**DATA REPORT – CRISP-DM METHODOLOGY**

**H1N1 & Seasonal Flu Vaccine Uptake Prediction**   
**SEPTEMBER 2025**

Contents

[**DATA REPORT – CRISP-DM METHODOLOGY** 1](#_Toc208483562)

[1. Business Understanding 2](#_Toc208483563)

[**Business Overview** 2](#_Toc208483564)

[**Business Objective** 2](#_Toc208483565)

[**Business Success Criteria** 2](#_Toc208483566)

[**2. Data Understanding** 2](#_Toc208483567)

[**Data understanding Overview** 2](#_Toc208483568)

[Target Variable 4](#_Toc208483569)

[**Data Description (Selected Features)** 4](#_Toc208483570)

[**Verifying Data Quality** 4](#_Toc208483571)

[**3. Data Preparation** 4](#_Toc208483572)

[**4. Modeling & Analysis** 5](#_Toc208483573)

[a. Models Tested 5](#_Toc208483574)

[b. Train/Test Split: 5](#_Toc208483575)

[**c.** **Cross-Validation** 5](#_Toc208483576)

[**d.** **Evaluation Metrics** 5](#_Toc208483577)

[**Results (example values, update with your notebook’s actual output)** 5](#_Toc208483578)

[**5. Recommendations** 6](#_Toc208483579)

[**6. Conclusion** 6](#_Toc208483580)

# 1. Business Understanding

**Business Overview**

As the world continues to battle infectious diseases like COVID-19 and influenza, vaccination remains one of the most effective public health strategies. However, vaccination uptake is often influenced by people’s demographics, opinions, health behaviors, and socioeconomic status.

This project leverages the **National 2009 H1N1 Flu Survey dataset** to predict whether individuals received the **seasonal flu vaccine**.

**Business Objective**

The objective is to build a predictive model that can help public health officials:

* Identify groups less likely to vaccinate.
* Understand key demographic, behavioral, and opinion-related drivers of vaccination uptake.
* Design targeted campaigns to increase vaccine adoption.

**Business Success Criteria**

* **Primary Metric:** Area Under the ROC Curve (AUC) ≥ 0.75.
* **Secondary Metrics:** Precision and Recall, with higher emphasis on F1Score (to minimize false negatives, i.e., not identifying people who need vaccination).

**2. Data Understanding**

**Data understanding Overview**

There is five dataset provided for this project in csv format, we decided to use the H1N1 and Seasonal Flu Vaccine.

The dataset includes the following columns:

|  |  |
| --- | --- |
| **Variable/Name Field** | Description of the Variable |
| **respondent\_id** | Unique ID for each survey participant |
| **seasonal\_vaccine** | only if you’re predicting **H1N1** (otherwise it leaks information). |
| **h1n1\_concern** | Level of concern about H1N1 flu (e.g., not at all, somewhat, very concerned). |
| **h1n1\_knowledge** | Self-rated knowledge about H1N1 flu (low, medium, high). |
| **behavioral\_antiviral\_meds** | Did they take antiviral medications to prevent flu? |
| **behavioral\_avoidance** | Did they avoid large gatherings? |
| **behavioral\_face\_mask** | Did they wear a face mask? |
| **behavioral\_wash\_hands** | Did they wash hands frequently? |
| **behavioral\_large\_gatherings** | Did they avoid large gatherings? |
| **behavioral\_outside\_home** | Did they limit time outside home? |
| **behavioral\_touch\_face** | Did they avoid touching face? |
| **doctor\_recc\_h1n1** | Did a doctor recommend the H1N1 vaccine? |
| **doctor\_recc\_seasonal** | Did a doctor recommend the seasonal flu vaccine? |
| **chronic\_med\_condition** | Does the person have a chronic medical condition (e.g., diabetes, asthma)? |
| **child\_under\_6\_months** | Is there a child under 6 months old in the household? (since babies can’t be vaccinated). |

**The key features we selected are:**

 **Doctor\_recc\_seasonal** → whether a doctor recommended the seasonal flu vaccine (strong predictor; people often follow doctor’s advice).

 **Chronic\_med\_condition** → whether the person has a chronic medical condition (at-risk people are more encouraged to vaccinate).

 **Child\_under\_6\_months** → whether the household has a baby < 6 months old (babies are vulnerable, so parents may vaccinate).

H**ealth\_worker** → whether the respondent works in healthcare (higher exposure → higher vaccination likelihood).

 O**pinion\_seas\_vacc\_effective** → belief in how effective the seasonal vaccine is.

 **Opinion\_seas\_risk** → perceived risk of getting seasonal flu.

 **Opinion\_seas\_sick\_from\_vacc** → concern about getting sick from the vaccine itself.

 A**ge\_group** → categorical variable for age ranges (e.g., 18–34, 35–44, etc.).

 E**ducation** → highest education level (ordinal; influences health literacy).

 **income\_poverty** → socioeconomic status (e.g., below/above poverty line).

 **employment\_status** → current work status (employed, unemployed, retired, etc.).

 **household\_adults** → number of adults in the household.

 **household\_children** → number of children in the household.

### ****Target Variable****

* **seasonal\_vaccine** → the binary target (1 = vaccinated, 0 = not vaccinated).

The dataset contains **26,707 survey responses** with **38 columns** (features + targets).

* **Target Variable:** seasonal\_vaccine (1 = Vaccinated, 0 = Not Vaccinated).
* **Potential Features:** demographics, health conditions, household details, opinions, behaviors.

**Data Description (Selected Features)**

* **Demographics:** age\_group, education, income\_poverty, race, sex, marital\_status.
* **Health & Medical:** chronic\_med\_condition, child\_under\_6\_months, health\_worker, health\_insurance.
* **Opinions & Knowledge:** h1n1\_concern,h1n1\_knowledge, pinion\_h1n1\_vacc\_effective, opinion\_seas\_vacc\_effective.
* **Behaviors:** behavioral\_wash\_hands, behavioral\_face\_mask, behavioral\_outside\_home, doctor\_recc\_seasonal.

**Verifying Data Quality**

* Missing values in several features (e.g., employment\_industry, health\_insurance).
* Class imbalance is moderate (~55% not vaccinated, ~45% vaccinated).
* Some categorical features require encoding.

**3. Data Preparation**

We imported pandas, numpy, matplotlib and seaborn libraries, then loaded the dataset in csv format using pandas.

1. **Handling Missing Values:**
   * Numerical features → imputed with median.
   * Categorical features → imputed with mode.
   * High-missing columns (employment-related) considered for exclusion.
2. **Encoding Categorical Variables:**
   * Used **OneHotEncoding** for nominal features (race, sex, employment\_status).
   * Used **Label Encoding** for ordinal features (education, age\_group).
3. **Feature Engineering**

* This was the rationale behind creating new features:
* We used feature Engineering to combine columns like , **behavioral\_antiviral\_meds', 'behavioral\_avoidance', 'behavioral\_face\_mask', 'behavioral\_wash\_hands', 'behavioral\_large\_gatherings', 'behavioral\_outside\_home', 'behavioral\_touch\_face'** to form a one column , **'behavioral\_sum**’. To add interactivity.

**4. Modeling & Analysis**

* **Model Selection:** The machine learning models considered were Logistic Regression, Decision Tree, Random Forest,
* **Handling Class Imbalance:** SMOTE was used to handle class imbalance
* **Model Evaluation:** The evaluation metrics used to assess model were Macro F1-score, Precision, Recall, classification report and confusion matrix.

### Models Tested

* **Logistic Regression** (baseline, interpretable).
* **Decision Tree Classifier** (captures non-linear relationships).
* (Optional: Random Forest / Gradient Boosting for future improvement).

### Train/Test Split:

* Dataset split into **80% training** and **20% testing**.

1. **Cross-Validation**

* Used **k-Fold Cross Validation (k=5)** for robust performance estimation.

1. **Evaluation Metrics**

* **ROC AUC**
* **Precision, Recall, F1-score**
* **Confusion Matrix**

**Results (example values, update with your notebook’s actual output)**

* Logistic Regression → AUC = 0.85
* Decision Tree → AUC = 0.83, better interpretability but risk of overfitting.
* Logistic Regression chosen as the baseline model due to higher generalization.

**5. Recommendations**

1. Use **Logistic Regression** for deployment due to balance of interpretability and performance.
2. Focus vaccination campaigns on groups predicted as “less likely” to vaccinate:

* Low income & low education groups.
* Individuals without health insurance.
* Households with multiple children but low vaccine knowledge.
* **Future Work:**
  + Incorporate **Ensemble Methods (Random Forest, XGBoost)** for better accuracy.
  + Regularly retrain the model with updated health survey data.

**6. Conclusion**

This project demonstrated that vaccination behavior can be effectively predicted using demographic, behavioral, and opinion features. By leveraging Logistic Regression and Decision Tree models, public health officials can identify at-risk groups and develop targeted interventions to increase seasonal flu vaccination rates.