A Three Level Prediction of Multidimensional Poverty in Elderly

Fabio D'Adda[1], Marco Cremaschi[1], Enza Messina[1], Marco Terraneo[2], Stefania Bandini[1] and Francesca Gasparini[1]

[1]Department of Computer Science, Systems and Communications, University of Milano - Bicocca, Italy
[2]Department of Sociology and Social Research, University of Milano - Bicocca, Italy

Abstract

Poverty is a multidimensional concept that is not only related to economic aspects but also to health status, consumption, social and context deprivation. In particular, older adults are likely to require help with some or everyday activities and the total costs of this help can be very high especially when they are alone and not in good health. In this work a heterogeneous dataset acquired to consider various aspects of health, environment, social networks, and quality of life of older people is considered as source of knowledge. A procedure to label this data, that also relies on the domain expert intervention, is here presented to overcome the lack of groundtruth data. On this labelled data, a three class classifier is proposed to predict a three level risk of poverty.

Keywords

Multidimensional Poverty, Elderly, XGBoost, Sustainability

1. Introduction

Poverty is one of the most significant social problems in Organization for Economic Cooperation and Development (OECD). The 2030 Agenda for Sustainable Development, approved in September 2015 by the United Nations[1], presented *End poverty in all its forms everywhere* as the first of the 17 Sustainable Development Goals to promote prosperity while protecting the planet.

The AMPEL project (Artificial intelligence facing Multidimensional Poverty in ELderly) focuses on the use of cutting-edge technologies in Artificial intelligence (AI), Machine Learning (ML), data analysis and data visualization to identify the risk of poverty in elderly people, relying on multidimensional indicators, learned from heterogeneous sources of information.

The project aims to define a poverty risk indicator, an alert semaphore (AMPEL in German), able to classify three levels of susceptibility to poverty, useful to identify where a prompt reaction would be needed, especially in emergencies. Poverty is a multidimensional concept: focusing on financial resources alone does not capture people's needs and quality of life. Being poor means, in fact, also a lack of access to resources enabling a minimum standard of living and participation in society. Elderly people are likely to require help with some or all everyday activities, and the total costs of this help can be very high and absorb a significant amount of their income, especially when they are alone and not in good health. Incomes of the elderly are generally low: 23% of older people are likely to be at risk of relative income poverty, and this phenomenon interests 25 out of 35 OECD countries [1].

To correctly identify the risk of poverty, the presented project relies on data directly referred to income and wealth and on material and social deprivation that are rarely collected or known by public welfare institutions, making it difficult to intercept those who require more support. Material deprivation captures the ability of individuals and households to afford specific types of goods and services. In contrast, social deprivation refers to systematically excluding individuals, families and groups from participation in economic, political and social activities.

In this paper, a classification approach to identify a three-level risk of poverty is presented, which faces the following issues:

1. Difficulties in finding labelled data that includes all the multi-facet aspects of poverty, especially in the case of the elderly;
2. Lack of quality and noise in the available dataset that makes it crucial to select robust features to feed ML algorithms;
3. Difficulties in understanding how different features can contribute to identifying poverty clusters, which is not obvious when using complex heterogeneous data.

The paper is organised as follows. In Section 2, a brief

Ital-IA 2023: 3rd National Conference on Artificial Intelligence, organized by CINI, May 29–31, 2023, Pisa, Italy
f.dadda4@campus.unimib.it (F. D'Adda); marco.cremaschi@unimib.it (M. Cremaschi); enza.messina@unimib.it (E. Messina); marco.terraneo@unimib.it (M. Terraneo); stefania.bandini@unimib.it (S. Bandini); francesca.gasparini@unimib.it (F. Gasparini)

[1]sdgs.un.org/goals

state-of-the-art on multidimensional poverty and machine learning approaches to face this problem, especially in the case of the elderly, is reported. In Section 3 the data and source of data considered are presented. In Section 4, the proposed framework of analysis is described. In Section 5 a brief description of the technologies used for the implementation is provided. Finally, we conclude this paper and discuss the future direction in Section 6.

2. Background

Populations in OECD countries are ageing rapidly, their health worsens, and they may struggle with everyday activities. The financial challenges faced by older people with Long-Term Care (LTC) needs can be very high and absorb a significant amount of their income. Home care and small out-of-pocket payments may be unaffordable without adequate social protection.

Most studies emphasize the economic facet of poverty on the basis of monetary income [2, 3, 4]. However, income-based indicators are poor proxies of material conditions among the elderly [5, 6] whereas non-monetary ones improve our understanding of who is poor, with a shift from a unidimensional to a multidimensional approach [7],[8]. Multidimensional measures of deprivation are composed of different indicators fitting into a synthetic scale [9, 10], which is deemed to reflect basic living standards and the exclusion from the minimum acceptable way of life in one's own society.

In 2010 the Multidimensional Poverty Index (MPI), was officially published by Oxford Poverty and Human Development Initiative[2] in collaboration with Human Development Report Office of the United Nations Development Programme (UNDP) [11]. The MPI considers poverty through ten indicators divided into three dimensions: health, education and standard of living. The dimensions are equally weighted, and so are the specific indicators.

Several methodologies to assess poverty from a multidimensional perspective exist, including methods aiming to implement aggregate data from different sources, and statistical approaches – *i.e.*, principal component analysis, or cluster analysis – which reflect the joint distribution of single deprivation indicators and aim for a bottom-up definition of synthetic scales [12]. Such approaches are adequate if they capture the joint distribution of deprivations, identify the poor ones (*i.e.*, dichotomising the population into poor and non-poor), and provide a single cardinal figure to assess poverty.

Only recently, thanks to big data availability and the development of data science techniques, ML and AI have been increasingly adopted to poverty estimation. Moreover, the COVID-19 pandemic led to a significant increase in extreme global poverty with respect to the last 20 years, according to reports published by the World Bank [13], increasing the effort in defining predictive models.

Besides the choice of the best algorithms, other crucial aspects, such as data quality and the presence of bias due to subjective and indirectly related data, exist, pushing to choosing other data sources such as remote sensing datasets [14]. A second issue is related to the difficulties in finding labelled data. Ensemble models were employed, assuming as ground truth for ML training the Proxy Means Test (PMT) labels, without verifying the accuracy of the PMT labels [15]. Among different ML techniques for poverty classification, decision tree [16], random forest [15], and ensemble approaches [17] are the most used.

3. Dataset

The approach presented here relies on the TAPAS dataset (Time and Places and Space in Aging dataset)[18], previously collected by Fondazione IRCCS Istituto Neurologico Carlo Besta[3] and Auser Lombardy[4], both currently involved in the AMPEL project[5]. To collect the dataset, a set of validated tools previously developed and used in two projects [19, 20] was administered to older people by trained interviewers to investigate various aspects of health, environment, social networks, and quality of life, resulting in a total of 744 features (or variables) for each individual. The questionnaires were administered to 429 people aged 50 or over.

The features can be grouped into three categories:i) Categorical Features: variables which assume a fixed range of values (*e.g.*, private health insurance coverage: yes or no); ii) Numeric Features: variables which assume numeric values, both discrete and continue (*e.g.*, heart rate: 84); and iii) Range Features: variables which assume range values, (*e.g.*, ages: 60-75).

For more details about the dataset please refer to [18].

4. The AMPEL approach

TAPAS represents an excellent dataset to investigate poverty from a multidimensional perspective, thanks to the large number of features used to describe each individual. However, the dataset is characterised by several issues, including the high number of *NULL* values, the low number of individuals (429) compared to the number of features (744) and the difficulty in identifying the most significant ones. To face these issues, a pipeline that firstly cleans the data and then develops a classification system able to identify those living in different levels of poverty has been developed. To solve the problem of

[2] ophi.org.uk

[3] www.istituto-besta.it

[4] www.auser.lombardia.it

[5] ampel.unimib.it

large number of features and small number of individuals, some heuristics were applied. The most significant characteristics have been selected with the involvement of a domain expert to identify a subset of 99 features to be used in the next step of classification. Each of these features could be assigned to one of the five dimensions here considered:

- **Maintenance Capacity**: the financial situation of an individual;
- **Consumption Deprivation**: organises some information related to the affordance capacity of the individual;
- **Health Status**: collects health status features related to the individual;
- **Housing Facilities**: contains all the information that describes the conditions of the dwelling and the neighbourhood in which the individual lives;
- **Social and Context Deprivation**: reports information about social relations.

4.1. Class labelling

Since the TAPAS dataset does not explain how poverty is distributed across the population, it is necessary to find an approach that can identify which individuals are in poverty condition or at risk of poverty. This issue has been solved by implementing the process explained by [21] with some changes to adapt the approach to the AMPEL case study. The approach aims to estimate the probability of each individual being poor by utilizing an array of vector weights to comprehend the characteristics of multidimensional poverty. Rather than selecting a single weight for each feature, this method enables the creation of an approximation of the entire space of feasible weights. These weights make it possible to label the individuals in the dataset according to the degree of poverty: i) high, ii) medium or iii) low.

The method starts by considering the initial dataset, as a n-by-p matrix $Y = [Y_{kj}]$, where n is the number of individuals (n=429) and p the size of the selected feature vector (p=99).

Each row of the matrix $y_k = (Y_{k1}, ..., Y_{kj}, ..., Y_{kp})$ gives the features of the k-th individual, while each column $y_j = (Y_{1j}, ..., Y_{kj}, ..., Y_{nj})$ gives the distribution of the j-th feature across individuals.

It is important to note that the values of the Y matrix are heterogeneous and can be ordinal, binary, or numeric. Matrix Y is thus mapped into a *deprivation matrix* $G = [G_{kj}]$ of values belonging to the range $[0, 1]$, with binary values associated to categorical features obtained applying proper thresholds and continuous real values for numeric ones. The mapping rules have been defined by the domain expert. This expert knowledge permits to define for each feature its deprivation level. An example of this mapping is reported below.

From G, summing over the rows, it is possible to calculate a deprivation score $c = (c_1, ..., c_k, ..., c_n)$ for each individual. The score represents the poverty associated to each individual, so the higher the score, the more deprived the individual will be. However to better reproduce the reality, a vector $v = (v_1, ..., v_j, ..., v_p)$ of weights should be defined to properly consider the contribution of each feature, where $\sum_{j=1}^{p} v_j = 1$ and $0 \leq v_j \leq 1$.

Thus the deprivation score c is obtained as:

$$c = G \times v^{\top} \tag{1}$$

Instead of defining a single vector of weights, following Liberati et al. [21], a set of m weight vectors are randomly generated from a uniform distribution, to explore the whole feasible weight space. In this work $m = 10.000$. The output of this step is a matrix $V = [V_{sj}]$ of size m-by-p, where m is the number of weight vectors. Each row $v_s = (V_{s1}, ..., V_{sj}, ..., V_{sp})$ corresponds to a vector of weights, while each column $v_j = (V_{1j}, ..., V_{sj}, ..., V_{mj})$ represents the distribution of the weights assigned to each feature.

From matrix G and V we can obtain a matrix of deprivation scores C as follows:

$$C = G \times V^{\top} \tag{2}$$

In the deprivation score matrix $C = [C_{ij}]$, of size n-by-m, each row $c_k = (C_{k1}, ..., C_{ks}, ..., C_{km})$ contains m deprivation scores for the k-th individual based on changes in weight vectors, while each column $c_s = (C_{1s}, ..., C_{js}, ..., C_{ns})$ represents the distribution of deprivation scores across individuals for a given weight vector s. This matrix can be considered as a "deprivation embedding matrix", where each vector $c_k = (C_{k1}, ..., C_{ks}, ..., C_{km})$ can be viewed as an embedded representation of the poverty level of k-th individual.

Starting from matrix C, the poverty rank matrix $R = [R_{ij}]$, with dimension n-by-m, is defined. Each vector $r_k = (R_{k1}, ..., R_{kj}, ..., R_{km})$ represents the poverty ranks of the $k - th$ individual with respect to his/her deprivation scores and can be defined on the basis of the following ranking function:

$$R_{ks} = 1 + \sum_{i \neq k} \rho [C_{ik} > C_{ks}] \; for s = 1, ..., m \tag{3}$$

where $\rho = 1$ when the condition in square brackets is met, and $\rho = 0$ when the condition is not met.

For each vector of weights, the rank R_{ks} of the k-th individual is obtained by summing the number of individuals whose deprivation score C_{ik}, is higher than the deprivation score C_{ks} of the considered individual, plus one. In other words, R_{ks} is equal to one plus the number of individuals who are "multidimensionally" poorer than

$$\begin{bmatrix} 2 & 2 & 2 & 8 \\ 2 & 2 & 2 & 7 \\ 1 & 2 & 3 & 7 \\ 2 & 1 & 2 & 5 \\ 1 & 2 & 1 & 1 \end{bmatrix} \longrightarrow \begin{bmatrix} 0 & 0 & 0 & 0.2 \\ 0 & 0 & 0 & 0.3 \\ 1 & 0 & 1 & 0.3 \\ 0 & 1 & 0 & 0.5 \\ 1 & 0 & 0 & 0.9 \end{bmatrix} \times \begin{bmatrix} 0.10 & 0.36 & 0.03 & 0.88 \\ 0.53 & 0.67 & 0.85 & 0.19 \\ 0.71 & 0.31 & 0.08 & 0.99 \\ 0.27 & 0.22 & 0.75 & 0.27 \end{bmatrix}^{\top} = \begin{bmatrix} 0.18 & 0.04 & 0.20 & 0.05 \\ 0.26 & 0.06 & 0.30 & 0.08 \\ 0.39 & 1.44 & 1.09 & 1.10 \\ 0.80 & 0.77 & 0.81 & 0.36 \\ 0.89 & 0.70 & 1.60 & 0.51 \end{bmatrix} \longrightarrow \begin{bmatrix} 5 & 5 & 5 & 5 \\ 4 & 4 & 4 & 4 \\ 3 & 1 & 2 & 1 \\ 2 & 2 & 3 & 3 \\ 1 & 3 & 1 & 2 \end{bmatrix}$$

the $k-th$ individual. Consequently, the higher the value of R_{ks}, the lower the poverty of $k-th$ individual.

Finally, a matrix of probabilities of being poor $B = [B_{kr}]$ with dimension n-by-n is defined. Each row $b_k = (B_{k1}, ..., B_{kr}, ..., B_{kn})$ gives for the k-th individual, his/her probabilities of occupying a rank from 1 to n in the poverty rank matrix R.

The probability B_{kr} that the $k-th$ individual occupies the poverty rank r in the considered population is defined as:

$$b_{kr} = \frac{V_{kr}}{m} \tag{4}$$

where:

$$V_{kr} = \sum_{s=1}^{m} \phi[R_{ks} = r]; \tag{5}$$

$\phi = 1$ every time $R_{ks} = r$ and $\phi = 0$ otherwise. An example of the creation of the B matrix is reported below.

$$\begin{bmatrix} 5 & 5 & 5 & 5 \\ 4 & 4 & 4 & 4 \\ 3 & 1 & 2 & 1 \\ 2 & 2 & 3 & 3 \\ 1 & 3 & 1 & 2 \end{bmatrix} \longrightarrow \begin{bmatrix} 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 0.5 & 0.25 & 0.25 & 0 & 0 \\ 0 & 0.5 & 0.5 & 0 & 0 \\ 0.5 & 0.25 & 0.25 & 0 & 0 \end{bmatrix}$$

The last step is to label the individuals within the three poverty levels: i) Elderly people with a high risk of poverty (red class); ii) Elderly people with a medium risk of poverty (yellow class); and iii) Elderly people with a low risk of poverty (green class). To this end, as a first attempt, the poverty ranks from 1 to n have been divided into three homogeneous groups. The cumulative probability of each individual for each of these three groups have been evaluated. The class of the considered individual is the one corresponding to the group with the highest cumulative probability. Depending on the initial choice of the three groups, the class assigned to the individuals can be different. The strategy related to this choice should be discussed with domain experts, stakeholders, public institutions and municipalities.

This matrix can be considered as a "deprivation embedding matrix", deprivation scores

In order to analyze the three groups of individuals obtained, the deprivation score matrix C, considered as a deprivation embedding matrix, is analyzed. The score

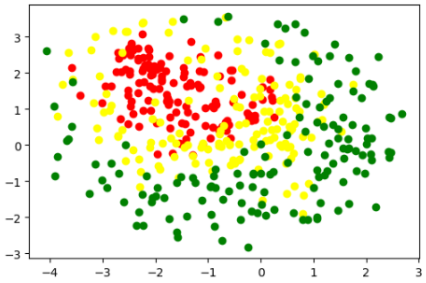

Figure 1: t-SNE representation of deprivation score matrix.

vectors with a dimension $m = 10000$ have been reduced by applying the dimensionality reduction algorithm *T-distributed Stochastic Neighbor Embedding* (t-SNE) [22] which uses the cosine similarity metric for calculating the distance between instances in a feature array. The corresponding vector space plot is represented in Figure 1. This representation shows how high poverty risk (red) and low poverty risk (green) individuals are clearly separated, while the medium poverty risk (yellow) ones in the representation are more scattered.

4.2. Poverty risk classification

Starting from the class labels assigned with the process described above, and the 99 Tapas features in matrix Y it is possible to train a model to infer the individual risk of poverty. The machine learning model adopted is the*XGBoost* model. This model is based on a gradient-boosting algorithm that uses a set of weak decision trees to make solid predictions. XGBoost also incorporates regularisation techniques to prevent overfitting and enhance the model generalisation ability. A 10-fold cross-validation technique was applied during the evaluation process to evaluate the model accuracy. Moeover, XGBoost permits to estimate the relative importance of each feature in the dataset to predict the target variable. Feature scores are calculated considering the number of times a feature is used to split the data across all decision trees in the model. The higher the number of splits on a feature, the higher its importance. In Figure 2 the first ten features are reported, with their relative importance.

In Figure 3, the model accuracy is reported by varying the number of features depending on their importance.

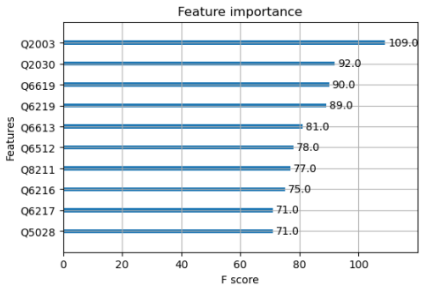

Figure 2: Feature importance with XGBoost.

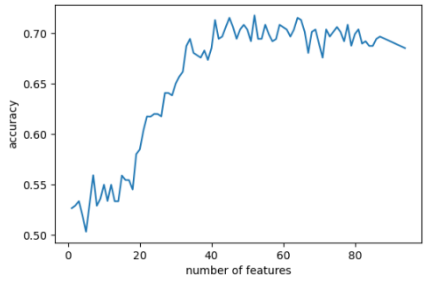

Figure 3: Accuracy score obtained using10-fold cross validation varying the number of features.

Plotting the XGBoost accuracy allows a visual understanding of which configuration can be adopted. In this work, the individuals for the three classes are 145 for the low risk class (green), 148 for the medium risk class, (yellow), and 136 for the high risk class (red). Considering 52 features, the model here proposed reached an accuracy of 72%. Figure 3 shows how starting from 40 features the accuracy begins to swing between 69% and 72%. A feature selection can be applied to decrease the quantity of starting variables selected by the domain expert by nearly fifty per cent. Eliminating unnecessary features is a key point to improve the model generalisation ability, increasing computational efficiency and providing a better understanding of the data.

5. Implementation

The entire pipeline has been implemented using Python. The libraries used to manipulate data and perform mathematical operations are respectively *pandas*[6] and *numpy*[7]. In order to perform some analysis and show the outcome of the labelling process, a dashboard to visualize data is available at the following link: Ampel Dashboard.

The overall software architecture of AMPEL project

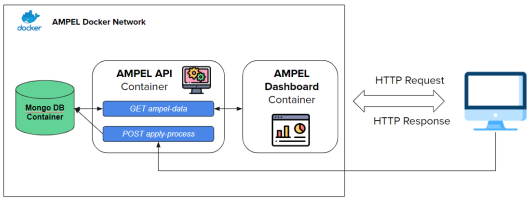

Figure 4: AMPEL Software architecture.

is represented in Figure 4 and AMPEL Repository [8] is publicly available, so the code can be downloaded and customised if needed.

The dashboard is composed of two parts:

- **Data Analysis**: It shows poverty statistics in the elderly population with respect to the three classes labelled. Labels can be modified changing the distribution of the ranks within this groups and consequently the cumulative probability of each of them. Four statistics have been added to the dashboard:
  1. **Qualification**: It shows statistics related to poverty in the elderly by showing results categorized on qualification levels, such as primary school and degree.
  2. **Age**: It shows statistics related to poverty in the elderly by showing poverty data per age.
  3. **Status**: It shows statistics related to poverty in the elderly by showing results categorized on status such as married, divorced etc.
  4. **Map**: Describes regions within a map that exhibit a higher concentration of poverty clusters.
- **Vector Space Representation**: This section shows the representation in order to dynamically modify the size of the representation space by modifying the ranks considered in the three groups adopted to define the poverty levels.

Finally, the **xgboost** library in python has been adopted for both feature selection and classification.

The *sklearn*[9] library, which provides tools for supporting machine learning tasks, has been used to implement the 10-fold cross-validation and apply it on the XGBoost model.

[6]https://pandas.pydata.org/
[7]https://numpy.org/

[8]https://gitlab.com/Fabio597/ampel
[9]https://scikit-learn.org/stable/

6. Conclusion and future works

This work proposes a strategy to label a dataset of heterogeneous sources of information, in order to classify older people into three classes of risk of poverty. The labelled data is here adopted in a traditional machine learning model that reaches an accuracy of about 72%. This labelled data can also be considered to develop a Bayesian Network (BN). The adoption of BN could be significant in the definition of multidimenisional poverty, as BN are self-explainable and they allow to know which variables led to specific result and to what extent each single data contribute to the final result. Moreover BN can easily integrate domain knowledge, keeping mutual interference among all the considered variables.

Acknowledgments

This research is supported by the FONDAZIONE CARIPLO "AMPEL: Artificial intelligence facing Multidimensional Poverty in ELderly" (Ref. 2020-0232).

References

[1] Health at a glance 2019: Oecd indicators, 2019.
[2] A. B. Atkinson, Income distribution in oecd countries, Evidence from Luxemburg income study (1995).
[3] M. Biewen, Income inequality in germany during the 1980s and 1990s, Review of Income and Wealth 46 (2000) 1–19.
[4] M. F. Förster, M. Mira D'Ercole, Income distribution and poverty in oecd countries in the second half of the 1990s, Income Distribution and Poverty in OECD Countries in the Second Half of the 1990s (February 18, 2005) (2005).
[5] M. Adena, M. Myck, Poverty and transitions in health in later life, Social science & medicine 116 (2014) 202–210.
[6] B. Nolan, C. T. Whelan, Measuring poverty using income and deprivation indicators: alternative approaches, Journal of European Social Policy 6 (1996) 225–240.
[7] P. Townsend, Poverty in the United Kingdom: a survey of household resources and standards of living, Univ of California Press, 1979.
[8] M. Terraneo, A longitudinal study of deprivation in european countries, International Journal of Sociology and Social Policy (2016).
[9] R. Boarini, M. M. d'Ercole, Measures of material deprivation in oecd countries (2006).
[10] S. P. Jenkins, L. Cappellari, Summarizing multiple deprivation indicators (2007).
[11] P. Prosperi, I. Peri, G. Vindigni, Problematiche aperte nell'analisi della povertà: questioni di misura e progressi nel raggiungimento degli obiettivi del millennio, Problematiche aperte nell'analisi della povertà: questioni di misura e progressi nel raggiungimento degli obiettivi del millennio (2011) 427–446.
[12] S. Alkire, J. E. Foster, S. Seth, M. E. Santos, J. Roche, P. Ballon, Multidimensional poverty measurement and analysis: chapter 3–overview of methods for multidimensional poverty assessment (2015).
[13] W. Bank, Poverty and shared prosperity 2020: Reversals of fortune, The World Bank, 2020.
[14] G. Li, Z. Cai, X. Liu, J. Liu, S. Su, A comparison of machine learning approaches for identifying high-poverty counties: Robust features of dmsp/ols night-time light imagery, International journal of remote sensing 40 (2019) 5716–5736.
[15] J. H. Mohamud, O. N. Gerek, Poverty level characterization via feature selection and machine learning, in: 2019 27th Signal Processing and Communications Applications Conference (SIU), IEEE, 2019, pp. 1–4.
[16] N. S. Sani, M. A. Rahman, A. A. Bakar, S. Sahran, H. M. Sarim, Machine learning approach for bottom 40 percent households (b40) poverty classification, Int. J. Adv. Sci. Eng. Inf. Technol 8 (2018) 1698.
[17] A. Abu, R. Hamdan, N. Sani, Ensemble learning for multidimensional poverty classification, Sains Malaysiana 49 (2020) 447–459.
[18] E. Guastafierro, et al., Social network and environment as determinants of disability and quality of life in aging: Results from an italian study, Frontiers in Medicine 9 (2022).
[19] M. Leonardi, et al., Determinants of health and disability in ageing population: the courage in europe project (collaborative research on ageing in europe), Clinical psychology & psychotherapy 21 (2014) 193–198.
[20] E. Guastafierro, et al., Identification of determinants of healthy ageing in italy: results from the national survey idagit, Ageing & Society 42 (2022) 1760–1780.
[21] P. Liberati, G. Resce, F. Tosi, The probability of multidimensional poverty: A new approach and an empirical application to eu-silc data, Review of Income and Wealth (2022).
[22] L. Van der Maaten, G. Hinton, Visualizing data using t-sne., Journal of machine learning research 9 (2008).