Vaccination uptake, happiness and emotions: using a supervised machine learning approach.

Abstract

The COVID-19 pandemic is an example of an immense global failure to curb the spread of a pathogen and save lives. To indirectly protect people against a deadly virus, a population needs to achieve herd immunity, which is attained either through vaccination or previous infection. However, achieving herd immunity by vaccination is preferable as it limits the health risks of disease. As the coronavirus mutated, vaccination estimates for achieving herd immunity went from 70% to 90%. In this study, we investigate the order of the importance of the variables to identify those factors that contribute most to achieving high vaccination rates. Secondly, we consider if subjective measures, including the level of happiness and different collective emotions of populations, contribute to higher vaccine uptake. We employ an XGBoost machine learning algorithm (and, as robustness tests, a Random Forest and a Decision Tree algorithm) to train our data. Our target output variable is the number of people vaccinated as a percentage of the population. We consider two thresholds of our output variable, the first at 70% of a country's population, corresponding to the initial suggestions to achieve herd immunity, and the second with a threshold of 90%, suggested later due to the highly infectious virus. We use a dataset that includes ten countries in the Northern and Southern Hemisphere and variables related to COVID-19, vaccines, country characteristics and the level of happiness and collective emotions within countries. We find that the most important variables listed in reaching the 70% and 90% thresholds are similar. These include the implemented vaccination policy, international travel controls, the percentage of the population in rural areas, the average temperature, and the happiness levels within countries. It is remarkable how the importance of subjective measures of people's emotions and moods play a role in attaining higher vaccination levels. As the vaccine threshold increases, the importance of subjective well-being variables rises. Therefore, not only the implemented policies and country characteristics but also the happiness levels and emotions play a role in compliance and achieving higher vaccination thresholds. Our results provide actionable policy insights to increase vaccination rates. Additionally, we highlight the importance of subjective measures such as happiness and collective emotions to increase vaccination rates and assist governments to be better prepared for the next global pandemic.

Keywords: COVID-19; vaccine; happiness; emotions, supervised machine learning

1. Introduction

The COVID-19 pandemic is an example of an immense global and national failure to curb the spread of the virus and save lives. The death toll due to COVID-19 proves the failure's magnitude. On 20 July 2023, the World Health Organisation (2023) reported that there had been a total of 768,237,788 confirmed cases of COVID-19, including 6,951,677 deaths. Europe has been the hardest hit region, with 2,245,217 deaths, and Africa has the least recorded deaths, with 175,408 deaths (although doubt is cast on the African numbers). The high death toll (lagging only behind the Spanish flu and HIV/AIDS) and the economic damage to countries, industries and individuals are unmeasurable (Baldwin, 2020; Ludvigson et al., 2020; Lu et al., 2020; Fetzer et al., 2020).

Furthermore, COVID-19 not only affected health but also had a profound impact on family functioning and well-being. For example, New Zealand found a significant increase in family violence reports to police, which ranged from 345 to 645 a day, compared to 271 to 478 a day in the same period in 2019 (Mental Health and Wellbeing Commission, 2023). Andrade et al. (2022) note that the fear and uncertainty of health risks, the stress from restrictions and constraints on everyday life, and financial concerns impacted emotional well-being.

During a pandemic, the aim is to stop the spread of the disease and protect individuals against a specific pathogen. We know that globalisation, the geography of economic relations and international travelling pose significant challenges in stopping the spread of a virus. A population must achieve herd immunity to protect people from the disease indirectly. Herd immunity is achieved when a population is immune through vaccination or immunity developed through previous infection. However, the World Health Organisation (WHO) supports achieving herd immunity through vaccination rather than exposing them to the pathogen. To safely achieve herd immunity against COVID-19, it was estimated at the early stages of the pandemic that a vaccination threshold of 70% should be achieved (Randolph & Barreiro, 2020; Bartsch et al., 2020; Goldblatt et al., 2022). However, as COVID-19 evolved, the virus mutated and became more infectious, and the estimated vaccination threshold increased to 90% (Plans-Rubió, 2022). According to Bloom et al. (2021), high vaccination uptake yields sizable and diverse health, economic, and social benefits, including herd protection, increased work hours and productivity, and potentially improved social equity. In other words, the faster the uptake, the fewer lives are lost, and the potentially devastating economic and social impact is minimised.

As of 10 July 2023, a total of 13,474,265,907 vaccine doses have been administered. This translates into 64.8% of the world population being fully vaccinated[1]. However, when we disaggregate the data, we see the stark inequality between high-income countries, 74.32%, and low-income countries, 27.54% (Mathieu et al. (2021). These low vaccination rates in developing and underdeveloped countries, despite global partnerships like COVAX, highlight the lack of international support and cooperation. As Sheikh et al. (2021) noted, most developing nations lack the financial and technological resources to invest in vaccine development. Therefore, relying on developed nations through global cooperation was instrumental in vaccinating their people. Unfortunately, in a shameful show of 'individuality', developed nations, constituting only 16% of the world population, bought more than half of the vaccines available at the start of 2021. This lack of international support and cooperation is seen as one of the biggest failures of the COVID-19 pandemic.

Greyling and Rossouw (2022) also argue that this immense failure is partly due to the inability at a global and national level to distribute and administer vaccines efficiently. Furthermore, at the national level, governments and the public health care systems did not only fail to stop the spread of the virus and protect human lives but also failed to adhere to basic norms of institutional rationality and transparency, breeding mistrust in governments (Paul et al., 2021; Sallam, 2021).

Considering the abovementioned, our primary aim is to retrospectively evaluate the COVID-19 pandemic and determine the most important factors to reach vaccination thresholds. ~~increase vaccine uptake.~~ Therefore, we will determine the most important factors for achieving herd immunity at the 70% vaccination threshold, estimated at the beginning of the COVID-19 pandemic and the 90% vaccination threshold, as estimated later in the pandemic. A secondary aim lies in determining those factors that differ between the 70% to 90% vaccination threshold to see which factors are responsible for advancing a population's decision to reach the higher vaccination level. Special consideration will be given to whether subjective well-being measures played a role in the decision to be vaccinated since we know that negative emotions, such as fear of the side effects of vaccines, influence peoples' attitudes towards receiving the vaccine (Greyling & Rossouw, 2022) and that happier people make better health-related decisions (Anik et al., 2009; Lyubomirsky et al., 2005).

To achieve the aforementioned, we use data from four datasets. The first dataset is extracted from Google COVID-19 Open Data[2]. It provides us with abundant information related to COVID-19 and information on population, geographical location, the economy, general health and climate. The other three time series datasets are derived from tweets and form part of the *Gross National Happiness.today*

[1] Total number of people who received all doses prescribed by the initial vaccination protocol, divided by the total population of the country.
[2] Available from https://health.google.com/covid-19/open-data/explorer

project[3]. These three unique datasets reflect i) the general sentiment and emotions within countries, ii) the sentiment and emotions towards vaccines and iii) the sentiment and emotions towards government institutions.

We use an Extreme Gradient Boosting (XGBoost) algorithm to build a model to determine the most important factors that can predict reaching vaccination thresholds. We chose the XGBoost model since it is more efficient, computationally much lighter and has been shown to outperform most supervised algorithms (Abdurrahim et al., 2020; Nielsen, 2016). However, we construct two other models using Random Forest and Decision Tree algorithms as robustness tests. After the model is built, we test the precision of our model's predictions using our test data and calculate the necessary test (fit) statistics, i.e., mean squared error (MSE), mean absolute error (MAE) and root mean square error (RMSE). In line with expectations, the XGBoost model gives the best-fit measures and delivers the best predictions ~~results~~ compared to the other two methods. Consequently, we discuss the results of the XGBoost model. Although we also present the results of the other models in Appendix C.

Our results on the importance of the factors that increase vaccine uptake at a 70% threshold and 90% threshold overlap with the following factors, vaccination policy implemented, international travel controls, the percentage of the population in rural areas and the average temperature. Interestingly, we find that the importance of happiness differs between the two thresholds. Happiness is less important in achieving the 70% threshold and can generally be reached with policy measures. However, to increase the threshold to 90%, the importance of happiness cannot be ignored. The results clearly show that if governments want higher levels of compliance and vaccine uptake, subjective well-being measures such as mood and emotions must be prioritised. Addressing how people feel, in general, towards vaccines and governments is vitally important when policymakers want to push beyond the lower 70% vaccine threshold and achieve the "golden standard" of 90% fully vaccinated.

Our study makes several contributions to the existing literature. First, this is the first study conducting a post-COVID-19 cross-country analysis of the most important variables to increase vaccine uptake. Second, we are the first study to include subjective measures of well-being in our estimations, such as happiness levels, people's emotions and their perceptions towards vaccines and governments, to establish whether subjective measures play a role in increasing vaccination uptake. Third, we are the first to apply supervised machine learning models to determine which factors matter most to achieve different vaccination thresholds (please note that our dependent variable is continuous, thus different to models in which a binary (mostly a "yes-no" response is used). ~~use various machine learning algorithms to train our data and determine which algorithm gives us the best fit, i.e., the most reliable predictions.~~

[3] Available from https://gnh.today/

Our XGBoost model can be used as a benchmark for future research related to the most important factors for increasing vaccination uptake. Furthermore, this study offers some actionable insights for policymakers on increasing vaccination rates to curb pandemics' health, economic and political effects.

The rest of the paper is structured as follows. The next section contains a literature review of studies investigating factors influencing COVID-19 vaccination rates. Section 3 describes the data and the selected variables, while section 4 outlines the methodology. The results and discussion follow in sections 5 and 6, while the paper concludes in section 7.

2. Literature review

Since increasing the uptake of the COVID-19 vaccine was fundamentally important to decrease the harm caused to human lives and livelihoods, many studies have focused on predicting factors associated with the uptake. However, there are not many studies that used machine learning to determine those factors that contribute to higher levels of vaccine uptake. Therefore, the literature review mainly discusses studies that relied on survey data and traditional empirical analysis, which also informs our discussion in the results section. Studies that used machine learning in their approach conclude this section.

2.1 Factors associated with vaccination uptake: Evidence from survey data

Regarding individual European country studies, Bajos et al. (2022) and Ward et al. (2020) focused on France and used data from the EpiCov survey, and self-collected data, respectively. Gomes et al. (2022) conducted a study in Portugal using a community-based survey called the COVID-19 Barometer: Social Opinion. These three studies generally concluded that the COVID-19 vaccine uptake was positively associated with age, educational attainment and income. According to Bajos et al. (2022), the least educated, those with the lowest incomes, and racial minority groups were less likely to accept the vaccine, and these differences were maintained or increased over time. Additionally, people's lack of trust in the government and scientists to manage the health crisis remained the primary reason for refusing to vaccinate. Ward et al.'s (2020) pre-vaccine study also found that individuals feeling close to a Far-Right party would refuse the vaccine when it became available. The primary reason any individual would refuse the vaccine was that it would not be safe. Gomes et al. (2022) also concluded that higher odds of hesitancy were associated with low confidence in Portugal's health services response to COVID-19 and non-COVID-19 and perceived the measures implemented by the government as inadequate.

Cross-country analysis was conducted by Bergmann et al. (2022) and Pronkina and Rees (2022), who used the 2021 summer SHARE Corona survey data (administered across 27 European countries). They confirmed the results of Bajos et al. (2022), Ward et al. (2020) and Gomes et al. (2022) by finding that the probability of being vaccinated increased with age, income, and educational attainment. Furthermore, Bergmann et al. (2022) concluded that prior illnesses were associated with a higher willingness to vaccinate. Interestingly, there was no clear and significant effect of subjective health and no strong effects with mental health issues were found. Pronkina and Rees (2022) argued that people who express trust in others are more likely to be vaccinated, while risk aversion and frequency of praying (a proxy for religiosity) were negatively correlated with the probability of being vaccinated against COVID-19. Furthermore, Europeans aged 50 and older did not base their decision to vaccinate against COVID-19 on case counts or excess mortality during the pandemic.

Corcoran et al. (2021), Czeisler et al. (2021), El-Mohandes et al. (2021), and Gatwood et al. (2021) found that Americans who express conservative political or religious beliefs are, on average, more vaccine-hesitant than those who do not although the relationship between political beliefs and COVID-19 vaccination hesitancy appears to be considerably more nuanced in Europe than it is in the United States (Ward et al., 2020; Lindholt et al., 2021; Raciborski et al., 2021; Bíró-Nagy & Szászi, 2022; Wollebæk et al., 2022). COVID-19 vaccine hesitancy is especially prevalent among individuals who express distrust in government and scientists (Kerr et al., 2021; Latkin et al., 2021; Lindholt et al., 2021; Rozek et al., 2021; Bajos et al., 2022).

2.2 Factors associated with vaccination uptake: Evidence from machine learning

In terms of previous machine learning studies, Lincoln et al. (2022) used Random Forest to probe for the optimum prediction accuracy for vaccine hesitancy and to find an economical model based on a selection of common global predictors. They used SHapley Additive exPlanations (SHAP) and permutation feature importance to estimate the importance of each variable in their model across their sample of five advanced countries (U.K., USA, Australia, Germany and Hong Kong). The authors found that by using only twelve variables (the combined most important variables from permutation feature importance and SHAP), they could achieve an 82% accuracy in predicting vaccine hesitancy, with the most crucial factors being vaccination conspiracy beliefs and a lack of confidence in governments, companies, and organisations in handling the pandemic (i.e., pandemic conspiracy beliefs).

Previous studies have successfully used XGBoost-based predictive models to predict influenza vaccine uptake. Shaham et al. (2020) used primary data for 250,000 Israelis collected between 2007 and 2017 to predict whether a patient would get vaccinated in the future. Their XGBoost-based predictive model

Commented [GT1]: Thus the outcome variable is at individual level - a yes no answer?

achieved a ROC-AUC[4] score of 0.91 with accuracy and recall rates of 90% on the test set. Prediction relied mainly on the patient's individual and household vaccination status in the past, age, number of encounters with the healthcare system, number of prescribed medications, and indicators of chronic illnesses. Using the XGBoost regressor, Cheong et al. (2021) used sociodemographic data to predict vaccine uptake across counties in the United States (U.S.). Their model predicted COVID-19 vaccination uptake across U.S. counties with 62% accuracy. The results from their permutation analysis and SHAP revealed the most important factors ~~top significant features found~~ to drive their predictive model were geographic location (longitude, latitude), education level (per cent of adults with less than a high school diploma, per cent of adults with a bachelor's or higher), and online access (households with broadband internet).

Also focusing on the US, Osman and Sabit (2022) use state-level vaccination rates to identify the most critical features that can predict which states will meet the vaccination threshold of 70%. Relying on Chi-square Automatic Interaction Detector (CHAID), a decision tree algorithm, the authors include several variables that may influence the state-specific vaccination rate. They categorise the variables into four groups: economic indicators, COVID-19-related indicators, Google mobility data, and COVID-19-related policy measures. After using three different model specifications, they discovered that workplace travel, the political affiliation of the governor, and the vaccine mandate in schools were the top three features of achieving the vaccination threshold.

In the above-mentioned studies related to machine learning applications, the outcome variables were binary variables, for example – the decision of a person to be vaccinated or not, or will a certain vaccination threshold be reached or not. In these studies, the most important factors to reach success (yes), during COVID-19 were determined. Our study differs from the previous literature in that we have the benefit of hindsight, thus we investigate the most important factors that contribute to reaching herd immunity (at different levels of 70 per cent or 90 per cent) and how the factors change when higher levels of herd immunity are to be reached. Our outcome variable is the percentage of the population that was vaccinated as a percentage of the population of a country (thus the measure that is used to determine herd immunity), it is a continuous variable, which represents a high level of variance and is not constricted to only a yes-no answer. Furthermore, our study includes a wide-reaching dataset including variables related to COVID-19 regulations, vaccination policies, country characteristics and very importantly subjective measures of well-being (not included by other studies) to highlight the importance of moods and emotions when higher vaccination thresholds must be attained.

Commented [GT2]: Once again a yes no answer - thus who will take up vaccines (at a certain level) and who will not - binary outcome variable

Commented [GT3]: Once again the question - yes/no will they make the 70 per cent

Formatted: Highlight

Formatted: English (New Zealand)

[4] Area Under the Curve of the Receiver Operating Characteristic curve.

3. Data and variables

3.1. Construction of datasets

The period under consideration is from 1 December 2020 to 16 September 2022. This period includes the first vaccine rollout and ends when new COVID-19 tests reach almost zero in all countries. Consequently, the main data source related to COVID-19, the *COVID-19 Government Response Tracker* dataset (Hale et al., 2021), was discontinued on 31 December 2022. We consider the data to find a retrospective view of those factors that mattered most for higher vaccination rates.

We use a merged dataset, including the Google COVID-19 Open Data[5] and our three constructed time-series datasets derived from tweets[6]. The three Twitter datasets reflect i) happiness levels and emotions of countries, ii) happiness levels and emotions towards vaccines and iii) happiness levels and emotions towards government institutions. The construction and validation of the Twitter datasets are explained in Appendix A.

This section briefly explains the Twitter data (see Appendix A for a full explanation). Tweets are extracted in real time based on a geographic bounding box corresponding to the country in question. Next, we use sentiment and emotion analysis to score the tweets. We aggregate the scores and derive indices for happiness and each of the eight emotions. For the Twitter datasets related to the government and COVID-19 vaccines, we used specific keywords to identify those tweets directly related to the topic.

To derive the dataset related to the COVID-19 vaccines, we extracted tweets using the keywords: *vaccinate, vacc, vaccine, Sputnik V, Sputnik, Sinopharm, Astrazeneca, Pfizer (if NEAR) vaccine, Pfizer-BioNTech, Johnson & Johnson,* and *Moderna*.

For the dataset related to governments, we extracted tweets using the keywords: *government, parliament, ministry, minister, senator, M.P.s, legislator, political, politics, prime minister.*

After extraction, we analysed the text of the tweets to determine the noise captured in the tweets. Subsequently, we found that the noise was minimal in both instances.

The Google COVID-19 Open dataset is rich and includes variables related to COVID-19 cases, deaths, vaccinations, demographic, economic, geographical, climate, health, health infrastructure and health care.

3.2 Data cleaning and validation

[5] Available from https://health.google.com/covid-19/open-data/explorer
[6] Available from https://gnh.today/

After merging the datasets from section 3.1, we had an initial merged dataset containing 145 variables. As a first instance, we set about to identify missing data. If the data was randomly missing with less than 3% overall missingness, we imputed the data, by either using the mean or the previous data point, for example, population size. Secondly, we dropped variables from our dataset with high missingness levels. For example, international support (67% missingness), emergency investment in health care (68% missingness) and mobility regulations (74% missingness), which reflects the strong regulations implemented during the first lockdowns in countries, such as access to retail and recreation, grocery stores, pharmacies and parks, were dropped. Thirdly, we removed highly correlated data so that only one of the variables remained in the dataset, for example, cumulative confirmed cases and cumulative tested cases; this eases the interpretation of the results.

Once the data was cleaned, we were left with 69 variables (including our outcome variable), which we classified into five categories (see section 3.4). Subsequently, these variables were used in the supervised algorithms (see sections 4.1-4.3) to train the models. We have 6530 observations which means we have 653 (just short of two years) observations per country in our sample.

In our study, the data comprising 69 variables are split into a training and testing dataset with an 80:20 split on all data, with the evaluation done on the unseen testing data.

~~6.3~~3.3 *Target/outcome variable*

Our main variable of interest is the country-level vaccination rate. We calculate vaccination rates as the percentage of the vaccinated population among those 18 and older as a percentage og the total population in the respective countries. This is in line with studies such as Randolph and Barreiro (2020), Bartsch et al. (2020) and Goldblatt et al. (2022).

In our sample, nine out of the ten countries met the lower threshold of 70% (see Table 1); South Africa lagged behind, reaching a mere 32.6%. Therefore, our 70% threshold model was reachable for the countries in the developed world but not for our developing country, South Africa (likely to be the same in other developing and underdeveloped countries). However, none of the countries in our sample achieved the higher 90% threshold, with Spain coming closest with 87%.

Table 1. Maximum vaccination rates on 16 September 2022

Country	Percentage of the population vaccinated on 16 Sept 2022
Australia	85.35
Belgium	76.19

Germany	76,43
Spain	86.58
France	80.07
Great Britain	76.15
Italy	79.46
The Netherlands	69.19
New Zealand	85.67
South Africa	32.64

Source: Authors' own calculations

3.4 Predictor variables/features

As mentioned in section 3.2, our models include 68 variables (apart from our outcome variable) to determine those factors most important for country-specific vaccination thresholds. We remind the reader that two variables, international support and emergency investment in health care, were not included as predictors in our models due to their high levels of missingness, 67% and 7468%, respectively. We acknowledge that these variables could have ranked among the most important variables and potentially have been included in the top ten. Therefore, when we report the results of our models, their absence should be kept in mind.

We categorise the variables into five groups: demographic, geographical, economic, COVID-19-related indicators and COVID-19-related policy measures. The COVID-19-related and policy data are high-frequency daily data, while the demographic, geographical and economic data are more stable over time. Table 2 gives an abbreviated list of the variables included in the models. For a full list, see Appendix B.

Table 2. An abbreviated list of variables

Variable	Description	Scale	Coding	Source
Vaccination policy	Policies for vaccine delivery for different groups	Ordinal scale	0 - No availability 1 - Availability for ONE of following: key workers/ clinically vulnerable groups (non-elderly) / elderly groups 2 - Availability for TWO of following: key workers/ clinically vulnerable groups (non-elderly) / elderly groups 3 - Availability for ALL of following: key workers/ clinically vulnerable groups (non-elderly) / elderly groups 4 - Availability for all three plus partial additional availability (select broad groups/ages) 5 - Universal availability	Hale et al. (2021)
Average temperature	Average temperature in the country	Celsius		World Bank (2023a)
Population density	People per square kilometre of land area			World Bank (2023b)

Workplace closing	Record closing of workplaces	Ordinal	0 - no measures 1 - recommend closing (or recommend work from home) or all businesses open with alterations resulting in significant differences compared to non-Covid-19 operations 2 - require closing (or work from home) for some sectors or categories of workers 3 - require closing (or work from home) for all-but-essential workplaces (e.g. grocery stores, doctors) Blank - no data	Hale et al. (2021)
Restrictions on gatherings	Record limits on gatherings	Ordinal	0 - no restrictions 1 - restrictions on very large gatherings (the limit is above 1000 people) 2 - restrictions on gatherings between 101-1000 people 3 - restrictions on gatherings between 11-100 people 4 - restrictions on gatherings of 10 people or less Blank - no data	Hale et al. (2021)
GNH	Happiness	Ordinal	Score per hour ranges from 0 to 10, with higher values indicating higher happiness. To generate daily data, the mean GNH per day is calculated.	Greyling et al. (2019)

Source: Multiple as specified within Table 2

4 Methodology

The methodology first explains the different machine-learning algorithms utilised and how we applied each algorithm to construct the models (training the models). We start with the XGBoost (our algorithm of choice) and include a Random Forest and Decision Tree algorithm as robustness measures. Next, ~~we explain how we trained the models, and lastly,~~ we describe the fit statistics used to evaluate the models.

4.1 Extreme Gradient Boosting (XGBoost)

To determine those factors most important in achieving our vaccination thresholds of 70% and 90%, we rely on the XGBoost method. It should be noted that traditionally, XGBoost models were used where only a binary outcome was considered (limiting the prediction to either an up or down or a yes and no option). However, we adjust the algorithm to consider a continuous dependent variable since our model's predictions are not limited to a binary outcome but to various rates.

XGBoost is an end-to-end tree-boosting system (Chen & Guestrin, 2016) and is a powerful supervised learning approach to classification and regression tree models based on ensemble methods. The scalability of XGBoost allows for a system that runs ten times faster than existing popular solutions on a single machine. As it is a gradient-boosting algorithm, XGBoost combines predictions from decision trees. XGBoost creates a better overall model while boosting it, continuously rebuilding it by focusing on the previous models' weaker points.

More specifically, the XGBoost algorithm works by assigning weights to all the independent variables, which are then fed into the decision tree as a simple method to recursively split the data into smaller groups to predict the target variable. A single decision tree would not work well on complex problems,

so through the boosting phase, the weights of variables predicted wrong by the first decision tree are increased, and these variables are then fed to the second decision tree. Therefore, an ensemble method combines multiple trees to build a single model sequentially, focusing on the decision trees that did not perform as well and combining these to create a stronger and more precise model. See Fig 1 for a brief illustration of how gradient tree boosting works.

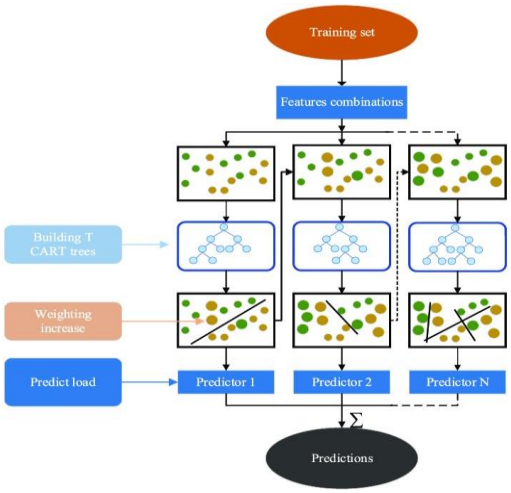


Fig 1. Illustration of how gradient tree boosting works

In all models, XGBoost minimises a regularised objective function which consists of two parts (see equation 1). The first part is a convex loss function (based on the predicted and target outputs) and measures how predictive our model is with respect to the training data. The training proceeds iteratively, adding new trees that predict the residuals or errors of prior trees that are then combined with previous trees to make the final prediction. In our model, gradient descent[7] optimisation is used to minimise the loss. The second part is a term inserted to reduce the risk of overfitting, which incorporate two steps: regularisation and pruning. The regularisation term for model complexity (in other words, the regression tree functions) penalises the model by adding extra terms (k times iteration) to the objective function, thus discouraging the model from becoming too complex. At the same time, pruning removes the nodes in decision trees that do not contribute significantly to the model's performance.

Therefore, following Liang et al. (2020), the objective function that is optimised in our model is specified in equation (1) as:

[7] Gradient descent is an iterative first-order optimisation algorithm used to find a local minimum/maximum of a given function. It trains machine learning models by minimising the loss incurred between predicted and actual results.

$$O(\theta) = \sum_{i}^{n} l(y_i, \hat{y}_i) + \sum_{1}^{k} \lambda(f_k) + c \tag{1}$$

Where $\sum_i^n l(y_i, \hat{y}_i)$ is the loss function, $\sum_1^k \lambda(f_k)$ is the regularisation term, and c is the constant. In turn, the regularisation term can be further explained in equation (2):

$$\lambda(f_k) = \delta H + \frac{1}{2}\psi \sum_{1}^{T} w_j^2 \tag{2}$$

Where T stands for the number of leaves, δ represents the complexity, and ψ is the penalty parameter (Liang et al., 2020).

Other than performance alone, XGBoost is computationally much lighter than, for example, the Random Forest method and has demonstrated greater accuracy over other methods. For example, Abdurrahim et al. (2020), comparing the accuracy of different predictive modelling algorithms, shows that XGBoost shows the highest accuracy score compared to other methods such as logistic regression, naive Bayes classifier, decision trees, and random forest. Although, in our study, we use Random Forest and Decision Trees to see if it gives us the same collection of relevant variables since the main focus of this study is to determine the major drivers to attain a vaccination threshold (see sections 4.3 and 4.4). Furthermore, Nielsen (2016) demonstrated that XGBoost learns better tree structures over decision tree models that use gradient boosting since XGBoost uses Newton boosting instead.

Multiple combinations were tested for the XGBoost model in our paper, and a tree depth of seven was selected since it delivered optimal results. Therefore, our XGBoost model is defined in equation (3) as:

$$F_M(x) = F_0 + v\beta_1 T_1(x) + v\beta_2 T_2(x) + \cdots + v\beta_M T_M(x) \tag{3}$$

Where M is the number of iterations. The gradient boosting model is a weighted $(B_1 \ldots \beta_M)$ linear combination of simple models $(T_1 \ldots T_M)$. $F_M(x)$ is the vaccination threshold as described in section 3.3.

Constructing~~Building~~ the XGBoost model

To build the model, we first started by using all the default settings of the XGBoost algorithm on the training data and refined the parameters afterwards. We started by refining the depth of the trees and tested depths between three and ten using the value with the lowest root mean square error (RMSE) (see section 4.4.3). The final tree depth was selected as seven, resulting in the lowest RMSE. The number of iterations is set to 100, with a termination clause added to stop the algorithm if the RMSE does not decrease after 5 iterations. After completing the refining stage, the model reaches the lowest RMSE at 16 iterations.

4.2 Random Forest

As mentioned in section 4.1, we want to see if the XGBoost model's results related to the most important variables are resilient. Therefore, we use an alternative tree-based machine learning approach to see if it gives us the same collection of relevant variables since the main focus is determining the major factors associated with reaching a vaccination threshold (70% and 90%, respectively). We are more concerned with whether we acquire the same set of variables than with the order in which these variables are important.

For this purpose, we employ the Random Forest and Decision Tree (see section 4.3) models. The Random Forest algorithm (Breiman, 2001) is an ensemble method using bootstrap aggregation to produce multiple independent models to be combined to finalise the predictions.

The Random Forest process uses subsets of the training data to build decision trees. The training data subsets are generated by sampling with replacement bootstrap samples from the training data. The decision tree created using the bootstrap sample can then only evaluate the parameters that are a part of the subset; this reduces the risk of overfitting. After training multiple decision trees, the predictions' averages are taken to get the final prediction output.

A critical characteristic of Random Forests is that they produce measures of variable importance that may be used to find the most important predictor variables (Hapfelmeier et al., 2014; Breiman, 2001). It also works well with small sample sizes and highly correlated sample features (Strobl et al., 2008). Random Forest ranks the variables in terms of a 'mean decrease in accuracy' (MDA). The MDA score indicates the accuracy lost when each variable is removed from the model. The variables are listed in order of decreasing relevance.

For the Random Forest model in this paper, we follow the specifications set out in the XGBoost model and set the depth of the decision trees as seven, with a total of 200 trees. We also rely on MSE, MAE and RMSE as evaluation metrics.

Constructing~~Building~~ the Random Forest Model

Similar to the XGBoost model, we first started by using the default Random Forest algorithm on the training data and refined the parameters afterwards. We set the depth of the trees to seven to be consistent with the XGBoost model. Initially, we used 20 trees, but after some investigation, the mean squared error (MSE) (see section 4.4.1) only started converging with 50 trees. We finally steeled on a total of 200 trees as it gives good results while not being overly computationally expensive.

4.3 Decision trees

Our last robustness measure uses a Decision Tree algorithm, a non-parametric supervised learning algorithm for classification and regression tasks. It has a hierarchical tree structure consisting of a root node, branches, internal nodes and leaf nodes. Please see Fig 2 for a brief illustration of how Decision Trees work.

Decision trees comprise many nodes that form a tree when put together; each of these nodes represents decisions made that split the data. The decision of which attribute to use for the split at each node is made by optimising some criterium; in our case, the mean square error is minimised.

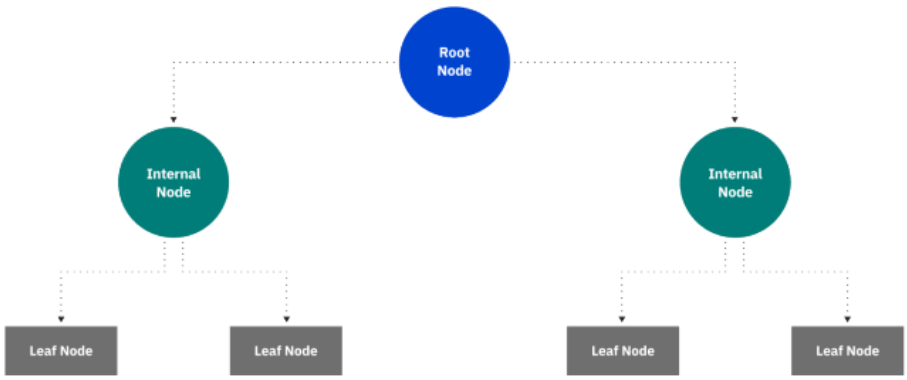


Fig 2. Illustration of how Decision Trees work

As shown in Fig 2, a decision tree starts with a root node with no incoming branches. The outgoing branches from the root node feed into the internal nodes, also known as decision nodes. Based on the available features, both node types conduct evaluations to form homogenous subsets denoted by leaf nodes or terminal nodes. The leaf nodes represent all the possible outcomes within the dataset.

Decision Tree learning employs a divide-and-conquer strategy by conducting a greedy search to identify the optimal split points within a tree. This process of splitting is then repeated in a top-down, recursive manner until all or the majority of records have been classified under specific class labels. Whether or not all data points are classified as homogenous sets largely depends on the decision tree's complexity. Smaller trees can more easily attain pure leaf nodes—i.e., data points in a single class. However, as a tree grows in size, it becomes increasingly difficult to maintain this purity, and it usually results in too little data falling within a given subtree. When this occurs, it is known as data fragmentation, which can often lead to overfitting.

As a result, decision trees prefer small trees, which is consistent with the principle of parsimony in Occam's Razor; that is, "entities should not be multiplied beyond necessity." Said differently, decision trees should add complexity only if necessary, as the simplest explanation is often the best. To reduce complexity and prevent overfitting, pruning is usually employed; this is a process which removes branches that split on features with low importance. The model's fit can then be evaluated through the process of cross-validation. Another way that decision trees can maintain their accuracy is by forming an ensemble via a random forest algorithm; this classifier predicts more accurate results, particularly when the individual trees are uncorrelated with each other.

Constructing~~Building~~ the decision tree~~s~~ model

We follow the specifications set out in sections 4.1 and 4.2 for the Decision Tree model in this paper. The number of nodes in the decision tree is 8. We also rely on MSE, MAE and RMSE as evaluation metrics.

4.4 Evaluation

Model evaluation is the process to use metrics to analyze the performance of the model, thus how well the model generalizes future predictions. There are many metrics like Accuracy, Precision, Recall, F1 score, Confusion Matrix, and various error calculations such as the Root Mean Square Error.

The first group is applicable to classification models and models in which the predicted outcome variable is discrete (or binary). However, our dependent variable (predicted variable) is a continuous variable (vaccinated population/total population) to evaluate the performance of the models we consider different error calculation measures as they summarise how close the prediction is to the actual value.

~~*4.4*~~

Formatted: Font: (Default) Times New Roman, 11 pt, Font color: Custom Color(RGB(39,50,57)), Expanded by 0,1 pt

Formatted: Font: (Default) Times New Roman, 11 pt, Font color: Custom Color(RGB(39,50,57)), Expanded by 0,1 pt

Formatted: Justified

Formatted: Font: (Default) Times New Roman, 11 pt, Font color: Custom Color(RGB(39,50,57)), Expanded by 0,1 pt

Formatted: Font: (Default) Times New Roman, 11 pt

Formatted: Font: (Default) Times New Roman, 11 pt

Formatted: Font: (Default) Times New Roman, 11 pt

Formatted: Font: (Default) Times New Roman, 11 pt

Formatted: Font: (Default) Times New Roman, 11 pt

Formatted: Font: (Default) Times New Roman, 11 pt

Formatted: Font: (Default) Times New Roman, 11 pt

Formatted: Font: (Default) Times New Roman, 11 pt

Formatted: Font: (Default) Times New Roman, 11 pt

Formatted: Normal, No bullets or numbering

WFor our models, we consider the use Mean Absolute Error (MAE), the mMean Ssquared E-error (MSE), mean absolute error and Rroot Mmean Ssquare Eerror (RMSE), as the usual metrics, such as accuracy and recall, cannot be used because we are predicting a continuous variable. Therefore, we do not have perfect predictions, as is the case when you use a discrete variable.

4.4.1 Mean squared error

The mean squared error (MSE) evaluates the proximity of a regression line to a group of data points. It is a risk function that corresponds to the predicted squared error loss value. MSE is computed by calculating the average of the squared mistakes resulting from a function's data, especially the mean. From equation (4), we see that the MSE is calculated by taking the observed value (y_i), subtracting the expected value ($\hat{y}_i$), and then squaring. Repeat for every observation. Afterwards, divide the total by the total number of occurrences (n) by the sum of the squares of the values.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{4}$$

Therefore, the MSE measures the error in prediction algorithms. This statistic quantifies the average squared variance between observed and predicted values. Squaring the differences removes negative mean squared error differences and guarantees that the squared mean error is always larger than or equal to zero. The value is usually always positive. When there are no errors in a model, the MSE equals 0.

Moreover, squaring magnifies the effect of greater inaccuracies. These computations punish greater mistakes disproportionately more than smaller ones, i.e., a model's worth increases proportionally to its degree of error. This attribute is necessary if we want our model's mistakes to be fewer.

The MSE in regression, for instance, might indicate the average squared residual. The MSE decreases as the data points align with the regression line, indicating less error in the model. A model with fewer errors yields more accurate predictions.

If the MSE is high, the data points are spread out quite a bit from the centre moment, while a low value implies the opposite. When the data points cluster tightly around their mean, the MSE will be modest (mean). It shows that our data values are distributed normally, that there is no skewness, and, most importantly, that there are fewer errors, where errors are defined as how far our data points are from the mean.

4.4.2 Mean absolute error

In the context of machine learning, absolute error refers to the magnitude of difference between the prediction of an observation and the true value of that observation. Mean absolute error (MAE) takes the average of absolute errors for a group of predictions and observations to measure the magnitude of errors for the entire group. MAE can also be referred to as the loss function specified in equation (1).

As one of the most commonly used loss functions for regression problems, MAE helps formulate learning problems into optimisation problems. It also serves as an easy-to-understand quantifiable measurement of errors for regression problems.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{5}$$

4.4.3 Root mean square error

The Root Mean Square Error (RMSE) represents the square root of the average squared differences between predicted and observed outcomes. It is a metric predominantly utilised in regression analysis and forecasting, where accuracy matters significantly. The lower the RMSE, the better the model's ability to predict accurately. Conversely, a higher RMSE signifies a greater discrepancy between the predicted and actual outcomes. RMSE initially computes the difference between each data point's observed and predicted value. This difference, known as the residual, is squared. The squared residuals are then summed up to obtain a cumulative figure divided by the number of data points to give the MSE. Finally, the square root of the MSE is calculated, resulting in the RMSE (see equation 6).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{6}$$

Where y_j is the true value of the dependent variable (vaccination rate), $\hat{y}_j$ is the predicted value of the dependent variable (vaccination rate), and n is the number of observations.

5 Results of the training of the models

In this section, we first discuss the results of the model construction~~building through iterations~~. Second, we discuss the evaluation of the fit of the models. Lastly, we discuss the application of the models to address our research questions.

5.1 Results of models through iterations

We consider Figs 3 and 4, which show the size of the RMSE over iterations for XGBoost and Random Forest, respectively (please note we do not have a similar Figure for the Decision Tree as it does not use an ensemble method). Figure 3 shows how the RMSE decreases over the number of iterations. It reaches a minimum at 16 iterations and remains constant up to 20 iterations.

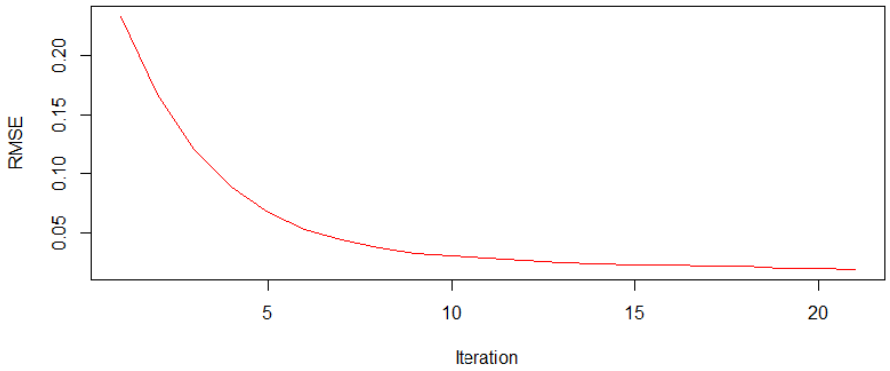


Fig 3. RMSE over iterations for XGBoost

The Random Forest model (section 4.2) took much longer to train compared to the effectiveness of the training of the XGBoost model. After 50 trees, it seemed as though the model converged, but upon further inspection, the results continued to improve with minute increments with each additional iteration. In Fig 4, an illustration of the Random Forest algorithm training is given. The MSE decreases, and we can see that after 50 iterations, the MSE is relatively small. The MSE becomes smaller with each iteration, but it does not converge to a specific value.

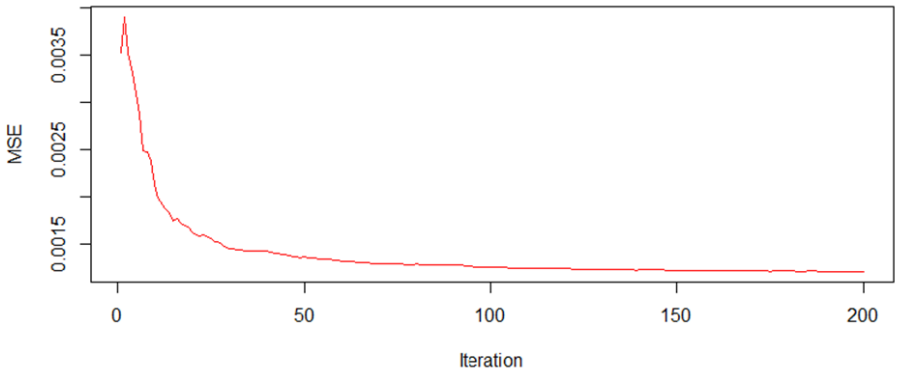


Fig 4. RMSE over iterations for Random Forest

5.2 Evaluation of the fit of the models

In this section, we discuss the evaluation of the three models that we built, XGBoost, Random Forest and Decision Tree. In Table 3, the fit statistics for the three models are given. We discuss the fit measures when constructing the model to reach ~~use the models to predict the output variable using the~~ a 90% threshold since this provides us with the largest possible test dataset (the fit measures ar also available for the 70 per cent level). ~~If we were to use the 70% threshold, we would encounter a reduction in the sample size since it would drop all target outcomes above 70%.~~ Our test data includes all variables (obviously ~~,~~excluding the target output variable). We notice that all measures of fit reveal very small errors, indicating a good-fitting model. Across all three of the fit statistics, the XGBoost performs the best with the lowest values. For the XGBoost, the MSE is 0.0014, the MAE is 0.0227, and the RMSE is 0.0375.

Table 3. Evaluation metrics across models

Model	MSE	MAE	RMSE
XGBoost	0.001412552	0.022707714	0.0375839
Random Forest	0.001861686	0.029981258	0.043147264
Decision tree	0.01222601	0.07180425	0.11057130

Source: Author's own calculations

Though the fit statistics indicate that the XGBoost model performed best when considering all models, a visual representation is also provided of all three models' predicted values in blue, with the true values

of the dependent variable displayed in red. In Fig 5, the XGBoost results are displayed. The predicted values (the blue line) are quite close to the true values (in red), reflecting a good fit.

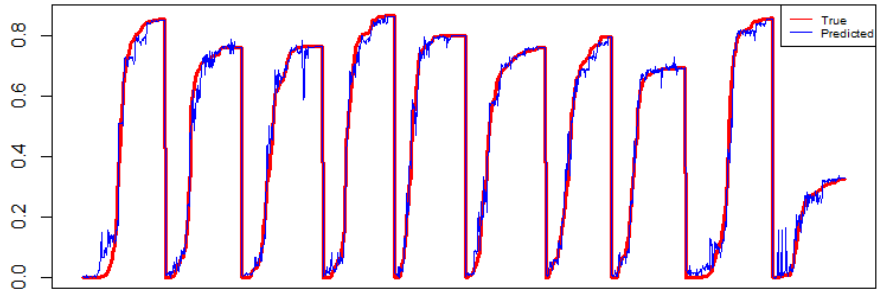


Fig 5. XGBoost - predicted values of the output variable against the true values

Considering Fig 6, we notice that using the Random Forest model, the predicted values (blue line) compared to the true values (red line) are not as good as in the case of the XGBoost model. This is also shown in the fit statistics (Table 3) with an MSE of 0.0018, an MAE of 0.0299 and an RMSE of 0.0431; thus, each fit statistic reveals bigger errors than in the case of the XGBoost model.

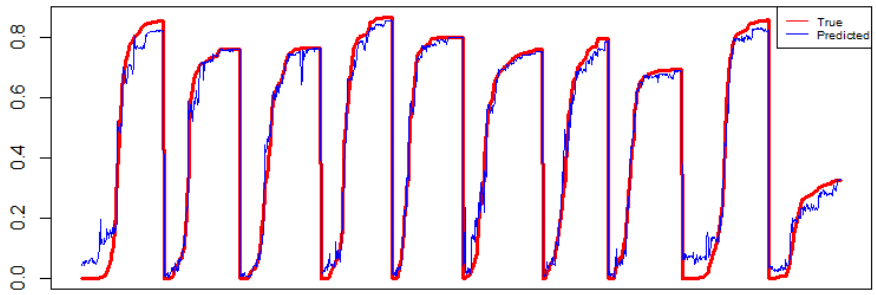


Fig 6 Random Forest - predicted values of the output variable against the true values

Looking at Fig 7, we notice that using the predicted values (blue lines) compared to the true values in the Decision Tree model, we could not reach similar levels of prediction as we did with either the XGBoost or the Random Forrest models. The fit statistics also reveal larger errors compared to the other models. The MSE is 0.0122, the MAE 0.0718 and the RMSE 0.1105 (Table 3).

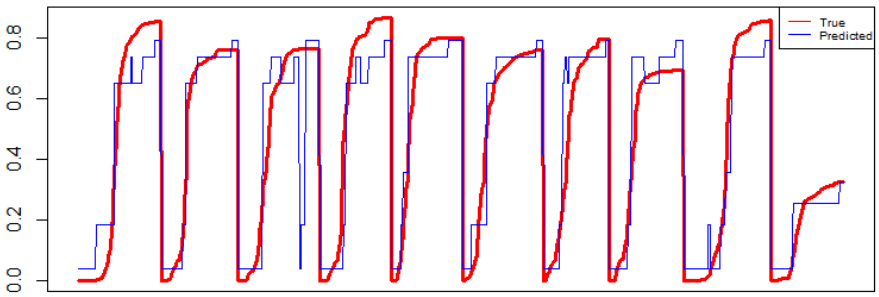


Fig 7 Decision Tree - predicted output variable values against the true values.

Fig 8 shows the predicted values of the outcome variable of all three models against the true value of the dependent variable. In Fig 8, the true value of the dependent variable is represented in red, while the predictions for the three models are represented in blue, XGBoost, green, Random Forest, and magenta, the Decision Tree. Fig 8 supports the results in Table 3 and Figs 5-7, as the XTBoost predictions (~~-~~blue line) is consistently closer to the true value (red line) compared to the values using the other two models ~~than the other two~~. This aligns with our expectations that the XGBoost model outperforms the other models. As mentioned previously the XGBoost model~~It shows both better~~ performs better ~~ance~~ and uses less computational power. Therefore, in discussing the~~the~~ application of the model to answer our research questions in sections 5.3 and 6~~, which determines the ranking of the importance of the variables~~, we interpret the XGBoost results[8].

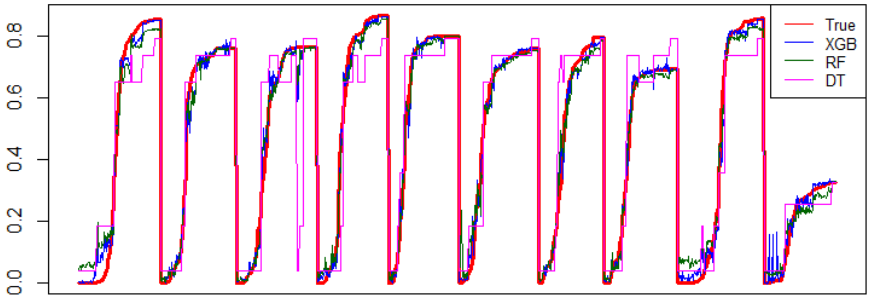


Fig 8. True value with all model predictions

[8] The reader should note that although the XGBoost outperforms the other models and is computationally less expensive, the Random Forest and Decision Tree Models have the benefit that they are easier to understand and visualise.

~~5.6~~5.3 *Results of the XGBoost model on variable importance*

Table 4 shows the results from our XGBoost model on ranking the importance of variables to reach a 70% and 90% vaccination threshold rate (see Appendix C for the Random Forest and Decision Tree). We also consider a third model at the 80% vaccination threshold level for comparative analysis.

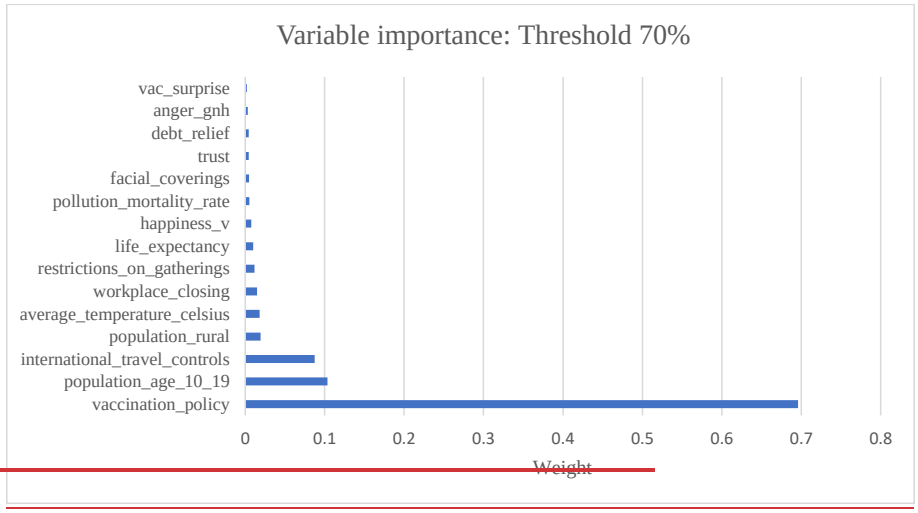
Considering the results from reaching the 70, 80 and 90% thresholds, we notice recurring factors among the five most important factors. The factors are related to the vaccination policies, the COVID-19 policies to limit the spread of the virus, and country characteristics such as the percentage of the population residing in rural areas and the average temperature in the countries. This implies that regardless of the vaccination threshold goal, governments should focus on their vaccination policy, international travel controls, the percentage of the population in rural areas and the average temperature to achieve their maximum vaccination rates.

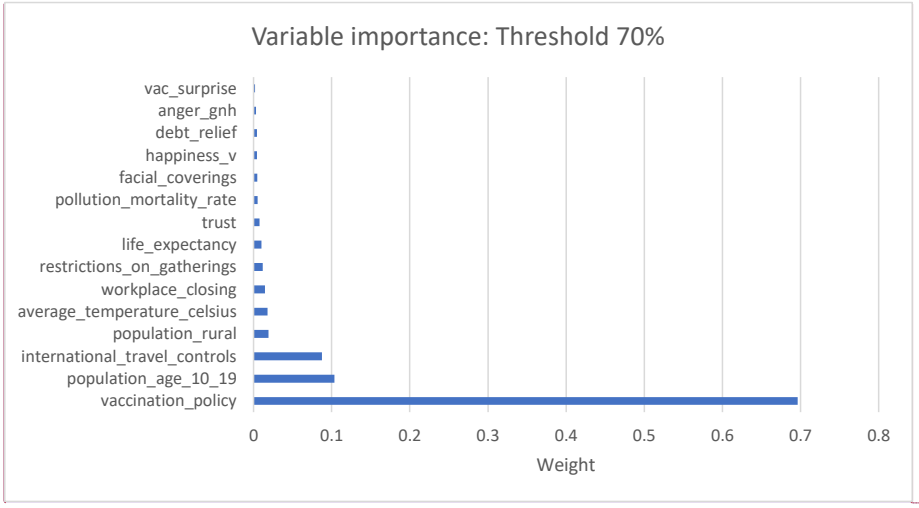
Table 4. Results on the order of the importance of the variables predicting vaccination thresholds of 70, 80 and 90%, respectively.

70% threshold	**80% threshold**	**90% threshold**
Vaccination policy	Vaccination policy	Vaccination policy
Population aged between 10-19	International travel controls	International travel controls
International travel controls	Percentage of population in rural areas	Percentage of population in rural areas
Percentage of population in rural areas	Restrictions on gatherings	**Happiness**
Average temperature	Average temperature	Average temperature
Workplace closing	Human Development Index	Population density
Restrictions on gatherings	**Happiness**	Human Development Index
Life expectancy	Workplace closing	Facial coverings
Happiness	Population aged between 10-19	Workplace closing
Pollution mortality rate	Out-of-pocket health expenditure	Restrictions on gatherings

Source: Authors' own calculations.

What is interesting to note is the important role subjective measures of well-being play in achieving vaccination goals. To gain a 70% vaccination (all countries met this threshold except S.A.), happiness was among the top ten important factors at number nine (Fig 9). If we increase our vaccination threshold to 80% (5 out of 10 countries met), happiness increases in importance and moves to the seventh place. However, to reach the vaccination threshold of 90% or more, we notice that people's happiness is again becoming increasingly important and reaches fourth place (Fig 10). Therefore, regardless of the threshold level~~model~~, happiness plays an important role, and the higher the vaccination threshold governments want to achieve, the more important it becomes.



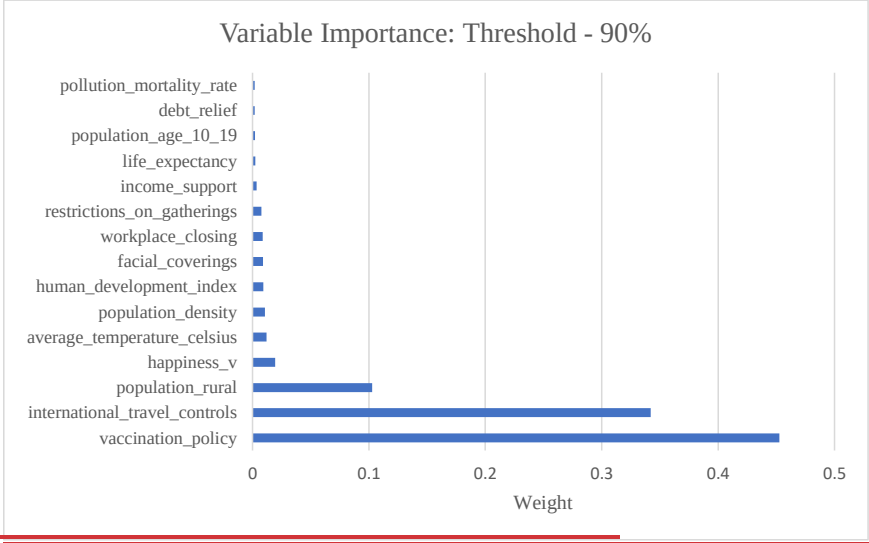


Commented [SR4]: Hierdie grafiek wys nog trust i.p.v happiness by nommer 9 en happiness is 12de.

Commented [GT5R4]: Moet ons dit verander - dit is die ware results - ons kan trust happiness maak?

Fig 9. Ranked variable importance - 70% vaccination threshold.

If we only consider the lowest threshold of 70% vaccination (Fig 9), most factors are objective and similar to the ones mentioned before. Although the share of the younger population also seems to be relatively important. From our sample, we note that all except one country managed to reach the 70% threshold, and therefore more attention should be paid to those factors from the 80% and 90% threshold models.



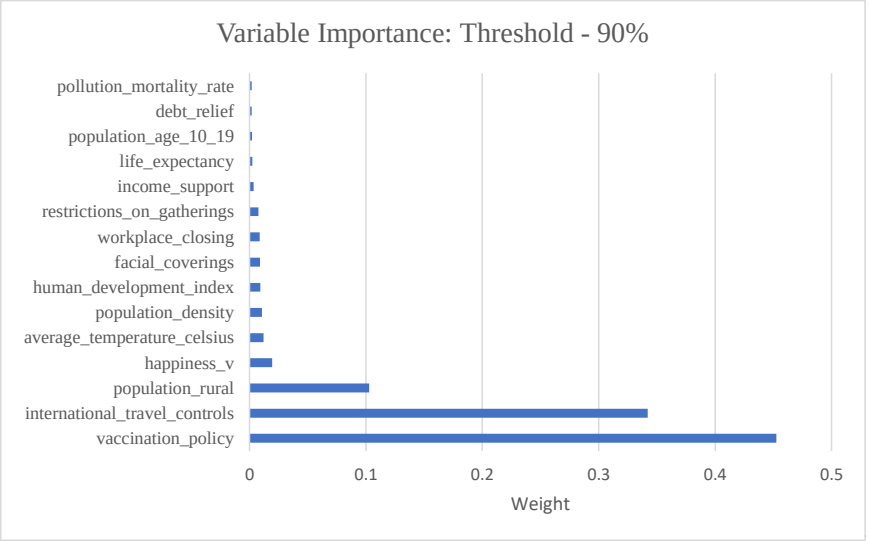


Fig 10. Ranked variable importance - 90% vaccination threshold.

If we move to the 80% threshold model, we again find similar factors important – though restrictions on gatherings become important. Therefore, if governments want to reach the 80% threshold, they should consider the role COVID-19-related policy measures such as restrictions on gatherings play. We note that the association between restrictions on gatherings and vaccination rates can be positive or negative, and a trade-off is implied. If we allow a policy that does not restrict gatherings, higher vaccination rates become more important. However, implementing restrictions on gatherings decreases the possibility of spreading the disease. This implies that more stringent measures can limit the spread instead of vaccinations. However, this is not ideal, given the abundant evidence of the negative effect on well-being resulting from stringent measures such as lockdowns (Smith et al., 2021; Abadi et al., 2021).

6 Discussion on the application

We will focus our discussion on the top 5 factors and use information from previous studies (see sections 2.2 and 2.3) to allude to the relationship with vaccination thresholds. Since we know from Plans-Rubió (2022) that more than 90% of a country's population would need to be vaccinated, given the infectiousness of the pathogen, to achieve herd immunity, our discussion will focus on achieving this "golden standard". Subsequent discussions will highlight where factors have significantly changed in ranking and discuss how happiness and collective emotions can increase vaccination rates. As far as we know, this is the first study that shows the importance of subjective well-being measures.

As noted in section 5.3, regardless of the vaccination threshold goal, governments should focus on their vaccination policy, international travel controls, the percentage of the population in rural areas and the average temperature to achieve their maximum vaccination rates (see Figs 9 and 10).

The vaccination policy implemented (groups that can access the COVID-19 vaccine) was shown by Greyling and Rossouw (2022) that when more groups of people can access the vaccine, for example, all age groups compared to fewer groups, it is positively related to attitude towards the vaccine. This means more people will vaccinate when more people have access to the COVID-19 vaccine.

Regarding international travel controls, we know that, for example, in New Zealand (one of the countries with the most stringent lockdowns and highest number of lockdowns), people were told to get vaccinated if they wanted their freedoms back. The then Prime Minister, Jacinda Ardern, clearly stated, "If you want summer […] get vaccinated." If you don't, "there will be everyday things you will miss out on". It wasn't until September 2022 that New Zealand fully opened up their international borders, allowing visitors ~~back in~~. Rossouw et al. (2021) found that international border controls acted as a dual shock, economic and social. Hospitality operators were impacted directly by the lack of international and domestic tourism and experienced a significant economic shock that negatively influenced their

livelihoods. Furthermore, being unable to travel the world is a social shock causing a decrease in happiness.

When it comes to the population percentage in rural areas, Barbieri et al. (2022) and Polašek et al. (2022) show that vaccine hesitancy is significantly higher in the rural than in the urban population. Additionally, De Boeck et al. (2020) and Oli et al. (2017) found that the complexity of the pipeline for vaccines from the regional depot to the facility level may create breaking points due to inadequate infrastructure and skills gap and that travelling to rural health facilities is more difficult than to urban health facilities. Rural populations, ~~ and~~ vulnerable and excluded people are among those for whom improved vaccination rates and access to care were~~are~~ urgently needed to prevent and treat COVID-19. Therefore, governments need to ensure that the rural populations receive targeted information related to the safety of ~~the~~ vaccines and that the rural population's access to ~~ the~~ vaccines is not hampered by procurement and capacity issues.

This study is the first to show the importance of subjective well-being in achieving vaccination thresholds. Concerning the vaccination threshold of 90%, happiness ranks fourth (seventh in the 80% threshold model and ninth in the 70% model) and is therefore important for governments to address. Measuring happiness, thus a subjective measure that captures people's evaluative mood, is very important in any decision-making process. In an ideal world, people make rational choices. The rational choice theory states that when humans are presented with various options under the conditions of scarcity, they will choose the option that maximises their individual satisfaction. Alas, humans are not rational, and their emotions drive them, and therefore they make irrational decisions. Therefore, emotions and happiness levels also drive decision-making processes in considering whether to get vaccinated. Additionally~~On the other hand~~, previous studies such as Kim et al. (2015) show that happier people make better health-related decisions since happier people are less inclined to engage in high-risk activities and take preventative action to mitigate risk. Also, happy people are not just self-centred or selfish; the literature suggests that happy individuals tend to be relatively more cooperative, prosocial, charitable, and "other-centred" (Kasser & Ryan, 1996; Williams & Shiaw, 1999).

Furthermore, Sarracino et al. (2023) showed that happiness and trust are positively correlated, meaning that as trust increases, so does happiness. Trust in others also promotes cooperation and solidarity with positive spillovers on compliance and well-being (Bargain & Aminjonov, 2020). The takeaway from trust and happiness is quite straightforward; the lower your vaccination rates, the more important people's levels of happiness and trust become. Happiness and trust are connected to compliance and doing something "for the greater good". Therefore, the more you want people to engage in a specific activity, such as getting vaccinated, the more important emotions and happiness levels become.

Average temperature ranks fifth important in all three of our threshold models. Jansson and Yamamoto (2022) studied five states in the U.S. to determine the relationship between average temperature, the

level of humidity and COVID-19 infection rates. The authors found that a higher-than-average temperature was consistently associated with a decreased relative risk of infection. Given that Fieselmann et al. (2022) found that one of the main reasons people do not get vaccinated is a perceived lower risk of infection, we can deduce that higher-than-average temperatures could lead to countries not meeting their maximum number of vaccine dosage uptake as a proportion of the population size of a country. Apart from the above, we know from studies conducted by Streefland et al. (1999a and b) that in developing countries, parents who do not adhere to vaccination schedules often do so because they are unable to go due to climatic conditions such as the weather being too hot, or roads being flooded from significant rainfall, or a crop needs to be to harvest before it withers in the heat. However, we note that the vaccine rollout was hampered in several European countries as well as the U.S. as severe snowstorms and unusual cold fronts caused inoculation centres, including mega facilities capable of vaccinating up to 20,000 people a day, to close (The Guardian, 2021; CBC News, 2021; John Hopkins Healthcare, 2021).

Factors rated among the top 5 in our 80% and 70% threshold models that did not appear in the 90% threshold model are restrictions on gatherings and the population aged between 10-19.

Although not top 5 in the 90% threshold model, restrictions on gatherings play an important role in the 80% threshold model. When Americans began receiving the COVID-19 vaccine at the end of December 2020, people started fantasising about the first thing they would do when the pandemic ended: go back to work, visit family, and hug friends (Marcus, 2021). From Greyling and Rossouw (2022), we also know that compliance with restrictions is negatively related to attitudes against the COVID-19 vaccine. When people are reluctant to comply with orders such as staying at home, then those individuals would be more willing to receive the COVID-19 vaccine. A study by Wright et al. (2022) investigated the relationship between vaccinated individuals' willingness to comply and the implemented behavioural regulations. The entire premise of the study is that vaccinated individuals believe they are less at risk because of their vaccination status. People think that when vaccinated, they do not need to comply with, for example, mask-wearing, social distancing etc., therefore creating a more positive attitude towards vaccines. This finding is informative to policymakers as a message of "less strict regulations" after vaccination can increase vaccine uptake.

As the percentage of the population between 10-19 decreases, the population rate increases since they were last to be vaccinated. Therefore, if only a small proportion were this age, more people would be allowed, according to vaccine policy, to get vaccinated, and the vaccination rate would increase. For example, for all developed countries in the sample groups between 10 and 19 were 12 per cent or less of the population – whereas in South Africa it was almost 18%... this is an indication of many things – also Western countries' populations are getting older – thus they were in a higher need to vaccinate the older bigger groups of people.

7 Conclusion

In this study, we used supervised machine learning to retrospectively evaluate the COVID-19 pandemic and determine the factors most important in increasing vaccine uptake. Therefore, we determined those factors associated with achieving herd immunity at the 70% vaccination threshold, estimated at the beginning of the COVID-19 pandemic and the 90% vaccination threshold, estimated later in the pandemic. By doing the aforementioned, we also determined those factors that differed between the 70% to 90% vaccination threshold, which were responsible for reaching the higher vaccination level. Throughout our analyses, we paid special attention to the role of subjective well-being measures in achieving vaccine thresholds since we know that negative emotions, such as fear of the side effects of vaccines, influence peoples' attitudes towards receiving the vaccine and that happier people make better health-related decisions.

We trained our models on the merged data set~~merged dataset consisting~~ of 6530 observations using an Extreme Gradient Boosting (XGBoost) algorithm and also used Random Forest and Decision Tree algorithms as robustness tests. After testing for precision, we found that the XGBoost model gave the best-fit measures and delivered the best results compared to the other two methods. Consequently, we discussed the results of the XGBoost model ~~applied to our test data~~ in determining the most important predicted factors that contributes to reaching different levels of herd immunity. ~~contributing to higher vaccination rates.~~

The above allowed us to make several contributions to existing literature. First, ours was the first study to conduct a post-COVID-19 cross-country analysis of the most important variables to reach different levels of herd immunity~~increase vaccine uptake~~. Second, we were also the first study to include subjective measures of well-being in our estimations. Third, we were the first study ot differentiate between the most important factors to reach different levels of herd immunity. Fourth we were the first to accomplish the [pre--mentioned using~~e~~ various machine learning algorithms to train our models~~data~~ and determine which algorithm gives us the best fit, i.e., the most reliable predictions. Subsequently, our XGBoost model can be used as a benchmark for future research related to the most important factors for reaching herd immunity levels~~increasing vaccination uptake~~. Furthermore, this study offered some actionable insights for policymakers on increasing vaccination rates to curb pandemics' health, economic and political effects.

Interestingly our ~~preferred~~ XGBoost model revealed similar important factors in predicting the 70% and 90% vaccination thresholds to reach different levels of herd immunity. These included the vaccination policy implemented, international travel controls, the percentage of the population in rural areas and the average temperature. Of significance was happiness's role in attaining the 90% vaccine threshold. Whereas happiness had a lower importance level in achieving the 70% threshold, the

importance of happiness in achieving the 90% vaccine threshold was clear. If governments want higher levels of compliance and vaccine uptake, subjective well-being measures such as mood and emotions must be prioritised. Addressing how people feel, in general, towards vaccines and governments is vitally important when policymakers want to push beyond the lower 70% vaccine threshold and achieve the "golden standard" of 90% fully vaccinated.

It would be negligent of us not to discuss our study's limitations. First, the sample of countries under investigation are mostly developed countries. It will be interesting to extend the sample to determine those policies, characteristics and subjective well-being measures deemed necessary to increase vaccination rates in developing countries and contrast those to the factors applicable to developed nations.

Second, although we know that lack of international support and cooperation played a significant role in procuring and disseminating vaccines in developing countries, we could not add variables reflecting international support or emergency investment in health care to our models due to high missingness. We acknowledge that these variables could have ranked among the most important variables and potentially have been included in the top five. The missingness of the observations of these variables is further proof of the failures of countries to prepare for pandemics and give international support. The missingness on international support was 67%, implying that international support was given infrequently. In the event where we added the amounts from the developed countries in our sample, it was still minimal. Furthermore, countries did not frequently invest in emergency health care. Of the observations in our dataset on this variable, 74% were missing. Note that these numbers are for developed countries; therefore, it is easy to imagine what the variable would reveal for developing countries. When we added these amounts, it was very little compared to the amounts spent on, for example, vaccines.

References

Abadi, D., Arnaldo, I., & Fischer, A. (2021). Anxious and Angry: Emotional Responses to the COVID-19 Threat. *Frontiers in Psychology, 12*, 676116.

Abdurrahim, Y., Ali, A. D., Sena, K., & Huseyin, U. (2020). Comparison of deep learning and traditional machine learning techniques for classification of pap smear images. *arXiv*, 2009.06366v1. Available from https://arxiv.org/pdf/2009.06366.pdf

Andrade, C., Gillen, M., Molina, J. A., & Wilmarth, M. J. (2022). The Social and Economic Impact of Covid-19 on Family Functioning and Wellbeing: Where do we go from here? *Journal of Family and Economic Issues,43*(2),205-212.

Anik, L., Aknin, L.B, Norton, M.I., &. Dunn, E.W. (2009). Feeling good about giving: The benefits (and costs) of self-interested charitable behavior. Harvard Business School Marketing Unit Working Paper No. 10-012, viewed 15 May 2020, from https://ssrn.com/abstract=1444831

Bajos, N., Spire, A., Silberzan, L., Sireyjol, A., Jusot, F., Meyer, L., Franck, J.-E., & Warszawski, J. (2022). When Mistrust in the Government and Scientists Reinforce Social Inequalities in Vaccination against Covid-19. Frontiers in Public Health, 10, 908152.

Baldwin, R. (2020). Keeping the lights on: Economic medicine for a medical shock. VoxEU.Org. 2020. Available from https://voxeu.org/article/how-should-we-think-about-containing-covid-19-economic-crisis

Barbieri, V., Wiedermann, C.J., Lombardo, S., Ausserhofer, D., Plagg, B., Piccoliori, G., Gärtner, T., Wiedermann, W., & Engl, A. (2022). Vaccine Hesitancy in the Second Year of the Coronavirus Pandemic in South Tyrol, Italy: A Representative Cross-Sectional Survey. *Vaccines, 10*:1584.

Bargain, O., & Aminjonov, U. (2020). Trust and compliance to public health policies in times of covid-19. *Journal of Public Economics, 192*, 104316.

Bergmann, M., Bethmann, A., Hannemann, T.-V., & Schumacher, A. T. (2022). Who are the Unvaccinated? Determinants of SARS-CoV-2 Vaccinations among Older Adults Across Europe. *Easy Social Sciences, Mixed 1*, 1-11. Available from https://doi.org/10.15464/easy.2022.01

Bíró-Nagy, A., & Szászi, A. J. (2022). The Roots of COVID-19 Vaccine Hesitancy: Evidence from Hungary. Journal of Behavioral Medicine, 1–16.

Bloom, D. E., Cadarette, D., Ferranna, M. (2021). The Societal Value of Vaccination in the Age of COVID-19. *American Journal of Public Health, 111*(6), 1049-1054.

Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5–32.

CBC News. (2021). Canada's Pfizer vaccine shipment delayed by winter weather in the U.S. Available from https://www.cbc.ca/news/politics/pfizer-delays-winter-weather-1.5915661

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). Association for Computing Machinery, New York, NY, USA, pp. 785–794. Available from https://doi.org/10.1145/2939672.2939785

Cheong, Q., Quon, S., Concepcion, K., & Kong, J. D. (2021). Predictive Modeling of Vaccination Uptake in U.S. Counties: A Machine Learning–Based Approach. *Journal of Medical Internet Research*, *23*(11).

Corcoran, K. E., Scheitle, C. P., & DiGregorio, B. D. (2021). Christian Nationalism and COVID-19 Vaccine Hesitancy and Uptake. *Vaccine, 39*(45), 6614– 6621.

Czeisler, M. E., Rajaratnam, S. W. M., Howard, M. E., & Czeisler, C. A. (2021). COVID-19 Vaccine Intentions in the United States—December 2020 to March 2021. Working paper, MedRxiv. Available from https://www.medrxiv.org/content/10.1101/2021.05.16.21257290v1.full.pdf

De Boeck, K., Decouttere, C., & Vandaele, N. (2020). Vaccine distribution chains in low- and middle-income countries: a literature review. *Omega, 97*,102097.

El-Mohandes, A., White, T. M., Wyka, K., Rauh, L., Rabin, K., Kimball, S. H., Ratzan, S. C., & Lazarus, J. V. (2021). COVID-19 Vaccine Acceptance among Adults in Four Major US Metropolitan Areas and Nationwide. Scientific Reports 11(1), 21844.

Fetzer, T. R., Witte, M., Hencel, L., Jachimowicz, J., Haushofer, J., Ivchenko, A., Caria, S., Reutskaja, E., Roth, C. P., Fiorin, S., Gómez, M., Kraft-Todd,. G, Götz, F. M., & Yoeli, E. (2020). Global Behaviors and Perceptions at the Onset of the COVID-19 Pandemic. National Bureau of Economic Research Working Paper No, 27082.

Fieselmann, J., Annac, K., Erdsiek, F. Yilmaz-Aslan, Y., & Brzoska, P. (2022). What are the reasons for refusing a COVID-19 vaccine? A qualitative analysis of social media in Germany. *BMC Public Health, 22*, 846.

Gatwood, J., McKnight, M., Fiscus, M., Hohmeier, K. C., & Chisholm-Burns, M. (2021). Factors Influencing Likelihood of COVID-19 Vaccination: A Survey of Tennessee Adults. *American Journal of Health-System Pharmacy, 78*(10), 879–889.

Gomes, I. A., Soares, P., Rocha, J. V., Gama, A., Laires, P. A., Moniz, M., Pedro, A. R., Dias, S., Goes, A. R., Leite, A., & Nunes, C. (2022). Factors Associated with COVID-19 Vaccine Hesitancy after Implementation of a Mass Vaccination Campaign. *Vaccines, 10*(2), 281.

Greyling, T., & Rossouw, S. (2022). Positive attitudes towards COVID-19 vaccines: A cross-country analysis. *PLOS ONE, 17*(3), 0264994.

Greyling, T., Rossouw, S., & Afstereo. (2019). *Gross National Happiness.today*. Available from http://gnh.today

Hale, T., Angrist, N., Goldszmidt, R., Kira, B., Petherick, A., Phillips, T., Webster, S., Cameron-Blake, E., Hallas, L., Majumdar, S., & Tatlow, H. (2021). A global panel database of pandemic policies (Oxford COVID-19 Government Response Tracker). *Nature Human Behaviour, 5*, 529–538.

Hapfelmeier, A., Hothorn, T., Ulm, K., & Strobl, C. (2014). A new variable importance measure for random forests with missing data. *Statistics and Computing, 24*, 21–34.

Jansson, M. K., & Yamamoto, S. (2022). The effect of temperature, humidity, precipitation and cloud coverage on the risk of COVID-19 infection in temperate regions of the USA—A case-crossover study. *PLOS ONE, 17*(9), e0273511.

John Hopkins Healthcare. (2021). Winter Storm Slows U.S. COVID Vaccine Rollout. Available from https://johnshopkinshealthcare.staywellsolutionsonline.com/RelatedItems/6,1650551452

Kasser, T., & Ryan, R. M. (1996). Further examining the American dream: Differential correlates of intrinsic and extrinsic goals. *Personality and Social Psychology Bulletin, 22,* 280–287.

Kerr, J. R., Schneider, C. R., Recchia, G., Dryhurst, S., Sahlin, U., Dufouil, C., Arwidson, P., Freeman, A. L. J., & van der Linden, S. (2021). Correlates of Intended COVID-19 Vaccine Acceptance Across Time and Countries: Results from a Series of Cross-Sectional Surveys. *BMJ Open, 11*(8), e048025.

Kim, E.S., Kubzansky, L.D., & Smith, J. (2015). Life satisfaction and use of preventive health care services. *Health Psychology*, *34*(7), 779– 782.

Latkin, C. A., Dayton, L., Yi, G., Konstantopoulos, A., & Boodram, B. (2021). Trust in a COVID-19 Vaccine in the U.S.: A Social-Ecological Perspective. *Social Science and Medicine, 270*, 113684.

Liang, W., Luo, S., Zhao, G., & Wu, H. (2020). Predicting Hard Rock Pillar Stability Using GBDT, XGBoost, and LightGBM Algorithms. *Mathematics*, *8*(5), 765.

Lincoln, T. M., Schlier, B., Strakeljahn, F., Gaudiano, B. A., So, S. H., Kingston, J., Morris, E. M., & Ellett, L. (2022). Taking a machine learning approach to optimise prediction of vaccine hesitancy in high income countries. *Scientific Reports*, *12*(1), 1-12.

Lindholt, M. F., Jørgensen, F., Bor, A., & Petersen, M. B. (2021). Public Acceptance of COVID-19 Vaccines: Cross-National Evidence on Levels and Individual-Level Predictors using Observational Data. *BMJ Open, 11*(6), e048172.

Lu, H., Nie, P., & Qian, L. (2020). Do Quarantine Experiences and Attitudes Towards COVID-19 Affect the Distribution of Psychological Outcomes in China? A Quantile Regression Analysis. Global Labor Organization Discussion Paper No, 512.

Ludvigson, S. C., Ma, S., & Ng, S. (2020). Covid19 and the Macroeconomic Effects of Costly Disasters. National Bureau of Economic Research Working Paper No, 26987. Available from https://doi.org/10.3386/w26987

Lyubomirsky, S., Sheldon, K.M., & Schkade, D. (2005). Pursuing happiness: The architecture of sustainable change. *Review of General Psychology, 9*(2), 111–131.

Marcus, J. (2021). Vaccinated People Are Going to Hug Each Other. *The Atlantic.* Available from https://www.theatlantic.com/ideas/archive/2021/01/giving-people-more-freedom-whole-point-vaccines/617829/

Mathieu, E., Ritchie, H., Roser, M., Hasell, J., Appel, C., & Giattino, C. (2021). A global database of COVID-19 vaccinations. *Nature Human Behaviour*, *5*(7), 947-953.

Mental Health and Wellbeing Commission. (2023). COVID-19 and safety in the home. COVID-19 Impact Insights Paper Number 4. Available from https://www.mhwc.govt.nz/assets/COVID-19-insights/Paper-4-COVID-and-safety-in-the-home/ENG_SafetyReport_Summary.pdf

Nielsen, D. (2016). Tree boosting with XGBoost: why does XGBoost win "every" machine learning competition? Norwegian University of Science and Technology. Available from https://ntnuopen.ntnu.no/ntnu-xmlui/bitstream/handle/11250/2433761/16128_FULLTEXT.pdf

Oli, A. N., Agu, R. U., Ihekwereme, C. P., & Esimone, C. O. (2017). An evaluation of the cold chain technology in South-East, Nigeria using Immunogenicity study on the measles vaccines. *Pan African Medical Journal, 27*:1–5.

Osman, S. M. I., & Sabit, A. (2022). Predictors of COVID-19 vaccination rate in USA: A machine learning approach. *Machine Learning with Applications, 10,* 100408.

Paul, E., Steptoe, A., & Fancourt, D. (2021). Attitudes towards vaccines and intention to vaccinate against COVID-19: Implications for public health communications. *The Lancet Regional Health-Europe, 1*(100012), 1–10.

Plans-Rubió P. (2022). Percentages of Vaccination Coverage Required to Establish Herd Immunity against SARS-CoV-2. *Vaccines, 10*(5), 736.

Polašek, O., Wazny, K., Adeloye, D., Song, P., Chan, K.Y., Bojude, D.A., Ali, S., Bastien, S., Becerra-Posada, F., Borrescio-Higa, F., et al. (2022). Research Priorities to Reduce the Impact of COVID-19 in Low- and Middle-Income Countries. *Journal of Global Health, 12*:09003.

Pronkina, E., & Rees, D. I. (2022). Predicting COVID-19 Vaccine Uptake. Institute of Labor Economics (IZA) Discussion Paper No. 15625.

Raciborski, F., Samel-Kowalik, P., Gujski, M., Pinkas, J., Arcimowicz, M., & Jankowski, M. (2021). Factors Associated with a Lack of Willingness to Vaccinate Against COVID-19 in Poland: A 2021 Nationwide Cross-Sectional Survey. *Vaccines, 9*(9), 1000.

Rossouw, S., Greyling, T., Adhikari, T. (2021). The evolution of happiness pre and peri-COVID-19: A Markov Switching Dynamic Regression Model. *PLoS ONE, 16*(12), e0259579.

Rozek, L. S., Jones, P., Menon, A., Hicken, A., Apsley, S., & King, E. J. (2021). Understanding Vaccine Hesitancy in the Context of COVID-19: The Role of Trust and Confidence in a Seventeen-Country Survey. *International Journal of Public Health, 66*, 636255.

Sallam, M. (2021). COVID-19 vaccine hesitancy worldwide: a systematic review of vaccine acceptance rates. *Vaccines, 9*(160), 1–14.

Shaham, A., Chodick, G., Shalev, V., & Yamin, D. (2020). Personal and social patterns predict influenza vaccination decisions. *BMC Public Health, 20*(222), 1–12.

Sheikh, A. B., Pal, S., Javed, N., & Shekhar, R. (2021). COVID-19 Vaccination in Developing Nations: Challenges and Opportunities for Innovation. *Infectious Disease Report, 14*(2),429–436.

Smith, L. E., Duffy, B., Moxham-Hall, V., Strang, L., Wessely, S., & Rubin, G. J. (2021). Anger and confrontation during the COVID-19 pandemic: a national cross-sectional survey in the U.K. *Journal of the Royal Society of Medicine, 114*(2), 77–90.

Streefland, P. H., Chowdhury, A. M. R., & Ramos-Jimenez, P. (1999a). Patterns of Vaccination Acceptance. *Social Science and Medicine, 49*, 1705–16.

Streefland, P. H., Chowdhury, A. M. R., & Ramos-Jimenez, P. (1999b). Quality of vaccination services and social demand for vaccinations in Africa and Asia. *Bulletin of the World Health Organization, 77*,722–30.

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics*, *9*(1), 1–11.

The Guardian. (2021). Severe snowstorm forces Greece to halt Covid vaccination drive. Available from https://www.theguardian.com/world/2021/feb/16/severe-snowstorm-forces-greece-to-halt-covid-vaccination-drive

United Nations, Department of Economic and Social Affairs, Population Division (2022). *World Population Prospects 2022, Online Edition*. Available from https://population.un.org/wpp/Download/Standard/MostUsed/

Ward, J. K., Alleaume, C., & Peretti-Watel, P. (2020). The French Public's Attitudes to a Future COVID-19 Vaccine: The Politicization of a Public Health Issue. *Social Science and Medicine, 265*, 113414.

Williams, S., & Shiaw, W. T. (1999). Mood and organisational citizenship behavior: The effects of positive affect on employee organisational citizenship behavior intentions. *Journal of Psychology, 133,* 656–668.

Wollebæk, D., Fladmoe, A., Steen-Johnsen, K., & Ihlen, Ø. (2022). Right-Wing Ideological Constraint and Vaccine Refusal: The Case of the COVID-19 Vaccine in Norway. *Scandinavian Political Studies, 45*(2), 253–278.

World Bank. (2018). The Human Capital Project. World Bank, Washington, DC. Available from https://data.worldbank.org/indicator/HD.HCI.OVRL.UB.MA?end=2020&start=2020&view=bar

World Bank. (2023a). Climate Change Knowledge Portal. Available from https://climateknowledgeportal.worldbank.org/

World Bank. (2023b). Food and Agriculture Organization and World Bank population estimates. Available from https://data.worldbank.org/indicator/EN.POP.DNST?view=chart

World Bank staff estimates based on the United Nations Population Division's World Urbanization Prospects: 2018 Revision. Available from https://data.worldbank.org/indicator/SP.RUR.TOTL.ZS

World Health Organisation. (2023). WHO Coronavirus (COVID-19) Dashboard. Available from https://covid19.who.int/ Accessed on 20 July 2023.

Wright, L., Steptoe, A., Mak, H. W., Fancourt, D. (2022). Do people reduce compliance with COVID-19 guidelines following vaccination? A longitudinal analysis of matched U.K. adults. *Journal of Epidemiology & Community Health, 76*, 109–115.

Appendix A

To derive our time-series data which captures sentiment and emotions, we construct variables using Big Data by extracting tweets from Twitter. In our analysis, we extracted two sets of tweets based on keywords, one related to COVID-19 vaccines and the other related to the government. The tweets containing these words amounted to 1,047,000 tweets. We extracted all tweets according to specific geographical areas (country).

For COVID-19 vaccines, we extract tweets using the keywords: *vaccinate, vacc, vaccine, Sputnik V, Sputnik, Sinopharm, Astrazeneca, Pfizer (if NEAR) vaccine, Pfizer-BioNTech, Johnson & Johnson,* and *Moderna*.

For the government, we extract tweets using the keywords: *government, parliament, ministry, minister, senator, M.P.s, legislator, political, politics, prime minister.*

The first step in our analysis is determining the tweets' language (we detected 64 different languages), and all non-English tweets were translated into English. After the translation process, we use NLP to extract the tweets' sentiment and underlying emotions. To test the robustness of coding the sentiment of the translated tweets, we use lexicons in the original language, if available, and repeat the process. We compare the coded sentiment of the translated and original text and find the results strongly correlated.

We make use of a suite of lexicons. Each differs slightly but primarily aims to determine the sentiment of unstructured text data. The two lexicons mostly used in our analysis are Sentiment140 and NRC (National Research Council of Canada Emotion Lexicon developed by Turney and Mohammad (2010)). The other lexicons are used for robustness purposes and are part of the Syuzhet package. The lexicons include Syuzhet, AFINN and Bing. The sentiment is determined by identifying the tweeter's attitude towards an event using variables such as context, tone, etc. It helps one form an entire opinion of the text. Depending on the lexicon used, the text (tweet) is coded. For example, if a tweet is positive, it is coded as 0; if neutral, 2 and if negative, 4.

We use the NRC lexicon to code the sentiment (as explained above) and analyse the underlying tweets' emotions. It distinguishes between eight basic emotions: anger, fear, anticipation, trust, surprise, sadness, joy and disgust (the so-called Plutchik (1980) wheel of emotions). NRC codes words with different values, ranging from 0 (low) to 8 (the highest score in our data), to express the intensity of an emotion or sentiment.

To construct the time-series data, we use the coding of the tweets and derive daily averages. In this manner, we derive a positive sentiment, a negative sentiment and eight emotion time series. We derive the sentiment time series using different lexicons as a robustness test and compare these results using correlation analyses. We perform additional robustness tests, for example, determining whether the sampling frequency significantly influences the results.

To test the robustness of the *frequency,* we construct the relevant index (time series) per day (the norm); we repeat the exercise but construct the time series per hour. We find similar trends in our hourly and daily time series, indicating that the timescale at which sampling occurs does not significantly influence the observed trend.

To test whether the *volume* of tweets affects the derived time-series data, we extract random samples of differing sizes from the daily text corpus of tweets. The time series based on these smaller samples (50 per cent and 80 per cent of the daily extracted tweets) are highly correlated to the original time series.

Appendix B

Full list of variables

Variable	Description	Scale	Coding	Source
Vaccination policy	Policies for vaccine delivery for different groups	Ordinal scale	0 - No availability 1 - Availability for ONE of following: key workers/ clinically vulnerable groups (non-elderly) / elderly groups 2 - Availability for TWO of following: key workers/ clinically vulnerable groups (non-elderly) / elderly groups 3 - Availability for ALL of following: key workers/ clinically vulnerable groups (non-elderly) / elderly groups 4 - Availability for all three plus partial additional availability (select broad groups/ages) 5 - Universal availability	Hale et al. (2021)
Population rural (Percentage of population in rural areas)	People living in rural areas as defined by national statistical offices. It is calculated as the difference between the total and urban populations.	Percentage		World Bank staff estimates based on the United Nations Population Division's World Urbanization Prospects: 2018 Revision.
Average temperature	Average temperature in the country	Celsius		World Bank (2023a)
Population density	People per square kilometre of land area			United Nations, Department of Economic and Social Affairs, Population Division (2022)
Workplace closing	Record closing of workplaces	Ordinal	0 - no measures 1 - recommend closing (or recommend work from home) or all businesses open with alterations resulting in significant differences compared to non-Covid-19 operations 2 - require closing (or work from home) for some sectors or categories of workers 3 - require closing (or work from home) for all-but-essential workplaces (e.g. grocery stores, doctors) Blank - no data	Hale et al. (2021)
Restrictions on gatherings	Record limits on gatherings	Ordinal	0 - no restrictions 1 - restrictions on very large gatherings (the limit is above 1000 people) 2 - restrictions on gatherings between 101-1000 people 3 - restrictions on gatherings between 11-100 people 4 - restrictions on gatherings of 10 people or less Blank - no data	Hale et al. (2021)
International travel controls	Restrictions on international travel	Ordinal	0 - no restrictions 1 - screening arrivals 2 - quarantine arrivals from some or all regions 3 - ban arrivals from some regions	Hale et al. (2021)

			4 - ban on all regions or total border closure Blank - no data	
Life expectancy	The average number of years a newborn would live if age-specific mortality rates in the current year were to stay the same throughout its life.	Years		United Nations, Department of Economic and Social Affairs, Population Division (2022)
Population age 10-19		Continuous		World Bank
Face coverings	Policies on the use of facial coverings outside the home	Ordinal	0 - No policy 1 - Recommended 2 - Required in some specified shared/public spaces outside the home with other people present or some situations when social distancing not possible 3 - Required in all shared/public spaces outside the home with other people present or all situations when social distancing not possible 4 - Required outside the home at all times regardless of location or presence of other people	Hale et al. (2021)
Income support	Record if the government provides direct cash payments to people who lose their jobs or cannot work. Note: only includes payments to firms if explicitly linked to payroll/salaries	Ordinal	0 - no income support 1 - government is replacing less than 50% of lost salary (or if a flat sum, it is less than 50% of median salary) 2 - government is replacing 50% or more of lost salary (or if a flat sum, it is greater than 50% of median salary) Blank - no data	Hale et al. (2021)
Pollution mortality rate				United Nations, Department of Economic and Social Affairs, Population Division (2022)
Debt relief	Record if the government is freezing financial obligations for households (e.g. stopping loan repayments, preventing services like water from stopping, or banning evictions)	Ordinal	0 - no debt/contract relief 1 - narrow relief, specific to one kind of contract 2 - broad debt/contract relief	Hale et al. (2021)
Surprise_Vac	The emotion surprise towards vaccines			Greyling et al. (2019)
Human capital index				World Bank staff calculations based on the methodology

				described in World Bank (2018)
Human development index				UNDP
GDP per capita (US$)				World Bank
Sadness_GNH	The emotion general sadness			Greyling et al. (2019)
Trust_GNH	The emotion general trust			Greyling et al. (2019)
Anticipation_Gov	The emotion anticipation towards government			Greyling et al. (2019)
Disgust_Gov	The emotion disgust towards government			Greyling et al. (2019)
Fear_Gov	The emotion fear towards government			Greyling et al. (2019)
Vac-VADER_sent	Sentiment towards the vaccine			Greyling et al. (2019)
Diabetes_prevalence				
Joy_Gov	The emotion joy towards government			Greyling et al. (2019)
Sadness_Gov	The emotion sadness towards government			Greyling et al. (2019)
Population age_0-9				
Anticipation_GNH	The emotion general anticipation			Greyling et al. (2019)
Fear_GNH	The emotion general fear			Greyling et al. (2019)
Vac_GNH	Happiness towards the vaccine			Greyling et al. (2019)
Joy_GNH	The emotion general joy			Greyling et al. (2019)
Anger_GNH	The emotion general anger			Greyling et al. (2019)
Disgust_GNH	The emotion general disgust			Greyling et al. (2019)
Contact tracing	Record government policy on contact tracing after a positive diagnosis	Ordinal scale	0 - no contact tracing 1 - limited contact tracing; not done for all cases 2 - comprehensive contact tracing; done for all identified cases	Hale et al. (2021)
Surprise_Gov	The emotion surprise towards government			Greyling et al. (2019)
GNH_Gov	Happiness towards government			Greyling et al. (2019)
Trust_Gov	The emotion trust towards government			Greyling et al. (2019)
Surprise_GNH	The emotion general surprise			Greyling et al. (2019)

Anger_Gov	The emotion anger towards government			Greyling et al. (2019)
Vac_anticipation	The emotion anticipation towards the vaccine			Greyling et al. (2019)
Vac_disgust	The emotion disgust towards the vaccine			Greyling et al. (2019)
Vac_sadness	The emotion sadness towards the vaccine			Greyling et al. (2019)
Vac_fear	The emotion fear towards the vaccine			Greyling et al. (2019)
Vac_anger	The emotion anger towards the vaccine			Greyling et al. (2019)
Vac_trust	The emotion trust towards the vaccine			Greyling et al. (2019)
Vac_joy	The emotion joy towards the vaccine			Greyling et al. (2019)
Testing policy	Record government policy on who has access to testing Note: this records policies about testing for current infection (PCR tests), not testing for immunity (antibody test)	Ordinal scale	0 - no testing policy 1 - only those who both (a) have symptoms AND (b) meet specific criteria (e.g. key workers, admitted to hospital, came into contact with a known case, returned from overseas) 2 - testing of anyone showing Covid-19 symptoms 3 - open public testing (e.g. "drive through" testing available to asymptomatic people) Blank - no data	Hale et al. (2021)
Infant mortality rate				
Out-of-pocket health expenditure				Hale et al. (2021)
Comorbidity mortality rate				Hale et al. (2021)
Smoking prevalence				Hale et al. (2021)
Physicians per 1000				Hale et al. (2021)
Population age_40-49				
Nurses per 1000				Hale et al. (2021)
Population age_60-69				
Population age_80_and_older				
GDP (USD)				
Public information campaigns				
Health expenditure (USD)				
Population age 20-29				

Population age 70-79				
Vader_pos_gov	Positive sentiment towards the government			Greyling et al. (2019)
Population age 50-59				
Vader_neg_gov	Negative sentiment towards the government			Greyling et al. (2019)
Population age 30-39				
Vac_vader_neg	Negative sentiment towards the vaccine			Greyling et al. (2019)
Vac_vader_pos	Positive sentiment towards the vaccine			Greyling et al. (2019)
Vac_vader_sent	Sentiment towards the vaccine			Greyling et al. (2019)

Appendix C

Ranking according to the importance of variables. XGBoost, Random Forest and Decision Tree – 90 % threshold

XGBoost – 90% threshold	Random Forest – 90% threshold	Decision Tree – 90% threshold
Vaccination policy	Vaccination policy	Vaccination policy
International travel controls	Restrictions on gatherings	Testing policy
Percentage of population in rural areas	International travel controls	Public information campaigns
Happiness	Debt relief	Contact tracing
Average temperature	Facial coverings	Facial coverings
Population density	Testing policy	International travel controls
Human Development Index	Income support	Income support
Facial coverings	Contact tracing	Restrictions on gatherings
Workplace closing	Comorbidity mortality rate	Population aged between 20-29
Restrictions on gatherings	Average temperature	Population aged between 0-9
Income support	Infant mortality rate	Population aged between 10-19
Life expectancy	Workplace closing	Human Development Index
Pollution mortality rate	Population aged 80 and older	Percentage of population in rural areas
Out-of-pocket health expenditure	GDP (USD)	Infant mortality rate
Debt relief	Public information campaigns	Out-of-pocket health expenditure
Trust (GNH)	Diabetes prevalence	Population aged between 30-39
Human capital index	Out-of-pocket health expenditure	Population aged between 40-49
Diabetes prevalence	Population density	Health expenditure (USD)
Human Development Index	Smoking prevalence	GDP per capita (USD)
DDP per capita (USD)	Life expectancy	Human capital index
Sadness (GNH) – lack of happiness	Disgust (GNH) – lack of happiness	Population density
Sentiment towards vaccines	Human capital index	Smoking prevalence
Smoking prevalence	Anger (GNH) – lack of happiness	Anger (GNH) – lack of happiness