# Introduction to Interpretability in Machine Learning

**Adrien Bibal** ADRIEN.BIBAL@UNAMUR.BE
**Benoît Frénay** BENOIT.FRENAY@UNAMUR.BE
PReCISE, Faculty of Computer Science, University of Namur, rue Grandgagnage 21, 5000 Namur, Belgium

**Keywords**: interpretability, comprehensibility, measures, machine learning models

Interpretability is considered as important in the literature. Yet, measuring interpretability seems challenging due to its subjective nature. This abstract presents the survey of the literature by Bibal and Frénay (2016).

Several terms are used in the literature alongside interpretability. Usability is one of those terms. A model is not usable if it is rejected despite an acceptable accuracy. In the medical domain, Freitas (2014) noted that a simple, easy to read, decision tree can be refused because medical doctors may consider that a simple model cannot represent complex medical situations. Usability depends on the interpretability of the model to be measured. The same situation can be observed with justifiability (Martens et al., 2011) which bridges the gap between the description of the data made by the model and the knowledge of the application domain: "does the model justify (or correspond to) the existing knowledge of the domain?".

Interpretability is more fundamental and corresponds to the ability of a human to comprehend (Giraud-Carrier, 1998) or to understand (Rüping, 2006) the model. Two ways to handle the measure of interpretability are proposed in the literature.

On the one hand, heuristics correspond to approximations of the human understanding made by the machine learning researcher (Rüping, 2006), e.g. the model complexity. This approach is easy to formalise but can hardly compare models of different types. For instance, one cannot compare the number of nodes of a decision tree with those of a neural network.

On the other hand, users can be considered in the evaluation of human comprehensibility. Some authors work with user-based surveys to assess the interpretability of models (Allahyari & Lavesson, 2011). In this methodology, measuring the interpretability consists in asking questions such as "do you find this model understandable?" or "is this model more understandable than that one?". This goes beyond the limits of the heuristics approach as users can compare models of distinct types. However, the user-based approach is limited by the difficulty to quantify interpretability from the answers of surveys and the existence of different model representations.

Bibal and Frénay (2016) highlights gaps in the literature. First, there are almost no links between the two approaches in the literature. The heuristics approach does not use the user-based approach for validation and the user-based approach does not try to extract heuristics that could be used to quantify interpretability. This is mostly due to the lack of research in the direction of the user-based surveys approach. Second, models considered as black boxes are neglected. The measure of interpretability mostly deals with white-boxes (e.g. decision trees and rule lists). However, one could argue that a very simple SVM may be more interpretable than a very complex decision tree. The measure of interpretability needs more investigation and a grey-scale measure taking advantage of the two approaches presented here could be developed.

## References

Allahyari, H., & Lavesson, N. (2011). User-oriented assessment of classification model understandability. *Proc. SCAI* (pp. 11–19).

Bibal, A., & Frénay, B. (2016). Interpretability of machine learning models and representations: an introduction. *Proc. ESANN 2016* (pp. 77–82).

Freitas, A. A. (2014). Comprehensible classification models: a position paper. *ACM SIGKDD Explorations Newsletter*, *15*, 1–10.

Giraud-Carrier, C. (1998). Beyond predictive accuracy: what? *Proc. ECML* (pp. 78–85).

Martens, D., Vanthienen, J., Verbeke, W., & Baesens, B. (2011). Performance of classification models from a user perspective. *Dec. Support Syst.*, *51*, 782–793.

Rüping, S. (2006). *Learning interpretable models*. Doctoral dissertation, Universität Dortmund.