**Understanding the Uptake of Vaccination among Demographic Feature**

**Overview**

This Analysis aims to understand the correlation between individuals' backgrounds, opinions, and health behaviors and their personal vaccination patterns. The research will use data collected in the National 2009 H1N1 Flu Survey to predict the likelihood of people receiving a vaccine. The results of this study will provide valuable insights for future public health efforts in the fight against COVID-19.

**Problem definition**

The global population's successful vaccination against COVID-19 is a major challenge facing public health efforts. An understanding of the factors that influence people's personal vaccination patterns, such as their backgrounds, opinions, and health behaviors, can help guide future public health strategies. The study will address this issue by using data from the National 2009 H1N1 Flu Survey to predict the probability of an individual receiving a vaccine, which will provide valuable information for future public health efforts.

**Data Understanding**

Data which contained 26707 entries was obtained from the public health website. The data is a collection of responses from people regarding the H1N1 and seasonal flu vaccines that they received. The data includes 35 columns for demographic information, such as age, gender, and location, as well as information on the respondents' opinions and health behaviors. The goal of this data is to understand how different factors, such as background and opinions, are related to personal vaccination patterns. This information can be used to guide future public health efforts in the distribution of vaccines. The data is organized as a CSV file, where each row represents one person who responded to the survey, and each column represents a different aspect of their response.

Below is a breakdown of the categorical variables:

**For all binary variables: 0 = No; 1 = Yes**.

**h1n1\_concern - Level of concern about the H1N1 flu**. 0 = Not at all concerned; 1 = Not very concerned; 2 = Somewhat concerned; 3 = Very concerned.

**h1n1\_knowledge** - Level of knowledge about H1N1 flu. 0 = No knowledge; 1 = A little knowledge; 2 = A lot of knowledge.

**behavioral\_antiviral\_meds** - Has taken antiviral medications. (binary)

**behavioral avoidance** - Has avoided close contact with others with flu-like symptoms. (binary)

**behavioral\_face\_mask** - Has bought a face mask. (binary)

**behavioral\_wash\_hands** - Has frequently washed hands or used hand sanitizer. (binary)

**behavioral\_large\_gatherings** - Has reduced time at large gatherings. (binary)

**behavioral\_outside\_home** - Has reduced contact with people outside of own household. (binary)

**behavioral\_touch\_face** - Has avoided touching eyes, nose, or mouth. (binary)

**doctor\_recc\_h1n1** - H1N1 flu vaccine was recommended by doctor. (binary)

**doctor\_recc\_seasonal** - Seasonal flu vaccine was recommended by doctor. (binary)

**chronic\_med\_condition** - Has any of the following chronic medical conditions: asthma or another lung condition, diabetes, a heart condition, a kidney condition, sickle cell anemia or other anemia, a neurological or neuromuscular condition, a liver condition, or a weakened immune system caused by a chronic illness or by medicines taken for a chronic illness. (binary) child\_under\_6\_months - Has regular close contact with a child under the age of six months. (binary)

**health\_worker** - Is a healthcare worker. (binary)

**health\_insurance** - Has health insurance. (binary)

**opinion\_h1n1\_vacc\_effective** - Respondent's opinion about H1N1 vaccine effectiveness. 1 = Not at all effective; 2 = Not very effective; 3 = Don't know; 4 = Somewhat effective; 5 = Very effective.

**opinion\_h1n1\_risk** - Respondent's opinion about risk of getting sick with H1N1 flu without vaccine. 1 = Very Low; 2 = Somewhat low; 3 = Don't know; 4 = Somewhat high; 5 = Very high.

**opinion\_h1n1\_sick\_from\_vacc** - Respondent's worry of getting sick from taking H1N1 vaccine. 1 = Not at all worried; 2 = Not very worried; 3 = Don't know; 4 = Somewhat worried; 5 = Very worried.

**opinion\_seas\_vacc\_effective** - Respondent's opinion about seasonal flu vaccine effectiveness. 1 = Not at all effective; 2 = Not very effective; 3 = Don't know; 4 = Somewhat effective; 5 = Very effective.

**opinion\_seas\_risk** - Respondent's opinion about risk of getting sick with seasonal flu without vaccine. 1 = Very Low; 2 = Somewhat low; 3 = Don't know; 4 = Somewhat high; 5 = Very high.

**opinion\_seas\_sick\_from\_vacc** - Respondent's worry of getting sick from taking seasonal flu vaccine. 1 = Not at all worried; 2 = Not very worried; 3 = Don't know; 4 = Somewhat worried; 5 = Very worried.

**age\_group** - Age group of respondents.

**education** - Self-reported education level.

**race** - Race of respondent.

**sex** - Sex of respondent.

**income\_poverty** - Household annual income of respondent with respect to 2008 Census poverty thresholds.

**marital\_status** - Marital status of respondent.

**rent\_or\_own** - Housing situation of respondent.

**employment\_status** - Employment status of respondent.

**hhs\_geo\_region** - Respondent's residence using a 10-region geographic classification defined by the U.S. Dept. of Health and Human Services. Values are represented as short random character strings.

**census\_msa** - Respondent's residence within metropolitan statistical areas (MSA) as defined by the U.S. Census.

**household\_adults** - Number of other adults in household, top-coded to 3.

**household\_children** - Number of children in household, top-coded to 3.

**employment\_industry** - Type of industry respondent is employed in. Values are represented as short random character strings.

**employment\_occupation** - Type of occupation of respondent. Values are represented as short random character strings.

**Data Preparation**

The data was categorical, no outliers were observed. However, columns such as census\_msa, hhs\_geo\_region, employment\_status, rent\_or\_own, marital\_status, income\_poverty, sex, age\_group,

Race and education were converted to a numerical format.

The missing values were filled using the mode of the respective column and all duplicates dropped.

By utilizing Lasso Regression and Decision Tree based feature selection, we were able to identify the most impactful features in our model. These methods analyze the relationships between the features and target variables and rank them based on their contribution to the model's predictive performance.

Lasso Regression is a type of regularization that adds a penalty to the size of the coefficients in the linear regression model. This technique helps us identify which features have the most impact on the target variable by shrinking the coefficients of less important features toward zero.

Decision Tree is an ensemble learning method that builds a decision tree and combines their predictions to produce a single result. This technique helps us identify the most important features by calculating the feature importance scores

In addition to the features highlighted by these methods, it is important to also consider the potential impact of other demographic factors such as health worker status, marital status, presence of chronic medical conditions, sex, race, education level, age group, and employment status. These variables may not be explicitly represented in the feature selection results, but their inclusion can further enhance the interpretability and accuracy of our model.

**Modeling and Evaluation**

Using the processed data, I was able to train a logistic model, decision tree model and a Random Forest model.

I evaluated each of the models’ performance using appropriate metrics such the Accuracy, Precision, Recall and F1 score. The best performing model was the Logistic model with the following metrics.

Accuracy: 0.7830400599026581

The accuracy of the model is 0.7830, which means that the model correctly predicts the target class 78.3% of the time. This is a good overall indicator of the model's performance, but it doesn't take into account false positives or false negatives, which can be important in certain applications.

Precision: 0.8020969855832241

Precision measures the proportion of positive predictions that are actually correct. In this case, the precision is 0.802, meaning that 80.2% of the positive predictions made by the model are correct. Precision is important when the cost of false positive predictions is high.

Recall: 0.9355414012738853

Recall measures the proportion of actual positive cases that the model correctly predicts. In this case, the recall is 0.935, meaning that the model correctly identifies 93.5% of the positive cases. Recall is important when the cost of false negative predictions is high.

F1 Score: 0.8636951664118545

The F1 score is the harmonic mean of precision and recall, and it provides a single score that balances both precision and recall. In this case, the F1 score is 0.864, which is a good overall indicator of the model's performance, especially when the costs of false positive and false negative predictions are similar.

In conclusion, the model has a relatively high accuracy, precision, and recall, which suggests that it is performing well. The high F1 score confirms this, as it provides a balance between precision and recall. However, these results should always be interpreted in the context of the problem and the specific application to fully understand the model's performance.

**Conclusion**

The results of this project provide valuable insights into the correlation between different factors and vaccine uptake. This information can be used to inform future public health efforts aimed at increasing the uptake of the COVID 19 Vaccine