U.PORTO
FEP FACULDADE DE ECONOMIA
UNIVERSIDADE DO PORTO

Predicting adherence to public health measures during the COVID-19 pandemic: a machine learning approach

Daniela Couto Botelho Monteiro

Dissertation

Master in Modelling, Data Analysis and Decision Support Systems

Supervised by

Prof. Doutor Patrício Costa
Prof. Doutor Jorge Valente

2023

Acknowledgments

I would like to acknowledge and express my sincere gratitude to my supervisors, Prof. Doutor Patrício Costa and Prof. Doutor Jorge Valente. This work would not have been possible without their guidance and incentive, while accommodating my desire to work on this project, not only from the data analysis and machine learning approach but also with the psychologist *goggles*.

I would also like to extend my appreciation to all the MADSAD students and professors with whom I was fortunate to work on this very enriching journey. Their different backgrounds, knowledge, and passions allowed me to learn so much and motivated me to explore new areas, methods, and tools and ultimately complete this cycle.

This was, therefore, not an individual achievement, but a collective accomplishment, only possible due to my family and friends' support, to whom I am so very grateful and ready to give back the time taken by this project.

This project uses data from SHARE Waves 1, 2, 3, 4, 5, 6, 7, 8, and 9 (DOIs: 10.6103/SHARE.w1.800, 10.6103/SHARE.w2.800, 10.6103/SHARE.w3.800, 10.6103/SHARE.w4.800, 10.6103/SHARE.w5.800, 10.6103/SHARE.w6.800, 10.6103/SHARE.w7.800, 10.6103/SHARE.w8.800, 10.6103/SHARE.w8ca.800, 10.6103/SHARE.w9ca800) see Börsch-Supan et al. (2013) for methodological details. The SHARE data collection has been funded by the European Commission, DG RTD through FP5 (QLK6-CT-2001-00360), FP6 (SHARE-I3: RII-CT-2006-062193, COMPARE: CIT5-CT-2005-028857, SHARELIFE: CIT4-CT-2006-028812), FP7 (SHARE-PREP: GA N°211909, SHARE-LEAP: GA N°227822, SHARE M4: GA N°261982, DASISH: GA N°283646) and Horizon 2020 (SHARE-DEV3: GA N°676536, SHARE-COHESION: GA N°870628, SERISS: GA N°654221, SSHOC: GA N°823782, SHARE-COVID19: GA N°101015924) and by DG Employment, Social Affairs & Inclusion through VS 2015/0195, VS 2016/0135, VS 2018/0285, VS 2019/0332, and VS 2020/0313. Additional funding from the German Ministry of Education and Research, the Max Planck Society for the Advancement of Science, the U.S. National Institute on Aging (U01_AG09740-13S2, P01_AG005842, P01_AG08291, P30_AG12815, R21_AG025169, Y1-AG-4553-01, IAG_BSR06-11, OGHA_04-064, HHSN271201300071C, RAG052527A) and from various national funding sources is gratefully acknowledged (see www.share-project.org).

Resumo

A crise pandémica relativa à COVID-19 desencadeou, na maioria dos países, a implementação de medidas de controlo epidémico para retardar a propagação do vírus. Embora a adesão às medidas de saúde pública tenha demonstrado efeitos positivos na diminuição da propagação do vírus e no número total de mortes, uma compreensão mais profunda do fenómeno poderá levar a uma melhor gestão da futuras crises. Assim, desenvolvemos uma estrutura de *Machine Learning* que nos permitisse prever a adesão às medidas de segurança.

Utilizamos dados do *Survey of Health, Ageing and Retirement in Europe* (SHARE), devido à sua riqueza intrinsecamente ligada à natureza longitudinal do estudo e às duas vagas especiais dedicadas à pandemia COVID-19. Além das variáveis específicas da COVID-19, outras variáveis do painel foram incluídas na análise através do EasySHARE.

O pré-processamento envolveu a seleção (usando o algoritmo MRMR) e criação de variáveis, o tratamento de *missing values* e a partição em dados de treino e teste, incluindo o uso de uma técnica de sobreamostragem para lidar com o desequilíbrio da variável dependente.

Sete algoritmos foram aplicados para criar os modelos de predição de classificação de uma variável categórica – a Adesão – construída a partir de comportamentos de saúde relacionados com a prevenção e risco. As métricas de desempenho dos modelos foram avaliadas e comparadas (*Accuracy, Precision, Recall, weighted average F1 score and Area Under the Receiver Operating Characteristic Curve One-vs-one*) e o melhor modelo foi identificado.

Também utilizámos *SHAP values* de forma a explicar a predição do melhor modelo, para uma melhor compreensão das contribuições de cada variável para o processo de classificação. O país do participante, a frequência de contacto com vizinhos, amigos ou colegas e a visita a um médico ou a um centro médico que não um hospital foram identificadas como sendo as variáveis com maior impacto.

Abstract

The pandemic crisis due to COVID-19 triggered, in most countries, the implementation of epidemic control measures to slow the spread of the virus. Although adherence to public health measures has shown positive effects in decreasing the spread of the virus and in the total number of deaths, a deeper understanding of the phenomenon may lead to better management of future crises. Thus, we aimed to develop a Machine Learning framework that allowed us to predict adherence to safety measures.

We used data from the Survey of Health, Ageing and Retirement in Europe (SHARE), due to its richness intrinsically linked to the longitudinal nature of the study and the two special waves dedicated to the COVID-19 pandemic. In addition to the COVID-19 specific variables, other panel variables were included in the analysis through EasySHARE.

Pre-processing involved feature selection (using the MRMR algorithm) and generation, handling of missing values, and partitioning of training and testing datasets, including an over-sampling technique to address class imbalance.

Seven algorithms were applied to create the classification prediction models of a categorical variable – Adherence – constructed from selected health behaviors related to prevention and risk. The models' performance metrics were evaluated and compared (Accuracy, Precision, Recall, weighted average F1 score, and Area Under the Receiver Operating Characteristic Curve One-vs-one) and the best model was identified.

We also used SHAP values to explain the best model's predictions, for a better understanding of the contributions of each feature to the classification process. Participant's country, contact frequency with neighbors, friends, or colleagues, and having visited a doctor or medical facility other than a hospital were the features with the highest impact.

Table of contents

List of figures

List of tables

1 Introduction

The pandemic crisis due to COVID-19 triggered, in most countries, the implementation of epidemic control measures to slow the spread of the virus, namely social distancing, quarantine periods and lockdowns, the adoption of personal hygiene measures such as the use of a mask, frequent hand washing, and respiratory measures, as well as COVID-19 vaccination plans.

The consequences of this pandemic are already well known. Still, its extension over time has revealed additional problems, with social isolation promoting feelings of loneliness, especially among the older population, and negative impacts on the health, care, and subjective well-being of populations (Atzendorf & Gruber, 2021).

Adherence to public health measures, such as mask-wearing, hand hygiene, and social distancing, has shown positive effects in decreasing the spread of the virus, as well as in the total number of deaths (Bonardi et al., 2020; Fischer et al., 2021; Juhn et al., 2021; Szwarcwald et al., 2020). Thus, a deeper understanding of the phenomenon may lead to better health crisis management.

To study this problem, we will use data from the SHARE Project that we briefly describe in the next section.

1.1 The SHARE-ERIC Project

The Survey of Health, Ageing, and Retirement in Europe (SHARE) is a multidisciplinary and cross-national panel database of microdata on health, socio-economic status, and social and family networks of approximately 140,000 participants aged 50 or older in 27 European countries and Israel.

SHARE was created as a response to a Communication by the European Commission calling to "examine the possibility of establishing, in co-operation with the Member States, a European Longitudinal Ageing Survey" (EUR-Lex - 32011D0166 - EN, 2011). Due to its importance, SHARE acquired, in March 2011, a legal status that constitutes itself as the first European Research Infrastructure Consortium (SHARE-ERIC). The project is harmonized with the U.S. Health and Retirement Study (HRS) and the English Longitudinal Study of Aging (ELSA) and has become a model for various research on aging in the world, as collected data from all waves is available to the research community.

Population ageing is a big challenge to societies, as the effects of population ageing include great challenges to pension systems and public policies. SHARE aims to generate research findings on these topics, based on a methodology with three core characteristics: longitudinal, with participants being interviewed every two years; ex-ante harmonization, where all countries use the same questionnaire in the same interview software, on the same schedule; and cross-national, with Wave 8 being held in 28 countries.

The SHARE project is also a large operation. More than 150 scientists around the world design SHARE, about 2000 interviewers conduct the interviews, more than 4000 scientists use SHARE data, and European and national politicians are advised based on SHARE research. Its dimension and impact also imply high-quality standards, as scientific validity and future funding depend on fulfilling those.

SHARE has been studying the life of the 50+ population across Europe for many years, accumulating a wealth of longitudinal data for research and strongly contributing to understanding the ageing process and its life-course determinants. The 8th Wave was planned to continue the project starting in October 2019. However, in March 2020, many European countries were so affected by the new COVID-19 pandemic that the continuity of the fieldwork was questioned, as security conditions could not be guaranteed. On the other hand, many countries started implementing social distancing measures, including lockdowns.

By March 2020, fieldwork was suspended in all countries, with about 70% of the interviews collected. However, stakeholders agreed that data about the health and living situation were extremely valuable to understanding the pandemic's short- and long-term implications. The framework for a new methodology adapted to public health circumstances emerged from two important findings. First, SHARE, as a longitudinal survey, provides an ideal infrastructure to put the implications of the pandemic in its proper context. SHARE's strength is using data about living conditions that were routinely recorded: labor market status, income, family and social contacts, income situation, and health history. Secondly, SHARE decided that this wealth of life-course data should be complemented by new data, collected during and immediately after the first lockdown, which would allow for the understanding of the circumstances experienced by the participants (Scherpenzeel et al., 2020). This study is known as the "SHARE Corona Survey" (SCS1). Due to the social distancing restrictions, SCS1 was conducted between June and September 2020 using telephone-administered interviews (CATI).

Aside from the new collection mode, a special COVID-19 questionnaire was developed, centered on the areas of health and behaviors associated with it, mental health, the effects of COVID-19, adjustments to work and the economy, and modifications to social networks. The key advantage of these new data is the link to the SHARE base panel study with its life course information on previous health conditions and economic and social living circumstances. The consequences of the pandemic, as well as the ones that arise from the health measures implemented, vary intensely with health, economic, and social preconditions. Therefore, this link between the pandemic data and the historical data may allow for a more segmented approach, crucial in the design of public policies, for example.

At the beginning of 2021, Wave 9 began to be prepared. Given the restrictions still imposed and the unpredictable nature of the health crisis, a second round of the SCS was held between June and August 2021 to investigate further the pandemic's impact in the previously studied areas. The questionnaire was revised to reflect new developments in health crisis management, such as vaccination. Data from SCS1 and SCS 2 has provided valuable insights into how the pandemic has affected different areas and contributes to evidence-based decision-making and policy formation.

1.2 Problem definition

Adherence to health behavior, which is also referred to, in some contexts, as compliance, medical adherence, or treatment adherence, is understood as the degree or extent to which a person follows the therapeutic prescription of health professionals (that may be a pharmacological therapy, diet or habits, and lifestyles) (Osterberg & Blaschke, 2005).

Adherence is a concept developed by Stanton in her model of adherence (1987), which implies the active involvement of the patient in the treatment process to produce a therapeutic result, and it is from this perspective that the term will be used (Turk & Meichenbaun, 1991). Despite being often used synonymously with adherence, the term compliance denotes passive behavior and non-participation in treatment; the term compliance, therefore, suggests obedience or even coercion, while adherence suggests conformity, negotiation, and therapeutic alliance (Shaffer and Yoon, 2001).

This phenomenon has been widely investigated in recent decades, especially due to the importance it has acquired in health care. Since the eighties, political and economic consequences of non-adherence have been studied. Therefore, research in this area, as well as the accumulated clinical knowledge on the prevalence of non-adherence, prompted the World Health Organization to create in 2003 the Adherence Project (WHO, 2003) and issue a set of recommendations. This document highlights the magnitude of the phenomenon of non-adherence, as well as the mortality and morbidity associated. It also highlights its economic impact on health systems and, above all, its negative consequences on the results of health indicators. Finally, it has the added value of not placing the burden exclusively on the patient, suggesting that health professionals and healthcare services, along with social, economic, and cultural factors, should be considered.

Adherence is often associated with social support, namely from family and people who are important to the patient, knowledge about treatment or disease process, motivation, and the relationship between the health professional and the patient (WHO, 2003). However, the importance/purpose of the health measures must be clearly present to the patient. Otherwise, non-adherence may occur because it lacks a sense for the patient (not understanding benefits or consequences) or because it becomes difficult to manage (too many measures, the complexity of use). The complexity of the regimen has been cited as the greatest barrier to therapeutic adherence (Martin et al., 2005).

Another factor that has been demonstrated to be a determinant of adherence is the knowledge that the subject has about the measures or the treatment. Older patients, prone to greater deterioration of health status and cognitive functions, often have multiple associated pathologies, making their adherence to health behaviors more complex (Smaje et al., 2018).

Lastly, patients' economic and social situation is also determinant in adherence behavior: support from family and friends, social support, isolation, and economic status may condition adherence (Murray, Morrow & Weiner, 2004).

The outbreak of COVID-19 has caused an unprecedented global health emergency and as such has been the focus of much research into understanding, predicting, and explaining adherence to safety measures – physical distancing, facial mask use, and respiratory measures. Interested in exploring the connection between individual behavior and the spread of disease, Zimmermann et al. (2020) highlighted the impact of human behavior on the spread of the pandemic, arguing that the most effective strategy to mitigate future pandemics is to grasp and understand the human factor and act quickly and, as far as possible, in advance to mitigate it at the beginning of an outbreak.

Overall, it is clear that research plays an integral role in understanding and predicting behaviors during pandemics. Data collected on pandemic policies unveiled some interesting patterns and indicates a high degree of uniformity in government responses to the COVID-19 pandemic in its initial stages (Hale et al., 2021). Although in early March, only a few countries had implemented rigorous containment and health measures, the majority of governments moved to a high level of response within two weeks from the middle of March, and intensive policy responses became widespread. However, after this initial global surge in policy responses, countries started to reduce their restrictions and, in some cases, re-imposed regulations as the epidemics surged and receded. Several recommendations and restrictions changed over time, making it difficult to understand to which degree individual behavior followed the national policy changes, due to many confounding factors. Only by deepening our knowledge of behavioral drivers can we hope to manage future pandemics by better anticipating and mitigating their effects. This underscores the importance of continuing research into understanding, predicting, and explaining human behavior during infectious outbreaks.

1.3 Dissertation framework

There are five key chapters in this dissertation. The literature review in Chapter 2 provides the theoretical framework for our research, on which we will base feature selection and our discussion of the findings. In Chapter 3, we cover ethical issues, the data we'll be using, the development of the target variable, and the pre-processing portion of the project. This includes feature generation and selection, handling missing values, and Train-Test partitioning. Additionally, we discuss the modelling strategy, hyperparametrization, and software. In Chapter 4, we present the results of using machine learning models to predict the target variable and compare their performance. We also present an analysis of the features' importance for the best model. In Chapter 5, which concludes this dissertation, we discuss the results while providing key conclusions.

2 Literature Review

In this section, we will explore relevant topics in framing the theme of the future dissertation. First, we describe the results of a review that was carried out on predictors of adherence during the COVID-19 pandemic. Following this more general topic analysis, we focused on a specific review of studies using machine learning approaches, presented in the second part of this section.

2.1 Predicting Adherence in the COVID-19 Pandemic

Recent studies have shown that adherence to public health measures such as social distancing and hygiene seems to depend on the traditional medication or health measures compliance predictors.

2.1.1 Social Determinants of Health

Growing evidence shows that socially disadvantaged populations face multiple barriers to healthy living, including limited capacity to engage in COVID-19 risk mitigation approaches optimally. Therefore, Social Determinants of Health (SDH) (WHO, 2008, 2018), such as income and social protection, education, unemployment, job insecurity, working life conditions, housing, and access to affordable health services of decent quality are still reported as a major predictor of compliance (Singu et al., 2020). Evidence shows that those with the lowest household income were less able and willing to self-isolate (Alkhaldi et al., 2021) and that the ability to adopt and comply with certain non-pharmaceutical interventions (NPIs) is lower in the most economically disadvantaged in society (Atchison et al., 2021).

2.1.2 Demographic variables

The most recent research shows that socio-demographic variables such as gender continue to predict adherence. Females were more likely than males to claim they were able to self-isolate, according to the COVID-19 survey conducted in Canada (Atchison et al., 2020), and a significant relationship between being female and the adoption of new health behaviors has also been found in the U.K. (Mant et al., 2021). Another study (Rayani et al., 2021) also showed that females stayed more at home, wore a mask, avoided party gatherings, and washed their hands more frequently than males. Social distancing and hygiene behaviors showed sub-optimal adherence, particularly for men and younger adults, in an American study of stress responses to the pandemic (Park et al., 2020), and men observed physical distancing less frequently than women (Fodjo et al., 2021). Lastly, in a Brazilian study aimed at analyzing the population's adherence to measures to restrict physical contact and the spread of the COVID-19 virus, it was identified that the group that did not adhere to the measures was composed mainly of men aged between 30 and 49 years, with a low level of education and who worked during the pandemic (Szwarcwald et al., 2020). A large pan-European study conducted in the early days of the pandemic also showed that adherence to preventive measures was higher among older, female, and highly educated respondents (Varghese et al., 2021).

2.1.3 Personality traits

Adherence to containment measures also seems to depend on individual factors, including personality traits. Evidence indicates that people with high levels of empathy tend to adhere more to public health measures (Carvalho & Machado, 2020). A study conducted using the Big Five personality traits (Bogg & Milad, 2020) also concluded that guideline adherence can be explained by individual differences in personality traits: conscientiousness ("to be reliable") and openness (more prone to "positive attitudes" towards the guidelines) were associated with superior guideline adherence.

2.1.4 Risk awereness

According to recent studies, older people, as well as those who had once been tested for COVID-19 (mostly contacts/exposed individuals), had higher odds of adhering to the preventive measures. A six-month Online National Survey in Cameroon by Fodjo and colleagues (2021) showed that these participants were keener to commit to rigorous COVID-19 preventive behaviors. Their results suggest that this happened because they were more aware that they stood a higher risk of becoming infected and, consequently, suffering greater consequences, such as becoming severely ill. Rayani et al. (2021) also linked perceived susceptibility and perceived severity with preventive behaviors in an online sample of Iranian students.

2.1.5 Information

Obtaining COVID-19 information from healthcare workers was also significantly associated with higher adherence levels (Ahmed et al., 2020). Knowledge of social distancing restrictions predicted intentions to adhere in specific situations, positive attitudes towards current restrictions, and a greater perceived ability to adhere (Sturman et al., 2020). Similar findings were revealed in a sample of pregnant women in Ghana (Apanga & Kumbeni, 2021). A multivariable logistic regression analysis revealed that knowledge of COVID-19 symptoms and transmission was associated with adherence to wearing a face mask. Also, women who knew that avoiding touching their eyes, nose, and mouth could prevent COVID-19 and that the virus could be transmitted by touching contaminated objects/surfaces presented increased handwashing and sanitizing and greater adherence to social distancing. A correlation between preventive behaviors and health information-seeking behaviors was also found in another study (Rayani et al., 2021). Individuals who actively seek out health information are more likely to engage in health-promoting behaviors due to a better understanding of the health risks associated with their behavior and the potential benefits of making healthier choices.

2.1.6 Beliefs

Beliefs regarding the prevalence of COVID-19 were also shown to predict behavior. People's individual beliefs about the severity of the virus, their susceptibility, and the efficacy of protective measures contribute to their health choices. The higher the believed fraction of the infected population, the more likely people were to report mask-wearing. This implies that the role of fear, and the beliefs/perceptions about COVID-19 prevalence, are important factors affecting self-protecting behaviors (van den Broek-Altenburg & Atherly, 2020). Also, illness perceptions toward COVID-19 significantly affected adherence to precautionary measures (Chong et al., 2020).

Nonetheless, the research by Bogg and Milad (2020) with components of the Health Belief Model and controlling for other variables such as demographics and personality traits (as referred before), showed that individuals who perceived greater risk of exposure or greater perceived health consequences weren't more likely to follow the guidelines. Individuals may base their perceptions of the severity of a health threat or of the benefits and barriers on the information they receive in information from different sources such as the media, family, friends, and health authorities. This variety of sources can provide a range of messages that may be conflicting (Nagler et al., 2020).

2.1.7 Clinical variables

Even in special populations (with chronic diseases), social determinants of health were associated with COVID-19 risk mitigation practices. This was found true for a population of adults with Cardiovascular Disease (Hagan et al., 2021), where individuals with a greater burden were found to practice more personal protection and social distancing. Logistic regression analysis showed, for a population of young adults with asthma (Vázquez-Nava et al., 2020), that being male, actively smoking, and believing that COVID-19 was not more dangerous for asthma patients was associated with non-adherence to all the basic preventive measures for COVID-19.

2.2 Machine Learning approaches

Although there is various research using machine learning in the context of the COVID-19 pandemic, from vaccine design (Ong et al., 2020) to predicting infected patients prognosis (Lopes et al., 2021) or the effectiveness of health measures (Lafzi et al., 2021), very limited research in adherence has used machine learning techniques.

Roma and colleagues (2020) tried to explain compliance with protective health measures using a Moderated Mediation Model and Machine Learning algorithms. They concluded that ML classification models' outcomes showed that the psychological and psychosocial variables considered could predict which individuals have high versus low compliance. Nevertheless, they also point out some limitations: the cross-sectional study design prevents drawing causal inferences. Individuals' psychological functioning before the virus spread could not be assessed. Furthermore, the data collection via a web-based survey relied on voluntary sampling and self-reported data; thus, the data may be distorted by selection or social desirability biases. One of the major strengths of using SHARE data is the link to previously collected data and the robust sampling of this survey.

Recently, Bailey and colleagues (2021) aimed to identify latent variables underlying adherence to COVID-19 guidelines and to examine demographic and psychological predictors of adherence. For this purpose, elastic net regression was used. Sixty-four demographic and psychological factors were analyzed, including emotion regulation skills and coping strategies. The authors debate that there are several advantages of Elastic Net over standard multivariate regression approaches, as this regularized regression method performs both regularization, by penalizing coefficient estimates, and variable selection, by decreasing the coefficients of irrelevant predictors to zero, dropping them out of the model (Zou & Hastie, 2005). This procedure also helps to prevent overfitting. The authors were able to identify demographic and psychological predictors for two forms of adherence: avoidance and cleaning. Less avoidance adherence is linked to religious affiliation, denial coping, full-time employment, substance use coping, and being 60 or older. On the other hand, behavioral and mindfulness emotion regulation skills, agreeableness, and democrat political affiliation predicted greater avoidance adherence. Cleaning adherence can be predicted by interpersonal and behavioral emotion regulation skills and conscientiousness.

Annual influenza vaccination is an important public health measure to prevent influenza infections and is strongly recommended for cardiovascular disease (CVD) patients, especially in the current coronavirus disease 2019 (COVID-19) pandemic. Therefore, Kim and colleagues (2021) conducted a study that aimed to develop a machine learning model to identify Korean adult CVD patients with low adherence to influenza vaccination. Separate models for participants under and over 65 years old were developed, as influenza immunization is free for the latter. The classification process was performed using logistic regression (LR), random forest (RF), support vector machine (SVM), and extreme gradient boosting (XGB) machine learning techniques. The models show comparable performance in classifying adult CVD patients with low adherence to influenza vaccination, with acceptable accuracy. For older patients, sex and insurance type significantly affect adherence to vaccination. For the younger participants, age and health status are the predictors.

Singh et al. (2022) used Natural Language Processing (NLP) and the Latent Dirichlet Allocation (LDA) modeling technique to study the behavioral response of agriculture stakeholders during the COVID-19 pandemic by extracting the most dominant topics from agriculture-related tweets from Twitter. Machine-learning-based methods were used to analyze the sentiments, emotions, and views of agriculture agents. Three phases of the lockdown measures were analyzed and showed that the early stages of the lockdown elicited a great amount of distress, indicating the widespread fear of insecurity felt. However, a decrease in distress was observed in the subsequent stage of the lock-down. Policymakers could use these findings to gain insights into the behavioral reactions of agricultural stakeholders and effectively initiate preventive measures to address similar issues in the future.

3 Methods

In this section, we describe the methodological elements of the project. We will address ethical concerns, describe the data we will use, the construction of the target variable, as well as the pre-processing stage, involving feature selection and generation, treatment of missing values, and Train-Test partitioning. We also present the modelling approach, the applied algorithms, hyperparametrization and the used software.

3.1 Ethics

SHARE data (SHARE Corona Survey and Easy SHARE) is anonymized and the participants' data match is only possible through an individual ID code. Access to data is granted for scientific use upon registration, subject to European Union and national data protection laws. The access also meets the requirements of the European Charter for Access to Research Infrastructures.

3.2 Data

This work uses the "SHARE Corona Survey" (SCS1 and SCS2) data as the main data source (Börsch-Supan, 2013, 2022a, 2022b). Additionally, data from previous waves (Easy SHARE) was used regarding personality traits, health behaviors, economic status, and general demographic characterization.

SHARE focuses on people 50 or older who live in the country where the study is being conducted. However, individuals in jail, hospitalized, out of the country during the entire survey period, or unable to speak the local language(s) are excluded from the study. Those who were interviewed in a previous wave are part of the longitudinal sample, and if they have a new partner living with them, that partner is also eligible for an interview. The study tracks and re-interviews participants who move within the country (but not abroad) and conducts end-of-life interviews for those who pass away. However, younger partners, new partners, and those who have never participated in SHARE will not be tracked if they move, nor are they eligible for end-of-life interviews.

Due to COVID-19 restrictions, interviews were conducted telephonically, using Computer Assisted Telephone Interview (CATI), unlike regular SHARE panel interviews, which are face-to-face in the respondents' home or living facility, using Computer Assisted Personal Interview (CAPI) (Scherpenzeel et al., 2020). SCS1 was conducted between June and September 2020, and SCS2 was collected between June and August 2021.

An adaptation of the sample design was also needed for the first round. Therefore, a sample was selected in each country, including two types of participants: panel members who had not been interviewed before the suspension of fieldwork and panel members who had already been interviewed in Wave 8 face-to-face interviews. In countries where fieldwork had not started (as with Portugal), all panel participants with phone numbers were interviewed. The second round included, for all countries, all the participants with a valid phone number.

The dataset has 49,253 participants in SHARE's Wave 9 SCS2 from 28 countries aged 50 to 107 (Figure 1).

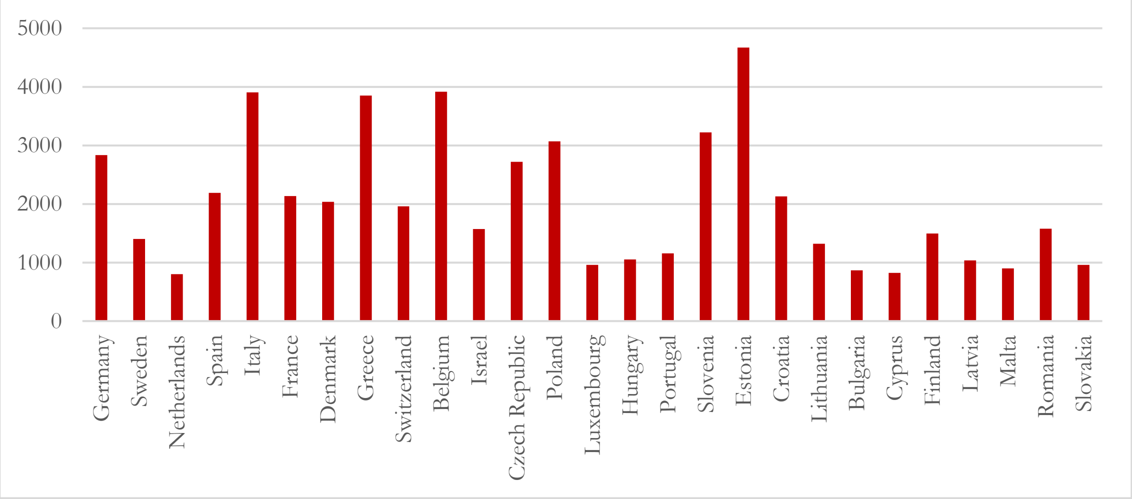


Figure 1 - Participants in Wave 9 by country

The dataset contains four ID variables that allow the link between this information and the panel datasets and seven characterization/confirmation variables.

COVID-19 specific variables related to safety measures adopted, health status before and since the outbreak, including mental health, COVID-19 related symptoms, hospitalizations, testing and deaths (self, family, friends), healthcare appointments forgone, postponed and denied, medical treatment and satisfaction with treatments, changes in work and economic situation, such as workload, place of work, internet connection, computer skills, unemployment or lay-off, economic variables such as financial support from employer or government, financial support to or from other family members, household's ability to make ends meet since the outbreak and, lastly, social support variables as changes in personal contacts with family and friends, help given and received, personal care given and received and volunteering.

In addition to the COVID-19 specific variables, panel variables were included in the analysis through EasySHARE. These included Demographic variables, Physical health, ADL (activities of daily living), Psychological variables, Financial/Income, and Personality traits.

3.3 Target variable

The COVID-19 questionnaire evaluates safety measures, which include 12 health behaviors that we selected, as they translate health behaviors related to prevention and risk and could, therefore, be used as a measure of adherence to recommendations.

Variables related to travel - cah111_11 “Health: used public transportation in the last three months” and cac142 “Health: traveled abroad for more than 48h since outbreak” – weren’t used as most participants weren’t traveling, and there were very different travel restrictions among countries.

The selected variables are evaluated in distinct ways (Table 1), such as 3 and 4-point Likert scales, categorical or dichotomic (Yes/No).

Table 1 – Measurement level of the 12 health behaviors

Variable	Description	Measurement Level
cah110_	Health: ever left home during the last 3 months	Dichotomic (Yes/No)
cah111_3	Health: met more than 5 people outside household during last 3 months	Ordinal: 4-point Likert scale (1 – Several times a week, 2 – About once a week, 3 – Less than once a week, 4 – Not at all)
cah111_6	Health: went shopping during the last 3 months	Ordinal: 4-point Likert scale (1 – Several times a week, 2 – About once a week, 3 – Less than once a week, 4 – Not at all)
cah111_7	Health: went to post office/bank/public office during the last 3 months	Ordinal: 4-point Likert scale (1 – Several times a week, 2 – About once a week, 3 – Less than once a week, 4 – Not at all)
cah111_8	Health: went to restaurant/pub during the last 3 months	Ordinal: 4-point Likert scale (1 – Several times a week, 2 – About once a week, 3 – Less than once a week, 4 – Not at all)
cah113_	Health: kept distance from others in public during the last 3 months	Ordinal: 4-point Likert scale (1 – Always, 2 – Often, 3 – Sometimes, 4 – Never)
cah116_	Health: covered cough/sneeze more during last 3 months compared to first wave	Ordinal: 3-point Likert scale (1 – More frequently, 2 – About the same, 3 – Less frequently)
cah017_	Health: took drugs or medicine as prevention against COVID-19	Dichotomic (Yes/No)
cahc117_	Health: has been vaccinated against Covid-19	Dichotomic (Yes/No)
cahc118_	Health: wants to get vaccinated against Covid-19	Categorical (1 – Yes, I already have a vaccination scheduled, 2 – Yes, I want to get vaccinated, 3 – No, I do not want to get vaccinated, 4 – I am still undecided)
cahc884_	Health: got flu vaccination in last 12 months	Dichotomic (Yes/No)
cahc119_	Health: had pneumonia vaccination within last 6 years	Dichotomic (Yes/No)

Even among the ordinal variables, distribution is very different, regarding skewness and kurtosis, and missing values (Table 2). Recoding of some of the answer codes (due to questionnaire routing) was necessary to obtain true missing values counting.

Table 2 - Descriptive statistics for ordinal health behaviors

	N	MV	% MV	Min	Max	M	STE	Me	Mo	SD	CV	Sk	SE Sk	Ku	SE Ku
cah111_3	49110	143	0.29%	1	4	2.91	0.005	3	4	1.13	38.96%	-0.569	0.011	-1.119	0.022
cah111_6	49213	40	0.08%	1	4	2.07	0.005	2	1	1.09	52.59%	0.639	0.011	-0.916	0.022
cah111_7	49185	68	0.14%	1	4	3.33	0.003	3	4	0.73	22.02%	-0.950	0.011	0.639	0.022
cah111_8	49110	143	0.29%	1	4	3.50	0.004	4	4	0.80	22.87%	-1.623	0.011	1.899	0.022
cah113_	49093	160	0.33%	1	4	1.71	0.005	1	1	1.04	60.70%	1.263	0.011	0.212	0.022
cah116_	48776	477	0.98%	1	3	1.83	0.002	2	2	0.48	26.40%	-0.408	0.011	0.485	0.022

N: Valid; MV: Missing values; Min: Minimum; Max: Maximum; M: Mean; STE: Standard Error Mean; Me: Median; Mo: Mode; SD: Standard Deviation; CV: Coefficient of variation; Sk: Skewness; SE Sk: Standard Error Skewness; Ku: Kurtosis; SE Ku: Standard Error Kurtosis.

After preliminary data analysis of the 12 variables these were dichotomized and recoded as a risk behavior (0) or an adherence behavior (1). The table in Appendix I summarizes the recoded values.

As countries were in different stages of vaccination, some with age group tiers, a new variable was computed by the recoding of the variables cahc117_ Health: has been vaccinated against Covid-19 and cahc118_ Health: wants to get vaccinated against Covid-19, coding as risk if someone wasn't vaccinated and didn't want to and as adherence if the person was vaccinated or wanted to be vaccinated.

The remaining 11 variables (Table 3) show different distribution of the participants regarding risk or adherence.

Table 3 – Results of the recode as risk or adherence

	Risk [0]		Adherence [1]	
	n	%	n	%
cah110__D (ever left home during the last 3 months)	44409	90.2%	4844	9.8%
cah111_3_D (met more than 5 people outside household during last 3 months)	21158	43.0%	28095	57.0%
cah111_6_D (went shopping during the last 3 months)	24115	49.0%	25138	51.0%
cah111_7_D (went to post office/bank/public office during the last 3 months)	6039	12.3%	43214	87.7%
cah111_8_D (went to restaurant/pub during the last 3 months)	10665	21.7%	38588	78.3%
cah113__D (kept distance from others in public during the last 3 months)	19552	39.7%	29701	60.3%
cah116__D (covered cough/sneeze more during last 3 months compared to first wave)	2733	5.5%	46520	94.5%
cah017__D (took drugs or medicine as prevention against COVID-19)	43095	87.5%	6158	12.5%
cahc884__D (got flu vaccination in last 12 months)	30265	61.4%	18988	38.6%
cahc119__D (had pneumonia vaccination within last 6 years)	42952	87.2%	6301	12.8%
cahc117_cahc118_D (wants to get vaccinated against Covid-19)	7544	15.3%	41709	84.7%

Most variables show small to no association, with some exceptions (Figure 2). Individuals who went to post office/bank/public office during the last 3 months or went to restaurant/pub during the last 3 months also reported meeting more than 5 people outside household during last 3 months (r=.754, r=.610, respectively). Going to post office/bank/public office is also associated with going shopping (r=.793), going to restaurant/pub (r=.876), but also with keeping distance from others in public during the last 3 months (r=.706). Lastly, getting the flu vaccination is positively correlated with having pneumonia vaccination (r=.647) and wanting to get vaccinated against Covid-19 (r=.601). Two variables presented only negative relation with the others: ever left home during the last 3 months and took drugs or medicine as prevention against COVID-19. Both are dichotomic and unbalanced, with 90% and 80% of the participants, respectively. The first was also used as a filter question for the subsequent ones, automatically coding the next answers. For this reasons, we decided not to use these two variables.

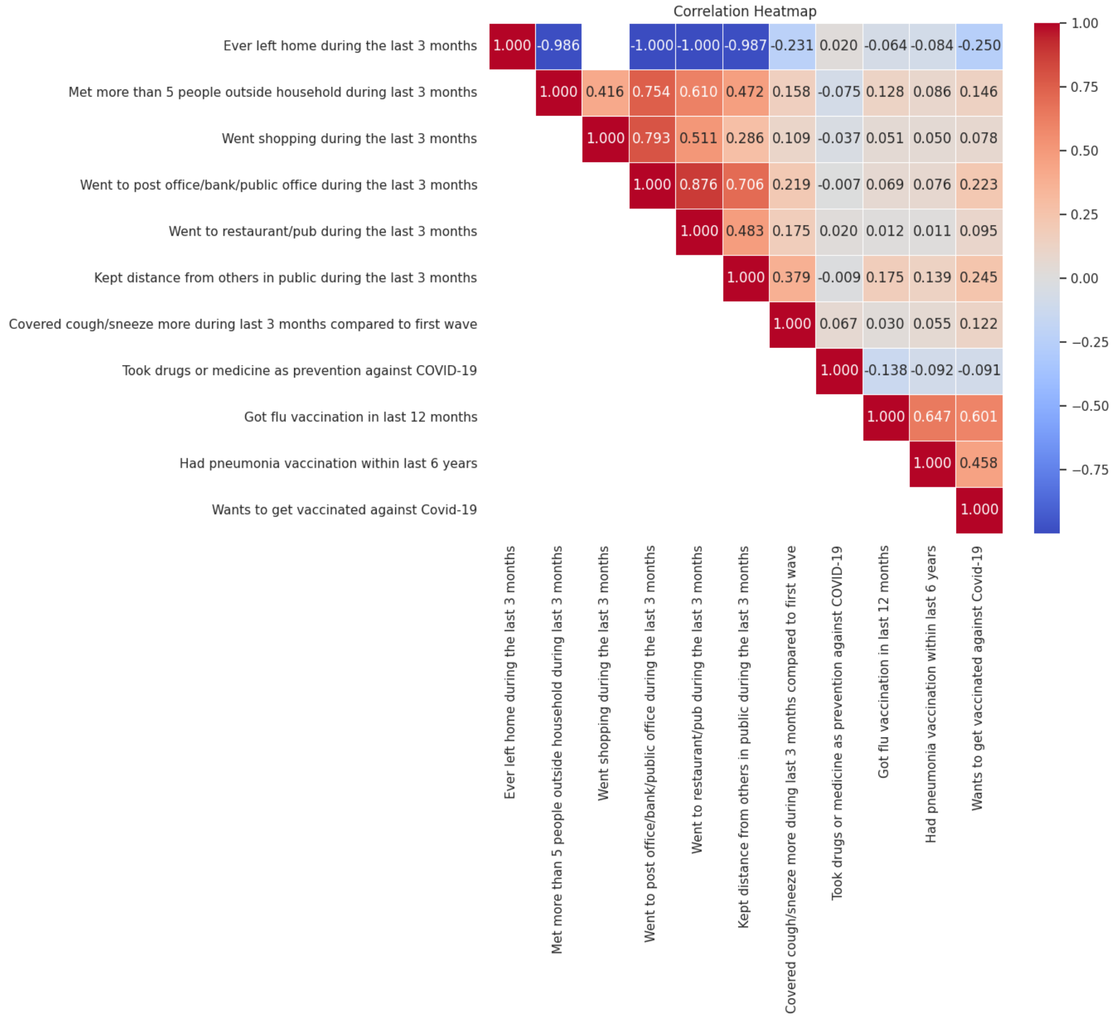


Figure 2 – Heatmap for Polychoric Correlations

This data analysis led us to create a composite measure of these nine behaviors to assess, as our outcome, which seems to be also the most usual option in the reviewed literature (Ahmed et al., 2022; Bailey et al., 2021; Bogg & Milad, 2020; Carvalho & Machado, 2020; Chong et al., 2020; Hagan et al., 2021).

A Categorical Principal Components (CATPCA) with oblique rotation and Kaiser Normalization was performed using the 9 variables, in order to identify the underlying patterns and interrelationships among the variables. The analysis extracted two components (using Kaiser's criterion), which explain only 44% of the total variance (Table 4).

Table 4 - CATPCA with oblique rotation: model summary

Dimension	Cronbach's Alpha	Total (Eigenvalue)	% of Variance	Rotation
1	.670	2.472	27.465	5.911
2	.371	1.493	16.588	3.343
Total	.841	3.965	44.053	

The first component (Comp 1) explains 27% of the variance and was characterized by high positive loadings on variables related to social behavior. The second component (Comp 2) explains 17% of the variance and was characterized by high positive loadings on variables related to vaccination. Cronbach's Alpha is acceptable for the first component (.670), but poor for the second (.371) (Hair et al., 2022).

The biplot (Figure 3) of the two components shows that the data points are clustered in the left lower quadrant with a seeming pattern of 4 lines.

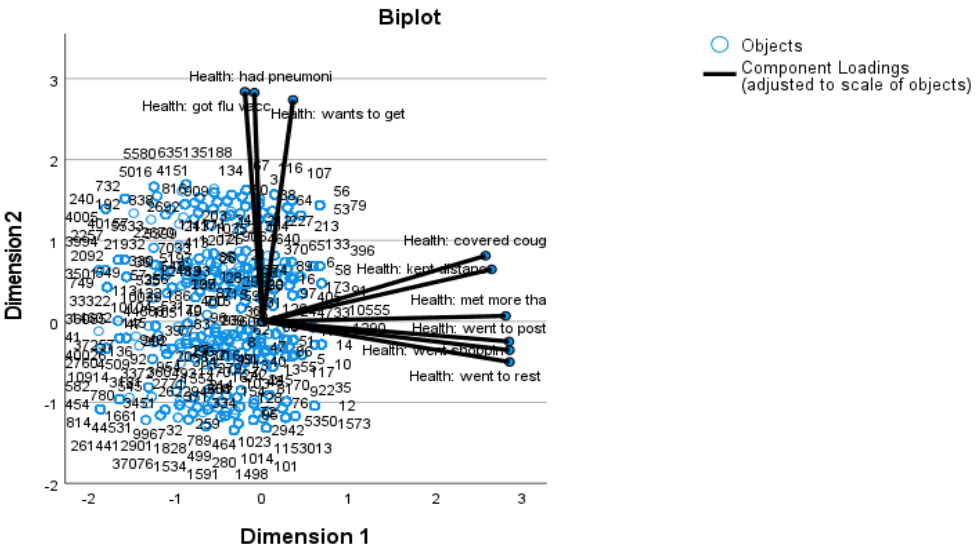


Figure 3 – Biplot representing the two components

Overall, the results of the CATPCA suggest a solution that retains two components but explains only 44% of the total variance. Given the loss of variance explained, we decided to proceed with the items.

Internal consistency of the nine items was assessed using the Kuder-Richardson Formula 20 (KR-20). The elimination of item cah116__D Health (covered cough/sneeze more during the last three months compared to first wave) increased the score, so we proceeded without it. The final KR-20 score of .629 indicates acceptable internal consistency among the items, suggesting that these items measure the same construct and are reliable, reason why we decided to continue the analysis, considering these could be computed into a one-dimensional measure of adherence.

A new variable was computed using the sum of the 8 remaining behaviors (Table 5).

Table 5 - Descriptive statistics for the computed adherence variable

	N	MV	Min	Max	M	STE	Me	Mo	SD	VC	Sk	SE Sk	Ku	SE Ku
Adherence	49253	0	0	8	4.70	0.008	5	5	1.82	38.69%	-0.596	0.011	-0.027	0.022

N: Valid; MV: Missing values; Min: Minimum; Max: Maximum; M: Mean; STE: Standard Error Mean; Me: Median; Mo: Mode; SD: Standard Deviation; VC: Variance Coefficient; Sk: Skewness; SE Sk: Standard Error Skewness; Ku: Kurtosis; SE Ku: Standard Error Kurtosis.

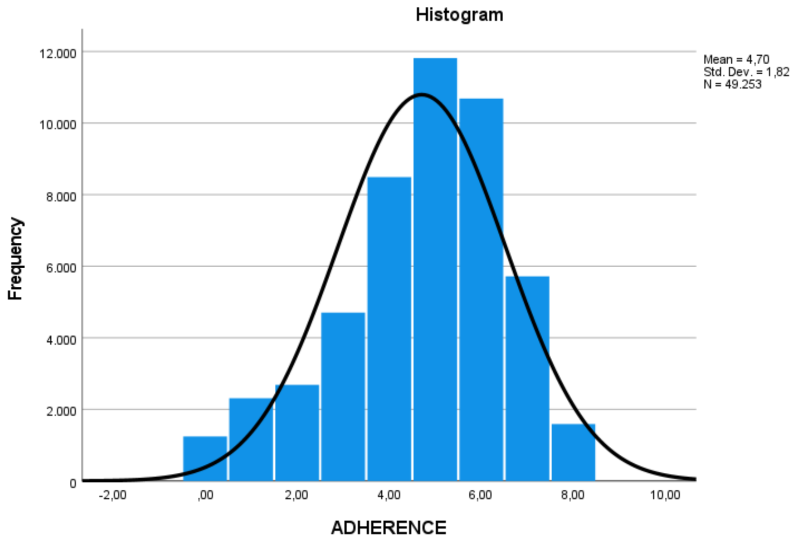
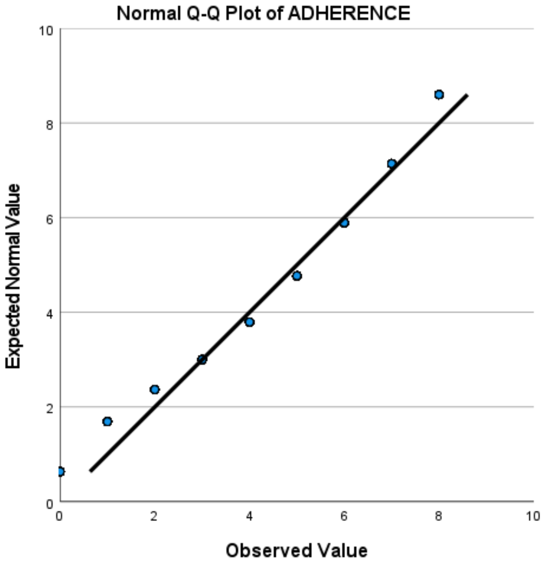


Figure 4 – Histogram of the computed adherence variable

In addition to visual inspection of the shape of the distribution (Figure 4) and analysis of skewness and kurtosis values, a Q-Q Plot, the most commonly used graphical test for normality (Sharma, 1996), is presented. Although, visually, it seems to have an adequate fit when compared to the normal distribution (Figure 5), the Kolmogorov-Smirnov normality test confirmed that it doesn’t follow a normal distribution, D(49253) = 0.17, p < 0.001.

Figure 5 – Q-Q Plot of the computed adherence variable



We then investigated the relation of this variable with COVID-19 infection, using the variable cac105_1 COVID-19: respondent tested positive, recoded so that the missing values from refusals to answer, not applicable and don't know would be considered as not having tested positive.

A chi-square test of independence was performed to evaluate the relationship between Adherence and testing positive for COVID-19. The relationship between these variables was significant, $X^2_{(8,\ n=49253)}$=114.96, p<.001. Those with higher values of Adherence (6 to 8) were less likely to have tested positive for COVID-19, while those with lower scores of Adherence (2 to 4) were more likely to have tested positive.

By discretization, we intended to transform this variable into a categorical one. Several methods for discretizing a continuous variable can be used, but we decided to evaluate two, as we believed that categories interpretation would be easier: equal width binning, dividing the range of the continuous variable into equal-width bins or intervals, and K-means clustering, grouping data into groups based on their similarity (Han et al., 2011).

To discretize a continuous variable with 9 levels (0 to 8) using equal-width binning, the range of the variable is divided into equal-width intervals or bins. We divided the range into 3 bins, which seemed to be the better interpretable solution: 0 thru 2 (13%), 3 thru 5 (51%) and 6 thru 8 (37%). This new variable (Adherence_3) also presents a significant relation with having tested positive for COVID-19, $X^2_{\text{(2, n=49253)}}$=70.248, p<.001, with participants of the third level (more adherence measures) being less likely to have tested positive and the ones in the second group being more likely.

In k-means clustering, the number of clusters k is determined based on the research question and the desired level of granularity and we tested solutions from 2 to 4 clusters. Due to interpretability and size (number of cases in each cluster), the solution of 3 clusters was chosen: C1 – M=4.28; C2 – M=1.23; C3 – M=6.49. As with the equal width binning variable the three clusters also present a significant relation with having tested positive for COVID-19, X^2 (2, 49253)=70.218, p<.001, with participants of C3 (higher adherence mean) being less likely to have tested positive and the ones in C1 being more likely.

With both strategies producing very similar results, we opted for the most straightforward and reproducible, proceeding to modelling with a 3-level adherence measure computed through equal-width binning.

3.4 Pre-processing

Due to the dimension and characteristics of the dataset, pre-processing involved three main steps aimed at transforming the original data into a format suitable for modeling: feature selection and generation, handling of missing values, and composition of training and testing datasets.

As underscored by Zhang et al. (2010), pre-processing plays an essential role in rectifying noise, irregularities, and redundancies in data. This process ensures data integrity, establishing a dependable foundation for further analyses, enhancing algorithmic efficiency (Mitchell, 1997).

Subsequently, we detail the approach taken for feature selection, employing the Maximum Relevance Minimum Redundancy (MRMR) algorithm, which serves as a discriminating filter in identifying the most salient features.

Addressing missing values, we will elaborate on elimination and strategic imputation. This involved the removal of observations and variables when appropriate, followed by testing and applying imputation techniques on the remaining data points.

Lastly, we will describe partitioning the dataset into training and testing subsets as the last step in the pre-processing pipeline. We explain the partitioning, along with the Synthetic Minority Over-sampling Technique (SMOTE), as a means to rectify class imbalance, effectively augmenting the dataset through synthetic instances and strengthening the robustness of subsequent modeling.

3.4.1 Feature selection and generation

The dataset included 587 variables, of which 222 were COVID-specific variables, posing a challenge in data dimensionality and relevance. Although each variable held theoretical relevance (because only theoretically relevant variables of the categories presented in the literature review were included), some were excessively specific or redundant. Consequently, an essential step of data refinement involved the aggregation of variables, creating new variables that encapsulated the essential information of their original counterparts (Appendix II). This process involved the dichotomization of each variable (0 – not present / 1 – present) and the computation of the new variable through the sum. Descriptive statistics for the new variables are presented in Table 6.

Table 6 - Descriptive statistics for the new computed variables

	N	Min	Max	M	STE	Me	Mo	Sk	SE Sk	Ku	SE Ku
CAH004_T Health conditions (sum)	49253	0	7	1.38	1.144	1	1	0.669	0.011	0.100	0.022
CAPH089_T Health - fragility (sum)	49253	0	4	0.87	1.138	0	0	1.214	0.011	0.527	0.022
CAH007_T Health - drugs (sum)	49253	0	7	1.39	1.329	1	0	0.853	0.011	0.247	0.022
CAC103_T COVID-19 - relatives had symptoms (sum)	49253	0	9	0.61	0.938	0	0	1.850	0.011	3.987	0.022
CAC105_T COVID-19 - relatives tested positive (sum)	49253	0	9	0.57	0.906	0	0	1.938	0.011	4.719	0.022
CAC120_T COVID-19 - symptoms attributed to Covid-19 (sum)	49253	0	9	0.19	0.876	0	0	5.481	0.011	32.527	0.022
CAC111_T COVID-19 - relatives hospitalized (sum)	49253	0	7	0.13	0.378	0	0	3.067	0.011	11.891	0.022

A strategic curation of variables was executed to further reduce the number of features. In instances where a broader, overarching variable existed, the associated specific sub-variables were excluded. For example, the variable "received financial support due to outbreak" was retained, while the sub-variables detailing the source of support were omitted. Similarly, the most granular variables within the same domain were removed. While variables like "children/parents/relatives/neighbors tested positive/were hospitalized" were retained, the more detailed variables such as "number of children/parents/relatives/neighbors tested positive/were hospitalized" were discarded. This step streamlined the dataset while preserving meaningful information.

Moreover, variables with substantial proportion of missing values, exceeding 50%, were excluded from the dataset. This strategic elimination ensured that only variables with comprehensive data were considered, upholding the analytical integrity of subsequent steps.

Following these preparatory steps, the dataset was effectively reduced to 62 variables, categorized according to predictor types (according to the literature review). Including the 7 new engineered variables, the dataset comprised 69 attributes. To refine this feature set further and identify the most informative variables, the Maximum Relevance Minimum Redundancy (MRMR) algorithm was employed.

MRMR is an automated feature selection algorithm that gauges the importance of features by assessing their relevance to the target variable while concurrently minimizing inter-feature redundancy (Peng et al., 2005). The algorithm operates in two phases: relevance computation and redundancy reduction. It scores each feature based on its relevance to the target variable, seeking to maximize the discriminative power of selected attributes. Additionally, MRMR scrutinizes the relationships between features, omitting those that provide redundant or duplicate information (Ding & Peng, 2005). The upside of MRMR lies in its capacity to systematically capture both the individual importance and collective synergy of features. It mitigates the risk of overfitting by promoting the inclusion of pertinent but uncorrelated attributes, expediting the feature selection process. However, MRMR may encounter challenges when dealing with highly correlated features or when presented with complex nonlinear relationships.

Upon applying MRMR to the dataset, the feature set was pruned, leading to a selection of 30 variables (Appendix III). This reduction streamlines subsequent analyses and ensures that the selected attributes are optimally informative and independent.

3.4.2 Missing values analysis and imputation

We addressed incomplete participant information by reducing the dataset to 30 variables, each with less than 26% missing values. Following Hair's (1998) guideline of excluding participants with more than 10% missing values, we filtered out such cases. Consequently, no variable harbored more than 6% missing values. However, recognizing that certain algorithms disallow any missing values, necessitating imputation (Gama et al., 2015), we proceeded with further analysis.

To comprehensively address missing values, the distribution of the absent data was assessed using the pyampute package (Schouten et al., 2022), specifically pyampute.exploration module, that includes mdPatterns, which displays unique missing data patterns, and MCARTest, which performs a statistical hypothesis test to evaluate whether it is likely that missing data has a Missing Completely At Random (MCAR) behavior.

Though statistically significant, an initial Little's MCAR test encountered challenges in elucidating patterns within the missing values (Figure 6).

The resulting visualization shows the missing data in red and the observed data in blue. The y axis on the left displays the count of rows that follow a pattern and the y axis in the right displays the number of missing values per pattern. The first row displays the data rows with no missing values.

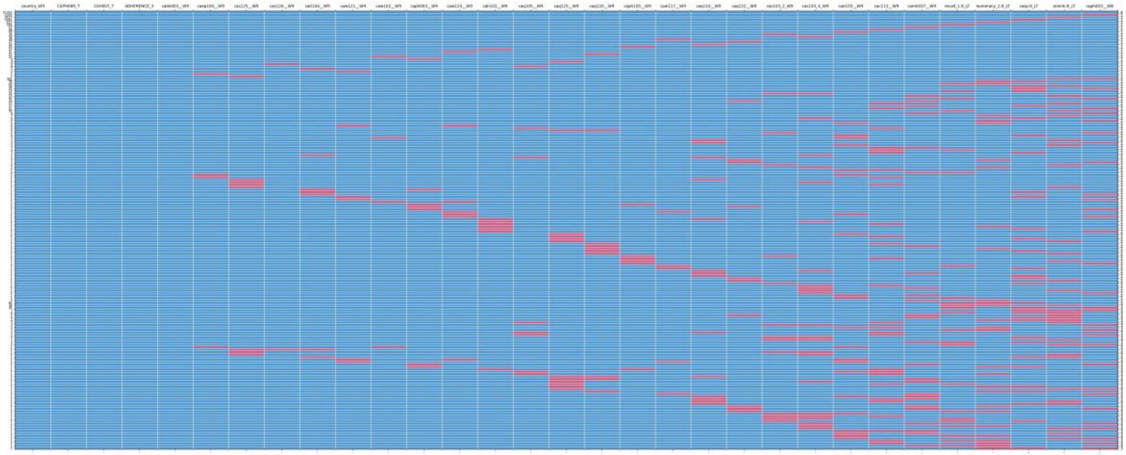


Figure 6 – Missing values pattern

A more focused examination was conducted exclusively on the five variables with over 1% missing values to enhance interpretability. The missing values are not completely at random, but the pattern results and plot (Figure 7) revealed that there were only 24 individuals, at most, following a specific pattern, and there were no patters with more than three missing values.

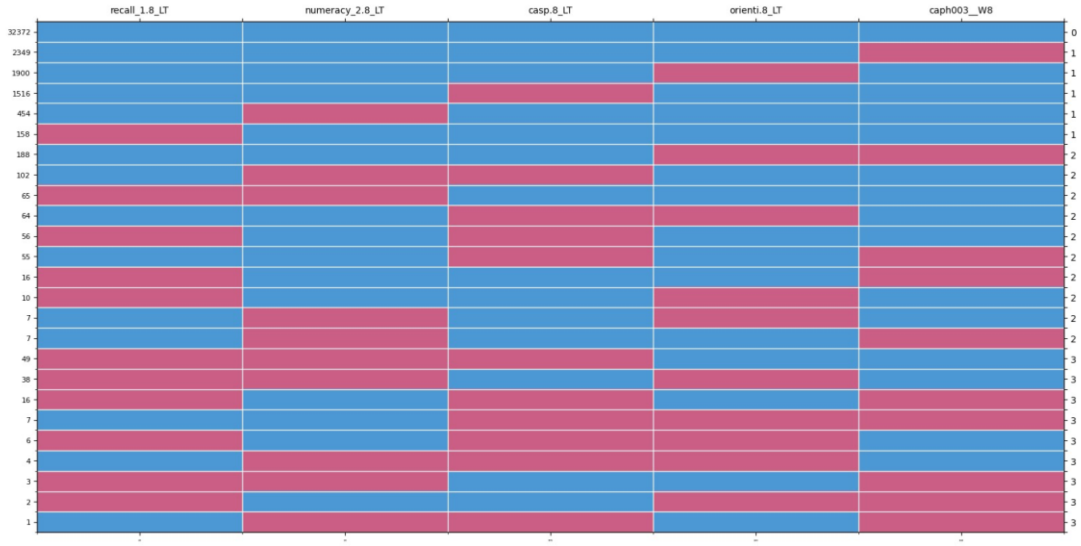


Figure 7 - Missing values pattern for variables with more than 1% missing values.

Considering these findings alongside the nature of remaining variables and their distributions, five imputation methods were tested, using a Decision Tree to compare performance: zero imputation, mode imputation, k-Nearest Neighbors (kNN) imputation, and the iterative imputation (sklearn.impute.IterativeImputer), with and without standardization (Figure 8).

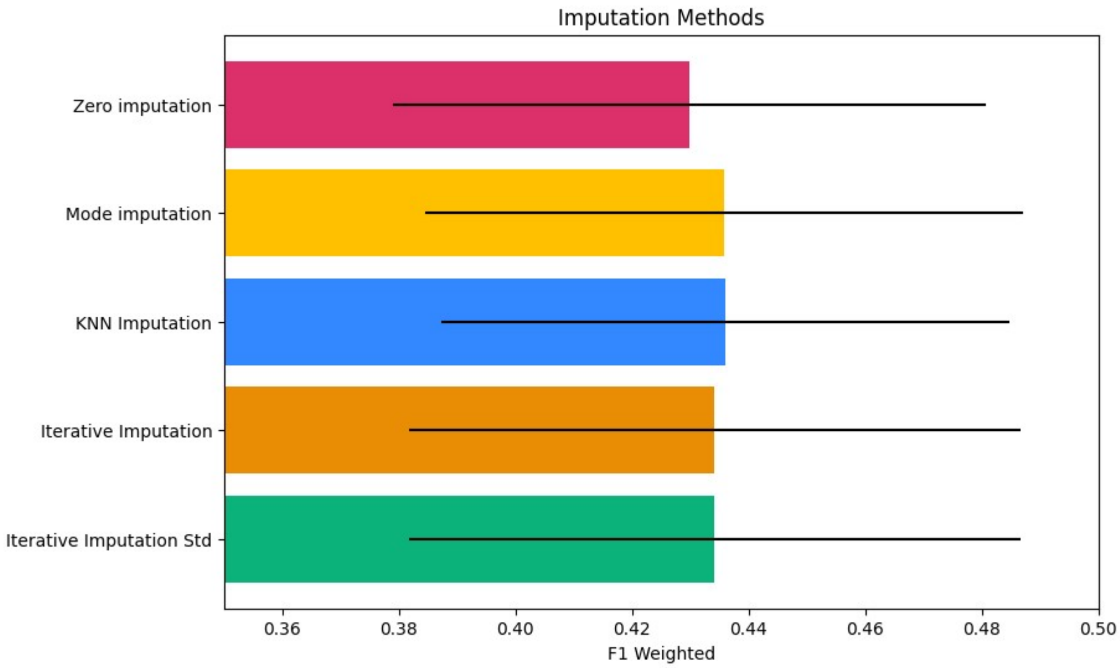


Figure 8 – Weighted average F1 score for the five tested imputation methods: mean and standard deviation

The weighted average F1 score was selected as the evaluation metric. This metric calculates the F1 score for each class separately and then takes a weighted average of these F1 scores, where the proportion of instances in each class determines the weights. Classes with more instances contribute more to the weighted average.

Given the minimal proportion of missing values, marginal performance differences were expected, an assumption confirmed by testing. Mode and kNN imputation generated identical scores. Consequently, Mode imputation was chosen for its clarity and reproducibility. This method's selection is in alignment with the prevailing low rate of missing values, reaffirming that the choice aligns with the dataset's characteristics.

3.4.1 Train-Test partitioning

To create the train and test sets, Partitioning (70%-30%) (Gholamy et al., 2018) with Stratified Sampling was applied to this dataset (test set has 11.834 observations). Some supervised learning algorithms require an equal class distribution to generalize well. To address the unbalanced target variable, SMOTE (Synthetic Minority Over-sampling Technique) was applied for over-sampling the minority classes, resulting in a training set with 41.709 observations (33.(3)% of each class) (Figure 9). SMOTE’s Oversample adjusts the class distribution by adding synthetic rows, and, as a result, the output contains the same number of rows for each of the possible classes.

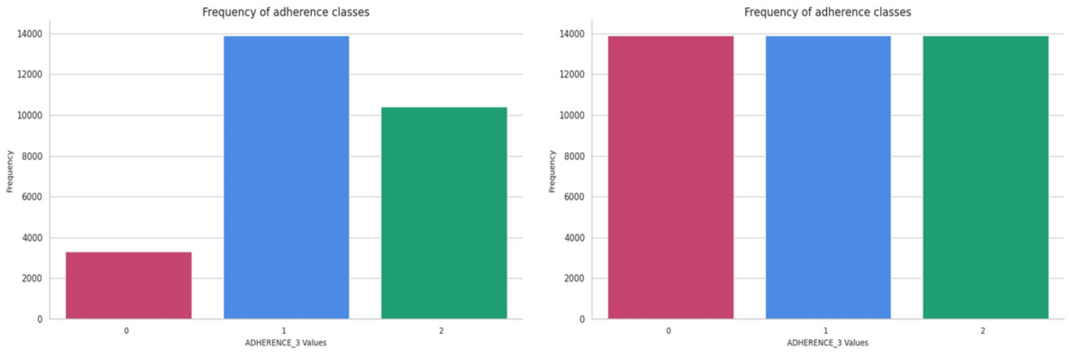


Figure 9 – Distribution of the target variable in the train set before and after oversampling

3.5 Data modelling

The problem consisted of a classification prediction task of a categorical variable. Diverse algorithms can be used in multi-class classification tasks. We decided to experiment with seven different ones: Decision Trees (DT), Naïve Bayes (NB), k-Neighbors Classifier (kNN), Random Forest (RF), Gradient Tree Boosting (GTB), Ada Boost (AB) and XGBoost (XGB), comparing their performance metrics, namely, Accuracy, Precision, Recall, weighted average F1 score and Area Under the Receiver Operating Characteristic Curve One-vs-one (ROC AUC OVO).

3.5.1 Decision Trees (DT)

The DT is a graphical model resembling an inverted tree composed of nodes and branches. It is alternatively referred to as a classification or regression tree based on whether the outcome variable is categorical or numeric. Employing a divide-to-conquer strategy, DT dissects intricate problems into simpler sub-problems and recursively applies the same tactic to each sub-problem. Its discriminative ability arises from segmenting attribute-defined space into sub-spaces, each linked to a class. Nodes hold attribute tests, dictating decisions; terminal nodes, or leaves, emerge when no subsequent node exists, signifying a class label. The root-to-leaf path constitutes a classification rule (Breiman, 1984; Gama et al., 2015).

DT exhibits efficacy with heterogeneous data, handling well missing values, mixed data, and irrelevant inputs, robustness for outliers in the input set, insensitivity to monotonic transformations in the input set and computational scalability.

The main parameters to tune are *max depth, min samples split* and *min samples leaf*.

3.5.2 Naive Bayes (NB)

The NB classifier is a probabilistic algorithm rooted in the assumption of attribute independence. Despite its simplistic independence assumption, which might not hold true for complex datasets, NB often outperforms other classifiers regarding classification accuracy, particularly on real-world datasets (Chandra et al., 2007). This approach is particularly valuable when training data is limited. By leveraging training data, the NB classifier learns and subsequently predicts the class of a test instance using the highest posterior probability. This method remains effective even with high-dimensional data due to its attribute-wise independence estimation (Rish, 2001). This allows the computation of the conditional probability, P(y_i | x), indicating the likelihood of object x belonging to class y_i.

3.5.3 k-Neighbors Classifier (kNN)

This classifier is an instance-based algorithm used for classification and regression tasks. In kNN, the class of an unclassified data point is determined by examining the class labels of its k-nearest neighbors from the training dataset. In this approach, distance metrics (such as Euclidean distance) are employed to measure the proximity between data points in the feature space. The kNN algorithm finds the k training examples that are closest to the new data point and assigns the class label that is most prevalent among these neighbors (Cover & Hart, 1967).

One notable characteristic of kNN is its simplicity and adaptability to various data patterns. However, it's sensitive to the choice of k, with this parameter influencing directly the algorithm's bias-variance trade-off (Gama et al., 2015).

The classifier in SciKit-Learn (sklearn.neighbors.KNeighborsClassifier) has an “auto” method to select the algorithm used to compute the nearest neighbors. KDTree was chosen and, therefore, parameters *leaf size* and *n neighbors* were tuned.

3.5.4 Random Forest (RF)

RF is an extension of bagging and consists of many individual decision trees that operate as an ensemble. RF utilizes a decision tree algorithm that selects optimal split points during tree construction, thereby increasing diversity among bagged trees and enhancing predictive accuracy. To ensure diversity, RF introduces feature randomness, also termed feature bagging or "the random subspace method." This technique involves forming a random feature subset, minimizing correlations among decision trees. Notably distinct from individual decision trees, RF doesn't consider all potential feature splits; rather, it employs a subset of features for each tree. The generalization error of a forest of tree classifiers depends on the strength of the individual trees in the forest and the correlation between them. RF combines the predictions of multiple decision trees through a voting mechanism (for classification) and can be efficiently parallelized, making it suitable for large datasets (Breiman, 2001).

The main parameters to tune arem similarly to DT, *max depth*, *min samples split*, *min samples leaf, and n estimators*.

3.5.5 Gradient Tree Boosting (GTB)

Boosting is a powerful learning model initially conceived for binary classification problems (Schapire, 1990) and can be extended to regression problems. In this algorithm, decision trees are formed sequentially, using information from the previous tree. The algorithm has an implementation in the SciKit-Learn package as sklearn.ensemble.GradientBoostingClassifier, which supports both binary and multi-class classification, builds an additive model in a forward stage-wise way, optimising arbitrary differentiable loss functions. The implementation is based on the seminal paper "Greedy Function Approximation: A Gradient Boosting Machine" (Friedman, 2001).

The parameter 'n_estimators' controls the quantity of regression trees. Tree size can be managed by either specifying the tree depth with 'max_depth' or determining the number of leaf nodes using 'max_leaf_nodes.' The 'learning_rate' is a hyper-parameter, constrained within the range [0.0, 1.0], which regulates overfitting through shrinkage.

3.5.6 Ada Boost (AB)

AdaBoost, short for Adaptive Boosting, is an ensemble learning technique in machine learning. It combines multiple weak learners, typically decision trees with limited predictive power, into a strong ensemble model. AdaBoost assigns varying weights to each weak learner's predictions and iteratively focuses on the instances that are misclassified by the ensemble. It sequentially trains weak learners to correct the errors of previous ones, giving more importance to the challenging data points until a predefined number of weak learners are trained, or a certain level of accuracy is achieved. The final model combines the individual weak learners' predictions to make accurate and robust predictions, effectively enhancing the model's overall performance.

The classifier in SciKit-Learn (sklearn.ensemble.AdaBoostClassifier) implements the algorithm known as AdaBoost-SAMME (Zhu et al., 2009). The parameter n_estimators control the number of weak learners, and the learning_rate parameter controls the contribution of the weak learners in the final combination. The main parameters to tune are n_estimators and the complexity of the base estimators (e.g., its depth max_depth or minimum required number of samples to consider a split min_samples_split).

3.5.7 Multi-layer perceptron (MLP)

Artificial Neural Networks (ANNs) are computational models inspired by the brain's information processing mechanisms. During the learning phase, ANNs adjust the connection weights between input and output units to predict correct class labels for input examples. The tested algorithm, MLP, is a supervised learning algorithm that learns a function $f(\cdot) = R^m \rightarrow R^o$ by training on a dataset, where m is the number of dimensions for input and o is the number of dimensions for output. It learns a non-linear function approximator for either classification or regression (Rosenblatt, 1958).

The implementation in SciKit-Learn (sklearn.neural_network.MLPClassifier) implements a multi-layer perceptron (MLP) algorithm that trains using Backpropagation and supports only the Cross-Entropy loss function (which it minimizes), allowing probability estimates by running the predict_proba method. MLPClassifier supports multi-class classification by applying Softmax as the output function.

3.5.8 XGBoost (XGB)

eXtreme Gradient Boosting (XGBoost) is a gradient boosting implementation, which is an ensemble learning technique aimed at improving model performance. In boosting, models are sequentially added to correct errors made by earlier models until further improvement is not feasible.

XGBoost employs an ensemble of decision trees, known as classification and regression trees (CART) because individual CARTs typically lack strong predictive power. The ensemble strategy aggregates predictions from multiple trees to enhance accuracy. Gradient boosting is the key technique used here, where new models predict the residuals or errors of previous models and their predictions are combined for the final outcome. This versatile algorithm is applicable to both regression and classification tasks, offering improved predictive capabilities (Chen & Guestrin, 2016).

The xgboost package for Python was used with its native interface. Parameters 'max_depth', 'n_estimators', 'gamma', 'learning_rate' and 'max_bin' were optimized.

3.5.9 Hiperparametrization and validation

Hiperparametrization was used to improve scores of each model. sklearn.model_selection. GridSearchCV was used to find the best combination of hyperparameters for each model by searching through a predefined grid of parameter values. One of the advantages of GridSearchCV is its integration with cross-validation. We used a 5-fold cross-validation strategy.

Cross-validation is a validation technique used to assess the model's performance more robustly, mitigating the risk of overfitting and providing a more accurate estimation of how the model will perform on unseen data. The model is trained on the training data for each fold while varying the hyperparameters specified in the grid. The parameters selected maximize the score of the left out data.

Optimized parameters are described in Table 7, along with the best parameters found.

Table 7 – Optimized parameters and optimal value

	Parameter	Best value		Parameter	Best value
AB	Max depth	30	**MLP**	Alpha	0,1
	Min samples split	10		Beta 1	0,2
	N estimators	400		Beta 2	0,01
DT	Max depth	10		Hidden layer sizes	(100, 50, 100)
	Min samples split	15	**NB**	Alpha	0,1
	Min samples leaf	6		Fit prior	True
GTB	Max depth	5	**RF**	Max depth	50
	Min samples split	20		Min samples split	2
	Learning rate	0,2		Min samples leaf	4
	N estimators	150		N estimators	300
	Subsample	0,8	**XGB**	Max depth	3
kNN (KDTree)	Leaf size	15		Learning rate	0,4
	N neighbors	3		N estimators	300
				Gamma	0,5
				Max bin	2

3.6 Software

IBM SPSS Statistics version 28.0 (IBM, 2021) was used for exploratory data analysis and computation of the target variable, while pre-processing, model training, optimization (hyperparametrization), and deployment were implemented in Python, using multiple libraries referenced later, highlighting the use of the SciKit-Learn package for most of the modeling (Pedregosa et al., 2011).

4 Results

In this section, we will discuss the results of the implemented machine learning models, including Decision Trees (DT), Naïve Bayes (NB), k-Neighbors Classifier (kNN), Random Forest (RF), Gradient Tree Boosting (GTB), Ada Boost (AB), Multi-layer perceptron (MLP) and XGBoost (XGB), comparing their performance on the given task and select the best-performing model for further analysis.

4.1 Performance Metrics

We've selected five performance metrics appropriate for multi-class classifications problems and implemented them with sklearn.metrics.

One of the most well-known performance measures to evaluate data mining models' performance is accuracy - the number of correct predictions from all predictions made. However, accuracy can be misleading. Especially in problems with a large class imbalance, a model can efficiently predict the value of the majority class and achieve a high classification accuracy, not being useful in the problem domain to predict other classes.

Therefore, along with Accuracy, we will compare the Precision and Recall. Precision quantifies the model's ability to make correct positive predictions among all positive predictions. The parameter "average" was set to "weighted", so that it calculates the metric for each label, and finds their average weighted by support (the number of true instances for each label). Recall, also known as Sensitivity or True Positive Rate, measures the model's ability to correctly identify all positive instances. Because Weighted Recall is equal to accuracy, we chose to compare Recall with the "average" parameter set to "macro", witch calculates metrics for each label, finding their unweighted mean.

We also compared the model's F1 score, balanced F-score or F-measure. The F1 score considers both precision and recall, providing a balance between these metrics. The weighted average F1 score is particularly useful for imbalanced datasets.

Lastly, we computed the Area Under the Receiver Operating Characteristic Curve (ROC AUC) from prediction scores. We used the weighted parameter, witch calculates metrics for each class and their average, weighted by support. We also chose the configuration One-vs-one (OVO), which computes the average AUC of all possible pairwise combinations of classes.

For the best performing model, we plotted the ROC AUC OVO and the ROC AUC OVR (One-vs-rest), considering the class 2 (most adherent) as the interest class.

4.2 Models' comparison

Table 8 summarizes the performance measures achieved with each of the previously described algorithms. Results indicate that the optimized XGBoost model outperformed all other models across all performance metrics except Recall. It achieved the highest accuracy, precision, weighted average F1 score, and ROC AUC OVO score. These findings underscore the effectiveness of XGBoost in this specific task.

Table 8 – Performance measures

	Accuracy	Precision	Recall	F1 score	ROC AUC OVO
MLP	0.300	0.520	0.422	0.279	0.682
NB	0.448	0.504	0.448	0.439	0.647
kNN	0.441	0.505	0.461	0.456	0.627
DT	0.568	0.561	0.506	0.562	0.702
RF	0.601	0.595	0.527	0.593	0.742
AB	0.603	0.598	0.527	0.594	0.741
GTB	0.609	0.603	0.609	0.601	0.745
XGB	0.616	0.610	0.534	0.608	0.750

We present the Receiver Operating Characteristic (ROC) curves to represent the model's classification performance visually. Figure 10 presents the ROC AUC OVR for class 2, the most adherent to public health measures (class 0 represents the less adherent and class 1 is the intermediate class of adherence).

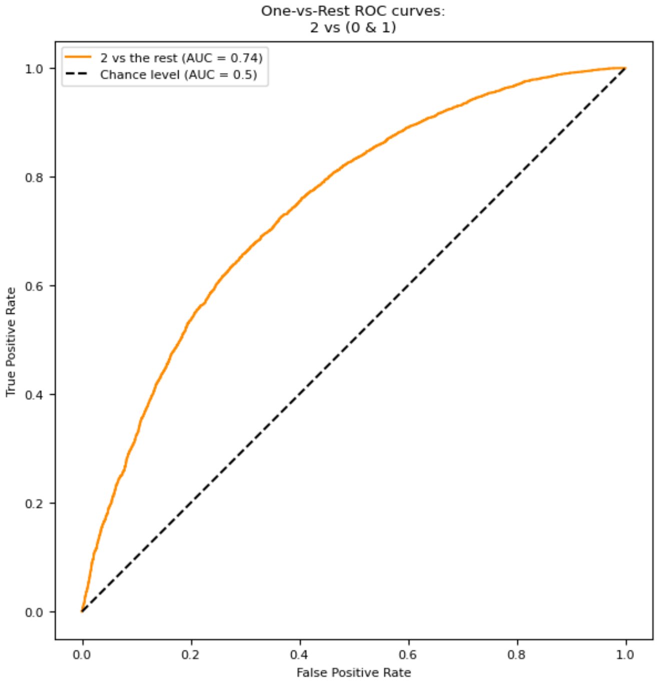
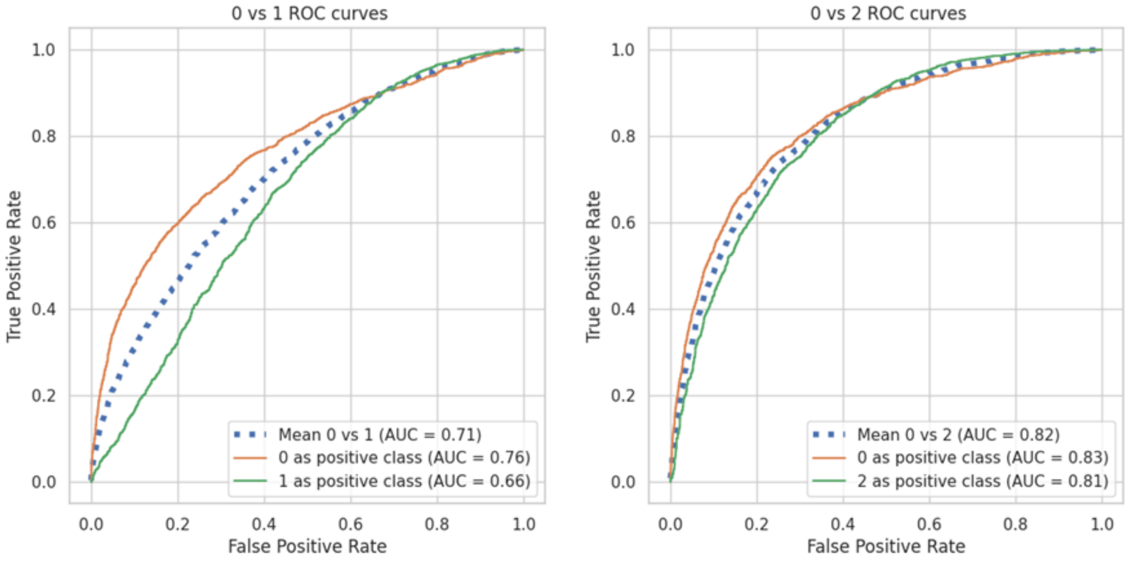


Figure 10 - ROC AUC OVR (2 vs rest) for the optimized XGBoost model

ROC AUC OVO were also plotted and are presented in Figure 11Figure 10. As expected, classes 0 and 1 are the least well identified. The model performs better in correctly identifying class 0 vs class 2, with the ROC AUC score (0.82) above the OVO macro average (0.75).



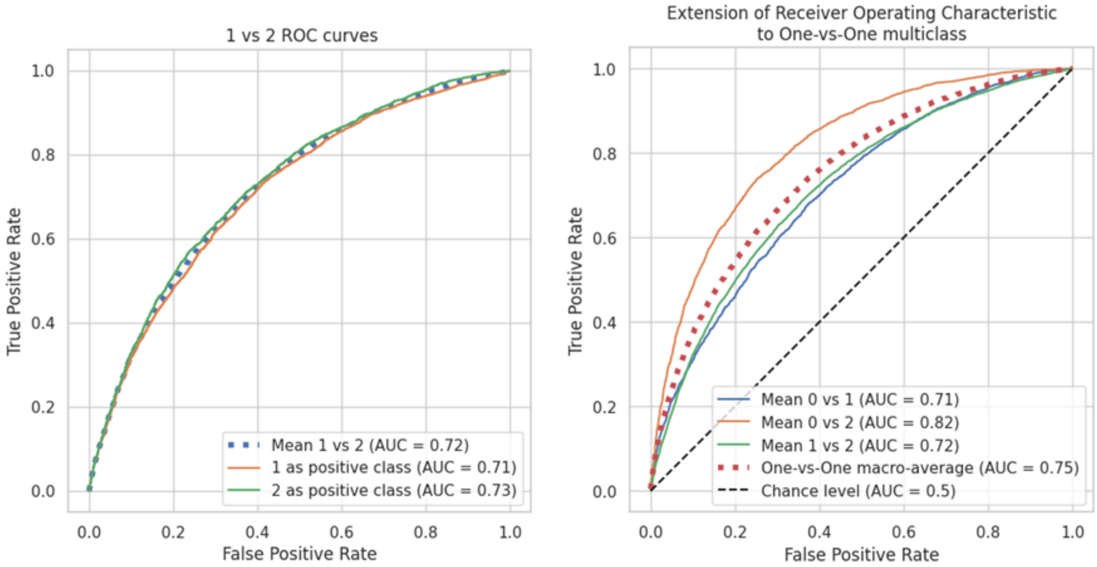


Figure 11 – ROC AUC OVO for the optimized XGBoost model

Using the available function of the xgboost package, we plotted the parameters' importance (global and the average gain across all splits the feature is used in) based on fitted trees.

The country (country_W9) of the participant is the feature with overall more importance, followed by their year of birth (cadn003__W9) and the quality of life and well-being index score (casp.8_LT). Also in the five most important features are the contact frequency with neighbors/friends/colleagues during last 3 months (cas103_4_W9) and the score for the 10-word recall test (recall_1.8_LT), a measure of cognitive function often used as indicators for the presence of cognitive impairment and dementia (Listabarth et al., 2022) (Figure 12).

The use of the internet for communications since the outbreak (cait104__W9) is the feature with the largest gain overall, followed by limitations due to health problems in last 6 months (caph105__W9) and satisfaction with treatment at a medical facility (caq122__W9) (Figure 13).

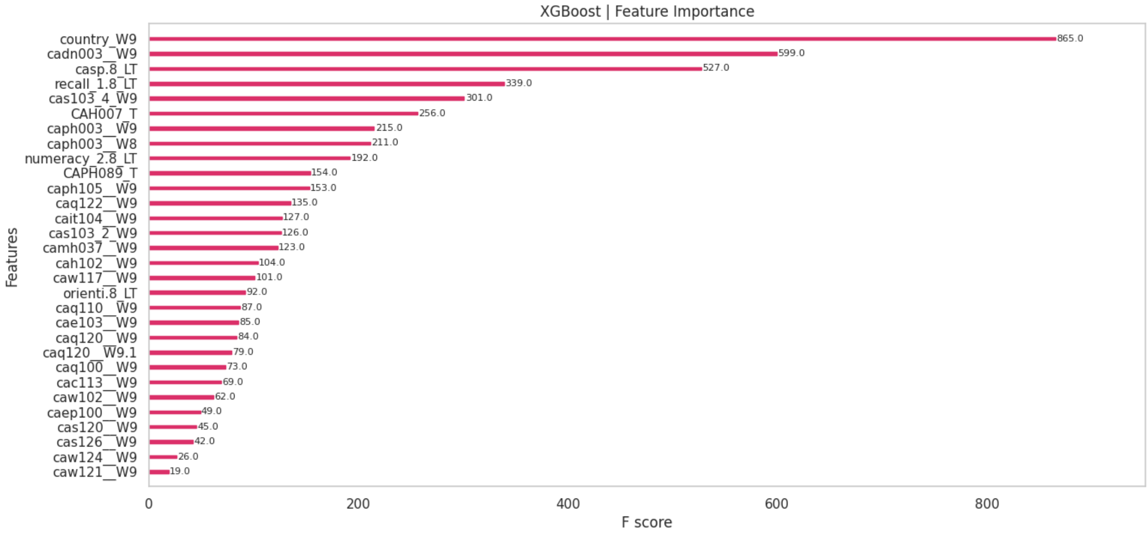


Figure 12 – Feature importance for the XGBoost model

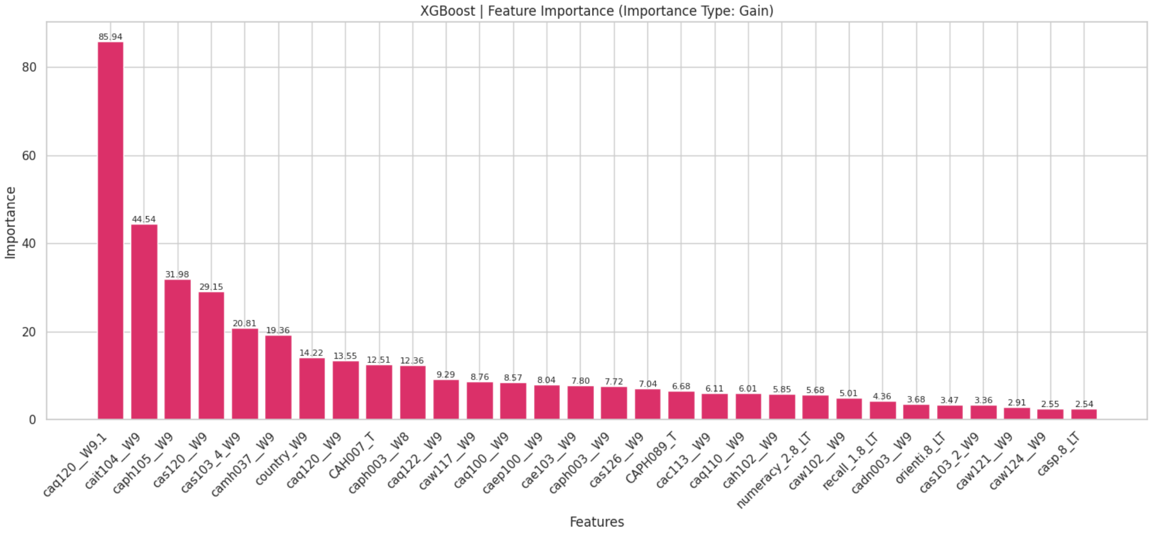


Figure 13 - Feature importance (gain) for the XGBoost model

Furthermore, we employed SHAP (SHapley Additive exPlanations) values to explain the XGBoost model's predictions. SHAP values offer insights into the contributions of each feature to the model's decision-making process, aiding in the interpretation of results and model transparency (Lundberg & Lee, 2017).

Figure 14 shows the participant's country (country_W9) is the feature with the highest impact. As this is a categorical variable, we will explore it further afterward. Contact frequency with neighbors/friends/colleagues during last 3 months (cas103_4_W9, 1: Daily - 5: Never), having visited doctor/medical facility other than hospital (caq120__W9.1, 0: No, 1:Yes), using the internet for communications since the outbreak (cait104__W9, 0: No, 1:Yes) and the year of birth (cadn003__W9) make up the five most impactful features. The amount of medication (CAH007_T Health - drugs (sum)), severe limitations due to a health problem in last 6 months (caph105__W9, 1: severely limited, 2: limited, but not severely, 3: not limited), health before the outbreak (caph003__W8, 1: Excellent – 5: Poor) and satisfaction with treatment at medical facility (caq122__W9, 1: very satisfied, 4: very dissatisfied) are the features with an impact over 0.2. We will now explore the impact of these features in each class. Our interest class is 2, the most adherent, but we'll present results for the three and explore further only for this class.

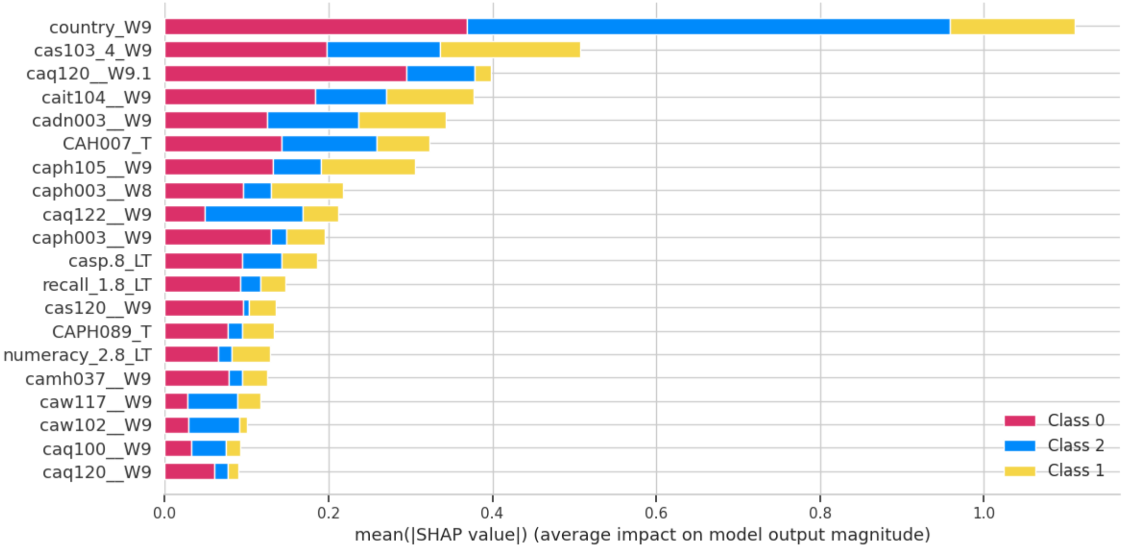


Figure 14 – Average impact of each feature on model output magnitude

For class 0 (the minority class, the less adherent) (Figure 16), the plot confirms and further details the average impact seen before.

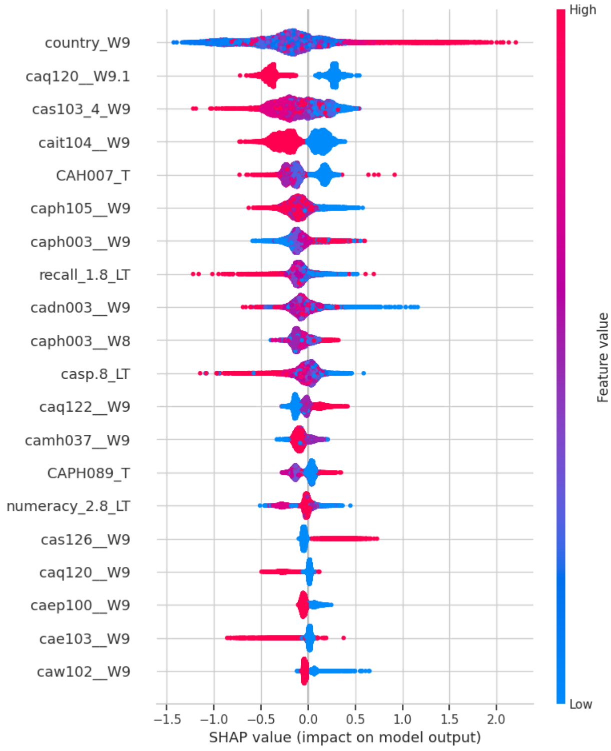


Figure 16 – SHAP Values plot for Class 0

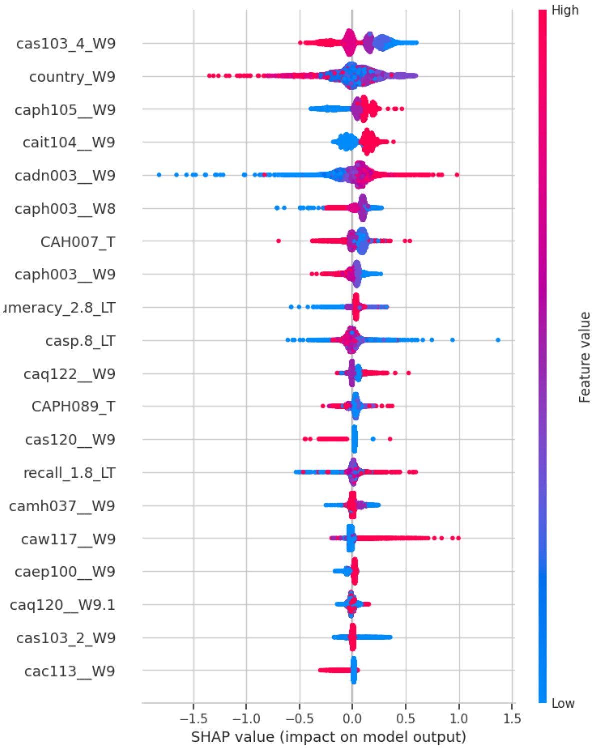


Figure 15 - SHAP Values plot for Class 1

Country (country_W9) is the feature with the highest impact. Using the internet for communications since the outbreak (cait104__W9) also impacts the classification as 0, along with higher satisfaction with treatment at medical facility (caq122__W9, 1: very satisfied, 4: very dissatisfied) and severe limitations due to a health problem in last 6 months (caph105__W9, 1: severely limited, 2: limited, but not severely, 3: not limited). The year of birth also impacts this classification (cadn003__W9), with older participants having a higher impact.

Regarding class 1 (Figure 15), the SHAP values plot shows that participants with higher contact frequency with neighbors/friends/colleagues during last 3 months (cas103_4_W9, 1: Daily, 5: Never) were more likely classified in this class. Contrary to the first class, for class 1, younger participants (cadn003__W9) and less limitations due to a health problem in last 6 months (caph105__W9, 1: severely limited, 2: limited, but not severely, 3: not limited) have a higher impact. Regarding the use of internet (cait104__W9), participants using it since the outbreak are more likely classified in this class, as well as the ones who felt the least safe at work (caw117__W9, 1: very safe, 4: very unsafe).

Finally, for class 2 (Figure 17), the most adherent, we can see, again, a strong impact of the participant's country (country_W9), which we will analyze more closely next. Participants with lower contact frequency with neighbors/friends/colleagues during last 3 months (cas103_4_W9, 1: Daily - 5: Never) were more classified as most adherent, but so were the ones that used the internet for communications since the outbreak (cait104__W9, 0: No, 1:Yes).

Participants with better cognitive function score (recall of words - recall_1.8_LT), higher quality of life and well-being scores (casp.8_LT) and younger participants (cadn003__W9) are also more classified as more adherent.

On the other hand, participants who take more medication (CAH007_T Health - drugs (sum)), the ones with higher health fragility (CAPH089_T), that visited doctor/medical facility other than hospital (caq120__W9.1, 0: No, 1:Yes), alongside with participants with an average satisfaction with treatment at medical facility (caq122__W9, 1: very satisfied, 4: very dissatisfied), and who didn't have to postpone a medical appointment (caq110__W9) were also more classified in this class.

Interestingly, participants who felt the least safe at work (caw117__W9, 1: very safe, 4: very unsafe) impact the classification both positively and negatively.

Those that received financial support due to outbreak (cae103__W9) also impacted positively the classification in this class.

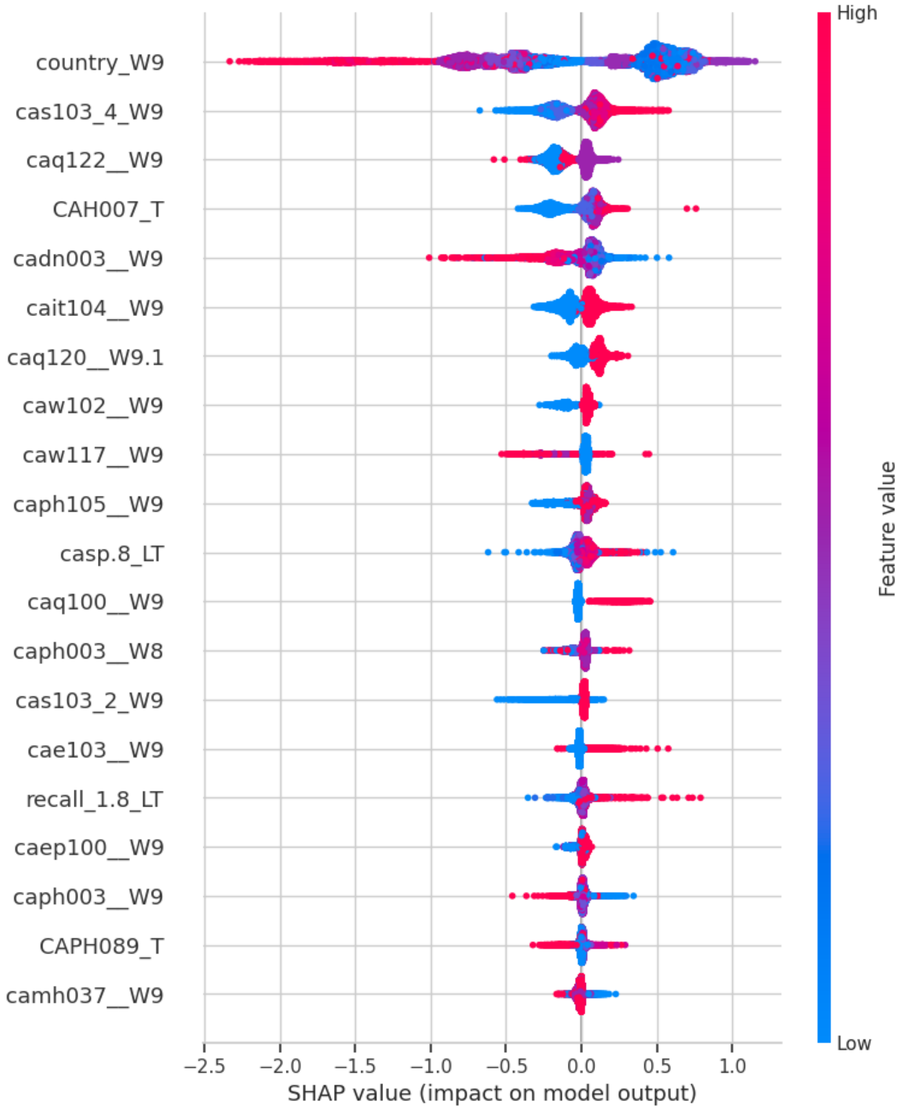


Figure 17 - SHAP Values plot for Class 2

To further explore the impact of this variables, we created dependence plots for the features with higher impact or an interesting behavior.

As previously seen, the participant's country was the most influential feature for class 2. In this dependence plot (Figure 18), we can see the interaction of the variables country and year of birth. Portugal is the country that has the most impact, followed by Luxembourg and Finland. For some countries, such as Austria, Spain and Denmark, the older participants (lower year of birth) have more impact than the younger ones. However, the younger participants are more impactful in others, such as Finland, Malta and Croatia.

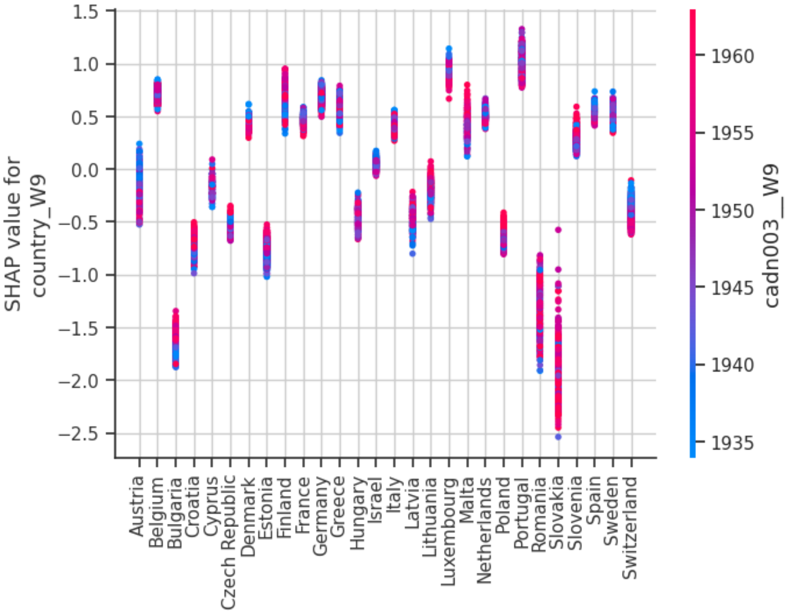


Figure 18 – SHAP dependence plot: Country – Year of birth

Figure 19 shows the interaction between country and medication taken (sum). Once again, some countries show a clear pattern of participants who take more medications having the most impact (e.g., Portugal and Spain). In contrast, in others there is an exact opposite pattern (Slovenia, Latvia and Finland).

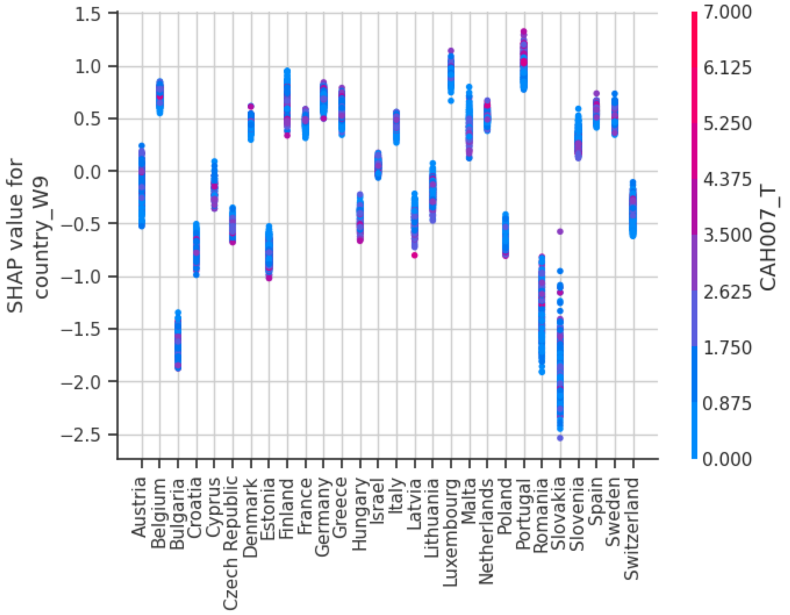


Figure 19 - SHAP dependence plot: Country – Medication taken

Figure 20 represents the dependence plot for country and CASP, quality of life and well-being index. Similarly, in Portugal, Austria and Spain, it is the people with the lowest quality of life and well-being index that most impact the classification in this class, whilst in Estonia, Lithuania and Luxembourg the opposite pattern is found.

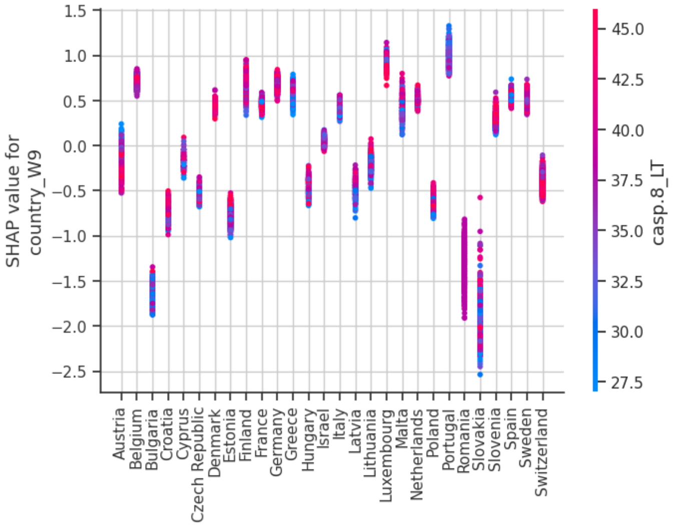


Figure 20 - SHAP dependence plot: Country – CASP: quality of life and well-being index

Figure 21 shows the rating of subjective health per country. In Portugal, Austria and Spain the participants with the lowest rating of subjective health are the ones with the highest impact (in accordance with previous results), while in Finland and Denmark, the ones with the highest rating are the ones who have more impact in the classification.

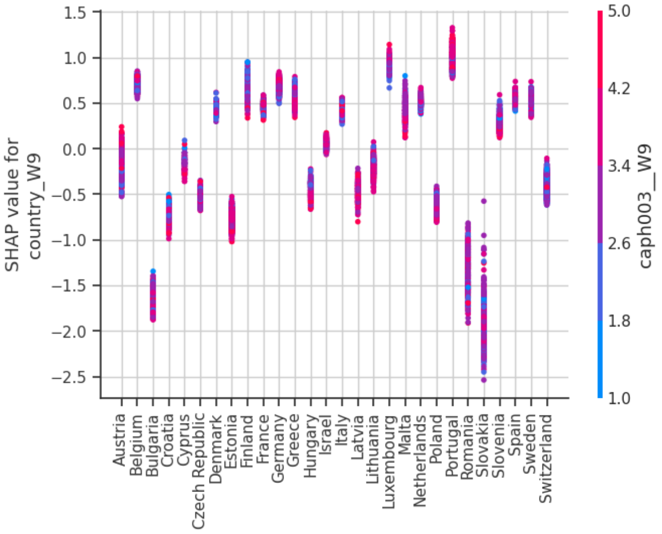


Figure 21 - SHAP dependence plot: Country – Rating of subjective health

Figure 22 depicts the interaction between year of birth and use of the internet for communications since the outbreak (cait104__W9, 0: No, 1:Yes). We can see that older participants are more significant for the model than the young ones (thus the curvilinear pattern), but also that, among the mean aged participants, the ones who use internet for communication are more influential. In comparison, among younger participants, almost all use the internet, yet those who don’t use it have a higher impact.

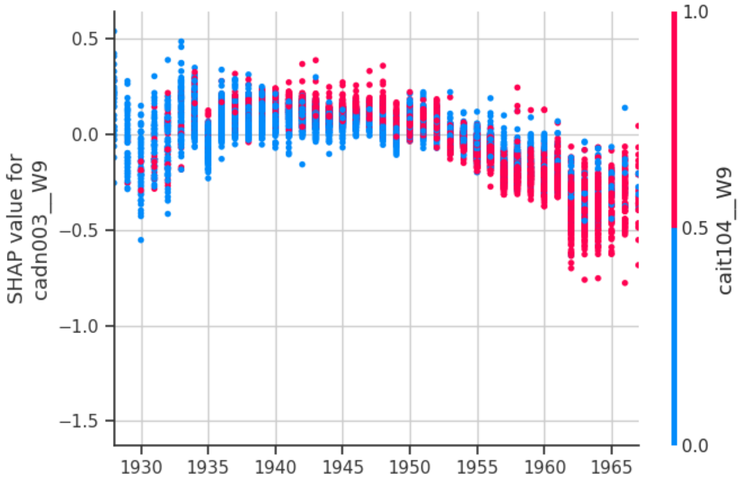


Figure 22 - SHAP dependence plot: Year of birth – Use of internet for communication since outbreak

Figure 23 illustrates the interaction between year of birth and limitation due to health issues. Mean aged participants with fewer limitations are more influential, while among the younger ones, having limitations seems to impact the model more.

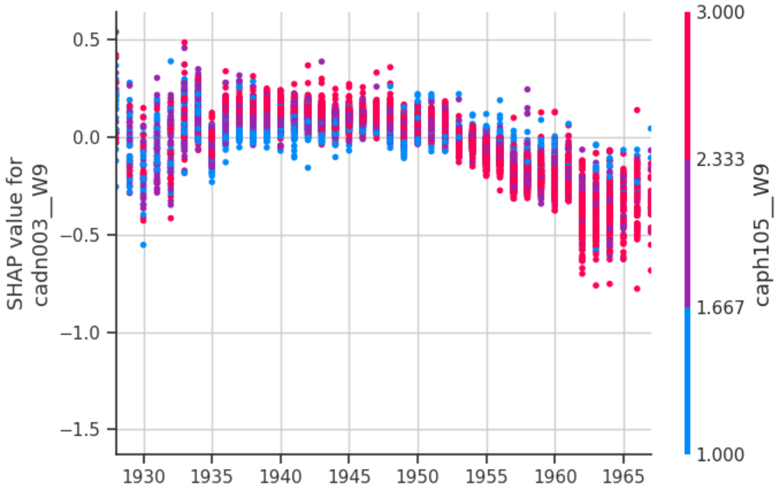


Figure 23 - SHAP dependence plot: Year of birth – Limitations because of a health problem

Lastly, Figure 24 clarifies the dependence of the Rating of subjective health and forwenting a medical treatment. Given the different patterns regarding subjective health ratings in different countries, we investigated other dependences of this feature. The plot shows that, among participants with lower subjective health ratings, giving up a medical treatment had a higher impact on the model than for those with higher subjective health ratings, among which not performing the medical treatment is less important.

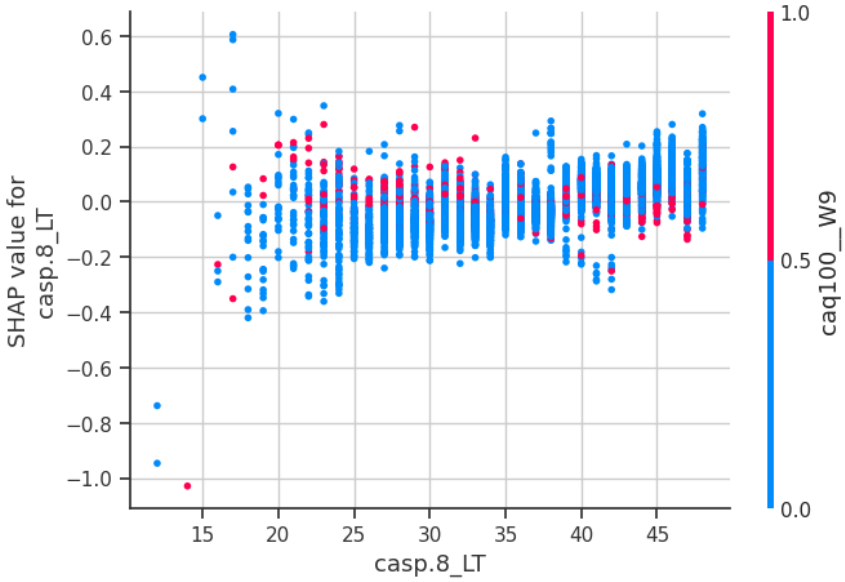


Figure 24 - SHAP dependence plot: Rating of subjective health – Healthcare: forwent medical treatment

5 Discussion and conclusions

Although guidelines and restriction altered throughout time, adherence to public health measures had a favorable influence on the spread of the infection and the overall number of deaths.

Predicting and explaining human behavior during infectious outbreaks represents the possibility of a better situation management.

Therefore, we aimed to develop a Machine Learning framework that allowed us to predict adherence to safety measures using data from the Survey of Health, Ageing and Retirement in Europe (SHARE). The COVID-19 questionnaire evaluated twelve health behaviors related to prevention and risk. Upon further analysis, we used nine of these to create a composite measure of adherence, later discretized into three classes: 0 – participants demonstrating lower levels of adherence (0 to 2 adherence behaviors), 1 – participants exhibiting intermediate levels (3 to 5 adherence behaviors) and 2 – participants displaying the highest levels of adherence (6 to 8 adherence behaviors).

Seven algorithms were used to create the classification prediction models and performance metrics were evaluated and compared.

The implementation and hyperparametrization of Xgboost, combined with the stratified train-test split and the over-sampling of minority classes, provided a good performance in classifying adherence, especially discriminating the 'most adherent' class (ADHERENCE_3=2) (AUC=0.82).

Investigating the feature importance for this implementation (using the xgboost parameters and SHAP values), we realised that the participants' country most impacted the model, but most noticeably, class 2. European countries implemented measures and restrictions and made vaccination available at different times and rates. Determining the causal effects of government policies is not simple, as it is impeded by numerous factors that can cause confusion and a multitude of possible sources of endogeneity (Hale et al., 2021). Nevertheless, Islam and colleagues (2020) evaluated data from 11 countries. They found that earlier implementation of lockdown was associated with a larger reduction in COVID-19 incidence compared with delayed implementation of lockdown after other physical distancing interventions were in place, such as school closures, workplace closures, and restrictions on mass gatherings. Likewise, implementing mask mandates in communal or public areas significantly decreased transmission rates compared to less strict policies that only required mask-wearing in certain public spaces. So, in future epidemics caused by airborne pathogens, mandating masks in almost all public areas from the outset would be a viable approach, considering the relatively minor social and economic impact of such an intervention (Sharma et al., 2021). Studying global governmental responses to the COVID-19 pandemic, Hale and co-authors established that variations in government regulations may have been the main factor influencing the spread of COVID-19, as seen in a 2009 influenza pandemic investigation. Early data from China supports this, and tighter regulations have been found to reduce SARS-CoV-2 spread. However, a review of 11 European countries' COVID-19 deaths shows that the effectiveness of recent interventions (third wave) is still uncertain (Hale et al., 2020).

But although mandatory public measures seem to impact the spread of the virus positively, compliance with these measures isn’t guaranteed. Shanka and Menebo (2022) found that trust in the government impacts compliance with COVID-19 safety protocols; problem awareness partially mediates this influence; individualistic orientation moderates the connection between trust in government and following COVID-19 precautions; and education level and health status are both linked to following safety guidelines. Similarly, perceptions of governmental communication as credible and honest positively predicted self-reported adherence in 8 countries (Lavallee et al., 2021). Our model was also differently impacted by other variables in different countries. While in some countries, the model was impacted by older, potentially sicker participants (more medication, lowest rating of subjective health) and participants with inferior quality of life and well-being scores in class 2, in other countries, the opposite happened.

We also found that participants who received financial support positively impacted the model’s classification in class 2. Research supports this finding, as financially secure workers comply more with COVID-19 restrictions, as they can stay home and enact social distancing behaviors, and financially insecure workers have less opportunity to do so (Probst et al., 2020). An extensive literature review of the economics of COVID-19, states that Temporary Paid Sick Leave has increased compliance with stay-at-home orders. At the same time, income level was shown to be a significant determinant, as low-income neighborhoods' residents are less likely to comply with stay at home recommendations during non-working hours. This is due to their higher likelihood of being front-line workers and need for frequent shopping trips. Additionally, those with lower income, usually have less flexible work arrangements, and limited indoor space and, therefore, are less likely to practice social distancing (Brodeur et al., 2021).

We also found a dependence on age and limitations, with younger participants who have limitations having a greater impact on the classification in class 2. A study in Macao examined adherence to six COVID-19 precautionary measures, revealing adherence was linked to perceived severity, benefit, barrier, cue-to-action, societal cynicism, and reward (Tong et al., 2020). Harper and colleagues also found that fear of COVID-19 was consistently the only predictor of adherence (e.g., social distance, enhanced hand hygiene), with no influence of political orientation or moral foundations (2020). In fact, diseases pose significant threats to physical health and economic well-being. Realistic and symbolic threats have different relationships with restricted public health behaviors. Realistic threats lead to higher self-reported adherence to social separation behaviors, while symbolic threats predict lower adherence and suggest innovative strategies to assert identity (Kachanoff et al., 2020). Correspondingly, illness perceptions toward COVID-19 also have a significant direct effect on adherence to precautionary measures (Chong et al., 2020), and perceived susceptibility has proven to be an important predictor of adherence to COVID-19 preventive behaviors (Yehualashet et al., 2021). These findings seem concordant with our results, as for someone with health limitations, a worse perceived health status, or a lower quality of life and well-being, COVID-19 presumably would pose a realistic health and/or economic threat. Although our results cannot be interpreted as suggesting causality and merely explain the model implemented, it seems appropriate to notice the same pattern.

This study has several limitations. Measuring adherence was not an objective of the SHARE Corona survey. Therefore, we created a potential measure of adherence with the information that was collected and although its internal consistency is acceptable and it correlates with testing positive for COVID-19, it doesn't have a conceptual basis and was not validated in any other way (such as criterion-related, construct or discriminant validity). Likewise, other strategies could have been employed for feature reduction. Ours was mainly based on theoretical knowledge of adherence predictors as presented in the literature review, and given the large number of available features, an automated technique was employed. Nevertheless, other feature reduction strategies could have been implemented (e.g., Principal Component Analysis), leading to different results. Finally, the model's performance is modest, and its application in real-world contexts may be limited.

Despite these limitations, we believe that predicting adherence (and explaining that prediction) in a reproducible way with data already collected to other ends is a relevant strength of this work. Moreover, the model’s explanation seems to align with recent research on COVID-19 safety measures effectiveness and adherence to those measures.

For further research, it would be very interesting to consider the cultural and social contexts of each country, as well as the measures in place and official communication on the pandemic at the tuime of the surveys.

We believe the patterns revealed highlight the importance of addressing the specific needs and vulnerabilities of individuals with health limitations, financial insecurity or a lower quality of life in public health interventions and policies related to COVID-19. By understanding these factors, policymakers can develop targeted strategies to mitigate these populations' potential health and economic risks during the pandemic. Tailoring interventions to address these unique challenges can lead to increased effectiveness and overall compliance with public health guidelines. Ultimately, prioritizing the needs of vulnerable populations will not only protect their well-being but also contribute to the overall success of pandemic response efforts.

References

Ahmed, M. Z., Fodjo, J. N. S., Gele, A. A., Farah, A. A., Osman, S., Guled, I. A., Ali, A. M., & Colebunders, R. (2020). COVID-19 in Somalia: Adherence to Preventive Measures and Evolution of the Disease Burden. *Pathogens*, *9*(9). https://doi.org/10.3390/pathogens9090735

Alagoz, O., Sethi, A. K., Patterson, B. W., Churpek, M., & Safdar, N. (2021). Effect of Timing of and Adherence to Social Distancing Measures on COVID-19 Burden in the United States. *Annals of Internal Medicine*, *174*(1), 50–57. https://doi.org/10.7326/m20-4096

Alkhaldi, G., Aljuraiban, G. S., Alhurishi, S., de Souza, R., Lamahewa, K., Lau, R., & Alshaikh, F. (2021). Perceptions towards COVID-19 and adoption of preventive measures among the public in Saudi Arabia: a cross-sectional study. *BMC Public Health*, *21*(1). https://doi.org/10.1186/s12889-021-11223-8

Apanga, P. A., & Kumbeni, M. T. (2021). Adherence to COVID-19 preventive measures and associated factors among pregnant women in Ghana. *Tropical Medicine & International Health*, *26*(6), 656–663. https://doi.org/10.1111/tmi.13566

Atchison, C., Bowman, L. R., Vrinten, C., Redd, R., Pristerà, P., Eaton, J., & Ward, H. (2021). Early perceptions and behavioral responses during the COVID-19 pandemic: a cross-sectional survey of UK adults. *BMJ Open*, *11*(1), e043577. https://doi.org/10.1136/bmjopen-2020-043577

Atzendorf, J., & Gruber, S. (2021). Depression and loneliness of older adults in Europe and Israel after the first wave of covid-19. *European Journal of Ageing*. https://doi.org/10.1007/s10433-021-00640-8

Bailey, B., Whelen, M. L., & Strunk, D. R. (2021). Adhering to COVID-19 health guidelines: Examining demographic and psychological predictors of adherence. *Applied Psychology: Health and Well-Being*, *13*(4), 968–985. https://doi.org/10.1111/aphw.12284

Bonardi, J., Gallea, Q., Kalanoski, D., & Lalive, R. (2020). Fast and local: How lockdown policies affect the spread and severity of covid-19. *Covid Economics*, *23*, 325–351.

Bogg, T., & Milad, E. (2020). Demographic, personality, and social cognition correlates of coronavirus guideline adherence in a U.S. sample. *Health Psychology*, *39*(12), 1026–1036. https://doi.org/10.1037/hea0000891

Börsch-Supan, A. (2022a). Survey of Health, Ageing and Retirement in Europe (SHARE) Wave 8. COVID-19 Survey 1. Release version: 8.0.0. SHARE-ERIC. Data set. DOI: 10.6103/SHARE.w8ca.800

Börsch-Supan, A. (2022b). Survey of Health, Ageing and Retirement in Europe (SHARE) Wave 9. COVID-19 Survey 2. Release version: 8.0.0. SHARE-ERIC. Data set. DOI: 10.6103/SHARE.w9ca.800

Börsch-Supan, A., M. Brandt, C. Hunkler, T. Kneip, J. Korbmacher, F. Malter, B. Schaan, S. Stuck, S. Zuber (2013). Data Resource Profile: The Survey of Health, Ageing and Retirement in Europe (SHARE). International Journal of Epidemiology. DOI: 10.1093/ije/dyt088.

Breiman, L. (1984). Classification And Regression Trees. In *Routledge eBooks*. https://doi.org/10.1201/9781315139470

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/a:1010933404324

Brodeur, A., Gray, D., Islam, A., & Bhuiyan, S. J. (2021). A literature review of the economics of COVID-19. *Journal of Economic Surveys*, *35*(4), 1007–1044. https://doi.org/10.1111/joes.12423

Carvalho, L. D. F., & Machado, G. M. (2020). Differences in adherence to COVID-19 pandemic containment measures: psychopathy traits, empathy, and sex. *Trends in Psychiatry and Psychotherapy*, *42*(4), 389–392. https://doi.org/10.1590/2237-6089-2020-0055

Chandra, B., Gupta, M., & Gupta, M. (2007). Robust approach for estimating probabilities in Naive-Bayes classifier. In *Springer eBooks* (pp. 11–16). https://doi.org/10.1007/978-3-540-77046-6_2

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Chong, Y. Y., Chien, W. T., Cheng, H. Y., Chow, K. M., Kassianos, A. P., Karekla, M., & Gloster, A. (2020). The Role of Illness Perceptions, Coping, and Self-Efficacy on

Adherence to Precautionary Measures for COVID-19. *International Journal of Environmental Research and Public Health*, *17*(18), 6540. https://doi.org/10.3390/ijerph17186540

Cover, T. M., & Hart, P. D. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, *13*(1), 21–27. https://doi.org/10.1109/tit.1967.1053964

Ding, C., & Peng, H. (2003). *Minimum redundancy feature selection from microarray gene expression data*. Computational Systems Bioinformatics. Proceedings of the 2003 IEEE Bioinformatics Conference, 523-528. https://doi.org/10.1109/csb.2003.1227396

Eastman, B., Meaney, C., Przedborski, M., & Kohandel, M. (2021). Modeling the impact of public response on the COVID-19 pandemic in Ontario. *PLOS ONE*, *16*(4), e0249456. https://doi.org/10.1371/journal.pone.0249456

EUR-Lex - 32011D0166 - EN (2011). 2011/166/EU: Commission Decision of 17 March 2011 setting up the SHARE-ERIC. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32011D0166

Fischer, C. B., Adrien, N., Silguero, J. J., Hopper, J. J., Chowdhury, A. I., & Werler, M. M. (2021). Mask adherence and rate of COVID-19 across the United States. *PLOS ONE*, *16*(4), e0249891. https://doi.org/10.1371/journal.pone.0249891

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*(5). https://doi.org/10.1214/aos/1013203451

Gama, J., Carvalho, A., & Faceli, K. (2015). *Extração de Conhecimento de Dados – Data Mining* (2nd ed.). Edições Sílabo.

Gholamy, A., Kreinovich, V., & Kosheleva, O. (2018). Why 70/30 or 80/20 Relation Between Training and Testing Sets: A Pedagogical Explanation. *Departmental Technical Reports (CS)*, 1209. https://digitalcommons.utep.edu/cgi/viewcontent.cgi?article=2202&context=cs_techrep

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, *3*, 1157–1182. https://doi.org/10.1162/153244303322753616

Hagan, K. K., Javed, Z., Cainzos-Achirica, M., Sostman, H. D., Vahidy, F. S., Valero-Elizondo, J., Acquah, I., Yahya, T., Kash, B., Andrieni, J. D., Dubey, P., Hyder, A.

A., & Nasir, K. (2021). Social Determinants of Adherence to COVID-19 Risk Mitigation Measures Among Adults with Cardiovascular Disease. *Circulation: Cardiovascular Quality and Outcomes*. https://doi.org/10.1161/circoutcomes.121.008118

Hair, J. F. (1998). *Multivariate data analysis: A Global Perspective*. Prentice Hall.

Hair, J. F., Babin, B. J., Anderson, R. E., & Black, W. C. (2022). *Multivariate data analysis*. Cengage Learning.

Hale, T., Angrist, N., Goldszmidt, R., Kira, B., Petherick, A., Phillips, T., Webster, S. B. G., 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*(4), 529–538. https://doi.org/10.1038/s41562-021-01079-8

Hale, T., Hale, A. J., Kira, B., Petherick, A., Phillips, T., Sridhar, D., Thompson, R. N., Webster, S., & Angrist, N. (2020). Global Assessment of the Relationship between Government Response Measures and COVID-19 Deaths. *medRxiv (Cold Spring Harbor Laboratory)*. https://doi.org/10.1101/2020.07.04.20145334

Han, J., Pei, J., & Kamber, M. (2011). *Data Mining: Concepts and techniques*. Elsevier.

Hannemann, T., & Atzendorf, J. (2021). Behavioral Risk Factors and Adherence to Preventive Measures: Evidence From the Early Stages of the COVID-19 Pandemic. Frontiers in Public Health, 9. https://doi.org/10.3389/fpubh.2021.674597

Harper, C. A., Satchell, L., Fido, D., & Latzman, R. D. (2020). Functional fear predicts public health compliance in the COVID-19 pandemic. *International Journal of Mental Health and Addiction*, *19*(5), 1875–1888. https://doi.org/10.1007/s11469-020-00281-5

IBM Corp. Released 2021. IBM SPSS Statistics for Windows, Version 28.0. Armonk, NY: IBM Corp

Islam, N., Sharp, S. J., Chowell, G., Shabnam, S., Kawachi, I., Lacey, B., Massaro, J. M., D'Agostino, R. B., & White, M. (2020). Physical distancing interventions and incidence of coronavirus disease 2019: natural experiment in 149 countries. *BMJ*, m2743. https://doi.org/10.1136/bmj.m2743

Juhn, Y. J., Wi, C. I., Ryu, E., Sampathkumar, P., Takahashi, P. Y., Yao, J. D., Binnicker, M. J., Natoli, T. L., Evans, T. K., King, K. S., Volpe, S., Pirçon, J. Y., Silvia Damaso, & Pignolo, R. J. (2021). Adherence to Public Health Measures Mitigates the Risk of COVID-19 Infection in Older Adults: A Community-Based Study. *Mayo Clinic Proceedings*, *96*(4), 912–920. https://doi.org/10.1016/j.mayocp.2020.12.016

Kachanoff, F., Bigman, Y. E., Kapsaskis, K., & Gray, K. (2020). Measuring realistic and symbolic threats of COVID-19 and their unique impacts on Well-Being and adherence to public health behaviors. *Social Psychological and Personality Science*, *12*(5), 603–616. https://doi.org/10.1177/1948550620931634

Kim, M., Kim, Y. J., Park, S. J., Kim, K. G., Oh, P. C., Kim, Y. S., & Kim, E. Y. (2021). Machine learning models to identify low adherence to influenza vaccination among Korean adults with cardiovascular disease. *BMC Cardiovascular Disorders*, *21*(1). https://doi.org/10.1186/s12872-021-01925-7

Lafzi, A., Boodaghi, M., Zamani, S., Mohammadshafie, N., & Hasti, V. R. (2021). Analysis of the effectiveness of face-coverings on the death ratio of COVID-19 using machine learning. *Scientific Reports*, *11*(1). https://doi.org/10.1038/s41598-021-01005-y

Lavallee, K. L., Brailovskaia, J., Scholten, S., Schneider, S., & Margraf, J. (2021). Perceptions of Macro- and Micro-Level factors predict COVID-19 Self-Reported Health and Safety Guidelines adherence. *European Journal of Psychology Open*, *80*(4), 152–164. https://doi.org/10.1024/2673-8627/a000016

Listabarth, S., Groemer, M., Waldhoer, T., Vyssoki, B., Pruckner, N., Vyssoki, S., Glahn, A., König-Castillo, D. M., & König, D. (2022). Cognitive decline and alcohol consumption in the aging population—A longitudinal analysis of the Survey of Health, Ageing and Retirement in Europe. *European Psychiatry*, *65*(1). https://doi.org/10.1192/j.eurpsy.2022.2344

Lopes, F., Kitamura, F., Prado, G., Kuriki, P., & Garcia, M. (2021). Machine learning model for predicting severity prognosis in patients infected with COVID-19: Study protocol from COVID-AI Brasil. *PLOS ONE*, *16*(2), e0245384. https://doi.org/10.1371/journal.pone.0245384

Lundberg, S., & Lee, S. (2017). A unified approach to interpreting model predictions. *Neural Information Processing Systems, 30*, 4768–4777.

https://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

Madley-Dowd, P., Hughes, R. A., Tilling, K., & Heron, J. (2019). The proportion of missing data should not be used to guide decisions on multiple imputation. *Journal of Clinical Epidemiology*, *110*, 63–73. https://doi.org/10.1016/j.jclinepi.2019.02.016

Mant, M., Holland, A., & Prine, A. (2021). Canadian university students' perceptions of COVID-19 severity, susceptibility, and health behaviours during the early pandemic period. *Public Health in Practice*, *2*, 100114. https://doi.org/10.1016/j.puhip.2021.100114

Martin, L. R., Williams, S. L., Haskard, K. B., & Dimatteo, M. R. (2005). The challenge of patient adherence. *Therapeutics and clinical risk management*, *1*(3), 189–199.

Murray, M., Morrow, D. & Weiner, M. (2004). A conceptual framework to study medication adherence in older adults. *The American Journal of Geriatric Pharmacotherapy*, *2* (1), 36-43.

Nagler, R. H., Vogel, R. I., Gollust, S. E., Rothman, A. J., Fowler, E. F., & Yzer, M. (2020). Public perceptions of conflicting information surrounding COVID-19: Results from a nationally representative survey of U.S. adults. *PLOS ONE*, *15*(10), e0240776. https://doi.org/10.1371/journal.pone.0240776

Ong, E., Wong, M. U., Huffman, A., & He, Y. (2020). COVID-19 Coronavirus Vaccine Design Using Reverse Vaccinology and Machine Learning. *Frontiers in Immunology*, *11*. https://doi.org/10.3389/fimmu.2020.01581

Osterberg, L., & Blaschke, T. (2005). Adherence to Medication. *New England Journal of Medicine*, *353*(5), 487–497. https://doi.org/10.1056/nejmra050100

Park, C. L., Russell, B. S., Fendrich, M., Finkelstein-Fox, L., Hutchison, M., & Becker, J. (2020). Americans' COVID-19 Stress, Coping, and Adherence to CDC Guidelines. *Journal of General Internal Medicine*, *35*(8), 2296–2303. https://doi.org/10.1007/s11606-020-05898-9

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). SciKit-Learn:

Machine Learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830. https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html

Peng, H., Long, F., & Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 27*(8), 1226–1238. https://doi.org/10.1109/tpami.2005.159

Probst, T. M., Lee, H. J., & Bazzoli, A. (2020). Economic stressors and the enactment of CDC-recommended COVID-19 prevention behaviors: The impact of state-level context. *Journal of Applied Psychology, 105*(12), 1397–1407. https://doi.org/10.1037/apl0000797

Pudjihartono, N., Fadason, T., Kempa-Liehr, A. W., & O'Sullivan, J. M. (2022). A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction. *Frontiers in Bioinformatics, 2*. https://doi.org/10.3389/fbinf.2022.927312

Rayani, M., Rayani, S., & Najafi-Sharjabad, F. (2021). COVID-19-related knowledge, risk perception, information seeking, and adherence to preventive behaviors among undergraduate students, southern Iran. *Environmental Science and Pollution Research*, *28*(42), 59953–59962. https://doi.org/10.1007/s11356-021-14934-y

Rish, I. (Ed.). (2001). An empirical study of the naive Bayes classifier. *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence*.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*(6), 386–408. https://doi.org/10.1037/h0042519

Schaffer, S. & Yoon, S. (2001). Evidence-based methods to enhance medication adherence. *Nurse Practice, 26* (12), 44-54.

Schapire, R. E. (1990). The strength of weak learnability. *Machine Learning*, *5*(2), 197–227. https://doi.org/10.1007/bf00116037

Scherpenzeel, A., Axt, K., Bergmann, M., Douhou, S., Oepen, A., Sand, G., Schuller, K., Stuck, S., Wagner, M., & Börsch-Supan, A. (2020). Collecting survey data among the 50+ population during the COVID-19 outbreak: The Survey of Health, Ageing and Retirement in Europe (SHARE). *Survey Research Methods*, *14*(2), 217–221. https://doi.org/10.18148/srm/2020.v14i2.7738

Schouten, R. M., Zamanzadeh, D., & Singh, P. (2022). *pyampute: a Python library for data amputation* [Dataset]. https://doi.org/10.25080/majora-212e5952-03e

Shanka, M. S., & Menebo, M. M. (2022). When and How Trust in Government Leads to Compliance with COVID-19 Precautionary Measures. *Journal of Business Research*, *139*, 1275–1283. https://doi.org/10.1016/j.jbusres.2021.10.036

Sharma, M., Mindermann, S., Rogers-Smith, C., Leech, G., Snodin, B., Ahuja, J., Sandbrink, J. B., Monrad, J. T., Altman, G. T., Dhaliwal, G., Finnveden, L., Norman, A. J., Oehm, S., Sandkühler, J. F., Aitchison, L., Gavenčiak, T., Mellan, T. A., Kulveit, J., Chindelevitch, L., . . . Brauner, J. M. (2021). Understanding the effectiveness of government interventions against the resurgence of COVID-19 in Europe. *Nature Communications*, *12*(1). https://doi.org/10.1038/s41467-021-26013-4

Sharma, S. (1996). Applied Multivariate Techniques. New York: John Wiley and Sons.

Siewe Fodjo, J. N., Ngarka, L., Njamnshi, W. Y., Nfor, L. N., Mengnjo, M. K., Mendo, E. L., Angwafor, S. A., Atchou Basseguin, J. G., Nkouonlack, C., Njit, E. N., Ahidjo, N., Chokote, E. S., Dema, F., Fonsah, J. Y., Tatah, G. Y., Palmer, N., Seke Etet, P. F., Palmer, D., Nsagha, D. S., . . . Njamnshi, A. K. (2021). COVID-19 Preventive Behaviours in Cameroon: A Six-Month Online National Survey. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3763763

Singh, M., Singh, A., Bharti, S., Singh, P., & Saini, M. (2022). Using Social Media Analytics and Machine Learning Approaches to Analyze the Behavioral Response of Agriculture Stakeholders during the COVID-19 Pandemic. *Sustainability*, *14*(23), 16174. https://doi.org/10.3390/su142316174

Singu, S., Acharya, A., Challagundla, K., & Byrareddy, S. N. (2020). Impact of Social Determinants of Health on the Emerging COVID-19 Pandemic in the United States. *Frontiers in Public Health*, *8*. https://doi.org/10.3389/fpubh.2020.00406

Smaje, A., Weston-Clark, M., Raj, R., Orlu, M., Davis, D., & Rawle, M. (2018). Factors associated with medication adherence in older patients: A systematic review. *Aging medicine (Milton (N.S.W))*, *1*(3), 254–266. https://doi.org/10.1002/agm2.12045

Stanton, A. L. (1987). Determinants of adherence to medical regimens by hypertensive patients. *Journal of Behavioral Medicine*, *10*(4), 377–394. https://doi.org/10.1007/bf00846477

Sturman, D., Auton, J. C., & Thacker, J. (2020). Knowledge of social distancing measures and adherence to restrictions during the COVID-19 pandemic. *Health Promotion Journal of Australia*, *32*(2), 344–351. https://doi.org/10.1002/hpja.443

Szwarcwald, C. L., Souza Júnior, P. R. B. D., Malta, D. C., Barros, M. B. D. A., Magalhães, M. D. A. F. M., Xavier, D. R., Saldanha, R. D. F., Damacena, G. N., Azevedo, L. O., Lima, M. G., Romero, D., Machado, S. E., Gomes, C. S., Werneck, A. D. O., Silva, D. R. P. D., Gracie, R., & Pina, M. D. F. D. (2020). Adesão às medidas de restrição de contato físico e disseminação da COVID-19 no Brasil. *Epidemiologia e Serviços de Saúde*,

Tamirat, T., & Abute, L. (2021). Adherence towards COVID-19 prevention measures and associated factors in Hossana town, South Ethiopia, 2021. *International Journal of Clinical Practice*, *75*(12). https://doi.org/10.1111/ijcp.14530

Tong, K. K., Chen, J. H., Yu, E. W., & Wu, A. M. S. (2020). Adherence to COVID-19 precautionary measures: applying the health belief model and generalised social beliefs to a probability community sample. *Applied Psychology: Health and Well-being*, *12*(4), 1205–1223. https://doi.org/10.1111/aphw.12230

Turk, D. & Meichenbaum, D. (1991). Adherence to Self-Care regimens: The Patient's Perspective. J. Sweet et al. (Eds.), *Handbook of Clinical Psychology in Medical Settings* (pp. 249-267). New York: Lenum Press.

Van Den Broek-Altenburg, E., & Atherly, A. (2020). Adherence to COVID-19 Policy Measures: Behavioral Insights from the Netherlands and Belgium. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3692644

Varghese, N. E., Sabat, I., Neumann-Böhme, S., Schreyögg, J., Stargardt, T., Torbica, A., van Exel, J., Barros, P. P., & Brouwer, W. (2021). Risk communication during COVID-19: A descriptive study on familiarity with, adherence to, and trust in the WHO preventive measures. *PLOS ONE*, *16*(4), e0250872. https://doi.org/10.1371/journal.pone.0250872

Vázquez-Nava, F., Vazquez-Rodriguez, E. M., Vazquez-Rodriguez, C. F., Betancourt, N. V. O., Ruiz, O. C., & Rodríguez-Castillejos, G. C. (2020). Risk factors of non-adherence to guidelines for the prevention of COVID-19 among young adults with asthma in a region with a high risk of a COVID-19 outbreak. *Journal of Asthma*, *58*(12), 1630–1636. https://doi.org/10.1080/02770903.2020.1818774

Wirth, R. & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining.

World Health Organization. (2003). *Adherence to Long-term Therapies: Evidence for Action*. World Health Organization.

World Health Organization. (2008). *Closing the Gap in a Generation: Health Equity through Action on the Social Determinants of Health* (1st ed.). World Health Organization.

World Health Organization. (2018). *WHO Housing and Health Guidelines*. World Health Organization.

Yehualashet, S. S., Asefa, K. K., Mekonnen, A. G., Gemeda, B. N., Shiferaw, W. S., Aynalem, Y. A., Bilchut, A. H., Derseh, B. T., Mekuria, A. D., Negash, W., Meseret, W., Shine, S., & Eshete, A. (2021). Predictors of adherence to COVID-19 prevention measure among communities in North Shoa Zone, Ethiopia based on health belief model: A cross-sectional study. *PLOS ONE*, *16*(1), e0246006. https://doi.org/10.1371/journal.pone.0246006

Zhu, J., Zou, H., Rosset, S., & Hastie, T. (2009). Multi-class AdaBoost. *Statistics and Its Interface, 2*, 349–360. https://doi.org/10.4310/sii.2009.v2.n3.a8

Appendices

Appendix I – Recode of adherence to health measures variables

Variable	Description	Original	Recode
cah110_	Health: ever left home during the last 3 months	-2 Refusal -1 Don't know 1 Yes	0 Risk
		5 No	1 Adherence
cah111_3	Health: met more than 5 people outside household during last 3 months	-2 Refusal -1 Don't know 1 Several times a week 2 About once a week	0 Risk
		3 Less than once a week 4 Not at all	1 Adherence
cah111_6	Health: went shopping during the last 3 months	-2 Refusal -1 Don't know 1 Several times a week	0 Risk
		2 About once a week 3 Less than once a week 4 Not at all	1 Adherence
cah111_7	Health: went to post office/bank/public office during the last 3 months	-2 Refusal -1 Don't know 1 Several times a week	0 Risk
		2 About once a week 3 Less than once a week 4 Not at all	1 Adherence
cah111_8	Health: went to restaurant/pub during the last 3 months	-2 Refusal -1 Don't know 1 Several times a week 2 About once a week	0 Risk
		3 Less than once a week 4 Not at all	1 Adherence
Variable	**Description**	**Original**	**Recode**

cah113_	Health: kept distance from others in public during the last 3 months		-2 Refusal -1 Don't know 2 Often 3 Sometimes 4 Never	0 Risk
			1 Always	1 Adherence
cah116_	Health: covered cough/sneeze more during last 3 months compared to first wave		-2 Refusal -1 Don't know 3 Less frequently	0 Risk
			1 More frequently 2 About the same	1 Adherence
cah017_	Health: took drugs or medicine as prevention against COVID-19		-2 Refusal -1 Don't know 5 No	0 Risk
			1 Yes	1 Adherence
cahc117_	Health: has been vaccinated against Covid-19	Recoded into a new variable: cahc117_cahc118_W9. Has been or wants to get vaccinated against COVID-19	-2 Refusal -1 Don't know cahc117__W9 5 No cahc118__W9 3 No, I do not want to get vaccinated cahc118__W9 4 I am still undecided	0 Risk
cahc118_	Health: wants to get vaccinated against Covid-19		cahc117__W9 1 Yes (has been vaccinated) cahc118__W9 1 Yes, I already have a vaccination scheduled cahc118__W9 2 Yes, I want to get vaccinated	1 Adherence
cahc884_	Health: got flu vaccination in last 12 months		-2 Refusal -1 Don't know 5 No	0 Risk
			1 Yes	1 Adherence
cahc119_	Health: had pneumonia vaccination within last 6 years		-2 Refusal -1 Don't know 5 No	0 Risk
			1 Yes	1 Adherence

Appendix II - Aggregated variables for feature reduction

Original variable	Original label	New variable	New label
cah004_1_W9	Health: hip fracture	**CAH004_T**	Health conditions (sum)
cah004_2_W9	Health: diabetes or high blood sugar		
cah004_3_W9	Health: high blood pressure or hypertension		
cah004_4_W9	Health: heart attack or other heart problem		
cah004_5_W9	Health: chronic lung disease		
cah004_6_W9	Health: cancer or malignant tumor		
cah004_7_W9	Health: other illness or health condition		
caph089_1_W9	Health: falling down in last 6 months	**CAPH089_T**	Health - fragility (sum)
caph089_2_W9	Health: fear of falling down in last 6 months		
caph089_3_W9	Health: dizziness, faints or blackouts in last 6 months		
caph089_4_W9	Health: fatigue in last 6 months		
cah007_1_W9	Health: high blood cholesterol drugs taken regularly	**CAH007_T**	Health - drugs (sum)
cah007_2_W9	Health: high bood pressure drugs taken regularly		
cah007_3_W9	Health: coronary or cerebrovascular diseases drugs taken regularly		
cah007_4_W9	Health: other heart diseases drugs taken regularly		
cah007_5_W9	Health: diabetes drugs taken regularly		
cah007_6_W9	Health: chronic bronchitis drugs taken regularly		
cah007_7_W9	Health: asthma drugs taken regularly		
cac103_2_W9	COVID-19: spouse or partner had symptoms	**CAC103_T**	COVID-19 - relatives had symptoms (sum)
cac103_3_W9	COVID-19: parent had symptoms		
cac103_4_W9	COVID-19: child had symptoms		
cac103_5_W9	COVID-19: other household member had symptoms		
cac103_6_W9	COVID-19: other relative outside household had symptoms		
cac103_7_W9	COVID-19: neighbor, friend or colleague had symptoms		
cac103_8_W9	COVID-19: caregiver had symptoms		
cac103_97_W9	COVID-19: other had symptoms		
cac105_2_W9	COVID-19: spouse or partner tested positive	**CAC105_T**	COVID-19 - relatives tested positive (sum)
cac105_3_W9	COVID-19: parent tested positive		
cac105_4_W9	COVID-19: child tested positive		
cac105_5_W9	COVID-19: other household member tested positive		
cac105_6_W9	COVID-19: other relative outside household tested positive		
cac105_7_W9	COVID-19: neighbor, friend or colleague tested positive		
cac105_8_W9	COVID-19: caregiver tested positive		
cac105_97_W9	COVID-19: other tested positive		
cac120_1_W9	COVID-19: fatigue attributed to respondent's Covid illness	**CAC120_T**	COVID-19 - symptoms attributed to Covid-19 (sum)
cac120_2_W9	COVID-19: cough, congestion, shortness of breath attributed …		
cac120_3_W9	COVID-19: loss of taste or smell attributed to respondent's …		
cac120_4_W9	COVID-19: headache attributed to respondent's Covid illness		
cac120_5_W9	COVID-19: body aches, joint pain attributed to respondent's …		
cac120_6_W9	COVID-19: chest or abdominal pain attributed to respondent's …		
cac120_7_W9	COVID-19: diarrhoea, nausea attributed to respondent's Covid illness		
cac120_8_W9	COVID-19: confusion attributed to respondent's Covid illness		
cac120_97_W9	COVID-19: other long-term or lingering effects attributed to …		
cac111_2_W9	COVID-19: spouse or partner hospitalized	**CAC111_T**	COVID-19 - relatives hospitalized (sum)
cac111_3_W9	COVID-19: parent hospitalized		
cac111_4_W9	COVID-19: child hospitalized		
cac111_5_W9	COVID-19: other household member hospitalized		
cac111_6_W9	COVID-19: other relative outside household hospitalized		
cac111_7_W9	COVID-19: neighbor, friend or colleague hospitalized		
cac111_8_W9	COVID-19: caregiver hospitalized		
cac111_97_W9	COVID-19: other hospitalized		

Appendix III - Variables selected through the use of MRMR (30 variables)

Feature	Label	Type
country_W9	Country identifier	Categorical
cadn003__W9	Intro: year of birth	Ordinal
caph003__W9	Health: rating of subjective health	Ordinal
cah102__W9	Health: change in health in the last 3 months	Ordinal
caph105__W9	Health: limitations because of a health problem in last 6 months	Ordinal
camh037__W9	Health: how often feelings of loneliness	Ordinal
cac113__W9	COVID-19: anyone died due to COVID-19	Dichotomic
caq105__W9	Healthcare: forwent medical treatment since last interview/July 2020	Dichotomic
caq110__W9	Healthcare: medical appointment postponed due to COVID-19 since last interview/J	Dichotomic
caq125__W9	Healthcare: treated in hospital since last interview/July 2020	Dichotomic
caq120__W9	Healthcare: visited doctor/medical facility other than hospital since last inter	Dichotomic
caq122__W9	Healthcare: satisfaction with treatment at medical facility	Ordinal
caw102__W9	Work: unemployed, laid off or business closed since last interview/July 2020	Categorical
caep100__W9	Work: retired after outbreak of Corona	Categorical
caw117__W9	Work: felt safe at work	Categorical
caw121__W9	Work: worked shorter hours since last interview/July 2020	Categorical
caw124__W9	Work: worked longer hours since last interview/July 2020	Categorical
cae103__W9	Economic: received financial support due to outbreak since last interview/July 2	Dichotomic
cas103_2_W9	Social: contact frequency with own parents during last 3 months	Ordinal
cas103_4_W9	Social: contact frequency with neighbors/friends/colleagues during last 3 months	Ordinal
cas125__W9	Social: received regular home care during last 3 months	Dichotomic
cas126__W9	Social: difficulties obtaining home care during last 3 months	Ordinal
cait104__W9	Social: use of internet for e-mailing, etc. since outbreak	Dichotomic
caph003__W8	Health: how was your health before the outbreak	Ordinal
casp.8_LT	casp.8: CASP: quality of life and well-being index	Scale
recall_1.8_LT	recall_1.8: Recall of words, first trial (based on cf008tot)	Scale
orienti.8_LT	orienti.8: Orientation to date, month, year and day of week	Ordinal
numeracy_2.8_LT	numeracy_2.8: Score of second numeracy test (subtraction)	Ordinal
CAPH089_T	Health - fragility (sum)	Scale
CAH007_T	Health - drugs (sum)	Scale