Supporting decision-makers in healthcare domain. A comparative study of two interpretative proposals for Random Forests

Massimo Aria, Corrado Cuccurullo, Agostino Gnasso

1. Introduction

Today, the availability of data is growing exponentially in all sectors, especially in the healthcare sector. Machine Learning (ML) techniques allow to analyze big data to exctrat knowledge and support healthcare activities (Miotto et al., 2018), such as models for the diagnosis of complex diseases (Dhillon and Singh, 2019), (Aria et al., 2020). Despite the use of ML is spreading in many applications, it is characterized by some limitations and disadvantages.

ML main drawback corresponds to its lack of interpretability which does not allow users to represent causal relationships and interactions between predictors and response. This leads to the inability to learn how particular decisions are made. From this problem derives the definition of the Black Box model, a highly accurate model with a large complexity that cannot be represented by a relational structure. In other words, it is not possible to visualize how it internally works.

Furthermore, the opaque nature of these models hinders application in various sectors, especially in critical ones such as healthcare. To undertake a decision-making process, having faith in a machine learning model is essential, to feel reassured when analyzing and using it.

Ribeiro et al. (2016) identify a different but at the same time-related definitions of trust: trust in a prediction and trust in a model. Trusting a prediction implies that the user will take a certain action based on it; it is important to determine this confidence given that the model will be used to make decisions think for example of the use of a decision-making process in the clinical field, the consequence of acting with absolute confidence on the predictions obtained without being able to understand how they are obtained. Having faith in a model is equivalent to evaluating the model as a whole and testing its ability to generalize with appropriate evaluation metrics. A problem that recurs in using data from real contexts is that they are often significantly different and the chosen metric may not be adequate, therefore an inspection procedure of individual predictions and their interpretations may be the optimal choice.

In this work, we pay attention to one of the most used, accurate, and performing models in Machine Learning, the Random Forest model (RF) (Breiman, 2001).

Random Forest is an evolution of Bagging which aims to reduce the variance of a statistical model, simulates the variability of data through the random extraction of bootstrap samples from a single training set, and aggregates predictions on a new record (see Breiman, 1996). Being an evolution of Bagging, Random Forest aims to obtain even more different and unrelated trees. It is known as an efficient ensemble learning model, as it ensures high predictive accuracy, flexibility, and immediacy; it is recognized as an intuitive and understandable approach to the construction process, but is also considered a Black Box model due to the large number of deep decision trees produced within it (Haddouchi and Berrado, 2019).

Massimo Aria, University of Naples Federico II, Italy, massimo.aria@unina.it, 0000-0002-8517-9411
Corrado Cuccurullo, University of Campania Luigi Vanvitelli, Italy, corrado.cuccurullo@unicampania.it, 0000-0002-7401-8575
Agostino Gnasso, University of Naples Federico II, Italy, agostino.gnasso@unina.it, 0000-0002-9220-9754

FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Massimo Aria, Corrado Cuccurullo, Agostino Gnasso, *Supporting decision-makers in healthcare domain. A comparative study of two interpretative proposals for Random Forests*, pp. 179-184, © 2021 Author(s), CC BY 4.0 International, DOI 10.36253/978-88-5518-461-8.34, in Bruno Bertaccini, Luigi Fabbris, Alessandra Petrucci (edited by), *ASA 2021 Statistics and Information Systems for Policy Evaluation. Book of short papers of the on-site conference*, © 2021 Author(s), content CC BY 4.0 International, metadata CC0 1.0 Universal, published by Firenze University Press (www.fupress.com), ISSN 2704-5846 (online), ISBN 978-88-5518-461-8 (PDF), DOI 10.36253/978-88-5518-461-8

The results deriving from the use of the Random Forest are valuable. Various studies have confirmed RF effectiveness in many sectors, such as biomedical for genetic selection (Díaz-Uriarte and De Andres, 2006). Breiman et al. (2001) states that Random Forest has A + performance but, having a prediction process that is difficult to understand, evaluates an F on interpretability. This leads to Occam's dilemma (Domingos, 1998) (Domingos, 1999).

The poor interpretability has prevented the adoption of the model in some sectors where there is little or no tolerance for errors, such as healthcare and clinical context (Ahmad et al., 2018). Having set the common goal of interpretability, in recent years the scientific community has fueled considerable interest in Interpretable Machine Learning, which today is an extremely open and active research field with numerous approaches that continually emerge every year (Adadi and Berrada, 2018) (Du et al., 2019) (Guidotti et al., 2018).

This research focuses on the comparison between two approaches proposed in the literature that attempt to overcome the interpretative problem. These approaches, Node Harvest by Meinshausen (2010) and inTrees by Deng (2019), are based on a post-processing interpretation method. They are also defined as Rule Extraction (Haddouchi and Berrado, 2019) approaches as they are focused on the extraction of rule sets. Both proposals use an understandable model based on the rules extracted from a Random Forest. The general idea is to identify a representative weak model to provide the interpretation. This one is selected from the sequence of weak models generated by the ensemble procedure. In particular, Node Harvest selects the set of rules through weights that are assigned based on quadratic programming with linear inequality constraints. Performing this task manages to coincide with two objectives, such as interpretability and accuracy in prediction.

Similarly, inTrees obtain interpretable information through the extraction and processing of rules deriving from a tree ensemble sequence. The extracted rules are used for the realization of a learner, which serves to make predictions on new data.

inTrees works through a series of algorithms that, at first, extract the rules and classify them; subsequently, they carry out a pruning phase on each rule, eliminating the rules that produce background noise or that are irrelevant. Subsequently, these algorithms select a compact set of rules considered relevant and not redundant. Frequent interactions are extracted and finally, everything is summarized in a learner that will be used to make predictions on new data.

2. Comparison Study

We compare Node Harvest and inTrees on four health datasets.
Comparison analysis is performed in an empirical context, where their performance is evaluated using performance metrics. These are obtained from the output and are compared to a reference standard (Aria et al., 2021).

The metrics that evaluate the performance of predictive models, when used for classification, are based on the confusion matrix, which contains the expected and observed class labels, as well as the predicted target category and the source category, as can be seen from Table 1 which represents the structure of a 2x2 confusion matrix.

Regarding comparison, the goal is to compare these approaches through the use of different health datasets. The analysis is conducted on four binary classification health datasets. These datasets are available in the UCI Machine Learning repository. They have different characteristics (see Table 2).

Table 1: Confusion Matrix

	Actual Positive Class	Actual Negative Class
Predicted Positive Class	*TP (True Positive)*	*FP (False Positive)*
Predicted Negative Class	*FN (False Negative)*	*TN (True Negative)*

Table 2: Main characteristics of the selected health datasets.

Datasets	Obs.	Qual. Feat.	Quant. Feat.	0/1 Response Rate	Unbalanced Response
Diabetic Retinopathy Debrecen	1151	3	16	118/120	False
EEG Eye State	14980	1	15	2375/1822	False
Cardiovascular Disease	10500	7	5	883/707	False
Pima Indians Diabetes	768	8	1	130/45	False

The analysis follows the following structure: we proceed with carrying out the random forest for each of the four datasets to obtain the performance of the standard model, in terms of the confusion matrix and prediction of the target variable; the extraction of the set of rules is carried out to investigate the paths taken by each observation, of which the most important and frequent rules of the set itself will also be shown.

Finally, the comparison of the various sets of rules obtained from the two investigated methodologies is performed. The final performance evaluation is conducted through nine parameters obtained from the confusion matrices: Accuracy, Precision, Sensitivity, Specificity, G-Mean, F1 Score, Youden's Index, Balanced Accuracy, Kappa (see Sokolova et al., García et al., Akosa).

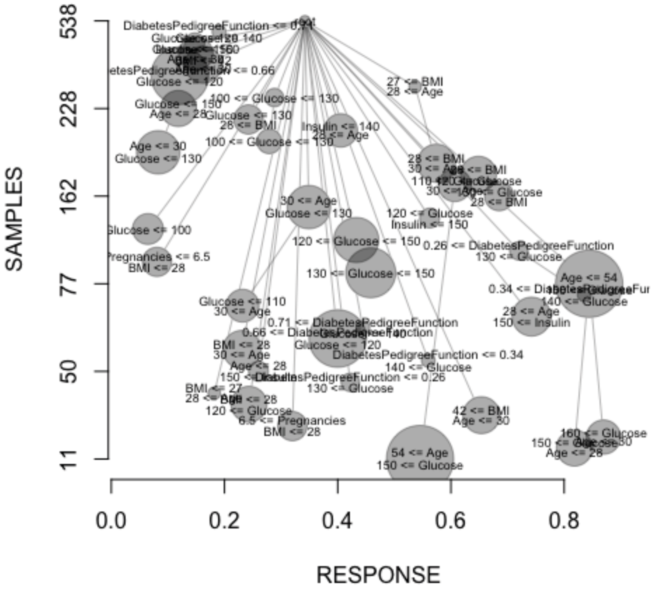
Examples are provided of the outputs obtained from the Node Harvest and inTrees approaches. These examples derive from the analysis conducted on Pima Indians Diabetes data: Node Harvest allows you to view the set of rules through an explanatory plot, provided in figure 1, while inTrees allows easy reading through summary tables that show the most frequent rule sets, such as in the table 3.

Table 3: inTrees (STEL) on Pima Indians Diabetes: set of decision rules that are easily applicable to new data. The impRRF value measures the relative percentage decrease in the Gini index for each rule derived from the random forest. The impRRF consider the length of each rule as a proxy of its complexity.

len	freq	err	condition	pred	impRRF
3	0.279	0.307	X[,2]>129.5 & X[,3]<=102 & X[,6]>27.2	1	1
2	0.326	0.366	X[,2]>114.5 & X[,8]>28.5	1	0.301
3	0.054	0.138	X[,1]>6.5 & X[,7]>0.6 & X[,7]<=1.41	1	0.162
4	0.134	0.278	X[,2]>96 & X[,5]<=34 & X[,6]>29.8 & X[,8]>30.5	1	0.144
2	0.84	0.282	X[,2]<=165.5 & X[,3]>39	0	0.139
1	0.553	0.219	X[,8]<=30.5	0	0.092
4	0.024	0.154	X[,1]<=4.5 & X[,2]<=168.5 & X[,5]>250 & X[,6]>29.85	0	0.088
3	0.184	0.232	X[,2]>127.5 & X[,6]>31.4 & X[,8]>24.5	1	0.073
2	0.786	0.261	X[,2]<=162 & X[,6]<=40.75	0	0.071
4	0.119	0.375	X[,3]<=77 & X[,5]<=118 & X[,6]>27.55 & X[,8]>30	1	0.060

Table 4 shows the nine performance metrics calculated on the four health datasets. The highest score, for each metric, is marked in bold. First of all, the interpretative solutions

Figure 1: Rule set plot obtained from Node Harvest on Pima Indians Diabetes.



proposed by Node Harvest (NH) and inTrees (STEL) represent an understandable approximation that provides an accurate summary of Random forest structure. All datasets show accurate measures very close to the reference value, provided by RF.

Focusing on the comparison, inTrees obtained higher scores in all the analyzed datasets. In particular, for EEG Eye State and Diabetic Retinopathy Debreceen, it shows much higher classification performances. It worth to noting, Node Harvest reports higher scores of sensitivity for all datasets. Maybe, it depends on the fact that this classifier can better recognize positive observations.

3. Conclusion

InTrees represents an excellent strategy for obtaining interpretative learners from Random Forest models.
The results deriving from this methodology are just as good, considering that the simplified rules based on the STEL classifier can be implemented in any programming language.

This work is a starting point for understanding the potential of Interpretable Machine Learning, which requires the development of innovative approaches that can meet the interpretative needs of each application context, such as the healthcare framework. A more complete comparative analysis should focus on analyzing data characterized by unbalanced responses and the presence of missing data (D'Ambrosio et al., 2012), and multiclass responses.

Table 4: Summary tables on the performance metrics performed on the four health datasets.

(a) *Diabetic Retinopathy Debrecen*

	RF	*NH*	*STEL*
Accuracy	0.64	0.64	**0.70**
Balanced Accuracy	0.64	0.65	**0.71**
Kappa	0.29	0.29	**0.41**
Specifity	0.65	0.48	**0.68**
Sensitivity	0.64	**0.82**	0.74
Precision	0.63	0.59	**0.65**
G-mean	0.64	0.63	**0.71**
F1	0.64	0.68	**0.69**
Youden's Index	0.29	0.30	**0.42**

(b) *EEG Eye State*

	RF	*NH*	*STEL*
Accuracy	0.92	0.68	**0.69**
Balanced Accuracy	0.92	0.67	**0.69**
Kappa	0.84	0.34	**0.38**
Specifity	0.89	0.48	**0.65**
Sensitivity	0.94	**0.85**	0.73
Precision	0.91	0.66	**0.72**
G-mean	0.92	0.64	**0.69**
F1	0.93	**0.75**	0.73
Youden's Index	0.83	0.33	**0.38**

(c) *Cardiovascular Disease*

	RF	*NH*	*STEL*
Accuracy	0.73	0.69	**0.71**
Balanced Accuracy	0.73	0.69	**0.71**
Kappa	0.47	0.38	**0.43**
Specifity	0.69	0.55	**0.66**
Sensitivity	0.78	**0.84**	0.77
Precision	0.71	0.64	**0.70**
G-mean	0.73	0.68	**0.71**
F1	0.74	0.72	**0.73**
Youden's Index	0.47	0.38	**0.43**

(d) *Pima Indians Diabetes*

	RF	*NH*	*STEL*
Accuracy	0.71	**0.74**	0.72
Balanced Accuracy	0.67	0.67	**0.69**
Kappa	0.35	**0.39**	0.38
Specifity	0.55	0.42	**0.57**
Sensitivity	0.80	**0.93**	0.80
Precision	0.78	0.74	**0.78**
G-mean	0.66	0.62	**0.68**
F1	0.79	**0.82**	0.79
Youden's Index	0.35	0.35	**0.38**

References

Adadi, A. and Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, **6**.

Ahmad, M. A., Eckert, C., and Teredesai, A. (2018). Interpretable machine learning in healthcare. In *Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics*, pp. 559–560.

Akosa, J. (2017). Predictive accuracy: A misleading performance measure for highly imbalanced data. In *Proceedings of the SAS Global Forum*, pp. 2–5.

Aria, M., Cuccurullo, C., and Gnasso, A. (2021). A comparison among interpretative proposals for random forests. *Machine Learning with Applications*.

Aria, M., D'Ambrosio, A., Iorio, C., Siciliano, R., and Cozza, V. (2020). Dynamic recursive tree-based partitioning for malignant melanoma identification in skin lesion dermoscopic images. *Statistical papers*, **61**(4).

Breiman, L. (1996). Bagging predictors. *Machine learning*, **24**(2):pp. 123–140.

Breiman, L. (2001). Random forests. *Machine learning*, **45**(1):pp. 5–32.

Breiman, L. et al. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, **16**(3):pp. 199–231.

D'Ambrosio, A., Aria, M., and Siciliano, R. (2012). Accurate tree-based missing data imputation and data fusion within the statistical learning paradigm. *Journal of classification*, **29**(2):pp. 227–258.

Deng, H. (2019). Interpreting tree ensembles with intrees. *International Journal of Data Science and Analytics*, **7**(4):pp. 277–287.

Dhillon, A. and Singh, A. (2019). Machine learning in healthcare data analysis: a survey. *Journal of Biology and Today's World*, **8**(6):pp. 1–10.

Díaz-Uriarte, R. and De Andres, S. A. (2006). Gene selection and classification of microarray data using random forest. *BMC bioinformatics*, **7**(1):pp. 3.

Domingos, P. (1998). Occam's two razors: the sharp and the blunt. In *KDD*, pp. 37–43.

Domingos, P. (1999). The role of occam's razor in knowledge discovery. *Data mining and knowledge discovery*, **3**(4):pp. 409–425.

Du, M., Liu, N., and Hu, X. (2019). Techniques for interpretable machine learning. *Communications of the ACM*, **63**(1):pp. 68–77.

García, V., Mollineda, R. A., and Sánchez, J. S. (2009). Index of balanced accuracy: A performance measure for skewed class distributions. In *Iberian conference on pattern recognition and image analysis*, pp. 441–448. Springer.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, **51**(5):pp. 1–42.

Haddouchi, M. and Berrado, A. (2019). A survey of methods and tools used for interpreting random forest. In *2019 1st International Conference on Smart Systems and Data Science (ICSSD)*, pp. 1–6. IEEE.

Meinshausen, N. (2010). Node harvest. *The Annals of Applied Statistics*, pp. 2049–2072.

Miotto, R., Wang, F., Wang, S., Jiang, X., and Dudley, J. T. (2018). Deep learning for healthcare: review, opportunities and challenges. *Briefings in bioinformatics*, **19**(6).

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144.

Sokolova, M., Japkowicz, N., and Szpakowicz, S. (2006). Beyond accuracy, f-score and roc: a family of discriminant measures for performance evaluation. In *Australasian joint conference on artificial intelligence*, pp. 1015–1021. Springer.