Introduction

The **goal** of this project is to build predictive models to estimate the probability of individuals receiving the **H1N1 vaccine** and the **seasonal flu vaccine** based on demographic, behavioral, and health-related factors. The dataset used for this analysis comes from the **2009 National H1N1 Flu Survey**, which includes responses from thousands of individuals regarding their attitudes towards vaccination, medical history, and risk perception.

Understanding vaccine uptake patterns is **crucial** for public health authorities to design targeted awareness campaigns and vaccination programs, especially during pandemic situations. By leveraging **machine learning models**, we aim to improve vaccine adoption strategies.

The models developed include:
✅ **Logistic Regression** – A simple baseline model.
✅ **Random Forest** – A tree-based model capturing non-linear relationships.
✅ **XGBoost** – An advanced boosting algorithm.
✅ **LightGBM** – A fast gradient boosting model that performed the best.

This report presents an exploratory **data visualization**, data preprocessing steps, and a comparison of **model performances using ROC-AUC scores**

Dataset Description

The dataset used in this analysis is the 2009 National H1N1 Flu Survey dataset, which contains responses from individuals regarding their vaccination status and various demographic, health-related, and behavioral factors.

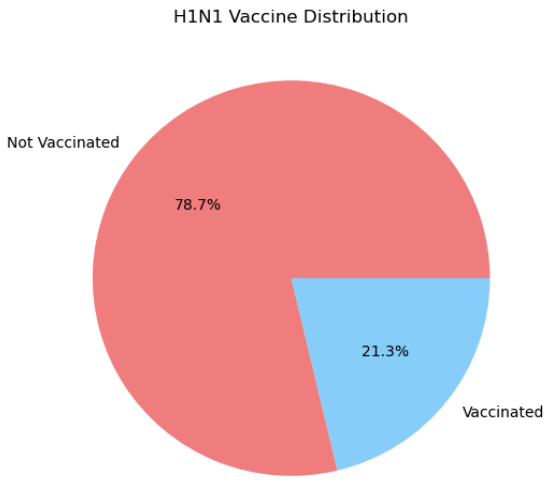
📌 Dataset Overview:

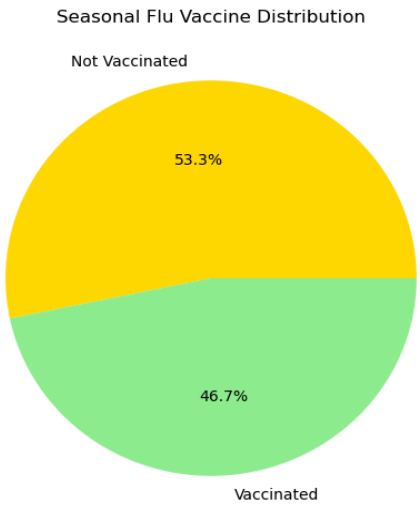
- Total Rows: 26708
- Total Columns: 36
- Target Variables:
    - h1n1_vaccine: 1 if the person took the H1N1 vaccine, 0 otherwise.
    - seasonal_vaccine: 1 if the person took the seasonal flu vaccine, 0 otherwise.
- Key Features:
    - Demographics: Age, gender, education level, income level, etc.
    - Health Conditions: Chronic diseases, doctor visits, general health conditions, etc.
    - Vaccination Attitudes: Concern about vaccine safety, effectiveness, trust in government, etc.
    - Behavioral Factors: Frequency of doctor visits, social distancing habits, etc.

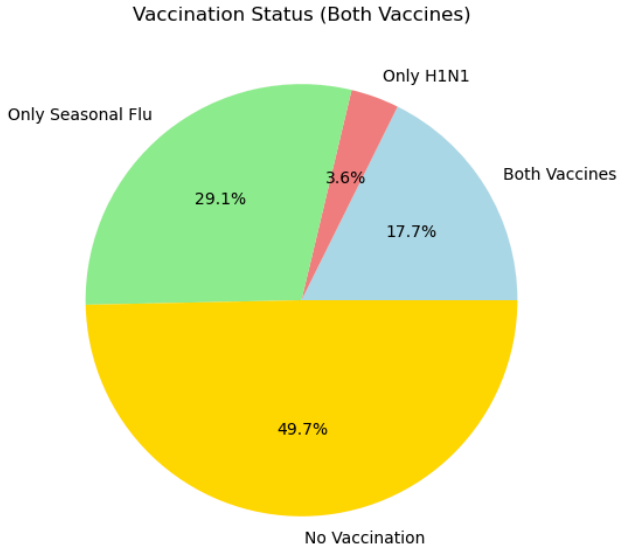
A total of 26,708 individuals are included in this dataset, providing a diverse sample for vaccine adoption modeling.

DATA VISUALIZATION

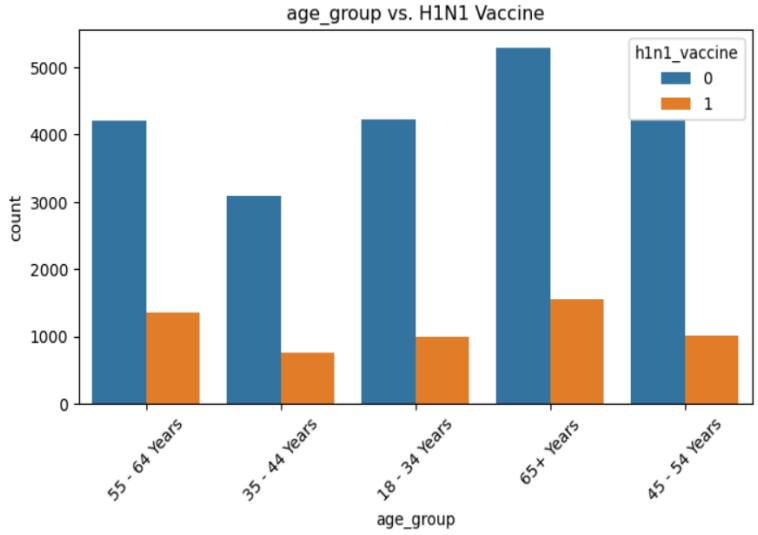
- PEOPLE VACCINATED STATS

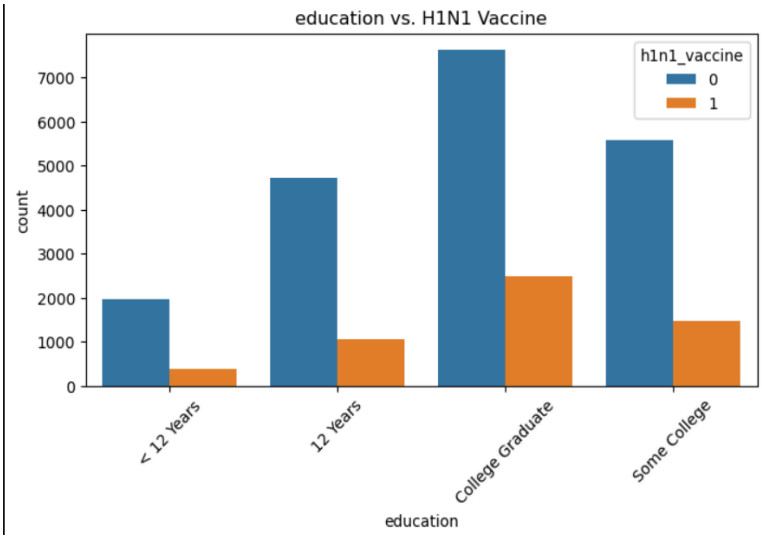


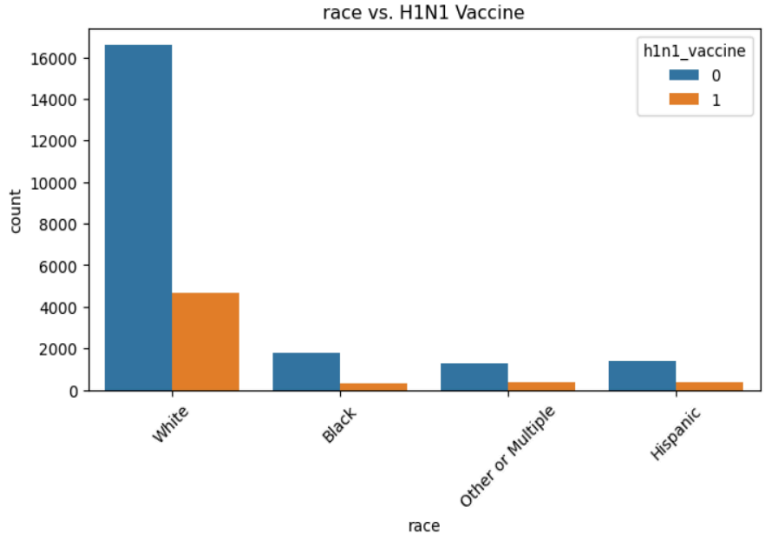


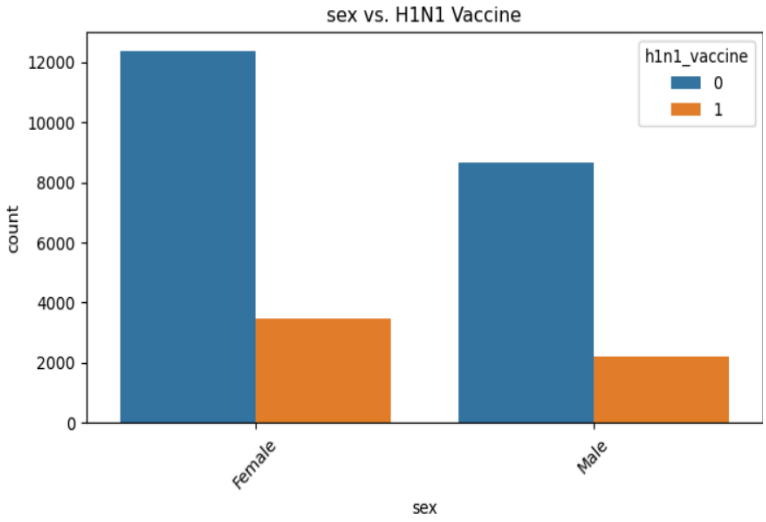


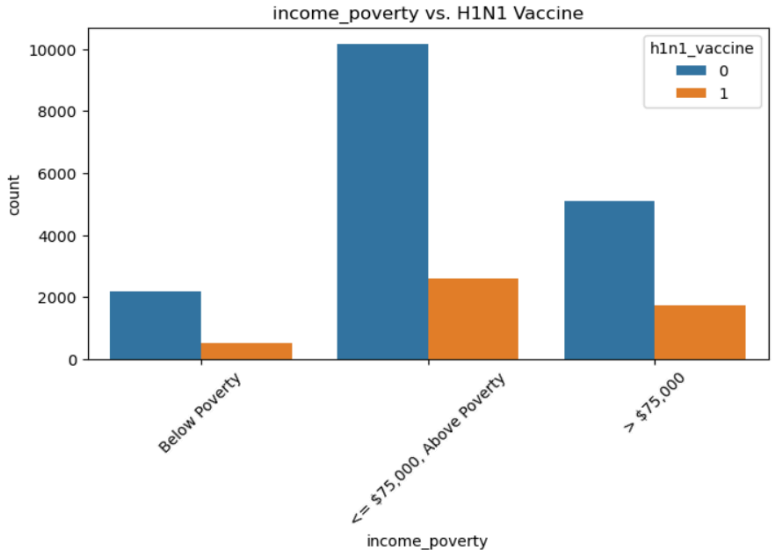
- FACTORS AFFECTING VACCINATION RATE

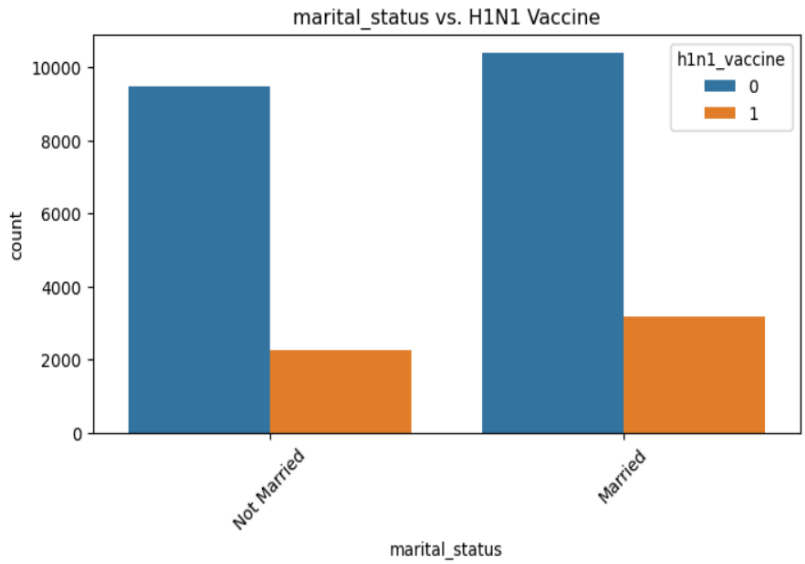


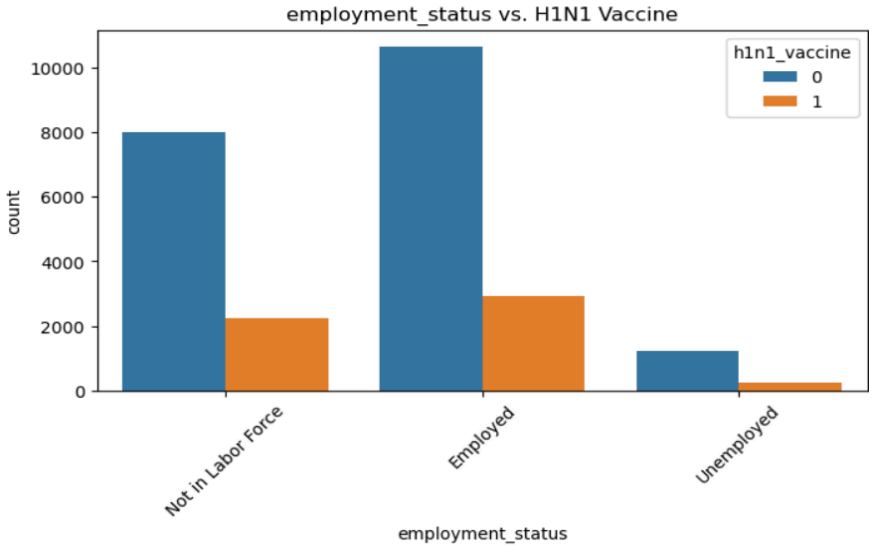


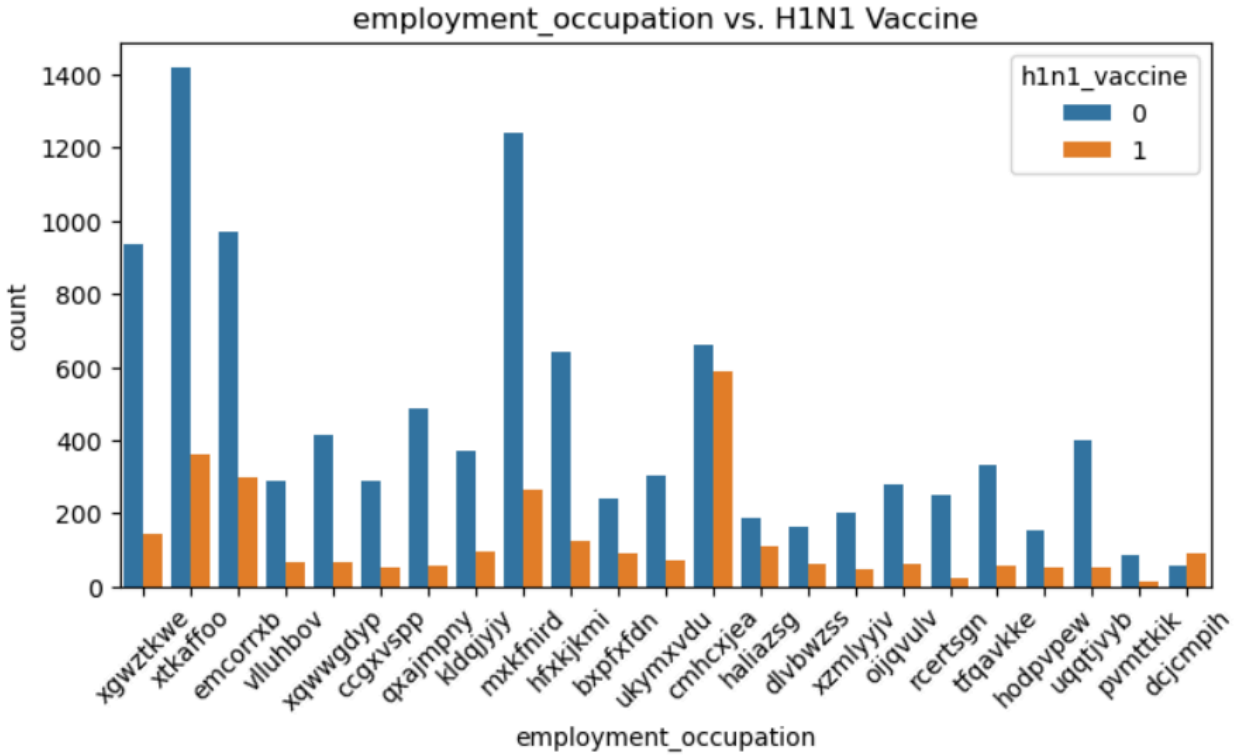


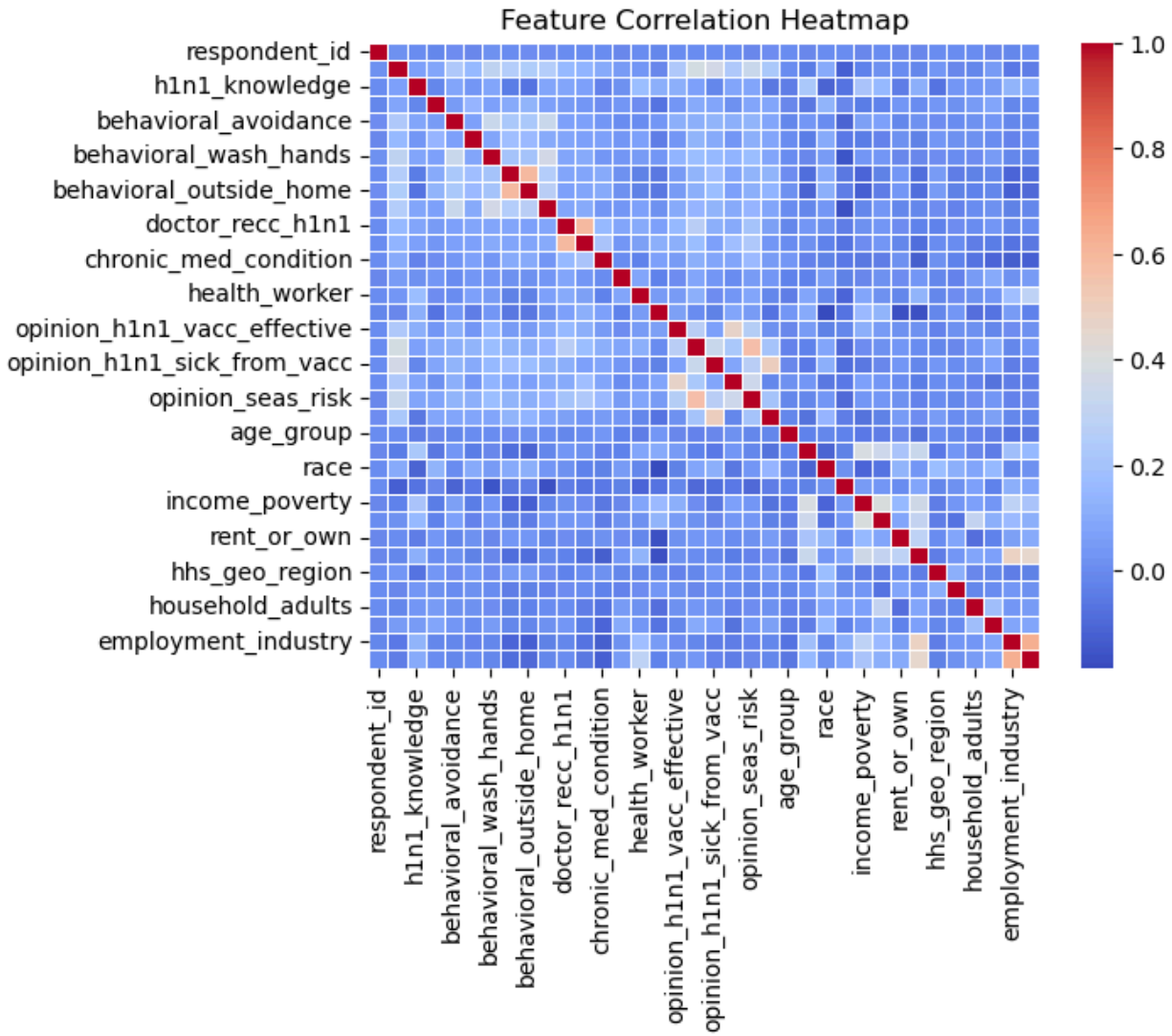

sex vs. H1N1 Vaccine
h1n1_vaccine
0
1
count
12000
10000
8000
6000
4000
2000
0
Female
Male
sex


income_poverty vs. H1N1 Vaccine
h1n1_vaccine
0
1
count
10000
8000
6000
4000
2000
0
Below Poverty
<= $75,000, Above Poverty
> $75,000
income_poverty


marital_status vs. H1N1 Vaccine
h1n1_vaccine
0
1
count
10000
8000
6000
4000
2000
0
Not Married
Married
marital_status


employment_status vs. H1N1 Vaccine
h1n1_vaccine
0
1
count
10000
8000
6000
4000
2000
0
Not in Labor Force
Employed
Unemployed
employment_status


employment_occupation vs. H1N1 Vaccine
h1n1_vaccine
0
1
count
1400
1200
1000
800
600
400
200
0
xgwztkwe
xtkaffoo
emcorrxb
vlluhbov
xqwwgdyp
ccgxvspp
qxajmpny
kldqjyjy
mxkfnird
hfxkjkmi
bxpfxfdn
ukymxvdu
cmhcxjea
haliazsg
dlvbwzss
xzmlyyjv
oijqvulv
rcertsgn
tfqavkke
hodpvpew
uqqtjvyb
pvmttkik
dcjcmpih
employment_occupation



Data Visualization Summary and Conclusion

After conducting exploratory data analysis (EDA) and visualizing the dataset, the following key insights were observed:

1. **Age Group Distribution:**
    - The dataset contains individuals from different age groups, with a higher proportion of middle-aged and older individuals.

    - The vaccination rates vary significantly across different age groups.
2. **Employment Status & Vaccination Rates:**
    - A noticeable trend was observed where unemployed individuals had different vaccination patterns compared to employed individuals.
3. **Education Level and Awareness:**
    - Higher education levels were generally correlated with higher vaccination rates, suggesting a possible link between awareness and vaccine acceptance.
4. **H1N1 and Seasonal Flu Vaccine Uptake:**
    - The proportion of people receiving the **H1N1 vaccine** was **lower** than those receiving the **seasonal flu vaccine**.
    - A significant number of individuals who took the **seasonal flu vaccine** also took the **H1N1 vaccine**, but the reverse was not as strong.

Conclusion from Data Visualization:

- **Demographic Factors:** Age, employment status, and education level play a crucial role in vaccine uptake.
- **Behavioral Trends:** Individuals who take the seasonal flu vaccine are more likely to take the H1N1 vaccine.
- **Healthcare Awareness:** People with higher awareness of flu risks show a higher vaccination rate.

Steps Followed in Building the Model

1. Data Preprocessing

- **Handling Missing Values:** Used **SimpleImputer** to fill missing values with the most frequent values.
- **Encoding Categorical Variables:** Converted categorical variables into numerical form using one-hot encoding.
- **Feature Selection:** Removed redundant or highly correlated features to improve model efficiency.
- Checked for outliers using boxplots and removed them without overfitting the model

2. Model Selection

Several models were trained and evaluated using **ROC-AUC** score:

- **Baseline Model:**
  - **Logistic Regression** achieved an **ROC-AUC of 0.8314 (H1N1)** and **0.8555 (Seasonal Flu).**
- **Tree-Based Models:**
  - **Random Forest:** Moderate improvement, but prone to overfitting.
  - **XGBoost & LightGBM:** Showed strong results due to their ability to handle categorical features and missing values.

3. Hyperparameter Tuning

- **LightGBM** was selected as the final model due to its high accuracy and efficiency.
- Optimized parameters:
  - num_leaves=20
  - n_estimators=200
  - max_depth=5

    - learning_rate=0.05
    - colsample_bytree=0.6
- Final **ROC-AUC Score for LightGBM:**
    - **H1N1 Vaccine: 0.865**
    - **Seasonal Flu Vaccine: 0.877**

The CSV file generated contained three columns:

1. **respondent_id** – The unique identifier for each individual in the test dataset.
2. **h1n1_vaccine** – The predicted probability that the individual received the H1N1 vaccine.
3. **seasonal_vaccine** – The predicted probability that the individual received the seasonal flu vaccine.

The **submission.csv** file provides key insights about the **likelihood of individuals receiving the H1N1 and seasonal flu vaccines** based on their demographic and behavioral features. Since the values in the h1n1_vaccine and seasonal_vaccine columns are probabilities (ranging from 0 to 1), they represent the model's confidence in whether each individual **received the vaccine**.

Key Insights from the CSV File

- A probability close to **1** means the model is very confident that the individual **took the vaccine**.
- A probability close to **0** means the model is confident that the individual **did not take the vaccine**.
- A probability around **0.5** means the model is uncertain and considers the individual as having a **50-50 chance** of taking the vaccine.