Health and Technology (2021) 11:875–885
https://doi.org/10.1007/s12553-021-00554-6

ORIGINAL PAPER

Cervical cancer risk prediction with robust ensemble and explainable black boxes method

Francesco Curia[1]

Received: 6 January 2021 / Accepted: 26 April 2021 / Published online: 14 May 2021

Abstract
Clinical decision support systems (CDSS) that make use of algorithms based on intelligent systems, such as machine learning or deep learning, they suffer from the fact that often the methods used are hard to interpret and difficult to understand on how some decisions are made; the opacity ofsome methods, sometimes voluntary due to problems such as data privacy or the techniques used to protect intellectual property, makes these systems very complicated. Besides this series of problems, the results obtained also suffer from the poor possibility of being interpreted; in the clinical context therefore it is required that the methods used are as accurate as possible, transparent techniques and explainable results. In this work the problem of the development of cervical cancer is treated, a disease that mainly affects the female population. In order to introduce advanced machine learning techniques in a clinical decision support system that can be transparent and explainable, a robust, accurate ensemble method is presented, in terms of error and sensitivity linked to the classification of possible development of the aforementioned pathology and advanced techniques are also presented of explainability and interpretability (Explanaible Machine Learning) applied to the context of CDSS such as Lime and Shapley. The results obtained, as well as being interesting, are understandable and can be implemented in the treatment of this type of problem.

Keywords Cervical cancer · Ensemble · Interpretable AI · Risk prediction

1 Introduction

1.1 Motivation

Machine learning and artificial intelligence more generally are playing a main role in every scientific field and the tools used by researchers in the clinical field to date have been based on classical statistics. The introduction of clinical problems solved by advanced machine learning tools in recent years has seen an exponential growth. These methods and algorithms are very powerful, but at the same time very complex in their interpretation; while if a linear or logistic regression is performed, we are able to give an even intuitive interpretation of the parameters estimated by the model, for many machine learning and deep learning algorithms the difficulty lies precisely in the fact of interpretability. Fortunately, research in the field of machine learning and AI interpretability has also progressed, providing tools that can help research apply complex models and interpret them. Another fairly well-known concept that arouses attention is the transparency of machine learning and deep learning techniques, which very often do not provide solutions in closed form but rather many methods of solving the cost functions associated with problems, which are heuristics, meta heuristics or simply approximations. Let's give an intuitive explanation: while in a linear regression for example, the estimation of the parameters takes place through the minimization of a cost function, of a quadratic type, which it is sufficient to derive twice, for some more complex functions used in machine learning or deep algorithms learning, these functions involve both many parameters to be estimated, and the non-linearity of the function itself and perhaps to find the optimum of these functions, non-punctual techniques are used. Clinical operators need interpretable solutions, explanations, especially as they are subject to legal responsibility from the point of view of the profession. This work intends to tackle a clinical problem, using machine learning algorithms; but without stopping only at the performing result of the model used, but giving an interpretation of the method and the components involved.

✉ Francesco Curia
francesco.curia@uniroma1.it

[1] Department of Statistical Sciences, Sapienza University of Rome, Piazzale Aldo Moro 5, 00185 Rome, Italy

1.2 Literacy review

Cervical cancer represents a very widespread disease for the world female population; like all diseases that afflict the population, the main purpose of medicine is to equip itself with tools that can help research and prevent the risk associated with this type of cancer. Machine learning can be a significant tool in fighting this disease. Having a decision support system in the clinical setting, based on advanced algorithms, is a valuable tool that has seen extensive use in the healthcare sector in recent years. The use of decision support systems in the clinical sector is widely known, Vidal et al. [1] present a Clinical Decision Support System (CDSS) based on the Analytic Hierarchy Process (AHP) method to assist pharmacists to choose a drug therapy in cancer patients, while always for the AHP Liberatore et al. [2] have implemented a CDSS 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). Still in the context of CDSS Dolan and Firisna [3] 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. In the context of cervical cancer and the study of this phenomenon by machine learning methods, in recent years several authors have produced significant works, such as Tseng et al. [4] who considered three different approaches including support vector machine (SVM), C5.0 and extreme learning machine, in order to identify risk factors useful for explaining and predicting the risk of cervical cancer. The experimental results demonstrate the model C5.0 is the most useful for identifying risk factors. Sharma [5] using data with 237 patients and 10 features (http://www.igcs.org) proposes in his work an algorithm based on a classification tree, achieving good results with the C5.0 method with accuracy of 67.5% using advance pruning option. More recent studies regarding the use of machine learning techniques for cervical cancer prevention, such as Wu and Zhou [6] and Geetha et al. [7] applying, respectively, an SVM algorithm on data composed of 32 risk factors and 4 target variables, the authors also use principal component analysis (PCA) to eliminate recursion of some characteristics. The most recent work by Geetha et al. [7] again using PCA they eliminate the recursion of the characteristics and balance the data sample through the SMOTE technique and then carry out the classification through a method based on decision trees known as random forest. Starting from the work of Sobar and Wijaya [8], the authors investigate how to predict a certain type of cervical cancer in advance. The data available to the authors come from a questionnaire distributed to 72 individuals, including 50 without cervical cancer and 22 with the disease. The study was conducted in Jakarta in Indonesia. 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 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, so starting from their work and with the data used, a new classification method was developed based on the methodology proposed in this work.

2 Methodology

2.1 Mathematical aspects

In this work, different supervised machine learning algorithms were built, specifically, logistic regression, classification trees, multilayer perceptron and an ensemble model were used which aggregate the models by the soft-voting method (see Equation 7).

Logistic regression Logistic regression (LR) is a widely used statistical-mathematical method. It is mainly used to estimate the probability $p = P(Y = 1|X = x)$ of a given (binary) phenomenon as a function of N-independent variables $x_1, ..., x_n$.

$$p = \frac{1}{1 + e^{-Z}} \tag{1}$$

where

$$Z = b_0 + \sum_i b_i x_i \tag{2}$$

The regression coefficients b_0 and b_i are usually estimated using maximum likelihood estimation [9]. In this case it is not possible to find a closed form expression for the values of the coefficients that maximize the likelihood function, so an iterative approximation method must be used. A first solution is initialized and at each step of the algorithm an evaluation is carried out in order to be improved and it is repeated until the algorithm does not converge to an optimum point.

Decision trees Decision trees can be applied to both regression and classification problems [10]. 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:

$$min_r \sum_{s=1}^{S} \sum_{i \in r_s} (y_i - \hat{y}_{r_s})^2 \tag{3}$$

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.

MultiLayer perceptron MLP is an artificial neural network (ANN) model which taking as input N-features $(x_1, ..., x_n)$ and associated N-weights, w_i,$i = 1, ..., N$ defining a function input $U(\cdot)$ which calculates the weighted sum of the input features.

$$U(x) = \sum_{i=1}^{N} w_i x_i \tag{4}$$

The result is then passed to an activation function f, which will produce the output of the perceptron.

$$y = f(U(x)) = 1 \quad if \quad x > \beta, 0 \quad otherwise \tag{5}$$

In the original perceptron, the activation function is a gradual function: where β is a threshold parameter. An example of a step function with $\beta = 0$. Learning in neural networks, more generally, in each learning phase, consists in adjusting the weights (updating) of its sensors in order to minimize the error on the training data. Generally, the [11] backpropagation algorithm is used as an optimization method, which minimizes the standard deviation. Of course, other learning algorithms can also be used.

Ensemble The basic idea is to build a classifier, which by definition, by aggregating models considered "weaker" creates a whole model with better qualities in correctly classifying the instances of the studied dataset [12]. For the classification problems we can compute a weighted majority vote by associating a weight w_j with classifier c_i

$$\hat{y}_i = \arg\max_i \sum_{j=1}^{m} w_j \cdot I_A\left(c_j(\mathbf{x}) = i\right) \tag{6}$$

where I_A is the indicator function and A is the set of unique class labels. Considering the predicted probabilities p_j for the j-th classifier, we can obtain the final ouput

$$\hat{y}_i = \arg\max_j \sum_{j=1}^{m} w_j \cdot p_j \tag{7}$$

In this experiment the weights w_j are set by the ratio $1/N$, where N is the number of classifiers c_i.

Evaluation metrics The algorithms used were evaluated by the following metrics, namely recall, precision, accuracy and Area Under Curve. The metrics defined are

$$precision = \frac{TP}{TP + FP} \tag{8}$$

and

$$recall = \frac{TP}{TP + FN} \tag{9}$$

For binary classification, accuracy can also be calculated in terms of positives and negatives classes

$$accuracy = \frac{TP + TN}{TP + FN + FP + TN} \tag{10}$$

2.2 Explainable machine learning: main methods

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. [13] 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 [14–16], failing to calibrate the concept of user trust by introducing this new type of error to the context as analyzed by [17] using these tools. Another example relating to the work of Bussone et al. [18] 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 [19]. In order to give the reader a general understanding of the main explainable machine learning techniques, some of the possibilities are illustrated below:

Lime Ribeiro et al. [20] introduces the concept of trade off between interpretability and loyalty LIME (Local Interpretable Model-Agnostic Explanations) formalized by the following optimization problem:

$$min_g \in G \quad L(f, g, \pi_x) + \Omega(g) \tag{11}$$

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 [21] 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 [22] 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) [23] 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 and Popescu presents another method, called *feature interaction* [24] 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

$$E_x\left(\frac{\partial^2 F(x)}{\partial x_j \partial x_k}\right)^2 > 0 \qquad (12)$$

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* [25], 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 [26] 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 $R(\theta) = \frac{1}{n}\sum_i L(z_i, \theta)$.

Shallow decision trees In the authors' work [27], the focus is on the problem of the trade off between the complexity of the model and its interpretability, therefore the authors consider decision trees as a modeling reference. This type of model is very easy to interpret, but for this very reason it often suffers from overfitting. The authors consider a deep learning problem and tackle it through the use of a decision tree called *shallow* in order to reduce the complexity and then make it interpretable by controlling its depth, to do this they use a micro-aggregation and cluster creation, the cluster size however determines the comprehensibility and privacy of the problem they face. Their method works on very large data sets and also on categorical data through ontologies for semantic consistency. For a more detailed discussion, from which various components of this chapter have been extracted, please refer to the excellent work of the authors [28].

3 Experimental results

3.1 Data analysis

The data (https://archive.ics.uci.edu) that were used by the authors did not require a processing as they are already encoded and without missing values.The study dataset suffers from the dimension (instances = 72, attributes = 19), as the basis of the learning process of a machine learning algorithm is the consistency in having as many examples possible in order to structure a meaningful learning process. Since the dataset is not very large and the percentage between the two classes is not too disproportionate (Class: Cancer Yes = 30%, No = 70%) no balancing technique was applied. The data did not require the need to be standardized or normalized, as they were scores obtained on an ordinal scale. The data were collected through a questionnaire administered in a specialized center in Jakarta (Indonesia). The attributes considered are 19 and as often happens in the medical-health context the dataset is not very large. The authors in the

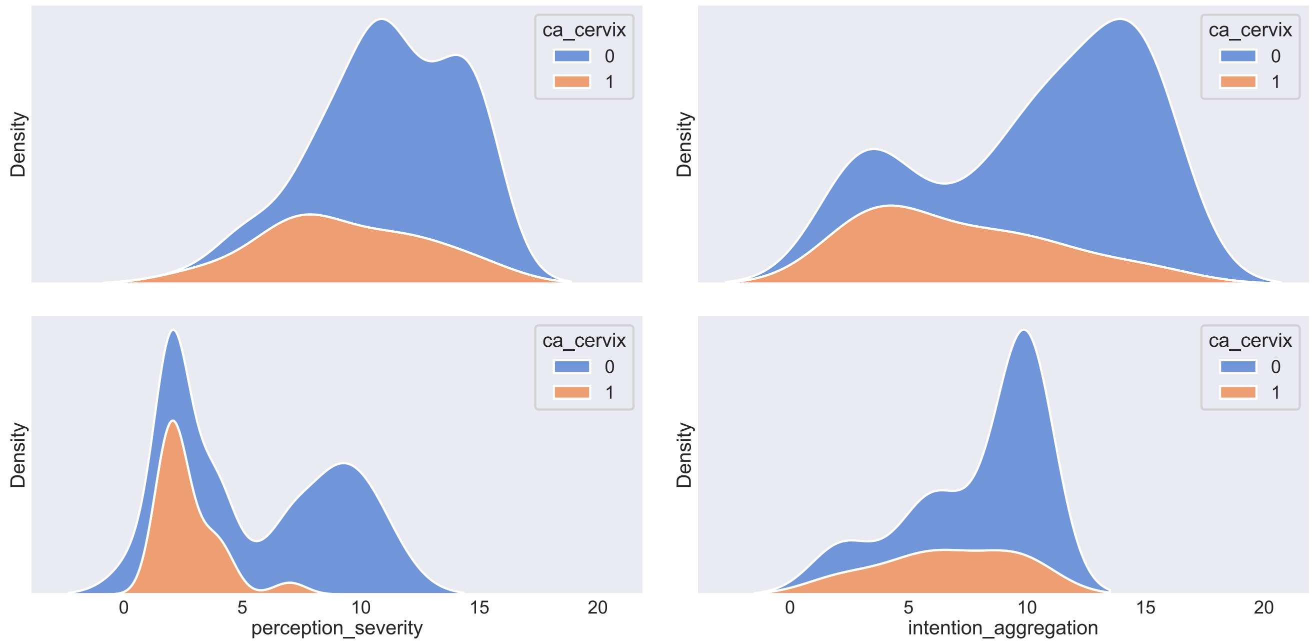
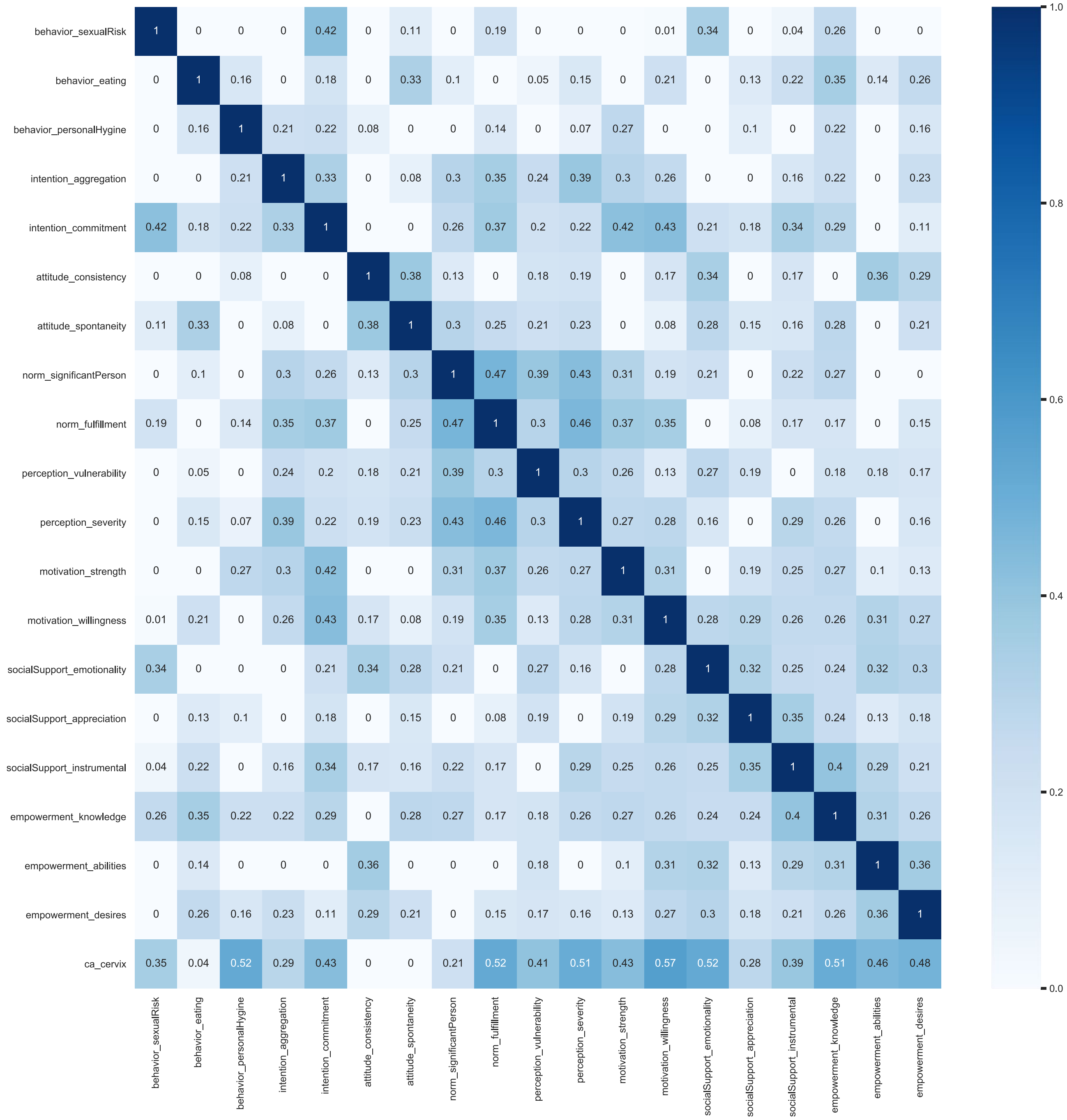


Fig. 1 Histrograms



	behavior_sexualRisk	behavior_eating	behavior_personalHygine	intention_aggregation	intention_commitment	attitude_consistency	attitude_spontaneity	norm_significantPerson	norm_fulfillment	perception_vulnerability	perception_severity	motivation_strength	motivation_willingness	socialSupport_emotionality	socialSupport_appreciation	socialSupport_instrumental	empowerment_knowledge	empowerment_abilities	empowerment_desires
behavior_sexualRisk	1	0	0	0	0.42	0	0.11	0	0.19	0	0	0	0.01	0.34	0	0.04	0.26	0	0
behavior_eating	0	1	0.16	0	0.18	0	0.33	0.1	0	0.05	0.15	0	0.21	0	0.13	0.22	0.35	0.14	0.26
behavior_personalHygine	0	0.16	1	0.21	0.22	0.08	0	0	0.14	0	0.07	0.27	0	0	0.1	0	0.22	0	0.16
intention_aggregation	0	0	0.21	1	0.33	0	0.08	0.3	0.35	0.24	0.39	0.3	0.26	0	0	0.16	0.22	0	0.23
intention_commitment	0.42	0.18	0.22	0.33	1	0	0	0.26	0.37	0.2	0.22	0.42	0.43	0.21	0.18	0.34	0.29	0	0.11
attitude_consistency	0	0	0.08	0	0	1	0.38	0.13	0	0.18	0.19	0	0.17	0.34	0	0.17	0	0.36	0.29
attitude_spontaneity	0.11	0.33	0	0.08	0	0.38	1	0.3	0.25	0.21	0.23	0	0.08	0.28	0.15	0.16	0.28	0	0.21
norm_significantPerson	0	0.1	0	0.3	0.26	0.13	0.3	1	0.47	0.39	0.43	0.31	0.19	0.21	0	0.22	0.27	0	0
norm_fulfillment	0.19	0	0.14	0.35	0.37	0	0.25	0.47	1	0.3	0.46	0.37	0.35	0	0.08	0.17	0.17	0	0.15
perception_vulnerability	0	0.05	0	0.24	0.2	0.18	0.21	0.39	0.3	1	0.3	0.26	0.13	0.27	0.19	0	0.18	0.18	0.17
perception_severity	0	0.15	0.07	0.39	0.22	0.19	0.23	0.43	0.46	0.3	1	0.27	0.28	0.16	0	0.29	0.26	0	0.16
motivation_strength	0	0	0.27	0.3	0.42	0	0	0.31	0.37	0.26	0.27	1	0.31	0	0.19	0.25	0.27	0.1	0.13
motivation_willingness	0.01	0.21	0	0.26	0.43	0.17	0.08	0.19	0.35	0.13	0.28	0.31	1	0.28	0.29	0.26	0.26	0.31	0.27
socialSupport_emotionality	0.34	0	0	0	0.21	0.34	0.28	0.21	0	0.27	0.16	0	0.28	1	0.32	0.25	0.24	0.32	0.3
socialSupport_appreciation	0	0.13	0.1	0	0.18	0	0.15	0	0.08	0.19	0	0.19	0.29	0.32	1	0.35	0.24	0.13	0.18
socialSupport_instrumental	0.04	0.22	0	0.16	0.34	0.17	0.16	0.22	0.17	0	0.29	0.25	0.26	0.25	0.35	1	0.4	0.29	0.21
empowerment_knowledge	0.26	0.35	0.22	0.22	0.29	0	0.28	0.27	0.17	0.18	0.26	0.27	0.26	0.24	0.24	0.4	1	0.31	0.26
empowerment_abilities	0	0.14	0	0	0	0.36	0	0	0	0.18	0	0.1	0.31	0.32	0.13	0.29	0.31	1	0.36
empowerment_desires	0	0.26	0.16	0.23	0.11	0.29	0.21	0	0.15	0.17	0.16	0.13	0.27	0.3	0.18	0.21	0.26	0.36	1
ca_cervix	0.35	0.04	0.52	0.29	0.43	0	0	0.21	0.52	0.41	0.51	0.43	0.57	0.52	0.28	0.39	0.51	0.46	0.48

Fig. 2 Cramer's V Matrix

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 Fig. 1 we can see with respect to the presence or absence of cervical cancer, how the values for the characteristics concerning intention, empowerment and behavior are distributed; the scores on the abscissas indicate the level of the ordinal variable.

Figure 1 shows the distributions of some characteristics through histograms; note that for the personal hygiene behavior in people without cervical cancer this factor is much higher than in those with cervical cancer, as well as if we look to the relative variable in the intention of social aggregation.

Table 1 Error comparison on train and test set: whole dataset

Evaluations - AUC: 0.98				
Algorithm	LR	DT	MLP	Ensemble
error on train set	0	0.33	0	0.075
error on test set	0.05	0.16	0.11	0.085
accuracy	0.94	0.83	0.88	0.945

Table 3 Error comparison on train and test set: restricted features

Evaluations - AUC: 0.84				
Algorithm	LR	DT	MLP	Ensemble
error on train set	0.018	0.33	0	0.087
error on test set	0.055	0.16	0.055	0.080
accuracy	0.94	0.83	0.94	0.945

By the Cramer's V [29] index application we can measure the association between the different features that make up the dataset; given the nature of the features of the ordinal categorical type, from Fig. 2 we can see several associations equal to 0 and some on average significant, but never exceeding 50%. The associations that report interesting values are those between the target that indicates or not the presence of cervical cancer (*ca_cervix*) and the variables concerning the area of motivation (willingness: 0.57), that of perception (severity: 0.51), behavior (personalHigyne: 0.52); we have other variables that present an association below 50% but still of interest such as that linked to sexual behavior (sexual risk: 0.35) and the enhancement of desire (desires: 0.48).

3.2 Main results

The results obtained on the available data showed that by the application of the single models, logistic regression, decision tree and multilayer perceptron, in the prediction of cervical cancer risk the accuracy values obtained, respectively are of 94%, 83% and 88%, while for the proposed ensemble model a binary classifier based on soft voting it was obtained a value equal to **94.5**%; it is interesting the AUC value which is equal to **0.98**, one percentage point higher than that of the authors [8], this means that on average, out of 100 cases examined, our ensemble classifier correctly classifies one instance more than the benchmark of the work done by the authors. All analysis was done using the Python 3.6 libraries i.e. Sklearn, Seaborn, Pandas, Numpy, Lime and Shap.

The proposed method for the classification of cervical cancer risk it is resulted to be robust and provide significant performances; from the table (table 1) that compares the single models used we can see how some of them, logistic regression (LR) and decision tree (DT) suffer from overfitting since the difference between classification error on training and test data is very high. By applying the ensemble (EC) classifier [24] this overfitting is drastically reduced and this is a first fundamental step in the construction of a classification algorithm. Secondly, from the statistics obtained (table 4) on the ensemble model it is possible to obtain other statistically significant information, such as *recall* and *precision* values while table 2 shows all the parameters for each algorithm used; the first metric expresses the sensitivity of the model and is represented by the ratio between the correct classifications for a given class, on the total of cases in which the event actually occurs, while the precision is a measure of the ratio between the number of correct classifications of a given event (class) on the total number of times that the algorithm classifies it.

Starting from these definitions, we can see that for the negative class (No Cancer) the model is very sensitive and precise in recognizing instances that do not present risk associated with the development of cervical cancer. The two metrics highlighted, such as the value of accuracy or AUC, which respectively measure the number of cases correctly classified by the model and the rate of true positives compared to the rate of false positives, which takes values between 0 and 1, they also suffer from the fact that the available dataset is very small, therefore the examples with which the model "learns" are very few. Despite this, as a starting point for building valid tools to support clinical decisions in the analysis and study of cervical cancer prevention, the proposed methodology is a viable path with a view to being able to collect larger datasets with which to train algorithms increasingly sophisticated in the recognition of certain clinical pathologies.

Figure 3 shows the normalized confusion matrixes between 0 and 1, representative of the classification capacity of the proposed ensemble algorithm. In the first (from the

Table 2 Main parameters for each model

Logistic Regression	Decision Tree	Multi Layer Perceptron
tollerance: 0.001	criterion: entropy	learning rate: adaptive
iterations: 250	max depth: 2	layer: (19,15,10,15,19)
solver: liblinear	fraction leaf: 0.25	solver: adam
C: 0.75	impurity split: 2	activation: tanh

Table 4 Classification report for Ensemble classifier

Statistics				
Target	precision	recall	macro avg acc.	weighted avg acc.
Cancer = No	0.94	1.00	0.97	0.83
Cancer = Yes	1.00	0.67	0.95	0.94

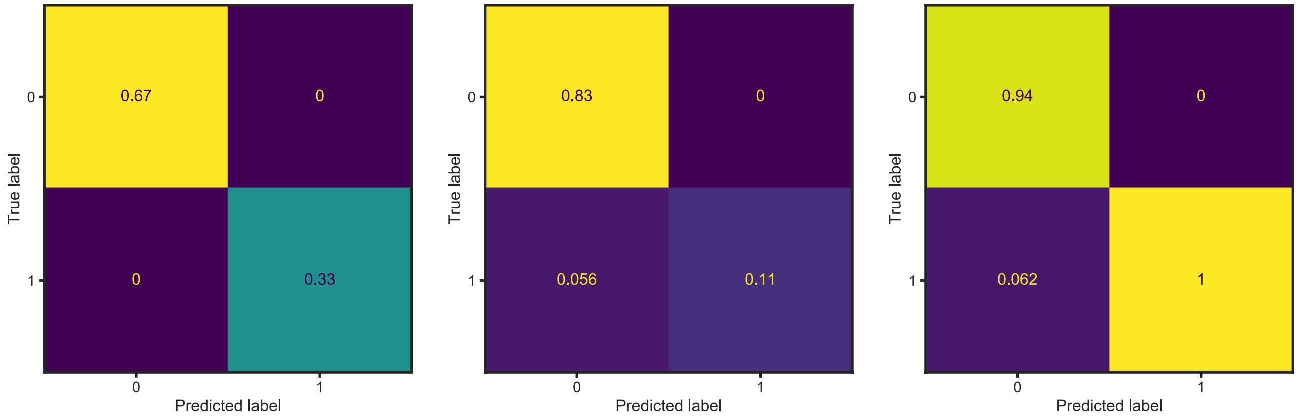


Fig. 3 Normalized confusion matrix : true values, predicted and on all sample observations

left) it is possible to observe the matrix relating to the true values, with a percentage error equal to 5% of incorrectly classifying the negative cases (cancer = Yes), i.e. that out of 100 cases 5 are classified as cancer but in reality they are not. The matrix in the center represents the classification capacity of the model on the predicted data, and the percentage of false positives increases to 6%, while the share of true positives also increases to 94%. In the matrix on the right,

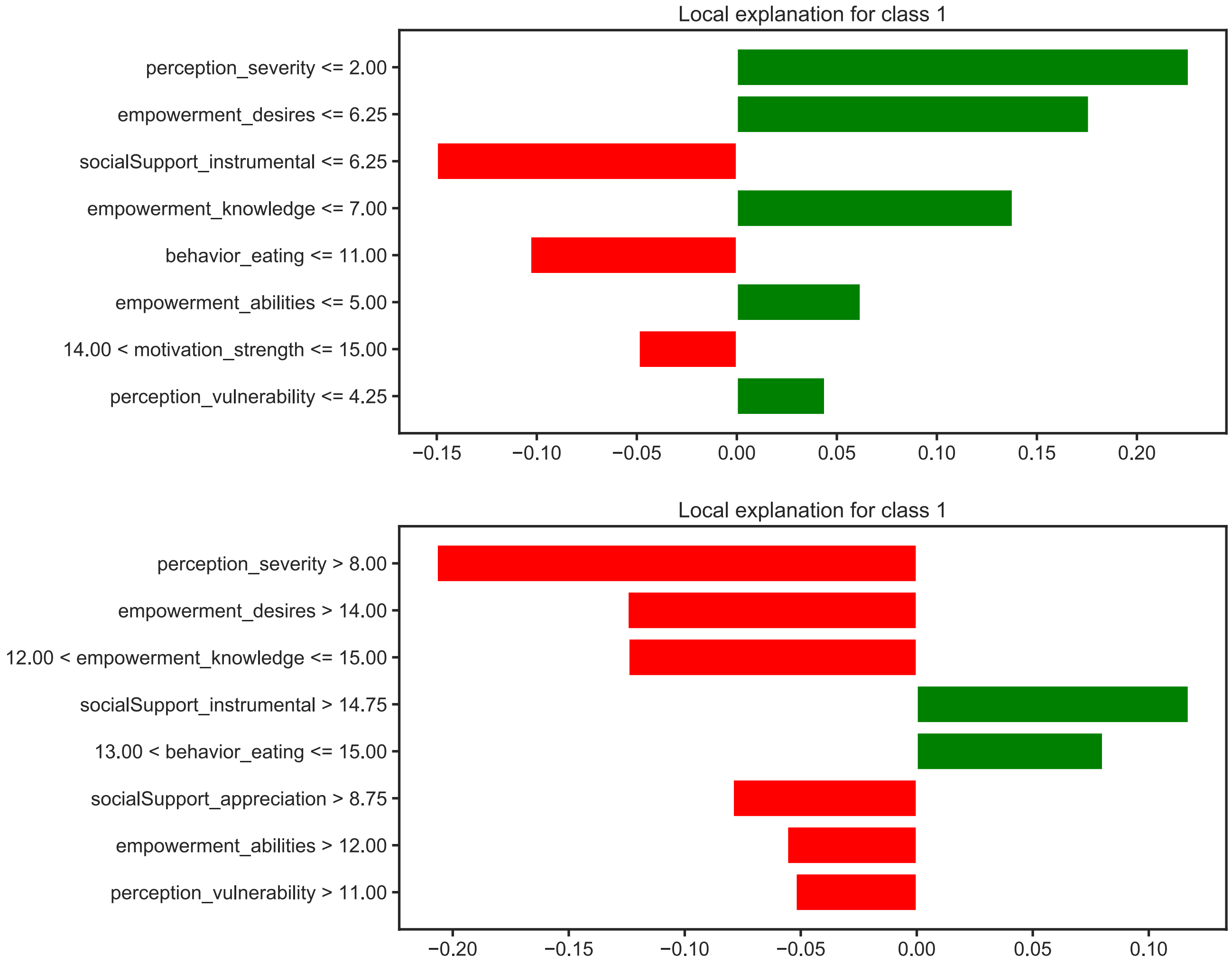


Fig. 4 Explainability with LIME

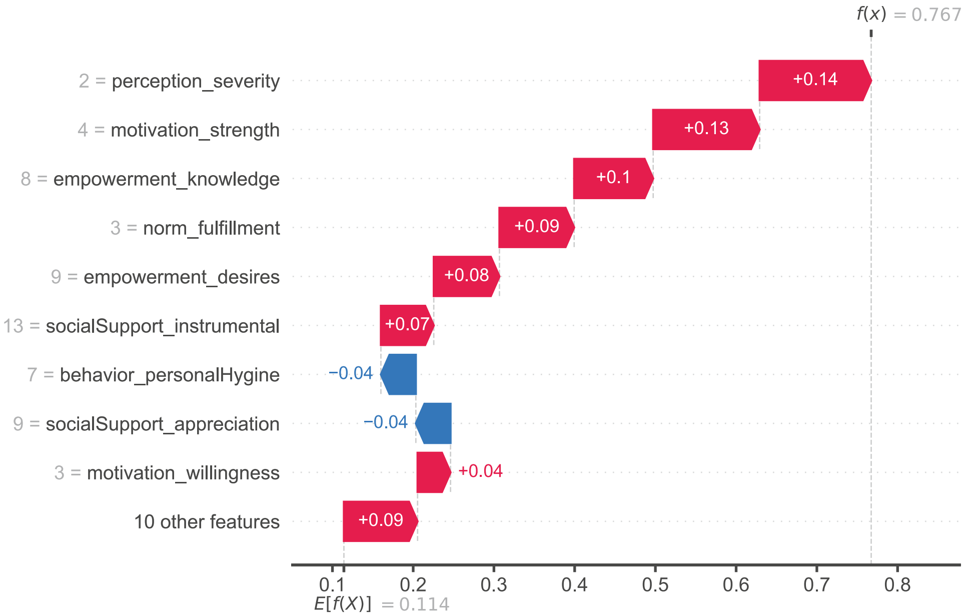


Fig. 5 Explainability with Shapley method: positive class, patient id: 15

on the other hand, we have the ability to classify all values, with a fairly interesting ability but which in any case tends to suffer from overfitting as there are neither false positives nor false negatives.

Table 3, on the other hand, shows the performances of the single algorithms and of the ensemble model relating to the initial dataset restricted by the features which, through the application of the Cramer's V index (Fig. 2), reported an association equal to 0 with the target variable. As you can see, however, the AUC value has drastically dropped from 0.98 to 0.84, the error discrepancy on train and test for the ensemble model is significantly lower and the accuracy has remained the same. It features a slight decrease in logistic regression overfitting and an improvement in accuracy for the MLP algorithm. The decision tree model, on the other hand, has remained unchanged. Although there is no strong statistical association between different features and the target variable, given the evaluation metrics, the Ensemble model is used on the initial dataset.

3.3 Clinical explainability

The part relating to the interpretability of the models used and the analysis of the features was conducted by the use of the LIME method [20], developed by Ribeiro et al. and with the Shapley method [25]. The first method provides interpretability results by perturbing a single instance. In Fig. 4 it is possible to observe the explanations for each of the features of interest of the model (patient id: 17 and patient id: 15, respectively); the figure specifically refers to a single instance (a specific patient) from which it is possible to obtain various information. The data are those of test and therefore we can make predictions on new instances, considering that we do not have certain information on his state of health but evaluate the risk that a particular subject may develop the disease. It is possible to observe his probability of developing the disease and in this case it is equal to **74**%; in the face of this risk we can see which factors influence the increase in statistical terms of developing the disease. The perception of severity and the empowerment of desire are factors that the model considered important, as they respectively increase the probability of developing the disease by 22% and 17%. This at a conceptual logical level makes sense, as a low perception of severity linked to an increase in sexual desire could push an individual to develop inadequate and compromising behaviors, as well as expose themselves to a higher risk. Diet-related behavior (*behavior_eating*) reduces the risk by 11%, and this is also a factor to be taken into consideration, as it is established that nutrition plays a significant role in the treatment of oncological pathologies. It is interesting to observe how social support (*socialSupport_instrumental*) reduces the risk by 16%, and this in a patient with a high probability of developing an oncological pathology is a not negligible factor from the point of view of psychological consequences that a disease causes. By selecting another patient, for whom the probability of developing the disease is equal to **11**%, we can also observe here that an increase in the perception of severity and perception of vulnerability can reduce respectively by 20% and 6% the risk. Therefore, the clinical operator from this information could intervene locally on the characteristics of interest that contribute to the development of the risk and using a black-box type machine learning algorithm, when he obtains a result he is aware of which factors motivated the model to the choice of the result.

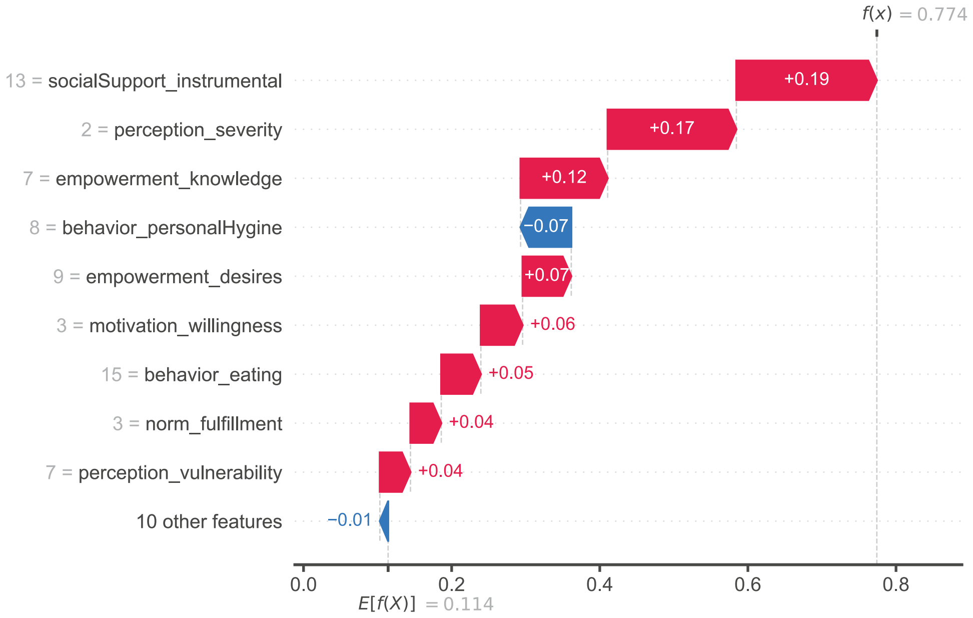
Figure 5 shows the explanation of the ensemble model via the shapley method. Considering the same instances on the test dataset classified by the model, we see the influence of the characteristics on the classification. On the first graph below for the positive class, relating to the risk of onset of the disease with probability equal to 76%, here too the perception of severity influences the prediction of 14%, while personal hygiene (personalHygine) lowers this probability of the 4%; the strength of motivation instead increases the probability by 13% and the empowerment desire by 8%. The bars in blue represent how much the probability of developing the disease is reduced for a given value of the affected feature.

Figure 6, relating to the 77% probability of developing cervical cancer, here too we can see which are the features that "weigh" the most in the prediction; personal hygiene (-7%) and eating behavior (+5%) therefore represent two risk factors to be kept under control in a clinical decision-making context associated with this pathology. Both methods of explainability produce interesting and very similar results. This tool can be of great use to the clinical decision-maker.

4 Conclusions

The cervical cancer problem that has been dealt with in this work by advanced machine learning algorithms is a complex problem but it has been possible to deduce procedures that have provided promising results in clinical research supported by artificial intelligence tools. This work is set in a fairly recent literature framework as regards the methodologies used in its prevention and risk prediction; in the future, the diagnosis and classification of the disease will be even faster by the development of new methods. The use of ensemble methods, as demonstrated by the results obtained, is a flexible, interpretable and significantly reliable tool in the treatment of this type of problem. The results obtained from the experiment carried out in this work are of a twofold nature: firstly it was possible to build an ensemble type classifier that reaches an accuracy of 94.5% and an AUC value of 0.98, higher values than the applications carried out on this dataset. Using advanced tools of Explainable ML it was possible to obtain both information on the black box models used (such as neural networks, MLP and the ensemble itself) and interpret them from the point of view of clinical observation and also interpret the features (i.e. the variables of the dataset) understanding the relationships

Fig. 6 Explainability with Shapley method: positive class, patient id: 17



between them and the target variable, which is the presence or absence of cervical cancer. In order to give concrete examples of interpretability, two units (two patients) were selected from the data test partition and analyzes were conducted on them, identifying which factors influenced the probabilistic predictions of the algorithms used. Different relationships have emerged as seen through the LIME and shapley methods difficult to extrapolate from a first glance at the data, therefore the clinical decision-maker using more advanced and more complex tools of machine learning algorithms is able to interpret the results and make accurate decisions, understand the nature of the predictions and how they were evaluated by the model used.

Funding Open access funding provided by Università degli Studi di Roma La Sapienza within the CRUI-CARE Agreement.

Declarations

Conflict of Interest The authors declare that they have no conflict of interest.

References

1. 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. Exp Syst Appl, 2010.
2. Liberatore MJ, Myers RE, Nydick RL, et al. Decision counseling for men considering prostate cancer screening. Comput Oper Res, 2003.
3. Dolan JG, Frisina S. Randomized controlled trial of a patient decision aid for colorectal cancer screening. Med Decis Making. 2002;22:125–39.
4. Tseng CJ, Lu CJ, Chang CC, Chen GD. Application of machine learning to predict the recurrence-proneness for cervical cancer. Neural Comput Appl. 2014;24(6):1311 1316.
5. Sharma S. Cervical cancer stage prediction using decision tree approach of machine learning. Int J Adv Res Comput Commun Eng. 2016;5(4):345 348.
6. Wu W, Zhou H. Data-driven diagnosis of cervical cancer with support vector machine-based approaches. IEEE Access 5:25189 25195, 201
7. Geetha R, Sivasubramanian S, Kaliappan M, et al. Cervical Cancer Identification with Synthetic Minority Oversampling Technique and PCA Analysis using Random Forest Classifier. J Med Syst. 2019;43:286.
8. Sobar MR, Wijaya AI. Behavior Determinant Based Cervical Cancer Early Detection with Machine Learning Algorithm. Adv Sci Lett, 22, 3120-3123, 2016.
9. Walker SH, Duncan DB. Estimation of the probability of an event as a function of several independent variables. Biometrika. 1967;54(1/2):167–78.
10. James G, Witten D, Hastie T, Tibshirani R. Tree-Based Methods (PDF). An Introduction to Statistical Learning: with Applications in R. New York: Springer. pp. 303–336, 2017.
11. Goodfellow I, Bengio Y, Courville A. 6.5 Back-Propagation and Other Differentiation Algorithms. Deep Learning. MIT Press. pp. 200-220, 2016.
12. Opitz D, Maclin R. Popular ensemble methods: An empirical study. J Artif Intell Res. 1999;11:169–98.
13. Muller J, Stoehr M, Oeser A, Gaebel J, Streit M, Dietz A, et al. A visual approach to explainable computerized clinical decision support. Comput Graph, 2020.
14. Schafer H, Hors-Fraile S, Karumur RP, 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.
15. Tonekaboni S, Joshi S, McCradden MD, Goldenberg A. What clinicians want: contextualizing explainable machine learning for clinical end use, 2019. arXiv:1905.05134
16. Ucla E. Outlining the design space of explainable intelligent systems for medical diagnosis, 2019.
17. 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.
18. 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.
19. Naiseh M. Explainability Design Patterns in Clinical Decision Support Systems. In: Dalpiaz F., Zdravkovic J., Loucopoulos P. (eds) Research Challenges in Information Science. RCIS. Lecture Notes in Business Information Processing, vol 385. Springer, 2020.
20. Ribeiro MT, Singh S, Guestrin C. Why should i trust you?: Explaining the predictions of any classifier. arXive, 2016.
21. Friedman JH. Greedy function approximation: A gradient boosting machine. Ann Stat. 2001;29:1189–232.
22. 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.
23. Apley DW. Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models. arXiv 2016. arXiv:1612.08468
24. Friedman JH, Popescu BE. Predictive learning via rule ensembles. Ann Appl Stat. 2008;2:916–54.
25. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems; MIT Press: Cambridge, MA, USA, pp. 4765-4774, 2017.
26. Koh PW, Liang P. Understanding black-box predictions via influence functions. ArXiv preprint arXiv:1703.04730, 2017.
27. Blanco-Justicia A, Domingo-Ferrer J, Martinez S, Sanchez D. Machine learning explainability via microaggregation and shallow decision trees. Knowl-Based Sys. 2020;194:105532.
28. Arrieta B, Rodriguez AD, Del Ser N, Bennetot J, Tabik A, Gonzalez SB, Garcia A, Gil-Lopez S, Molina S, Daniel Benjamins V, Chatila R, Raja HF. Explainable Artificial Intelligence (XAI): Concepts. Opportunities and Challenges toward Responsible AI: Taxonomies; 2019.
29. Cramer, Harald. Mathematical Methods of Statistics. Princeton: Princeton University Press, page 282, 1946.