UTM UNIVERSITI TEKNOLOGI MALAYSIA

FACULTY OF COMPUTING
UTM Johor Bahru

SECB3203-01(PROGRAMMING FOR BIOINFORMATIC)

Semester 01 2025/2026
Section 01

Progress 3

Faculty of Computing

Dataset:
https://www.kaggle.com/datasets/miadul/tuberculosis-x-ray-dataset-synthetic

Project Title:
Tuberculosis Disease Classification Using Synthetic Chest X-Ray Images and Machine Learning Techniques

---

Student Name	Matric Number
NATIJAH BINTI HUDA	A23CS0142
NUR IMAN BINTI MOHAMAD ZAHARI	A23CS0158
LIANA DARWISYAH BINTI AZMAN	A23CS0102

LECTURER'S NAME : DR. SEAH CHOON SEN

SUBMISSION DATE :

Table of contents

1.0 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is performed to understand the distribution, relationships, and patterns within the synthetic tuberculosis dataset. Pandas and NumPy are used to analyze patient attributes and identify differences between tuberculosis and normal cases.

1.1 Descriptive Statistics

Purpose

To summarize the basic characteristics of numerical variables such as Age, Cough_Severity, and Breathlessness.

Method

Descriptive statistics are generated using Pandas to compute:

- Mean
- Median
- Standard deviation
- Minimum and maximum values

Explanation

This analysis provides an overview of patient demographics and symptom severity levels. It helps identify data spread, potential outliers, and whether normalization is required before model development.

1.2 Basic Grouping

Purpose

To compare symptoms between Tuberculosis (TB) and Normal patient groups.

Method

The dataset is grouped using the Class attribute (TB / Normal) with Pandas groupby() to analyze:

- Average cough severity
- Presence of fever and night sweats

- Differences in smoking history

Explanation

Grouping helps identify symptom patterns that differ between TB and normal cases. This supports understanding which features may contribute most to classification.

1.3 Analysis of Variance (ANOVA)

Purpose

To determine whether symptom differences between TB and Normal groups are statistically significant.

Method

ANOVA is applied to numerical features such as:

- Age
- Cough_Severity
- Breathlessness
- Fatigue

Explanation

ANOVA evaluates whether the mean values of these features differ significantly between classes. Features with significant variation are considered important for model development.

1.4 Correlation Analysis

Purpose

To identify relationships between patient symptoms.

Method

Correlation coefficients are calculated for numerical features using Pandas correlation functions.

Explanation

Correlation analysis reveals relationships such as:

- Cough severity vs breathlessness
- Fever vs night sweats
- Fatigue vs weight loss

Highly correlated features may indicate redundancy, while weak correlations suggest independent predictors.

2. Summary of Findings

Exploratory Data Analysis reveals meaningful patterns in patient symptoms and demographic data. Differences between tuberculosis and normal cases are observed through grouping and ANOVA, while correlation analysis highlights relationships between clinical features. These findings guide feature selection and model development in later project stages.