Finally, I took a step further and tried ensemble methods that combine the predictions of multiple models to see if I could improve the performance of my model even further. I compared the performance of individual models, including logistic regression, decision tree, and random forest, and found that the ensemble method showed improved accuracy, precision, recall, and F1 score compared to the individual models.

After considering the trade-off between complexity and interpretability, I concluded that the improved performance of the ensemble model was worth the trade-off in complexity and decided to use this as my final model for solving the problem. I will continue to monitor its performance and make any necessary adjustments to ensure the best results.

# Predicting H1N1 and Seasonal Flu Vaccine Uptake

# INTRODUCTION

The project aims to identify the correlation between various factors and vaccine uptake. By understanding the factors that influence an individual's decision to receive a vaccine, public health efforts can be better directed toward increasing vaccine coverage and reducing the spread of flu-related illnesses.

The project focuses on a binary classification problem and the target variable is the receipt of either H1N1 or seasonal flu vaccine. The project will utilize a combination of logistic regression, decision trees, random forests, and ensemble methods to develop the machine-learning models and evaluate their performance.

# FEATURE SELECTION

In this project, I aimed to build a model to solve a problem. I started by performing feature selection, which is the process of selecting the most important features in the data that have the greatest impact on the target variable. I used three techniques for feature selection: Lasso Regression, Random Forest, and XGBoost.

However, after further analysis and consideration, I have decided to add additional features to this list. I believe that the information represented by these features is important for accurately predicting the outcome, and will further strengthen the model. These features include 'sex', 'child_under_6_months', 'chronic_med_condition', 'health_insurance', and 'race'. These features provide additional context to the data and can help capture patterns that may be missed by relying solely on the automatically selected features.

# MODEL TRAINING AND EVALUATION

After performing feature selection, I trained different models, including Logistic Regression, Decision Tree, and Random Forest, using these selected features. I then evaluated the performance of each model using accuracy, precision, recall, F1 score, and other relevant evaluation metrics.

As I was working on the classification problem, I trained three different models, which are Logistic Regression, Decision Tree, and Random Forest. These models were trained using the feature set that was selected through the intersection of Lasso, XGBoost and Random Forest methods.

The performance of each model was evaluated using various metrics such as Accuracy, Precision, Recall, and F1 Score. The results showed that the Decision Tree model had the highest Accuracy of 0.837, followed by Logistic Regression and Random Forest with 0.836.

In terms of Precision, Logistic Regression had a score of 0.697, while Decision Tree and Random Forest had scores of 0.693 and 0.682, respectively. In terms of Recall, Decision Tree had a score of 0.413, while Logistic Regression and Random Forest had scores of 0.401 and 0.426, respectively.

The F1 Score, which is the harmonic mean of Precision and Recall, was highest for Decision Tree with a score of 0.518, followed by Logistic Regression and Random Forest with scores of 0.509 and 0.524, respectively.

Based on these results, I concluded that the Decision Tree model performed the best among the three models, with the highest Accuracy, Precision, and F1 Score. Hence, I decided to use this model as my final model for solving the problem. However, I will continue to monitor the performance and make any necessary adjustments to ensure the best results.

# CONCLUSION

After evaluating the performance of multiple machine learning models, including logistic regression, decision tree, and random forest, I found that using an ensemble method provided the best results in predicting whether a person received the H1N1 or seasonal flu vaccine. The model accurately considered various factors such as personal information, opinions, and health behaviors, in making its predictions.

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 vaccine coverage and reducing the spread of flu-related illnesses.

In conclusion, the ensemble method proved to be an effective approach in accurately predicting vaccine uptake based on the factors considered in the National 2009 H1N1 Flu Survey.