Mobile Networks and Applications (2021) 26:159–173
https://doi.org/10.1007/s11036-020-01696-z

Machine Learning Based Approach for Sustainable Social Protection Policies in Developing Societies

Zahid Mumtaz[1] · Peter Whiteford[1]

Accepted: 19 November 2020 / Published online: 7 January 2021
© Springer Science+Business Media, LLC, part of Springer Nature 2021

Abstract

Machine learning has been increasingly used for making informed public policy decisions, however, its application in the area of social protection in developing societies has been largely overlooked. We have employed unsupervised machine learning K-means clustering technique for exploring a big data that comprised of 88 attributes and 570 instances for better targeting of households that are in urgent need of welfare from the government. The clusters formed showed common patterns relating to insecurities in terms of loss of income and property, unemployment, disasters and disease etc. faced by households in each cluster. We found that households falling in rural areas jurisdictions face severe insecurities compared to other localities and are in urgent need of social protection interventions. We concluded that by employing K-means clustering unsupervised machine learning approach big data (even if it is limited) can be explored effectively for better targeting of social protection interventions for both developing and smart societies. The unsupervised machine learning technique presented in this study is an efficient approach because it can be used by societies that are facing data constraints and can achieve optimal results for increasing the welfare of poor by using the said approach.

Keywords Artificial intelligence · Machine learning · K-means clustering · Big data · Social protection · Smart and developing societies

1 Introduction

We are in an era of information revolution where data are produced and stored in every field at an unprecedented rate [1, 2]. This provides social scientists and policy makers with an opportunity to build and test theories using these latest data analysis techniques [3]. By combining social theory with computer science, we can utilise big data to predict and hopefully answer major problems faced by different societies [4]. The use of artificial intelligence (AI), in the area of public policy is relatively new [5, 6]. One of the AI instruments, which is becoming widely popular is machine learning (ML), [7]. ML techniques emerged primarily from computer science and engineering but recently its application in the area of public policy has increased because of the development of the availability of data, open-source software and sophisticated ML data analysis techniques [8]. Social policy is an important area of public policy which deals with poverty reduction and increasing the welfare of the population [9]. It is aimed at improving the well-being and livelihoods of those who are disadvantaged in a society through a range of mechanisms such as social protection and provisions of health and education etc.[1] [9]. Social policy mechanisms in most high-income countries are well developed and play a vital role in combatting poverty and are considered an essential tool for economic development in these countries [10]. However, social policy outside the developed world is fragmented and governments in developing countries are faced with fiscal, data and capacity constraints which are a major impediment to the effective implementation of social policy instruments such as social protection [11, 12]. In addition, poor targeting of the people

✉ Zahid Mumtaz
Zahid.mumtaz@anu.edu.au

Peter Whiteford
peter.whiteford@anu.edu.au

[1] Crawford school of public policy, College of Asia and Pacific, Australian National University, Canberra, Australian Capital Territory 2601, Australia

[1] As the purpose of this paper is not to explain the social policy therefore a very brief explanation is provided here.

who are in need of welfare and institutional weaknesses is a major impediment in the successful formulation and implementation of social protection programmes in developing countries.

Pakistan is a developing country that has the fifth largest population in the world. Its GDP per capita in 2018 was USD 1482 in purchasing power terms or 132nd out of 189 countries and regions. Around 31% of the population are estimated to live in poverty and its 2018 ranking on the United Nations Development Programme (UNDP), Human Development Index is 152 out of 189 countries. While there are a range of official data sources on living conditions and wellbeing in Pakistan, many of these are dated representing a major data constraint. Therefore, for the purpose of this study, a survey of 570 households from 14 different cities in Pakistan was conducted, which shows more in-depth data[2] on living conditions and their relationships to different forms of social protection. The complexity of this big data collected motivates the use of the latest data analysis techniques to more comprehensively explore and target the disadvantage for the provision of welfare. As a result, this article will propose a novel methodology to explore this large data by using unsupervised machine learning (UML), K-means clustering technique and argue that through the application of this technique, better targeting of the population for social protection interventions can be achieved that will not only be useful for its formulation but also assist in overcoming institutional weaknesses in the developing countries.

The article is structured as follows. First, a review of the literature showing the application of ML techniques in various areas of public policy will be presented. Second, based on the findings from the literature, the lack of the application of ML techniques in the area of social protection is identified as a major gap. Thereafter, we will explain that how this paper will bridge this gap by using UML K-means clustering technique to explore a survey data followed by a brief explanation of the concept of social protection. Third, the methodology of survey data collection used in this article is explained along with reasons for using UML K-means clustering technique to explore this survey data. We will then explain UML K means clustering technique and use this approach to explore the said survey data. We will also compare the results of K-means clustering technique with UML DBSCAN (Density-based spatial clustering of applications with noise) clustering approach for the purpose of utilising the best results. Fourth, the results of four clusters formed by using the UML K-means clustering technique will be explained by using descriptive statistics to show the need and priority of social protection interventions for the households surveyed. Finally, we present conclusion and implications of this study for future research.

2 Use of machine learning in public policy – A review of the literature

ML algorithms such as decision trees, dimension reduction methods, K-means nearest neighbour, support vector models, and penalized regression can be used to improve the effectiveness of public policies that have significant social and economic implications and can go beyond policy management to have a theoretical impact [13–15]. Several studies indicate that ML techniques have been used for making informed decisions in policy areas such as improving health policy, reforming education sector, improving tax policy and addressing climate change issues [14]. The advantage of ML methods over traditional statistical tools is that they provide new approaches to improve estimation of causal effects, which can reduce the reliance of these estimates on modelling assumptions and thereby enhancing the credibility of policy analysis. In addition, ML places great emphasis on model checking (through holdout samples and cross-validation) and model shrinkage (adjusting predictions toward the mean to reduce overfitting) making it a better approach for policy analysis [16]. In the succeeding paragraphs, a review of the studies conducted in the various areas of public policy where ML has been used for policy analysis will be presented.

Burscher et al. [17] used a ML approach for the automatic coding of policy issues to apply it on news articles and parliamentary questions and compared it with human annotations. The results showed that ML algorithms performed better than human coders and generalizations can be made across contexts highlighting implications for methodological advances and empirical theory testing. Andini et al. [14] argue that effectiveness of tax rebate scheme in Italy can be improved by selecting the beneficiaries of the scheme through using ML algorithms. This use of ML approach for targeting the beneficiaries helped in saving 29.5% (about 2 billion euro) of the funds earmarked to the scheme. Kasy [18] in his qusai experiment combines optimal taxation and insurance theory with ML and nonparametric Bayesian decision theory to propose a framework based on a standard social welfare function by using a data set of a health insurance experiment. When the ML algorithms were applied to the dataset, the values obtained for the optimal policy choice through ML were substantially different from those obtained using the standard statistics approach. The results obtained through ML algorithms points toward a large area of potential applications for these methods in informing policy decisions. Ballestar at al. [7] conducted a study in the area of higher education for identifying the long-term effects of research conducted by university researchers by using six years of program data developed in Madrid. They design a ML multilevel model: automated nested longitudinal clustering, to discover on whom, when, and for how long the policies adopted as a result of the research have an effect. They argue that the findings of this study are relevant for

[2] 88 attributes were collected against a household.

government agencies and universities to understand the productivity of academics working under long-term incentive-based programs and for maximizing the generation of knowledge. Chalfin et al. [19] in a similar study used data on teacher tenure decision to show that large social welfare gains can be achieved from using ML tools to predict worker productivity.

Kleinberg et al. [15] and Ashrafian and Darzi [20] argue that ML approach can be utilised for achieving the objectives and social welfare gains of health policy such as creating the conditions that ensure good health, social care for an entire population through preventive strategies, protection from disease, promotion of healthy lifestyles, and population screening through knowledge capture. Brady et al. [21] used a large data set from the US census bureau to compare the performance of ML algorithms with manual classification of public health expenditures to determine, if ML approach could provide a faster and a cheaper alternative. Compared with manual classification, the ML algorithms produced more accurate estimates showing that ML is a time and cost saving tool for estimating public health spending in the US that can be used in public health organizations to evaluate the impact of evidence based public health resource allocation. Pan et al. [22] in their study used administrative data for 6457 women collected by the department of human services, Illinois, for a period of one year to develop a model for adverse birth prediction and improve upon the existing paper-based risk assessment by using ML approach. ML algorithms developed and then compared with paper-based risk assessment for early assessment of adverse birth risk among pregnant women as a means of improving the allocation of social services. ML algorithms outperformed the current paper-based risk assessment by up to 36%. It was estimated that improvements obtained as a result of ML algorithms will allow 100 to 170 additional high-risk pregnant women screened for program eligibility each year to receive services that would have otherwise been unobtainable which shows potential for machine learning to move government agencies toward a more data informed approach to evaluating risk and providing social services. Benites-Lazaro et al. [23] argue that ML algorithms can be a very powerful tool to provide a different approach of handling complex issues such as climate change and energy. Using a mixed method approach, an unsupervised probabilistic modelling was combined with discourse analysis to examine the changes in debates related to ethanol production in Brazil and its relationship with climate change and food security. The approach was useful in explaining; the discourse of the various actors on climate change, ethanol, and food security issues in Brazil and the narrative of various actors over a period of ten years. Hino et al. [24] argue that public agencies aiming to enforce environmental regulation with limited resources can use ML algorithms to achieve their objectives such as predicting the likelihood of a facility failing a water-pollution inspection and proposing inspection for high-risk facilities. Despite all the advantages that ML provides for informed policy decisions, Athey [8] argue that ML driven policies may deprive stakeholders of the knowledge about how and why policies are made, raising issues like transparency, interpretability, fairness, or discrimination, therefore, public should be informed of the processes that are undertaken by public agencies.

3 Finding and gaps in literature

As discussed in the previous section, various studies indicate the use of ML based approach in different areas of public policy such as health, education, tax and climate change policy, for making and improving policy decisions. In addition, the ML approach has been combined with other interpretative research techniques such as discourse analysis for a more in-depth examination of a policy problem. Ballestar at al. [7] argue that for big data to achieve its full potential in policy studies, multi-disciplinary approaches are needed that build on new computational algorithms from the ML literature, but also that bring in the methods and practical learning from decades of multi- disciplinary research using empirical evidence to inform policy decisions. Despite the fact that ML techniques have been applied in various areas of public policy, however, its application in the area of social protection - a major field of social policy in developing countries, for the better identification and targeting of populations within a country who require immediate social protection interventions has been largely overlooked, which presents a major gap in the literature. The next section of the paper highlights that how this paper will fill in this gap, by first explaining the concept of social protection and then proposing a methodology by using UML K-means clustering technique, to accurately identify populations present in various regions of a country, who are in urgent need of social protection interventions.

4 Social protection and data constraints in developing countries

Poverty is a social problem and in the absence of active redistributive governmental policies coupled with widely shared economic growth, it can continue to span over generations giving rise to serious health, education and other societal problems [25]. Social policies are a subset of public policies that includes state actions to protect weakest members of a community in particular, as well as responding to the social needs of all the members of a society in general [9, 26]. Social protection is an important social policy tool that has been adopted by several developing countries and international donor agencies to combat poverty and increasing the welfare of the poor [27, 28]. However, developing countries are faced with financial constraints, which limits not only their capacity to fund large-scale social protection programmes but also reduces their coverage [12]. In addition, factors such as poor targeting and lack of data availability remains a major impediment towards the successful implementation of social protection programmes [29–31]. By applying the latest data analysis techniques some data constraints can be overcome, which can lead to the improvements in the well-being of individuals in developing countries [32].

5 Methodology for data collection and reasons for using K-means clustering technique

This article uses a cross sectional survey dataset collected as part of the first author's PhD research.[3] This survey was conducted in 14 different cities in Pakistan including 570 households that were receiving informal assistance from religious institutions. The cities were randomly selected based on the multi-dimensional poverty index (MPI).[4] From every decile of MPI at least one city was randomly selected. Three to eight religious institutions from each said city's rural and urban areas were randomly selected and from the record of every religious institution, at least four to eight households were randomly selected for the survey. The questions in the survey were based on household characteristics, income and jobs/activities of households members, their assets, risks and shocks faced by the households, different kinds and duration of formal social protection received by the households and kinds and duration of informal support received by households through means such as family, friends, landlord, non-governmental organisations (NGOs), religious institutions and employer etc. There were 88 attributes (variables) against which the responses of each household was recorded. Based on the research objectives, which are to identify the dimensions of need in a developing country context and use this for determining better targeting of social protection interventions, this study chose a UML K-means clustering technique. In addition, at his stage, the purpose of the study is not to make predictions, therefore a UML clustering technique best suits the desired outcome of this study. An advantage of using UML clustering is that it requires no parameters (explicit labels), to be provided to the UML algorithms that targets to optimally minimize the human bias while forming clusters. Whereas, other statistical software such as SPSS or STATA require input parameters (explicit labels), in order to form clusters. It is because of this very reason that this study is using UML algorithms to explore a large survey data set. As far as we are aware, this is the first study where the UML K-means clustering technique has been used to explore a survey data in order to identify population and regions within a country for social protection interventions.

5.1 K-means clustering

Clustering [33], an UML approach, determines the way data is distributed in some space called "Density Estimation". In other words, clustering is the process of grouping together the similar instances based on similarity of their features or attributes without using training-base and assigning labels to instances.[5] There are various ways [34] for measuring the feature similarity based on attribute types such as: cosine similarity for vector-based data, jaccard similarity for set based data or euclidean distance for point data. This article employs euclidean distance-based similarity measures since the available data can only be interpreted as independent points. Clustering algorithms have different variations [35] that can be chosen based on desired output, nature of data and experimental parameters. The types of clustering algorithms that were evaluated for this study are: K-means clustering, DBSCAN (Density-based spatial clustering of applications with noise) clustering, hierarchical clustering and gaussian mixture clustering. After comprehensive analysis, K-means clustering was selected for clustering and DBSCAN for comparison. K-means clustering is also called exclusive clustering that necessarily assigns each instance to a cluster value (leaving no outlier). An instance in the dataset that has been made part of one cluster can never be part of another cluster (non-overlapping). In K-means clustering, it is a crucial aspect to decide upon how many clusters can be made over the existing distribution of dataset instances. So, the Elbow method [36] was exploited to have "inertia value" i.e. optimal number of clusters (here we got 4 clusters to model the instances optimally). Moreover, the metric of silhouette distance was used that is the measure of inter-cluster distances. K-mean clustering model which has the maximum silhouette distance is regarded as best [37–39], which in this case was 0.44063. A detailed view of evaluating different parameters of K-means clustering is provided in Table 1.

DBSCAN [4] is another technique that is used in this paper for comparing with K means clustering.[6] In DBSCAN approach, a cluster is initiated, if reasonable number of points is placed in a region else the points are regarded as noise (esp.

[3] Australian National University ethics protocol: 2019/377

[4] In Pakistan, the Multidimensional poverty Index (MPI) is a way of measuring poverty. MPI combines various deprivations that affect a household across three dimensions: education, health, and living standards and 11 indicators spread across these 3 dimensions. A household is considered multidimensionally poor if it is deprived in at least 33% of the weighted indicators [40]. Details of the cities is provided in Table A of the appendix

[5] In various ML techniques labels are assigned to instances and trainig data is used for constructing models. However, in k means clustering no training data is used and no labels are assigned to to instances for forming clusters.

[6] In DBSCAN dense region is a proximity, where the minimum number of instances are accumulated to establish a new cluster.

Table 1 Evaluation of Parameters for K-Means Clustering

Sr No	No of Clusters	Iterations	Average Silhouette Distance
1	2	100	0.31167404372524425
	2	200	0.42167404372524425
	2	300	0.37167404372524425
2	3	100	0.19571252544888507
	3	200	0.19571252544888507
	3	300	0.19571252544888507
3	**4**	**100**	0.4406300642801672
	4	200	0.29063007427901271
	4	300	0.3022564279111273
4	5	100	0.1696563875761801
	5	200	0.1696563875761801
	5	300	0.1696563875761801
5	6	100	0.19121033960102746
	6	200	0.12121033960102746
	6	300	0.12121033960102746
6	7	100	0.12121033960102746
	7	200	0.12122664652729666
	7	300	0.12122664652729666
7	8	100	0.2625846808150078
	8	200	0.11625846808150078
	8	300	0.11625846808150078
8	9	100	0.11584201611858196
	9	200	0.21584201611858196
	9	300	0.11584201611858196
4	10	100	0.22116262184732944
	10	200	0.12116262184732944
	10	300	0.12116262184732944
Best results	4	100	0.4406300642801672

the ones in low-density regions). Density of cluster is determined by the points through neighbourhood called Epsilon (also called disk size). The parameters used for DBSCAN clustering including optimal ones, are given in Table 2.

From Tables 1 and 2, it is clear that the maximum silhouette distance (i.e. 0.44063) is obtained by using K-means clustering technique. In addition, the number of clusters in K-means is four with no outliers as compared to DBSCAN that has formed 3 clusters with 87 outliers that are treated as 87 independent clusters, therefore, K-means clustering technique stands a better choice for the analysis of this data.

Table 2 Evaluation of Parameters for DBSCAN Clustering

Sr No	Epsilon	Min Points	Clusters	Outliers	Average Silhouette Distance
1	5	3	7	288	−0.21423170029391006
2	5	4	4	319	0.09579946839607748
3	5	5	3	335	−0.0863223531547006
4	5	10	3	407	−0.10238636017417664
5	7	3	5	72	−0.013542326770429502
6	7	4	3	87	0.20213496492473393
Best results					
	7	4	3	87	0.20213496492473393

Table 3 Number of households in each cluster

Cluster	Number of households
0	390
1	47
2	95
3	38

5.2 Description of clusters

Four clusters; cluster0, cluster1, cluster2 and cluster3 were formed by using a UML K-means clustering algorithm. Tables 3, 4 and 5 shows the number of households in each cluster, the percentage of total rural urban[7] areas in each cluster and the number of households of every city in each cluster. In addition, the percentage of the number of households that every city has within a particular cluster and the percentage share of rural and urban households of each city in each cluster is given in Tables B and C of the appendix. Tables 3, 4 and 5 indicate that cluster0 and cluster3 have the majority of the households belonging to rural areas with 67.40% and 84.20% respectively. Whereas, cluster1 and cluster3 have majority of the households belonging to urban areas with 91.40% and 69.50% respectively. Cluster0 and Cluster3 has the majority of the representation of households from cities that have high MPI (Table A of the appendix). This is more prominent in case of cluster3, which has no households in cities such as Lahore, Chakwal, Gujranwala and Faisalabad: the cities have less then 20% of MPI (Table A of the appendix). Cluster1 and cluster2 has majority of the households that belong to cities with low MPI. The case is more prominent in cluster2 which has almost no to negligible presence of households from cities such as Barkhan, Bajor, UpperDir and Bhakkar; these cities have high MPI (Table A of the appendix). The brief analysis shows that cluster0 and 3 are similar in terms rural-urban based households and MPI, however, cluster3 is a more extreme case with majority of its households falling in high MPI rural areas. Similarly, cluster1 and 2 have similar number rural- urban households and MPI based presence (Tables B

[7] In Pakistan, rural and urban areas are present within the geographical limits of a city. Rural areas are generally referred to villages where the process the urbanization is limited or has not taken place and people rely on informal employment mechanisms such as agriculture etc. Whereas, urban areas are generally referred to cities where process of urbanization has taken place and there are opportunities of formal employment. For administering rural areas government has formed union councils and for urban areas municipal corporations are present.

Table 4 Percentage of rural and urban areas of all cities in each cluster

Cluster0			Cluster1			Cluster2			Cluster3		
Rural	Urban	Total	Rural	Urban	Total	Rural	Urban	Total	Rural	Urban	Total
67.40%	32.50%	100%	8.50%	91.40%	100%	30.50%	69.50%	100%	84.20%	15.70%	100%

and C of the appendix). However, cluster2 have majority of households belonging to low MPI urban areas (Tables B and C of the appendix).

6 Results

As mentioned earlier, during the survey the data was collected against 88 attributes for each household and a total of 570 households were surveyed from 14 different cities. These attributes have been grouped into household characteristics, assets held by the households, risks and shocks faced by the households, formal social protection received by the households through various sources, informal social protection received by the households through various sources and benefits received from religious institutions such as madrassas. In the following paragraphs, we will explain the results of four clusters in terms of said grouped attributes.

6.1 Household characteristics of clusters

The average household size of cluster0 and cluster3 is 8.7 and 16 respectively, whereas the cluster1 and cluster2 has an average household size of 9.1 and 7 respectively (Table 6). The details of the average number of adult male and children and the adult female and children (in terms of percentage) are given in Tables D and E of the appendix. Cluster0 and cluster3 has the lowest household income of 100,000 to 15,0000 PKR and 150,000 to 200,000 PKR per year respectively (Table 7). Whereas, cluster1 and 2 has the household income of 1.3 million PKR and 500,000 to 550,000 PKR per year respectively (Table 7). Table 7 further indicates that each household member in cluster0, cluster1, cluster2 and cluster3 is living on approximately 18 to 28 cents, 2.4 US dollar, 1.2 to 1.3 US dollar and 14 to 19 cents per day respectively. The details of the percentage number of households that fall in different income brackets is given in the Table E1 of the appendix. In terms of employment, the majority of the heads of the

Table 5 Number of households in each cluster – City wise

	Cluster0			Cluster1			Cluster2			Cluster3			Total number of households surveyed
	Rural	Urban	Total	Rural	Urban	Total	Rural	Urban	Total	Rural	Urban	Total	
Toba tek singh	9	5	14	1	5	6	2	4	6	1	0	1	27
Barkhan	54	0	54	0	0	0	0	0	0	2	0	2	56
Bajor	36	0	36	0	0	0	0	0	0	12	0	12	48
Upper Dir	26	0	26	0	0	0	1	0	1	5	0	5	32
Vehari	12	14	26	0	2	2	4	3	7	3	0	3	38
Faisalabad	16	12	28	1	3	4	2	5	7	0	0	0	39
Multan	9	6	15	0	3	3	0	1	1	1	0	1	20
Gujrawala	6	1	7	0	3	3	4	6	10	0	0	0	20
Chakwal	3	1	4	0	1	1	10	11	21	0	0	0	26
Lodhran	15	15	30	1	0	1	1	2	3	1	0	1	35
Okara	18	16	34	0	0	0	1	0	1	1	2	3	38
Kasur	21	19	40	0	0	0	0	1	1	0	1	1	42
Bakkhar	23	26	49	1	1	2	3	0	3	3	1	4	58
DG khan	15	10	25	0	0	0	1	4	5	3	2	5	35
Lahore	0	2	2	0	25	25	0	29	29	0	0	0	56
Total	263	127	390	4	43	47	29	66	95	32	6	38	570

Table 6 Average Household Size

	Average household size Cluster0	Average household size Cluster1	Average household size Cluster2	Average household size Cluster3
Adult Male	2.6	3.7	2.52	4.57
Adult female	2.1	2.8	1.95	3.84
Children male	1.9	1.2	1.38	4.26
Children female	2	1.4	1.11	3.71
Household size – Total	8.7	9.1	7	16.44

households are employed, with an 85.13% employment rate in cluster0, 100% employment rate in cluster1, 98% employment rate in cluster2 and 84.21% employment rate in cluster3 (Table 8). Further analysis about the nature of employment of the head of the households indicate that almost 74% of the heads of the households of cluster0 are working in the informal sector such as daily wages labour, road side vendors, tenants, small scale farming on own land for subsistence, or child labour, while almost 15% are not employed because of some disability or disease. The case is even more prominent in cluster3 where almost 90% of heads of the household are either unemployed or working in low paid informal sector jobs (Table 9). Cluster1 has almost 80% of the households working in formal sector which includes government, formal private sector jobs and working in the defence with no case of child labour. Similarly, cluster2 has almost 55% of formal sector employment (Table 9). The other members of the households of cluster0 and cluster3 are involved in informal sector activities such as working for others on their farms, daily wages, looking after livestock etc. (Tables F and I of the appendix). Informal employment amongst other members of the households in cluster1 and cluster2 is relatively low as compared to cluster0 and cluster1 (Tables F and I of the appendix). Almost 42% of households in cluster0 and 56% of the households in cluster3 faced the problem of disability or disease. However, this situation is relatively better in cluster1 and cluster2, where only 6% and 10% of the households face such a problem (Tables G and H of the appendix).

6.2 Assets

The majority the households in cluster0, cluster1, cluster2 and cluster3, own a house at 77.69%, 91.49%, 80% and 81.58% respectively (Table 10). However, Table 11 shows that the conditions of the dwelling are substantially different in different clusters. The majority of the houses in cluster0 and cluster3 are partially cemented or brick or a mud house. Approximately 5% of the households in cluster0 are living in a temporary shelter, 25% in partially cemented or bricked and 42% in mud house (Table 11). The situation is even more striking in cluster3, where 21% households are residing in partially cemented house, 50% in mud houses and approximately 3% in a temporary shelter (Table 11). Cluster1 and cluster2 have relatively better positions with 96% and 80% of households living in completely cemented houses (Table 11). The number of rooms of each house in each cluster also vary as cluster1 and cluster3 have 80% and 50% respectively of the houses that have either one or two rooms (Table 12). However, approximately 81% of houses in cluster1 have either 3,4 5 or 6 rooms (Table 12). Similarly, approximately 87% of households in cluster2 have either 2, 3 or 4 rooms (Table 12). From the details mentioned above, it is

Table 7 Annual Household Income

Average of Annual income cluster0	100,000–150,000 PKR	600–900 US dollar per year approximately	68–103 US dollar per person per year approximately	18–28 cents per person per day approximately
Average of Annual income cluster1	1.3 million PKR	8000 US dollar per year approximately	879 US dollar person per year approximately	2.4 US dollar per person per day approximately
Average of Annual income cluster2	500,000–550,000 PKR	3050–3350 US dollar per year approximately	440 – 480US dollar person per year approximately	1.2–1.3 US dollar per person per day approximately
Average of Annual income cluster3	150,000–200,000 PKR	900–1200 US dollar per year approximately	54–72 US dollar person per year approximately	14–19 cents per person per day approximately

Table 8 Occupation and Employment - (%age)

	Employed	Unemployed
Employment status of the head of household cluster0	85.13%	14.87%
Employment status of the head of household cluster1	100%	0
Employment status of the head of household cluster2	97.89%	2.11%
Employment status of the head of household cluster3	84.21%	15.79%

clear that the density of people in a house is highest in cluster3 with 16 people living in 2 rooms approximately, followed by cluster1 where 9 people are living in 2 rooms approximately, followed by cluster2 where 7 people are living in 3 rooms, and least in cluster2 where 9 people are living in 5 rooms (Table 13). The details of the agricultural land, livestock held and purpose of keeping livestock is given in Tables J, K and L of the appendix. In terms of holding bank accounts cluster0 and cluster3 have approximately 90% of the households that never had a bank account. Cluster1 and cluster2 have 85% and 65% of the households respectively that have access or are holding a bank account (Table 14). This shows that the majority of the households in cluster zero and three do not have access to formal credit, whereas, cluster1 and cluster2 have the majority of the households that have access to the formal credit.

6.3 Risks and shocks

Table 15 shows the percentage of households that have ever faced any risk or shock in their life course. It can be observed that almost 97% of the household members in cluster0 and 100% of the households in cluster3 have experienced a risk or shock in their life course. However, the case is less prominent in cluster1 and cluster2 with 68% and 75% of the households having suffered any risk or shock in their life course (Table 15). The most common shocks that are faced by the households in cluster0 is illness, disability, death of an adult family member resulting in loss of income, infant mortality and unemployment. Some households did face natural disasters, loss of job/business and migration because of conflict (Tables M and M1 of the appendix). The incidence of unemployment amid these shocks is highest with over 70% of the households have either 1, 2 or 3 members faced unemployment (Tables M and M1 of the appendix). There is also the prevalence of disease such as polio, hepatitis and TB in cluster0 (Tables M and M1 of the appendix). The households in cluster3 also faced shocks such as illness, disability, death of a family member resulting in loss of income, infant mortality, unemployment and disease such as Hepatitis, Polio and TB (Tables P and P1 of the appendix). In addition, almost 42% of the households in cluster3 lost their job or business because of conflict in their area, 48% of the households migrated from their houses because of conflict and almost 48% faced natural disasters such as floods and earthquakes, which has resulted in the loss of property (Tables P and P1 of the appendix). The incidence of these shocks is relatively less in cluster1 and 2 (Tables N, N1, O and O1 of the appendix). The most common shock faced by households in cluster1 is marriage[8] with 42% of households needed assistance to face this shock. In cluster2 almost 30% of the households faced unemployment and marriage (Tables N, N1, O and O1 of the appendix).

In terms of the means available for the households to bear the shocks, the households in cluster1 and cluster3 mainly rely on help from the community, informal credit loan taken from friends or the community, and in some cases, help was given from the extended family (Table 16). Only 3% of the households in cluster0 and 5% of the households in cluster3 received some sort of help from the government to bear the shocks (Table 16). Only 3% households in cluster0 and 5% in cluster1 did not require any assistance to bear the shock (Table 16). 1.5% of the households in cluster1 received some sort of assistance from local and international NGOs, whereas, 5% of the households in cluster three received support from international NGOs only (Table 16). 80% and 60% of the households of cluster1 and cluster2 respectively were in no need of any assistance to bear the shock faced by them. However, reliance on the informal support mechanisms is also present in cluster1 and cluster2. 12% of the households in cluster1 and 6% in cluster2 relied on informal mechanism called "committee"[9] to bear the shocks (Table 16).

6.4 Formal social protection received by the households through various sources

Table 17 shows the various kinds of formal social protection mechanisms that are currently operating in Pakistan along with the percentage of surveyed households that are receiving benefits from them. Only 28% of the households in cluster0 and 31% of the household in cluster3 are receiving the

[8] During the survey, it revealed that in order to fulfill the expenses of marriage some households took loans and some had to use their savings, therefore marriage is considered as a shock.

[9] Insurance provided in shape of rotating savings and credit associations where every member contributes towards this fund and get the cash in time of need.

Table 9 Occupation of the head of the household – (%age)

Occupation - head of household (Number of households	Formal sector government job	Armed Forces	Formal sector private job	Informal sector own small business	Informal sector work for other (includes daily wages labour)	Informal sector own land agriculture	Informal sector work on other land (tenant)	Unemployed	Child labour	Retired government servant	Formal private sector business	Working overseas (labour)	Formal business	Grand total
cluster0	1.79%	1.28%	5.13%	21.28%	34.62%	12.05%	4.87%	15.64%	1.79%	1.03%	0.26%	0.26%	0	100.00%
cluster1	42.55%	2.13%	12.77%	14.89%	2.13%	2.13%	0	0	2.13%	0	0	4.26%	17.02%	100%
cluster2	29.47%	4.21%	10.53%	18.95%	14.74%	9.47%	2.11%	2.11%	–	5.26%	–	2.11%	1.05%	100%
cluster3	7.89%	–	7.89%	15.79%	18.42%	26.32%	5.26%	10.53%	5.26%	–	–	2.63%	–	–

Table 10 Dwelling

	Live in a Rented accommodation	Own a house	Living on someone else land in a temporary shelter	Living on someone's property without rent	Government accommodation
Number of households Cluster0	8.21%	77.69%	4.87%	8.97%	0.26%
Number of households Cluster1	4.26%	91.49%	0	2.13%	2.13%
Number of households Cluster2	11.58%	80.00%	0	4.21%	4.21%
Number of households Cluster3	2.63%	81.58%	7.89%	7.89%	

benefits of BISP which is the largest social protection programme in Pakistan. However, coverage of BISP in cluster1 is 0% and in cluster2 is only 4.21% (Table 17). The coverage of another social assistance programme: zakat and bait-ul-mal scheme, is less than 3% in every cluster (Table 17). The coverage of other social protection programmes is also relatively low in all other clusters. For example, coverage of private sector pension schemes, youth loan, public lunger and panagah (free food and shelter programme), in all the four clusters is almost 0% (Table 17). Only 3% of the households in cluster0, 5% in cluster3, 15% in cluster1 and 10% in cluster2 are receiving the benefits of public pension scheme, which indicates the presence of more formal sector employment in cluster1 and cluster2 (Table 17). The coverage of labour markets measures: rural and national support programme, is almost non-existent in all the four clusters (Table 17). Coverage of sehat card; a health insurance programme is 15% in cluster0, 23% in cluster3, 2% in cluster1 and 8% in cluster2 (Table 17). The coverage of free universal public education provided through public schools is 85% and 92% in cluster0 and cluster3 respectively. In cluster3, only 28% percent of the households received public education for 10 years and above, 39% of the households received public education only for five to 10 years and the remaining households receive less than five years or no education (Table Q of the appendix). Similarly, in cluster0 only 25% of the households received public education for 10 years and above, while the remaining households receive for 10 years and below (Table Q2 of the appendix). In cluster2, 77% of the population received public education for over 10 years and above and a number of households are sending their children to the private schools as well (Table Q1 of the appendix). In cluster1, 49% of the households have received public education for 10 years and above, however, a substantial number of households do send their children to private schools as well (Table Q3 of the appendix). The coverage of universal health provided through government hospitals and dispensaries is relatively better in all the four clusters with 94% in cluster0, 89% in cluster3, 55% in cluster1 and 88% in cluster2 (Table 17). The relatively low utilization of government health facilities in cluster1 indicates that households opt for private hospitals when they have the resources. Further details about the duration for which households are receiving the benefits of social protection programmes are provided in Tables Q, Q1, Q2 and Q3 of the appendix.

Table 11 Condition of house

	Completely cemented and bricked	Partially cemented or bricked	Mud house	Temporary shelter
Cluster0	28.21%	24.87%	42.31%	4.62%
Cluster1	95.74%	2.13%	2.13%	0
Cluster2	80.00%	15.79%	4.21%	0
Cluster3	26.32%	21.05%	50.00%	2.63%

6.5 Informal social protection received by the households through various sources

A majority of the surveyed households are receiving informal assistance through sources such as charity and zakat,[10] extended family and friends, help from employer and landlord, help through remittances and assistance from national and international NGOs. Table 18 shows that 72% of the households in cluster0 and 55% in cluster3 are receiving help from their family friends and community in shape of charity and zakat. Support from landlord in all the clusters is almost non-existent and a very little support is received from local and international NGOs (Table 18). Assistance received in shape of remittances is more common in cluster1 and cluster2 with almost 12% of households receive such assistance. Approximately, 3% households in cluster0 and cluster3 receive assistance

[10] Zakat is one of the five pillars of Islam and is mandatory on every Muslim who is financially stable. According to Islamic teachings, zakat is paid @2.5% of the wealth to the poor and needy Muslims as an obligation. It is applicable on every Muslim who owns 613.35 g of silver, or 87.49 g of gold or who owns one or more assets liable, equal in value to 613.35 g of silver or 87.49 g of gold. Zakat is given to Muslims: who are poor and not have any income source etc.

Table 12 Number of rooms

	No room	1 Room	2 Rooms	3 Rooms	4 Rooms	5 Rooms	6 Rooms	7 Rooms	8 Rooms	9 rooms	10 Rooms	12 Rooms
Cluster0	5.13%	20.26%	58.46%	8.21%	5.13%	2.31%	0.51%	–	–	–	–	–
Cluster1	–	–	–	21.28%	25.53%	19.15%	17.02%	6.38%	6.38%	–	2.13%	2.13%
Cluster2	–	5.26%	40.00%	33.68%	14.74%	4.21%	2.11%	–	–	–	–	–
Cluster3	2.63%	5.26%	44.74%	36.84%	5.26%	–	2.63%	2.63%	–	–	–	–

from remittances (Table 18). Assistance received from local and international NGOs is non-existent in cluster1and cluster2, however, their marginal coverage can be seen in cluster0 and cluster3 (Table 18). Support from employer is present in cluster0 and cluster3 with 7% and 16% respectively and almost non-existent in cluster1 and cluster2 with 0 and 2% respectively. The duration of the informal social protection received by these households through these informal sources is given in Tables R, R1, R2 and R3 of the appendix.

6.6 Madrassa benefits

It is important to see the kind of benefits which the households are receiving from the religious institutions (madrassas) because the surveyed households were randomly selected from madrassa records. Table 19 shows that majority of the households do send their children to madrassas. 48% household's children in cluster0, 55% in cluster3, 47% in cluster2 and only 27% in cluster1 are receiving boarding facilities from the madrassas (Table 19). Tables S, S1, S2 and S3 of the appendix shows the details of the kinds of benefits received by households from madrassas. A range of benefits such as free food, clothing, accommodation, stipend, free health treatment, religious and public-school education, assistance during festive seasons and technical training is given to the children of households that are going to madrassas. As many as eight members of a household are receiving these benefits, for instance 0.5% of the households in cluster0 are receiving free food from madrassas (Table S of the appendix). It is also evident that households in cluster0 are receiving highest percentage of madrassa benefits, followed by cluster3, cluster2 and cluster1(Tables S, S1, S2 and S3).

Table 13 Density

Cluster0	Cluster1	Cluster2	Cluster3
1.9 rooms/8.7 persons	5 rooms /9.1 persons	2.7 rooms /7 persons	2.60 rooms /16.44 persons

7 Conclusion and implications for future research

We have presented a novel methodology of exploring a data set by combining UML K-means clustering approach with descriptive statistics for the purpose of identifying the differences between areas of Pakistan and better target assistance to the population within a country that are in urgent need of government support through various forms of social protection programmes in the presence of data constraints. We have utilised UML K-means clustering technique for exploring a survey data of 570 households recorded against 88 attributes (variables) and compared it with UML DBSCAN for obtaining the optimal results. We used metric of silhouette distance to measure inter-cluster distances and obtained best results by using K means clustering: four clusters were formed leaving no outliers with the maximum silhouette distance of 0.4406300642801672 [37–39] (Tables 1 and 2). The advantage of using UML K-means clustering technique is that no labels are assigned to the instances for forming the clusters which reduces human bias in the formation of clusters. The results of four clusters formed: cluster0, cluster1, cluster2 and cluster3, were then explored further by using descriptive statistics to identify the common patterns and the insecurities faced by households in each cluster.

We formed summary tables of each cluster for almost every attribute against a given instance. These attributes are grouped into household characteristics, assets held by households in each cluster, risks and shocks faced by households in each cluster, formal social protection received by households in each cluster, informal assistance received by households in each cluster and benefits received by households through

Table 14 Bank Account

	No response	Yes (Any member of household have bank account)	No member of household has bank account
Cluster0	1.03%	9.49%	89.49%
Cluster1	–	85.11%	14.89%
Cluster2	–	65.26%	34.74%
Cluster3	–	10.53%	89.47%

Table 15 Risks and Shocks

	Households that have ever faced any shock	Households that never faced any shock
Cluster0	97.69%	2.31%
Cluster1	68.09%	31.91%
Cluster2	75.79%	24.21%
Cluster3	100.00%	0%

religious institutions. We found that in cluster0 and cluster3, majority of the households are residing in rural areas, whereas in cluster1 and cluster3 majority of the households belong to urban areas (Tables 3, 4 and 5 and Tables B and C of the appendix). The average income per household member per day is least in cluster3 followed by cluster0, cluster1 and is highest in cluster2. This indicates that households in cluster0 and cluster1 are living in extreme poverty and are in need of more support than households in cluster1 and cluster2 (Table 7). Unemployment, informal employment, disability and disease is more prominent in cluster3 and cluster0. However, the situation is relatively better in cluster1 and cluster2, which have less unemployment, more formal sector employment and less prevalence of disease and disability (least in cluster2) (Table 9).

A majority of the households in cluster0 and cluster3 are living in abysmal conditions which includes poor condition of houses and high living density within a house; worst in cluster3 (Table 11). The living conditions of the households of cluster1 and cluster2 are relatively better (Table 13). Majority of the households in cluster0 and cluster3 do not have access to formal credit, whereas this condition is relatively better in cluster1 and cluster2 (Table 14). Households of cluster3 faced maximum number shocks followed by cluster0, cluster1 and cluster2 (Tables P and P1, N, N1, O and O1 of the appendix). In terms of means available to face the shocks, households of cluster3 and cluster0 mainly rely on informal mechanisms and in some cases did not receive any support from any source to bear the shock (Table 16). Whereas, majority of the households of cluster1 and cluster2 had either savings, receiving remittances or have devised informal mechanisms to counter the effects of the shocks faced by them (Tables 17, 18 and 19).

The analysis stipulates that households in cluster3 face severe insecurities followed by cluster0. Both the clusters have households belonging to rural areas and common districts are Bajor, Upper Dir, Bharkhan, DG Khan. It can therefore be concluded that the out of the 14 cities surveyed, the rural areas of these four districts require urgent social protection interventions in shape of cash transfers to supplement their income, skill training programmes, which can improve

Table 16 Means to face the shock used by the households

	No need of assistance	Government help	Selling of assets	Help from Community	Informal credit -(Loan from community, friends etc.)	Committee (Informal Insurance – people in a community join together to credit fund)	Did not receive any help from any source to bear the shock	Formal Loan from Bank	Help form family (extended family)	Savings - own	Help from NGOs local	Help form NGOs international
Cluster0	3.85%	2.31%	3.33%	22.31%	23.59%	8.46%	27.44%	0.26%	5.38%	1.54%	0.77%	0.77%
Cluster1	80.85%	–	–	–	4.26%	12.77%	–	–	–	2.13%	–	–
Cluster2	60.00%	–	5.26%	3.16%	14.74%	5.26%	4.21%	–	7.37%	–	–	–
Cluster3	5.26%	5.26%		7.89%	26.32%	5.26%	42.11%		2.63%			5.26%

Table 17 Formal social protection received by the households through various sources

	Public programmes of Zakat/Bait-ul-mal (Social Assistance)		BISP		Public lunger and panagah (free food and shelter programme)		Sehat Card (Social Insurance)		Retirement pension – public sector		Retirement pension – Private sector (Social insurance, Labour market measure)		Rural and national support programmes		Free technical education from government		Youth loan		Free Public education		Free health treatment from government hospital or dispensary	
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No
Cluster0	2.05%	97.95%	28.21%	71.79%	–	100%	15.13%	84.62%	3.33%	96.67%	–	100%	0.26%	99.74%	1.28%	98.72%	0	100%	85.90%	14.10%	94.62%	5.13%
Cluster1	2.13%	97.87%	0	100%	–	100%	2.13%	97.87%	14.89%	85.11%	–	100%	–	100%	–	100%	–	100%	63.83%	36.17%	55.32%	44.68%
Cluster2	1.05%	98.95%	4.21%	95.79%	–	100%	8.42%	91.58%	10.53%	89.47%	–	100%	1.05%	98.95%	4.21%	95.79%	–	100%	89.47%	10.53%	88.42%	11.58%
Cluster3	2.63%	97.37%	31.58%	68.42%	–	100%	23.68%	76.32%	5.26%	94.74%	–	100%	–	100%	–	100%	–	100%	92.11%	7.89%	89.47%	10.53%

Table 18 Informal social protection received by the households through various sources

	Informal Social assistance		Informal assistance, Informal insurance		Informal assistance, Informal insurance, Labour market measure		Informal assistance, Informal insurance, Labour market measure		Informal assistance, Informal insurance, Labour market measure			(Informal assistance, Informal insurance, Labour market measure
	Assistance - family, friends or community, ZAKAT and charity		Landlord		Remittances		Local NGOs /organisations such as Edhi, Akhuwat etc.		International NGOs such as USAID			Employer
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No
Cluster0	72.05%	27.95%	0.51%	99.49%	3.59%	96.41%	2.56%	97.44%	2.31%	97.69%	6.67%	93.33%
Cluster1	4.26%	95.74%	0	100%	12.77%	87.23%	0	100%	0	100%	0	100%
Cluster2	29.47%	70.53%	0	100%	12.63%	87.37%	3.16%	96.84%	0	100%	3.16%	96.84%
Cluster3	55.26%	44.74%	0	100.00%	2.63%	97.37%	5.26%	94.74%	2.63%	97.37%	15.79%	84.21%

Table 19 Day Scholar or Boarder

	Day Scholar	Boarder
Day Scholar or boarder – Households Cluster0	51.28%	48.72%
Day Scholar or boarder – Households Cluster1	72.34%	27.66%
Day Scholar or boarder – Households Cluster2	52.63%	47.37%

their technical skills that would be helpful for employment generation and micro credit schemes for increasing access to formal credit etc. In addition, contributory insurance schemes can be introduced in urban areas of cluster1 and cluster2 as the households in these clusters are already using such mechanisms informally to support themselves in times of need. Since, the purpose of this paper is not to provide any concrete recommendations about the nature of social protection interventions required for each cluster, therefore this aspect presents an avenue for future research as to what kind of social protection interventions best suit each cluster. Furthermore, the causes of these insecurities within each cluster can be explored in future research.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11036-020-01696-z.

Acknowledgments The authors would like to thank Mr. Sohail Sarwar for their valuable contribution in providing technical support for the successful completion of this paper.

References

1. Freeman C, Louçâ F (2001) As time goes by: from the industrial revolutions to the information revolution. Oxford University Press, New York
2. Buchel O (2015) Big data: a revolution that will transform how we live, work, and think. J Inf Ethics 24(1):132
3. Patty JW, Penn EM (2015) Analyzing big data: social choice and measurement. PS: Polit Sci Polit 48(1):95–101
4. Grimmer J (2015) We are all social scientists now: how big data, machine learning, and causal inference work together. PS: Polit Sci Polit 48(1):80–83
5. Thierer AD, Castillo O'Sullivan A, Russell R (2017) Artificial intelligence and public policy. George Mason University, Mercatus Center
6. Naudé W, Dimitri N (2019) The race for an artificial general intelligence: implications for public policy. AI & Soc:1–13
7. Ballestar MT, Doncel LM, Sainz J, Ortigosa-Blanch A (2019) A novel machine learning approach for evaluation of public policies: an application in relation to the performance of university researchers. Technol Forecast Soc Chang 149:119756
8. Athey S (2017) Beyond prediction: using big data for policy problems. Sci (New York, N.Y.) 355:483–485
9. Kemshall H (2002) Risk, social policy and welfare. Open University Press, Buckingham
10. Kangas O, Palme J (2009) Making social policy work for economic development: the Nordic experience: making social policy work. Int J Soc Welf 18:S62–S72
11. Gough I, Wood GD (2004) Insecurity and welfare regimes in Asia, Africa, and Latin America: social policy in development contexts. Cambridge University Press, Cambridge
12. Mumtaz Z, Whiteford P (2017) Social safety nets in the development of a welfare system in Pakistan: an analysis of the Benazir income support Programme. Asia Pac J Public Adm 39(1):16–38
13. Hindman H (2015) Building better models: prediction, replication, and machine learning in the social sciences. Ann Am Acad Pol Soc Sci 659(1):48–62
14. Andini M et al (2018) Targeting with machine learning: an application to a tax rebate program in Italy. J Econ Behav Organ 156:86–102
15. Kleinberg J, Ludwig J, Mullainathan S, Obermeyer Z (2015) Prediction policy problems. Am Econ Rev 105(5):491–495
16. Athey S, Imbens GW (2017) The state of applied econometrics: causality and policy evaluation. J Econ Perspect 31(2):3–32
17. Burscher B, Vliegenthart R, De Vreese CH (2015) Using Supervised Machine Learning to Code Policy Issues: Can Classifiers Generalize across Contexts? Ann Am Acad Pol Soc Sci 659(1):122–131
18. Kasy M (2018) Optimal taxation and insurance using machine learning — sufficient statistics and beyond. J Public Econ 167: 205–219
19. Chalfin A, Danieli O, Hillis A, Jelveh Z, Luca M, Ludwig J, Mullainathan S (2016) Productivity and selection of human capital with machine learning. Am Econ Rev 106(5):124–127
20. Ashrafian H, Darzi A (2018) Transforming health policy through machine learning. PLoS Med 15(11):1002692
21. Brady ES et al (2017) Machine-learning algorithms to code public health spending accounts. Public Health Rep (1974) 132(3):350–356
22. Pan I, Nolan LB, Brown RR, Khan R, van der Boor P, Harris DG, Ghani R (2017) Machine learning for social services: a study of prenatal case Management in Illinois. Am J Public Health 107(6): 938–944
23. Benites-Lazaro LL, Giatti L, Giarolla A (2018) Topic modelling method for analyzing social actor discourses on climate change, energy and food security. Energy Res Soc Sci 45:318–330
24. Hino M, Benami E, Brooks N (2018) Machine learning for environmental monitoring. Nat Sustain 1(10):583–588
25. Boran A (2012) Poverty: malaise of development. University of Chester Press, Chester
26. Aspalter C, Pribadi KT (2017) Development and social policy: the win-win strategies of developmental social policy. In: London;New York. Routledge, Taylor and Francis Group
27. Barrientos A (2013) Social assistance in developing countries. Cambridge University Press, Cambridge
28. World Bank (2018) The State of Social Safety Nets 2018. World Bank, Washington, DC
29. Abu Sharkh M, Gough I (2010) Global welfare regimes: a cluster analysis. Glob Soc Policy 10(1):27–58
30. Monchuk V (2013) Reducing poverty and investing in people: the new role of safety nets in Africa. The World Bank, Washington, DC
31. Pritchard B (2014) Feeding India: livelihoods, entitlements and capabilities. Routledge, New York and Oxfordshire
32. Hilbert M (2016) Big data for development: a review of promises and challenges. Dev Policy Rev 34(1):135–174
33. Rodriguez MZ et al (2019) Clustering algorithms: a comparative approach. J PLOS ONE 14(1). https://doi.org/10.1371/journal.pone.0210236
34. Sze-To A, Wong AK (2018) Discovering patterns from sequences using pattern-directed aligned pattern clustering. IEEE Trans Nanobioscience 17(3):209–218

35. Xu D, Tian Y (2015) A comprehensive survey of clustering algorithms. Ann Data Sci 2(2):165–193

36. Ester M, Kriegel HP, Sander J, Xu X (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. In: Kdd, vol 96, no 34. AAAI Press, Portland, pp 226–231

37. Rousseeuw PJ (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math 20:53–65

38. Lovmar L, Ahlford A, Jonsson M, Syvänen AC (2005) Silhouette scores for assessment of SNP genotype clusters. BMC Genomics 6(1):35

39. Kodinariya TM, Makwana PR (2013) Review on determining number of cluster in K-means clustering. Int J 1(6):90–95

40. UNDP (United Nations Development Programme) (2016) Multidimensional poverty in Pakistan, Islamabad: UNDP. Accessible at: http://www.pk.undp.org/content/pakistan/en/home/library/hiv_aids/Multidimensional-Poverty-in-Pakistan.html. Accessed 22 Apr 2020

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Zahid Mumtaz is a PhD candidate at the Crawford School of Public Policy, Australian National University (ANU). He holds a maters' degree in public policy and a master's degree in political science. He is the recipient of ANU university research scholarship and HDR fee merit scholarship for PhD. He has published in Asia pacific journal of public administration. He is a career civil servant and has held important portfolios in public sector.

Peter Whiteford is a Professor in the Crawford School of Public Policy at The Australian National University, Canberra. Between 2008 and 2012 he worked at the Social Policy Research Centre at the University of New South Wales (UNSW) in Sydney. He previously worked as a Principal Administrator in the Directorate of Employment, Labour and Social Affairs of the Organisation for Economic Co-operation and Development in Paris. He has published extensively on various aspects of the Australian and international systems of income support.