International Journal of Computational Intelligence Systems
Vol. **13(1)**, 2020, *pp.* 974–987
DOI: https://doi.org/10.2991/ijcis.d.200715.003; ISSN: 1875-6891; eISSN: 1875-6883
https://www.atlantis-press.com/journals/ijcis/

Research Article

Teaching Explainable Artificial Intelligence to High School Students

Jose M. Alonso*,

Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), Universidade de Santiago de Compostela (USC), Santiago de Compostela, Spain

ARTICLE INFO

Article History
Received 03 May 2020
Accepted 14 Jul 2020

Keywords
Explainable artificial intelligence
Interpretable computational intelligence
Decision trees
Fuzzy rule-based classifiers
Educational AI resources

ABSTRACT

Artificial Intelligence (AI) is part of our everyday life and has become one of the most outstanding and strategic technologies. Explainable AI (XAI) is expected to endow intelligent systems with fairness, accountability, transparency and explanation ability when interacting with humans. This paper describes how to teach fundamentals of XAI to high school students who take part in interactive workshop activities at CiTIUS-USC. These workshop activities are carried out in the context of a strategic plan for promoting careers on Science, Technology, Engineering and Mathematics. Students learn (1) how to build datasets free of bias, (2) how to build interpretable classifiers and (3) how to build multi-modal explanations.

1. INTRODUCTION

The European Commission (EC) has identified Artificial Intelligence (AI) as "the most outstanding and strategic technology for the 21st century" [1]. Accordingly, many future jobs will require employees with background on AI. This is a good motivation for high school students to think about enrolling into careers on Science, Technology, Engineering and Mathematics, i.e., the so-called STEM careers. Unfortunately, there is a lack of professionals with STEM background and many new positions are hard to be filled. In addition, AI is a multi-disciplinary field where people with background on STEM careers meet people with background on cognitive and social sciences such as Psychology, Linguistics, etc. This fact makes even harder to find qualified professionals for open positions. Hence, there is a need for promoting STEM careers since the earliest.

As we have seen with the global crisis produced by COVID-19, we move toward a new society where new technology (such as Telecommuting, Internet of Things or AI) evolves quickly to fit with our everyday life. In consequence, the time needed for a new technology to become a commodity pervasive to the entire society is shorter and shorter. Thus, in a near future, the social and economic impact of AI will be huge and everyone will need at least basic knowledge on AI. This is the reason why fundamentals of AI are beginning to be disseminated to the society, i.e., to the general public beyond technical audience, e.g., via TED talks.[1]

The EC also emphasizes the importance of Explainable AI (XAI[2] in short) in order to develop an AI coherent with European values [1]: "to further strengthen trust, people also need to understand how the technology works, hence the importance of research into the explainability of AI systems."

This is in accordance with the European General Data Protection Regulation [2] which remarks that European citizens have the "right to an explanation" of decisions affecting them, no matter who (or what AI system) makes such decision. In consequence, XAI attracts attention of researchers not only in STEM careers but also in Ethics [3], Law, Psychology or Social Sciences in a broad sense [4].

Moreover, XAI is in the core of human-centric computing applications, e.g., decision-support and recommender systems for e-Health or e-Learning [5]. Indeed, one of the main challenges of XAI is how to build conversational agents able to provide humans with semantically grounded, persuasive, trustworthy and effective interactive explanations [6,7]. Explanations are expected to be presented as a narrative/story in Natural Language (NL) because this aids human comprehension. Moreover, with the aim of becoming effective, explanations should be communicated to users through multi-modal (i.e., a mixture of textual and graphical but also other modalities) interactive interfaces.

Prof. Zadeh made many highly valuable contributions to the Fuzzy Logic (FL) field and beyond, e.g., the definition of fuzzy sets [8], the concept of linguistic variables and their application to approximate

* *Email: josemaria.alonso.moral@usc.es*

[1] TED stands for Technology, Entertainment and Design. TED talks (https://www.ted.com/) have become a viral video phenomenon with a worldwide community. They include talks for dissemination of AI.

[2] The acronym XAI was introduced by the USA Defense Advanced Research Projects Agency (DARPA) through the so-called XAI DARPA challenge (https://www.darpa.mil/program/explainable-artificial-intelligence).

reasoning [9], the paradigm of computing with words [10] or the computational theory of perceptions [11]. Many of these contributions were pioneer ideas and/or challenging proposals with a lot of potential to be fully developed later by other researchers [12]. It is worth noting that interpretability is one of the most valuable properties of fuzzy sets and systems. Accordingly, interpretability issues have been thoroughly addressed by researchers in the FL field for years. Moreover, interpretability is deeply rooted in the fundamentals of FL [13] and interpretability issues are in the core of XAI too. Therefore, it is not surprising that, as described in the bibliometric analysis presented in [14], about 30% of publications in Scopus related to XAI, dated back to 2017 or earlier, came from authors well recognized in the FL field. Nowadays, XAI is a prominent and fruitful research field where many of Zadeh's contributions can become crucial if they are carefully developed and thoroughly applied.

This paper presents the XAI4ALL dissemination initiative which is developed at the Research Center in Intelligent Technologies (CiTIUS) of the University of Santiago de Compostela (USC). CiTIUS-USC organizes periodically (about once per month) thematic workshops with the 2-fold goal of (1) making Science closer to society and (2) motivating high school students to opt for STEM careers. This initiative is aligned with the vision of the European Higher Education Area (EHEA)[3] which has the origin in the so-called Bologna Process. The EHEA comprises 48 countries which agree to and adopt reforms on higher education with the aim of increasing staff and students' mobility as well as facilitating employability in Europe.

In practice, the XAI4ALL initiative takes place in the form of an interactive workshop with a group of about 20 students (in the range from 6 to 17 years old). Each workshop takes about 60 minutes. It starts with a brief but motivating introduction to XAI and its applications (10 minutes). Then, students play with software tools while following a step-by-step tutorial in which the relevance of FL for XAI is highlighted (30 minutes). Then, students are asked about the rationale they think is behind the provided explanations (5 minutes). Then, they are taught how such explanations are actually generated (10 minutes). The workshop ends with a brief summary of learned lessons (5 minutes).

XAI4ALL was initially conceived (academic year 2018/2019) as XAI4KIDS (see Ref. [15]) with students between 6 and 17 years old. Depending on the age, we slightly changed the introductory part but the interactive part was always the same, with a Scratch game to recognize basketball players. This was especially appealing for teenagers but was hard to follow by the youngest kids. Therefore, in the academic year 2019/2020, we opted for splitting XAI4ALL into two independent workshops (XAI4KIDS and XAI4TEENS) adapted to the age of the students. We changed the focus of the game in case of kids (from 6 to 12 years old) while we kept the basketball game with teenagers (from 13 to 17 years old) and made it more interactive in the sense that now the students are assisted to code their own Scratch program.

The rest of the manuscript is organized as follows: Section 2 introduces the related software. Section 3 sketches the development of an XAI4TEENS workshop. Finally, Section 4 concludes the paper.

[3]http://www.ehea.info/

2. MATERIALS AND METHODS

There are some educational AI resources for children online but as far as we know they do not include any specific XAI resources. For example, "Machine Learning for Kids"[4] is a useful tool for teaching children (1) how to learn AI systems from data and (2) how to use them to make games in Scratch. Another related initiative is "Teens in AI"[5] which organizes activities (such as hackathons, boot camps or mentoring programs) to promote AI among teenagers.

The rest of this section introduces the software that is used in the context of the XAI4ALL initiative.

2.1. Scratch

Scratch[6] is a project of the Lifelong Kindergarten Group at the Massachusetts Institute of Technology (MIT) Media Lab. It is available in more than 40 languages and used worldwide in schools of more than 150 different countries. Kids can learn to code in Scratch and share their creations with others in an online community. As a side effect, they learn strategies for solving problems, designing projects and communicating ideas. Accordingly, students can learn with Scratch from elementary school to college (Scratch is designed especially for ages 8 to 16) and across heterogeneous disciplines (not only STEM but also Linguistics, Art, Social studies, etc.). ScratchED[7] gathers an online community of educators who share their initiatives.

The Scratch editor can run online and offline. It has a user-friendly interface (see Figure 1) that makes coding a very intuitive task. Several predefined "blocks" are available to code procedures (related to movements, events, control actions, etc.) by drag and drop. Additional blocks can be defined by the user what makes Scratch a very simple but powerful programming tool. There are also extensions to connect Scratch programs to specific hardware and/or software. In XAI4ALL we use Scratch 3.0.

The Scratch editor includes three main panels:

- **The execution panel.** On top of the right part of the screen, the execution panel displays the running scene (e.g., it is the

Figure 1 | Scratch editor with the XAI4ALL project.

[4]https://machinelearningforkids.co.uk/

[5]https://www.teensinai.com/

[6]https://scratch.mit.edu/

[7]http://scratched.gse.harvard.edu/

XAI4ALL welcome scene in the case of Figure 1). It is worth noting that Scratch programs are not compiled. Instructions are just interpreted and executed on the fly. The execution begins when clicking the green flag that is on top of this panel. It can be aborted at any time just by clicking the red button that is located just to the right of the green flag.

- **The sprite panel.** It is located just below the execution panel. The characters playing a role in a Scratch program are called sprites. In XAI4ALL, the sprite *X* (which is selected in Figure 1) initializes the program. In addition, the sprite *Guaje* guides the user through the workshop. Notice that, once selecting one sprite, the related information is displayed on the left-hand side of the screen.
- **The programming panel.** It includes all tools needed to visually (just with drag and drop actions) code a Scratch program. There are three different views that change by selecting the corresponding tag (*Code*, *Costumes*, *Sound*) on top. Code is created as a mixture of blocks which represent different kinds of programming instructions (e.g., if-then, control loops, etc.). In Figure 1, Motion blocks are displayed (e.g., move, turn, go to, etc.). The user can build up new blocks as a combination of previous ones and use them in a similar way to functions in other programming languages. It is also possible to define Scratch extensions that are aimed at creating new blocks beyond the available ones. It is worth noting that XAI4ALL does not use any extension. Thus, everything is coded just by reusing blocks available in the editor.

2.2. ExpliClas

ExpliClas[8] is a web service for XAI [16,17]. It is made up of a REST API[9] along with a web client which offers two modes of running (see Figure 2): (1) Beginner and (2) Expert.

In XAI4ALL, we use ExpliClas in the Beginner mode where the user is guided step by step to climb the DIKW[10] pyramid [18,19]:

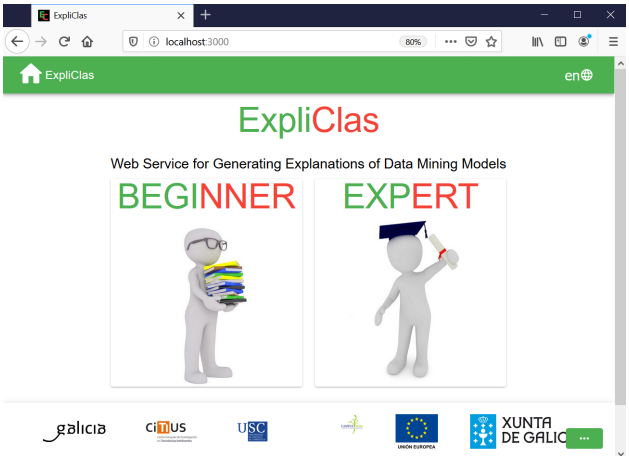

Figure 2 | Home page of the ExpliClas web client.

(1) Creating information from data; (2) creating knowledge from information and (3) creating wisdom from knowledge. In practice, ExpliClas assists users in the following tasks:

- Selecting one out of a pool of preloaded datasets (or uploading a new dataset). Several benchmark datasets (e.g., iris, wine, etc.) are available for illustrative purpose.
- Building one or more interpretable classifiers from the selected dataset. These classifiers automatically extract useful information from data. In addition, the extracted information is represented in a way that is easy to interpret by humans.
- Analyzing the behavior of each classifier regarding both accuracy and explainability, at local and global level. Local analysis involves the classification of one single data instance while global analysis pays attention to the entire dataset (i.e., to the confusion matrix regarding both training and test data). This task is facilitated by means of multi-modal explanations (which are actually human-friendly visualizations along with textual interpretations in NL) of the knowledge embedded into the classifier under consideration.
- Comparing several classifiers in terms of their balance between accuracy and explainability.

ExpliClas makes transparent and intuitive the generation and use of Weka[11] classifiers [20], even by nonexpert users. It is worth noting that Weka is a well-known open source Data Mining project, leaded by researchers affiliated to the University of Waikato (New Zeland) and with a community of users and developers worldwide. The current version of ExpliClas provides users with the following six classification algorithms:

- **RandomForest** (RF). This is an ensemble learning method that creates a combination of decision trees which are generated with the C4.5 algorithm first introduced by Quinlan [21,22]. Even though single classifiers are deemed as interpretable white boxes, their random combination is hardly interpretable and is considered as a black box. RF is included in ExpliClas because, as explained in [23], it is able to get high accuracy in most classification problems and therefore it is commonly taken as baseline from the point of view of accuracy.
- **J48**. This is the implementation in Weka of the Quinlan's C4.5 algorithm. It actually generates pruned C4.5 classifiers which are deemed as interpretable white boxes because by traversing the trees from root to leaves it is possible to understand the classification of every single data instance.
- **RepTree** (REPT). This is a variant of the Quinlan's C4.5 algorithm. It uses regression tree logic and creates multiple trees in different iterations. At the end, it selects the best tree among all the generated trees. The generated classifiers are considered as white boxes because they are decision trees with the same format than those generated by J48.
- **RandomTree** (RANDT). This is another variant of the Quinlan's C4.5 algorithm. It applies bagging to produce a random set of training data instances for the generation of

[8] ExpliClas Web Client: https://demos.citius.usc.es/ExpliClas/

[9] ExpliClas API: https://demos.citius.usc.es/ExpliClasAPI/

[10] DIKW stands for Data, Information, Knowledge and Wisdom.

[11] The Waikato Environment for Knowledge Analysis: https://www.cs.waikato.ac.nz/ml/weka/

several decision trees. At the end, similarly to REPT, it provides users with only one individual tree model, and because of that the generated classifier is also deemed as a white box.

- **Fuzzy Hoeffding Decision Tree** (FHDT) [24]. This is a decision tree whose structure resembles a directed acyclic graph, with internal nodes representing a test on a feature, branches denoting the outcome of the test and terminal nodes (leaves) containing instances belonging to one or more class labels. It is actually a fuzzy extension of the original Hoeffding Decision Tree and it is deemed as a gray box because it is interpretable at certain degree depending on the complexity of the fuzzy tree. Moreover, FHDT is enriched with global semantics and linguistic interpretability what makes an outstanding difference with respect to the previous C4.5 trees which have local semantics and lack of linguistic interpretability. Thus, in addition to the usual visualization of the tree structure, it is also possible to visualize the fuzzy inference and naturally verbalize (in NL) classifications made by FHDT [17].
- **Fuzzy Unordered Rule Induction Algorithm** (FURIA) [25]. This algorithm generates fuzzy IF-THEN classification rules with fuzzy sets of trapezoidal shape in the antecedent of each rule. It is worth noting that in contrast with FHDT, FURIA deals with fuzzy sets which have local semantics and lack of linguistic interpretability. Therefore, classifiers generated by FURIA are deemed as gray-box classifiers because they are made up of a set of rules which can be interpreted (at certain degree) by users. If there is no rule with activation degree greater than zero then FURIA offers to the user three options: (1) Abstain (i.e., no output class is given); (2) voting for the a-priori most frequent class in the dataset and (3) rule stretching (i.e., a new set of rules is dynamically created from the initial rule base by removing antecedents in order one by one, rule by rule, until the instance is covered).

Once one of the algorithms enumerated above is used to build an interpretable classifier, then the Explainer module in ExpliClas is in charge of the classifier analysis, i.e., automatically interpreting classifications made by the given classifier.

It is worth noting that in case of RF the analysis is made by comparison with the other algorithms but no explicit explanation is generated. On the other hand, in the case of interpretable classifiers, two kinds of explanation are provided: (1) Global and Local Explanation of the selected classifier (which is trained by Weka with 10-fold cross-validation) and (2) Explanation of the confusion matrix (regarding both training and test data).

Explanations comprise multiple modalities, i.e., they include graphical visualization along with sentences in NL. Branches of C4.5 decision trees (J48, REPT and RANDT) are first translated into crisp IF-THEN rules by traversing each tree from the root to the leaves. In addition, FHDT and FURIA produces naturally fuzzy IF-THEN rules. Then, decisions are verbalized and justified in terms of the fired rules (no matter the nature, either crisp or fuzzy, of the selected classifier). In the case of FHDT and FURIA, several rules can be fired with different activation degrees (for a given test instance) what makes easier handling naturally imprecision and uncertainty but requires paying attention to co-fired rules (while in crisp trees there is always only one fired rule per test instance). Rule co-firing is visualized by means of fuzzy inference-grams (fingrams) [26]. In the case of crisp trees, even if only one branch can be activated for a given test instance, ExpliClas analyzes potential alternative branches when data values are close to the borderline split values in the decision nodes of the trees. This information is useful to analyze the robustness of the classifier because slight variations in a given test instance are likely to activate these alternative branches.

Since only FHDT is endowed with global semantics, in order to guarantee semantic grounding for classifiers generated by the rest of algorithms, global semantics is enforced (no matter if the selected algorithm is either crisp or fuzzy) by means of defining beforehand strong fuzzy partitions with the suitable number of linguistic terms (e.g., small, medium, large) for each decision variable. Then, either split values in decisions trees or fuzzy intervals in FURIA are interpreted in terms of the closer linguistic terms previously defined. Those terms are used to verbalize decisions in the form of textual explanations. Moreover, no matter the classifier under study, the same linguistic terms are used to build all the textual explanations. In consequence, it is feasible to make a comparison, at linguistic level, of the explanations associated to different classifiers.

Such explanations are generated by means of a Natural Language Generation (NLG) module which implements the NLG pipeline (Macro Planner + Micro Planner + Surface Realizer) proposed by Reiter and Dale [27]. This NLG pipeline is recognized as the most popular in the related literature [28]. The ExpliClas NLG pipeline actually includes a multilingual linguistic realizer (English,[12] Spanish[13] and Galician[14]) which is a mixture of templates and text dynamically generated.

The interested reader is kindly referred to [17] for further details about ExpliClas and how the linguistic approximation and realization are carried out.

3. THE XAI4TEENS WORKSHOP

This section describes step-by-step a XAI4ALL workshop session adapted for teenagers, i.e., high school students from 13 to 17 years old. Basic knowledge about fundamentals of Scratch is required for students attending this workshop. Fortunately, most Spanish high school students are used to code with Scratch. Otherwise, they are asked to follow a selection of Scratch tutorials in advance.

3.1. Introduction to XAI in Scratch

The workshop starts with a brief introduction to XAI that is coded as a Scratch story ready to play. Once clicked the "START" button (see Figure 1), the sprite *Guaje* tells a story about XAI (see Figure 3).

The story begins with an informal definition of intelligence. The relevance of five words (i.e., autonomy, curiosity, learning, knowledge and creativity) is emphasized in connection with intelligence. Thus, intelligence can be defined as "the ability to learn or understand...

[12] SimpleNLG [29] github: https://github.com/simplenlg/simplenlg

[13] SimpleNLG-ES [30] github: https://github.com/citiususc/SimpleNLG-ES

[14] SimpleNLG-GL [31] github: https://github.com/citiususc/SimpleNLG-GL

the ability to apply knowledge to manipulate the environment... the ability to make reasoning and autonomous decisions." Then, *Guaje* refers to the theory of multiple intelligences [32] and the need to develop carefully all of them via education, training or experience. Then, it is time to introduce AI as the ability of a digital computer (computer-controlled robot or any other electronic device) to perform tasks commonly associated with intelligent beings. Indeed, AI is frequently associated to developing systems endowed with the intellectual processes characteristic of humans (e.g., the ability to reason, discover meaning, generalize or learn from past experience). However, in agreement with the incompatibility principle postulated by Zadeh [33], the more powerful AI-based systems become, the more complex and the less comprehensible they are. Thus, XAI emerges with the aim of facing the challenge of building self-explanatory AI-based systems.

3.2. Climbing the DIKW Pyramid

The path to build XAI-based systems goes all over from data to wisdom (see the DIKW pyramid in Figure 4).

Figure 3 | A story about explainable artificial itelligence (XAI) in Scratch.

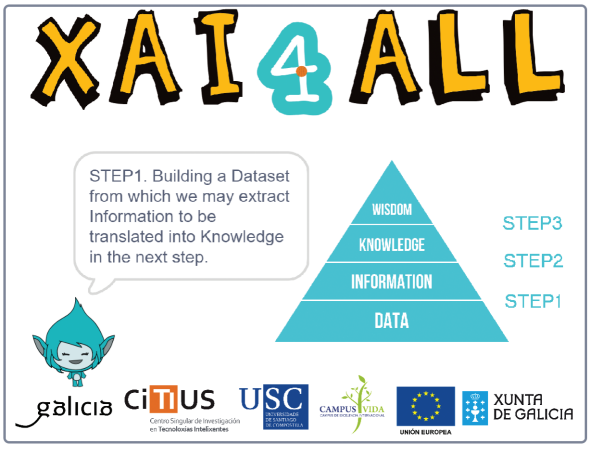

Figure 4 | How to build explainable artificial itelligence (XAI)-based systems?

First, raw data need to be processed and translated into useful information (STEP 1). Then, information is translated into knowledge thanks to tools provided by knowledge engineering (STEP 2). Then, knowledge acquired by the user is assimilated and put into context of his/her previous background and thus it becomes wisdom (STEP 3).

As an illustrative use case we have coded in Scratch an interactive game for classifying roles of basketball players.

Raw data were manually collected from the statistics available online in the websites of the Spanish Basketball Leagues for 50 male[15] and 50 female[16] basketball players. Thus, we created the basketball-players classification dataset.[17] The dataset is made up of 100 instances corresponding to five classes (Point-Guard, Shooting-Guard, Small-Forward, Power-Forward, Center) which are linked to 11 attributes (Height, Minutes, Points, 2-points Field Goals Percentage, 3-points Field Goals Percentage, Free Throws Percentage, Rebounds, Assists, Blocks, Turnovers and Global Assessment). It is worth noting that we have taken statistics averaged all over the entire career of each player.

Table 1 shows the distribution of data instances regarding the output class (i.e., the player role). It is easy to appreciate how data are unbalanced when segmenting by human skin color (BLACK *versus* WHITE) even if the initial dataset was carefully built to be perfectly balanced with 20 instances (10 male and 10 female players) belonging to each class. Notice that, we are aware that human skin color ranges from the darkest (BLACK) to the lightest (WHITE) hues and just referring to BLACK and WHITE is an over generalization. We overgeneralize here for the sake of simplicity and for illustrating the effect of over generalization when learning from data. Unfortunately, over generalization takes place in the preparation of many datasets and it is not always easy to note.

In addition, Table 2 shows the distribution of data instances in the entire basketball-players dataset regarding gender (MEN *versus* WOMEN) and skin color (BLACK *versus* WOMEN). It is easy to appreciate that there are many more white than black players. This fact is in agreement with the distribution of players in Spanish basketball teams.

It is worth noting that gender and skin color information is provided implicitly by the photos depicted in Figure 5 but is not explicitly coded in the dataset for the sake of fairness [34]. Actually, when running XAI4TEENS with different groups of students and asking them to select WHITE *versus* BLACK players there is always an

Table 1 | Distribution of data instances in the basketball-players dataset (regarding player role: PG stands for Point-Guard, SG is Shooting-Guard, SF is Small-Forward, PW is Power-Forward and CE is Center).

	PG	SG	SF	PF	CE
MEN	10	10	10	10	10
WOMEN	10	10	10	10	10
BLACK	1	4	6	11	8
WHITE	19	16	14	9	12
ALL	20	20	20	20	20

[15] http://www.acb.com/

[16] http://competiciones.feb.es/estadisticas/

[17] https://gitlab.citius.usc.es/jose.alonso/basketballplayers-dataset

interesting discussion. In practice, the perception of color changes from one person to other, therefore some players are arbitrarily tagged as WHITE or BLACK only because we force students to do a binary classification. As a result, they understand the utility of considering not only binary but fuzzy classification.

Before addressing the STEP1, i.e., extracting information from data, we must first create and cure a dataset that is free of bias. In order to illustrate the importance of this initial stage, we ask students to build their own dataset. To do so, they have to select a pool of players among those available (see Figure 5). Once the photo of one player is clicked, then statistics are displayed and the player can be ticked as selected or not (see Figure 6).

Once students have selected a list of players, then the workshop chair generates a reduced version of the basketball-players dataset which contains only the selected players. Then, this dataset is uploaded to ExpliClas with the aim of climbing step by step the DIKW pyramid (see Figure 4).

Table 2 Distribution of data instances in the basketball-players dataset (regarding gender and skin color).

	BLACK	WHITE	ALL
MEN	12	38	50
WOMEN	18	32	50
ALL	30	70	100

Figure 5 Building a dataset in Scratch.

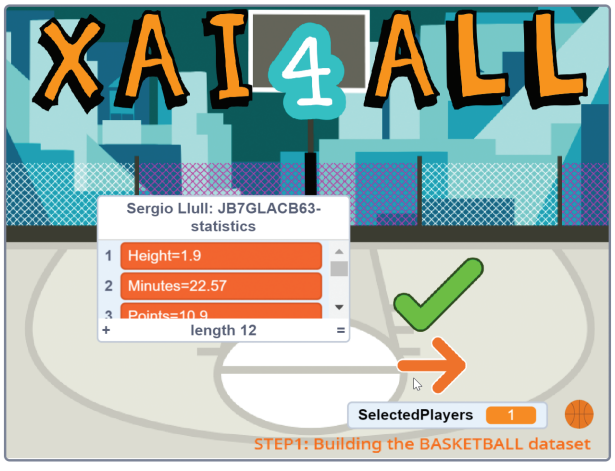

Figure 6 Selecting a basketball player in Scratch.

In the rest of this section, for the sake of simplicity and with illustrative purpose, we will summarize the results reported with ExpliClas for the five subsets of data instances given in Table 1. They are actually five different versions of the basketball-players dataset. The aim of these experiments is to show the effect on data-driven modeling of arbitrary segmenting datasets. The goodness (and fairness) of a model extracted from data depends on the goodness (and fairness) of the available data. First, we will pay attention to the goodness of the generated models in terms of their accuracy, then we will focus on interpretability issues and finally we will analyze the interpretability-accuracy trade-off for each model.

We have generated classifiers with all the algorithms introduced in Section 2.2. In all cases, we have directly applied the default values suggested by ExpliClas because finding out the most accurate classifiers is out of the scope of this work. In the case of FHDT we considered strong fuzzy partitions with 3, 5 and 7 linguistic terms what produced FHDT3, FHDT5 and FHDT7, respectively.

With the aim of analyzing the balance between accuracy and interpretability in each classifier, we have measured:

- Accuracy in terms of: (1) The ratio of correctly classified instances (CCIs) which is reported in Table 3; (2) the F-score (see Table 4) which is the harmonic mean of Precision and Recall; (3) the Mean Absolute Error (MAE) which is reported in Table 5 and (4) the Root Mean Square Error (RMSE) which is detailed in Table 6.
- Interpretability in terms of: (1) The number of leaves (in decision trees) or rules (in fuzzy classifiers), NR in short, as detailed in Table 7; (2) the length of tree branches (in decision trees) or rules (in fuzzy classifiers), TRL in short, which is reported in Table 8 and (3) the number of distinct concepts (NC) which are embedded in the classification model (see Table 9).

All reported results correspond to average values for 10-fold cross-validation.

Tables 3–9 share the same format. They have 5 columns (one per dataset) and 8 rows (one per classifier). The best classifier (according to the quality index reported in the table) is highlighted in bold for each dataset. On the one hand, in the case of accuracy indexes, the best classifier is the one with the highest CCI (see Table 3), the

Table 3 Percentage of CCIs.

	MEN	WOMEN	BLACK	WHITE	ALL
RF	70.00	58.00	46.67	**55.71**	**63.00**
J48	58.00	**60.00**	33.33	42.86	55.00
REPT	62.00	52.00	23.33	51.43	46.00
RANDT	54.00	58.00	33.33	47.14	54.00
FHDT3	58.00	46.00	50.00	44.29	39.00
FHDT5	68.00	48.00	40.00	28.57	43.00
FHDT7	**74.00**	40.00	**56.67**	35.71	46.00
FURIA	68.00	48.00	33.33	41.43	53.00

CCI, correctly classified instance; RF, RandomForest; REPT, RepTree; RANDT, RandomTree; FHDT, Fuzzy Hoeffding Decision Tree; FURIA, Fuzzy Unordered Rule Induction Algorithm.

Table 4 F-score.

	MEN	WOMEN	BLACK	WHITE	ALL
RF	0.700	0.585	0.342	**0.555**	**0.631**
J48	0.579	**0.601**	0.328	0.410	0.544
REPT	0.623	0.446	0.174	0.417	0.454
RANDT	0.529	0.572	0.356	0.471	0.538
FHDT3	0.442	0.435	0.257	0.303	0.330
FHDT5	0.621	0.463	0.247	0.233	0.390
FHDT7	**0.701**	0.367	**0.441**	0.331	0.443
FURIA	0.676	0.468	0.183	0.387	0.510

RF, RandomForest; REPT, RepTree; RANDT, RandomTree; FHDT, Fuzzy Hoeffding Decision Tree; FURIA, Fuzzy Unordered Rule Induction Algorithm.

Table 5 MAE.

	MEN	WOMEN	BLACK	WHITE	ALL
RF	0.176	0.212	0.252	0.224	0.209
J48	0.173	**0.166**	0.273	0.227	0.193
REPT	0.178	0.218	0.316	0.240	0.239
RANDT	0.184	0.168	0.267	**0.211**	**0.184**
FHDT3	0.264	0.281	0.255	0.290	0.285
FHDT5	0.218	0.256	0.280	0.280	0.271
FHDT7	0.192	0.251	0.253	0.264	0.271
FURIA	**0.131**	0.215	**0.248**	0.223	0.210

MAE, mean absolute error; RF, RandomForest; REPT, RepTree; RANDT, RandomTree; FHDT, Fuzzy Hoeffding Decision Tree; FURIA, Fuzzy Unordered Rule Induction Algorithm.

Table 6 RMSE.

	MEN	WOMEN	BLACK	WHITE	ALL
RF	**0.282**	**0.322**	0.367	**0.335**	**0.316**
J48	0.381	0.379	0.484	0.435	0.402
REPT	0.342	0.373	0.443	0.377	0.391
RANDT	0.429	0.410	0.516	0.460	0.429
FHDT3	0.345	0.365	0.358	0.375	0.368
FHDT5	0.314	0.354	0.396	0.377	0.366
FHDT7	0.300	0.366	**0.357**	0.365	0.367
FURIA	0.328	0.426	0.434	0.433	0.408

RMSE, root mean square error; RF, RandomForest; REPT, RepTree; RANDT, RandomTree; FHDT, Fuzzy Hoeffding Decision Tree; FURIA, Fuzzy Unordered Rule Induction Algorithm.

Table 7 Number of leaves/rules (NR).

	MEN	WOMEN	BLACK	WHITE	ALL
RF	–	–	–	–	–
J48	8	9	7	13	17
REPT	5	**3**	**3**	**2**	8
RANDT	14	15	13	27	28
FHDT3	**3**	**3**	**3**	5	**5**
FHDT5	5	5	5	5	**5**
FHDT7	7	7	7	7	7
FURIA	7	8	4	10	10

RF, RandomForest; REPT, RepTree; RANDT, RandomTree; FHDT, Fuzzy Hoeffding Decision Tree; FURIA, Fuzzy Unordered Rule Induction Algorithm.

highest F-score (see Table 4), the lowest MAE (see Table 5) and the lowest RMSE (see Table 6). On the other hand, in the case of interpretability indexes, the best classifier is associated to the smallest NR (see Table 7), the shortest TRL (see Table 8) and the smallest NC (see Table 9). It is worth noting that we have not reported values of interpretability indexes for the RF algorithm because it generates black-box classifiers.

Table 8 Total branch/rule length (TRL).

	MEN	WOMEN	BLACK	WHITE	ALL
RF	–	–	–	–	–
J48	20	27	16	52	70
REPT	10	5	5	**2**	28
RANDT	56	55	51	120	118
FHDT3	**3**	**3**	**3**	8	8
FHDT5	5	5	5	5	**5**
FHDT7	7	7	7	7	7
FURIA	13	15	5	20	23

RF, RandomForest; REPT, RepTree; RANDT, RandomTree; FHDT, Fuzzy Hoeffding Decision Tree; FURIA, Fuzzy Unordered Rule Induction Algorithm.

Table 9 Number of distinct concepts (NC).

	MEN	WOMEN	BLACK	WHITE	ALL
RF	–	–	–	–	–
J48	13	15	11	23	29
REPT	7	4	4	**2**	13
RANDT	26	27	24	49	52
FHDT3	**3**	**3**	**3**	6	6
FHDT5	5	5	5	5	**5**
FHDT7	7	7	7	7	7
FURIA	13	15	5	20	23

RF, RandomForest; REPT, RepTree; RANDT, RandomTree; FHDT, Fuzzy Hoeffding Decision Tree; FURIA, Fuzzy Unordered Rule Induction Algorithm.

To sum up with, regarding accuracy, RF turns up the most accurate classifier for ALL and WHITE regarding CCI, F-score and RMSE. It is also the most accurate classifier for MEN and WOMEN with respect to RMSE. J48 is the most accurate classifier for WOMEN regarding CCI, F-score and MAE. FHDT7 is the most accurate classifier for MEN and BLACK regarding CCI and F-score. It is also the most accurate classifier for BLACK with respect to RMSE. In addition, in case of considering MAE, RANDT turns up as the most accurate classifier for ALL and WHITE while FURIA is the most accurate classifier for MEN and BLACK.

With respect to interpretability issues, FHDT3 turns up the simplest classifier for all three interpretability indexes under consideration (NR, TRL and NC) in MEN, WOMEN and BLACK. In addition, REPT and FHDT5 are the simplest classifiers for WHITE and ALL, respectively.

Of course, Tables 3–9 contain many details that high school students cannot appreciate. Therefore, we show to the students a picture like the ones depicted in Figures 7–9 (but regarding only one dataset, i.e., the one generated by the students) instead of the previous tables. This kind of pictures show the balance between interpretability (measured in terms of NC; see axis x) and accuracy (measured in terms of CCI; see axis y). They help students understand how to select the best classifier when keeping in mind two conflicting goals (i.e., interpretability and accuracy).

In agreement with the no free lunch theorem [35], there is no classification algorithm to produce always the best classifier. On the contrary, the best classifier depends on each dataset but also on the quality index to consider. Thus, we must explore the interpretability-accuracy trade-off when looking for the best classifier for a specific dataset.

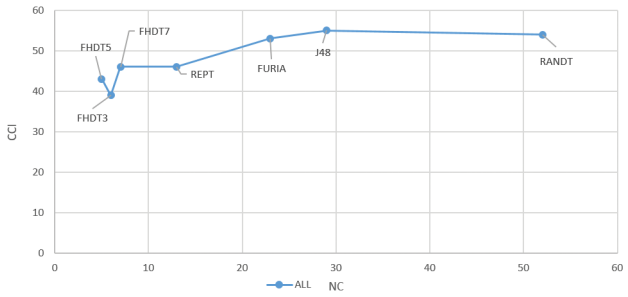

Figure 7 | Interpretability-accuracy trade-off chart (ALL).

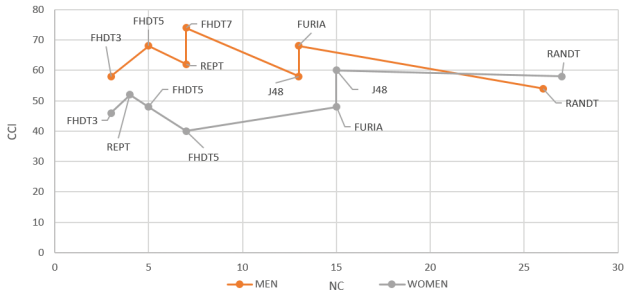

Figure 8 | Interpretability-accuracy trade-off chart (MEN *versus* WOMEN).

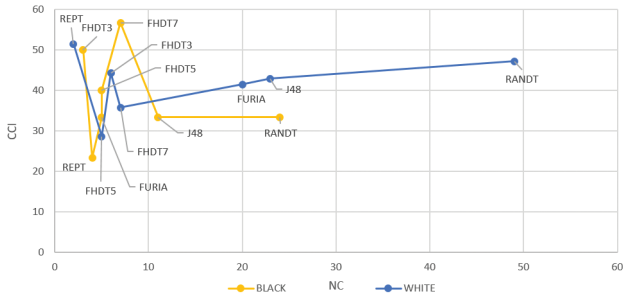

Figure 9 | Interpretability-accuracy trade-off chart (BLACK *versus* WHITE).

In Figure 7, J48 emerges as the classifier with the best interpretability-accuracy trade-off regarding the entire basketball-players dataset (pay also attention to column ALL in Tables 3 and 9). Nevertheless, it is worth noting that the percentage of CCI (55%) is quite low. In addition, the most accurate classifier (i.e., RF) also achieves a low CCI (63%). This fact is probably due to the fact that we are considering together men and women while their statistics are quite different. For example, the height of male players is in [1.81, 2.2] while the height of female players is in [1.66, 1.94]. As a result, we identify two groups of players with small overlapping among them. Even worse, the meaning of short/tall is context dependent and varies in each group. Notice that, Height is by far the most relevant attribute in the dataset. Indeed, it is used in the 80% of the generated classifiers (see the last column in Table 10 where we report the mean value for all the five datasets). Moreover, Height becomes crucial in case of segmenting the dataset by gender. It is used by all classifiers (100%) in the case of MEN and by 87.5% of classifiers in case of WOMEN. As it can be seen in Table 10, the other most relevant attributes (i.e., 3-points Field Goals Percentage, Rebounds and Assists) are only used by about 50% of the classifiers.

Table 10 | Relevance of attributes in the dataset. 2P is 2-points Field Goals Percentage; 3P is 3-points Field Goals Percentage; 1P is Free Throws Percentage and GA is Global Assessment.

Attribute	MEN	WOMEN	BLACK	WHITE	ALL	Mean
Height	**100**	**87.5**	62.5	**75**	**75**	**80**
Minutes	25	37.5	37.5	50	50	40
Points	25	50	12.5	37.5	37.5	32.5
2P	62.5	37.5	25	50	50	45
3P	37.5	37.5	**75**	50	62.5	52.5
1P	25	37.5	12.5	50	37.5	32.5
Rebounds	62.5	62.5	12.5	62.5	50	50
Assists	50	50	37.5	50	62.5	50
Blocks	12.5	12.5	12.5	**75**	50	32.5
Turnovers	25	37.5	50	37.5	50	40
GA	12.5	37.5	37.5	37.5	**75**	40

In the light of the reported results, we observe that it seems a good thing to segment the initial dataset (ALL) into two distinct datasets (MEN and WOMEN) in order to get better accuracy. As it can be appreciated in Figure 8, FHDT7 and J48 are the classifiers with the best interpretability-accuracy trade-off for MEN and WOMEN, respectively.

In case we were segmenting the dataset by skin color instead of gender (see Figure 9), FHDT7 and REPT would emerge as the classifiers with the best interpretability-accuracy trade-off for BLACK and WHITE, respectively. However, both classifiers achieve low accuracy (i.e., CCI around 50%; see Table 3), probably due to the unbalanced and arbitrary segmentation that we did. Accordingly, we can conclude that segmenting the dataset this way makes no sense in the context of the problem under consideration.

Of course, it is arguable if an AI-based system which segmented subjects by gender or skin color would be ethically admissible. Anyway, let us remark that here we are only illustrating how XAI facilitates understanding why the creation of AI-based systems which are personalized for groups of people (e.g., segmenting a population by gender or skin color) can get better or worse balance between accuracy and interpretability. Indeed, XAI faces the challenge of building AI-based systems which are not only accurate but also explainable, fair and accountable intelligent systems. One of the main goals of the XAI4ALL initiative is to make general public (and especially high school students) aware of the need to keep in mind ethical issues when dealing with AI-based systems. Nevertheless, a deeper discussion about Ethics in AI is out of the scope of this paper. For further details, the interested reader is kindly referred to the Ethics guidelines for trustworthy AI which have been issued by the independent high-level expert group on AI set up by the EC [3].

3.3. Explaining Interpretable Classifiers

Once we have evaluated the goodness of the generated classifiers in terms of both accuracy and interpretability, then it is time to use ExpliClas to go deeper in how these classifiers work. Figure 10 shows the decision tree associated to the J48 classifier generated from the WOMEN dataset. Only 6 out of the 11 attributes in the dataset are considered. The tree is made up of 8 decision nodes and 9 leaf nodes.

With a quick visual inspection of the tree (see Figure 10), we observe that Rebounds is in the root of the tree while Height and 3P-Field-Goal-Perc (i.e., the percentage of 3-points goals) appear twice each. Looking a bit more carefully, we can appreciate that Center is associated with the tallest players. In addition, Point-Guard is associated with those players who give more assists. In addition, Height seems to be the key attribute to distinguish between Shooting-Guard and Small-Forward. Likewise, the percentage of three points goals is a good indicator to distinguish between Small-Forward and Power-Forward.

Even if this tree is small, figuring out how it works for a specific data instance requires some expertise and effort. We need to traverse the tree from root to leaves in order to identify the activated output. Fortunately, this task is easier with the help of ExpliClas. Figure 11 shows the branch activated in the tree for a given data instance.

It is worth noting that ExpliClas provides the user with alternative plausible classifications for the given instance and the activated branch is highlighted either in green (right output; see Figure 11) or in red (wrong output; see Figure 12).

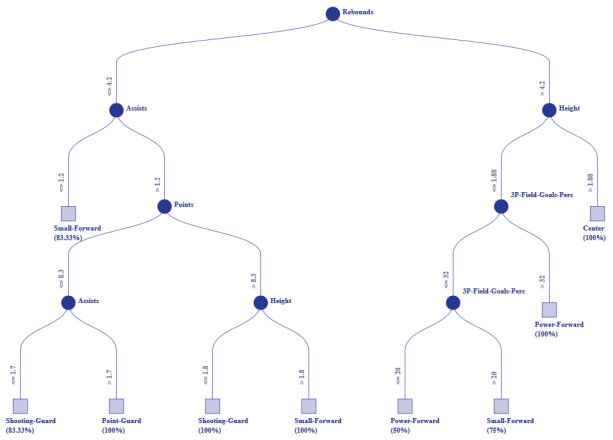

Figure 10 | J48 decision tree for WOMEN.

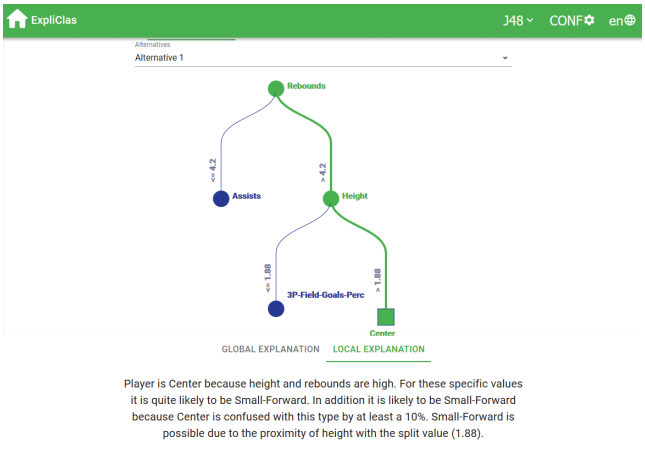

Figure 11 | Screenshot of ExpliClas with local explanation associated to the classification of a test data instance.

A panel with textual (global or local) explanations is displayed at the bottom of the screen (see Figure 11). The global explanation summarizes the classifier behavior, including the list of classes, a comment related to accuracy and one or more comments related to the confusion among classes: "*There are 5 types of player: Point-Guard, Shooting-Guard, Small-Forward, Power-Forward and Center. This classifier is quite confusing because CCIs represent a 60%. There may be confusion related to some types of player. Specifically, when we talk about types Point-Guard, Small-Forward and Shooting-Guard.*" In addition to the visual explanation that is depicted on top of Figure 11, the related local explanation in NL is as follows: "*The player is Center because Height is extremely high and Rebounds is medium. There is also a minor chance that it is Small-Forward.*" The numerical conditions "Height > 1.88" and "Rebounds > 4.2" are verbalized as "*Height is extremely high*" and "*Rebounds is medium,*" respectively. The explanation also remarks that the actual output (Center) may be confused with a wrong alternative output (Small-Forward), what is illustrated in Figure 12. This is consistent with the analysis of the confusion matrix in Figure 13. Namely, the actual Height (1.91) is close to the split value (1.88) in the related decision node. Indeed, the same thing happens for the 10% of cases where Center is confused with Small-Forward in 10-fold cross-validation.

In the case of MEN, the FHDT7 classifier can be visualized either as a fuzzy decision tree (see Figure 14) or as a fuzzy rule base (see Figures 15 and 16). On the one hand, Figure 15 shows the rule view. On the other hand, Figure 16 shows the fingrams view.

All these visualizations come with textual explanations. In the case of local explanations associated to fuzzy rules, in addition to the textual explanation that verbalizes the knowledge embedded into the winner rule, a bar graph (see Figure 15) shows the activation degree associated to each class. It is also possible to visualize the membership functions related to each rule premise. In addition, the fingrams view illustrates the interaction among rules at both global and local inference level.

Since the FHDT7 classifier is created with a strong fuzzy partition of 7 fuzzy sets per attribute, and Height turns up as the most relevant attribute in the dataset (see Table 10), then the rule base is made up

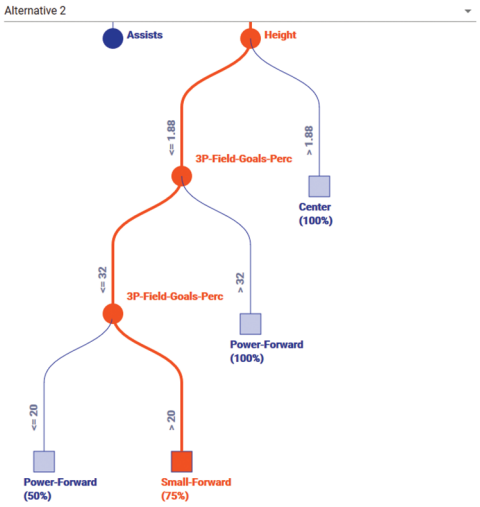

Figure 12 | Visualizing an alternative wrong classification.

Figure 13 | Analysis of the confusion matrix.

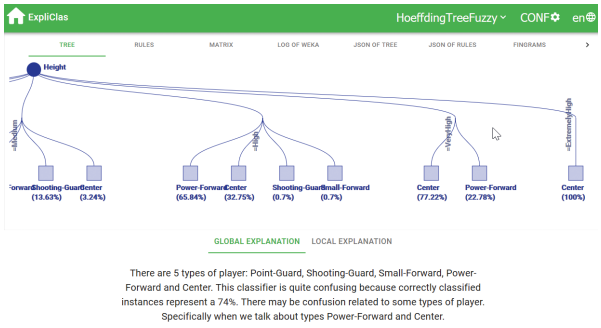

Figure 14 | Screenshot of ExpliClas with the tree view (and global explanation) of FHDT7 for MEN.

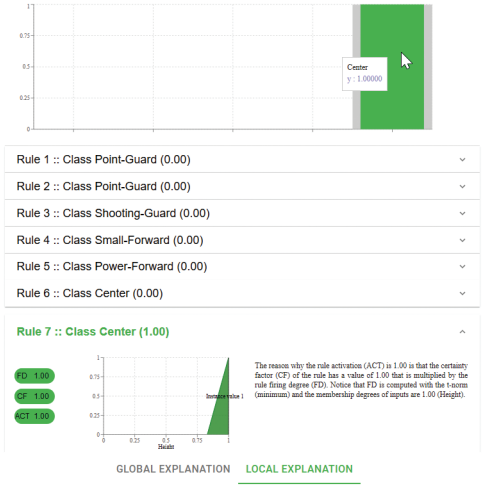

Figure 15 | ExpliClas screenshot with the rule view (and local explanation) of FHDT7 for MEN.

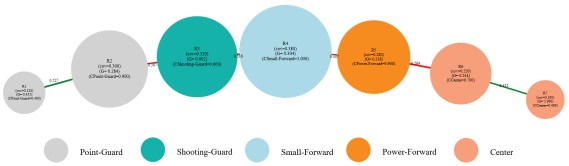

Figure 16 | ExpliClas fingrams view of FHDT7 for MEN.

of 7 rules which relate the height of a player with his role in a basketball team. The linguistic description of these rules is as follows:

- R1. IF Height is Extremely low THEN Player is Point-Guard
- R2. IF Height is Very low THEN Player is Point-Guard
- R3. IF Height is Low THEN Player is Shooting-Guard
- R4. IF Height is Medium THEN Player is Small-Forward
- R5. IF Height is High THEN Player is Power-Forward
- R6. IF Height is Very high THEN Player is Center
- R7. IF Height is Extremely high THEN Player is Center

According to the fingrams view (see Figure 16), it seems to be a linear relation between Height and the output class (what is confirmed just reading the list of rules given above). Thus, Point-Guard is the shortest player and Center is the tallest one. In addition, there may be confusion among players with a borderline height between two adjacent classes (e.g., Point-Guard *versus* Shooting-Guard or Small-Forward *versus* both Shooting-Guard and Power-Forward). Therefore, R4 that identifies Small-Forward players is represented by the biggest node which is also placed in the center of Figure 16. In addition, we observe that rules R1 and R2 (likewise R6 and R7) are redundant (i.e., they cover the same output class) and could be merged. Therefore, the rule base can be reduced to the following five rules:

- R12. IF Height is Extremely low OR Very low THEN Player is Point-Guard
- R3. IF Height is Low THEN Player is Shooting-Guard
- R4. IF Height is Medium THEN Player is Small-Forward
- R5. IF Height is High THEN Player is Power-Forward
- R67. IF Height is Very high OR Extremely high THEN Player is Center

3.4. The Interactive Basketball Game in Scratch

Once we have understood the behavior of the selected classifiers, it is time to come back to Scratch. In the final part of the XAI4TEENS workshop, students are taught how to code their own explainable classifier in Scratch. The challenge is to enhance a pre-coded interactive game that is taken as the starting point. As it can be seen in Figure 17, the game consists in identifying the role of a basketball player in terms of its statistics.

Once the photo of one player is clicked then statistics are displayed on the right-hand side of the screen (see Figure 18). Afterward, the result of classification along with the related explanation are shown in the screen. In this illustrative example, we have selected the same player related to the data instance that was classified as a Center by the tree in Figure 11.

It is worth noting that the naturalness and effectiveness of the generated explanations depend on how well semantically grounded the selected linguistic terms are.

As we explained in the previous section, Height is the most relevant attribute and its meaning depends on the context of the problem. Building explainable classifiers is a matter of careful design. In practice, the meaning of Height is to be characterized by a fuzzy partition. Even if considering uniformly distributed fuzzy sets may look rather artificial, this way of doing is common in the literature. Therefore, instead of looking for fuzzy partitions tuned in accordance with the data distribution we opted for creating a strong fuzzy partition with 5 fuzzy sets uniformly distributed in the universe of discourse. Each fuzzy set has associated one linguistic term in the ordinal "short-tall" linguistic scale (short, medium-height, tall, very tall, extremely tall). The meaning of these linguistic terms varies in accordance with the gender of the player. More precisely, the meaning of each linguistic term is described by a triangular fuzzy set in terms of three parameters (a, b, c) such as $[(\mu(a) = 0)/\ a; (\mu(b) = 1)/\ b; (\mu(c) = 0)/\ c]$, with $\mu(x)$ the membership degree of value x, $x \in U$ (U is the universe of discourse of the attribute X), $\mu(x) = 0$ if $x < a$ or $x > c$, $\mu(x) = (b - x)/(b - a)$ if $x \in [a, b]$, $\mu(x) = 1 - (c - x)/(c - b)$ if $x \in [b, c]$. Obviously, only 2 parameters are required in case of the first and last fuzzy sets in the partition, because they have semi-trapezoidal shape.

- Height $_{WOMEN} \in [1.66, 1.94]$
  - Short = [1/1.66, 0/1.73]
  - Medium-height = [0/1.66, 1/1.73, 0/1.8]
  - Tall = [0/1.73, 1/1.8, 0/1.87]
  - Very tall = [0/1.8, 1/1.87, 0/1.94]
  - Extremely tall = [0/1.87, 1/1.94]
- Height $_{MEN} \in [1.81, 2.2]$
  - Short = [1/1.81, 0/1.908]
  - Medium-height = [0/1.81, 1/1.908, 0/2.005]
  - Tall = [0/1.908, 1/2.005, 0/2.103]
  - Very tall = [0/2.005, 1/2.103, 0/2.2]
  - Extremely tall = [0/2.103, 1/2.2]

Regarding the rest of attributes (Minutes, Points, etc.), we also considered strong fuzzy partitions but with only three linguistic terms in linguistic scales such as "low-high" or "few-many" and without making any difference between MEN and WOMEN.

The XAI4ALL Scratch code includes the decision tree associated to the J48 classifier (see Figure 19) for WOMEN and the fuzzy rules associated to the FHDT7 classifier for MEN (see Figure 20).

This Scratch code is created by means of drag and drop actions in the Scratch editor, and it is provided to the students. The *GetLinguisticTerm* block (see Figure 21) searches for the most meaningful linguistic terms to describe in NL the numerical intervals (J48) or fuzzy sets (FHDT7) appearing in the activated branch of the tree (J48) or the winner fired fuzzy rule (FHDT7), respectively. This

Figure 17 | Screenshot of the interactive game in Scratch.

Figure 18 | Example of a multi-modal explanation in Scratch.

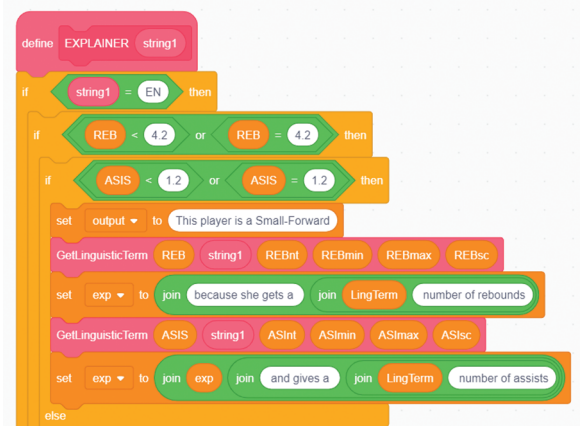

Figure 19 | Snippet of the Scratch code for J48 (WOMEN).

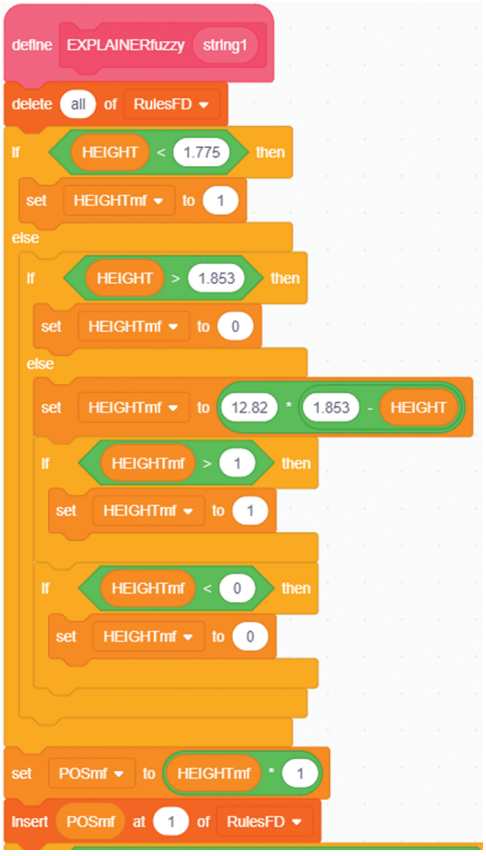

Figure 20 | Snippet of the Scratch code for FHDT7 (MEN).

Scratch block is a very naive implementation of the ExpliClas NLG module for linguistic approximation and realization.

Running the provided Scratch project, students can see in practice how a classification task, either crisp or fuzzy, is carried out and how textual explanations are generated. In addition, students are asked for editing the given code and producing an enhanced version of FHDT7. For example, as we described in the previous section, FHDT7 includes 7 partially redundant rules that can be reduced to only 5 rules; thus getting better interpretability while keeping the same accuracy. Moreover, students are invited to design a new more realistic and meaningful fuzzy partition for Height, instead of the uniform fuzzy partition that we provide them with the Scratch code.

Afterward, students experiment how the classifier taught to recognize the role of male players may fail to classify properly the role of female players, and vice versa. Finally, they face the challenge of building an explainable classifier able to recognize properly basketball players no matter their gender or skin color. With this practical work, students learn, among other things, to be aware of the need to keep in mind ethical issues when designing intelligent systems.

It is worth noting that the XAI4ALL Scratch studio project (including all the source code and pictures used to develop the XAI4TEENS workshop described in this paper) is available online.[18] In addition, the ExpliClas software is online, as well as all what is needed to reproduce the experiments described in this

[18]The XAI4ALL Scratch Project: https://scratch.mit.edu/projects/330579801/

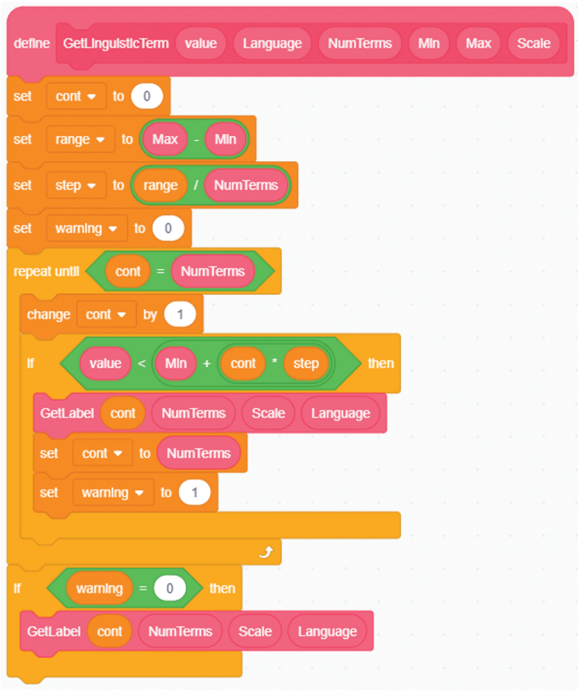

Figure 21 | The GetLinguisticTerm Scratch block.

manuscript.[19] Also, the interested reader is kindly referred to the following videos for getting further insights about how XAI4ALL works: XAI4ALL-P1,[20] XAI4ALL-P2,[21] and XAI4ALL-P3.[22]

4. CONCLUSIONS AND FUTURE WORK

This paper has presented the XAI4ALL initiative which is aimed at disseminating the fundamentals of XAI for the general public. More precisely, we have described the XAI4TEENS workshop which has the focus on high school students.

First, a brief introduction to XAI is done with an interactive story in Scratch. Second, the ExpliClas beginner module is used to teach students how to climb the pyramid from data to wisdom, i.e., how to pave the way from raw data to self-explainable classifiers. Third, students learn how to integrate an explainable classifier in an interactive game coded in Scratch. All educational resources needed to run the XAI4TEENS workshop are available online so it is ready to take place not only in person but also as a webinar in the context of e-Learning.

Even if a thorough dissertation about ethical issues in AI is out of the scope of the workshop, it is important to make students think about the potential consequences of letting AI-based systems to segment a population by gender or skin color. Actually, as described in [34], we are surrounded by many examples of gender bias that affects our everyday life. Indeed, for a long time many things (e.g., cars, workplaces, etc.) have been designed for a stereotype of white man who has a height of 1.76 m and a weight of 77 kg.

[19]https://gitlab.citius.usc.es/jose.alonso/xai

[20]XAI4ALL on Scratch (Part 1): https://youtu.be/Ri1g4D6id1I

[21]XAI4ALL on Scratch (Part 2): https://youtu.be/Ri1g4D6id1I

[22]XAI4ALL on Scratch (Part 3): https://youtu.be/zg1Y85fPJM8

In addition to XAI4ALL, other similar workshops (regarding other AI topics such as robotics, image processing, augmented reality, etc.) are developed by other colleagues in CiTIUS-USC. Moreover, once per year, all research centers in USC organized together a day for science. We are satisfied with the overall outcome and evaluation of these activities. The general public provides us with qualitative valuable feedback in order to improve the program from one year to the next. A prove of the success of this initiative is the fact that every year more schools demand to take part.

We don't ask either students or teachers to fill in formal surveys about their satisfaction with each workshop, but we receive valuable informal qualitative feedback from the teachers after each session. For example, XAI4ALL was initially conceived as XAI4KIDS but after talking with school teachers we identified the need to adapt it for teenagers and therefore we created XAI4TEENS. Accordingly, XAI4KIDS evolved to offer kids Scratch games to play while XAI4TEENS offers students the opportunity to develop small Scratch programs by themselves.

As future work, we plan to adopt a *fabula* model as well as argumentation schemes in order to make explanations even more natural, narrative and persuasive. In addition, explanations will be enriched with causal relations and counterfactual facts. Finally, we will explore the chance of running the XAI4ALL project with a physical/virtual embodiment robotic platform.

CONFLICT OF INTEREST

There are no "conflict of interest" to report.

AUTHORS' CONTRIBUTIONS

This manuscript is the author's original work. The content of this manuscript or a major portion thereof has not been copyrighted, published, accepted for publication or submitted simultaneously elsewhere. This is an **EUSFLAT 2019 post-conference paper** which has been extended with more than 40% of new content as required.

ACKNOWLEDGMENTS

Jose M. Alonso is a *Ramón y Cajal* Researcher (RYC-2016-19802). This research is partially supported by the Spanish Ministry of Science, Innovation and Universities (grants RTI2018-099646-B-I00, TIN2017-84796-C2-1-R, TIN2017-90773-REDT, RED2018-102641-T), the Galician Ministry of Education, University and Professional Training (grants ED431F2018/02, ED431C2018/29, ED431G2019/04). Some of the previous grants were co-funded by the European Regional Development Fund (ERDF/FEDER program).

REFERENCES

[1] European Commission, Artificial Intelligence for Europe, Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions, Technical Report, Brussels, Belgium, 2018. https://ec.europa.eu/digital-single-market/en/news/communication-artificial-intelligence-europe

[2] Parliament and Council of the European Union, General Data Protection Regulation (GDPR), 2016. http://data.europa.eu/eli/reg/2016/679/oj

[3] EU High-Level Expert Group on Artificial Intelligence, Ethics guidelines for trustworthy AI, 2019. https://ec.europa.eu/digital-single-market/en/news/draft-ethics-guidelines-trustworthy-ai

[4] T. Miller, Explanation in artificial intelligence: insights from social sciences, Artif. Intell. 267 (2019), 1–38.

[5] N. Tintarev, J. Masthoff, Explaining recommendations: design and evaluation, in: F. Ricci, L. Rokach, B. Shapira (Eds.), Recommender Systems Handbook, Springer, Boston, MA, USA, 2015, pp. 353–382.

[6] A. Abdul, J. Vermeulen, D. Wang, B.Y. Lim, M. Kankanhalli, Trends and trajectories for explainable, accountable and intelligible systems: an HCI research agenda, in CHI Conference on Human Factors in Computing Systems, Montreal, Canada, 2018, vol. 582, pp. 1–18.

[7] O. Biran, C. Cotton, Explanation and justification in machine learning: a survey, in IJCAI Workshop on Explainable AI, Melbourne, Australia, August 19-25, 2017, pp. 8–13.

[8] L.A. Zadeh, Fuzzy sets, Inf. Control. 8 (1965), 338–353.

[9] L.A. Zadeh, The concept of a linguistic variable and its application to approximate reasoning I, Inf. Sci. 8 (1975), 199–250.

[10] L.A. Zadeh, From computing with numbers to computing with words - from manipulation of measurements to manipulation of perceptions, IEEE Trans. Circuits Syst. I. Fundam. Theory Appl. 46 (1999), 105–119.

[11] L.A. Zadeh, A new direction in AI: toward a computational theory of perceptions, Artif. Intell. Mag. 22 (2001), 73–84.

[12] J.M. Alonso, From Zadeh's computing with words towards explainable artificial intelligence, in: R. Fuller, S. Giove, F. Masulli (Eds.), WILF2018 - 12th International Workshop on Fuzzy Logic and Applications, Springer, Cham, Switzerland, 2019, pp. 244–248.

[13] J.M. Alonso, C. Castiello, C. Mencar, Interpretability of fuzzy systems: current research trends and prospects, in: J. Kacprzyk, W. Pedrycz (Eds.), Handbook of Computational Intelligence, Springer, Berlin, Heidelberg, Germany, 2015, pp. 219–237.

[14] J.M. Alonso, C. Castiello, C. Mencar, A bibliometric analysis of the explainable artificial intelligence research field, in 17th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU), Cádiz, Spain, 2018, vol. CCIS853, pp. 3–15.

[15] J.M. Alonso, Explainable artificial intelligence for kids, in 11th Conference of the European Society for Fuzzy Logic and Technology (EUSFLAT), Prague, Czech Republic, September 9-13, Atlantis Press, 2019, pp. 134–141.

[16] J.M. Alonso, A. Bugarin, ExpliClas: automatic generation of explanations in natural language for Weka classifiers, in IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), New Orleans, LA, USA, 2019, pp. 1–6.

[17] J.M. Alonso, P. Ducange, R. Pecori, R. Vilas, Building explanations for fuzzy decision trees with the ExpliClas software, in IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Glasgow, Scotland, UK, July 19-24, 2020, pp. 1–8. https://2020.wcci-virtual.org/presentation/poster/building-explanations-fuzzy-decision-trees-expliclas-software

[18] R.L. Ackoff, From data to wisdom, J. Appl. Syst. Anal. 16 (1989), 3–9.

[19] J. Rowley, The wisdom hierarchy: representations of the DIKW hierarchy, J. Inf. Commun. Sci. 33 (2007), 163–180.
[20] I.H. Witten, E. Frank, M.A. Hall, C.J. Pal, Data Mining: Practical Machine Learning Tools and Techniques, Elsevier, 2016.
[21] J.R. Quinlan, Induction of decision trees, Mach. Learn. 1 (1986), 81–106.
[22] J.R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers, San Mateo, CA, USA, 1993.
[23] M. Fernández-Delgado, E. Cernadas, S. Barro, D. Amorim, Do we need hundreds of classifiers to solve real world classification problems?, J. Mach. Learn. Res. 15 (2014), 3133–3181.
[24] R. Pecori, P. Ducange, F. Marcelloni, Incremental learning of fuzzy decision trees for streaming data classification, in 11th Conference of the European Society for Fuzzy Logic and Technology (EUSFLAT), Prague, Czech Republic, September 9-13, Atlantis Press, 2019, pp. 748–755.
[25] J. Hühn, E. Hüllermeier, FURIA: an algorithm for unordered fuzzy rule induction, Data Mining Knowl. Discov. 19 (2009), 293–319.
[26] D. Pancho, J.M. Alonso, L. Magdalena, Enhancing fingrams to deal with precise fuzzy systems, Fuzzy Sets Syst. 297 (2016), 1–25.
[27] E. Reiter, R. Dale, Building Natural Language Generation Systems, Cambridge University Press, Cambridge, UK, 2000.
[28] A. Gatt, E. Krahmer, Survey of the state of the art in natural language generation: core tasks, applications and evaluation, J. Artif. Intell. Res. 61 (2018), 65–170.
[29] A. Gatt, E. Reiter, SimpleNLG: a realisation engine for practical applications, in European Workshop on Natural Language Generation (ENLG), Athens, Greece, 2009, pp. 90–93.
[30] A. Ramos-Soto, J. Janeiro-Gallardo, A. Bugarin, Adapting SimpleNLG to spanish, in 10th International Conference on Natural Language Generation, ACL, Santiago de Compostela, Spain, 2017, pp. 142–146.
[31] A. Cascallar-Fuentes, A. Ramos-Soto, A. Bugarin, Adapting SimpleNLG to galician language, in 11th International Conference on Natural Language Generation, ACL, Tilburg, The Netherlands, 2018, pp. 67–72.
[32] H. Gardner, Frames of Mind: The Theory of Multiple Intelligences, Basic Books, New York, USA, 1983.
[33] L.A. Zadeh, Outline of a new approach to the analysis of complex systems and decision processes, IEEE Trans. Syst. Man Cybern. SMC-3 (1973), 28–44.
[34] C. Criado-Perez, Invisible Women: Exposing Data Bias in a World Designed for Men, Chatto & Windus, London, UK, 2019.
[35] D.H. Wolpert, W.G. Macready, No free lunch theorems for optimization, IEEE Trans. Evol. Comput. 1 (1997), 67–82.