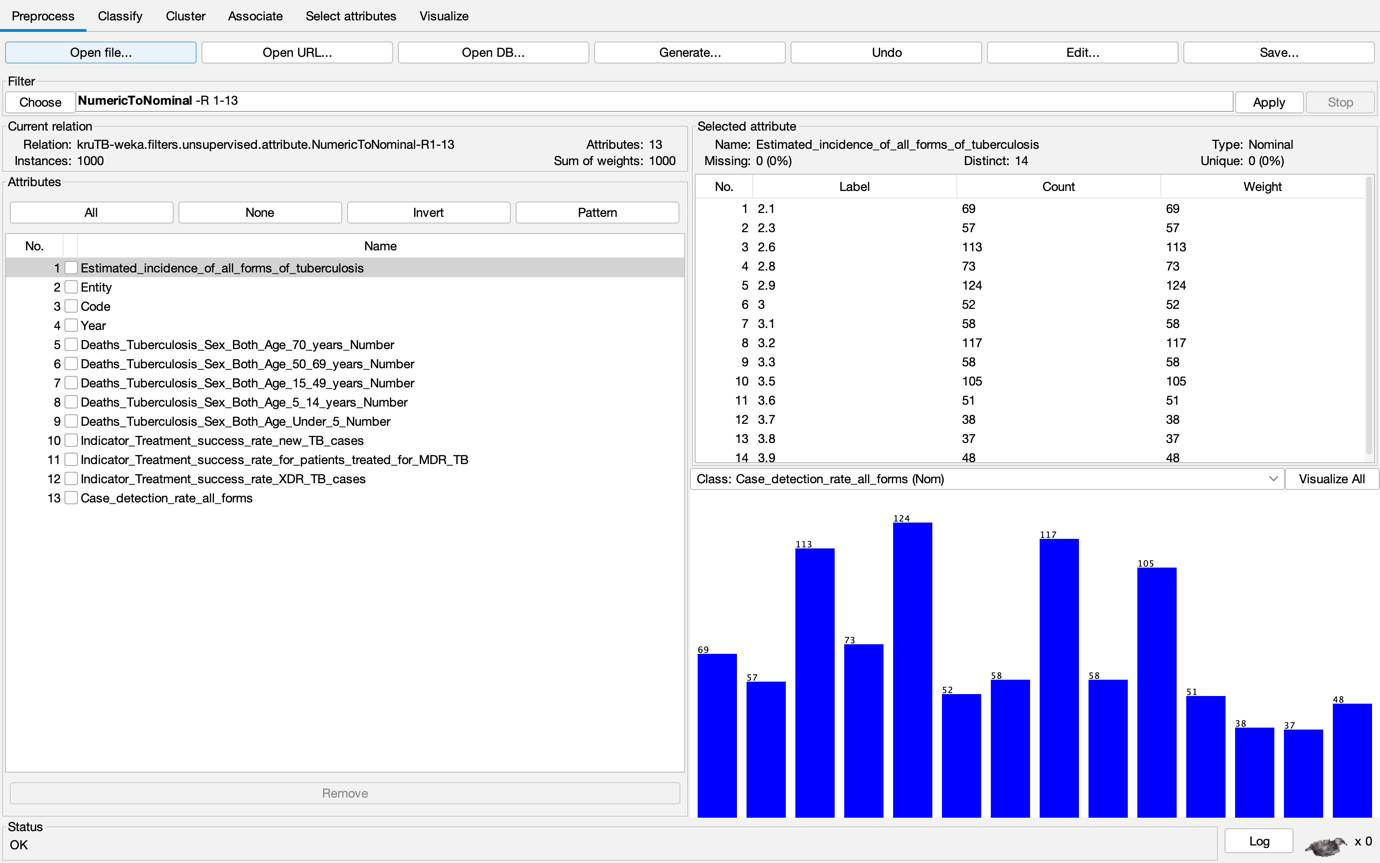
**Insights**: Understanding which treatment types have the highest success rates can inform public health policies on which treatments to prioritize.

### **Predictive Modeling and Classification Using Weka**

I selected the TOP 1000 subset from dbo.TB saved as a CSV file and loaded in to WEKA.

* 1. **Load Dataset into Weka**:

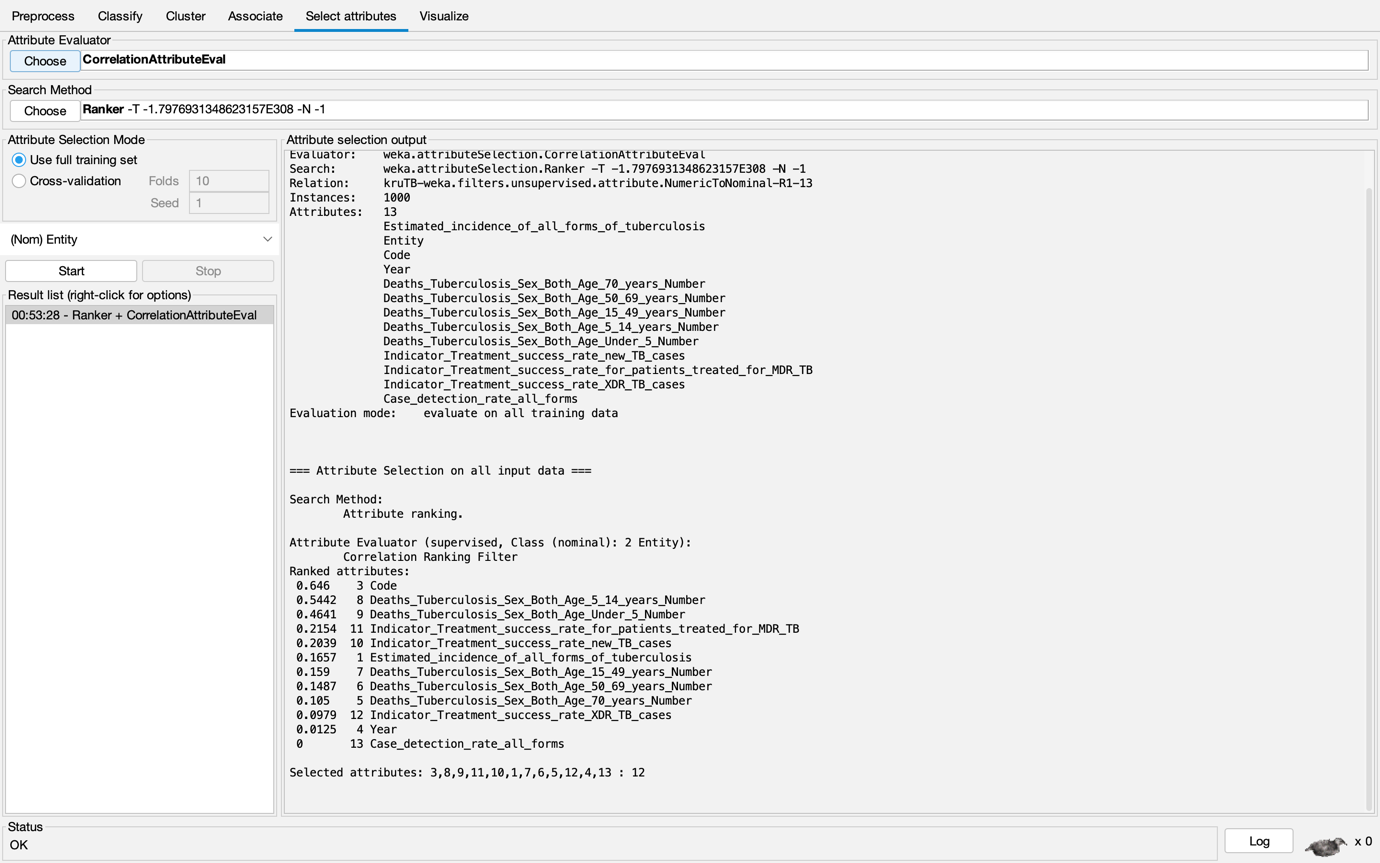
Convert all the variables from numeric to nominal.



#### **Feature Selection**

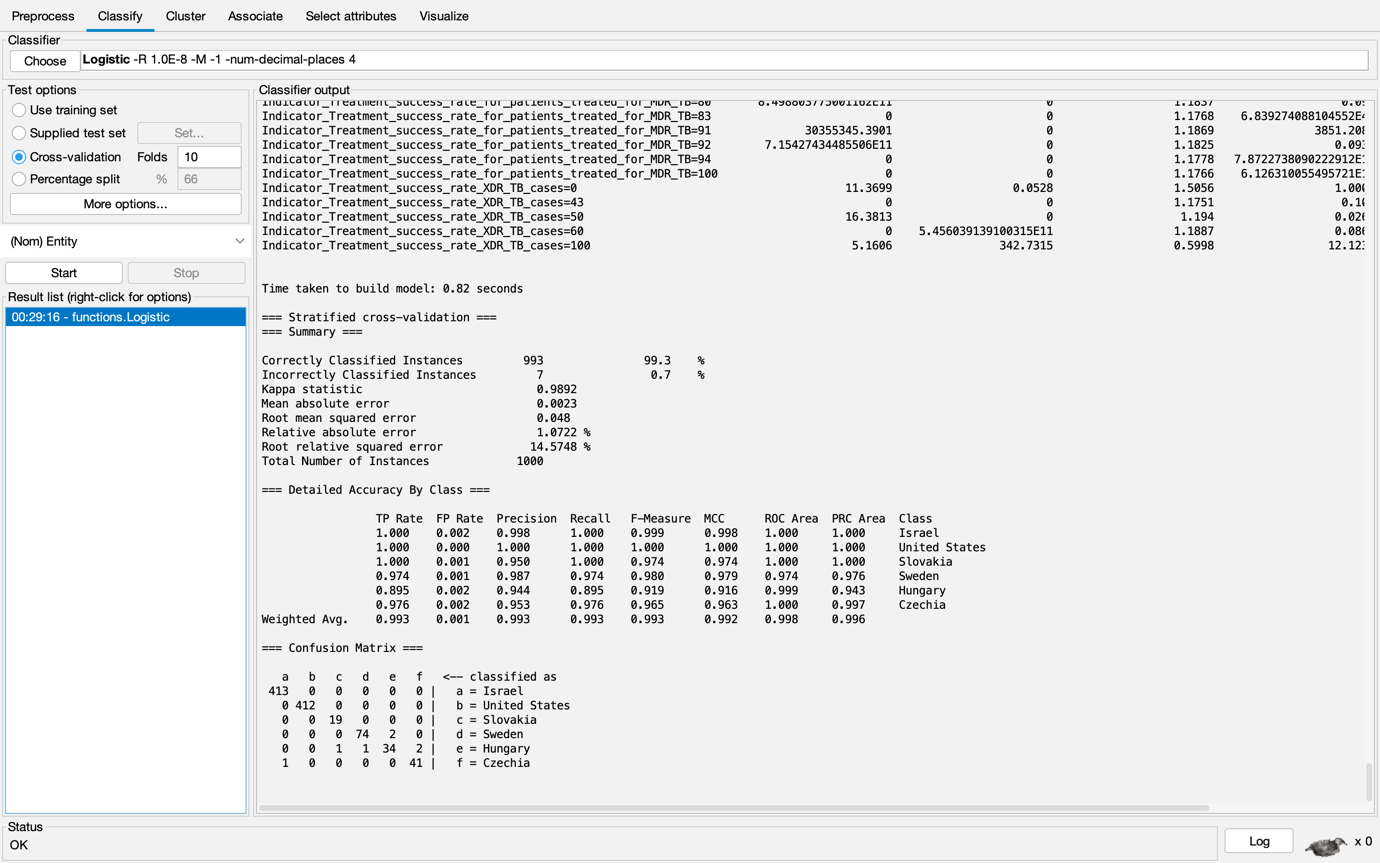
Feature selection was performed to enhance the model's performance, reduce the risk of overfitting, improve interpretability, and decrease computational costs. By using CorrelationAttributeEval with a Ranker search method, the most relevant attributes were identified and selected, ensuring that the final model would be more effective and interpretable. This step is crucial for building a robust model that accurately reflects the underlying patterns in the tuberculosis dataset.

* In the below image, CorrelationAttributeEval was used as the attribute evaluator, combined with a Ranker search method. Here's what the process involved:
* Attributes Evaluated: A total of 13 attributes related to tuberculosis statistics were evaluated. These attributes included various metrics such as the estimated incidence of tuberculosis, the number of deaths by age group and sex, and treatment success rates for different types of tuberculosis.
* Ranking of Attributes: The attributes were ranked based on their correlation with the target variable ("Entity"). The top three attributes with the highest correlation were:
* Country Code (0.646): This attribute had the highest correlation with the target variable, indicating it is a strong predictor.
* Deaths\_Tuberculosis\_Sex\_Both\_Age\_5\_14\_Years\_Number (0.544): This attribute also showed a strong correlation, suggesting that the number of deaths in this age group is an important factor.
* Indicator\_Treatment\_success\_rate\_for\_patients\_treated\_for\_MDR\_TB (0.441): This attribute was also highly ranked, reflecting its importance in predicting tuberculosis outcomes.
* Selected Attributes: Out of the 13 attributes, 12 were selected based on their ranking. This selection process indicates that most of the attributes contribute valuable information for distinguishing between entities. The selection was inclusive, meaning it retained most of the features, which could help in capturing a comprehensive understanding of the data.



* 1. **Model Training**:

**Cross-validation**: Use cross-validation (e.g., 10-fold) to evaluate model performance on the training set.

**LOGISTIC REGRESSION:**

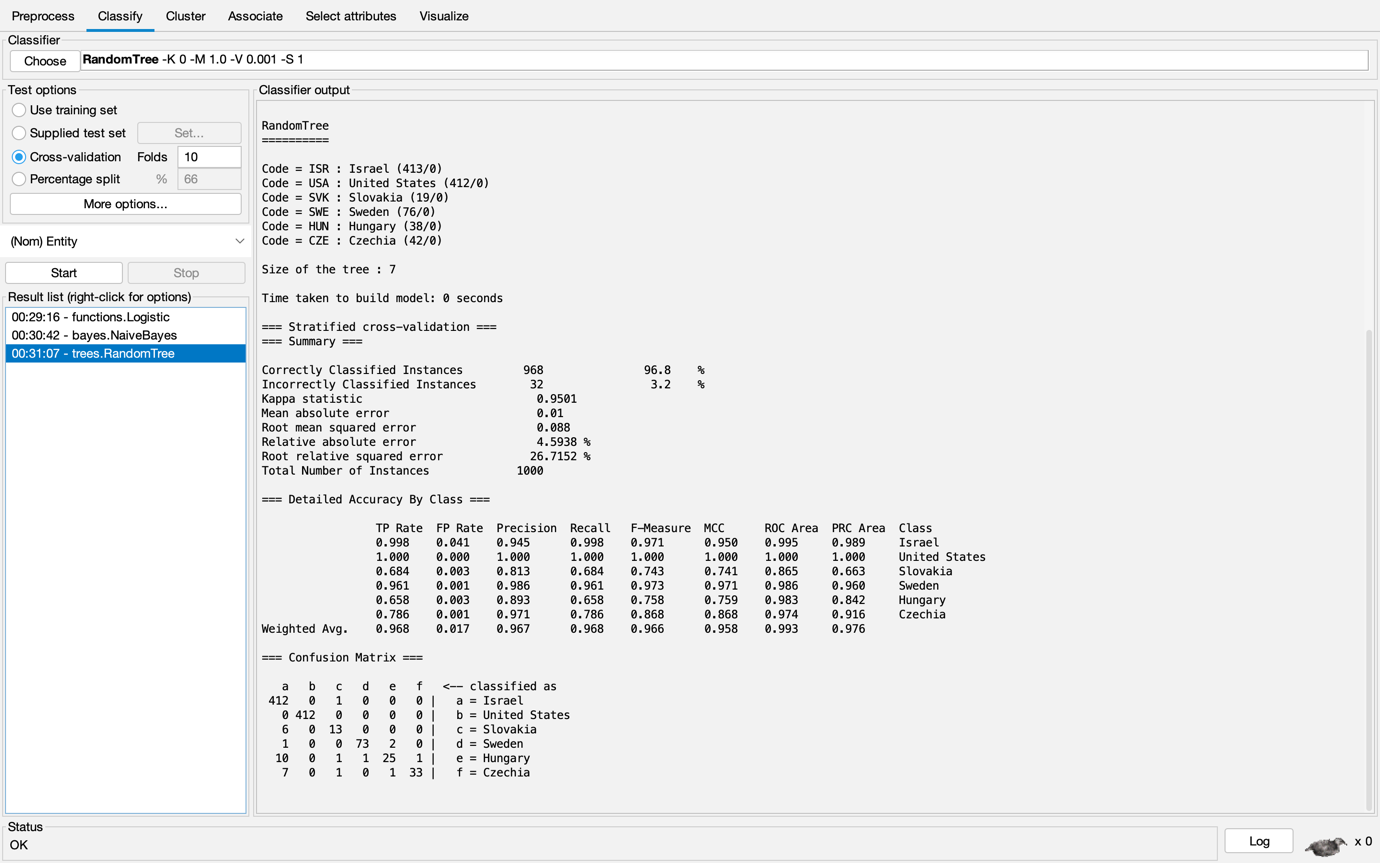
The Logistic Regression model used in this project demonstrated a high level of accuracy and reliability. Specifically, the model achieved an accuracy of 99.3%, correctly classifying 993 out of 1000 instances. The Kappa statistic, which measures the agreement between predicted and actual classifications, was 0.9892, indicating a very strong agreement.

#### **Error Metrics:**

* **Mean Absolute Error (MAE):** 0.0023
* **Root Mean Squared Error (RMSE):** 0.0843

The confusion matrix for this model shows that there were only 7 misclassified instances, indicating strong performance across all classes.

**RANDOM TREE:**



The RandomTree model used in this project achieved an accuracy of 96.8%, correctly classifying 968 out of 1000 instances. The Kappa statistic for this model was 0.9501, indicating strong agreement between the predicted and actual classifications, though slightly lower than that of the Logistic Regression model.

#### **Error Metrics:**

* **Mean Absolute Error (MAE):** 0.01
* **Root Mean Squared Error (RMSE):** 0.0888

The confusion matrix for the RandomTree model shows 32 misclassified instances, reflecting a lower performance in classifying some of the classes compared to the Logistic Regression model.

* 1. **Model Evaluation**:

|  |  |  |
| --- | --- | --- |
| **Metric** | **Logistic Regression** | **RandomTree** |
| **Accuracy** | 99.3% (993/1000) | 96.8% (968/1000) |
| **Kappa Statistic** | 0.9892 | 0.9501 |
| **Mean Absolute Error (MAE)** | 0.0023 | 0.01 |
| **Root Mean Squared Error (RMSE)** | 0.0843 | 0.0888 |
| **Misclassified Instances** | 7 | 32 |
| **True Positive Rate (TP Rate)** | 0.944 - 1.000 (varies by class) | 0.743 - 1.000 (varies by class) |
| **Precision** | 0.944 - 1.000 (varies by class) | 0.743 - 1.000 (varies by class) |
| **Confusion Matrix** | 7 misclassifications | 32 misclassifications |
| **Overall Conclusion** | Higher accuracy and reliability | Lower accuracy, higher error rates |

 **Logistic Regression** outperforms **RandomTree** in terms of accuracy, Kappa statistic, and error metrics, making it the more reliable model for this dataset.

 **RandomTree** shows more misclassifications and slightly higher error rates, indicating it is less effective at accurately classifying instances within this dataset.

**CLUSTER ANALYSIS:**

Cluster analysis, also known as clustering, is an unsupervised machine learning technique for identifying and grouping related data points in large datasets without concern for the specific outcome.

I conducted clustering to compare the results and get insights from the dataset.

Results of a SimpleKMeans clustering algorithm applied to tuberculosis data, dividing the dataset into four distinct clusters. The clusters are based on features like tuberculosis incidence, deaths by age group, and treatment success rates, providing insights into regional differences in these metrics. Each cluster's centroid represents the average characteristics of the data points within that cluster.

**Final Clusters**:

* **Cluster 0**: Contains 286 instances (28.6% of the data)
* **Cluster 1**: Contains 335 instances (33.5% of the data)
* **Cluster 2**: Contains 253 instances (25.3% of the data)
* **Cluster 3**: Contains 126 instances (12.6% of the data)

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

The Logistic Regression classifier was applied to a dataset that had been divided into four clusters. The model achieved a high overall accuracy of 95.2%, with strong performance metrics across all clusters, including a high Kappa statistic (0.9339), indicating that the model was effective at correctly classifying instances into their respective clusters. The confusion matrix reveals minimal misclassification between clusters, further supporting the model's effectiveness.



When clustered model is compared with the non-clustered model under Logistic regression, I have observed the following:

The **non-clustered model** has a higher accuracy (99.3% vs. 95.2%) and a higher Kappa statistic (0.9892 vs. 0.9339) compared to the clustered model. This suggests that the non-clustered model is better at predicting the correct class for each instance.

The **non-clustered model** also has lower MAE and RMSE values, indicating more precise predictions compared to the clustered model. This further supports the conclusion that the non-clustered model performs better overall.

|  |  |  |
| --- | --- | --- |
| **Metric** | **Without Clustering** | **With Clustering** |
| **Accuracy** | 99.3% (993/1000) | 95.2% (952/1000) |
| **Kappa Statistic** | 0.9892 | 0.9339 |
| **Mean Absolute Error (MAE)** | 0.0023 | 0.0238 |
| **Root Mean Squared Error (RMSE)** | 0.0843 | 0.1511 |
| **Misclassified Instances** | 7 | 48 |
| **True Positive Rate (TP Rate)** | 0.979 - 1.000 (varies by class) | 0.905 - 0.979 (varies by cluster) |
| **Precision** | 0.944 - 1.000 (varies by class) | 0.933 - 0.976 (varies by cluster) |
| **Confusion Matrix** | 7 misclassifications | 48 misclassifications |
| **Overall Conclusion** | Higher accuracy and reliability | Lower accuracy, increased complexity |

**Results** - Clustering can introduce additional complexity, as it attempts to group similar data points together before classification. While this can be useful in certain contexts, in this case, it appears to have reduced the overall performance of the model. The clustered model may have struggled to accurately classify data within each cluster, leading to a higher number of misclassifications and lower accuracy.

### **Conclusion**

This project highlights the ongoing global challenge posed by tuberculosis (TB), particularly in low- and middle-income countries where the burden of the disease is most severe. Through comprehensive data analysis, including descriptive, comparative, and predictive methodologies, several key insights were drawn:

* **Global Trends and High-Risk Regions**:

The analysis revealed significant variability in TB incidence rates across different regions, with countries in southern Africa, such as Eswatini and South Africa, showing particularly high rates. This suggests a need for targeted public health interventions in these regions.

* **Age-Specific Vulnerabilities**:

Mortality analysis by age group indicated that older populations, especially those aged 50-69 and 70+, are the most affected by TB. In contrast, younger age groups, particularly those under 5 years, exhibit much lower mortality rates. This insight underscores the need for age-specific health interventions to protect vulnerable populations.

* **Impact of HIV on TB Incidence**:

A strong correlation was found between HIV prevalence and TB incidence, particularly in regions with high HIV infection rates. This emphasizes the importance of integrated healthcare strategies that address both TB and HIV to effectively reduce the burden of these co-epidemics.

* **Treatment Success Rates**:

There is considerable variation in the success rates of TB treatment, especially for multi-drug-resistant (MDR) and extensively drug-resistant (XDR) TB. Some countries, like Benin, have achieved high success rates, while others struggle, particularly with resistant strains. This highlights the need for improved treatment protocols and resources in regions with lower success rates.

* **Predictive Modeling for Future Trends:**
* The predictive models developed, particularly using techniques like Logistic Regression demonstrated high accuracy in forecasting TB incidence and treatment outcomes.
* These models can be crucial tools for public health planning, allowing policymakers to anticipate and mitigate future outbreaks.
* Regional Prioritization: Focus efforts on high-incidence regions, particularly in southern Africa, to curb the spread of TB.
* Targeted Interventions: Develop age-specific interventions, particularly for older populations, who are most vulnerable to TB mortality.
* Strengthening Treatment Protocols: Invest in improving treatment protocols and access to care, especially in regions struggling with drug-resistant TB strains.
* Utilizing Predictive Models: Leverage predictive modeling to guide future public health interventions and optimize resource allocation.

The above results were obtained using the country or region ("Entity") as the target variable, with other factors such as age-specific mortality, treatment success rates, and HIV prevalence used as predictive variables. This approach allowed for a detailed analysis of how different factors contribute to the TB burden across various regions.

|  |  |
| --- | --- |
| **Key Insight** | **Brief Description** |
| **High-Risk Regions** | Southern Africa (e.g., Eswatini, South Africa) shows high TB incidence.  targeted interventions needed. |
| **Age Vulnerabilities** | Older populations (50-69, 70+) most affected; younger groups (under 5)  have lower mortality rates. |
| **HIV and TB Correlation** | Strong link between high HIV prevalence and increased TB incidence.  integrated healthcare strategies crucial. |
| **Treatment Success Rates** | Varied success rates for MDR and XDR TB; some countries excel,  others need improved protocols. |
| **Predictive Modeling** | Logistic Regression models highly accurate; crucial for anticipating and  mitigating future TB outbreaks. |
| **Entity as Target Variable** | Country or region ("Entity") used as target variable,  with other factors as predictors. |

**REFERENCES:**

* Abdool Karim SS, Naidoo K, Grobler A, Padayatchi N, Baxter C, Gray A, et al. Timing of initiation of antiretroviral drugs during tuberculosis therapy. New England Journal of Medicine. 2010;362(8):697–706. [[PubMed](https://pubmed.ncbi.nlm.nih.gov/20181971)]
* <https://www.kaggle.com/datasets/gggupta2012/tuberculosisclassification?resource=download> - Link for Data set.
* <https://www.ncbi.nlm.nih.gov/books/NBK430746/>
* <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4548428/#:~:text=Several%20antibiotic%20regimens%20are%20proposed%20for%20treatment%20of%20MDR%20g,due%20to%20ESBL%2Dproducing%20bacteria>.
* <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7983091/>
* <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10200164/>
* <https://developer.ibm.com/articles/os-weka2/>

<https://www.hain-lifescience.de/en/products/microbiology/mycobacteria/facts-about-resistant-tuberculosis.html>

<https://www.who.int/news-room/fact-sheets/detail/tuberculosis>