Table of Contents

Executive Summary

It's Q1 2022, and COVID is still rampaging through Canada. Vaccination rates have stalled, and the Government of Canada is looking for ways to overcome the hump. To this end, they are evaluating the possibility of a nation-wide vaccine mandate for all adults.

Our Data Science team at the Public Health Agency of Canada, have been contracted to consult the government on determining public sentiment regarding the implementation of legislation for a vaccine mandate. Using COVID behavior data obtained from Institute of Global Health Innovation Imperial College London, we develop a model that allows us to predict with over 68% accuracy whether an individual will be for or against a vaccine mandate for adults. Our model can be deployed by the Government of Canada to create a concise survey without including the potentially divisive and direct question.

Our model suggests that by analyzing responses to these specific questions, the government can predict whether an individual will support or oppose a vaccine mandate:

Question	Response Option
Age (Adults over 18 only)	Numeric (integer)
Gender	Male or Female (binary)
Number of people in your household	Numeric (integer)
How many days a week do you feel depressed or down?	1 (Not at all) to 4 (almost every day) and 5 (Prefer not to answer)
How often do you wear a face mask inside the grocery store?	1 (Always) to 5 (Not at all)
Have you avoided attending public events, such as sports matches, festivals, theaters, clubs, or going to religious services?	1 (Always) to 5 (Not at all)
Did you avoid mixing with other households indoors?	1 (Always) to 5 (Not at all)
How well or badly do you think the Government are handling the issue of the Coronavirus (COVID-19)	1 (Very well) to 4 (Very badly) and 5 (Don't know)
I believe government health authorities in my country will provide me with an effective COVID19 vaccine.	1 (Strongly agree) to 5 (Strongly disagree)
A vaccine for coronavirus (COVID-19) will protect me against any variants, strains, or mutations of coronavirus.	1 (Strongly agree) to 5 (Strongly disagree)

Introduction

Background

The COVID19 pandemic caused much concern across the world for years. Due to the nature of the threat caused by the virus's spread, governments were tasked with creating legislation and mandates to serve as the framework within which the contagion was to be controlled. Arguably the most impactful factor in containing the virus was the distribution of an effective vaccine for their population.

The Government of Canada approached this problem from two angles. One being the logistics problem: ensuring the procurement and physical availability of the vaccine for its citizens. The second, and arguably more challenging, was the enforcement problem. With power to create legislation and orders "to protect the health and well-being of Canadians" (Government of Canada), the Government was faced with the problem of solving an even bigger problem that the logistics of the vaccine. To figure out, the extent to which the government can use its power to ensure the highest vaccination rates possible.

With a great degree of fear and skepticism in the air, the availability of vaccine for a Canadian did not mean they would be receiving it. Individuals have agency over their bodies, and receiving the vaccine is an individual-level decision. There is a degree of soft power a government can try to leverage to push its narrative upon the public. However, the receptiveness of the masses can vary vastly based on many factors.

The Government has the power to create a vaccine mandate through legislation to force individuals into receiving the vaccine. However, this is a risky option, as it overrules the individual's self-determination of a bodily choice, at a time of intense emotion. There are many consequences of such a decision, as it would risk public trust, potentially erode democratic integrity, and many more ripple effects to come.

Objective

As part of the Public Health Agency of Canada, we are tasked with advising the government on what the public sentiment will be for a vaccination mandate. Through our paper, we present a model that can accurately predict on an individual level, whether a person will be for or against a legislated vaccine mandate for adults. Our model yields a short survey for individuals to fill out, without directly asking this sensitive question, to predict how they will respond.

Methodology

To construct and train our model, we use data collected by the Institute of Global Health Innovation Imperial College London. The set we use contains COVID-related behavioural survey responses from Canadians in 2022. After cleaning, analyzing, and modeling this data, we identified ten insightful questions that yield insight on basic characteristics of a person, their lifestyle during the pandemic, and thoughts on perspectives on the role of government at these times. With these carefully chosen and validated response categories, we create a model with ten predictors which is highly interpretable and provides an accurate prediction. The final model is shared in the Model Construction section as we work through our process in the following sections.

Data Cleaning & Processing

Identifying target variable from raw data

The complete dataset had 6431 observations, with 511 response variables each. After a preliminary screening regarding missing values and inapplicable columns, we were able to narrow the list of model variables to about a tenth of the original dataset. With our objective being to advise the government on sentiment regarding vaccine mandate policy, we easily identify vac_man_6 as the target variable.

Column Key	Value	Refined Question	Response Option
vac_man_6	All adults	Do you think vaccinations should be mandatory for all adults?	Yes or No

Overcoming the slight interpretability challenge of the column's name (*vac_man_6*) and data key definition being vague (*All adults*), through observing the labels and definitions of other questions asked in this category -especially vac_man_99 (*I don't think vaccinations should be mandatory for anyone*) , we derive the full definition of our target variable observations as being "Do you think vaccinations should be mandatory for all adults?".

Identifying candidates for predictor variables

From the now reduced pool of columns in our questionnaire data, we qualified predictor variables through their ability to collectively paint a picture of one's sentiment on living in a COVID dominated environment. The ones our team identified as being meaningfully important and statistically valid -at a glance, were:

Column Key	Refined Question	Response Option
age	Age (Adults over 18 only)	Numeric (integer)
Gender	Gender	Male or Female (binary)
household_size	Number of people in your household	Numeric (integer)
PHQ4_2	How many days a week do you feel depressed or down?	1 (Not at all) to 4 (almost every day) and 5 (Prefer not to answer)
i12_health_22	How often do you wear a face mask inside the grocery store?	1 (Always) to 5 (Not at all)
i12_health_26	Have you avoided attending public events, such as sports matches, festivals, theaters, clubs, or going to religious services?	1 (Always) to 5 (Not at all)
i12_health_27	Did you avoid mixing with other households indoors?	1 (Always) to 5 (Not at all)
WCRex1	How well or badly do you think the Government are handling the issue of the Coronavirus (COVID-19)	1 (Very well) to 4 (Very badly) and 5 (Don't know)
vac2_3	I believe government health authorities in my country will provide me with an effective COVID19 vaccine.	1 (Strongly agree) to 5 (Strongly disagree)
vac2_7	A vaccine for coronavirus (COVID-19) will protect me against any variants, strains, or mutations of coronavirus.	1 (Strongly agree) to 5 (Strongly disagree)

Processing for analysis

With our variables of interest selected, we engaged in three key data cleaning and preparing activities to prepare our dataset for analysis:

Map categorical data

Most data of predictors like PHQ4_2, WCRex1, i12_health_22, i12_health_26, i12_health_27, vac2_3, vac2_7 and gender are categorical data and are predefined, so we need to map them with number 1, 2, 3, 4 ,5. Given that each predictor does not have too many levels (less than 5), we do not consider reducing categories, which ensures the accuracy of our data. Also, the evaluation standards for each predictor are different. For example, one in variable PHQ4_2 means positive result, but one in variable i12_health_22 represents negative result. To unify these evaluation criteria, we decided to use 1 to represent the most negative rating and 5 to represent the best rating, and so on.

Create dummy variables

For gender and the response, only two types (male and female: yes or no), so we choose to dummy them by mapping response values of Male to 1, and Female to 0. Similarly, we dummy our target variable vac_man_6 so that the value is 1 for a response of Yes, and 0 for No.

Drop data

As stated earlier, we first remove columns without values. These columns are present in the dataset, as the data is obtained from a global survey which housed many locally instanced surveys across the world. In the household size column, there are two categories called ‘prefer not to say’ and ‘I do not know’, which only occupy no more than 4% in the whole data and we drop these observations as it does not cost us much to do so with respect to data size and result accuracy. Finally, the total number of the data size is 6170, which is statistically significant to perform modeling.

Exploratory Data Analysis

We begin with a correlation matrix (Fig. [1]) to identify predictor variables that could threaten the integrity of our model.

Generally, variables paired with each other are not correlated (fig. [1]). The highest two correlations are between vac2_3 and vac2_7 (0.63), and between i12_health_26 and i12_health_27 (0.6). Since they are not perfectly correlated to each other, we should not omit any of them at this stage. No collinearity issue here. For further investigation, we might choose to apply principal component analysis to these variables.

The target variable has the strongest correlation with variable vac2_3 (-0.32), followed by vac2_7 (-0.3) and WCRex1 (0.26). It has little correlation with gender (-0.0034) and PHQ4_2 (0.052).

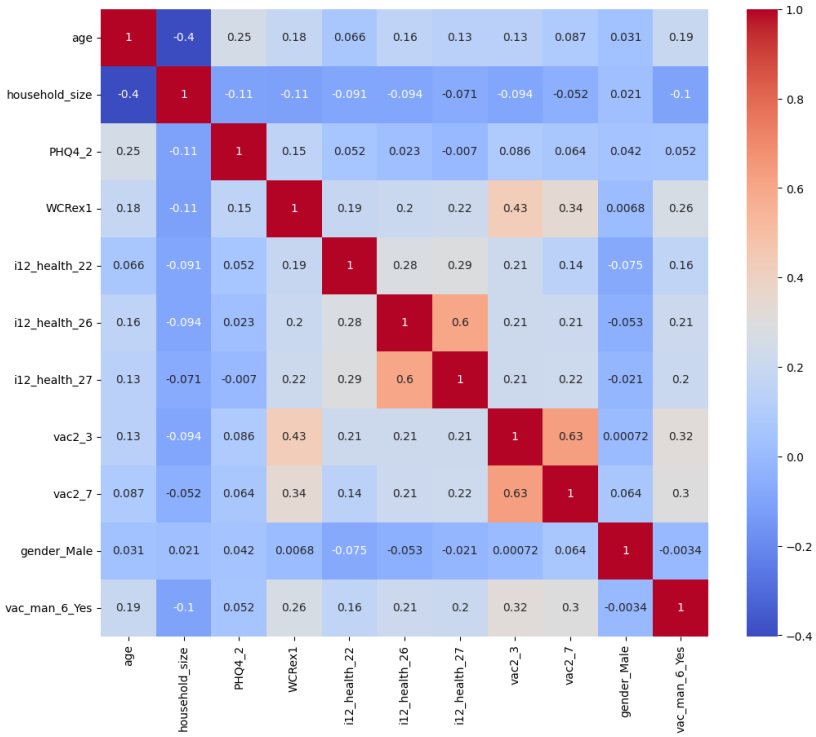


Fig. [1] - Correlation heatmap of all variables

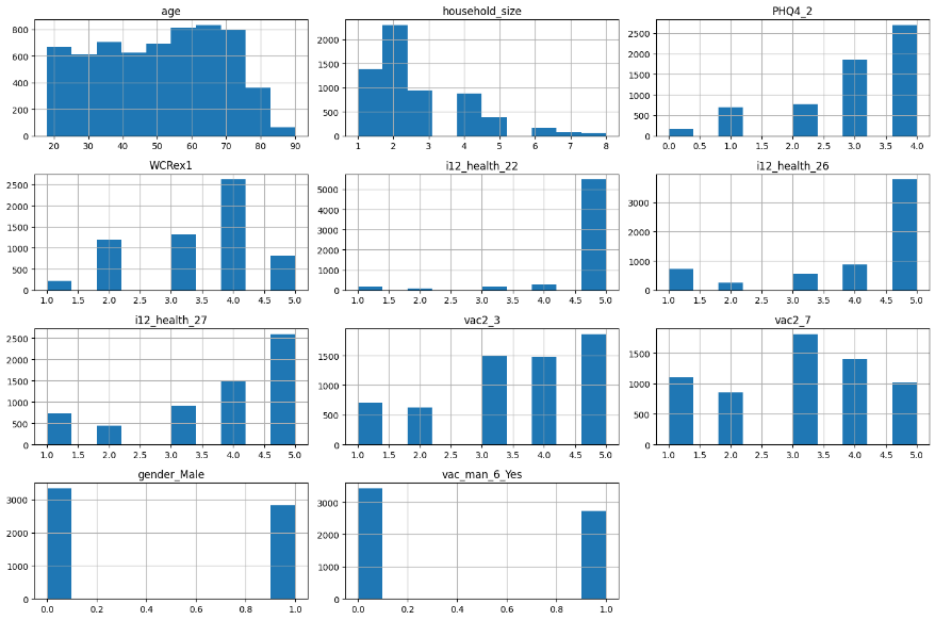


Fig. [2a] - Histograms of all variables

Now absent of any critical collinearity problems, we continue to investigate the distribution of observations our variables. We do this through the help of two forms of visualizations: Histograms (Fig. [2a]) and Density Plots (Fig. [2b]) of all selected variables -latter on the next page.

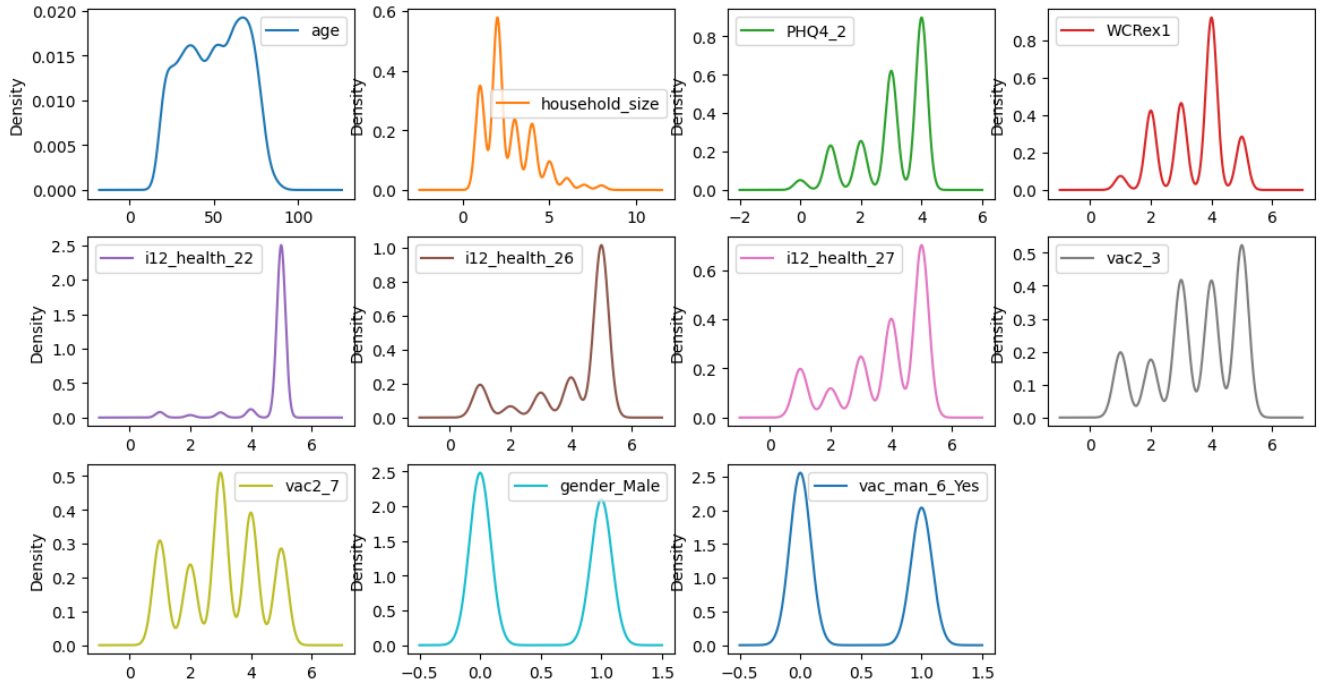


Fig. [2b] - Density plots of all variables

Several variables (vac_man_6_Yes, WCRex1, vac2_3, and vac2_7) show a cyclical or periodic distribution, suggesting they might be ordinal or categorical in nature. Variables like age and household_size provide insights into their real-world meanings. For instance, fewer large households and older individuals. Some variables, such as PHQ4_2 and i12_health_27, show multimodal distributions, indicating multiple groupings or categories within them. As the responses align with what would be expected behaviour within the COVID environment, we proceed to investigate how the distribution of responses to our target variable looks through normalized response histograms:

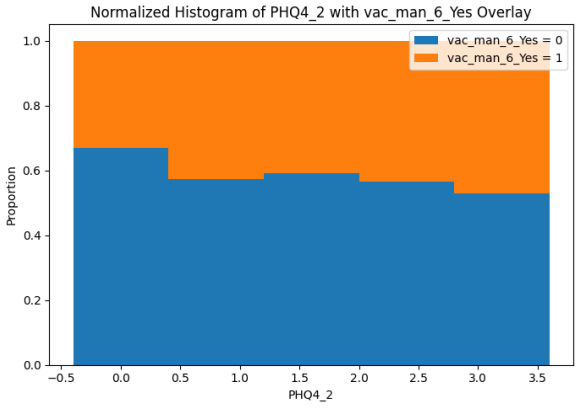


Fig. [3a] – Normalized Response Histogram, PHQ4_2 (depression/downness frequency)

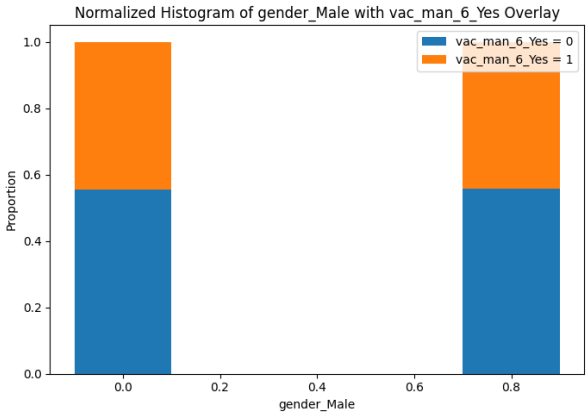


Fig. [3b] – Normalized Response Histogram, gender

Based on these ten normalized histograms, the change in gender does not have a significant effect on the target variable as well as the household size variable and PHQ4_2 variable.

For the other variables, we see that as the response values increase, we can see the number of people who favour the vaccine mandate shows an increasing trend. Older people are more likely to agree to the mandate, as are those who avoided public events, and those who avoided more mixing with other households indoors. From a government sentiment standpoint, those who believe that the government will provide support the vaccine mandate. As are those who believe in the vaccine's efficacy against the virus and any variants. Due to space constraints, you may find the rest of the normalized response histograms in the Appendix (Fig. 3c).

Model Construction

Balanced Dataset Split

Prior to modeling, it was essential to ensure the quality and reliability of our dataset. One of the critical checks performed was to ascertain the balance of our data, particularly in terms of the distribution of target variable classes. Through thorough analysis, we confirmed that the dataset is indeed balanced for the target variable's classes and proportions are relatively close.

The dataset is split into a train set and test set. 70% of observations are the training set and the remaining 30% are the test set.

Training Set	Test Set
70%	30%

Normalization

KNN relies on calculating distances between data points to determine the nearest neighbors. Features on large scales can disproportionately influence the distance metric, making the algorithm biased towards those features. However, by normalizing, we can ensure that the computed distances are not dominated by one or more features due to their scale, providing a more balanced and accurate classification result.

We split the training set and testing set before normalization because some normalization methods use sample statistics, such as mean and variance of the dataset to scale, which means the information from the test set is being used to scale the training set and vice versa. By splitting the first and then normalizing, we can prevent the 'data leakage' phenomenon from influencing the training data processing.

In choosing an appropriate normalization method, it is often necessary to make judgments based on the characteristics and distribution of the data. Based on the EDA plot before, we decided to use Box-Cox transformation and Z_score normalization to do normalization.

1. Box-Cox transformation

   Predictors 'household size', WCRex1', 'i12_health_22','i12_health_26' and 'i12_health_27' exhibit severe skewness, so we need to normalize these columns with the Box-Cox method. This method is effective for normalizing data that does not follow a normal distribution.

2. Z_score normalization

   Given that the range of the predictors 'age', 'vac2_3' and 'vac2_7' varied significantly, and they were near normally distributed, Z_score normalization is applied. After transformation, these data will obey a normal distribution with mean 0 and standard deviation 1.

Model Construction

In this dataset, we used many-to-one classification, and it is a two-classification problem. K-nearest neighbors model is applied to make classification. The framework of KNN is to find points in the dataset that are closest to an observation we want to classify. The model will label this observation based on what are the closest observations to it, with respect to parameters we give it, and how close these points are to it.

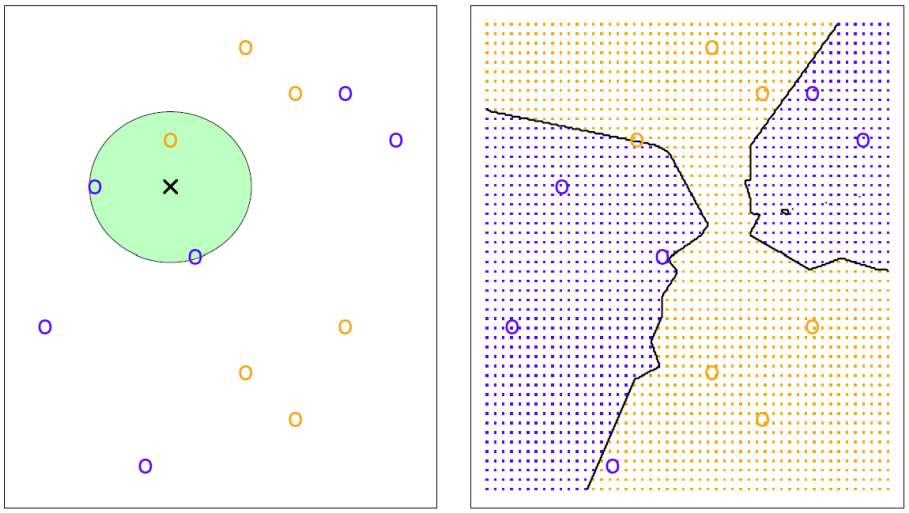


Fig. [4] - Framework of the KNN model ((James, Witten, Hastie, & Tibshirani, 2021, p. 40)

We have almost 6170 observations in our dataset. Given that we want to avoid large biases in the model, we do not want to choose a large value of k. So, we set our k-fold number as 10 to do cross validation.

Model Result

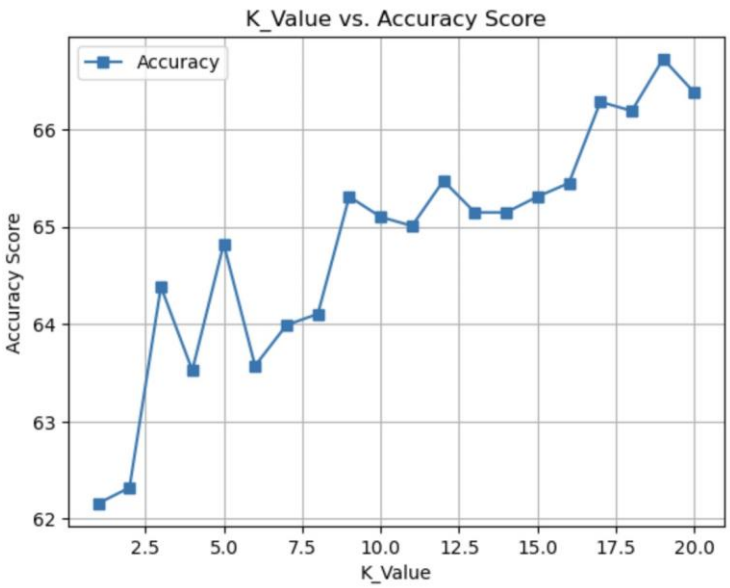


Fig. [5] - K Values Vs. Accuracy Score

We use the model's accuracy as the primary metric under evaluation. The figure above illustrates the accuracy score for different k values. From the results in Fig. 5, we can see that the model demonstrates accuracy levels peaking around 68% when k=19.

In the KNN model, the selection of k is pivotal, as it dictates the number of neighbors factored in for classification. Commonly, a k value less than 10 is favored to mitigate the effect of noise, optimize computational efficiency, and strike a harmonious balance between overfitting and underfitting. In our case, we might be tempted to test higher values for k as we are seeing a strong trend of increase in accuracy. However, our dataset in not large enough to viably consider using a larger value for k.

To decide on the optimal k for this problem, we rely not just on accuracy, but also on supplementary metrics: 'sensitivity', 'specificity', and 'precision'. These metrics provide a more comprehensive view of model performance.

Given our objective — determining whether an individual believes all adults should be mandated to get vaccinated, we lean heavily on the sensitivity metric. This decision stems from the presumption that the government promotes vaccination, and hence, detecting true positive cases (those who favor vaccine mandates) is paramount.

Observing Fig. 6 below, we see that the precision rate drops dramatically when k = 8, so we pass it first. Given our objective — determining whether an individual believes all adults should be mandated to get vaccinated, we lean heavily on the sensitivity metric. This decision stems from the presumption that the government promotes vaccination, and hence, detecting true positive cases (those who favor vaccine mandates) is paramount. Therefore, after evaluating all the metrics holistically, and considering a balance between sensitivity and precision, we select k = 5 as the most appropriate choice for our model and proceed as such.

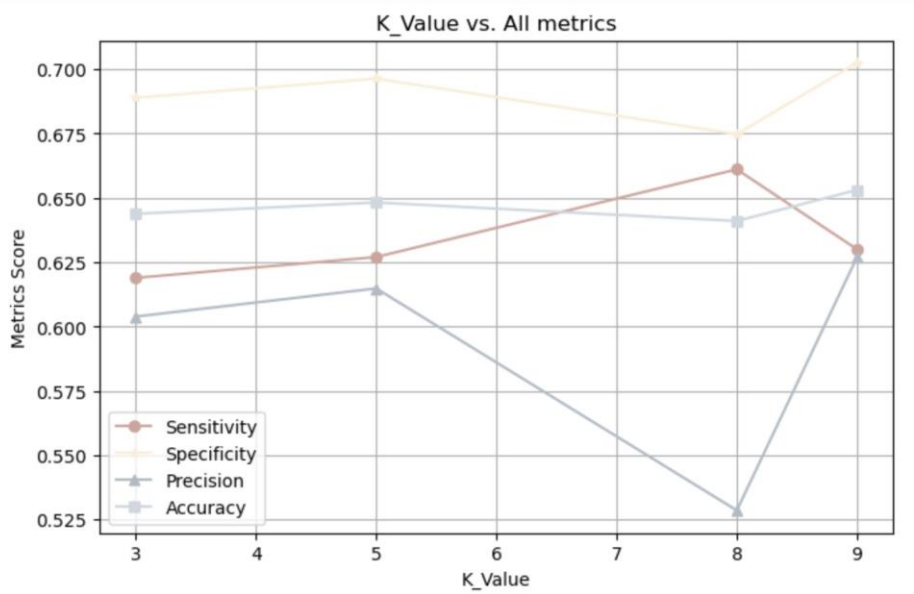


Fig. [6] - K values vs. All metrics

Finally, we present the Confusion Matrix for the test data in Fig. 7, resulting from using the best model parameters (k = 5). Key model metrics of accuracy, sensitivity and precision rate are 65, 63 and 61.

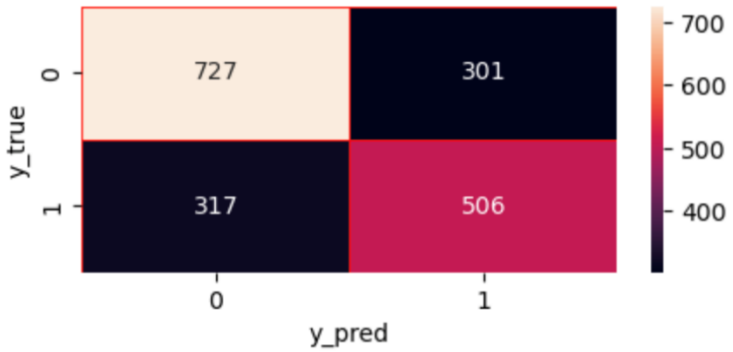


Fig. [7] - K values vs. All metrics

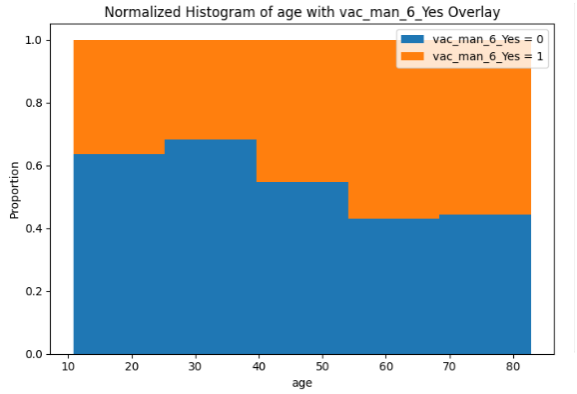
Conclusion

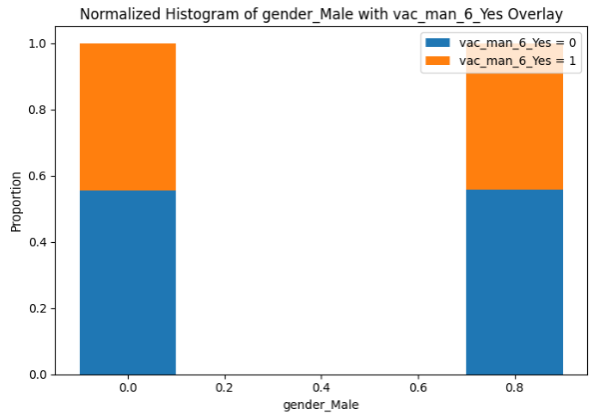
The accuracy of the model indicated room for improvement. Upon reflection, some adjustments to strive towards this would include dropping the “How often do you wear a face mask inside the grocery store?” predictor (key: i12_health_22). This is due to its limited utility, as the density plot reveals a concentration of observations with little variation. When we add intuition and context to the data, we can recall that masks were mandated in grocery stores around Q1 2022, which the window of time within the survey was presumably conducted. A preliminary run with this predictor removed shows an accuracy of 4% at k=5, lifting the model to 72%.

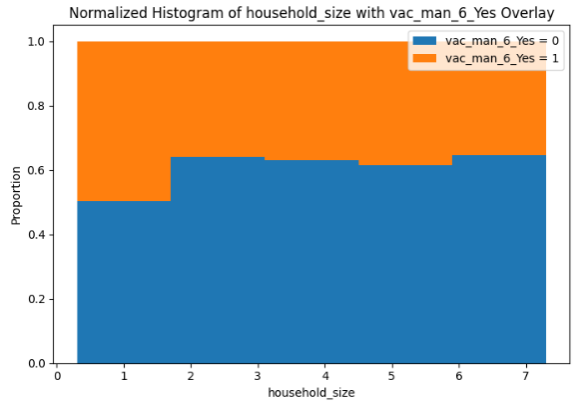
From a strategic perspective, our results yield a model that could pave the way for a brief questionnaire, used to predict an individual’s sentiment on adult vaccine mandates. Perhaps more importantly, we establish that through roundabout sentiment-oriented questions, we can indirectly get answers to tougher-to-ask questions. Furthermore, we emphasize the importance of incorporating individuals' sentiments and opinions into the legislative process within a healthy democracy.

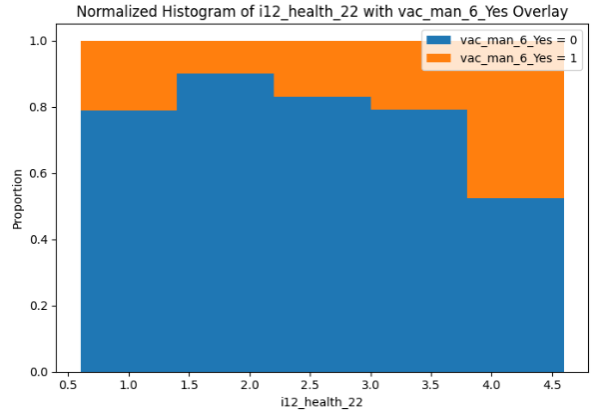
Appendix

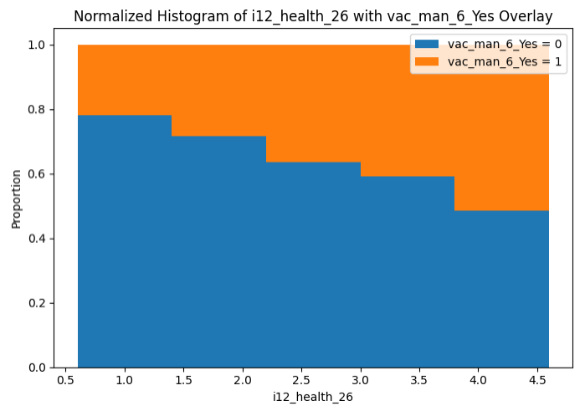
- Fig. [3c] - Normalized Response Histograms for all predictor variables:

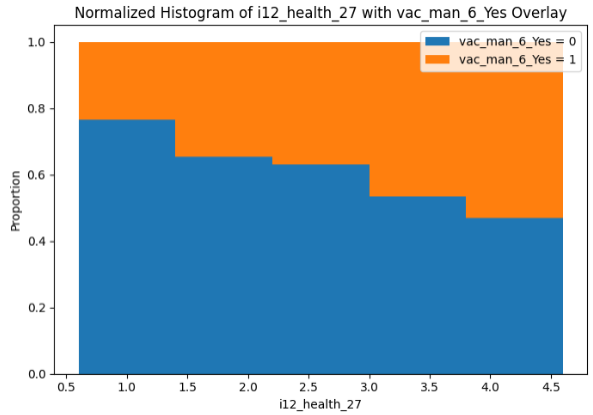


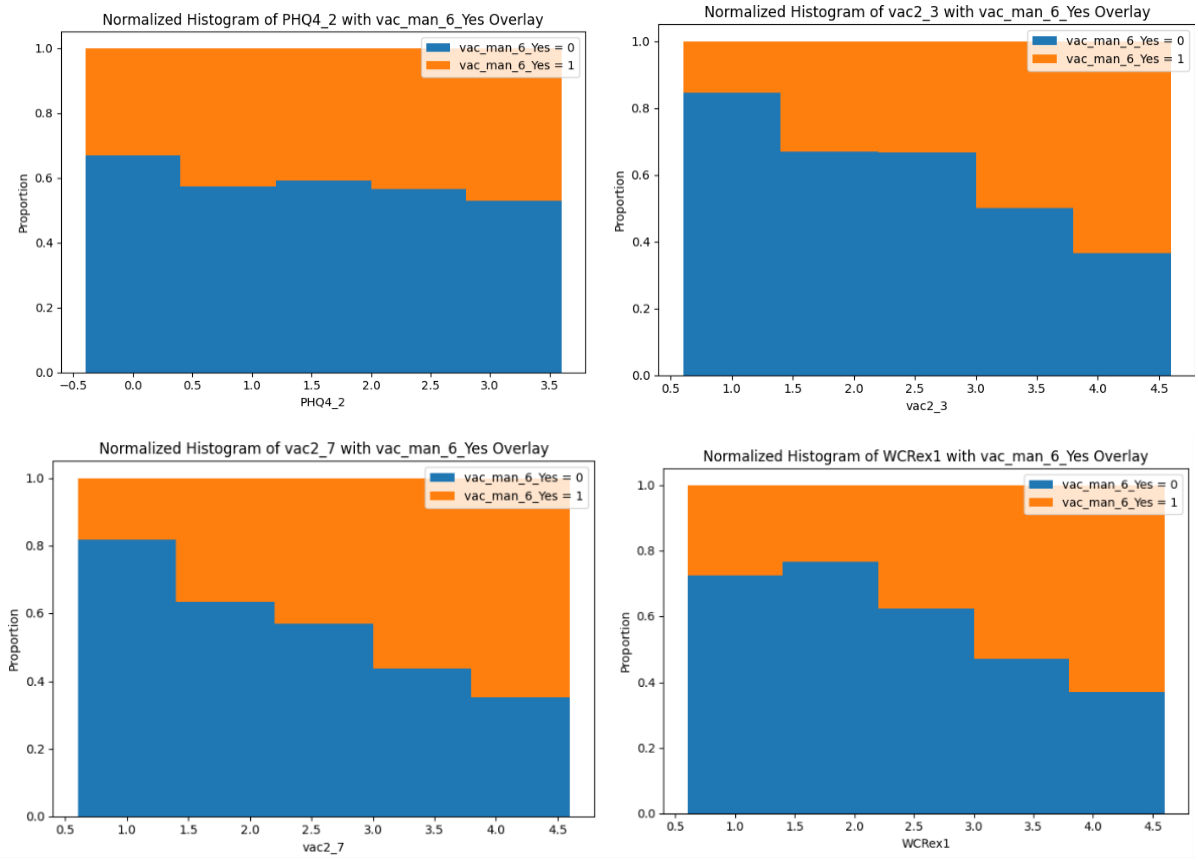












References

- Government of Canada. *Government of Canada's response to COVID-19*. Department of Justice, https://www.justice.gc.ca/eng/csj-sjc/covid.html (Accessed October 22nd 2023).

- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An Introduction to Statistical Learning (2nd ed.). Springer.