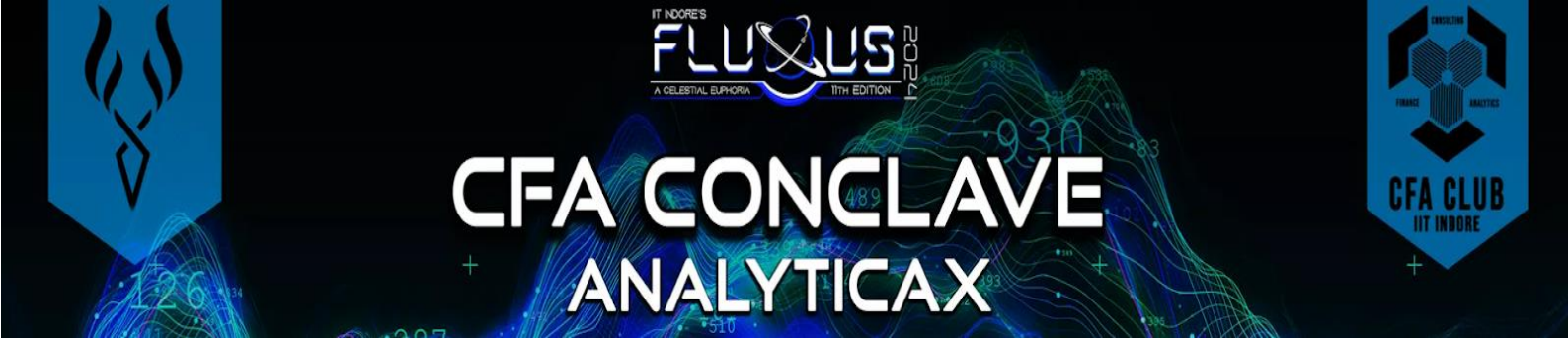


Project Report

→ Team

Neural Knights

1. Shreyans Katariya
2. Aditya Ghai

Github Link

→ Problem Statement

Anticipate the likelihood of individuals contracting **H1N1** and receiving their **yearly flu** vaccine. This entails forecasting two probabilities: one for vaccine_h1n1 and one for vaccine_seasonal. Each row in the dataset corresponds to an individual from the 2009 National H1N1 Flu Survey (NHFS) conducted by the CDC.

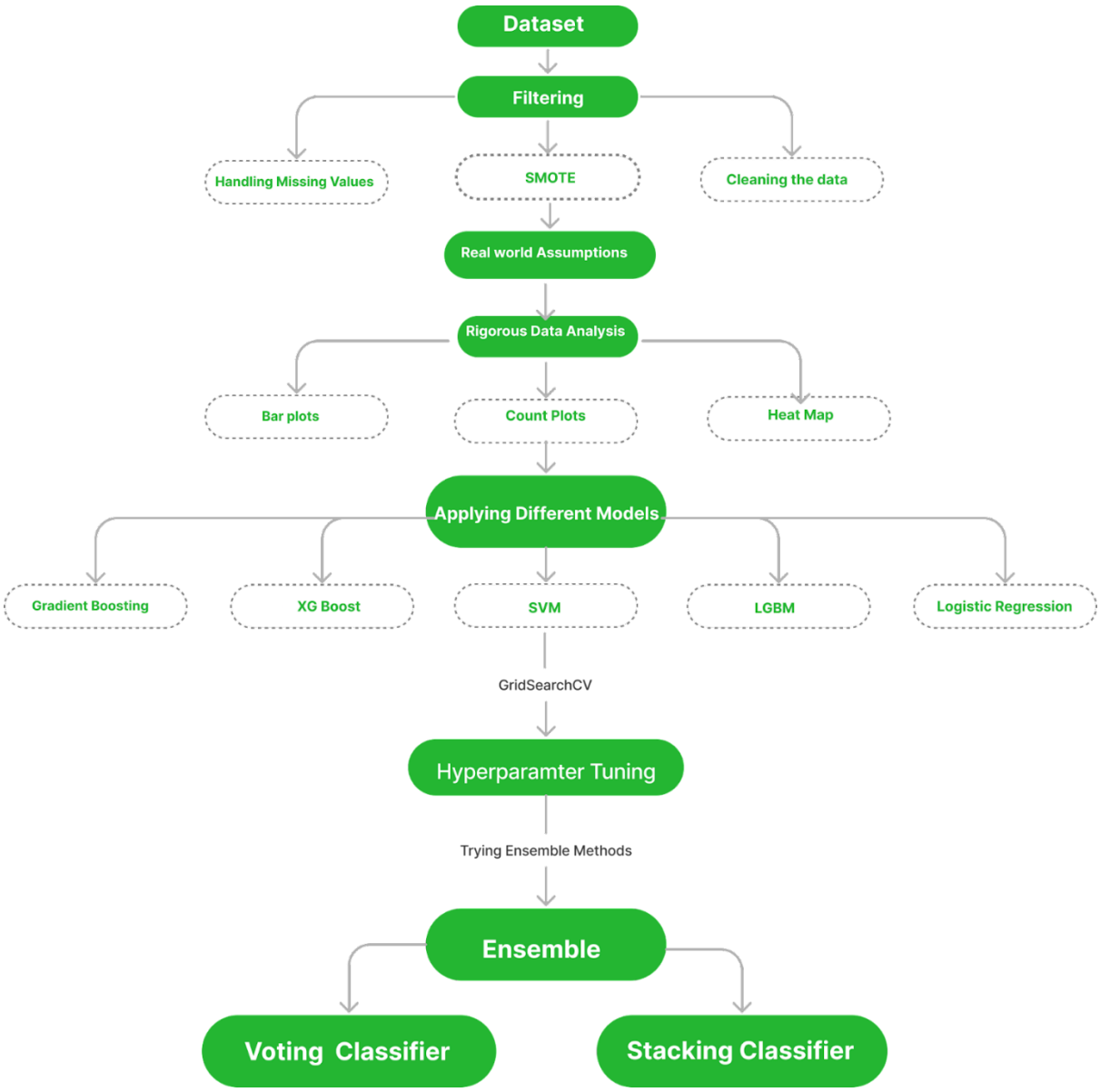
→ Required Prerequisites-

1. **Training set Features :(**Link**)**
2. **Training set Labels :(**Link**)**
3. **Test set Features:(**Link**)**

→ Dataset Description-

We were given three datasets for this task: **two training datasets** and **one testing dataset**. The training datasets consist of features and labels, with one containing the training features and the other containing the corresponding training labels. The objective is to predict the probabilities of individuals receiving vaccines.

Architecture

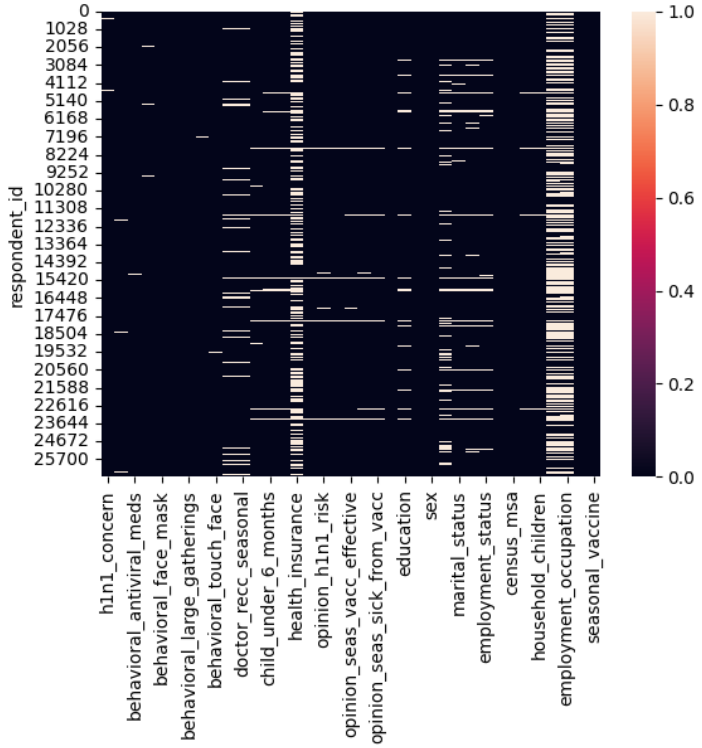


→ Data Handling

First we have seen the data through heatmap to see the missing values and we found some of the conclusions by just looking at it.

Heatmaps-

The heatmap analysis reveals several insights:



1)Columns like health_insurance, employment_industry, and employment_occupation have **many missing values.**

2)Income_poverty also shows a significant amount of missing data.

3)Doctor_recc_h1n1 and doctor_recc_seasonal have the same proportion of missing values.

After some **Analysis of missing values** we discovered -

1. Missing values in education often coincide with missing values in income_poverty, marital_status, rent_or_own, employment_status, employment_industry, and employment_occupation columns.

2. Rows with missing values in doctor_recc_seasonal also have missing values in doctor_recc_h1n1.

3. Missing values in employment_status lead to missing values in employment_industry and employment_occupation columns, especially when categorized as "**Not in Labor Force**" or "**Unemployed**."

→ Data preprocessing

- In the data preprocessing phase, we first address missing values in the employment_status column by leveraging the observed relationship with employment_industry and employment_occupation columns.

- Given that missing values in employment_status correspond to missing values in these related columns, we confidently impute "**Not in Labor Force**" as the replacement value, aligning with similar cases of "**Not in Labor Force**" or "**Unemployed**" statuses. This choice is supported by its status as the second most frequent value in the column. To avoid introducing bias, missing values in respondent opinion columns are uniformly imputed with the value 3, representing "Don't know." For the remaining missing values in both the train and test datasets, which are mostly missing at random (MAR), we fill them with the mode of each respective column.**Also Encoding is done for Nominal and Ordinal Columns.**

- Since the dataset consists of categorical variables, using the mode as the imputation strategy helps maintain the categorical nature of the data. To ensure compatibility with various machine learning algorithms and statistical techniques, columns with dtype float64 are converted to int64, effectively transforming these variables from continuous to discrete.

- The presence of a single duplicated row is considered negligible and is safely ignored without compromising the integrity or accuracy of the results. Additionally, normalization is deemed unnecessary as the dataset exclusively consists of categorical columns, which do not possess a natural numerical order or scale. When categorical variables are represented as ranges or intervals, such as **age_group**, assigning a single value to each category reduces information loss and provides a more precise estimate for analysis, treating the variable as if it takes on a continuous scale with some level of discretization.

→ Correlating with real world scenarios-

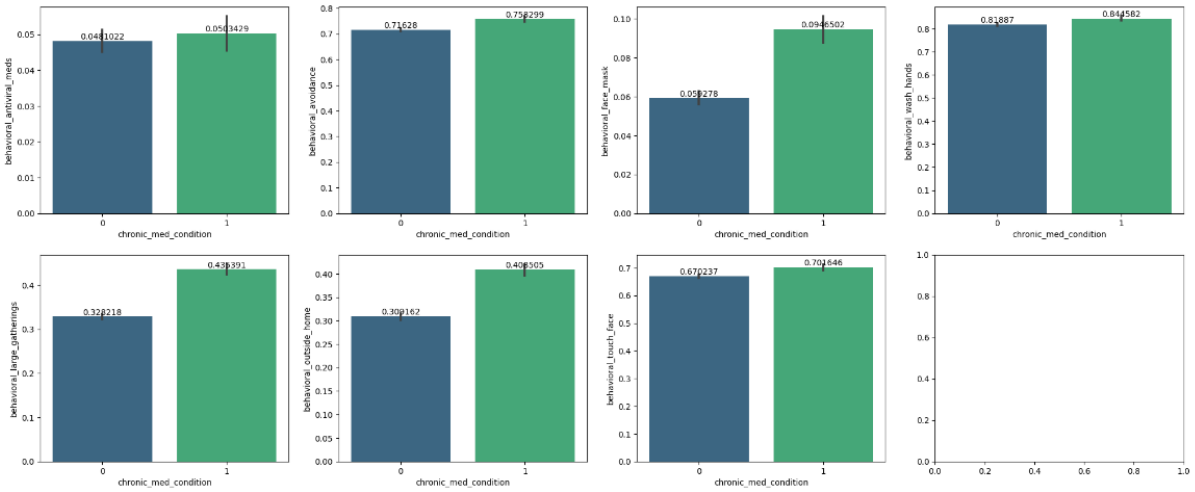
The initial step involved understanding the dataset and interpreting the meaning of each column. We then established logical connections, such as the likelihood of healthcare workers being more inclined to receive vaccines. This deduction was based on real-life scenarios and assumptions drawn from the data.We logically made many assumptions and then performed **EDA** to see that the assumptions would apply on the dataset or not.

→ Rigorous data analysis

Here are the key observations regarding factors influencing vaccine uptake after doing sharp data analysis :

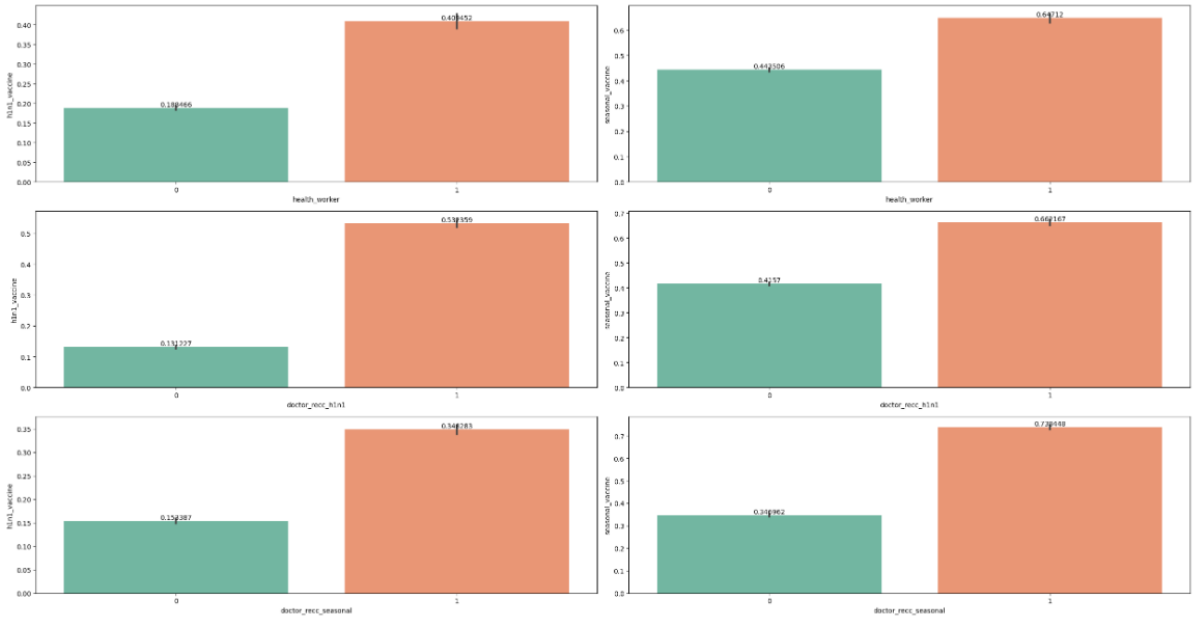
- **Chronic Medical Conditions:** Individuals with chronic medical conditions are more inclined to take preventive measures against flu-like illnesses, indicating a higher likelihood of vaccine uptake.

- **Specific Employment Industries and Occupations:** Certain employment industries and occupations exhibit higher vaccine uptake rates for both H1N1 and seasonal flu vaccines, suggesting a correlation between occupational factors and vaccine acceptance.



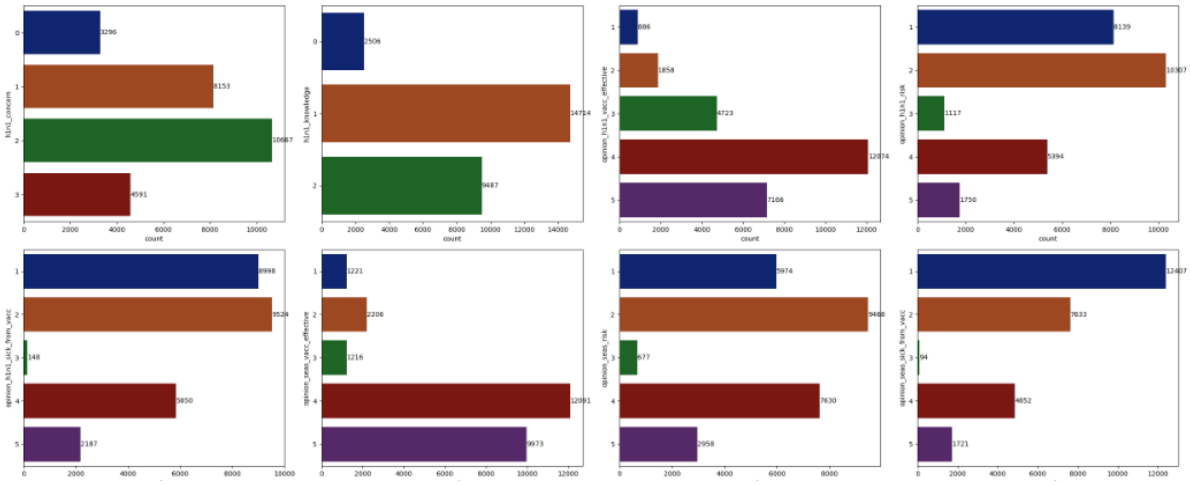
- **Healthcare Workers:** Healthcare workers demonstrate a higher propensity to receive both vaccines, underscoring the influence of profession on vaccine uptake.

- **Doctor Recommendations:** Vaccine uptake is significantly influenced by doctor recommendations, highlighting the importance of healthcare providers in promoting vaccination.



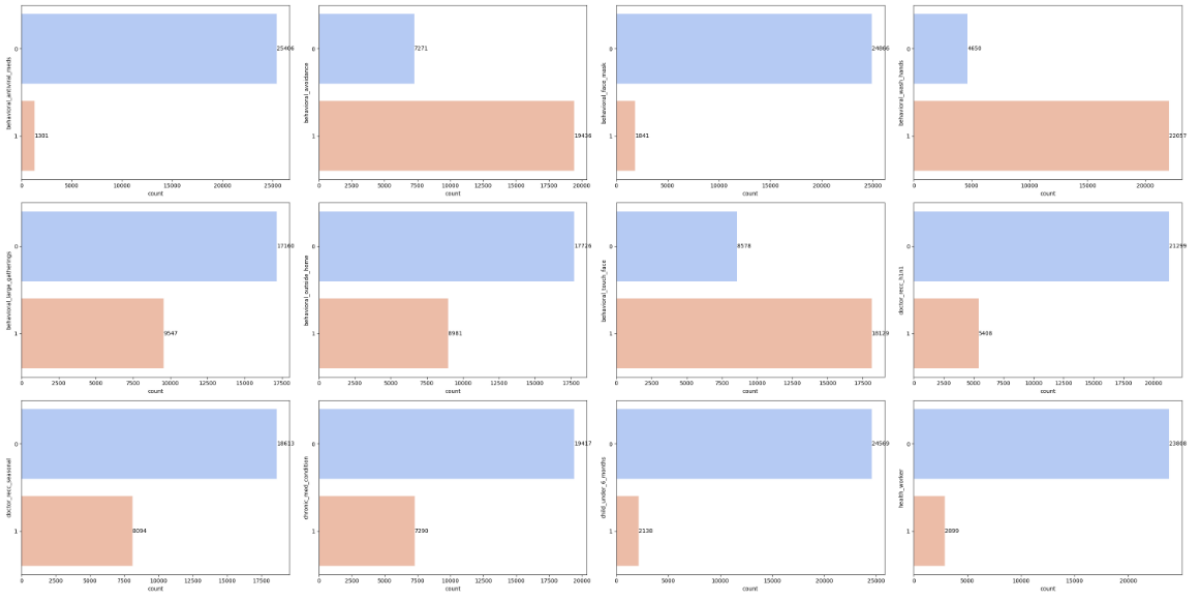
- **Perceived Vaccine Effectiveness and Risk Perception:** Individuals' perceptions of vaccine effectiveness and their perceived risk of illness play pivotal roles in vaccine acceptance.

- **Other Factors**: Additional factors such as concern about H1N1, older age, not being in the labor force, and absence of children in the household are associated with higher seasonal flu vaccine uptake, indicating diverse influences on vaccination behavior.



These observations collectively emphasize the multifaceted nature of vaccine acceptance, influenced by various individual, occupational, and healthcare-related factors.

Bivariate Analysis



The bar plot indicates a notable pattern where individuals with specific chronic medical conditions, such as asthma, diabetes, heart conditions, kidney conditions, sickle cell anemia, neurological or neuromuscular conditions, liver conditions, or weakened immune systems due to chronic illnesses or medications, tend to exhibit more proactive behavioral precautions to prevent the occurrence of flu-like illnesses.

→ Applying the models

After rigorously exploring diverse modeling approaches, including both traditional algorithms and advanced ensembling techniques, we witnessed remarkable variations in predictive performance. Among the arsenal of models deployed, our analysis uncovered insights into the effectiveness of certain methodologies. For instance, our foray into **Gradient Boosting** revealed its exceptional performance, boasting a 10-fold ROC-AUC score of 0.8698 for the H1N1 vaccine prediction task. This technique, known for its ability to sequentially minimize errors, showcased robust predictive power, harnessing the collective strength of individual decision trees. Similarly, the **Light Gradient Boost model**, a variant of Gradient Boosting, exhibited comparable efficacy with a score of 0.8696, underscoring the versatility and reliability of boosting algorithms. These findings underscore the importance of model selection and parameter tuning in optimizing predictive accuracy, reaffirming our commitment to employing a diverse array of methodologies to unlock deeper insights from the dataset.

→ The summary of the models -

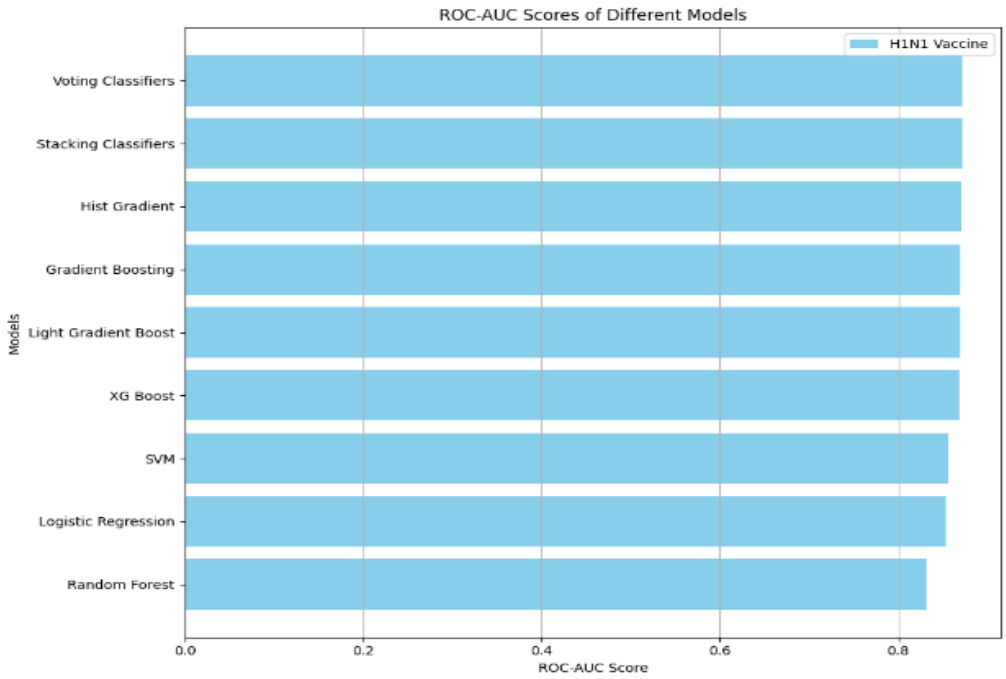
The 10 fold ROC-AUC score of different models which we applied are as follows -

Model name	H1N1_Vaccine	Seasonal_vaccine
Random Forest	0.8319	0.8210
Logistic Regression	0.8534	0.8459
SVM	0.8559	0.8520
Gradient Boosting	0.8698	0.8635
XG Boost	0.8683	0.8617
Hist Gradient	0.8701	0.8629
LGBM	0.8704	0.8631

Model name	H1N1_Vaccine	Seasonal_vaccine

Stacking Classifiers	0.8716	0.8647
Voting Classifiers	0.8717	0.8646

To visualize it better



→ <u>Conclusion</u>

In conclusion, our comprehensive analysis of the 2009 National H1N1 Flu Survey dataset, augmented by rigorous data preprocessing and insightful correlation with real-world scenarios, has provided invaluable insights into the factors influencing individuals' likelihood of contracting the H1N1 virus and receiving seasonal flu vaccines. Through meticulous data cleaning and preprocessing steps, we addressed missing values and ensured the integrity of the dataset for further analysis. By correlating dataset attributes with real-world scenarios, such as the influence of healthcare professions on vaccine uptake, we gained a deeper understanding of the underlying dynamics driving vaccination behavior.

Our exploration of various predictive modeling techniques, ranging from traditional algorithms to advanced ensembling methods, yielded promising results. Notably, gradient boosting techniques demonstrated exceptional predictive performance, with the Light Gradient Boost model showcasing a **10-fold ROC-AUC score** of 0.8696 for the H1N1 vaccine prediction task. These findings underscore the efficacy of boosting algorithms in capturing complex patterns within the dataset.

Additionally, our ensemble models, including **Stacking Classifiers** and **Voting Classifiers**, exhibited further improvements in predictive accuracy, achieving ROC-AUC scores of

0.8716 and **0.8717**, respectively. This highlights the effectiveness of combining multiple models to leverage their individual strengths and enhance overall predictive performance.

In essence, our analysis underscores the multifaceted nature of vaccine acceptance, influenced by a myriad of individual, occupational, and healthcare-related factors. By leveraging advanced analytical techniques and domain knowledge, we have not only developed accurate predictive models but also deepened our understanding of vaccination behavior, paving the way for more targeted public health interventions and strategies to combat infectious diseases.

----------Thank You----------