

# Data Driven Diagnosis Of Pesticide Poisoning

M Vinod<sup>1</sup>, Annabemoju Hasini<sup>2</sup>, Gorla Nandhu<sup>3</sup>

<sup>1</sup>Associate professor, Department of CSE, Bhoj Reddy Engineering College for Women, India

<sup>2,3</sup>B.Tech Students, Department of CSE, Bhoj Reddy Engineering College for Women, India

**ABSTRACT** On a Data Science project, it is crucial to assess the significance of the data and discern patterns that inform decision-making grounded on domain-specific expertise. Moreover, a precise description of techniques and the compilation of documentation to direct a project's progress from conception to conclusion are crucial components. This paper introduces a Data Science model intended to direct the process, including data gathering and training to enhance knowledge discovery. Driven by shortcomings in current Data Science approaches, especially the absence of realistic, step-by-step instructions for data preparation to achieve the production phase. The suggested approach, termed "Data Refinement Cycle with Supervised Machine Learning (DRC-SML)," was created in response to the evolving In recent years, Data Science has made significant progress beyond its roots in traditional fields of Statistics. One of the most striking indications of this tendency is the exponential growth in the amount of data that is being generated and stored all over the globe. According to Cremin et al. [1,] this number reached around 44 zettabytes at the beginning of the year 2020. Furthermore, forecasts show that by the year 2025, the daily global data production would approach 463 exabytes. The phrase "big data" is often used to refer to this enormous data aggregation, which is differentiated by its substantial volume and many types of data.

Despite the significant progress that has been made in the field of data science, the successful execution of efforts within this industry continues to be plagued with a great deal of difficulty. According to

requirements of a Data Science project designed to aid healthcare practitioners in identifying pesticide toxicity in rural laborers. The dataset used in this study was derived from scientific research including the collection of 1027 samples, encompassing data pertinent to toxicity biomarkers and clinical assessments. We attained an accuracy of 99.61% using just 27 principles for diagnostic determination. The outcomes enhanced healthcare practices and elevated the quality of life in rural regions. The project results proved the efficacy of the suggested model.

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

## INTRODUCTION

the findings of Saltz and Krasteva [2], around 87 percent of Data Science projects do not make it to the production phase.

Due to the fact that it is an interdisciplinary discipline, data science has applications in a variety of different categories of interest. Throughout the process, it is vital to engage with domain experts that have a comprehensive understanding of the relevant issues in order to achieve the results that are intended. It is necessary for data scientists to have a comprehensive understanding of Knowledge Discovery in Databases (KDD), which necessitates competence in the fields of statistics, computer science, computing, databases, and machine learning.

It is common for models in the field of data science to conform to cyclical frameworks that are known as

the data lifecycle. These frameworks provide data scientists direction as they go through the KDD process. In this context, one of the most persistent challenges is the lack of interpretability in complex models, which, when combined with the presence of data of poor quality or noise, has the potential to weaken the effectiveness and reliability of the models [3].

The authors Jain and Kushagra [5] highlight the fact that the quality of a model that is developed is substantially tied to the data that is provided. These include the identification of relevant data, the consolidation of databases, the cleaning of data, the generation of new data, and the extraction of unique features from previously collected data. When it comes to the lifetime of a Data Science project, the phase that requires the biggest amount of effort and is maybe the most significant is the phase that involves the preparation of data.

According to the findings of De Bie and colleagues [6], the use of machine learning methods is an essential part of the arsenal of a data scientist. Over the course of the last twenty years, these techniques have significantly increased in importance. They range from really fundamental processes to more complex ones, such as deep learning. Nevertheless, it is of the utmost importance to emphasize that these techniques often imply the availability of substantial volumes of high-quality data, which, in practice, presents additional challenges.

There are three basic classifications that may be used to machine learning: supervised learning, unsupervised learning, and reinforcement learning. The process of supervised learning involves the annotation of data by specialists, and the instances are described by a dataset and a class label that corresponds to it. The key purpose is to create a classifier by making use of existing examples, which will allow the computer to successfully classify new

occurrences that have not been labeled [7].

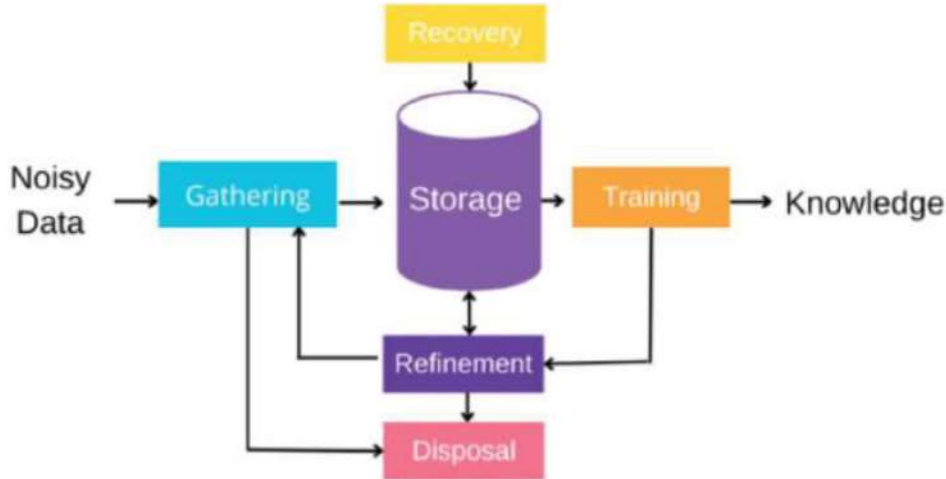
A major mathematical tool that serves the purpose of addressing a certain category of uncertainty and imprecision is known as Rough Set Theory (RST). This theory was first presented by Pawlak and was further investigated by Achariva and Abraham. When it comes to tackling real machine learning difficulties, RST has shown its effectiveness, regardless of whether it is used on its own or in conjunction with other machine learning models. The emphasis of a project in the field of data science is on data, which is often positioned within a specific context. There are a variety of problems that have benefited from the implementation of Data Science projects [10], including healthcare analytics, marketing, and financial analysis.

The purpose of this study is to investigate a public health problem that has not been well investigated within the field of data science. According to Peña-Fernández et al. [11], the issue of pesticide poisoning among agricultural workers is a significant concern that may result in significant social and economic implications on a worldwide scale. This is due to the lack of standardized tests for biological diagnosis as well as a shortage of competent healthcare providers who are equipped to handle patients with such conditions.

Nevertheless, when we first started our investigation, we discovered that there was no existing Data Science model that provided instructions that were both pragmatic and sequential for the preparation of data in order to reach the production phase. The current models need effective resources for the preparation of data, which includes the individual evaluation of each data point, the removal of unnecessary information, the transformation of data, the development of new data, and the selection of training and testing datasets.

## METHODOLOGY

### System Architecture



### System Architecture

The system architecture for a novel data science model using supervised learning for diagnosing pesticide toxicity in rural laborers is structured to accommodate diverse inputs, analyze them effectively, and provide precise predictions. This architecture incorporates data collection, preprocessing, model training, assessment, and deployment to guarantee dependable and actionable insights for medical practitioners serving rural populations. The data gathering layer underpins the system, collecting pertinent information from many sources, including health records, surveys, environmental monitoring systems, and direct contributions from rural laborers. The data generally comprises a synthesis of demographic information, symptomatology, pesticide exposure history, environmental variables, and medical history. The information may include worker age, gender, period of pesticide exposure, kind of pesticide used, symptoms such as dizziness or headaches, and pre-existing health issues such as asthma or diabetes. Furthermore, external variables such as meteorological conditions, environmental pesticide

concentration, and application patterns are also documented to provide a thorough understanding of the elements that may contribute to pesticide poisoning. Upon data collection, the data preparation layer is tasked with cleansing, standardizing, and converting raw data into a format suitable for model building. This phase addresses absent values, anomalies, and guarantees that categorical variables (e.g., pesticide kinds) are accurately recorded using techniques such as one-hot encoding or label encoding.

Numerical characteristics are normalized or standardized to provide uniform scaling, which is crucial for several machine learning techniques. In supervised learning, the dataset is labeled, including historical data that indicates whether the worker suffered from pesticide poisoning (i.e., the goal variable). Feature engineering is a crucial phase that employs domain expertise to generate additional variables potentially enhancing model accuracy, such as total pesticide exposure or symptom severity.

The fundamental component of the system is the supervised learning model layer. Utilizing labeled

data, diverse machine learning algorithms, including Decision Trees, Random Forests, Support Vector Machines (SVM), and Neural Networks, are applied to discern the correlation between input features (e.g., pesticide exposure, symptoms) and the output (i.e., diagnosis of pesticide poisoning).

The program is trained on past data to identify trends and forecast the probability of poisoning. The supervised learning methodology guarantees that the model can effectively generalize to novel data, yielding precise predictions for new instances based on acquired patterns.

Moreover, cross-validation methods are used to mitigate overfitting and evaluate the model's efficacy on novel data. The model assessment layer is essential for evaluating the performance of the learned model post-training. Multiple measures, including accuracy, precision, recall, F1-score, and ROC-AUC, are used to assess the model's efficacy in diagnosing pesticide toxicity. For instance, accuracy and memory are crucial in medical diagnostics, since false negatives (failing to identify poisoning when it happens) may lead to severe repercussions. The assessment findings inform the selection of the optimal model or suggest the need for further hyperparameter adjustment to enhance performance.

### **Proposed Machine Learning-Based Model**

The proposed machine learning model for detecting pesticide poisoning in rural laborers utilizes supervised learning approaches to estimate the probability of poisoning based on many characteristics. The model initiates by gathering extensive data from many sources, including demographic information (age, gender), pesticide exposure history (type, duration, frequency), symptoms (headache, dizziness, nausea), and environmental variables (temperature, humidity, pesticide concentration). The data points are

meticulously preprocessed to guarantee their appropriateness for machine learning.

This entails addressing absent values, normalizing numerical attributes such as age and exposure length to a uniform scale, and encoding categorical variables like pesticide kind and symptoms using techniques such as one-hot encoding. The objective is to provide a pristine, organized dataset in which the characteristics are suitably represented for model training. After the data is prepared, the subsequent stage involves choosing the suitable machine learning method. Algorithms like Decision Trees, Random Forests, Support Vector Machines (SVM), and Neural Networks are assessed to identify the most effective model for forecasting pesticide toxicity. These algorithms are trained on historical data to discern patterns and correlations between input characteristics (e.g., pesticide exposure and symptoms) and the target variable (diagnosis of pesticide poisoning). Post-training, the model is assessed using criteria including accuracy, precision, recall, and F1-score to verify its capability to accurately identify pesticide poisoning in novel instances. Upon selection and validation of the optimal model, it is deployed for practical use, allowing healthcare practitioners to enter pertinent data and get timely forecasts. This machine learning model facilitates early diagnosis of pesticide toxicity and establishes a basis for further enhancements via regular data updates and retraining, so assuring sustained accuracy.

### **Dataset**

The dataset used for detecting pesticide toxicity in rural laborers comprises diverse data points gathered from individuals potentially exposed to pesticides. The data points include details on demographics, exposure history, symptoms, and medical problems. This dataset aims to facilitate the development of a machine learning model capable of predicting

pesticide poisoning in workers based on the attributes included within the data. This is an overview of the primary characteristics in the dataset:

**Demographics:** Data including age, gender, and employment that may affect the probability of poisoning. Illustration: Age, Gender, Type of Occupation.

**Details about the worker's exposure to pesticides,** including the pesticide kind, exposure duration, and frequency of exposure. Example: Pesticide Classification (e.g., Organophosphates), Exposure Duration (e.g., 5 hours), Exposure Frequency (e.g., Weekly).

**Symptoms:** Physical manifestations that the worker may encounter, including dizziness, nausea, headaches, etc. Dizziness, nausea, headache, vomiting.

**Medical History:** Details on prior health problems, including asthma, respiratory disorders, or other chronic ailments that may increase a worker's susceptibility to pesticide exposure. Illustration: Asthma, Diabetes, Respiratory Disorders.

**Environmental Factors:** Information on environmental variables during pesticide application, including temperature, humidity, and airborne pesticide concentration.

Illustration: Temperature, Humidity, Pesticide

### Concentration.

Feature Category	Feature Description	Example
1. Demographics	Information about the worker's basic characteristics	Age, Gender, Occupation
		Example: Age (45), Gender (Male), Occupation (Farmer)
2. Pesticide Exposure	Details related to the worker's exposure to pesticides	Type of pesticide, Duration, Frequency
		Example: Type (Organophosphate), Duration (5 hours), Frequency (Weekly)
3. Symptoms	Physical symptoms exhibited by the worker	Dizziness, Nausea, Headache, Vomiting
		Example: Dizziness, Nausea
4. Medical History	Pre-existing health conditions that may affect susceptibility	Asthma, Respiratory issues, Diabetes
		Example: Asthma, Diabetes
5. Environmental Factors	Environmental conditions influencing pesticide exposure	Temperature, Humidity, Pesticide Concentration
		Example: Temperature (30°C), Humidity (60%), Pesticide concentration (High)

Figure 2.:Dataset

### Training

At this step, quality was evaluated based on the satisfaction derived from the outcomes achieved during training in conjunction with the domain expert. This phase pertains to the machine learning process, whereby several training sessions are executed, and the outcomes of these iterations are evaluated. Revisiting the refining phase is often necessary to enhance training outcomes.

Consequently, the DRC–SML model advocates for the documentation of training sessions, aiding the data scientist in evaluating the iterations that yielded optimal outcomes.

The DRC–SML model was originally developed to function alongside the RST algorithm as a machine learning instrument. Subsequent research assessing various machine learning methodologies, as outlined in Section II, may be used to improve the efficacy of

the suggested model. Models based on RST are recognized for generating explicit and interpretable decision rules, facilitating the study and validation of these rules by domain experts, and thereby improving transparency in the decision-making process.

The RST algorithm is acknowledged for its capacity to manage faulty, ambiguous, or noisy data, rendering it proficient at detecting pertinent information—typical situations faced in real-world applications, such as medical diagnostics, as shown in the PHH project setting. This scenario exhibits uncertainty in samples that include characteristics irrelevant to the diagnosis, null values, or those classified as "Not Informed," along with a diversity of kinds, where 16% are continuous attributes and 11% are multivalued attributes.

The RST is a supervised machine learning model that requires a precise delineation of condition characteristics and decision attributes. The training process using the RST algorithm comprises two separate phases: the generation of reducts and the extraction of rules. During the reduct generation phase, a search is performed for attribute subsets that possess the same equivalence relation and decision-making capability as the initial set of attributes provided for training. The capacity to make choices with fewer characteristics is advantageous, particularly when specific information is lacking. This allows for the selection of subsets containing the necessary information to get precise diagnoses using a more succinct data set. The elimination of extraneous features is crucial for streamlining the training process and producing coherent decision rules. This enhanced method enhances the analysis's efficacy, even with limited datasets, yielding more succinct rules and equally strong judgments.

### Recovery

At this point, fresh samples are included to provide further training aimed at refining the rules established for decision-making. This approach is conducted while maintaining adherence to the principles of privacy, copyright, and quality, as highlighted during the first data collecting phase. A data refinement form is used to get information, prevent the repeated collection of previously rejected data, and assure adherence to the discretization categories and criteria set during the refinement phase.

### STORAGE

The suggested architecture employs a relational database schema for data storage. In this framework, data are organized into subsets referred to as tables, which create relationships among them based on their intrinsic data.

Relational Database Management Systems (RDBMS) are essential in this setting, since they inherently support the four primary pillars defined below:

Privacy: RDBMS facilitates safe data access for numerous users with the use of passwords and access limits.

Integration: Data integration is accomplished by linkages formed between tables inside the RDBMS. Moreover, it may be seamlessly integrated with other storage models because to its capability to import and export files in text format.

Quality: The relational database model adheres to standards to guarantee data consistency and integrity. The RDBMS employs Structured Query Language (SQL) for data querying and management, facilitating rapid, secure, and high-quality information retrieval.

Preservation: The RDBMS includes a metadata file that offers descriptive details about the data. This function ensures efficient maintenance over time, consequently enhancing data comprehension and



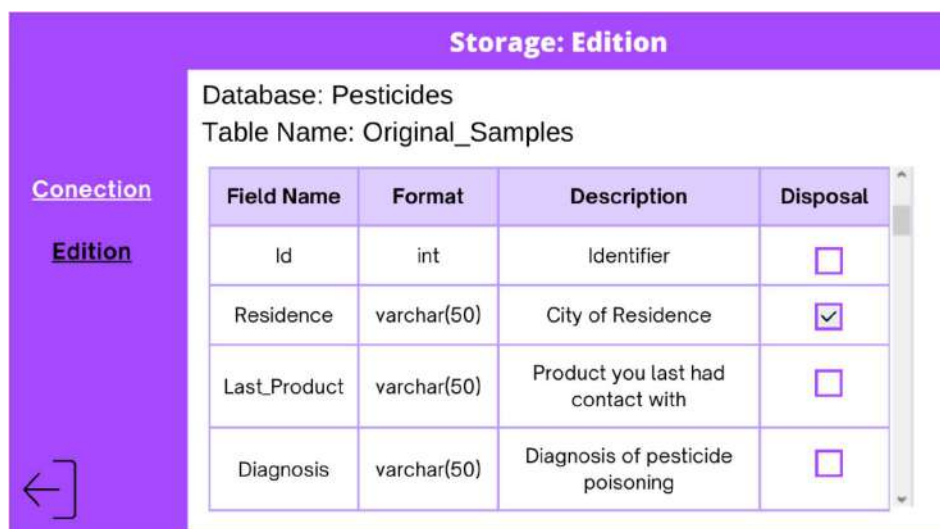
preservation.

Consequently, the DRC–SML methodology utilizes the relational database architecture with an RDBMS to guarantee the security, integration, quality, and preservation of stored data, facilitating analyses of the information inside the repository. This analysis facilitates the discovery of extraneous data that may be eliminated and, when required, the generation of data for absent but crucial information pertinent to the inquiry.

In the PHH project, the data acquired during the collecting phase were recorded in a spreadsheet. Nonetheless, this storage structure results in a lack

of interaction among the data, complicating the creation of visuals that facilitate diverse data analyses. To achieve the goals of the storage phase, the spreadsheet data was migrated to a relational database. Consequently, it became feasible to exclude information deemed extraneous to the diagnosis process, including address particulars and assessments of medical services.

Furthermore, critical components were integrated, namely a "ID" for the unique identification of each sample and "Diagnosis" for the classification and labeling of each sample.



**Storage: Edition**

Database: Pesticides  
Table Name: Original\_Samples

Conection  
Edition

Field Name	Format	Description	Disposal
Id	int	Identifier	<input type="checkbox"/>
Residence	varchar(50)	City of Residence	<input checked="" type="checkbox"/>
Last_Product	varchar(50)	Product you last had contact with	<input type="checkbox"/>
Diagnosis	varchar(50)	Diagnosis of pesticide poisoning	<input type="checkbox"/>

FIGURE 3. Example of Storage Form.

## CONCLUSION

This study introduces an innovative method for detecting pesticide toxicity in rural agricultural laborers using a supervised learning algorithm. Utilizing Random Forest and Decision Tree algorithms, we want to develop a prediction model that precisely identifies pesticide poisoning cases based on many criteria, including pesticide exposure levels, worker demographics, and reported symptoms. The suggested approach is expected to enhance the diagnostic process by providing a

quicker, more dependable, and data-driven alternative to conventional procedures. Furthermore, the model's capacity to provide real-time forecasts might assist healthcare professionals in remote regions, where resources and knowledge may be constrained. The data preparation, feature extraction, and model-building phases guarantee the system's ability to manage intricate and diverse datasets, hence facilitating improved health

outcomes for agricultural laborers susceptible to pesticide poisoning.

1. Z. Pawlak, *Rough Sets: Theoretical Aspects of Reasoning About Data*, 1st ed. Dordrecht, The Netherlands: Springer, 1991.
2. 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.
3. 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.
4. 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.
5. 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.
6. 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.
7. 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.
8. B. Mondal, *Artificial Intelligence: State of the Art*. Cham, Switzerland: Springer, 2020, pp. 389–425, doi: [10.1007/978-3-030-32644-9](https://doi.org/10.1007/978-3-030-32644-9).
9. 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.
10. 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.
11. 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.
12. J. D. Kelleher and B. Tierney, *Data Science*. Cambridge, MA, United States: MIT Press, 2018.
13. C. Silva, M. Saracee, and M. Saracee, “Data science in public mental health: A new analytic framework,” in *Proc. IEEE Symp. Comput. Commun. (ISCC)*, Jun. 2019, pp. 1123–1128.