Available online at www.jcsonline.in
Journal of Current Science & Humanities
11 (1), 2023, 1-08.

Impact Factor- 2.05

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

[1] G.Venkatesh, [2] B.Raghupathi,[3] K.Lakshmi Kalpana, [4] B.SAIKUMAR VARMA

[1,2,3] Assistant Professors,Department of Computer Science and Engineering, Kasireddy Narayanreddy College Of Engineering And Research, Abdullapur (V), Abdullapurmet(M), Rangareddy (D), Hyderabad - 501 505

[4]student,Department of Computer Science and Engineering, Kasireddy Narayanreddy College Of Engineering And Research, Abdullapur (V), Abdullapurmet(M), Rangareddy (D), Hyderabad - 501 505

ABSTRACT

Pesticide poisoning in rural labourers is being diagnosed in this research using ML algorithms. Because of its ease of use and readability, the Decision Tree algorithm has become a favourite. It creates a tree-like structure that can anticipate outcomes by splitting the data into subsets based on feature values. Because it works well with both numerical and categorical data, this model is great for pinpointing the factors that lead to pesticide toxicity. The RF technique, in contrast, is an ensemble approach that constructs several DT and then merges their predictions in order to decrease overfitting and increase accuracy. The improved prediction performance and stability offered by RF are the consequence of its capacity to aggregate the data from several trees. Providing a more stable and dependable model for the diagnosis of pesticide poisoning, it shines when dealing with complicated datasets with many characteristics. In order to mitigate the dangers that pesticide exposure poses to rural workers' health, both algorithms help build a better diagnostic tool.

I. INTRODUCTION

Data science and machine learning have recently become popular tools in the healthcare industry, with many seeing its use as a means to enhance decision-making and diagnostic precision. Predictive models have the potential to revolutionise many different fields, but public health is one that is particularly benefiting from these approaches. Pesticide poisoning is one of the most serious health problems that rural workers, especially those in the agricultural industry, confront. Acute poisoning and long-term chronic diseases are among the potential outcomes of pesticide exposure, which may occur by inhalation, skin contact, or consumption. Reducing the negative impacts of pesticide exposure requires prompt identification and treatment. A supervised learning model for the diagnosis of pesticide toxicity in rural labourers is proposed in this study. Classifying and predicting the risk of pesticide poisoning in persons working in rural agricultural areas will be accomplished by training the model using a mix of clinical data, environmental exposure measures, and health-related markers. We will use supervised learning, more especially classification algorithms, to look at data from employees who have been exposed to pesticides and had different degrees of poisoning symptoms. This project aims to develop a reliable diagnostic tool that may be used in healthcare settings in rural areas, where resources are typically few and where patients can benefit greatly from a quick diagnosis. By bringing new insights to healthcare providers and allowing for more focused treatments, data science has the ability to transform the way pesticide poisoning is diagnosed and treated.The study's results will hopefully help agricultural workers' health and the public health sector as a whole by providing a data-driven, scalable method for identifying occupational disorders in rural areas.By using various methods, particularly RF and DT algorithms, the proposed system seeks to improve the detection of pesticide toxicity in rural labourers. To forecast the probability of pesticide poisoning, this data-driven method trains a model using past health records, levels of environmental exposure, and employee demographics. In comparison to more conventional approaches, the system's ability to examine intricate data patterns and relationships will allow for a more rapid and precise diagnosis. An ensemble of decision trees will be created using Random Forest, which is recognised for its resilience against overfitting and its ability to handle huge datasets. These trees will then vote on the final categorisation. In contrast, decision trees will provide a straightforward and interpretable method of comprehending the interplay between many elements (such as worker health history or pesticide exposure) and the diagnosis. In rural areas where resources are limited, the suggested method will provide a robust and scalable approach for diagnosing pesticide toxicity by merging the two algorithms.

II. PROPOSED MODEL

A. Study Data

A variety of data points obtained from employees who could have been exposed to pesticides make up the dataset used for the diagnosis of pesticide poisoning in rural labourers. In this data set, you can find details on their demographics, medical history, symptoms, and exposure history. A machine learning model that can use the characteristics in this dataset to determine whether a worker has been poisoned by pesticides is the purpose of this dataset. A summary of the dataset's most salient characteristics follows:

Personal details: Factors that could affect the probability of poisoning, include gender, age, and profession. For instance, we may look at age, gender, and occupation type.

Information on the worker's exposure to pesticides, including the kind of pesticide, the amount of time exposed, and how often this occurred. Pesticide Type (e.g., organophosphates), Exposure Duration (e.g., 5 hours), and Exposure Frequency (e.g., weekly) are some examples.

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)

Physical symptoms that the employee can encounter include lightheadedness, queasy stomach, headaches, and nausea. Headache, Vertigo, Nausea, and Vomiting are among examples.

Workers with asthma, respiratory problems, or other long-term diseases that increase their risk of pesticide exposure should have their medical history reviewed. Case in point: respiratory illnesses, diabetes, and asthma. Environmental Considerations: Information on the weather, humidity, and pesticide concentration in the air that were present during the application of the pesticide. Consider the following: pesticide concentration, humidity, and temperature.

B) System Architecture

In order to diagnose pesticide toxicity in rural labourers, a new data science model is being developed. This model will use supervised learning and is intended to take many inputs, analyse them rapidly, and make accurate predictions. To guarantee trustworthy and practical insights for healthcare providers serving rural communities, this architecture incorporates data gathering, preprocessing, model training, assessment, and deployment.

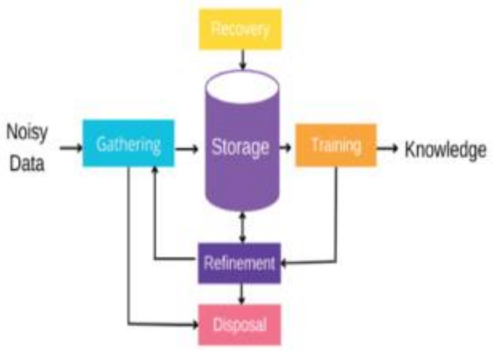

Fig1.System Architecture

The data collecting layer serves as the system's backbone, compiling pertinent information from a variety of sources including health records, surveys, environmental monitoring devices, and the hands-on experience of rural workers. Information gathered usually comprises a mix of demographics, symptoms, environmental variables, medical history, and history of pesticide exposure. Worker demographics such as age and gender, exposure time and pesticide type, reported symptoms like lightheadedness or headaches, and medical history (including asthma or diabetes) might all be included of the information. We also take into account environmental elements like weather, pesticide dosage, and use trends to give you a full picture of what might cause pesticide poisoning.Data cleaning, normalisation, and transformation into a format suitable for model building are the responsibilities of the data preprocessing layer when data collection is complete. This part of the process deals with outliers, missing data, and makes ensuring that categorical information (such pesticide kinds) are encoded correctly using techniques like one-hot encoding or label encoding. Because many machine learning methods rely on characteristics that are uniform in size, it is necessary to normalise or standardise numerical features. For supervised learning to take place, the dataset must be labelled with the worker's history of pesticide toxicity serving as the goal variable. Another crucial stage is feature engineering, which involves using domain knowledge to build additional variables (such total pesticide exposure or disease intensity) that might increase model accuracy.

Layers of supervised learning models form the backbone of the system. To understand the association between input characteristics (e.g., symptoms, pesticide exposure) and output (e.g., diagnosis of pesticide poisoning), mL algorithms like DT, RF, and SVM are used with labelled data. The probability of poisoning may be predicted by training the algorithm on past data that shows trends. By applying the learnt patterns to new instances, the supervised learning method guarantees that the model can generalise well to unseen data. To further evaluate the model's performance on new data and avoid overfitting, cross-validation methods are used.

To determine how well a taught model performed after training, an evaluation layer is essential. To measure how well the model can identify cases of pesticide poisoning, many measures are used, including recall, accuracy, precision, F1-score, and ROC-AUC. Because of the gravity of false negatives (the failure to recognise poisoning when it happens), accuracy and memory are of the utmost importance in medical diagnosis. You may use the evaluation results to choose the best model or see if you need to tweak the hyperparameters further.

C) Proposed Machine Learning-Based Model

Use of supervised learning methods to forecast the probability of poisoning from a variety of characteristics is the main emphasis of the suggested machine learning-based model for diagnosing pesticide poisoning in rural labourers. First, the model gathers extensive data from several sources, such as age, gender, environmental variables (such as temperature and humidity), symptoms (such as headache, dizziness, and nausea), and a history of pesticide exposure (including the kind, length, and frequency of exposure). To make sure these data points are machine learning ready, they are preprocessed with care. This involves tasks such as filling in missing values, encoding categorical variables (such as pesticide kind and symptoms) using approaches like one-hot encoding, and scaling numerical features (such as age and exposure time) to a consistent range. In order to train the model, it is necessary to have a clean, organised dataset with well represented characteristics. Choosing the right ml algorithm follows data preparation. To find out which algorithm works best for forecasting pesticide toxicity, we compare DT, RF, SVM, and NN. Input characteristics (such as pesticide exposure and symptoms) and the goal variable (such as a diagnosis of pesticide poisoning) are learnt by these algorithms from the previous data. After the best model has been evaluated and chosen, it is put into action in the field so that healthcare providers may enter pertinent data and get prompt forecasts. This ML model not only aids in the early diagnosis of pesticide toxicity, but it also lays the groundwork for further advances via retraining and constant data updates, guaranteeing its accuracy throughout time.

III. METHODOLOGY

Gathering Information: The first step is to collect information from pesticide-exposed rural labourers. Common sources for this information include health evaluations, medical records, and questionnaires. Workers' ages, occupations, pesticide exposure durations, ambient factors, and symptoms (such as vertigo, nausea, and headaches) are all important components. The labels of cases of pesticide poisoning are also gathered, which are essential for supervised learning. The capacity of the model to generalise and perform properly is heavily dependent on the quantity and quality of the data that was gathered.

An essential first step in developing machine learning models is data preparation, which involves cleaning and structuring raw data. At this stage, problems including noise, discrepancies, and missing data are addressed. Depending on the amount of missing data, missing values are either removed or impedanced. Using techniques like one-hot encoding or label encoding, numerical forms are created for categorical data like exposure level or work type. In addition, numerical features are standardised or normalised to make sure they are all on the same scale, which keeps the model from being too dependent on any one characteristic.

The extraction and selection of important characteristics for the prediction of pesticide toxicity is carried out at this step. To create a "risk factor" based on exposure history or symptoms, for example, feature extraction would include generating new features from domain information. In order to find and keep the characteristics that have the largest impact on the model's predicted accuracy, feature selection methods like correlation analysis, chi-squared tests, or mutual information are used. To make the model work better and use less computing power, unnecessary or redundant features are removed.

Building the Model: After the data has been cleaned up, ML methods are used to train the models. Based on the values of various attributes, the Decision Tree model branches the data into numerous branches, with judgements being made at each node. Using measures like Gini impurity or information gain, it selects the feature that efficiently partitions the data during training. Overfitting is a problem with Decision Trees, even if they are easy to comprehend and interpret. This happens especially with complicated models. This is addressed by using Random Forest.

The last step before a model is put into use in a real-world application is its deployment, which occurs after training and evaluation. This concept is used in a method that healthcare personnel or experts may use to diagnose pesticide toxicity. Predicting pesticide poisoning in real-time using worker profiles, exposure history, and symptoms is made possible by this approach. The system that was put into place is easy to use, can grow with the business, and can process data in real time. To make sure the model works well in the long run and can react to new data, it may be needed to retrain it and keep an eye on it.

IV.CONCLUSION

this experiment proves that ml can be used to diagnose pesticide toxicity in rural labourers using algorithms like Random Forest and Decision Tree. We have created an effective framework for forecasting health hazards associated with pesticides by following a methodical strategy that includes data collecting, preprocessing, feature extraction, model construction, and deployment. The input data was refined to concentrate on the most important variables by feature extraction and selection, which improved the model's accuracy. Overfitting was less of a concern since ensemble approaches like Random Forest improved the model's generalisability and resilience.

To improve worker safety and allow for quick interventions, the implemented model provides a viable and scalable solution for real-time pesticide poisoning diagnosis. Other rural health concerns where environmental exposure is a major factor may be addressed by expanding and adapting this strategy. Maintaining the system's accuracy and responsiveness to new data patterns will also need constant data collecting and model retraining. Early diagnosis and preventative health care for rural workers might be greatly improved with the integration of ml into public health monitoring systems.

V.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, pp. e862, Feb. 2022.

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

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

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. 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, [online] Available: https://www.sciencedirect.com/science/article/pii/S0952197620302529.

8. 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, pp. 377, Sep. 2021.

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

10.B. Mondal, Artificial Intelligence: State of the Art, Cham, Switzerland:Springer, pp. 389-425, 2020.