**ML Report**

**Introduction:**

The National 2009 H1N1 Flu Survey was conducted to collect information on personal opinions, health behaviors, and demographic characteristics of individuals regarding the H1N1 and seasonal flu vaccines. The goal of this project is to develop a machine-learning model that can accurately predict whether a person received the H1N1 or seasonal flu vaccine based on the data collected in the survey.

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.

The following report provides a detailed overview of the data pre-processing, model development, and evaluation steps taken to achieve the project's goal.

**Exploratory Data Analysis**

The data consists of two data frames, features\_df and labels\_df, where features\_df represents the independent variables and labels\_df represents the dependent variables. The features\_df data frame has a shape of 26707 rows and contains 35 columns with information about the respondents. The data types in the features\_df data frame include float64 for 23 columns and object for 12 columns.

Here's a brief description of each column in the **features\_df** data frame:

* **h1n1\_concern**: Represents the level of concern a person has regarding the H1N1 virus.
* **h1n1\_knowledge**: Represents a person's level of knowledge about the H1N1 virus.
* **behavioral\_antiviral\_meds**: Indicates whether a person has taken antiviral medications in response to the H1N1 virus.
* **behavioral\_avoidance**: Indicates the extent to which a person has altered their behavior in response to the H1N1 virus.
* **behavioral\_face\_mask**: Indicates whether a person is using face masks in response to the H1N1 virus.
* **behavioral\_wash\_hands**: Indicates whether a person is washing their hands more frequently in response to the H1N1 virus.
* **behavioral\_large\_gatherings**: Indicates whether a person is avoiding large gatherings in response to the H1N1 virus.
* **behavioral\_outside\_home**: Indicates whether a person is spending less time outside their home in response to the H1N1 virus.
* **behavioral\_touch\_face**: Indicates whether a person is touching their face less frequently in response to the H1N1 virus.
* **doctor\_recc\_h1n1**: Indicates whether a person's doctor has recommended they receive the H1N1 vaccine.
* **doctor\_recc\_seasonal**: Indicates whether a person's doctor has recommended they receive the seasonal flu vaccine.
* **chronic\_med\_condition**: Indicates whether a person has a chronic medical condition.
* **child\_under\_6\_months**: Indicates whether a person has a child under 6 months of age.
* **health\_worker**: Indicates whether a person is a health worker.
* **health\_insurance**: Indicates whether a person has health insurance.
* **opinion\_h1n1\_vacc\_effective**: Represents a person's opinion on the effectiveness of the H1N1 vaccine.
* **opinion\_h1n1\_risk**: Represents a person's perception of the risk posed by the H1N1 virus.
* **opinion\_h1n1\_sick\_from\_vacc**: Represents a person's opinion on the risk of getting sick from the H1N1 vaccine.
* **opinion\_seas\_vacc\_effective**: Represents a person's opinion on the effectiveness of the seasonal flu vaccine.
* **opinion\_seas\_risk**: Represents a person's perception of the risk posed by the seasonal flu.
* **opinion\_seas\_sick\_from\_vacc**: Represents a person's opinion on the risk of getting sick from the seasonal flu vaccine.
* **age\_group**: Represents a person's age group.
* **education**: Represents a person's level of education.
* **race**: Represents a person's race.
* **sex**: Represents a person's gender.
* **income\_poverty**: Represents a person's income relative to the poverty line.
* **marital\_status**: Represents a person's marital status.
* **rent\_or\_own**: Indicates whether a person rents or owns their home.

The labels\_df data frame has a shape of 26707 rows and 2 columns and includes information about whether the respondents received the H1N1 and seasonal flu vaccines, represented by the h1n1\_vaccine and seasonal\_vaccine columns, respectively.

**Dealing with missing values:**

The data had a couple of missing values as shown below:

| **Column Name** | **Number of Missing Values** | **Percentage of Missing Values** |
| --- | --- | --- |
| **h1n1\_concern** | **92** | **0.34%** |
| **h1n1\_knowledge** | **116** | **0.43%** |
| **behavioral\_antiviral\_meds** | **71** | **0.27%** |
| **behavioral\_avoidance** | **208** | **0.78%** |
| **behavioral\_face\_mask** | **19** | **0.07%** |
| **behavioral\_wash\_hands** | **42** | **0.16%** |
| **behavioral\_large\_gatherings** | **87** | **0.33%** |
| **behavioral\_outside\_home** | **82** | **0.31%** |
| **behavioral\_touch\_face** | **128** | **0.48%** |
| **doctor\_recc\_h1n1** | **2160** | **8.10%** |
| **doctor\_recc\_seasonal** | **2160** | **8.10%** |
| **chronic\_med\_condition** | **971** | **3.65%** |
| **child\_under\_6\_months** | **820** | **3.08%** |
| **health\_worker** | **804** | **3.01%** |
| **health\_insurance** | **12274** | **46.08%** |
| **opinion\_h1n1\_vacc\_effective** | **391** | **1.47%** |
| **opinion\_h1n1\_risk** | **388** | **1.46%** |
| **opinion\_h1n1\_sick\_from\_vacc** | **395** | **1.48%** |
| **opinion\_seas\_vacc\_effective** | **462** | **1.73%** |
| **opinion\_seas\_risk** | **514** | **1.93%** |
| **opinion\_seas\_sick\_from\_vacc** | **537** | **2.01%** |
| **education** | **1407** | **5.28%** |
| **income\_poverty** | **4423** | **16.59%** |
| **marital\_status** | **1408** | **5.28%** |
| **rent\_or\_own** | **2042** | **7.65%** |
| **employment\_status** | **1463** | **5.46%** |
| **household\_adults** | **249** | **0.93%** |
| **household\_children** | **249** | **0.93%** |
| **employment\_industry** | **13330** | **49.98%** |
| **employment\_occupation** | **13470** | **50.38%** |

The total number of values in the features dataframe is 26,707**.**

The mode, which is the most frequent value in a column, is a commonly used method for filling missing values because it does not greatly impact the distribution of the data. In this case, as the missing values were filled with the mode, it is likely that the resulting data accurately represents the majority of the values in each column.

However, it is important to consider the percentage of missing values in each column before deciding to fill with the mode. If the percentage of missing values is high, the mode may not accurately reflect the true distribution of the data and may result in skewed or inaccurate conclusions. In such cases, it may be more appropriate to consider alternative methods for handling the missing data, such as imputation methods or removing the rows with missing values.

As a result, the **employment\_occupation** and **employment\_industry** were dropped.

**Outliers**

As I was exploring the data, I noticed that there were outliers present. Outliers are data points that lie significantly away from the rest of the data and can have a significant impact on the results of the model. In this case, removing the outliers could result in the loss of important information, as they may represent a separate group or subclass of data.

Given that the data is entirely categorical feature encoded, it was important to consider the implications of removing the outliers. After careful consideration, I decided to keep the outliers in the data, as they could provide valuable insight into the problem I am trying to solve.

It is worth noting that outliers should always be handled with caution and a thorough understanding of the implications of removing or keeping them. In this case, keeping the outliers was the best decision given the nature of the data and the problem I was trying to solve.

**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.

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.

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

XGBoost is a gradient-boosting algorithm that combines multiple weak models to produce a strong model. This technique also helps us identify the most important features, but instead of calculating feature importance scores, it directly optimizes the model's performance.

I combined the features selected from each of the three techniques to get an intersection of the most important features. This intersection represents the features that were deemed important by all three techniques, providing a more robust and reliable feature selection process.

After performing automatic feature selection using the intersection of Lasso, XGBoost, and Random Forest, I ended up with a list of seven important features: ['h1n1\_knowledge', 'opinion\_h1n1\_vacc\_effective', 'opinion\_seas\_risk', 'health\_worker', 'opinion\_h1n1\_risk', 'age\_group', 'doctor\_recc\_h1n1'].

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.

By adding these features, I am making a conscious effort to increase the granularity of the data and provide the model with more information to work with. I am confident that this decision will result in a more robust and accurate model, and will enable me to make more informed predictions.

**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.

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.

**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.