Local explainability of fuzzy and classic models focusing on disaster management in Brazilian municipalities

Renata Ribeiro [1], Norma Valencio[2], Heloisa Camargo[1]

[1]Computing Department – Federal University of São Carlos (UFSCar)

[2]Department of Environmental Sciences – Federal University of São Carlos (UFSCar)

renataribeiro11787@gmail.com, {norma.valencio,heloisacamargo}@ufscar.br

Abstract. *Disasters are socio-environmental processes of losses and damages related to severe or extreme events. The lower the socio-spatial self-protection to deal with such events, the greater the chances of a disaster occurring. In Brazil, when the local damages and losses incurred exceed the own resources from the municipal administration to assist the affected people, the triggering of declaration of emergency issued by Brazilian municipalities is required as a legal instrument to obtain appropriate external support. In this article, we develop a study that applies machine learning and explainable AI techniques to generate fuzzy and classic classification models and interpret predictions with a data set that relates the declaration of emergencies and indicators corresponding to Sustainable Development Goals 1, 3, 6 and 10, from 2016 to 2022. A qualitative analysis was performed on the results provided by the explainable AI techniques that led to the identification of indicators that have the greatest influence on predictions and provided additional support to field researchers and decision makers in the context of disaster response.*

1. Introduction

Disasters can be understood as social processes related to physical hazardous events that cause profound disruptions in affected communities [Perry and Quarantelli 2005], [Quarantelli 1998], bringing human suffering and damages of various types, such as human, material, economic, and environmental. Due to the potential to act as a disturbing and disruptive factor in the dynamics and reality of those involved, aspects such as income, dwelling, health, work, education, and basic sanitation can be affected. The degree to which affected actors will be harmed and the time it will take to restore normal daily life depends on the quality of institutional support received [Valencio et al. 2022]. In the face of disasters, if the damage caused exceeds the capacity of local actors to respond to the situation, the local public authority declares an emergency. The duration of emergency decrees allows the local administration to take or induce measures to repair the situation more strategically and urgently, enabling certain social and economic routines, territorial flows and infrastructures to be resumed, even if dealing with a certain precariousness for a longer period of time so that a full recovery can be envisaged.

In an attempt to investigate multidimensional factors associated with disasters, Machine Learning (ML) has been explored by researchers, often accompanied by techniques to explain the behavior of models. The ability of a ML model to explain its behavior has been increasingly demanded by all stakeholders involved in the development and use of such systems. Recent advances in ML systems using black-box models have prompted an increasing demand for these models to offer a minimum level of transparency, especially in domains where the harm resulting from incorrect conclusions is critical. This topic has been addressed within the scope of Explainable Artificial Intelligence (XAI) [Barredo Arrieta et al. 2020], which aims to propose and apply methods that provide the interpretability required for such models to be reliable. In the topic of disasters, XAI has been applied successfully [Albahri et al. 2024].

In general, methods for explaining the functioning of ML models can be aimed at global interpretability - of the model as a whole - or local - where the explanation focuses on specific instances. Although XAI methods are aimed at explaining black-box models, models that are considered naturally transparent can also be difficult to understand, depending on their complexity. Thus, models based on rules, both classical and fuzzy, or trees, can benefit from XAI techniques. When it comes to Fuzzy Systems (FS), additional challenges are posed, such as the semantic meaning of fuzzy sets [Stepin et al. 2024]. In general, the synergy between FS and XAI has been investigated from different approaches, such as the proposal to use fuzzy rules to provide interpretability of opaque systems [Cao et al. 2024], to seek advances in the explanation of type-2 FS [D'Alterio et al. 2020], to explain inferences of FS [Mendel and Bonissone 2021] or to apply methods in different domains [Upasane et al. 2024].

In a broader context, our work focus on analyzing how social, economic, environmental, and territorial indicators are (or are not) related to the declaration of emergency situation or public calamity in Brazilian municipalities. In a previous work [Silva et al. 2024] we used ML and SHAP to explain ensemble models and find the variables with the greatest influence on the classification of disaster types, and thus determine patterns that relate municipal characteristics to types of decrees. In this article, this general objective continues to be pursued, with the exploration of the flexibility offered by fuzzy rules. The problem was formulated as a binary classification, in which municipalities are characterized by their economic, social, spatial, sanitation and health indicators, modeled as input variables, related to the Sustainable Development Goals (SDG) 1 (End poverty in all its forms everywhere), 3 (Ensure healthy lives and promote well-being for all at all ages), 6 (Ensure availability and sustainable management of water and sanitation for all) and 10 (Reduce inequality within and among countries) [United_Nations 2023]. The instances are labeled as *yes* (municipality declared an emergency in a year) or no (municipality did not declare an emergency in a year). After generating two Fuzzy Rule-Based Classification Models (FRBCM) using FARC-HD [Alcala-Fdez et al. 2011] and CHI [Chi et al. 1996] algorithms, one Decision Tree (DT) model using C4.5 [Quinlan 1993] and one ensemble model using Random Forest (RF) [Breiman 2001], we applied two methods for local interpretability of FARC-HD and RF models, with the aim of discovering which variables/indicators have the greatest influence on the classification of instances from the test set. Based on the results of the interpretation techniques, analyzes were performed and evidence was extracted about how to provide support for field researchers and decision makers in the context of disaster response.

This paper is organized as follows: in section 2, the local explanation techniques used in this study are briefly reviewed; in section 3, the dataset and the generated ML models are described; the local interpretations are discussed in Section 4 and Section 5 contains the conclusions and future work.

2. Local Explanation Methods

Explainable AI methods can be classified into a large number of categories, according to different criteria [Barredo Arrieta et al. 2020]. In a simple way, we can differentiate between methods for global explanation, which provide explanations for a model in general, and methods for local explanation, which provide explanations for a specific input data. We also differentiate between model-agnostic methods, which can be used with any model, and model-dependent methods, which are designed for use with a specific model. In this section, we briefly describe the local explanation techniques used in this paper: a technique to explain fuzzy rules by means of linguistic terms and LIME.

2.1. Local Interpretability of Fuzzy Rules

The authors of [Mendel and Bonissone 2021] argue that it is not valid to explain the output of Mamdani or Takagi–Sugeno–Kang (TSK) fuzzy systems using IF_THEN rules. Instead, it is valid to explain the output of such a FS as an association of the antecedents of a small subset of the original larger set of rules, using a phrase such as *"These" linguistic antecedents are symptomatic of "this" output*.

Given a FS, the first step in the method proposed there is to define a codebook, composed of easily understandable words associated to fuzzy sets in partitions on the variable domains that will be used to explain predictions. The reason for defining a codebook is that fuzzy rule learning algorithms, such as FARC-HD [Alcala-Fdez et al. 2011], can, as part of the learning process, adjust the parameters of the fuzzy sets to improve the model performance metrics, which can change the shape of the fuzzy sets to the point of obscuring their semantics and impairing the natural transparency of the rules. Still with the aim of improving performance, it may be necessary to define a large number of sets in each partition, compromising the interpretability of the rules. Defining a codebook composed of fuzzy sets in an adequate number and format to represent easy-to-understand words allows explanations to be presented in a clear and understandable way. Using Jaccard similarity, the fuzzy sets appearing in the rules antecedents are mapped to the word in the codebook with the highest similarity.

According to the explanation method proposed in [Mendel and Bonissone 2021], given a specific instance, the following steps are applied:

- Present the instance to FS to get the model prediction
- Calculate the percentage of contribution of each rule fired
- Select a small set of rules with the highest contribution to the output
- Map the linguistic values of the input variables to the linguistic term in the Codebook with higher similarity
- Create the explanation for each rule and perform the interpretation analysis

The method selecting the small set of rules depends on the particular type of FS, which can be Mamdani FS with COS or centroid defuzzification. In the end, explanation comes up in the form of

x_1 is *low*, and x_2 is *moderate* and x_3 is *high* are symptomatic of $f(x)$ = y,

where x is the input instance and y is the output and f represents the fuzzy inference system.

In this work, we advocate that the above interpretation method, designed for Mamdani FS, can be applied to FRBCM, with the modification of the rule selection step to suit the form and inference of fuzzy classification rules. Therefore, the rule selection criterion proposed in [Mendel and Bonissone 2021] for Mamdani models was adapted accordingly, as described in section 4.1.

2.2. LIME

Local Interpretable Model-agnostic Explanations (LIME)[Ribeiro et al. 2016] is a technique that provides local explanations for prediction models. The idea behind LIME is to train a local surrogate model to explain individual predictions that are usually made by black-box models. This technique assumes that complex models establish nonlinear associations between inputs and outputs and are difficult to explain globally. The proposal is then to create local explanations from specific instances based on the hypothesis that, locally, data points in the neighborhood of a given instance are linearly separable. Local surrogate models are interpretable models that are trained on instances generated to explain individual predictions of black-box machine learning models and are trained to approximate the predictions of the underlying black-box model. Instead of training a global surrogate model, LIME focuses on training local surrogate models. LIME is model-agnostic, which means that it can be applied to any machine learning model. The technique attempts to understand the model by perturbing the input data samples and understanding how the predictions change.

3. Dataset and Models Generation

3.1. Dataset

Data were collected covering the period from 2016 to 2021 for each of the Brazilian municipalities. Brazil has 5,570 municipalities, which led to a total of 33,421 instances, representing each municipality in each year of the period covered. The data were modeled according to a binary classification problem, with the features representing different indicators and the class indicating whether the municipality had a decree in the year or not. The value *yes* for the class indicates that the municipality declared the occurrence of one or more disasters during the year. The value *no* means that no disaster occurred during the year in the municipality. The information on emergency decrees is in accordance with the records of the Integrated Disaster Information System (S2iD)[1], a federal government platform focused on the management of disasters in the country.

We considered thirteen input variables that reflect economic, social, spatial, sanitation, and health indicators, related to the SDGs 1, 3, 6, and 10 [United_Nations 2023]. All variables assume continuous numerical values. Most variables values refer to municipal granularity and some of them refer to state granularity. The data sources for the indicators are websites made available by the Brazilian government such as the National Sanitation Information System (SNIS)[2], Department of Information Technology of the Unified Health System (DATASUS)[3] and the Automatic Recovery System of the Brazil-

[1] https://s2id.mi.gov.br/

[2] https://www.gov.br/cidades/pt-br/acesso-a-informacao/acoes-e-programas/saneamento/snis

[3] http://tabnet.datasus.gov.br/cgi

ian Institute of Geography and Statistics (SIDRA/IBGE) [4]. A summary of the variables and their meaning is shown in Table 1.

Table 1. Variables and their meanings.

Name	Description	Granularity	Source
Year	Year	annual	S2ID
Water	Total population served with water supply	municipal	SNIS
Sewage	Number of households with sanitary installations	municipal	Datasus
Hosp	Number of hospitalizations by place of residence	municipal	Datasus
U5-Mort	Under-5 mortality rate	state	SIDRA
Ill-Mort	Mortality rate from diseases of the circulatory system, malignant tumors, diabetes mellitus and chronic respiratory diseases	state	SIDRA
S-Mort	Suicide mortality rate	state	SIDRA
N-Deaths	Number of deaths, missing persons and people directly affected by disasters per 100,000 inhabitants	state	SIDRA
Births	Live births, occurring in the year, by month of registration, sex, place of birth, number of births per birth, mother's age at the time of birth and mother's place of residence	municipal	SIDRA
Pop	Total population	municipal	SNIS
Pop-Dens	Population density	municipal	SIDRA
Pop-Chan	Absolute population change	municipal	
Geo-rate	Geometric growth rate	municipal	SIDRA

3.2. Classification Models Generation

The first step in the experiments was to generate the classification models using the well-known algorithms: FARC-HD [Alcala-Fdez et al. 2011] and CHI[Chi et al. 1996], which are FRBM, the tree-based algorithm C4.5 [Quinlan 1993] and the ensemble algorithms Random Forest (RF)[Breiman 2001]. The algorithms FARC-HD, CHI and C4.5 were run using Keel software[5][Triguero et al. 2017], a platform that provides various Machine Learning algorithms for multiple applications. RF was executed on the Weka platform [Frank and Hall 2016]. The dataset was partitioned into training and test sets, according to the k-fold cross-validation strategy, with k=10. The performance of the experiments was assessed using the metrics of accuracy, precision, recall, and F1-Score. Table 2 shows the mean values for each evaluation metric and each algorithm.

The model with the best performance, according to all metrics, is RF. Between the two fuzzy models, FARC-HD generated the one with the highest performance. These two models were selected to be further analyzed by means of local interpretability techniques.

4. Explaining the Models by Local Interpretability

The local interpretations of the two models selected in the previous section are described.

4.1. Local Interpretability Analysis of Fuzzy Rules

The interpretability analysis of fuzzy rules generated by the FARC-HD algorithm was done using the technique proposed in [Mendel and Bonissone 2021]. We selected the

[4] https://sidra.ibge.gov.br

[5] www.keel.es

Table 2. Performance metrics of the generated models.

	Class	C4.5	RF	CHI	FARC-HD
Accuracy		88%	89%	77%	84%
Precision	no	86%	86%	75%	83%
	yes	91%	91%	81%	86%
	mean	88%	90%	78%	84%
Recall	no	94%	89%	76%	90%
	yes	81%	83%	61%	76%
	mean	87%	89%	75%	83%
F1-score	no	90%	91%	81%	86%
	yes	85%	87%	70%	80%
	mean	88%	89%	76%	83%

model generated by the partition of the cross-validation sampling strategy that had the highest accuracy (85%). The partition has 30,079 instances in the training set and 3,342 instances in the test set. The model generated from this fold has 35 rules, and each variable domain was partitioned into five fuzzy sets. With a high number of fuzzy rules, the natural interpretability favored by the linguistic terms can be obscured. FARCD-HD adjusts the fuzzy sets by lateral tuning, which can generate sets with large overlapping areas and with different supports, further impairing interpretability. The local interpretability method chosen to understand specific predictions meets the initial objectives of our research, that is, to discover the possible relationship between the indicators selected as input variables for the model (see section 3) and the occurrence of disasters in municipalities.

Some adaptations were made to the original method proposed in [Mendel and Bonissone 2021]. First, since the five fuzzy sets created and tuned during the process overlap in such a way that their central points maintain the order of the sets with respect to the domain of the variable, thus minimally maintaining the semantic meaning of the sets, we did not use a Codebook. The meaning of the rules fired in the classification inference is described using the linguistic terms *Very Low*, *Low*, *Medium*, *High* and *Very high*, associated to the five fuzzy sets generated by FARC-HD. Second, the selection of the five rules with the highest contribution to the classification takes into account the rules with the predicted class in the consequent. The third and last modification regards the form to express the interpretation. The rules generated by FARC-HD have at most three variables in the antecedent, making the appearance of the same variable in more than one of the five most influential rules very rare. Therefore, the analyses are not expressed in a specific format, but instead presented in a textual form.

The test instances used in the local explanation technique were selected from the test set, being two instances belonging to the *yes* class and two instances belonging to the (*no*) class. All four instances were correctly classified by the FARC-HD model. Table 3 presents the variable values for each instance considered. The following steps were applied to each of the test instances.

- Present the instance to the FRBCM generated by FARC-HD to get the prediction
- Calculate the percentage of contribution of each rule fired with the predicted class in the consequent
- Select the five rules with the highest contribution to the classification
- Create the explanation for each rules and perform the interpretation analysis

Table 4 presents the results generated by the interpretability process described here for each of the 4 selected instances. The third column of the table (Contrib.(%)) indicates the percentage to which the rule contributed to the classification and the last column (CF) shows the certainty factor of the rule, found by FARC-HD, which is a measure of how confident one can be in the classification predicted by that rule. Note that, as the rules are generated by FARC-HD, it is not expected to get the same variables in all rules.

Table 3. Values of selected instances for interpretability.

	Instance 1	Instance 2	Instance 3	Instance 4
Year	2021	2021	2019	2016
Water	27696	1731	6908	5560
Sewage	366	24	43	68
Hosp	2052	147	984	611
U5-Mort	16.5	15.1	19.3	19.1
Ill-Mort	4502	10424	7641	7412
S-Mort	8	9.8	5.1	4.6
N-Deaths	466.6	253.8	505.2	34.2
Births	422	32	299	198
Pop	27696	2462	15732	16226
Pop-Dens	11.76	2.74	22.65	8.7
Pop-Chan	2687	-331	-146	1237
Geo-rate	0.83	-0.88	-0.09	0.63
Class	yes	yes	no	no

Table 4. Interpretability of fuzzy rules generated by FARC-HD.

Inst.	Fuzzy rules	Contrib.(%)	CF
1	1: IF Year IS High THEN yes	16.70	0.5078
	2: IF Water IS Very Low AND U5-Mort IS Medium AND Geo-rate IS Low THEN yes	15.68	0.4768
	3: IF Year IS High AND Sewage IS Very Low THEN yes	15.13	0.4601
	4: IF U5-Mort IS Medium AND Ill-Mort IS Low AND Geo-rate IS Medium THEN yes	12.88	0.3915
	5: IF S-Mort IS Medium AND N-Deaths IS Low AND Pop-Chan IS Medium THEN yes	12.45	0.3785
2	1: IF Water IS Very Low AND U5-Mort IS Medium AND Geo-rate IS Low THEN yes	18.21	0.6491
	2: IF U5-Mort IS Medium AND Ill-Mort IS Low AND Geo-rate IS Low THEN yes	17.22	0.6139
	3: IF U5-Mort IS Low AND Ill-Mort IS Low AND S-Mort IS Medium THEN yes	15.42	0.5497
	4: IF Year IS High THEN yes	14.25	0.5078
	5: IF Year IS High AND Sewage IS Very Low THEN yes	12.90	0.4599
3	1: IF Year IS Medium THEN no	17.27	0.7119
	2: IF Year IS Medium AND U5-Mort IS Medium THEN no	17.27	0.7119
	3: IF U5-Mort IS Very Low AND Pop-Dens IS Very Low AND Geo-rate IS Medium THEN no	15.92	0.6562
	4: IF Year IS Medium AND Sewage IS Very Low THEN no	15.64	0.6448
	5: IF Hosp IS Very Low AND U5-Mort IS Very Low AND S-Mort IS Low THEN no	14.29	0.5893
4	1: IF U5-Mort IS Very Low AND Pop-Dens IS Very Low AND Geo-rate IS Medium THEN no	26.01	0.8413
	2: IF Year IS Very Low THEN no	20.16	0.6519
	3: IF Year IS Very Low AND Water IS Very Low THEN no	14.93	0.4828
	4: IF Hosp IS Very Low AND U5-Mort IS Very Low AND S-Mort IS Low THEN no	13.77	0.4454
	5: IF Water IS Very Low AND U5-Mort IS Very Low AND S-Mort IS Low THEN no	13.08	0.4231

From Table 4, experts in the disaster domain can extract evidence that contributes to understanding the relationships between indicators and emergency decrees and between these and the SDGs. In this sense, having in mind SGDs 3 (Ensure healthy lives and promote well-being for all at all ages), 6 (Ensure availability and sustainable management of water and sanitation for all), and 11 (Make cities and human settlements inclusive, safe, resilient, and sustainable), we highlight the following.

4.1.1. Interpretation of the rules fired for Instances 1 and 2 - Class *yes*

Three of the rules are the same for both instances: Rules 1, 2, and 3 for instance 1 and Rules 1, 4 and 5 for instance 2. This is an indication that the variables and their values appearing in these rules are among the most influential in the prediction of class *yes*. Let us highlight the main variables that appear in the rules.

- The variable *Year* appears in two rules for instance 1 (Rules 1 and 3) and two for instance 2 (Rules 4 and 5) with value *Very High*
- The variable *Water* appears in one rule for instance 1 (Rule 2) and one for instance 2 (Rule 1) with value *Very Low*
- The variable *Sewage* appears in one rule for instance 1 (Rule 3) and one for instance 2 (Rule 5) with value *Very Low*
- The variable *U5-Mort* appears in two rules for instance 1 (Rules 2 and 4) and two for instance 2 (Rules 1 and 2) with value *Medium*
- The variable *Geo-rate* appears in two rules for instance 1 (Rules 2 and 4) with values *Low* and *Medium* and two for instance 2 (Rules 1 and 2) with value *Low*
- The variable *Ill-Mort* appears in one rule for instance 1 (Rule 4) and in two for instance 2 (Rules 2 and 3) with value *Low*

This examination summarizes as the linguistic explanation shown in Table 5.

Table 5. Linguistic Explanation

Instances belonging to class *yes*		Instances belonging to class *no*	
Variable	Linguistic Value	Variable	Linguistic Value
Year	Very High	Year	Medium \ Very Low
Water	Very Low	U5-Mort	Very Low
Sewage	Very Low	Pop-Dens	Very Low
U5-Mort	Medium	Geo-rate	Medium
Gen-rate	Medium \ Low	Hosp	Very Low
Ill-Mort	Low	S-Mort	Low

These results indicate that the Brazilian municipalities most susceptible to the hazards faced during the period studied (floods, droughts, diseases) were those that did not provide their populations, even if they were demographically stable or reduced, with access to a decent drinking water service. The precariousness of water supply infrastructure and services for the local population directly affects the poor health of one of the most vulnerable age groups, such as early childhood, which makes them prone to mortality, and people with mental health problems and a lack of appropriate support, both with average values. What may protect these children from a higher risk to their lives may be a better water supply in the educational establishments where they are housed, something that will be investigated later. It is surprising that the variable Ill-Mort (Mortality rate from diseases of the circulatory system, malignant tumors, diabetes mellitus and chronic respiratory diseases), with a low score, appears in one rule for one instance and twice in another instance, associating it with the occurrence of disasters. In this case, the factors involved in the failure of municipalities to address hazards extend beyond effective local health management. A case-by-case check may eventually reveal that these municipalities have a good quality of life. However, they are geographically located in areas prone to large-scale natural events, such as strong winds, hail, and large storms, among others.

4.1.2. Interpretation of the rules fired for Instances 3 and 4 - Class *no*

Two of the rules are the same for both instances: Rules 3 and 5 for instance 3 and Rules 1 and 4 for instance 4. This is an indication that the variables and their values appearing in these rules are among the most influential in the prediction of class *no*. The main variables that appear in the rules are commented on next.

- The variable *Year* appears in three rules for instance 3 (Rules 1, 2 and 4) with value *Medium* and two for instance 4 (Rules 2 and 3) with value *Very Low*
- The variable *U5-Mort* appears in two rules for instance 3 (Rules 3 and 5) and two for instance 4 (Rules 1 and 4) with value *Very Low*
- The variable *Pop-Dens* appears in one rule for instance 3 (Rule 3) and one for instance 4 (Rule 1) with value *Very Low*
- The variable *Geo-rate* appears in one rule for instance 3 (Rule 3) and one for instance 4 (Rule 1) with value *Medium*
- The variable *Hosp* appears in one rule for instance 3 (Rule 5) and one for instance 4 (Rule 4) with value *Very Low*
- The variable *S-Mort* appears in one rule for instance 3 (Rule 5) and one for instance 4 (Rule 4) with value *Low*

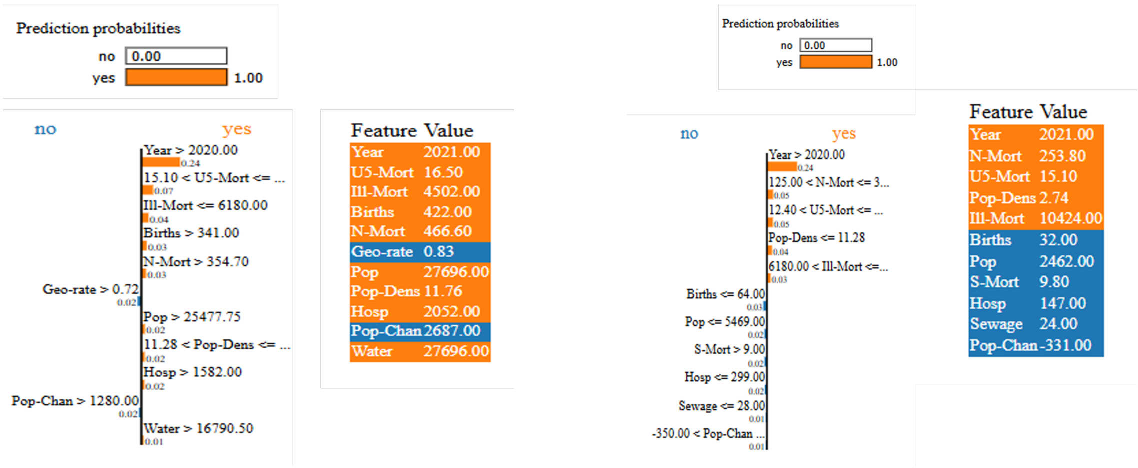
The summary of the linguistic explanation can be found in table 5.

If municipalities that focus on primary maternal and child health care are less prone to disasters, this raises hypotheses that should be investigated in future studies. More directly, the reduced infant mortality rate suggests that a local public policy of appropriate pregnancy care and monitoring of early childhood development has been effective. The commitment of the local administration to the health and well-being of the vulnerable age group from 0 to 5 years could be aligned with other welfare policies, which include consistent civil protection. Indirectly, this could also signal the positive contribution of appreciable sociocultural and economic factors. This hypothesis requires new efforts to associate with other datasets, such as those related to GDP per capita and the population's education levels. The variables in the two rules that fired for both instances confirm that effective local policies for the physical and mental health of the general population, indicated by low hospital admissions and suicide mortality, are more easily implemented when the population is tiny and growing only moderately. This makes territorial planning more feasible and civil protection actions more governable, avoiding the need to declare emergencies to address hazards.

4.2. Local Interpretability using LIME

In this section, we describe the analysis generated by LIME [Ribeiro et al. 2016] to explain the ensemble model generated by RF. An ensemble model is considered a non transparent model, for it is not possible to understand its functioning directly from the model itself, as can be done with rule-based or tree-based models. Therefore, the aid of interpretation techniques is appropriate when the purpose besides learning a model is to reveal the most influential variables in the output. The LIME technique was selected because it provides local interpretation, as the fuzzy method described in the previous section.

The default parameters for LIME were maintained as follows: Lasso regression was used as the surrogate model and an exponential smoothing kernel was used to calculate weights for the generated instances with a kernel width of 0.75 times the square root of the number of columns of the training data. Figures 1a, 1b, 2a and 2b, show the interpretability results generated by LIME implementation[6] for instances 1, 2, 3 and 4, respectively, presented in Table 3. On the left side of the figures one can see the probabilities of classification in each class and the importance of the variables for the prediction. On the right-hand side appear the instance values.



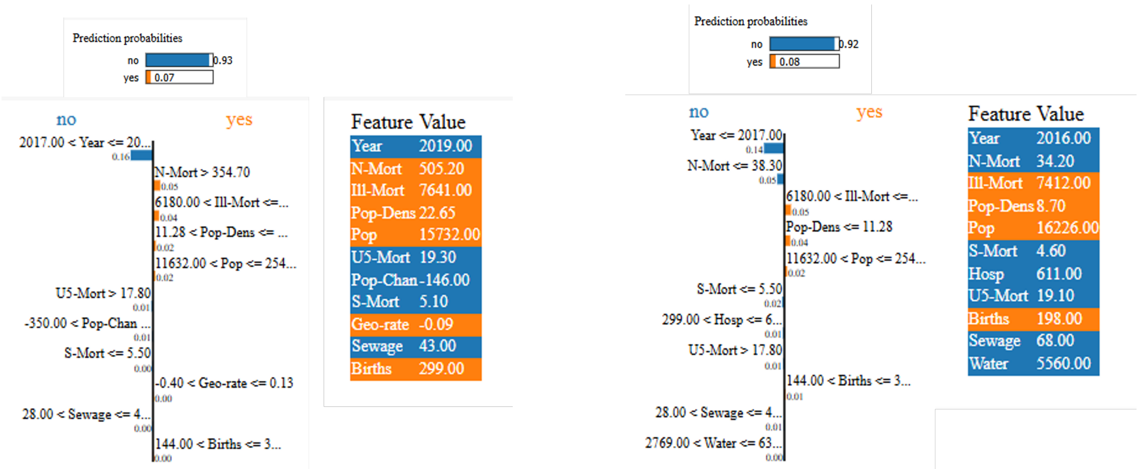
Feature	Value
Year	2021.00
U5-Mort	16.50
Ill-Mort	4502.00
Births	422.00
N-Mort	466.60
Geo-rate	0.83
Pop	27696.00
Pop-Dens	11.76
Hosp	2052.00
Pop-Chan	2687.00
Water	27696.00

(a) LIME Explanation - Instance 1.

Feature	Value
Year	2021.00
N-Mort	253.80
U5-Mort	15.10
Pop-Dens	2.74
Ill-Mort	10424.00
Births	32.00
Pop	2462.00
S-Mort	9.80
Hosp	147.00
Sewage	24.00
Pop-Chan	-331.00

(b) LIME Explanation - Instance 2.

Figure 1. LIME Explanation for Instances of Class *Yes*.



Feature	Value
Year	2019.00
N-Mort	505.20
Ill-Mort	7641.00
Pop-Dens	22.65
Pop	15732.00
U5-Mort	19.30
Pop-Chan	-146.00
S-Mort	5.10
Geo-rate	-0.09
Sewage	43.00
Births	299.00

(a) LIME Explanation - Instance 3.

Feature	Value
Year	2016.00
N-Mort	34.20
Ill-Mort	7412.00
Pop-Dens	8.70
Pop	16226.00
S-Mort	4.60
Hosp	611.00
U5-Mort	19.10
Births	198.00
Sewage	68.00
Water	5560.00

(b) LIME Explanation - Instance 4.

Figure 2. LIME Explanation for Instances of Class *no*.

From Figures 1a–2b, we can extract evidence similar to the ones presented in section 4.1. This analysis provides new insights that complement the evidence obtained in the explanation of the fuzzy model.

The ranking for the most important variables in the prediction are the weights for the linear regressor playing the role of surrogate model. We highlight the following. Figure 1a shows that, in the RF model, five variables in the top ten most important for

[6] https://github.com/marcotcr/lime

classification of Instance 1 were also identified as relevant by the fuzzy technique described in section 4.1 Year, U5-Mort, Ill-Mort, Geo-rate and Water). Similarly, Figure 1b shows that, in the RF model, four variables in the top ten most important for classification of Instance 1 were also identified as relevant by the fuzzy technique.

Analyzing the predictions of the instances from class *No* one can notice that, in the RF model prediction, six variables in the top ten most important for classification of Instance 3 (Year, Ill-Mort, Pop-Dens, U5-Mort, S-Mort and Geo-rate) (Figure 2a) and five variables in the top ten most important for classification of Instance 4 (Year, Por-Dens, S-Mort, Hosp and U5-Mort) (Figure 2b) were identified by the fuzzy technique (2a)

5. Conclusions

In this paper, we investigate the application of local explanation techniques from explainable AI to interpret predictions made by a fuzzy rule-based model and an ensemble model. The models were generated from a dataset that associates spatial, sanitation and health variables/indicators of Brazilian municipalities, related to the Sustainable Development Goals (SDG) 1, 3, 6 and 10 with the occurrence of disasters. The explanations for the fuzzy rule model were obtained with a technique proposed in the literature that uses the linguistic terms of the rule itself to obtain the explanations. The explanations for the ensemble model were generated by the LIME technique. The results of the two techniques are complementary and the qualitative analyses performed shed light on possible points of refinement of the feature engineering and model generation stages, in addition to providing support for actors involved in dealing with disasters. This research will continue with experiments that can bring improvements both in the performance of the model and in the results of the explanation techniques, such as the inclusion and/or exclusion of variables that represent indicators related to other SDGs, execution of other classification algorithms, application of the LIME technique in the fuzzy rule-based model and exploration of global explanation techniques. Another path to be explored, which might bring useful analyzes, is to apply the same strategy to data from each macroregion of the country.

References

Albahri, A. S., Khaleel, Y. L., Habeeb, M. A., Ismael, R. D., Hameed, Q. A., Deveci, M., Homod, R. Z., Albahri, O. S., Alamoodi, A. H., and Alzubaidi, L. (2024). A systematic review of trustworthy artificial intelligence applications in natural disasters. *Computers and Electrical Engineering*, 118(Part B):Article number: 109409.

Alcala-Fdez, J., Alcala, R., and Herrera, F. (2011). A fuzzy association rule-based classification model for high-dimensional problems with genetic rule selection and lateral tuning. *IEEE Transactions on Fuzzy Systems*, 19(5):857–872.

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., and Herrera, F. (2020). Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115.

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

Cao, J., Zhou, T., Zhi, S., Lam, S., Ren, G., Zhang, Y., Wang, Y., Dong, Y., and Cai, J. (2024). Fuzzy inference system with interpretable fuzzy rules: Advancing explainable artificial intelligence for disease diagnosis—a comprehensive review. *Information Sciences*, 662:120212.

Chi, Z., Yan, H., and Pham., T. (1996). *Fuzzy Algorithms: With Applications To Image Processing and Pattern Recognition*. World Scientific.

D'Alterio, P., Garibaldi, J. M., and John, R. I. (2020). Constrained interval type-2 fuzzy classification systems for explainable ai (xai). In *2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–8.

Frank, E. and Hall, M. A. (2016). *The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques"*. Morgan Kauffman, fourth edition.

Mendel, J. M. and Bonissone, P. P. (2021). Critical thinking about explainable ai (xai) for rule-based fuzzy systems. *Trans. Fuz Sys.*, 29(12):3579–3593.

Perry, R. W. and Quarantelli, E. L. (2005). *What is a disaster? New answers to old questions*. Xlibris Press.

Quarantelli, E. L. (1998). *What is a disaster? Perspectives on the question*. Routledge.

Quinlan, J. (1993). *C4.5: Programs for Machine Learning*. Morgan Kauffman.

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*, KDD '16, page 1135–1144, New York, NY, USA. Association for Computing Machinery.

Silva, L. G. T., Matos, A. L., Carvalho, G. G., Valencio, N. F. L. S., and Camargo, H. A. (2024). Explainability of machine learning models with xgboost and shap values in the context of coping with disasters. In *Proceedings of the Brazilian Conference on Intelligent Systems*, BRACIS 2024, pages 152–166, Berlim. Springer.

Stepin, I., Suffian, M., Catala, A., and Alonso-Moral, J. M. (2024). How to build self-explaining fuzzy systems: From interpretability to explainability [ai-explained]. *Comp. Intell. Mag.*, 19(1):81–82.

Triguero, I., Gonzalez, S., Moyano, J. M., Alcalá-Fdez, S. G. J., Fernandez, J. L. A., del Jesús, M. J., Sanchez, L., and Herrera, F. (2017). Keel 3.0: An open source software for multi-stage analysis in data mining. *Int J Comput Intell Syst*, 10:1238–1249.

United_Nations (2023). *The Sustainable Development Goals Report 2023*. United Nations Publications, New York, special edition.

Upasane, S. J., Hagras, H., Anisi, M. H., Savill, S., TAYLOR, I., and Manousakis, K. (2024). A type-2 fuzzy based explainable ai system for predictive maintenance within the water pumping industry. *IEEE Trans. on Artificial Intelligence*, 5(2):490–504.

Valencio, N., Valencio, A., and da Silva Baptista, M. (2022). What lies behind the acute crises: The social and infrasystems links with disasters in brazil. In Iossifova, D., Gasparatos, A., Zavos, S., Gamal, Y., Long, Y., and Yin, Y., editors, *Urban Infrastructuring: Reconfigurations, Transformations and Sustainability in the Global South*, pages 35–52. Springer Nature.