Flu Shot Learning: Predicting Vaccination Patterns

Project Overview

This project focuses on building machine learning models to predict whether individuals received the H1N1 and seasonal flu vaccines. Using demographic data, social factors, personal beliefs, and behavioral indicators, the analysis aims to uncover the drivers behind vaccination decisions. By identifying these key factors, the project contributes to addressing an important public health challenge: improving the effectiveness of future vaccination campaigns

Project Goals

- Develop classification models to predict two target variables:
    - **H1N1 vaccine —** whether an individual received the H1N1 vaccine
    - **Seasonal flu vaccine —** whether an individual received the seasonal flu vaccine
- Identify and analyze the factors that most strongly influence vaccination decisions
- Create clear, engaging visualizations to communicate insights
- Deploy a simple dashboard showcasing models, predictions, and findings

---

Data Description

Dataset Source

The dataset comes from the **National 2009 H1N1 Flu Survey (NHFS)**, conducted by the CDC between **October 2009 and June 2010** during the H1N1 “swine flu” pandemic.

Features Overview

The dataset includes **36 input features**, covering:

- **Demographics**: age, income, education, etc.
- **Social indicators**: geographic region, household composition
- **Opinions & perceptions**: attitudes toward vaccines, perceived risks
- **Behaviors**: preventive health measures and health-seeking behavior

Target Variables

- h1n1_vaccine → 0 = No, 1 = Yes
- seasonal_vaccine → 0 = No, 1 = Yes

Project Structure

Sprint 1: Exploratory Data Analysis (3 weeks)

- Clean and preprocess the dataset
- Handle missing values and categorical variables
- Visualize distributions, patterns, and correlations
- Deliverable → **EDA report with visualizations and initial insights**

Sprint 2: Model Development (3 weeks)

- Feature engineering and selection
- Train baseline classification models for both targets
- Apply cross-validation and hyperparameter tuning
- Evaluate using **ROC AUC** metric
- Deliverable → **Trained models with documented performance metrics**

Sprint 3: Insights & Deployment (3 weeks)

- Analyze feature importance and model explanations
- Build a **Streamlit dashboard** to visualize predictions and insights
- Document findings and recommendations for public health strategies
- Deliverable → **Interactive dashboard and final presentation**

Technical Requirements

Technologies

- Python (pandas, scikit-learn, matplotlib, seaborn)
- Jupyter Notebooks for EDA and documentation
- GitHub for version control and collaboration
- Streamlit for dashboard deployment

Performance Metric

- **ROC AUC** (area under the Receiver Operating Characteristic curve)
- Final score = Average ROC AUC across both prediction targets

Resources

Dataset Files

- training_set.csv: Training data with labels
- test_set.csv: Test data for predictions