

PHASE 3

PREDICTING H1N1 AND SEASONAL FLU VACCINE UPTAKE

By DENNIS MUENDO



TABLE OF CONTENTS

01. Project Overview

02. Business Problem

03. Data Analysis

04. Model Results

05. Conclusion

06. Recommendations

Project Overview

In this analysis, the focus is on leveraging data-driven insights to design personalized vaccination campaigns. The goal is to employ predictive models to identify individuals at a higher risk of not receiving these vaccines and tailor interventions to address their specific concerns or barriers.

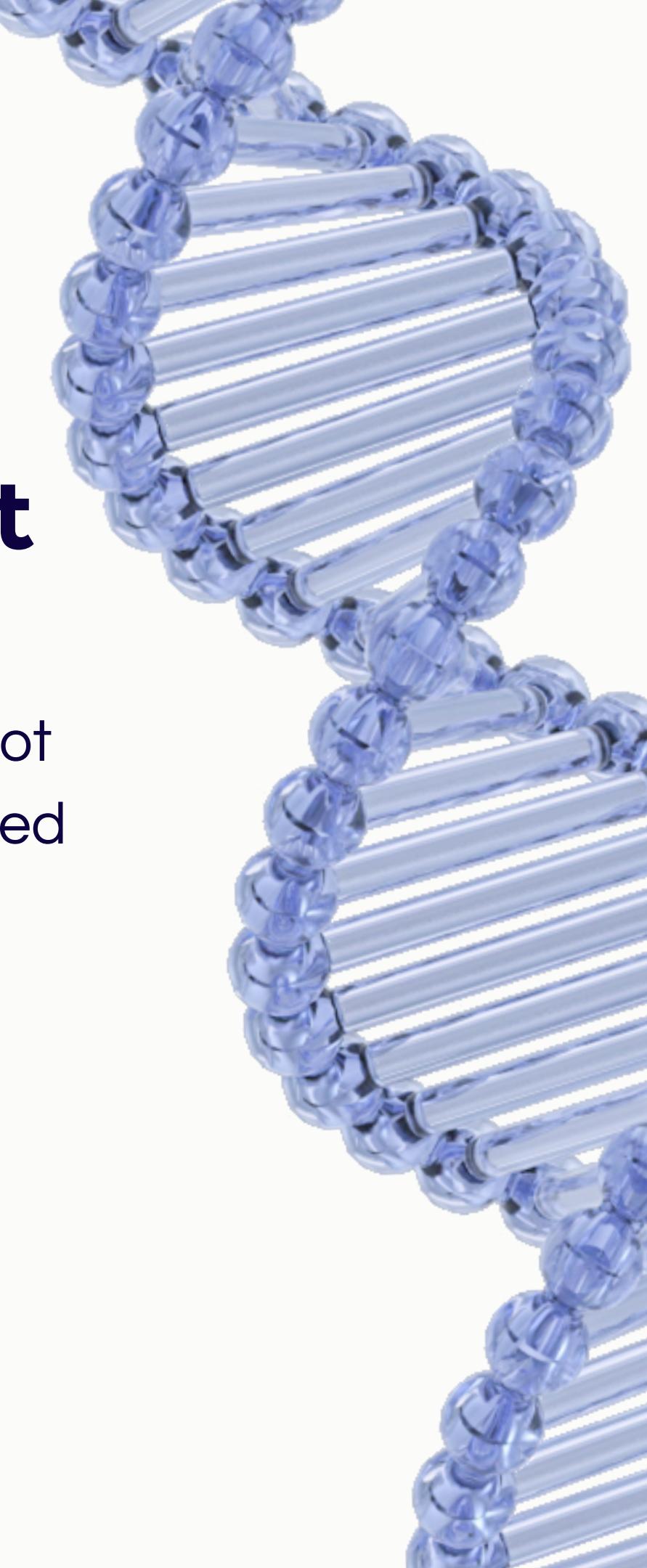


Business Problem

What are the factors that lead one to getting vaccinated or not getting vaccinated

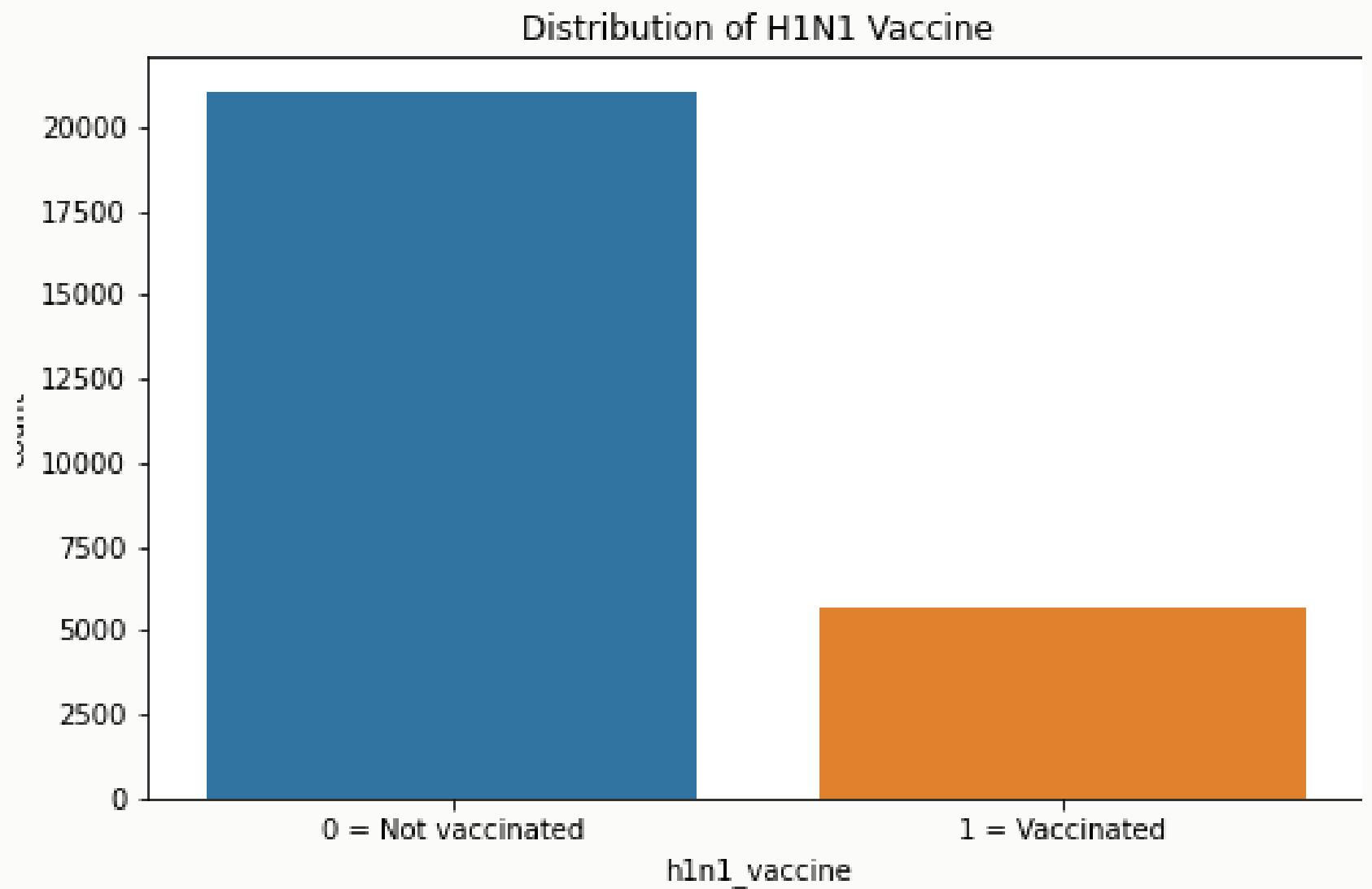
How can we predict if an individual is vaccinated or not vaccinated?

The goal of utilizing data-driven insights to improve public health outcomes and prevent the rapid transmission of pandemic influenza strains, ultimately reducing fatality rates linked with these viral infections.



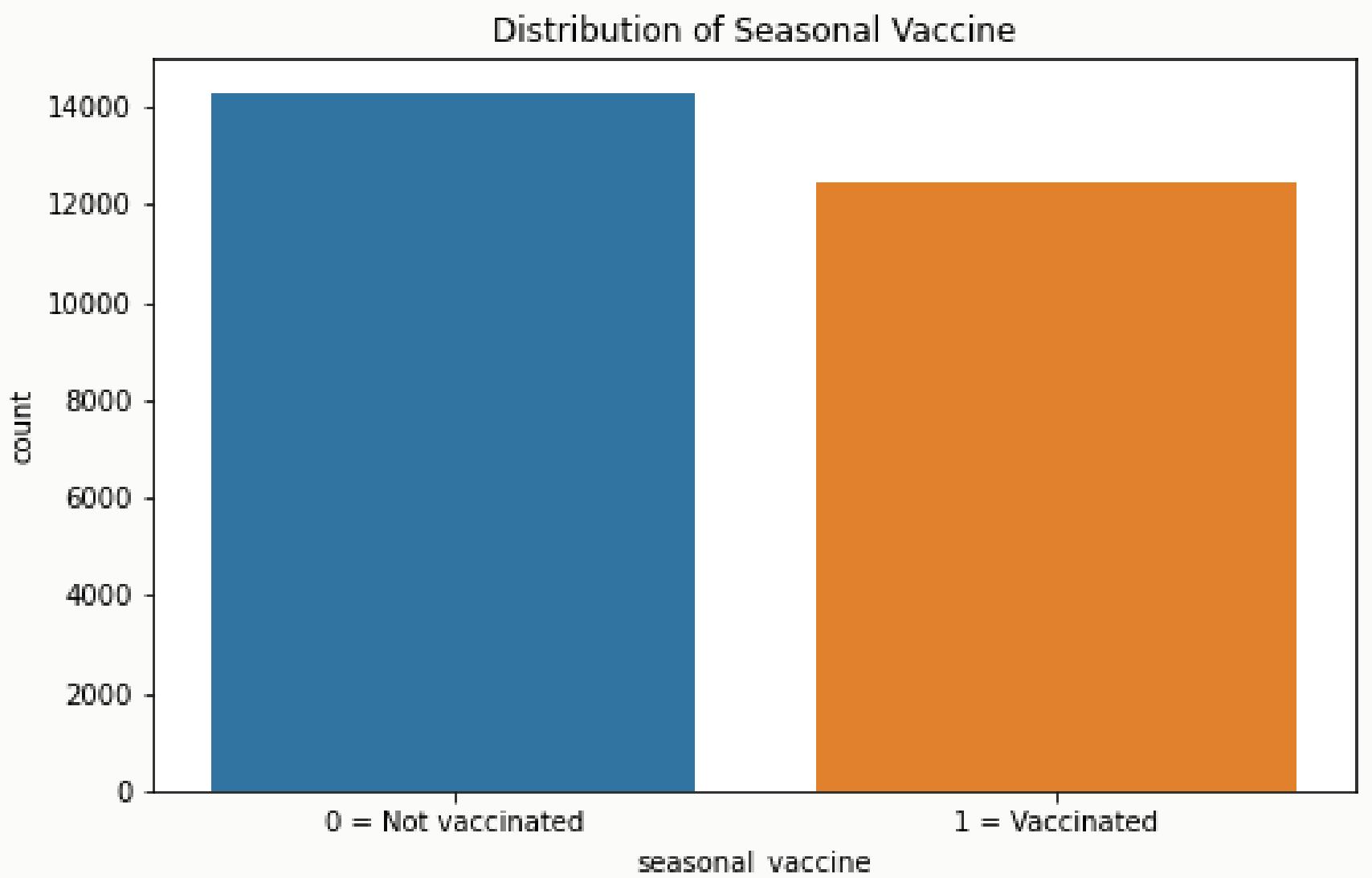
Data Analysis - unbalanced target

There is an imbalance in the 'H1N1_vaccine' as the bar for not vaccinated is very high compared to the one for vaccinated.



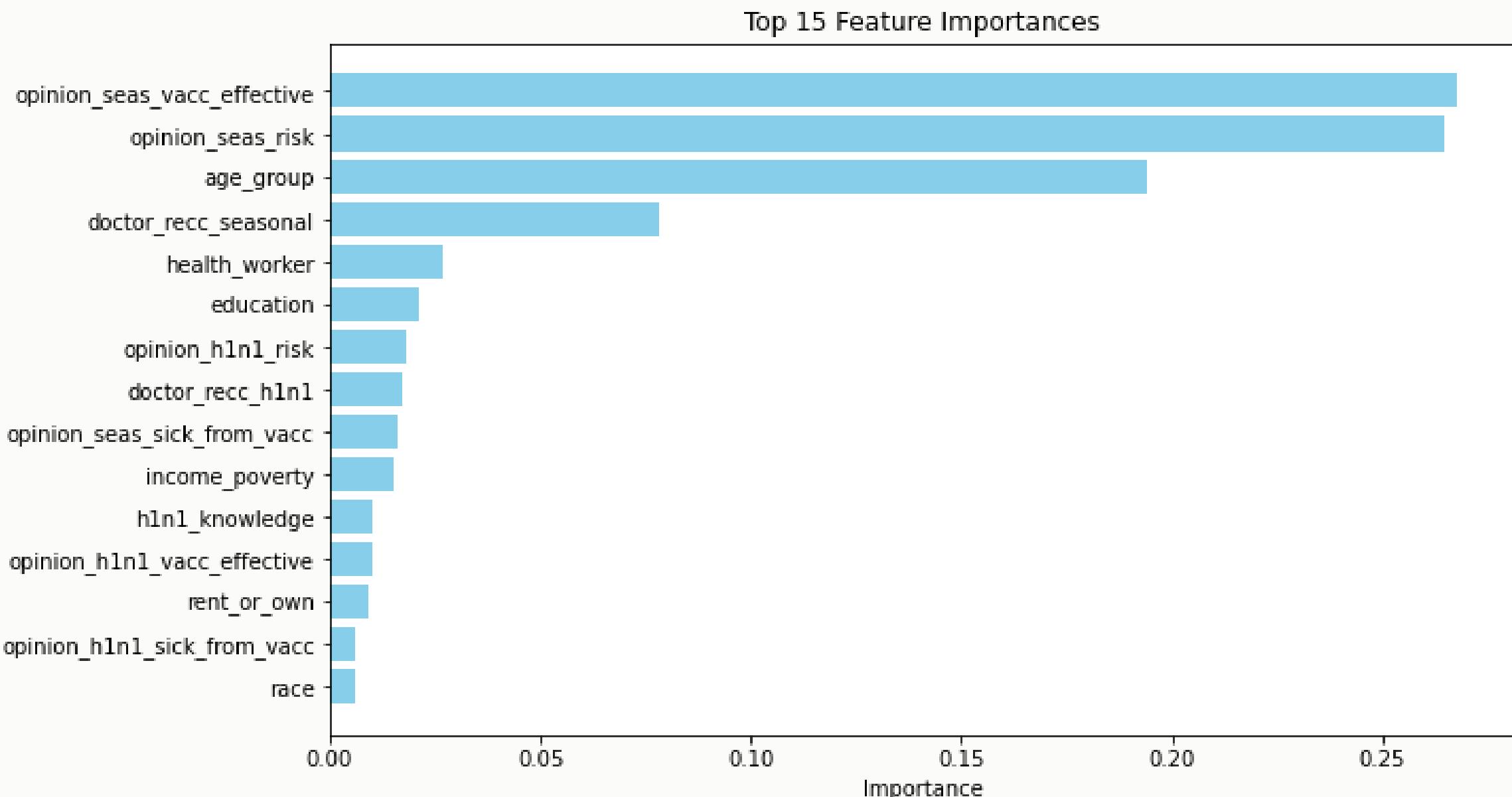


Data Analysis - unbalanced target



The distribution for 'seasonal_vaccine' seems to be evenly distributed as the number of not vaccinated individuals is slightly bigger than vaccinated.

Model: Decision Tree

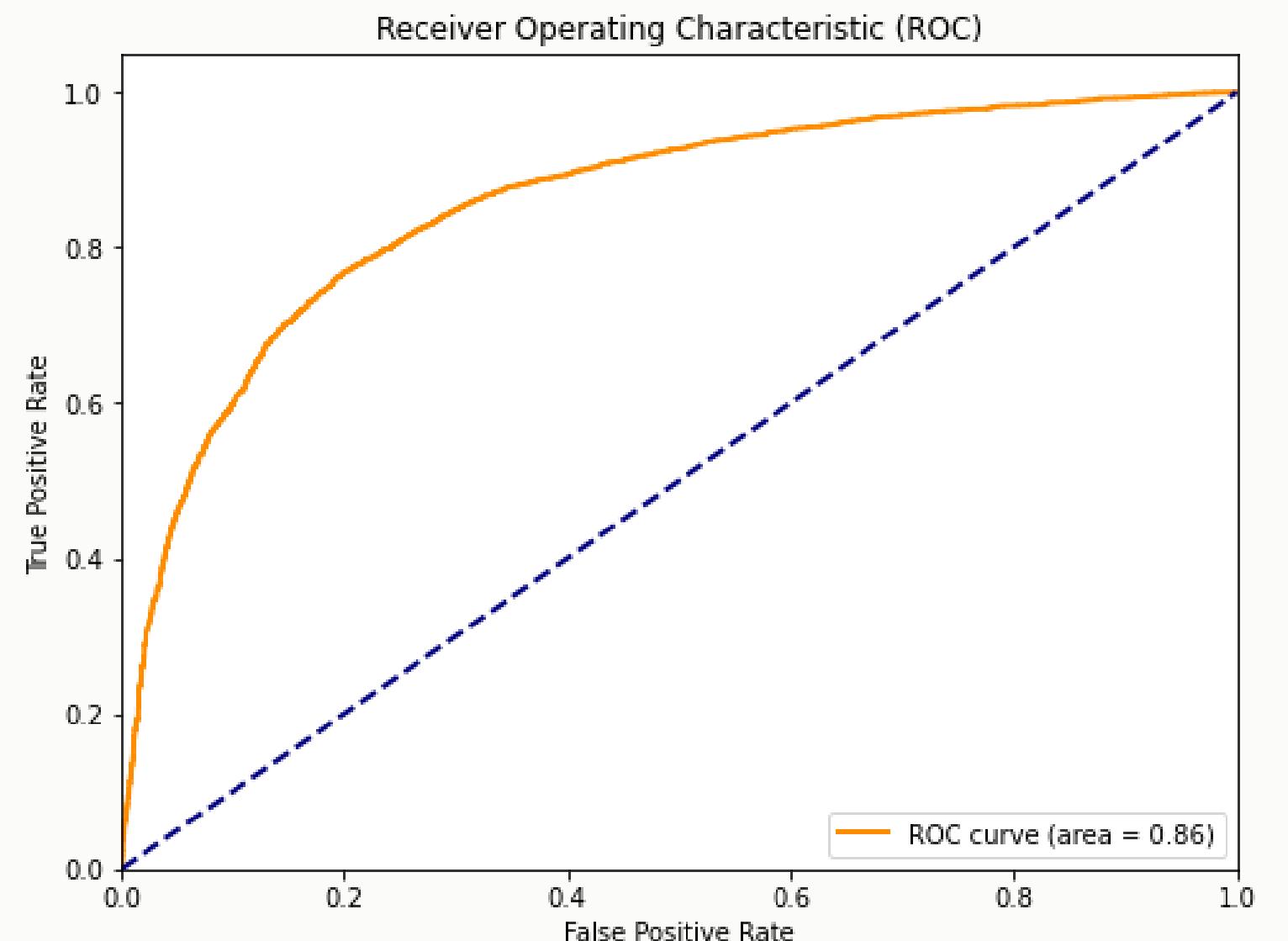


Decision Tree Model has accuracy, precision, recall, f1-score, and specificity in the upper 60% range, which is not a good performance. According to the model's assessment, the three most influential features are opinion_seas_vacc_effective, opinion_seas_risk and age_group respectively.



Model: Random Forest

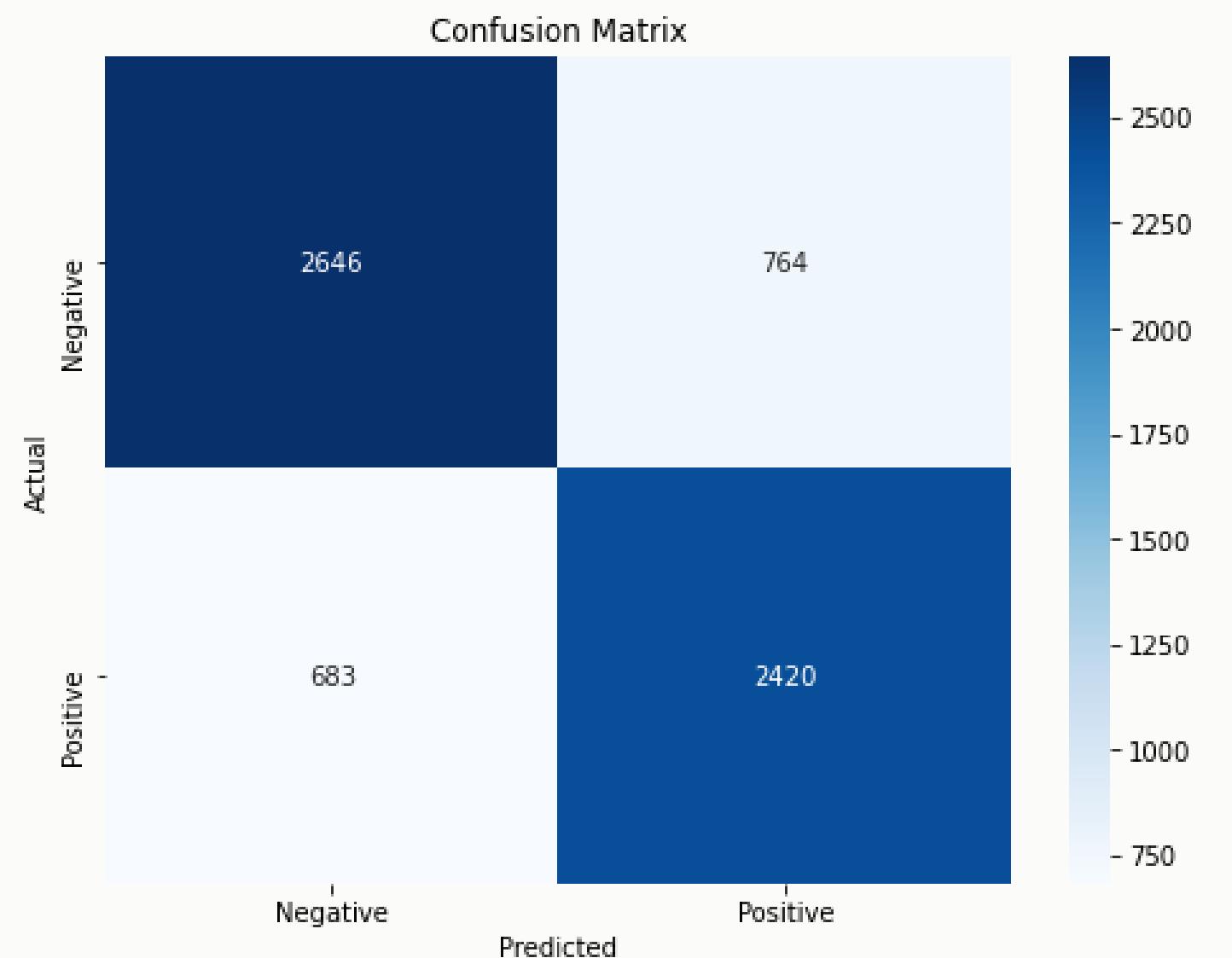
The ROC-AUC score is a measure of the model's ability to distinguish between positive and negative classes. The results indicate an accuracy of approximately 78% and an ROC-AUC score of approximately 0.86, suggesting reasonably good model performance.

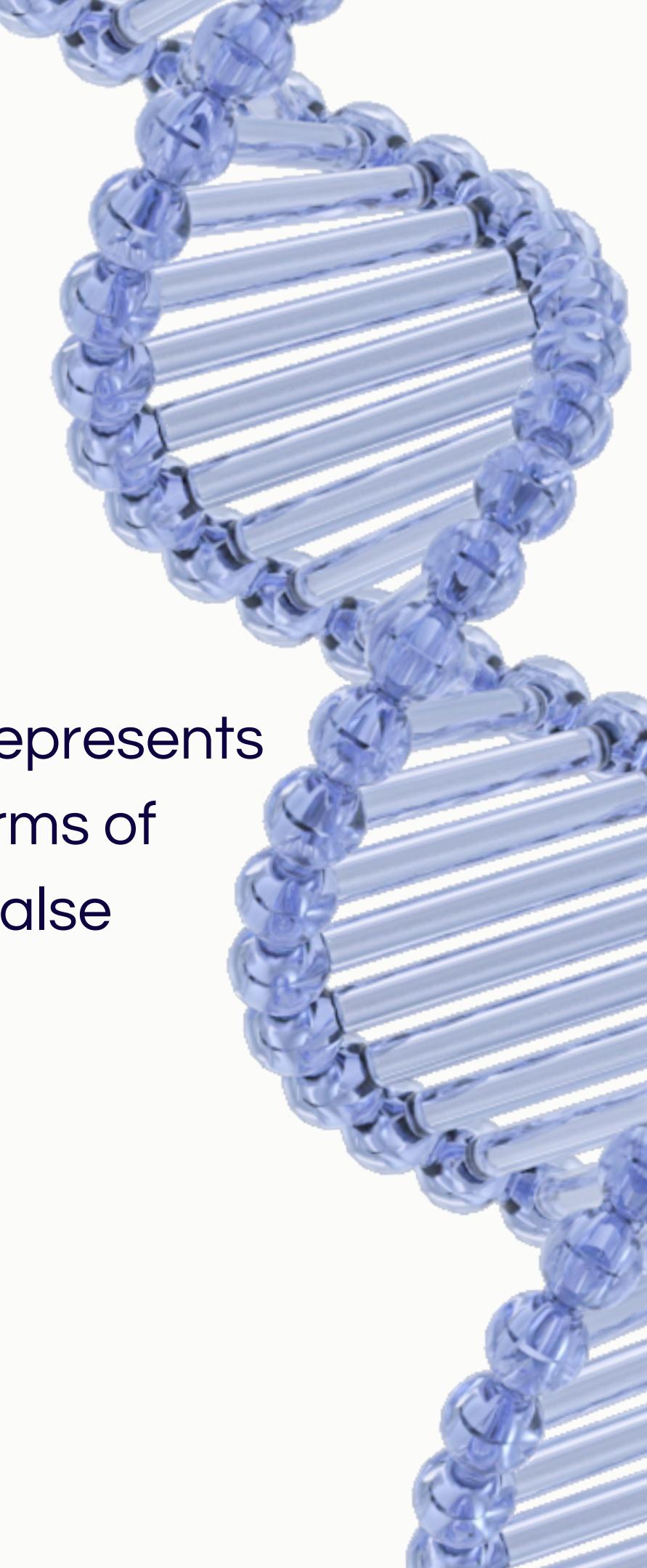




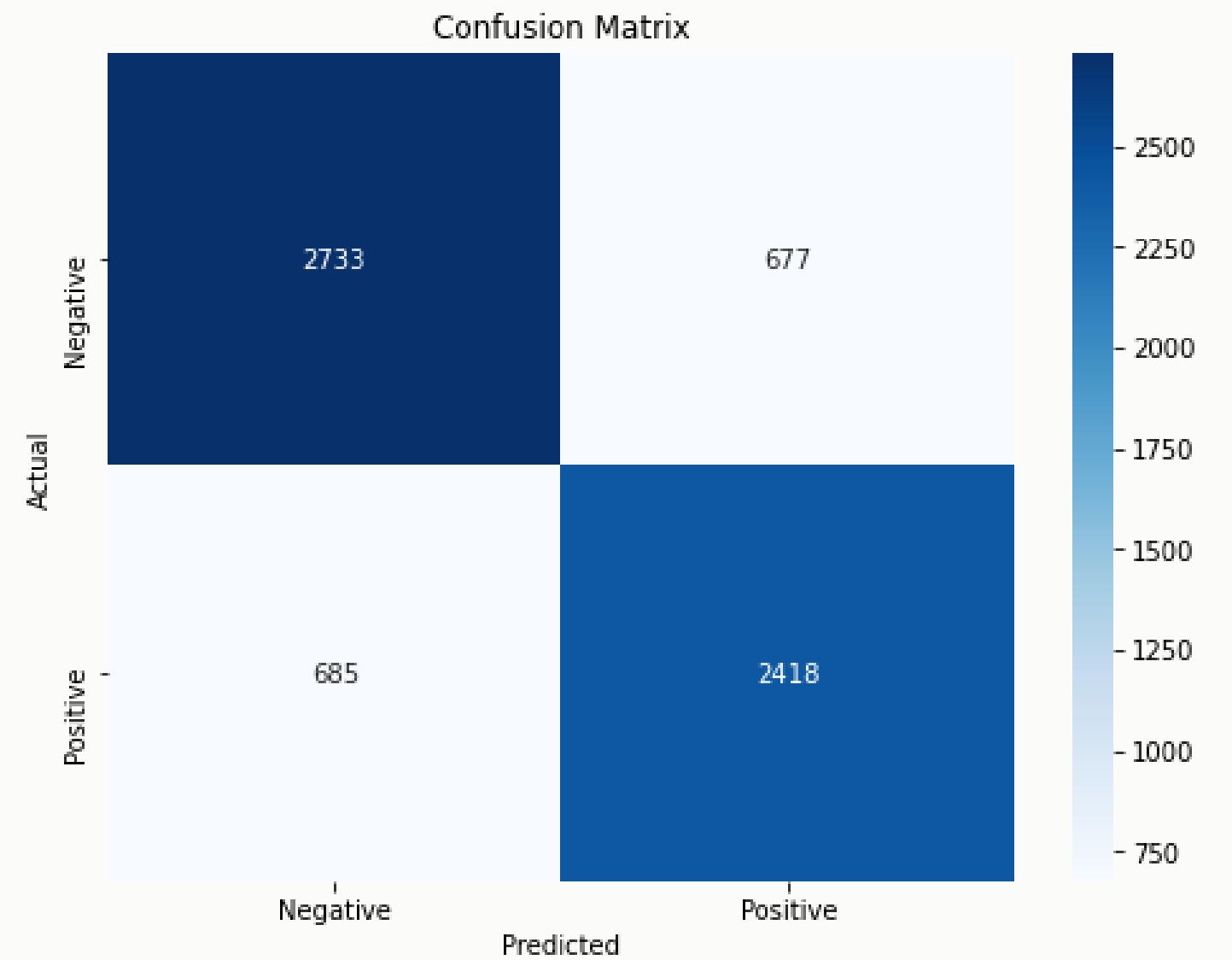
Model: Random Forest

Random Forest algorithm's model predicts the outcome based on the votes of the majority of all the decision trees in the forest, reducing overfitting.



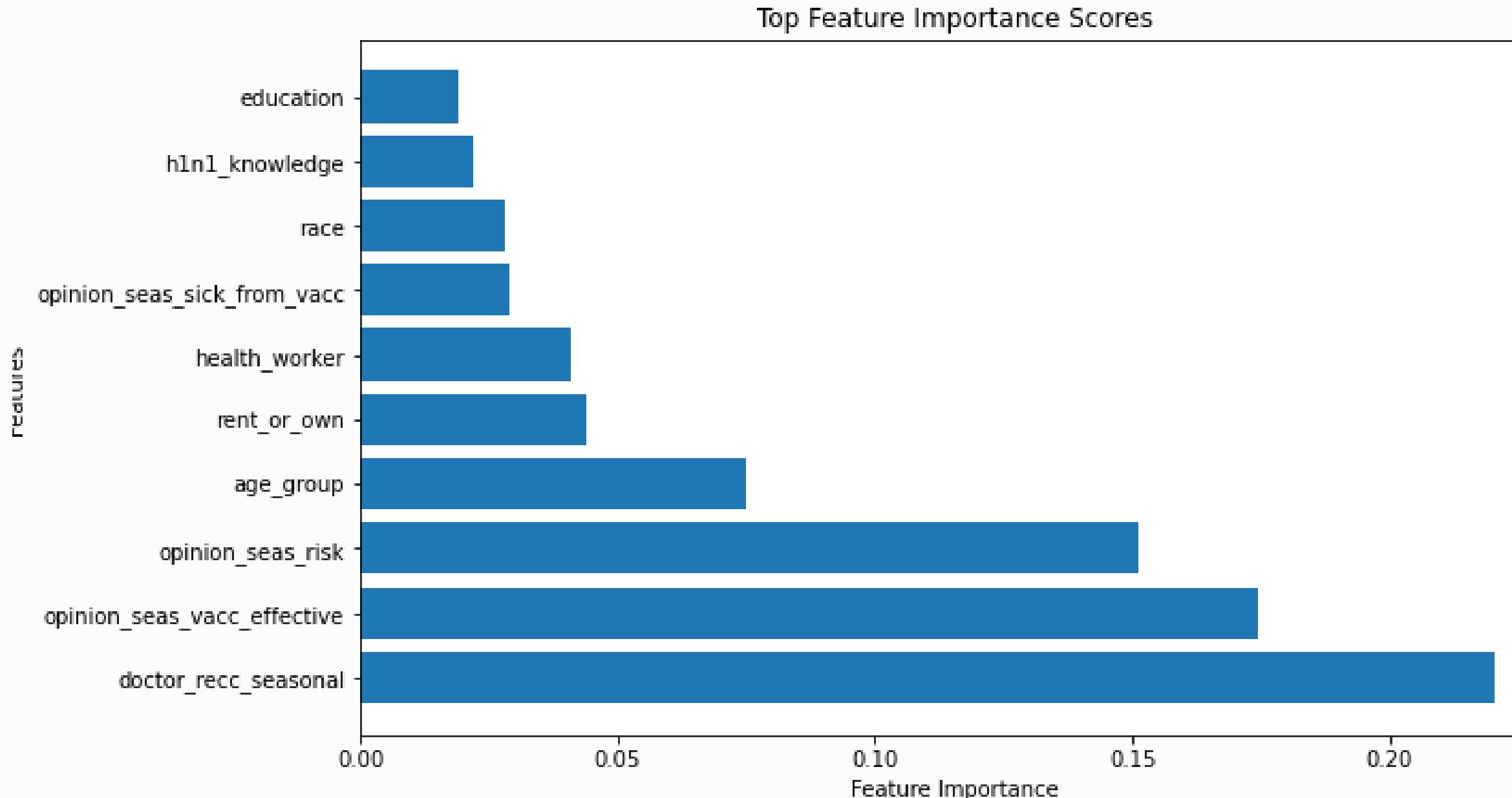


Model: XGBoost

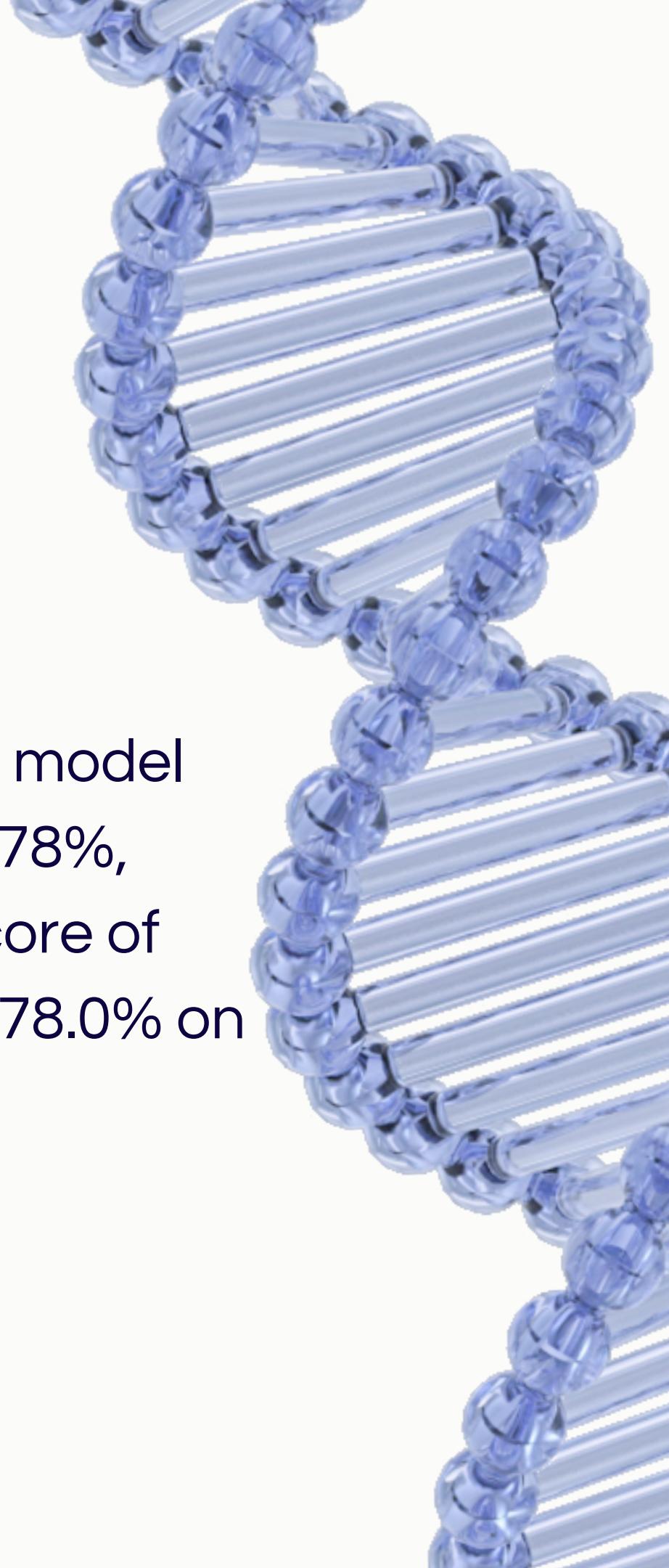


The confusion matrix visually represents the model's performance in terms of true positives, true negatives, false positives, and false negatives.

Model: XGBoost

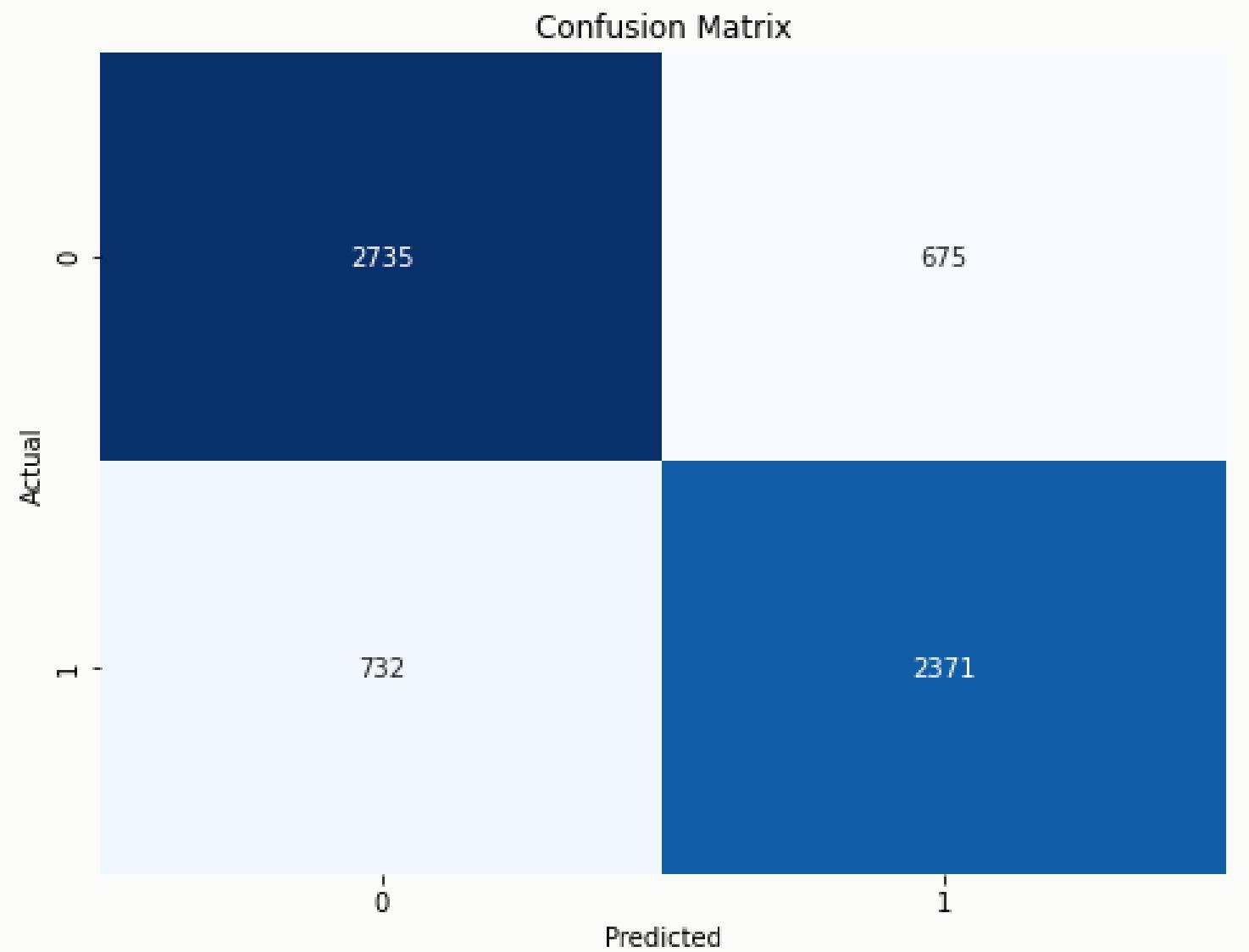


The results indicate an accuracy of approximately 0.79, and the top 10 important features are displayed in a bar plot, helping to understand the model's feature importance in the classification task.



Model: Logistic regression

The initial Logistic Regression model achieved an accuracy rate of 78%, precision 78%, recall 76%, f-score of 77% on the training data and 78.0% on the ROC AUC Score.



Conclusions

It was observed that XGBoost had the highest accuracy and balanced precision and recall, making it a strong performer in this binary classification task.

XGBoost demonstrated the best overall performance among the models tested. It achieved a higher accuracy, precision, and recall compared to other models. Therefore, it is recommended to use XGBoost for making predictions in this binary classification task.



Recommendations

Create a dashboard that includes the model's suggestions and a prediction of whether a patient will receive a vaccination.

Create persuasive information for healthcare professionals to tell their patients about the risk of the seasonal flu and the H1N1 flu

Patients are asked to complete a survey before seeing a doctor so that the doctor's office can determine if the patient is expected to get vaccinated or not.

