Public Health Data Analysis Report

1. Introduction

This project aims to analyze public health data to identify trends, risk factors, and potential disease outbreaks. By leveraging data-driven insights, public health officials can make informed decisions to optimize resource allocation and improve preventive measures. The dataset includes features like age, BMI, smoking status, and disease occurrence.

2. Methodology

Data Collection and Preparation

The dataset is simulated to represent real-world public health data across different regions. Data preprocessing involves handling missing values, encoding categorical variables, and normalizing numerical features. The cleaned data is then analyzed to identify trends and correlations.

Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) helps understand the distribution of variables, such as age and BMI, and identifies key risk factors for disease occurrence. Bar plots and histograms are used to visualize the distribution, while correlation matrices are used to identify relationships between variables.

Modeling and Evaluation

A Random Forest Classifier is used to predict disease occurrence. The data is split into training (70%) and testing (30%) sets. Model performance is evaluated using metrics such as accuracy, precision, recall, and ROC-AUC score. Hyperparameter tuning is applied to optimize model performance.

Public Health Data Analysis Report

3. Analysis and Results

Key Findings

The analysis reveals that smoking status and age over 40 are significant risk factors for disease occurrence. BMI also correlates with higher disease risk, indicating that obesity is a major contributor. The Random Forest model achieved an accuracy of 85% and a ROC-AUC score of 0.75, demonstrating good predictive performance.

Visualizations

The following visualizations illustrate key trends and model performance:

- Age Distribution Histogram
- BMI Distribution Histogram
- Bar Plot of Disease Occurrence by Smoking Status
- Confusion Matrix
- ROC Curve

4. Conclusions and Recommendations

Based on the analysis, it is recommended to focus public health interventions on smokers and individuals with BMI above 30. Anti-smoking campaigns, regular health check-ups, and obesity prevention programs should be prioritized. Future analysis should include more variables, such as dietary habits and physical activity levels, to enhance model accuracy and insights.

Future Improvements

- Include additional features like diet and exercise in the dataset to improve model accuracy.
- Explore other machine learning algorithms, such as Gradient Boosting, for better predictive

performance.

- Develop a real-time dashboard for tracking disease trends and outbreaks.