Prediction Model for H1N1 Vaccine Uptake

Business Understanding

Vaccines provide immunization for individuals, and enough immunization in a community can further reduce the spread of diseases through "herd immunity." In Katsiroumpa A. et al.,(2023) Predictors of Seasonal Influenza Vaccination Willingness among High-Risk Populations-Three Years after the Onset of the COVID-19 Pandemic, it was observed that a majority of the participants expressed reluctance or hesitation towards getting vaccinated.

Vaccine hesitancy is a well-known phenomenon, and the World Health Organization recognizes it as one of the top ten threats to global health. Various factors such as social, economic, and demographic background, opinions on risks of illness and vaccine effectiveness, and behaviors towards mitigating transmission may contribute to vaccine hesitancy. Having comprehensive knowledge of the traits linked to individuals' vaccination behavior can aid in the planning and implementation of future public health initiatives.

The goal of this project is to build a predictive model determining whether people got H1N1 vaccines using information shared about their backgrounds, opinions, and health behaviors.

Objectives

The main objective of this project is to build a model that will predict the whether individuals took the H1N1 Vaccine or not based off the information shared.

- To identify the most significant features in determining whether an individual is vaccinated against H1N1.
- To build a Decision Tree, Random Forest and SMV model that can accurately predict whether an individual is vaccinated against H1N1.
- To assess the performance of the predictive models and identify the best one and potential areas for improvement.

Data Understanding

Data Source

The dataset used for this project was obtained from Drivendata.

Data Description

The data was in csv format and contained training, test and label data. The dataset contained 26,707 rows and 38 columns. The 38 columns contained 26 numerical features and 12 categorical features.

Data Preparation

Loading the data

The necessary libraries were imported and then the training and label dataset was loaded onto the notebook.

Reading and checking the data

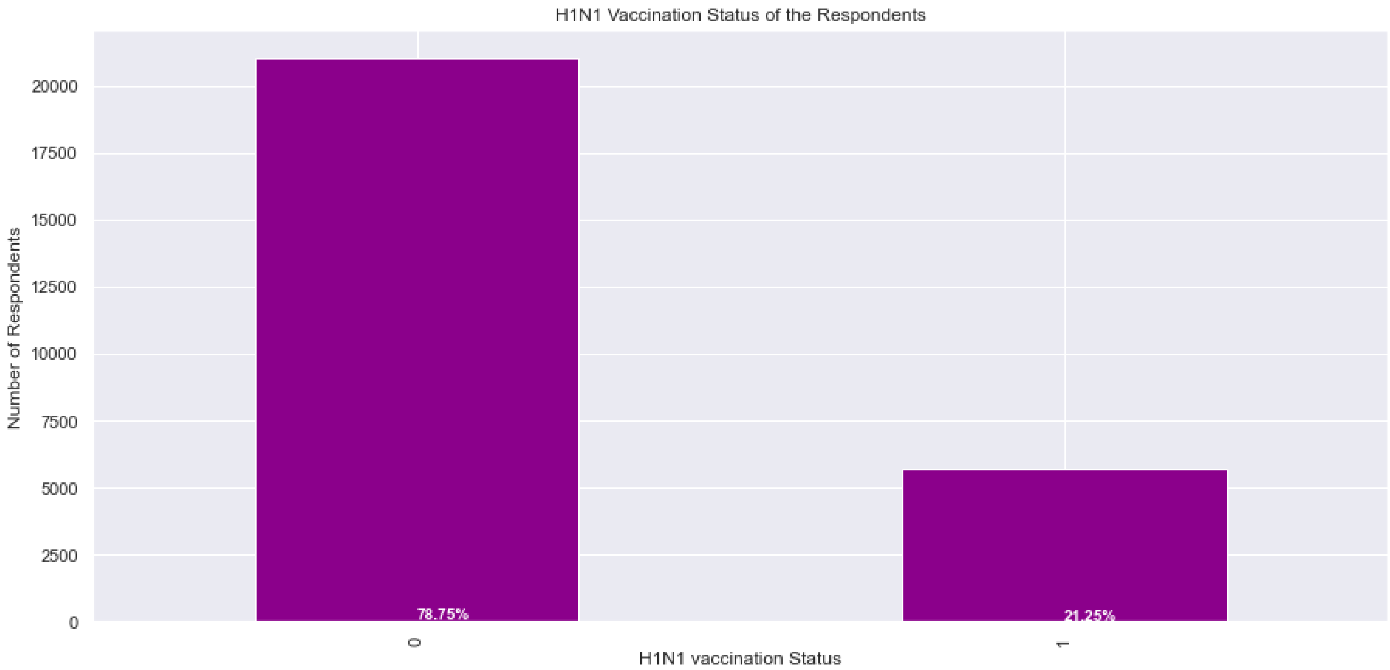
The training and label dataset was first merged, read and then checked to ensure that each column has an appropriate datatype. This also highlighted the columns with null values which would be tidied in the next few steps. Lastly, the various statistical measures in the data frame were checked for each feature.

Cleaning the data

The data was then cleaned and pre-processed to ensure that it was complete for modeling purposes. This included identifying the null values, computing the percentage of the missing values and dropping columns with missing data greater than 10%. Later, a data frame of the remaining columns with missing values was created and the numerical and categorical columns were cleaned separately. For the numerical columns, the null values were replaced with the mode while for the categorical columns null values were replaced with 'missing'. With the null values taken care off, all duplicated rows were identified and dropped accordingly and all columns relating to seasonal_vaccine were dropped. This was because H1N1 was selected as the target variable for the project.

External data source validation

The data set was measured against a reliable external data source Katsiroumpa A. et al.,(2023) *Predictors of Seasonal Influenza Vaccination Willingness among High-Risk Populations-Three Years after the Onset of the COVID-19 Pandemic*, where was noted that among participants, 39.4% were willing to accept the seasonal influenza vaccine, 33.9% were unwilling, and 26.8% were hesitant.



From the plot above, it is observed that 78.75% of the respondents did not receive the H1N1 vaccination.

The plot aligned with the results of the research conducted by Katsiroumpa A. et al.,(2023) on vaccine adoption in the Greek population. It is also worth mentioning that the study took demographic factors into account as possible predictors, which strengthens the data used in this analysis, which includes various demographic characteristics such as gender, age, number of children in the household, etc.

Exploratory Data Analysis

The data sets were analyzed and trends found by using statistics and visualizations to aid in comprehending the data set. There were several questions that were answered in this step by comparing the predictor variables with the target variable using data visualization tools. The questions answered and variable relationships established included:

- Does the sex of a respondent determine H1N1 vaccine uptake?
- Does the sex of a respondent determine H1N1 vaccine uptake?
- Does the sex of a respondent determine H1N1 vaccine uptake?

- Does the employment_status of a respondent determine H1N1 vaccine uptake?
- Does the marital_status of a respondent determine H1N1 vaccine uptake?

Modeling

Data modeling process involved pre processing the data first. This involved feature selection which is the process of selecting a subset of the most important features or variables from a larger set of features. By conducting feature selection and deleting noisy, redundant, or irrelevant characteristics from the data, feature selection aims to improve the performance of machine learning systems. A heat map showing the correlations between the various features and target variable was plotted. Afterwards, all unecessary columns and features with low correlation with the target variable were dropped. Lastly, all categorical features were encoded using ordinal encoding and the dataframe split and scaled.

The next steps involve building models using Decision Tree, Random Forest and SMV classifiers. The steps for each of these classifiers were similar and are broken down as follows:

i) Training a classifier.

ii) Making predictions for test data.

iii) Calculating accuracy, precsion and recall.

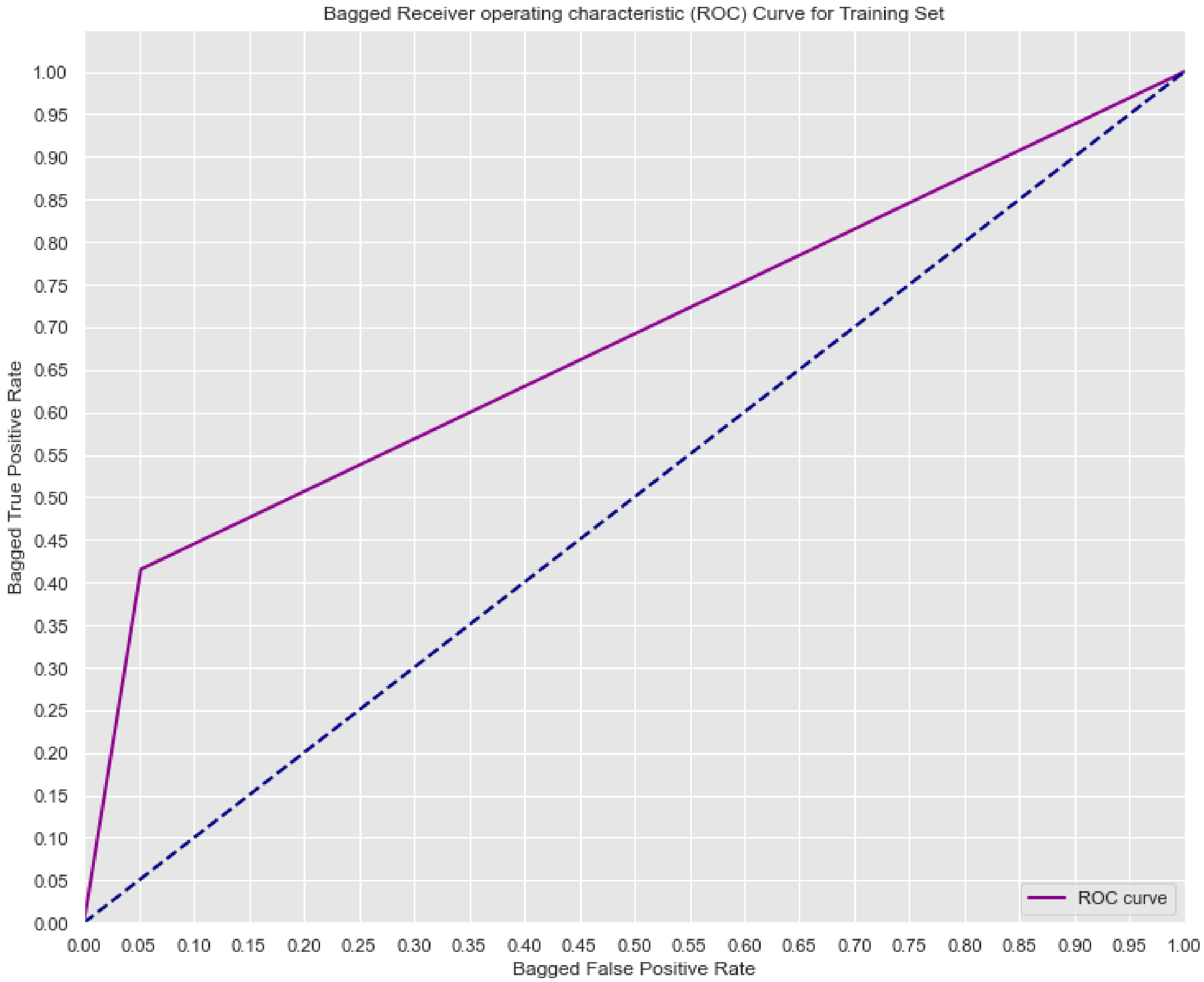
iv) Using GridSearchCV to tune the classifier’s hyperparamenters.

v) Insantiating the model with the best parameters from grid search.

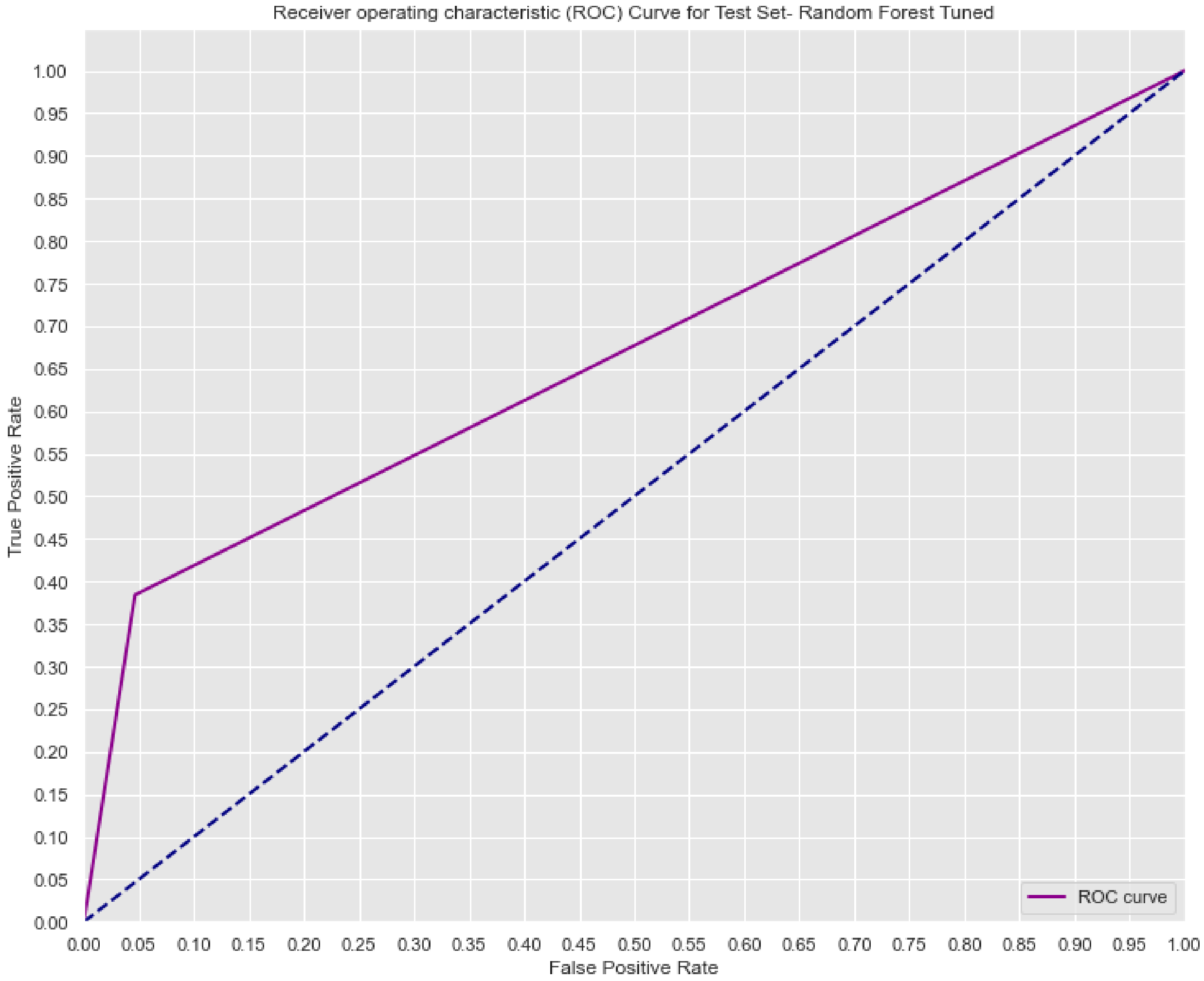
vi) Making predictions for test data using the tuned model.

vii) Calculating accuracy, precsion and recall once more.

These steps were used in all models with the exception of decision tress where Bootdtrap Aggregation (Bagging) was carried out to reduce the variance of the model and improve its generalization performance.



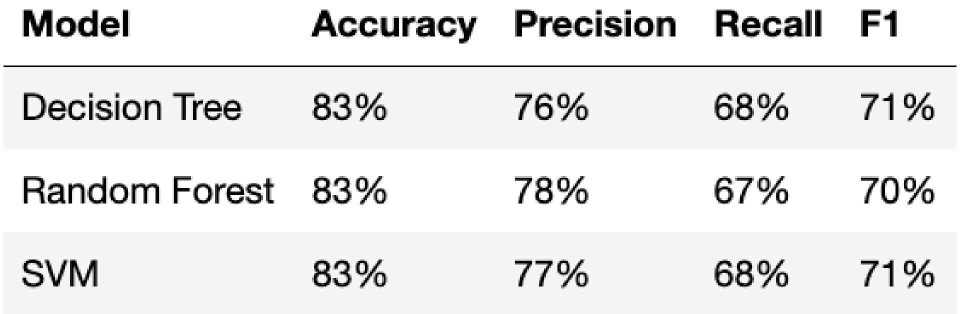
When using the decision tree classifier the models accuracy score increased from 76.17% to 83.57% which it states that 83.57% of the data was correctly predicted. Additionally, the recall is 64%, precision 65% and f1 score is 64% which were fairly good scores and no further tuning was done. The ROC curve was plotted and is shown below.



When using the random forest classifier remained the same at 83.17%. Moreover, the recall increases to 67% from 66% and the f1 score to 70% from 69% while precision decreases to 77% from 78%.

Lastly, Support Vector Machine was the final model classifier used. After, a long wait of loading the optimum hyperparamenters , the various scores were inserted in the model and a good working model created. Even though some scores decrease, the main focus was on precision which is very important when dealing with the data set on H1N1.

Conclusion



Model	Accuracy	Precision	Recall	F1
Decision Tree	83%	76%	68%	71%
Random Forest	83%	78%	67%	70%
SVM	83%	77%	68%	71%

The table above shows a summary of the various scores from the best working models. It was important to establish what score is more compatible with the data set on the H1N1 vaccine. As discussed earlier, this model will be used by governments and NGOs during vaccine drives to better help them plan their resources better and even carry out awareness campaigns. During such preparations, it would be crucial to ensure that resources are not wasted and that teh campaigns have an impact. It is for this reasons that precision is important as maximizing precision will minimize number of false positives which could lead to resource wasting.

However, the precision scores are pretty close and thus the recall score will also guide in selecting a model. From a careful review of the table, SVM presents a well balanced scores and would serve the purpose of accurately predicting whether or not individuals would get vaccinated.