THE

Proposal

Jun 22, 2025

NYC Mortality Insights

Presented by:

Christian De La Torre
Araceli Castelan
Jonathan Suconota
Kevin Sookram

01 EXECUTIVE SUMMARY

Overview

NYC Mortality Insights is a data-driven project aimed at classifying causes of death in New York City into high and low mortality categories based on year, sex, race/ethnicity, and cause of death. By leveraging public health data and machine learning classification models, the project seeks to identify health disparities and significant mortality trends within NYC's diverse population.

Our approach involves creating a binary classification target based on the median Age Adjusted Death Rate, followed by extensive exploratory data analysis, feature selection, and modeling using logistic regression, decision trees, SVM, KNN, and ensemble methods. The results will be evaluated and visualized with ROC-AUC curves and relevant metrics.

Key features and capabilities

- Systematic data exploration of NYC's leading causes of death dataset
- Advanced feature engineering including label and one-hot encoding
- Feature selection using Chi-Squared tests, Pearson correlation, and RFE
- Multiple classification models and ensemble methods for robust prediction
- Comparative evaluation using accuracy, recall, precision, F1 score, and ROC-AUC graphs

Benefits

This project addresses the critical need to understand mortality disparities in NYC, enabling public health officials and policymakers to target interventions effectively. By classifying mortality risk based on demographic and cause-of-death factors, we provide actionable insights for resource allocation and health equity efforts. Our analytical framework can be extended to other datasets and cities, enhancing broader public health surveillance.

02

PROBLEM

Public health depends on understanding mortality patterns. NYC's diverse population experiences varied mortality rates, but relations between death rates and demographics such as gender and race/ethnicity are not clearly classified. This limits targeted interventions.

SOLUTION

Classify causes of death into high and low mortality categories by demographic and cause variables, using machine learning classification models trained on a processed dataset. This systematic approach allows identification of significant health disparities.

03

Approach

Data Exploration

- Find out if any variable is highly correlated with a new response variable *High Mortality*.

Data Preparation

- Label encoding for *Sex*
- One-hot encoding for *Leading Cause of Death* and *Race Ethnicity*

Feature Selection

- Chi-Squared Test
- Pearson Correlation
- Wrapper Method: Recursive Feature Elimination (RFE)

Models

- Logistic Regression
- Decision Tree Classifier
- Support Vector Machine (SVM)
- K-Nearest Neighbors (KNN)
- Two ensemble methods

Graphics

- We plan to generate ROC-AUC for each model and create graphs to compare accuracy, recall, precision, and F1 score.

04

TEAM

- **Christian De La Torre** – Experienced in exploratory data analysis and data preparation, ensuring data integrity and meaningful initial insights.
- **Araceli Castelan** – Skilled in statistical feature selection and dimensionality reduction techniques.
- **Jonathan Suconota** – Expert in machine learning model creation, evaluation, and comparative analysis.
- **Kevin Sookram** – Strong background in project revision, quality assurance, and professional presentation.