IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 27 February 2024, accepted 5 March 2024, date of publication 11 March 2024, date of current version 22 March 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3375764

RESEARCH ARTICLE

A New Data Science Model With Supervised Learning and Its Application on Pesticide Poisoning Diagnosis in Rural Workers

JAQUELINE C. S. CARVALHO[1,2], TALES C. PIMENTA[1], (Senior Member, IEEE), ALESSANDRA C. P. SILVERIO[2], MARCOS A. CARVALHO[1,2], AND JOAO PAULO C. S. CARVALHO[3]

[1]Institute of Systems Engineering and Information Technology, Federal University of Itajubá, Itajubá 37500-903, Brazil
[2]Department of Computer Science, José do Rosário Vellano University, Alfenas 37130-000, Brazil
[3]Mathematics and Natural Sciences Division, Brescia University, Owensboro, KY 42301, USA

Corresponding authors: Jaqueline C. S. Carvalho (jaqueline.carvalho@unifenas.br) and Tales C. Pimenta (tales@unifei.edu.br)

This work was supported in part by CAPES, in part by CNPq, and in part by FAPEMIG.

ABSTRACT In a Data Science project, it is essential to determine the relevance of the data and identify patterns that contribute to decision–making based on domain–specific knowledge. Furthermore, a clear definition of methodologies and creation of documentation to guide a project's development from inception to completion are essential elements. This study presents a Data Science model designed to guide the process, covering data collection through training with the aim of facilitating knowledge discovery. Motivated by deficiencies in existing Data Science methodologies, particularly the lack of practical step–by–step guidance on how to prepare data to reach the production phase. Named "Data Refinement Cycle with Supervised Machine Learning (DRC–SML)", the proposed model was developed based on the emerging needs of a Data Sciense project aimed at assisting healthcare professionals in diagnosing pesticide poisoning among rural workers. The dataset used in this project resulted from scientific research in which 1027 samples were collected, containing data related to toxicity biomarkers and clinical analyses. We achieved an accuracy of 99.61% with only 27 rules for determining the diagnosis. The results optimized healthcare practices and improved quality of life in rural areas. The project outcomes demonstrated the success of the proposed model.

INDEX TERMS Data science, decision support system, machine learning, pesticide poisoning diagnosis.

I. INTRODUCTION

Data Science has significantly transcended its origins in traditional Statistics. A striking indicator of this evolution is the exponential growth in the amount of data globally generated and stored. According to Cremin et al. [1], this volume reached approximately 44 zettabytes in early 2020, and it is projected that by 2025, the daily amount of data generated will reach 463 exabytes on a global scale. This massive collection of data is commonly referred to as "big data" and is characterized by a substantial volume and a wide variety of data types.

Despite the remarkable growth in the field of Data Science, the successful execution of projects in this domain continues to pose significant challenges. According to information from Saltz and Krasteva [2], approximately 87% of Data Science projects fail to reach the production phase.

As a multidisciplinary field, Data Science has applications in various areas of interest. Collaboration with domain experts who can deeply understand the issues at hand is crucial throughout the process to achieve the proposed results. Data scientists must possess comprehensive skills in Knowledge Discovery in Databases (KDD) that presupposes knowledge in statistics, computer science, databases, and machine learning [3], [4].

Data Science models often follow cyclical structures known as the data lifecycle, which serves as a guide for data

The associate editor coordinating the review of this manuscript and approving it for publication was Asad Waqar Malik.

scientists throughout the KDD process. A recurring challenge in this context is the lack of interpretability in complex models as well as the presence of low–quality or noisy data, which can compromise the effectiveness and reliability of the models [3].

As highlighted by Jain and Kushagra [5], the quality of a developed model is intrinsically linked to the data provided. This involves identifying relevant data, integrating datasets, cleaning data, creating new data, and extracting new features from existing data. In summary, data preparation is the most time-consuming and possibly the most critical step in the lifecycle of a Data Science project.

De Bie et al. [6] found that machine learning methods play a predominant role in a data scientist's toolbox. These methods have gained prominence over the last two decades, spanning from relatively simple techniques to complex approaches, such as deep learning. However, it is essential to emphasize that such methods often assume the availability of substantial volumes of high–quality data, which, in practice, presents additional challenges.

Machine learning can be categorized into three main types: supervised, unsupervised, and reinforcement learning. In supervised learning, data are labeled by experts, and examples are described by a dataset and an associated class label. The central goal is to build a classifier based on these examples, allowing the machine to classify new unlabeled examples [7].

Rough Set Theory (RST), proposed by Pawlak [8] and reviewed by Achariva and Abraham [9], represents a valuable mathematical tool for managing a specific type of uncertainty and imprecision. RST, whether adopted alone or in conjunction with other machine learning models, has demonstrated its effectiveness in solving real–world machine learning problems.

A Data Science project is centered on data that are usually embedded in a specific context. Many problems, such as financial analysis, marketing, and healthcare analytics, have benefited from Data Science projects [10].

This study addresses a public health problem that has not yet been explored in Data Science. According to Peña-Fernández et al. [11], pesticide poisoning in rural workers is a matter of great concern, and results in significant social and economic losses worldwide. This is due to of the lack of standardized tests for biological diagnosis and the shortage of trained healthcare professionals to deal with such cases.

However, we noticed when starting this study that no existing Data Science model provided practical step–by–step guidance on how to prepare data to reach the production phase. Existing models do not offer streamlined resources for data preparation, including the individual analysis of each piece of information, exclusion of irrelevant data, data transformation, creation of new data, and selection of training and testing sets.

Thus, the initial goal of assisting in the diagnosis of a public health problem evolved into a new purpose: to develop a practical Data Science model capable of addressing any supervised machine learning problem. The motivation for the proposal of the model lies in the pressing need for practical and targeted approaches to data preparation, filling a crucial gap in Data Science project management. This model not only aims to address the identified deficiencies but also to boost effectiveness and transparency in Data Science projects, providing a systematic and transparent approach throughout the project's lifecycle. The diagnosis of pesticide poisoning in rural workers became an example used to demonstrate this model.

Therefore, this article presents two innovative contributions:

1) A new Data Science model called "Data Refinement Cycle with Supervised Machine Learning (DRC–SML)". The uniqueness of this model lies in its ability to meticulously analyze and monitor each piece of information present in the dataset through standardized forms. This provided resources for categorizing data with multiple values or missing values. Additionally, the model suggests documenting each step of the process, including all training sessions conducted. This simplifies progress verification and task tracking, allowing for a retrospective analysis of documented training sessions and promoting a more transparent and traceable approach.
2) A Data Science project called "Planting and Harvesting Health (PHH)", which aims to assist in the diagnosis of pesticide poisoning in rural workers, uses the DRC–SML model and RST as a machine learning tool. This project not only contributes to healthcare professionals but also serves as a scenario for the creation of the new Data Science model DRC–SML, which was developed and applied throughout this research.

The rest of this article is organized as follows: Section II reviews related research on the challenges encountered in Data Science projects and the applications of supervised machine learning. Section III introduces and demonstrates the proposed DRC–SML model in the context of a PHH project. In Section IV, an analysis of the developed model and the final results of the project are presented. Finally, Section V summarizes the study and suggests directions for future research.

II. RELATED WORKS

A. DATA SCIENCE SCENARIO

The field of Data Science is intrinsically multidisciplinary, combining elements of computer science, mathematics, and statistics to extract knowledge and value from large volumes of data [12]. A survey conducted by [13] involving industry professionals and nonprofit organizations revealed that 85% of the respondents believed that the adoption of more refined and consistent processes can improve the effectiveness of Data Science projects.

Various authors and companies have proposed approaches to Data Science project management, introducing new tools and processes to address these issues. Some of the most recognized models include the Agile Data Science Lifecycle [14], Cross–Industry Standard Process for Data Mining (CRISP–DM) [15], Microsoft Team Data Science Process (Microsoft–TDSP) [16], and Domino Data Science Lifecycle [17]. A detailed critical analysis of Data Science methodologies was presented in [18].

Martinez et al. [17] presented empirical data obtained from a survey of 237 Data Science professionals, highlighting the predominance of the Agile Data Science Lifecycle over the traditional CRISP–DM methodology. However, only 25% of the participants claimed to follow a specific methodology, underscoring the lack of a clearly defined model for Data Science project management, which has been identified as one of the main challenges in this field.

Gm et al. [19] and Sarker [10] addressed the importance of fundamental data science operations, such as data cleaning, data processing, data modeling, data visualization, and data presentation techniques. However, they did not provide practical guidance on their implementation.

Yao [20] presented a conceptual Data Science model that addresses perspectives from information science, management science, cognitive science, and computer science, but also does not establish specific guidelines for documenting and implementing the conceptual model.

B. APPLICATIONS OF SUPERVISED MACHINE LEARNING

In the domain of Machine Learning, Laturiuw and Singgalen [21] conducted a comprehensive comparison of algorithms such as Support Vector Machines (SVM), Naive Bayes (NB), Decision Trees (DT), and k–Nearest Neighbors (k–NN) algorithm, identifying that DT and SVM, with the assistance of SMOTE [22] operators to handle imbalanced data, demonstrated superior performance in their study.

The article by [23] reviews various Machine Learning algorithms, including DT, SVM, k–NN, and NB, highlighting the superior performance of SVM in terms of accuracy. It should be noted that the performance of these algorithms highly depends on the domain and specific dataset, which can vary in different contexts.

In a comprehensive machine learning review conducted by Velayutham et al. [24], algorithms including logistic regression, KNN, decision tree, and SVM were evaluated on datasets related to breast cancer, heart disease, iris, and wine quality. The results were consolidated, and the algorithms were ranked based on their performance, as presented in Table 1. Logistic regression emerged as the top–performing algorithm overall, especially in heart disease, iris, and wine quality. However, its drawback became evident in suboptimal performance on the breast cancer dataset. Logistic regression demonstrated advantages in handling both two–class and multiclass datasets, performing well in both low and high–dimensional environments. The decision tree secured the second position overall, excelling in breast cancer and adapting to binary classification and multidimensional environments. SVM ranked third, exhibiting superior performance in breast cancer, while KNN and Naïve Bayes followed suit, standing out in diverse environments and contexts.

TABLE 1. Comparison and Ranking of Algorithm – Data Centric [24].

Algorithm/ Dataset	Breast cancer	Heart disease	Iris	Wine quality	Total	Ranking
Logistic Regression	4	5	5	5	19	1
KNN	3	3	4	4	14	4
Decision Tree	4	4	4	4	17	2
SVM	5	3	4	4	16	3
Naïve Bayes	2	3	4	4	13	5

Saravanan and Charan [25] used Convolutional Neural Networks (CNN) and SVM as Machine Learning tools in a dataset for human activity recognition, achieving accuracy rates of 92.46% and 90.19%, respectively. These results indicate the superiority of the CNN classifier for the analysis of human activities.

However, it is important to note that Rough Set Theory (RST) is rarely mentioned in reviews of Machine Learning algorithms, despite several studies establishing a solid foundation for its application in Machine Learning problems. Singh and Yao [26] emphasized that previous research addressed pneumonia identification using Machine Learning algorithms, particularly Deep Learning, in a binary classification approach. However, they introduced RST during the machine learning process to mitigate uncertainty, achieving a remarkable accuracy rate of 96.25% in pneumonia detection.

Nayak et al. [27] conducted a study using the RST to reveal latent information in imprecise datasets, focusing on the accurate identification of malaria symptoms as conditional attributes.

The potential of RST in the multidisciplinary field of Machine Learning, involving the selection of relevant attributes and decision–making, has been demonstrated in [28], [29], [30], [31], [32], [33], and [34].

C. CONCEPTS OF GRANULAR COMPUTING

Granular Computing, an innovative approach in computational intelligence, addresses the inherent complexity and imprecision of real–world systems. Grounded in the concept of granularity, where information is organized hierarchically, researchers explore methodologies for decision system approximation, particularly in managing large–scale datasets.

In the referenced study [35], Cybulski and Artiemjew investigates the use of random sampling to approximate decision systems, emphasizing its application to Big Data challenges. The paper advocates for concept–dependent granulation as a reference method, with experiments on real–world data revealing consistent insights into effective random sampling for rapid decision system size reduction.

Another relevant contribution [36] explores the competitiveness of granulating training data into subgroups with post–granulation object fusion. The study evaluates the efficiency, in terms of runtime and classification, comparing this granulation approach with employing a complete training system. Findings offer valuable insights into trade–offs and advantages of different granulation strategies in training data contexts.

In this article, the granulation of training data was performed by forming training subgroups based on the collaboration and interests of the domain expert.

III. PROPOSED MODEL

A. STUDY DATA

In the scope of the PHH project, we utilized data, which originated from a scientific research study in the field of Biomedicine. This study was approved by the Research Ethics Committee of Jose do Rosario Vellano University in collaboration with the Federal University of Alfenas (protocol numbers 149718 and 415856). In this scientific research in Biomedicine, 1,027 samples were collected from rural workers in southern Minas Gerais, Brazil. Each sample contained 121 attributes, encompassing socioeconomic information, a history of pesticide poisoning and hospitalization, in addition to the use of personal protective equipment. Blood samples were also collected for the evaluation of pesticide exposure biomarkers as well as the detection of possible renal and hepatic sequelae. All the attributes of this dataset are described in the data repository available in [37].

It is important to emphasize that after receiving information, all volunteers provided their consent to participate in the study by signing an Informed Consent Form [38].

B. DRC-SML MODEL IN THE PHH PROJECT

The proposed model comprehensively encompasses all the stages of the KDD process. The general framework of the model is presented in Figure 1(a), in which we delineate the sequence of steps that constitute this process. In this figure, the pipeline of the model is observable, illustrating distinct stages and their sequencing. In summary, the model adheres to a cyclical approach that incorporates data collection, exclusion of irrelevant data, data storage, and continuous data refinement, with the capability of eliminating extraneous information and incorporating new data as needed. Subsequently, during the training process, insights may emerge indicating the necessity for additional refinements. At the end of the cycle, the model imparts knowledge to facilitate decision–making. In the subsequent phase, data retrieval was based on the outcomes of the refinement stage, enabling the execution of fresh training processes.

This model is underpinned by the fundamental pillars of privacy, integration, quality, respect for copyrights, and data preservation. These pillars are illustrated in Figure 1(b) and guide the stages of the model to ensure its performance and integrity.

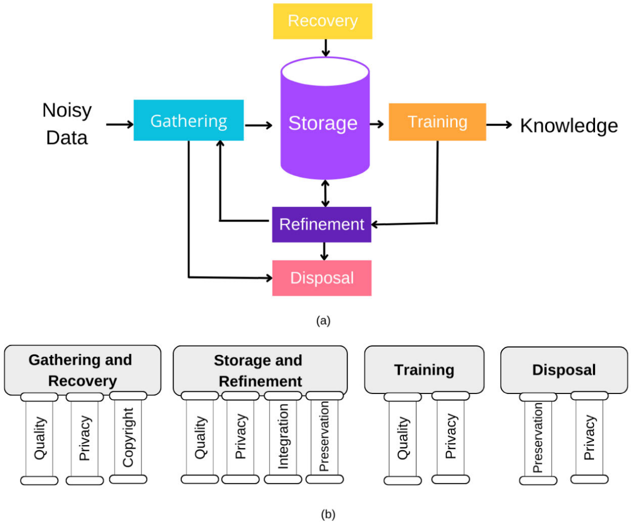


FIGURE 1. (a) Model DRC–SML (b) Model pillars DRC–SML.

Next, we detail all stages of the proposed model, addressing the pillars to be considered in each stage, as well as the application of each stage in the context of the PHH project, whose goal is medical diagnosis assistance.

1) GATHERING

Data acquisition can occur through data research (manual collection) or from existing data sources (digital collection). Regardless of the chosen method, the following pillars must be considered.

- Privacy: In manual collection, access to data should be limited to individuals responsible for the collection who should possess appropriate technical training. In the context of digital collection, it is essential to define access levels to prevent breaches of confidentiality and ensure compliance with laws regulating the protection of digital data.
- Copyright: It is essential to obtain documented authorization from the data owner for use in accordance with the ethical guidelines and current legislation.
- Quality: Planning is crucial in manual collection, particularly when data do not yet exist and must be collected. This may involve creating standardized questions and answers to ensure data collection quality and avoiding multiple responses to the same question. In the case of digital collection, which is based on existing data, it is essential to identify the repositories and determine whether the data are properly standardized.

In the PHH project, digital data were acquired using pre–existing information provided by a domain expert and recorded in a spreadsheet format. It is crucial to highlight that a significant portion of these data exhibited levels of noise, with only a few data points characterized by the presence of consistent response patterns.

2) DISPOSAL

Irrelevant information can potentially divert and decrease computational efficiency. Therefore, when certain elements are identified as irrelevant in the dataset used to support decision–making, their exclusion from the set becomes essential. This information should be archived in a separate repository to ensure privacy and to preserve data for possible future use.

3) STORAGE

The proposed model utilizes a relational database model for data storage. In this paradigm, data are structured into subsets known as tables, which establish connections between them based on their own data.

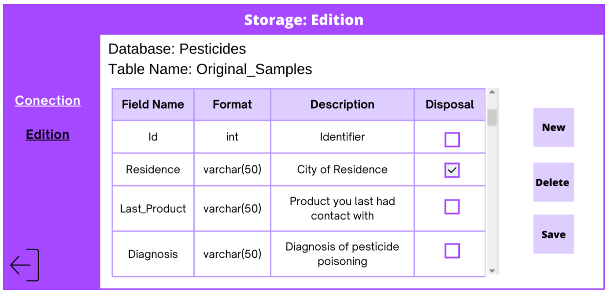
Relational Database Management Systems (RDBMS) play a crucial role in this context, as they already provide support for the four main pillars outlined below:

- Privacy: RDBMS enable multiple users to access data securely by implementing passwords and access restrictions.
- Integration: Data integration is achieved through relationships established between tables within the RDBMS. Furthermore, it is possible to integrate it with other storage models easily, owing to the ability to import and export files in text format.
- Quality: The relational database model follows standards to ensure data consistency and quality. The RDBMS uses Structured Query Language (SQL) for data querying and maintenance, which ensures quick, secure, and high-quality information retrieval.
- Preservation: The RDBMS incorporates a metadata file that provides descriptive information regarding the data. This feature enables effective maintenance over time, thereby facilitating data understanding and preservation.

Therefore, the DRC–SML approach employs the relational database model in conjunction with an RDBMS to ensure the security, integration, quality, and preservation of stored data, enabling analyses of the information contained in the repository. This analysis allows for the identification of irrelevant data that can be subsequently discarded and, when necessary, the creation of data for missing but essential information for the research at hand.

In the PHH project, the data obtained during the collection stage were stored in a spreadsheet. However, in this storage format, the data lack interaction between them, making it difficult to create visualizations that allow for data analysis in different ways. To fulfill the objectives of the storage stage, the content of the spreadsheet was transferred to a relational database. Subsequently, it was possible to delete information considered irrelevant to the diagnostic process, such as address details and medical service evaluations. Additionally, essential elements were incorporated, namely, an "ID" used for the unique identification of each sample and "Diagnosis" employed for the categorization and labeling of each sample.

Figure 2 shows the form recommended by the DRC–SML model for documenting the storage stage and showcasing data related to the PHH project. This form provides the names of fields designated for the storage of each piece of information within the chosen database, with the purpose of enabling the data scientist to input textual descriptions for each field, mark fields for deletion, and add new fields as necessary.



Storage: Edition

Database: Pesticides
Table Name: Original_Samples

Field Name	Format	Description	Disposal
Id	int	Identifier	□
Residence	varchar(50)	City of Residence	☑
Last_Product	varchar(50)	Product you last had contact with	□
Diagnosis	varchar(50)	Diagnosis of pesticide poisoning	□

FIGURE 2. Example of Storage Form.

4) REFINEMENT

In Data Science, the data refinement process is a critical step involving essential procedures, such as handling missing values, standardization, elimination of irrelevant information, and transformation when necessary. To maintain data preservation and quality over time, the DRC–SML model proposes the use of a carefully crafted form to document the refinement process of each stored data while data privacy and integration are maintained by the RDBMS. This form covers crucial information such as the nature of the data (continuous, discretized, multivalued, or null) and the data status during the refinement process, indicating whether it is completed or should be discarded.

Notably the nature of the data significantly influences our refinement strategy. For continuous data, it is imperative to establish discretization rules to create defined intervals, making them suitable for subsequent analyses. For discretized data, it is essential to document the previously applied rules, enabling the understanding and replication of the process. Lastly, when dealing with null or multivalued data, it is advisable to assess the feasibility of returning to the collection stage, and in the case of multivalued data, we suggest disaggregating the information into individual units and assigning a new classification to each singular piece of information. This resultes in a more transparent and effective structure for further analysis.

At this stage, the collaboration of the domain expert plays a crucial role, indicating which data can be discarded, and guiding the refinement according to their knowledge, ensuring proper data representation.

Figure 3(a) presents the refinement form of the proposed model applied to PHH project data. In this context, "Creatinine" is highlighted as an example of continuous data that underwent discretization, as illustrated in Figure 3(b).

Furthermore, as an example of multivalued data in the PHH project, we identified information related to the last product with which the rural worker had contact. Respondents provided a list of products, such as "Polyram, Supera, Bitrin," in a comma–separated text format, indicating the brand names of these products. However, for analysis

TABLE 2. Types of Diagnostics, specifications and corresponding acronyms.

Type	Specification	Diagnosis
Intoxication Acute	CH_P: With Decreased Activity	IA–CHPD
	CH_P: With Decreased Activity High Concentrations of Pesticides	IA–HP–CHPD
	CH_P: With Decreased Activity Low Concentrations of Pesticides	IA–LP–CHPD
	CH_P and CH_E: With Decreased Activity Altered Creatinine (Kidney Problem)High Concentrations of Pesticides	IA–HP–CHPD–CHED–KPCR
	CH_P: With Decreased Activity Altered Ast and Alt (Liver Injury)	IA–CHPD–ASTALT–LI
Intoxication Chronic	CH_E: With Decreased Activity	IC–CHED
	CH_E: With Decreased Activity Altered Ast (Liver Injury)	IC–CHED–AST–LI
	No intoxication	NI
	No diagnosis, sample missing information	ND

purposes, we sought the chemical categorization of these products rather than their brand names. We initiated new data collection, introducing specific categories, as shown in Figure 3(c), to represent the chemical classification. Each response was individually evaluated and associated with the respective chemical classification, as presented in Figure 3(d).

However, even after assigning chemical classification to the products, some responses remained multivalued. An example of the response "Polyram, Supera, and Bitrin" was one such case, where the corresponding chemical classes were defined as "Carbamate, Organophosphorus, Triazole". In this context, it was up to the domain expert to select the most relevant chemical class for the question, as illustrated in Figure 3(e).

Another relevant example, as shown in Figure 3(a), pertains to information related to the diagnosis associated with sample labeling in the PHH project. Initially, this information was absent. Table 2 presents nine diagnostic categories, with five for acute poisoning cases, two for chronic poisoning, one for the absence of poisoning, and one for reporting the lack of necessary information for diagnosis. Domain experts defined the categories and applied them to diagnosis. The proposed model enabled the selection of the necessary data for the assignment of diagnosis related to each sample, resulting in the generation of two forms, due to the complexity of diagnostic assignment criteria, as illustrated in Figure 4.

5) TRAINING

During this stage, quality was assessed based on the satisfaction of the results obtained during training in collaboration with the domain expert. This stage corresponds to the machine learning process, during which multiple training sessions are conducted, and the results of these iterations are analyzed. Returning to the refinement stage is often required to improve the training results. Therefore, the DRC–SML model proposes keeping record of the training conducted, assisting the data scientist in analyzing the iterations that produced the best results.

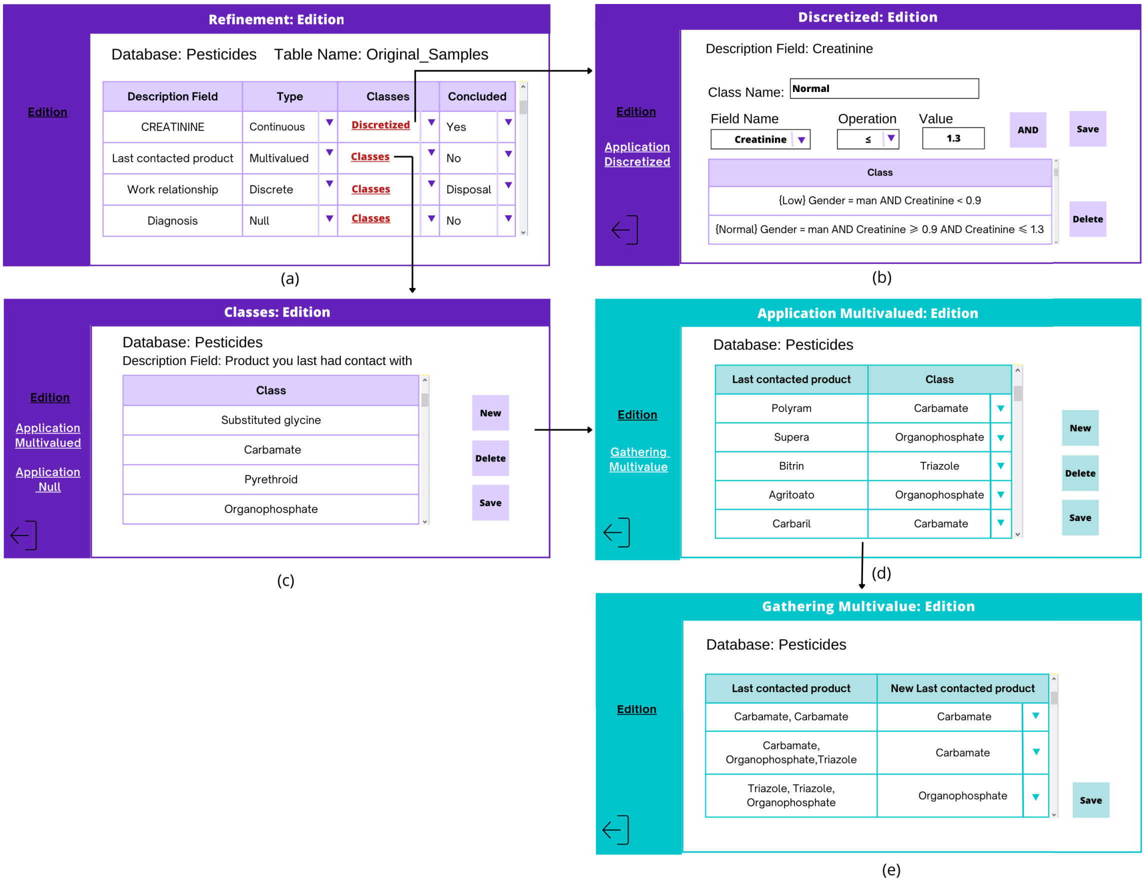
The DRC–SML model was initially designed to operate in conjunction with the RST algorithm as a machine learning tool. Future work involving the evaluation of different machine learning techniques, as presented in Section II, can be employed to enhance the performance of the proposed model. RST–based models are known to produce clear and interpretable decision rules, simplifying the analysis and validation of rules by domain experts and thereby enhancing transparency in the decision–making process. The RST algorithm is recognized for its ability to handle imperfect, uncertain, or noisy data, making it effective in identifying relevant data — common scenarios encountered in real–world applications, such as medical diagnostics, as observed in the context of the PHH project. In this scenario, uncertainty can be observed in samples that present attributes unnecessary for the diagnosis, null values, or of the type "Not Informed," in addition to the variety of types, with 16% being continuous attributes and 11% being multivalued attributes.

The RST is a supervised machine learning model that requires an appropriate definition of condition attributes and decision attributes. Training using the RST algorithm involves two distinct phases: generating reducts and extracting rules. In the reduct generation phase, a search is conducted for subsets of attributes that have the same equivalence relation and decision–making power as the original set of attributes submitted for training. The ability to make decisions with a reduced number of attributes is valuable, especially in the absence of certain information, where we have the opportunity to choose subsets in which the information is present, in order to obtain accurate diagnoses with a more concise set of data. The exclusion of irrelevant attributes is essential for simplifying the training process and generating consistent decision rules. This optimized approach contributes to the effectiveness of the analysis, even with less extensive datasets, resulting in more concise rules and equally robust decisions.

It is important to document the training conducted in the reduct generation, providing the following information:

- Data source and the number of examples or samples used in training.
- Set of attributes selected as the condition set and decision attribute. The condition attribute set can consist of all condition data or be partial, consisting of subsets of the condition attributes.

The DRC–SML model adopts the Cross–Validation (CV) technique, specifically employing k–fold cross–validation [39], allowing for the definition of the number k of partitions and the volume of samples in each partition. The dataset is divided into k subsets. The model is then trained k times, with each iteration using a different subset as the validation set, while the remaining subsets collectively form the training set. This iterative process provides a more robust evaluation of the model's performance, ensuring that each



Refinement: Edition

Edition

Database: Pesticides Table Name: Original_Samples

Description Field	Type	Classes	Concluded
CREATININE	Continuous	Discretized	Yes
Last contacted product	Multivalued	Classes	No
Work relationship	Discrete	Classes	Disposal
Diagnosis	Null	Classes	No

(a)

Discretized: Edition

Edition

Application Discretized

Description Field: Creatinine

Class Name: Normal

Field Name: Creatinine — Operation: ≤ — Value: 1.3 — AND — Save

Class
{Low} Gender = man AND Creatinine < 0.9
{Normal} Gender = man AND Creatinine ⩾ 0.9 AND Creatinine ⩽ 1.3

Delete

(b)

Classes: Edition

Edition

Application Multivalued

Application Null

Database: Pesticides

Description Field: Product you last had contact with

Class
Substituted glycine
Carbamate
Pyrethroid
Organophosphate

New

Delete

Save

(c)

Application Multivalued: Edition

Edition

Gathering Multivalue

Database: Pesticides

Last contacted product	Class
Polyram	Carbamate
Supera	Organophosphate
Bitrin	Triazole
Agritoato	Organophosphate
Carbaril	Carbamate

New

Delete

Save

(d)

Gathering Multivalue: Edition

Edition

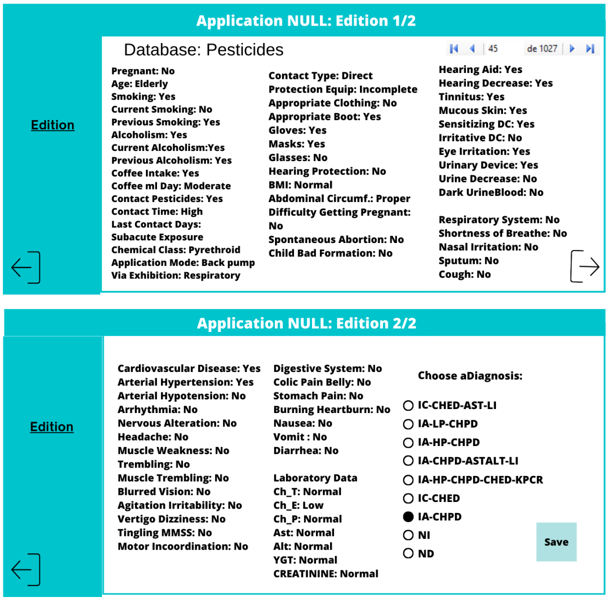
Database: Pesticides

Last contacted product	New Last contacted product
Carbamate, Carbamate	Carbamate
Carbamate, Organophosphate,Triazole	Carbamate
Triazole, Triazole, Organophosphate	Organophosphate

Save

(e)

FIGURE 3. (a)Example of a data refinement form. (b)Example of a form for discretizing continuous data. (c)Example of a form for defining categories. (d)Example of a form for collecting multivalued data. (e)Example of a form for selecting multivalued data.



Application NULL: Edition 1/2

Edition

Database: Pesticides

45 de 1027

Pregnant: No
Age: Elderly
Smoking: Yes
Current Smoking: No
Previous Smoking: Yes
Alcoholism: Yes
Current Alcoholism:Yes
Previous Alcoholism: Yes
Coffee Intake: Yes
Coffee ml Day: Moderate
Contact Pesticides: Yes
Contact Time: High
Last Contact Days:
Subacute Exposure
Chemical Class: Pyrethroid
Application Mode: Back pump
Via Exhibition: Respiratory

Contact Type: Direct
Protection Equip: Incomplete
Appropriate Clothing: No
Appropriate Boot: Yes
Gloves: Yes
Masks: Yes
Glasses: No
Hearing Protection: No
BMI: Normal
Abdominal Circumf.: Proper
Difficulty Getting Pregnant: No
Spontaneous Abortion: No
Child Bad Formation: No

Hearing Aid: Yes
Hearing Decrease: Yes
Tinnitus: Yes
Mucous Skin: Yes
Sensitizing DC: Yes
Irritative DC: No
Eye Irritation: Yes
Urinary Device: Yes
Urine Decrease: No
Dark UrineBlood: No

Respiratory System: No
Shortness of Breathe: No
Nasal Irritation: No
Sputum: No
Cough: No

Application NULL: Edition 2/2

Edition

Cardiovascular Disease: Yes
Arterial Hypertension: Yes
Arterial Hypotension: No
Arrhythmia: No
Nervous Alteration: No
Headache: No
Muscle Weakness: No
Trembling: No
Muscle Trembling: No
Blurred Vision: No
Agitation Irritability: No
Vertigo Dizziness: No
Tingling MMSS: No
Motor Incoordination: No

Digestive System: No
Colic Pain Belly: No
Stomach Pain: No
Burning Heartburn: No
Nausea: No
Vomit : No
Diarrhea: No

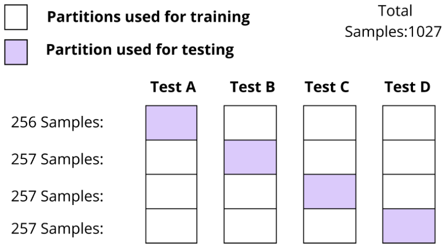
Laboratory Data
Ch_T: Normal
Ch_E: Low
Ch_P: Normal
Ast: Normal
Alt: Normal
YGT: Normal
CREATININE: Normal

Choose aDiagnosis:

○ IC-CHED-AST-LI
○ IA-LP-CHPD
○ IA-HP-CHPD
○ IA-CHPD-ASTALT-LI
○ IA-HP-CHPD-CHED-KPCR
○ IC-CHED
● IA-CHPD
○ NI
○ ND

Save

FIGURE 4. Example of a form for gathering null data.



□ Partitions used for training

▨ Partition used for testing

Total Samples:1027

	Test A	Test B	Test C	Test D
256 Samples:	▨	□	□	□
257 Samples:	□	▨	□	□
257 Samples:	□	□	▨	□
257 Samples:	□	□	□	▨

FIGURE 5. Tests for cross validation.

data point is used for both training and validation throughout the k iterations.

In the PHH project, we performed CV following these steps:

- We selected samples according to the type of diagnosis, resulting in nine distinct selections.
- Next, we distributed these samples to create four partitions of proportional sizes, each containing approximately the same percentage for each diagnosis type.

These partitions were used for conducting four distinct trainings, alternating the test partition denominated Tests A, B, C, and D, while the remaining partitions underwent training, as illustrated in Figure 5.

The refinement stage provides a dataset with 79 condition attributes and one decision attribute for the training stage. We organized 12 different groupings with condition attributes in collaboration with a domain expert, as presented in Table 3. We created 11 partial groupings, each with the purpose

TABLE 3. Training Groups.

Group	Description of condition set attributes	Total attributes
1	Contact data with pesticides, laboratory tests and chemical class	22
2	Personal data and laboratory test data	23
3	Chemical class and lab test data	9
4	Clinical data regarding cardiovascular symptoms and laboratory tests	12
5	Clinical data regarding nervous system changes and laboratory tests	18
6	Clinical data regarding changes in the digestive system and laboratory tests	15
7	Clinical data regarding changes in the respiratory system and laboratory tests	13
8	Clinical data regarding changes in the hearing system and laboratory tests	11
9	Clinical data regarding skin changes and laboratory tests	12
10	Clinical data regarding changes in the urinary tract and laboratory tests	11
11	Data regarding the diagnosis of cancer and laboratory tests	10
12	Complete set with all condition data	80

of analyzing laboratory test information alongside different data. These included laboratory test data combined with information about specific organ systems, such as the nervous system, cardiovascular system, among others. Additionally, groups were obtained by combining laboratory test data with pesticide exposure information, personal data, and chemical classes. In all of these groupings, diagnosis was used as the decision attribute. Additionally, a grouping encompassing all condition data (Group 12) was created.

These 12 groups were employed as selection criteria in the training partitions resulting from CV for conducting tests A, B, C, and D.

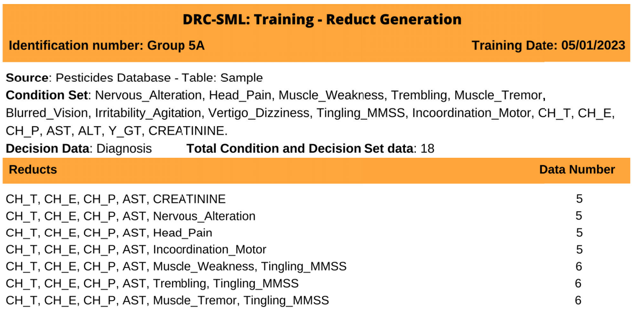
Figure 6 presents the form of the DRC–SML model used to document the results of training in reduct generation. For example, 18 condition attributes related to Group 5 were selected and applied to the training partitions for test A. During this training, we identified seven subsets with the same decision–making power as the original set.

Among the 12 trained groupings, Group 5 confirmed the importance of Central Nervous System alterations in individuals exposed to pesticides, according to the results presented in the scientific research conducted by Silverio et al. [38].

The reduct–generation phase was followed by an analysis conducted by the domain expert, who selected the reducts considered relevant for subsequent rule extraction. Maintaining a historical record of these relevant reducts, along with their respective rules is of utmost importance for comparison and future analyses.

Figure 7 presents the extraction form of the proposed model, documenting, as an example, the selection of the reduct "CH_T, CH_E, CH_P, AST, CREATININE" from the results of Group 5 training in Test A, which resulted in the generation of 27 decision–making rules. These rules are presented in Table 5.

At the end of training, the model validated the rules on the test set. The results of these validations are presented in Section IV.



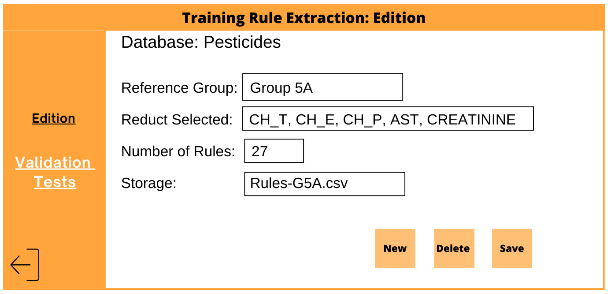
DRC-SML: Training - Reduct Generation

Identification number: Group 5A **Training Date: 05/01/2023**

Source: Pesticides Database - Table: Sample
Condition Set: Nervous_Alteration, Head_Pain, Muscle_Weakness, Trembling, Muscle_Tremor, Blurred_Vision, Irritability_Agitation, Vertigo_Dizziness, Tingling_MMSS, Incoordination_Motor, CH_T, CH_E, CH_P, AST, ALT, Y_GT, CREATININE.
Decision Data: Diagnosis **Total Condition and Decision Set data**: 18

Reducts	Data Number
CH_T, CH_E, CH_P, AST, CREATININE	5
CH_T, CH_E, CH_P, AST, Nervous_Alteration	5
CH_T, CH_E, CH_P, AST, Head_Pain	5
CH_T, CH_E, CH_P, AST, Incoordination_Motor	5
CH_T, CH_E, CH_P, AST, Muscle_Weakness, Tingling_MMSS	6
CH_T, CH_E, CH_P, AST, Trembling, Tingling_MMSS	6
CH_T, CH_E, CH_P, AST, Muscle_Tremor, Tingling_MMSS	6

FIGURE 6. Example of form for generating reducts.



Training Rule Extraction: Edition

Database: Pesticides

Reference Group: Group 5A
Reduct Selected: CH_T, CH_E, CH_P, AST, CREATININE
Number of Rules: 27
Storage: Rules-G5A.csv

FIGURE 7. Exemple of Rule Extraction Form.

All training procedures were conducted using RStudio software, version 1.4.1103, in conjunction with the Rough-Sets library, version 1.3-7 [40]. In the context of the 11 partial groupings, we used the functionalities offered by this library to identify all the possible reducts. Regarding the complete group (Group 12), the training focused on finding a single reduct, resulting in the identification of the set "CH_T, CH_E, CH_P, AST, Last_ContactDays," thus validating the relevance of laboratory information in this study context.

6) RECOVERY

In this stage, new samples are included to conduct additional training to enhance the rules generated for decision–making. This procedure is carried out while upholding the commitment to the pillars of privacy, copyright, and quality, as emphasized during the initial data collection phase. Additionally, a data refinement form is employed to retrieve information, avoid redundant acquisition of previously discarded data, and ensure compliance with the discretization categories and rules established in the refinement stage.

IV. RESULTS AND DISCUSSION

A. DRC–SML MODEL AS A KDD PROPOSAL

The DRC–SML model demonstrated its effectiveness in addressing all stages of KDD according to the Data Science hierarchy, as illustrated in Figure 8. The proposed model provides a documented methodology that aligns with each stage of the KDD pyramid.

As emphasized by Kelleher and Tierney [3], it is of paramount importance that the results of a Data Science

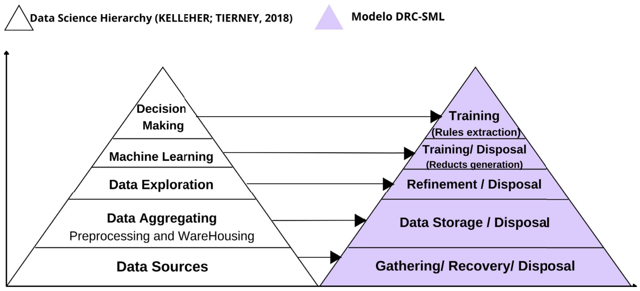


FIGURE 8. DRC–SML model applied to the KDD hierarchy in Data Science.

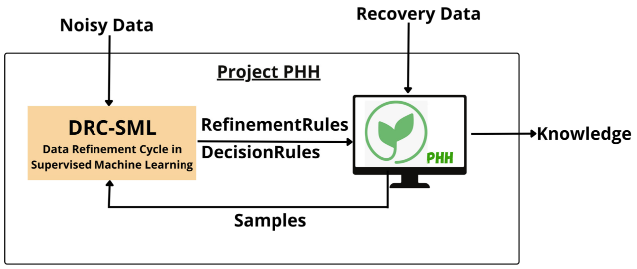


FIGURE 9. Interaction among the data from the scientific research and the DRC–SML model in PHH project.

project are communicated in an accessible manner so that even team members without technical backgrounds can understand them. In this context, the DRC–SML model stands out because of its simplicity,avoiding complex and unnecessary documentation, and focusing on essentials. Additionally, the model offers practical implications, such as:

- Effective Data Refinement: The model analyzes, discretizes, identifies, and eliminates redundancies, inconsistencies, and noise, resulting in cleaner and more accurate data.
- Dimensionality Reduction: The model can contribute to reducing dimensionality by identifying less relevant attributes, which is useful in extensive datasets.
- Transparent Decision–Making: The transparency of the model facilitates user acceptance and trust, as the generated rules are comprehensible.

The use of documentation proposed by the model can be applied to datasets in various domains such as health, marketing, finance, among others, making the model valuable for a wide range of professionals contributing to Data Science projects.

Figure 9 illustrates how the model received noisy data from previous scientific research and produced the following results:

- A set of refinement rules, that established patterns for data recovery, thereby expanding the available sample set for future training.
- A set of decision rules, that played a crucial role in guiding decision–making within the scope of the project. The model significantly contributed to the development of the PHH project, incorporating these rules as machine learning resources.

TABLE 4. Better results from group training.

Smallest Reducts	Accuracy of tests				Average
	A (%)	B (%)	C (%)	D (%)	± Standard Deviation
CH_T, CH_E, CH_P, AST, Last_ContactDays	95,75	96,89	98,05	98,05	97,19 ± 1,10
CH_T, CH_E, CH_P, AST, Mode_Application	96,91	98,44	98,05	98,44	97,96 ± 0,72
CH_T, CH_E, CH_P, AST, Type_Contact	98,07	99,22	98,05	98,83	98,54 ± 0,58
CH_T, CH_E, CH_P, AST, CREATININE	99,61	99,61	98,44	97,67	98,83 ± 0,95
CH_T, CH_E, CH_P, AST, Circumference_Abdominal	98,46	99,22	98,83	98,83	98,84 ± 0,31
CH_T, CH_E, CH_P, AST, Spontaneous_Abortion	99,61	99,22	98,44	98,83	99,03 ± 0,50
CH_T, CH_E, CH_P, AST, Nervous_Alteration	99,61	99,61	98,44	98,83	99,13 ± 0,58
CH_T, CH_E, CH_P, AST, Head_Pain	98,84	99,22	99,22	98,83	99,03 ± 0,22
CH_T, CH_E, CH_P, AST, Incoordination_Motor	99,23	99,22	99,22	98,83	99,13 ± 0,20
CH_T, CH_E, CH_P, AST, Vomit	98,46	99,61	99,22	98,44	98,93 ± 0,58

The DRC–SML model, with its cyclical methodology, specific documentation for each stage, and the use of machine learning, is capable of addressing challenges faced by data science, such as a large volume of data, noisy data, and, especially, the lack of direction leading to project completion, ensuring its effectiveness over time. In the PHH project, where patterns of pesticide poisoning may evolve due to changes in agricultural practices, types of pesticides used, among other factors, the DRC–SML, with its ability to recover and adapt to changes in data, can handle this dynamic, ensuring that the model remains relevant and effective in assisting healthcare professionals.

B. RESULTS OF THE APPLICATION OF THE DRC-SML MODEL TO THE DATA OF THE PHH PROJECT

Through the analysis of the results obtained in the application of the model to the PHH project data, we observed a progressive reduction in the amount of information from the data collection stage to the completion of the training stage. Initially, we had a set of 121 attributes, of which 41 were excluded throughout the process.

In all 48 training sessions, based on the 12 groupings applied in Tests A, B, C, and D, we identified minimal reducts consisting of only five attributes. Table 4 presents these reducts along with the accuracy of validation of the generated rules.

In the context of the analysis presented in Table 4, it is evident that all reducts achieved considerably high accuracy. Any of these reducts can be used to classify the new samples. However, in this specific scenario, the specialist opted for the "CH_T, CH_E, CH_P, AST, CREATININE" reduct because of its remarkable accuracy as well as its inclusion of information from laboratory analyses.

Table 5 presents the rules associated with the application of this reduct, which were generated during Test A. For example, the first rule in this table can be interpreted as follows:

TABLE 5. Decision Table resulting from the reduct "CH_T, CH_E, CH_P, AST, CREATININE" in Test A.

CH_T	CH_E	CH_P	AST	CREATININE	DIAGNOSIS
Low	Low	Low	Normal	Normal	IA–HP–CHPD
Low	Low	Low	Normal	Low	IA–HP–CHPD–CHED–KPCR
Normal	Normal	Low	Normal	Normal	IA–LP–CHPD
Off	Normal	Low	Off	Off	IA–CHPD
Low	Normal	Low	Normal	Normal	IA–CHPD
Normal	Normal	Low	Off	Off	IA–CHPD
Normal	Normal	Low	Low	Normal	IA–CHPD–ASTALT–LI
Low	Low	Normal	Normal	Normal	IC–CHED
Normal	Low	Normal	Normal	Normal	IC–CHED
Low	Normal	Normal	Normal	Normal	IC–CHED
Normal	Low	Normal	Off	Off	IC–CHED
Low	Low	Normal	Normal	Low	IC–CHED
Normal	Low	Normal	Normal	Low	IC–CHED
Low	Low	Normal	Low	Normal	IC–CHED–AST–LI
Low	Normal	Normal	Low	Normal	IC–CHED–AST–LI
Normal	Low	Normal	Low	Normal	IC–CHED–AST–LI
Normal	Normal	Normal	Normal	Normal	NI
Normal	Normal	Normal	Low	Normal	NI
Normal	Normal	Normal	Low	Low	NI
Normal	Normal	Normal	Normal	Off	NI
Normal	Normal	Normal	Normal	Low	NI
Normal	Normal	Normal	Off	Off	NI
Low	Normal	Normal	Off	Off	ND
Off	Off	Off	Low	Normal	ND
Off	Off	Off	Off	Off	ND
Off	Off	Off	Normal	Normal	ND
Low	Low	Normal	NI	NI	IC–CHED

If CH_T = "Low, " CH_E = "Low, " CH_P = "Low, " AST = "Normal," and CREATININE = "Normal," then the diagnosis is "IA–HP–CHPD."

These decision rules are understandable and interpretable for humans, providing transparency as explainable Artificial Intelligence (AI) techniques.

Recall and accuracy (according to Equations (1) and (2)) were determined based on the values obtained in the classifications of true positives (TP), true negatives (TN), and false negatives (FN).

$$Recall = \frac{TP}{TP + FN} \times 100 \tag{1}$$

$$Accuracy = \frac{TP + TN}{N} \times 100 \tag{2}$$

In our analysis, correctly diagnosed samples are considered true positives, while unclassified samples are considered false negatives, as they should have been categorized and were not due to the absence of rules to classify them. In our diagnostic case, no rule with more than one possible diagnosis occurred, so no samples were found to be considered false positives or true negatives.Due to the absence of false positive samples, Precision calculations yielded undefined values with division by zero errors, thereby impeding other calculations such as the F1–Score.

Table 6 provides a detailed presentation of accuracy for the reduct:"CH_T, CH_E, CH_P, AST, CREATININE" in each test, considering all types of diagnoses.

In the chosen reduct, calculations were performed to determine the recall and precision of each diagnosis, and the results are presented in Tables 7 and 8.

TABLE 6. Accuracy by the "CH_T, CH_E, CH_P, AST, CREATININE" in each test, considering all types of diagnoses.

Tests	N	TP	TN	FP	FN	Accuracy(%)
A	256	255	0	0	1	99,61
B	257	256	0	0	1	99,61
C	257	253	0	0	4	98,44
D	257	251	0	0	6	97,67

TABLE 7. Recall of each diagnosis referring to the rules generated by the "CH_T, CH_E, CH_P, AST, CREATININE."

Diagnosis	Recall of tests (%)			
	A	B	C	D
IA–HP–CHPD	100,00	100,00	100,00	100,00
IA–HP–CHPD–CHED–KPCR	100,00	100,00	100,00	100,00
IA–LP–CHPD	100,00	100,00	100,00	100,00
IA–CHPD	100,00	0,00	100,00	0,00
IA–CHPD–ASTALT–LI	100,00	100,00	100,00	100,00
IC–CHED	98,04	100,00	94,12	100,00
IC–CHED–AST–LI	100,00	100,00	100,00	100,00
NI	100,00	100,00	100,00	98,43
ND	100,00	100,00	66,67	33,33

TABLE 8. Accuracy of each diagnosis referring to the rules generated by the "CH_T, CH_E, CH_P, AST, CREATININE."

Diagnosis	Accuracy of tests (%)			
	A	B	C	D
IA–HP–CHPD	100,00	100,00	100,00	100,00
IA–HP–CHPD–CHED–KPCR	100,00	100,00	100,00	100,00
IA–LP–CHPD	100,00	100,00	100,00	100,00
IA–CHPD	100,00	99,61	100,00	99,61
IA–CHPD–ASTALT–LI	100,00	100,00	100,00	100,00
IC–CHED	99,61	100,00	98,83	100,00
IC–CHED–AST–LI	100,00	100,00	100,00	100,00
NI	100,00	100,00	100,00	98,83
ND	100,00	100,00	99,61	99,22

In the set of 1027 samples, there was oversampling for some diagnoses, and undersampling for others. During cross–validation (CV), an attempt was made to achieve a uniform distribution of these samples among different diagnoses. However, in Tests B and D, it was observed that the diagnoses in "IA–CHPD" recorded a recall of 0% due to the inability to generate effective rules for the classification of this diagnosis in the test set. Considering that new sample collections are underway as part of the PHH Project and that the initial results are already benefiting healthcare professionals, we do not deem it necessary to resort to simulation methodologies for class balancing. Adjustments to the samples through the recovery process will be necessary in the future to address this imbalance.

As established by the DRC–SML model, in the recovery stage, new samples are collected based on all information obtained during the refinement stage. Table 9 presents a portion of this information, specifically the refinement rules of the data present in the reducts listed in Table 4.

Upon completion of the recovery stage, it will be possible to conduct new training sessions with the aim of incorporating an increasing number of diverse cases of pesticide poisoning,

TABLE 9. DRC–SML: Refinement Rules for Recovery.

Original name	Refinement Rules
Last time you had contact (in days) with a pesticide	Uninformed, [0, 7 days] -Acute Exposure [8, 30 days] - Subacute Exposure, [31, 90 days] - Subchronic Exposure, [91, *] - Chronic Exposure
Product application method	Back pump, Hose, Tractor without cabin, Closed cabin tractor, Off
Contact Type	Direct, Indirect, No contact
Nervous system alteration, Headache, Motor incoordination, Vomit, Spontaneous abort	Yes / No
CH_T	[Val $<$ 15.5] Low,[Val $\geq$ 15.5] Normal
CH_E	[Val $<$ 32] Low, [Val $\geq$ 32] Normal
CH_P	[Val $<$ 1.3] Low, [Val $\geq$ 1.3] Normal
AST	[Val $<$ 4] Low, [Val $\geq$ 4 and Val $\leq$ 36] Normal
CREATININE	Men: [Val $<$ 0.9] Low, [Val $\geq$ 0.9 and Val $\leq$ 1.3] Normal Women: [Val $<$ 0.6] Low, [Val $\geq$ 0.6 and Val $\leq$ 1.1] Normal
Abdominal circumference	Men: [*, 101] Adequate, [102, *] Inadequate Women: [*, 87] Adequate, [88, *] Inadequate

thereby improving the model's ability to handle various diagnoses with greater accuracy.

V. CONCLUSION AND FUTURE RESEARCH

The DRC–SML model demonstrated its effectiveness in addressing all stages of the KDD process according to the hierarchy of Data Science, as illustrated in Figure 8. Its simplicity and practicality make it suitable for professionals from various fields involved in Data Science projects. Furthermore, it successfully handled noisy data and remove unnecessary information, resulting in reduced uncertainty in the dataset. This led to a dataset prepared for training, and consequently, the derivation of well–validated decision rules. These results significantly contributed to the PHH project, improving its efficiency and usefulness, as 27 decision rules were obtained with 99.61% diagnostic accuracy. These rules serve as support for healthcare professionals' decision–making and contribute to the health of agricultural workers, which is crucial for ensuring agricultural productivity and product quality.

We identified Data Science models with purely theoretical descriptions, and separately, we found machine learning applications assuming that the data is already prepared for use, which hindered the presentation of a comparative analysis with the proposed model.

In future work, it is recommended:

- To explore the connection of the DRC–SML model with object-oriented databases and graph-oriented databases, allowing for greater flexibility in data storage and retrieval.
- Expanding the scope by incorporating the necessary documentation into the DRC–SML modeling for tracking additional machine learning methods. This approach will enable data scientists to assess and choose the most suitable method based on the specific characteristics of their problems, allowing for comparative analyses on the same dataset.
- To improve the accessibility and practical utility of the model, a possible extension would be to create an API to implement a mobile data collection application.

These efforts have the potential to expand the impact of the model and broaden its applicability in Data Science projects.

REFERENCES

[1] C. J. Cremin, S. Dash, and X. Huang, "Big data: Historic advances and emerging trends in biomedical research," *Current Res. Biotechnol.*, vol. 4, pp. 138–151, Jan. 2022.

[2] J. Saltz and I. Krasteva, "Current approaches for executing big data science projects a systematic literature review," *PeerJ Comput. Sci.*, vol. 8, p. e862, Feb. 2022.

[3] J. D. Kelleher and B. Tierney, *Data Science*. Cambridge, MA, United States: MIT Press, 2018.

[4] C. Silva, M. Saraee, and M. Saraee, "Data science in public mental health: A new analytic framework," in *Proc. IEEE Symp. Comput. Commun. (ISCC)*, Jun. 2019, pp. 1123–1128.

[5] S. Jain, "Comprehensive survey on data science, lifecycle, tools and its research issues," in *Proc. Int. Conf. Mach. Learn., Big Data, Cloud Parallel Comput.*, vol. 1, May 2022, pp. 838–842.

[6] T. D. Bie, L. D. Raedt, J. Hernández-Orallo, H. H. Hoos, P. Smyth, and C. K. I. Williams, "Automating data science: Prospects and challenges," *Commun. ACM*, vol. 65, no. 2, pp. 76–87, 2022.

[7] S. Uddin, A. Khan, M. E. Hossain, and M. A. Moni, "Comparing different supervised machine learning algorithms for disease prediction," *BMC Med. Informat. Decis. Making*, vol. 19, no. 1, pp. 281–296, Dec. 2019.

[8] Z. Pawlak, *Rough Sets: Theoretical Aspects of Reasoning About Data*, 1st ed. Dordrecht, The Netherlands: Springer, 1991.

[9] D. P. Acharjya and A. Abraham, "Rough computing—A review of abstraction, hybridization and extent of applications," *Eng. Appl. Artif. Intell.*, vol. 96, Nov. 2020, Art. no. 103924. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0952197620302529

[10] I. H. Sarker, "Data science and analytics: An overview from data-driven smart computing, decision-making and applications perspective," *Social Netw. Comput. Sci.*, vol. 2, no. 5, p. 377, Sep. 2021.

[11] A. Peña-Fernández, M. Peña, M. Lobo, and M. Evans, "Interventions to enhance the teaching of toxicology at a U.K. University," in *Proc. EDULEARN Conf.*, Palma, Spain, Jul. 2018, pp. 7126–7130.

[12] B. Mondal, *Artificial Intelligence: State of the Art*. Cham, Switzerland: Springer, 2020, pp. 389–425, doi: 10.1007/978-3-030-32644-9.

[13] J. Saltz, N. Hotz, D. Wild, and K. Stirling, "Exploring project management methodologies used within data science teams," in *Proc. Americas Conf. Inf. Syst.*, 2018, pp. 1–5.

[14] J. Russell, *Agile Data Science 2.0: Building Full-Stack Data Analytics Applications With Spark*. Sebastopol, CA, USA: O'Reilly Media, 2017.

[15] R. Wirth and J. Hipp, "CRISP-DM: Towards a standard process model for data mining," in *Proc. 4th Int. Conf. Practical Appl. Knowl. Discovery Data Mining*, vol. 1. London, U.K.: Springer-Verlag, 2000.

[16] F. Foroughi and P. Luksch, "Data science methodology for cybersecurity projects," 2018, *arXiv:1803.04219*.

[17] I. Martinez, E. Viles, and I. G. Olaizola, "A survey study of success factors in data science projects," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2021, pp. 2313–2318.

[18] I. Martinez, E. Viles, and I. G. Olaizola, "Data science methodologies: Current challenges and future approaches," *Big Data Res.*, vol. 24, May 2021, Art. no. 100183. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2214579620300514

[19] M. Gm, A. Alameen, M. Kolhar, and M. Rahmath, "Data science techniques, tools and predictions," *Int. J. Recent Technol. Eng.*, vol. 8, pp. 5661–5668, Mar. 2020.

[20] Y. Yao, "Symbols-Meaning-Value (SMV) space as a basis for a conceptual model of data science," *Int. J. Approx. Reasoning*, vol. 144, pp. 113–128, May 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0888613X2200024X

[21] A. K. Laturiuw and Y. A. Singgalen, ''Sentiment analysis of raja ampat tourism destination using CRISP-DM: SVM, NBC, DT, and k-NN algorithm,'' *J. Inf. Syst. Informat.*, vol. 5, no. 2, pp. 518–535, May 2023.

[22] T. Kaewwiset, P. Temdee, and T. Yooyativong, ''Employee classification for personalized professional training using machine learning techniques and SMOTE,'' in *Proc. Joint Int. Conf. Digit. Arts, Media Technol. ECTI Northern Sect. Conf. Electr., Electron., Comput. Telecommun. Eng.*, Mar. 2021, pp. 376–379.

[23] B. Abdualgalil and S. Abraham, ''Applications of machine learning algorithms and performance comparison: A review,'' in *Proc. Int. Conf. Emerg. Trends Inf. Technol. Eng.*, Feb. 2020, pp. 1–6.

[24] V. Sarveshwaran, J. J. Jayakanth, and Y. Renu, ''Comparative analysis of diverse classification algorithms of machine learning by using various quality metrics,'' in *Proc. IEEE 5th Int. Conf. Cybern., Cognition Mach. Learn. Appl. (ICCCMLA)*, Oct. 2023, pp. 551–556.

[25] M. S. Saravanan and S. Charan, ''Prediction of insufficient accuracy for human activity recognition using convolutional neural network in compared with support vector machine,'' in *Proc. 5th Int. Conf. Contemp. Comput. Informat. (IC3I)*, Dec. 2022, pp. 1915–1919.

[26] S. Singh and J. Yao, ''Pneumonia detection with game-theoretic rough sets,'' in *Proc. 20th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2021, pp. 1029–1034.

[27] S. K. Nayak, S. K. Pradhan, S. Mishra, S. Pradhan, and P. K. Pattnaik, ''Rough set technique to predict symptoms for malaria,'' in *Proc. 8th Int. Conf. Comput. Sustain. Global Develop. (INDIACom)*, Mar. 2021, pp. 312–317.

[28] M. M. Mafarja and S. Mirjalili, ''Hybrid binary ant lion optimizer with rough set and approximate entropy reducts for feature selection,'' *Soft Comput.*, vol. 23, no. 15, pp. 6249–6265, Aug. 2019.

[29] L. Zhang, J. Zhan, and J. C. Alcantud, ''Novel classes of fuzzy soft-coverings-based fuzzy rough sets with applications to multi-criteria fuzzy group decision making,'' *Soft Comput.*, vol. 23, pp. 5327–5351, Jun. 2019.

[30] Z. Chelly Dagdia, C. Zarges, B. Schannes, M. Micalef, L. Galiana, B. Rolland, O. de Fresnoye, and M. Benchoufi, ''Rough set theory as a data mining technique: A case study in epidemiology and cancer incidence prediction,'' in *Machine Learning and Knowledge Discovery in Databases*, U. Brefeld, E. Curry, E. Daly, B. MacNamee, A. Marascu, F. Pinelli, M. Berlingerio, and N. Hurley, Eds. Cham, Switzerland: Springer, 2019, pp. 440–455.

[31] A. Fiedukowicz, ''The use of rough rules in the selection of topographic objects for generalizing geographical information,'' *Polish Cartographical Rev.*, vol. 52, no. 1, pp. 1–15, Mar. 2020, doi: 10.2478/pcr-2020-0001.

[32] N. Kumari and D. P. Acharjya, ''A decision support system for diagnosis of hepatitis disease using an integrated rough set and fish swarm algorithm,'' *Concurrency Comput., Pract. Exper.*, vol. 34, no. 21, p. e7107, Sep. 2022, doi: 10.1002/cpe.7107.

[33] M. S. Pathan, Z. Jianbiao, D. John, A. Nag, and S. Dev, ''Identifying stroke indicators using rough sets,'' *IEEE Access*, vol. 8, pp. 210318–210327, 2020.

[34] G. Yuan, J. Zhou, and Q. Chen, ''Rough-fuzzy clustering based on adaptive weighted values and three-way decisions,'' in *Proc. IJCRS*, Suzhou, China. Berlin, Heidelberg: Springer, 2022, pp. 420–429, doi: 10.1007/978-3-031-21244-4.

[35] R. Cybulski and P. Artiemjew, ''Application of random sampling in the concept-dependent granulation method,'' in *Proc. Ann. Comput. Sci. Inf. Syst.*, Sep. 2022, pp. 3–11.

[36] R. Cybulski and P. Artiemjew, ''Accelerating concept-dependent granulation technique using data decomposition,'' in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2022, pp. 6195–6201.

[37] J. Carvalho, (2023), ''Cases of pesticide exposure,'' *IEEE Dataport*, doi: 10.21227/kn78-jf53.

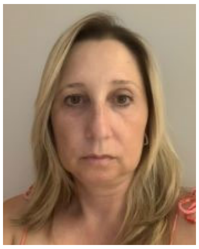
[38] A. C. P. Silvério, I. Martins, D. A. Nogueira, M. A. S. Mello, E. A. C. D. Loyola, and M. M. D. C. Graciano, ''Assessment of primary health care for rural workers exposed to pesticides,'' *Revista de Saúde Pública*, vol. 54, p. 9, Jan. 2020, doi: 10.11606/s1518-8787.2020054001455.

[39] D. Berrar, ''Cross-validation,'' in *Encyclopedia of Bioinformatics and Computational Biology*, S. Ranganathan, M. Gribskov, K. Nakai, and C. Schnbach, Eds. New York, NY, USA: Academic, 2019, pp. 542–545. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B978012809633820349X

[40] L. S. Riza, A. Janusz, C. Bergmeir, C. Cornelis, F. Herrera, D. Ś. Zak, and J. M. Benítez, ''Implementing algorithms of rough set theory and fuzzy rough set theory in the R package 'RoughSets,''' *Inf. Sci.*, vol. 287, pp. 68–89, Dec. 2014.

JAQUELINE C. S. CARVALHO received the B.Sc. degree in computer science from José do Rosário Vellano University, in 1995, and the M.Sc. degree in electrical engineering from the Federal University of Itajubá, in 2000, where she is currently pursuing the Ph.D. degree. She is a Professor of computer science with José do Rosário Vellano University. Her research interests include data science and artificial intelligence.

TALES C. PIMENTA (Senior Member, IEEE) received the B.Sc. and M.Sc. degrees in electrical engineering from the Federal University of Itajubá, in 1985 and 1988, respectively, and the Ph.D. degree in electrical and computer engineering from Ohio University, in 1992. He held the first visiting scholar position in low-voltage analog integrated circuits with The Ohio State University, in 1997, the second position in the area of ultra-high frequency integrated circuits with Virginia Polytechnic Institute and State University, in 2005, and the third position in the area of devices for biomedical applications with North Florida University. He is currently a Professor with the Federal University of Itajubá. His research interests include circuits and systems for biomedical applications.



ALESSANDRA C. P. SILVERIO received the B.Sc. degree in biochemical pharmacy and the Ph.D. degree in pharmaceutical sciences (toxicology) from the Federal University of Alfenas, in 1995 and 2016, respectively. She is currently a Professor in medicine, pharmacy, biomedicine, nutrition, and physical education with José do Rosário Vellano University. She coordinates the CNPq-accredited research group in collective health with a greater interest in rural worker health and diabetes.

MARCOS A. CARVALHO received the B.Sc. degree in computer science from José do Rosário Vellano University, in 1995, and the M.Sc. degree in computer science from the Federal University of Itajubá, in 2010, where he is currently pursuing the Ph.D. degree. He is a Professor of computer science with José do Rosário Vellano University. His research interests include digital image processing and artificial intelligence.

JOAO PAULO C. S. CARVALHO received the B.Sc. degree in computer science and mathematics from Brescia University, Owensboro, KY, USA, in 2023. His research interests include data science, artificial intelligence, and web/mobile programming languages.