A Fairness-Oriented Visual Extension for the PSyKE Platform

Federico Sabbatini[1,*]

[1]National Institute for Nuclear Physics – Section in Florence, Sesto Fiorentino, Italy

Abstract

This paper introduces a visual extension for the PSyKE platform, with the goal of supporting fairness assessment and bias detection for predictive models based on symbolic knowledge bases. The extension focuses on the analysis of knowledge bases expressed in Prolog, the default representation adopted within PSyKE, and it is effective also to study the knowledge distilled from opaque predictive models through symbolic knowledge-extraction techniques. The tool generates a heatmap linking knowledge items and user-defined protected/sensitive groups with the proportion of individuals belonging to the groups and covered by each rule. This visual representation allows for immediate identification of fairness and bias issues.

Keywords

Algorithmic fairness, Explainable artificial intelligence, Symbolic knowledge, PSyKE

1. Introduction

In recent years, the pursuit of fairness in artificial intelligence (AI) and machine learning (ML) has gained significant momentum, especially in high-stakes domains such as finance, healthcare and criminal justice [1, 2, 3], where biased models can exacerbate social inequalities [4, 5, 6]. Since predictive tools behaving as black boxes, e.g., deep neural networks and ensemble models, are becoming increasingly pervasive, the demand for interpretable and accountable AI systems is growing in parallel [7, 8, 9, 10, 11].

Several complementary approaches have been proposed in the literature to improve transparency and/or detect and mitigate bias in ML models. These include the adoption of inherently interpretable models [12, 13] and post-hoc explanation techniques or other distillation methods that approximate complex models with simpler, more understandable ones [14]. Fairness auditing techniques, on the other hand, assess the outcomes of a model with respect to protected groups and can be based on group metrics (e.g., demographic parity, equal opportunity; [15]) or performed at an individual level (e.g., through counterfactual analysis; [16, 17]). While many of these methods focus on numeric or graphical summaries, symbolic representations such as logic rules offer a compact and human-readable way to reason about decision logic. This makes them particularly interesting not only for interpretation, but also for formal fairness inspection and compliance with legal frameworks.

One promising approach to improve transparency is symbolic knowledge extraction (SKE), which provides interpretable surrogate models mimicking the behaviour of opaque ones [18, 19, 20, 21]. Surrogate models are usually based on human-intelligible symbolic representations of the acquired knowledge, e.g., lists of logic rules or shallow decision trees. Symbolic knowledge bases enable human users to audit decisions, verify compliance with ethical or legal standards and detect the presence of potential bias and unfairness in the decision-making process [22].

To this purpose, the PSyKE Python platform [23, 24] was proposed to support SKE from arbitrarily complex black boxes in the form of Prolog theories. PSyKE also offers an extension for Semantic Web interoperability [25] and several tools for trustworthy AI [26] and explainable clustering [27]. Unfortunately, dedicated support for fairness assessment and bias detection and mitigation is currently missing in the platform. More in detail, it is possible to detect evidences of unfairness by inspecting the

AEQUITAS 2025: Workshop on Fairness and Bias in AI | co-located with ECAI 2025, Bologna, Italy

*Corresponding author.

f.sabbatini1@campus.uniurb.it (F. Sabbatini)

0000-0002-0532-6777 (F. Sabbatini)

Prolog theories produced by PSyKE, e.g., by checking if protected or sensitive features are present in the preconditions of the symbolic rules. However, despite being an essential requirement for ethical AI systems [28], in PSyKE there are no more sophisticated mechanisms to measure if individual rules disproportionately affect some protected groups, for instance identified by sex, gender or ethnicity.

To address this limitation of PSyKE, in this work a visual extension aimed at supporting bias detection in symbolic knowledge bases is presented. The tool basically processes a knowledge base in Prolog format to perform a group-wise assessment of the rules' coverage, with protected and unprotected groups identified by the user. Each rule is analysed to determine the proportion of individuals from each group that satisfy its conditions. The results are visualised in a heatmap where rows represent rules, columns correspond to groups and cell values indicate the relative impact of a rule on a group. Starting from the assumptions that if rules disproportionately affect certain groups, the knowledge base is likely biased and that, conversely, if the impact is uniform across groups, the knowledge can be considered fair, this visualisation allows practitioners to easily identify biased rules and unfair knowledge bases according to the presence of a strong disparity in the groups' coverage.

This contribution enriches the interpretability capabilities of the PSyKE framework with a fairness-aware analysis, enabling users to better understand and mitigate hidden biases in opaque and symbolic explanations. Accordingly, in Section 2 related works on fairness assessment and bias detection are reported, in Section 3 the contribution of this paper is detailed, in Section 4 experiments on real-world data sets to assess the applicability of the proposed extension are shown and in Section 5 conclusions are drawn.

2. Related Works

2.1. Fairness and Bias in Machine Learning

Fairness in ML algorithms and systems has been studied from multiple perspectives, including group fairness, individual fairness, and causal fairness [15, 16, 29, 30]. Group fairness metrics assess whether a model treats different demographic groups (e.g., defined by race or gender) equally in terms of positive/negative predictions or predictive error rates. Common examples include demographic parity, equalised odds, and predictive parity. In contrast, individual fairness requires that similar individuals receive similar outcomes. This latter is often formalised through similarity metrics and/or counterfactuals.

Bias can affect ML systems at various stages, including data collection, labelling, training, or deployment. Numerous fairness auditing tools have been developed to detect and mitigate such biases, ranging from pre-processing approaches to balance data sets, to in-processing techniques that regularise model training, to post-hoc methods that analyse and modify outcomes [31, 32, 33, 34].

Despite this rich landscape, due to high task complexity and data set dimensionality, fairness assessments are often performed at the model level or using aggregate statistical measures. There remains a significant gap in the literature when it comes to tools that visually illustrate the extent to which the predictions of a symbolic model impact different demographic groups in a specific ML context.

2.2. Existing Fairness and Explainability Tools

A number of toolkits have been developed for fairness auditing in ML models. Amongst them, Aequitas [35], AI Fairness 360 [36], and Fairlearn [37] are comprehensive libraries that compute group fairness metrics and possibly support bias mitigation techniques. These frameworks, however, typically operate on opaque models and produce only aggregated statistics without associating them to symbolic rules.

Visual and interactive tools such as FairVis [38] and the What-If Tool [39] allow users to inspect fairness properties through data exploration and counterfactual reasoning. Yet, they lack support for symbolic or rule-based explanations.

Table 1
Comparison of fairness and explainability tools. Only the approach proposed in this work combines symbolic rule inspection with fairness analysis and heatmap-based visualisation.

Tool	Symbolic support	Fairness auditing	Visualisation	Group-level analysis
Aequitas [35]	✗	✓	✓ (metrics dashboard)	✓ (aggregated)
AI Fairness 360 [36]	✗	✓	✓ (not rule-level)	✓ (aggregated)
Fairlearn [37]	✗	✓	✓ (dashboard)	✓ (aggregated)
FairVis [38]	✗	✓	✓ (interactive)	✓ (intersectional)
What-If Tool [39]	✗	✓	✓	✓ (counterfactuals)
RuleMatrix [40]	✓	✗	✓	✗
SIRUS [41]	✓	✗	✗	✗
Kamiran et al. [42]	✗	✓	✗	✗
PSyKE	✓	✓	✓ (heatmap)	✓ (per rule)

Some efforts in interpretable ML have focused on rule visualisation. RuleMatrix [40], for instance, displays classifiers as rule tables to enhance evaluation and interpretability. SIRUS [41] provides stable rule sets derived from ensemble models, but no visualisation tools are provided. Furthermore, these tools do not address group fairness nor allow users to assess differential impacts across protected groups.

Another line of research focuses on modifying models to mitigate bias. Kamiran et al. [42] proposed a decision-theoretic framework for discrimination-aware classification. Their approach leverages the reject option in probabilistic classifiers and the disagreement region in ensemble methods to reduce unfair decisions. While the method effectively decreases discrimination without altering training data or the classification algorithm itself, it does not produce symbolic explanations or provide visual tools for assessing group-level impacts.

The properties of all mentioned tools are summarised in Table 1. To knowledge, no existing solution combines symbolic knowledge base analysis with visual group-level fairness inspection. The PSyKE extension proposed in this work fills this gap by providing a heatmap-based visualisation tool where each rule is analysed in terms of its impact on different privileged and unprivileged groups. This allows users to identify biased or fair rules with fine granularity, and is applicable across multiple pedagogical rule extractors.

2.3. Symbolic Knowledge Extraction and the PSyKE Framework

Knowledge extraction refers to the process of approximating an arbitrarily complex opaque predictive model with a symbolic representation, for instance list or trees of propositional rules or logic programs, these latter both human- and machine-interpretable [43]. A wide range of different post-hoc explainability procedures fall under this approach. All these techniques allow users to better understand, trust and potentially debug ML-based systems.

The PSyKE framework is a comprehensive platform supporting a broad range of tools for explainable AI. Currently, it includes the implementation of several pedagogical SKE algorithms, i.e., Rule-Extraction-as-Learning [44], Trepan [45], CART [46], Iter [47], GridEx [48], GridREx [49], HEX [50] and CReEPy [51], all of which learn symbolic surrogates from a trained black-box model. In addition to knowledge extraction, PSyKE provides modules for explainable clustering based on the EXACT and CREAM algorithms [52, 53], supervised discretisation of features, hyper-parameter tuning for knowledge extractors [54], and various metrics to assess the quality and similarity of symbolic knowledge bases [55, 56, 57, 58].

The modular and extensible design of PSyKE facilitates a wide range of explainable AI workflows. However, to date, the platform does not offer any tool explicitly designed to fairness inspection and bias detection, despite being an increasingly important requirement for ethical AI.

3. A Visual Extension for Unfairness and Bias Detection in PSyKE

3.1. Combining Fairness, Explainability and Symbolic AI

Recent work in the field of explainable AI has begun to explore the relationship between model interpretability and fairness. Relying upon an interpretable model does not guarantee that it is fair. A bias may be present in the training data and thus propagated through the final model. A bias may be developed during the learning phase of the predictive model or may become noticeable or even more explicit when the model behaviour is translated into a symbolic representation.

Several approaches and metrics exist in the literature to check fairness constraints in symbolic knowledge bases or models. Some of these focus on studying whether certain rules disproportionately cover specific protected or unprotected groups. However, these analyses are based on numeric assessment that may lack an immediate understanding and interpretation. Effective visual interfaces to facilitate the analysis of domain experts are commonly absent in fairness-specific frameworks.

This work addresses this limitation by proposing a visual inspection tool dedicated to group fairness assessments for symbolic knowledge bases, integrated into the PSyKE framework. By mapping rule coverage across protected and unprotected groups into a heatmap, it is possible to visually detect disparities and potential sources of bias and unfairness, thus supporting both interpretability and ethical analysis.

3.2. Extending PSyKE for a Fairness-Aware Analysis

Starting from version 0.8.14, PSyKE is empowered with the visual extension proposed in this work. Currently, the visual fairness analysis is only available for knowledge bases obtained through SKE, even though in the future it will be extended to support generic knowledge bases and possibly opaque predictive models.

Figure 1 shows the workflow to generate with PSyKE a heatmap expressing the group-level fairness extent of a symbolic knowledge base extracted from an opaque predictor. More in detail, after having trained an opaque ML model on a given data set, one of the SKE algorithms included in PSyKE can be applied to distil a symbolic knowledge base mimicking the behaviour of the underlying model. The symbolic knowledge is composed of an ordered list of Prolog clauses.

To build the heatmap, users have to define a set of relevant groups to be analysed. Indeed, the identification of privileged or unprivileged groups, as well as the detection of protected and sensitive features, is not performed automatically by PSyKE's routines. The heatmap is generated according to the following rationale:

- Knowledge items, corresponding to logic rules, are evaluated in order and represented as heatmap rows, from the top to the bottom;
- User-defined groups are represented as heatmap columns, following the same order given by the user;
- The heatmap cell corresponding to row r and column c shows the percentage of individuals of group c affected by rule r with respect to the cardinality of group c.

Formally, for each rule $k_r \in \mathcal{K}$ composing the knowledge base $\mathcal{K}$, the subgroup of instances $d^r \in \mathcal{D}$ of data set $\mathcal{D}$ covered by k_r (i.e., those satisfying the preconditions of k_r) is calculated. Then, for each group $d_c \in \mathcal{D}$, the percentage of individuals also appearing in d^r is calculated and placed in the cell $h_{r,c}$ of heatmap $\mathcal{H}$:

$$h_{r,c} = \frac{\mid d^r \cap d_c \mid}{\mid d_c \mid} \cdot 100. \tag{1}$$

Clearly, all individuals are represented in $\mathcal{H}$, i.e.:

$$\sum_{r \in \mathcal{H}} h_{r,c} = 100\% \qquad \forall c \in \mathcal{H}. \tag{2}$$

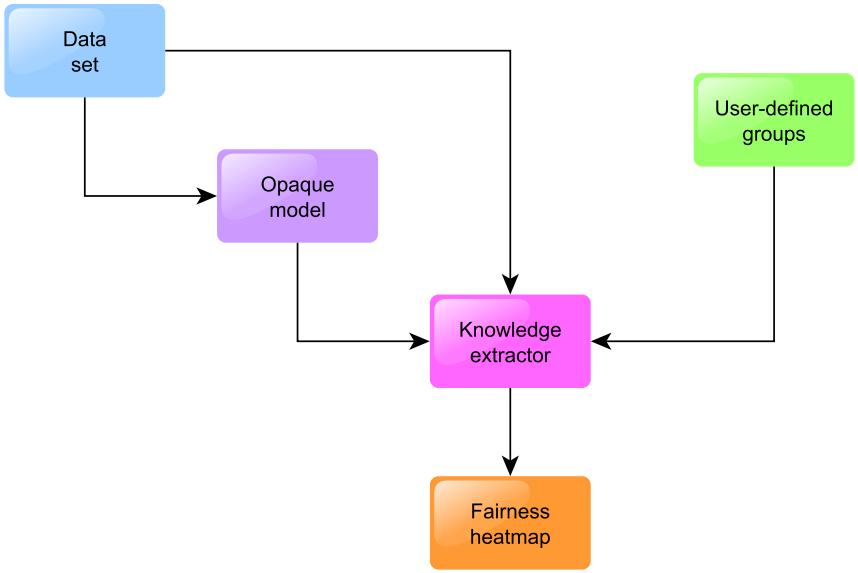


Figure 1: Abstract workflow to obtain the heatmap expressing the group-level fairness extent of a symbolic knowledge base extracted from an opaque predictor.

The knowledge base $\mathcal{K}$ is expected to be fair if all heatmap cells on the same rows contain similar values, meaning that all groups are covered evenly by the corresponding rules:

$$\forall r \in \mathcal{H} \;\; \exists p_r \text{ s.t. } h_{r,c} \simeq p_r \;\; \forall c \in \mathcal{H}. \tag{3}$$

The heatmap can be customised by users, however some automated elaborations are performed by the visualisation routine to enhance the resulting readability. For instance, two kinds of colourbars are supported, one for binary classification and one for the multi-class case. For this latter, values range between 0% and 100%, with the intensity and/or the colour visually identifying the corresponding percentage. A more sophisticated output is available for binary classification tasks, as depicted in Figure 2. In this case a diverging colormap is adopted and associated with a colourbar having symmetric values. Two different colours are associated with the possible outcomes, which intensity expresses a percentage value. Rules covering no data set individuals correspond to white background in the heatmap cells. Differentiating the colours for the possible outcomes enables a quick interpretation of the knowledge fairness. Indeed, it is straightforward to check if negative or positive outcomes are generally given to some groups. Additionally, it is possible to identify rules that can be pruned due to a low data coverage across groups.

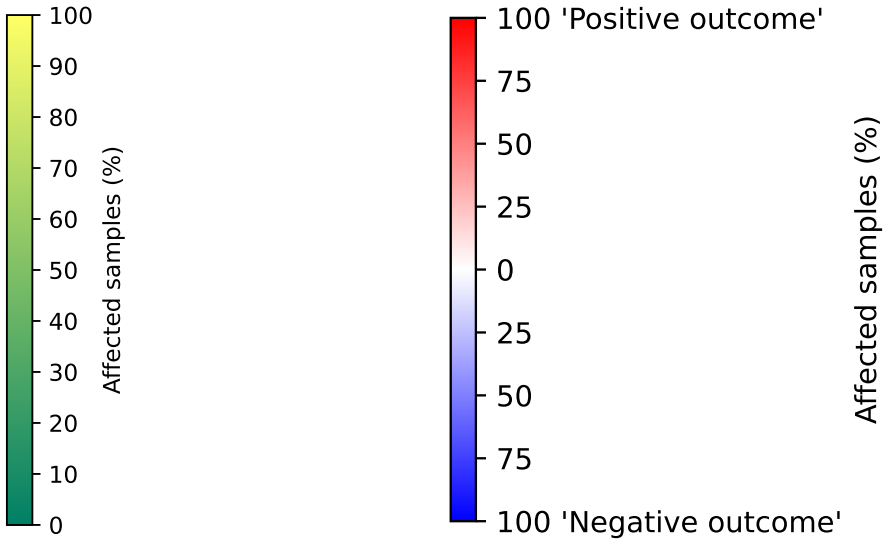
It is pointed out here that knowledge bases producing fair heatmaps may still present biases or unfairness. Conversely, when unfair heatmaps are generated, the knowledge is always biased.

3.3. Advantages, Limitations and Future Developments

The proposed visual extension for fairness offers several significant advantages in the context of fairness and bias detection in symbolic knowledge bases.

Intuitive visualisation. By representing the impact of each rule across different groups in a heatmap format, the tool introduced in this paper enables users to quickly identify disparities and potential biases. This visual approach simplifies the complex analysis of symbolic models, making fairness assessment more accessible.

Group-level analysis. The heatmap explicitly associates rules with the proportion of individuals affected in each protected group, allowing fine-grained detection of differential impacts. This facilitates the identification not only of overt biases in model behaviour, but also of subtle ones.



(a) Multi-class classification task.

(b) Binary classification task.

Figure 2: Example of colourbar associated with a multi-class classification task (left). Values range between 0% and 100%. Example of colourbar associated with a binary classification task (right). Values are symmetric with respect to the bar centre, representing 0%. In this case the heatmap is also a quick tool to highlight if positive outcomes are most likely given to a privileged group and if some rules can be pruned due to low coverage (white background colour for the whole heatmap row).

Integration with symbolic models. Unlike many fairness tools that operate on black-box models or raw data, this extension works directly on symbolic knowledge bases generated through a rule extraction process, possibly within an interpretable ML workflow. This makes it highly suitable for explainable AI applications where symbolic representations are preferred and/or needed.

User-driven flexibility. Users can specify which groups should be analysed, tailoring the fairness assessment to the specific context and regulatory requirements of their application.

The proposed tool also has some limitations, that will be tackled in the future.

Dependency on the knowledge-extraction process. The effectiveness of the analysis is inherently tied to the quality and completeness of the extracted symbolic knowledge. If the rules poorly represent the underlying model or data, fairness assessments may be inaccurate or misleading.

Limitation to symbolic models. The PSyKE extension focuses on symbolic knowledge bases and does not directly support black-box models or raw data. Therefore, its applicability is constrained to contexts where symbolic extraction is feasible and meaningful.

Static analysis. The current implementation provides a snapshot of rule impacts but does not incorporate dynamic or causal analyses that could deepen understanding of bias origins or model behaviour over time.

Scalability. For knowledge bases composed of many rules and/or in presence of numerous user-defined groups, the heatmap visualisation may become complex to interpret or computationally expensive to generate without further optimisation or interactive filtering.

Future efforts will be also devoted to provide a dynamic, interactive tool where users can customise the final heatmap, for instance by filtering rules and groups or by combining multiple sensitive attributes.

4. Experiments

The PSyKE extension for fairness visual inspection has been evaluated on several case studies. Results of experiments are reported in the following.

4.1. The Loan Data Set Case Study

The first case study involves the Loan data set.[1] It includes 11 input variables that represent possibly important factors influencing the decision to approve or deny a loan. The outcome is a binary variable indicating the final loan decision. Additionally, the data set contains a unique identifier for each loan application. It consists of 614 records in total, but only 480 are complete and free of missing values. During the experiments reported here, records with missing data were removed, and categorical attributes were transformed into discrete numerical features, in compliance with the majority of SKE models implemented in PSyKE. The data set is shown in Figure 3a. Only the credit history and loan amount input features are reported in the plot, since they appear to be the most discriminative for the decision. Male applicants are represented by the ♂ symbol, whereas female ones are identified with ♀. Blue and red symbols are associated with a positive outcome, whereas green and violet symbols are used for negative outcomes. These different colours and symbols highlight that there is not an evident bias in granting or denying loans based on the applicant gender. Conversely, the decision appears strongly correlated with the credit history.

Three different techniques were used to objectively evaluated the fairness extent of the ML models trained upon these data. First, the disparate impact index was calculated [59]. Second, an ML model was trained and symbolic knowledge was distilled and analysed for fairness inspection. Finally, the PSyKE heatmap was generated.

It is recalled here that the disparate impact metric evaluates the level of equal or unequal treatment between two groups by comparing the proportion of individuals from each group who receive favourable outcomes. The disparate impact index thus provides a quantitative assessment of the differential treatment experienced by different demographic groups. The index may be calculated on the raw data set to check the fairness extent of the data used in the ML workflow, on the opaque predictions provided by the trained ML model and on the symbolic knowledge extracted from the black box. The disparate impact calculated on the raw data set[2] was equal to 0.990, corresponding to a fair treatment between males and females.

As for the ML workflow, the whole data set was split into training and test set with a 75%:25% ratio. The former was used to train a random forest classifier composed of 25 base trees with maximum depth equal to 4. The corresponding accuracy score measured on the test set was of 0.83. The disparate impact measured for the black-box predictions was equal to 0.996, meaning equal treatment for male and female applicants, aligned with the fairness extent of the underlying data set.

The CART extractor [46] was employed to extract symbolic knowledge from the random forest. A maximum depth of 2 was set for the extractor. Predictions drawn on the basis of the extracted knowledge exhibited a fidelity of 0.99 with respect to the random forest predictions and an accuracy of 0.82 with respect to the data set output feature. The corresponding disparate impact was of 0.984, close to that observed for the black box and the data. The decision boundaries identified with CART are shown in Figure 3b and the equivalent symbolic knowledge is reported in Listing 1. CART outputs basically confirm the evidence according to which the credit history feature plays a major role for the final decision, whereas the gender attribute is not considered.

Despite clarifying the absence of gender inspection in taking the loan decision, the Prolog rules obtained through CART give no information about the impact of individual rules on the male and female groups, for instance due to other input features related with gender but not identified as sensitive

[1] https://www.kaggle.com/datasets/burak3ergun/loan-data-set

[2] All fairness and performance numeric assessment were carried out on the test set. Results may thus slightly differ depending on the performed splitting into training and test sets

Listing 1: Rules extracted with CART for the Loan data set.

```
loan(LoanAmount, Gender, Married, ..., CreditHistory, PropertyArea, YES) :-
   CreditHistory > 0.5.
loan(LoanAmount, Gender, Married, ..., CreditHistory, PropertyArea, NO) :-
   LoanAmount <= 547.5.
loan(LoanAmount, Gender, Married, ..., CreditHistory, PropertyArea, YES).
```

(a) Data set samples.

(b) CART decision boundaries.

Figure 3: Loan data set instances projected according to the credit history and loan amount input features. Male applicants are represented by the ♂ symbol, whereas female applicants are identified with ♀. Blue and red symbols are associated with a positive outcome, whereas green and violet are used for negative outcomes. The different colours and symbols clearly show how there is not an evident bias in granting or denying loans based on the applicant gender. Conversely, the decision appears strongly correlated with the credit history. CART decision boundaries basically confirm this evidence. Instances are randomly scattered around discrete values of the x-axis in order to limit the visual hindrances of superposition.

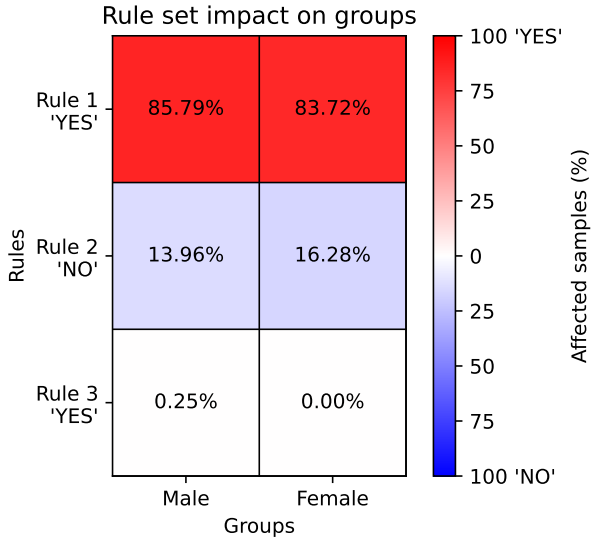


Figure 4: Heatmap generated with PSyKE for the Loan data set.

or protected. On the other hand, this information is clearly captured in the heatmap shown in Figure 4, where each rule is estimated to evenly affect both groups of individuals.

From the heatmap it is possible to notice that:

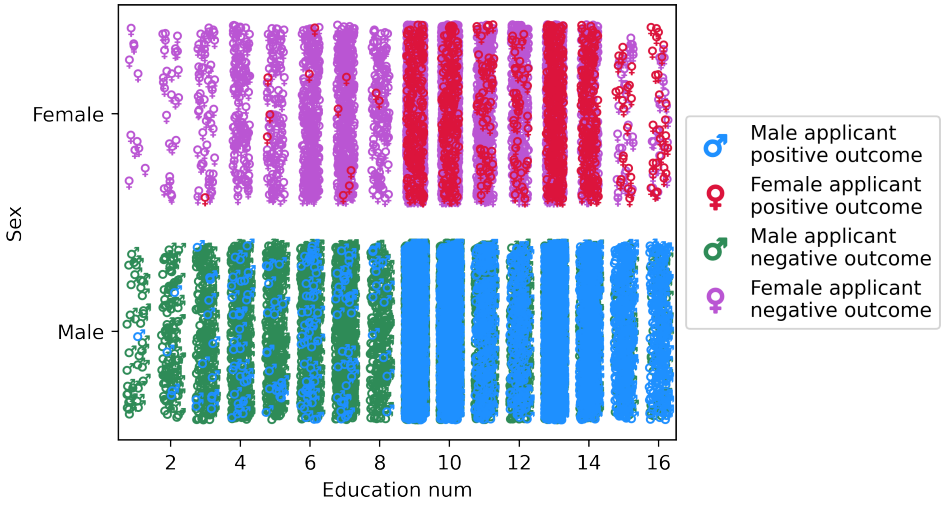
1. There are 3 rules and 2 groups of interest, according with the knowledge base extracted with CART and the data set under analysis;
2. The first rule, based on the credit history, is the one with largest coverage and affects in the same manner both males and females;
3. Analogously, also the second rule appears fair, with a similar impact on both groups;
4. The last rule has a negligible impact on female applicants as well as on male ones, since it covers a very limited subregion of the data set domain;
5. Female and male applicants exhibit a similar probability of being granted or denied a loan.

In conclusion, the data set appears to be fair as well as the random forest trained upon it and, in turn, the symbolic knowledge extracted with CART. This evidence may be obtained via plots, disparate impact assessments or the heatmap generated by PSyKE, without differences in the final results of the fairness analysis.

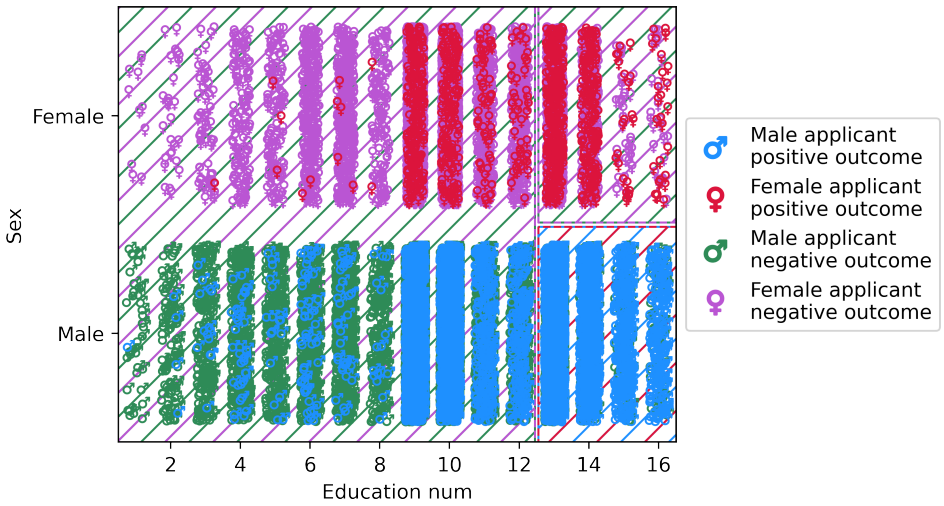
4.2. The Adult Data Set Case Study

The second case study presented in this work is based on the Adult data set.[3] It includes 14 numeric and categoric input variables to be used to predict whether the income of an adult person is above or below an annual threshold of $50 000. The outcome is thus a binary variable. Data set instances affected by missing data were removed, and categorical attributes were converted into discrete ones, as for the previous case study. The data set is shown in Figure 5a. Displayed features are the individual education expressed as a number and the individual sex. The plot follows the same logic of the previous case study, with different colours and symbols to discern between male/female individuals and positive/negative outcomes. It is possible to notice that the decision is mainly based on the education values, however the behaviour for male and female individuals appear not the same, with a higher propensity to give positive outcomes to male adults.

[3] https://archive.ics.uci.edu/dataset/2/adult



(a) Data set samples.



(b) CART decision boundaries.

Figure 5: Subset of the Adult data set instances (50%) projected according to the education (numeric) and sex input features. Symbols and colours follows the same logic as in Figure 3. CART decision boundaries clearly appear unfair, since they are based on the sex attribute.

As before, three methods were employed to evaluate fairness. The disparate impact calculated on the raw data was equal to 0.358, highlighting a severe fairness issue related to a diverse treatment between male and female individuals.

A K-nearest neighbours classifier parametrised with K=150 was trained on the data set (training:test set ratio = 75%:25%). The corresponding accuracy score measured on the test set was of 0.83, whereas the disparate impact observed for the black-box predictions was very low and equal to 0.066. This value implies that the unequal treatment present in the training data was magnified in the ML model.

Symbolic knowledge was extracted with the CART algorithm parametrised with a maximum depth of 4. The resulting predictions shown a fidelity of 0.91 with respect to the ML predictions and an accuracy of 0.79 with respect to the original data. The corresponding disparate impact was of 0.000. Indeed, according to the knowledge extracted with CART, it is not possible for female individuals to obtain a positive prediction (see the decision boundaries identified with CART in Figure 5b and the equivalent symbolic knowledge included in Listing 2). CART outputs magnify the bias contained in the data and learnt with the opaque model, according to which male individuals are favoured over female ones.

The unfair scenario emerging from the aforementioned considerations can be easily detected through

Listing 2: Rules extracted with CART for the Adult data set.

```
income(Age, 'Education-num', ..., Race, Sex, 'Capital-gain', '<=50K') :-
   'Education-num' =< 12.5.
income(Age, 'Education-num', ..., Race, Sex, 'Capital-gain', '>50K') :-
   Sex = 'M'.
income(Age, 'Education-num', ..., Race, Sex, 'Capital-gain', '<=50K').
```

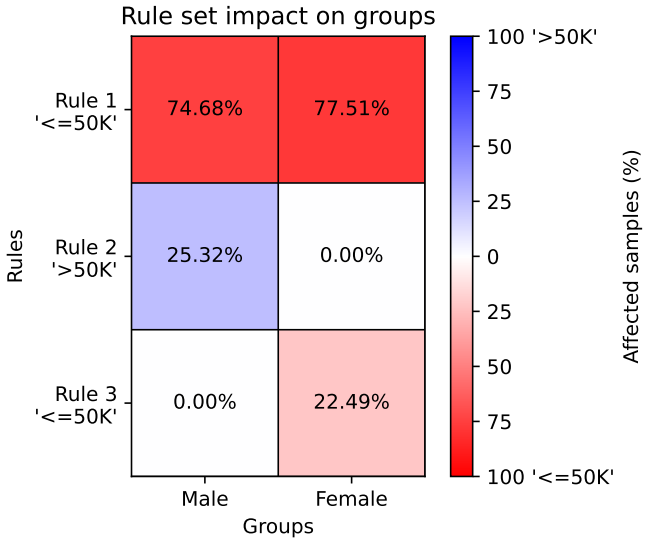


Figure 6: Heatmap generated with PSyKE for the Adult data set.

the heatmap generated with PSyKE and shown in Figure 6, where two rules out of three only affect single demographic groups. From the heatmap it is possible to notice that:

1. There are 3 rules and 2 groups of interest, as expected;
2. The first rule, based on the individual education, has a high coverage and affect in the same manner both the male and female population;
3. Conversely, the second rule only affects male individuals, by assigning a positive outcome;
4. Analogously, the third rule only affects female individuals, but by assigning a negative prediction;
5. From the analysis of the second column, it is clear that for females it is not possible to obtain a positive outcome, regardless of the values of other input features.

In conclusion, the data set of this case study is strongly biased, exhibiting an evident unequal treatment between female and male groups, and this evidence may be highlighted via traditional methods as well as thanks to the heatmap generated by PSyKE.

5. Conclusions

In this work, a visual extension of the PSyKE platform designed to support the fairness-aware inspection of symbolic knowledge bases is introduced. Through the analysis of the impact of individual symbolic rules at a group level and the visualisation of their coverage through a heatmap, the proposed approach enables users to detect potential biases and disparities across protected demographic groups. This enhancement contributes to bridging the gap between interpretability and fairness in post-hoc explainable AI models, particularly in settings where symbolic rule representations are used to approximate complex predictive models.

The presented tool allows practitioners to go beyond global fairness metrics, towards the inspection of how each extracted rule affects different subpopulations. This capability is essential for ethical auditing, regulatory compliance and the development of responsible AI systems.

Future work will focus on extending this framework to support additional fairness criteria and, more concretely, the integration of bias mitigation techniques in the SKE workflow, in order to obtain not only an explainable model out of an opaque one, but also a fair predictor from a biased one. Further investigations may involve the exploitation of PSyKE to obtain counterfactuals that can be used for fairness assessments and the study of causality to empower the PSyKE framework.

Declaration on Generative AI

During the preparation of this work, the author used GPT-4 in order to: Grammar and spelling. After using this tool, the author reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] K. Morik, Medicine: Applications of machine learning, in: C. Sammut, G. I. Webb (Eds.), Encyclopedia of Machine Learning and Data Mining, Springer, 2017, pp. 809–817. URL: https://doi.org/10.1007/978-1-4899-7687-1_530. doi:10.1007/978-1-4899-7687-1_530.

[2] J. Angwin, J. Larson, S. Mattu, L. Kirchner, Machine Bias (2016). URL: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[3] F. Martínez-Plumed, C. Ferri, D. Nieves, J. Hernández-Orallo, Fairness and missing values, CoRR abs/1905.12728 (2019). URL: http://arxiv.org/abs/1905.12728. arXiv:1905.12728.

[4] V. Eubanks, Automating inequality: How high-tech tools profile, police, and punish the poor, St. Martin's Press, 2018.

[5] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A survey on bias and fairness in machine learning, ACM Computing Surveys 54 (2022) 115:1–115:35. URL: https://doi.org/10.1145/3457607. doi:10.1145/3457607.

[6] C. O'neil, Weapons of math destruction: How big data increases inequality and threatens democracy, Crown, 2017.

[7] F. Doshi-Velez, B. Kim, Towards a rigorous science of interpretable machine learning, arXiv preprint arXiv:1702.08608 (2017).

[8] Z. C. Lipton, The mythos of model interpretability, Queue 16 (2018) 31–57. doi:10.1145/3236386.3241340.

[9] European Commission, Directorate-General for Communications Networks, Content and Technology, Ethics guidelines for trustworthy AI, Publications Office, 2019. doi:10.2759/346720.

[10] European Commission, AI Act – Proposal for a regulation of the european parliament and the council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative acts, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206, 2021.

[11] D. Pedreschi, S. Ruggieri, F. Turini, Discrimination-aware data mining, in: Y. Li, B. Liu, S. Sarawagi (Eds.), Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, Nevada, USA, August 24–27, 2008, ACM, 2008, pp. 560–568. URL: https://doi.org/10.1145/1401890.1401959. doi:10.1145/1401890.1401959.

[12] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nature Machine Intelligence 1 (2019) 206–215. doi:10.1038/s42256-019-0048-x.

[13] C. Rudin, C. Chen, Z. Chen, H. Huang, L. Semenova, C. Zhong, Interpretable machine learning: Fundamental principles and 10 grand challenges, CoRR abs/2103.11251 (2021). URL: https://arxiv.org/abs/2103.11251. arXiv:2103.11251.

[14] E. M. Kenny, C. Ford, M. Quinn, M. T. Keane, Explaining black-box classifiers using post-hoc explanations-by-example: The effect of explanations and error-rates in XAI user studies, Artificial Intelligence 294 (2021) 103459. doi:10.1016/j.artint.2021.103459.

[15] J. Baumann, C. Heitz, Group fairness in prediction-based decision making: From moral assessment to implementation, in: 9th Swiss Conference on Data Science, SDS 2022, Lucerne, Switzerland, June 22–23, 2022, IEEE, 2022, pp. 19–25. URL: https://doi.org/10.1109/SDS54800.2022.00011. doi:10.1109/SDS54800.2022.00011.

[16] P. George John, D. Vijaykeerthy, D. Saha, Verifying individual fairness in machine learning models, in: R. P. Adams, V. Gogate (Eds.), Proceedings of the Thirty-Sixth Conference on Uncertainty in Artificial Intelligence, UAI 2020, virtual online, August 3–6, 2020, volume 124 of *Proceedings of Machine Learning Research*, AUAI Press, 2020, pp. 749–758. URL: http://proceedings.mlr.press/v124/george-john20a.html.

[17] M. J. Kusner, J. Loftus, C. Russell, R. Silva, Counterfactual fairness, in: I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 30, Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf.

[18] R. Andrews, J. Diederich, A. B. Tickle, Survey and critique of techniques for extracting rules from trained artificial neural networks, Knowledge-Based Systems 8 (1995) 373–389. doi:10.1016/0950-7051(96)81920-4.

[19] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, ACM Computing Surveys 51 (2018) 1–42. doi:10.1145/3236009.

[20] G. Ciatto, F. Sabbatini, A. Agiollo, M. Magnini, A. Omicini, Symbolic knowledge extraction and injection with sub-symbolic predictors: A systematic literature review, ACM Computing Surveys 56 (2024) 161:1–161:35. doi:10.1145/3645103.

[21] F. Sabbatini, Four decades of symbolic knowledge extraction from sub-symbolic predictors. A survey, ACM Computing Surveys (2025). doi:10.1145/3749097.

[22] F. Sabbatini, R. Calegari, Unmasking the shadows: Leveraging symbolic knowledge extraction to discover biases and unfairness in opaque predictive models, in: R. Calegari, V. Dignum, B. O'Sullivan (Eds.), Proceedings of the 2nd Workshop on Fairness and Bias in AI co-located with 27th European Conference on Artificial Intelligence (ECAI 2024), Santiago de Compostela, Spain, October 20th, 2024, volume 3808 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024. URL: https://ceur-ws.org/Vol-3808/paper13.pdf.

[23] F. Sabbatini, G. Ciatto, R. Calegari, A. Omicini, On the design of PSyKE: A platform for symbolic knowledge extraction, in: R. Calegari, G. Ciatto, E. Denti, A. Omicini, G. Sartor (Eds.), WOA 2021 – 22nd Workshop "From Objects to Agents", volume 2963 of *CEUR Workshop Proceedings*, Sun SITE Central Europe, RWTH Aachen University, 2021, pp. 29–48. URL: https://ceur-ws.org/Vol-2963/paper14.pdf, 22nd Workshop "From Objects to Agents" (WOA 2021), Bologna, Italy, September 1–3, 2021. Proceedings.

[24] F. Sabbatini, G. Ciatto, R. Calegari, A. Omicini, Symbolic knowledge extraction from opaque ML predictors in PSyKE: Platform design & experiments, Intelligenza Artificiale 16 (2022) 27–48. doi:10.3233/IA-210120.

[25] F. Sabbatini, G. Ciatto, A. Omicini, Semantic Web-Based interoperability for intelligent agents with PSyKE, in: D. Calvaresi, A. Najjar, M. Winikoff, K. Främling (Eds.), Explainable and Transparent AI and Multi-Agent Systems, volume 13283 of *Lecture Notes in Computer Science*, Springer, 2022, pp. 124–142. doi:10.1007/978-3-031-15565-9_8.

[26] R. Calegari, F. Sabbatini, The PSyKE technology for trustworthy artificial intelligence, in: A. Dovier, A. Montanari, A. Orlandini (Eds.), AIxIA 2022, volume 13796 of *Lecture Notes in Computer Science*, Springer, Cham, Switzerland, 2023, pp. 3–16. doi:10.1007/978-3-031-27181-6_1, XXI International Conference of the Italian Association for Artificial Intelligence, AIxIA 2022, Udine, Italy, November 28 – December 2, 2022, Proceedings.

[27] F. Sabbatini, R. Calegari, Unlocking insights and trust: The value of explainable clustering algorithms for cognitive agents, in: R. Falcone, C. Castelfranchi, A. Sapienza, F. Cantucci (Eds.),

Proceedings of the 24th Workshop "From Objects to Agents", Roma, Italy, November 6–8, 2023, volume 3579 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 232–245. URL: https://ceur-ws.org/Vol-3579/paper18.pdf.

[28] S. Verma, J. Rubin, Fairness definitions explained, in: Y. Brun, B. Johnson, A. Meliou (Eds.), Proceedings of the International Workshop on Software Fairness, FairWare@ICSE 2018, Gothenburg, Sweden, May 29, 2018, ACM, 2018, pp. 1–7. URL: https://doi.org/10.1145/3194770.3194776. doi:10.1145/3194770.3194776.

[29] D. Madras, E. Creager, T. Pitassi, R. S. Zemel, Fairness through causal awareness: Learning causal latent-variable models for biased data, in: D. Boyd, J. H. Morgenstern (Eds.), Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29–31, 2019, ACM, 2019, pp. 349–358. URL: https://doi.org/10.1145/3287560.3287564. doi:10.1145/3287560.3287564.

[30] J. R. Loftus, C. Russell, M. J. Kusner, R. Silva, Causal reasoning for algorithmic fairness, CoRR abs/1805.05859 (2018). URL: http://arxiv.org/abs/1805.05859. arXiv:1805.05859.

[31] J. J. Ward, X. Zeng, G. Cheng, FairRR: Pre-processing for group fairness through randomized response, in: S. Dasgupta, S. Mandt, Y. Li (Eds.), International Conference on Artificial Intelligence and Statistics, 2–4 May 2024, Palau de Congressos, Valencia, Spain, volume 238 of *Proceedings of Machine Learning Research*, PMLR, 2024, pp. 3826–3834. URL: https://proceedings.mlr.press/v238/john-ward24a.html.

[32] M. K. Duong, S. Conrad, Towards fairness and privacy: A novel data pre-processing optimization framework for non-binary protected attributes, in: D. Benavides-Prado, S. M. Erfani, F. Philippe, Y. L. Boo, Y. S. Koh (Eds.), Data Science and Machine Learning – 21st Australasian Conference, AusDM 2023, Auckland, New Zealand, December 11–13, 2023, Proceedings, volume 1943 of *Communications in Computer and Information Science*, Springer, 2023, pp. 105–120. URL: https://doi.org/10.1007/978-981-99-8696-5_8. doi:10.1007/978-981-99-8696-5_8.

[33] M. Wan, D. Zha, N. Liu, N. Zou, In-processing modeling techniques for machine learning fairness: A survey, ACM Transactions on Knowledge Discovery from Data 17 (2023) 35:1–35:27. URL: https://doi.org/10.1145/3551390. doi:10.1145/3551390.

[34] F. Petersen, D. Mukherjee, Y. Sun, M. Yurochkin, Post-processing for individual fairness, in: M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, J. Wortman Vaughan (Eds.), Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6–14, 2021, virtual, 2021, pp. 25944–25955. URL: https://proceedings.neurips.cc/paper/2021/hash/d9fea4ca7e4a74c318ec27c1deb0796c-Abstract.html.

[35] P. Saleiro, B. Kuester, A. Stevens, A. Anisfeld, L. Hinkson, J. London, R. Ghani, Aequitas: A bias and fairness audit toolkit, CoRR abs/1811.05577 (2018). URL: http://arxiv.org/abs/1811.05577. arXiv:1811.05577.

[36] R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic, S. Nagar, K. Natesan Ramamurthy, J. T. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, Y. Zhang, AI fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias, IBM Journal of Research and Development 63 (2019) 4:1–4:15. URL: https://doi.org/10.1147/JRD.2019.2942287. doi:10.1147/JRD.2019.2942287.

[37] H. J. P. Weerts, M. Dudík, R. Edgar, A. Jalali, R. Lutz, M. Madaio, Fairlearn: Assessing and improving fairness of AI systems, Journal of Machine Learning Research 24 (2023) 257:1–257:8. URL: https://jmlr.org/papers/v24/23-0389.html.

[38] Á. A. Cabrera, W. Epperson, F. Hohman, M. Kahng, J. Morgenstern, D. H. Chau, FAIRVIS: visual analytics for discovering intersectional bias in machine learning, in: R. Chang, D. A. Keim, R. Maciejewski (Eds.), 14th IEEE Conference on Visual Analytics Science and Technology, IEEE VAST 2019, Vancouver, BC, Canada, October 20–25, 2019, IEEE, 2019, pp. 46–56. URL: https://doi.org/10.1109/VAST47406.2019.8986948. doi:10.1109/VAST47406.2019.8986948.

[39] J. Wexler, M. Pushkarna, T. Bolukbasi, M. Wattenberg, F. B. Viégas, J. Wilson, The What-If tool: Interactive probing of machine learning models, IEEE Transactions on Visualization and Computer Graphics 26 (2020) 56–65. URL: https://doi.org/10.1109/TVCG.2019.2934619. doi:10.1109/TVCG.

2019.2934619.

[40] Y. Ming, H. Qu, E. Bertini, RuleMatrix: Visualizing and understanding classifiers with rules, IEEE Transactions on Visualization and Computer Graphics 25 (2019) 342–352. URL: https://doi.org/10.1109/TVCG.2018.2864812. doi:10.1109/TVCG.2018.2864812.

[41] C. Bénard, G. Biau, S. Da Veiga, E. Scornet, Sirus: Stable and interpretable rule set for classification, Electronic Journal of Statistics 15 (2021) 427–505. doi:10.1214/20-EJS1792.

[42] F. Kamiran, A. Karim, X. Zhang, Decision theory for discrimination-aware classification, in: 12th IEEE International Conference on Data Mining, ICDM 2012, Brussels, Belgium, December 10–13, 2012, IEEE Computer Society, 2012, pp. 924–929. URL: https://doi.org/10.1109/ICDM.2012.45. doi:10.1109/ICDM.2012.45.

[43] F. Sabbatini, Exploiting Explainable Artificial Intelligence for Space Weather Investigations on board LISA and Future Space Interferometers for Gravitational Wave Detection, Ph.D. thesis, University of Urbino, 2025.

[44] M. W. Craven, J. W. Shavlik, Using sampling and queries to extract rules from trained neural networks, in: Machine Learning Proceedings 1994, Elsevier, 1994, pp. 37–45. doi:10.1016/B978-1-55860-335-6.50013-1.

[45] M. W. Craven, J. W. Shavlik, Extracting tree-structured representations of trained networks, in: Advances in Neural Information Processing Systems 8. Proceedings of the 1995 Conference, The MIT Press, 1996, pp. 24–30. URL: http://papers.nips.cc/paper/1152-extracting-tree-structured-representations-of-trained-networks.pdf.

[46] L. Breiman, J. Friedman, C. J. Stone, R. A. Olshen, Classification and Regression Trees, CRC Press, 1984.

[47] J. Huysmans, B. Baesens, J. Vanthienen, ITER: An algorithm for predictive regression rule extraction, in: Data Warehousing and Knowledge Discovery (DaWaK 2006), Springer, 2006, pp. 270–279. doi:10.1007/11823728_26.

[48] F. Sabbatini, G. Ciatto, A. Omicini, GridEx: An algorithm for knowledge extraction from black-box regressors, in: Explainable and Transparent AI and Multi-Agent Systems. Third International Workshop, EXTRAAMAS 2021, Virtual Event, May 3–7, 2021, Revised Selected Papers, volume 12688 of *LNCS*, Springer Nature, Basel, Switzerland, 2021, pp. 18–38. doi:10.1007/978-3-030-82017-6_2.

[49] F. Sabbatini, R. Calegari, Symbolic knowledge extraction from opaque machine learning predictors: GridREx & PEDRO, in: Proceedings of the 19th International Conference on Principles of Knowledge Representation and Reasoning, KR 2022, Haifa, Israel. July 31 - August 5, 2022, 2022. URL: https://proceedings.kr.org/2022/57/. doi:10.24963/kr.2022/57.

[50] F. Sabbatini, R. Calegari, Hierarchical knowledge extraction from opaque machine learning predictors, in: AIxIA 2024 – Advances in Artificial Intelligence, volume 15450 of *Lecture Notes in Computer Science*, Springer, Cham, Switzerland, 2025, pp. 257–273. doi:10.1007/978-3-031-80607-0_20, XXIII International Conference of the Italian Association for Artificial Intelligence, AIxIA 2024, Bolzano, Italy, November 25–28, 2024, Proceedings.

[51] F. Sabbatini, R. Calegari, Unveiling opaque predictors via explainable clustering: The CReEPy algorithm, in: G. Boella, F. A. D'Asaro, A. Dyoub, L. Gorrieri, F. A. Lisi, C. Manganini, G. Primiero (Eds.), Proceedings of the 2nd Workshop on Bias, Ethical AI, Explainability and the role of Logic and Logic Programming co-located with the 22nd International Conference of the Italian Association for Artificial Intelligence (AIxIA 2023), Rome, Italy, November 6, 2023, volume 3615 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 1–14. URL: https://ceur-ws.org/Vol-3615/paper1.pdf.

[52] F. Sabbatini, R. Calegari, ExACT explainable clustering: Unravelling the intricacies of cluster formation, in: C. K. Baker, L. Gómez Álvarez, J. Heyninck, T. Meyer, R. Peñaloza, S. Vesic (Eds.), Joint Proceedings of the 2nd Workshop on Knowledge Diversity and the 2nd Workshop on Cognitive Aspects of Knowledge Representation co-located with 20th International Conference on Principles of Knowledge Representation and Reasoning (KR 2023), Rhodes, Greece, September 3–4, 2023, volume 3548 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023. URL: https://ceur-ws.org/Vol-3548/paper3.pdf.

[53] F. Sabbatini, R. Calegari, Explainable Clustering with CREAM, in: Proceedings of the 20th International Conference on Principles of Knowledge Representation and Reasoning, 2023, pp. 593–603. URL: https://doi.org/10.24963/kr.2023/58. doi:10.24963/kr.2023/58.

[54] F. Sabbatini, R. Calegari, Untying black boxes with clustering-Based symbolic knowledge extraction, Intelligenza Artificiale 18 (2024) 21–34. doi:10.3233/IA-240026.

[55] F. Sabbatini, R. Calegari, On the evaluation of the symbolic knowledge extracted from black boxes, AI and Ethics 4 (2024) 65–74. doi:https://doi.org/10.1007/s43681-023-00406-1.

[56] F. Sabbatini, R. Calegari, Symbolic knowledge-Extraction evaluation metrics: The FiRe score, in: K. Gal, A. Nowé, G. J. Nalepa, R. Fairstein, R. Radulescu (Eds.), ECAI 2023 - 26th European Conference on Artificial Intelligence, September 30 – October 4, 2023, Kraków, Poland – Including 12th Conference on Prestigious Applications of Intelligent Systems (PAIS 2023), volume 372 of *Frontiers in Artificial Intelligence and Applications*, IOS Press, 2023, pp. 2033–2040. doi:10.3233/FAIA230496.

[57] F. Sabbatini, R. Calegari, ICE: An evaluation metric to assess symbolic knowledge quality, in: AIxIA 2024 – Advances in Artificial Intelligence, volume 15450 of *Lecture Notes in Computer Science*, Springer, Cham, Switzerland, 2025, pp. 241–256. doi:10.1007/978-3-031-80607-0_19, XXIII International Conference of the Italian Association for Artificial Intelligence, AIxIA 2024, Bolzano, Italy, November 25–28, 2024, Proceedings.

[58] F. Sabbatini, C. Sirocchi, R. Calegari, Symbolic knowledge comparison: Metrics and methodologies for multi-agent systems, in: M. Alderighi, M. Baldoni, C. Baroglio, R. Micalizio, S. Tedeschi (Eds.), Proceedings of the 25th Workshop "From Objects to Agents", Bard (Aosta), Italy, July 8–10, 2024, volume 3735 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024, pp. 202–216. URL: https://ceur-ws.org/Vol-3735/paper_17.pdf.

[59] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, S. Venkatasubramanian, Certifying and removing disparate impact, in: L. Cao, C. Zhang, T. Joachims, G. I. Webb, D. D. Margineantu, G. Williams (Eds.), Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, August 10–13, 2015, ACM, 2015, pp. 259–268. URL: https://doi.org/10.1145/2783258.2783311. doi:10.1145/2783258.2783311.