Tuberculosis (TB) Diagnosis Using Logistic Regression

# 1. Introduction

Tuberculosis (TB) is a serious infectious disease, and early diagnosis is crucial for effective treatment and control. In this project, we developed a machine learning model using Logistic Regression to predict the likelihood of TB based on clinical and demographic features. The dataset initially contained high class imbalance, which was addressed using SMOTE (Synthetic Minority Oversampling Technique).

# 2. Dataset Overview

- Number of samples (original): 492  
- Target variable: TB\_Diagnosis (0 = No TB, 1 = TB)  
- Class distribution (original): 0 (95.5%), 1 (4.5%)  
- Balanced distribution after SMOTE: 0 (50%), 1 (50%)

# 3. Methodology

1. Data preprocessing: Renamed columns, created derived target, applied SMOTE.  
2. Model: Logistic Regression.  
3. Metrics: Confusion Matrix, Accuracy, Precision, Recall, F1-score, ROC Curve, Precision-Recall Curve.

# 4. Results

Confusion Matrix:  
[[90 4]  
 [ 0 94]]  
  
Classification Report:  
Precision (No TB: 1.00, TB: 0.96)  
Recall (No TB: 0.96, TB: 1.00)  
F1-score (0.98)  
Accuracy: 0.98  
ROC AUC: 0.99  
Average Precision: 0.99

# 5. Discussion

The model achieved 98% accuracy with strong precision and recall. Recall = 1.00 for TB means nearly all TB-positive patients were identified. Precision = 0.96 minimizes false positives. Balancing the dataset with SMOTE was essential.

# 6. Conclusion

Logistic Regression with SMOTE-balanced data effectively predicts TB diagnosis, achieving high accuracy and recall. This model can support healthcare professionals in TB screening and early detection.