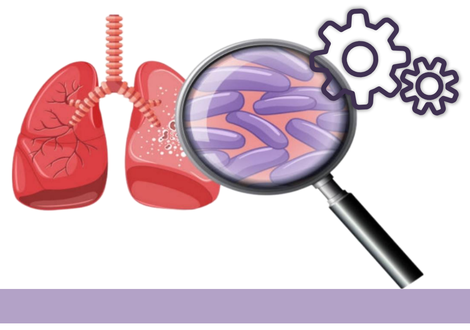
THE APPLICATION OF MACHINE LEARNING ALGORITHMS IN THE PREDICTION OF TUBERCULOSIS DISEASE IN ESWATINI

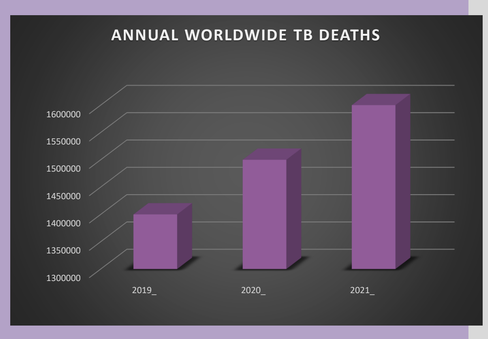


Dlamini Tifeziwe and Fashoto Stephen Gbenga

Department of Computer Science, University of Eswatini, Kwaluseni, Eswatini

INTRODUCTION

Tuberculosis remains one of the leading causes of death worldwide, with more than a million deaths recorded in 2021 alone ("Tuberculosis," 2022), and this infection and mortality rate remains alarmingly high today. The Kingdom of Eswatini is not shielded from the adverse effects of this infectious disease as it is currently considered a Tuberculosis high burdened country(Haumba et al., 2015), garnering an estimated 348 incidences of tuberculosis per 100 000 people in 2021, according to statistics gathered by the World Health Organization World Bank (2021). Despite the advancements in healthcare and the availability of a cure for this disease, Tuberculosis still remains a predominant infectious-disease killer, largely due to the fact that a majority of the TB cases are identified and treated far too late(Owoseye, 2019). There is evidently a need for a system that will effectively and efficiently identify and detect TB in patients early in order to curb the rapid spread of the disease and ultimately plummet the TB related mortality rate in the kingdom of Eswatini. This system can be achieved through the use of machine learning solutions. This paper therefore seeks to predict Tuberculosis disease in patients based on the initial symptoms presented by a patient, utilizing various machine learning algorithms



(World Health Organisation, 2023)

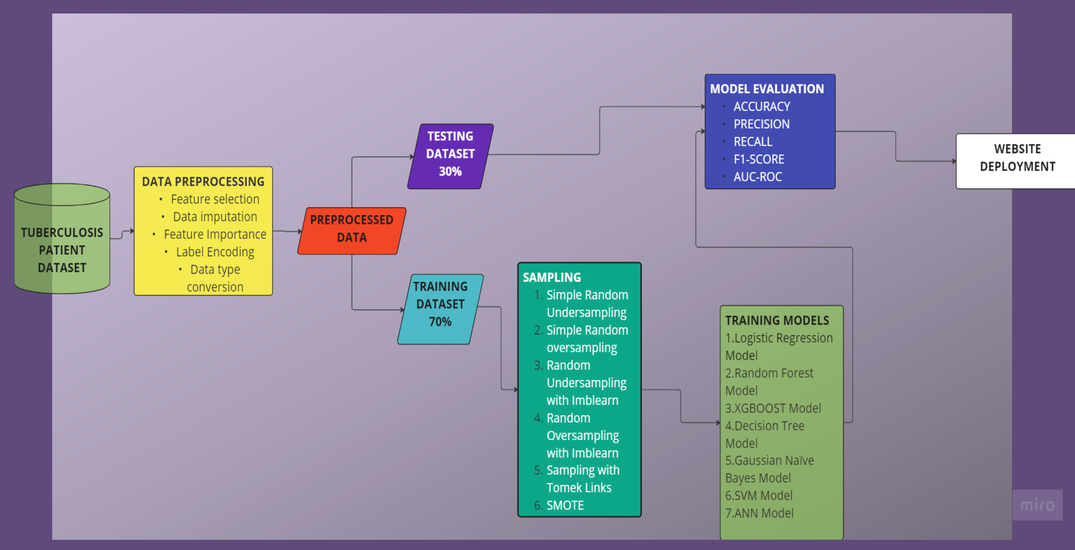
AIM

To employ machine learning techniques to predict Active Tuberculosis disease in patients in Eswatini.

PROBLEM STATEMENT

The mortality rate of Tuberculosis is alarmingly high, and the number of cases continue to show an upward trend in recent years, making TB one of the world's deadliest infectious diseases ("Tuberculosis," 2022), impacting the healthcare of citizens of many countries, including the Kingdom of Eswatini, which has been considered a Tuberculosis high burdened country (Haumba et al., 2015). This is largely owed to the fact that many of these cases are not identified or detected on time, and this delay in diagnosis and treatment ultimately contributes to the severity, risk of mortality and rapid transmission of the disease in the country (Abdu et al., 2020).

TUBERCULOSIS PREDICTION MODEL FRAMEWORK



SAMPLING METHODS

1. Simple Random Undersampling
2. Simple Random oversampling
3. Random Undersampling with Imblearn
4. Random Oversampling with Imblearn
5. Sampling with Tomek Links
6. SMOTE

ML PREDICTION MODELS

1. Logistic Regression Model
2. Random Forest Model
3. XGBOOST Model
4. Decision Tree Model
5. Gaussian Naive Bayes Model
6. Support Vector Machine Model
7. Artificial Neural Network Model

MATERIALS & METHODOLOGY

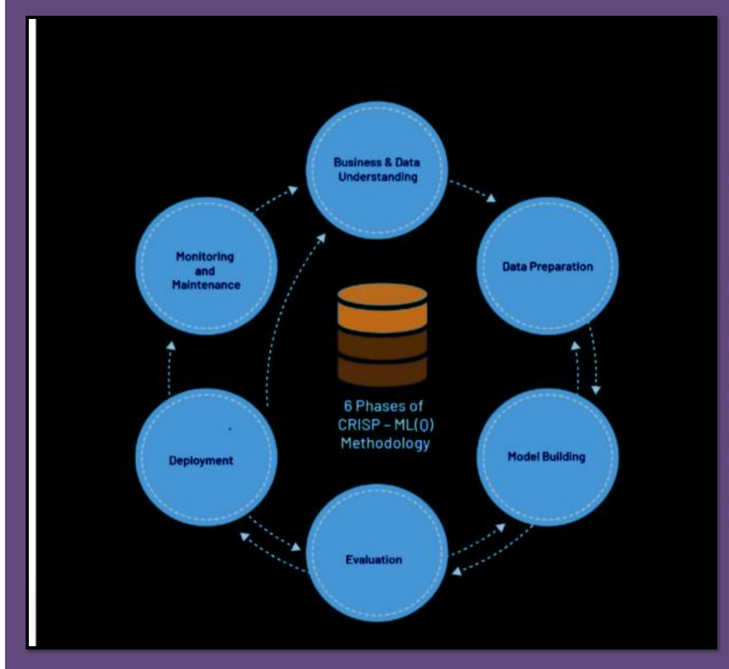
This study aims to employ various machine learning techniques to accurately and efficiently predict the existence of the Tuberculosis disease in the patients in Eswatini, using a local medical Tuberculosis dataset

DATA SOURCE

- Dataset used is a local medical Tuberculosis dataset collected from the national Health Management Information System (HMIS) facility
- contains a total of 84,388 TB patient records, maintained between the years 2018- 2022.
- It contains a total of 19 columns

CRISP-ML(Q) METHODOLOGY

This is a variation of the CRISP-DM methodology, as it utilizes the principles of CRISP-DM, but is modified to the particular requirements of the machine learning applications. It is a framework that highly emphasizes on quality of the model(Studer et. al.,2021).



1. *Business and Data Understanding*
2. *Data Preparation*
3. *Modelling*
4. *Evaluation*
5. *Deployment*
6. *Monitoring and Maintenance*

PRELIMINARY RESULTS

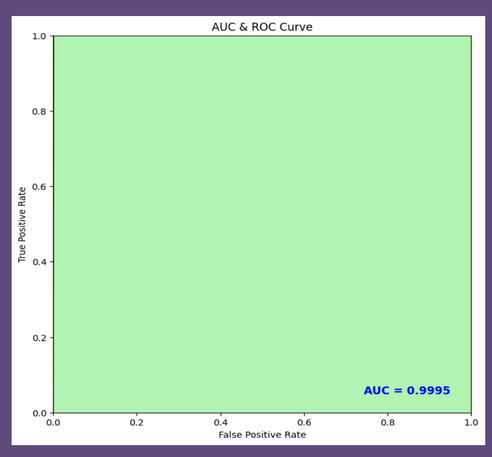
The performance of each model was determined using:

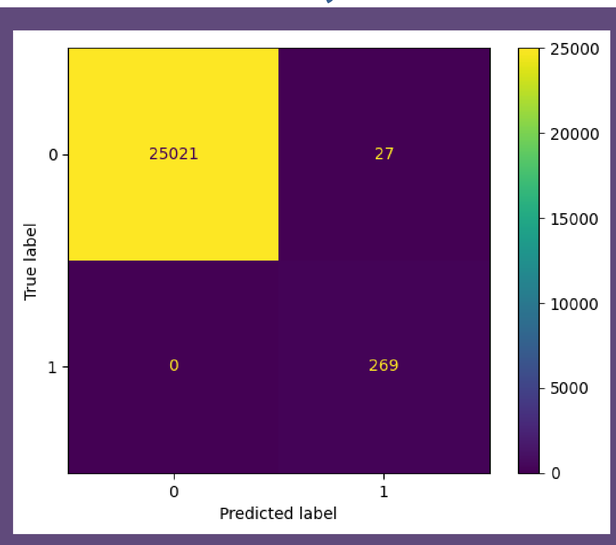
1. **Accuracy** ACCURACY= (TP+TN)/(TP+FP+TN+FN)
2. **Precision** PRECISION=TP/(TP+FP)
3. **Recall** RECALL=TP/(TP+FN)
4. **F-1 score** F1-SCORE=2*(PRECISION*RECALL)/(PRECISION+RECALL)
5. **AUC-ROC score**

In the preliminary results obtained so far, the best performing ML model in classifying the local TB dataset is the XGBOOST model, with the Random Oversampling class imbalance handling technique

XGBOOST+ RANDOM OVERSAMPLING WITH IMBLEARN

```
Accuracy:    0.9988940237784888
Precision:   0.9057239057239057
Recall:      1.0
F1-Score:    0.950530035335689
```





REFERENCES

Haumba, S., Dlamini , T., Calnan, M., Ghazaryan, V., Smith-Arthur, A., Preko, P., & Ehrenkranz, P. (2015). Declining tuberculosis notification trend associated with strengthened TB and expanded HIV care in Swaziland. *Public Health Action*, *2*(5), 103-105. http://dx.doi.org/10.5588/pha.15.0008

Tuberculosis. (2022, October 27).World Health Organisation. https://www.who.int/news-room/factsheets/detail/tuberculosis#:~:text=In%202021%2C%20an%20estimated%2010.6,TB%20is%20curable%20and%20preventable.

Owoseye, A. (2019, August 15). WHO, Global Funds, Stop TB want global action against tuberculosis . Premium Times. https://www.premiumtimesng.com/health/health-news/346763-who-global-funds-stop-tb-want-immediate-action-on-world-leaders-commitment-to-fight-tuberculosis.html?tztc=1

Abdu, A., Balchut, A., Girma, E., & Mebratu, W. (2020). Patient Delay in Initiating Tuberculosis Treatment and Associated Factors in Oromia Special Zone, Amhara Region . *Pulm Med*, 10.1155/2020/6726798

Studer, S., Bui, T., Drescher, C., Hanuschkin, A., Winkler, L., Peters, S., & Müller, K. (2021). Towards CRISP-ML(Q): A Machine Learning Process Model with Quality Assurance Methodology. *Mach. Learn. Knowl. Extr*, (3), 392-413. https://doi.org/10.3390/make3020020

World Health Organisation . (2023, April 12). Tuberculosis . World Health Organisation . https://www.who.int/news-room/fact-sheets/detail/tuberculosis#:~:text=Tuberculosis%20mostly%20affects%20adults%20in,%2D%20and%20middle%2Dincome%20countries