Machine Learning Model to Predict the Uptake of H1N1 (Swine Flu) Vaccine

# **Report by:**

Gabriel Tenesi

Zipporah Muriithi

Brian Kimathi

Wesley Kipsang

Sharon Wathiri

# **Business Understanding**

**Business Overview**

A vaccine is a medical tool that helps the body build immunity against diseases. Vaccines not only protect individuals but also protect communities through herdimmunity, where enough people are immunized to reduce the overall spread of disease.

There are different types of vaccines, such as the seasonal flu vaccine (for common flu strains) and the H1N1 vaccine (for swine flu). These play a key role in preventing large outbreaks and saving lives.

In 2009, the world faced a pandemic caused by the H1N1 influenza virus (swine flu), which led to an estimated 151,000–575,000 deaths globally in its first year. A vaccine for H1N1 was introduced in October 2009. Shortly after, the U.S. National 2009 H1N1 Flu Survey was conducted to measure who received the H1N1 and seasonalflu vaccines.

The survey also collected information on people’s demographics**,** healthstatus**,** behaviors**,** andopinions. Studying this data helps us understand why some groups chose vaccination while others did not, and provides guidance for future publichealthefforts.

**Problem Statement**

Although vaccines like the seasonal flu and H1N1 were available in 2009, uptake was low, especially for H1N1. This reflects vaccine hesitancy, which weakens herd immunity and increases disease risk.

The key challenge is to understand the factors influencing vaccination decisions—such as demographics, health beliefs, and doctor recommendations—and to build predictive models. These insights can help identify hesitant groups and support better public health strategies in future pandemics.

**Business Objectives**

**Main Objective**

To build a machine learning predictive model that identifies the likelihood of getting the H1N1 vaccine based on the demographic data, health status, and individual’s opinions and behaviors.

**Specific Objectives**

1. To analyze the effect of demographic factors and opinions and beliefs on vaccine uptake.
2. To investigate the influence of health status and behaviors Doctor's recommendations in influencing vaccination decisions
3. To identify the Top features that drive vaccination decisions and hesitancy.

**Success Criteria**

Business success criteria:

* Gain a clear understanding of H1N1 vaccine uptake patterns across different groups of people.
* Identify key factors that influence vaccine decisions.
* Provide insights that can guide public health communication strategies to reduce vaccine hesitancy.

Data success criteria:

* Perform thorough EDA to identify patterns and relationships between the various factors.
* Build at least two machine learning classification models to predict H1N1 vaccine uptake.
* Identify and rank the most important features influencing H1N1 vaccination.
* Ensure results are interpretable and clearly communicated for both technical and non-technical audiences.

# **Data Understanding**

The data is on National 2009 H1N1 Flu Survey in USA. The data has two binary classification targets, which are H1N1 and Seasonal flu vaccination. For this project, H1N1 is our binary classification target because H1N1 vaccine is specific to the swine flu which was the pandemic at the time as compared to the seasonal flu vaccine that catered for various flu strains. The data has **26707 entries** and **38 columns** that include demographic data, health status of the individuals together with their opinions and behaviors towards the vaccine. The demographic information includes age, education, race, gender, marital status, income and employment. The health status includes the chronic medical conditions, health insurance status, and specific behaviors like avoiding close contact or using face masks. The Opinions and knowledge include level of concern about the H1N1 flu, perceived risk of illness, and opinions on vaccine effectiveness and safety. Majority of the columns are either categorical, discrete or binary.

# **Data Preparation**

The 5 columns related to seasonal vaccine were dropped as well as 3 columns that did not have useful information on the H1N1 vaccine and had the most missing values

There were no duplicates in the dataset.

Missing values in majority of the columns which were imputed using mode and “Unknown value” as most of them were categorical.

# **EDA**

From the dataset, above 20,000 people did not receive the vaccine while only about 6,0000 received the vaccine.

**Demographic Factors impact on H1N1 vaccination**

Age Group: Older respondents (45-64, 65+ years) show a higher vaccine uptake than the younger groups (18-34 years), who are less likely to be vaccinated.

Sex: Females have a slightly higher uptake than males. The difference seems small but noticeable.

Education: Higher education levels (College Graduate) have a higher uptake of the vaccine than those with ≤12 years of education.

Income poverty: Average income levels <= $75,000, Above Poverty had the highest vaccine uptake as compared to the other in come levels

**Opinion Impact on H1N1 vaccination**

The opinion on the vaccine effectiveness has the largest positive coefficient (+0.73), meaning the more effective people believe the vaccine is, the more likely they are to take it. This will in turn increase public trust in vaccine effectiveness could substantially boost uptake.

**Health status impact on H1N1 vaccination**

People with chronic medical conditions were slightly more likely to take the H1N1 vaccine (r = 0.10).

Protective behaviors such as mask use (r = 0.07), frequent handwashing (r = 0.07), and avoiding crowds (r = 0.05) also showed small positive links to vaccination.

Those who received doctors’ recommendation had a significantly higher average rate of vaccination. This suggests that trusted medical advice is a powerful driver of vaccination uptake.

# **Modeling**

The categorical variables were encoded using one hot encoding to prepare them for model development

Due to class imbalance of the target variable H1N1 vaccine, SMOTE oversampling technique was used to balance the classes.

The logistic Regression was first used as a baseline classification model, then the Random Forest and an optimization to the random forest, Gradient Boosting was utilized to improve the model performance. The three models were compared against each other to select the best fit model.

# **Evaluation**

Logistic Regression Evaluation:

F1 Score: 0.7938

Precision: 0.8111

Recall: 0.7838

ROC AUC Score: 0.7942

Tuned Random Forest Evaluation:

F1 Score: 0.4929

Precision: 0.6798

Recall: 0.4523

ROC AUC Score: 0.5316

Gradient Boosting Evaluation:

F1 Score: 0.8173

Precision: 0.816

Recall: 0.8308

ROC AUC Score: 0.8247

Having tried the 3 models, it became clear that the Gradient Boosting model significantly outperforms both Logistic Regression and Random Forest in both key metrics of accuracy, F1 score, and robustness towards class imbalance. Logistic Regression, while simple and interpretable, was insufficient to capture the complexity in the data and thus reflected lower performance metrics.

The high mean F1 score of 0.8173 implies that the model has a good balance of precision and recall, which is something that will be very useful in this instance, since it's a multi-class problem.

**Feature Importance**

Key predictors derived from the feature importance analysis are doctor recommendation, opinion of the vaccine effectiveness and opinion of the vaccine risk, further evidencing that they are the major factors that contribute to vaccine uptake.

# **Conclusion**

The Gradient Boosting model, developed in this research for estimating H1N1 vaccine uptake has shown excellent performance. It has achieved an F1 score of 81.73% on the test set, proving fairly efficient in estimating H1N1 vaccine uptake. Overall, the model performs with quite high accuracy and balanced metrics, which suggests that it reliably predicts the vaccine uptake without considerable bias.

# **Recommendation**

1. Doctor’s role: Encourage doctors to actively recommend the H1N1 vaccine.
2. Effectiveness Messaging: Share clear evidence that the vaccine works.
3. Risk Awareness: Emphasize the dangers of H1N1 infection.
4. Close Knowledge Gaps: Provide simple education and target groups with hesitancy and low awareness.