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

There were 14 duplicates that were dropped and the missing values in majority of the columns which were imputed using a KNN imputer.

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

# **EDA**

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

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

**Correlation**

The correlation plot places doctor recommendation to the h1n1 vaccine as a strong influencer of the vaccine uptake as it has a correlation coefficient of 0.39, followed by a correlation coefficient of 0.32, 0.27 on opinion of the risk of getting the vaccine and the opinion of the vaccine efficiency respectively. Additionally, the health behaviors do not contribute much to the uptake of h1n1 vaccine as compared to the opinions of the persons since they have a very low correlation of 0.07.

# **Modeling**

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

**Logistic Regression**

The logistic Regression was first used as a baseline classification model and Random Forest was later compared to the Logistic regression since it caters for complex patterns.

The Logistic Regression model has fair scores in correctly predicting the majority class; its misclassifications are few across all classes.

*Accuracy:* 0.774574, fair.

*F1 Score:* 0.78919, showing fair balance between precision and recall.

*Precision:* 0.822089 indicates that when the model predicts a target, it is often correct.

*Recall:* 0.774574, implies that the model misses few positive instances.

*ROC AUC:* 0.83038, it suggests a reasonable discriminative ability, though failing to effectively predict minority classes.

**Random Forest**

The Random Forest model outperforms Logistic Regression especially after applying hyperparameter tunning, in terms of classification accuracy, correctly classifying most samples.

*Accuracy:* 0.825314, nearly unchanged from the untuned model.

*F1 Score:* 0.8155942, close to the untuned performance.

*Precision:* 0.812540, strong but slightly higher than before tuning.

*Recall:* 0.825314, consistent with accuracy.

*ROC AUC:* 0.819959, showing minimal impact from tuning on class differentiation.

Having tried both the Logistic Regression and Random Forest models, it became clear that the tuned Random Forest model significantly outperforms Logistic Regression 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.

Therefore, the Random Forest algorithm performed much better in modeling such complex relations between variables of the dataset and was more suitable for this project.

# **Evaluation**

The selected model performs quite well on average, with the mean value of 0.8254, whereby most of the features are rightly predicted. This gives quite promising results toward h1n1 vaccine prediction with respect to features.

The high mean F1 score of 0.8095 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.

With an average precision of 0.8086, the model does generally well in predicting the actual positives correctly. This is important for correctly classifying features.

A high recall would mean that the model is good at finding examples from all features, which is important to ensure all relevant cases are detected. Mean Recall: 0.8254.

# **Conclusion and Recommendations**

The Random Forest-based model, developed in this research for estimating h1n1 vaccine uptake has shown excellent performance. It has achieved 82.14% accuracy and an F1 score of 81.06% 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 h1n1 vaccine uptake without considerable bias.

Key predictors derived from the feature importance analysis are opinion of the vaccine effectiveness, opinion of the vaccine risk, doctor recommendation, age group, education, opinion on getting sick from vaccine and, income poverty, further evidencing that they are the major factors that contribute to vaccine uptake.

Despite challenges with class imbalance, the Random Forest model presents a robust tool in h1n1 vaccine uptake predictions, which would be refined further by adding more focus on class balance.