Health Information Science and Systems

RESEARCH

A divisive hierarchical clustering methodology for enhancing the ensemble prediction power in large scale population studies: the ATHLOS project

Petros Barmpas[1]*, Sotiris Tasoulis[1], Aristidis G. Vrahatis[1], Spiros V. Georgakopoulos[27], Panagiotis Anagnostou[1], Matthew Prina[2,3], José Luis Ayuso-Mateos[4,5,6], Jerome Bickenbach[7,8], Ivet Bayes[4,13], Martin Bobak[9], Francisco Félix Caballero[10,11], Somnath Chatterji[12], Laia Egea-Cortés[13], Esther García-Esquinas[10,11], Matilde Leonardi[14], Seppo Koskinen[15], Ilona Koupil[16,17], Andrzej Pająk[18], Martin Prince[3,19], Warren Sanderson[20,21], Sergei Scherbov[20,22,23], Abdonas Tamosiunas[24], Aleksander Galas[25], Josep Maria Haro[4,13], Albert Sanchez-Niubo[4,13], Vassilis P. Plagianakos[1] and Demosthenes Panagiotakos[26]

Abstract

The ATHLOS cohort is composed of several harmonized datasets of international groups related to health and aging. As a result, the Healthy Aging index has been constructed based on a selection of variables from 16 individual studies. In this paper, we consider additional variables found in ATHLOS and investigate their utilization for predicting the Healthy Aging index. For this purpose, motivated by the volume and diversity of the dataset, we focus our attention upon data clustering, where unsupervised learning is utilized to enhance prediction power. Thus we show the predictive utility of exploiting hidden data structures. In addition, we demonstrate that imposed computation bottlenecks can be surpassed when using appropriate hierarchical clustering, within a clustering for ensemble classification scheme, while retaining prediction benefits. We propose a complete methodology that is evaluated against baseline methods and the original concept. The results are very encouraging suggesting further developments in this direction along with applications in tasks with similar characteristics. A straightforward open source implementation for the R project is also provided (https://github.com/Petros-Barmpas/HCEP).

Keywords: Clustering, Prediction enhancement, ATHLOS cohort, Ensemble methods

Introduction

Health Informatics has received much attention in the past years, since it permits Big Data collection and analysis, as well as the extraction of patterns that are free of the strict methodological assumptions of statistical modeling [1, 2]. Recent advances in the biomedical domain that contribute in the early and accurate disease detection, patient care and community services, generate data at an increasing rate. These complex datasets belong to the Big Data field, containing various variable types with different scales or experimental setups, in many cases incomplete [3]. The large data volume on each biomedical research field offers the opportunity to open new avenues for exploring the various biomedical phenomena. Machine Learning (ML) methods are considered as the first choice for the analysis of this data, as they can tackle their volume and complexity. In recent years, both unsupervised and supervised ML methods have been applied to biomedical challenges with reliable results.

A large category on this perspective is the population studies for aging and health analysis, where they offer a

*Correspondence: petrosbarmpas@uth.gr
[1] Department of Computer Science and Biomedical Informatics, University of Thessaly, Lamia, Greece
Full list of author information is available at the end of the article

plurality of large scale data with high diversity and complexity. Aging and health indicators are an important part of such research as the population aging observed in most developed countries leads to an increasing interest in studying health and aging, since the elderly are nowadays the fastest-growing segment in large regions, such as Europe, Asia and the USA [4–8]. As such, discovering health-related factors in an attempt to understand how to maintain a healthy life is of crucial importance. Meanwhile, it has long been reported that Sociodemographic factors are significant determinants of various health outcomes, such as healthy aging [9, 10], while evidently aging involves interactions between biological and molecular mechanisms with the environment, and as a result, it is a multifactorial phenomenon that everyone experiences differently [11].

The EU-funded ATHLOS (Ageing Trajectories of Health: Longitudinal Opportunities and Synergies (EU HORIZON2020-PHC-635316, http://athlosproject.eu/)) Project produces a large scale dataset in an attempt to achieve a better understanding of aging. The produced harmonized dataset includes European and international longitudinal studies of aging to identify health trajectories and determinants in aging populations. Under the context of ATHLOS, a metric of health (the Healthy Aging index) has been created using an Item Response Theory (IRT) approach [12] delivering a common metric of health across several longitudinal studies considered in ATHLOS. Interestingly, there is a plethora of available variables within the harmonized dataset that have not been considered when generating the aforementioned metric of health, encouraging the further exploration of associated factors through the utilization of Pattern Recognition and ML approaches. Nevertheless, the imposed data volume and complexity generate challenges for ML related to Big Data management and analytics.

There exist many recently published studies based on the ATHLOS dataset with promising results in several scientific fields, such as cardiovascular disease evaluation [13–17], demographic studies about sociodemographic indicators of Healthy Aging index [18] and the impact of socioeconomic status [19–21], nutrition science studies such as nutrition effects on health [22–24], alcohol drinking patterns effects on health [25, 26] and even psychology studies assessing the impact of depression and other psychological disorders related to aging and health [27–29]. Nevertheless, the ATHLOS data specifications require analysis through state of the art ML methods to uncover hidden patterns, tackle the data complexity and help to better interpret the characteristics that affect the state of human health. Predicting the Healthy Aging index can be considered one of the greatest challenges of ATHLOS projects in the health informatics domain. Previously, members of the ATHLOS consortium published studies [30, 31] by applying various supervised ML algorithms on part of ATHLOS data (ATTICA and ELSA study respectively). While these studies have shown remarkable results, a study of the Healthy Aging index prediction in the unified and harmonized ATHLOS data utilizing all additional information has not yet been done.

In this study, we proposed a hybrid framework that includes the integration of unsupervised and supervised ML algorithms to enhance prediction performance on large-scale complex data. More precisely, we developed a divisive hierarchical clustering framework for ensemble learning to enhance the prediction power on ATHLOS large-scale data regarding its Healthy Aging index. We focus our attention upon the clustering for the prediction scheme. Evidently, unsupervised learning enhances the prediction power, exploiting the structure of the data. We show that imposed computation bottlenecks can be surpassed, when using appropriate hierarchical clustering within a clustering for ensemble classification scheme, while retaining prediction benefits. We propose a complete methodology that is evaluated against other baseline methods. The results are very encouraging, suggesting further developments in this direction along with applications in tasks with similar characteristics.

Related work

In the last decade, several studies have been published regarding the integration of unsupervised and supervised learning strategies, most of which concern the incorporation of clustering models to classification algorithms for the improvement of the prediction performance. Although there has been a remarkable progress in this area, there is a need for more robust and reliable frameworks under this perspective given the ever-increasing data generation in various domains. Clustering can be considered as a preprocessing step in a classification task, since in complex data with non-separable classes the direct application of a classifier can be ineffective. In [32] the authors provided evidence that the training step in separated data clusters can enhance the predictability of a given classifier. In their approach the k-means and a hierarchical clustering algorithm were utilized to separate the data, while neural networks were applied for the classification process.

The utilization of clustering in an attempt to gain more information regarding the data and subsequently reducing errors in various prediction tasks has been previously explored. The clustering outcome can be considered as a compressed representation of the dataset and has the potential to exploit information about its structure. In [33], the authors examine the extent to which analysis of clustered samples can match predictions made by analyzing the entire dataset at once. For this purpose, they compare prediction results using regression analysis on original and clustered datasets. It turned out that clustering improved regression prediction accuracy for all examined tasks. Additionally, the authors in [34] also investigated whether clustering can improve prediction accuracy by providing the appropriate explanations. They proposed the coordination of multiple predictors through a unified ensemble scheme. Furthermore, in [35], the authors integrated the semi-supervised fuzzy c-means (SSFCM) algorithm into the Support Vector Machine (SVM) classifier offering promising results regarding the improvement of SVM's prediction power. Their hypothesis lies on the fact that unlabeled data include an inner structure, which can be efficiently uncovered by data clustering tools. This has been shown to be a crucial preprocessing step to enhance the training phase of a given classifier.

Following a similar perspective, the SuperRLSC algorithm utilizes a supervised clustering method to improve classification performance of the Laplacian Regularized Least Squares Classification (LapRLSC) algorithm [36]. Their motivation is based on the intuition that the clustering process contributes to the identification of the actual data structure by constructing graphs that can reflect more refined structures. A step further is to incorporate ensemble clustering before the classification stage, since an ensemble approach can elucidate the data structure in a more realistic manner [37]. The authors applied this framework to identify breast cancer profiles and provide reliable results, since ensemble clustering algorithms capable to deal with the biological diversity are essential for clinical experts. Other approaches such as the work in [38] utilize a clustering process to reduce the number of instances used by the imputation on incomplete datasets. The unsupervised learning part in this method offered better results not only in the classification accuracy, but also in terms of computational execution time. Given that the most population-based studies include a plethora of missing values, this framework has a great potential to export reliable results in cases.

Although several hybrid approaches, including supervised and unsupervised ML techniques, have been recently proposed, the rise of Big Data challenges along with the diversity issues on population studies, necessitates further research in this direction. Since the ATHLOS dataset has a strong inherent complexity and diversity with a plethora of missing values, developments in clustering-assisted classification should be beneficial. Next, we will show that an incorporated clustering methodology enhances the prediction power of various classification and regression tools.

Background material

Data description

The ATHLOS harmonized dataset [39] includes European and international longitudinal studies of aging. It contains more than 355,000 individuals who participated in 17 general population longitudinal studies in 38 countries. We specifically used 15 of these studies, which are the 10/66 Dementia Research Group Population-Based Cohort Study [40], the Australian Longitudinal Study of Aging (ALSA) [41], the Collaborative Research on Ageing in Europe (COURAGE) [42], the ELSA [43], the Study on Cardiovascular Health, Nutrition and Frailty in Older Adults in Spain (ENRICA) [44], the Health, Alcohol and Psychosocial factors in Eastern Europe Study (HAPIEE) [45], the Health 2000/2011 Survey [46], the HRS [47], the JSTAR [48], the KLOSA [49], the MHAS ([50]), the SAGE [51], the SHARE [52], the Irish Longitudinal Study of Ageing (TILDA) [53] and the Longitudinal Aging Study in India (LASI) [54].

Ensemble learning

Ensemble methods have seen rapid growth in the past decade within the ML community [55]. An ensemble is a group of predictors, each of which gives an estimate of a response variable. Ensemble learning is a way to combine these predictions with the goal that the generalization error of the combination is smaller than each of the individual predictors. The success of ensembles lies in the ability to exploit the predictors' diversity. Thus, if a predictor exhibits enhanced generalization performance regarding a class or subspace of the dataset, then the aggregated strength of all the predictors can form a global, more reliable one.

A significant portion of research outcomes in ensemble learning aims towards finding methods that encourage diversity in the predictors. Mainly, there are three reasons for which ensembles perform better than the individual predictors [56]. The first reason is statistical. A learning algorithm can be considered a way to search the hypotheses space to identify the best one. The statistical problem is caused due to insufficient data. Thus, the learning algorithm would give a set of different hypotheses with similar accuracy on the training data. With ensembling, the risk of choosing the wrong hypothesis would be averaged out to an extend. The second reason is of computational nature. Often, while looking for the best hypothesis, the algorithm might be stuck in a local optimum, thus giving the inferior results. By considering multiple such hypotheses, we can obtain a much better approximation of the true function. The third reason is representational. Sometimes the underlining function might not be any hypothesis in the hypotheses space. With the ensemble method, the representational space might be expanded to give a better approximation of the unknown function.

Ensemble learning also coincides with the task of clustering, since the performance of most clustering techniques is highly data-dependent. Generally, there is no clustering algorithm, or algorithm with distinct parameter settings, which performs well for every set of data [57]. To overcome the difficulty of identifying a proper alternative, the methodology of cluster ensemble has been continuously developed in the past decade.

Projection based hierarchical divisive clustering

Hierarchical clustering algorithms create clusters either in a divisive (top-down) or an agglomerative (bottom-up) approach. In the former, the whole dataset is considered a single cluster in the first step. According to predefined split criteria, a succession of splits produces subsets down to the final clustering scheme. On the other hand, the agglomerative approach starts by considering each data point as its own unique pre-cluster. In contrast, according to some similarity measures, successive joints of pre-clusters produce hyper-sets of points that finally lead up to the clustering structure. Despite the promising performance of hierarchical clustering algorithms in uncovering the data structure [58, 59], the computational overhead opposed impedes their usage in Big Data scenarios. However, more recent advancements in both agglomerative [60, 61] and divisive strategies [62, 63] have exposed their broad applicability and robustness. In particular, it has been shown that, when divisive clustering is combined with integrated dimensionality reduction [59, 64, 65], we can still get methods capable of indexing extensive data collections. In contrast to agglomerative methodologies, such indexes allow fast new sample allocation to clusters.

In more detail, several projection-based hierarchical divisive algorithms try to identify hyper-planes that best separate the clusters. This can be achieved with various strategies, more notably by calculating the probability distribution of the projected space and avoid separating regions with high-density [66–68]. The latter presents computational challenges in the calculation of the density of each neighborhood. Motivated by the work of [69], instead of finding the regions with high density, the authors in [59, 64] try to identify regions with low density to create the separating hyper-planes.

The dePDDP [64] algorithm builds upon the Principal Direction Divisive Partitioning (PDDP) [70], a divisive hierarchical clustering algorithm defined by the compilation of three criteria, for cluster splitting, cluster selection, and algorithm termination, respectively. These algorithms incorporate information from the projections p_i: $p_i = u_1(d_i - b), \quad i = 1, 2, \ldots, n$ onto the first principal component u_1 to produce the two subsequent partitions at each step. In more detail, dePDDP splits the selected partition $\mathcal{P}^{\updownarrow}$ by calculating the kernel density estimation $\hat{f}'\left(x; h'\right)$ of the projections p_i^l and the corresponding global minimizer x^* defined as the best local minimum of the kernel density estimation function. Then constructs $P_1^l = \left\{d_i \in \mathcal{D} : p_i^l \leq x^*\right\}$ and $P_2 = \left\{d_i \in \mathcal{D} : p_i^l > x^*\right\}$. Now, let $\mathcal{P}$ a partition of the dataset D into k sets. Let F be the set of the density estimates $f_i = \hat{f}\left(x_i^*; h\right)$ of the minimizers X_i^* for each $C_i \in \mathcal{P}$. The next set to split is C_j, with $j = \arg\max_i \left\{f_i : f_i \in \mathcal{F}\right\}$. Finally, the algorithm allows the automatic determination of clusters by terminating the splitting procedure as long as there are no minimizer for any of the clusters $C_i \in \mathcal{P}$.

By using techniques like the fast Gauss transform, linear running time for the kernel density estimation is achieved, especially for the one-dimensional case. To find the minimizer, only the density at n positions (between the projected data points) needed to be evaluated, since those are the only places with valid splitting points. Thus, the total complexity of the algorithm remains $\mathrm{O}(k_{\max}(2 + k_{SVD})(s_{nz}na))$.

The Minimum Density Hyper-planes (MDH) algorithm [59] follows a similar clustering procedure. However, instead of using the First Principal Component for the calculation of the splitting hyper-plane that minimizes the density, MDH follows a projection pursuit formulation of the associated optimization problem to find minimum density hyper-planes. Projection pursuit methods optimize a measure of interest of a linear projection of a data sample, known as the projection index, in this case the minimum value of the projected density. Although this is a theoretically justified approach, it is more computationally intensive, mainly due to the optimization procedure. Thus, when the clustering efficiency is not of crucial importance (data indexing) or the computational resources are limited or real-time performance is required, the dePDDP approach can be considered as a satisfactory approximation of MDH.

The proposed ensemble methodology

The concept proposed in [34] showed that an ensemble learning predictor based on different clustering outcomes can improve the prediction accuracy of regression techniques. The performance gains are associated with the change in locality features when training prediction models for individual clusters, rather than the whole dataset. Different clustering outputs $\mathcal{P}$ are retrieved by providing various k values to the k-means clustering algorithm, increasing the diversity of the outcomes. For $k = 1, 2, \ldots, L$, we retrieve L $\mathcal{P}_k$ individual partitionings. Then, for each cluster

$C_k^i \in \mathcal{P}_k$, with $i = 1, 2, \dots, k$ and $k = 1, 2, \dots, L$, a model is trained. The final predictions for each data point are calculated by averaging among the predicted values retrieved by the models that correspond to the clusters C_k^i that falls within. Selecting a cutoff L for k (how many individual partitionings $\mathcal{P}_k$ should be calculated) is not straightforward, but data dependent heuristics can be estimated.

There is a crucial trade-off, however, for this methodological framework with respect to the computational complexity, imposed by the number of predictors that need to be trained. Even though each model is trained upon a subset of the original dataset, we still need to train $\frac{L\times(L+1)}{2}$ predictors. As a result, the computational complexity increases exponentially. Large scale prediction tasks similar to the one studied here can prohibit the extensive utilization of this concept, in particular when combined with computationally demanding predictors, such as Neural Networks and Support Vector Machines.

In this work, motivated by recent advantages in projection-based divisive hierarchical algorithms, we proposed an ensemble algorithmic scheme able to surpass the aforementioned computational burden, while retaining prediction benefits. The key idea is to generate the L partitionings by iteratively expanding a binary tree structure. Divisive clustering algorithms allow us to stop the clustering procedure, when the predefined number of clusters k has been retrieved. Then, to retrieve the partitioning for $k = k + 1$, we only need to split one of the leaf nodes. In practice, L partitionings can be retrieved by a single execution of the algorithm, where k is set to the threshold value L. By monitoring the order of binary splits, we retrieve $\mathcal{P}$ constituted by the individual partitionings that correspond to the $k = 1 \dots L$ values.

Arguably, we sacrifice some of the diversity between the individual partitionings $\mathcal{P}_k$, since each two consecutive partitionings only differ with respect to the portion of the dataset that constitutes the selected for splitting leaf node, but simultaneously benefit greatly by only having to train $2L + 1$ models. Again, to provide the final prediction for each data point, we need to average the predicted values retrieved by the models that correspond to the clusters C_k^i. This means that we need to combine information retrieved by the nodes (clusters) appearing along the path each sample followed from the root node (containing the full dataset) the the leaf node that lies within. Note that this divisive structure not only allow us to interpret the ensemble procedure, but is also straightforward to efficiently assign new observations to the tree structure providing the corresponding predictions for new arriving samples.

Algorithm 1: Clustering for Ensemble Prediction Framework

```
Result: Hierarchical Clustering for Ensemble Prediction (HCEP)
Given the full dataset $\mathcal{D}$ and the maximum number of clusters $L$;
Cluster($\mathcal{D}$,$L$); Extract the $L$ partitionings $\mathcal{P}_L$;
Given $Trainset$ and $Testset$;
for $k = 1, 2, \dots, L$ do
    foreach $cluster\ i\ in\ \mathcal{P}_k$ do
        Let $tr_i \subset Trainset$ be the collection of training samples $\in i$;
        Train a Prediction Model $PM_i^k$ using $tr_i$;
    end
end
foreach $Sample\ n\ in\ Testset$ do
    Find $i, k$ for which $n \in C_i^k$;
    Predict the response variable $\hat{y}$ based on $PM_i^k$;
    $count$ = Number of $C_i^k$ clusters;
end
Average the $count$ $\hat{y}$ predictions;
```

In Algorithm 1, we present the complete proposed algorithmic procedure entitled Hierarchical Clustering for Ensemble Prediction (HCEP). In summary, the first step is to execute the projection based divisive clustering algorithm of choice and retrieve the complete resulting binary clustering tree. Keep in mind that the response variable is not taken into account for this step, as such, this is an unsupervised procedure. Then for each node of the tree, we train the selected prediction algorithm based only on samples belonging to the particular training set. For every sample belonging to the test set, we can now provide final predictions by averaging across the individual predictions of this particular sample retrieved by the corresponding nodes of the tree that lies within. For each new arriving sample, we initially pass it through the tree structure until reaching the appropriate leaf node. This is done by projecting the new sample onto the one dimensional vector retrieved for each node of the tree and deciding whether it should be assigned at the right or the left child. Then the prediction mechanism is applied as before.

Naive clustering for prediction

We are also interested in investigating the effectiveness of clustering in prediction, when used as a single preprocessing step [33]. We expect that the characteristics of the ATHLOS dataset employed in this work, such as its large scale and the imposed complexity by the appearance of both continuous and categorical variables, present a unique opportunity to expose the benefits, if any, in training individual models for sub-populations of samples belonging to the same cluster.

In practice, this procedure can be achieved utilizing any clustering algorithm. Here, we employ both the k-means and projection based divisive clustering as representatives of partitioning and hierarchical clustering, respectively. The algorithmic procedure is presented in Algorithm 2. The clustering takes place initially for a given number of clusters, which is subject to further investigation. Then, a prediction model is trained for each cluster utilizing the respective train samples, while for each sample in the train set, the final prediction is provided by the model that corresponds to the cluster it lies within. The new arriving sample is initially allocated to a cluster and then a similar procedure is followed to provide predictions. Notice that this procedure should be significantly more computationally efficient than the ensemble methodology, since we only need to train L models. In addition, for particular prediction algorithms with close to exponential complexity with respect to the number of samples, we also expect a significant computational boost against their application on the full dataset $\mathcal{D}$.

```
Algorithm 2: Clustering for Prediction Framework
  Result: Naive Clustering for Prediction
  Given the full dataset $\mathcal{D}$ and the number of clusters $L$;
  Cluster($\mathcal{D}$,$L$); Retrieve $\mathcal{P}_L$;;
  Given $Trainset$ and $Testset$;
  foreach $i$ cluster in $\mathcal{P}_L$ do
      Let $tr_i \subset Trainset$ be the collection of training samples $\in i$;
      Train a Prediction Model $PM_i$ using $tr_i$;
  end
  foreach Sample $n$ in $Testset$ do
      Find $i$ for which $n \in C_i$;
      Predict the response variable $\hat{y}$ based on $PM_i$;
  end
```

Data preprocessing

The aforementioned studies' dataset described in Sect. 3.1 consist of 990,000 samples from more than 355,000 individuals, who participated in 17 general population longitudinal studies in 38 countries. The dataset contains 184 variables; two response and 182 independent variables. Response variables are the raw and the scaled Healthy Aging index of each patient. Regarding the independent variables (see supplementary material sheet S1), nine variables were removed including various indexes (sheet S2), 13 variables were removed including obviously depended variables that cannot be taken into account (sheet S3), and six variables were removed including information that cannot be considered within the prediction scheme (sheet S4). Furthermore, the 47 variables (sheet S5), which used to originally calculate the Healthy Aging index [12] are also excluded. Not only these features create a statistical bias regarding the Healthy Aging index, which is the response variable in our analysis, but in this work we aim to uncover new insights for external variables that have previously been considered not significantly relevant. Removing any samples for which the Healthy Aging index is not available, the resulting data matrix is constituted by 770,764 samples and 107 variables.

To this end, we have to deal with the critical step of missing value imputation. For this purpose, we utilized the Vtreat [71] methodology, a cutting-edge imputation tool with reliable results. Vtreat is characterized by a unique strategy for the creation of the dummy variables, which resulted in the construction of 458 dummy variables. Next, a significance pruning process step took place, where each variable was evaluated based on its correlation with the Healthy Aging index (response variable).

Experimental analysis

In the first part of our experimental analysis, we compare the proposed ensemble scheme based on Projection Based Hierarchical Clustering (HCEP) against the original one (based on the k-means partitioning clustering). We also examine if there are any benefits when compared against the naive clustering for prediction scheme presented in Sect. 4.1, utilizing both aforementioned clustering approaches. For this set of experiments, the divisive algorithm of choice is dePDDP, while the maximum number of clusters L is set to 40, which is a value greater than the average optimal number of clusters retrieved by dePDDP to effectively examine the behavior of the proposed methodology. For every run of k-means and dePDDP, the number of clusters k is given as input. k-means is allowed to choose the most appropriate convergence among 10 random initializations [72, 73], while for the dePDDP algorithm, the bandwidth multiplier parameter is set to 0.05, a relative small value to guarantee enough binary splits that will lead to the required number of leafs (clusters). Finally, to avoid highly unbalanced tree structures, we set a threshold so that clusters with less than N/k points are not allowed to be split [74], where N is the total number of points in the dataset. All methodologies are implemented using the *R-project* open source environment for statistical computing, while specifically for dePDDP we utilize a native efficient implementation and for k-means we employed the implementation provided by the "biganalytics" package, called BigKmeans [75], which benefit from the lack of memory overhead by not duplicating the data. The choice for the employed clustering algorithms is based not only on their satisfying performance, but also on their simplicity and the structural ability to create an index that can be used to allocate future observations. For dePDDP algorithm, new instances are pushed into the tree until it reaches the respective leaf node. For the k-means algorithm, we allocate every instance of the testing set to the closest cluster by calculating the minimum distance to the cluster centroids.

For the prediction task, we employ the traditional Linear Regression (LR) and Random Forests (RF) algorithms. Again, the default parameter values are those provided by the corresponding implementations found in [76]. For the RF, we used 50 trees to guarantee its low computational complexity, due to hardware imposed restrictions, and the M_{try} variable was defined as $p/3$, where p are the number of variables. The regression performance is evaluated with respect to the Root Mean Square Error and the R-squared (RSQ). The mean squared error (MSE) is a measure of an estimator's quality, with values closer to zero indicating better performance. The MSE is the second moment of the error. Thus, it incorporates both the variance of the estimator (how widely spread the estimates are from one data sample to another) and its bias (how far off the average estimated value is from the truth). MSE has the same units of measurement as the square of the quantity being estimated. In an analogy to standard deviation, taking the square root of MSE yields the root-mean-square error RMSE [77]. R-squared (R^2) is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model.

$$\mathrm{R}^2 = 1 - \frac{Unexplained Variation}{Total Variation} \tag{1}$$

R-squared indicates to what extent the variance of one variable explains the variance of the second variable. So, if the R^2 of a model is 0.50, then approximately half of the observed variation can be explained by the model's inputs.

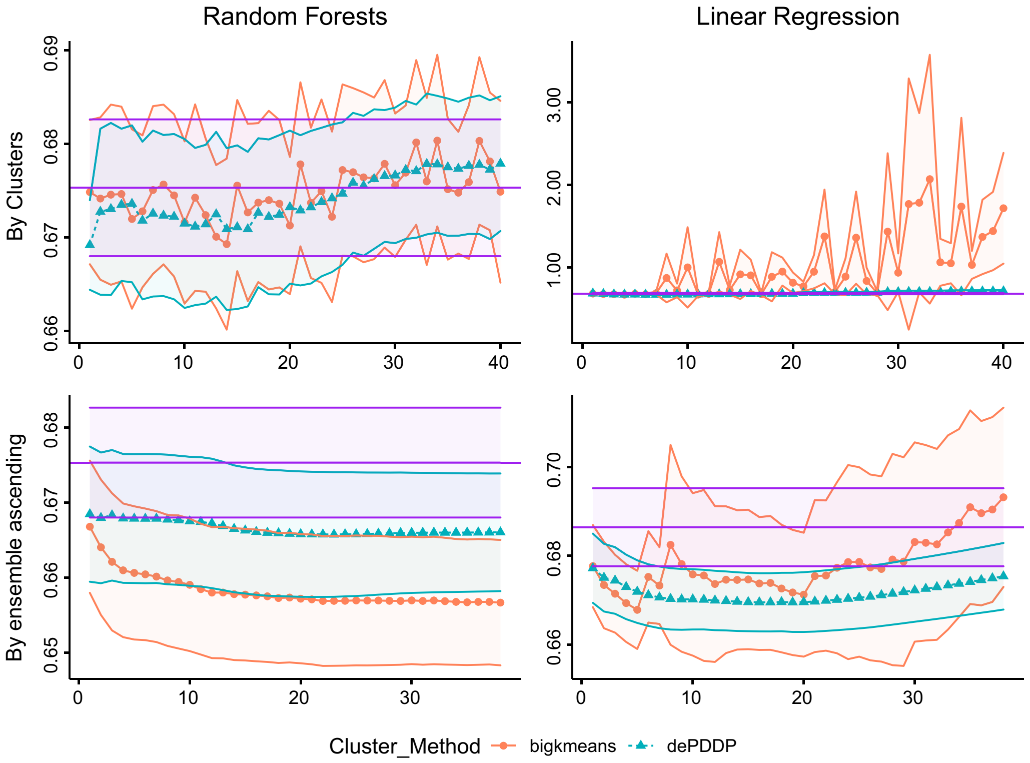

Fig. 1 RMSE metric results with respect to the Healthy Aging index prediction for Random Forests (first column) and Linear Regression (second column). Circular points with continuous red lines represent the results (vertical axes) when bigKmeans algorithm is utilized, while triangular points with blue dashed lines represent the results for the dePDDP algorithm, respectively. Each row of the plots depicts the clustering for different prediction strategies; naive methodology (first row) and Ensemble on ascending range of clusters (second row). The horizontal purple shaded area represents the corresponding values for the catholic models (training a single predictor in the entire dataset). Mean values are reported according to the utilized bootstrapping, while colored ribbons present the standard deviation between the experiments

The results with respect to the RMSE metric regarding the prediction of Healthy Aging index are reported in Fig. 1. To achieve robust validation of the results, while maintaining reasonable execution times, we utilize a bootstrapping technique by randomly sampling (with replacement) 50,000 samples for training and 1000 samples for testing [78]. The procedure is repeated 10 times, with different subsets for training and testing, respectively. Then, we estimate the performance of each model by computing the average score and the corresponding standard deviation. These are reported using line plots for mean values and shaded ares for standard deviation respectively. The top row of figures corresponds to the naive methodology, while the bottom row corresponds to the ensemble approaches, respectively. For both cases we report the performance of the catholic models indicated by the straight purple shaded area, parallel to axes X. Orange and green shaded areas indicate the performance of the *k*-means and the dePDDP algorithms, when combined with either Random Forests (left column) or Linear regression (right column), respectively. Notice that performance is reported with respect to the number of clusters (X axes). For the naive methodology, each number of clusters L corresponds to the RMSE value retrieved for this particular value of L, while for the ensemble models for each L value we observe the RMSE resulting by aggregating predictions for $k = 1, 2, \ldots, L,$.

In Fig. 1, we observe a performance boost compared to the catholic regression models that is more evident and robust for the ensemble methodologies (Algorithm 1). For up to 20 clusters, the naive models also appear to improve prediction performance, at least when utilizing RF, but when *k*-means is selected there is no consistency. For the ensemble models best performance is achieved by *k*-means combined to RF.

Finally, for the Ensemble Linear Regression we observe that the utilization of dePDDP leads to a more robust behavior where the prediction performance is not significantly affected for minor variations in the number of clusters.

This is most likely due to over-fitting, since for a high enough number of retrieved clusters, we expect to end up

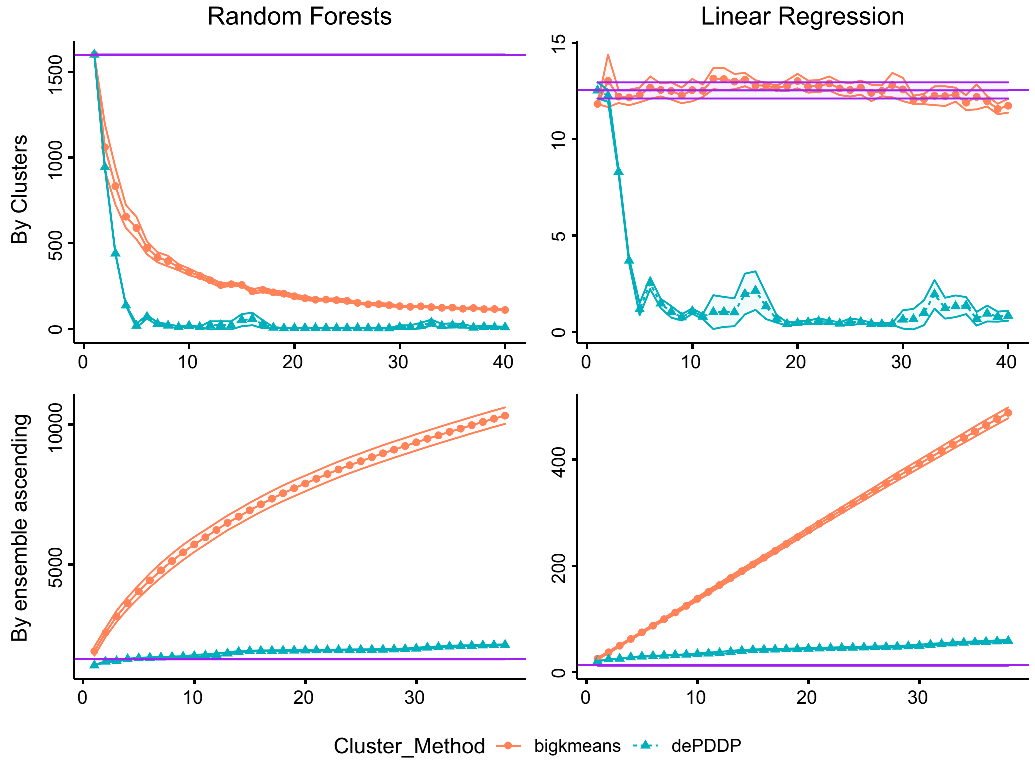

Fig. 2 Computational cost in seconds utilizing Random Forest (first column) and Linear Regression (second column) prediction models, respectively. Circular points with continuous red lines represent the results (vertical axes) when bigKmeans algorithm is utilized, while triangular points with blue dashed lines represent the results for the dePDDP, respectively. Each row of the plots depicts the clustering for different prediction strategies. Naive methodology (first row) and Ensemble on ascending range of clusters (second row). The horizontal purple shaded area represents the corresponding values for the catholic models (training a single predictor in the entire dataset). Mean values are reported according to the utilized bootstrapping, while colored ribbons present the standard deviation between the experiments

Table 1 Mean RMSE and R^2 for different regression models

	RMSE (std)	R^2(std)
LR	0.6851586(0.01815245)	0.5420653(0.02738484)
RF	0.6753074(0.01462793)	0.5551348(0.02472105)
XGboost	0.6937884(0.0138055)	0.5494156(0.02426911)
KNN	0.7858604(0.01970342)	0.4205703(0.02504982)
DNN1	0.6891531(0.01597495)	0.5364436(0.03170498)
DNN2	0.6923526(0.016212242)	0.5105284(0.03162947)
ENS-LR-dePDDP	0.6774225(0.04522381)	0.5505334(0.06650373)
ENS-LR-Kmeans	0.6783292(0.02592916)	0.5516001(0.02380914)
ENS-RF-dePDDP	0.6671423(0.01575666)	0.5659424(0.02244701)
ENS-RF-Kmeans	**0.6583103**(0.01672183)	**0.577264**(0.02440015)

The models presented are: Linear Regression (LR), Random Forrests (RF), XGboost, Deep Neural Network with 1 (DNN1) and 2 (DNN2) hidden layers, Hierarchical Ensemble method (HCEP) using Linear Regression or RF based on dePDDP (ENS-LR-dePDDP and ENS-RF-dPDDP, respectively) and ensemble method using LR or RF based on *k*-means (ENS-LR-Kmeans and ENS-RF-Kmeans, respectively). In parentheses are the Standard Deviation of the metrics across their 10 individual executions and the best performing method for every metric is denoted in bold font

with clusters characterized by low sample size compared to the number of variables.

Having concluded that the HCEP framework is able to enhance prediction performance compared to the catholic models and the naive approach, its computational overhead is studied, as well. Figure 2 is devoted to the computational time comparisons. As expected, the naive approach can reduce computational time, at least for complex method such as the RF that are greatly affected by samples size. More importantly, the computational complexity comparison between the ensemble approaches indicates that the utilization of the proposed method is justified. It is evident that consistent prediction power benefits can be achieved with minimal computational overhead. Notice here that the aforementioned computational times for RF have been achieved by implementing a parallel execution strategy accommodated by the "foreach" R package [79]. Experiments took place on a PC with Intel i9 7920x processor and 32 GB of RAM running the Ubuntu Linux operating system.

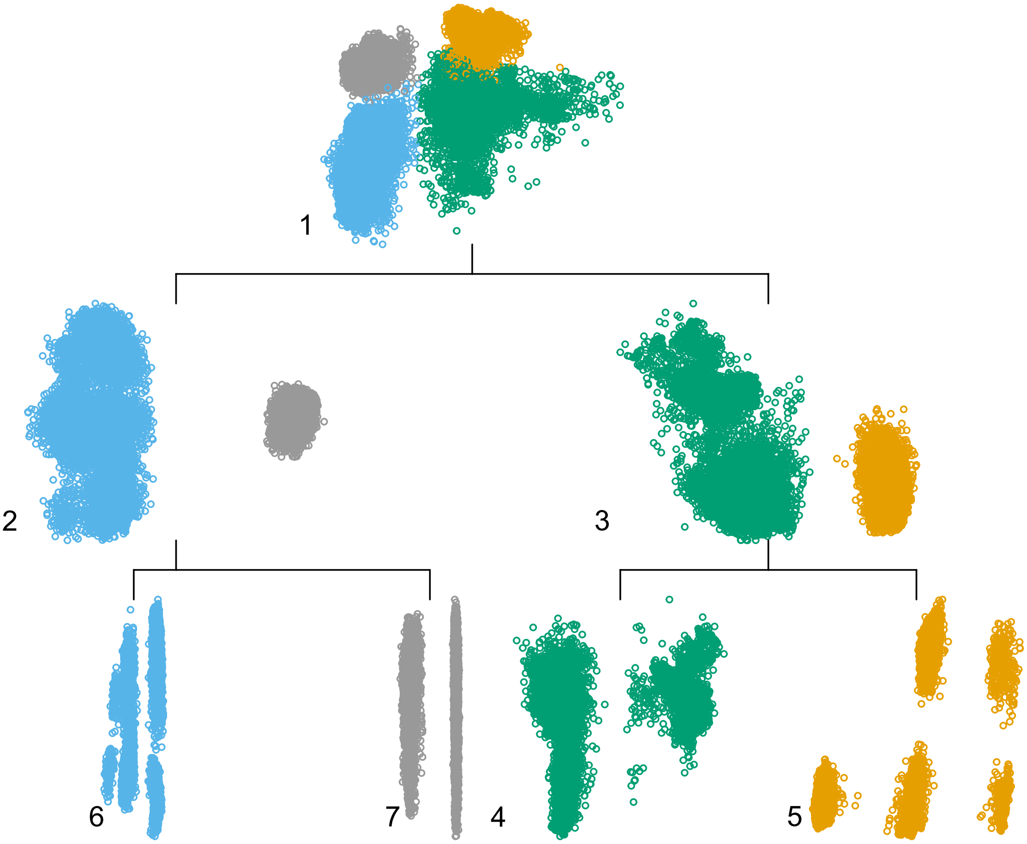

Fig. 3 Tree structure example of the hierarchical MDH experiment. In this instance, each step of the algorithm is presented in a top-down order, with every level indicating the corresponding cut for a total of 4 clusters. The data points are colored according to the final clustering and each subset is indicated with numbering from 1 (original dataset) to 6 and 7 (last split producing two of the final clusters)

Extended comparisons

The performance of the proposed HCEP methodology is further evaluated by comparing it against state-of-the-art regression models in predicting the Healthy Aging index using the same bootstrapping technique. In detail, six regression models have been applied, namely the Linear Regression (LR) model, the Random Forests (RF), the k nearest neighbors (kNN), the XGboost [80], and two Deep Neural Network architectures (DNN_1 and DNN_2).

Briefly, in kNN regression, the average of the Healthy Aging index values of the five Nearest Neighbors of a given test point is calculated. The RF regression performs the RF process by calculating the average output of all trees in the final prediction for each test sample. We applied 100 trees and the M_{try} variable was defined as $\sqrt{p}$, where p are the number of variables. Extreme Gradient Boosting (XGBoost) is a cutting-edge classifier, based on an ensemble of classification and regression trees [80]. Given the output of a tree $f(x) = wq(x_i)$, where x is the input vector and wq is the score of the corresponding leaf q, the output of an ensemble of K trees will be: $y_i = \sum_{k=1}^{K} f_k(x_i)$. The first DNN ($DNN_1$) is constructed with two hidden layers of 100 neurons and one output layer of one neuron, while the second (DNN_2) with one hidden layer with 100 neurons. The ReLU activation function is utilized in hidden layers to control the gradient vanishing problem. The Backpropagation (BP) training algorithm is applied with the learning rate set to 0.001. We selected these two DNN architectures to deal with both the over- and under-fitting challenges of ATHLOS dataset after an extensive parameter optimization search was conducted utilizing the Adam algorithm [81].

The results are summarized in Table 1. For both ensemble methods we chose to present the values when the maximum number of clusters is set to $L = 30$, which is the average estimated value provided by dePDDP algorithm, when utilized for cluster number determination with its default parameters. Notice that, computational limitations do not allow the extensive use of traditional approaches for this purpose [82, 83], while minor

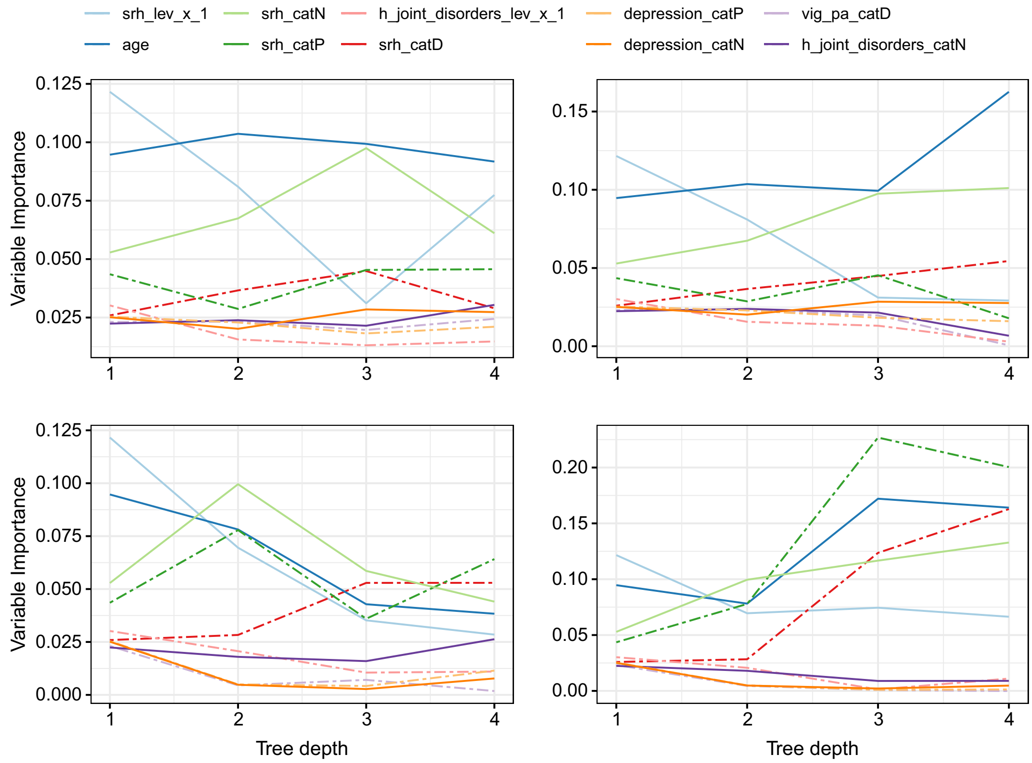

Fig. 4 Variable importance propagation of the prediction model's ten most influencing variables across each node in the different paths of the tree structure

variations to this number do not significantly alter the comparison outcome. As shown, the ensemble methodologies outperform any other confirming the prediction enhancement assumption. More precisely, the k-means based method combined with RF achieve the best score with respect to both metrics. However, the proposed HCEP performs better when LR is utilized for prediction. In general, HCEP comes second to the original k-means based scheme, something to be expected due to the loss of clustering diversity, as previously discussed in Sect. 4. However, the added value of HCEP arises when considering the minimal computational overhead.

Tree visualization and variable importance

In Fig. 3, we visually investigate the clusterability of the dataset at hand through projection based hierarchical clustering. to this end, we used the MDH implementation [84] provided for the R package. For this experiment, we utilized HCEP, where the maximum number of cluster is conveniently set to $L = 4$. Through the iterative 2D visualization for each node of the tree, we visually identify clear patters indicating visually separable clusters. Apparently, the algorithms performs well in identifying clusters, confirming the prediction performance boost we observed previously, even for the naive clustering approach. Note that, sample coloring across the tree structure is following the categorization of the four clusters retrieved at the leaf nodes. For example, for train samples that fall in cluster 5, the predicted Healthy Aging index is retrieved by averaging the corresponding predictions of the models fitted for clusters 1, 3 and 5.

The straightforward interpretability of the HCEP approach motivated us to further investigate the potential of utilizing it in describing an innovating variable importance analysis. Notice here, that this is an uncommon task for most ensemble prediction approaches or even impossible in many cases. For this purpose we utilized the Percentage Increase in MSE (PiMSE) metric [76] through the Random Forests model for very node of the tree. Then for every path from the root node to each one of the leaf nodes, we investigate the PiMSE metric of the nodes within the path, since every point in the test set will be eventually predicted based on one of these paths. For the example at hand (Fig. 3), we consider the 10 most important variables, calculated by averaging PiMSE across all aforementioned paths. Figure 4 illustrates how these variables differentiate for each one of the four paths (1-3-4, 1-3-4, 1-2-7 and 1-2-6), according to the changes in PiMSE from the root to the leaf nodes. In more detail, each subplot depicts one of the four different paths. The PiMSE score is presented in the vertical axes, with the horizontal axes indicating the corresponding node in each path. Larger values in a variable indicate greater PiMSE score, thus expressing a more significant influence of that variable in that particular node. More specifically, the most important variables depicted here were the “srh” (Respondent’s self-rated/self-reported health, with “$catP''$”, “$catN''$”, etc. being their transformation after the statistical prepossessing), the “h-joint-disorders” (History of arthritis, rheumatism or osteoarthritis), “depression” (Current depressive status) and “age” (Age at time of measure). Another observation we can make through this visualization is that for 2 paths, “age” significance drops as tree depth is increasing, in contrast to the other two paths for which it grows. This finding may lead to the conclusion that there exist sub-populations for which a particular variable is relevant in predicting the response variable.

Concluding remarks

Population studies for aging and health analysis offer a plurality of large scale data with high diversity and complexity. Aging and health indicators are an important part of such research, while predicting the Healthy Aging index can be considered one of the greatest challenges. Motivated by the volume and diversity of the ATHLOS dataset, we focus our attention upon the clustering for prediction scheme, where unsupervised learning is utilized to enhance the prediction power. We show that imposed computation bottlenecks can be surpassed, when using appropriate hierarchical clustering within a clustering for ensemble classification scheme, while retaining prediction benefits. In addition, we investigated in depth the interpretability of the proposed architecture exposing additional advantages, such as a novel variable importance analysis. The proposed methodology is evaluated against several regression methods and the original concept exhibits very encouraging results, suggesting further developments in this direction are possible. Additionally, a straightforward open source implementation for the R project is provided. The direct expansion of the proposed methodology in classification could suggest a promising future direction, while the utilization of random space transformations to increase diversity of ensemble schemes [85, 86] seems also feasible.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1007/s13755-022-00171-1.

Below is the link to the electronic supplementary material.Supplementary file1 (XLS 76 kb)

Acknowledgements

This work is supported by the ATHLOS (Aging Trajectories of Health: Longitudinal Opportunities and Synergies) project, funded by the European Union’s Horizon 2020 Research and Innovation Program under Grant Agreement Number 635316.

Author details

[1]Department of Computer Science and Biomedical Informatics, University of Thessaly, Lamia, Greece. [2]Social Epidemiology Research Group. Health Service and Population Research Department, Institute of Psychiatry, Psychology & Neuroscience, King’s College London, London, UK. [3]Global Health Institute, King’s College London, London, UK. [4]Centro de Investigación Biomédica en Red de Salud Mental, CIBERSAM, Madrid, Spain. [5]Department of Psychiatry, Universidad Autónoma de Madrid, Madrid, Spain. [6]Hospital Universitario de La Princesa, Instituto de Investigación Sanitaria Princesa (IIS Princesa), Madrid, Spain. [7]Swiss Paraplegic Research, Guido A. Zäch Institute (GZI), Nottwil, Switzerland. [8]Department of Health Sciences & Health Policy, University of Lucerne, Lucerne, Switzerland. [9]Department of Epidemiology and Public Health, University College London, London, UK. [10]Department Preventive Medicine and Public Health, Universidad Autónoma de Madrid, Idipaz, Madrid, Spain. [11]Centro de Investigación Biomédica en Red de Epidemiología y Salud Pública, CIBERESP, Madrid, Spain. [12]Information, Evidence and Research, World Health Organization, Geneva, Switzerland. [13]Research, Innovation and Teaching Unit. Parc Sanitari Sant Joan de Déu, Sant Boi de Llobregat, Spain. [14]Fondazione IRCCS Istituto Neurologico Carlo Besta, Milan, Italy. [15]National Institute for Health and Welfare (THL), Helsinki, Finland. [16]Centre for Health Equity Studies, Department of Public Health Sciences, Stockholm University, Stockholm, Sweden. [17]Department of Public Health Sciences, Karolinska Institutet, Stockholm, Sweden. [18]Department of Epidemiology and Population Studies, Jagienllonian University, Krakow, Poland. [19]Centre for Global Mental Health. Health Service and Population Research Department, Institute of Psychiatry, Psychology & Neuroscience, King’s College London, London, UK. [20]International Institute for Applied Systems Analysis, World Population Program, Wittgenstein Centre for Demography and Global Human Capital, Laxenburg, Austria. [21]Department of Economics, Stony Brook University, Stony Brook, NY, USA. [22]Austrian Academy of Science, Vienna Institute of Demography, Vienna, Austria. [23]Russian Presidential Academy of National Economy and Public Administration (RANEPA), Moscow, Russian Federation. [24]Lithuanian University of Health Sciences, Kaunas, Lithuania. [25]Department of Epidemiology and Preventive Medicine, Jagiellonian University, Krakow, Poland. [26]Department of Nutrition and Dietetics, School of Health Science and Education, Harokopio University, Athens, Greece. [27]Department of Mathematics, University of Thessaly, Lamia, Greece.

Received: 13 September 2021 Accepted: 30 March 2022
Published online: 18 April 2022

References

1. Lee K-S, Lee B-S, Semnani S, Avanesian A, Um C-Y, Jeon H-J, Seong K-M, Yu K, Min K-J, Jafari M. Curcumin extends life span, improves health span, and modulates the expression of age-associated aging genes in drosophila melanogaster. Rejuvenation Res. 2010;13(5):561–70.
2. Mathias JS, Agrawal A, Feinglass J, Cooper AJ, Baker DW, Choudhary A. Development of a 5 year life expectancy index in older adults using predictive mining of electronic health record data. J Am Med Inform Assoc. 2013;20(e1):e118–24.
3. Herland M, Khoshgoftaar TM, Wald R. A review of data mining using big data in health informatics. J Big data. 2014;1(1):1–35.
4. Eurostat, Population structure and ageing. statistics explained.
5. Mather M, Jacobsen LA, Pollard KM. Aging in the united states, Population Reference Bureau; 2015.

6. Organization WH, et al. Men, ageing and health: achieving health across the life span. Tech. rep. Geneva: World Health Organization; 2001.
7. DESA U. World population ageing 2015, in: United Nations DoEaSA, population division editor; 2015.
8. Alwan A, et al. Global status report on noncommunicable diseases 2010. Geneva: World Health Organization; 2011.
9. Seeman TE, Crimmins E, Huang M-H, Singer B, Bucur A, Gruenewald T, Berkman LF, Reuben DB. Cumulative biological risk and socio-economic differences in mortality: Macarthur studies of successful aging. Soc Sci Med. 2004;58(10):1985–97.
10. Wu M-S, Lan T-H, Chen C-M, Chiu H-C, Lan T-Y. Socio-demographic and health-related factors associated with cognitive impairment in the elderly in Taiwan. BMC Public Health. 2011;11(1):22.
11. Wagner K-H, Cameron-Smith D, Wessner B, Franzke B. Biomarkers of aging: from function to molecular biology. Nutrients. 2016;8:338. https://doi.org/10.3390/nu8060338.
12. Caballero FF, Soulis G, Engchuan W, Sánchez-Niubó A, Arndt H, Ayuso-Mateos JL, Haro JM, Chatterji S, Panagiotakos DB. Advanced analytical methodologies for measuring healthy ageing and its determinants, using factor analysis and machine learning techniques: the athlos project. Sci Rep. 2017;7:43955.
13. Higueras-Fresnillo S, Guallar-Castillón P, Cabanas-Sanchez V, Banegas JR, Rodríguez-Artalejo F, Martinez-Gomez D. Changes in physical activity and cardiovascular mortality in older adults. J Geriatr Cardiol: JGC. 2017;14(4):280.
14. Martinez-Gomez D, Guallar-Castillon P, Higueras-Fresnillo S, Garcia-Esquinas E, Lopez-Garcia E, Bandinelli S, Rodríguez-Artalejo F. Physical activity attenuates total and cardiovascular mortality associated with physical disability: a national cohort of older adults. J Gerontol: Ser A. 2018;73(2):240–7.
15. Graciani A, García-Esquinas E, López-García E, Banegas J. Ideal cardiovascular health and risk of frailty in older adults. Circulation. 2016;9(3):239–45.
16. Tyrovolas S, Panagiotakos D, Georgousopoulou E, Chrysohoou C, Tousoulis D, Haro JM, Pitsavos C. Skeletal muscle mass in relation to 10 year cardiovascular disease incidence among middle aged and older adults: the attica study. J Epidemiol Community Health. 2020;74(1):26–31.
17. Kollia N, Panagiotakos DB, Chrysohoou C, Georgousopoulou E, Tousoulis D, Stefanadis C, Papageorgiou C, Pitsavos C. Determinants of healthy ageing and its relation to 10-year cardiovascular disease incidence: the Attica study. Cent Eur J Public Health. 2018;26(1):3–9.
18. Kollia N, Caballero FF, Sánchez-Niubó A, Tyrovolas S, Ayuso-Mateos JL, Haro JM, Chatterji S, Panagiotakos DB. Social determinants, health status and 10-year mortality among 10,906 older adults from the English longitudinal study of aging: the athlos project. BMC Public Health. 2018;18(1):1357.
19. Soler-Vila H, García-Esquinas E, León-Muñoz LM, López-García E, Banegas JR, Rodríguez-Artalejo F. Contribution of health behaviours and clinical factors to socioeconomic differences in frailty among older adults. J Epidemiol Community Health. 2016;70(4):354–60.
20. Doménech-Abella J, Mundó J, Moneta MV, Perales J, Ayuso-Mateos JL, Miret M, Haro JM, Olaya B. The impact of socioeconomic status on the association between biomedical and psychosocial well-being and all-cause mortality in older spanish adults. Soc Psychiatry Psychiatr Epidemiol. 2018;53(3):259–68.
21. Hossin M, Koupil I. Early life social and health determinants of adult socioeconomic position across two generations. Eur J Public Health. 2018;28(4):cky213.
22. Machado-Fragua MD, Struijk EA, Graciani A, Guallar-Castillon P, Rodríguez-Artalejo F, Lopez-Garcia E. Coffee consumption and risk of physical function impairment, frailty and disability in older adults. Eur J Nutr. 2019;58(4):1415–27.
23. Tyrovolas S, Haro JM, Foscolou A, Tyrovola D, Mariolis A, Bountziouka V, Piscopo S, Valacchi G, Anastasiou F, Gotsis E, et al. Anti-inflammatory nutrition and successful ageing in elderly individuals: the multinational medis study. Gerontology. 2018;64(1):3–10.
24. Stefler D, Malyutina S, Nikitin Y, Nikitenko T, Rodriguez-Artalejo F, Peasey A, Pikhart H, Sabia S, Bobak M. Fruit, vegetable intake and blood pressure trajectories in older age. J Hum Hypertens. 2019;33(9):671–8.
25. León-Muñoz LM, Guallar-Castillón P, García-Esquinas E, Galán I, Rodríguez-Artalejo F. Alcohol drinking patterns and risk of functional limitations in two cohorts of older adults. Clin Nutr. 2017;36(3):831–8.
26. Ortolá R, García-Esquinas E, Galán I, Guallar-Castillón P, López-García E, Banegas J, Rodríguez-Artalejo F. Patterns of alcohol consumption and risk of falls in older adults: a prospective cohort study. Osteoporos Int. 2017;28(11):3143–52.
27. de la Torre-Luque A, Ayuso-Mateos JL, Sanchez-Carro Y, de la Fuente J, Lopez-Garcia P. Inflammatory and metabolic disturbances are associated with more severe trajectories of late-life depression. Psychoneuroendocrinology. 2019;110:104443.
28. de la Torre-Luque A, de la Fuente J, Sanchez-Niubo A, Caballero FF, Prina M, Muniz-Terrera G, Haro JM, Ayuso-Mateos JL. Stability of clinically relevant depression symptoms in old-age across 11 cohorts: a multi-state study. Acta Psychiatr Scand. 2019;140(6):541–51.
29. de la Torre-Luque A, de la Fuente J, Prina M, Sanchez-Niubo A, Haro JM, Ayuso-Mateos JL. Long-term trajectories of depressive symptoms in old age: relationships with sociodemographic and health-related factors. J Affect Disord. 2019;246:329–37.
30. Panaretos D, Koloverou E, Dimopoulos AC, Kouli G-M, Vamvakari M, Tzavelas G, Pitsavos C, Panagiotakos DB. A comparison of statistical and machine-learning techniques in evaluating the association between dietary patterns and 10-year cardiometabolic risk (2002–2012): the attica study. Br J Nutr. 2018;120(3):326–34.
31. Engchuan W, Dimopoulos AC, Tyrovolas S, Caballero FF, Sanchez-Niubo A, Arndt H, Ayuso-Mateos JL, Haro JM, Chatterji S, Panagiotakos DB. Sociodemographic indicators of health status using a machine learning approach and data from the English longitudinal study of aging (elsa). Med Sci Monit. 2019;25:1994.
32. Alapati YK, Sindhu K. Combining clustering with classification: a technique to improve classification accuracy. Lung Cancer. 2016;32(57):3.
33. Rouzbahman M, Jovicic A, Chignell M. Can cluster-boosted regression improve prediction of death and length of stay in the ICU? IEEE J Biomed Health Inform. 2017;21(3):851–8. https://doi.org/10.1109/JBHI.2016.2525731.
34. Trivedi S, Pardos ZA, Heffernan NT. The utility of clustering in prediction tasks, arXiv:1509.06163.
35. Gan H, Sang N, Huang R, Tong X, Dan Z. Using clustering analysis to improve semi-supervised classification. Neurocomputing. 2013;101:290–8.
36. Belkin M, Niyogi P, Sindhwani V. Manifold regularization: a geometric framework for learning from labeled and unlabeled examples. J Mach Learn Res. 2006;7:2399–434.
37. Agrawal U, Soria D, Wagner C, Garibaldi J, Ellis IO, Bartlett JM, Cameron D, Rakha EA, Green AR. Combining clustering and classification ensembles: a novel pipeline to identify breast cancer profiles. Artif Intell Med. 2019;97:27–37.
38. Tran CT, Zhang M, Andreae P, Xue B, Bui LT. Improving performance of classification on incomplete data using feature selection and clustering. Appl Soft Comput. 2018;73:848–61.
39. Sanchez-Niubo A, Egea-Cortés L, Olaya B, Caballero FF, Ayuso-Mateos JL, Prina M, Bobak M, Arndt H, Tobiasz-Adamczyk B, Pająk A, et al. Cohort profile: the ageing trajectories of health-longitudinal opportunities and synergies (athlos) project. Int J Epidemiol. 2019;48(4):1052–1053i.
40. Prina AM, Acosta D, Acosta I, Guerra M, Huang Y, Jotheeswaran A, Jimenez-Velazquez IZ, Liu Z, Llibre RJ, Salas JA. Cohort profile: the 10/66 study. Int J Epidemiol. 2017;46(2):406.
41. Luszcz MA, Giles LC, Anstey KJ, Browne-Yung KC, Walker RA, Windsor TD. Cohort profile: the Australian longitudinal study of ageing (alsa). Int J Epidemiol. 2016;45(4):1054–63.
42. Leonardi M, Chatterji S, Koskinen S, Ayuso-Mateos JL, Haro JM, Frisoni G, Frattura L, Martinuzzi A, Tobiasz-Adamczyk B, Gmurek M, et al. Determinants of health and disability in ageing population: the courage in Europe project (collaborative research on ageing in europe). Clin Psychol Psychother. 2014;21(3):193–8.
43. Steptoe A, Breeze E, Banks J, Nazroo J. Cohort profile: the English longitudinal study of ageing. Int J Epidemiol. 2013;42(6):1640–8.
44. Rodríguez-Artalejo F, Graciani A, Guallar-Castillón P, León-Muñoz LM, Zuluaga MC, López-García E, Gutiérrez-Fisac JL, Taboada JM, Aguilera MT, Regidor E, et al. Rationale and methods of the study on nutrition

and cardiovascular risk in Spain (enrica). Revista Española de Cardiología (English Edition). 2011;64(10):876–82.

45. Peasey A, Bobak M, Kubinova R, Malyutina S, Pajak A, Tamosiunas A, Pikhart H, Nicholson A, Marmot M. Determinants of cardiovascular disease and other non-communicable diseases in central and eastern Europe: rationale and design of the hapiee study. BMC Public Health. 2006;6(1):255.
46. KS, Health 2000 and 2011 surveys-thl biobank. National Institute for Health and Welfare (2018). Accessed 18 July 2008.
47. Sonnega A, Faul JD, Ofstedal MB, Langa KM, Phillips JW, Weir DR. Cohort profile: the health and retirement study (hrs). Int J Epidemiol. 2014;43(2):576–85.
48. Ichimura H, Shimizutani S, Hashimoto H. Jstar first results 2009 report. Research Institute of Economy, Trade and Industry (RIETI): Tech. rep; 2009.
49. Park JH, Lim S, Lim J, Kim K, Han M, Yoon IY, Kim J, Chang Y, Chang CB, Chin HJ, et al. An overview of the Korean longitudinal study on health and aging. Psychiatry Investig. 2007;4(2):84.
50. Wong R, Michaels-Obregon A, Palloni A. Cohort profile: the Mexican health and aging study (MHAS). Int J Epidemiol. 2017;46(2):e2–e2.
51. Kowal P, Chatterji S, Naidoo N, Biritwum R, Fan W, Lopez Ridaura R, Maximova T, Arokiasamy P, Phaswana-Mafuya N, Williams S, et al. Data resource profile: the world health organization study on global ageing and adult health (Sage). Int J Epidemiol. 2012;41(6):1639–49.
52. Börsch-Supan A, Brandt M, Hunkler C, Kneip T, Korbmacher J, Malter F, Schaan B, Stuck S, Zuber S. Data resource profile: the survey of health, ageing and retirement in Europe (SHARE). Int J Epidemiol. 2013;42(4):992–1001.
53. Whelan BJ, Savva GM. Design and methodology of the Irish longitudinal study on ageing. J Am Geriatr Soc. 2013;61:S265–8.
54. Arokiasamy P, Bloom D, Lee J, Feeney K, Ozolins M. Longitudinal aging study in India: vision, design, implementation, and preliminary findings. In: Smith JP, Majmundar M, editors. Aging in Asia: findings from new and emerging data initiatives. Washington: National Academies Press; 2012.
55. Seetharaman P, Wichern G, Le Roux J, Pardo B. Bootstrapping single-channel source separation via unsupervised spatial clustering on stereo mixtures. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2019. pp. 356–360.
56. Dietterich TG, Ensemble methods in machine learning. In: International workshop on multiple classifier systems, Springer, 2000; pp. 1–15.
57. Boongoen T, Iam-On N. Cluster ensembles: a survey of approaches with recent extensions and applications. Comput Sci Rev. 2018;28:1–25.
58. Saraçli S, Doğan N, Doğan İ. Comparison of hierarchical cluster analysis methods by cophenetic correlation. J Inequal Appl. 2013;2013(1):1–8.
59. Pavlidis NG, Hofmeyr DP, Tasoulis SK. Minimum density hyperplanes. J Mach Learn Res. 2016;17(1):5414–46.
60. Murtagh F, Legendre P. Ward's hierarchical agglomerative clustering method: which algorithms implement ward's criterion? J Classif. 2014;31(3):274–95.
61. Zhang W, Zhao D, Wang X. Agglomerative clustering via maximum incremental path integral. Pattern Recogn. 2013;46(11):3056–65.
62. Sharma A, López Y, Tsunoda T. Divisive hierarchical maximum likelihood clustering. BMC Bioinform. 2017;18(16):546.
63. Tasoulis S, Cheng L, Välimäki N, Croucher NJ, Harris SR, Hanage WP, Roos T, Corander J. Random projection based clustering for population genomics. IEEE Int Conf Big Data (Big Data). 2014;2014:675–82. https://doi.org/10.1109/BigData.2014.7004291.
64. Tasoulis SK, Tasoulis DK, Plagianakos VP. Enhancing principal direction divisive clustering. Pattern Recogn. 2010;43(10):3391–411.
65. Hofmeyr DP. Clustering by minimum cut hyperplanes. IEEE Trans Pattern Anal Mach Intell. 2016;39(8):1547–60.
66. Azzalini A, Torelli N. Clustering via nonparametric density estimation. Stat Comput. 2007;17(1):71–80.
67. Stuetzle W, Nugent R. A generalized single linkage method for estimating the cluster tree of a density. J Comput Graph Stat. 2010;19(2):397–418.
68. Menardi G, Azzalini A. An advancement in clustering via nonparametric density estimation. Stat Comput. 2014;24(5):753–67.
69. Ben-David S, Lu T, Pál D, Sotáková M. Learning low density separators. In: Artificial Intelligence and Statistics; 2009, pp. 25–32.
70. Boley D. Principal direction divisive partitioning. Data Min Knowl Disc. 1998;2(4):325–44.
71. Zumel N, Mount J vtreat: a data. frame processor for predictive modeling, arXiv:1611.09477.
72. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math. 1987;20:53–65.
73. Baker FB, Hubert LJ. Measuring the power of hierarchical cluster analysis. J Am Stat Assoc. 1975;70(349):31–8.
74. Tasoulis S, Pavlidis NG, Root T. Nonlineardimensionality reduction for clustering. Pattern Recogn. 2020;107:107508.
75. Emerson J, Kane M. biganalytics: Utilities for "big. matrix" objects from package "bigmemory", J Stat Softw.
76. Liaw A, Wiener M, et al. Classification and regression by randomforest. R News. 2002;2(3):18–22.
77. Chai T, Draxler RR. Root mean square error (RMSE) or mean absolute error (MAE)?-Arguments against avoiding RMSE in the literature. Geosci Model Develop. 2014;7(3):1247–50.
78. Kim J-H. Estimatingclassification error rate: repeated cross-validation, repeated hold-out and bootstrap. Comput Stat Data Anal. 2009;53(11):3735–45. https://doi.org/10.1016/j.csda.2009.04.009.
79. Microsoft, S. Weston, foreach: provides Foreach Looping Construct, r package version 1.4.7 url = https://CRAN.R-project.org/package=foreach (2019).
80. Chen T, Guestrin C. 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.
81. Kingma DP, Ba J. Adam: a method for stochastic optimization, arXiv:1412.6980.
82. Rousseeuw PJ, Kaufman L. Finding groups in data, Hoboken: Wiley Online Library 1.
83. Tibshirani R, Walther G, Hastie T. Estimating the number of clusters in a data set via the gap statistic. J R Stat Soc Ser B. 2001;63(2):411–23.
84. Hofmeyr D, Pavlidis N. Ppci: an r package for cluster identification using projection pursuit. R J Appear. 2019. https://doi.org/10.32614/RJ-2019-046.
85. Tasoulis SK, Vrahatis AG, Georgakopoulos SV, Plagianakos VP. Biomedical data ensemble classification using random projections. In: 2018 IEEE International Conference on Big Data (Big Data), IEEE; 2018, pp. 166–172.
86. Cannings TI, Samworth RJ. Random-projection ensemble classification. J R Stat Soc Ser B. 2017;79(4):959–1035.