SAPIENZA
UNIVERSITÀ DI ROMA

Explainable Clinical Decision Support System

Opening black-box meta-learner algorithm Expert's based

ABRO PhD

P.hD in Operation Research – XXXIII Cycle

Candidate

Dr. Francesco Curia
ID number 1557544

Thesis Advisor

prof. Ing. Paolo Dell'Olmo

31 October 2020

Thesis defended on 19 April 2021
in front of a Board of Examiners composed by:

Prof. Fabio Tardella, Sapienza Università di Roma (chairman)

Prof. Giuseppe Baselli, Politecnico di Milano

Prof. Stefano Panzieri, Università di Roma Tre

Explainable Clinical Decision Support System
Ph.D. thesis. Sapienza - University of Rome

Faculty of Information Engineering, Computer Science and Statistics.

Author's email: francesco.curia@uniroma1.it

In memoria di mia madre Lio Domenica

Abstract

Mathematical optimization methods are the basic mathematical tools of all artificial intelligence theory. In the field of machine learning and deep learning the examples with which algorithms learn (*training data*) are used by sophisticated cost functions which can have solutions in closed form or through approximations. The interpretability of the models used and the relative transparency, opposed to the opacity of the black-boxes, is related to *how* the algorithm learns and this occurs through the optimization and minimization of the errors that the machine makes in the learning process. In particular in the present work is introduced a new method for the determination of the weights in an ensemble model, supervised and unsupervised, based on the well known Analytic Hierarchy Process method (AHP). This method is based on the concept that behind the choice of different and possible algorithms to be used in a machine learning problem, there is an *expert* who controls the decision-making process. The expert assigns a complexity score to each algorithm (based on the concept of complexity-interpretability trade-off) through which the weight with which each model contributes to the training and prediction phase is determined. In addition, different methods are presented to evaluate the performance of these algorithms and explain how each feature in the model contributes to the prediction of the outputs. The interpretability techniques used in machine learning are also combined with the method introduced based on AHP in the context of clinical decision support systems in order to make the algorithms (black-box) and the results interpretable and explainable, so that clinical-decision-makers can take *controlled* decisions together with the concept of "right to explanation" introduced by the legislator, because the decision-makers have a civil and legal responsibility of their choices in the clinical field based on systems that make use of artificial intelligence. No less, the central point is the interaction between the expert who controls the algorithm construction process and the domain expert, in this case the clinical one. Three applications on real data are implemented with the methods known in the literature and with those proposed in this work: one application concerns cervical cancer, another the problem related to diabetes and the last one focuses on a specific pathology developed by HIV-infected individuals. All applications are supported by plots, tables and explanations of the results, implemented through Python libraries. The main case study of this thesis regarding HIV-infected individuals concerns an unsupervised ensemble-type problem, in which a series of clustering algorithms are used on a set of features and which in turn produce an output used again as a set of *meta*-features to provide a set of labels for each given cluster. The *meta*-features and labels obtained by choosing the best algorithm are used to train a Logistic regression *meta*-learner, which in turn is used through some explainability methods to provide the value of the contribution that each algorithm has had in the training phase. The use of Logistic regression as a *meta*-learner classifier is motivated by the fact that it provides appreciable results and also because of the easy explainability of the estimated coefficients.

Contents

Chapter 1

Interpretability

In this work it is dealt with the problem of interpretability of the algorithms used in machine learning, a subset of the techniques used in artificial intelligence. The main methods of explainability and interpretability known in the literature have are used for the explanation of the results of the applications and new methods i'll be introduced in the work to support the decision-making phase. The work is based on the interaction between experts for the choice of models to be used in the clinical decision and on the results obtained, which must necessarily be validated by the clinical domain expert. The method introduced is based on the choice of n-algorithms by an expert, who assigns a score to each based on the complexity and interpretability of the model, this score through the methodology inherited from the AHP method leads to obtaining weights for each model, which contributes to the evaluation and to the predictions obtained from the models. In the first two applications, the results obtained were very good, the first on cervical cancer produced better results than those currently present in the literature. As regards the application on diabetes data the results were not higher than those known in some of the works cited but contextualizing the methodology used in the work, it may be considered important because the authors cited for the works on the dataset relating diabetes didn't use methods of interpretability. In the third application, n-clusterizers are trained, chosen by the expert, each with its own score and relative weight. The first training takes place on the original data set and the second on the labels obtained from each model used. Subsequently, a *meta learner* based on Logistic regression is trained considering as features the single models used in the second training layer and as a target the label obtained by applying the *mode* function (6.16) on the outputs obtained from each model. Subsequently, the application of the methods of interpretability and of features importance led to the evaluation of the importance of each single clustering model used. In the post-analysis phase, the weights given by the expert are compared with the (optimal) obtained by solving a quadratic optimization problem. In the last application on the DED problem for HIV patients there was no benchmark for comparison but rather than the overall result of the models used; the most interesting result obtained was that of being able to introduce a new approach to the explainability and interpretability of algorithms and the evaluation of each single model used. This work could be in the future starting point for the CDSSs in the context of the explainable AI.

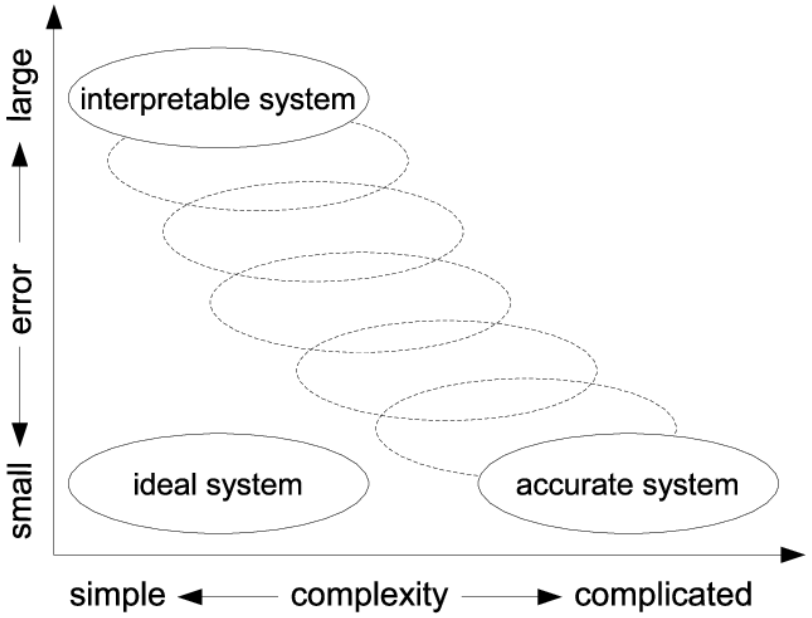

Figure 1.1. Source: Wang, Di *et al.* (2018). An Interpretable Neural Fuzzy Inference System for Predictions of Underpricing in Initial Public Offerings, Neurocomputing

1.1 Introduction

The concept of interpretability is very broad and in the literature we find several definitions, borrowed from mathematical logic [1] this definition we could partially define it as a "relationship" between formal theories, which express the possibility of interpreting or translating the one in the other. Through this definition, interpretability can be clearly extended to machine learning (ML) problems considering both supervised and unsupervised problems. According to Cain's definition [2], a mathematical model is therefore a non-perfect but very faithful quantitative representation of a natural phenomenon. Loyalty is not a term chosen at random and subsequently this word will also occur in ML problems. Starting from these definitions, the interpretability of a model is a fundamental aspect for evaluating *how* decisions were made and evaluating the predictions made by the model a posteriori. Miller [3] defines interpetability as the "degree to which a human being can understand the cause of a decision", therefore the more interpretable a model, the more it is possible to understand decision logic. Going into more detail of interpretability in the problems related to machine learning Doshi-Velez and Kim [4] give a very exhaustive definition of interpretability defining it as "the ability to explain or present in a comprehensible way to a human being".

1.2 Lack of transparency

Although a machine learning model provides predictions with high precision and profound accuracy, what matters most to a decision-maker is how it came to that

decision and therefore how it came about the prediction was determined or how a particular instance is classified by the algorithm. Doshi-Velez and Kim [4] answer this question by stating that: "the problem is that a single metric, like accuracy, it is an incomplete description of most of the real world tasks". Always the same authors explain it anyway, there are cases and applications where it is not necessary to provide an explanation in terms of interpretability. In the work of Carvalho *et al.*[12] provides an exhaustive survey of the problem of interpretability, the authors citing Doshi-Velez and Kim, divide the situations in which explanations are not necessary into two categories:

(a) in the absence of significant impact or serious consequences for incorrect results and

(b) when the problem is sufficiently studied and validated in real applications that we trust the decisions of the system, even if the system is not perfect

Certainly there are situations where it is necessary to provide an explanation on why one choice is preferable to another, think, as an example, to the machine learning applications related to the clinical field. Ahmad *et al.* [13] address the problem of the global and local interpretability of a classification tree used in classifying an individual as diabetic or not, and pose the problem that a global model may not capture some nuances that instead at the local level it would be possible to identify and explain at the individual level, rather than on the whole population. Interpretability is also deal with through the LIME (Local Interpretable Model-Agnostic Explanations) an method introduced by Ribeiro [6] which provide a local linear interpretation of the model. For the healthcare sector the aim of interpretability will provide valid motivations to the physicians who are called to respond to important problems, from an ethical and professional point of view and social.

According to Ribeiro, the problem of interpretability therefore lies in the incompleteness of the formalization of the problem. The scenarios that are impacted are different, among which we find:

(a) **Security**: the artificial intelligence system is never fully testable, as it is not possible to create a complete list of scenarios in which the system may not work.

(b) **Ethics**: the concept of morals and ethics for an artificial intelligence system is a concept too abstract to be codified entirely by the system.

(c) **Objectives**: since the algorithm could optimize conflicting objectives a situation of disagreement would arise between what you want and what you get.

(d) **Multi-objective**: two well-defined wishes in ML systems can compete with each other, think of the accuracy of the prediction required and the problem of privacy and create a multi-objective disadvantage

Once you understand the problem of interpretability, you can proceed to define the types, properties and methods that make up this vast and controversial topic of interpretability.

1.3 Type of interpretability

Starting from the work of Tjoa and Guana [14], we can give some definitions concerning the types of interpretability:

(A) **Perceptual interpretability**: In this category human perception is considered as interpretation. We find a series of dedicated subcategories for this purpose

1. **Saliency**: This method explains the decisions of an algorithm through assigned values that reflect the importance of the input components and their contribution to the decision. These values can take the form of probability or in the recognition of images through heat maps.
2. **Signal method**: These are interpretability methods that observe the stimulation of neurons (therefore in a deep learning approach) or a set of neurons called signal methods. The activated values of neurons can be manipulated and/or transformed into interpretable forms. For example, the activation of neurons in a layer can be used to reconstruct an image similar to input as happens with autoencoders.
3. **Verbal interpretability**: In this typology it is assumed that there are verbal structures that a human can understand immediately. Logical declarations can be formed by the correct concatenation of predicates, connectives etc. An example of logic may be that of conditional education. A clear example of this type of interpretability is shown in Townsend *et al.* [15] in which this concept can be seen from a symbolic and relational point of view.

(B) **Interpretability through mathematical structures**

1. **Predefined model**: In the study of complex systems the idea is that parametric mathematical model can be helpful in explaining the phenomenon. The interpretation of the parameters is better if consistent with the hypothesis behind the assumptions, for example in a linear regression model, if the basic hypotheses are respected and the OLS estimates are consistent then the interpretation of the parameters is certainly clearer, adding more complex components the model can be improved. In the perspective of more complex models such as neural networks the concept of trade off is the basis of the reasoning that leads to the explainability of the model.
2. **Features extraction**: Is one of the best known techniques in the literature it is an approach based on correlations and associations that allows

to discriminate which features are less relevant than others or serves to find internal patterns that can help explain and translate the complex components of the models.

3. **Sensitivity**: In this category we find the methods based on gradient analysis, perturbation and localization are considered. They are therefore considered infitesimal neighborhoods of nearby points by evaluating how the function changes in a neighborhood, this is the concept behind the analysis of the gradient, while at the local level all those points are considered all the closer to the prediction made by the model that do not involve losses high.
4. **Optimization**: This approach is considered to be the basis of the desirable properties of the algorithms, or with which method if exact or heuristic, a mathematical programming problem is solved. As it happens in the LIME method already mentioned, a function of "infidelity" is minimized and qualitatively analyzed. The breakdown of the problem into many subproblems leads to an approximate approach of the problem but anyway of simpler interpretation.

1.4 Properties of interpretable models

1.4.1 Trasparency

After seeing that there is no single definition for interpretability, another fundamental concept is that of properties that an interpretable model must possess. Again different meanings are possible and studies produce, step by step, more definitions and answers to different questions. Several works deal with the theme and in Lipton [5] we find a good and comprehensive survey of the desirable properties of interpretable models; among the main properties there is the very important one of the "transparency" of the models, that is, how the algorithm actually works. Specifically, transparency concerns the mechanism underlying the model that is used and how each element or component of the model works, in a mathematical sense we can identify these components in the parameters that make up the model used. Within this macro concept we find several components, Arrieta *et al.*[25] provide a good definition of transparency: "a model is considered to be transparent if by itself it is understandable". We can characterize different degrees of understandability

(A) **Simulability**: The author sets this sub-feature as follows: "a model is transparent if a person can contemplate a model simultaneously" or in the meaning of Ribeiro *et al.* [6]: "a model is interpretable if it can be presented visually and understood intuitively". The concept of transparency is also applicable as the algorithm training is provided, the necessary steps and each step which output it produces. Of course, by definition an interpretable model is a simple model, but a problematic problem that is added to that of interpretation is surely the concept of compromise between complexity and interpretability, the challenge therefore remains methods that work intuitively dare a fairly

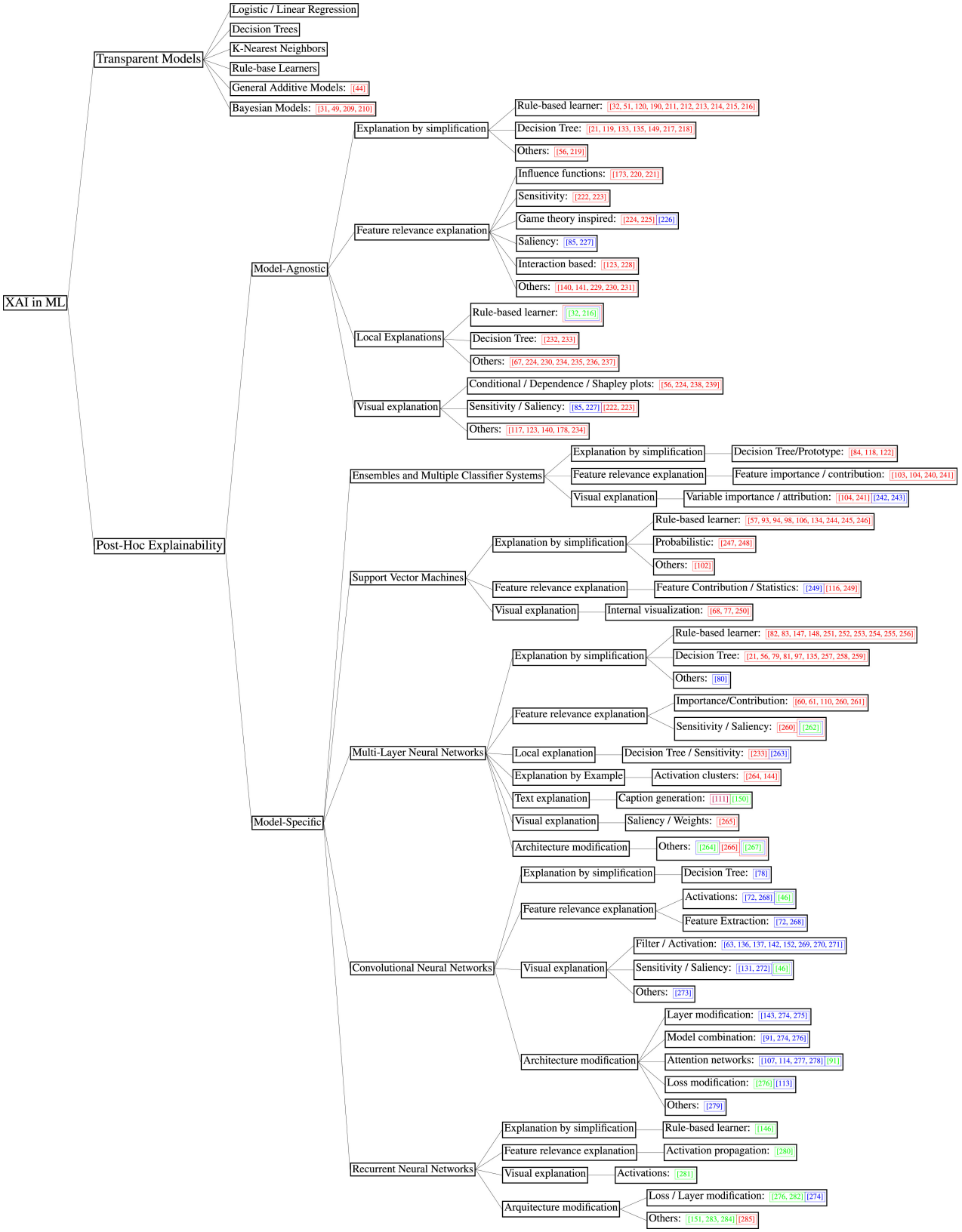

Figure 1.2. Taxonomy of XAI models. Source: A.B.,Arrieta *et al.* [25]

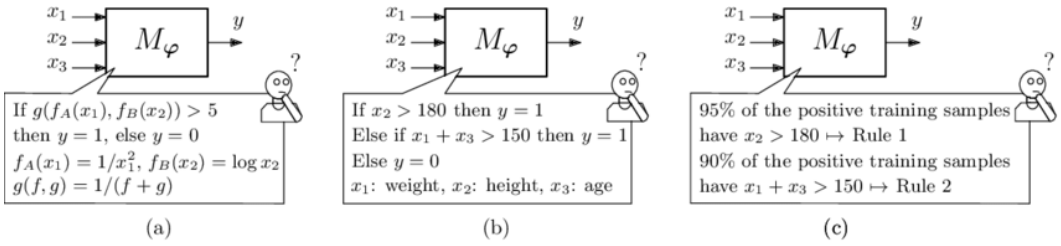

Figure 1.3. Conceptual diagram exemplifying the different levels of transparency characterizing a ML model M, with ϕ denoting the parameter set of the model at hand: (a) simulatability; (b) decomposability; (c) algorithmic transparency. References [25]

simple explanation of complex models, such as which on average lead to very predictive results more accurate.

(B) **Decomposition**: Returning to Lipton, in his work he also introduces the decomposition property, i.e. each input element of a model must be individually interpretable, just think of the inputs of a classification tree or a linear model. This property discriminates between models to which the inputs are engineered or anonymous, this property could also be affected not only by the number of features considered and by their engineering, even if certain indicators are contained in the data set or not.

(C) **Algorithmic transparency**: At the algorithm level the notion of interpretability can be applied in the phase in which the algorithm learns. The author sets the example in the case of linear models, investigating the surface of the error obtained by minimizing the loss function. The convergence to an excellent global also for test data, can introduce the concept of trust in the learning method, which in black-boxing like deep neural networks does not happen because, in the training phase, often the associated cost functions to learning, they are optimized through heuristic methods and therefore the solutions produced are not excellent overall, but at best with approximations. Therefore the concept of algorithmic transparency lies precisely in the very way in which the algorithm works.

1.4.2 Post-Hoc Interpretability

Another element on which the concept of interpretability introduced by Lipton's work is examined is that of Post-Hoc interpretability, in which the evaluation takes place after the application of the model. This interpretation does not clarify very precisely how the model works. The main approaches which we will examine shortly mainly concern visualizations and representations of trained models. One of the main advantages of the a posteriori approach is that of being able to use more complex models without therefore sacrificing high predictive performance.

(A) **Textual Explanation**: McAuley and Leskovec [7], in the context of recommendation systems use textual interpretation to explain decisions made through latent factors, therefore not directly observable. The methodology

suggested by the work is based on the idea of training latent factors for the predictions minimizing the square of the error and maximizing the probability of likelihood. Since it is necessary that the latent space be interpretable and a quantitative assessment can be made, Chang *et al.* [8] propose quantitative methods for which measure semantic meaning in inferred arguments, showing that they capture aspects of the model that they are not recognized by previous model quality measures based on likelihood; it is interesting to note that in their work the models of arguments that perform best with a certain propensity could infer less semantically significant arguments.

(B) **Local Explanation**: Starting from the concept of local optimality on the concept of the "nearest neighbor", again in the work of Ribeiro is considered the analysis of a point locally close to another reference point in a given region, where learning occurs through a separate scattered linear model. Plumb *et al.* [10] present a new method, called MAPLE, which using Random Forest (RF) presents itself as an extremely accurate predictive model that provides very faithful explanations, eluding the typical compromise between accuracy and interpretability. The idea is based on the use of RF to select the characteristics for locally linear models introduced in the work of Kazemitabar *et al.* [11], for the calculation of the importance of the variables calculated through the use of classification trees considering the impurity of the nodes. All these methods are therefore based on the nearest neighbors algorithm and give a local assessment of the problem.

(C) **Visualization**: A very interesting approach on the evaluation of a posteriori interpretability is to exploit the visualization of the results of the post-training model. A qualitative method based on the analysis of disturbed inputs that can give clues to what the model has learned. This approach is surely inherited from the works and from the methodology related to the problems of computer vision and image recognition, among the main works that exploit this methodology we have Mahendran *et al.* [9] in which a discriminative convolutional neural network is trained to generate representation through image input. The original image can be reconstructed with high fidelity starting from very high levels of representations, optimizing the cost function through a gradient-descent randomly initialized on the pixels of the training images.

1.5 Interpretability: methods and models

This section presents the main methods and the latest works with the greatest impact in the field of interpretability.

1.5.1 Linear regression

A linear regression model predicts the dependent variable as a weighted sum of the independent variables, or also called covariates. The linearity of the relationship clearly facilitates interpretation. These models are used to model the dependence of a y (target) variable on some x variables. It is assumed that the data are affected

by noise (of the Gaussian type with mean μ and variance equal to σ) and the relationship can be formalized as follows:

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \epsilon \tag{1.1}$$

where ϵ is the error, which is the difference between the forecast and the actual result. These errors are assumed to follow a Gaussian distribution, which means that we make errors in both a negative and positive direction and we make many small errors and a few big errors. One of the main methods in estimating model parameters is the well-known "ordinary least squares method" (OLS) used to find the weights β_j that minimize the differences squared between the actual and estimated results:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\beta_0,\ldots,\beta_p} \sum_{i=1}^{n} \left(y_i - \left(\beta_0 + \sum_{j=1}^{p} \beta_j x_j^{(i)} \right) \right)^2 \tag{1.2}$$

Linearity simplifies the estimation procedure and, above all, linear equations have an easy-to-understand interpretation, for example, in the clinical field one of the fundamental aspects is to quantify the influence of a drug or therapy and at the same time take into account sex , age and other characteristics in an interpretable way and linearity leads to interpretable models. Linear effects are easy to describe. The assumption of normality makes the estimates consistent and the estimators enjoy the optimality properties of the estimators, in case the assumptions are not satisfied, the estimates are distorted and the results invalidated. Another important property is the one concerning constant variance (homoschedasticity) where the error terms are constant. Another important property is the absence of multicollinearity, that is when a variable can be expressed as a linear combination of the other variables. Depending on the type of variable, the interpretation, while remaining simple, also changes. Numerical Variable: increasing a unit modifies the estimated result based on its weight. Binary variable: it is a variable that takes only two possible values for each observation. Categorical variable with multiple multiclasses: it is a variable with a fixed number of possible values. For a categorical variable with k-classes, the interpretation for each category is therefore the same as the interpretation for binary variables. The optimization of the loss function which produces least squares estimates being an unconstrained quadratic function, produces a global optimum in the parametric space, therefore the transparency required in the interpretability of the models is obtained. One of the main problems, already discussed, is the understanding and regulation of algorithms; linear models are easy to interpret for this reason, if we take for example non-linear model functions, the functions to be optimized often require non-global methods and therefore the achievement of the optimal is considered "opaque". Two models that introduce sparsity are LASSO and RIDGE, which starting from the mathematical formalization of the cost function for the OLS estimate (1.2) add a term of *penalty*, which aims to introduce the concept of selection of variables. In the LASSO case, the penalty term that is added is the

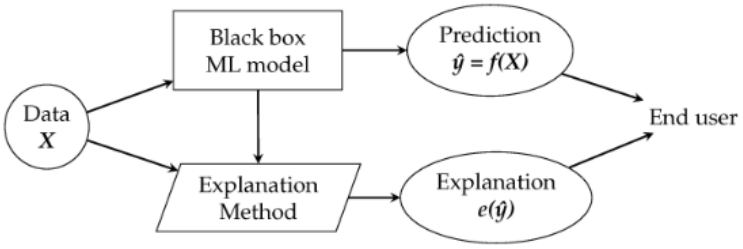

Figure 1.4. Black Box Algorithms. Source: [12]

norm L-1 so the (1.2), considering the constraint $\sum_j^n |\beta| \leq t$, becomes

$$\min_{\boldsymbol{\beta}} \left(\frac{1}{n} \sum_{i=1}^{n} (y_i - x_i^T \boldsymbol{\beta})^2 + \lambda ||\boldsymbol{\beta}||_1 \right) \tag{1.3}$$

this norm L-1 cancels many of the parameters in the estimation procedure thus reducing the dimensionality of the characteristic number, hence the concept of *sparsity*. The λ parameter is the regularization parameter and is calculated through cross-validation. For $\lambda \to \infty$ many weights become 0. By introducing the L-2 standard instead, we obtain the Tikhonov regularization model, known as RIDGE Regression. The problem (1.2) then becomes

$$\min_{\boldsymbol{\beta}} \left(\frac{1}{n} \sum_{i=1}^{n} (y_i - x_i^T \boldsymbol{\beta})^2 + \lambda ||\boldsymbol{\beta}||_2 \right) \tag{1.4}$$

This method aims to reduce the number of variables by obviating the problem of multicollinearity; it is also used to solve and manage problems in the literature known as *Ill -Posed*, where $||\beta||_2 = \sum_j \beta^2$ is deriving from the constraint $\sum_j^n \beta^2 \leq t$.

Regarding the importance of variables in regression, the absolute value of the test statistic $t_{\hat{\beta}}$ is used, defined as:

$$t_{\hat{\beta}_j} = \frac{\hat{\beta}_j}{\sigma_{\hat{\beta}}} \tag{1.5}$$

where $\sigma_{\hat{\beta}}$ is the standard error of the estimator. The importance of a variable therefore increases with increasing weight and the more high is the variation in the estimated weight and the less important the features is for the model. The estimated coefficients increase the value of the predicted variable y by one unit of measurement, therefore the interpretation of linear regression models is very simple. In the explanation of the models, an important measure of how much variance is explained by the model is the index R^2.

1.5.2 Decision trees

Decision trees can be applied to both regression and classification problems. Let's start by illustrating the problems related to regression and discuss the process of building a tree which takes place through two main phases:

1 The feature space is divided, that is, the set of possible values for $X_1, X_2, ..., X_k$: in S distinct and non-overlapping regions, $r_1, r_2, ..., r_s$.

2 For each sample i that falls in the r_s region, we calculate the average of the response values for the values of the training set r_j.

3 The goal is to find regions $r_1, ..., r_s$ that minimize RSS, or:

$$\min_r \sum_{s=1}^{S} \sum_{i \in r_s} (y_i - \hat{y}_{r_s})^2 \tag{1.6}$$

where $\hat{y}_{r_s}$ is the average response for training observations within the s-th region. The problem is combinatorial, therefore it is onerous to calculate all the possible partitions of the space of the features in s-regions, therefore the function is optimized through a heuristic method of type *greedy*, called "recursive binary division". The approach starts from the top down because it starts at the top of the tree (at that point all observations belong to a single region) and then subdivide subsequently the space of the features: each division is indicated by two new branches (binary) lower on the tree. It's *greedy* because at each (recursive) iteration of the tree-building algorithm, the best division is done in that particular step. To perform the recursive binary division, we first select the predictor X_k and the cutoff point m such as to divide the space of the input variables in the regions $\{X : X_k \geq m\}$ and $\{X : X_k \leq m\}$ leading to the reduction of RSS. Considering everything the space $X = \{X_1, ..., X_k\}$ and all the possible values of the cutting points for each of the predictors, therefore the predictor and the cutting point are chosen so that the resulting tree has the Lower RSS. The process described can produce good predictions but the problem of data overfitting is likely to occur, leading to poor performance in the testing phase. This is because the resulting tree may be too complex. A tree with fewer divisions (i.e. fewer regions $r_1, ..., r_s$) could lead to a decline of variance and better interpretation at the expense of a small bias. A possible solution to the overfitting problem is to build the tree until the reduction of the RSS due to each division exceeds a certain (high) threshold.

We now describe classification trees, very similar to a regression tree, except that it is a classification method used to predict a qualitative rather than quantitative response. For a classification tree, the prediction concerns the observation belonging to the most frequent training class in the region to which it belongs. In interpreting the results of a classification tree, we are interested not only in class prediction corresponding to a particular region of the terminal node, but also in the class proportions between training observations falling within that region. The task of growing a classification tree is quite similar to the task of grow a regression tree. Just like in the regression setting, we use recursive binary division to grow a classification tree. However, in classification, RSS cannot be used as a criterion for carrying out binary divisions and the classification error rate is therefore used. The goal of classification is to assign an observation in a certain region with respect to the most frequent error rate in the set of training data in a specific region; the classification

error rate is equal to the fraction of the training observations in that region that do not belong to the most common class, in formulas we can write:

$$E = 1 - \max_k \hat{p}_{sk} \tag{1.7}$$

Where $\hat{p}_{sk}$ represents the proportion of observations in the training dataset in the s-th region which belongs to the k class. In this case, however, the classification error is not sufficiently sensitive for the growth of trees, therefore there are some measures preferable to the simple classification error rate, starting from this, the Gini index is defined by:

$$G = \sum_{k=1}^{K} \hat{p}_{sk}(1 - \hat{p}_{sk}) \tag{1.8}$$

which represents a measure of the total variance between the k-classes. The Gini index is used as a measure of node purity: a small value indicates that a node contains predominantly observations from a single class.

Another measure derived from the classification error rate is the well-known entropy index, defined as:

$$e = -\sum_{k=1}^{K} \hat{p}_{sk} \cdot log(\hat{p}_{sk}) \tag{1.9}$$

it is shown that entropy will assume a value close to zero if the values of $\hat{p}_{sk}$ are all close zero or close to one. Therefore, like the G index, entropy will take on a small value if the node is pure. The interpretation of a decision tree is very simple, therefore widely used model even if it falls within the trade-off paradigm between complexity and interpretability. On average, predictions are less accurate than more sophisticated methods such as neural networks; its simplicity lies in the fact that starting from the main node (father), we move on to the next nodes and the edges indicate which subsets are being displayed. Once a leaf node is reached, this node returns the expected result. Its simplicity and power at the same time is in the explanation and importance of the variables, the importance is calculated through all the subdivisions of the space of the features for which the variable was used, we calculate how much the entropy or the Gini index has reduced compared to the parent node of all importations normalized between 0 and 100. Therefore the importance can be interpreted as an overall share of the model. Each prediction of a decision tree (of classification) can be explained by the decision decomposition in one component for each variable.

1.5.3 Artificial neural networks

According to Molnar's work [27] deep learning (DL) has had a remarkable use especially in the field of image recognition, computer vision, image classification and Natural Language Processing problems. To use a neural network, data input is passed through several layers of multiplication with the weights learned and through

nonlinear transformations. A single prediction can involve millions of mathematical operations depending on the architecture of the neural network. We should consider millions of weights interacting complexly to understand a prediction from a neural network. To interpret the behavior and predictions of neural networks, we need specific interpretation methods. The methods mainly used for explainability are those illustrated in the following paragraphs relating to methods for nonlinear models. Molnar also introduces concepts and methods for the explainability of deep neural networks, such as Feature Visualization, which is basically a mathematical optimization problem that aims to find the input that maximizes the activation of this unit, in formulas defined as follows:

$$img^* = \arg\max_{img} \sum_{x,y} h_{n,x,y,z}(img) \tag{1.10}$$

the previous problem represents the search for a new image that maximizes the activation (average) of a unit in this case a single neuron. Another approach to this type of problem, in the case of deep networks, concerns the "Dissection of the network", again presented in the work of Molnar [27] originating in the work of Bau *et. al.* [28], where in this case the interpretation of a unit of a convolutional neural network is quantified. The resolution algorithm has three main steps, set out below:

1 Identify a broad set of human-labeled visual concepts

2 Gather hidden variables 'response to known concepts

3 Quantify alignment of hidden variable - concept pairs

For a more details of the method see the work [28].

1.5.4 Optimization based methods

In Bertsimas *et. al.* [16] the concept of interpretability is formalized through a mathematical optimization problem in which given an interpretability function $L(m)$ for all paths $m \in P$, a path m is considered more interpretable than another path m' if and only if $L(m) \leq L(m')$. By setting an interpretability level l, the minimum constraint problem becomes the following:

$$\begin{aligned} \min_{m \in P} \quad & c(m) \\ \text{s.t.} \quad & L(m) \leq l \end{aligned} \tag{1.11}$$

This problem becomes a problem on the Pareto frontier as a trade off between interpretability and precision. By solving this problem for each l, the authors produce the "price" of the interpretability of the model given a class of candidate models. Defining the complexity of the model as L-complexity

$$\begin{aligned} L_{complexity}(m) = \min \quad & |m| \\ & m \in P(m) \end{aligned} \tag{1.12}$$

for $l = K$ the (1.2) become

$$\min_{m \in P_K} \quad c(m_K) \tag{1.13}$$

Bertimas *et al.* generalize the problem of interpretability for problems already known in the literature, ($L0$-constrained sparse regression) for linear models, listed by best classification of a data dimension or the k problem means to find the k cluster best chance. Once these elements are defined, the authors arrive at the formulation of a multi-objective problem defined in the following way

$$\min_{K \geq 0} \quad \left(\min_{m \in P_k} c(m_K) + \lambda \cdot L_\alpha(m) \right) \tag{1.14}$$

where $\lambda \in R$, is the trade off parameter between interpretability and cost, in which $L_\alpha = \sum_{K=1} \alpha_k \cdot c(m_K)$, different K and α can be calculated different pareto-excellent solutions.

1.5.5 Methods for nonlinear problems

Montavon *et al.* [17] presents a very interesting method for the interpretation of nonlinear classification models such as deep neural networks, described on a problem decomposition approach, specifically the authors define the problem as "Deep Taylor Decomposition". Intuitively, the approach starts from Taylor's series expansion for the treatment of a complex ML or DL problem. The method proposed by the authors is based on the known "backward propagation" method, calculating the contribution of each individual input within the network starting from the output obtained. As part of the computer vision or image recognition algorithms, the classifier should not only indicate whether the image is included in a certain category but should also indicate which inputs (i.e. pixels) they included in the decision. Below we show the idea of the decomposition method shown in the work of Montovan *et al.* [18]. Formally a simple $ReLU$ output neuron of the type

$$x_j = \max(0, \sum_i x_i \cdot w_i + b_i) \tag{1.15}$$

has been defined, a real-value input vector with $b_i < 0$. It is possible to note that in a subset of the input space the neuronal function is linear (therefore in terms of 'local' decomposition), in this subset it is therefore possible to write the outuput as an expansion in Taylor series in the first order, i.e.

$$x_j = \sum_i \frac{\Delta x_j}{\Delta x_i}|_{(x_i)_i = (\tilde{x})_i} \cdot (x - \tilde{x}) \tag{1.16}$$

where $(\tilde{x}_i)_i$ is called root point of the active set. Starting from the sum elements, the decomposition of the output on the input variables can be written as:

$$[x_j]_i = \frac{\Delta x_j}{\Delta x_i}|_{(x_i)_i = (\tilde{x})_i} \cdot (x - \tilde{x}) \tag{1.17}$$

preserving the conservation property through the following constraint

$$\sum_i [x_j]_i = x_j \tag{1.18}$$

Assuming that for all neurons $(x_j)_j$ whose contribution is x_i, we can write $[x_f]_i = x_j \cdot c_j$ product of an activation neuron and a constant, we show that it holds for every $[x_f]_i$, we have

$$c_j = \sum_j \frac{w_{ij}^+ [x_f]_j}{\sum_i x_i \cdot w_{ij}} \tag{1.19}$$

the decomposition $([x_f]_p)_p$ on the input variables $(x_p)_p$ can be easily calculated by applying these propagation rules with a step backwards. The redistribution rule for each neuron is always conservative, therefore we can say that even the decomposition for the whole network is conservative and we get $\sum_p [x_f]_p = x_f$.

1.5.6 Other methods

In addition to these two very recent methods proposed by Bertsimas and Montavon, there are other methods used in the problems of interpretability, among them:

LIME Ribeiro *et. al.* [6] in this regard, introduces the concept of trade off between interpretability and loyalty LIME (Local Interpretable Model-Agnostic Explanations) formalized through the following optimization problem:

$$\min_{g \in G} L(f, g, \pi_x) + \Omega(g) \tag{1.20}$$

where $\Omega(g)$ can be defined as a measure of complexity (as opposed to interpretability) of the model g, for example the number of parameters, or the depth of a tree in the case g is a Classification Trees, or for a linear model the number of non-zero weights, for example in the Lasso - Ridge approach. So a model g, belonging to the wider class of models G, minimizes the L, which is a loss function which measures the infidelity of the model considering the proximity measure π_x. Infidelity is defined by the authors as "the predictive behavior of the model near the instance to be predicted", therefore a discrepancy between what is expected and what is predicted.

Partial Dependence Plot In Friedman's work [19] some methods for the interpretation of models are presented. PDP is focuses on visualization, one of the most powerful interpretative tools and the display is limited to small topics. Functions of a single variable with real value can be plotted as a graph of the values of $\hat{F}(x)$ against each corresponding value of x. The functions of a single categorical variable can be represented by a bar chart, each bar represents one of its values and the bar height the value of the function. Viewing functions of higher-dimensional topics is more difficult. Is therefore useful to be able to visualize the partial dependence of the approximation $\hat{F}(x)$ on small selected subsets of the input variables. The

functional form of $\hat{F}$ depends on the chosen values of the input subset z_l, if the dependency is not very strong the expected value of $\hat{F}(x)$, that is $\mathrm{E}[\hat{F}(x)]$ can represent a good synthesis of the partial dependence of the chosen variables of the subset z_l, a value such that $z_l \cup z_i = x$ where z_l is the complement subset of size l and z_i is a chosen target subset. Dependencies can be different, as additive or multiplicative, for example in classification problems the author suggests that partial dependence diagrams of each $\hat{F}_k(x)$ on subsets of variables z_l most relevant for a given class provide information on how input variables affect the respective probabilities of individual classes.

Individual Condition Expectation ICE [20] is a tool to visualize the model estimated by any supervised learning algorithm. While the PDP helps to visualize the partial average relationship between the estimated response and one or more features, in the presence of substantial interaction effects, the partial response relationship can be heterogeneous, therefore an average like the PDP, can blur the complexity of the relationship modeled, instead the ICE improves the partial dependence diagram by graphically representing the functional relationship between the expected response and the characteristic for the individual observations. In particular, the ICE graphs show the variation of the values adapted in the range of a variable suggesting where and to what extent heterogeneity can exist.

Accumulated Local Effects Plot Compared to PDP, which is the most popular approach to visualizing the effects of predictors with supervised learning models with black box, which produces erroneous results if predictors are strongly correlated, since the extrapolation of the response to predictive values that are far outside the multivariate endowment of the training data is required, the Accumulated Local Effects (ALE) [21] does not require this unreliable extrapolation with related predictors, therefore the ALE method is substantially less computationally expensive than PDPs, which only requires $2^{|J|} \times n$ supervised learning model evaluations $f(x)$ to calculate each $\hat{f}(x_J)_{ALE}$, compared to $K^{|J|} \times n$ evaluations to calculate each model $\hat{f}(x_J)_{PDP}$.

Feature Interaction Starting from his work on the PDP method, Friedman *et. al* presents another method, called Feature Interaction [22] which assumes that a function $F(x)$ has an interaction between two of its variables x_j and x_k if the difference in the value of $F(x)$ as a result of changing the value of x_j depends on the value of x_k. Such an assumption can be formalized as

$$\mathrm{E}_x \left(\frac{\partial^2 F(x)}{\partial x_j \partial x_k} \right)^2 > 0$$

or by an analogous expression for categorical variables implying finite differences. If there is no interaction between these variables, the function $F(x)$ it can be expressed as the sum of two functions, that is $F(x) = f_j(x_j) + f_k(x_k)$ one of which does not depend on x_j and the other independent of x_k.

Shapley Value Among the important works to refer to we mention the Shapley Values [23], an innovative method in which an additive method assesses the importance of variables through the expected conditional value of the original model, we mention the work of Koh and Liang [24] in which the authors measure the importance of the variables through the Influence Function, i.e. starting from the minimization of a risk function of the following type $R(\theta) = \frac{1}{n}\sum_i L(z_i, \theta)$. For a more detailed discussion, from which various components of this chapter have been extracted, please refer to the excellent work of the authors [25].

1.5.7 Interpretable unsupervised learning

In the field of interpretability, even in unsupervised models such as clustering, several papers have proposed methods of interpreting results and characteristics. Bertsimas *et. al.* [29] propose an unsupervised interpretability method, based on optimization formulated as a mixed integer programming problem, through the generation of decision trees, their method approximates the optimal solution on a global level that partitions the feature space. The algorithm is optimized using a procedure called coordinates descent and through different metrics for the evalutation of the perfomances such as the Silhouette Index and the Dunn's index. The method exploits the interpretability of decision trees used with supervised learning to explain the characteristics and logic used in cluster formation. Another method based on decision trees for interpretable clustering is proposed by the authors [30]; the approach is based on unsupervised binary trees. It is a three-step procedure, the first involves the recursive binary partition within the data in order to increase homogeneity. During the second phase, the pruning phase, the aggregation of adjacent nodes is evaluated, while in the last step, the union phase, similar clusters are joined. Also for the hierarchical clustering method the authors [31] base their method on decision trees, as flexible mathematical tools that lead to understandable explanations for decision-makers. The unsupervised decision tree is interpretable in terms of rules: the authors state that each leaf node represents a cluster and the path from the root node to a leaf node represents the rule. The decision to branch in each node of the tree is made based on the grouping trend of the data available in the node. The authors introduce four different measures to select the most appropriate attribute to use to divide the data in each branch node and also propose two algorithms for partitioning the data in each node. Many methods used in the interpretation of clustering methods are based on visual approaches, such as graphs, of the scatterplot type for example, to visualize both the groups and the errors in the partition. The authors [32] propose a visualization approach, in which is placed the objects on a grid and add a continuous topography to the background, expressing the distribution of uncertainty across all clusters. In the work of Park and Choi [33], the authors propose an interpretability method linked to Gaussian processes (GP), used in non-parametric and probabilistic modeling; in the context of an interpretable system linked to the clinical setting, a decision-maker could have difficulty in understanding the results and parameters. In this work, the authors propose a method that he uses multiple GP transition models capable of describing multimodal dynamics. They apply the method to some case studies such as air traffic control and on a flight simulator. Vitku *et. al.* [34] in the field of unsupervised reinforcement learning

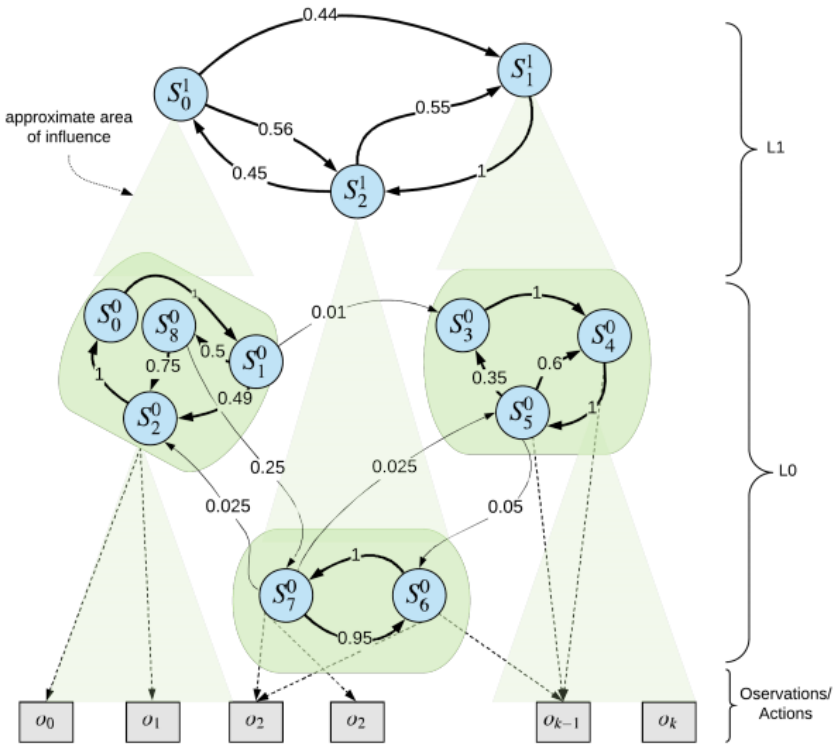

Figure 1.5. Source: [34]

propose a work based on a very simple but highly interpretable layered architecture, defined TOY, for unsupervised hierarchical problems. The representation of the data can be interpreted both in a symbolic key to the sub-symbolic one. The architecture is shown in fig. 1.5 and shows the learning ability of the method. The method is based on two main properties, (citing the authors): "the learned model is stored in the module of hierarchical representations with the following properties: 1) they are always more abstract, but can retain details when necessary, and 2) they are easy to manage in their local e symbolic-like form, thus also allowing to observe learning process at every level of abstraction". For a broader discussion of methods related to the interpretation of unsupervised problems, the work of Chen [35] is an excellent dissertation on the problem.

Bibliography

[1] Japaridze, G., and De Jongh, D. (1998) "The logic of provability" in Buss, S., ed., Handbook of Proof Theory. North-Holland: 476–546.

[2] John W. Cain, "Mathematical Models in the Sciences", in Molecular Life Sciences, 2014

[3] Miller, Tim. "Artificial Intelligence Explained: Social Science Insights." arXiv Preprint arXiv: 1706.07269. (2017).

[4] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning, 2017.

[5] Zachary Chase Lipton. The mythos of model interpretability. CoRR,abs/1606.03490,

[6] Ribeiro, Marco Tulio, Singh, Sameer, and Guestrin, Carlos. "Why should i trust you?": Explaining the predictions of any classifier

[7] McAuley, Julian and Leskovec, Jure. Hidden factors and hidden topics: understanding rating dimensions with review text. In RecSys. ACM, 2013.

[8] Chang, Jonathan, Gerrish, Sean, Wang, Chong, BoydGraber, Jordan L, and Blei, David M. Reading tea leaves: How humans interpret topic models.

[9] Mahendran, Aravindh and Vedaldi, Andrea. Understanding deep image representations by inverting them. In CVPR, 2015.

[10] Plumb, Molitor, Talwalkar. Model Agnostic Supervised Local Explanations". arXiv preprint arXiv: 1807.02910 (2019)

[11] Jalil Kazemitabar, Arash Amini, Adam Bloniarz, and Ameet S Talwalkar. Variable importance using decision trees. In Advances in Neural Information Processing Systems, pages 425–434,2017

[12] Carvalho, D.V.; Pereira, E.M.; Cardoso, J.S. Machine Learning Interpretability: A Survey on Methods and Metrics. Electronics 2019, 8, 832.

[13] Muhammad Aurangzeb Ahmad, Ankur Teredesai, Carly Eckert. "Interpretable Machine Learning in Healthcare". IEEE International Conference on Healthcare Informatics (ICHI) 2018.

[14] A Survey on Explainable Artificial Intelligence (XAI): Towards Medical XAIErico Tjoa, Cuntai Guan, arXiv: 2019

[15] J. Townsend, T. Chaton, and J. M. Monteiro. Extracting relational explanations from deep neural networks: A survey from a neuralsymbolic perspective. IEEE Transactions on Neural Networks and Learning Systems, pages 1–15, 2019.

[16] Bertsimas, D., Delarue, A., Jaillet, P., Martin, S. The Price of Interpretability. ArXiv, abs/1907.03419, 2019.

[17] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, and K.-R. Muller. Explaining nonlinear classification decisions with deep Taylor decomposition. Pattern Recognition, 65:211–222, 2017.

[18] Montavon, Grégoire Lapuschkin, Sebastian Binder, Alexander Samek, Wojciech Müller, Klaus-Robert. (2016). Deep Taylor Decomposition of Neural Networks.

[19] Friedman, J.H. Greedy function approximation: A gradient boosting machine. Ann. Stat. 2001, 29, 1189–1232

[20] Goldstein, A.; Kapelner, A.; Bleich, J.; Pitkin, E. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. J. Comput. Gr. Stat. 2015, 24, 44–65.

[21] Apley, D.W. Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models. arXiv 2016, arXiv:1612.08468

[22] Friedman, J.H.; Popescu, B.E. Predictive learning via rule ensembles. Ann. Appl. Stat. 2008, 2, 916–954

[23] Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems; MIT Press: Cambridge, MA, USA, 2017; pp. 4765–4774.

[24] P.W. Koh, P. Liang. "Understanding black-box predictions via influence functions". ArXiv preprint arXiv:1703.04730, 2017

[25] B.,Arrieta, A.D.,Rodriguez, N.,Del Ser, J., Bennetot, A., Tabik, S. B., González, A., García, S., Gil-López, S., Molina, D., Benjamins, V. R., Chatila, R. H., Francisco. (2019). "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI"

[26] "An Introduction to Statistical Learning" ,Book published 2013 in Springer Texts in Statistics,Authors: Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani

[27] Molnar, Christoph. "Interpretable machine learning. A Guide for Making Black Box Models Explainable", 2019. https://christophm.github.io/interpretable-ml-book/.

[28] Bau, David, et al. "Network dissection: Quantifying interpretability of deep visual representations." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017.

[29] Bertsimas, D., Orfanoudaki, A. and Wiberg, H. "Interpretable clustering: an optimization approach". Machine Learning (2020).

[30] Fraiman, R., Ghattas, B. and Svarc, M. "Interpretable clustering using unsupervised binary trees". Advances in Data Analysis and Classification 7, 125–145 (2013)

[31] Basak, Jayanta and Krishnapuram, Raghu. (2005). "Interpretable hierarchical clustering by constructing an unsupervised decision tree". Knowledge and Data Engineering, IEEE Transactions on. 17. 121- 132.

[32] Kinkeldey, C., Korjakow, T., Benjamin, J. (2019). "Towards Supporting Interpretability of Clustering Results with Uncertainty Visualization".

[33] Park, Y.J., Choi, H.L. (2018)." InfoSSM: Interpretable Unsupervised Learning of Nonparametric State-Space Model for Multi-modal Dynamics".

[34] Vítku, Jaroslav et al. “ToyArchitecture: Unsupervised learning of interpretable models of the environment.” PLoS ONE 15 (2020)

[35] Chen, J. "Interpretable clustering methods". Thesis dissertation, August 2018, Department of Electrical and Computer Engineering Northeastern University Boston, Massachusetts

Chapter 2

Mathematical aspects of decision making

2.1 Introduction

The multi-criteria decision making process provides a very important tool for making the best decisions, when the set of alternatives is also very large and complex. Having quantitative tools available that allow you to choose the best possible alternative, among those present, is a prerogative that decision-makers must considered as fundamental. Within the Multi Criteria Decision Making Analysis (MCDMA) the weights assigned to the criteria are provided by 'experts' and built on the basis of their experience. Over the years, several works have contributed strongly to the literature related to this branch of mathematics. The multi-criteria decision making process deals with the structure and resolution of decision problems and planning related to multiple criteria considered. When considering decisions for multi-objective problems, in the concept of Pareto optimality (is not possible to improve one element without making the other worse), we find the root of the concept of dominated choices and non-dominance in the strict sense.

2.1.1 Key elements

In order to have a complete vision of the problem and the methodology, we provide some essential definitions:

Definition (Non-dominance) A feasible (alternative) solution $x \in X$ is efficient (not dominated, Pareto optimal) if and only if there is no $x \in X$ such that $f_j(x) \geq f_j(x_0)$ for every $j \in K = \{1, ..., k\}$ and $f_k(x) \neq fj(x_0)$ for at least one $j \in K$.

Definition (Dominance - discrete alternatives) Given two distinct alternatives A_i and A_k, A_i dominates A_k ($A_i \leq A_k$) if and only if $x_{ij} \leq x_{kj}$ for each $j = 1, ..., n$.

Definition (Multicriteria Decision) Choice of an action or alternative from a set of eligible alternatives carried out on the basis of two or more criteria.

Definition (Criterion) Indication on how to evaluate a type of performance measured for the different alternatives, or how the most efficient alternative to that performance should be chosen.

Definition (Attribute) Measurement of a performance of an alternative; it is a parameter; is provided in the case of discreet alternatives.

Definition (Objective) Function that measures a performance for an alternative defined as a point in the space of decision variables; used in the case of continuous alternatives.

Definition (Rule of decision) Rule used to order alternatives according to the information acquired and the preferences of the decision maker.

Two general types of decision rules:

1. Optimizing rule, in which a complete order among all possible alternatives is established (global optimum)

2. Satisficing rule, in which a satisfactory alternative is determined (excellent local)

The alternatives considered can be:

1. discrete and finished (enumerable)

2. continuous and infinite (not enumerable)

2.1.2 General steps of a decision making process

A general decision-making process consists of the following elements:

1. **Definition of the problem**

2. **Problem formulation** Specification of attributes or objectives and criteria

3. **Model building** Identification of decision variables, constraints, formalization of structural properties, use of representation techniques such as graphs

4. **Analysis, evaluation and decision** Generation of the set of eligible alternatives and estimation of attribute values or goals; collection of information on the state of nature and preferential judgments by the decision maker

5. **Implementation** of the decision and reassessment of the decision

General characteristics of a multi criteria problem

1. There are many criteria
2. There are many attributes/goals
3. Conflict between criteria
4. Incommensurability between attributes/objectives
5. Choice between a finite set of explicitly defined alternatives or a Infinite set of implicitly defined alternatives

Components of a decision-making process

1. Objectives/attributes
2. Criteria
3. Decision maker(s) and any supports for information processing
4. Decision rule

Formulation
A multi criteria problem can be mathematically formalized as follows

$$\begin{aligned} \max_{x \in X} \quad & F(X) = [f_1(x), ..., f_k(x)]^T \\ \text{s.t.} \quad & x \in X \supset \mathrm{R}^n \end{aligned} \tag{2.1}$$

where x is the vector of the decision variables, $f_j(\cdot)$, $j = 1, ...k$, is the objective j-th and X is the set of feasible alternatives.

2.2 Multi criteria decision making methods

In this section the main and most adopted methods in the context of multi-criteria decisions will be presented, it is not an exhaustive form but an overview of the methods and their applications, in order to give the reader the main elements to understand the methodologies used in decision problems. In the next paragraphs we will divide the methods into two broad classes, the first relating to the Multi attribute decision making (MADM) methods and the second relating to the Multi objective decision making (MODM) methods.

2.2.1 Analytic Hierarchy Process (AHP)

From the definition of Saaty (1980), author of the [1] method developed between 1971 and 1975, the AHP method is a general theory of measurement which is used to derive ratio scales from pair comparisons, both discrete and continuous. Paired comparisons can be made through measurements or through a fundamental scale that reflects the relative importance of preferences. The method has found its broadest applications in decision making, planning and allocation of multi criteria resources and in conflict resolution. AHP can be defined, according to the definition of R.W. Saaty [2], as "a nonlinear framework to implement both deductive and inductive thinking", all this without using *syllogism* in which several factors are taken into consideration. In particular, the AHP allows to assign priorities to a series of alternatives, reporting the assessments (both qualitative and quantitative), which otherwise would not be directly related to each other comparable, combining multidimensional scales of measures in a single priority ranking. The methodology is based on a series of pairwise comparisons between the criteria which assigns a relative importance to them, assigning a percentage weight. The sum of the weights is 1. Below are presented the main elements of the AHP method:

With A_i we define the single stimulus and a_{ij} the numerical value associated with the comparison between the i and j criteria whose number of criteria is equal to n. All comparisons in total will be $\frac{n(n-1)}{2}$, the associated matrix will therefore be a $A_{n\times n}$ matrix used to create the vector of the priorities (percentage weights) of every single criterion. Starting from the rating scale between 1 and 9, each level corresponds to the rating:

Values (a_{ij})	**Interpretation**
1	*i and j are equally important*
3	*i is slightly more important than j*
5	*i is quite more important than j*
7	*i is definitely more important than j*
9	*i is absolutely more important than j*
1/3	*i is slightly less important than j*
1/5	*i is quite less important than j*
1/7	*i is far less important than j*
1/9	*i is absolutely less important than j*

When using AHP to model a problem, you need to have a hierarchical or graph structure to represent that problem and compare a couple of relationships within the structure. In the case of discrete comparisons, they conduct a domain matrix and in the continuous case a nucleus of Fredholm operators [3], from which the ratio scales are derived in the form of main eigenvectors, or eigenfunctions, as appropriate. These matrices or kernels are positive and reciprocal, and the following relationships hold:

(a) $a_{ij} = \frac{1}{a_{ji}}$.

(b) if A_i it is of equal intensity (relative) to A_j , then $a_{ij} = a_{ji} = 1$

(c) the main diagonal of the matrix A is composed entirely of values unit

Once the A matrix of pairwise comparisons is obtained, the vector of the weights **w** (percentages) from assigning to each stimulus can be calculated simply by determining the maximum eigenvalue λ and its eigenvector **v**. Moving on to normalization, so that the sum of its elements is equal to 1, we obtain the vector of the priorities relating to the stimuli A_i, defined as follow:

$$W = \frac{\mathbf{v}_\lambda}{\sum_{i=1}^{n} \mathbf{v}(i)} \tag{2.2}$$

The vector of the weights W maintains the order of the rows of the matrix of the comparisons in pairs that you open is set by the decision maker and once the vector of the priorities has been determined it will therefore be necessary to understand whether the matrix of the comparisons a pairs is consistent, i.e. we should measure whether the subjective judgments of the decision maker in any comparison are consistent or not, about that we introduce the metrics adopted to determine the consistency of a matrix and the tolerance thresholds which are used to determine if a matrix of pairwise comparisons it may be well placed or not, so it is define the Consistency Index (CI):

$$CI = \frac{\lambda - n}{n - 1} \tag{2.3}$$

where λ is the maximum eigenvalue of the matrix A. Setting the Random Consistency Index (RI) in which the relative value of RI is associated with the size of the matrix A, from following table:

RI	0.0	0.0	0.58	0.9	1.12	1.24	1.32	1.41	1.45	1.49
n	1	2	3	4	5	6	7	8	9	10

we can finally define the last evaluation metric, the Consistency Ratio through the simple ratio:

$$CR = \frac{CI}{RI} \tag{2.4}$$

2.2.2 PROMETHEE Methods

The methods called PROMETHEE I and PROMETHEE II were developed by J.P. Brans [5] in 1982 a few years after J.P. Brans and B. Mareschal developed PROMETHEE III and PROMETHEE IV. The same authors in support of the

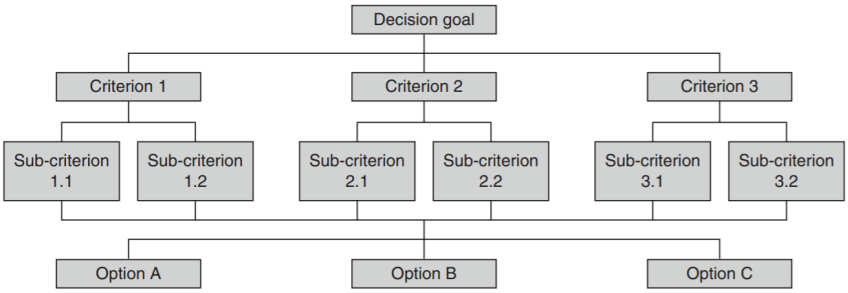

Figure 2.1. AHP structure, Source: [27]

models mentioned, have also developed a visualization tool, called GAIA. These methods have seen numerous applications and in various sectors such as banking, industrial, workforce planning, water resources, investments, medicine and others. Considering the following multicriteria problem:

$$\max\{g_1(a), g_2(a), g_3(a), ..., g_k(a) | a \in A\} \tag{2.5}$$

where A is a finite set of alternatives a_j, for $j = 1, ..., n$ and $g_i(\cdot)$, for $i = 1, ..., k$ a set of evaluation criteria, both maximization and minimization are possible. The decision maker's expectation is to identify an all-criteria optimization alternative. The following problem is part of the aforementioned (subsection *1.5.1*) *ill - posed* problems, as there is no point alternative optimizing all criteria simultaneously. The solution of a multicriteria problem therefore does not depend only on the starting data represented in the evaluation table but also by the decision-maker himself. The best (compromise) solution also depends on individual preferences and additional information is therefore required. The PROMETHEE method is a method belonging to the family of outranking methods and aims to define relationships according to which the comparisons between alternatives will be made, this occurs through the preference functions. These functions represent an intensity of preference between two different alternatives; taking two alternatives a and b we define these relations:

(a) a is preferred to b if and only if $g(a) \geq g(b)$

(b) a is indifferent to b if only if $g(a) = g(b)$

An alternative is therefore considered preferred to another if the difference between the two alternatives for the function considered is enough great, in formulas we can write: $g(b)$ - $g(a) \geq \epsilon$, where ϵ is a positive parameter. The importance of each individual criterion is given by the weight assigned with respect to the weights attributed to the other criteria. Weights are established prior to decision maker in the case of domain mastery otherwise they can also be provided by different sources. These weights w_j have the following properties:

(a) $w_j > 0$

(b) $\sum_j w_j = 1$

Each criterion has its own preference function, which assumes values between 0 and 1; the parameters are chosen a priori by the decision maker on the basis to the problem. Defined a generic preference function, $g(\cdot)$, for a generic c criterion. Further defining for convenience $g(b) - g(a) = x$, we show the main functions used, in according to the definition of Brans and Vincke [6], we have:

Usual Criterion

$$d(x) = \begin{cases} 1 & x \geq 0 \\ 0 & x \leq 0 \end{cases} \tag{2.6}$$

This type of function has an area of indifference if and only if $g(a) = g(b)$, outside the decision maker will have a preference for only one of the two alternatives. The area of indifference is already determined and there is therefore no need to choose parameters.

Quasi-criterion

$$d(x)_\gamma = \begin{cases} 1 & x \geq \gamma \\ 0 & x \leq \gamma \end{cases} \tag{2.7}$$

The function $g(\cdot)$ has an area of indifference defined by the parameter γ. In the interval $(-\gamma, \gamma)$ the decision maker finds the two alternatives indifferent. While outside of this range, one of the alternatives is strictly preferred. The decision maker will have to choose the value of the parameter γ.

Criterion with linear preference

$$d(x)_z = \begin{cases} 1 & x \geq z \\ \frac{x}{z} & x \leq z \end{cases} \tag{2.8}$$

The decision maker, if the preference for a given alternative increases progressively, can use this type of function, with z defined parameter of *progression*, and this parameter to be defined is z which defines the speed with which the indifference zone grows up to 1.

Gaussian criterion

$$d(x)_g = \begin{cases} 1 - e^{-x^2/2\sigma^2} & x \geq 0 \\ 0 & x \leq 0 \end{cases} \tag{2.9}$$

This function derives from the well-known function used in statistics and the only parameter to be determined is σ.

Comparison of alternatives
After choosing the functions to be used for the comparison between the alternatives, it will be necessary to establish a ranking. So it is define a preference index. We will therefore have:

$$\phi(a,b) = \frac{1}{K}\sum_{h=1}^{K} D_h(a,b) \tag{2.10}$$

$\phi(\cdot)$ therefore indicates how preferred the a alternative to the b alternative compared to all the h-criteria. If the decision maker considers it appropriate to weigh the criteria differently from the final decision we should introduce another criterion, defined as follows:

$$\phi(a,b)_w = \frac{1}{K}\sum_{h=1}^{K} D_h(a,b)\cdot w_h \tag{2.11}$$

Once the preference for each of the alternatives over a specific one has been defined h criterion, the decision maker for all criteria will have to calculate a ranking of the currencies alternatives, in order to define the *outranking flows*, given by the following formula:

$$\eta(a)^+ = \frac{1}{N-1}\sum_{x\in K} \phi(a,x) \tag{2.12}$$

and

$$\eta(a)^- = \frac{1}{N-1}\sum_{x\in K} \phi(a,x) \tag{2.13}$$

Where $\eta(\cdot)^+$ is positive outranking, respectively, and $\eta(\cdot)^-$ is negative. Then there are (N-1) alternatives with which we compared the alternative *to* and K is the space of alternatives, while x instead represents the *deviation* of the specification preference function $g(\cdot)$ for a over the same preference function for the other alternatives.

2.2.3 ELECTRE methods

ELECTRE (**EL**imination **E**t **C**hoix **T**raduisant la **RE**alité) methods are a series of methods developed by Bernard Roy [7] since the 1960s, born from the need to solve complex problems in the corporate application field. Over the years, several methods belonging to this family of methodologies have been developed, giving rise to the French school of Decision Making. The method tries to build on the A set of alternatives a more complete R outclassing relationship than basic dominance. The global preference model admits

(a) incomparability between alternatives

(b) non-transitivity

The incomparability therefore results in certain situations in which they do not exist sufficient information to establish whether a situation of clear preference turns out to be $a_i R a_j$ or $a_j R a_i$. Several *Electre* methods have been proposed that depend strictly on the decision problem:

(a) selection of preferable alternatives

(b) sorting of alternatives

(c) classification of alternatives

The common basis is the construction of the outclassing relationship from comparisons in pairs and by defining a pair of alternatives a_i, a_j we can define the following sets:

(a) $I^+(a,b) = \{k \in I : g_k(a_i) \geq g_k(a_j)\}$ $s.t$ $W^+(i,j) = \sum_k w_k, k \in I^+$

(b) $I^-(a,b) = \{k \in I : g_k(a_i) \leq g_k(a_j)\}$ $s.t$ $W^-(i,j) = \sum_k w_k, k \in I^-$

To implement the method there are some steps to be performed, firstly the construction of the outclassing relationships (R) then the tracing of a graph on R, and the identification of a subset of the alternatives (N) defined *kernel*. Once the first step has been carried out, in order to build the outclassing relationship, the concordance test and the discordance test are carried out. The concordance test is a concordance index for each pair (a_i, a_j), where $c(a_i R a_j)$ expresses the advantage that the decision maker has in choosing a_i rather than a_j. This index is defined as:

$$c(a_i R a_j) = \frac{W^+}{W^+ + W^-} \tag{2.14}$$

One of the problems for which all subsequent versions of ELECTRE have been developed is that of ranking, for an exhaustive discussion of all the methods see the work of Figueira *et al.* [8].

2.2.4 TOPSIS

The TOPSIS model is a multi-criteria decision analysis method developed by Ching-Lai Hwang and Yoon with further developments by Yoon between [9], [10] 1981 and 1987 and subsequently by Hwang, Lai and Liu in 1993 [11]. TOPSIS is based on the concept that the alternative chosen should have the smallest geometric distance from the ideal positive solution and the longest from the negative ideal solution. The method compares a set of alternatives by identifying the weights for each criterion normalizing the scores for each criterion and calculating the geometric distance between each alternative and the ideal alternative, i.e. the best score for each criterion. The criteria in the TOPSIS method increase and decrease monotonically. The different scale units of alternatives and criteria lead to the normalization of

values in order to standardize the analysis. The compensatory methods therefore allow to exchange the criteria, in which a poor result can be compensated by a good result in another criterion. In the classic TOPSIS method it is assumed that the ratings of the alternatives and the weights are represented by numerical data and the problem is solved by a single decision maker. The problem arises when there are multiple decision makers as the optimal (preferred) solution must be agreed upon by the interest of the group usually with different objectives. The TOPSIS model is articulated through a series of steps described below:

Step 1. Decision matrix and weights
Define the decision matrix $X_{i,j}$ and the vector of the weights $\tilde{W}$ such that $\tilde{W} \cdot \tilde{W}^T = 1$ and the function criteria can be either benefit functions or cost functions.

Step 2. Normalized decision matrix
The normalization of the values can be carried out through one of the well-known standardization formulas.

$$x_{s1} = \frac{x_{ij}}{\sqrt{\sum_{ij} x^2}} \tag{2.15}$$

$$x_{s2} = \frac{x_{ij}}{max_i \; x_{ij}} \tag{2.16}$$

for $i = 1, ..., m$ and $j = 1, ..., n$

Step 3. Weighted normalized decision matrix
The matrix normalized is obtained by $v_{ij} = w_j \cdot x_{ij}$, for $i = 1, ..., m$ and $j = 1, ..., n$, s.t. $\sum_j w_j = 1$

Step 4. Positive ideal and negative ideal solutions
This step identifies the ideal positive alternative and the ideal negative alternative is identified. The ideal positive solution is the solution that maximizes the benefit criteria and *minimizes* the cost criteria while the ideal negative solution *maximizes* the cost criteria and minimizes the benefit criteria.

Step 5. Separation measures
In this step of the TOPSIS method it is possible to apply a series of distance metrics in order to obtain the separation of each alternative from the positive ideal solution.

Step 6. Relative closeness to the positive ideal solution
The relative closeness of the i-th alternative a_j with respect to a^+ is defined by

$$R_i = \frac{d_i^-}{d_i^+ + d_i^-} \tag{2.17}$$

where $d_i^+ = \sqrt{\sum_j v_{ij} - v_j^+}$ and $d_i^- = \sqrt{\sum_j v_{ij} - v_j^-}$, $R_i \in [0, 1]$

Step 7. Rank the preference
In this last step a set of alternatives now can be ranked by the descending order of the value of R_i and order or select the alternative closest to 1.

There are several extensions of the method with determination of the attributed values as an interval for a single decision maker and a method with certain attributed values as intervals for group decision making. Another version concerns the quantitative and qualitative criteria in the TOPSIS method with expressed weights by linguistic variable, for further details, see the works [12], [13], [14].

In the work of Pirdashti *et al.* [15] we find an exhaustive taxonomy of the MCDM methods in which we find the already examined AHP, ELECTRE, PROMETHEE and TOPSIS; furthermore, Multi-objective decision-making methods (MODM) are explored as most real-life decision problems involve multiple and conflicting goals; this method is used to solve this type of problem characterized by multiple objective functions to be maximize and minimize at the same time [16], [17]. A MODM model considers a vector of decision variables, objective functions and different constraints the objective remains that of optimize the objective functions while the decision maker chooses a solution from a series of efficient solutions since MODM problems rarely have a single optimal solution [18]. The main problem remains that related to the identification or approximation of a set of points known as the Pareto-optimal border. Several approaches have been proposed for solving these problems, simple approaches that require very little information, or methods based on mathematical programming techniques. Most multi-objective methods fall into two macro categories: those that use mathematical programming techniques and those that use evolutionary algorithms (EA)

2.2.5 Multi-objective mathematical programming

In this method, a set of linear functions are optimized with respect to a series of linear constraints. If at least one objective function or constraint is not linear, we get a multi-objective nonlinear programming problem. The resolution of these problems occurs through a priori methods, interactive methods and a posteriori methods [19]. When using an a priorir method the decision maker expresses his preferences before the solution process while in interactive methods, the dialogue phase with the decision maker it is exchanged with the calculation phase and the process usually converges after some iterations to the optimal solution. Instead with regard to a posteriori methods, efficiency of the solutions it is generated that the decision maker intervenes and selects the best solution.

2.2.6 Goal programming

In according with *et al.*[15], the goal programing was first proposed in the 1950s by Charnes *et al.* [20] and is a method that aims to solve problems where there are mainly conflicts between decision makers and also used when multiple attributes are assigned. The main objective of this methodology is to minimize the errors

committed in the failure to achieve the objectives. Within this class of methods we find both the deterministic and the stochastic part in the parameters; while for decision variables they can be integer, continuous and mixed, just as objective functions and constraints can be both linear and nonlinear. For further information, see the main works [21].

2.2.7 Evolutionary algorithms

In this wide class of multi-objective programming problems, the search for the global optimum is a difficult task, since the space of solutions is complex and given the nature of the number of objectives often a factor of conflict also comes into play. To solve this type of problem several methods have been developed including the Evolutionary Algorithms (EA) [22]. EAs work on one population of potential solutions based on two principles: selection and variation. This method belongs to the class of metaheuristic algorithms which produce lower level heuristic procedures that can be performed to perform one partial search. It is applicable to various optimization problems with limited computing capacity and insufficient imperfect information. In this situations, these methods provide adequate solutions.

2.2.8 Genetic algorithm

This method of local research has proven to be the most popular over time in the field of mathematical optimization as the research technique of the excellent is based on the principles of genetic evolution and natural selection. It is a probabilistic research method that uses research techniques inspired by the works of the biologist Darwin regarding the evolutionary theory of natural selection and species survival. The method was developed by Holland [23] in the mid-1970s and later popularized by Goldberg in 1989 [24]. The method is based on a random but direct search to find the optimal global solution without the objective function being derivable. Furthermore, the search is not distorted towards any locally optimal solution. Goldberg has shown that the method can computationally solve very large combinatorial problems.

2.2.9 Simulated annealing

This method is also part of the Evolutionary Algorithm and the Genetic Algorithm of the class of probabilistic meta-heuristic models used to find the global optimum [25]. The name of the method draws inspiration from the metallurgical field, a technique that involves controlled heating and cooling of materials to increase the size of its crystals and reduce their defects. For each step, the algorithm replaces the current solution with a "random" near point and this point is chosen with a probability that depends both on the difference between the values of the corresponding function and also on a certain parameter (called temperature), which is gradually reduced during the iteration. The dependence is such that the current solution changes randomly if the temperature is high, but decreases more and more if the temperature moves towards zero.

2.2.10 Tabu Search

The Tabu Search [26] is an algorithm also meta-heuristic that belongs to the class of local search techniques which is also nourished to solve combinatorial optimization problems when the space of solutions is very complex; the method improves the execution of a local search method using memory structures. When a candidate solution has been determined, it is marked as "tabu" so that the algorithm does not visit that solution again. The algorithm uses a procedure to iteratively move from one solution to another until some stopping criteria are satisfied.

2.3 Learning formulated as optimization problems

Mathematical optimization represents a fundamental element in machine learning and deep learning problems. The use of these techniques in recent years has seen an exponential growth, in different fields of application, from clinical to industrial, the most used artificial intelligence systems are all based on the same concept, learning by minimizing error; this error is determined through the minimization of an objective function, defined cost function or loss function. From supervised to unsupervised methods, all of these techniques use linear and nonlinear methods in order to instruct a program to perform a certain task. In the first chapter of this work we introduced about the interpretability and transparency of the algorithms used as well as the ethical problems associated with these algorithms. The transparency of black-boxes derives from the fact that very often, as seen in the previous chapter, many algorithms do not have optimal global solutions and many solution methods are based on approximations, or as defined in the literature, heuristic or metaeuristic methods. Therefore it is difficult to interpret a solution that was first obtained heuristically and secondly in a translation of reality to mathematical formalization in a non-linear way. Over the years, various techniques have been developed, in this dedicated chapter these techniques and the main supervised and unsupervised methods will be discussed in order to have a general vision. The optimization techniques over time have made important contributions to the development of new machine learning and deep learning algorithms, the main techniques from which to start can be divided into three large macro areas: first order optimization methods that work on the gradient of the objective function, like the well-known Gradient Descent [28]. We find also we the higher-order methods of which Newton's method is one of the best known and finally we have the class of so-called methods free-derivative or heuristics. For the first category the evolutions are also very well known, such as the descent of the stochastic gradient for example [29]. Compared to first order optimization methods, higher order ones have a better convergence in terms of computational performances [30]. One of the main problems in these higher-order methods is that related to the inverse of the Hessian matrix and over time several works have proposed solutions to overcome the problem [31] using approximations of the Newton method. We will see below that therefore the problems of machine learning and deep learning can be treated and therefore solved as problems of mathematical optimization. The nature of the ML problem to be solved will lead to a different optimization problem from case to case.

2.3.1 Supervised problems

In supervised learning given a pair of examples $(y_1, x_1), (y_2, x_2), ..., (y_k, x_k)$ the goal is to find a function $f(x_1, x_2, ..., x_k)$ that minimizes the loss function:

$$\frac{1}{N}\sum_{i=1}^{N} l(y^i, f(x^i)) \tag{2.18}$$

where N is the size of training samples, x_i is the vector of the variables (or also called features) of the sample, y_i is the dependent variable (or *target*) which is the corresponding label and l is the loss of function. What makes the type of learning different is the fact that in the supervised we have a target object of study; then depending on the type of problem we have to face within the supervised field, we have different types of loss functions. The functional form can be quadratic, as occurs with linear regression in which loss is nothing more than squared deviation, we saw in the first chapter, that from the classic problem of estimating parameters in the regression model, we can introduce constraints which lead to forms of the type (1.3 and 1.4) models known as $LASSO$ and $RIDGE$. There are other types of loss functions that are widely used, such as in the case of classification problems, in which the variable y_i can take either binary values or multiclass values, in this case we use functions such as Cross-Entropy or Hinge Loss or Generalized smooth hinge loss, Tangent Loss and Savage Loss to name a few of the most used.

2.3.2 Semi-Supervised problems

As for Semi-Supervised learning, it can be seen as a combination of supervised and unsupervised learning in which a part of the data is labeled, while a part is not. Therefore we will have $X_l = \{(x_1, y_1), ..., (x_k, y_k))\}$ and $X_u = \{x_{k+1}, ..., x_N\}$, that is, a pair of features and targets X_l is a set of only features X_u with $N = k + m$. Historically Vapnik can be considered one of the founding fathers of this type of problem [32], in the 70s he introduced the concept of *transductive learning* in which he inferred on the values of the target variable for unlabelled features, while in *inductive learning* the aim is to map the functional form $f(\cdot)$; in this regard, the authors [33] introduces assumptions so that unlabeled data can be used and at least one of the following hypotheses must be satisfied:

(a) Continuity - Points that are close to each other are more likely to share a label

(b) Cluster - The data tend to form discrete clusters and points in the same cluster are more likely to share a label

(c) Manifold - The data lie approximately on a manifold of much lower dimension than the input space

Among the methods used mainly for this class of problems we have: Generative Models, Low-density separation, Graph-based methods and Heuristic approaches. In order to provide an explanatory example, as well known in the literature we present a classic SVM problem:

$$\begin{aligned} \min_{x,w,\xi} \ & \frac{1}{2}||w||^2 + C\sum_{i=1}\xi_i \\ \text{s.t. } & y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i \\ & \xi_i \geq 0, i = 1, ...N \end{aligned} \tag{2.19}$$

Using the information of unlabeled data, it's need add further constraints on the unlabeled data to the original objective of SVM with slack variables ξ_i. Now define ϵ_j as the misclassification error of the unlabeled samples; if its true label is positive and η_j as the misclassification error of the unlabeled samples. Obtaining the following formulation for the S^3VM (Semi-Supervised-Support Vector Machines) in sense of Bennet and Demiriz [45]:

$$\begin{aligned} \min_{x,w,\xi} \ & \frac{1}{2}||w||^2 + C\left[\sum_{i=1}\xi_i + \sum_j \min(\epsilon_j, \eta_j)\right] \\ \text{s.t. } & y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i \\ & \xi_i \geq 0, i = 1, ...N \\ & w^T x_i + b + \epsilon_j \geq 1 \\ & -(w^T x_i + b) + \eta_j \geq 1 \\ & \eta_j, \epsilon_j \geq 0, j = N+1, .., m \end{aligned} \tag{2.20}$$

where C is defined as a penalty coefficient.

2.3.3 Unsupervised problems

In unsupervised problems as mentioned in the previous subsection, there is a set of features $x_1, ..., x_k$, and there is no *target* variable. The algorithms of this learning class are based on the well-known techniques of *Clustering* which aim to divide the dataset into specific groups. The best-known algorithm is the k - means, which divides the instances present in the data into k distinct groups. In terms of mathematical optimization the problem can be formulated by minimizing the following function:

$$\min_{M} \sum_{k=1}\sum_{x\in M_k} ||x - \mu_k||_2^2 \tag{2.21}$$

where K is the number of clusters, x is the features vector of samples, μ_k is the center of cluster k (centroids) and M_k is the sample set of the k-th cluster.

Other techniques based on machine learning are the known PCA (Principal Component Analysis) which is used in the context of dimensional reduction; as well as another non-linear method [35] based on unsupervised neural networks or *Autoencoders* which is also a method based on the minimization of a loss function in order to minimize the reconstruction error between the input vector x and the estimated output $\hat{x}$.

2.3.4 Reinforcement Learning

Reinforcement Learning is another class of learning methods belonging to artificial intelligence. In recent years has seen considerable interest in various areas of application, such as energy [36], in the field of financial trading [37], in the clinical field [38] and obviously robotics [39]. Like the other learning methods we have seen even in the case of Reinforcement Learning, learning takes place through mathematical optimization; specifically the goal is to find an optimal strategy between a set of possible solutions and the optimization problem in reinforcement learning can be formulated as maximizing the cumulative *return* after executing a series of actions which are determined by the policy function, we define certain essential elements that make up an optimization problem in Reinforcement Learning, we have the target that is defined as the probability of performing a specific action a in a space of possible states s, through the function of *policy* $a = \pi(s)$, the maximization will therefore be:

$$\max_{\pi} V_{\pi}(s) = \mathrm{E}\left[\sum_{k=0}^{\infty} \phi^k r_{t+k} | S_t = s\right] \tag{2.22}$$

where $\mathrm{E}(\cdot)$ is the expected value of random state s and $V_{\pi}(s)$ is the value function of state s under policy $\pi(\cdot)$, r is the reward and $\phi \in [0, 1]$ is the discount factor.

2.4 Ensemble Learning

Ensemble methods are a set of machine learning techniques that combine a series of algorithms. The idea behind the method is that multiple algorithms combined between them can be more accurate than each of them taken individually. The concept of to combine together *opinions* is the basis of the human decision-making process that we already find in the times of the ancient Greeks and the theorem of the judges of Condocert is a clear example of how one can precisely "combine" together one's opinions on the basis of specific criteria examined. Ensemble methods are therefore more efficient, dividing the problem into many multiple sub-problems, easier to treat from a computational point of view and easier to understand. Without loss of generality in this chapter, the main techniques and a broad overview of the reasons for using these methods will be presented. The aim always remains to introduce the reader to a rather comprehensive overview and enable him to understand the applications and experimental results that will be treated later.

From a mathematical point of view, the ensemble methods aim to minimize the total error of the learners in order to enhance accuracy, depending on the type of problem treated (see section 2.3) this method have different mathematical properties and structures. We can divide the methods into two categories, depending on the type of learning. In the case of supervised learning we can consider combine by *learning*, while for unsupervised learning we have combine by *consensus*; for semi-supervised learning we can consider a further combination of the two approaches. For the first type, among the advantages we certainly have that we can have information from the labeled data and we can potentially improve accuracy; among the disadvantages

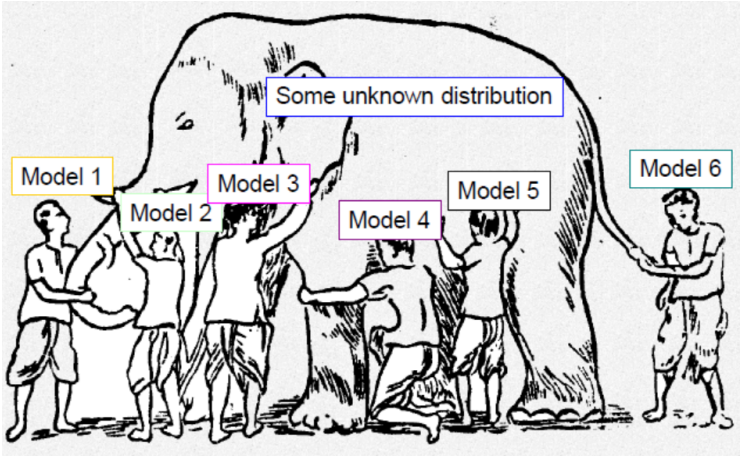

Figure 2.2. Source: Data Mining and Machine Learning for Astronomical Applications

we can consider that it is necessary to have all the labels during the training so on unlabeled data it may not work well. For the second type of method it can obviously be used for unlabeled data and in terms of accuracy it improves the general performance of the algorithm, among the disadvantages there is no feedback from the data.

So, considering the supervised learning given a data set $X = (x_1, x_2, \ldots, x_n)$ and their corresponding labels $Y = (y_1, y_2, \ldots, y_n)$ the ensemble approach computes:

(a) a set of learners $(c_1, c_2, \ldots, c_k)$, each of which maps data $c_j(x) = y, j = 1, .., k$

(b) combination of learners $c^*(\cdot)$ which minimizes the global error: $\min_w c^*(x) = \sum_{j=1}^{k} w_j c_j(x)$

Relative to the unsupervised learning instead given an unlabeled data set $X = (x_1, x_2, \ldots, x_n)$ an ensemble approach for this problem results to be:

(a) a set of clustering solutions $(C_1, C_2, \ldots, C_k)$, each of which maps data to a cluster $c_j(x) = z$

(b) a unified clustering solutions c^* which combines base clustering solutions by consensus.

There are several strategies for training *learners*, for simplicity we'll refer to them by calling them classifiers rather than regressors. One method is to sample a subset of the data and train k - classifiers on it; another strategy is to use all the examples of the dataset by training different learners. Another possibility is to train k - classifiers on a subset of the features of the dataset and another method involves inserting randomness into the training procedure; the main techniques related to these training strategies will be presented below. Following the definition of [40] the main idea of a ensemble learning consists of two main steps: generating predictions or clusters using multiple *weak* classifiers or clustering methods and (b) integrating

multiple results into a function to get the final output with voting or consensus schemes.

2.4.1 Supervised ensembles

As we have already mentioned, ensemble methods are considered in many machine learning problems like the cutting-edge solution to solve several complex problems where a single model fails to be highly accurate; the combination of multiple models leads on average to better results. Ensemble methods are now a fundamental element of machine learning and their flexibility and accuracy has made it one of the most used methods. They are used in various areas such as in Yu *et. al.* [55] in which the authors apply ensemble algorithms for multi-class classification problems in the biochemical field, while Daliri [56] uses a combination of SVM for the classification of breast mammograms in the biomedical field. Zhang and Sun respectively use ensemble techniques in transportation, [57], [58]. Hu *et. al.* in the field of computer security studying the case of intrusions through the use of classifiers based on ensemble methods [59]. Fersini *et al.* [60] apply Bayesian ensemble techniques to sentiment analysis problems. The list of works is very vast, therefore here we limit ourselves to a brief review, not exhaustive, but explanatory of the various possible areas of application of these techniques.

Bagging
As part of the ensembles methods supervised among the most well-known algorithms there is the *Bagging* [41], which generates subsets of the training data through random sampling. The formation of the basic models in the integration model is performed in parallel. Considering a training dataset D of size N, the method of bagging generates M dataset $\tilde{D}$ of size $\tilde{N}$ through a simple random sampling with uniform replication; replication sampling will introduce repeated observations and for $N = \tilde{N}$ there will be a fraction $(1 - 1/e)$ of distinct observations; the type of sample is known in literature like **Bootstrap**, therefore we would have $\tilde{M}$ datasets which will be used to train $\tilde{M}$ algorithms which i'll be combined together in order to produce a combined output which in the case of the regression it will be the *average* and in the case of the classification it will be *voting*. This technique primarily aims to reduce variance through the variance-bias tradeoff, considering the bagging estimator $\bar{c}_j(x) = \mathrm{E}(c_j(X))$

$$\mathrm{E}[Y - c_j(X)] = \mathrm{E}[Y - \bar{c}_j(x) + \bar{c}_j(x) - c_j(X)]^2 \tag{2.23}$$

the bagging reduces variance and leaves bias unchanged

$$= \mathrm{E}[Y - \bar{c}_j(x)]^2 + \mathrm{E}[\bar{c}_j(x) - c_j(X)]^2 \geq \mathrm{E}[Y - \bar{c}_j(x)]^2 \tag{2.24}$$

Boosting
Another technique widely used in the field of ensembles is the *Boosting*; in this algorithm more (weaks) models are generated consecutively giving more and more weight to the errors made in the previous models. Through this series of iteration, the error is minimized, the models are created more accurately which takes into account the aspects that caused errors in the previous models, finally obtaining a

model with better accuracy than each model that constitutes it. Considering for an a classifier $c_j(x)$ the error is:

$$\epsilon_j = \frac{\sum_{j=1}^{k} w_j c_j(x)_{I_{[c_j(x) \neq y]}}}{\sum_{j=1}^{k} w_j} \tag{2.25}$$

while the the classifier's importance is formulated as follow:

$$\phi_j = \frac{1}{2} log \left(\frac{1 - \epsilon_j}{\epsilon_j} \right) \tag{2.26}$$

where $log(\cdot)$ is the natural logarithm, then the final combination solve the following optimization problem

$$\arg \max_{y} c(x)^* = \sum_{j} \phi_j \cdot c_j(x)_{I_{[c_j(x)=y]}} \tag{2.27}$$

Gradient Boosting
An evolution of the method is linked to the *Gradient Boosting* which produces a predictive model in the form of a set of (weaks) predictive models, generally through the use of decision trees. It constructs the model gradually and generalizes it by allowing the optimization of a differentiable loss function to minimize the general error. Like with Boosting, Breiman [41] assumes that enhancement can be interpreted as an optimization algorithm on a given loss function. Mason *et al.* [42] have introduced a method to enhance algorithms such as iterative gradient descent or by optimizing a cost function in the function space by choosing iteratively a function that aims in the direction of the negative gradient. Considering the boosting problem, then the gradient boosting the method tries to find an approximation $\hat{C}(x)$ that minimizes the average value of the loss function on the training set and considering the equation (5.5):

$$C_m(x) = C_{m-1}(x) + \arg \min_{f_m \in \mathcal{F}} \left[\sum_{i} L(y_i, C_{m-1}(x_i)) + f_m(x_i) \right] \tag{2.28}$$

where $f_m \in \mathcal{F}$ is a base learner function, $L(\cdot)$ is some specified loss function and I_x is the indicator function.

Below is the generalization of the algorithm [43]:

input : Sample distribution $\mathcal{D}$
Learning algorithm $\mathcal{L}$
Number of learning rounds $\mathcal{T}$
output : $\mathcal{H}(x) = combine\ outputs(h_1(x), ..., h_t(x))$
Process:

1. $\mathcal{D}_1 = \mathcal{D}$: initial distribution
2. **For**: $t = 1, ..., T$
3. $h_t(x) = \mathcal{L}(D_t)$: Train a weak learner from $\mathcal{D}_t$
4. ϵ_t: evaluate error of h_t
5. $\mathcal{D}_{t+1} =$ *adjust distribution* $(\mathcal{D}_t, \epsilon_t)$
6. **End**

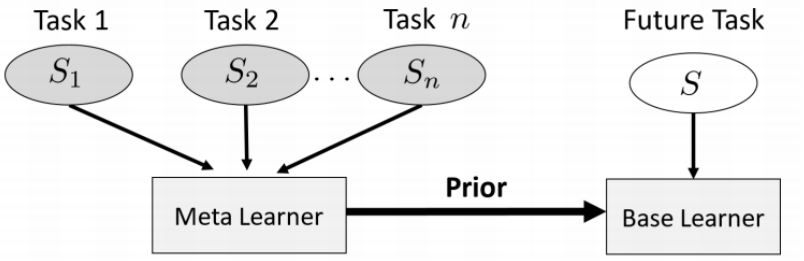

Figure 2.3. The architecture of the Stacking algorithm, Source:[68]

Stacking
Another method [44] widely used in literature linked to the previous ones is the *Stacking* which combines a series of heterogeneous models (could also be weaks), training a meta-lerner on the outputs of the predictions obtained by weak learners. Therefore there are a couple of elements that characterize the staking methods, the first is the number and type of algorithms to be combined and the second element is the meta-learner to combine them together. For example,thinking to combine algorithms such as classification *Tree*, *Logistic* regression and a *KNN* algorithm for a classification problem and then use a *Neural network* as a meta-leaner that takes the outputs of the previous models as input and combines them in order to obtain a more accurate prediction of every single prediction obtainable from weak models. An evolution of the method concerns the multi-level extension in which stacking is carried out with multiple levels. The algorithm can be defined as follows:

input : $\mathcal{D}(x_i, y_i)$, $x_i \in R^M, y_i \in N$
output : An ensemble learner $\mathcal{H}(x)$
Process:

1. **Step 1** : learn first level learner
2. **For**: $t = 1, ..., T$ **Do**:
3. learn a base learner $h_t(x)$ on $\mathcal{D}$
4. **Step 2**: construct new dataset from $\mathcal{D}$
5. **For**: $t = 1, ..., m$ **Do**:
6. construct new dataset that contains (x_{new}, y) where $x_{new} = h_j(x)$ for $j = 1, ..., T$
7. **Step 3**: learn second level learner
8. learn new learner h^{new} based on the newly constructed dataset
9. **Return** $\mathcal{H}(x) = h^{new}(x)(h_1(x), ..., h_T(x))$
10. **End**

2.4.2 Semi-Supervised and Unsupervised

The methods related to semi-supervised ensemble techniques have had a lot of follow-up in recent years. Unlike those that have been previously discussed, these methods aim to expand the set of information available through the data. The reference algorithm in this case will use only partially labeled data for training and secondly, this learner is used to assign so-called *pseudo* - tags to data without labels. The original data together with the pseudo-tags are used in order to update the previously trained models, subsequently the results obtained combined to obtain the final prediction using a certain schema; in the case of classification the *voting* while for the regression problems is the simple average. Therefore, in the case of semi-labeled data the ensemble methods have higher accuracy than models taken individually. As for the unsupervised ensemble methods we find several methods such as Bootstrap samples, Different subsets of features, Different clustering algorithms, Random number of clusters, Random initialization for K-means and Varying the order of given in on-line methods like a BIRCH. These methods can be combined with different approaches, one is called *direct* where the correspondence between the labels and the partitions is found and then merges them with clusters with the same labels. In the *undirect* or also known as *meta-clustering*, where each output is treated as a categorical variable corresponding to a cluster in a new feature space. The consensus clustering framework can be presented as follow [47]: given a set of N data points $\mathbf{X} = (x_1, x_2, ..., x_N)$ and a set of C clusterings $\phi = (\phi_1, \phi_2, ..., \phi_C)$ of the data points in $\mathbf{X}$. Each clustering ϕ_i is a mapping from $\mathbf{X}$ to $(1, ..., n_{\phi_i})$ where n_{ϕ_i} is the number of clusters in ϕ_i. The problem of clustering consensus is to find a new clustering ϕ^* of the data $\mathbf{X}$ that best summarizes the clustering ensemble Φ. The

goal of unsupervised ensemble methods is to combine results from multiple sources. The arduous part of this task is the absence of labeled data and also the search for a meta-learner or an aggregation function that provides accurate and reliable results. In real problems this does not happen and often the ensemble algorithms are limited to sub-optimal solutions and this problem can affect the precision of the models. This class of models has been applied to several real problems, such as in [52] genetics or to problems related to electricity consumption [53], rather than features selection [54] problems. This brief overview will illustrate the main methods and algorithms used in this class of unsupervised problems.

Cluster-based Similarity Partitioning
The clustering of objects with this method occurs through the measurement of the dissimilarity between two objects, defined as the percentage of objects in common in the same cluster, therefore the more two objects are similar, the higher the probability that the two observations will be placed in the same cluster. Defining the similarity between v_i and v_j as:

$$s(v_i, v_j) = \frac{\sum_{k=1}^{K} I_{(c_k(v_i)-c_k(v_j))}}{K} \tag{2.29}$$

where $I(\cdot)$ is the indicator function and $c_k(\cdot)$ the k- th cluster.

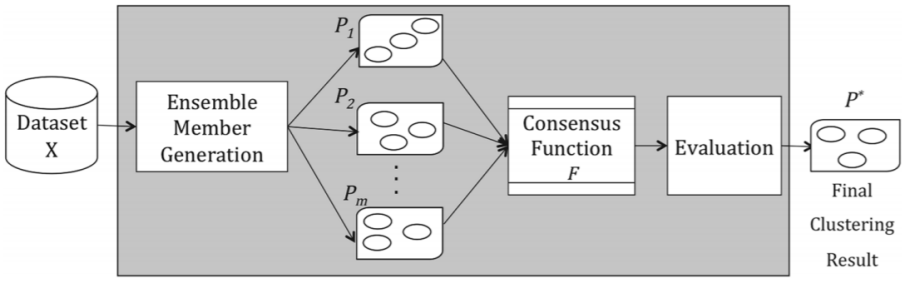

Figure 2.4. Source: A generic clusteringensemble framework, Source:[45]

HyperGraph-Partitioning Algorithm
In the work of Strehl and Ghosh[46] the HGPA algorithm is defined as a direct method whose group is partitioned using data clusters as strong bonds; by cutting a minimum number of hyper-edges the problem is formulated in the form of a hypergraph. The authors propose in this method all hyperedges have the same weight and all vertices are weighted equally in the same way. The algorithm aims to divide the hypergraph into k - components approximately disjoint that have the same size. By the following constraint with a maximum imbalance of 5% obtaining equal dimensions:

$$k \cdot \max_{l \in 1,\ldots,k} \frac{n_l}{n} \leq 1.05 \tag{2.30}$$

Under certain assumptions:

1. X, set of data
2. C_l, $l = 1, ..., k$, partitioning of n-data into k-clusters
3. λ, label vector representing a partition
4. ϕ, cluster function s.t. $\phi : X \rightarrow \lambda$

The authors also propose a consensus function defined as follows:

$$\Gamma : [\lambda^{(1,...,r)}] \rightarrow \lambda \tag{2.31}$$

This function has no knowledge of the original features X and of the clustering algorithm ϕ.

Meta-Clustering Algorithm
Once we have seen the direct approaches, now we also provide a view on the main indirect approach for this class of problems. In the MCLA method [47] like for the stacking in the context of supervised ensembles, the clustering level 0 output is used to train a goal-learner which in turn is always a cluster technique. The cluster correspondence problem is solved by grouping the clusters identified in the individual clusterings of the ensemble. As seen above, each cluster is also represented here as a hyperedge. The algorithm groups and compresses related hyperedges into k - clusters and assigns each data point to the compressed hyperedge in which it participates most strongly. In the work of Jurek *et al.* [48] the authors propose a method for classifying instances based on meta-clustering; starting from a set of classifiers, instances of a validation set are initially classified. The output of each classifier comes next considered as a new set of features to train another learner. Next, a validation set is clustered based on the new attributes for each cluster, its own centroid is calculated and also the label of the class to which it belongs. In this case, the method takes advantage of clustering to partition instances for later training, but it remains a method for supervised problems. Gionis *et al.* [49] propose a work based on cluster aggregation presenting a series of algorithms that from the computational point of view guarantee the quality of the solution despite many of the problems treated are of the NP-Hard type. The proposed methods are based on the variation of the correlation between clusters and they also show how through sampling it is possible to scale over a large amount of data. The proposed methods can use categorical, heterogeneous data, identify outliers and also identify the correct number of clusters to use. In this method the authors solve the task of minimizing the disagreement between a set of cluster inputs across the optimal cluster.

2.4.3 Pairwise Similarity Approach

Based on equation (5.7) this method also makes use of the concept of dissimilarity with similarity matrix $\mathcal{S}$, given the appropriate metric option, a consensus cluster can be made.

Furthest Consensus
The authors [47] describe the method in which the algorithm's goal is to locate the most distant k clusters, starting with a search for the pair of distant points in order to find the most distant k cluster centers. The algorithm begins with finding a pair of data points that are more distant and assign them as cluster centers. In an iterated way a subsequent storage center further away from the previous centers found. Finally, all points are awarded to the nearest centroid.

Hierarchical Agglomerative Clustering Consensus
This method is a standard algorithm regarding the correlation clustering problem. Start by placing all the data pointing to singleton clusters, iteratively the algorithm joins two clusters that have the closest mean similarity measure that is given in the similarity matrix $\mathcal{S}$. The procedure stops when K clusters remain

2.4.4 Other approaches

Mutual Information
An EM (Expectation - Maximization) algorithm is used to optimize quadratic mutual information in order to find the optimal consensus function, therefore the algorithm works by defining an objective function as mutual information. The procedure requires multiple iterations to avoid minimal poor quality premises

Mixture Model
The [50] authors propose a probabilistic consensus algorithm using a finite mixture of multinomial distributions. Also in this case, obviously the solution is undertaken through the application of the EM algorithm for the clustering ensemble, maximizing the probability of belonging.

Cluster Correspondence
In the work [51] the authors, through the formulation of a linear programming problem, approach and solve the problem of finding clusters of different clusterings in the ensemble for a constrained and unconstrained optimization problem. Finally, a simple *voting* procedure is applied to assign data points to clusters.

2.4.5 Combinations methods

In this section we'll now consider the main methods for combining the outputs of the individual algorithms used; there are two main methods, the *weighting* methods and the *meta - learning* methods as presented in Rokach's work, [61]. The author describes the weighting methods as useful if the base classifiers perform the same task and if the results (or successes) can be compared, while the meta-learning methods

are more suitable for cases where some classifiers consistently classify correctly or consistently classify examples incorrectly.

Weighting methods
In the case of the weighting method, the weight can be dynamic or fixed, the weight of the weighted classifier is proportional to the assigned weight. In general, the weights in an ensemble algorithm are assigned either proportionally to the number of learners by setting $w_i = 1/N$ for $i = 1, ..., N$ or with an optimization-based approach, going to minimize the loss function or are initialized uniformly and subsequently updated at each iteration during training, minimizing the deviation (i.e. exponential loss) or maximizing the accuracy (performance-based) of the global model.

Majority voting
As part of the combination of ouputs in ensemble methods one of the most used methods is that of *majority* voting in which the classification of an observation without label is assigned to the class that obtains the highest number of votes. Its mathematical formulation can be expressed, following the work of [61]:

$$C(x) = \arg \ \max_{i \in y} \sum_{k} I_{(y_k, c_i)} \tag{2.32}$$

where y_k is the classification of the k-th classifier and I_x is the indicator function. For the probabilistic classifiers, the (5.10) become:

$$C(X) = \arg \ \max_{i \in y} \sum_{k} p(y = c_i|X) \tag{2.33}$$

and $p(y = c|x)$ is the probability of class c given an instance x.

Performance weighting
In the work of Optiz and Shavlik [62] the weights are proportional to the accuracy obtained on the test data, mathematically it is defined as follows:

$$w_i = \frac{1 - E_i}{\sum_j (1 - E_j)} \tag{2.34}$$

where E_i is a normalization factor based on the classifier's performance evaluation i-th on a validation dataset.

Bayesian combination
In the Bayesian approach the combination occurs through the weight associated with each classifier with the posterior probability of data distribution. In Buntine's work [63] the method is defined as follows:

$$C(X) = \arg \ \max_{i \in y} \sum_{k} p(y = c_i|X) \cdot p(f_k|\mathcal{D}) \tag{2.35}$$

$p(f_k|\mathcal{D})$ is the probability that the classifier f_k is correct given the training set $\mathcal{D}$.

Logarithmic opinion pool

Hansen [64] in 2000 proposed a method for combining learners called *log opinion pool*, which selects the classifiers as follows:

$$C(X) = \arg \ \max_{i \in y} e^{\sum_k w_k log(p(y=c_j|X))} \tag{2.36}$$

w_k is the weight for the k-th classifier s.t. $\sum_k w_k = 1$, $w_k \geq 0$

Meta-combination methods

We saw earlier that stacking [65] uses a meta-learner as a combinator which is an algorithm that combines the predictions of the outputs of the combined models as inputs (see Wolpert 1992). The basic idea is to create a metadata dataset containing one tuple for each tuple in the original dataset. The target attribute remains as in the original training set. Usually the method works by partitioning the data into two subsets. The first subset is reserved to form the meta-data set e the second subset is used to create the base level classifiers. The author suggests that the method can achieve better performance if the probabilities of output are used for each label in the class from the basic level. Merz, [66] introduces stacking and correspondence analysis strategies. Correspondence analysis is a statistical method widely used to geometrically model the relationship between the rows and columns of a matrix whose variables are categorical and is used in the ensemble context to explore the relationship between the training examples and their classification for a set of classifiers. The problem of meta-learning can be formalized as follow; given an a loss function $\mathcal{L}$, a set of parameters θ and dataset $\mathcal{D} = (\mathbf{x}, \mathbf{y})$ then the optimization problem result:

$$\theta^* = \arg \min_{\theta} \mathrm{E}_{\mathcal{D} \sim p(\mathcal{D})}[\mathcal{L}_\theta(\mathcal{D})] \tag{2.37}$$

where $p(\cdot)$ is the distribution of learning tasks which are minimized respect to the parameters θ and $\mathrm{E}(\cdot)$ is the expected value with respect to the distribution p. Each task is associated with a dataset $\mathcal{D}$ containing both feature vectors and true labels. In the first level of training step we obtain from classifier f_θ an output (probabilities) from data point belonging to the class y_i given the feature vector $\mathbf{x}$ from $\mathcal{D}$ that is $p(y_i|\mathbf{x})$ used in the second learn level for training the algorithm with $\mathcal{D}^{'} = (\mathbf{x}, p(y_i|\mathbf{x}))$ for minimized the (5.15).

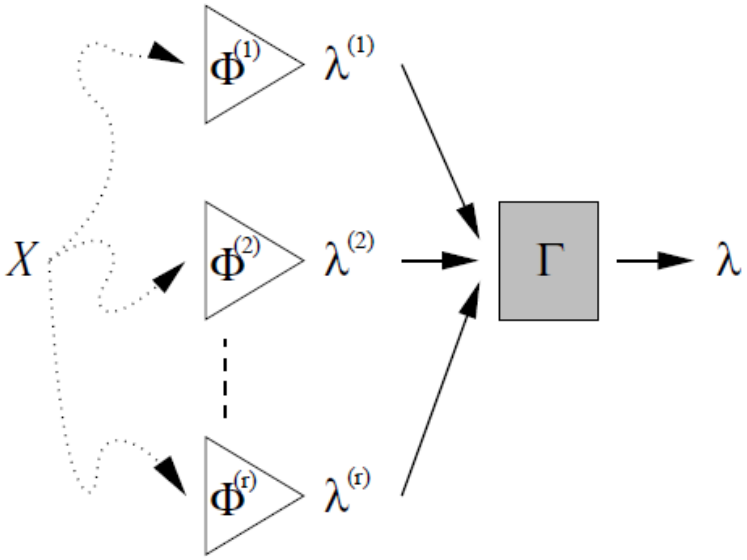

Figure 2.5. The Cluster Ensemble: a consensus function Γ combines clusterings from a variety of sources, without resorting to the original object features, Source:[46]

Bibliography

[1] Saaty, T.L.,"The Analytic Hierarchy Process", McGraw-Hill, New York, 1980

[2] R.W. Saaty, "The analytic hierarchy process—what it is and how it is used", Mathematical Modelling Volume 9, Issues 3–5, 1987, Pages 161-176

[3] T. L. Saaty, "Decision Making for Leaders", Wadsworth, Belmont, Calif., 1982.

[4] Watrobski, Jaroslaw and Ziemba, Pawel and Jankowski, Jaroslaw and Ziolo, Magdalena. (2016). Green Energy for a Green City - A Multi-Perspective Model Approach. Sustainability

[5] J.P. Brans (1982). "L'ingénierie de la décision: élaboration d'instruments d'aide à la décision. La méthode PROMETHEE". Presses de l'Université Laval.

[6] Brans, J.P. and Vincke, P. (1985) A Preference Ranking Organisation Method: (The PROMETHEE Method for Multiple Criteria. Decision-Making). Management Science, 31, 647-656.

[7] Bernard Roy, Classement et choix en présence de points de vue multiples (la méthode ELECTRE), in La Revue d'Informatique et de Recherche Opérationelle (RIRO), n. 8, 1968, pp. 57-75

[8] Figueira, José; Salvatore Greco; Matthias Ehrgott (2005). Multiple Criteria Decision Analysis: State of the Art Surveys. New York: Springer Science Business Media, Inc. ISBN 0-387-23081-5

[9] Hwang, C.L.; Yoon, K. (1981). Multiple Attribute Decision Making: Methods and Applications. New York: Springer-Verlag

[10] Yoon, K. (1987). "A reconciliation among discrete compromise situations". Journal of the Operational Research Society

[11] Hwang, C.L.; Lai, Y.J.; Liu, T.Y. (1993). "A new approach for multiple objective decision making". Computers and Operational Research

[12] Hung C.C., Chen L.H. (2009): "A Fuzzy TOPSIS Decision Making Model with Entropy Weight under Intuitionistic Fuzzy Environment". Proceedings of the International Multi-Conference of Engineers and Computer Scientists IMECS, Hong Kong.

[13] Jahanshahloo G.R., Lofti F.H., Izadikhah M. (2006b): Extension of the TOPSIS Method for Decision-Making Problems with Fuzzy Data. "Applied Mathematics and Computation", 181, pp. 1544-1551.

[14] Jadidi O., Hong T.S., Firouzi F., Yusuff R.M., Zulkifli N. (2008): TOPSIS and Fuzzy Multi-Objective Model Integration for Supplier Selection Problem. "Journal of Achievements in Materials and Manufacturing Engineering of Achievements in Materials and Manufacturing Engineering", 31(2), pp. 762-769

[15] Mohsen Pirdashti, Madjid Tavana, Mimi Haryani Hassim, Majid Behzadian, I.A. Karimi, "A taxonomy and review of the multiple criteria decision-making literature in chemical engineering", Int. J. Multicriteria Decision Making, Vol. 1, No. 4, 2011

[16] Ehrgott, M. (2005) Multicriteria Optimization, 2nd ed., Springer, New York.

[17] Ehrgott, M. and Wiecek, M.M. (2005) 'Multiobjective programming', in J. Figueira, S. Greco and M. Ehrgott (Eds.): Multiple Criteria Decision Analysis: State of The Art Surveys, Springer Science Business Media Inc., pp.667–722.

[18] Zhang, J.L.G., Ruan, D. and Wu, F. (2007) Multi-Objective Group Decision Making Methods, Software and Applications with Fuzzy Set Techniques, Imperial College Press.

[19] Hwang, C.L. and Masud, A.S. (1979) Multi Objective Decision Making, Methods and Applications, Berlin, Springer.

[20] A Charnes, WW Cooper, R Ferguson (1955) Optimal estimation of executive compensation by linear programming, Management Science, 1, 138-151

[21] Stelios H Zanakis and Sushil K Gupta, (1985), A categorized bibliographic survey of goal programming, Omega, 13, (3), 211-222

[22] Vikhar, P. A. "Evolutionary algorithms: A critical review and its future prospects". Proceedings of the 2016 International Conference on Global Trends in Signal Processing, Information Computing and Communication (ICGTSPICC). Jalgaon: 261–265

[23] J.H. Holland (1975) Adaptation in Natural and Artificial Systems, University of Michigan Press, Ann Arbor, Michigan

[24] Goldberg, D. E. (1989). Genetic algorithms in search, optimization, and machine learning. Reading, MA: Addison-Wesley

[25] Aarts E. and Korst J., Simulated Annealing and Boltzmann Machines, John Wiley and Sons, 1990.

[26] Fred Glover (1986). "Future Paths for Integer Programming and Links to Artificial Intelligence". Computers and Operations Research. 13 (5): 533–549.

[27] Dolan, J.G. Multi-Criteria Clinical Decision Support. Patient-Patient-Centered-Outcome-Res 3, 229–248 (2010)

[28] Curry, Haskell B. (1944). "The Method of Steepest Descent for Non-linear Minimization Problems". Quart. Appl. Math. 2 (3): 258–261

[29] Botton Leon (1998). "Online Algorithms and Stochastic Approximations". Online Learning and Neural Networks. Cambridge University Press

[30] J. Hu, B. Jiang, L. Lin, Z. Wen, and Y.-x. Yuan, "Structured quasinewton methods for optimization with orthogonality constraints," SIAM Journal on Scientific Computing, vol. 41, pp. 2239–2269, 2019

[31] J. E. Dennis, Jr, and J. J. More, "Quasi-Newton methods, motivation and theory," SIAM Review, vol. 19, pp. 46–89, 1977

[32] "Semi-Supervised Learning", Chapelle, Schölkopf and Zienin 2006

[33] Vapnik, V.; Chervonenkis, A. (1974). Theory of Pattern Recognition (in Russian). Moscow: Nauka cited in Chapelle, Schölkopf and Zienin 2006, p. 3

[34] K. P. Bennett and A. Demiriz, "Semi-supervised support vector machines," in Advances in Neural Information processing systems, 1999, pp. 368–374.

[35] Hinton, G. E. and Zemel, R. S. (1994). Autoencoders, minimum description length and Helmholtz free energy. In Advances in neural information processing systems 6(pp. 3-10)

[36] Lazic, N., Boutilier, C., Lu, T., Wong, E., Roy, B., Ryu, M., and Imwalle, G. (2018). Data center cooling using model-predictive control. In NeurIPS

[37] Bacoyannis, V., Glukhov, V., Jin, T., Kochems, J., and Song, D. R. (2018). Idiosyncrasies and challenges of data driven learning in electronic trading. In NeurIPS 2018 Workshop on Challenges and Opportunities for AI in Financial Services: the Impact of Fairness, Explainability, Accuracy, and Privacy

[38] Komorowski, M., Celi, L. A., Badawi, O., Gordon, A. C., and Faisal, A. A. (2018). The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. Nature Medicine, 24:1716–1720.

[39] Kober, J., Bagnell, J. A., and Peters, J. (2013). Reinforcement learning in robotics: A survey. International Journal of Robotics Research, 32(11):1238–1278.

[40] Dong, Xibin, "A survey on ensemble learning", Frontiers of Computer Science Volume: 14 Issue 2 (2020) ISSN: 2095-2228

[41] Breiman L. Bagging predictors. Machine Learning, 1996, 24(2): 123– 140

[42] Mason, L.; Baxter, J.; Bartlett, P. L.; Frean, Marcus (May 1999). "Boosting Algorithms as Gradient Descent in Function Space"

[43] Zhou Z H. Ensemble Methods: Foundations and Algorithms. Chapman and Hall-CRC, 2012

[44] P.Smyth and D.Wolpert. Linearly combining density estimators via stacking . Machine Learning, 36(1-2):59-83, 1999.

[45] Alqurashi, T. and Wang, W.,“Clustering ensemble method”

[46] Strehl, A. and Ghosh, J., “Cluster Ensembles–A Knowledge Reuse Framework for Combining Partitionings,“Journal of Machine Learning Research “ (2002)

[47] Nguyen, N. and Caruana, R., “Consensus Clusterings. Proceedings” - IEEE International Conference on Data Mining 2007

[48] A. Jurek, Y. Bi, S. Wu and C. D. Nugent. "Clustering-Based Ensembles as an Alternative to Stacking," in IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 9, pp. 2120-2137, Sept. 2014

[49] Gionis, A., Mannila, H., and Tsaparas, P." Clustering aggregation. ACM Transactions Knowledge Discovery Data, 1, 1, Article 4 (March 2007)

[50] A. Topchy, A. K. Jain, and W. Punch. A mixture model for clustering ensembles. In Proceedings SIAM International Conference on Data Mining, 2004

[51] C. Boulis and M. Ostendorf. Combining multiple clustering systems. In The 8th European conference on Principles and Practice of Knowledge Discovery in Databases(PKDD), LNAI 3202, pages 63–74, 2004

[52] Monti, S., Tamayo, P., Mesirov, J., Golub,. T. “Consensus Clustering: A Resampling-based Method for Class Discovery and Visualization of Gene Expression Microarray Data” . Machine Learning 2003, 52:91–118

[53] Laurinec, P., Lóderer, M., Lucká, M. et al. Density-based unsupervised ensemble learning methods for time series forecasting of aggregated or clustered electricity consumption. J Intell Inf Syst 53, 219–239 (2019)

[54] Elghazel, H., Aussem, A. Unsupervised feature selection with ensemble learning. Mach Learn 98, 157–180 (2015)

[55] Yu, G., Rangwala, H., Domeniconi, C., Zhang, G., Yu, Z. “Protein function prediction using multilabel ensemble classification”. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2013

[56] Daliri MR. “Combining extreme learning machines using support vector machines for breast tissue classification”. Computer Methods in Biomechanics and Biomedical Engineering, 2015

[57] Zhang B. “Reliable classification of vehicle types based on cascade classifier ensembles”. IEEE Transactions on Intelligent Transportation Systems, 2013

[58] Sun, S., Zhang, C. “The selective random subspace predictor for traffic flow forecasting”. IEEE Transactions on Intelligent Transportation Systems, 2007

[59] Hu, W., Hu, W., Maybank, S. “AdaBoost-based algorithm for network intrusion detection”. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 2008

[60] Fersini, E., Messina, E., Pozzi, FA. “Sentiment analysis: Bayesian ensemble learning”. Decision Support Systems, 2014

[61] Rokach, L., “Ensemble-based classifiers”,Artificial Intelligence Review, February 2010

[62] Opitz, D., Shavlik, J. “Generating accurate and diverse members of a neural network ensemble”. Advances in neural information processing systems, vol 8. 1996

[63] Buntine, W. “A theory of learning classification rules. Doctoral Dissertation”. School of Computing Science, University of Technology, Sydney, Australia 1990

[64] Hansen, J. "Combining predictors. Meta machine learning methods and bias, variance and ambiguity decompositions". PhD Dissertation. Aurhus University 2000

[65] Wolpert, D.H. " Stacked generalization". Neural Networks, vol 5. Pergamon Press, Oxford, pp 241–259, 1992

[66] Merz, C.J. "Using correspondence analysis to combine classifier". Machine Learning 36(1–2):33–58, 1999

[67] Hospedales, T.M., Antoniou, A., Micaelli, P., Storkey, A.J., “Meta-Learning in Neural Networks: A Survey”, ArXiv, 2020

[68] Amit, R., and Meir, R. (2018). "Meta-Learning by Adjusting Priors Based on Extended PAC-Bayes Theory". ICML.

Chapter 3

Proposed approach for CDSS

Clinical decision support systems (CDSS) play a very important role in the health sector, since every action taken by a decision maker is crucial by an operational point of view and for the an ethical and legal point of view. Decision makers can be of various types in a framework of CDSS i.e. a Medicine Doctor (M.D.), minister, or scientist task force. The results of a decision support system on the one hand make it possible to take informed choices, since it is assumed that there is an expert who *supervises* the decision-making process and on the other hand they can lead to poorly interpretable results, to depending on the models that are used; therefore not very useful for the intended purposes. As described in the first chapter on the interpretability of black-box algorithms, a CDSS based on these methodologies carries with them a great responsibility. The output of a model can concern for example a drug therapy, the administration of a drug, the experimentation of a vaccine rather than surgery or compatibility on organ transplants. A M.D. who adopts a CDSS is clearly subject to legal liability (see appendix A.2), obviously in addition to the ethical-professional, therefore the interpretability and the *transparency* of the models used must guarantee the full explainability of the results; why one model was preferred over another and how this model was used. In this chapter, a new method is proposed for the construction of the decision-making process applicable to the clinical context: three case studies are presented, data are analyzed, mathematical models are built and the results are validated by an experimental phase of the analysis, furthermore some mathematical methods for the interpretability of the machine learning algorithms that are applied, both in the context of supervised and unsupervised problems

3.1 State of art

Wyatt and Spiegelhalter [1] define medical decision aids as: "active knowledge systems which use two or more items of patient data to generate case-specific advice". For this reason the Clinical DSSs are typically designed to integrate a medical knowledge base, patient data and an inference engine to generate case specific advice. The authors [1] continue defining some principles that can exempt the actors involved in a clinical support system from legal liability in the event that there are legal problems, specifically they could:

1. The system has been carefully evaluated in the laboratory studies
2. The system provided its user with explanations, well calibrated probabilities or the opportunity to participate in the decision-making process
3. No misleading claims had been made for the system
4. Any error was in the design or specification rather than in the coding or hardware
5. Users had been adequately trained and had not modified the system

Point 2 is exactly what we said previously regarding the responsibility linked to the interpretability and transparency of the solutions obtained by the algorithms used in the CDSS. From the 70s onwards there have been several clinical support systems based on artificial intelligence, for example a work from 1972 by de Dombal *et. al.* [2] in which a first attempt is made to implement automatic reasoning in conditions of uncertainty. The system was developed by Leeds University, designed to support the diagnosis of acute abdominal pain and on the basis of analysis the need for surgery. System decision making was based on the Bayesian approach. Miller *et. al.* [3] developed INTERNIST-1, one of the first clinical decision support systems designed to support diagnosis, in 1970 the CDSS was a rule-based expert system designed by the University of Pittsburgh in 1974 for the diagnosis of complex diagnosis of complex problems in general internal medicine. It uses patient observations to deduce a list of compatible disease states (based on a tree-structured database that links diseases with symptoms). In the work of Shortliffe [4], (MYCIN), a rule-based expert system designed to diagnose and recommend treatment for some blood infections (antimicrobial selection for patients with bacteremia or meningitis) was proposed. It was later extended to manage other infectious diseases. Clinical knowledge in the CDSS is represented as a set of rules **IF - THEN**.

3.1.1 Cancer

The use of decision support systems in the clinical sector is widely known, Vidal *et. al.* [5] present a CDSS based on the AHP method to assist pharmacists to choose a drug therapy in cancer patients, while always for the AHP Liberatore *et. al.* [7] have implemented a DSS relating to the protocols necessary for prostate cancer and the study has indicated that the decision counseling protocol is appropriate in primary care only if it is well structured and coordinated by an expert analyst (decision-maker). Carter *et. al.* [8] considered three models for post-lumpectomy treatment; the treatment alternatives considered in the study were the observation, radiation and combination of tamoxifen, radiation and tamoxifen and simple mastectomy; the methods used were a Markov Process, AHP and ANP; among the three AHP it was the fastest in producing the expected results. The study shows that the choice of a particular method depends on the context and the requirement established by the analysis. Still in the context of CDSS Dolan [6] used the AHP method in the choice of five types of screening for colon cancer, 50% of the patients on which the model was tested produced positive results such as use the CDSS in a clinical sector.

3.1.2 Diabetes

Several studies have also been conducted in the context of diabetes treatment using the CDSS. Specifically, [9] present a work based on the evaluation of the impact of an electronic system to support clinical decisions on diabetes, relating to medical records on the control of glycated hemoglobin A1c, blood pressure and cholesterol levels (LDL) in adults with diabetes. The study is relative to the period 2006-2007 on 2,556 diabetic patients. The CDSS was designed to improve care for those patients whose hemoglobin A1c, blood pressure or LDL levels were higher than the target through the application of general and generalized linear mixed models with repeated time measurements. In [10] Georga *et. al.* it is present a clinical diabetes management system to support the follow-up and treatment processes of diabetic patients and also the authors propose a data mining of time models as a tool to predict and explain the long-term course of the disease. In the context of methods for multi-criteria decisions, Rung-Ching *et. al.* [11] propose a TOPSIS based method to calculate the ranking of anti-diabetic drugs; the CDSS presents a utility of 87% through a recommendation system for outpatients. The authors also discuss the fact that in addition to helping the clinical diagnosis of doctors, the system can not only serve as a guide for specialist doctors, but it can also help non-specialist doctors and young doctors to prescribe medications.

3.1.3 Cardiovascular diseases

In the context of ischemic stroke, precisely in patients who have thrombolysis, Lee *et. al.* [12] have developed a clinical decision support system to customize treatment in patients who present stroke. A series of 958 patients hospitalized within 12 hours of the onset of ischemic stroke from a representative clinical center in Korea, was used to establish a prognostic model through a multivariate logistic regression, which was used to develop the model for overall and safety results. In the model the authors considered age as a predictor, the score of the Rankin scale modified previously; initial score from the National Institutes of Health Stroke Scale (NIHSS); previous stretch; diabetes; previous use of antiplatelet treatment, antihypertensive drugs and statins; gaps; thrombolysis; from start to treatment time and systolic blood pressure. The predictors of the final safety outcomes were age, initial NIHSS score, thrombolysis, start of treatment time, systolic blood pressure and glucose level. A new computerized model for predicting results for thrombolysis after ischemic stroke was therefore developed within the framework of the CDSS using large amounts of clinical information. The model was validated by the experts (decision-maker) and the model's performance was deemed clinically satisfactory. Another work that refers to a clinical decision support system is that proposed by Andersen *et. al.* [13] about secondary stroke prevention. The multidisciplinary team was funded received by the Office of Nursing Services of the Veterans Health Administration. This paper presents the *alpha test* results obtained while using an integrated model for the development of the clinical decision support system which emphasizes the prospects of the end user throughout the development process. The clinical support tool involved the development of several integrated functions, among the main features of the tool we found the automated request and the documentation of the

secondary stroke prevention guidelines in the electronic medical record. Usability of the system was assessed with a questionnaire developed by the investigator and an open-ended question. The prototype resulted in a significant increase with a *p-value* less than 0.5 of the provider documentation for six of the 11 guidelines compared to the basic documentation when using the standard system . The authors conclude that the results produced support the fact that the guideline suggestion was successfully designed to produce a usable and useful clinical decision support system for the prevention of secondary stroke. Arts *et. al.* [14] propose a method to improve adherence to guidelines with a non-invasive clinical decision support system integrated into the workflow. The proposed sampling framework is a randomized controlled cluster study in Dutch general practices. A support system has been developed that implements properties positively associated with efficacy: in real time, non interruptive and based on electronic medical records data, furthermore the recommendations were based on the guidelines of the Dutch general practitioners for atrial fibrillation using CHA2DS2 -VAsc for stroke risk stratification. As an assessment metric regarding the effectiveness of the method, adherence to the guideline was measured. The authors propose an association assessment approach using a chi-square to check group differences and a mixed effects model to correct clusters and basic adhesion. Out of 781 individuals, the authors report the following results: of the total, 76 notifications received a response: 58% of dismissals and 42% of acceptance. By the end of the study, groups had improved by 8% and 5% respectively. There was no statistically significant difference between the groups (control: 50%, intervention: 55% P = 0.23). Cluster analysis revealed similar results. Only one of the useful reasons for non-adherence was captured. Therefore as the same authors say the study has not been able to demonstrate the effectiveness of a decision support system probably due to the lack of use, because in view of real and complex problems it is not possible to use CDSS that lead to satisfying results.

3.1.4 Other applications

Not only problems related to heart attacks, cancer and diabetes, CDSS have also been used very recently following the 2019 coronavirus epidemic (COVID-19). Until now many drugs and methods have been used in the treatment of the disease. However no effective therapeutic options have been found. The study proposed by [15] aims to evaluate COVID-19 treatment options using multi-criteria decision techniques, in particular PROMETHEE, Fuzzy and VIKOR. This technique is based on the evaluation and comparison of complex and multiple criteria to evaluate the most appropriate alternative. Among the treatments analyzed are those related to anti retrovirals used for HIV patients such as favipiravir (FPV), lopinavir/ritonavir and other drugs used during the most acute initial phase of the pandemic as hydroxychloroquine, interleukin-1 blocker, immunoglobulin intravenous (IVIG) and plasma exchange. Among the criteria the authors include and use for the analysis are side effects, drug delivery method, cost, plasma turnover, fever level, age, pregnancy, and renal capacity. The results of the analysis conducted showed that plasma exchange was the best alternative, followed by FPV and IVIG, while hydroxychloroquine was the least favorable. Weights could be assigned based on the opinions of decision makers (medical-clinicians). The DSS described in chapter 3 have been used in

the development of very important decision problems such as drug treatment, the choice of the best therapy and predictions about a given health event but in a CDSS framework were not only used for these purposes, but also for the selection of the best algorithms as in the study of Khanmohammadi and Rezaeiahari [16] in which the authors, through the determination of the criteria and sub-criteria concerned have evaluated the performance of a series of machine learning algorithms through the AHP method. The authors, using features such as computational complexity, accuracy, memory used in algorithm training, draw up a ranking of the best algorithms. Therefore the use of a CDSS can be direct, like the works analyzed previously and indirectly as in the case in which you have to choose a model that is always used in the clinical sector but of different use.

Elementary Methods	Value-based Measurement Methods	Goal Programming and Reference	Outranking Methods
• Maximin • Maximax • Hurwicz • Disjunctive • Conjunctive • Lexicographic	• SLAM • MAVT • MAUT • **AHP*** • Weighted Sum	• **Goal Programming*** • Heuristics • Meta-heuristics	• ELECTRE • PROMETHEE • **TOPSIS*** • **VIKOR***

Figure 3.1. The MCDA methods used in healthcare decision-making Source:[17]

3.1.5 Explainability

As described in the opening chapters the problem of the explainability of results in a support system based on artificial intelligence is crucial. The best techniques in terms of algorithms, accuracy, efficiency and computational complexity can be used but without the interpretation of the output obtained everything becomes useless. So in the context of systems that support clinical decisions, this requirement it is essential. Clinical Decision Support Systems (CDSS) provide assistance to clinicians in decision making. In the work of Muller *et. al.* [18] by definition these systems are based on patient-specific evidence and representations of clinical knowledge modeled by algorithms and mathematical models by experts and provide recommendations in finding the right diagnosis or optimal therapy. In this paper is proposed an approach based on *visualization*. The authors present a glyph-based approach coordinated multiple views to support explainable computerized clinical decisions: "inspired by common decision making in clinical routine". Specifically, this type of methodology is very intuitive and therefore offers explainable and understandable results. The authors show that multiple views show the certainty of the calculation result like the recommendation and a series of clinical scores. About the model used, the authors presented an approach for a CDSS based on a bayesian causal network representing the therapy of laryngeal carcinoma. The results were evaluated and validated by two experienced otolaryngologists. Several other studies have addressed the question of the explainability of CDSS, such as in [19], [20], [21], failing to calibrate the concept of user trust by introducing this new type of error to the context as analyzed by [22]

using these tools. Another example relating to the work of Bussone *et al.* [23] who studied the effect of the explanation on *trust* and dependence. The authors state: "neglecting human factors and user experience in designing the CDSS explanation could lead to over-reliance on medical professionals in these recommendation systems, even when it is wrong", which the authors define an "excessive reliance". There is also another possible problem when the explanation that does not provide enough information could lead to users who *reject* the suggestions, for example self-sufficiency or low confidence as described in the work of [24].

3.2 Proposed Methodology

To support of what has been described and analyzed in the previous chapters, in this phase of the work a new methodology is proposed for the explainability of the clinical decision support systems. Specifically it is an approach based on ensemble methods and methods for multi-criteria decisions. We have seen in the section dedicated to the *ensemble* methods, especially for the *stacking* methods that the combination of algorithms through the training of a *meta-learner* (see section 2.4) who uses the outputs of the trained models as input to combine the results in order to obtain a final result it is better than the results taken individually. In this work the meta-learner used is based on the AHP method (Analytic Hierarchy Process), described in the chapter dedicated to the methods for multi-criteria decisions. Below we'll describe the approach of this new method, and subsequently we will provide results based on experiments with clinical data. In each of the three applications to the case studies examined the data are documented and the analyzes performed were carried out using Python 3.7 and related machine learning libraries (i.e sklearn, lime, pandas, numpy, eli5, seaborn, scipy).

3.2.1 Methodology

We assume there is an *expert* who supervises the whole decision-making process. Starting from the definition of the rating scale for alternatives and preferences in this case the expert defines an evaluation rating with respect to the usable algorithms. In accordance with the trade-off between complexity and interpretability, the expert supervisor considers five scores, which identify the complexity and interpretability of the algorithm. By table 3.1, the value 1 identifies that class of models that are considered simple in terms of interpretability like results and parameters that define the model; in this group we find for example the linear regression model, rather that a logistic regression and so on up to the value of 9, which represents according to the expert, the most complex class of models in terms of explainability but in terms of performances.

Introduction
Once we have chosen the algorithms that i'll be used training on dataset $\mathbf{x}$ we would therefore have $l_1(\mathbf{x}), ..., l_k(\mathbf{x})$ learners (can be classifiers o regressors) each will produce an output (estimated) $\hat{o}_i = l_i(\mathbf{x})$, $i = 1, ..., k$. Using pairwise comparisons the relative importance one of criterion (learner $l_i, i = 1, ..., k$) over another can be expressed by the following matrix associated:

Expert Rating Scale		
Intensity	**Definition**	**Interpretation**
1	**Simple model**	Easy interpretation, usually linear model i.e. regression
3	**Simple but effective model**	Good trade off between interpretation and complexity i.e. SVM, SVC
5	**Not too simple model**	Models that can be interpreted but still more advanced i.e. random forest, classification trees
7	**Complex model**	Non-linear models that on average give better results but begin to be more complex in interpretation i.e. neural networks
9	**Very complex model**	In this category more complex, deep neural networks, can be considered, i.e. LSTM, RBM, autoencoder, etc..

Table 3.1. Rating Scale

$$\mathcal{L} = \begin{array}{c} \\ l_1 \\ l_2 \\ l_3 \\ ... \\ l_k \end{array} \begin{array}{c} \begin{array}{ccccc} l_1 & l_2 & l_3 & ... & l_k \end{array} \\ \begin{bmatrix} 1 & 3 & 5 & ... & 9 \\ 1/3 & 1 & 3/5 & ... & 1/3 \\ 1/5 & 5/3 & 1 & ... & 5/9 \\ ... & ... & ... & ... & ... \\ 1/9 & 3 & 9/5 & ... & 1 \end{bmatrix} \end{array}$$

Starting from a $\tilde{\mathcal{L}} = \mathcal{L}\mathcal{L}^{'}$ comparison matrix it is necessary obtain the priority vector $\mathcal{W}$ which is the normalized eigenvector of the matrix. The method is an approximation of the eigenvector of a reciprocal matrix. Considering the following normalization:

$$\tilde{a}_{ij} = \frac{a_{ij}}{\sum_j^n a_{ij}} \tag{3.1}$$

where a_{ij} is the entry of matrix $\tilde{\mathcal{L}}$ for the elements (ij), $i = 1, ..., m$ rows and $j = 1, ...n$ columns, and $\tilde{a}_{ij}$ is the element normalized s.t. $\sum_j \tilde{a}_{ij} = 1$. The sum of all elements in priority vector is 1 and this shows relative weights among the things that we compare.

Then the priority vector result:

$$\mathcal{W} = \frac{1}{k}\begin{bmatrix} \tilde{a}_{11}+ & \tilde{a}_{12}+ & ... & +\tilde{a}_{1m} \\ \tilde{a}_{21}+ & \tilde{a}_{22}+ & ... & +\tilde{a}_{2m} \\ \tilde{a}_{31}+ & \tilde{a}_{32}+ & ... & +\tilde{a}_{3m} \\ \tilde{a}_{n1}+ & \tilde{a}_{n2}+ & ... & +\tilde{a}_{nm} \end{bmatrix}$$

in compact way:

$$\bar{\mathcal{W}} = [w_1, w_2, ..., w_k]' \tag{3.2}$$

From the relative weight, we can also check the consistency of the matrix. Principal eigenvalue is obtained by the sum of products between each element of eigenvector and the sum of columns of the reciprocal matrix

$$\lambda_{max} = \sum_{i}^{k}\sum_{j}^{n} a_{ij} \cdot w_i \tag{3.3}$$

the comparison matrix $\mathcal{L}$ is consistent if $a_{ij} \cdot a_{ji} = a_{ik}\ \forall i, j, k$ and Saaty proved that for consistent reciprocal matrix the largest eigenvalue is equal to the size of comparison matrix, $\lambda_{max} = n$. Then he gave a measure of consistency called *Consistency Index* defined by:

$$CI = \frac{\lambda_{max} - n}{n - 1} \tag{3.4}$$

Once we have obtained the weights $w_i \in [0, 1]$ of the vector $\mathcal{W}$ we can construct our ensemble model $\mathcal{M}$, it is therefore necessary to define the set of learners l_i, $i = 1, ..., n$ according to the problem to be faced which whether classification, regression or clustering, in the following paragraphs the associated methodology is highlighted.

3.2.2 Regression and classification problems

For the regression problem considering for each learner l_i, $i = 1, ..., k$, the output $o_i = l_i(\mathbf{x})$

$$\mathcal{M}(l_i, \mathbf{w}) = \frac{1}{k}\sum_{i}^{k} l_i(\mathbf{x}) \cdot w_i \tag{3.5}$$

that is the final ouput of ensemble algorithms. For the classification problem we can compute a weighted majority vote by associating a weight w_j with classifier l_i:

$$\mathcal{M}(l_i, \mathbf{w}) = \arg\max_{i} \sum_{j=1}^{m} w_j \cdot I_c\left(l_j(\mathbf{x}) = i\right) \tag{3.6}$$

where I_c is the indicator function and C is the set of unique class labels. Considering the predicted probabilities p_j for the j-th classifier, we can write:

$$\mathcal{M}(p_j, \mathbf{w}) = \arg\max_j \sum_{j=1}^{m} w_j \cdot p_j \tag{3.7}$$

This method ensures that the weights usually assigned in the ensemble methods are determined by the AHP method, under the supervision of the expert who can attribute a weight based on experience and not only on the performance of the algorithm, incorporating within itself the knowledge of the phenomenon in a more expert-driven approach than just black-box or data-driven. Once the final ouput is obtained, the interpretability of the CDSS must be considered and in this work are presented different approaches based on what have just seen.

3.2.3 Clustering problems

For the clustering problems in the context of unsupervised learning, in this work the proposed methodology is similar to the one detailed in the previous paragraph on classification and regression. Starting from table 3.1 and from the $\mathcal{L}$ matrix we consider $c_i(\mathbf{x})$ learners, $i = 1, ..., k$ where $\mathbf{x}$ is the features space then we can define the Cluster Stacking Algorithm (CSA) considering at the first level the application of n - learners $c_i(\mathbf{x})$, each weighted with the weight determined through steps 3.1-3.4 such that

$$C(\mathbf{x}) = \sum_i w_i \cdot I_{(c_i(\mathbf{x})=k)} \tag{3.8}$$

as done in the case of classification and regression once we have obtained the labels for each model we move on to the second level where we use this output as an input for a *meta clusterizer*, defined $c_m(\mathbf{x})$. Once the meta-clusterizer is applied we may solve the following problem

$$l^* = \arg\max_r [c_1(\mathbf{x}|s_1), ..., c_k(\mathbf{x}|s_k), c_m(\mathbf{x}|s_j)] \tag{3.9}$$

with

$$s_j = \frac{(I_j / \sum_j I_j)}{v_j \cdot w_j} \tag{3.10}$$

where l^* is the final label for the $i-$th sample in the dataset $\mathcal{D}$ and v_i, $i = 1, ..., m$ is the *v-measure* obtained by

$$v_i = \frac{1 + \gamma \cdot h \cdot c}{\gamma \cdot h + c} \tag{3.11}$$

I_j is the intensity (table 3.1) gives by the expert for each model j, w_j is the weight obtained by (3.1-3.4).

For $\gamma > 1$ it will be given more weight to the completeness, if less homogeneity i'll be more important (fur details see [25]). A further method that can be used is based on majority voting, already discussed in 2.4 whose solution is given by the following consensus function

$$l^m = \text{mode}(c_1(\mathbf{x}), ..., c_k(\mathbf{x}), c_m(\mathbf{x})) \tag{3.12}$$

Defining

$$z_i = \begin{cases} 1 & \text{if } l_j(\mathbf{X}) = \text{mode}(l_1, ..., l_k) \\ 0 & \text{otherwise} \end{cases} \tag{3.13}$$

where $l_j(\mathbf{X}) = C_{j,k}$ is the k-th cluster's label for the j-th clusterizer. $\text{mode}(l_1, ..., l_k) = C_k^m$ is the mode cluster's label obtained by the (3.12). Finally, the chosen algorithm will be the one that maximize the following decision function:

$$d(z_i, w_i) = \frac{1}{N} \sum_i z_i \cdot w_i \tag{3.14}$$

where w_i the expert's weights for the i-th clusterizer and $C_j = l_j(\mathbf{X})$ for $i \neq j$ is the j-th cluster's label.

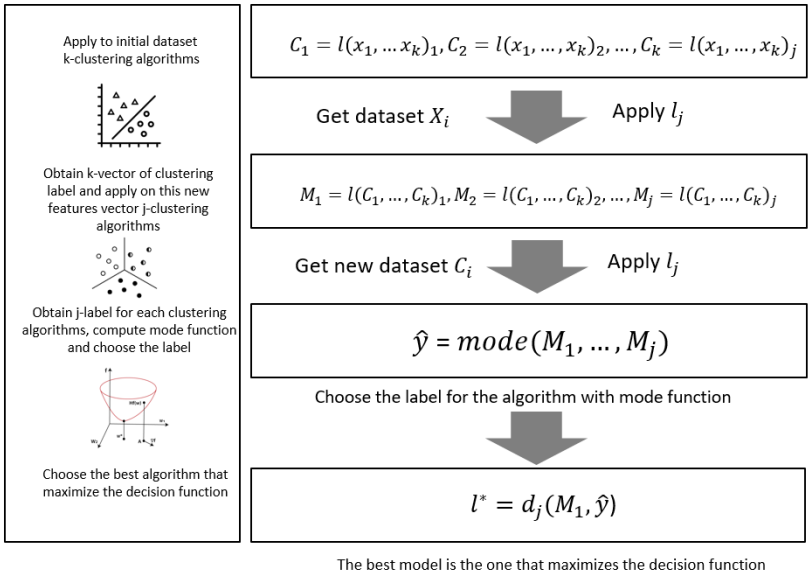

Figure 3.2. framework clustering ensemble method

3.2.4 Evaluation

In order to evaluate the weight given by the expert to each model in the pre-analysis phase (both classification regression and clustering problems), we introduce the quantity ϕ_i which measures the (coherence) difference squared of the weight of the expert and the importance $\beta_{i,j}$ of the model i-th for the j-th cluster obtained by the meta-learner (i.e. Logistic regression, Decision Trees, XGBoost, etc.) in the post-analysis phase

$$\phi(\mathbf{w}, \beta) = \frac{1}{w_i} \sum_{j}^{n} (\beta_{i,j} - w_i)^2 \tag{3.15}$$

To evaluate the correctness of the assignment of the weights to the individual models that the expert has carried out, we introduce the following problem of constrained minimum:

$$\begin{aligned} \min_{w_i^*} & \sum_i (I_i w_i - \tilde{w}_i)^2 \\ \text{s.t.} & \sum_i w_i = 1 \\ \text{and } & w_i \geq 0 \end{aligned} \tag{3.16}$$

The final evaluation result

$$w_i^{ratio} = \frac{|w_i - w_i^*|}{w_i} \tag{3.17}$$

parameter	**description**
$\tilde{w}_i$	expert's weight for the model i-th
I_i	expert's intensity for the model i-th
variable	**description**
w_i^*	optimal value that minimize the bias between complexity and expert's weight

3.2.5 Explainability proposed methods

For a classification problem, binary or multiclass, starting from the confusion matrix consider the accuracy value of each single learner l_i denoted by $\mathbf{z}_i$, for each feature $\mathbf{x}_i$ standardized with the following formula $\mathbf{x}_s = (\mathbf{x}_i - \bar{\mathbf{x}})/\sigma_{\mathbf{x}}$ in the dataset $\mathcal{D}$ in this work is introduced the explainability e_i, that can be computed by

$$e_{\mathbf{x_i}} = \sum_i \mathbf{x}_s \cdot w_i \cdot \mathbf{z}_i^{-1} \tag{3.18}$$

the (3.18) indicates how important a feature is to accuracy, in proportion to each criterion used (learner l_i) with the weight w_i. For the evalutation on the overall accuracy of the ensemble then

$$\tilde{e}_{\mathbf{x_i}} = \sum_i \mathbf{x}_s \cdot w_i \cdot \mathbf{Z}^{-1} \tag{3.19}$$

where $\mathbf{Z}$ is the general accuracy obtained by the ensemble such that $\mathbf{Z} \geq z_i$, for $i = 1, ..., k$ and further more is possible consider the proportion for each learner l_i of the quote of accuracy on the overall

$$e_{l_i} = z_i \cdot w_i \cdot \mathbf{Z}^{-1} \tag{3.20}$$

defined PAR (Proportion Accuracy Rate).

Another method of explainability of the features presented in this work considers the class of binary problems where the probabilities of belonging to class 0 or 1 are produced for the ensemble $\mathcal{M}$. By indicating with $p_i = P(Y|C = i)$ the probability that the output Y given the class C is equal to i with $i = 0, 1$, for each classifier l_i defining the following Features Function Explanation (FFE)

$$\mathcal{E}(\mathbf{x_i}|C = i) = \sum_i p_i \cdot \mathbf{x_i^T} \cdot (w_j \cdot \mathbf{z_j})^{-1} \tag{3.21}$$

where $\mathbf{z_j}$ is the accuracy value for the j-th classifier l_i, w_j is the weight for the classifier j and $\mathbf{x_i}$ is the i-th feature (standardized) in the dataset $\mathcal{D}$.

3.3 Models and results for cervical cancer detection

Starting from the work of Sobar *et. al.* [26], the authors investigate how to predict a certain type of cervical cancer in advance. The data available to the authors come from a questionnaire distribuited to 72 individuals, including 50 without cervical cancer and 22 with the disease. The study was conducted in Jakarta in Indonesia. In order to predict the type of cancer studied in advance, the authors apply two well-known machine learning algorithms, Logistic regression and the Naïve Bayes classifier. The authors achieve very good results for each model the accuracy value is 91.67% and 87.5% respectively, while the AUC values 0.96 and 0.97 respectively. However the analysis conducted by the authors was not based on the interpretation and explainability of the results.

Data Processing
The data that were used by the authors did not require a processing as they are already encoded and without missing values. The data available to the authors come from a questionnaire administered in a specialized center in Jakarta (Indonesia) distribuited to 72 individuals, including 50 without cervical cancer and 22 with the disease. The attributes considered are 19 and as often happens in the medical-health context the dataset is not very large. The authors in the questions posed in the questionnaire consider behavior from the point of view of social science and psychology. The areas considered are: the theory related to common behavior (The Health Belief Model) or the theory of protection motivation (PMT), the theory of behavior planning (TPB), the social cognitive theory (SCT) and others.

In figure 3.3 we can see with respect to the presence or absence of cervical cancer, how the values for the features concerning attitude, empowerment and behavior are distributed, the scores on the abscissas indicate the level of the ordinal variable. We note that for sexual behavior awareness in people without cervical cancer is much higher, as well as for personal hygiene and the attitude to spontaneity.

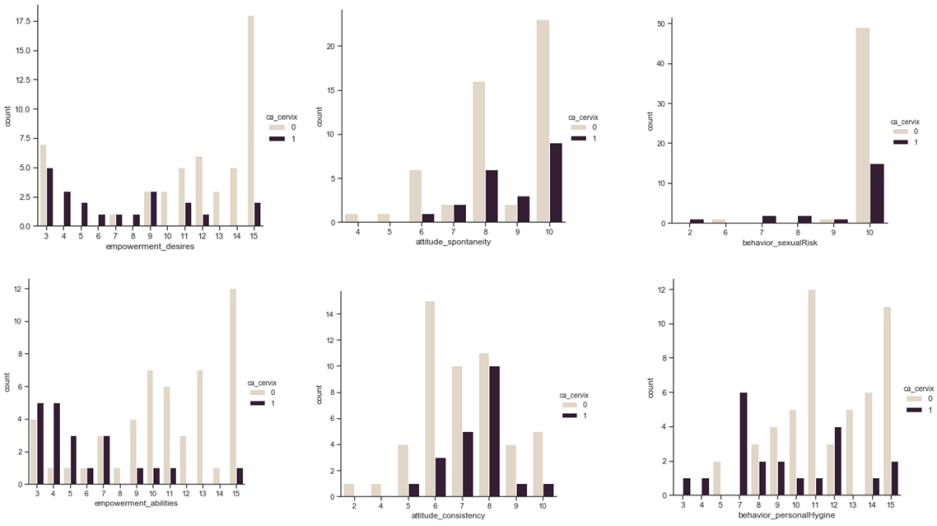

Figure 3.3. features representation for attitude, empowerment and behavior

Methodology

So starting from their work and with the data used, a new classification method was developed based on the methodology proposed in this thesis. An ensemble classifier was built using the weights, for each algorithm, obtained by applying the AHP method. Specifically, three algorithms were considered in order to show the validity of the proposed methodology and evaluate its use in the clinical-health field. As first model it was used the logistic regression (LR) was considered, starting from the criteria table 3.1 defined, is was assigned the value 1 because the model is very explainable and interpretable, based on the assumptions made earlier and discussed extensively in previous chapters.

Evaluations				
Algorithm	**LR**	**DT**	**MLP**	**EM**
error on train set	0	0.3333	0	0.0751
error on test set	0.0555	0.1666	0.1111	0.0859
weight assigned	0.6788	0.2254	0.0956	
intensity	1	3	7	

Table 3.2. Error comparison on train and test set

The second model choosen, a slightly more complex but still widely used, is a decision tree (DT), which the expert has assigned 3; the third and last model is a Multi Layer Perceptron (MLP) which the expert has instead assigned 7 due to its complexity given by the non-linearity and complexity in the parameters. It is assumed that the values assigned to the algorithms have been assigned by the expert, who supervises the entire decision-making process together with clinical-health personnel, who will then be the user of the model and who will validate the results. Considering what has been said the priority weights were obtained by the table 3.1 and applying (3.1-3.2);

once the weights and probabilities of each class for each of the three algorithms used were obtained the ensemble model (EM) was obtained by (3.6).

Statistics				
	Precision	**Recall**	**F1-Score**	**Support**
Cancer = No	0.94	1.00	0.97	15
Cancer = Yes	1.00	0.67	0.80	3
Accuracy			0.945	18
macro avg	0.97	0.83	0.88	18
weighted avg	0.95	0.94	0.94	18

Table 3.3. classification report for Meta-Classifier

The results obtained were interesting, for each single classifier used a very high accuracy value was not obtained, while through the use of the ensemble model value is was obtained a precision of **94.5**% and a value AUC of **0.98**; one point higher than that obtained in the work of the authors [26]. Also in terms of accuracy we can observe almost 3 percentage points more than that obtained with the Logistic regression used by the authors. Each model individually obtained an accuracy value equal to 94% for the LR, equal to 83 % for the DT and 88 % for the MLP. The result obtained by the ensemble is therefore very valid.

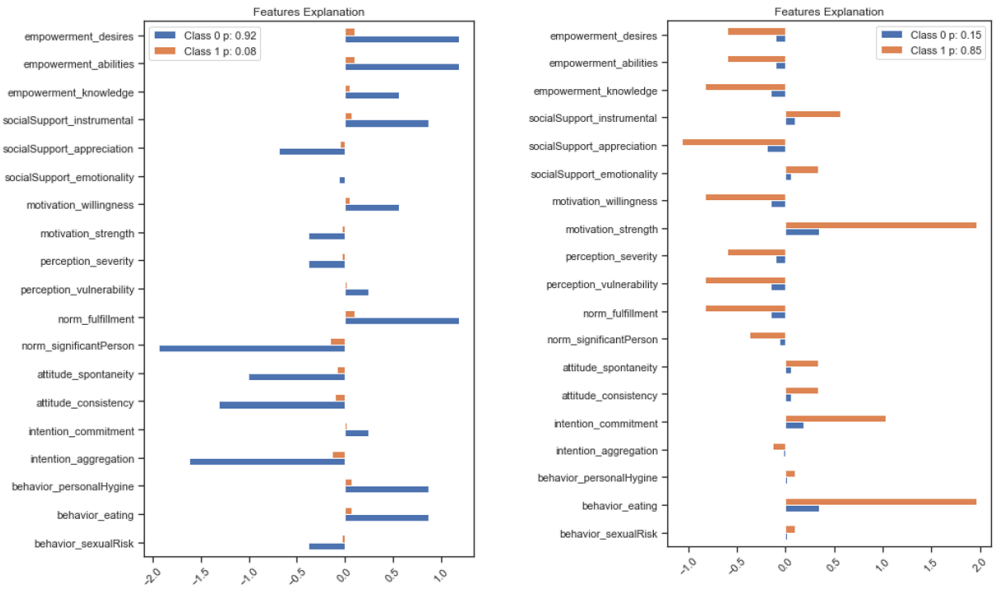

Figure 3.4. features explanation for two examples on test set

Explanation

The authors considered 8 variables, like a behavior, intention, attitude, motivation and other social characteristics. Starting from these variables, by the application of (3.21) for each algorithm used, it is possible to compare the individual results for two distinct observations of the explainability for the features shows in figure 3.4.

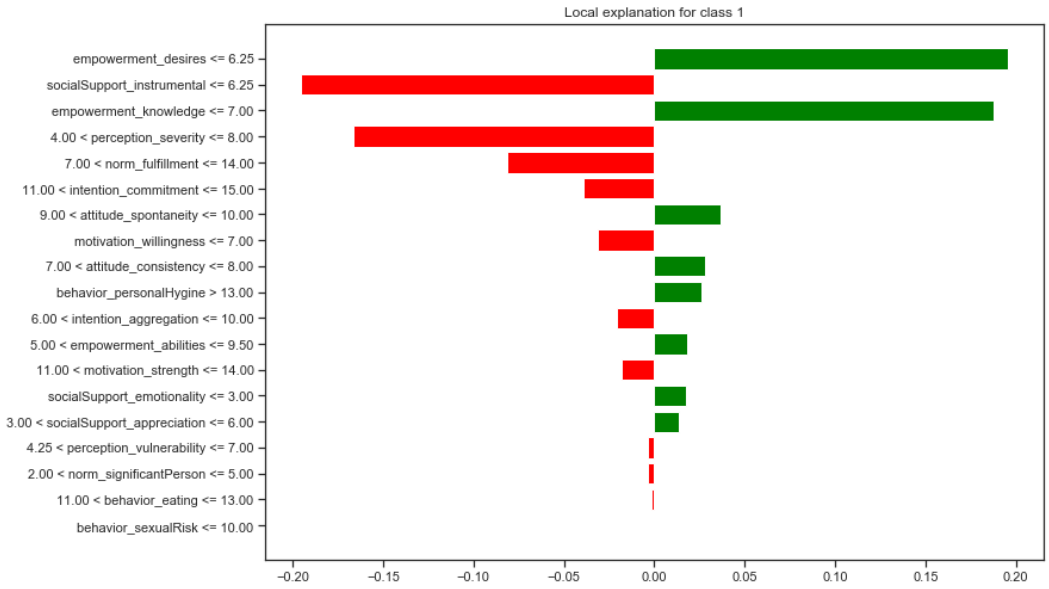

Figure 3.5. features explanation with LIME

Considering the consistency matrix used by the expert for the assignment of the values for each model, the consistency index value (CI) (see formula 3.4) is **0.008** and therefore according to Saaty's definition we can consider it consistent. Using the value of the Random Index (RI) to relate the CI we obtain the value of the coherence coefficient (CR) which is equal to **1.48**%, being less than 10% by definition there is no inconsistency in the attribution of the subjective judgment to expert algorithms. From table 3.2 we can deduce the error values on the train set and on the test set. As we can see in the LR algorithm we are in the presence of a slight overfitting value while for the DT the error is high, about double on the training data. As for the MLP here too we notice a slight overfitting but the interesting thing is the result on the ensemble (EM), as by definition this ensemble method has both removed the overfitting as the value of the error on the train and test data (**0.075**, **0.085** respectively) it is very close, and it has drastically reduced the overall error compared to single models. In plot 3.4 we can observe the contribution of each features with respect to the predicted class 0-1 in terms of probability, weight assigned by the expert to the model and accuracy of the EM classifier. The values are normalized, i.e. we note how some features have a highly negative contribution such as the *attitude_spontaneity* for the class 0 (left side plot) in which there is a no risk of cervical cancer, instead is positive for the class 1 (right side plot) in agreement with LIME method. Figure 3.5 and 3.6 shows the features values for the positive class (Cancer: Yes) obtained through the **LIME** method. The values refer to the local explainability in a neighborhood of a given instance and the assumed values, negative or positive reported on the abscissa axis indicate how much would describe or increase respectively, the probability of onset of cervical cancer given a given probability predicted.

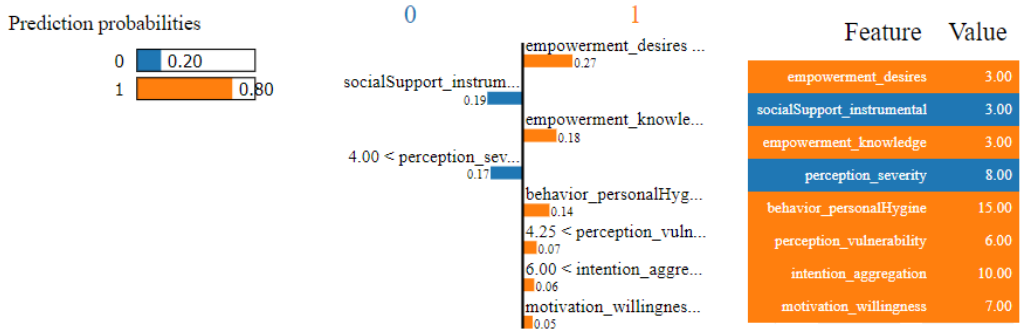

Figure 3.6. features explanation with LIME

Final remarks

Cervical cancer is a problem that affects the female population worldwide, constituting a very high risk factor. In this application, starting from works that have already dealt with the subject, a new method for the prediction of this disease has been proposed. Through the application of three supervised machine learning algorithms, a binary classifier was built capable of classifying with an accuracy of 94.5 % the new instances that present risk characteristics related to the development of the disease. The expert assigned each algorithm individually a score which, through the methodological procedures described in the dedicated section, has become a weight attributed to the complexity and interpretability of the model itself. Compared to the [26] authors' benchmark, the ensemble classifier based on the expert's knowledge led to better results in terms of accuracy and AUC (equal to 98 %). From the techniques used to explain the results associated with the models used (figure 3.4) it emerges that behaviors related to nutrition or social support, rather than the perception of vulnerability, can be considered characteristics that influence the possibility of developing this disease. There are also limitations that must be considered: the data are not many and therefore the accuracy of the model could be improved through more examples despite the very low classification error on the train and test data set (0.075 and 0.085 respectively) and since the difference between the two is very small (<1 %) this implies that the model is not affected by overfitting; it would have been possible to use many more algorithms and then combine them, but this could have been a problem in the weighting phase, therefore limiting to a maximum of five models could represent a good trade off, also in light of the determination of the weights through the method AHP, also considering the size of the comparison matrix. The proposed method showed that the matrix of scores assigned by the expert is consistent (<0.001) with his judgment and also the relative CI (<1.5 %). Overall, the proposed tool is easy to implement and in the clinical context it could represent a valid modeling choice.

3.4 Models and results for diabetes predictions

Diabetes is a chronic disease characterized by an excess of glucose in the blood, the International Diabetes Federation has estimated an alarming rise in the number of diabetics by the year 2030 [27],[28]. This disease is divided into two forms, type 1

diabetes and type 2 diabetes. Hyperglycemia can be caused by insufficient insulin production (i.e. the hormone that regulates the level of glucose in the blood) or by its inadequate action. Type 1 diabetes is characterized by the total absence of insulin secretion, while type 2 diabetes is determined by a reduced sensitivity of the organism to insulin and this disease can progressively worsen over time and is established on the basis of a pre-existing condition of insulin resistance. Type 2 diabetes is a disease with a high spread all over the world also due to the lifestyle of today, such as an unhealthy diet and/or little or no physical activity. In type 1 diabetes, affected people must necessarily take insulin by injection, in type 2 diabetes an appropriate drug therapy associated with a healthy lifestyle allows to contain the negative effects of the disease. Often the presence of hyperglycemia does not give any symptoms or signs, for this reason diabetes is considered a subtle disease. The associated symptomatology in acute cases is characterized by fatigue, increased thirst (polydipsia), increased diuresis (polyuria), unsolicited weight loss, sometimes even concomitant with increased appetite, malaise, abdominal pain, up to to arrive, in the most serious cases, to mental confusion and loss of consciousness. The major complications deriving from diabetes can cause the patient various damages, which are divided into:

1. Ocular (retinopathy): caused by chronic hyperglycemia and hypertension leading to alteration of blood vessels with consequent worsening of vision up to blindness
2. Cardio-cerebrovascular: myocardial infarction or ischemic heart disease, stroke
3. Renal (nephropathy): damage to the filtering structures of the kidney which can lead in extreme cases to dialysis
4. Neurological (neuropathy): anatomical and functional alteration of the central, peripheral and Voluntary nervous system, sensory, motor, visual, acoustic deficits

According to scientific studies, the individuals who are most likely to develop diabetes are:

(a) Fasting blood glucose between 100 and 126 mg/dl

(b) First degree family members for type 2 diabetes

(c) Body Mass Index, i.e. weight ratio in kilos/height in m2, with a value > 25 kg/m2

Data processing
Starting from the Pima Indian Diabetes Database (PIDD) provided by the National Institute of Diabetes, Digestive and Kidney Diseases, several authors have proposed algorithms and methods to predict and classify diabetes. The data set consists of 768 patients (called *examples*) each with 9 numerical features and the data refer to women aged 21 to 81 years. The target variable under study is the class variable

(diabetes = 1 (yes), diabetes = 0 (no)) [29]. Deepti and Dilip [30] use PIDD data for the classification of diabetes through three machine learning algorithms, the authors specifically apply Naive Bayes (NB), Decision tree and Support Vector Machine, obtaining respectively in terms of accuracy a value of **76.30**% for the NB, 73.82% for the DT and the lowest for the SVM equal to 65.10% and the maximum recall value is reached by the NB equal to 0.763. Han *et al.* Han [31] again on this PIDD dataset apply an algorithm based on two steps, in the first they apply an improved k-means and in the second step a Logistic regression. Through this method the authors reach an accuracy of 3% higher than the results present in other works (**95.42**%), such as that of Patil *et. al.* [32]; the authors obtain an accuracy result equal to 92.38% through their method called Hybrid Prediction Model (HPM) which uses the Simple k-means clustering algorithm aimed at validating the chosen class label data (incorrectly classified instances are removed, the model is extracted from the original data) and then apply the classification algorithm to the resulting data set. The C4.5 algorithm is used to create the final classification model using the k-fold cross-validation method. The purpose of this case study is not to obtain a better classifier in terms of metrics like accuracy, recall and precision, but rather to provide a valid method in terms of explainability of these results that in the authors cited in literature examined have not provided. The variables of PIDD data are:

1. Number of times pregnant
2. 2-hour OGTT plasma glucose
3. Diastolic blood pressure
4. Triceps skin fold thickness
5. 2-hour serum insulin
6. BMI
7. Diabetes pedigree function
8. Age
9. Status ((diabetes = 1 (yes), diabetes = 0 (no))

Methodology

Starting from the results obtained from the work carried out on this type of dataset (PIDD) is was built an ensemble-type classifier based on six algorithms. Specifically the models considered were a Support Vector Classifier (**SVC**) to which the expert gave a complexity (intensity) score of 4, while the second model considered is a **KNN** to which 1 was assigned due to its simplicity in the explanation of the parameters and in the interpretation of the results. The third model considered is a feed forward neural network (**NN**) to which the expert assigned a complexity score of 7, due to its complexity in the parameters. A Gradient Boosting (**GB**) with a complexity value of 6 was also considered, being an ensemble of random forests which in turn are fairly simple models in interpretation. Subsequently a Gaussian Process Classifier (**GPC**)

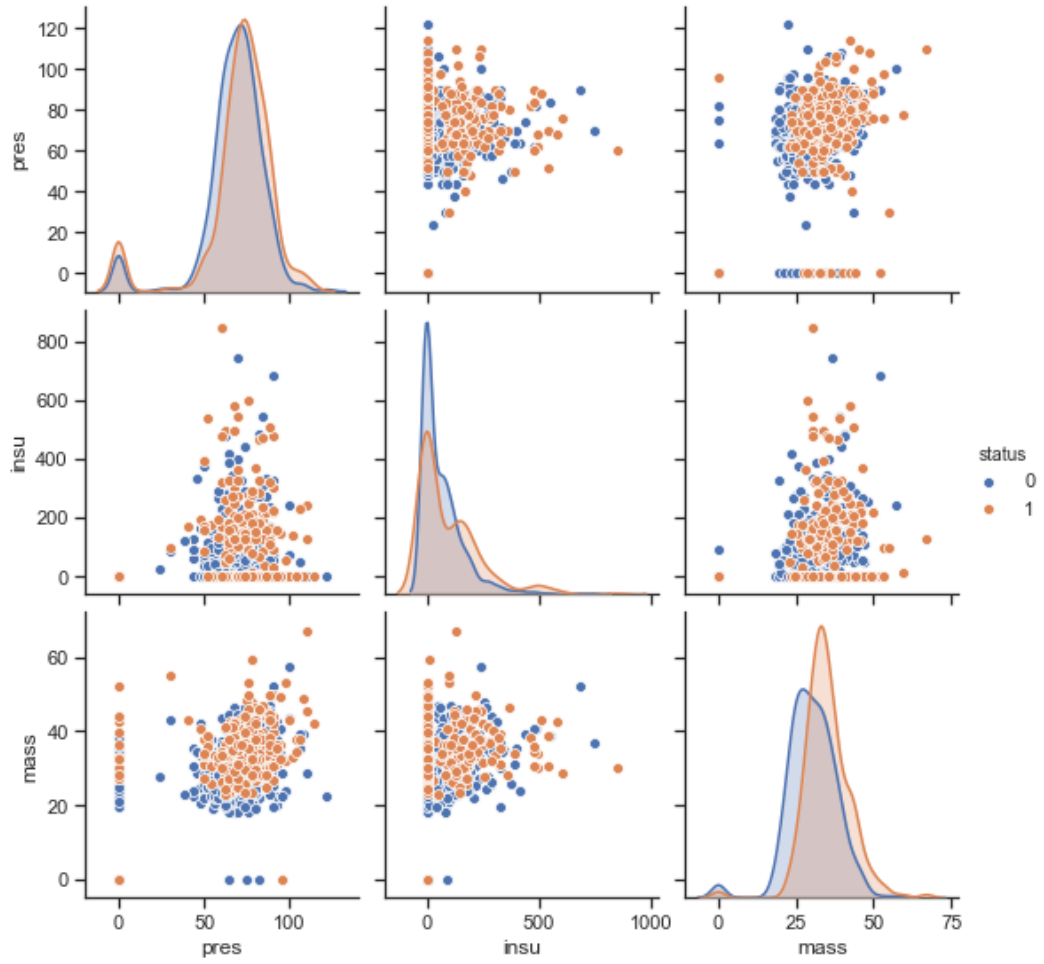

Figure 3.7. features distributions

is was inserted in the construction of the ensemble to which was given a complexity score of 9, being a model that works with radial Kernel functions therefore quite complex to interpret locally. The last model inserted is a Logistic regression (**LR**) to which the expert assigned a score of 2. By determining the weights for each model through the AHP-based methodology and by the scores assigned by the expert the ensemble is was trained on the training data and by the Voting Classifier method (hard voting) the value obtained in terms of accuracy is equal to **78**%.

From plot 3.7 it is possible to observe the distribution of some of the main variables such as body mass index, insulin values and blood pressure, respect to the status of diabetic or not indicated by the binary variable 0-1. It is possible to note that there are linearly no important relationships between the variables shown in the plot.

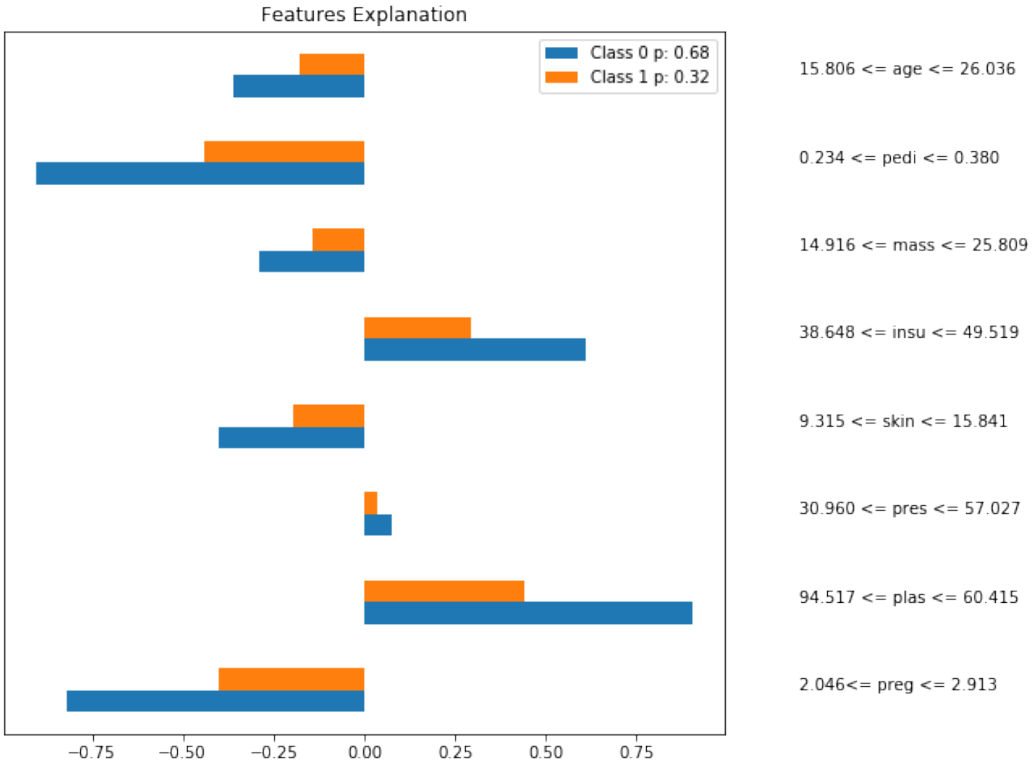

Figure 3.8. Explanation with *FFF* see 3.12

Algorithm	**SVC**	**KNN**	**NN**	**GB**	**GPC**	**LR**	**error**	**EC**
accuracy on train	0.76	0.77	0.75	0.733	0.77	0.75	0.755 ±0.012	0.78
accuracy on test	0.75	0.75	0.74	0.731	0.76	0.748	0.746 ± 0.009	0.753
gap	0.01	0.02	0.01	0.002	0.01	0.001	0.008 ± 0.006	0.027
gap %	1.6	2.30	1.5	0.28	1.6	0.2	1.24 ± 0.76	3.46
recall	0.72	0.70	0.68	0.60	0.69	0.71	0.68 ± 0.04	0.716
precision	0.71	0.71	0.70	0.72	0.72	0.73	0.715 ± 0.009	0.753
weight assigned	0.126	0.446	0.070	0.105	0.035	0.214	-	-
intensity	4	1	7	6	9	2	-	-

Table 3.4. performances metrics

Explanation

In table 3.4 we can observe the metrics obtained for each model used at single level, at medium level and compared with respect to the metrics obtained by the ensemble-classifier (EC) obtained by combining the single models weighted with a hard voting procedure with the assigned weights by the expert through the determination with the AHP method. The classic metrics for the evaluation of classification algorithms have been calculated like accuracy, on the train data and on the test data, the recall that indicates a measure of sensitivity of the model and represents the ratio between the correct classifications for a class on the total of cases and the precision which indicates the ratio between the number of correct classifications of a given event (in this case a class) over the total number of times that the model classifies it. Metrics that measure the error between train and test (relative and percentage)

were also considered to give a global dimension of the error; moreover, all these evaluations for each model were compared on an average level with those of the ensemble classifier proposed in this work. The following case study compared to the benchmarks achieved by the other studies carried out on this dataset is placed in an intermediate way since, in terms of accuracy, it does not exceed the results obtained but the purpose of the work is to interpret the results and predictions from the point from a clinical point of view, assuming the interaction between the expert (who chooses the weights and models) and the expert in the clinical domain, thus making the decision-making process more transparent. As reported in table 3.4, taking the models individually, such as KNN or GPC, we exceed 76.30% of the work of [30] with an accuracy of 77% but we are almost 2 percentage points below compared to other works, on the whole the EC ensemble weighted model reaches an accuracy of **78**% higher than the performances of Deepti and Dilip, in the works cited the explainability of the results and predictions is not considered. We can also note that the EC model has higher precision and recall than the single models used. As regards the assigned weights, the matrix of the pairwise comparisons shows a consistency value (CI) equal to **0.067** as regards the consistency of the judgments given by the expert measured with the CR index, we have that the percentage value is equal to **5.4**%, therefore less than 10% thus the intensity assigned as a score to the individual algorithms, in terms of trade-off between complexity and explainability is fully consistent and coeherent with the evaluations of the expert.

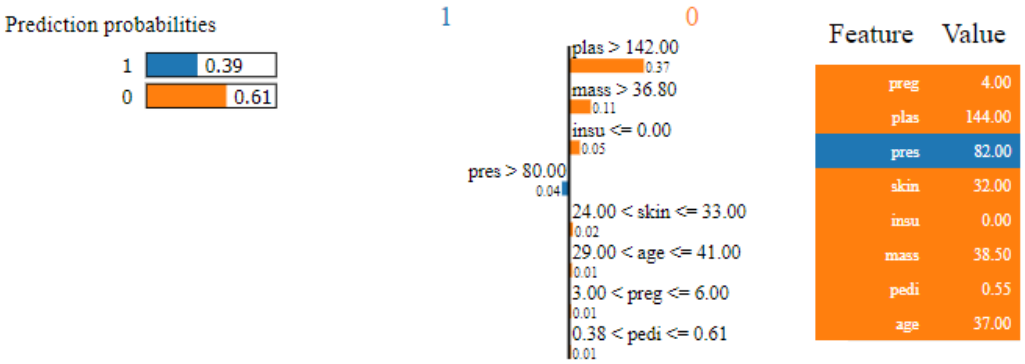

Figure 3.9. LIME Explanations

Figure 3.9 shows the results obtained by the **LIME** method for a given observation selected from the predictions made on the test dataset. It's can see, for this particular instance (patient) the probability of developing diabetes is 39% and this probability decreases by 4% when the blood pressure value (feature *pres*) is greater than 80. When the body mass value measured through the BMI feature is greater than 36.80, for this patient the probability of not developing diabetes (class 0) increases by 11%. From figure 3.10 we can see how the values increase and decrease the probability of not developing diabetes as a function of the values assumed by the features involved. In figure 3.8 we can observe the results obtained in terms of explainability of the method implemented through formula 3.21 relating to a specific instance (patient)

selected randomly in order to make a comparison with the LIME method represented in figure 3.9.

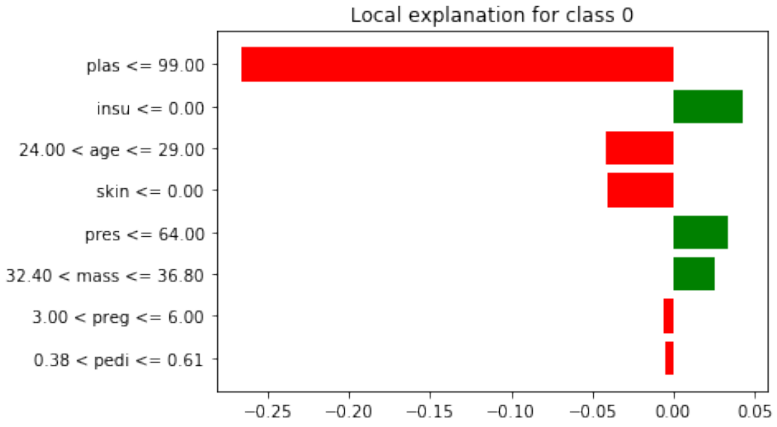

Figure 3.10. features explanation with LIME

Final remarks

In a clinical context, in addition to evaluating the efficiency of the models used, it is also important to be able to break down the elements that constitute the problem. Through the application of the methods of explainability and interpretability of the results, Lime and the one proposed in this work (see formula 3.21), we were able to validate the results obtained by considering individually the features present in the data set. The prediction of diabetes through the proposed ensemble method has an accuracy of 78% and several variables contribute to the pathological development, as shown in figure 3.8 and figure 3.9. In both figures, for a classified instance, the methods show the same phenomenon; an increase in plasma in a given range contributes with a high probability to the development of the pathology in an individual with probability of 32 % (fig. 3.8) and 39 % (fig. 3.9) respectively in the two methods used. The proposed approach also allows through the use of the weights determined by the expert to assign a specific weight to each decision maker (ie SVC, KNN, LR, etc.) and therefore to be able to carry out a global assessment based on both predictive and interpretable ability. models. The consistency in the assignment of weights is supported by the results obtained in terms of CI (<0.06) and CR (<5.4 %). The method proposed in this thesis is validated by the results obtained and therefore fully usable within a clinical decision-making process.

3.5 Models and results for clustering DED patients

In order to complete the application of the methods proposed in this work, in this last case study a dataset concerning a case-control study by Agrawal *et. al.* [33] will be treated on HIV-infected patients (type 1). The authors review and compare data from 34 HIV-infected patients and 32 control patient observations, in order to: "study the tear cytokine profile in HIV-infected patients with HIV Disease Dry Eye (DED) and study the association between the severity of ocular

inflammatory complications and tear cytokine levels". The methodology proposed by the authors however, does not concern a study conducted through machine learning methodologies but rather a parametric study based on a classic statistical epidemiological approach; in this application the aim is to find meaningful patterns by the ensemble clustering procedure developed in the thesis work. The method is therefore unsupervised despite the presence (if desired) of a binary variable that indicates whether the patient is HIV-infected or not. It is not a discussion of this application to predict whether a patient with certain characteristics may be affected by the disease although it is not excluded that it may be a topic for later discussion. The study involved the comparison of 41 features inherent in cytokine levels using the Luminex bead assay; the authors collected the data through recruitment in a Singapore referral eye center. The authors used Logistic regression for the study in order to understand the correlations and the statistical significance of the relationships. As mentioned, the intent of this work is to find significant patterns in the data, through an unsupervised ensemble method in accordance with the results obtained by the authors, they setate that specifically: "statistically significant differences were observed in the mean epithelial growth factor (EGF), growth-related oncogene (GRO) and gamma-induced interferon values protein 10 (IP-10) ". They also state that" EGF and IP-10 levels were higher and GRO levels were lower in DED tear HIV-infected patients compared with DED patients without HIV infection. The authors found: "no significant association between varying levels of ocular surface parameters and cytokine concentrations in HIV patients with DED", for a p-value greater than 0.05. The authors therefore conclude that: "the EGF and IP-10 values were significantly elevated and the GRO levels were lower in the tear profile of HIV patients with DED versus immunocompetent patients with DED".

Data processing

The data concern 41 cytokine-related characteristics of HIV-infected patients with DED (n = 34) and unaffected patients (n = 32) for a total of 126 observations and 44 features, these data were acquired through analyzes carried out at a clinical facility in Singapore. The data were processed by excluding the features that presented a percentage of missing values > 10 % (fig. 3.12). Therefore the following have been excluded: Eotaxin 49 %, IL-17A 17 %, IL-2 13 % and IL-3 62 %, IL-9 56 % and MIP-1a 86 %. The variables that had values lower than 10 % were imputed through the mean of the variable.

	Agglomerative	**k-Means**	**Spectral**	**Birch**
I_j	1	3	5	7
v_j	0.67	0.40	0.38	1
w_i	0.60	0.11	0.13	0.16
s_j	0.15	4.17	6.16	2.65
$d(z_i, w_i)$	0	0.08	0.03	0.05

Table 3.5. clustering algorithms

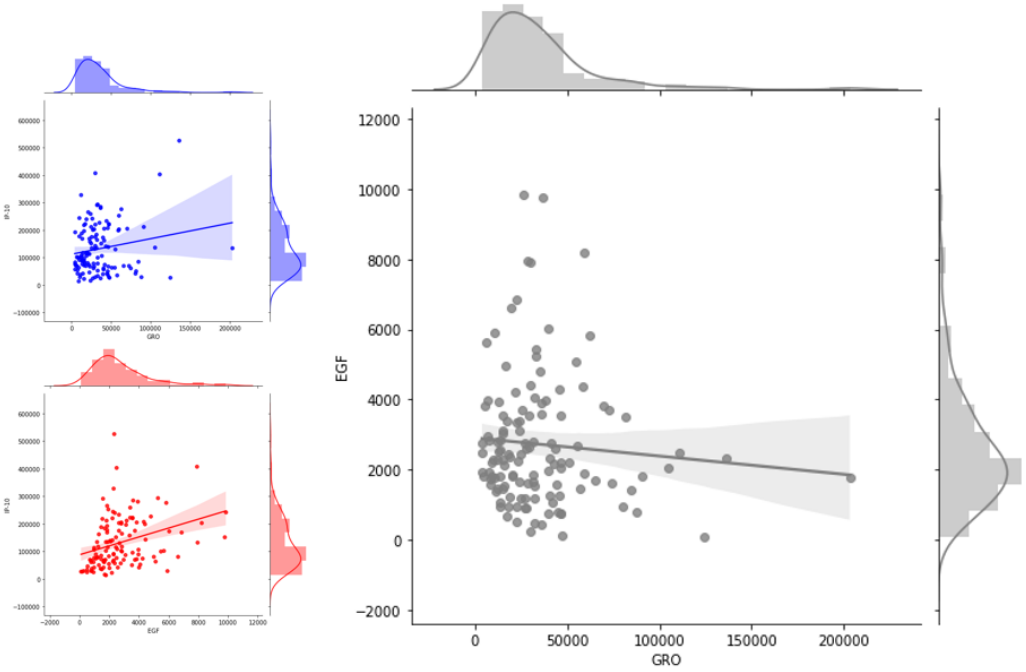

Figure 3.11. relations between EGF,GRO and IP-10

Methodology

The method proposed in this application involves the application of n-clusterizers on the initial dataset $\mathbf{X}$, once the labels have been obtained we build the new dataset $\tilde{\mathbf{X}}$; these are used to train other m-clusterizers, which in turn will produce a new output and by the application of functions (3.12-3.13) and the score function (3.14) we obtain a ranking of the *meta*-clusterizers. The table 3.5 shows the values for each clusterizer used.

$\mathbf{C_1}$	$\mathbf{C_2}$	...	$\mathbf{C_j}$	$\mathbf{C^m}$
1	1	...	0	1
0	2	0	0	0
0	0	...	1	0
2	0	2	1	2
1	1	...	2	1
...	...	...	...	...
2	1	1	$c_{j,k}$	c_k^m

Table 3.6. example of methodology

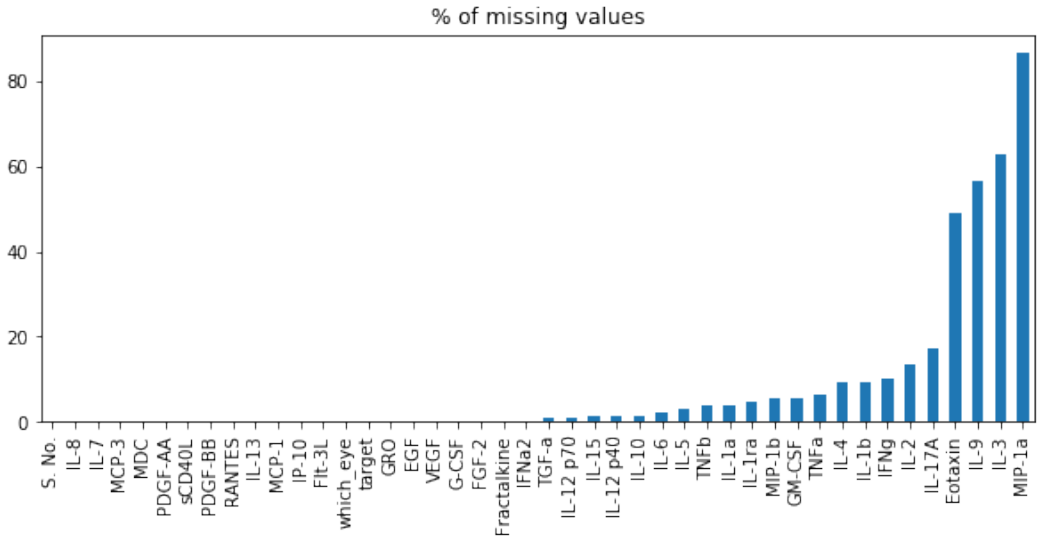

Figure 3.12. missing values

Again from the table of results 3.5 on the performance of the algorithms we can see that the Birch model obtains a v-measure value equal to 1 with perfect homogeneity, while the Spectral clustering model is the one that obtains a higher s_j score than the other algorithms clustering. The largest weight w_i assigned by the expert is that relating to the Agglomerative model equal to 0.60 despite having an intensity of importance equal to 1, thus considering it a more "simple" model in the way it works. This simplicity derives from the transparency of the way it works, we know that this method starts from the insertion of each element in a different cluster and then proceeds to the gradual unification of clusters two by two at each iteration. By determining the decision function d_j, the Agglomerative method obtains an error value equal to 0 (maximizing the number of pairs of equal labels between two models l_j and l^m, is equivalent to minimizing the difference between the different labels $\epsilon = 1 - d_j$), therefore better than the other clusterizers. The decision function compares the label of model l_j, with the value of the model label obtained by majority voting using mode function. The analysis of the eigenvalues of the consistency matrix of the scores assigned by the expert shows that the CI (consistency index) value is 0.36, while the CR index value is 40%, 4 times higher than 10% as a limit for the consistency of the expert's judgments, therefore exactly for this reason, since there is a lot of discrepancy between the weights assigned by the expert and the results obtained is introduced the method 3.15-3.17. Below is the pseudocode of the ensemble algorithm proposed in this application:

Data: Features set $\mathbf{X}$ and weights w_i
Result: Final cluster's label $\hat{C}_i$
Step1. Expert assign intensity score for each algorithm l_j, $j = 1, ..., n$
Step2. Train each algorithm on dataset $\mathbf{X}$
Step3. From each algorithm obtain the cluster's label $C_{i,k}$
Step4. Use cluster's label $C_{i,k}$ as new features set $\tilde{\mathbf{X}}$
Step5. Train $l_1, ..., l_k$ algorithms on a new meta-features set $\tilde{\mathbf{X}}$
Step6. Compute with intensity score I_j for each algorithms l_j the weight w_i
Step7. From each algorithm obtain the cluster's label $\tilde{C}_{i,k}$
Step8. Compute $l^m = \text{mode}(l_1(\tilde{C}_{i,k}), ..., l_k(\tilde{C}_{i,k}))$
Step9. Compute decision function $d(z_i, w_i)$ and choose the best algorithm
Step10. Using meta-features $\tilde{\mathbf{X}}$ and optimal label l_i^{opt} train supervised *meta*-learner and get probabilities class

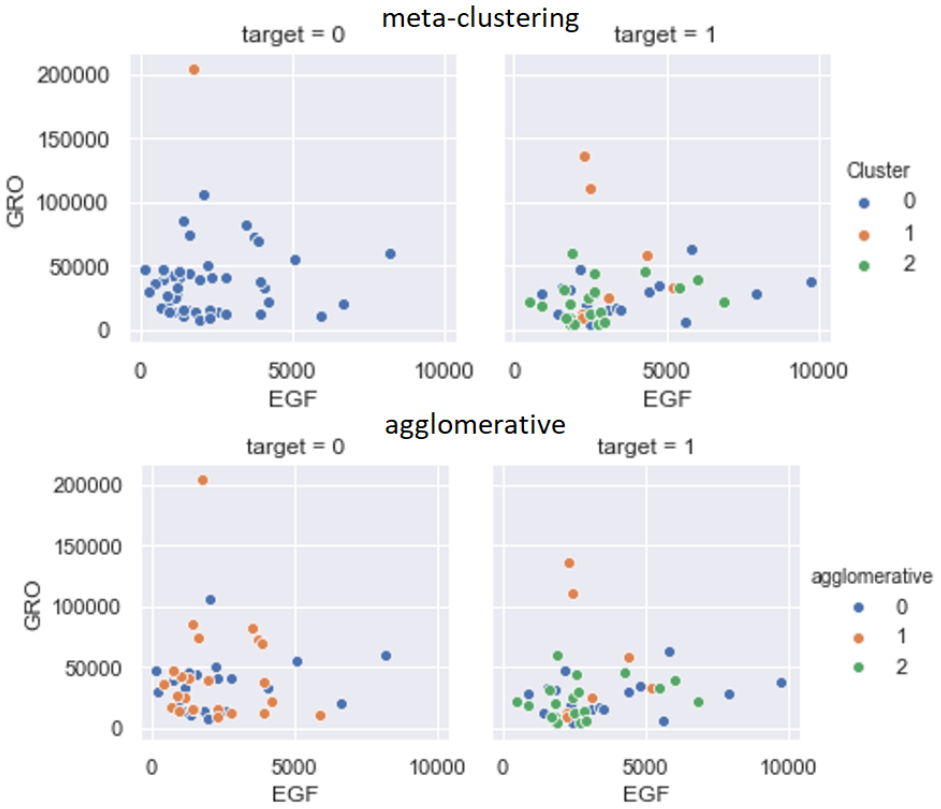

Figure 3.13. clustering results comparison

Explanation
Through Figure 3.14 we observe the results of applying the LIME method for a given observation. There are three clusters and the value of the features considered with respect to the positivity or negativity of the patients examined in the study is highlighted. From the results we note that the observation (patient ID) in question

belongs to Cluster 2 with a high probability, the values of the variable GRO and the VEGF variable are very high and each of them contributes, respectively with 7% and 10% probability on the negativity of a subject with respect to the affected disease. According to the study conducted by Agrawal *et. al.* [33] the variables involved such as GRO, EGF and IP-10, also through LIME, play an important role in the clustering algorithms used. A decrease in the value for the IP-10 variable decreases the probability of DED disease by 13%. While in the figure 3.13 the clusters obtained for the variables involved were represented and the comparison concerned the Agglomerative algorithm, which is the one that reported better performance metrics with the meta-clustering algorithm obtained by applying the 3.13. We note that the meta-clustering method manages to group the observations in a better way (left side) than the other method (bottom left) especially for Cluster 0 relative to patients with target = 0, i.e. not affected by HIV disease. Specifically, in cluster 0 for status = 0 those observations are included that have mean EGF values equal to 2235 with a range equal to (497; 3973), while for the variable GRO there are values included on the average between (13261; 58985).

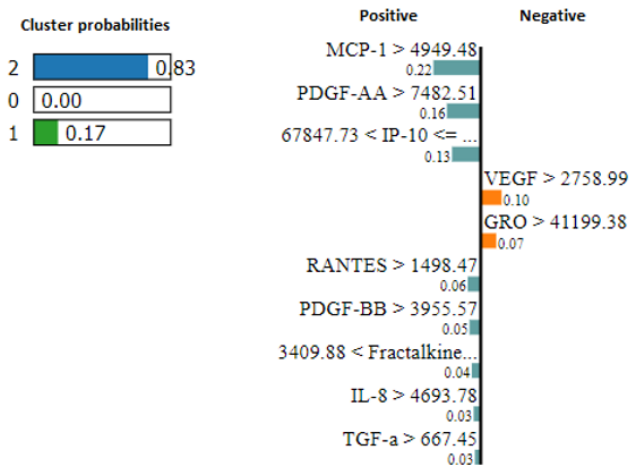

Figure 3.14. clustering explanation with LIME

To obtain the cluster probabilities, i.e. that a new predicted observation belongs to a certain cluster with a certain probability, the Logistic regression is was used as *meta*-learner motivated by the fact that this model is the most used in the analysis of both clinical, epidemiological and health phenomena, which was trained through the set of *meta*-features obtained in the second layer of the proposed methodology and the (optimal) label obtained by choosing the best algorithm evaluated with the decision function $d_j(\cdot)$. Figure 3.2 shows the logical schema of the proposed methodology, in which it is observed that in the first layer a series of algorithms chosen by the expert through the assignment of an intensity of importance and the deduction of the weights, are applied to the initial dataset. Once the labels are obtained from each model, this become the *meta*-features used in the second layer,

in which once again are applied the same algorithms of the first layer. From the results of each applied algorithm the label is deduced by applying the *mode* function, it is then compared through the decision function with the output of each algorithm.

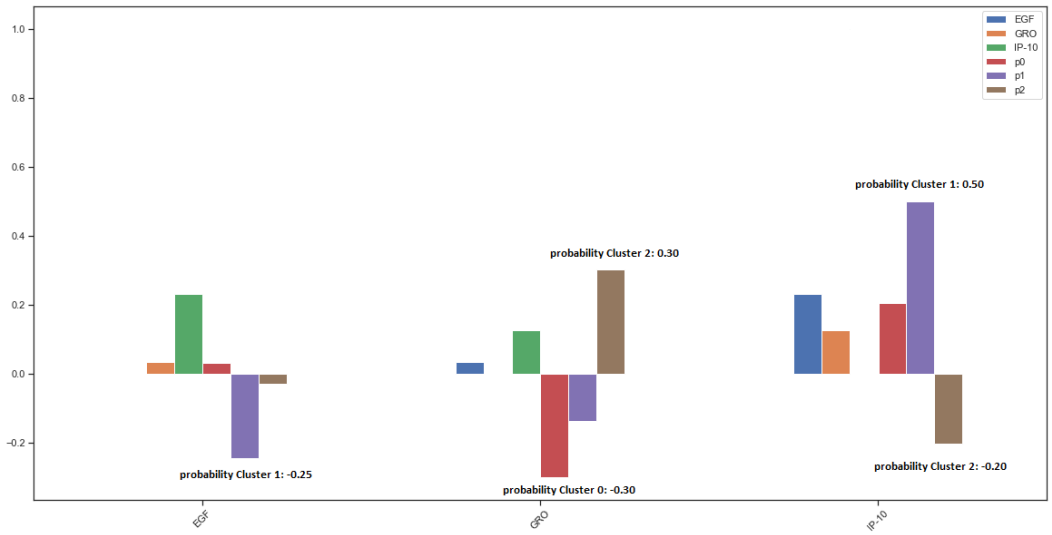

Figure 3.15. correlations for cluster's probabilities

Figure 3.15 shows the values of the linear correlation between the features considered, EGF, GRO, IP-10 and the respective probabilities of belonging of the observations to the respective three clusters. From this barplot it can be seen that for example for the IP-10 variable, the probability that a given unit belongs to Cluster 1 is on average equal to 50% and decreases linearly always for the same variable by 20% for the Cluster 2. Considering the EGF variable, the probability of belonging of the unit to Cluster 1 decreases linearly by 25% and for the variable GRO it decreases by 30% in the case of Cluster 0 but increases by 30% in the case of Cluster 2.

y=0 (probability 0.012, score -1.805) top features			y=1 (probability 0.032, score -0.793) top features			y=2 (probability 0.956, score 2.598) top features		
Contribution?	Feature	Value	Contribution?	Feature	Value	Contribution?	Feature	Value
+2.858	<BIAS>	1.000	+1.413	sp_mc	2.000	+2.721	km_mc	2.000
+0.589	km_mc	2.000	+0.782	birch_mc	1.000	+2.018	agg_mc	2.000
-0.891	birch_mc	1.000	+0.488	<BIAS>	1.000	+1.097	sp_mc	2.000
-1.852	agg_mc	2.000	-0.166	agg_mc	2.000	+0.109	birch_mc	1.000
-2.509	sp_mc	2.000	-3.310	km_mc	2.000	-3.346	<BIAS>	1.000

Figure 3.16. explanations with ELI5 of meta-features and meta-learner

Figure 3.16 reports the results obtained by the implementation of the LIME method through the Python ELI5 library; for a given observation we have that the meta-features with the greatest contribution are those relating to the k-means (km_mc) and *Agglomerative* (agg_mc) clusterizers, respectively with a contribution of +2.72 and +2.01 for the probability of belonging to Cluster 2 of 0.956, whose score is equal to 2.6. The same figure also shows the bias values for each cluster. The table then explains how each model used in the second layer contributes to the next meta-clustering phase, specifically, we calculate how much each single clustering

model contributes in the phase of to attribute of the observation to the cluster. For example in the case of an observation, the probability that the Cluster is 0, is equal to 0.012 and the contribution of the k-means model is +0.58 while the same model to which the expert has given intensity equal to 3, in the case of probability that the cluster is 2 is +2.72. This indicates that the k-means algorithm is better able to clusterizer the observations that with a certain probability belong to Cluster 2.

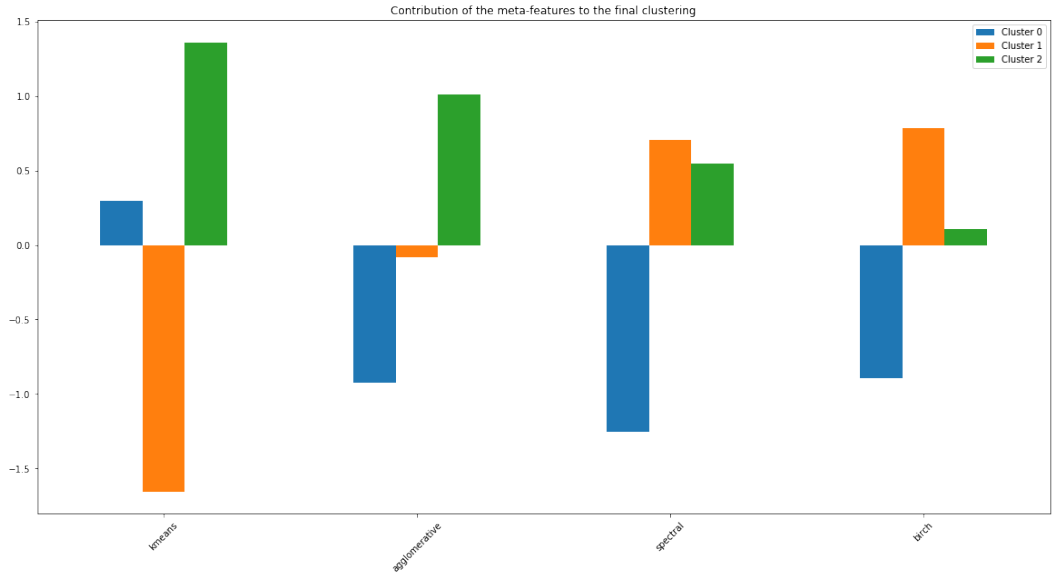

Figure 3.17. meta-features importance with logistic regression

Figure 3.17 shows the values of the importance of the features obtained by the Logistic regression meta-learner, which calculates the importance through the product of the amplitude of the coefficient by its standard deviation. It is very interesting to observe as for the assignment of the observation to one of the three clusters and each model provides a certain contribution. For example, the *Spectral* clustering contributes in a greater (but negative) way than the other models, in the assignment towards Cluster 0; while k-means in a greater (but positive) way in Cluster 2 as also highlighted with the model implemented through the library ELI5.

	Cluster 0	**Cluster 1**	**Cluster 2**	w_i	$\phi(\mathbf{w}, \beta)$
agglomerative	-0.925	-0.083	1.008	0.60	4.95
birch	-0.891	0.782	0.108	0.16	9.10
spectral	-1.254	0.706	0.548	0.13	18.27
kmeans	0.295	-1.655	1.36	0.11	42.40

Table 3.7. ranking of algorithms importance

From table 3.7 applying (3.15) we have a consistency ranking between the complexity score (I_j) initially attributed by the expert for the corresponding weight w_i and the importance of the model obtained after the application of the Logistic *meta*-learner.

It can be noted that the clustering model of the *Agglomerative* type to which the expert attributed intensity 1, is the most consistent with what was obtained with the *meta*-learner as it presents the lowest discrepancy value (inconsistency). The ϕ function assumes values between 0 and infinity a high value indicates poor consistency with what is attributed by the expert. By dividing by the maximum value we get the normalized value between 0 and 1, where 0 indicates maximum consistency and 1 total inconsistency.

	w_i	w_i^*	I_i	w_i^{ratio}
agglomerative	0.60	0.84	1	0.40
birch	0.16	0.06	7	0.625
spectral	0.13	0.05	5	0.615
kmeans	0.11	0.04	3	0.636

Table 3.8. results of the optimization

The results obtained from the solution of (3.16) and by formula (3.17) are reported in table 3.8, which also shows the value of the complexity intensity I_i, the weights w_i and the optimal values obtained for each clustering model w_i^*. It can be noted that the lowest value is always related to the *Agglomerative* model in accordance with the values assigned by the expert, the method is therefore a good way to assign and evaluate the choices made by the expert in the construction and interpretation decision-making process of the models used.

Final remarks

In this last experimental part of the thesis an unsupervised problem was faced; starting from a case-control study conducted with a clinical statistical approach, the problem was transformed into a clustering problem in order to identify consistent patterns within the data. The proposed methodology also allowed to identify the best clustering algorithm and to interpret the results. Interpretability was carried out through the use of the LIME method, highlighting the importance of some factors, such as VEGF and GRO, whose respective values increase the probability of belonging to cluster 2 by 10% and 7% respectively, of a selected patient with 83 % probability of belonging to this group. The selected groups were 3, in the third group (cluster 2) we find patients with higher concentrations of EGF and GRO values, with a high probability of developing DED (see figure 3.13). Through the use of a Logistic *meta*-classifier it was possible, using the values of the cluster methods used, to define which methods were predominant in the analysis. Using the metrics proposed in this methodology (3.15,3.17) it was possible to compare the judgment expressed by the expert in the attribution of weights, this evaluation was introduced as the CR value was very high, equal to 40 %, four times the consistency limit. By taking this problem into a quadratic programming problem it was possible to determine the optimal weights w_i^* (3.16) which minimize the error between the

expert's assignment and the real value; in this regard, the concept of *coherence* (3.15) was also introduced, which allows to evaluate the discrepancy between the weight assigned by the expert and the value relative to the importance of the model assigned by the *meta*-learner (in our case the Logistic regression), since in our case the features at level 0 of our Clustering Staking Algorithm (CSA) in the next level becomes a *meta*-features, but which represents the model used for clustering at level 0 (for the logical schema see figure 3.2). Four different types of clustering have been used in order to demonstrate the validity of the proposed methodology, the k-means has been introduced as it is of rapid convergence and of considerable simplicity in the explanation of how the groups are created. In this case the concept of *transparency* (see 1.2) is related to *how* these models work. The *Agglomerative* cluster, *Birch* and *Spectral* cluster were also introduced in order to review the most well-known clustering methods and insert them in the context of the determination and assignment of the weight to each model by the expert; first because we assume that each model represents a different decision maker and secondly to understand how their results can be interpreted, transforming a numerical information into a qualitative choice in the clinical context. We can however conclude that the method proposed can be of great interest in a clinical decision-making process and support other approaches, even of a more classical type such as that of the authors from whom the problem treated in the application was borrowed.

3.6 Conclusions and future research

In possession of powerful data analysis tools that are at the same time understandable and interpretable in the clinical context is a challenge that Artificial Intelligence has set itself in recent years. Machine learning can be a valuable ally in the clinical decision-making process that scientists, doctors and researchers are called upon to use every day in order to make our lives better, and contribute to the development and protection of the health of every single individual as it is mentioned in our constitution in article 32. Clinical staff have the ethical, social and legal responsibility for every action that is taken, as amply documented in the appendix A.1 of this work, and therefore it becomes essential to be able to interpret the output of an intelligent system in order to to make delicate decisions on which human life and health depend. It is also evident that the interaction between man and machine, the latter understood as software, robot or artificial intelligence system, has been a reality for some years and will constitute normality in the near future: it is our social and moral duty as researchers to contribute to the development and improvement of methods and models that can lead to ever better solutions and support for specific areas such as the clinical-medical one. In this thesis various topics have been addressed, all of which are fundamental in the decision-making process. In particular, in the first part the problem of interpretability and transparency of algorithms was dealt with from the point of view of models and methods, and then in the second chapter deepen the mathematical aspects of decision-making, ensemble methods and optimization as a learning problem. In the third chapter, however, a new methodology was presented for the problem of interpretability of machine learning in the clinical context. The logical schema that starts from the realistic hypothesis that an expert is enabled to

choose a set of possible algorithms (both supervised and unsupervised) that can be used to solve a clinical problem is experimented in the thesis through three possible case studies, each addressed with the proposed methodology. The expert attributes a value based on his experience, on a scale from 1 to 9 whose values determine the trade off between the interpretability and complexity of the model (table 3.1). Interpretability is based on how these models arrive at a solution, or if in closed or exact form, through the optimization of a loss function, which represents the learning process, or if through an approximation, heuristic or meta-heuristics methods, as occurs for various artificial intelligence algorithms whose loss functions are sometimes particularly complex and strongly non-linear. Subsequently, the expert through the AHP method (discussed in chapter 2) determines a weight for each model. The different models are then combined with each other in order to provide an aggregate prediction or classification, as in methods known as ensembles, where however the weights are usually determined empirically. Once the output of the combined model has been obtained, the latter must first be evaluated by the expert (Data Analyst) who assigned the values and subsequently together with the clinical expert the results obtained must be validated. In order to evaluate the models used, some methods have been proposed (see paragraphs 3.14-3.17). It is first assessed whether the weights given by the expert are consistent, through two indices known as CI (Consistency Index) and CR (Consistency Ratio) inherited from the AHP method; if the values obtained are consistent, the models can be subsequently evaluated with the classic indicators (accuracy, recall and precision for classification, R^2 for regression problems, v-measure and homogeneity index for clustering), otherwise if values of CI and especially of CR are not considered acceptable, the proposed methods are used (see paragraphs 3.14-3.17), in order to determine the optimal weight for the model and evaluate how much the expert's judgment has deviated from the optimal one. The above applies to both supervised and unsupervised problems. Once the weights and algorithms used have been evaluated, it is necessary to move on to the phase of interpreting the models and features involved in the analysis, this phase inevitably involves the interaction between the expert and the clinical staff. In order to make the models and features explainable, known techniques were used when suitable, such as LIME, and some new methods, proposed in the thesis work, were introduced. These methods remain valid for all types of learning problems, be it regression, classification or clustering (3.18-3.21). Specifically, these methods weigh each features of the model by the weight attributed by the expert and report it by the metric used in the evaluation (i.e. accuracy or R^2) depending on the type of problem: in this regard, it is been introduced the Function Features Explanation (see formula 3.21) which was then used also in the proposed applications, by appropriately modifying some elements; in the case of classification, probabilities will be used, in regression the predicted value and in clustering the number of elements belonging to a given cluster compared to the total number of elements. A new method was also presented for the clustering problem, based on *stacking* ensemble methods. Once the clustering models to be used (level 0) have been determined, then the output of each one, in this case the labels obtained, is used as input for a subsequent model (level 1), which is defined by a *meta*-learner. Through the use of this *meta*-learner, which can be a Logistic regression, a Linear regression or a Decision tree or XGboost, a *features* importance is carried out on the output in order to determine the most

important variables in the prediction, which in the case ours are represented by the algorithms used; in this way we are able to interpret and evaluate each model as if it were a variable, obtaining a ranking of importance. Therefore, once the methodology proposed in the work was defined, it was tested through three different applications in order to show the validity of the method and its use. The three case studies involved two classification problems and one clustering. The results obtained showed the validity of the proposed method. In the first case, the method has led to overcoming the benchmarks known in the literature in terms of accuracy by applying an ensemble algorithm on data concerning the possibility of developing cervical cancer. In the second case study, diabetes was treated as a pathology of interest, not settling at higher levels of accuracy than what is present in the literature, but presenting the problem in a different way, interpreting both the variables involved and giving an explanation as to why some models are preferable to others and how the output can be interpreted. The third problem was treated as unsupervised, providing empirical evidence on some variables that make up the object of study, in the case treated it was the development of the DED pathology in HIV-infected patients. The proposed clustering method has brought interesting results, using the proposed explainability methods it has been highlighted which clustering algorithms work better than others and overall the results obtained validate those obtained in the case-control study from which the data and the problem to face. The results of the thesis therefore appear encouraging and it is hoped that they will form the basis for new approaches, enrich existing ones and make machine learning easier to contextualize in the clinical setting. The appendix A.4 of this work deals with the problem of the ethical responsibility of Artificial Intelligence and that of the clinical staff who find themselves using these very complex mathematical tools, a very important problem that must not be overlooked when the decisions that are made concern the our health.

Bibliography

[1] Wyatt, J., Spiegelhalter, D. "Field trials of medical decision-aids: potential problems and solutions". Annual Symposium on Computer Application in Medical Care. Symposium on Computer Applications in Medical Care. 3-7. 1991

[2] de Dombal, FT., Leaper, D.J., Staniland, J.R., McCann, A.P., Horrocks, J.C. "Computer-aided diagnosis of acute abdominal pain". Br Med J. 1972

[3] Miller, R.A, Pople, H.E. Jr, Myers, J.D. "Internist-1, an experimental computer-based diagnostic consultant for general internal medicine". N Engl J Med. 1982

[4] Shortliffe, E. "Computer-based medical consultations: MYCIN". Artificial Intelligence - AI. 388. 1976

[5] Vidal, L., Sahin, E., Martelli, N., Berhoune, M., Bonan, B. "Applying AHP to select drugs to be produced by anticipation in a chemotherapy compounding unit". Expert Systems With Applications, 2010

[6] Dolan J.G, Frisina S. "Randomized controlled trial of a patient decision aid for colorectal cancer screening". Medical Decision Making, 2002; 22: 125–139

[7] Liberatore, M.J., Myers, R.E., Nydick, R.L. et al. "Decision counseling for men considering prostate cancer screening". Computers and Operations Research, 2003

[8] Carter, K.J., Ritchey, N.P., Castro, F., Caccamo, L.P., Kessler, E., Erickson, B.A. "Analysis of Three Decision-making Methods A Breast Cancer Patient as a Model". Medical Decision Making, 1999

[9] O'Connor, P. J., Sperl-Hillen, J. M., Rush, W. A., Johnson, P. E., Amundson, G. H., Asche, S. E., Ekstrom, H. L., and Gilmer, T. P. "Impact of electronic health record clinical decision support on diabetes care: a randomized trial". Annals of family medicine, 2011

[10] Georga, E., Protopappas, V., Arvaniti, E., Fotiadis, D. (2019). "The Diabino System: Temporal Pattern Mining from Diabetes Healthcare and Daily Self-monitoring Data". 2019

[11] Rung-Ching, C. ,Hui Qin, J., Chung-Yi, H. ,Cho-Tsan, B. "Clinical Decision Support System for Diabetes Based on Ontology Reasoning and TOPSIS Analysis". Artificial Intelligence in Medical Applications, 2017

[12] Lee JS, Kim CK, Kang J, et al. A Novel Computerized Clinical Decision Support System for Treating Thrombolysis in Patients with Acute Ischemic Stroke. J Stroke. 2015

[13] Anderson, J.A. ,Willson,P. , Peterson, N.J., Murphy, C., Kent, T.A., "Prototype to Practice: Developing and Testing a Clinical Decision Support System for Secondary Stroke Prevention in a Veterans Healthcare Facility", CIN: Computers, Informatics, 2010

[14] Arts, D.L., Abu-Hanna, A., Medlock S.K., van Weert, H.C.P.M. "Effectiveness and usage of a decision support system to improve stroke prevention in general practice: A cluster randomized controlled trial", PLOS ONE, 2017

[15] Ozsahin, I. " A Clinical Decision Support System for the Treatment of COVID-19 with Multi-Criteria Decision-Making Techniques", (Preprint) 2020.

[16] Khanmohammadi, S., Rezaeiahari, M. "AHP based Classification Algorithm Selection for Clinical Decision Support System Development". Procedia Computer Science. 2014

[17] Öztürk, N., Tozan, H., Vayvay, Ö. " A New Decision Model Approach for Health Technology Assessment and A Case Study for Dialysis Alternatives in Turkey". International Journal of Environmental Research and Public Health 2020

[18] Müller, J., Stoehr, M., Oeser, A., Gaebel, J., Streit, M., Dietz, A., et al. "A visual approach to explainable computerized clinical decision support. Computers and Graphics", 2020

[19] Schafer, H., Hors-Fraile, S., Karumur, R.P., Calero Valdez, A., Said, A., Torkamaan, H., Ulmer, T., Trattner, C.. "Towards health (aware) recommender systems". In: Proceedings of the 2017 international conference on digital health. pp. 157–161 (2017)

[20] Tonekaboni, S., Joshi, S., McCradden, M.D., Goldenberg, A. "What clinicians want: contextualizing explainable machine learning for clinical end use". arXiv preprint arXiv:1905.05134 (2019)

[21] Ucla, E. "Outlining the design space of explainable intelligent systems for medical diagnosis" (2019)

[22] Naiseh M. "Explainability Design Patterns in Clinical Decision Support Systems". In: Dalpiaz F., Zdravkovic J., Loucopoulos P. (eds) Research Challenges in Information Science. RCIS 2020. Lecture Notes in Business Information Processing, vol 385. Springer

[23] Bussone, A., Stumpf, S., O'Sullivan, D." The role of explanations on trust and reliance in clinical decision support systems". In: 2015 International Conference on Healthcare Informatics. pp. 160–169. IEEE (2015)

[24] Naiseh, M., Jiang, N., Ma, J., Ali, R." Explainable recommendations in intelligent systems: Delivery methods, modalities and risks". In: The 14th International Conference on Research Challenges in Information Science. Springer (2020)

[25] Rosenberg, A., Hirschberg, J. (2007)." V-Measure: A Conditional Entropy-Based External Cluster Evaluation Measure"

[26] Sobar, Machmud, R., and Wijaya, A.I. (2016). Behavior Determinant Based Cervical Cancer Early Detection with Machine Learning Algorithm. Advanced Science Letters, 22, 3120-3123.

[27] Amatul, Z., Asmawaty, A.K., and AznanM.A., M. (2013). "A comparative study on the pre-processing and mining of Pima Indian diabetes dataset".

[28] International Diabetes Federation, http://www.idf.org/diabetesatlas/5e/regional-overviews

[29] Pima Indian Diabetes Database, Url: www.ics.uci.edu/ mlearn/MLRepository.html

[30] Deepti, S., Dilip, S.S.,"Prediction of Diabetes using Classification Algorithms". Procedia Computer Science,Volume 132,2018,Pages 1578-1585

[31] Han W., Shengqi Y., Zhangqin H., Jian H., Xiaoyi W."Type 2 diabetes mellitus prediction model based on data mining". Informatics in Medicine Unlocked,Volume 10,2018,Pages 100-107

[32] B.M. Patil, R.C. Joshi, Durga T. "Hybrid prediction model for Type-2 diabetic patients". Expert Systems with Applications,Volume 37, Issue 12,2010,Pages 8102-8108

[33] Agrawal R, Balne PK, Veerappan A, et al. "A distinct cytokines profile in tear film of dry eye disease (DED) patients with HIV infection". Cytokine. 2016;88:77-84.

Appendix A

The ethics of AI

A.1 Legals aspects

The concepts of model interpretability and algorithmic transparency discussed in chapter 1 undoubtedly involve the problem of ethics. The new regulations on data processing and information security places machine learning and artificial intelligence in general, under observation as a fundamental tool in technological evolution. The new data protection regulation (GDPR) entered into force in 2018 [1] across the EU has defined the guidelines that the whole technological world is facing and respecting as a law. Many decisions, as we all know, are made through machine learning algorithms, these decisions also have an indirect impact on who suffer the decisions made through the outputs obtained from these artificial intelligence systems. The regulation introduces the "right to explanation" through which a user can ask for an explanation of an algorithmic decision that has been made. Article 22 of the regulation states (quoting verbatim):

Automated individual decision making, including profiling

1. The data subject has the right not to be subject to a based decision exclusively on automated processing, including profiling, which produces legal effects that concern him or that affects him in a similar way in a significant way
2. Paragraph 1 does not apply if the decision:
   (a) is necessary for the conclusion or execution of a contract between the data subject and a data controller
   (b) is authorized by Union or Member State law to which the controller is subject and which also provides for appropriate measures to safeguard the rights and freedoms of the data subject and legitimate interests o
   (c) is based on the explicit consent of the interested party
3. In the cases referred to in paragraph 2, letters a) and c), the data controller implements suitable measures to safeguard the rights and freedoms of the data subject and the legitimate interests, at least the right to obtain human intervention by controller, to express their point of view and to contest the decision.

4. The decisions referred to in paragraph 2 are not based on special categories of personal data referred to in Article 9 (1), unless Article 9, paragraph 2, letter a) or letter g), and adequate measures to safeguard the rights and freedoms of the data subject and the interests are in place.

In the light of these new [2] provisions, it is crucial, today and more and more in the future, to build algorithms that, on the one hand, respect the rules of the legislation, and on the other, be interpretable. On this topic there are several works that have dealt with the topic, to name one we have that of Goodman and Flaxman [3], in which some definitions are taken from article 4 of the regulation:

(a) Personal data: "any information relating to an identified or identifiable natural person"

(b) Data subject: "the natural person to whom data relates"

(c) Processing: "any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means"

(d) Profiling: "any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person"

A.2 Public sector

As regards this issue linked to AI transparency and ethics, the Association for Computing Machinery US Public Policy Council (USACM) was released overseas, or a "Declaration on algorithmic transparency and responsibility" in 2017, in which states that explainability is one of the seven principles for transparency and algorithmic responsibility and the same is particularly important in public policy contexts [4]. Considering also the other countries, they have made public the request for interpretability of AI. The Dutch Artificial Intelligence Manifesto [5] states that the utmost importance of artificial intelligence systems is not only accurate, but also able to explain how the system came to its decision. Another example is the French strategy for artificial intelligence, presented in 2018 by the then President of the Republic, whose content focuses on a series of proposals including the transparency of the algorithm, which involves the production of more explainable and interpretable models, then the user interface (GUI) and understanding of mathematical-methodological mechanisms in order to produce interpretable explanations [**?**]. Returning to Europe, in April 2018 the European Commission published a communication to many European official bodies, i.e. the European Parliament and the European Council, on the strategic importance of AI, which underlines the importance of the interpretation of algorithms and decision support systems (DSS) trust and awareness in people. Furthermore, these systems must be designed in a way that allows "humans" to include their actions in order to maximize the impartiality of AI systems [7]. The European Commission, in a report [8] on AI, also identifies the risk of non-transparency of the black box algorithms and the associated interpretability risk in two performance

risks for II [9]. In the document "Ethical guidelines for a trustworthy AI" published in 2019 [10] by the group of experts on AI, or AI HLEG (group of integrated experts established by the European Commission), are defined seven requirements that the systems of artificial intelligence must respect to be considered reliable, responsible and not least transparent.

A.2.1 Ethical requirements

Below we define the main requirements that must be met by AI system operators, as outlined by AI HLEG [10].

Human agency and oversight AI systems support human autonomy and decision making by acting as enablers for a democratic system and fair society by promoting fundamental rights and enabling human surveillance.

(A) **Fundamental rights**: AI systems can also enable and hinder fundamental rights. They allow people to trace their personal data or increase accessibility to education, therefore in support of their right to education. Prior to the development of the system, consideration should be given to the possibility of reducing or justifying the risks of a negative impact of AI systems on fundamental rights, in a democratic society in order to respect the rights and freedoms of others. An externally active feeback service is also proposed in order to monitor AI systems that potentially violate fundamental rights.

(B) **Human agency**: Users should be able to make autonomous informed decisions about AI systems. Artificial intelligence systems should support people in order to make better and more informed choices in accordance with their goals. Sometimes it is possible to distribute artificial intelligence systems modeling and influencing human behavior through mechanisms that can be difficult to detect, since they can exploit subconscious processes, including various forms of manipulation, deception. The general principle of user autonomy must be central to the functionality system. The key to this is the right not to be subject to a decision based solely on automated processing when this has legal effects on users or also significantly affects them.

(C) **Human oversight**: Human supervision helps ensure that an artificial intelligence system does not compromise human autonomy or cause negative effects. Supervision can be achieved through governance mechanisms such as a human-in-theloop (HITL), human-in-the-loop (HOTL) or human-in-command (HIC) approach. HITL refers to the ability of human intervention in every decision-making cycle of the system, which in many cases is neither possible nor desirable. HOTL refers to the capacity for human intervention during the system design cycle and the monitoring of system operation. HIC refers to the ability to control the overall activity of the artificial intelligence system (including its wider economic, social, legal and ethical impact) and the ability to decide when and how to use the system in any particular situation.

Technical robustness and safety Technical solidity is configured as a necessary element for the interpretability of AI, ML and DL systems, closely linked to the principle of damage prevention. Technical solidity requires that AI systems be developed with a preventive approach in order to behave reliably by minimizing involuntary damage and preventing unacceptable damage. In addition, the conception of physical and mental integrity of humans is defined.

(A) **Attack resilience and security**: Artificial intelligence systems, like all software systems, must be protected from vulnerabilities that can allow them to be exploited by opponents i.e. the hacking. Attacks can affect data, the model (loss of the model) or underlying infrastructure, both software and hardware. Insufficient security processes can also lead to errors decisions or even physical damage.

(B) **Accuracy**: Accuracy refers to the ability of an artificial intelligence system to make correct judgments, for example to classify correct information into appropriate categories or its ability to formulate predictions, controls or decisions based on data or models. When inaccurate occasional forecasts cannot be avoided it is important that the model (system) can indicate the probability of these errors. A high level of precision is particularly crucial in which the artificial intelligence system directly affects human life.

Privacy and data governance Privacy is closely linked to the principle of damage prevention. The prevention of damage to privacy also requires adequate data governance that covers the quality and integrity of the data used, access to protocols and the ability to process data in order to protect privacy.

(A) **Privacy and data protection**: Artificial intelligence systems must guarantee privacy and data protection in the whole system of the life cycle. This concerns the information initially provided by the user, as well as the information generated on the user during his interaction with the system (i.e. the user's response to recommendation systems). In order for people to trust the data collection process, it is necessary to make sure that the data collected about them will not be used to discriminate illegally or unjustly against them.

(B) **Quality and integrity of data**: The quality of the data sets used is fundamental for the performance of artificial intelligence systems. When data is collected, it can contain socially constructed distortions, inaccuracies and errors. This must be addressed before training with a specific data set. Furthermore, data integrity must be guaranteed. Entering malicious data into an artificial intelligence system can change its behavior, particularly with self-learning systems. Therefore the data quality and processing phase plays a fundamental role.

(C) **Access to data**: Data protocols governing data access should be established in an organization that manages personal data. These protocols should outline who can access data and under what circumstances. Only duly qualified personnel with competence and need for access, therefore data governance is an essential element of AI systems.

Transparency This requirement is closely linked to the principle of interpretability and includes the transparency of the elements relevant to an artificial intelligence system: data, algorithms and decisions.

(A) **Traceability**: The data sets and processes that determine the decision of the artificial intelligence system, including data collection and data labeling, as well as the algorithms used, should be documented according to the best possible standards to allow the traceability and transparency. This also applies to decisions made by the artificial intelligence system. Traceability facilitates verifiability and explainability.

(B) **Explainability**: The explainability concerns the ability to explain both the technical processes of an artificial intelligence system and the related human decisions. Whenever an AI system has a significant impact on people's lives, it should be possible to request an adequate explanation of the decision making process of the system.

(C) **Communication**: AI systems should not represent themselves as human to users; humans have the right to be informed that they are interacting with an artificial intelligence system. This implies that artificial intelligence systems must be identifiable as such. This could include communicating the level of accuracy of the artificial intelligence system, as well as its limitations.

Diversity, non-discrimination and equity In order to achieve reliable artificial intelligence, we must allow for inclusion and diversity throughout the lifecycle of the artificial intelligence system. In addition to the consideration and involvement of all stakeholders in the process, this too involves ensuring fair access through inclusive design processes as well as equal treatment. This requirement is closely linked to the principle of equity.

For a complete and rigorous discussion of the elements that make up the official document, see the work [10].

A.3 Private sector

As for private companies operating in the field of artificial intelligence, Google has made public their recommendations and responsibilities in the field of AI [11], in which the main one is based on the interpretability of AI systems. Among the main recommendations that have been provided by the American giant, we find: planning of interpretation, design of the interpretable model, understanding of qualified personnel and explanations communicating to model users. Platforms that offer AI solutions, such as H2O.ai, of interpretability have made one of the main features. In addition to the recommended strategies and practices, interpretation is also one of the main objectives in ML solutions and products currently marketed. H2O Driverless AI, an automated machine the learning platform offered by H2O.ai provides interpretation as one of its distinctive characteristics [12], the same for other platforms with the well-known IBM Watson or Paxata [13], [14], also among the

main companies in the field of AI that offers solutions based on the interpretability of the models. In the work of Carvalho *et al.* [15] several companies are mentioned that contribute in the private sector to the interpretability and explainability of machine learning. Principal authors cite *Kindy* which provides an Explanable AI platform for governments, such as financial services and healthcare. The authors also explain that since this product is intended for regulated business domains, the main characteristic of this product is, in fact, its explainability [16]. Google has long recognized the need for interpretability as the main pillar and has developed important research that has led to new tools, such as *Google Vizier*, a service for optimizing black boxes [8]. The Silicon Valley giant led by Zuckerberg, Facebook, in collaboration with Georgia Tech, has published an article in which it shows a visual element as a tool for exploring Deep Neural Networks models on an industrial scale [17]. Another private company, active at 360 degrees in the world of intelligence is Uber, which carries out very advanced artificial intelligence projects, such as self-driving vehicles, recently announced *Manifold*, an agnostic model as a tool for visual debugging for machine learning [18].

A.4 Ethical problems

Technological progress, in addition to bringing many positive things, inevitably drags some concerns behind it especially in the audience of people who do not have a thorough knowledge of the world of AI. There are therefore some questions that seem to be legitimate to ask, we give a general overview of the issues raised in recent years [19].

Social inequality and jobs In a study by the McKinsey Global Institute [20] it emerges that among the main problems, job loss is one of the most feared on the one hand, and most practical on the other, in fact it is expected that by 2030, as many as 800 millions of people will lose their jobs due to the work done by robots based on artificial intelligence systems. On the job issue, however, it also emerges the fact that more employment would be given by people who could work in the field of AI. A personal opinion on the issue is that social inequality could grow, as only those who would have access to an adequate level of education could work in the field of AI and last but not least this would discriminate against those who prefer, by ability or opportunity, work less qualified. Another problem is that related to taxes, in that a robot would not receive a salary like a human, would not pay taxes and would enter a zero flow of taxes necessary for the social and democratic life of a state.

Human and non-human errors In 2016 Teka Microsoft's chatbot through training on Twitter learning from the examples of human users in turn began to insult other users with racist and xenophobic comments. The error is therefore human on the one hand, but indirect for a certain, as the examples provided were totally wrong, on the one hand also the AI system (Tay) made a mistake and for this Microsoft immediately removed the chatbot from the network, the question in this case is who is responsible, on Tay [21], on Microsoft or on Twitter users, or on Twitter itself? This is also one of the issues raised in the AI.

AI and technological singularity The Guardian [22] has published an article dated 2014 which deals with the problem of technological singularity, that is, the hypothetical moment when technological growth becomes uncontrollable and irreversible leading to unpredictable changes in human civilization [23], [24] . This to date represents one of the causes of concern for many people namely the machine that surpasses the human.

AI bias issues Artificial intelligence has become increasingly inherent in facial and voice recognition systems, some of which have real commercial implications and directly affect people. Artificial intelligence systems, inevitably, are vulnerable to prejudices and errors introduced by its human creators. The data used to train these algorithms can have bias. For example, the facial recognition algorithms developed by Microsoft, IBM and Megvii all had prejudices in detecting the gender of people. These systems have been able to detect the sex of white men more accurately than the gender of darker skinned men. The question raised then is the following: can artificial intelligence become discriminatory? It should be ensured that AI systems do not have the same moral and ethical 'flaws' as their creators/trainers, but if the algorithms develop some prejudice towards someone or against a race, gender, religion or ethnicity, blame it will mainly reside on how the system was taught and formed so people who work in the world of AI, ML and DL, must bear in mind the prejudice in determining which data to use, or could one day, have to respond personally from a legal point of view.

A.5 Right to explanation

According to Recital 71 EU GDPR: "The data subject should have the right not to be subject to a decision, which may include a measure, evaluating personal aspects relating to him or her which is based solely on automated processing and which produces legal effects concerning him or her or similarly significantly affects him or her, such as automatic refusal of an online credit application or e-recruiting practices without any human intervention". In the field of transparency of the algorithms, there are situations in which, however, you are actually subject to decisions made by algorithms. For example consider the case of profiling in the granting of a mortgage, a loan or in any case areas where there is a need to be based on personal factors, such as in the insurance context, in which certain regions or areas are discriminated against compared to others [25] in the calculation of the annual premium. At present, the "right to explanation" is a highly debated topic, according to Edwards and Veale [26], the search for a "right to an explanation" in the general data protection regulation at best can be distracting while in the worst, feeding a new type of "transparency error". Also within the GDPR there are also (i) the right to erasure and the right to data portability and (ii) to privacy based on the design, the impact assessments on data protection and the certifications and seals on privacy. Through these points we could proceed towards a more responsible, more interpretable XAI at the center of which is the social, economic and ethical life of man. Several leading personalities from the scientific and industrial world have expressed concern regarding the future of artificial intelligence. Stephen Hawking in 2014 told the BBC [27]: "The primitive

forms of artificial intelligence that we already have, have proved very useful. But I think that the development of full artificial intelligence could mean the end of the human race". Elon Musk was also of the same opinion, who in 2014 instead said [28]: "I think we should be very attentive to artificial intelligence. If I had to guess what our greatest existential threat is, it is probably that. So we have to be very careful". Bill Gates (2015) joins the previous chorus stating [29]: "A few decades after that though the intelligence is strong enough to be a concern. I agree with Elon Musk and some others on this and don't understand why some people are not concerned". The right to explanations therefore remains an important topic within the world of AI, from an ethical, social and economic point of view.

Bibliography

[1] Regulation (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) [2016] OJ L119/1.

[2] Parliament and Council of the European Union (2016). General Data Protection Regulation

[3] Goodman, B., Flaxman, S. (2017). European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation". AI Magazine, 38(3), 50-57.

[4] ACM US Public Council. Statement on Algorithmic Transparency and Accountability, 2017

[5] IPN SIG AI. Dutch Artificial Intelligence Manifesto, 2018

[6] Cédric Villani. AI for Humanity—French National Strategy for Artificial intelligence, 2018

[7] European Commission. Artificial Intelligence for Europe. 2018

[8] European Commission. Algorithmic Awareness-Building. 2018

[9] Rao, A.S. Responsible AI and National AI Strategies. 2018

[10] High-Level Expert Group on Artificial Intelligence (AI HLEG). Ethics Guidelines for Trustworthy Artificial Intelligence, 2019

[11] Google. Responsible AI Practices - Interpretability

[12] H2O.ai. H2O Driverless AI

[13] https://www.datarobot.com/wiki/interpretability/

[14] https://www.ibm.com/blogs/research/2019/08/ai-explainability-360/

[15] Carvalho, D.V.; Pereira, E.M.; Cardoso, J.S. Machine Learning Interpretability: A Survey on Methods and Metrics. Electronics 2019, 8, 832.

[16] Kyndi. Kyndi AI Platform

[17] Kahng, M.; Andrews, P.Y.; Kalro, A.; Chau, D.H.P. ActiVis: Visual exploration of industry-scale deep neural network models. IEEE Trans. Vis. Comput. Gr. 2018, 24, 88–97.

[18] Zhang, J.; Wang, Y.; Molino, P.; Li, L.; Ebert, D.S. Manifold: A Model-Agnostic Framework for Interpretation and Diagnosis of Machine Learning Models. IEEE Trans. Vis. Comput. Gr. 2019, 25, 364–373.

[19] https://kambria.io/blog/the-7-most-pressing-ethical-issues-in-artificial-intelligence/

[20] https://www.mckinsey.com/featured-insights/future-of-work/jobs-lost-jobs-gained-what-the-future-of-work-will-mean-for-jobs-skills-and-wages

[21] https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist

[22] https://www.theguardian.com/technology/2014/feb/22/robots-google-ray-kurzweil-terminator-singularity-artificial-intelligence

[23] "Collection of sources defining "singularity"". singularitysymposium.com. Retrieved 17 April 2019.

[24] Eden, Amnon H.; Moor, James H. (2012). Singularity hypotheses: A Scientific and Philosophical Assessment.

[25] https://www.ilfattoquotidiano.it/2012/07/17/assicurazione-auto-in-italia-se-sei-straniero-costa-di-piu/296043/

[26] Edwards, L., Veale, M., "Slave to the Algorithm? Why a 'Right to an Explanation' Is Probably Not the Remedy You Are Looking For" (May 23, 2017). 16 Duke Law and Technology Review 18 (2017)

[27] https://www.theguardian.com/science/2014/dec/02/stephen-hawking-intel-communication-system-astrophysicist-software-predictive-text-type

[28] https://aeroastro.mit.edu/video-categories/2014-centennial-symposium-videos

[29] https://www.theguardian.com/technology/2015/jan/29/artificial-intelligence-strong-concern-bill-gates