TOBACCO USE AND MORTALITY ANALYSIS USING STATISTICAL MODELING

AARYA SHUKLA

UNIFIED MENTOR

TOBACCO USE AND MORTALITY ANALYSIS USING STATISTICAL MODELING

Abstract

Tobacco use is a leading cause of preventable deaths worldwide. This project explores the relationship between tobacco consumption and mortality rates using a data-driven approach. Through exploratory data analysis (EDA), feature engineering, and predictive modeling, this analysis uncovers patterns in tobacco-related mortality and aims to support public health policy formulation with evidence-based insights.

1. Introduction

Tobacco-related illnesses account for millions of deaths each year, making it a significant global health crisis. While awareness campaigns and regulations exist, understanding historical data trends can highlight how tobacco usage impacts mortality over time and across regions. This project leverages structured datasets to build statistical models that explain and forecast mortality rates based on tobacco consumption data.

2. Dataset Description

While the exact dataset filename is not specified in the notebook, based on the code, it appears to include the following attributes:

- **Year**: The year of record
- **Country**: Country name
- **Tobacco_Use**: Percentage of tobacco users
- **Mortality_Rate**: Death rate due to tobacco use
- Other derived or calculated features such as averages or log-transformed data

Data Preprocessing Steps:

- Handled missing values
- Normalized or scaled features where needed
- Checked for correlation between features and target variable

3. Methodology

The following steps were carried out:

- **Loading and Cleaning**:
  - Read CSV data
  - Dropped or filled missing values
  - Filtered relevant columns and sorted chronologically
- **Exploratory Data Analysis (EDA)**:
  - Visualized trends in tobacco use and mortality over time
  - Compared across countries
  - Plotted correlation matrix to understand relationships
- **Modeling Techniques**:
  - Implemented **Linear Regression** for prediction

- Applied **Logarithmic Transformation** to tackle skewed distributions
- Evaluated with metrics like R^2, MAE, and RMSE

4. Exploratory Data Analysis

Insights from EDA:

- **Positive Correlation**: Strong positive correlation between Tobacco_Use and Mortality_Rate.
- **Country-wise Variation**: Certain countries consistently show higher mortality despite lower use—suggesting healthcare or demographic influence.
- **Year-wise Trends**: Gradual decline in tobacco use in many regions, but mortality remains high, possibly due to long-term effects.

Visualizations Used:

- Line plots of mortality rate vs. year
- Bar plots of tobacco use by country
- Heatmaps of correlation matrices

5. Modeling and Predictions

Model Used: Scikit-learn's LinearRegression

- **Features**: Tobacco_Use, Year, Country (encoded), transformed mortality values
- **Target**: Mortality_Rate

- **Model Performance**:
  - R^2 Score: High, indicating good fit
  - RMSE and MAE: Acceptable levels, confirming linearity

Results:

- Model was able to predict mortality rates with reasonable accuracy using tobacco use levels.
- Log-transformation improved model interpretability and performance in cases of skew.

6. Results and Discussion

- **Prediction Accuracy**: Linear regression works well for aggregated national-level data, though further improvements are possible with non-linear models.
- **Interpretability**: The model highlights the strong influence of tobacco use on mortality but may not account for confounding factors like income, healthcare, or environment.

7. Conclusion

This study reinforces the established public health concern: higher tobacco use leads to increased mortality. Modeling this relationship quantitatively provides a foundation for:

- **Policy evaluation** over time
- **Forecasting** future trends in mortality

- **Identifying high-risk regions** or populations

8. Future Work

Apply **Random Forests** or **XGBoost** for non-linear modeling

- Include other variables like GDP, healthcare access, and age demographics
- Perform **time-series forecasting** (e.g., ARIMA or LSTM)

9. References

- WHO Tobacco Fact Sheets
- Scikit-learn Documentation
- Python Libraries: Pandas, Seaborn, Scikit-learn, Matplotlib