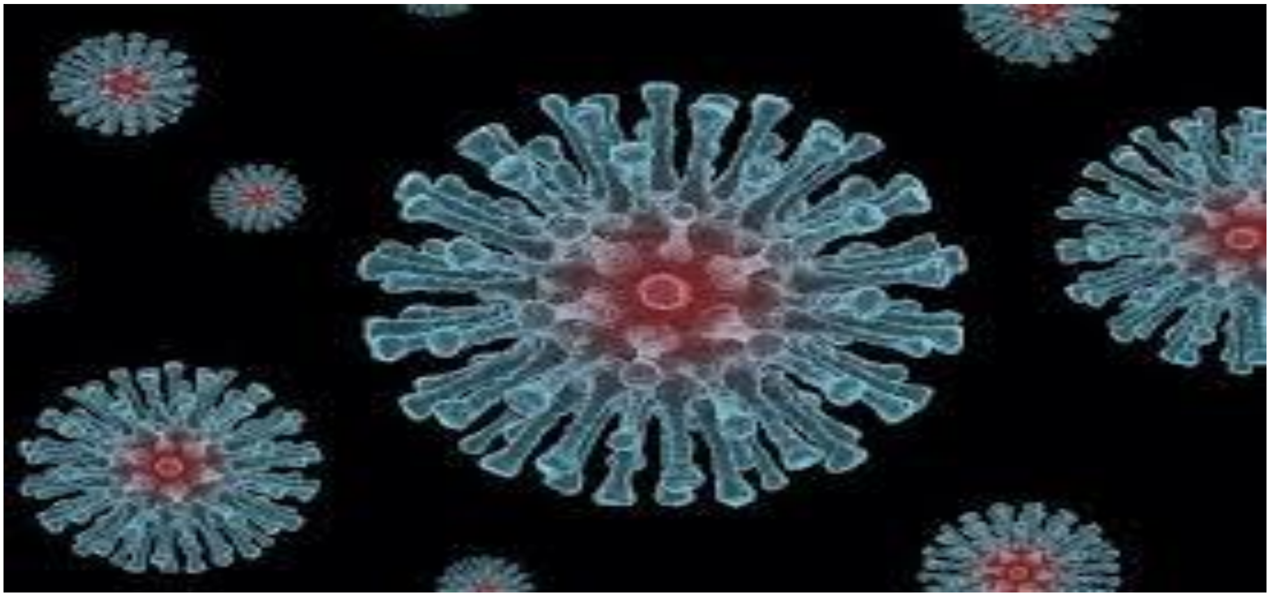
Exploring Determinants of H1N1 Vaccine Uptake: A Data-Driven Public Health Approach Using Classification Models



Background Information

1. What happened in 2009 that shook the world?

In the early months of the 2009, the first case of H1N1 virus (also called Swine flu) was recorded in Mexico and then due to the airline travel, the virus later rapidly all spread around the world and later was declared a pandemic by the World Health Organization. The virus caused an estimated 284,400 deaths globally and was found that this 2009 H1N1 strain is a unique combination of human, swine and avian(bird) influenza A viruses.

Children and young adults became the most infected and affected. However, it was found that complications leading to hospitalization and the need for intensive care were prevalent in :

- Very young children
- Pregnant women
- Those who are morbidly obesity
- Those with underlying medical conditions such as chronic lung and cardiac diseases, diabetes, weak immunity
- Those with Bacterial coinfection were among the fatal cases

The 2009 H1N1 influenza pandemic highlighted the importance of understanding vaccination patterns to improve public health.

2. What is H1N1 Flu?

The H1N1 flu, also called the swine flu, is a type of influenza A virus.

Symptoms usually start quickly and can include:

- Fever for some patients
- Muscle Aches
- Chills and sweating
- Cough
- Sore throat
- Runny or stuffy nose
- Watery and even red eyes
- Eye pain
- Body aches
- Headache
- Tiredness and general weakness
- Diarrhea
- Stomach problems, vomiting which is more common in children

These swine flu symptoms develop about 1 to 4 days after exposure to the virus.

With time, the H1N1 flu strain from the pandemic became one of the strains that cause seasonal flu. Most people with the flu get better on their own but the complications are much worse and can be much more deadly, especially for people at high risk.

To curb this type of flu, the seasonal flu vaccine is used to protect against the H1N1 flu and other seasonal flu viruses.

Problem Statement

Immunization/Vaccination is thus a matter that cannot be ignored as an important tool in managing the spread of influenza.

As seen during the COVID-19 pandemic, it was clear that personal vaccination decisions are influenced by multiple factors, including background, health behaviour and opinions toward vaccines.

The National 2009 H1N1 Flu Survey provides rich data for analyzing these factors and how they influence vaccination uptake.

This analysis should support public health experts on what to do to provide better vaccination outreach.

Objectives

- The **main objective** is building a predictive model that can accurately forecast whether a person received the H1N1 vaccine based on a variety of features such as their concerns about H1N1, health behaviors, opinions about the vaccine, demographics, and more.

- Analyse the various factors through visualizations to analyse the behaviour of each when it comes to receiving vaccinations. For example, to see how various opinions are distributed.

- Evaluate the predictive model and provide actionable insights to inform future public health vaccination strategies.

Metrics of Success

In classification problems such as predicting vaccination status, several evaluation metrics are essential to assess model performance effectively. Each metric provides unique insights and is used based on the specific requirements of the analysis.

4.1 Accuracy

Target: Between 70% and 80%

Accuracy measures the proportion of correct predictions out of all predictions made by the model. A high accuracy score indicates that the model is making correct predictions most of the time. However, accuracy can be misleading when the dataset is imbalanced, as it might favour the majority class.

Reason for use:

Accuracy is a straightforward and intuitive metric to get a general understanding of overall model performance. It serves as a baseline before analysing more nuanced metrics.

4.2 Precision

Target: Greater than 80%

Precision evaluates the correctness of positive predictions. It measures the proportion of true positive predictions relative to all predicted positives. A high precision score signifies that the model has a low false positive rate.

Reason for use:

Precision is particularly useful when the cost of false positives is high. For example, in this context, misclassifying unvaccinated individuals as vaccinated could lead to errors in decision-making or public health strategies.

4.3 Recall (Sensitivity)

Target: Greater than 80%

Recall measures the proportion of actual positives correctly identified by the model. A high recall score means that the model minimizes false negatives, ensuring that most of the positive cases are captured.

Reason for use:

Recall is critical in scenarios where missing positive cases has significant consequences. For vaccination analysis, ensuring that vaccinated individuals are correctly identified can provide a clearer understanding of coverage and gaps.

4.4 F1 Score

Target: Greater than 0.8

The F1 score is the harmonic mean of precision and recall, offering a balanced measure that accounts for both metrics. It is particularly valuable when the dataset is imbalanced, ensuring that neither precision nor recall is disproportionately emphasized.

Reason for use:

The F1 score provides a comprehensive view of model performance by balancing the trade-offs between precision and recall, making it ideal for cases where both false positives and false negatives need to be minimized.

4.5 Area Under the ROC Curve (AUC-ROC)

Target: Greater than 0.85

The AUC-ROC metric evaluates the model’s ability to distinguish between classes across different thresholds. A higher AUC score indicates that the model effectively separates the positive and negative classes.

Reason for use:

AUC-ROC is a robust metric for assessing classification performance across a range of decision thresholds. It is particularly useful in understanding the model’s capability to differentiate between classes beyond a single fixed threshold.

Using a combination of accuracy, precision, recall, F1 score, and AUC-ROC, it is possible to build a model that performs well and aligns with the goals of the analysis.

Data Understanding

Data source and guidelines of the use of this data

The data for this competition comes from the National 2009 H1N1 Flu Survey (NHFS).

In their own words:

The National 2009 H1N1 Flu Survey (NHFS) was sponsored by the National Center for Immunization and Respiratory Diseases (NCIRD) and conducted jointly by NCIRD and the National Center for Health Statistics (NCHS), Centers for Disease Control and Prevention (CDC). The NHFS was a list-assisted random-digit-dialing telephone survey of households, designed to monitor influenza immunization coverage in the 2009-10 season.

The target population for the NHFS was all persons 6 months or older living in the United States at the time of the interview. Data from the NHFS were used to produce timely estimates of vaccination coverage rates for both the monovalent pH1N1 and trivalent seasonal influenza vaccines.

The NHFS was conducted between October 2009 and June 2010. It was one-time survey designed specifically to monitor vaccination during the 2009-2010 flu season in response to the 2009 H1N1 pandemic. The CDC has other ongoing programs for annual phone surveys that continue to monitor seasonal flu vaccination.

The source dataset comes with the following data use restrictions:

- The Public Health Service Act (Section 308(d)) provides that the data collected by the National Center for Health Statistics (NCHS), Centers for Disease Control and Prevention (CDC), may be used only for the purpose of health statistical reporting and analysis.

- Any effort to determine the identity of any reported case is prohibited by this law.

- NCHS does all it can to ensure that the identity of data subjects cannot be disclosed. All direct identifiers, as well as any characteristics that might lead to identification, are omitted from the data files. Any intentional identification or disclosure of a person or establishment violates the assurances of confidentiality given to the providers of the information.

Therefore, users will:

- ➢ Use the data in these data files for statistical reporting and analysis only. Make no use of the identity of any person or establishment discovered inadvertently and advise the Director, NCHS, of any such discovery (1 (800) 232-4636). Not link these data files with individually identifiable data from other NCHS or non-NCHS data files. By using these data, you signify your agreement to comply with the above requirements.

Link:

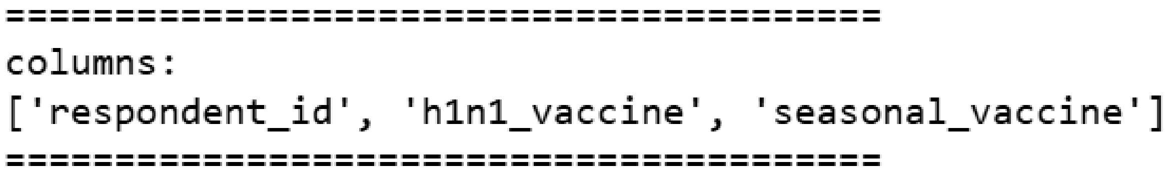
https://webarchive.loc.gov/all/20140511031000/http://www.cdc.gov/nchs/nis/about_nis.htm#h1n1

The data to be used for this analysis is broken into four parts. Below is a breakdown of what each type of data provides:

1. Training set features - Dataset contains the features associated with each respondent in the training data
2. Training set labels - A unique identifier for each respondent in the training set.
3. Test set features - Contains features for respondents in the test set. This is what you'll use to make predictions.
4. Submission Format - This is the format in which you need to submit your predictions

Columns and what the meaning

For the Submission Format data and the Training set labels, they have the same columns as follows:



```
==========================================
columns:
['respondent_id', 'h1n1_vaccine', 'seasonal_vaccine']
==========================================
```

The other remaining data sets also have the same columns (Training set features and Test set features data sets). These 35 columns are as follows:
Note that for all binary variables: 0 = No; 1 = Yes.

Feature Descriptions

1. **respondent_id**
   A unique and random identifier.
2. **h1n1_concern**
   Level of concern about the H1N1 flu.
   - 0 = Not at all concerned
   - 1 = Not very concerned
   - 2 = Somewhat concerned
   - 3 = Very concerned
3. **h1n1_knowledge**
   Level of knowledge about H1N1 flu.
   - 0 = No knowledge
   - 1 = A little knowledge
   - 2 = A lot of knowledge
4. **behavioral_antiviral_meds**
   Has taken antiviral medications. (binary)
5. **behavioral_avoidance**
   Has avoided close contact with others with flu-like symptoms. (binary)
6. **behavioral_face_mask**
   Has bought a face mask. (binary)
7. **behavioral_wash_hands**
   Has frequently washed hands or used hand sanitizer. (binary)
8. **behavioral_large_gatherings**
   Has reduced time at large gatherings. (binary)
9. **behavioral_outside_home**
   Has reduced contact with people outside of own household. (binary)
10. **behavioral_touch_face**
    Has avoided touching eyes, nose, or mouth. (binary)
11. **doctor_recc_h1n1**
    H1N1 flu vaccine was recommended by doctor. (binary)
12. **doctor_recc_seasonal**
    Seasonal flu vaccine was recommended by doctor. (binary)
13. **chronic_med_condition**
    Has any of the following chronic medical conditions: asthma or other lung conditions, diabetes, heart conditions, kidney conditions, sickle cell anemia, neurological or neuromuscular conditions, liver conditions, or weakened immune systems due to chronic illnesses or medications. (binary)
14. **child_under_6_months**
    Has regular close contact with a child under the age of six months. (binary)
15. **health_worker**
    Is a healthcare worker. (binary)

16. **health_insurance**
    Has health insurance. (binary)
17. **opinion_h1n1_vacc_effective**
    Respondent's opinion about H1N1 vaccine effectiveness.
    - 1 = Not at all effective
    - 2 = Not very effective
    - 3 = Don't know
    - 4 = Somewhat effective
    - 5 = Very effective
18. **opinion_h1n1_risk**
    Respondent's opinion about the risk of getting sick with H1N1 flu without a vaccine.
    - 1 = Very Low
    - 2 = Somewhat low
    - 3 = Don't know
    - 4 = Somewhat high
    - 5 = Very high
19. **opinion_h1n1_sick_from_vacc**
    Respondent's worry of getting sick from taking the H1N1 vaccine.
    - 1 = Not at all worried
    - 2 = Not very worried
    - 3 = Don't know
    - 4 = Somewhat worried
    - 5 = Very worried
20. **opinion_seas_vacc_effective**
    Respondent's opinion about seasonal flu vaccine effectiveness.
    - 1 = Not at all effective
    - 2 = Not very effective
    - 3 = Don't know
    - 4 = Somewhat effective
    - 5 = Very effective
21. **opinion_seas_risk**
    Respondent's opinion about the risk of getting sick with seasonal flu without a vaccine.
    - 1 = Very Low
    - 2 = Somewhat low
    - 3 = Don't know
    - 4 = Somewhat high
    - 5 = Very high
22. **opinion_seas_sick_from_vacc**
    Respondent's worry of getting sick from taking the seasonal flu vaccine.
    - 1 = Not at all worried
    - 2 = Not very worried
    - 3 = Don't know
    - 4 = Somewhat worried
    - 5 = Very worried
23. **age_group**
    Age group of the respondent.

24. **education**
    Self-reported education level.
25. **race**
    Race of the respondent.
26. **sex**
    Sex of the respondent.
27. **income_poverty**
    Household annual income of the respondent with respect to 2008 Census poverty thresholds.
28. **marital_status**
    Marital status of the respondent.
29. **rent_or_own**
    Housing situation of the respondent.
30. **employment_status**
    Employment status of the respondent.
31. **hhs_geo_region**
    Respondent's residence using a 10-region geographic classification defined by the U.S. Dept. of Health and Human Services. Values are represented as short random character strings.
32. **census_msa**
    Respondent's residence within metropolitan statistical areas (MSA) as defined by the U.S. Census.
33. **household_adults**
    Number of other adults in the household, top-coded to 3.
34. **household_children**
    Number of children in the household, top-coded to 3.
35. **employment_industry**
    Type of industry the respondent is employed in. Values are represented as short random character strings.
36. **employment_occupation**
    Type of occupation of the respondent. Values are represented as short random character strings.

Given this data we can see that they fall under the following categories:

- **Demographics**
  These columns describe basic demographic information about the respondents.
  - age_group, education, race, sex, income_poverty, marital_status, etc.
- **Health Behaviors**
  These columns represent various health-related behaviors and actions taken by the respondents to prevent illness.
  - behavioral_avoidance, behavioral_face_mask, behavioral_touch_face, etc.
- **Health Status**
  These columns provide information about the health conditions and situations of the respondents.

    - chronic_med_condition, child_under_6_months, health_worker, health_insurance.
- **Opinions About Vaccines**
  These columns reflect the respondents' views on the effectiveness and risks associated with the H1N1 and seasonal flu vaccines.
    - opinion_h1n1_vacc_effective, opinion_h1n1_risk, opinion_h1n1_sick_from_vacc, opinion_seas_vacc_effective, opinion_seas_risk, opinion_seas_sick_from_vacc.
- **Doctor Recommendations**
  These columns indicate whether the respondents' doctors recommended them to get the H1N1 or seasonal flu vaccine.
    - doctor_recc_h1n1, doctor_recc_seasonal.
- **Geographic and Employment Data**
  These columns provide information about where the respondents live and their employment details.
    - hhs_geo_region, employment_status, employment_industry, census_msa.
- **Respondent ID**
  This column is a unique identifier for each respondent, linking these features to other datasets.

Libraries used

1. Data Manipulation and Analysis

- **pandas**: Provides data structures like DataFrames for efficient manipulation, analysis, and storage of structured data.
- **numpy**: Offers support for numerical computations, including working with arrays and performing mathematical operations.

2. Data Visualization

- **seaborn**: High-level interface for creating attractive and informative statistical graphics. Commonly used for visualizing data distributions and relationships.
- **matplotlib.pyplot**: A foundational library for creating static, animated, and interactive visualizations in Python.

3. Machine Learning Models and Utilities

- **sklearn.model_selection**:
    - train_test_split: Splits data into training and testing subsets.
    - GridSearchCV: Performs hyperparameter tuning using cross-validation to optimize model performance.
    - StratifiedKFold: Ensures balanced class distributions across folds in cross-validation.
- **sklearn.preprocessing**:
    - OneHotEncoder: Encodes categorical data as binary (one-hot) arrays.

  - StandardScaler: Scales features to have a mean of 0 and standard deviation of 1.
  - RobustScaler: Scales features robust to outliers using median and interquartile range.
- **sklearn.impute**:
  - SimpleImputer: Handles missing data by replacing it with statistical values like the mean, median, or a constant.
- **sklearn.ensemble**:
  - RandomForestClassifier: A versatile ensemble learning method using decision trees.
- **sklearn.linear_model**:
  - LogisticRegression: A statistical model for binary classification problems.
- **sklearn.tree**:
  - DecisionTreeClassifier: A simple model that uses decision trees for classification tasks.
- **sklearn.metrics**:
  - accuracy_score, confusion_matrix, classification_report: Tools for evaluating classification model performance.
  - roc_auc_score, roc_curve: Metrics to assess model performance using the ROC curve and AUC score.
- **sklearn.utils.class_weight**:
  - compute_class_weight: Computes class weights to handle imbalanced datasets.
- **sklearn.pipeline**:
  - Pipeline: Combines preprocessing steps and a model into a single workflow for streamlined processing.

4. Feature Engineering and Dimensionality Reduction

- **sklearn.feature_selection**:
  - RFE: Recursive Feature Elimination for selecting important features.
- **sklearn.decomposition**:
  - PCA: Principal Component Analysis for dimensionality reduction and capturing variability in the data.

5. Handling Imbalanced Data

- **imblearn.over_sampling**:
  - SMOTE: Synthetic Minority Over-sampling Technique for generating synthetic samples to balance datasets.

6. Statistical Analysis

- **statsmodels.api**: Provides tools for statistical models, hypothesis testing, and data exploration.
  - variance_inflation_factor: Evaluates multicollinearity in regression models.
  - add_constant: Adds a constant term for regression analysis.

7. Warnings Management

- **warnings**: Suppresses unnecessary warnings, especially useful during development or when dealing with deprecated functions.

8. Additional Utilities

- **io**: Provides tools for input and output operations, like handling streams of data.
- **import_ipynb**: Allows importing Jupyter Notebook files into other notebooks.

Data cleaning and data processing

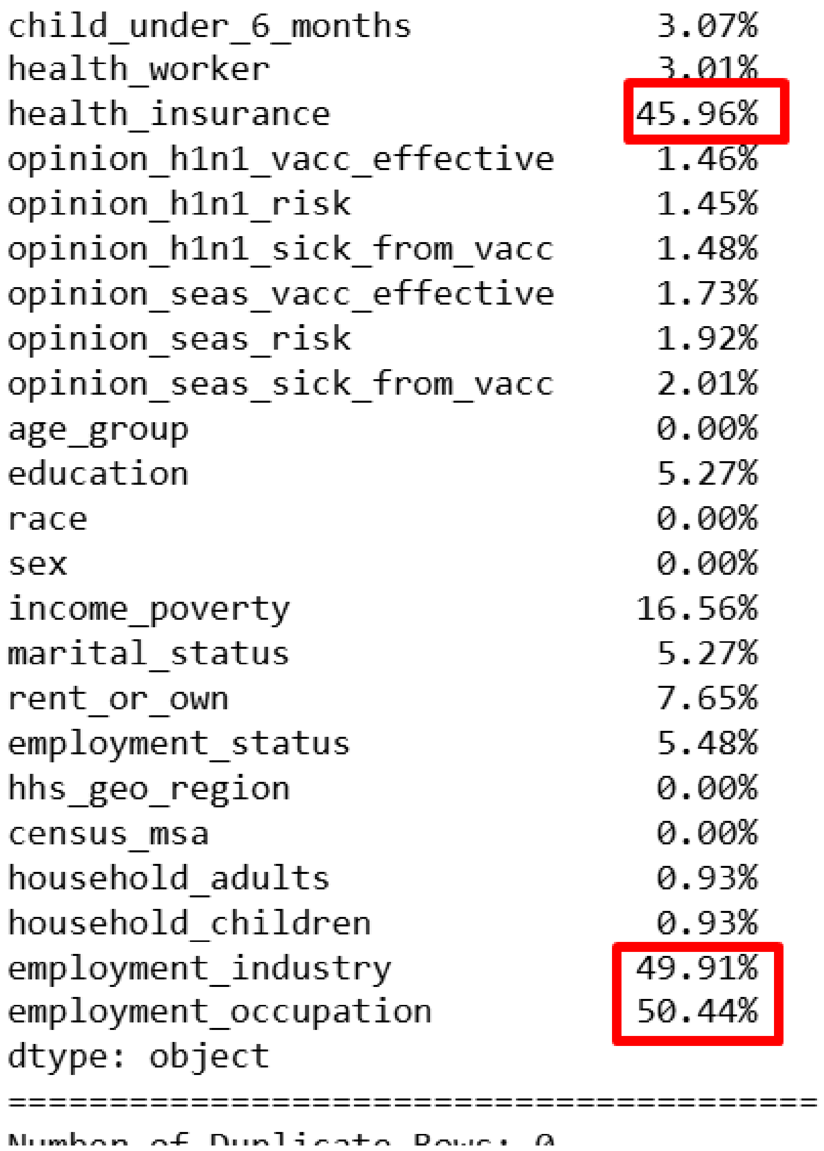
Handling of missing data

The only two data sets that had missing value were:

- Training set features
- Testing set features

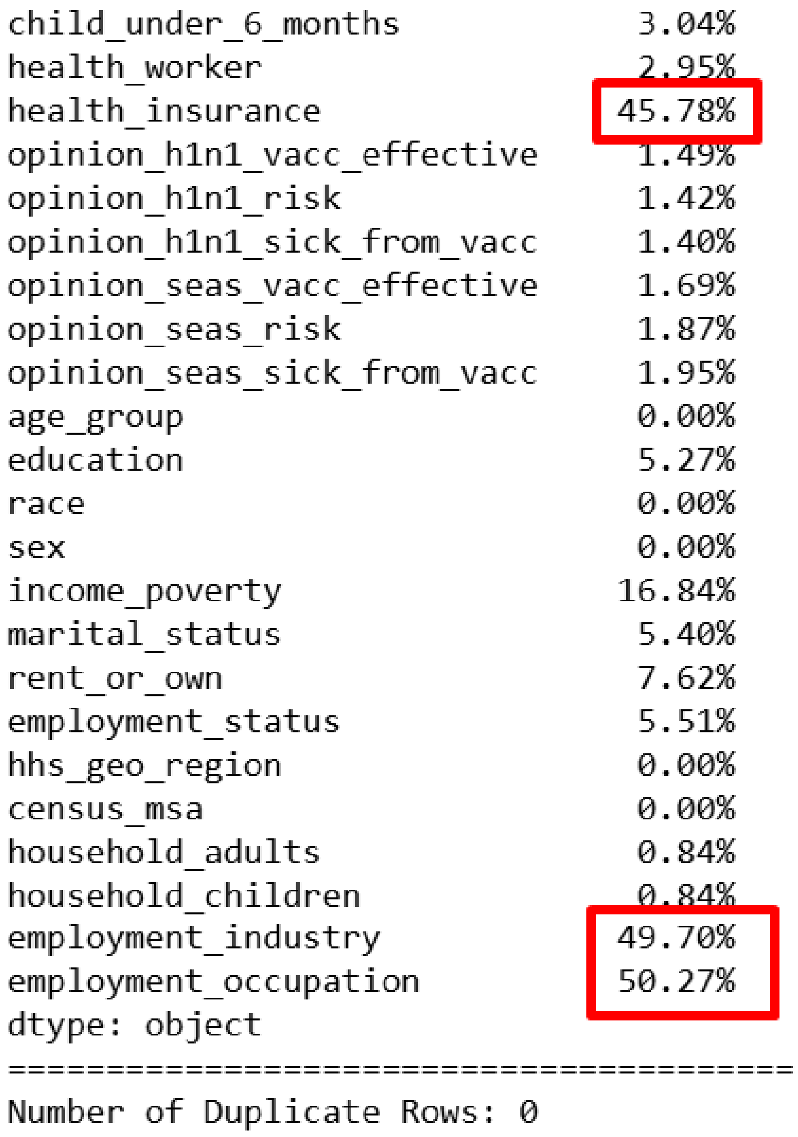
The percentage of missing values is as shown below:

- Training set features



```
child_under_6_months          3.07%
health_worker                 3.01%
health_insurance             45.96%
opinion_h1n1_vacc_effective   1.46%
opinion_h1n1_risk             1.45%
opinion_h1n1_sick_from_vacc   1.48%
opinion_seas_vacc_effective   1.73%
opinion_seas_risk             1.92%
opinion_seas_sick_from_vacc   2.01%
age_group                     0.00%
education                     5.27%
race                          0.00%
sex                           0.00%
income_poverty               16.56%
marital_status                5.27%
rent_or_own                   7.65%
employment_status             5.48%
hhs_geo_region                0.00%
census_msa                    0.00%
household_adults              0.93%
household_children            0.93%
employment_industry          49.91%
employment_occupation        50.44%
dtype: object
========================================
Number of Duplicate Rows: 0
```

- Testing set features



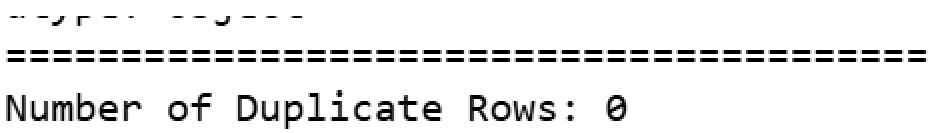
```
child_under_6_months          3.04%
health_worker                 2.95%
health_insurance             45.78%
opinion_h1n1_vacc_effective   1.49%
opinion_h1n1_risk             1.42%
opinion_h1n1_sick_from_vacc   1.40%
opinion_seas_vacc_effective   1.69%
opinion_seas_risk             1.87%
opinion_seas_sick_from_vacc   1.95%
age_group                     0.00%
education                     5.27%
race                          0.00%
sex                           0.00%
income_poverty               16.84%
marital_status                5.40%
rent_or_own                   7.62%
employment_status             5.51%
hhs_geo_region                0.00%
census_msa                    0.00%
household_adults              0.84%
household_children            0.84%
employment_industry          49.70%
employment_occupation        50.27%
dtype: object
========================================
Number of Duplicate Rows: 0
```

Missing values were replaced by the modal category in each column where there were missing values.

Columns that had over 45% of missing values were dropped from the data.

Handling of duplicates

There were no duplicates in all data sets.



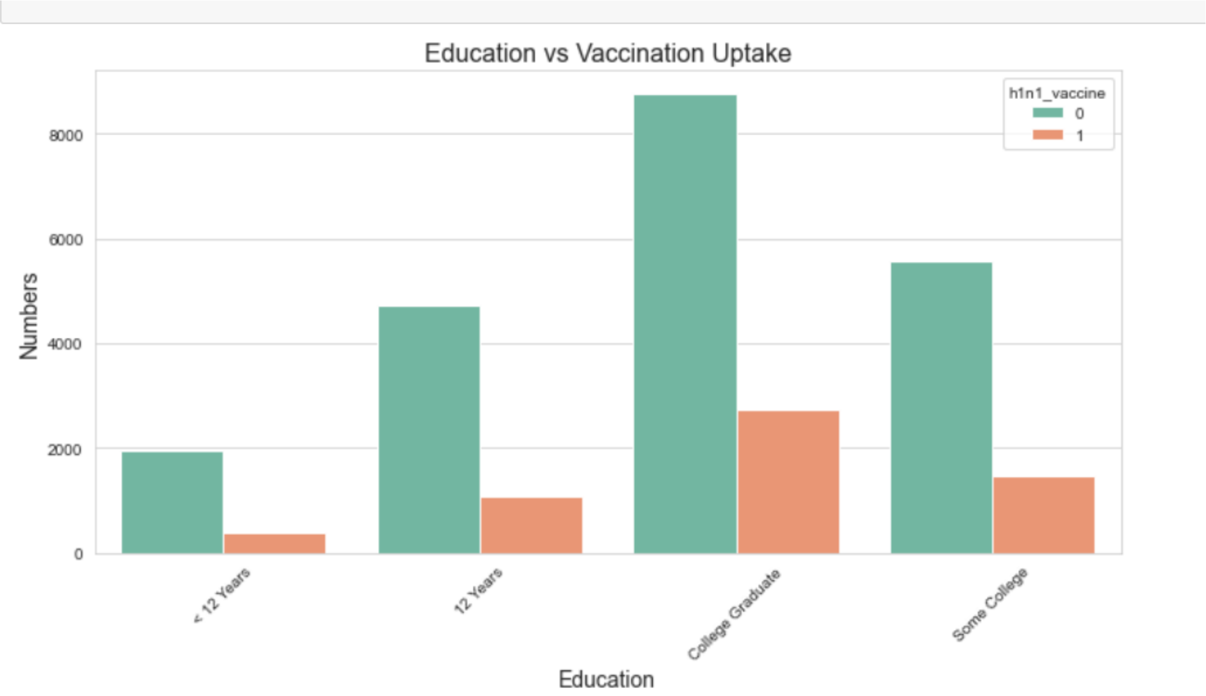
```
========================================
Number of Duplicate Rows: 0
```

Handling of categorical data to be used for logistic regression
One hot encoding approach was applied to remove categorical information to numerical values for the purpose of modeling.

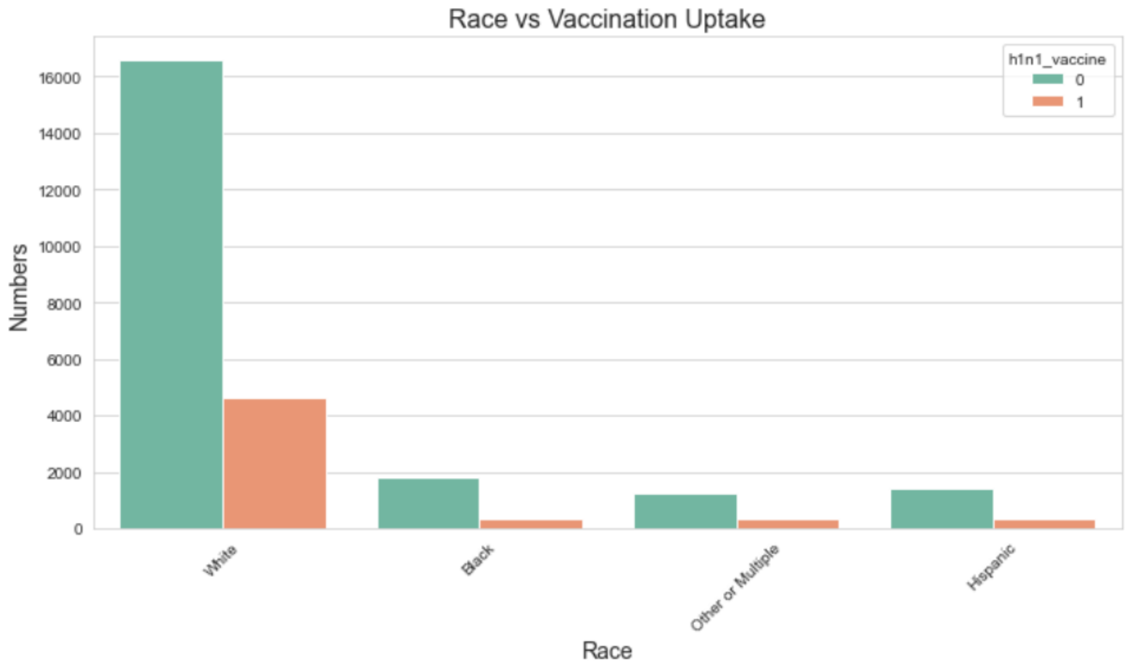
However, this method significantly increased the columns.

Exploratory Data Analysis

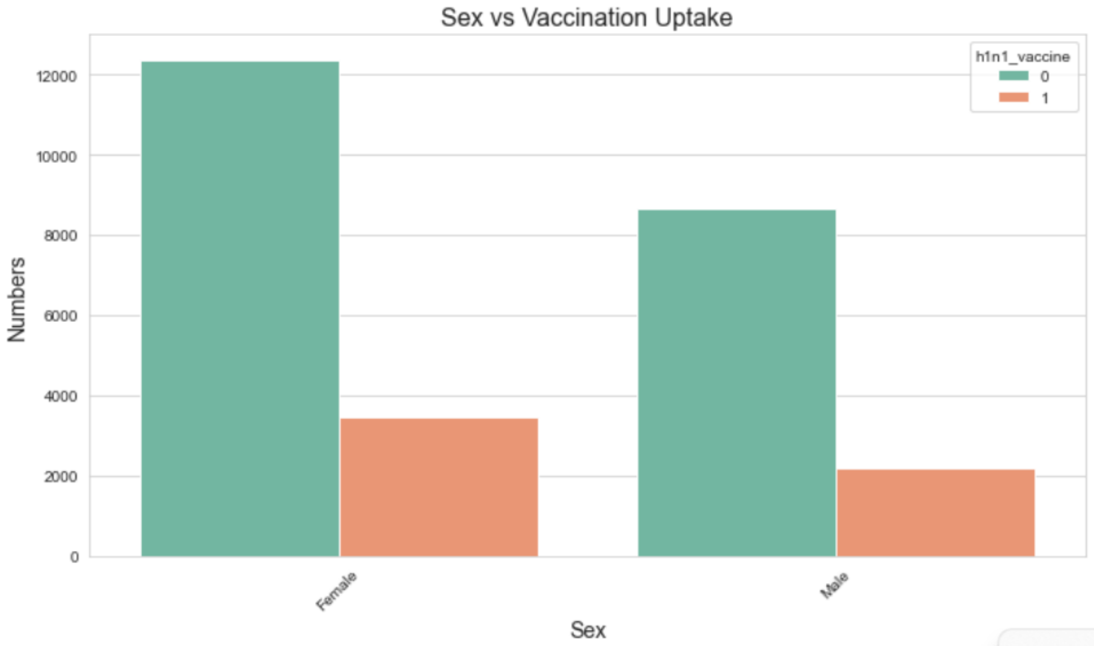
The graph below shows the relationship between three common demographic factors and vaccination uptake. These demographic features are Education, Age, Sex



- From the graph we see when it came to education more college students did not take the vaccine



- The graph indicates that among racial groups, White individuals are more likely to both get vaccinated and not get vaccinated. This suggests that White people represent the majority group in the dataset, aligning with the observed outcomes.
- The graph below highlights that women are less likely than men to receive vaccinations, indicating a noticeable gender difference in vaccination avoidance.



A correlation matrix was also plotted, which shows the large data set that also has very minimal correlation and multicollinearity (effect of one hot encoding).

MODELING

Model Overview and Optimization

The modelling process for predicting H1N1 vaccination status involved several machines learning techniques, including Logistic Regression, Decision Trees, and Random Forest models. Here's a step-by-step summary:

1. Data Preparation

- **Features and Target**: The data was prepared by merging the training features with the target variable (`h1n1_vaccine`). The feature set (`X`) included all columns except `respondent_id` and the target, while the target (`y`) was the vaccination status (`h1n1_vaccine`).
- **Class Imbalance**: Since the dataset was imbalanced (with more unvaccinated individuals), **SMOTE** (Synthetic Minority Over-sampling Technique) was applied to balance the dataset. This step helped to create synthetic samples of the minority class to improve the model's ability to predict both classes.
- **Train-Test Split**: The data was split into an 80/20 ratio for training and validation, ensuring that the class distribution in both sets was similar using **stratification**.

---

2. Model Training and Hyperparameter Tuning

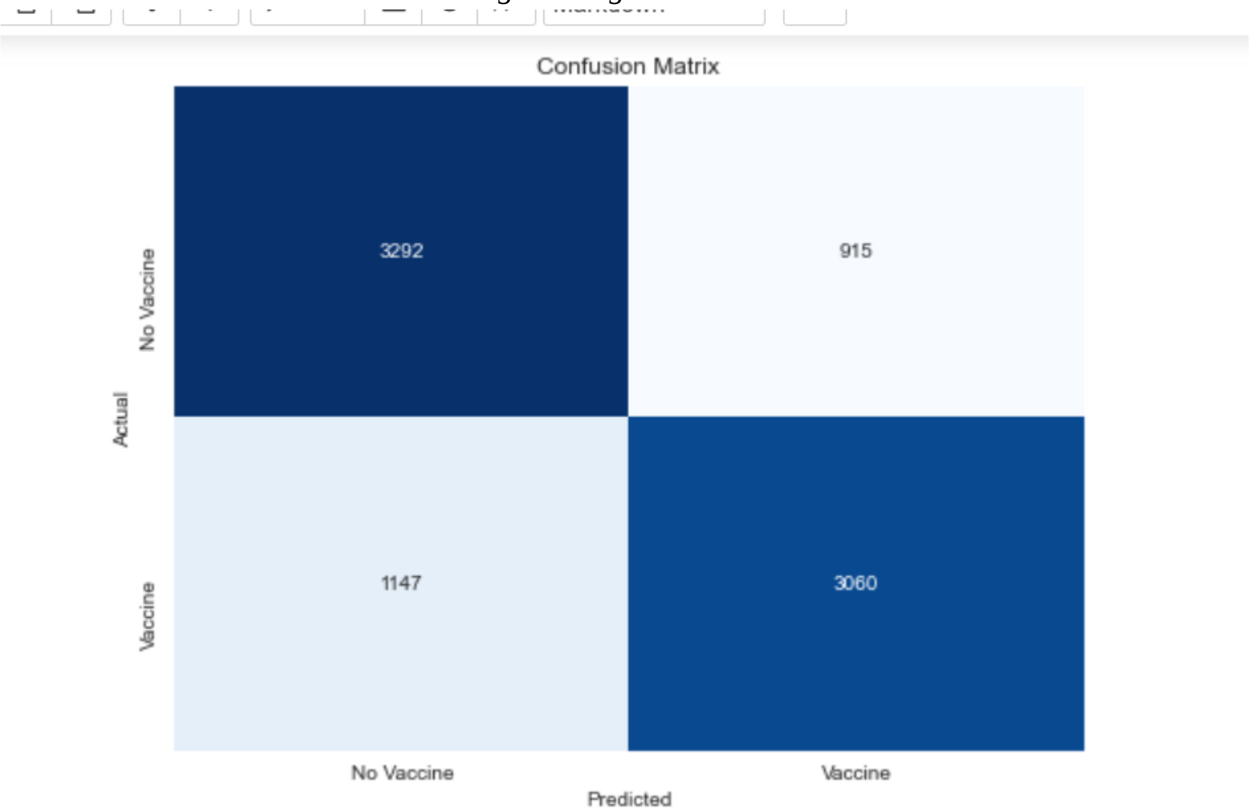
Three models were used to predict vaccination status:

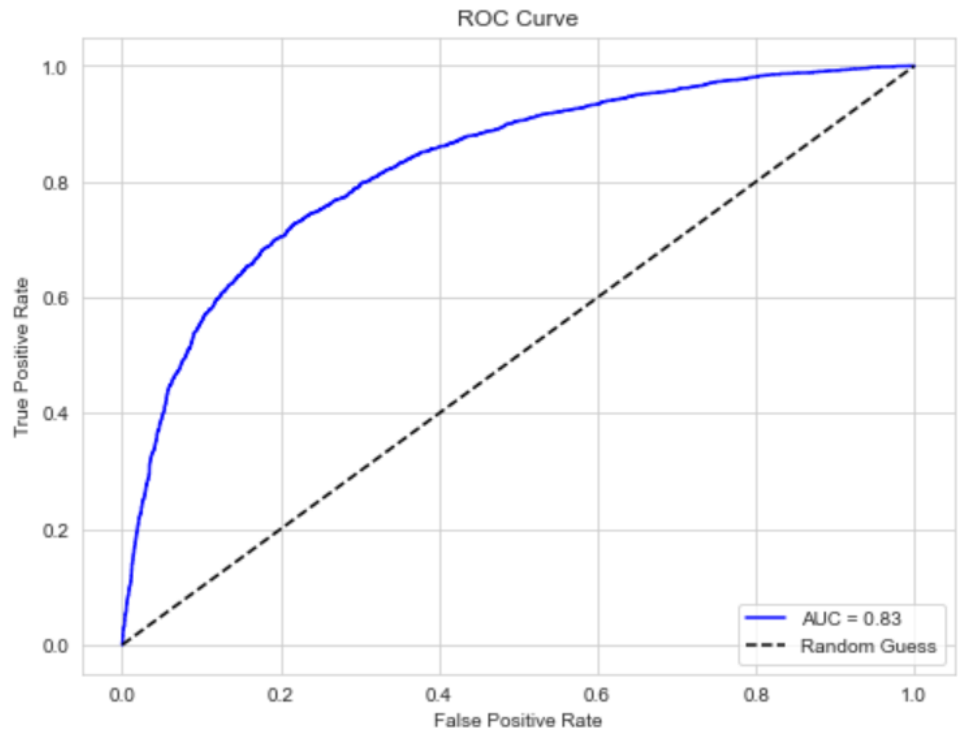
- **Logistic Regression**:
  - A pipeline was created with **RobustScaler** for scaling features and **Logistic Regression** for modeling. The model was tuned using **GridSearchCV** with hyperparameters like `C` (regularization strength), `penalty` (type of regularization), and `solver` (optimization method).
  - The goal was to identify the best combination of parameters that would give the highest performance on validation data.
- **Decision Tree**:
  - A **Decision Tree Classifier** was trained and optimized using **GridSearchCV**. Hyperparameters like `max_depth`, `min_samples_split`, `min_samples_leaf`, and `criterion` (the splitting criterion, either 'gini' or 'entropy') were tuned.
  - The model was evaluated based on its ability to separate vaccinated from unvaccinated individuals, using cross-validation and AUC-ROC as the performance metric.

- **Random Forest**:
    - The **Random Forest Classifier** was trained using 100 trees, with balanced class weights to handle the imbalanced dataset. This model's performance was evaluated using the standard accuracy, confusion matrix, classification report, and AUC-ROC score.
    - **Test Predictions**: The trained Random Forest model was also used to make predictions on the test data, providing probability scores for each class.

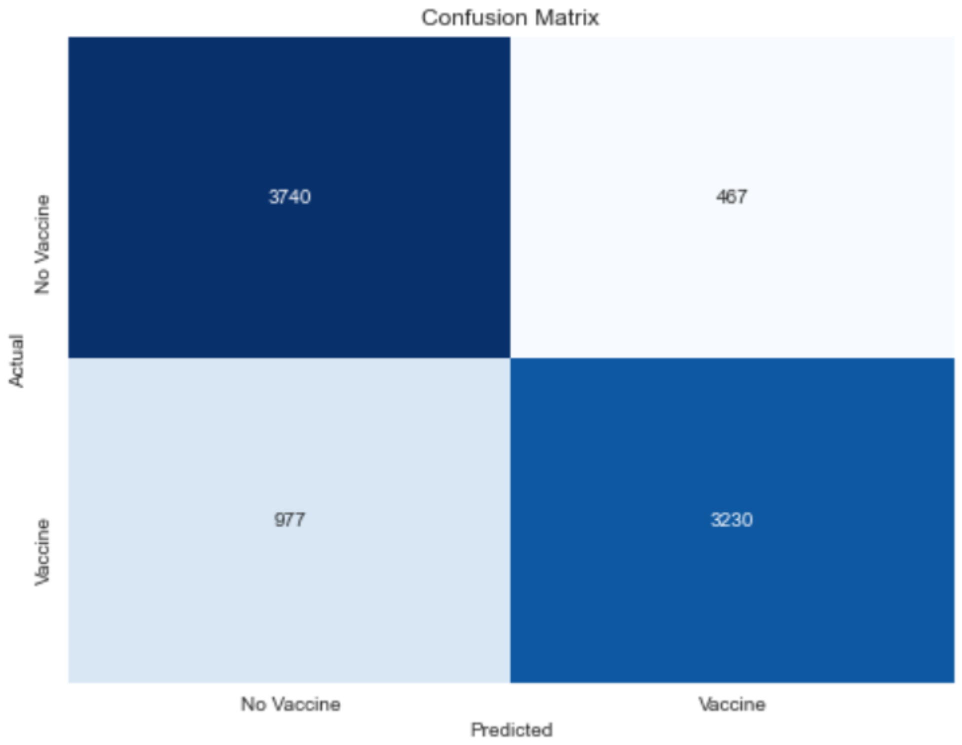
EVALUATION

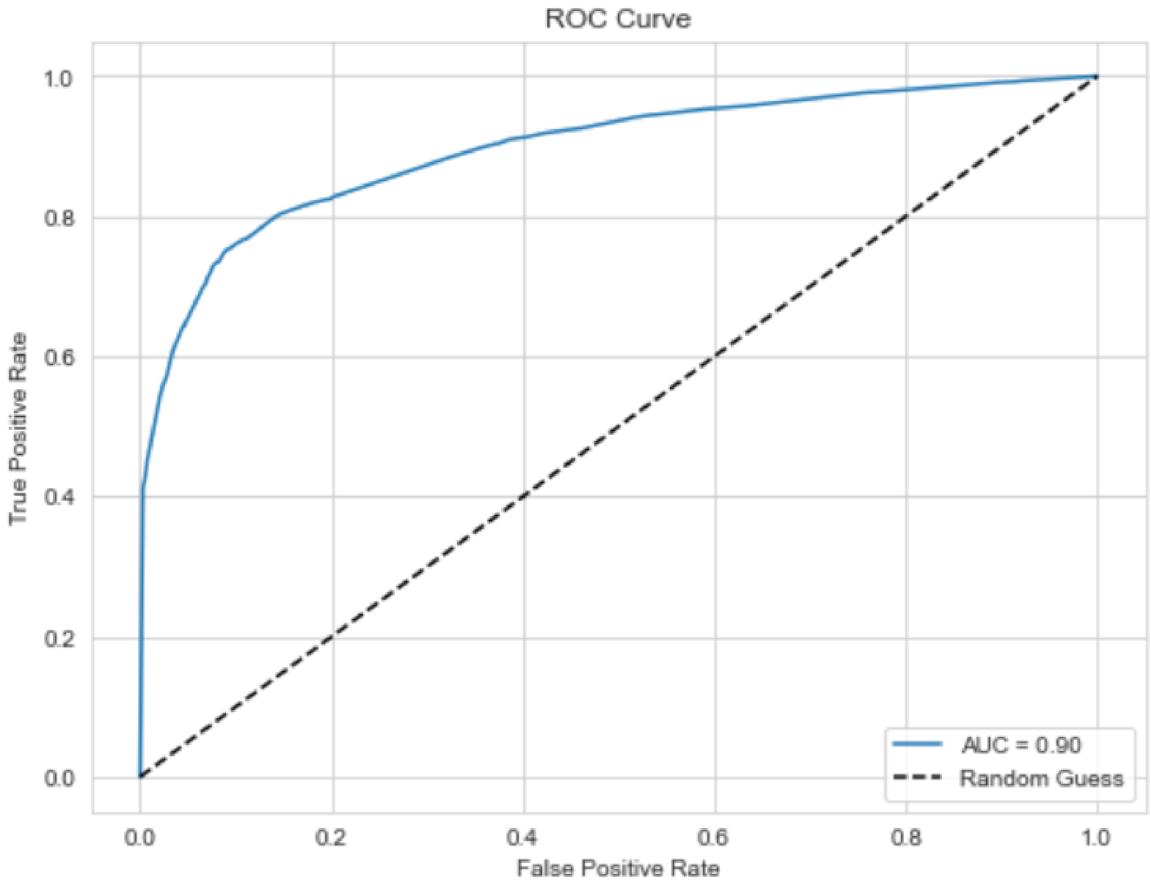
Confusion matrix and ROC curve – *Logistic Regression*





Confusion Matrix and ROC curve - *Decision Tree*





Evaluation of Model Results

1. Logistic Regression

- Best Parameters: {'logreg__C': 1, 'logreg__penalty': 'l2', 'logreg__solver': 'liblinear'}
- Performance Metrics:
  - ➢ Accuracy: 0.75 (Good as it is within the acceptable range of 70-80%, but not exceptional).
  - ➢ Precision: 0.77 (Below the benchmark of >80%, indicating some false positives).
  - ➢ Recall: 0.73 (Also below the benchmark, meaning some vaccinated individuals are not being identified).
  - ➢ F1 Score: 0.75 (Suggests a moderate balance between precision and recall).
  - ➢ AUC-ROC Score: 0.828 (Close to the desired threshold of 0.85, but leaves room for improvement).

Logistic regression performed moderately well. While accuracy and F1 scores are reasonable, precision and recall could be improved to achieve a better balance. This model may struggle slightly with imbalanced datasets.

2. Decision Tree Classifier

- Best Parameters: {'criterion': 'entropy', 'max_depth': 10, 'min_samples_leaf': 4, 'min_samples_split': 10}
- Performance Metrics:
    - ➢ Accuracy: 0.83 (Above the benchmark and quite strong).
    - ➢ Precision (Class 0): 0.79
    - ➢ Precision (Class 1): 0.87 (Class 1 precision is better than 80%, but the average is slightly below).
    - ➢ Recall: 0.83 (Good balance across classes).
    - ➢ F1 Score: 0.83 (Meeting the desired threshold).
    - ➢ AUC-ROC Score: 0.896 (Exceeds the benchmark of 0.85, showing strong separability between classes).

The decision tree model demonstrates strong performance across most metrics. The AUC-ROC score is excellent, and the F1 score highlights a good balance between precision and recall. However, the slightly lower precision for one class suggests there may still be occasional false positives.

3. Random Forest Classifier

Performance Metrics:

- ➢ Accuracy: 0.83 (Strong performance).
- ➢ Precision (Class 0): 0.85
- ➢ Precision (Class 1): 0.69 (Falls below the benchmark, indicating many false positives for this class).
- ➢ Recall: 0.38 for Class 1 (Very low, indicating a significant issue with false negatives).
- ➢ F1 Score: Weighted average of 0.81 (Drops due to low recall for Class 1).
- ➢ AUC-ROC Score: 0.800 (Falls short of the desired benchmark of 0.85).

Decision Tree Classifier emerges as the best model for this task. It provides a strong balance across key performance metrics, such as accuracy (0.83), recall (0.83), precision (Class 1: 0.87), F1 score (0.83), and AUC-ROC (0.896). The AUC-ROC score exceeds the benchmark of 0.85, indicating its strong ability to separate the classes effectively, particularly for the vaccinated class. It also has a good balance between precision and recall, making it suitable for identifying vaccinated individuals without significant false positives or negatives.

While random forest achieved high accuracy and precision for one class, its poor recall for Class 1 (vaccinated individuals) is a significant drawback. This indicates the model is missing many vaccinated individuals, which is critical for this prediction task.

CONCLUSION

- The models evaluated—Logistic Regression, Decision Trees, and Random Forest—revealed useful insights into predicting vaccination status.
- Decision Trees and Random Forest showed better performance in terms of AUC-ROC and overall accuracy compared to Logistic Regression.
- Random Forest had slightly lower precision and recall for the vaccinated class, suggesting difficulty in predicting the minority class, but it still performed well in distinguishing between the classes.
- Logistic Regression, while effective, provided more modest performance with an AUC-ROC score of 0.83 and 75% accuracy.
- Decision Trees had a good balance between precision and recall, with an AUC-ROC score of 0.89, demonstrating effectiveness in identifying vaccinated individuals while minimizing prediction errors.
- The SMOTE technique helped address class imbalance, improving the performance of Decision Trees and Random Forest, which are generally better at handling imbalanced classes.

RECOMMENDATIONS

- Decision Trees and Random Forest are the most promising models.
- Decision Trees are recommended when interpretability is important, as they offer a transparent decision-making process.
- Random Forest is better suited for higher generalization and might be preferable if performance on new, unseen data is critical, despite slightly reduced precision for the vaccinated class.

- Logistic Regression, while suitable as a baseline model, does not outperform Decision Trees and Random Forest and is not recommended as the primary model for this task.

NEXT STEPS

- **Hyperparameter tuning** should be further refined, particularly for Decision Trees and Random Forest, to achieve better performance.
- Exploring models like **Gradient Boosting** or **XGBoost** could provide better results, as these often outperform Decision Trees and Random Forest in classification tasks.
- **Feature engineering** should be explored to create or modify features that better capture important relationships in the data. Interaction terms or transformations of variables might enhance model performance.
- Addressing **multicollinearity** in the dataset could help stabilize models, especially Logistic Regression, which is sensitive to such issues.
- Once the best model is chosen, focus on **model interpretability** using tools like SHAP or LIME to understand which features most influence predictions.
- The next step would be to **deploy the model**, ensuring it is scalable, maintainable, and capable of handling new data effectively.
- Continuous **monitoring** and **retraining** of the model will be necessary to ensure it remains accurate as new data is processed.