🚬 Tobacco Use & Mortality Prediction - Machine Learning Report

1. Introduction

This project analyzes the relationship between tobacco use and mortality rates using Machine Learning. It predicts mortality rates based on tobacco-related factors.

2. Data Sources

The datasets used in this project:

- admissions.csv → Hospital admission records related to tobacco-related diseases
- fatalities.csv → Mortality data from tobacco-related illnesses
- metrics.csv → Economic and tobacco pricing data
- prescriptions.csv → Records of smoking cessation prescriptions
- smokers.csv → Smoking prevalence across different age groups

3. Data Preprocessing

- Missing Values Handling:
  - Numerical Features → Filled with median values
  - Categorical Features → Encoded using Label Encoding
- Merging Data: Combined all datasets using Year as the key
- Feature Selection: Selected top features affecting mortality

4. Exploratory Data Analysis (EDA)

- Histograms: Show distribution of smoking prevalence
- Scatter Plots: Show the relationship between smoking rates and mortality
- Heatmaps: Show correlation between features

5. Model Selection & Training

- Algorithm Used: RandomForestRegressor
- Train-Test Split: 80% training, 20% testing
- Hyperparameter Tuning: Used n_estimators=100 for best performance

6. Model Evaluation

- Mean Absolute Error (MAE): 100.00

• Mean Squared Error (MSE): 138934.15

• R^2 Score: 0.9998

7. Streamlit Web App

• User Inputs: Year, ICD10 Code, Diagnosis Type, Smoking Data

• Predicted Output: Estimated mortality based on inputs

8. Findings & Insights

• Higher tobacco prices = Lower smoking prevalence

• Older populations have higher mortality risks

• Government policies on taxation impact smoking rates significantly

9. Future Improvements

• Implement deep learning models (LSTM, XGBoost)

• Deploy model on AWS / Streamlit Cloud