H1N1 AND SEASONAL FLU VACCINES

Final Project Submission

Student names: Annah Mukethe

Student pace: Part Time

Scheduled project review date/time: Phase 3

Instructor name: Winnie Anyoso

PRESENTATION OBJECTIVES

INTRODUCTION

- Project Overview
- Business Understanding
- Project Methodology

DATA PREPARATION

- Data inspection
- Data preprocessing
- Feature engineering

MODELING

- Feature selection
- Model Training
- Model Valuation

SUMMARY

- Visualizations
- Recommendation
- Conclusion

PROJECT OVERVIEW

A vaccine for the H1N1 flu virus became publicly available in October 2009. In late 2009 and early 2010, the United States conducted the National 2009 H1N1 Flu Survey. This phone survey asked respondents whether they had received the H1N1 and seasonal flu vaccines, in conjunction with questions about themselves. These additional questions covered their social, economic, and demographic background, opinions on risks of illness and vaccine effectiveness, and behaviors towards mitigating transmission. A better understanding of how these characteristics are associated with personal vaccination patterns can provide guidance for future public health efforts.

BUSINESS UNDERSTANDING

INTRODUCTION

The COVID-19 pandemic has highlighted the critical importance of vaccination campaigns in controlling the spread of infectious diseases. Public health efforts are more effective when they are informed by an understanding of the factors that influence individuals' decisions to get vaccinated. By analyzing past vaccination patterns, such as those during the 2009 H1N1 flu pandemic, we can gain valuable insights that can help design better vaccination strategies, improve communication campaigns, and ultimately increase vaccine uptake in future public health crises.

BUSINESS PROBELM

The task is to predict whether individuals received the H1N1 flu vaccine using data from the 2009 National H1N1 Flu Survey. This binary classification problem involves analyzing various factors, such as demographics, opinions, and health behaviors, to determine their relationship with vaccination behavior. The outcome of this analysis will help public health authorities identify key factors influencing vaccine acceptance and inform strategies to enhance vaccine coverage in the population.

OBJECTIVES

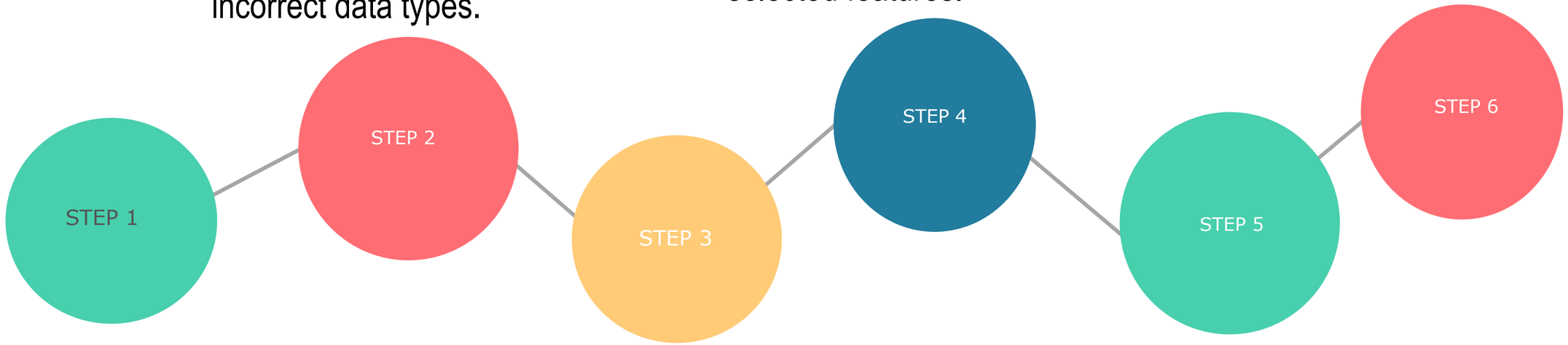
1. Identify Key Predictors: Determine the most significant factors that influence whether an individual received the H1N1 vaccine.

2. Develop a Predictive Model: Build a binary classification model to accurately predict whether a survey respondent received the chosen vaccine.

3. Evaluate Model Performance: Assess the model's performance using appropriate metrics such as accuracy, precision, recall, AUC and F1 score to ensure its reliability in predicting vaccination behavior.

4. Provide Actionable Insights: Analyze the model's findings to provide public health authorities with actionable insights that can guide future vaccination campaigns and strategies, particularly in the context of managing public health responses to pandemics.

PROJECT METHODOLOGY



DATA COLLECTION AND INSPECTION

Gathering the necessary data from the provided dataset.

DATA PREPROCESSING

Cleaning the data to handle missing values, outliers, and incorrect data types.

ANALYSIS

Analyzing the data to find patterns, relationships, and insights.

MODELING

Building predictive models to estimate house prices based on selected features.

MODEL EVALUATION

Assessing the models' performance using appropriate metrics.

NEXT STEPS

Drawing conclusions from the model results and providing recommendations.

DATA INSPECTION AND UNDERSTANDING

Dataset Overview:

Source:

The data was sourced from this website https://www.drivendata.org/competitions/66/flu-shot-learning/

Content:

The dataset contains responses given 0 being No and 1 being yes to vaccine, along with various attributes such as the h1n1 knowledge, concerns, opinion about H1N1 vaccine effectiveness and more. A separate file was created **feature_description.txt** that provides a description of the column names

Data Size:

The dataset has 26,707 rows and 37 columns.

DATA PREPROCESSING

Handling Missing Values:

Identified missing values in 21 columns with employment_occupation having the highest number of missing values accounting for 50% followed by employment_industry with 49.9 % missing values. We took steps to handle missing by dropping some of the columns whose data did not have any significant meaning. We also imputed the rest of the columns with the most frequent value and others with a specific values ‘unknown’ to avoid any manipulations to the individual responses given.

Handling Outlier:

Upon examining the outliers in the dataset, particularly in the opinion_seas_vacc_effective, opinion_seas_sick_from_vacc, household_adults, and household_children columns, we decided not to drop or cap these outliers. The key justification was that, the variables opinion_seas_vacc_effective and opinion_seas_sick_from_vacc capture individuals' perceptions and opinions regarding the effectiveness and potential side effects of the seasonal flu vaccine. Outliers in these columns likely represent strong opinions or unique perspectives that are critical to understanding the full spectrum of public sentiment. Removing these outliers could lead to a loss of valuable insights, particularly in understanding the range of beliefs and attitudes that influence vaccination decisions.

Feature Engineering:

Our dataset did not have any columns that could aid in generating new columns

DATA ANALYSIS

Descriptive Statistics:

- Calculated summary statistics for the dataset columns.
- Analyzed the distribution of the key features.

Correlation Analysis:

- Examined correlations between key variables such as age group, race, marital status, employment status h1n1 knowledge, doctor recommendation on h1n1 against the target variable which is h1n1 vaccine as well as against other features.

Multivariate Analysis:

- Conducted multivariate analysis to understand how variables like opinion on h1n1 vaccine effective, opinion on h1n1 risk influence the housing price.

VISUALIZATIONS

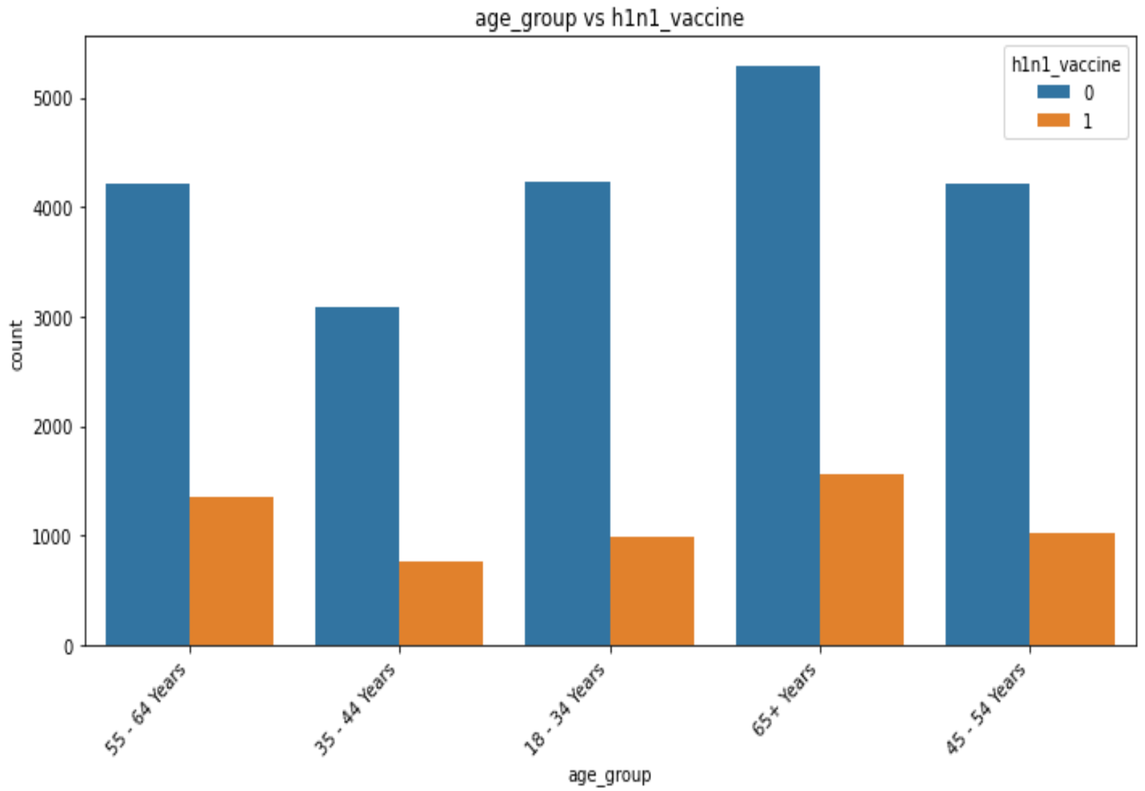
Distribution Plots:

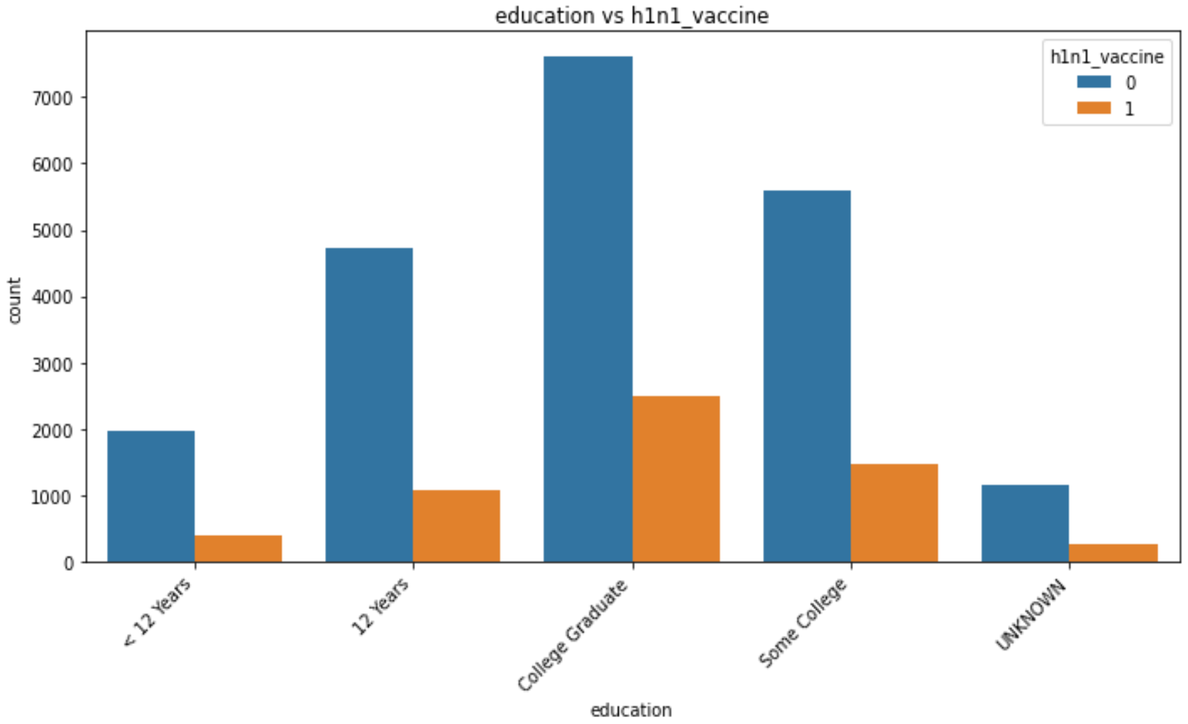
- Used bar plot to show counts of each age group, education, race, sex, marital status, rent or own, employment status on vaccine uptakes.
- Created box plot to show h1n1 vaccine uptake based on opinion of effectiveness of the vaccine, opinion on risk, respondent being a health worker.

Features Analysis:

- Used pair plot to understand the correlation between respondents features and h1n1 vaccine.
- Also plotted heatmap to get feature that are highly correlated to our target value.

H1N1 VACCINE UPTAKE BASED ON AGE GROUPS AND EDUCATION LEVELS

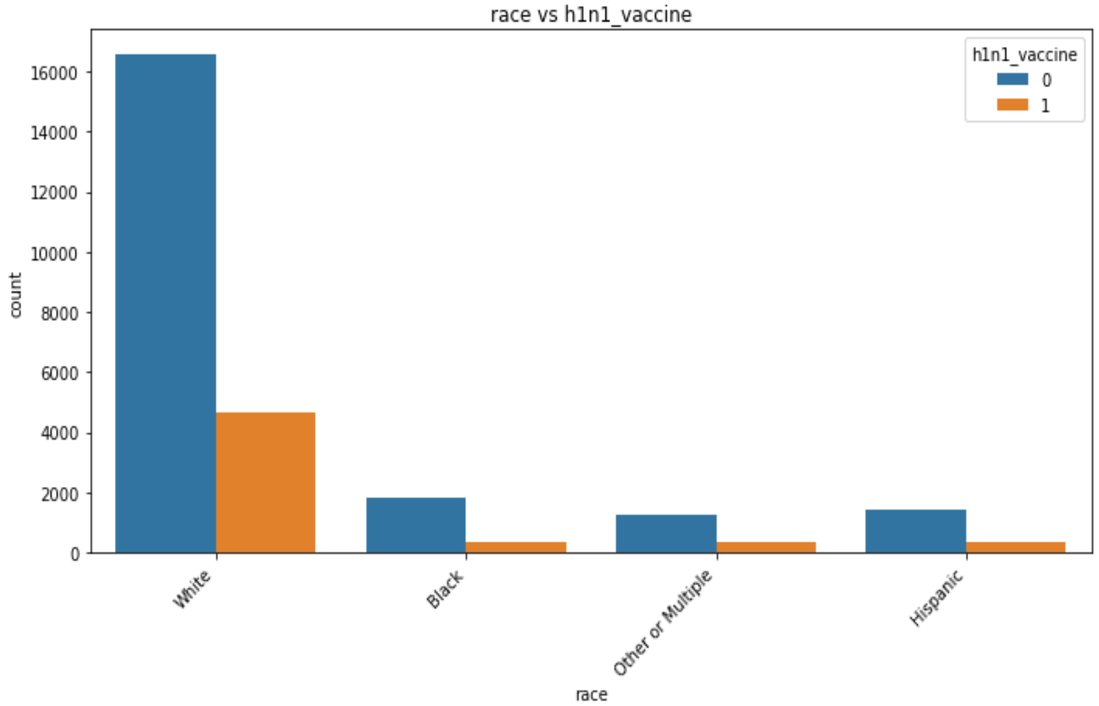


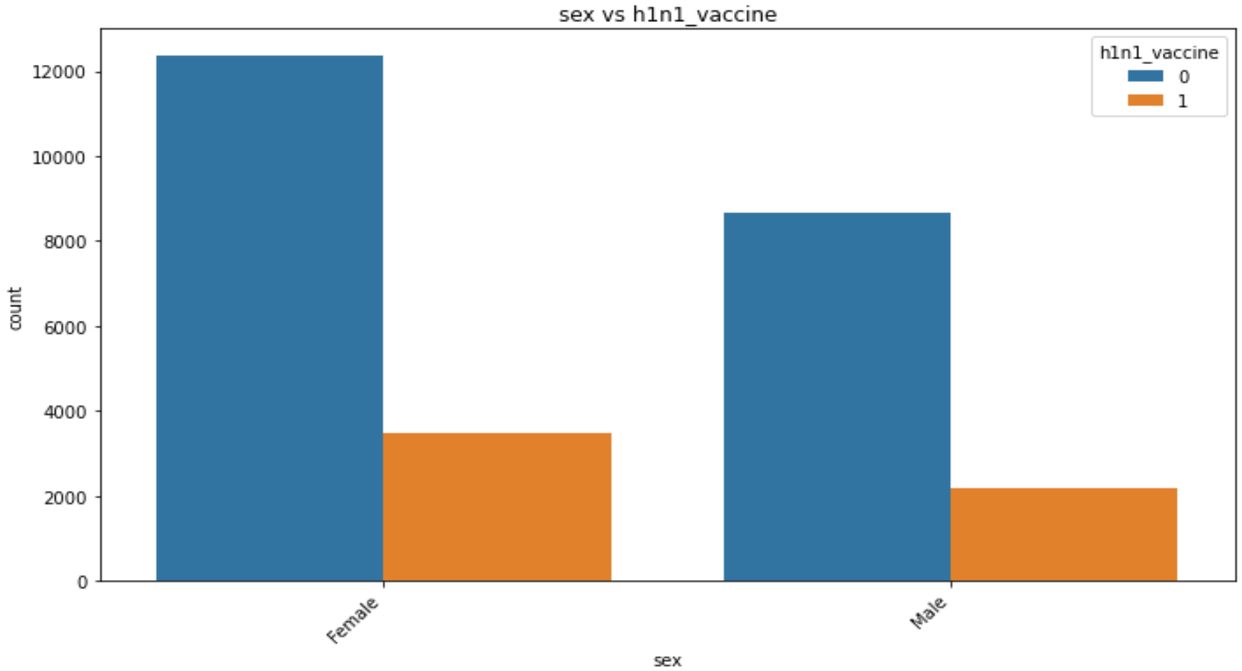


The general population seems to have a low uptake of the H1N1 vaccine across all these age groups. However the 65+ years age group does have the highest number of populations not interested in the vaccine. The same age group, along with the 55-64 age group also showed a reasonably high number in receiving the vaccine as compared to all other groups.

From this bar plot, college student categories indicates high number of people did not get the vaccine as compared to all other categories.

H1N1 VACCINE UPTAKE BASED ON RACE AND SEX

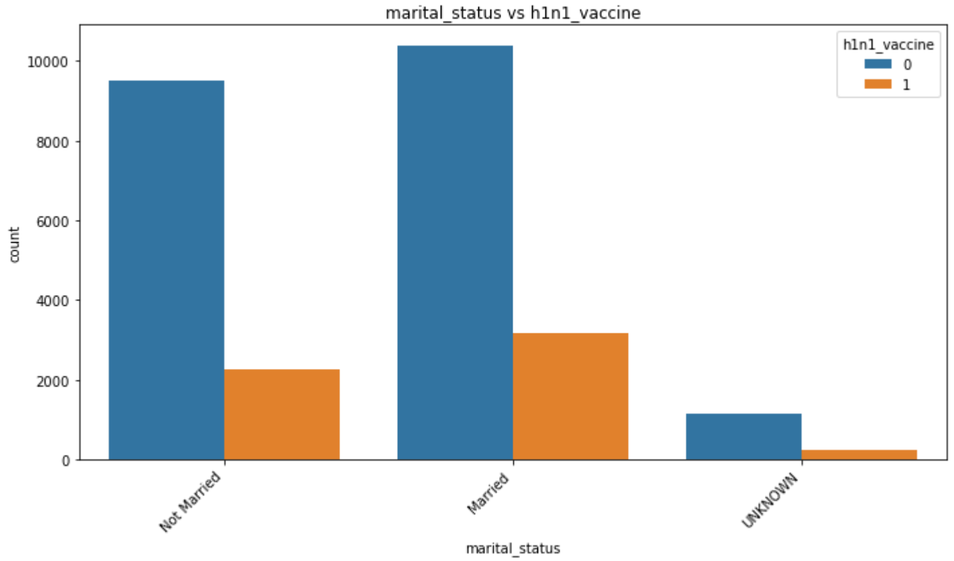


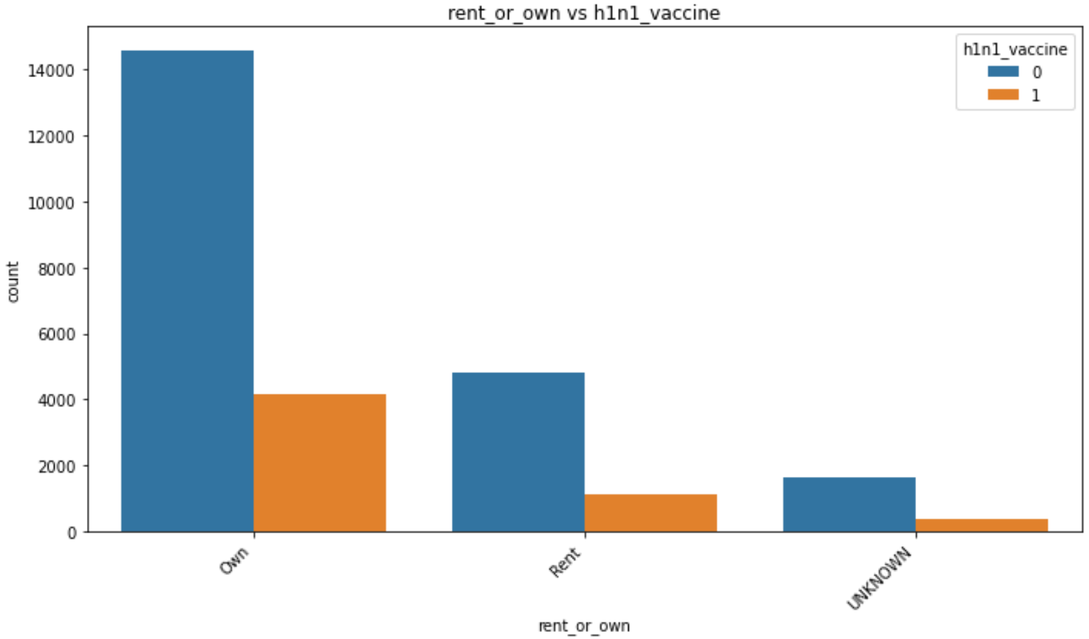


Population belonging to the white race category takes the majority number in regards to not taking the vaccine as compared to the other race categories. Also the white category did have a higher number of vaccinated individuals.

The female gender has the highest number of individuals who are not vaccinated as compared to the male category. In addition, the female category also has a slightly higher number of individuals who are vaccinated.

H1N1 VACCINE UPTAKE BASED ON MARITAL STATUS AND RENT OR OWN HOUSING SITUATION

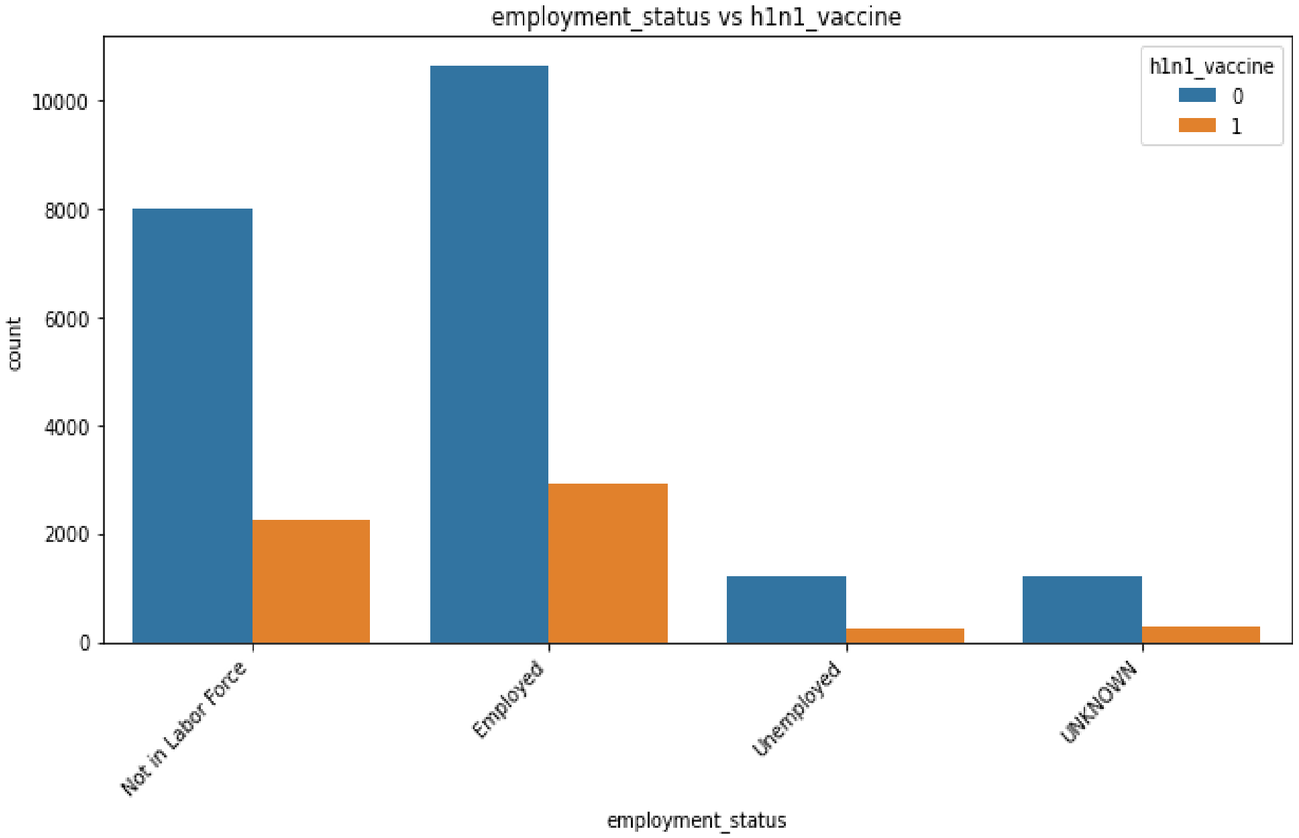


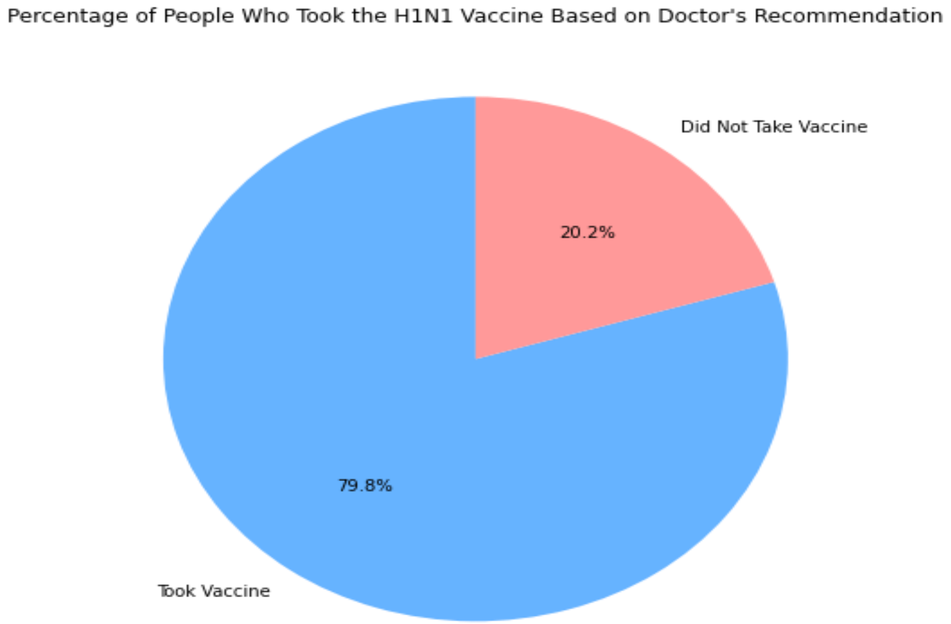


Married category has the highest number of individuals who are not vaccinated as compared to the other categories. In addition, the same category also has a slightly higher number of individuals who are vaccinated.

Based on the housing situation, the individuals with their own houses, had majority number of non-vaccinated individuals while on the other hand the rented category had fewer numbers.

H1NI VACCINE UPTAKE BASED ON EMPLOYMENT STATUS AND DOCTORS RECOMMEDATION

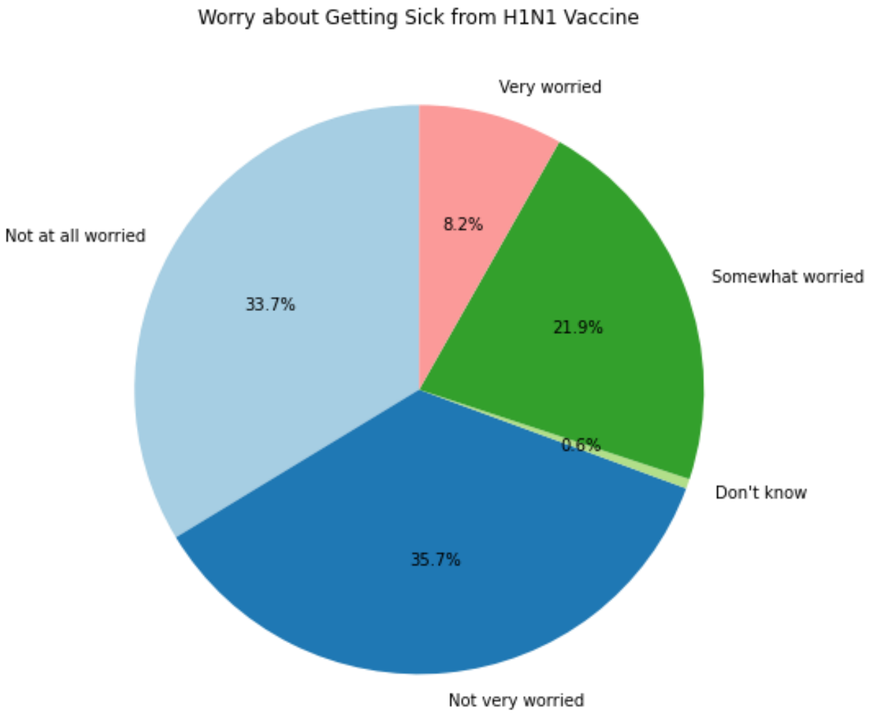




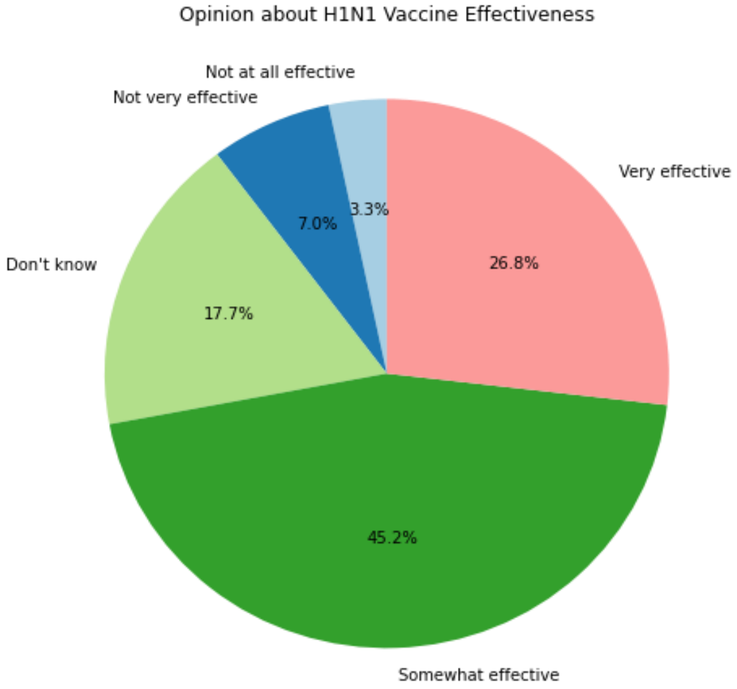
The employed followed for the people who are not in the labor force had high number of individuals who did not take the vaccine compared to those whom did.

The majority of respondents trusted and acted on their doctor's recommendation to take the H1N1 vaccine, indicating high compliance with medical advice regarding vaccination.

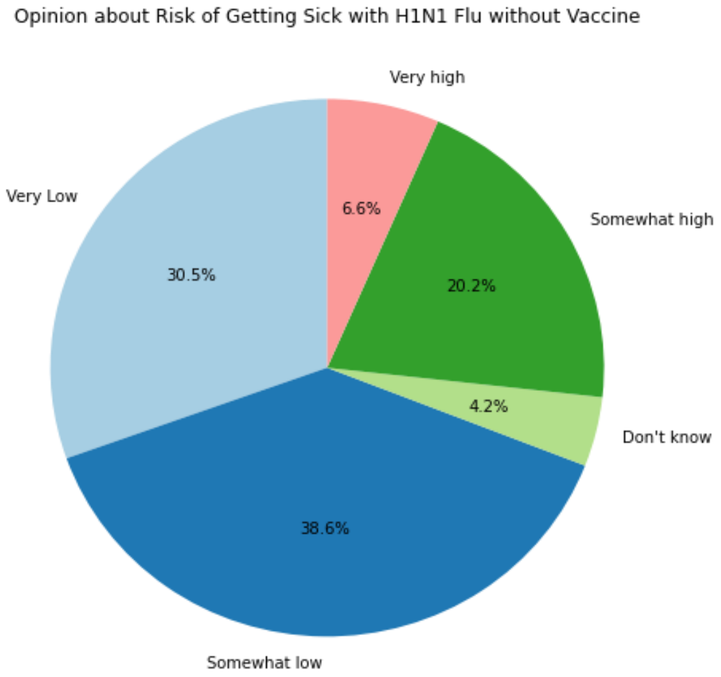
H1N1 VACCINE UPTAKE AGAINST RESPONDENTS OPINIONS



The largest segments of the chart show that most people are either "Not very worried" (35.7%) or "Not at all worried" (33.7%), totaling 69.4% of respondents. This suggests that a significant majority have low levels of concern about adverse effects from the vaccine.



The majority of respondents (72%) have a positive perception of the vaccine’s effectiveness (somewhat to very effective), while a smaller portion either has doubts or believes the vaccine is not effective.



Most respondents (69.1%) perceive a high risk of getting sick without the H1N1 vaccine, which could explain why many are inclined to follow recommendations to get vaccinated. However, there is still a significant portion (20.2%) who are uncertain about the risk.

GROUP 16-PT7

MODELING APPROACH

Feature selection

- Feature were selected based on their values on the heatmap and domain knowledge.

Preprocessing

- Standardization: Features normalized for consistent scaling
- Encoding: One-hot encoding applied to categorical features

Data splitting

- The dataset was split into 80% training set and 20 % the test set.

Modeling Techniques

- Trained the model on logistic regression and decision trees, then used did hyper parameter tuning to get the best parameters and improve the model,

Model Evaluation

- We evaluated the models based on accuracy, precision, recall, F1 score and AUC
- Model performance comparisons and results interpretation.

MODELING EVALUATION AND RESULTS INTERPRETATION

Below are the results from both logistic and decision trees models:

Results Comparison

Metric	Logistic Regression	Decision Trees
Accuracy	0.84	0.84
Precision (Class 1)	0.70	0.69
Recall (Class 1)	0.40	0.42
F1-Score (Class 1)	0.51	0.52
AUC	0.82	0.81

Based on the results table, both models perform similarly, with Decision Tree showing a slight edge in recall and F1-score for Class 1. Logistic Regression has a marginally higher AUC, suggesting it may be preferable for overall discriminative ability.

CONCLUSION

This project aims to provide a thorough analysis of the H1N1 vaccine uptake to support public health officials and policymakers in making informed decisions regarding vaccination strategies and public outreach. Through detailed exploratory data analysis and model development, we have identified the key factors that significantly influence individuals' decisions to receive the H1N1 vaccine.

RECOMMEDATIONS

Below are some of the recommendations to public health authorities:

1. Targeted Interventions: Focus educational campaigns on older adults, especially those aged 65 and above, to increase vaccine uptake in this demographic.

2. Enhance Doctor Involvement: Encourage healthcare providers to play a more active role in recommending the vaccine, as their influence is crucial for increasing vaccination rates.

3. Address Misconceptions: Develop public health campaigns to dispel myths and boost confidence in the vaccine's effectiveness, emphasizing clear and evidence-based information.

4. Conduct Further Analysis: Explore additional health-related factors or interactions that might influence vaccination decisions, particularly for those with chronic medical conditions.

THANK YOU