International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

A NEW DATA SCIENCE MODEL WITH SUPERVISED LEARNING AND ITS APPLICATION ON PESTICIDE POISONING DIAGNOSIS IN RURAL WORKERS

AKKILI JAYAGAYATHRI, Dr. C. MANJUNATH
MCA Student, Professor & Head of the Department
DEPT OF MCA
PVKK INSTITUTE OF TECHNOLOGY(AUTONOMOUS), Anantapuramu – 515001 (A.P)
a.gayathri4554@gmail.com, cpmnc15@gmail.com

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 Science 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.

1. INTRODUCTION

1.1. 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 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.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

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..

1.2. PROJECT OBJECTIVE

A critical issue for data owners is how to efficiently and securely grant privilege level-based access rights to a set of data. Data owners are becoming more interested in selectively sharing information with data users based on different levels of granted privileges. The desire to grant level-based access results in higher computational complexity and complicates the methods in which data is shared on the cloud. Research in this field focuses on finding enhanced schemes that can securely, efficiently and intelligently share data on the cloud among users according to granted access levels. Based on a study conducted by the National Institute of Standards and Technology (NIST), Role-Based Access Control (RBAC) models are the most widely used to share data in hierarchical enterprises of 500 or more individuals. RBAC models aim to restrict system access to authorized users as they provide access control mechanisms.

2. SYSTEM ANALYSIS

2.1. EXISTING SYSTEM

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,

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

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.

2.1.1. Disadvantages

- 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

2.2. PROPOSED SYSTEM

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.

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.

2.2.1. Advantages

- The proposed model comprehensively encompasses all the stages of the KDD process.
- In the scope of the PHH project, we utilized data, which originated from a scientific research study in the field of Biomedicine.

3. SYSTEM DESIGN

3.1. ARCHITECTURE DESIGN

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

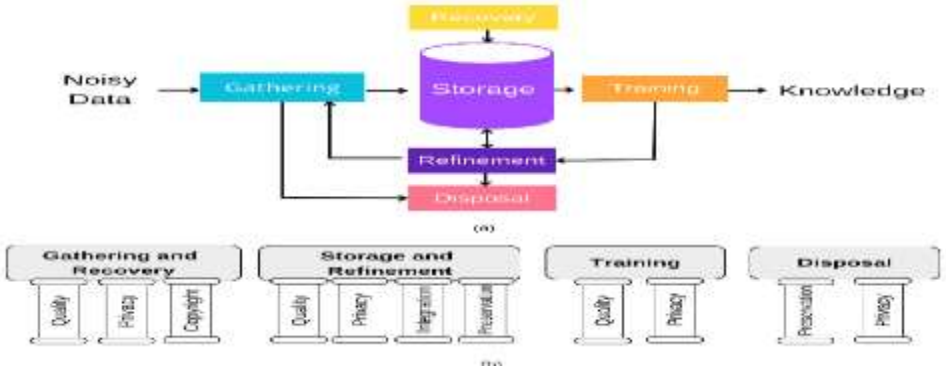


Fig : Architecture Design

3.2. MODULES DESCRIPTION

1)Upload Dataset: using this module we will upload disease diagnosis dataset and dataset

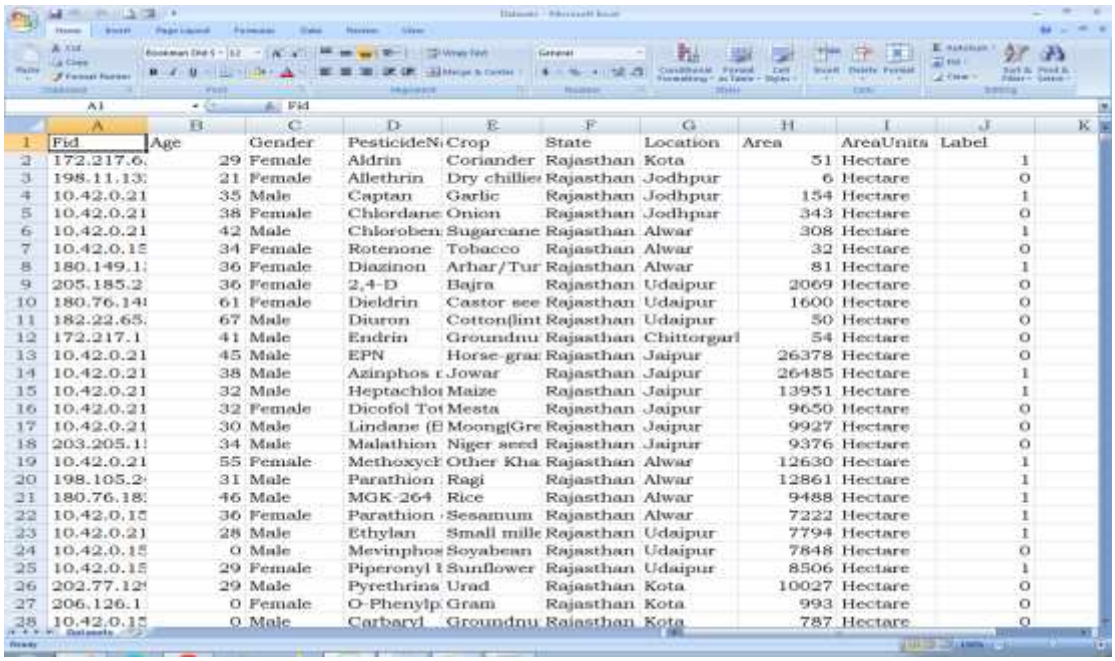
2)Extract Features from Dataset: using this module we will extract features from both datasets and then build training dataset

3)Train Algorithm: using above train dataset we will train svm model and then build a trained model and this model can be used to predict disease from any new test test files

4)svm Accuracy & Loss Graph: using this module we will display comparison graph between accuracy and loss of svm trained model

5)Upload Test &Predict : using this module we will upload test files and then apply trained model on that test to pesticides disease.

4. SCREEN SHOTS



	A	B	C	D	E	F	G	H	I	J	K
1	Fid	Age	Gender	PesticideN	Crop	State	Location	Area	AreaUnits	Label	
2	172.217.6.	29	Female	Aldrin	Coriander	Rajasthan	Kota	51	Hectare	1	
3	198.11.13:	21	Female	Allethrin	Dry chillie	Rajasthan	Jodhpur	6	Hectare	0	
4	10.42.0.21	35	Male	Captan	Garlic	Rajasthan	Jodhpur	154	Hectare	1	
5	10.42.0.21	38	Female	Chlordane	Onion	Rajasthan	Jodhpur	343	Hectare	0	
6	10.42.0.21	42	Male	Chloroben	Sugarcane	Rajasthan	Alwar	308	Hectare	1	
7	10.42.0.15	34	Female	Rotenone	Tobacco	Rajasthan	Alwar	32	Hectare	0	
8	180.149.1:	36	Female	Diazinon	Arhar/Tur	Rajasthan	Alwar	81	Hectare	1	
9	205.185.2	36	Female	2,4-D	Bajra	Rajasthan	Udaipur	2069	Hectare	0	
10	180.76.14(	61	Female	Dieldrin	Castor see	Rajasthan	Udaipur	1600	Hectare	0	
11	182.22.65.	67	Male	Diuron	Cotton(lint	Rajasthan	Udaipur	50	Hectare	0	
12	172.217.1	41	Male	Endrin	Groundnu	Rajasthan	Chittorgarl	54	Hectare	0	
13	10.42.0.21	45	Male	EPN	Horse-grai	Rajasthan	Jaipur	26378	Hectare	0	
14	10.42.0.21	38	Male	Azinphos (	Jowar	Rajasthan	Jaipur	26485	Hectare	1	
15	10.42.0.21	32	Male	Heptachlo	Maize	Rajasthan	Jaipur	13951	Hectare	1	
16	10.42.0.21	32	Female	Dicofol Tot	Mesta	Rajasthan	Jaipur	9650	Hectare	0	
17	10.42.0.21	30	Male	Lindane (E	Moong(Gre	Rajasthan	Jaipur	9927	Hectare	0	
18	203.205.1!	34	Male	Malathion	Niger seed	Rajasthan	Jaipur	9376	Hectare	0	
19	10.42.0.21	55	Female	Methoxycl	Other Kha	Rajasthan	Alwar	12630	Hectare	1	
20	198.105.2	31	Male	Parathion	Ragi	Rajasthan	Alwar	12861	Hectare	1	
21	180.76.18:	46	Male	MGK-264	Rice	Rajasthan	Alwar	9488	Hectare	1	
22	10.42.0.15	36	Female	Parathion	Sesamum	Rajasthan	Alwar	7222	Hectare	1	
23	10.42.0.21	28	Male	Ethylan	Small mille	Rajasthan	Udaipur	7794	Hectare	1	
24	10.42.0.15	0	Male	Mevinphos	Soyabean	Rajasthan	Udaipur	7848	Hectare	0	
25	10.42.0.15	29	Female	Piperonyl I	Sunflower	Rajasthan	Udaipur	8506	Hectare	1	
26	202.77.12!	29	Male	Pyrethrins	Urad	Rajasthan	Kota	10027	Hectare	0	
27	206.126.1	0	Female	O-Phenylp	Gram	Rajasthan	Kota	993	Hectare	0	
28	10.42.0.15	0	Male	Carbaryl	Groundnu	Rajasthan	Kota	787	Hectare	0	

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

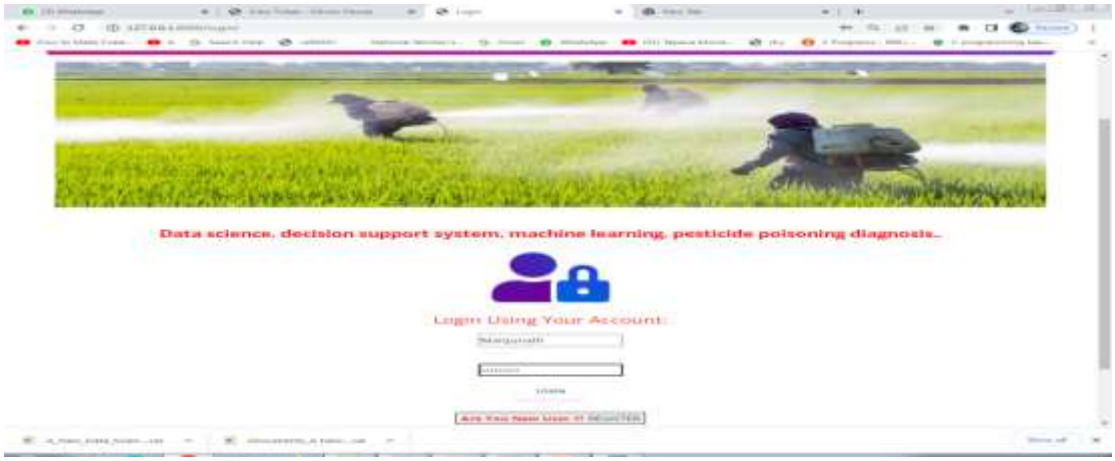
Screen: Datasets Page

Screen :View Detected Pesticides

Screen :Ratio of Pesticides

Screen :ML Ratio Accuracy in Line Chart

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper



Screen :User Login Page

Screen :User Menu Page

Screen :User Prediction Page

5. CONCLUSION

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.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

6. FUTURE ENHANCEMENTS

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.

BIBLIOGRAPHY

[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,’’ PeerJComput. 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..