

PREDICTING HOW LIKELY INDIVIDUALS ARE TO RECEIVE H1N1 VACCINES

By Caren Chepkoech

INTRODUCTION

The prevention of infectious diseases depends heavily on vaccination which is a critical component of public health. But not everyone reacts to immunization in the same way. People react differently, they may experience different side effects such as fever, feel dizzy or soreness.

Our objective is to create a reliable classification model that can correctly evaluate a person's response to h1n1 vaccine based on specific characteristics. The end result of this initiative will help healthcare practitioners to make decisions about the delivery of the vaccine.



CONTENT

i.Business Understanding

li.Exploratory Data Analysis

lii.Modelling

Iv.Conclusion


v.Recommendations and Future Steps



BUSINESS UNDERSTANDING

Problem Statement

Vaccination has become a key public health measure that is used to fight and curb infectious diseases. The aim of this project is to build a model that can predict how individual will respond to vaccine based on various factors such as age, sex, health status and their knowledge on h1n1 vaccine. This information will help medical practitioners make informed decisions about who should get the vaccination and who should not receive vaccination and how best to manage its administration.




OBJECTIVES

To build a classification model that can predict the reaction of individuals to a vaccine based on certain factors.

Identify which factors affects individual's reaction to vaccine

To accurately predict the general response of individual's response to a new vaccine

Build a classification model that accurately predicts the response of individuals to new vaccines and provide actionable insights on how to prevent infectious diseases.



METRIC OF SUCCESS

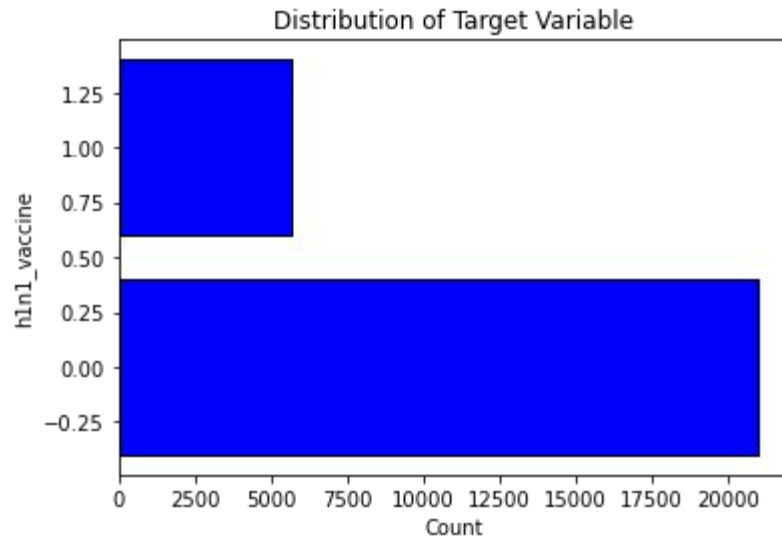
The final model will be considered a success if it has an accuracy and f1 score of not less than 75%. The goal is to make as accurate as possible predictions, that's why the choice of success metrics is the accuracy score and f1 score.



Exploratory Data Analysis

Univariate Analysis

The goal of this analysis is to establish the proportion target variable `h1n1_vaccine`. from the graph only 5000 people out of 20000 took the vaccine

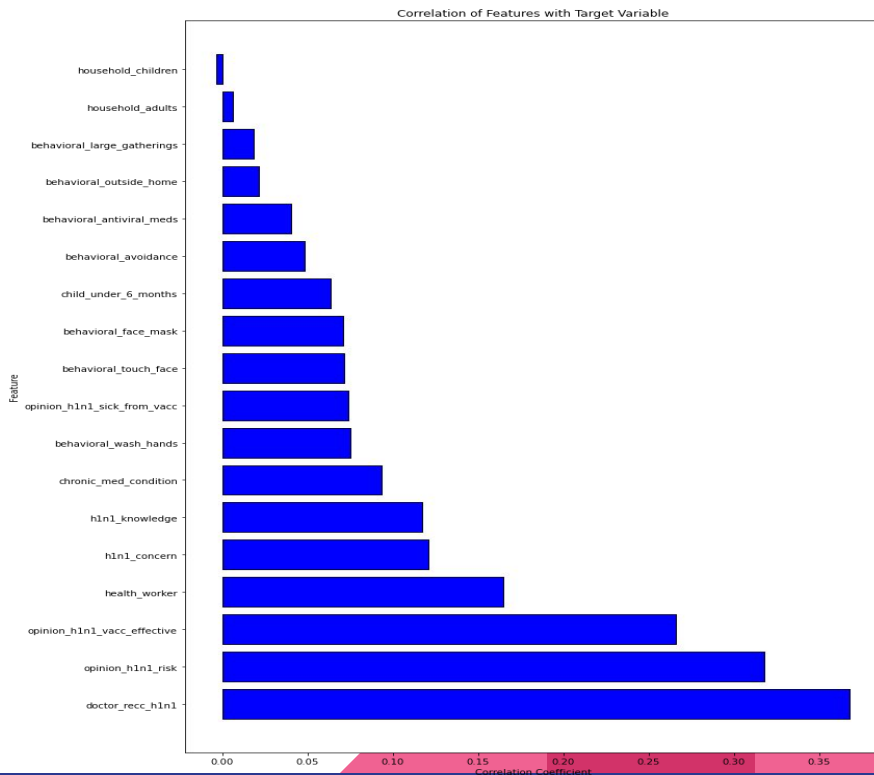


Bivariate Analysis

The correlation coefficient is generally not high for most of the variables

However, there are some that are more correlated to the target variable like doctor_recc_h1n1, opinion_h1n1_risk, and opinion_h1n1_vacc_effective.

These seem like great variables to work with increase prediction accuracy.



MODELING

Model 1 - Logistic Regression

Logistic regression model was fitted on the training dataset and this resulted in an accuracy score of 81% and an f1-score of 43%. This model is not the best model to use for predictions because with an f1-score that is low, the model is vulnerable to making more wrong predictions than right on



Model 2 - K Nearest Neighbour

Before this model was fitted, feature selection was performed in an attempt to improve the outcome. Features that were not highly correlated with the target variable were dropped and only 10 were left to be used in training and fitting the model. The outcome was an accuracy score of 79% and an f1-score of 36%. This model was performing worse than the previous one.




Model 3 - Decision Tree

The decision tree was fitted to the dataset and the outcome was an accuracy score of 78% and an f1-score of 36%. This model was also not doing well because of the low f1-score.

Hyperparameter Tuning

This was carried out in order to improve the model performance by introducing hyperparameters and tuning them until the best combination was found and applied to the model. Grid search cv was used to carry this out on the decision tree model and the outcome was a model with 80% accuracy score and 78% f1-score. This model surpassed the success metric before and would therefore be a well performing model in terms of prediction accuracy.



CONCLUSION

After experimenting with different models using different techniques, the final model (the decision tree), which was tuned by hyperparameters was selected. This is because it met the success criteria by recording an accuracy score of 80% and f1 score of 78%. The first three models did not do so well because they were not complex enough to handle the nature of data presented. This is why while the accuracy score was high, the f1 score remained low. By adjusting the hyperparameters, we were able to find the optimal complexity that balanced overfitting and underfitting.



RECOMMENDATIONS AND FUTURE STEPS

More research on other vaccines should be carried out to supplement this on h1n1.

Future surveys to be carried out in person.

An improvement in data preprocessing techniques is paramount for future future projects.

A research on why people are averse to vaccination should be carried out

