Intelligenza Artificiale 18 (2024) 51–65
DOI 10.3233/IA-240027
IOS Press

Facing multidimensional poverty in older adults: An artificial intelligence approach that reveals the variable relevance

Lorenzo Olearo*,[1], Fabio D'Adda, Enza Messina, Marco Cremaschi, Stefania Bandini and Francesca Gasparini
Department of Computer Science, Systems and Communications, University of Milano - Bicocca, Milan, Italy

Abstract. Despite the rapid development in very recent years of Artificial Intelligence models to predict poverty risk, this problem still remains an unsolved open challenge, especially from a multidimensional perspective. One of the main challenges is related to the scarcity of labelled and high-quality data for training models coupled with the lack of a general reference model to build good predictors. This results in the proposal of a variety of approaches tailored to specific contexts. This paper presents our proposal to address multidimensional poverty prediction, starting from an unlabelled dataset. We focus on the case of a fragile population, the older adults; our approach is highly flexible and can be easily adapted to various scenarios. Firstly, starting from expert knowledge, we apply a stochastic method for estimating the probability of an individual being poor, and we use this probability to identify three levels of risk. Then, we train an XGBoost classification model and exploit its tree structure to define a ranking of feature relevance. This information is used to create a new set of aggregated features representative of different poverty dimensions. An explainable novel Naive Bayes model is then trained for predicting individuals' deprivation level in our particular domain. The capacity to identify which variables are predominantly associated with poverty among older adults offers valuable insights for policymakers and decision-makers to address poverty effectively.

Keywords: Multidimensional poverty, poverty prediction, older adults, feature ranking, naive bayes, XGBoost

1. Introduction

Poverty, as highlighted by the Organization for Economic Cooperation and Development (OECD), is undeniably one of the most significant social problems of our global community. It is crucial to recognise that poverty impacts not only developing countries but also vulnerable groups within all societies. The Council of Europe[2] reveals the profound correlation between poverty and human rights, emphasising that extreme wealth imbalance results in extreme inequality in the realisation of fundamental human rights. Therefore, addressing poverty is tied to mitigating this inequality, and being able to predict it becomes a crucial first step in defeating it. The application of AI to the problem of poverty prediction represents a recent development, with significant early work published in 2016 [26], and a rapid increase in interest in 2021 [44]. The increasing use of AI in addressing poverty, and more broadly, in achieving the Sustainable Development Goals, particularly over the past two years, is closely related to the cumulative impacts of the COVID-19 pandemic, the Russia-Ukraine conflict, and climate change [9, 40]. At first, poverty and its correspond-

*Corresponding author: Lorenzo Olearo, Department of Computer Science, Systems and Communications, University of Milano - Bicocca, Milan, Italy. E-mail: lorenzo.olearo@unimib.it.

[1]This research is supported by the Fondazione CARIPLO "AMPEL: Artificial intelligence facing Multidimensional Poverty in ELderly" (CUP H45F20000840007, Ref. 2020-0232).

[2]www.coe.int.

ing indicators were considered solely in relation to monetary aspects, mainly measuring income-related indicators [7, 10, 21]. However, recent developments recognise that poverty is the product of a multifaceted interaction of various aspects, requiring the definition of multidimensional indicators [42, 43]. The concept of multidimensional poverty includes several definitions and measurement methodologies, reflecting different approaches to data collection and analysis. Moreover, addressing the needs of vulnerable populations requires the identification of suitable measures capable of capturing their peculiarities.

The research presented in this paper is related to the AMPEL Project (Artificial intelligence facing Multidimensional Poverty in ELderly) that faces the challenge of predicting poverty within a particularly vulnerable demographic: older adults. In this context, income-based indicators are recognised as inadequate proxies for assessing material conditions [3, 34] whereas non-monetary indicators offer enhanced insight into identifying those experiencing deprivation. The primary goal of the project is to establish a three-level poverty risk classifier capable of effectively identifying individuals requiring urgent assistance, particularly in the case of emergencies. As a secondary outcome of no less importance, the project aims to identify which indicators related to the elderly population are relevant for describing multidimensional poverty. To this end, a questionnaire has been crafted to capture correlated dimensions such as economic status, needs, health conditions, loneliness, and social interactions, among others. With the invaluable help of Auser[3] volunteers, the questionnaires have been distributed to approximately 500 senior citizens in Lombardy. This represents a unique and valuable dataset, acquired at a individual-based level, which the authors will provide upon request.

In this paper, we present our classification approach aimed at identifying three levels of poverty risk, considering the following key issues:

1. Pre-process the collected data to ensure data quality and remove noise from the features used to feed ML algorithms;
2. Apply a procedure to automatically assign labels to the collected data, starting from preliminary information acquired from domain experts and taking into account the multifaceted aspects of poverty;
3. Identify which are the features that are mostly relevant to predict poverty risk;
4. Apply different supervised Machine Learning models on a set of aggregated features selected from the relevant ones;
5. Identify the variables that are mainly responsible for predicting poverty in the elderly population, thereby directing the efforts of policymakers and municipalities to prevent and effectively address it.

The paper is organised as follows. A preliminary analysis of the state of the art is reported in Section 2, considering the role of AI in predicting multidimensional poverty. Subsection 3.1 presents the AMPEL dataset and its sources. In Subsection 3.2, the proposed framework for the analysis is described, with a particular emphasis on data pre-processing and data labelling (Subsections 3.3 and 3.4), on the feature relevance analysis (Subsection 3.5), and on different classification strategies (Subsection 4). In Section 5, a brief description of the data visualisation dashboard that has been implemented is reported. Eventually, we conclude this paper by discussing the advantages and limits of our work in Section 6 and drawing the future research directions in Section 7.

2. State of the art

Despite poverty being a well-recognised issue, the use of AI models for predicting poverty remains relatively recent, with notable developments observed since 2016 [11, 26], and experiencing a considerable acceleration starting from 2021. Literature reviews or surveys systematically evaluating the contribution of AI within this field are scarce [25, 44]. Particularly, the contribution of AI models has become increasingly significant in recent years, in conjunction with a shift in perspective on poverty, which has progressively been acknowledged as a multidimensional issue intricately tied to various facets and unique characteristics of the targeted population, including factors such as age and geographic location. Multidimensional measures of deprivation include diverse indicators fitting into a synthetic scale [12, 14], which is deemed to reflect basic living standards and the inability to meet the minimum acceptable way of life in one's own society. Several methodologies for assessing poverty from a multidimensional perspective exist, including those aiming to aggregate data from different sources, and statistical approaches, *i.e.*,

[3] Auser is an Italian voluntary and social promotion association committed to promoting active ageing of the elderly and enhancing their role in society (www.auser.it).

principal component analysis, or cluster analysis [4], which are considered appropriate when they effectively capture the joint distribution of deprivations, identify those experiencing poverty, and yield a single cardinal figure for evaluating poverty levels.

One major challenge in predicting poverty using AI lies in the scarcity of high-quality labelled data. Various proxies, such as the Proxy Means Test (PMT) labels, can be considered as ground truth for ML training. However, these proxies are not easily verifiable [33]. Traditionally, datasets for poverty prediction are related to demographic and livelihood indicators [1], household surveys [5, 39], or remote sensing data, such as satellite images or geospatial data [23, 26, 28], but to the best of our knowledge, no dataset at a subject-based level has been shared with the research community. Among different ML techniques for poverty classification, decision tree [39, 47], random forest [13, 23, 33], and ensemble approaches [1, 29, 48] are the most used.

The pioneering work by Jean et al. [26] proposed to analyse high-resolution satellite images to predict poverty considering a Convolutional Neural Network (CNN) model pre-trained on ImageNet, and the more recent work of Wijaya et al. [46] applied a deep neural network to estimate city-level poverty starting from e-commerce data, considered as indicators of consumption and purchasing power. Despite these advancements, deep learning models have not been widely adopted, instead, traditional machine learning approaches based on hand-crafted features and explainable AI models [49] are preferred. These approaches excel in identifying specific characteristics associated with poverty, which is essential for developing targeted poverty prevention strategies. In line with this, AI models have been applied for feature selection [41] and feature ranking [6] to create models that rely only on the most important variables [39].

3. Materials and methods

3.1. The AMPEL Dataset

A questionnaire designed by domain experts was administered to senior adults by trained interviewers, to gather data on various aspects of their living conditions. We have considered several measures and multidimensional poverty indexes as well as standardised questionnaires. Finally, we decided to follow the social inclusion approach introduced in [42]. An individual can be poor, that is, socially excluded, despite having adequate income [35], when they do not participate in key activities of the society in which they live. In fact, the concept of exclusion refers to the systematic exclusion of individuals, families, and groups from economic, political, and social activities that are fundamental to the quality of life [22, 28]. Based on these theoretical justifications, therefore, it was decided to prepare a questionnaire aimed at measuring the many factors that contribute to the material and social deprivation of older people, namely: socioeconomic conditions (economic stress, material deprivation, housing conditions); difficulty in accessing health care and services; health (general health conditions, physical and sensory limitations, chronic diseases and conditions, risk factors, psychological well-being) [36]; daily life (social support, generalised trust, safety, social relationships, participation in social activities); and subjective well-being.

Table 1
Target municipalities for the AMPEL dataset

AUSER senior center	Number of elderly
Lecco	71
Lodi	62
Cantù	51
Cinisello	61
Cernusco sul Naviglio	31
Pioltello	59
Prealpi-Milano	33
Cologno	62
Gallarate	66
TOTAL	**496**

The resulting questionnaire thus covered topics such as monetary aspects, environment, social networks, and quality of life, with a total of 125 variables (features) for each individual. The questionnaires were administered to 496 people, (336 females) with an average age of 76.4 (standard deviation = 8.9). The questionnaire in the original Italian version, together with its English translation are available to the research community[4]. Both the questionnaire and the protocol chosen to administer it have been reviewed and approved by the Research Ethics Committee at The University of Milano-Bicocca, Italy.
The geographical distribution of the elderly subjects considered in our study is illustrated in both Table 1 and Fig. 1.

The collected variable type can be i) Categorical variables, which assume a fixed set of values and include binary variables (*e.g.*, private health insurance coverage: yes or no), or ii) Numerical variables, which can be either discrete (*e.g.*, the number of indi-

[4] ampel.inside.disco.unimib.it/dataset.

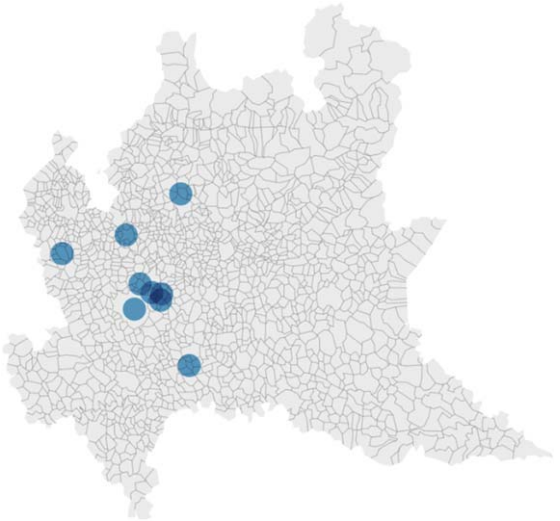

Fig. 1. Map of the considered municipalities in Lombardy - Italy.

viduals that live in a house) or continuous (*e.g.*, the height and weight of a person). It should be noted that in the collected dataset, the numerical features are considerably less than the categorical ones, a topic that will be addressed in Section 3.3.

The collected variables captured by the questionnaire can be grouped in 5 different categories as suggested in [42]:

- **Maintenance capacity**: measures the economic stress and a household's financial situation, emphasising the importance of household income in measuring deprivation;
- **Consumption deprivation**: represents the long-term financial hardship of the household. It does not reflect the capacity for current spending; instead, it includes indicators related to limited ownership of goods;
- **Health status**: is the most significant category influenced by social conditions, manifesting in human bodies as the cumulative result of socio-economic challenges encountered throughout individuals' life courses;
- **Housing facilities**: are related to the quality of housing, such as basic amenities, state of repair and so on;
- **Social and context deprivation**: reports information about social relations.

3.2. The proposed framework of analysis

The prediction of multidimensional poverty using innovative AI based solutions requires addressing methodological challenges to face the complexity of the task, especially in cases in which the dataset comprises a limited sample size (496 individuals) and includes missing values due to participant information gaps. In this Section, a methodology is presented to address these critical data issues and the lack of labelling. This approach is illustrated through a pipeline of four distinct phases, as in Fig. 2:

1. **Data cleaning**: This step focuses on identifying and rectifying errors or inconsistencies in the dataset to improve its quality, accuracy, and reliability. It plays a crucial role in the subsequent steps, as the data quality directly impacts the validity and effectiveness of the machine learning models. Moreover, feature engineering is required to correct type heterogeneity within the dataset;
2. **Data labelling**: as the ground truth needed to train the model is not provided, this step proposed a stochastic approach to assign labels to the collected data, starting from domain expert knowledge;
3. **Feature relevance**: The algorithm relies on multidimensional poverty to consider the different levels of deprivation of an individual. This phase aims to find which features have the most significant impact on the definition of poverty;
4. **Poverty prediction**: The last step aims to build a machine learning model using labelled data and feature relevance information to classify new, previously unseen data instances and, ultimately, define a three-level poverty predictor.

3.3. Data cleaning

This dataset requires essential preliminary data pre-processing steps to clean and normalise the data

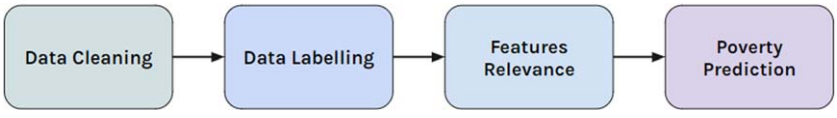

Fig. 2. The proposed framework.

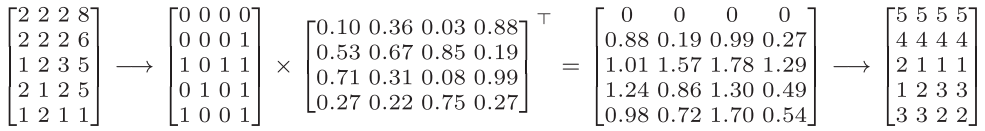

Fig. 3. Data labelling: the first matrix represents for each row a subject, for each column a feature. The second matrix is the Deprivation Matrix, obtained by applying the cut-offs defined by the domain expert. This matrix is multiplied with a matrix of weights to obtain the Deprivation Score Matrix. Our procedure extracts 10000 different weight vectors from a uniform distribution. In the example reported in figure, $m = 4$. The last matrix is the Poverty Indicator Matrix, which reports the poverty rank, of each individual within the considered population.

entries. Firstly, generalities and irrelevant features, such as those with an excessive number of missing or noisy values, are removed. When possible, missing values are filled with relevant ones. For instance, if an individual lacks a specific condition and does not answer the relevant question in the form, all such entries are thus filled with a categorical value representing the absence of that specific condition.

The dataset comprises data collected from a survey, including features that are answers to open-ended questions, presenting considerable name heterogeneity, with answers that are semantically equivalent but syntactically different. To mitigate this effect, whenever feasible, semantically similar values are mapped to a common one.

Related and dependent features represent another significant issue in this dataset. To reduce redundancy, the dataset is simplified by grouping redundant features into a single meaningful one.

The majority of the features in our dataset are categorical, and only a few are numerical, such as age or number of cigarettes smoked weekly.

Considering the small number of observations in this dataset, namely 496, the numerical features are simplified, reducing their information by *binarising* their values. For example, instead of representing the number of people living with an individual, a possibly challenging numerical value to handle in a categorical dataset, this information is transformed into whether this individual lives alone. This type of feature engineering is applied with the same logic over all the numerical features remaining in the dataset, thus transforming it into an entirely categorical dataset. The steps illustrated above are repeated on the whole dataset and allow reducing the number of features from 125 to 103. From a computational perspective, this feature reduction may not be significant, but it could help to eliminate some noise, enhancing the quality of the classification results and providing better explainability.

3.4. Data labelling

Due to the absence of a standard definition and a specific indicator of multidimensional poverty, the challenge lies in determining the poverty risk level of each individual in the dataset to build a ground truth on which Machine Learning prediction models can be trained.

This issue has been solved by applying the framework described in a previous study [20] and originally proposed by Liberati et al. [30]. This approach assesses the likelihood of poverty for each individual by considering a weighted sum of the deprivation indices associated with each feature (established through a cut-off value). Instead of defining a fixed importance weight vector, this technique builds an embedded representation at the individual level, obtained by randomly sampling the weight distribution and evaluating an aggregated index. This procedure is crucial for classifying individuals within the dataset into three poverty levels: high (represented with red colour), medium (yellow colour), or low (green colour) depending on the resulting index distribution. As mentioned above, the implementation follows the methodology reported in [20], with adjustments made to the initial deprivation cut-offs to align with the characteristics of the new dataset. All the details of the implementation of this labelling method are reported in [20], but to understand how the vector space has been built, all the steps have been summarised below (and illustrated in Fig. 3):

1. **Load dataset**: This preliminary step is necessary to load the dataset and the relative cut-offs, and its purpose is also to prepare the data for the subsequent steps of the pipeline;
2. **Deprivation matrix**: The first phase entails mapping the initial dataset onto a *deprivation matrix* by applying the binary cut-offs, adequately defined by a domain expert for this specific data. The resulting binary matrix indi-

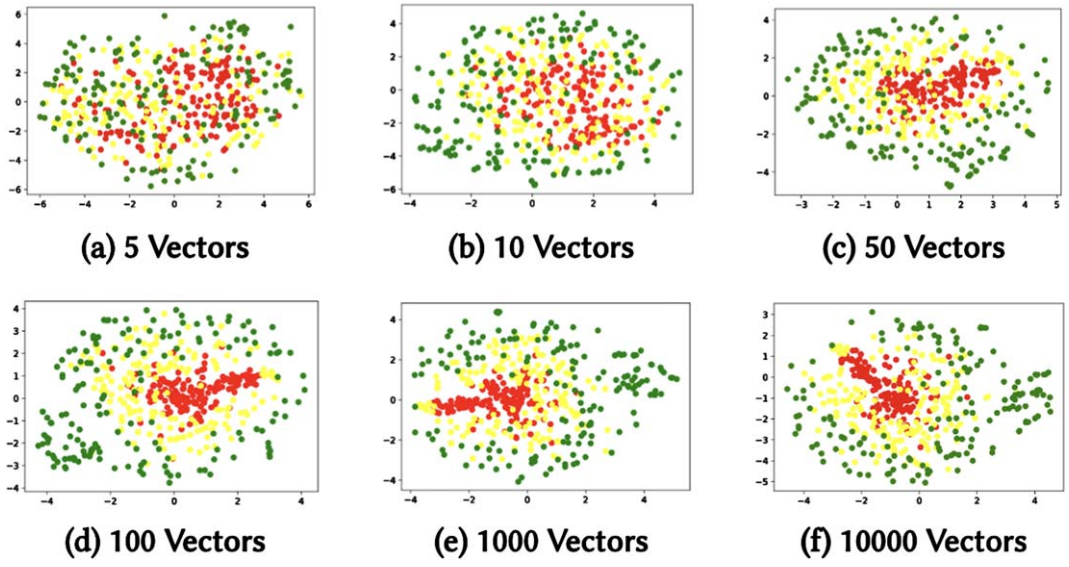

Fig. 4. TSNE representation of vector spaces by varying the number of weight vectors.

cates whether an individual is deprived or not, with respect to each feature;

3. **Deprivation score matrix**: The *deprivation matrix* calculates the poverty score for each individual by weighting each feature with a predefined score. Instead of defining a single vector of weights, Liberati et al. [30] propose to randomly sample a set of m weight vectors from a uniform distribution, allowing us to assess the index distribution depending on different importance factors; The final matrix contains m deprivation scores for each individual, computed by using the m samples of weight. Each vector can be considered as an embedding representation of the poverty level of each individual. In this work, we use $m = 10000$;
4. **Poverty indicator matrix**: Starting from the *deprivation score matrix*, the *poverty indicator matrix* is defined. This matrix represents the poverty ranks of each individual based on their deprivation scores within the considered population. For each weight vector, an individual's rank is computed by adding the number of individuals with a higher deprivation score than the current individual's score, plus one. In simpler terms, the score is one plus the count of individuals who are multidimensionally poorer than the current individual. Therefore, the higher the score, the lower the poverty level;
5. **Probability matrix**: The last phase builds a *probability matrix*, delineating the probabilities of individuals being in poverty. Each entry in the matrix denotes the likelihood of each individual occupying a rank from 1 to N in the final poverty matrix, where N is the number of individuals considered.

Using their ranking distribution, reported in the Poverty Indicator Matrix, individuals can be classified into three levels of poverty, high (visually coded with red colour), medium (visually coded with yellow colour) or low (visually coded with green colour).

To qualitatively analyse the results of this labelling procedure, we show, in Fig. 4, the distribution of individuals belonging to different classes, for increasing sample size m. Here the 2D spatial representation of the embedding vectors using TSNE [45] (T-distributed Stochastic Neighbor Embedding) is visualised. TSNE is a dimensionality reduction technique, commonly used in ML and data analysis for visualising high-dimensional data in a lower-dimensional space. In particular, we report how vectors representing individuals of different classes are distributed in the embedding representation space considering m equal to 5, 10, 50, 100, 1000 and 10000 respectively. We note that by increasing m, the three groups of individuals tend to cluster and become visually distinguishable in the representation space.

This behaviour better explains the advantages of employing a set of randomised vectors:

– **Exploration of solution space**: randomised vectors allow the exploration of a broader range

Table 2
Groups size depending on number of weight vectors

# Vectors	Red	Yellow	Green
2	174	163	159
4	167	168	161
5	162	169	165
10	167	162	167
50	164	165	167
100	163	166	167
1000	162	167	167
10000	162	167	167
100000	162	167	167

of solutions within the feasible weight space. In contrast to having a unique vector of weights, which would lock the solution into a single configuration, randomised vectors provide a more comprehensive view of the solution space. This flexibility can be beneficial in cases where the optimal solution is not known in advance;
– **Robustness**: randomised vectors contribute to the robustness of the algorithm by making it less sensitive to variations in data;
– **Efficient exploration**: randomisation facilitates efficient exploration of the solution space by enabling the rapid exploration of a wide range of weight vectors, which is particularly useful when searching for good solutions in large and complex optimisation problems.

Using a sample size m adequately high allows the model to reach stable results in terms of label assignment to individuals, as shown in Table 2.

3.5. Feature relevance

One of the relevant aspects of this project is to understand the contribution of features belonging to different domain dimensions, to an individual's poverty. To address this issue, we analysed the feature correlation through Cramer's V [18], a measure of association between two nominal variables based on Pearson's χ^2 test. We observed that the only significant correlations are observed among features related to health status (HS). Therefore, to obtain an estimation of the feature relevance, we trained a machine learning model on the labelled dataset. In particular, the Information Gain value of each feature in the tree-based model XGBoost [15] was investigated.

Once fitted to the data, an XGBoost model can return the importance of each feature based on its contribution in predicting the right label. This importance can be defined following various metrics; in this case, particular interest lies in the gain score of each feature, defined as the average gain across all splits the feature is used in. The gain of a specific feature represents its relative contribution to the classification of each tree in the model; it quantifies the amount of information gained about the target class after adding the relative split in the model.

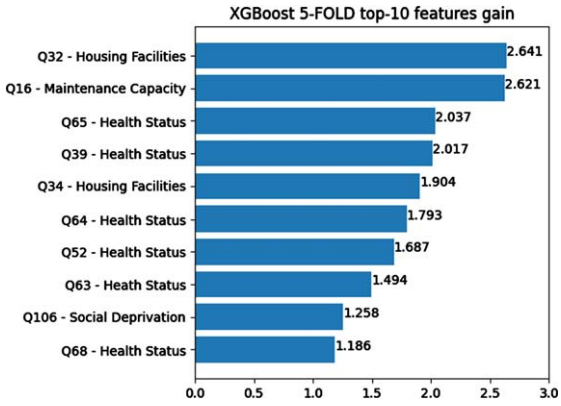

Fig. 5. Top 10 features by their gain scores obtained as the average of the scores across the 5 cross-validated models.

The information gain of a tree-based model is a powerful metric, but it can be biased, especially in case where the dataset has a large number of features (103 after the pre-processing step) and a low number of observations (496). In such circumstances, the model is at high risk of overfitting, as verified by our experiment[5].

To address this issue, a 5-fold cross-validation process is adopted. The dataset is initially divided into two subsets: 80% is used for training, while the remaining 20% is retained for testing. In the training subset of the dataset, a 5-fold cross-validation of the XGBoost model is carried on, and all of the 5 models fitted in the process are saved, along with their gain scores. To avoid overfitting, these scores are then averaged across the 5 models. Having computed these scores, it is now possible to rank each feature.

In Fig. 5 the variables in the top-10 positions are reported, considering the poverty dimension to which they were assigned by the domain expert (see Section 3.1). In Table 3 a detailed description of each of them is provided, sorted by their averaged gain scores across the cross-validated XGBoost models. The associated questions are reported in the last col-

[5]Gitlab page with experiment code: gitlab.com/Fabio597/ampel.

Table 3
Questions corresponding to the first top-10 features

Code	Dimension	Full question [Translated to English]
Q32	Health status [HS]	Are you affected by one or more of the following long-term diseases or pathological conditions? [*e.g.*, Neurodegenerative diseases (Alzheimer's disease, Parkinson's disease, or other senile dementias)]
Q16	Maintenance capacity [MC]	How do you evaluate the overall quality of your life?
Q65	Health status [HS]	Are you affected by one or more of the following long-term diseases or pathological conditions? [*e.g.*, Mood disorders (such as depression)]
Q39	Health status [HS]	Do you have difficulty walking on a flat surface for 500 meters without the help of another person or the use of aids (such as canes, crutches, wheelchairs, etc.)?
Q34	Housing facilities [HF]	Are you affected by one or more of the following long-term diseases or pathological conditions? [*e.g.*, Disorders affecting the nervous system (epilepsy, multiple sclerosis)]
Q64	Health status [HS]	Regarding the housing where you live, do one or more of the following conditions occur? [*e.g.*, The housing is in poor condition]
Q52	Health status [HS]	In the last 12 months, have you had to forgo healthcare services because you could not afford them? If yes, what services? [*e.g.*, Medical tests or treatments (excluding dental or orthodontic)]
Q63	Health status [HS]	Are you affected by one or more of the following long-term diseases or pathological conditions? [Recurrent headaches]
Q106	Social deprivation [SC]	In the last 12 months, have you been on vacation for at least one week (even if not continuous)?
Q68	Health status [HS]	Regarding the housing where you live, do one or more of the following conditions occur? [The housing is too small]

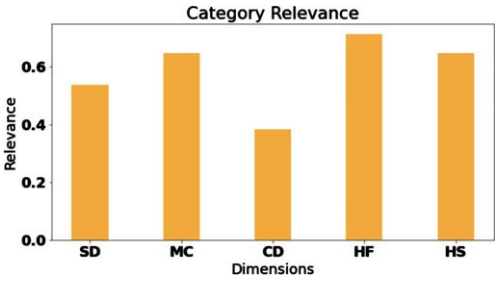

Fig. 6. Category Relevance computed using the gain scores obtained from XGBoost. The score has been normalised by the size of each dimension MC = Maintenance Capacity (10), CD = Consumption Deprivation (12), HS = Health Status (41), HF = Housing Facilities (11), and CD = Social and Context Deprivation (12).

umn. The complete list of the survey questions is made available for the research community in Italian (original version) and in English through the AMPEL Dashboard presented in Section 5.

From the analysis of Fig. 5 and Table 3, it emerges that 7 out of 10 among the top-10 features are assigned to the Health Status dimension. This finding is not surprising, as poverty is associated with health-related outcomes, including physical and mental disorders [24]. The presence of multiple conditions, *i.e.*, the comorbidity and multimorbidity status, negatively impacts wealth as people age: the presence of comorbidities was associated with 20–22% wealth loss over three-years [27] and such a negative effect extends from individuals to their networks [32]. Additionally, prevalent mental health conditions, such as depression, act as a booster for multimorbidity and total healthcare expenditure among older adults [17]. Moreover, a clear bi-directional relationship exists between adverse economic situations, health status and disability [8, 19].

The second most relevant variable is related to an overall evaluation of the quality of life, which is consistent with studies that highlight the strong relation between poverty, life satisfaction and happiness [2].

To give more insight, we also computed the relevance of each poverty dimension by averaging the gains of their features. Results, reported in Fig. 6, show a relatively balanced distribution. Despite the abundance of features related to the Health Status in the top-10 rank, depicted in Fig. 5, this dimension is not dominant among the five considered, and the most prominent one is the Housing Facilities.

Computing the relevance of both single features and poverty dimensions is of paramount importance, as it offers to municipal policymakers valuable insights, guiding them to identify appropriate interventions to effectively address and prevent poverty. In addition, this analysis enables to focus on a subset of questions that should be prioritised when administering a new questionnaire to a different sample of individuals.

Table 4
Performance of the XGBoost classifier

Class	Accuracy	Recall	F1-Score	Support
Green	0.73	0.90	0.81	21
Yellow	0.72	0.70	0.71	37
Red	0.89	0.81	0.85	42

Table 5
Performance of the Categorical Naive Bayesian classifier

Class	Accuracy	Recall	F1-Score	Support
Green	0.89	0.76	0.81	21
Yellow	0.72	0.89	0.80	37
Red	0.94	0.81	0.87	42

4. Experimental results of multidimensional poverty prediction

Three distinct classification models are examined in this study: XGBoost [16], Categorical Naive Bayes [37] and a novel hybrid approach that combines XGBoost feature gain scores with a Naive Bayes classifier, taking into account the five poverty dimensions defined in Section 3.1.

Through the following analysis, the same 80% portion of the dataset previously used for feature relevance is employed to train the presented models, while the remaining 20% is retained for the testing. Thus, all the models are trained on identical data and assessed on the same test set.

In this Section, the proposed models are evaluated considering three different metrics: *accuracy*, *recall* and *F1-score*.

4.1. The XGBoost classification model

Building upon our previous work [20], we seek to leverage the insight gained into the impact of each feature. We first examine XGBoost model in its categorical classification setting. This model provides the baseline for subsequent predictive models, and it is chosen because of several factors: it offers substantial interpretability by visualising the aforementioned feature importance, and both training and inference on the model are remarkably fast. With these considerations in mind, our XGBoost model achieves 79% of overall accuracy across the test set as reported in Table 4.

4.2. The categorical naive bayes classification model

To gain more explainablity, we also considered a Naive Bayes Classifier. Naive Bayes models assume conditional independence between the various features of the dataset. Such an assumption is reasonable given that the questions asked to the individuals cover different subjects, albeit sometimes related to the same topic. Once again, the model is trained on the entire training set, achieving slightly better results than the XGBoost model with a global accuracy of 84% as shown in Table 5.

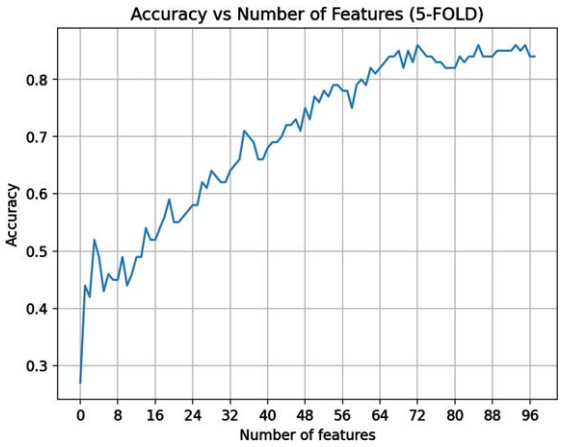

Fig. 7. Accuracy of the Categorical Naive Bayes classifier concerning the number of features selected from a list, ordered for the gain scores obtained applying the XGBoost model.

Increasing the number of features tends to add more information to the model. Ideally, this can lead to better discrimination between classes. However, this also has its downsides: as the number of features grows, so does the risk of overfitting the train data: the model might learn unrelated noise while losing the ability to generalise on the test set later.

Naive Bayes classifiers are known for their simplicity and effectiveness in many classification tasks [37]. However, they tend to be less informative as the number of features increases. To create a model as explainable as possible, it can be interesting to study how the model's performance changes while increasing the number of features it is trained with.

In Fig. 7, we present the performance evaluation of the model employing a 5-fold strategy, showcasing the model performance as we increment the number of features. Note that features have been selected following the order of the average gain score computed on the 5 cross-validated XGBoost models as discussed in Section 3.5.

The accuracy trend appears irregular, varying the number of features considered, probably due to the small number of observations in the dataset. How-

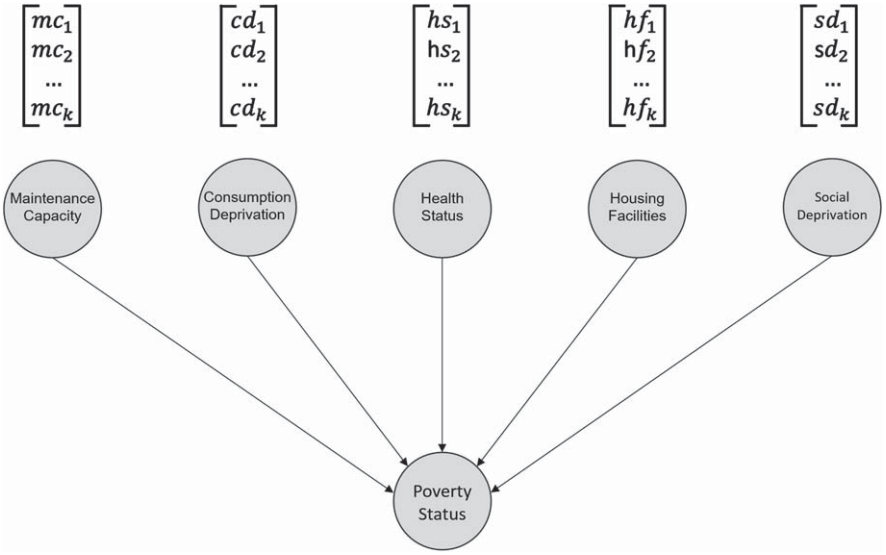

Fig. 8. Structure of the Naive Bayes classifier, where each of the 5 parent nodes represents the score in that poverty dimension for each individual.

ever, it is still possible to appreciate the incremental tendency that stabilises around the upper 60 ones.

4.3. The naive bayes multidimensional poverty classifier

The purpose of this model's introduction is to provide a predictive model that is as understandable as possible and that can function in a multidimensional environment using a reduced number of features. These dimensions correspond to the five groups defined in Section 3.1: *Social Deprivation, Maintenance Capacity, Consumption Deprivation, Housing Facilities and Health Status*.

For each dimension, a new feature is computed by linearly combining the XGBoost information gain of each feature with its relative cut-off defined by the domain expert. To improve the generalisation power of the model, the gain used for this linear combination is the one averaged from the 5 cross-validated models in Section 3.5.

Starting from the binary deprivation matrix defined in Section 3.4, each row describes the features where an individual is considered deprived. These rows are here referred to as *deprivation vectors*. Each column represents, for a given feature, the distribution of individuals that exceeds the cut-offs defined by the domain expert.

Let us define a as the vector of the feature information gains, i.e. each a_i *with* $i \in \{1, \cdots, n\}$ is the gain score of the i-th feature, and let b be the deprivation vector where each b_j *with* $j \in \{1, \cdots, n\}$ represents if and individual is deprived or not for the j-th feature. Then we define the individual deprivation score as:

$$a^\top b = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} \begin{bmatrix} b_1 \; b_2 \; \cdots \; b_n \end{bmatrix} = \textit{deprivation score} \tag{1}$$

If we group the elements of the deprivation vector into the five poverty dimensions, the resulting values represent the 'magnitude' of deprivation experienced by the individual in each of the five dimensions.

At this point, each individual is encoded into 5 values that will now be used to determine their poverty class. To this end, a Naive Bayes classifier is applied, having a parent node for each of the 5 classes of features as shown in Fig. 8.

Table 6
Performance of the Naive Bayes multidimensional poverty classifier

Class	Accuracy	Recall	F1-Score	Support
Green	0.95	0.95	0.95	21
Yellow	0.81	0.92	0.86	37
Red	0.95	0.83	0.89	42

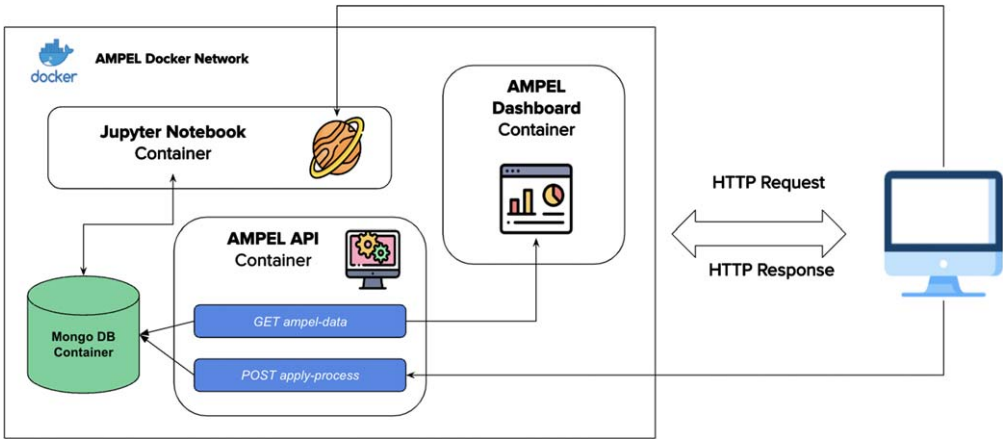

Fig. 9. AMPEL Software Architecture.

The classifier built on top of this pipeline shows good performance as reported in Table 6, reaching an overall accuracy of 89% on the test set. It is essential to mention that the test set has never been used in the entire pipeline except for assessing this metric. This result shows how well the model can generalise using the average gain extracted from the cross-validated XGBoost models introduced in Section 3.5.

5. The AMPEL Project: Dashboard

The project implementation includes multiple independent software modules that cooperate by exchanging data resulting from the analyses described above. Figure 9 shows the different software components implemented and deployed using *Docker containers*. A brief description of each part is reported below:

- **MongoDB**: *MongoDB* has been used to store AMPEL data. It is an open-source NoSQL database management system[6]. It falls under the category of document-oriented databases and is designed to handle large volumes of unstructured or semi-structured data;
- **API**: The deployable code runs in the *AMPEL API* container, facilitating the integration of various applications and technologies;
- **Jupyter notebook**: Most of the analyses have been implemented in a *Jupyter Notebook* as it allows to create and share documents containing live code, equations and visualisations;
- **Dashboard**: The *Dashboard* service, depicted in Fig. 9, serves as an important tool for data visualisation. It has been developed to provide the research community access to the collected dataset and analysis code. The dashboard is instrumental as it enables the visual analysis of the results carried out in the previous Sections.

Among the services outlined above, the AMPEL Dashboard(publicly available[7]) serves as a visual representation of essential information, data, metrics and performance indicators, organised and displayed in a centralised and easily accessible manner. The dashboard is designed to provide a quick and concise overview of critical information, allowing experts to monitor, analyse, and make informed decisions based on real-time and historical data. The Dashboard presents four pages: Home, Graphs, Dataset and Feature Relevant. i) The Home page serves as the starting point or landing page when users access the dashboard. Its primary purpose is to offer a centralised and organised view of key information, metrics, or features that users may need to quickly access or monitor. Additionally, it presents and explains the project's purpose (Fig. 10). ii) The Graphs page displays the information related to the analysis conducted on the elderly population. Its purpose is to provide a visual exploration of the analysis conducted by domain experts (Fig. 11). iii) The Dataset page presents all the questions, both in Italian and English versions, included in the questionnaire provided to each individual. iv) The last page, which is Feature Relevance, showcases the most important

[6] www.mongodb.com/it-it.

[7] ampel.inside.disco.unimib.it.

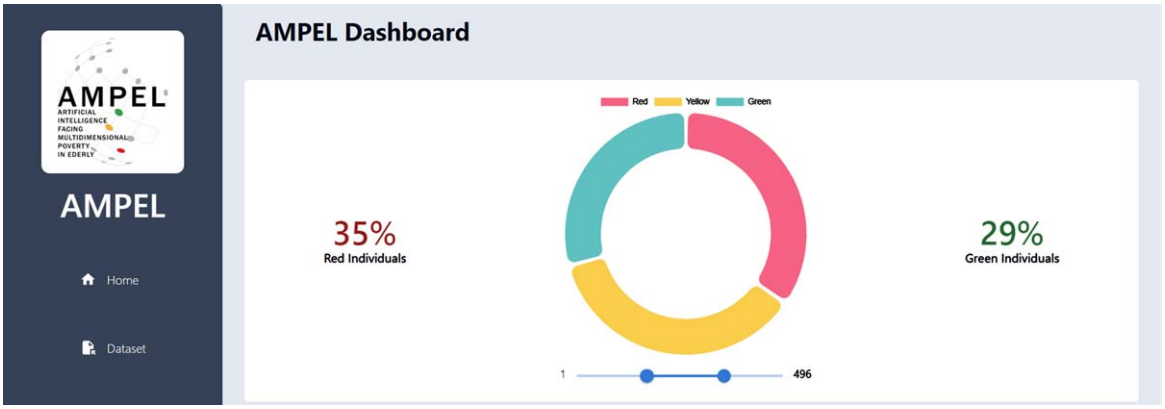

Fig. 10. Panel utilised by the domain expert for examining poverty within the population.

Fig. 11. Examples of different data visualisations for data analysis. Top left: distributions of subjects within the three poverty classes, considering different levels of education. Top right: distribution of the subjects within the three poverty classes, considering the gender. Bottom: geographic distribution of the considered subjects

questions in the dataset by understanding the contribution of each feature in making predictions.

Among the charts available on the Graph page of the dashboard, the main contributions are:

- **Poverty on degree**: poverty is analysed by categorising individuals according to their academic level;
- **Poverty on gender**: poverty is analysed by categorising individuals according to gender;
- **Poverty map**: poverty is analysed by highlighting on a geographical map the areas with the highest concentrations of poverty.

6. Discussion

This work introduces a novel approach to predicting multidimensional poverty using unlabelled data. While encouraging results were achieved, it's important to acknowledge both the limitations and strengths of this methodology.

Although the dataset is valuable because it focuses on individual subjects rather than demographic or aggregated information such as satellite images or geospatial data, it is important to emphasise the limits of observations.

On the contrary, the data collected refers to subjects distributed within a localised area in Lombardy,

Italy. While this may be considered a further limitation, it actually could increase the richness of the information contained in the dataset. It is noteworthy that multidimensional poverty can be described by variables that are strictly related to the peculiarities of the considered population, and a broader sample of subjects may hide some important aspects specific to smaller groups of individuals.

Another notable strength of the proposed approach is its reliance on domain experts' knowledge, which, while not limiting machine learning in data exploration, reveals new relationships and meaningful indicators in the definition of multidimensional poverty. Specifically, by leveraging the tree structure of XGBoost, it becomes possible to understand which variables are the most relevant for predicting poverty in the target population. Moreover, starting from their relevance, it is feasible to explore high-level poverty dimensions that can better encapsulate the primary aspects involved in poverty creation. In this work, we extracted the feature ranking from cross-validated XGBoost models. However, other approaches like SHAP [31] could be considered and potentially combined together. From preliminary comparisons, our ranking shows partial agreement with the feature rank obtained using SHAP, but more in-depth analysis is needed to better understand the differences.

Furthermore, considering each feature separately might not always be practical, for this reason, the *Multidimensional Classifier* proposed in Section 4.3 enables prediction according to a lower number of poverty dimensions defined by the domain expert. This approach is particularly useful when limited resources for supporting the elderly are available as it facilitates the identification of the main categories for intervention.

The framework here presented is specifically designed to predict various levels of poverty within the elderly population. However, its main advantage resides in its adaptability and potential for generalisation across different contexts. For example, it can be extended to encompass data concerning poverty among various vulnerable populations. In particular, even if different groups of subjects for age or geographic locations are characterised by different features, the same framework can be applied to new data, revealing the relevance of variables directly related to the considered scenario. The framework is not limited to poverty prediction within the elderly population and can be applied in diverse contexts. For instance, it could be employed for the analysis of digital health records to identify different levels of specific pathologies, such as mental decline.

Lastly, a dashboard has been developed as a prominent tool for data visualisation, accessible to the research community along with the collected dataset and codes developed for the analysis.

7. Conclusion and future work

The proposed AI approaches for predicting multidimensional levels of poverty have demonstrated promising performance. The feature relevance analysis enables the definition of poverty scores based solely on five poverty dimensions, derived from appropriately linearly combined features. This approach allows model training on feature vectors of a reduced dimensionality while retaining valuable information from the original feature set.

However, it is worth acknowledging both the potential biases in the final results introduced by the labelling procedure and the possibility of revising and extending the definition of the five dimensions. To further validate and generalise our proposal, we plan to leverage other labelled datasets available in the literature. This approach will allow us to assess both the labelling procedure and the effectiveness of the proposed Naive Bayes multidimensional classifier.

Conflict of interest

There is nothing to declare.

Acknowledgements

We want to give our thanks to Alberto Raggi and Alessia Marcassoli for their supporting work during the experimentation and Marco Terraneo for his supporting work during data analysis. Lastly, we would like to thank Giulia Rosemary Avis for her precious revision.

References

[1] A. Abu, R. Hamdan and N. Sani, Ensemble learning for multidimensional poverty classification, *Sains Malaysiana* **49**(2) (2020), 447–459.

[2] M. Adena and M. Myck, Poverty and transitions in key areas of quality of life, *Active ageing and solidarity between generations in Europe: First results from SHARE after the economic crisis*, (2013), pp. 55–73.

[3] M. Adena and M. Myck, Poverty and transitions in health in later life, *Social Science & Medicine* **116** (2014), 202–210.
[4] S. Alkire, J.E. Foster, S. Seth, M.E. Santos, J. Roche and P. Ballon. Multidimensional poverty measurement and analysis: chapter 3–overview of methods for multidimensional poverty assessment. 2015.
[5] A. Alsharkawi, M. Al-Fetyani, M. Dawas, H. Saadeh and M. Alyaman. Poverty classification using machine learning: The case of jordan, *Sustainability* **13**(3) (2021), 1412.
[6] D. Arribas-Bel, J.E. Patino and J.C. Duque, Remote sensingbased measurement of living environment deprivation: Improving classical approaches with machine learning, *PLoS one* **12**(5) (2017), e0176684.
[7] A.B. Atkinson, Income distribution in oecd countries, *Evidencefrom Luxemburg Income Study*, 1995.
[8] S.W. Bae, S. Yun, Y.S. Lee, J.-H. Yoon, J. Roh and J.-U. Won, Income changes due to disability ratings and participation in economic activities caused by industrial accidents: a population-based study of data from the fourth panel study of workers compensation insurance (pswci), *International Journal of Environmental Research and Public Health* **15**(11) (2018), 2478.
[9] W. Bank, *Poverty and shared prosperity 2020: Reversals of fortune*. The World Bank, 2020.
[10] M. Biewen, Income inequality in germany during the 1980s and 1990s, *Review of Income and Wealth* **46**(1) (2000), 1–19.
[11] J.E. Blumenstock, Fighting poverty with data, *Science* **353**(6301) (2016), 753–754.
[12] R. Boarini and M.M. d'Ercole, Measures of material deprivation in oecd countries. 2006.
[13] C. Browne, D.S. Matteson, L. McBride, L. Hu, Y. Liu, Y. Sun, J. Wen and C.B. Barrett, Multivariate random forest prediction of poverty and malnutrition prevalence, *PloS one* **16**(9) (2021), e0255519.
[14] L. Cappellari and S.P. Jenkins, Summarizing multiple deprivation indicators. Technical report, *ISER Working Paper Series*, 2006.
[15] T. Chen and C. Guestrin, Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, (2016), pp. 785–794.
[16] T. Chen, T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho, K. Chen, R. Mitchell, I. Cano, T. Zhou, et al., Xgboost: extreme gradient boosting, *R package version 0.4-2* **1**(4) (2015), 1–4.
[17] S. Choi, S. Lee, J. Matejkowski and Y.M. Baek, The relationships among depression, physical health conditions and healthcare expenditures for younger and older americans, *Journal of Mental Health* **23**(3) (2014), 140–145.
[18] H. Cramer, Mathematical methods of statistics, princeton univ. *Press, Princeton, NJ*, 27, 1946.
[19] A.L. Danielewicz, E. dOrsi and A.F. Boing, Contextual income and incidence of disability: results of epifloripa elderly cohort, *Revista de saude publica* (2019), 53.
[20] F. DAdda, M. Cremaschi, E. Messina, M. Terraneo, S. Bandini and F. Gasparini, A three level prediction of multidimensional poverty in elderly. In *Proceedings of the Italia Intelligenza Artificiale - Thematic Workshops co-located with the 3rd CINI National Lab AIIS Conference on Artificial Intelligence (Ital IA 2023)*, volume 3486, (2022), pp. 544–549. CEUR.
[21] M.F. Förster and 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.
[22] C.G. Gore, J.B. Figueiredo, et al., Social exclusion and antipoverty policy: A debate. *(No Title)*, 1997.
[23] S. Hu, Y. Ge, M. Liu, Z. Ren and X. Zhang, Village-level poverty identification using machine learning, high-resolution images, and geospatial data, *International Journal of Applied Earth Observation and Geoinformation* **107** (2022), 102694.
[24] M. Huisman, A.E. Kunst and J.P. Mackenbach, Socioeconomic inequalities in morbidity among the elderly; a european overview, *Social Science & Medicine* **57**(5) (2003), 861–873.
[25] R. Isnin@Hamdan, A.A. Bakar and N.S. Sani, Does artificial intelligence prevail in poverty measurement? volume 1529, pp. 042082. IOP Publishing, apr 2020.
[26] N. Jean, M. Burke, M. Xie, W.M. Davis, D.B. Lobell and S. Ermon. Combining satellite imagery and machine learning to predict poverty, *Science* **353**(6301) (2016), 790–794.
[27] H. Kim and J. Lee, The impact of comorbidity on wealth changes in later life, *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences* **61**(6) (2006), S307–S314.
[28] G. Li, Z. Cai, Y. Qian and F. Chen, Identifying urban poverty using high-resolution satellite imagery and machine learning approaches: Implications for housing inequality, *Land* **10**(6) (2021), 648.
[29] Q. Li, S. Yu, D. Échevin and M. Fan, Is poverty predictable with machine learning? a study of dhs data from kyrgyzstan, *Socio-Economic Planning Sciences* **81** (2022), 101195.
[30] P. Liberati, G. Resce and F. Tosi, The probability of multidimensional poverty: A new approach and an empirical application to eu-silc data, *Review of Income and Wealth*, 2022.
[31] S.M. Lundberg and S.-I. Lee, A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30* (2017), pp. 4765–4774. Curran Associates, Inc.
[32] B. Malmberg, E. Andersson and S. Subramanian, Links between ill health and regional economic performance: evidence from swedish longitudinal data, *Environment and planning A* **42**(5) (2010), 1210–1220.
[33] J.H. Mohamud and O.N. Gerek, Poverty level characterization via feature selection and machine learning. In *2019 27^{th} Signal Processing and Communications Applications Conference (SIU)*, (2019), pp. 1–4. IEEE.
[34] B. Nolan and C.T. Whelan, Measuring poverty using income and deprivation indicators: alternative approaches, *Journal of European Social Policy* **6**(3) (1996), 225–240.
[35] B. Perry, The mismatch between income measures and direct outcome measures of poverty, *Social Policy Journal of New Zealand* (2002), pp. 101–127.
[36] M. Pinilla-Roncancio, Disability and poverty: two related conditions. a review of the literature, *Revista de la Facultad de Medicina* **63** (2015), 113–123.
[37] I. Rish, et al., An empirical study of the naive bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, volume 3, (2001), pp. 41–46.
[38] G. Rodgers, C. Gore and J.B. Figueiredo, *Social exclusion: Rhetoric, reality, responses.* 1995.
[39] N.S. Sani, M.A. Rahman, A.A. Bakar, S. Sahran and H.M. Sarim, Machine learning approach for bottom 40 percent

households (b40) poverty classification, *Int. J. Adv. Sci. Eng. Inf. Technol* **8**(4-2) (2018), 1698.

[40] S.K. Satapathy, S. Saravanan, S. Mishra and S.N. Mohanty, A comparative analysis of multidimensional covid-19 poverty determinants: An observational machine learning approach, *New Generation Computing* **41**(1) (2023), 155–184.

[41] T.P. Sohnesen and N. Stender, Is random forest a superior methodology for predicting poverty? an empirical assessment, *Poverty & Public Policy* **9**(1) (2017), 118–133.

[42] M. Terraneo, A longitudinal study of deprivation in european countries, *International Journal of Sociology and Social Policy* 2016.

[43] P. Townsend, *Poverty in the United Kingdom: a survey of household resources and standards of living*. Univ of California Press, 1979.

[44] A. Usmanova, A. Aziz, D. Rakhmonov and W. Osamy, Utilities of artificial intelligence in poverty prediction: a review, *Sustainability* **14**(21) (2022), 14238.

[45] L. Van der Maaten and G. Hinton, Visualizing data using t-sne, *Journal of Machine Learning Research* **9**(11) 2008.

[46] D.R. Wijaya, N.L.P.S.P. Paramita, A. Uluwiyah, M. Rheza, A. Zahara and D.R. Puspita, Estimating city-level poverty rate based on e-commerce data with machine learning, *Electronic Commerce Research* (2022), pp. 1–27.

[47] J. Xu, J. Song, B. Li, D. Liu and X. Cao, Combining night time lights in prediction of poverty incidence at the county level, *Applied Geography* **135** (2021), 102552.

[48] M.H.A. Zamzuri, N. Sofian and R. Hassan, The forecasting of poverty using the ensemble learning classification methods, *International Journal on Perceptive and Cognitive Computing* **9**(1) (2023), 24–32.

[49] W. Zhang, T. Lei, Y. Gong, J. Zhang and Y. Wu, Using explainable artificial intelligence to identify key characteristics of deep poverty for each household, *Sustainability* **14**(16) (2022), 9872.