H1N1 AND SEASONAL FLU VACCINE

ADMINISTRATION DATA

Presented by: Wallace Ouma

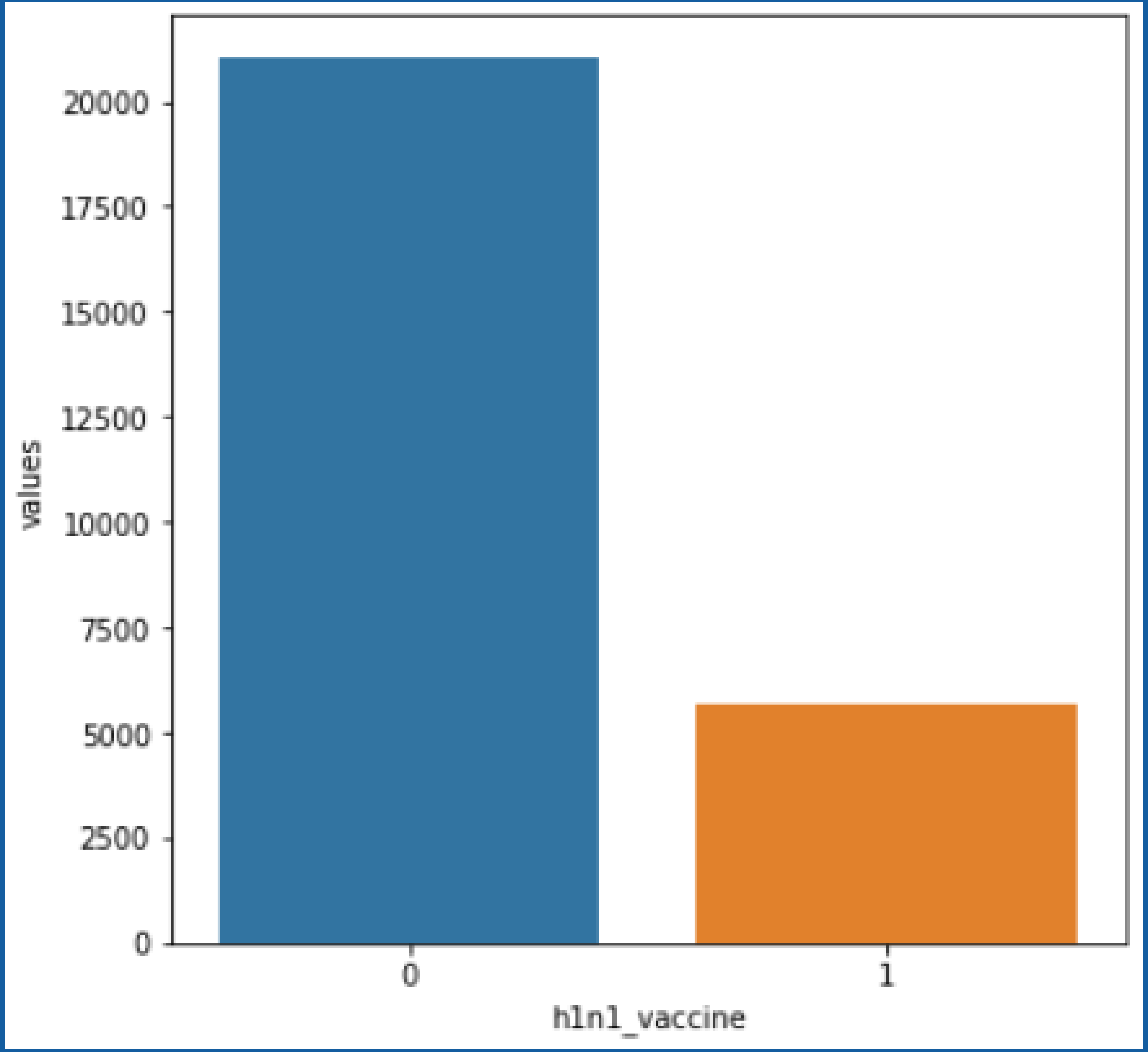
OVERVIEW

The National 2009 H1N1 Flu Survey (NHFS) was funded by the National Center for Immunization and Respiratory Diseases (NCIRD) and carried out collaboratively by NCIRD and the National Center for Health Statistics (NCHS), both part of the Centers for Disease Control and Prevention (CDC). The NHFS involved a telephone survey of households, utilizing a list-assisted random-digit-dialing method, aimed at tracking influenza immunization coverage during the 2009-10 season.

PROJECT OBJECTIVE

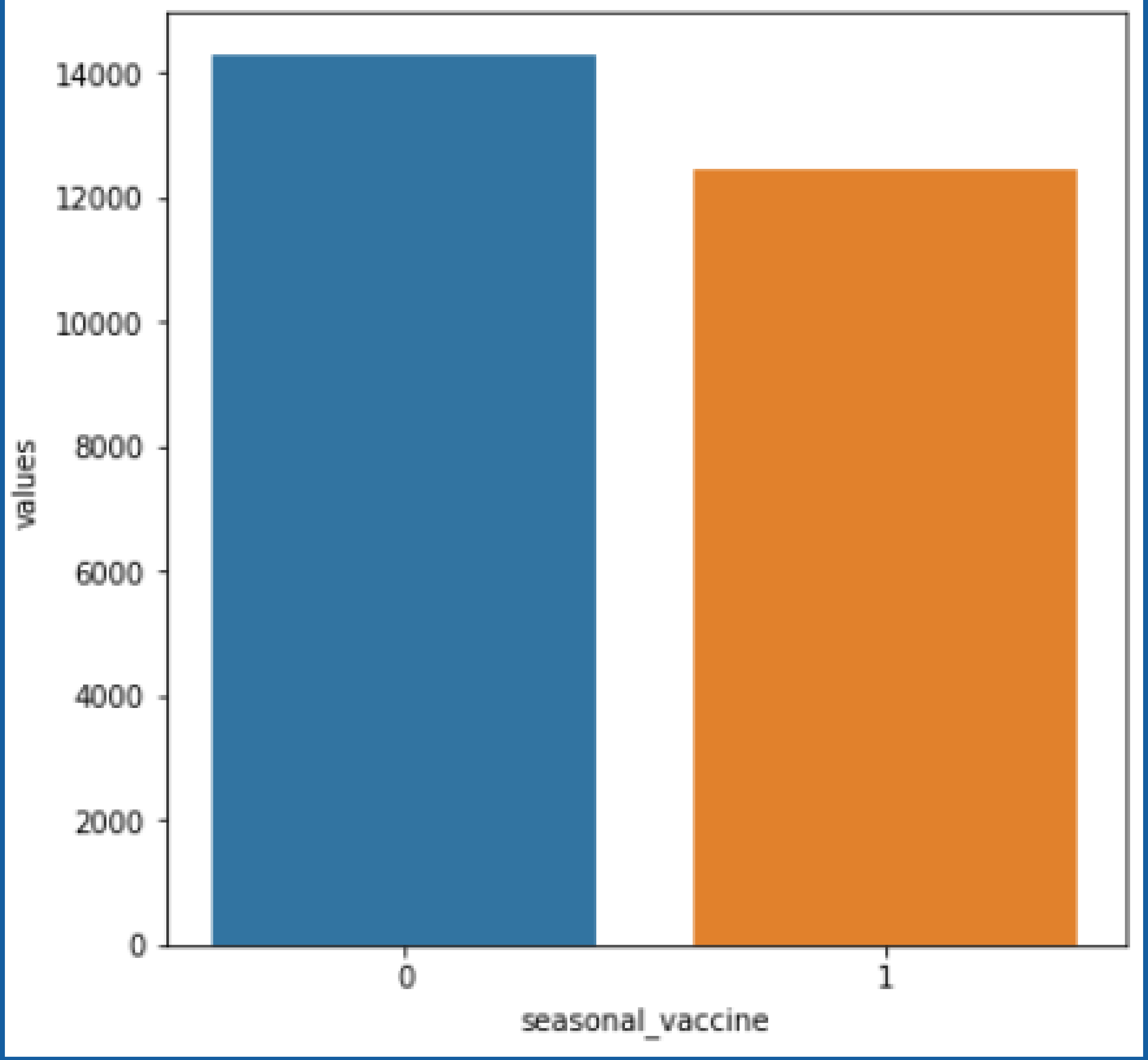
Predict whether a participant will obtain the H1N1 and/or seasonal flu vaccinations based on the details they provided regarding their backgrounds, perspectives, and health practices.

TARGET VARIABLES



H1N1 (h1n1_vaccine)

Given that there are nearly four times more individuals who did not receive the H1N1 vaccine compared to those who did, it's necessary to utilize different metrics to evaluate the fit of the models. This imbalance in the dataset may require the use of specialized evaluation techniques to account for the disproportionate class distribution.



Seasonal Flu Vaccine (seasonal_vaccine)

The number of individuals who received the flu shot is nearly as high as the number of those who did not. This balance contrasts with the difference in H1N1 vaccine recipients, which is more pronounced.

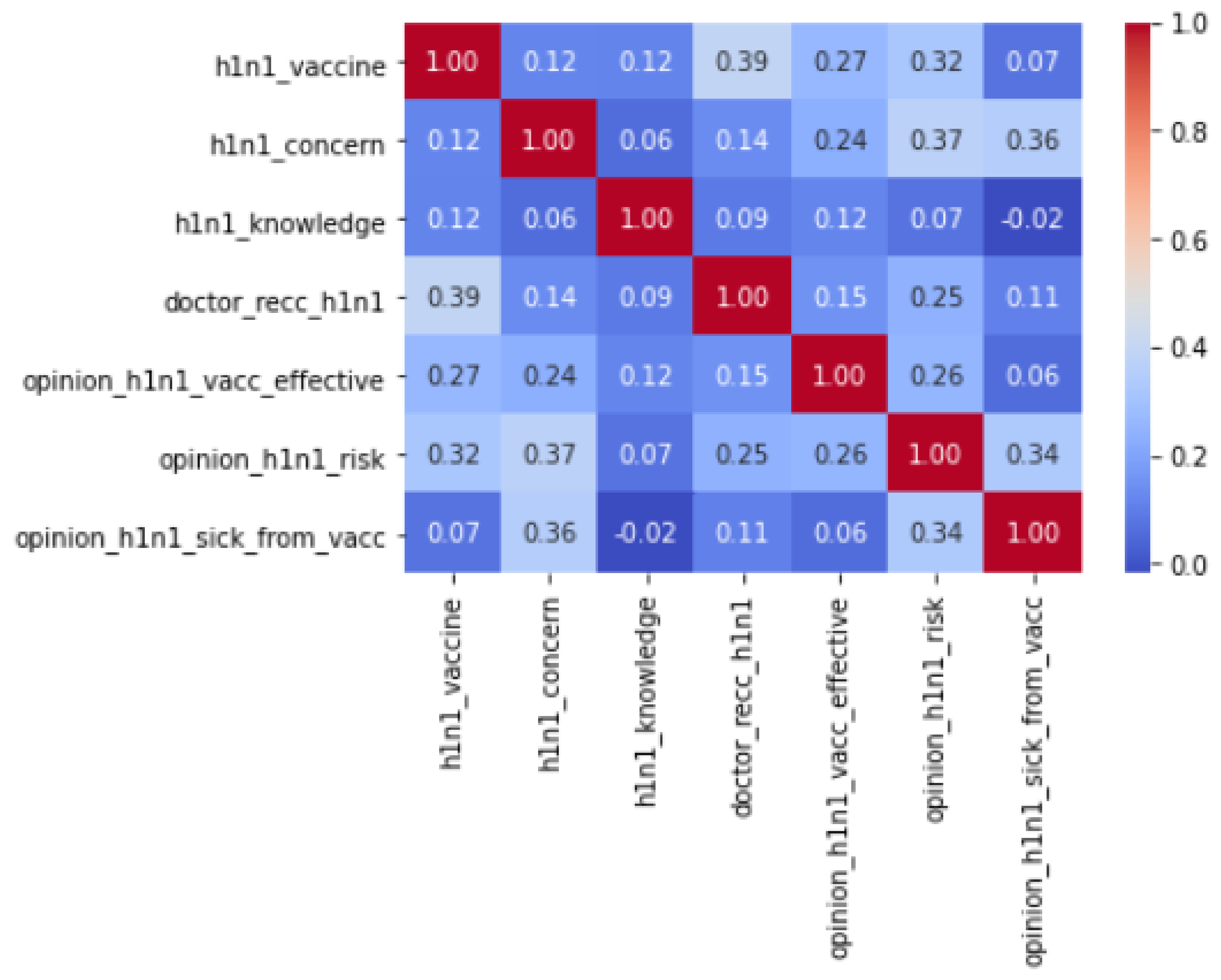
EXPLORATORY DATA ANALYSIS

EXPLORATORY DATA ANALYSIS (EDA)

Before proceeding with modeling data for prediction, it's essential to conduct exploratory data analysis (EDA). EDA involves observing and exploring variables along with their interactions to gain insights and understand the underlying patterns in the data.

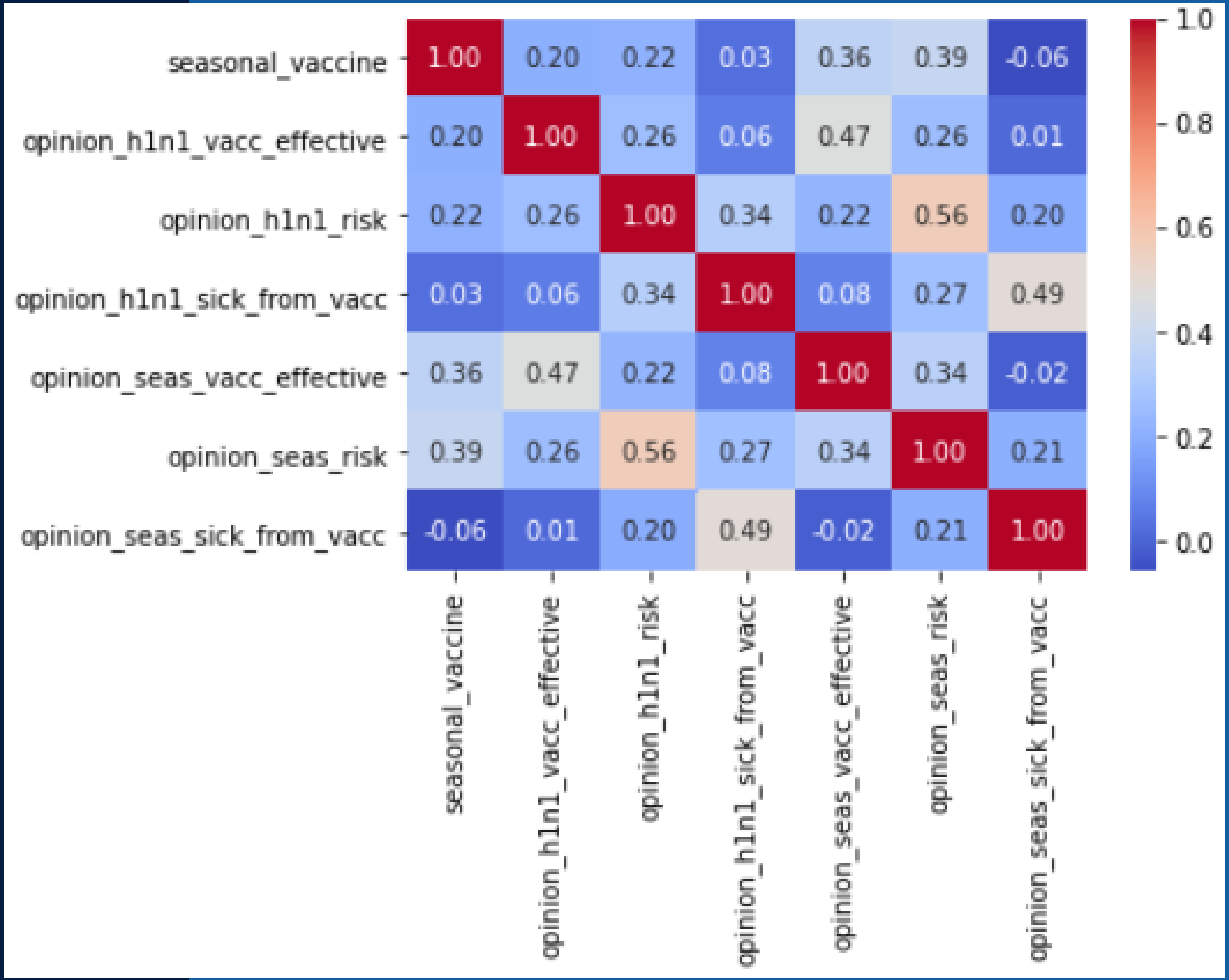
FEATURE INTERACTIONS

H1N1



	h1n1_vaccine	h1n1_concern	h1n1_knowledge	doctor_recc_h1n1	opinion_h1n1_vacc_effective	opinion_h1n1_risk	opinion_h1n1_sick_from_vacc
h1n1_vaccine	1.00	0.12	0.12	0.39	0.27	0.32	0.07
h1n1_concern	0.12	1.00	0.06	0.14	0.24	0.37	0.36
h1n1_knowledge	0.12	0.06	1.00	0.09	0.12	0.07	-0.02
doctor_recc_h1n1	0.39	0.14	0.09	1.00	0.15	0.25	0.11
opinion_h1n1_vacc_effective	0.27	0.24	0.12	0.15	1.00	0.26	0.06
opinion_h1n1_risk	0.32	0.37	0.07	0.25	0.26	1.00	0.34
opinion_h1n1_sick_from_vacc	0.07	0.36	-0.02	0.11	0.06	0.34	1.00

SEASONAL FLU



	seasonal_vaccine	opinion_h1n1_vacc_effective	opinion_h1n1_risk	opinion_h1n1_sick_from_vacc	opinion_seas_vacc_effective	opinion_seas_risk	opinion_seas_sick_from_vacc
seasonal_vaccine	1.00	0.20	0.22	0.03	0.36	0.39	-0.06
opinion_h1n1_vacc_effective	0.20	1.00	0.26	0.06	0.47	0.26	0.01
opinion_h1n1_risk	0.22	0.26	1.00	0.34	0.22	0.56	0.20
opinion_h1n1_sick_from_vacc	0.03	0.06	0.34	1.00	0.08	0.27	0.49
opinion_seas_vacc_effective	0.36	0.47	0.22	0.08	1.00	0.34	-0.02
opinion_seas_risk	0.39	0.26	0.56	0.27	0.34	1.00	0.21
opinion_seas_sick_from_vacc	-0.06	0.01	0.20	0.49	-0.02	0.21	1.00

FEATURE ENGINEERING

NULL VALUES

Null values were handled by replacing them with the mode (most frequent value) of each feature. This approach was particularly suitable given the prevalence of categorical variables in the dataset. However, due to a significant number of missing values in 'health_insurance', 'income_poverty', 'employment_industry', and 'employment_occupation', these variables were excluded from the analysis.

ONE-HOT ENCODING

One-hot encoding is a method of representing categorical variables as binary vectors, where each category is represented by a binary value (0 or 1) in a separate dimension. This approach enables the inclusion of more features in the model, strengthening its predictive power compared to traditional methods.

DATA PREPARATION

TRAINING AND TEST DATA SPLIT

To accurately assess model performance, the dataset will be divided into training and test datasets. The models will be trained using the training data and then evaluated by comparing predictions against the actual results from the test set. In this scenario, the dataset was split into 80% for training and 20% for testing purposes.

TRAINING AND TEST DATA SPLIT

Two distinct training and test datasets were divided. One set was designated for training and testing with the target variable labeled as 'h1n1_vaccine', while another set was created with the target variable labeled as 'seasonal_vaccine'.

FEATURE SCALING

Feature scaling was implemented on both the training and test datasets. This process normalizes the data, making it easier to compute during modeling, and ensures standardization of the functionality range within the input dataset.

FEATURE SELECTION

LASSO (L1) REGULARIZATION

Lasso (L1-Regularization) was used for feature selection, which identifies the most significant variables for analysis while nullifying coefficients for less important features to prevent overfitting. This approach enhances model robustness and was applied separately to both the H1N1 and Seasonal Flu datasets.

MODELING

MODELS TESTED

- Logistic Regression
- K-Nearest Neighbors
- Gradient Boosting Classifier
- Decision Tree
- Decision Tree with Tuning
- Random Forest
- Random Forest with Tuning
- Voting Classifier Ensemble

METRICS

Accuracy

Precision

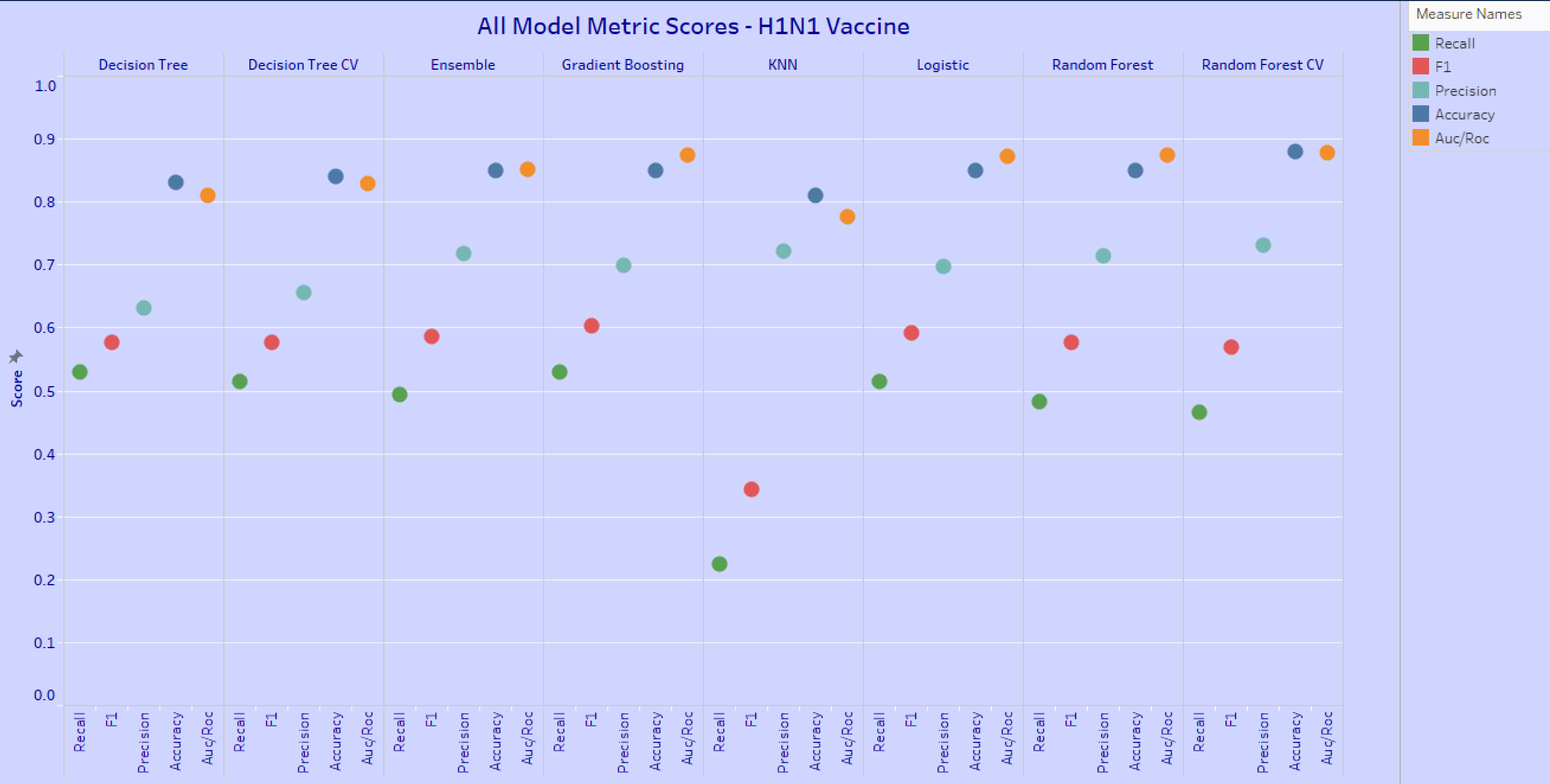
Recall

F1 Score

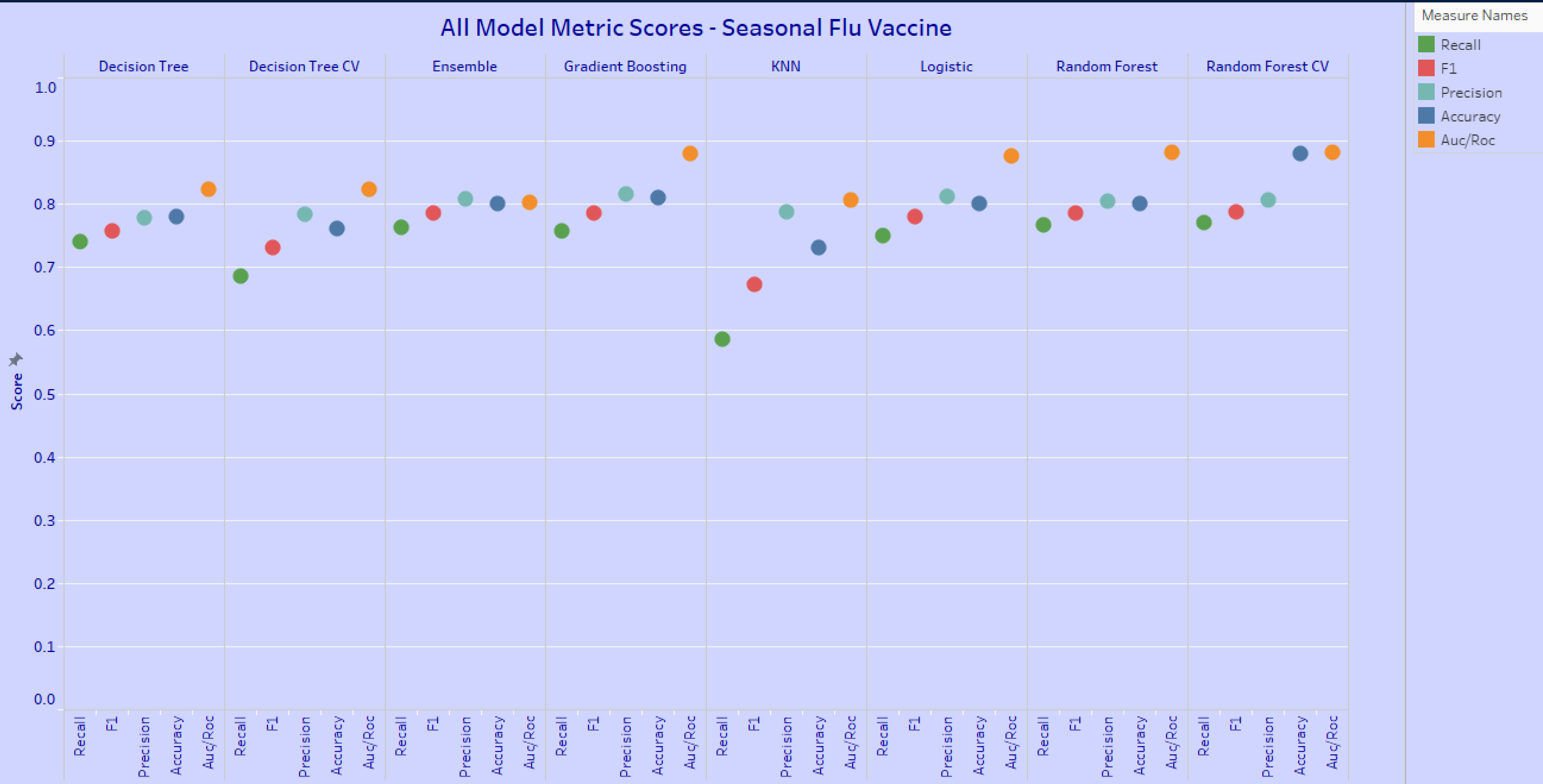
Misclassified Samples

AUC/ROC Score

All Model Metric Scores - H1N1 Vaccine

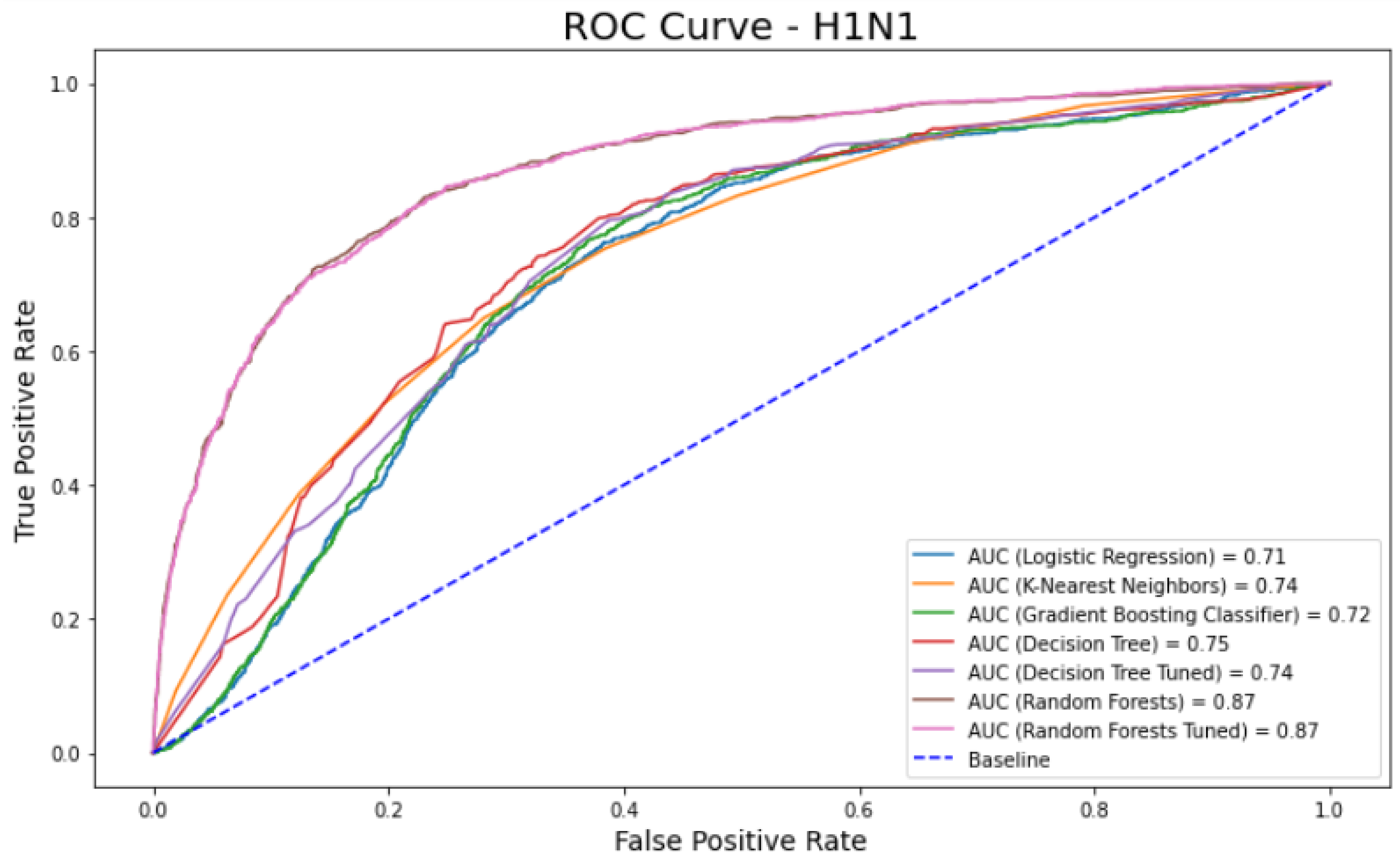


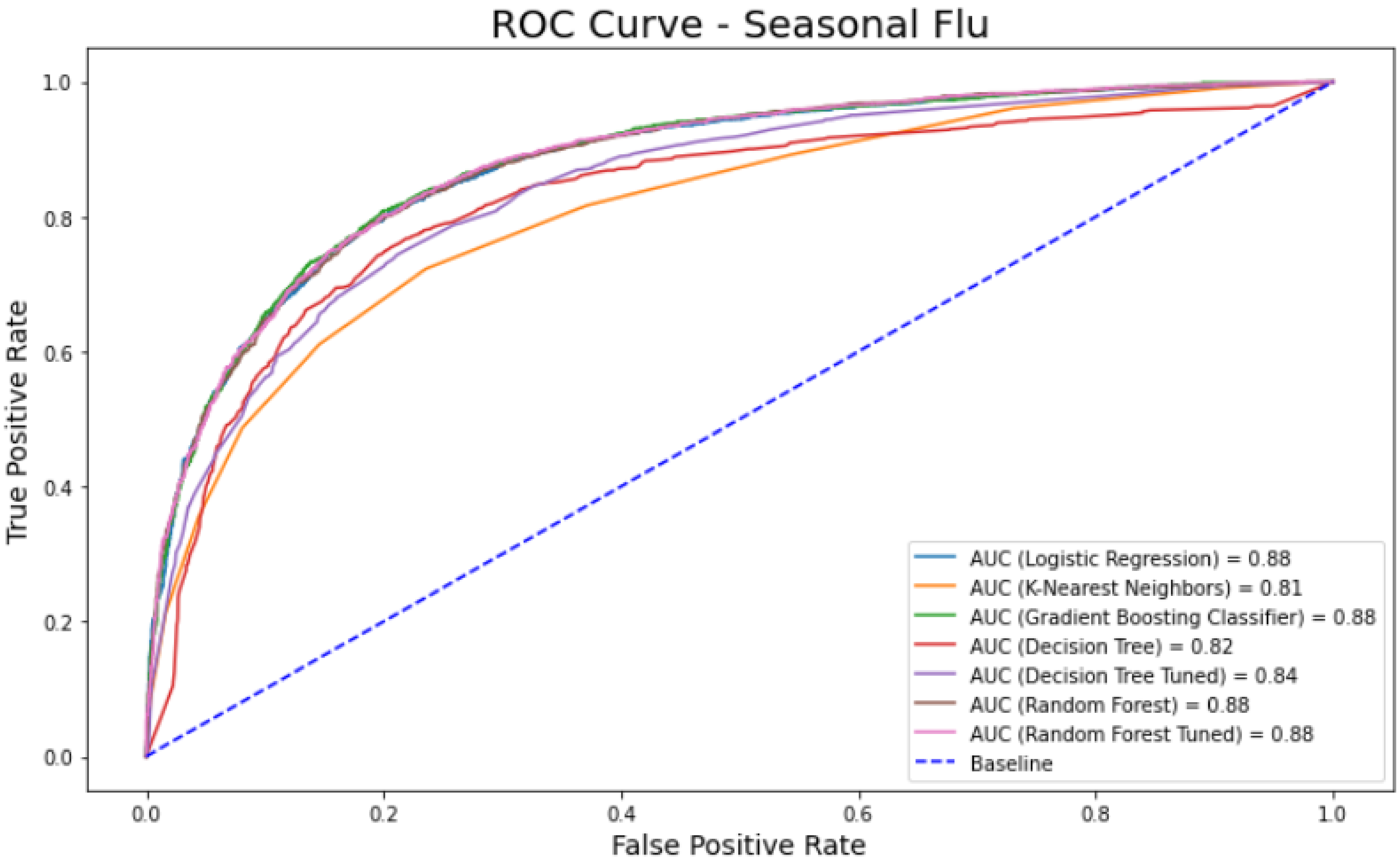
All Model Metric Scores - Seasonal Flu Vaccine



WHY USE ROC/AUC TO DETERMINE THE BEST MODEL?

The ROC is plotted between true positive (1.0) and false positive (0.0) rates.
The area under the curve (AUC) is the summary of this curve that tells generally how well a model is performing.
The AUC measures the quality of the model's predictions irrespective of what classification threshold is chosen - ideal for an unbalanced problem like ours.


ROC Curve - H1N1
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
AUC (Logistic Regression) = 0.71
AUC (K-Nearest Neighbors) = 0.74
AUC (Gradient Boosting Classifier) = 0.72
AUC (Decision Tree) = 0.75
AUC (Decision Tree Tuned) = 0.74
AUC (Random Forests) = 0.87
AUC (Random Forests Tuned) = 0.87
Baseline


ROC Curve - Seasonal Flu
True Positive Rate
1.0
0.8
0.6
0.4
0.2
0.0
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
AUC (Logistic Regression) = 0.88
AUC (K-Nearest Neighbors) = 0.81
AUC (Gradient Boosting Classifier) = 0.88
AUC (Decision Tree) = 0.82
AUC (Decision Tree Tuned) = 0.84
AUC (Random Forest) = 0.88
AUC (Random Forest Tuned) = 0.88
Baseline

Proposed Models

H1N1	Random Forest with Cross Validation
Seasonal Flu	Random Forest with Cross Validation

QUESTIONS?