Identifying the Effects of COVID-19 on Psychological Well-being through Unsupervised Clustering for Mixed Data

Katharina Lingelbach[1,2*], Sabrina Gado[3], Doris Janssen[3], Daniela Piechnik[3], Martin Eichler[4], Dennis Knopf[5], Leopold Hentschel[4], Markus Schuler[6,7], Daniel Sernatinger[5], and Matthias Peissner[3]

[1] University of Stuttgart, Institute of Human Factors and Technology Management (IAT), 70569 Stuttgart, Germany
[2] Department of Applied Neurocognitive Psychology, University Oldenburg, 26129 Oldenburg, Germany
[3]Fraunhofer IAO, Fraunhofer Institute for Industrial Engineering IAO, 70569 Stuttgart, Germany
[4]National Center for Tumor Diseases (NCT/UCC), University Hospital Dresden, Technical University Dresden, 01307 Dresden, Germany
[5]seracom GmbH, 70178 Stuttgart, Germany
[6]Helios Hospital Emil von Behring, Department of Oncology, 14165 Berlin, Germany
[7]Medical Faculty, Technical University Dresden, 01307 Dresden, Germany
*katharina.lingelbach@iat.uni-stuttgart.de

Abstract. The COVID-19 pandemic has a strong worldwide impact on not only the health and economic sector but also the (socio-)psychological functioning of individuals. Since psychological health is an important protective factor to prevent diseases, it is crucial to identify individuals with increased vulnerability during the crisis. 275 adults participated in a German online survey from April until August 2020 which investigated health-related, social, behavioral, and psychological effects of the COVID-19 pandemic. We here introduce an unsupervised clustering approach suitable for mixed data types by combining the Gower distance with the Partitioning Around Medoids (PAM) algorithm k-Medoids on this real-world dataset. We could identify three clusters differing significantly in subjects' well-being, psychological distress and current financial and occupational concerns. The clusters also differed in age with younger persons reporting greater financial and occupational concerns, increased anxiety, higher psychological distress and reduced subjective well-being. To explore which features had a strong impact on the clustering, we applied supervised machine learning algorithms for a three-class classification problem and examined the features using a wrapping method and the feature importance implemented in the random forest. The models were able to recognize cluster affiliation with a high accuracy. Particularly, answers regarding the financial and occupational concern, psychological distress and current well-being were decisive for the assignment to a cluster. In sum, the clustering approach can identify persons with weakened psychological protective factors allowing to provide tailored recommendations for preventive actions based on the cluster affiliation, e.g., via a web application.

Keywords: Psychological well-being, COVID-19 pandemic, Machine learning, Mixed data clustering, k-Medoids, Gower distance, Data mining.

1 Introduction

The COVID-19 pandemic not only has a major worldwide impact on health and economic sectors, but also on psychological well-being and quality of life, e.g., due to social isolation. The study of these effects is particularly relevant, since social and psychological resources are known to be important protective factors for diseases [1,2]. The possibility to identify subgroups in the population especially vulnerable during or due to the consequences of the COVID-19 pandemic would allow for fast implementation of targeted preventive interventions. Almost immediately after the outbreak several national and international projects started investigating not only health-related and epidemiological effects but also the impact on the economic, social and psychological status within longitudinal surveys [3–6]. Since most of the surveys are embedded in larger household panel surveys such as the German Socio-Economic Panel (SOEP), they further allow to compare effects of the pandemic with results from former survey-waves as a reference. Pierce et al. [5] examined longitudinal changes in mental health from 53,351 adults in the UK using regression models. The study revealed an increasing prevalence of clinically significant mental disorders in the population from 18.9 % in 2018 and 2019 to 27.3 % in April 2020. In addition, values of the established General Health Questionnaire (GHQ-12) increased from 11.5 in 2018 and 2019 to 12.6 in April 2020 indicating more stress and lower mental health. The increase was particularly large among individuals between 18 and 34 years, women and people living with young children.

Li et al. [7] proposed an alternative approach to examine the psychological effects of the COVID 19 pandemic by analyzing posts in the Chinese online social network Weibo of 17,865 active members using machine learning (ML) methods. They compared relevant features including word frequency and content-based scores for emotional and psychological constructs (e.g., anxiety, depression, and life satisfaction) before and after the outbreak of the pandemic. The authors observed that negative emotion scores (e.g., anxiety, depression and indignation) and sensitivity to social risks increased, while positive emotion scores (e.g., happiness) and life satisfaction decreased after the onset of the pandemic. While online users reported increased concern regarding their health and family, leisure-related topics were addressed less often [7].

Jha et al. [8] proposed a promising approach to examine the impact of the COVID-19 pandemic on emotional, psychological and social well-being using Bayesian networks and classical ML methods. They used data of a representative sample of 17,764 adults in the United States of America to identify key factors for mental health during the COVID-19 pandemic. The results indicate that women perceived higher stress levels compared to men. In addition, individuals aged between 18-29 years showed a higher vulnerability to anxiety than other age groups. Subjects who already suffered from chronic mental disorders showed a higher susceptibility to develop further mental disorders during the COVID-19 pandemic. Using ML, the authors were able to estimate the level of mental health with an accuracy of ~80% [8].

Based on these findings, there is a great need for further research to confirm preliminary findings regarding the psychological, emotional and social effects of the COVID-19 pandemic and to identify individuals vulnerable in their psychological well-being and with a special disposition for mental disorders during and due to the current living conditions under COVID-19.

ML approaches in particular have the ability to learn rules which identify complex relationships between variables, further called features, in large datasets.

Cluster analyses are data-driven statistical classification techniques to reveal which elements in a dataset are similar with respect to previously selected features [9,10]. Most clustering procedures assign elements to clusters based on their distance to the cluster centers called centroids. Many well-established algorithms are suited to cluster numerical and independent data [11]. However, real-world datasets often comprise mixed data types also including ordinal (categorical but in an ordered structure) and nominal (categorical without an ordered structure) features [12]. Major challenges for clustering techniques of large mixed datasets are to a) define an adequate measure of distance for the combination of numerical and categorical data types in a multi-dimensional space, b) choose an appropriate cluster center initialization method, and c) identify the optimal number of clusters [13].

The Gower distance is suggested to be a suitable metric to estimate similarity between elements in mixed datasets [14,15].

It comprises several distance metrices, which are calculated depending on the data type ranging from 0 to 1 between each element. For numerical interval and ordinal data (ranked and adjusted for ties), the range-normalized Manhattan distance is typically used [14,16]. For nominal data, the variables are dummy coded first by assigning either a 0 or 1 for each category depending on the absence or presence of the qualitative attribute. In the second step, the Dice coefficient representing the distance between categories (0 = equal) is calculated using the following formula

$$Dice\ distance = \frac{NNEQ}{NTT+NNZ} \tag{1}$$

with *NNEQ* being the number of non-equal dimensions which is the sum of the number of dimensions in which the first value is 1 (true) and the second is 0 (false); *NTT* representing the number of dimensions in which both values are true and *NNZ* being the number of nonzero dimensions which is the sum of the number of dimensions in which (1) the first value is 1 (true), (2) the second is 0 (false), and (3) both values are 1 (true) [14,17,18].

The individual distances between the *k* variables of two elements calculated above are combined to the Gower distance. It is defined as the average similarity S_{ij} between the two elements *i* and *j* for each variable *k* [14]

$$S_{ij} = \frac{\sum_k^n \delta_{ijk} s_{ijk}}{\sum_k^n \delta_{ijk}} \tag{2}$$

where δ_{ijk} represents a quantity, which is equal to 1 if the *k*-th variable can be compared for the elements *i* and *j*; otherwise it is set to 0.

The matrix of the Gower distance contains the scores of all pair-wise element comparisons.

Combining the Gower distance with a Partitioning Around Medoids (PAM) algorithm such as the k-Medoids [19] is suggested to be well suited to identify evenly distributed clusters in mixed data sets [12,20–22]. The medoid of a cluster in PAM algorithms represents the most centrally located object, which is the median, in a cluster [19]. Consequently, the objects with the smallest sum of relative distances to all other objects are selected as initial medoids and the initial clusters are obtained by assigning each object to the nearest medoid [19].

In this study an explorative data analysis is used to cluster subjects differing in their psychological well-being during COVID-19. For this purpose, we applied an unsupervised clustering approach combining the Gower distance with the PAM algorithm k-Medoids to a real-world mixed dataset.

2 Methods

2.1 Description of the survey and data collection

The data was extracted from a still running German online survey called *WIBCE* (German: *Was ich bei Corona erlebe*; English: *What I experience during Corona*) which started at the 1st of April 2020. Here, responses until the 24th of August were analyzed. Participants could voluntarily fill in the online questionnaire as many times as they liked even on a daily basis. However, some modules (e.g., items examining depression) of the questionnaire were provided at a defined time interval (e.g., with at least seven days in-between). Since answers to the questions were not mandatory, the number of provided data differs for each feature. The items of the online questionnaire comprise epidemiological and health-related information (general as well as COVID-19 specific, e.g., whether a person is at risk for severe illness), the behavior in everyday life during the pandemic, social factors (e.g., the number and form of contacts with others) as well as psychological factors.

To access psychological well-being and related concepts, we used established scales and questionnaires, that are the Patient Health Questionnaire-4 (PHQ-4, [23]) for psychological distress including an anxiety [24] and depression [25] dimension as well as the EQ-5D-5L examination of health-related quality of life [26–28]. In addition, we asked participants to rate their current well-being, concerns regarding an infection with SARS-CoV-2 and their future financial and occupational situation. To quantify participant's contact behavior, we computed a score based on contact related items (e.g., numbers of contacts to other persons in the last 24 hours, contacts at work).

An overview of the features, their ranges, and meaning is provided in **Table 1**. For the cluster analysis, only the first entry of a person and only participants providing answers to all variables related to psychological well-being were included.

Table 1. Overview of the features.

Feature name	Interpretation
Features used in the Cluster Analyses	
Current well-being	Values range from 0 (bad) to 4 (excellent).*
Occupational concern	Values range from 1 to 4 where 1 indicates no concerns at all and 4 representing major concerns regarding the future occupational situation.*
Financial concern	Values range from 1 to 4 where 1 indicates no concerns at all and 4 representing major concerns regarding the future financial situation.*
Concern regarding an infection	Values range from 0 (not concerned at all) to 4 (very concerned).*
Quality of life (EQ-5D-5L Index)	Values range from 0 to 1 where 0 indicates low perceived health-related quality of life.**
Depression (PHQ-2)	Values range from 0 to 6 with higher values indicating greater perceived depression.*
Anxiety (GAD-2)	Values range from 0 to 6 with higher values indicating greater perceived anxiety.*
Additional Features used in the Prediction of Cluster Affiliation	
Age	Values ranged from 18 to 74.*
Sex	Dummy variable (0 and 1). The value 0 is assigned to a female while 1 is assigned to a male participant.
Psychological Distress (PHQ-4)	It is a sum score of the PHQ-2 and GAD-2 with a range from 0 to 12. Higher values indicate greater psychological distress. *
Education	Values range from 0 representing no school leaving certificate to 4 indicating a qualification for university entrance.*
Income	Values range from 0 to 8. A value of 0 indicates a net household income below 500€ and 8 a net household income above 7000€.*
COVID-19 risk factor	Dummy variable (0 and 1). A value of 0 indicates that the person is not at a greater risk for a severe infection.
Numbers of risk factors	Values ranged from 0 to 4. It is the sum of reported risk factors for one person.*
Contact to Others	Values ranged from 2 to 25. A higher score indicates more active and frequent contact with other people.*
Multiple Participation in WIBCE	Values ranged from 1 to 85 representing the number of participations of a person in the survey since April 2020.*

*step size = 1, **step size = 0.01

2.2 Description of the sample

A total of 173 participants (39.9 % male) were included in the cluster analysis with an average age of M = 44.81 years (SD = 13.31; range: 18 to 74). On average, they shared a household with three (2.9) other persons and 37.0 % reporting to live together

with under-age children. 5.8 % of the participants reported symptoms associated with COVID-19 and 40.5% reported criteria assigning them to a group at higher risk of severe COVID-19 based on the definition of the Robert Koch Institute [29]. The educational level was rather high compared to the average in the German population with 79.8% graduated at least high school (here the German Abitur). 58.4% reported a net household income of more than 3,000 € per month.

2.3 Cluster analysis

Cluster identification. To identify subsets of people with similar characteristics regarding their psychological protective factors, mental well-being and perceived distress due to or during the COVID-19 pandemic, an unsupervised cluster analysis was used. Since the here investigated dataset comprises not only numerical variables (e.g., age or scores with sufficiently granulated scaling and meaningful order but also ordinal variables (e.g., scores with a short range (0 – 5)) and nominal variables (e.g., sex), we used a clustering approach suitable for mixed datasets. Therefore, the Gower distance as a similarity measure was combined with the PAM algorithm k-Medoids[17] [30,31]. Numerical data was scaled using the MinMaxScaler from sklearn (0.23.2) in Python (3.8.6). The features were further z-standardized within the calculation of the Gower distance using the Gower package (0.0.5). Clusters were identified using the k-Medoids algorithm implemented sklearn_extra (0.1.0b2).

Determination of number of clusters. To determine the number of clusters, several methods have been introduced [32]. Here, we compared the Elbow method and Silhouette coefficient.

In the Elbow method, we calculated the sum of squared errors of the distance between the sample points in each cluster and the centroid of the cluster iteratively for the different number of clusters *k.* Smaller values indicate more homogeneous data within the clusters. The choice of *k* is typically done by inspecting the Elbow plot showing the overall error as a function of the cluster number. Typically, the error drops sharply for small cluster numbers and then suddenly only little error reduction can be gained by further increasing the cluster number. This elbow point is used as number of clusters.

As an alternative, the Silhouette coefficient is a measure of how similar an element (e.g., a person) is to the associated cluster compared to another adjacent one. The silhouette coefficient ranges from -1 to +1, with a high value indicating that the element has a high similarity with its own cluster and a low similarity with the adjacent.
In case of deviating results, we would have explored which number of clusters is more suitable for the dataset by analyzing the clusters' features statistically.

Cluster description. For the description of the resulting clusters and their properties, explorative statistical comparisons for the demographical and psychological features were performed. Differences between the clusters were examined using 5,000-fold bootstrapped 95 % confidence intervals (CI) of the mean value with no overlap between the CI indicating a statistical significance of $p < .01$ and a partial overlap without including the means a statistical significance of $p < .05$ [33].

Prediction of the cluster affiliation. We were further interested which features were particularly informative for the clustering and whether we can predict to which

cluster an individual belongs based on the given responses in the questionnaire. Therefore, supervised ML algorithms were applied using the cluster affiliation as labels and the data of 156 persons (participants with missing values in any of the features were excluded). Four different ML algorithms were compared: (1) a k-nearest neighbor classifier, (2) a logistic regression, (3) a random forest, and (4) a support vector machine. The models were implemented in Python (3.8.6) using sklearn (0.23.2) [17]. The dataset was divided into a training and a test set with a stratified 80:20 split. Hyperparameters of the algorithms were optimized using the training data in a 5-fold cross-validated grid search and evaluated with the performance measure accuracy balanced for the number of samples per class. The feature wrapping method sequential feature selection using the library mlxtend [34] with accuracy as evaluation metric was used to reduce redundant features. The relevance of each feature in the prediction was calculated from an embedded method of the random forest accessible via the attribute *feature importance*. A 5-fold cross-validation with balanced accuracy as performance metric was chosen to evaluate the prediction. For the comparison of the model performance, we used the random level based on the number of classes ($\frac{100}{n_{classes}}$) and a random level estimated by a dummy classifier [17] with the method *stratified* as a reference.

3 Results

3.1 Determination of number of clusters.

We observed a consistent suggestion of k = 3 clusters for the adequate number of clusters with both, the Elbow method and Silhouette coefficient (see **Fig. 1** and **Fig. 2**). **Table 2** provides an overview of the descriptive analysis of the clusters regarding their demographic information. The clusters are comparable with respect to their size as well as numbers of persons in a household, gender and belonging to a COVID-19 risk group.

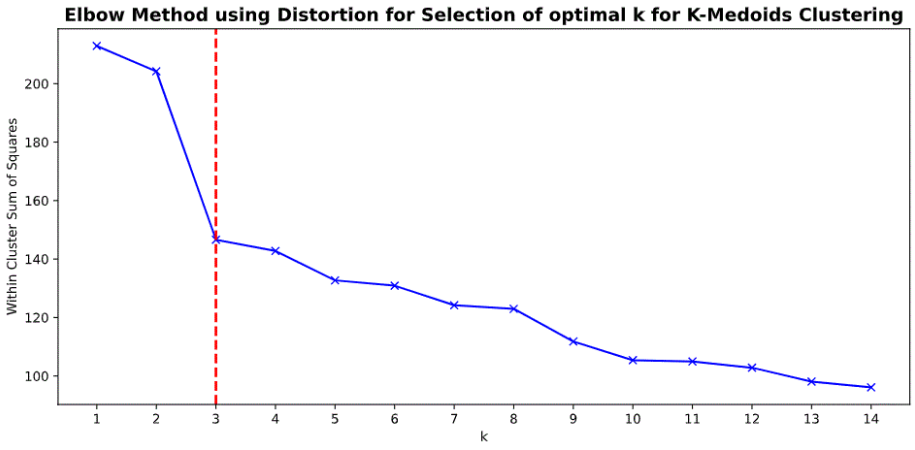


Fig. 1. Elbow plot shows the overall error using the within cluster sum of squares (y-

axis) as a function of the cluster number *k* (x-axis). A bend comparable to an elbow is visible at *k* of three indicating a suitable number of clusters.

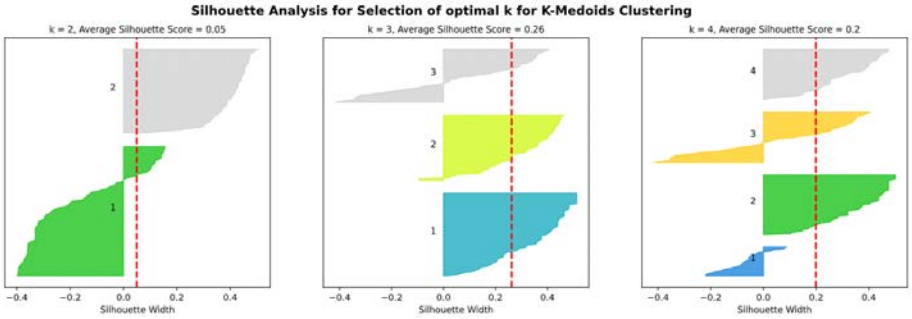


Fig. 2. The Silhouette analysis returns a score of 0.26 at $k = 3$ indicating that three clusters are a good choice. Here the Euclidean distance was used as metric.

Table 2. Overview of the descriptive statistic for the clusters.

Cluster	C1	C2	C3
N	71	56	46
Sex (male in %)	33.8	46.4	41.3
Mean age in years (*SD*)	47.85 (13.09)	44.05 (11.81)	41.04 (14.10)
Mean number of persons in household (under-age)	2.83 (0.72)	3.02 (0.79)	3 (0.5)
Belonging to a COVID-19 risk group in %	40.8	37.5	43.5

3.2 Statistical analysis of the clusters

To explore the differences between the three clusters (C1, C2 and C3), cluster means and their 95% confidence intervals (CI) were calculated for each feature via bootstrapping with 5,000 repetitions. Results are provided in **Fig. 3** and **Table 3**. A significant difference for age with $p < .05$ between C1 (older) and C2, C3 (younger) C1 was found. Significantly more female participants were assigned to C1 compared to C2 ($p < .05$) but not C3. There is no difference in the distribution of female and male participants between C2 and C3. The clusters differ noticeably in their ratings regarding the current wellbeing with the highest values in C1, decreased well-being in C2 (C1 vs. C2, $p < .05$) and lowest well-being in C3 (C3 vs. C1 and C2, $p < .01$). C1 reported the least concern regarding the occupational and financial situation followed by C2 and C3 (all with $p < .01$). C3 showed the significantly highest concern. Interestingly, C3 on average reported a significantly lower income than C1 and C2 (with $p < .05$). There was a

difference between C3 and C1, C2 indicating higher concerns regarding an infection with the SARS-CoV-2 in C3 (with $p < .05$). Regarding the quality of life, there was no difference in the ratings between C1 and C2 but a significant difference between both and C3 (with $p < .01$) with the least quality of life in the latter. Analyzing the two dimensions (1) depression with the PHQ-2 and (2) anxiety with the GAD-2, we observed a significant difference between C1 and C2 ($p < .05$) as well as between C1 and C2 to C3 ($p < .01$). C1 reported the lowest values for depression and anxiety and C3 reported to be significantly more depressed and anxious. For the sum score PHQ-4 of the two dimensions indicating psychological distress, there was a significant difference between all clusters at $p < .01$ with the lowest values in C1 and highest psychological distress in C3. There are no differences between the clusters for education, belonging to a COVID-19 risk group, and contact to others. However, in case the participants reported to belong to a COVID-19 risk group, participants of C3 reported on average more factors indicating their belonging to the risk group compared to C2 ($p < .05$) but not C1. Finally, participants of the C1 filled in the survey more often compared to C3 ($p < .05$). To conclude, C3 revealed the least well-being and quality of life, higher depression, anxiety and psychological distress. Furthermore, it reported higher concerns regarding a) the occupational and b) financial situation as well as c) a possible infection with the SARS-CoV-2.

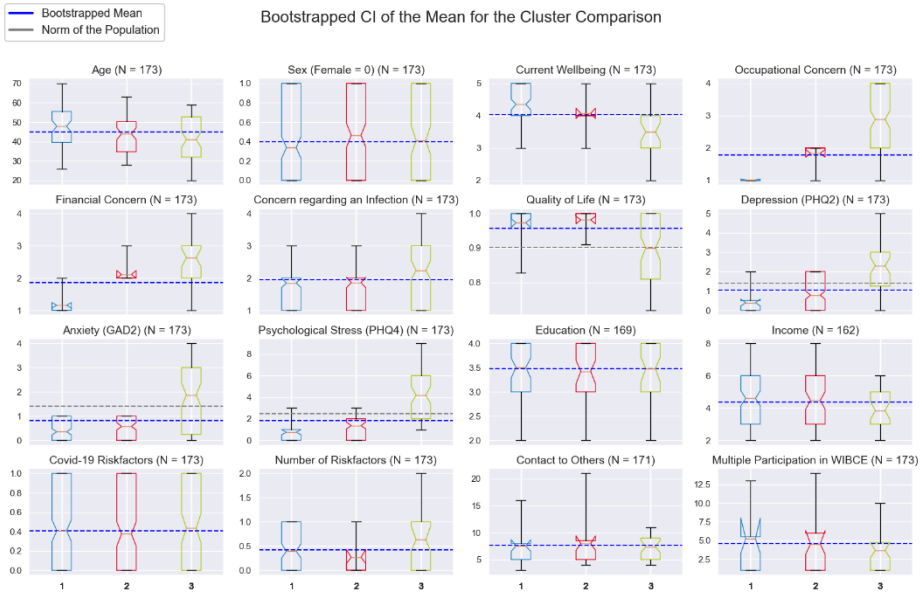


Fig. 3. Bootstrapped means and 95% confidence intervals (CI) of the clusters (blue, red and green) for each feature. Notches in the boxes of the plot visualize the upper and lower boundary of the CI with the orange line representing the mean of each cluster and the blue line representing the overall mean of the clusters. The grey line represents a population norm as reference with *M* = 0.902 for the EQ-5D-5L quality of life index [35], *M* = 1.4 for the PHQ-2 [25], *M* = 1.4 for the GAD-2 [24], and *M* = 2.5 for the PHQ-4 [23]. The box comprises 50% of the distribution from the 25th to the 75th quartile. The ends of the whiskers represent the 5th and 95th quartile of the distribution.

Table 3. Bootstrapped means and 95 % CI with the lower and upper bound for the features and clusters.

Feature	C1	C2	C3
Age	47.84 [44.73, 50.92]	44.00 [40.98, 47.14]	41.03 [36.98, 45.09]
Sex	0.34 [0.23, 0.45]	0.46 [0.34, 0.59]	0.41 [0.26, 0.57]
Current Well-being	4.35 [4.18, 4.51]	4.07 [3.91, 4.23]	3.50 [3.22, 3.78]
Occupational concern	1.01 [1, 1.04]	1.84 [1.71, 1.95]	2.89 [2.61, 3.15]
Financial concern	1.16 [1.07, 1.24]	2.11 [2, 2.23]	2.63 [2.35, 2.91]
Concern regarding an infection	1.84 [1.70, 1.99]	1.86 [1.70, 2.02]	2.24 [1.96, 2.54]
Quality of life (EQ-5D-5L Index)	0.97 [0.96, 0.98]	0.98 [0.97, 0.99]	0.90 [0.86, 0.93]
Depression (PHQ-2)	0.40 [0.23, 0.58]	0.78 [0,57, 1]	2.30 [1.89, 2.72]
Anxiety (GAD-2)	0.35 [0.23, 0.48]	0.57 [0.43, 0.73]	1.87 [1.39, 2.37]

Psychological Distress (PHQ-4)	0.75 [0.52, 1]	1.36 [1.05, 1.70]	3.41[4.18, 5.02]
Education	3.50 [3.29, 3.70]	3.43 [3.17, 3.65]	3.48 [3.20, 3.72]
Income	4.60 [4.13, 5.07]	4.47 [3.92, 5.02]	3.84 [3.30, 4.40]
COVID-19 risk factor	0.41 [0.30, 0.52]	0,38 [0.25, 0.5]	0.43 [0.28, 0.59]
Numbers of risk factors	0.39 [0.25, 0.55]	0.27 [0.14, 0.43]	0.63 [0.39, 0.91]
Contact to Others	7.53 [6.52, 8.68]	7.97 [6.73, 9.40]	7.36 [6.27, 8,6]
Multiple Participation in WIBCE	5.25 [3.30, 8.00]	4.54[3.16, 6.32]	3.66 [2.70, 4.74]

3.3 Prediction of the Cluster affiliation

For an explainable machine learning approach, it is of particular value to know the features' relevance in the clustering. Therefore, we applied supervised machine learning to examine whether we can predict the cluster affiliation with a high accuracy and use a feature wrapping method and the feature importance implemented in the random forest. **Table 4** and **Fig. 4** show the performances of the here explored algorithms and their hyperparameters chosen by the grid search. The chance level estimated by the dummy classifier was set to 0.313. The feature wrapping method selected (1) *current well-being, concern* regarding the (2) *occupational* and (3) *financial situation* and (4) *psychological distress (PHQ-4)* as informative features for almost all algorithms. Only the logistic regression algorithm chose an additional feature (5) *age*. For the prediction of cluster label affiliation to the three clusters, the random forest performed best with a high accuracy score of 94.4% (*SD*=0.07). The estimated feature importance ranked *concerns regarding the occupational situation* as feature with the highest impact on the prediction (score of 0.392), followed by *concerns regarding the financial situation* (score of 0.254), *psychological distress (PHQ-4;* score of 0.226), and *current well-being* (score of 0.128).

Table 4. Overview of the average model performance estimated from the cross-validation with the balanced accuracy metric for the training and test set for each algorithm for the prediction of cluster affiliation. KNN = k-Nearest Neighbor, LR = Logistic Regression, RF = Random Forest, SVM = Support Vector Machine.

Algorithm	*M (SD)* Training	*M (SD)* Test	*Hyperparameter*
KNN	0.845 (0.052)	0.878 (0.194)	n_neighbors = 4, leaf_size = 8, algorithm = ball_tree
LR	0.867 (0.081)	0.644 (0.103)	C = 0.234, penalty = l1, solver = saga
RFC	0.851 (0.057)	0.944 (0.070)	max_depth = 7, max_features = auto, n_estimators = 200, criterion = gini
SVM	0.879 (0.032)	0.889 (0.141)	C = 1, gamma = 0.001, kernel = linear

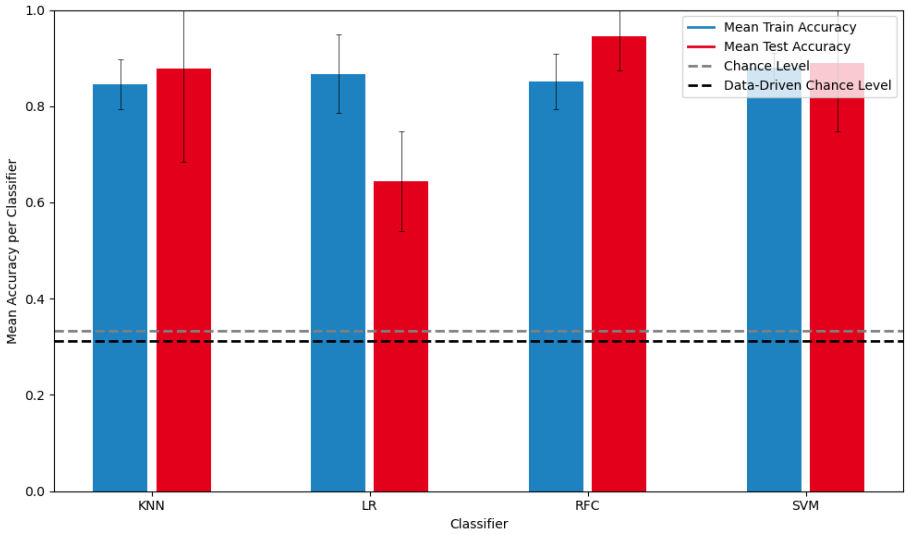


Fig. 4. Overview of the average model performance using the training set (blue) and test set (red). The grey dotted line marks the chance level based on the number of clusters (0.33). The black dotted line indicates the chance level (0.313) estimated by the dummy classifier. KNN = k-Nearest Neighbor, LR = Logistic Regression, RF = Random Forest, SVM = Support Vector Machine.

4 Discussion

By combining the Gower distance with the PAM algorithm k-Medoids, we could identify three clusters differing significantly in subjects' well-being, psychological distress and current financial and occupational concerns during the COVID-19 pandemic. Further, our cluster comparison revealed that younger persons were assigned to the cluster reporting greater financial and occupational concerns, increased anxiety, higher psychological distress, reduced well-being and quality of life. Especially the financial and occupational concerns, psychological distress and current well-being were decisive variables for the cluster affiliation. Since classifiers' accuracies were high, we can assume that the features used to predict the cluster labels were also most informative for the clustering. While the sequential feature selection extracts the relevant features independently from each other, the random forest's feature importance includes interdependencies in its selected subset. The here proposed clustering approach performed well on the mixed dataset. However, further research could investigate other clustering algorithms (e.g., DBSCAN or fuzzy k-modes) and non-random initialization methods. Previous studies reported increased mental disorders, distress and decreased mental health [5,8]; especially in individuals between 18 and 34 years [5,8], women [5,8], persons already diagnosed with mental disorders [8] and people living with young children [5]. We observed comparable findings in our C3. In comparison to reference values of

a norm population [23,25], the scores for depression (PHQ-2) and psychological distress (PHQ-4) were significantly higher in C3 but not in C1 and C2. The increased concerns regarding the financial and occupational situation might be explained by the prevailing poor hiring opportunities and the economic recession in Germany. Interestingly, the clusters did not differ in their number of social contacts to other persons. Hence, either the reduced contact to other persons due to social distancing actions did not have a strong influence on psychological health and could not explained differences in psychological health between the clusters; or the C3 suffered particularly strongly due to the social distancing. However, whether the low psychological well-being and increased perceived distress in C3 is a consequence of the COVID-19 pandemic or a phenomenon observed in the young population in general is not clearly distinguishable. Nevertheless, our unsupervised clustering approach could identify individuals with reduced psychological protective factors and increased psychological distress. This allows to provide cluster-tailored recommendations, e.g., via a web-application, for preventive actions potentially helping to alleviate concerns, decrease perceived distress and improve psychological well-being as well as quality of life.

References

1. Luanaigh, C.O., Lawlor, B.A.: Loneliness and the health of older people. International journal of geriatric psychiatry **23**, 1213–1221 (2008).
2. Valtorta, N.K., Kanaan, M., Gilbody, S., Ronzi, S., Hanratty, B.: Loneliness and social isolation as risk factors for coronary heart disease and stroke: systematic review and meta-analysis of longitudinal observational studies. Heart (British Cardiac Society) **102**, 1009–1016 (2016).
3. Kühne, S., Kroh, M., Liebig, S., Zinn, S.: The Need for Household Panel Surveys in Times of Crisis: The Case of SOEP-CoV. 195-203 Pages / Survey Research Methods, Vol 14 No 2 (2020): Survey Research Methods During the COVID-19 Crisis / Survey Research Methods, Vol 14 No 2 (2020): Survey Research Methods During the COVID-19 Crisis (2020).
4. COSMO Germany: German COVID-19 Snapshot MOnitoring (COSMO Germany). PsychArchives Preprints. https://www.psycharchives.org/handle/20.500.12034/2398 (2020).
5. Pierce, M., Hope, H., Ford, T., Hatch, S., Hotopf, M., John, A., Kontopantelis, E., Webb, R., Wessely, S., McManus, S., Abel, K.M.: Mental health before and during the COVID-19 pandemic: a longitudinal probability sample survey of the UK population. The Lancet Psychiatry **7**, 883–892 (2020).
6. Panchal, N., Kamal, R., Orgera, K., Cox, C., Garfield, R., Hamel, Liz, Munana, C., Chidambaram, P.: The Implications of COVID-19 for Mental Health and Substance Use. https://www.kff.org/coronavirus-covid-19/issue-brief/the-implications-of-covid-19-for-mental-health-and-substance-use/ (2020). Accessed 19 November 2020.

7. Li, S., Wang, Y., Xue, J., Zhao, N., Zhu, T.: The Impact of COVID-19 Epidemic Declaration on Psychological Consequences: A Study on Active Weibo Users. International journal of environmental research and public health **17** (2020).
8. Jha, I.P., Awasthi, R., Kumar, A., Kumar, V., Sethi, T.: Learning the Mental Health Impact of COVID-19 in the United States with Explainable Artificial Intelligence (2020).
9. Clatworthy, J., Buick, D., Hankins, M., Weinman, J., Horne, R.: The use and reporting of cluster analysis in health psychology: a review. British journal of health psychology **10**, 329–358 (2005).
10. Romesburg, H.C.: Cluster analysis for researchers. Lulu Press, S.l. (2004).
11. Nagpal, A., Jatain, A., Gaur, D.: Review based on data clustering algorithms. In: 2013 IEEE Conference on Information & Communication Technologies (ICT 2013). Thuckalay, Tamil Nadu, India, 11 - 13 April 2013, pp. 298–303. IEEE, Piscataway, NJ (2013).
12. van de Velden, M., Iodice D'Enza, A., Markos, A.: Distance-based clustering of mixed data. WIREs Comput Stat **11**, e1456 (2019).
13. Ahmad, A., Khan, S.S.: Survey of State-of-the-Art Mixed Data Clustering Algorithms. IEEE Access **7**, 31883–31902 (2019).
14. Gower, J.C.: A General Coefficient of Similarity and Some of Its Properties. Biometrics **27**, 857 (1971).
15. Kesh, S.: How to Calculate Gower's Distance Using Python. https://medium.com/analytics-vidhya/concept-of-gowers-distance-and-it-s-application-using-python-b08cf6139ac2 (2020). Accessed 21 September 2020.
16. Grané, A., Salini, S., Verdolini, E.: Robust multivariate analysis for mixed-type data: Novel algorithm and its practical application in socio-economic research. Socio-Economic Planning Sciences, 100907 (2020).
17. 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.: Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011).
18. Dice, L.R.: Measures of the Amount of Ecologic Association Between Species. Ecology **26**, 297–302 (1945).
19. Park, H.-S., Jun, C.-H.: A simple and fast algorithm for K-medoids clustering. Expert Systems with Applications **36**, 3336–3341 (2009).
20. Madhulatha, T.S.: Comparison between K-Means and K-Medoids Clustering Algorithms. In: Wyld, D.C., Wozniak, M. (eds.) Advances in computing and information technology. First international conference, ACITY 2011, Chennai, India, July 15 - 17, 2011 ; [the First International Conference on Advances in Computing and Information Technology] ; proceedings, vol. 198, pp. 472–481. Springer, Berlin (2011).
21. Budiaji, W., Leisch, F.: Simple K-Medoids Partitioning Algorithm for Mixed Variable Data. Algorithms **12**, 177 (2019).

22. Foss, A.H., Markatou, M., Ray, B.: Distance Metrics and Clustering Methods for Mixed-type Data. International Statistical Review **87**, 80–109 (2019).
23. Kroenke, K., Spitzer, R.L., Williams, J.B.W., Löwe, B.: An ultra-brief screening scale for anxiety and depression: the PHQ-4. Psychosomatics **50**, 613–621 (2009).
24. Kroenke, K., Spitzer, R.L., Williams, J.B.W., Monahan, P.O., Löwe, B.: Anxiety disorders in primary care: prevalence, impairment, comorbidity, and detection. Annals of internal medicine **146**, 317–325 (2007).
25. Löwe, B., Kroenke, K., Gräfe, K.: Detecting and monitoring depression with a two-item questionnaire (PHQ-2). Journal of psychosomatic research **58**, 163–171 (2005).
26. Ludwig, K., Graf von der Schulenburg, J.-M., Greiner, W.: German Value Set for the EQ-5D-5L. PharmacoEconomics **36**, 663–674 (2018).
27. Janssen, M.F., Szende, A., Cabases, J., Ramos-Goñi, J.M., Vilagut, G., König, H.H.: Population norms for the EQ-5D-3L: a cross-country analysis of population surveys for 20 countries. The European journal of health economics : HEPAC : health economics in prevention and care **20**, 205–216 (2019).
28. Huber, M.B., Reitmeir, P., Vogelmann, M., Leidl, R.: EQ-5D-5L in the General German Population: Comparison and Evaluation of Three Yearly Cross-Section Surveys. International journal of environmental research and public health **13** (2016).
29. Robert Koch Institute: Informationen und Hilfestellungen für Personen mit einem höheren Risiko für einen schweren COVID-19-Krankheitsverlauf. https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/Risikogruppen.html;jsessionid=D38B99E8897126E54CAA737B4B612B53.internet052?nn=13490888 (2020).
30. Filaire, T.: Clustering on Mixed Type Data: A Proposed Approach Using R. https://towardsdatascience.com/clustering-on-mixed-type-data-8bbd0a2569c3. Accessed 21 September 2020.
31. Shendre, S.: Clustering Datasets Having Both Numerical and Categorical Variables. https://towardsdatascience.com/clustering-datasets-having-both-numerical-and-categorical-variables-ed91cdca0677. Accessed 21 September 2020.
32. Yuan, C., Yang, H.: Research on K-Value Selection Method of K-Means Clustering Algorithm. J **2**, 226–235 (2019).
33. Cumming, G., Finch, S.: Inference by eye: confidence intervals and how to read pictures of data. The American psychologist **60**, 170–180 (2005).
34. Raschka, S.: MLxtend: Providing machine learning and data science utilities and extensions to Python's scientific computing stack. JOSS **3**, 638 (2018).
35. Cabases, J., Szende, A., Janssen, B.: Self-Reported Population Health: An International Perspective based on EQ-5D. Springer, s.l. (2014).