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 (Naamba et al., 2015), garnering an estimated 346 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 (Swisse, 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.

AIM

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

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 fit 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

TUBERCULOSIS PREDICTION MODEL FRAMEWORK

PRELIMINARY RESULTS

The performance of each model was determined using:
1. Accuracy
2. Precision
3. Recall
4. F1 score
5. AUC-ROC score

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

REFERENCES

https://doi.org/10.3390/PhA.5010008


