H1N1 VACCINE PREDICTION

PRESENTED BY:
MERCY SILALI

PROJECT OVERVIEW

The goal of this project is to predict whether individuals received the H1N1 flu vaccine based on data collected from the National 2009 H1N1 Flu Survey. By analyzing various features from the dataset, we aim to build a predictive model that can identify the likelihood of someone having received the H1N1 vaccine. This information is valuable for public health officials seeking to understand vaccine uptake patterns and tailor future vaccination campaigns.

PROJECT OBJECTIVES

1. FEATURE ENGINEERING AND SELECTION: Choosing the features that are likely to affect the h1n1 vaccine uptake
2. MODEL SELECTION: Comparing different models to find the best one.
3. PERFORMANCE EVALUATION: Measuring how well a model performs on training data

The overall goal is to predict vaccine uptake.

DATA OVERVIEW

This particular dataset is part of a competition designed to predict flu vaccination patterns, drawing from the National 2009 H1N1 Flu Survey. The data includes comprehensive details about survey respondents, encompassing their health behaviors, opinions on flu vaccines, and demographic information.

The dataset includes various features related to respondents' backgrounds and behaviors.

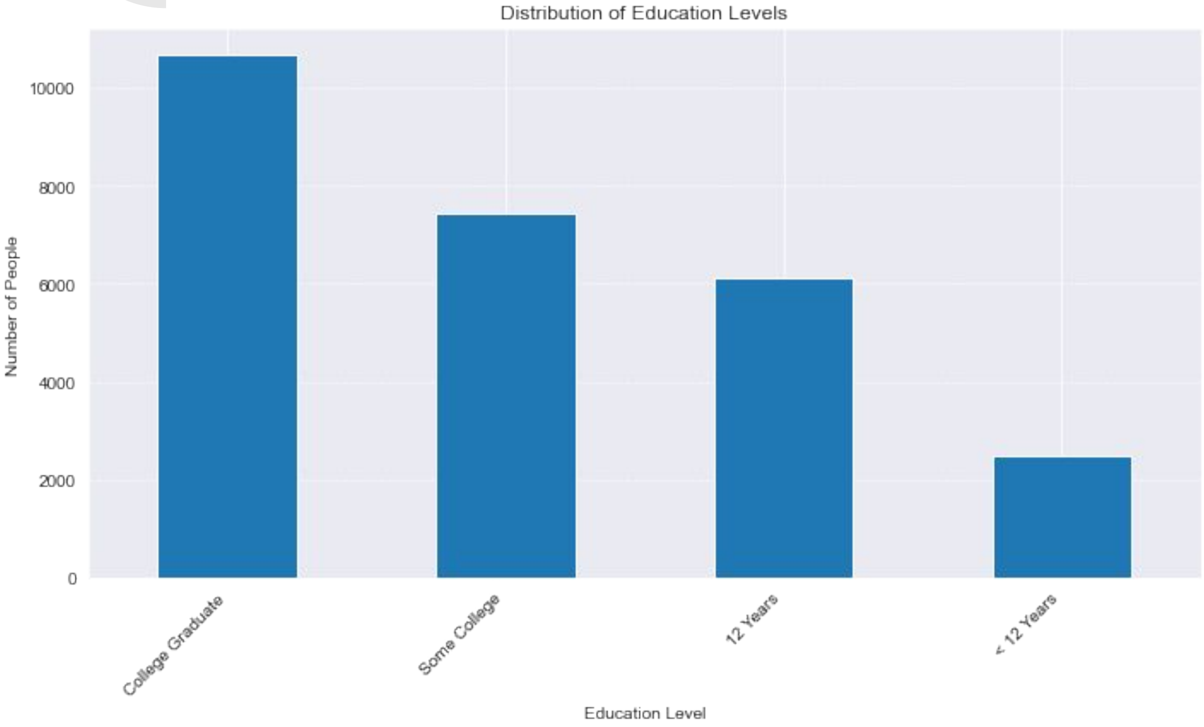
Target Variable: H1N1 Vaccine Receipt

DATA CLEANING

The data was cleaning by imputing missing values in numerical columns with the median while the categorical columns were filled using the forward fill method.

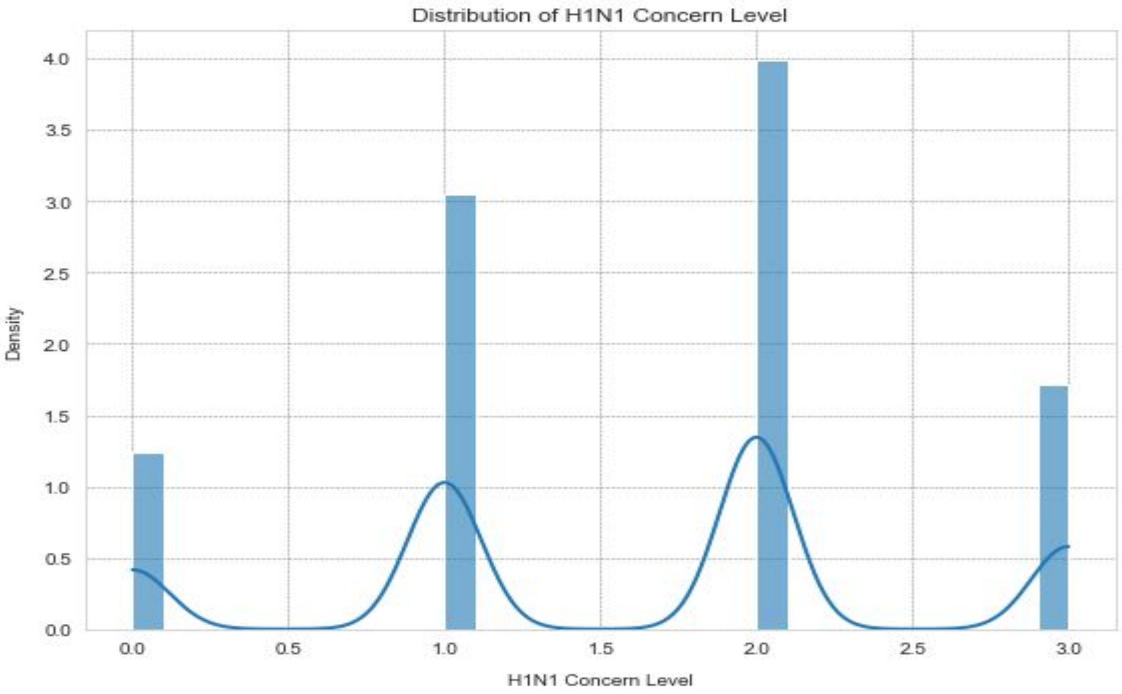
All the columns were retained as they were all important in the prediction of H1N1 vaccine uptake.

EXPLEORATORY DATA ANALYSIS



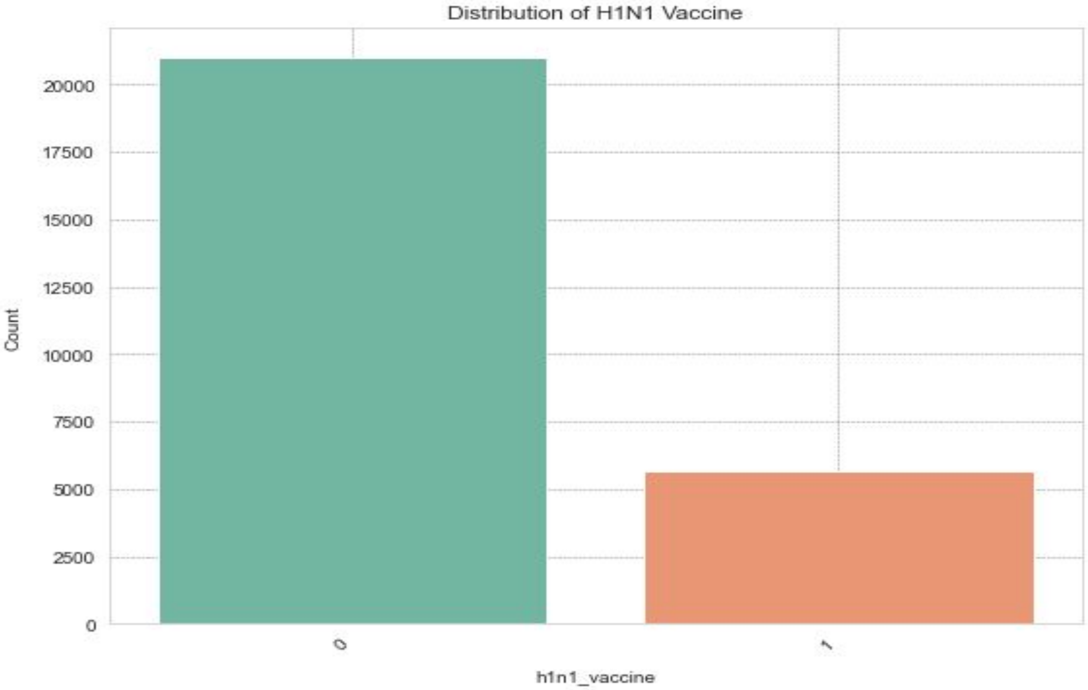
Most of the individual have graduated from college while a few have less that 12 year of school experience.

H1N1 CONCERN DISTRIBUTION



The individuals have different concern for h1n1_vaccine. Most people have little to no knowledge and the other half are somewhat knowledgeable on h1n1_vaccine.

DISTRIBUTION OF H1N1 VACCINE



From the countplot, majority of the people did not receive the H1N1 vaccine. The bar corresponding to '0'(no vaccine) is significantly taller the the bar '1'(received vaccine). This indicates that majority of the individual in this dataset did not receive the H1N1 vaccine.

MODELLING

The models employed in this prediction analysis were: **Logistic Regression and Decision Trees**

The data was first preprocessed by scaling and one hot encoding in order for it to be used in a machine learning algorithm.

LOGISTIC REGRESSION

Model Performance:

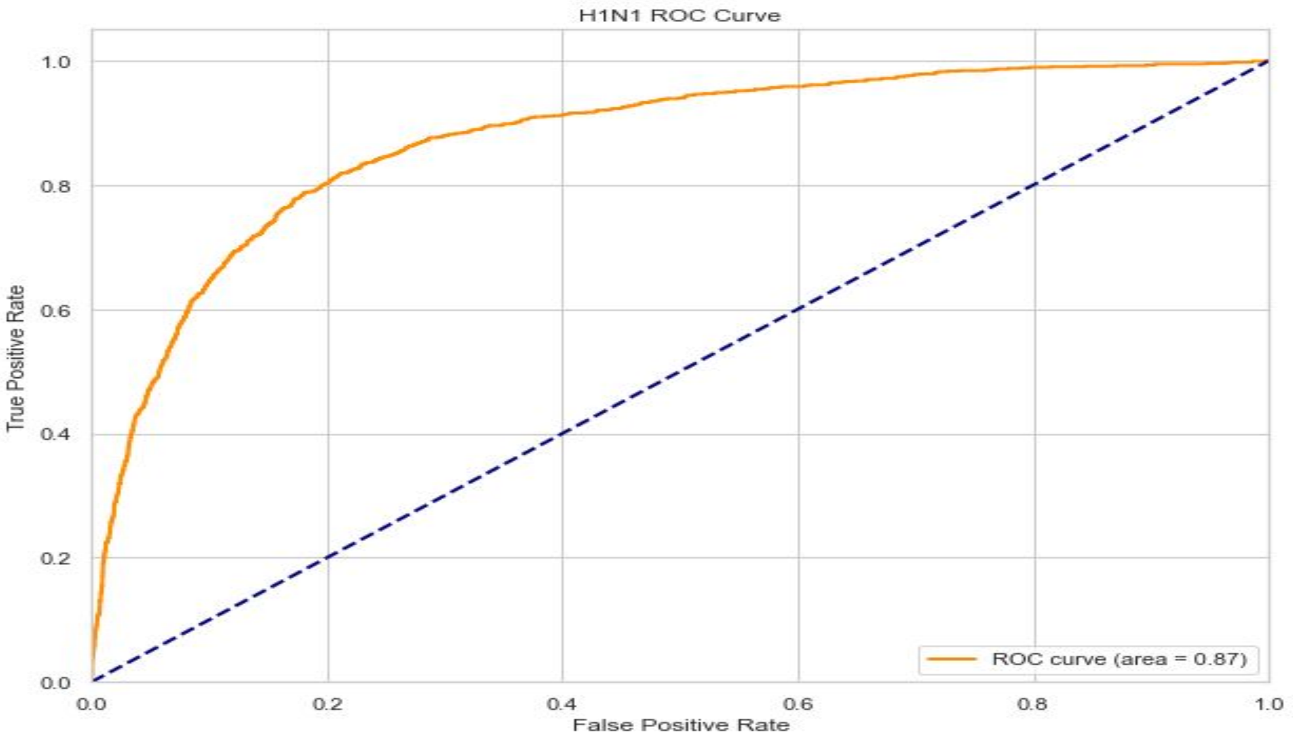
- The model is generally accurate but might miss some positive cases.
- It's good at correctly predicting negative cases.

Confusion Matrix Breakdown:

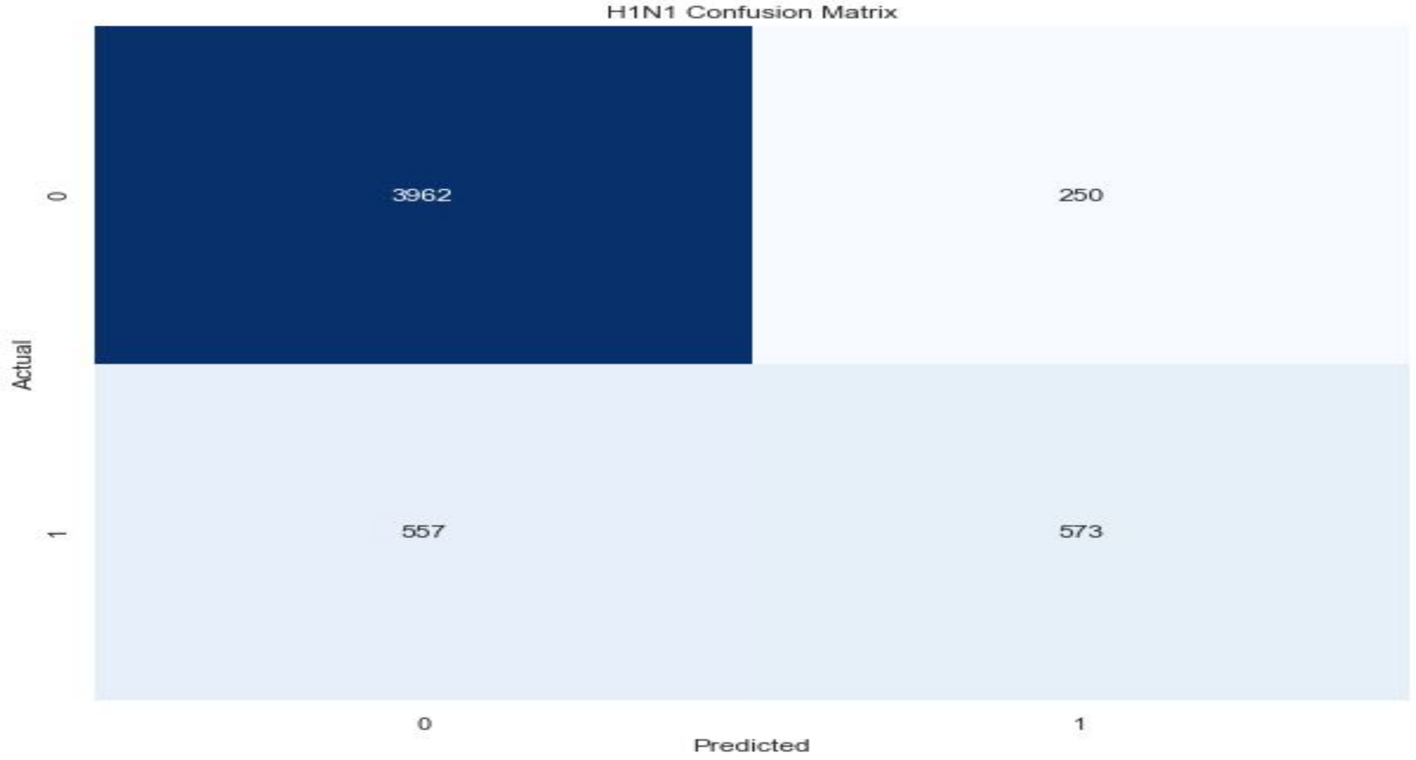
- Many cases were correctly predicted.
- Some cases were incorrectly predicted.

These is a shown in the visualisations below.

ROC-AUC CURVE FOR LOGISTIC REGRESSION



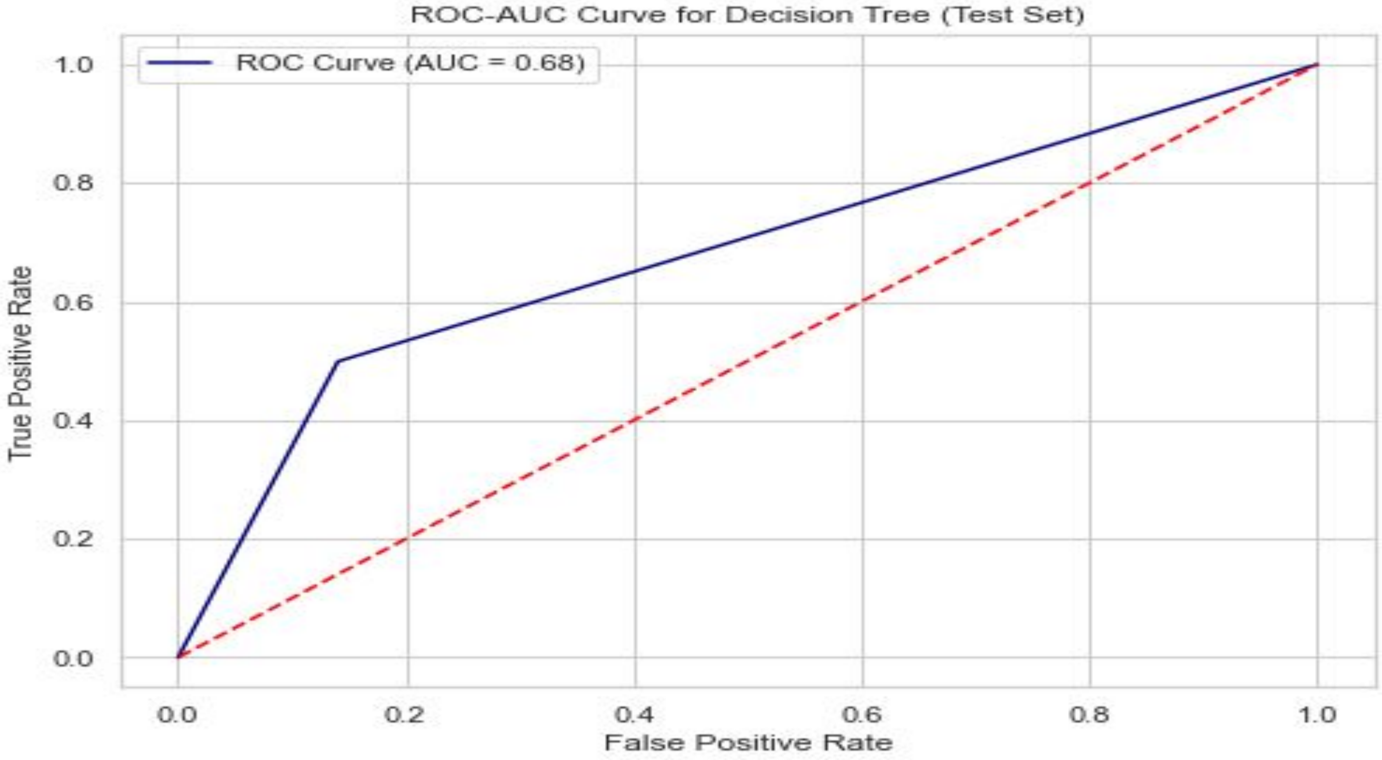
CONFUSION MATRIX



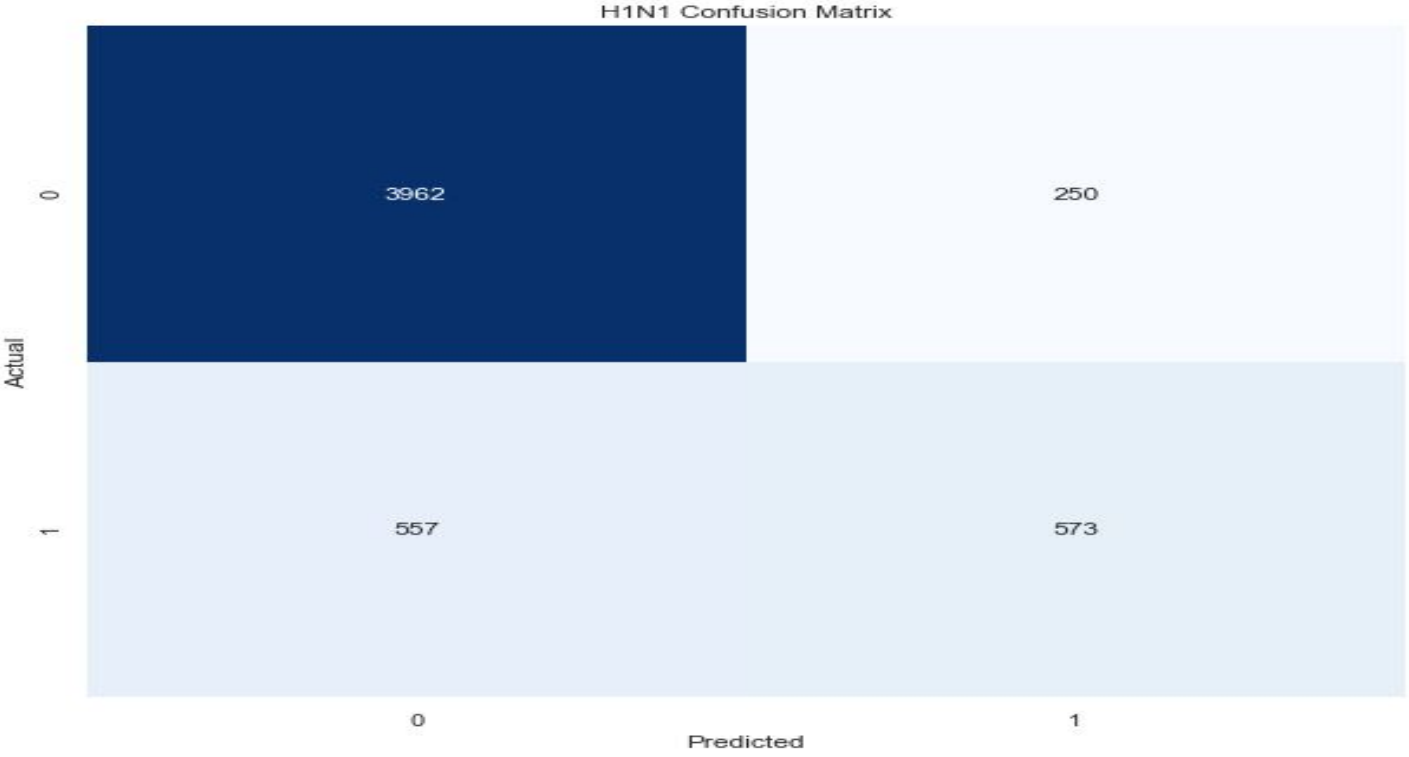
DECISION TREES

From the confusion matrix: Accuracy: 81% of the samples were correctly classified. Precision: 49% of the samples predicted as positive were actually positive. Recall: 50% of the actual positive samples were correctly predicted. F1-score: The model's overall performance is considered moderate, balancing precision and recall.
AUC = 0.68: This indicates that the model's performance is moderately good. Since the model's curve lies above the random guess line, it confirms that the model is performing better than random chance.

ROC-AUC CURVE



CONFUSION MATRIX



MODEL COMPARISON

Overall the first model (logistic regression model) performed better in predicting the likelihood of an individual to get h1n1 vaccine than the other models.

The results of the prediction of the likelihood of an individual to get h1n1 vaccine using logistic regression model are as follows: The model had a high accuracy of 0.85 while the decision trees model are as follows: The model had a high accuracy of 0.78. However, the logistic regression model outweighs both models with an accuracy of 0.85.

RECOMMENDATIONS

1. Hyperparameter Tuning: Conduct a more thorough hyperparameter optimization process to identify the optimal values for parameters such as regularization strength, learning rate, and number of iterations.
2. Feature Engineering: Explore creating new features or transforming existing ones to better capture relevant patterns and improve predictive power.
3. Collecting more up to date data.
4. Model Selection: Consider alternative models, such as Random Forest.

NEXT STEPS

1. Continuous Monitoring: Implement a system to monitor the model's performance in a production environment and identify any degradation over time.
2. Retraining: Regularly retrain the model on new data to ensure it remains accurate and up-to-date.
3. Further analysis to determine the specific features that affect the vaccination of individuals
4. Deployment of the best performing model