**H1N1 PREDICTIVE ANALYSIS: MACHINE LEARNING FUNDAMENTALS**

Mary Immaculate Kithei Kyalo

Moringa School

Data Science certification

**DATA REPORT**

**BUSINESS UNDERSTANDING**

**Overview**

The 2009 H1N1 flu pandemic highlighted the importance of effective vaccination campaigns in controlling the spread of infectious diseases. Public health officials have long recognized the need to understand the factors that influence vaccination rates, such as individual demographics, opinions, and health behaviors. By leveraging data from the National 2009 H1N1 Flu Survey, we aim to provide insights that can guide public health efforts to optimize vaccine outreach and uptake.

As the global community continues to combat COVID-19, lessons learned from previous pandemics can inform strategies for future vaccination campaigns. Predictive modeling of vaccination patterns enables a data-driven approach to identify individuals or groups most likely to benefit from targeted interventions.

**Problem Statement**

The challenge is to predict whether individuals received the H1N1 flu vaccine or the seasonal flu vaccine based on their background, health behaviors, and opinions. This process is challenging especially because we have two target variables. This binary classification problem will focus on predicting people who received H1N1 flu vaccine using the available datasets. Insights derived from this analysis can enhance our understanding of vaccination drivers and barriers.

**Objectives**

1. **Predictive Modeling:** Develop a robust machine learning model to predict whether individuals received the H1N1 vaccine based on their demographic, behavioral, and opinion-based features.
2. **Insight Generation:** Identify key factors that influence vaccination decisions and their relative importance in prediction.
3. **Public Health Guidance:** Provide actionable recommendations to public health authorities for designing effective vaccination campaigns by understanding vaccination behaviors.
4. **Data-driven Strategies:** Facilitate the design of targeted interventions for groups with low vaccination rates to improve uptake in future public health efforts.

**Success Criteria**

1. **Model Performance:**
   * Achieve a high accuracy, F1-score, or AUC-ROC score for the predictive model to ensure reliable predictions.
2. **Feature Importance:**
   * Identify and communicate the most influential factors affecting vaccination decisions.
3. **Public Health Insights:**
   * Translate model findings into actionable strategies that public health officials can implement in real-world vaccination campaigns.
4. **Stakeholder Utility:**
   * Provide clear and concise recommendations to public health stakeholders that align with the dataset's findings and analysis.

**DATA UNDERSTANDING**

**Overview of the Data**

The dataset comes from the **National 2009 H1N1 Flu Survey**, which collected information on individuals' health behaviors, opinions about vaccines, and demographic backgrounds. It helps us understand the patterns behind who got vaccinated during the H1N1 flu pandemic and why.

**Key Data Points**

The data includes two main types of information:

1. **Personal and Demographic Information**
   * **Age Group:** Age category of the person.
   * **Education Level:** The highest level of education attained.
   * **Income Level:** Income status categorized as above or below the poverty line.
   * **Race and Gender:** Self-reported details about race and gender.
2. **Health and Behavioral Information**
   * **Health Conditions:** Whether the person has chronic medical conditions.
   * **Doctor’s Recommendation:** Whether a doctor recommended the flu vaccine.
   * **Health Behaviors:** Actions like wearing masks, washing hands, or avoiding large gatherings.
   * **Opinions about Vaccines:** Perceptions of vaccine effectiveness, risks, or side effects.

**Focus**

We are working to predict **whether someone received the H1N1 flu vaccine**. This will help us find patterns and identify factors that influenced vaccination decisions.

Understanding this data can help public health teams:

* Identify groups of people who might need more support or targeted vaccine campaigns.
* Understand what opinions or behaviors drive decisions about getting vaccinated.
* Learn from the H1N1 pandemic to improve strategies for future vaccine rollouts (like for COVID-19 or flu vaccines).

**Data Challenges**

Some parts of the data are missing or unclear, which means we need to handle it carefully to ensure accurate results. For example:

* Missing values in health insurance status or income levels.
* Opinions or behaviors that might be self-reported and subject to personal bias.

By addressing these challenges, we aim to build a reliable model that can help public health efforts become more targeted and effective. This project is about turning complex data into actionable knowledge for better health outcomes.

**DATA PREPARATION**

To make sense of the data and use it effectively for predicting who received the H1N1 vaccine, data preparation was done. The data preparation involved several steps:

1. **Cleaning the Data**

Cleaning the data also involved checking and handling duplicate values to ensure that we avoid handling similar records

The data had missing or incomplete information. For example, not everyone answered all the questions in the survey.

* 1. For numerical features we filled in missing values with averages or the most common answers using the imputer from sklearn Library
  2. For categorical data we used "Unknown" as a placeholder for missing answers.

1. **Selecting the Right Information**

Not all data is equally useful. Including unnecessary details could make the model less accurate or harder to interpret.

* 1. We focused on information that is most relevant to predicting vaccination, such as health behaviors, doctor's recommendations, and opinions about vaccines.
  2. We removed information that was very similar (highly correlated), so we don’t duplicate insights or confuse the model.

1. **Making the Data Usable for Analysis**

The computer cannot understand text or categories like "Male" or "Female" directly. These need to be converted into numbers

1. We used a method called **One-Hot Encoding** to turn categories into number where a value of 1 or 0 indicates the category.

**MODELLING**

Once we prepared the data, the next step was to build a **model** to predict whether someone received the H1N1 vaccine. The modelling involved the following steps:

Our goal was to predict whether a survey respondent got the H1N1 vaccine based on things like:

* How concerned they were about H1N1.
* Whether their doctor recommended the vaccine.
* Their health behaviors, like wearing a face mask or washing hands.

**Train\_Test\_Split Model from sklearn**

We used the prepared data to train the model in three key steps:

* **Train:** We gave the model a portion of the data where we already knew whether someone got vaccinated.
* **Test:** We checked how well the model performs using another portion of the data it hadn’t seen before.
* **Refine:** If the model wasn’t accurate enough, we adjusted it by trying different techniques or giving it better input data such as scaling and hyper parameter tuning to improve model performance

The prediction focused on a binary classification model due to the nature of the target variable. The machine learning models involved include **Logistic Regression:** Simple and effective, like drawing a line between people who got vaccinated and those who didn’t and the DecisionTreeClassifier to determine the best prediction.

**EVALUATING**

After building our predictive model, the next step is **model evaluation** which involves checking how well the model performs in making predictions. We need to ensure it’s accurate and reliable before making important decisions based on its predictions. Model evaluation tells us how well it works.

**Model Evaluation**

We tested the model by asking it to predict vaccinations for a set of survey data where we already knew the actual outcomes. Then, we compared the model’s predictions to the actual results. This allowed us to measure its performance.

**Key Performance Metrics for Model Evaluation**

* **Accuracy:**  
  This tells us the percentage of correct predictions. For example, if the model predicts vaccination correctly 72% of the time, its accuracy is 72%.  
  *Question it answers:* *"How often does the model get it right?"*
* **Precision:**  
  Precision focuses on when the model predicts "Yes" (vaccinated). It tells us how often these predictions are correct.  
  *Question it answers:* *"When the model says someone got vaccinated, how often is it right?"*
* **Recall:**  
  Recall tells us how many of the actual vaccinated people the model correctly identified.  
  *Question it answers:* *"How many vaccinated individuals did the model correctly predict?"*
* **F1-Score:**  
  This combines Precision and Recall into a single number, giving us a balanced view of the model’s performance.

Model evaluation allows us to determine when to improve the model performance either by adding more data or better quality data, try a different modeling approach or fine-tune its parameters to improve performance.

Top of Form

Bottom of Form