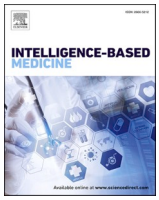
Intelligence-Based Medicine 9 (2024) 100125

Contents lists available at ScienceDirect

Intelligence-Based Medicine

journal homepage: www.sciencedirect.com/journal/intelligence-based-medicine



Machine learning classification of vitamin D levels in spondyloarthritis patients

Luis Ángel Calvo Pascual [a,*], David Castro Corredor [b], Eduardo César Garrido Merchán [a]

[a] Comillas Pontifical University, ICADE, Alberto Aguilera, 25, 28015, Madrid, Spain
[b] Department of Rheumatology, Hospital General Universitario de Ciudad Real, Spain

ARTICLE INFO

Key Indexing Terms:
Ankylosing spondyloarthritis
Machine learning
Arthritic psoriasis
1 alpha
25 dihydroxy 20 epi vitamin d3

ABSTRACT

Objectives: Predict the 25 dihydroxy 20 epi vitamin d3 level (low, medium, or high) in spondyloarthritis patients.
Methods: Observational, descriptive, and cross-sectional study. We collected information from 115 patients. From a total of 32 variables, we selected the most relevant using mutual information tests, and, finally, we estimated two classification models using machine learning.
Result: We obtain an interpretable decision tree and an ensemble maximizing the expected accuracy using Bayesian optimization and 10-fold cross-validation over a preprocessed dataset.
Conclusion: We identify relevant variables not considered in previous research, such as age and post-treatment. We also estimate more flexible and high-capacity models using advanced data science techniques.

1. Introduction

In recent years, much attention has been paid to machine learning applications in Rheumatology (see Refs. [1,2]). Machine learning is a subfield of computer science that deals with algorithms that estimate the optimal values for the parameters of statistical models to minimize the generalization error of the prediction for a given dataset [3]. Machine learning models obtain greater accuracy than classical models in predicting the values of a qualitative or quantitative variable in different populations of individuals and also find more complex relations between variables. Using machine learning, we offer a new perspective to address a controversial issue: the relation between serum levels of 25-hydroxy-vitamin D and inflammatory activity in spondyloarthritis patients.

Spondyloarthritis (SpA) is a group of chronic inflammatory rheumatic diseases mainly affecting the axial skeleton and the peripheral joints. Within this group are included ankylosing spondylitis (AS), non-radiographic axial spondylitis (nr-axSpA), psoriatic arthritis (PsA), reactive arthritis (RA), spondyloarthropathy associated with inflammatory bowel disease and undifferentiated spondyloarthropathy [4]. These pathologies share similar pathogenic mechanisms, clinical manifestations and a strong association with the HLA-B27 antigen. It should be noted, mainly in patients with axial spondyloarthropathy, that inflammation of the spine and sacroiliac joints leads to structural damage of these areas due to new bone formation as the body attempts to repair the damage. These structural modifications affect the patient's general condition by limiting their mobility and, as a result, disrupting their quality of life [5].

The Assessment of SpondyloArthritis international Society (ASAS) provides the precise definition and classification of SpA [6]. According to ASAS criteria, the diagnosis of spondyloarthritis is made if a patient presents two or more of the following features: the presence of the HLA-B27 gene, inflammatory low back pain, dactylitis, enthesitis, arthritis, uveitis, psoriasis, family history, elevated CRP, etc. In clinical diagnosis, machine learning models are not carried out; doctors just measure with different tests the variables and check if two or more measurements are above healthy values. Therefore, the power of machine learning methods has not been used to improve this disease's diagnosis and treatment. In this context, minimizing the generalization error would produce models with more accurate clinical predictions.

Current epidemiological evidence shows a significant association between vitamin D deficiency and a higher incidence of activity in autoimmune diseases, such as systemic lupus erythematosus (SLE), type 1 diabetes mellitus, multiple sclerosis, and rheumatoid arthritis (RA), a link that is not as clear with spondyloarthritis. Some authors consider that there is a correlation between vitamin D deficiency and the inflammatory activity in SpA patients (see Refs. [7,8,9,10,11]).

* Corresponding author.
E-mail addresses: lacalvo@comillas.edu (L.Á. Calvo Pascual), d.castrocorredor@gmail.com (D. Castro Corredor), ecgarrido@icade.comillas.edu (E.C. Garrido Merchán).

https://doi.org/10.1016/j.ibmed.2023.100125
Received 4 August 2022; Received in revised form 13 March 2023; Accepted 22 November 2023
Available online 6 December 2023

However, since this correlation is weak, other authors believe that these variables are independent (see Refs. [12,13,14]).

Although all these previous studies were valuable, they are based on classic explanatory models, such as logistic regression. Therefore, they are more focused on explaining how the independent variables modify the dependent variable than on minimizing the generalization error of their predictions, which is why they generally have low capacity. Consequently, they do not obtain the best model fitting all the data in terms of predictive accuracy or any other loss function when the accurate latent distribution of the data is complex. For example, these studies do not mention age as a relevant factor in explaining vitamin D levels in SpA patients. By contrast, we found that age is an important factor; in fact, it formed part of one of our machine learning models.

We aim to provide a sound methodology to estimate the parameters of a machine learning model to minimize the estimation of generalization error, in terms of its accuracy, in the problem of classifying vitamin D levels in SpA patients. In addition, since this method is a black box from a causal point of view, we also estimate an entirely interpretable decision tree, although it has less predictive power. Our procedure can be directly applied to other variables and clinical problems, especially in those where traditional statistics do not shed light on establishing precise regressions and classifications, such as the influence on a disease of factors such as sex, age, climate conditions, etc.

This paper is organized as follows: firstly, we do a brief review of the state-of-the-art. In Section 2, we present the patients, the variables, and the methodological foundations of our research. In Section 3, we describe the results. Using mutual information and a chi-square test, we study the level of dependence of the variables for vitamin D levels, selecting the most relevant variables to compute machine learning classification models. In addition, we describe the technical features and the accuracy of the models we obtained. We finish this paper with a discussion in Section 4, reviewing previous publications and comparing our results.

2. Methods

2.1. Study population

We collected information from 115 spondyloarthritis patients (according to ASAS 2009 criteria, see Ref. [15]) treated during outpatient visits at the Rheumatology Department of Hospital General Universitario de Ciudad Real between June 2018 and June 2019. The demographic information of our dataset can be consulted in Table 1.

Table 1
Descriptive information about patients.

Variable	Range	Number	Percentage
Age	0–20	4	3.48
	20–40	35	30.43
	40–60	58	50.43
	>60	18	15.65
Sex	M	64	55.65
	F	51	44.35
Family History	Yes	32	27.83
	No	83	72.17
Years of evolution	0–1	23	20.00
	1–4	32	27.83
	4–10	31	26.96
	10–20	20	17.39
	>20	9	7.83
SpA Subtype	nr-axSpA	12	10.43
	AS	59	51.30
	PsA	24	20.87
	ASpA	9	7.83
	Reactive Arthritis	11	9.57

2.2. Variables

We recorded 33 categorical and numerical variables for each patient grouped into five categories: descriptive, primary clinical manifestation, therapeutic options, inflammatory activity, and vitamin D levels.

The descriptive variables are sex, age, the subtype of spondyloarthritis, family history, and the evolution of SpA in years (Table 1).

As far as clinical manifestation is concerned, we assessed: the presence of axial affectations, peripheral arthritis/synovitis, enthesitis, dactylitis, uveitis, psoriasis, and inflammatory bowel disease, as well as imaging findings if sacroiliitis was evident based on the New York criteria, syndesmophytes, and bone edema as detected by MRI.

This study also considers the therapeutic options that the patient was given: corticosteroids, nonsteroidal anti-inflammatory drugs (NSAIDs), synthetic disease-modifying anti-rheumatic drugs (methotrexate and sulfasalazine), biologic drugs such as anti-TNF alpha (etanercept, adalimumab, golimumab, infliximab, and certolizumab), anti-IL17 and anti-IL-12/23. We used two (yes/no)-categorical variables to indicate if the patient received calcium-vitamin D/vitamin D in the pre-treatment or post-treatment.

The inflammatory activity variables selected for this study were BASDAI (Bath Ankylosing Spondylitis Diseases Activity Index) for patients with axial spondyloarthritis and DAPSA (Disease Activity for Psoriatic Arthritis) for patients with psoriatic arthritis. A BASDAI and DAPSA below four were defined as control of the inflammatory activity of the disease. Likewise, elevated acute phase reactants were recorded using ESR and CRP.

Finally, we measured serum parathyroid hormone (PTH) concentration, as well as calcium, and phosphate concentrations. The target variable in our study, serum levels y of 25-OH-vitamin D was divided into three groups: deficit ($y < 20$ ng/ml), insufficiency ($20 \leq y \leq 30$ ng/ml) and desirable-optimal ($y > 30$ ng/ml).

2.3. Methodology

2.3.1. Feature selection

Firstly, we discriminate whether any of the variables x_i independently explain the vitamin D levels y or not. To do so, we estimate, for any explicative variable x_i, the mutual information $i(x_i, y)$ that measures the difference between the joint distribution $p(x_i, y)$ and the products of the marginal distributions $p(x_i)\,p(y)$. This quantity explains the amount of information gained about y when x_i is known. In particular, it is given by the following expression:

$$i(x_i,y) = KL(p(x_i,y), p(x_i)\,p(y)) \tag{1}$$

where KL is the Kullback-Leibler divergence function. Observe that if a variable x_i does not add any information to y, then the joint distribution $p(x_i, y)$ of the variables would be equal to the product of the marginal distributions $p(x_i)\,p(y)$ which implies $p(x_i|y) = p(y)$, i.e. y doesn't affect x_i. The mutual information regressor relies on entropy estimation from k-nearest neighbor distances (see equation (8) of [16]). We computed this information regressor 1000 times and estimated the mutual information as the means, using the scikit-learn library [15]. We used the mutual information computation of vitamin D levels with respect to the other variables as a feature selector in our dataset (see Fig. 1 and section 3.1).

On the other hand, to obtain a complete perspective on the problem of choosing predictor variables, we also performed a classical chi-square test [17] using the Matlab function *fscchi2* [18]. With this function we obtained $-log(p_i)$, where p_i is the p-value of the chi-squared test for categorical data of each variable x_i with respect to the vitamin D levels. The chi-squared test for categorical data is based on that if the components of the vectors x_i, y_i are x_i^j, y^j respectively, then

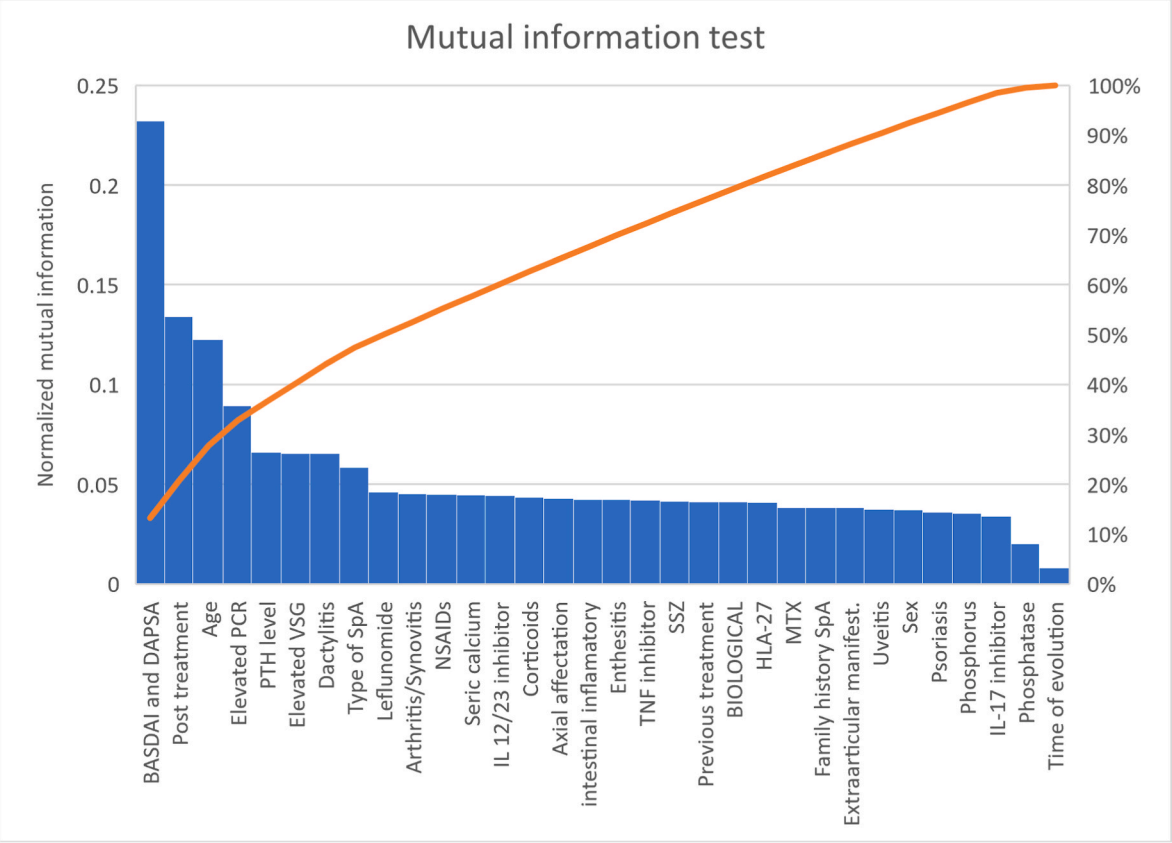


Fig. 1. Pareto chart of mutual information of the variables with respect to vitamin D levels.

$$\sum_j \left(x_i^j - y^j\right)^2 \Big/ y^j \qquad (2)$$

approximates to a $\chi^2_{1,114}$ distribution (the numbers 1, 114 correspond to 2 variables, 115 entries). We used the chi-square test as a feature selector in our dataset, to complement the mutual information test (see

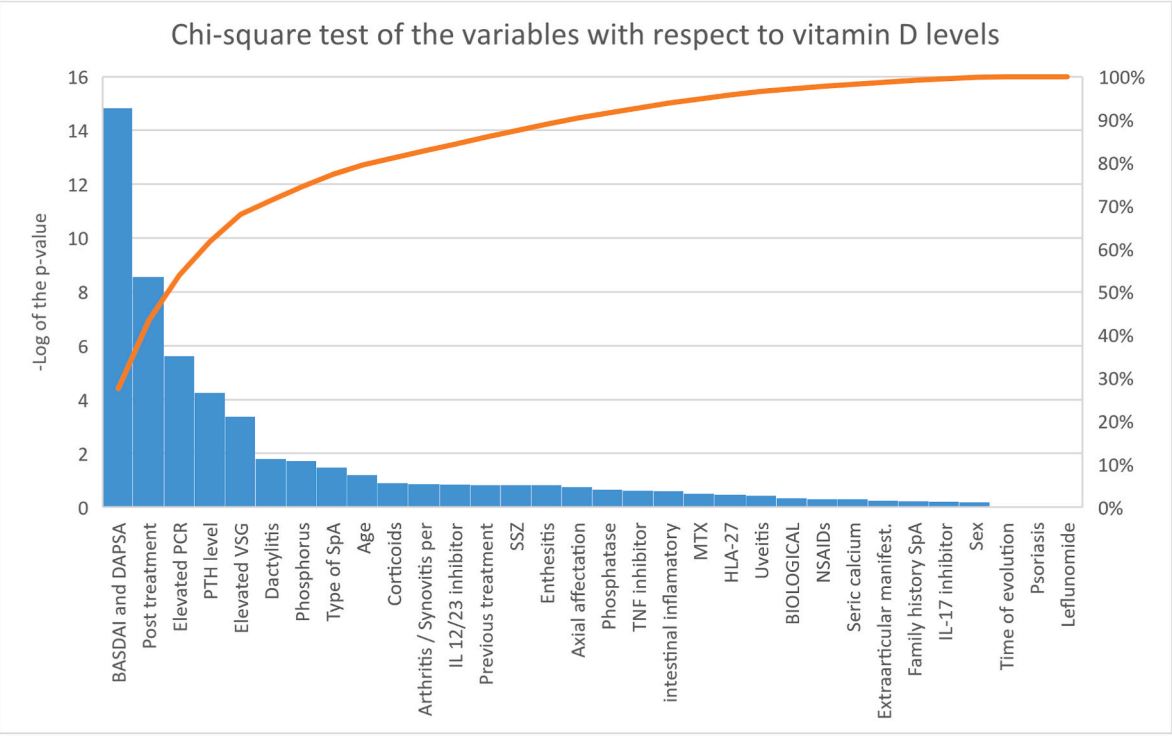


Fig. 2. Pareto chart of -log of the p-value of the chi-square test of the variables with respect to vitamin D levels.

Fig. 2 and section 3.1).

To sum up, from our dataset of 33 variables, we selected the features with the highest mutual information with respect to vitamin D levels (Fig. 1). Those variables will form part of a machine learning model. Complementary, we also consider the variables with higher chi-squared tests with respect to vitamin D levels (Fig. 2), to obtain the variables that the authors should have considered in their classical regression models.

2.3.2. Interpretable model

Decision forests try to minimize the impurity, estimated via the entropy criterion

$$\mathrm{H(D)} = -\sum_i log\ (p_i)\ p_i, \tag{3}$$

of a given dataset D where p_i are the proportions of each variable value by splitting the dataset hierarchically according to the value of the variables that most minimizes the impurity of the dataset D. Decision trees have several hyper-parameters θ_i such as the maximum depth of the tree or the value of the regularizer, *i.e.,* the pruning hyper-parameter. To minimize the generalization error, estimated by repeated 10-fold cross-validation, with respect to the hyper-parameter space Θ of the decision trees, we use a random search procedure of 1000 iterations using the Random Forest library of R [19].

We computed an interpretable model to predict vitamin D levels in SpA patients (see Fig. 3 and section 3.2).

2.3.3. Machine learning methodology

We estimate the machine learning model M with the best-expected accuracy a^* classifying Vitamin D levels y depending on BASDAI and DAPSA levels, post-treatment, CRP and ESR levels, and PTH concentration, summarized in dataset $\mathrm{D} = (\mathbf{X}, \mathbf{y})$. Hence, we solve the following optimization problem in an N-space of machine learning models with their hyper-parameters θ_M

$$a^* = Max_{M \in N} E(f(M, \theta_M, D))M, \theta_M, D))f \tag{4}$$

where f is an estimator that retrieves the computed accuracy of model M in dataset D using k-fold cross-validation. Recall that a^* is the optimal accuracy over all the N-space of machine learning models with their hyper-parameters. In other words, if the model M hyper-parametrized with θ_M is the optimal one, then, a^* is just its accuracy, that is, its percentage of correctly classified instances.

To compute a^* and find the machine learning model $M \in N$, we use the Classification Learner library of Matlab [18], which trains models fitting the data with 10-fold cross-validation on a total of 24 different models (see Table 4). We provide a short description of each of them. Neural networks apply sequentially non-linear transformations

$$y = a_n\ (\boldsymbol{W_n}(\ldots a_1(\boldsymbol{W_1 x} + \boldsymbol{b_1})) + \boldsymbol{b_n}) \tag{5}$$

to a point $\mathbf{x}$ to determine the value y that minimizes a loss function optimizing the parameters $\boldsymbol{W_i}, \boldsymbol{b_i}$ with a gradient descent method. Trees create linear decision frontiers iteratively minimizing an impurity criterion on the data like the entropy iteratively, hence maximizing the leaves of the tree where the data lies after being split. The naive Bayes method simply models that all the explanatory variables are independent with respect to y and estimates p(y $|\mathbf{X}$). SVMs (Support Vector Machines) use the kernel trick to create a feature space where the data is linearly separable and compute the hyperplane whose support vectors are at a maximal distance hence splitting the data and being able to classify it robustly. The k-nearest neighbor is a lazy non-parametric learner algorithm whose logic is to assign the mean value to a new instance based on its k-nearest points according to some distance metric. Finally, ensembles are combinations of classifiers whose generalization error estimation has a lower variance, making them robust models. More information about machine learning models and algorithms can be found in Ref. [3].

To solve the previously mentioned optimization problem, the 24 different models are included in the N set of machine learning models. Each has an associated θ_M set of hyper-parameters. The classification learner Matlab library uses the iterative Bayesian optimization algorithm for such a task. Bayesian optimization solves the hyper-parameter tuning problem by iteratively using the predictive distribution of a

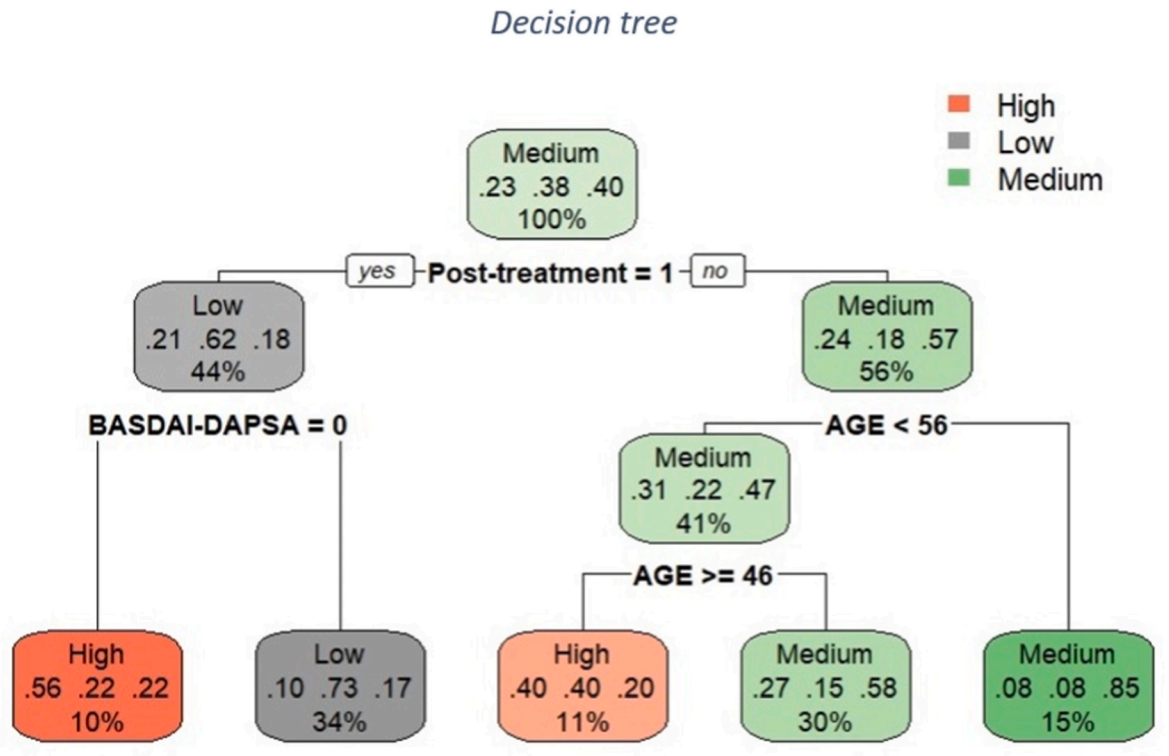


Fig. 3. Decision tree computed by a random search that minimizes the generalization error estimated via repeated 10-fold cross-validation.

probabilistic surrogate model over the hyper-parameter space N of machine learning models that can be a random forest or a Gaussian process model. This Gaussian predictive distribution $N(\mu, \Sigma)$, analytical in the case of a Gaussian process and empirical in the case of a random forest, contains a prediction of how every hyper-parameter and model (M, θ_M) would perform over the dataset $E(f(M, \theta_M, D)$. Bayesian optimization uses an analytical acquisition function $\alpha(\cdot)$ over $E(f(M, \theta_M, D)$, where f is the prediction algorithm of model M, which represents a trade-off between exploration and exploitation. In each iteration, Bayesian optimization maximizes the acquisition function $\alpha(\cdot)$ using a cheap optimization procedure such as the L-BFGS algorithm, described in Ref. [20], to suggest the hyper-parameters and the model (M, θ_M) that maximize the exploration and exploitation trade-off. The L-BFGS (Limited Memory Broyden Fletcher Goldfarb Shanno algorithm) is a quasi-Newton method, that iteratively optimizes a function using first and second-order derivatives, approximating the BFGS method with less computer memory. It optimizes the acquisition function, which we can differentiate, approximating the Hessian matrix H using a set of representative vectors for gradient updates with respect to the input space of the optimization problem. If the acquisition function is not differentiable, Bayesian optimization can use a genetic algorithm, or another metaheuristic, to estimate a local optimum. Finally, once the iterations of the procedure are finished, Bayesian optimization recommends the best-observed combination of hyper-parameters and models (M, θ_M) to solve the problem. As a result, in the context of vitamin D classification, this procedure will return the machine learning models that maximize the expected accuracy of the classification problem.

Please observe that no validation set is needed in this methodology as we do not use the train-test partition to fit the models. In particular, the metric that we optimize is the 10-fold cross-validation estimation of the generalization error, which is performing 10 different partitions of all the dataset, where the test set in each one of the partitions is one portion of the dataset and the estimation is giving as the average between the 10 models in each of the iterations of the Bayesian optimization process. Consequently, we are fitting the models with respect to all the datasets, minimizing the probability of overfitting as the k-fold cross-validation methodology has less bias and variance as an estimator of the generalization error that the train-validation-test methodology and making data leakage not possible. In other words, if the Bayesian optimization has T iterations as budget and in every iteration, we estimate the generalization error via 10-fold cross-validation over an ensemble of K models we estimate the parameters of a total of T*10*K models over 10 different partitions of the dataset minimizing the bias and variance of our generalization error estimator.

It is important to emphasize that the estimated accuracy loss function that the hyperparameter-tuned machine learning model (M, θ_M) incurs with respect to the dataset D does not encode, or represent, an exploration-exploitation trade-off in the hyper-parameter space N. The criterion that encodes the exploration-exploitation search behavior in the dataset is the acquisition function $\alpha(\cdot)$ of the Bayesian optimization method. For example, the upper confidence bound acquisition function $\alpha(\cdot)$ balances the prediction of the loss function $\mu(M, \theta_M)$ that a particular tuned model (M, θ_M) does and the uncertainty about that prediction $\sigma(M, \theta_M)$, and is given, for each configuration θ_M, by the following expression

$$\alpha(N(\mu,\Sigma)) = \mu(M,\theta_M) + \lambda\sigma(M,\theta_M), \tag{6}$$

where λ is a Bayesian optimization hyper-parameter that balances exploitation and exploration whose most used default value is 0.1. The acquisition function is computed over a grid that covers all the space N using the Gaussian predictive distribution $N(\mu, \Sigma|D)$ of the Gaussian process model over previous observations D. Lastly, we focus on the accuracy loss function instead of other popular classification loss functions such as recall or F-measure since what is most important for us is the success in the prediction. In other words, false positives and false negatives have equal value for us, what is important is the sum of true positives and true negatives with respect to all the data. And, in particular, that is what accuracy encodes.

We used this machine learning methodology to compute our non-interpretable machine learning model to predict vitamin D levels in SpA patients (see Section 3.3).

3. Results

3.1. Feature selection

We estimate mutual information between all the variables x_i with respect to the target variable, the vitamin D levels y (low, medium, high). As shown in Fig. 1, the variables most dependent on y are BASDAI and DAPSA levels (0.23), post-treatment (0.13), age (0.12), and CRP level (0.09).

On the other hand, we also rank the importance of the predictors using chi-square tests (Fig. 2). In this case, the most correlated variables with y, ordered by -log of the p-value, are BASDAI and DAPSA levels (14.83), post-treatment (8.56), CRP level (5.61), and PTH concentration (4.25).

From a clinical point of view, as explained in the introduction, there is a controversy on the relation between vitamin D levels and the activity of SpA, measured with BASDAI, DAPSA and CPR, because the classical linear correlation is weak. We can confirm a functional dependency between both variables thanks to the mutual information test (Fig. 1). On the other hand, we found that age, although it has a low linear correlation with Vitamin D (Fig. 2), it presents higher mutual information than other variables (Fig. 1) and that is why age participates in the explicative machine learning model. Post-treatment and PTH were also not considered in previous studies. These variables are more weakly correlated than activity with respect to Vitamin D levels. Despite this, they are important from a machine learning point of view, and also form part of our models.

3.2. Interpretable model

We obtain a decision tree with an accuracy of 0.5185, where the majority rule criterion is 0.423, meaning that the model covers some variability about the target using the explanatory variables. The optimized hyperparameters and their ranges can be consulted in Table 2. The shape of the decision tree is illustrated in Fig. 3, where high values of vitamin D levels occur in patients who received post-treatment and have low BASDAI-DAPSA inflammatory activity. Each node of Fig. 3 contains the three probabilities of the different classes of Vitamin C levels (high, low, medium) in the training data belonging to each node. Each node also contains the percentage of patients based on the conditions of the tree.

Concerning the performance, in this case, the F1 score is 0.5792, the kappa error is 0.064, and Cohen's kappa is 0.33. (Table 5). This performance is low for a 3-class classification problem and we believe that it is due to the nature of the problem and the low levels of mutual information.

3.3. Predictive model

The best machine learning model obtained is a subspace discriminant

Table 2
Optimal hyperparameters for the Random Forest model.

Hyperparameter	Value
Ensemble method	RUSBoost
Number of learners	29
Learning rate	0.0137
Maximum number of splits	2

ensemble that presents an accuracy of 61.7 % in the confusion matrix (Fig. 4). The optimized hyperparameters and their ranges can be consulted in Table 3. An ensemble is a set of weak classifiers, such as decision trees of width, which jointly predict the class's label. For example, an ensemble of 1000 decision trees of length one will create an empirical predictive distribution of the vitamin D level. The critical factor of an ensemble is that each of the classifiers is specialized in a particular region of the feature space, generating negative correlations in their prediction. As a result, the variance of the generalization error of the ensembles is lower than the one of a single classifier, creating more robust predictors that generally perform better in practice in terms of the loss function of the problem, for example, the accuracy.

The errors in predicting high vitamin D levels being low and vice versa are minor (3 and 4 cases out of 115). The F1-score is 0.6281, the kappa error 0.07, and Cohen's kappa is 0.40. (Table 5). The ROC curve of this model (Fig. 5) shows the goodness of our model, where the area under the curve is 0.75 and, in addition, our model is balanced, meaning that the ensemble model, represented by the red dot, is approximately located in the middle of the curve. Finally, we should mention that the performance is low for a 3-class classification problem and we believe that it is due to the nature of the problem and the low levels of mutual information.

4. Discussion

Some publications pointed out the relationship between inflammatory activity in SpA-patients and vitamin D levels. In Ref. [7], an association was observed between low vitamin D levels and higher disease activity levels, measured by BASDAI and DAPSA. In Ref. [10], the authors proved that patients with vitamin D deficiency had a higher activity of disease measured by ASDAS, although the differences observed were small. In line with this, in Ref. [11], there is a systematic review of patients with ankylosing spondylitis (AS), where the vitamin D levels in AS patients were lower than healthy, although the cause is unclear. Finally, in Ref. [9], patients with ankylosing spondylitis and undifferentiated spondyloarthritis negatively correlated with vitamin D levels and ESR and CRP. Still, they did not find any relation between vitamin D levels and BASDAI-DAPSA activity.

On the other hand, in Ref. [21], the authors did not find a clear relationship between vitamin D deficiency and the worse course of the disease in patients with spondyloarthritis compared to a healthy control

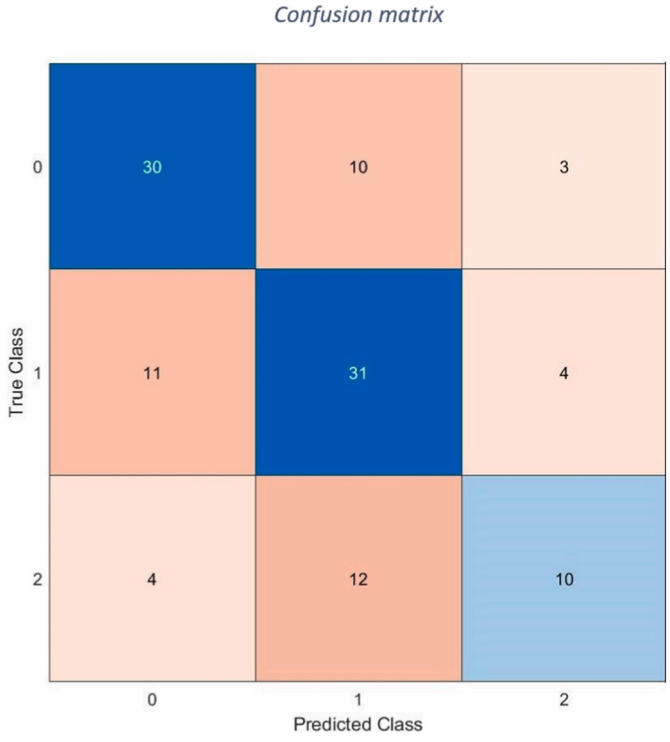


Fig. 4. Confusion matrix of the Ensemble method with 10-fold cross-validation.

Table 3
Optimal hyperparameters for the Ensemble model.

Classification models	Subtypes
Classification Trees	Fine, Medium and Coarse
Discriminant Analysis	Linear and Quadratic
Naive Bayes	Gaussian and Kernel
Nearest Neighbors	Fine, Medium, Coarse, Cosine, Cubic and Weighted
Support Vector Machine Classification	Linear, Quadratic, Cubic, Fine Gaussian, Medium Gaussian and Coarse Gaussian
Classification Ensembles	Subspace discriminant, Subspace KNN, RUSBooted Trees, Boosted Trees and Bagged Trees
Neuronal Networks	Narrow, Medium, Wide, Bilayered and Trilayered

Table 4
Classification Learner App models.

Hyperparameters	Range
Maximum depth	[3,10]
Minimum number of samples in each terminal node of the tree	[3,10]
Penalty parameter	[0.01, 0.1]
Minimum number of pieces needed to do a split	[3, 20]

Table 5
Comparison of the performance of the models.

	Interpretable model	Predictive model
Accuracy	0.58	0.63
F1 score	0.06	0.07
Kappa error	0.33	0.4
Cohen kappa	0.52	0.62
AUC	0.58	0.75

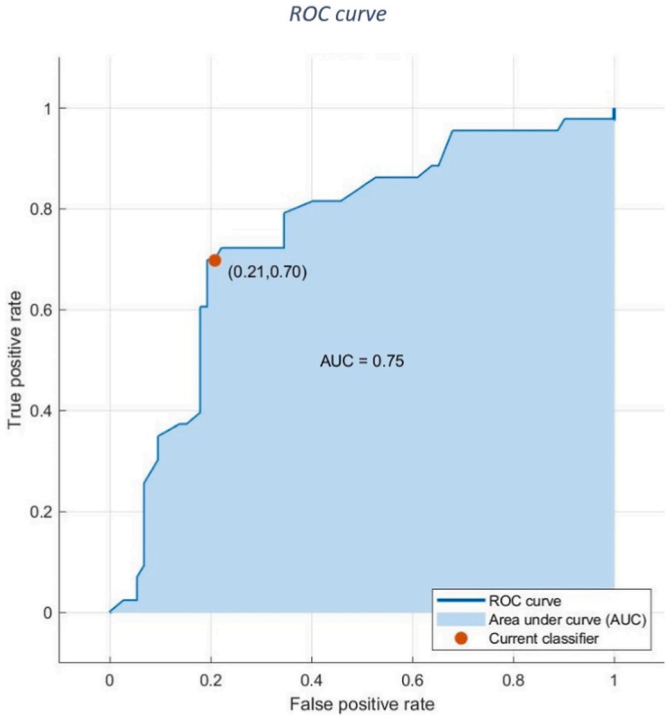


Fig. 5. Roc curve of the Ensemble method with 10-fold cross-validation.

group. Additionally, in Ref. [12], there is no association between vitamin D levels and disease activity in axial and peripheral spondyloarthritis patients. Finally, in Ref. [13], the authors conclude no significant correlation between vitamin D levels and the disease's severity.

In this paper, we show how a Bayesian optimization methodology can be applied to predict vitamin D levels in SpA patients, and in general, could be applied in any clinical problem involving a variable to predict y with respect to other variables X. Our Bayesian method optimized the used machine learning method m and the value of its associated hyper-parameters $\boldsymbol{H}^*,m^*$ In other words, we have solved the hierarchical black-box non-convex global optimization problem

$$\boldsymbol{H}^*,m^* = argmin\ L(\boldsymbol{H},m|\boldsymbol{D}), \tag{7}$$

where L is the loss function used to encode how good is the model m performance and $\boldsymbol{D}$ is the dataset. We want to emphasize that any clinical problem can be wrapped in the supervised learning framework. Any situation where we want to fit the joint probability distribution $p(\boldsymbol{X},y)$ of a variable y with respect to other variables $\boldsymbol{X}$, will have its performance boosted with a Bayesian optimization methodology such as the one presented in Ref. [22]. Bayesian optimization can be used to optimize any set of models, perfectly wider than the one shown in Ref. [23], and also includes deep learning techniques [22] and for their hyper-parameter configuration, providing better predictive results that can be useful as a support for taking decisions by the professional. Finally, Bayesian optimization could also be used not just to minimize the precision or any other performance loss function but also to simultaneously optimize other objectives such as the recall, f-measure, a personalized loss function, prediction time, size of the models, risk or other interesting goals using many objective Bayesian optimizations [24].

In particular, in this paper, we selected the most relevant functions to explain vitamin D levels (See Figs. 1 and 2 and Section 3.1), obtaining age as a relevant factor, not considered in the previous research. From a medical point of view, that makes sense since there is a decrease in the cutaneous capacity of synthesis of vitamin D in older patients. Other factors such as inflammatory activity, PTH concentration and post-treatment were not considered, since, individually, they present a low linear correlation with respect to vitamin D levels. However, when these variables are simultaneously considered, we obtain more complete models, such as an ensemble with 61.7 % of accuracy in the confusion matrix (Fig. 4) and a decision tree (Fig. 3) with an accuracy of 51.85 %, in which age, post-treatment and PTH play an essential role. The performance obtained by our methodology, an accuracy of 61.7 % in the confusion matrix and an AUC of 0.75, suggests that more data needs to be considered to optimize these loss functions. More extensive randomized controlled trials are required. Recall that machine learning methods encode the patterns that explain a target variable **y** with respect to a set of explanatory variables. In probability terms, they infer the value of the parameter tensor **W** of a joint probability distribution $p(\mathbf{y},\mathbf{X}|\mathbf{W},\mathbf{D})$, from dataset **D**. Consequently, this probability distribution does not encode causality, but highly complex non-linear correlations. Hence, it is useful to infer the causality between the variables from a methodology like randomized controlled trials.

Finally, we emphasize that we do not have full access to the theoretical joint probability distribution of the data $p(\boldsymbol{X},y)$, but only to a representative dataset **D**=(**X, y**) sample of it. As a result, we optimize the hyper-parameter values of the machine learning algorithms with the most practical unbiased methodology coming from the machine learning community to estimate the generalization error of a machine learning algorithm [22,3]: Bayesian optimization of the k-fold cross-validation of the accuracy loss function of the machine learning algorithm. If **D** is representative of the theoretical joint probability distribution of the data $p(\boldsymbol{X},y)$, then, the estimated generalization error will be accurate. The only way to improve the properties of this estimator would be to perform one-leave-out cross-validation, but we did not choose this estimator as it is unfeasible to do it in practice, as it requires to estimate as many models as data points in the sample. For further work, to assess the generalization capability of this methodology and to research why the majority of errors belong to the misclassification of one label value, it would be useful to perform randomized controlled trials or interventions in a causality model.

Support and grants

No.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Kim K-J, Tagkopoulos I. Application of machine learning in rheumatic disease research. Korean J. Intern. Medv. 2019;34(4):708–22.

[2] Shoop-Worrall S, Cresswell K, Bolger I, Dillon B, Hyrich K, Geifman N. Nothing about us without us: involving patient collaborators for machine *learning, Applications in Rheumatology*. Ann Rheum Dis 2021;80:1505–10.

[3] Murphy KP. Machine learning: a probabilistic perspective. MIT Press; 2012.

[4] Collantes-Estevez E. New paradigms in the diagnosis and classification of spondylarthritis. Reumatol Clínica 2013;9:199–200.

[5] Rios V, Poddubnyy D. Old and new treatment targets in axial spondyloarthritis. RMD Open 2015;1:e000054.

[6] Rudwaleit M, et al. The Assessment of SpondyloArthritis international Society (ASAS) handbook: a guide to assess spondyloarthritis. Ann Rheum Dis 2009;68 (Suppl 2):ii1–44. https://doi.org/10.1136/ard.2008.104018.

[7] Castro Corredor D, Ramírez Huaranga MA, Mínguez Sánchez MD, Anino Fernández J, Mateos Rodríguez JJ, Rebollo Giménez AI, González Peñas M, Seoane Romero J, Luque Zafra M, de Lara Simón IM, Cuadra Díaz JL. Vitamin D, an inflammatory activity marker for spondyloarthritis? Arch Osteoporosis 2020;15(1): 126.

[8] Castro D, Ramírez MA, Mínguez MD, Luque M, Cuadra JL. El déficit de vitamina D en pacientes con espondiloartritis en un hospital de Castilla-La Mancha. Rev Colomb Reumatol 2020. https://doi.org/10.1016/j.rcreu.2020.09.004.

[9] Erten S, Kucuksahin O, Sahin A, Altunoglu A, Akyol M, Koca C. Decreased plasma vitamin D levels in patients with undifferentiated spondyloarthritis and ankylosing spondylitis. Intern. Med. 2013;52:339–44.

[10] Fernandez S, Etcheto A, van der Heijde D, Landewé R, van den Bosch F, Dougados M, Moltó A. Vitamin D status in spondyloarthritis: results of the ASAS-COMOSPA international study. Clin Exp Rheumatol 2018;36(2):210–4.

[11] Pokhai GG, Bandagi S. Abrudescu A Vitamin D levels in ankylosing spondylitis: does deficiency correspond to disease activity? Rev. Bras. Rheumatol. 2014;54: 330–4.

[12] Guła Z, Kopczyńska A, Hańska K, Słomski M, Nowakowski J, Kwaśny-Krochin B, Gasowski J, Korkosz M. Vitamin D Serum concentration is not related to the activity of spondyloarthritis. Reumatologia 2018;56(6):388–91.

[13] Kolahi S, Khabbazi A, Kazemia N, Mahdavi AM. Does vitamin D deficiency contribute to higher disease activity in patients with spondyloarthritis? Immunol Lett 2019;212:1–5. https://doi.org/10.1016/j.imlet.2019.06.005.

[14] Koçyiğit BF, Akyol A. Vitamin D levels in patients with ankylosing spondylitis: is it related to disease activity? Pakistan J Med Sci 2018;34(5):1209–14.

[15] Pedregosa F, et al. Scikit-learn: machine learning in Python. J Mach Learn Res 2011;12(Oct):2825–30.

[16] Kraskov A, Stogbauer H, Grassberger P. Estimating mutual information. Phys Rev E 2004;69.

[17] McHugh ML. The chi-square test of independence. Biochemical medica 2013;23(2): 143–9.

[18] MATLAB and fscchi2 function R. Natick, Massachusetts, United States: The MathWorks, Inc.; 2021. https://es.mathworks.com/help/stats/fscchi2.html.

[19] RStudio Team. Random forest library. Boston, MA: RStudio: Integrated Development for R. RStudio, PBC; 2020. https://cran.r-project.org/web/packages/randomForest/index.html.

[20] Liu D, Nocedal J. On the limited memory BFGS method for large scale optimization. Math Program 1989;45(3):503–28. Ser. B).

[21] Crotti C, Becciolini A, Biggioggero M, Favalli EG. Vitamin D and spondyloarthritis: review of the literature. Open Rheumatol J 2018;12(Suppl-1):214–25. M3.

[22] Garrido Merchán EC. Advanced methods for Bayesian optimization in complex scenarios. PhD thesis. UAM; 2021.

[23] Wang Z, Zoghi M, Hutter F, Matheson D, De Freitas N. Bayesian optimization in high dimensions via random embeddings. In: Ijcai; 2013, August. p. 1778–84.

[24] Martín LA, Garrido-Merchán EC. Many objective bayesian optimization. 2021. *arXiv preprint arXiv:2107.04126*.