**Recent Development and Futuristic Trends of Data Mining in Agriculture**

Vishwa Gohil, Nitin Varshney, Alok Shrivastav, Yogesh Garde

Department of Agricultural Statistics, N. M. College of Agriculture, Navsari Agricultural University, Navsari, Gujarat

1. **INTRODUCTION**

Agriculture is a business with risk that is influenced by geography, politics, economics, climate, and other elements. There are some risks that can be estimated using sophisticated computational, statistical and mathematical techniques. Information extraction from agricultural databases is a challenge. Data mining is a method that can meet this challenge and provide the knowledge needed for agricultural improvement.

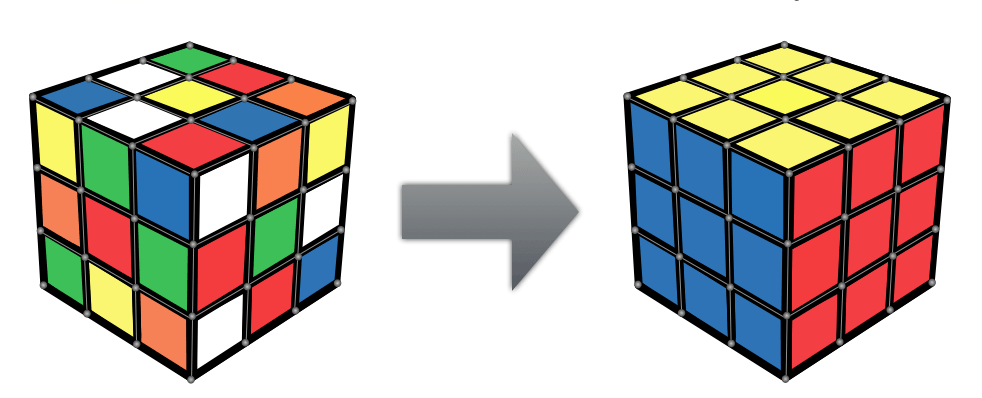
There is an essential desire to improve the  management of agricultural resources in light of the rising population and the corresponding requirement for rising agricultural productivity. Agriculture is the backbone of the Indian economy since it directly supports about two-thirds of the population. Given that only one-third of the area under cultivation is irrigated, agriculture has very low production. Therefore, to meet the rising demand for food, farmers, agricultural researchers and the government are working harder than ever to create methods to increase food production. But nowadays, only a few farmers use modern agricultural methods, tools and techniques for better agriculture productivity; instead of the majority of farmers still carry out agriculture-related duties by hand. Moreover, in traditional crop field management, uniform input application not only consider the concept of spatial and temporal variability within a crop field, but also results in environmental pollution and reduction of farm profits. The need of site-specific management or precision agriculture has been advocated by researchers, producers and farmers in the worldwide. Advanced information technology that can provide quick and cost-effective ways to identify spatial variability within crop fields is the basis of precision agriculture. Moreover, remote sensing technologies have advanced rapidly in recent years and have become effective tools for site-specific management in crop protection and production (Sindhu and Sindhu, 2017).

Agriculture's raw data is extremely diversified. For the development of an agricultural information system, it is important to assemble, organize and integrate it. In order to accumulate knowledge and trends, it is necessary to use information and communications technology, which will make it possible to extract important data from agriculture. In addition, it will play a role in the complete elimination of manual jobs. The production cost will be lower, the yield will be higher, and the market price will increase owing to easier data extraction straight from electronic sources and its transmission to secure electronic documentation systems. Data mining technique will provide information about crops and enable agricultural enterprises to predict trends about customer‟s conditions or their behaviour. It analyzes the data from different perspectives and helps in finding relationships in seemingly unrelated data. The computational needs of agriculture data and how data mining techniques can be used as a tool for knowledge management in agriculture should be considered by researchers. Data warehouses can be prepared to hold agriculture data, which makes transaction management, information retrieval and data analysis much easier (Sindhu and Sindhu, 2017).

Data mining has emerged as a valuable tool in agriculture, enabling farmers, agronomists and researchers to leverage the power of data to enhance agricultural practices and improve overall productivity. As data from sensors, satellites, weather stations, and other sources become more and more readily available, data mining techniques have become crucial in transforming agriculture into a data-driven industry (Majumdar,2017).

Data mining is a process of extracting useful patterns, insights and knowledge from large datasets. It involves analyzing vast amounts of data to discover hidden relationships, trends and patterns that can be valuable for making informed business decisions or gaining insights into various domains. The goal of data mining is to transform raw data into actionable information, , aiding decision-making and revealing valuable trends that may not be apparent through traditional data analysis methods. (Han *et al*., 2011)

Data mining plays a critical role in various industries, including finance, healthcare, retail, telecommunications, marketing and more. The insights gained from data mining can lead to more informed decision-making, enhanced business strategies, improved customer satisfaction and increased efficiency. In Data Mining, the goal is to split data in different categories, each of them representing some feature the data may have. (Han *et al*., 2011)



Data Mining is the knowledge discovering technique by analyzing the potentially interesting and unknown patterns in a large volume of datasets from summarizing the useful information from various perspectives.

Data Mining is the process of extracting useful and important information from large sets of data. Data Mining in agriculture field is a relatively novel research field. Yield prediction is a very important agricultural problem. Any farmer is interested in knowing how much yield he is about to expect. In the past, yield prediction was performed by considering farmer's experience on particular field and crop. Consider that data are available for some time back to the past, where the corresponding yield predictions have been recorded. In any of Data Mining procedures the training data is to be collected from some time back to the past and the gathered data is used in terms of training which has to be exploited to learn how to classify future yield predictions. (Ramesh and Vardhan, 2013).

Data Mining becomes popular in the field of agriculture for soil classification, wasteland management and crop and pest management. In assessed the variety of association techniques in Data Mining and applied into the database of soil science to predict the meaningful relationships and provided association rules for different soil types in agriculture. Similarly, agriculture prediction, disease detection and optimizing the pesticides are analyzed with the use of various data mining techniques earlier. (Rajeswari, and Arunesh, 2016)

Extracted knowledge from different data sets are used for better decision making in critical situations. Knowledge discovery in a database (KDD) is often called data mining. Extract knowledge which is included in the dataset and transferring that data into human understandable form to further use is an objective of the data mining. By doing data mining we can able to generate and build meaningful and knowledgeable data.

Data mining techniques are used in many fields to enhance the quality and importance of useful data. It plays a vital role in each respective field. Data mining especially in field of agriculture is a vast and valuable concept. Because people who belong to the agriculture field faces several issues like decreasing production due to unsuitable environmental conditions like flood, drought and many alternative natural reasons and rarely factors, so there is more scope for data mining.

Data mining tasks can be classified into two categories:

1. Descriptive data mining
2. Predictive data mining

Descriptive data mining tasks characterize the general properties of the data in the database while predictive data mining is used to predict the direct values based on patterns determined from known results. Prediction involves using some variables or fields in the database to predict unknown or future values of other variables of interest. As far as data mining technique is concern, in the most of cases predictive data mining approach is used. Predictive data mining technique is used to predict future crop, weather forecasting, pesticides and fertilizers to be used, revenue to be generated and so on. ( Kodeeshwari and Ilakkiya, 2017)

Databases today are huge. Now a day they are counted in Peta and exabytes. The corporate world is a cut-throat world so to challenge others Decisions must be made with maximum knowledge. This can be done by data mining.

**GOALS OF DATA MINING** (Anno., 2023a)

1. **Pattern Discovery:** To identify meaningful patterns, relationships and trends within the data that may not be immediately evident, helping organizations gain valuable insights.
2. **Predictive Modeling:** To develop accurate and reliable predictive models that can forecast future trends or outcomes based on historical data, enabling proactive decision-making.
3. **Classification:** To categorize data instances into predefined classes or categories based on their attributes, making it easier to make decisions and take appropriate actions.
4. **Clustering:** To group similar data instances together based on their characteristics, aiding in data exploration and revealing natural structures within the dataset.
5. **Anomaly Detection:** To identify rare or unusual patterns or data points that deviate significantly from the norm, which can be critical for detecting fraud or abnormal behavior.
6. **Recommendation Systems:** To provide personalized recommendations to users based on their preferences and past interactions, enhancing user experience and engagement.
7. **Optimization:** To optimize processes and resources by analyzing data and finding the best strategies or configurations for specific objectives.
8. **Text Mining and Sentiment Analysis:** To extract valuable information from unstructured text data, such as customer reviews, social media posts, or news articles, and understand sentiments or opinions.
9. **Association Rule Mining:** To discover interesting relationships or correlations between different items or attributes in transactional data.
10. **Decision Support:** To provide decision-makers with data-driven insights and support, facilitating informed and effective decision-making.

**DATA MINING PROCESS** (Anno., 2023b)

The data mining process typically involves several key steps:

1. **Data Collection:** The first step is to gather relevant data from various sources, such as databases, spreadsheets, web logs, social media, or any other data repositories.
2. **Data Cleaning:** Also known as data cleansing. Raw data is often messy, containing errors, missing values, or inconsistencies. Data cleaning involves preprocessing the data to remove noise and correct errors, ensuring data quality and reliability.
3. **Data Exploration:** In this stage, analysts explore the data visually and statistically to gain initial insights into its characteristics. They may use techniques like data visualization and summary statistics to understand the distribution and patterns within the dataset.
4. **Data Transformation:** Data is transformed into a suitable format for analysis. This may include normalizing or scaling numerical values, encoding categorical variables, or creating new features to enhance the data mining process.
5. **Choosing Data Mining Techniques:** Based on the objectives of the analysis and the nature of the data, appropriate data mining techniques are selected. These techniques can include association rule mining, classification, clustering, regression, and more.
6. **Data Mining Algorithms:** In this step, specific algorithms are applied to the prepared data to extract patterns and insights. Each algorithm is designed to solve specific types of problems and may require fine-tuning to achieve the best results.
7. **Interpreting Results:** The discovered patterns and insights are interpreted to gain meaningful knowledge and actionable information. Analysts need domain expertise to interpret the results accurately.
8. **Validation and Evaluation:** The mined patterns and models are validated and evaluated to ensure their accuracy and reliability. This step helps in assessing the performance of the data mining process.
9. **Deployment and Implementation:** Once the data mining process proves successful, the results are implemented into real-world applications, such as improving business strategies, enhancing customer experiences, optimizing processes, or making informed decisions.

**IMPORTANCE OF DATA MINING**

* **Precision Agriculture**

Data mining facilitates precision agriculture, where farmers can make informed decisions at a highly granular level. By analyzing data from various sources, such as soil sensors, weather forecasts, satellite imagery and crop yield data. Farmers can optimize irrigation, fertilization and pest control strategies tailored to specific areas of their fields. This targeted approach minimizes resource wastage and maximizes crop yields.

* **Crop Disease Detection**

Data mining techniques can analyze historical and real-time data related to crop health and environmental conditions. By identifying patterns and anomalies, data mining helps in early detection of diseases, pests, or nutrient deficiencies, allowing timely intervention and preventing widespread crop losses.

* **Crop Yield Prediction**

Through the analysis of historical data on weather patterns, soil quality and crop yields, data mining models can predict potential crop yields for future seasons. Such predictions aid in better planning, budgeting and marketing strategies for farmers and agribusinesses.

* **Market Analysis**

Data mining can analyze market trends, pricing data and consumer preferences, enabling farmers to make informed decisions about what crops to grow and how to adapt to market demands effectively.

* **Livestock Management**

Data mining techniques can be applied to monitor and analyze data from livestock, such as cattle or poultry. This helps in optimizing feed allocation, identifying health issues early and improving overall animal welfare.

* **Climate Resilience**

Data mining can be used to analyze climate data, identifying long-term trends and changes in weather patterns. This knowledge can help farmers adapt their practices to climate change and build climate-resilient agricultural systems.

* **Resource Management**

By analyzing data on water usage, energy consumption and resource allocation, data mining can assist in optimizing resource management on farms, leading to increased sustainability and cost-efficiency.

**SOURCES OF DATA FOR MINING** (Anno., 2023c)

1. **Data Warehouse**
   1. A data warehouse is defined as the collection of data integrated from multiple sources that will queries and decision making. There are three type datawarehouse: Enterprise  datawarehouse, Data Mart and Virtual Warehouse. Two approaches can be used to update data in DataWarehouse: Query-driven Approach and Update-driven Approach.
   2. Application in Business decision making, Data mining, etc.
2. **Transactional Database**
   1. Transactional databases is a collection of data organized by time stamps, date, etc to represent transaction in databases. This type of database has the capability to roll back or undo its operation when a transaction is not completed or committed. Highly flexible system where users can modify information without changing any sensitive information. Follows [ACID property](https://www.geeksforgeeks.org/acid-properties-in-dbms/) of DBMS.
   2. Application in Banking, Distributed systems, Object databases, etc.
3. **Multimedia Databases**
   1. Multimedia databases consists audio, video, images and text media. They can be stored on Object-Oriented Databases. They are used to store complex information in a pre-specified formats.
   2. Application: Digital libraries, video-on demand, news-on demand, musical database, etc.
4. **Spatial Database**
   1. Store geographical information. Stores data in the form of coordinates, topology, lines, polygons, etc.
   2. Application: Maps, Global positioning, etc.
5. **Time-series Databases**
   1. Time series databases contains stock exchange data and user logged activities.Handles array of numbers indexed by time, date, etc. It requires real-time analysis. This type of data is collected over time, such as stock prices, weather data, and website visitor logs.
   2. Application: eXtremeDB, Graphite, InfluxDB, etc.
6. **Multimedia databases**
7. A Multimedia database (MMDB) is a collection of related [multimedia](https://en.wikipedia.org/wiki/Multimedia) [data.](https://en.wikipedia.org/wiki/Data_%28computing%29) The [multimedia](https://en.wikipedia.org/wiki/Multimedia) data include one or more primary media [data types](https://en.wikipedia.org/wiki/Data_types) such as [text](https://en.wikipedia.org/wiki/Written_text), [images](https://en.wikipedia.org/wiki/Image), [graphic](https://en.wikipedia.org/wiki/Graphic) [objects](https://en.wikipedia.org/wiki/Object_%28computer_science%29)(including [drawings](https://en.wikipedia.org/wiki/Drawing), [sketches](https://en.wikipedia.org/wiki/Sketch_%28drawing%29), and [illustration](https://en.wikipedia.org/wiki/Illustration) [animation](https://en.wikipedia.org/wiki/Animation)s sequences, [audio](https://en.wikipedia.org/wiki/Sound) and [video](https://en.wikipedia.org/wiki/Video).
8. **WWW**
9. WWW refers to World wide web is a collection of documents and resources like audio, video, text, etc which are identified by Uniform Resource Locators (URLs) through web browsers, linked by HTML pages, and accessible via the Internet network. It is the most heterogeneous repository as it collects data from multiple resources. It is dynamic in nature as Volume of data is continuously increasing and changing.
10. Application: Online shopping, Job search, Research, studying, etc.
11. **Structured Data**
    1. This type of data is organized into a specific format, such as a database table or spreadsheet. Examples include transaction data, customer data and inventory data.
12. **Semi-Structured Data**
13. This type of data has some structure, but not as much as structured data. Examples include XML and JSON files and email messages.
14. **Unstructured Data**
15. This type of data does not have a specific format, and can include text, images, audio and video. Examples include social media posts, customer reviews, and news articles.
16. **External Data**
17. This type of data is obtained from external sources such as government agencies, industry reports, weather data, satellite images, GPS data, etc.
18. **Streaming Data**
19. This type of data is generated continuously, such as sensor data, social media feeds and log files.
20. **Relational Data**
21. This type of data is stored in a relational database and can be accessed through SQL queries.
22. **NoSQL Data**
23. This type of data is stored in a NoSQL database, and can be accessed through a variety of methods such as key-value pairs, document-based, column-based or graph-based.
24. **Cloud Data**
25. This type of data is stored and processed in cloud computing environments such as AWS, Azure, and GCP.
26. **Big Data**
27. This type of data is characterized by its huge volume, high velocity, and high variety, and can be stored and processed using big data technologies such as Hadoop and Spark.

**DATA MINING TECHNIQUES**

1. **Classification**

In this method, data is categorized into predefined classes or labels. It is used for tasks like email spam detection, sentiment analysis, or disease diagnosis, where the model needs to assign instances to specific classes based on their features.

1. **Clustering**

Clustering algorithms group similar data points together based on their features, without any predefined class labels. It helps in segmenting data and finding natural groupings within the dataset.

1. **Regression Analysis**

Regression is used to predict numerical values based on historical data. It's employed in tasks like sales forecasting, stock price prediction, or estimating the impact of variables on a particular outcome.

1. **Association Rule Mining**

This technique identifies interesting relationships or correlations between different items in a dataset. For example, in retail, it can help discover patterns like "Customers who buy product A are likely to buy product B."

1. **Anomaly Detection**

This method focuses on identifying unusual patterns or outliers in the data that deviate significantly from the norm. Anomaly detection is crucial for fraud detection, fault diagnosis, or identifying abnormal behavior in cybersecurity.

1. **Sequential Pattern Mining**

It discovers sequential patterns or temporal relationships within the data. This is often applied to areas like analyzing customer behavior patterns in a sequence of events.

**Fig. 1 Data mining techniques**

**ALGORITHMS**

1. **Supervised learning**

* A supervised learning algorithm analyzes the training data and produces an inferred function, which can be used for mapping new examples. An optimal scenario will allow for the algorithm to correctly determine the class labels for unseen instances. This requires the learning algorithm to generalize from the training data to unseen situations in a "reasonable" way.
* Eg.: Regression models, k-Nearest-Neighbor, Decision trees, Neural networks

1. **Unsupervised learning**

* It is the [machine learning](https://en.wikipedia.org/wiki/Machine_learning) task of inferring a function to describe hidden structures from unlabeled data. Since the examples given to the learner are unlabeled, there is no error or reward signal to evaluate a potential solution. This distinguishes unsupervised learning from [supervised learning](https://en.wikipedia.org/wiki/Supervised_learning) and [reinforcement learning.](https://en.wikipedia.org/wiki/Reinforcement_learning)
* Eg.: K-means clustering, Self-organized map, Association rule mining

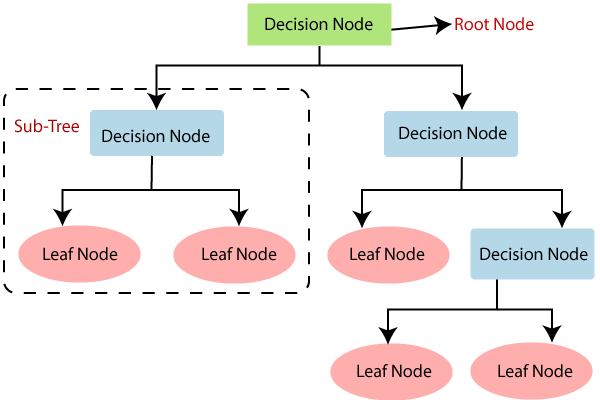
1. **CLASSIFICATION TECHNIQUES**

It is a classification technique based on machine learning. One of a predefined set of groups are classified into each time in a set of data. Software is developed in classification that can acquire information about how are data items classified into a group. The results of the classification and prediction models are evaluated according to their scalability, speed, predictive accuracy, robustness and interpretability ( Kodeeshwari and Ilakkiya, 2017).

The different classification techniques for discovering knowledge are Rule-Based Classifiers, Bayesian Networks (BN), Decision Tree (DT), Nearest Neighbor (NN), Artificial Neural Network (ANN), Support Vector Machine (SVM), Rough Sets, Fuzzy Logic, and Genetic Algorithms.

* 1. **Decision Tree**

The decision tree is the most robust classification technique in data mining. It is a flowchart similar to a tree structure. Here, every internal node refers to a test on a condition and each branch stands for an outcome of the test (whether it’s true or false). Every leaf node in a decision tree holds a class label. You can split the data into different classes according to the decision tree. It would predict which classes a new data point would belong to the created decision tree. Its prediction boundaries are vertical and horizontal lines.



**Fig. 2 Decision tree**

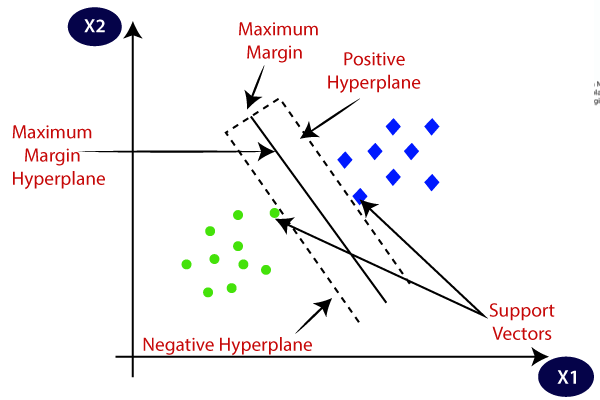
* 1. **Naive Bayes**

The naive bayes algorithm assumes that every feature is independent of each other and that all the features contribute equally to the outcome. Another assumption this algorithm relies upon is that all features have equal importance. It has many applications in today’s world, such as spam filtering and classifying documents. Naive bayes only require a small quantity of training data for the estimation of the required parameters.

A naive bayes classifier is significantly faster than other sophisticated and advanced classifiers. The naive bayes classifier is notorious for being poor at estimation because it assumes all features are of equal importance, which is not true in most real-world scenarios.

* 1. **Support Vector Machine**

The support vector machine algorithm also known as SVM, represents the training data in space differentiated into categories by large gaps. New data points are then mapped into the same space, and their categories are predicted according to the side of the gap they fall into. This algorithm is especially useful in high dimensional spaces and still effective in cases where a number of dimensions is greater than the number of samples.This algorithm lags in providing probability estimations. You have to need to calculate them through five-fold cross-validation, which is highly expensive.



**Fig. 3 Support vector machine**

* 1. **K-Nearest Neighbour**

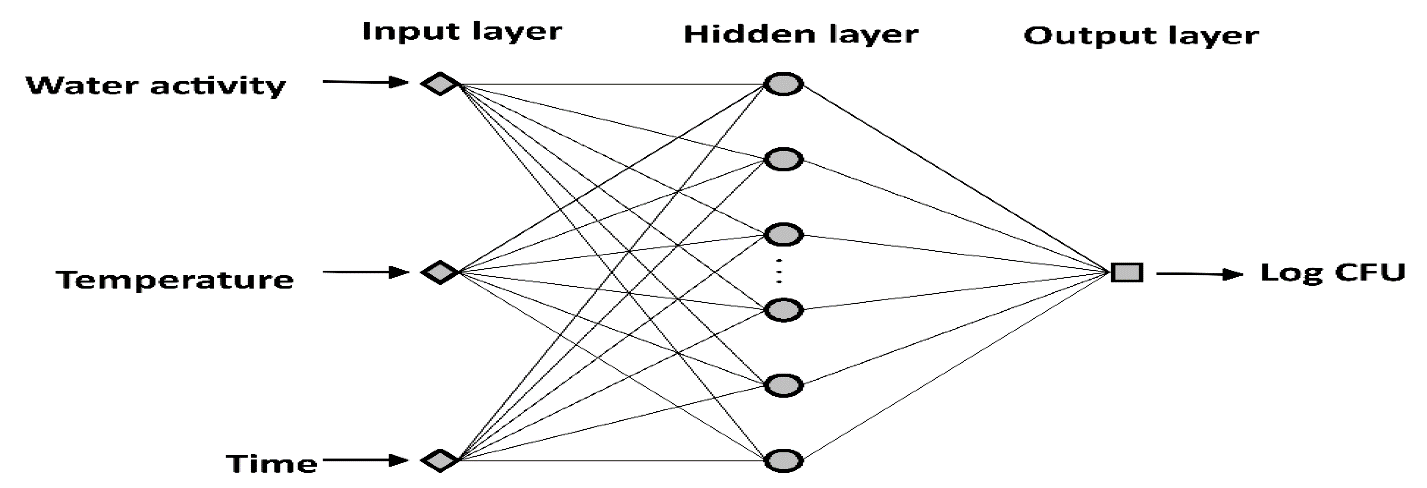
The k-nearest neighbor algorithm has a non-linear classifier. It predicts the class of a new test data point by finding its k nearest neighbors' class. You have selected the k nearest neighbors of a test data point by using the Euclidean distance. In the k nearest neighbors, you have to count the number of data points present in different categories, and you have to assign the new data point to the category with the most neighbors. It’s quite an expensive algorithm as finding the value of k takes a lot of resources. It also has to calculate the distance of every instance to every training sample, which further enhances its computing cost.

* 1. **Artificial Neural Network**

It can also help in the prediction of rainfall using artificial neural networks and other applications of artificial neural networks in hydrology are forecasting daily water hassle and flow forecasting

Inoduction

Modern agriculture needs to have a high production efﬁciency combined with a high quality of obtained products. This applies to both crop and livestock production. To meet these requirements, advanced methods of data analysis are more and more frequently used, including those derived from artiﬁcial intelligence methods. Artiﬁcial neural networks (ANNs) are one of the most popular tools of this kind. They are widely used in solving various classiﬁcation and prediction tasks. For some time they have also been used in the broadly deﬁned ﬁeld of agriculture. They can form part of the precision farming and decision support systems. Artiﬁcial neural networks can replace the classical methods of modeling and are one of the main alternatives to classical mathematical models. The spectrum of the applications of artiﬁcial neural networks is very wide. For a long time now, researchers from all over the world have been using these tools to support agricultural production, making it more efﬁcient and providing the highest-quality products possible. The purpose of this Special Issue was to publish high-quality research and review papers that cover the application of various types of artiﬁcial neural networks in solving relevant tasks and problems of what is widely deﬁned as agriculture.



**Fig. 4** **Artificial neural network**

1. **CLUSTERING TECHNIQUES** (Kodeeshwari and Ilakkiya, 2017)

Clustering is a data mining technique which is used to group the set of data objects into multiple clusters [meaning sub-classes] . A sub-set of object which are similar is called a cluster. High similarity occurs when the objects are in same clusters and the objects are dissimilar in other clusters. Similarities and dissimilarities are evaluated by describing the objects based on attribute value. Algorithm of clustering are used in following steps such as for identifying the data, analyze the data, data refinement, model construction, detection of out structure.

Clustering is a data mining technique which maps the similar instance together, and dissimilar instance together, and dissimilar instance belong to diverse group based on data instance. The data instance are divided into subsets. To identify different information clustering technique is used because it correlates with examples where similarities and ranges agree. In this technique there is no need of prior knowledge about data. Clustering technique comes under unsupervised learning that takes unlabeled data records and differentiate them and for processing data. Clustering consists a cluster centre that contains all the clusters. A well defined clustering method will generate a high quality clusters.

1. Inter class[similarity low]
2. Intra class[similarity high]

A standalone data mining tool in cluster analysis or preprocessing step for various algorithms In order to achieve data distribution. The term clustering is also called as unsupervised learning. “hidden patterns” are used in cluster analysis for machine learning. Clustering is simply defined as more number of attributes with large datasets. Clustering algorithm was brought into life for rapid growth in text mining. Spatial database and information retrieval into various clusters. Since on spatial data for optimum clusters there undergoes continuous research in data mining. Because of this clustering is an issue till dated in data mining. One of the first step in data mining analysis is clustering. For example, in an industry with a group of employees may need to know about the various works in their projects in order to check what are all products are completed and to be delivered and which are the project yet to be modified and delivered to the customers. Clustering is a technique mainly used in agriculture science, monitors the quality of water change, and in precision agriculture to produce high yield. clustering is classified based on various methods such as

* + - 1. Density based method
      2. Partition based method
      3. Hierarchical based method

To make the concept clearer, we can take book management in the library as an example. In a library, there is a wide range of books on various topics available. The challenge is how to keep those books in such a way that readers can take several books on a particular topic without hassle. By using the clustering technique, we can keep books that have some kinds of similarities in one cluster or in one shelf and then label it with a meaningful name. If reader's want to grab books in that topic, they would only have to go to that shelf instead of looking for the entire library

The following are some characteristics that should have a good clustering algorithm.

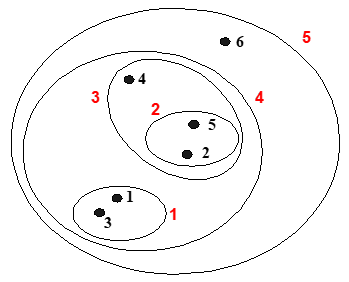
1. A clustering algorithm should easily interpretable and usable.
2. It should have high dimensionality.
3. It can partition constrained-oriented data.
4. It should be scalable.
5. It can produce uninformed-shaped clusters.
6. It is capable to deal the data which has different kinds of attributes.
7. It should take less domain knowledge.
8. It should handle noise data.

The clustering is used to determine a new set of groups, these groups are of significance in themselves, and their assessment is intrinsic.

* 1. **Hierarchical Clustering**

It involves creating clusters that have a predetermined ordering from top to bottom. **Agglomerative methods (bottom-up)**: The two ‘closest’ (most similar) clusters are then combined and this is done repeatedly until all subjects are in one cluster.

**Divisive methods (top-down):** In which all subjects start in the same cluster and the above strategy is applied in reverse until every subject is in a separate cluster. Produces a set of nested clusters organized as a hierarchical tree. Can be visualized as a dendrogram.

**Fig. 5: Agglomerative methods (bottom-up) Fig. 6: Divisive methods (top-down)**

* 1. **Partitioning Clustering**

It decomposes a data set into a set of disjoint clusters. Given a data set of N points, a partitioning method constructs K (N>=K) partition of the data, with each partition representing a cluster.

* 1. **Density Base Clustering**

A cluster is a dense region of points, which is separated by low-density regions, from other regions of high density. Used when the clusters are irregular or intertwined, and when noise and outliers are present.

**Fig. 7: Density base clustering**

1. **ASSOCIATION RULE MINING** (Kodeeshwari and Ilakkiya, 2017)

Association rule mining is a technique in data mining was developed by agrawal, imielinski and swami in 1993. This is one of the well organized technique of data mining to search the hidden or desired pattern among of data. The main focus in this method is to find relationship between various item in the relational database. Association rules are used to discover rules and to find the elements which occur recursively in a dataset consisting more absolute selections of element.

Association is a data mining technique that determines the possibility of the items which are co-occurred in a collection of data. Association rules are defined as the relationship between the co-occurring items. Hence the “sales transactions” are frequently analyzed using this technique. For processing numerical data association is the best technique. In a given set of transactions, find rules that will indicate the occurrences of an item based on the occurrences of other item. The issue to find all associated rules that satisfy minimum support for which user has specified. Association technique is known best and a straight forward data mining technique. Strength measures of rules can be defined using two rules

For example, Association expression between the items A and B are in the form A => B, the meaning of such a rule is that transactions that contain A also contain B within the database. When we use association rules, we should consider two important basic rule measures such as support and confidence. These two measures have some threshold value to mine interesting patterns from a huge collection of data. Naturally, the support is the real value and confidence is the probability value. The rules which do not satisfy the threshold value are called uninteresting. To measure the interestingness of an association rule, the rule must be simple, with certainty and utility.

The rule basic measure support and confidence between the items A and B is defined as

In association rule mining, the confidence value is 100% then the analyzed data is always correct; such rules are called exact. Strong association rules have the minimum support threshold and minimum confidence threshold.

The association rules are categorized depending on the following principles:

1. Data dimensions involved in the association rule.
2. Type of data values in the rule.
3. Level of abstraction of data incurred in the rule.
4. Depending on other extensions of data.

An association rules exposed in various areas like, market basket analysis, customer segmentation, catalog design, store layout and telecommunication alarm prediction.

1. **REGRESSION** (Kodeeshwari and Ilakkiya, 2017)

Regression is a data mining function that is used to predict a numeric or continuous value. Regression analysis is a predictive modelling technique which gives the relation between the independent variable (y) and dependent variable (x). The variable that is been predicted are dependent variable and the variable which are predicted is used to predict the values of dependent variable is called independent variable. The important tool for analyzing and modelling data is the regression.

Regression also represents the variations that occurred in one variable following variations in another variable. One important issue regarding regression is it describes the relationship among the variables in detail correlation – it represents the strength of the relation between the variables. Regression tasks are often treated as classification tasks with quantitative class labels.

The methods for prediction are linear regression (LR) and nonlinear regression (NLR).

Where, a = Intercept

b = Slop

ε = Error

y = Dependent Variable

X = Independent Variable

**COMPARISON OF DATA MINING TECHNIQUES**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Differentiator** | **Classification** | **Clustering** | **Association Rules** | **Regression** |
| Methods | Predictive method | Descriptive method | Descriptive method | Predictive method |
| Usage | Used to predict the instance class from a pre- labeled instance. | Used to find the "natural" grouping of instances given unlabeled data. | Used to discover interesting relations between variables in large DB's. | Used to predict the instance class from a pre-labeled instance. |
| Algorithm | \*Decision trees.  \*ANN  \*Bayesian Classifier  \*K-nearest  \*Support vector  \*LAD Tree \*LWL  \*J48  \*Jrip | \*Hierarchical clustering  \*Partition clustering  \*Density clustering  \*Model based clustering methods  \*Grid-based methods  \*Soft-computing methods  \*Elbow method  \*Silhotte method  \*Gap statistics method | \*Multi – rule mining  \*Apriori algorithm,  \*Dynamic Hashing Pruning  \*Dynamic Item set counting  \*FP Growth (improve version of Apriori algorithm) | \*Linear regression  \*Non-Linear regression |
| Data needs | Labelled samples | Unlabeled samples | Labelled samples | Labelled samples |
| Learning method | Supervised learning | Unsupervised learning | Unsupervised learning | Supervised learning |

|  |  |  |  |
| --- | --- | --- | --- |
| **Area of Application** | **Major Contributions** | **Tools Used** | **Year** |
| Web Based Tomato expert Information System | Based on information from different species the expert system decides the disease and displays its control measure of disease | ID3 Decision Tree Algorith  Optimization Algorithm | 2010 |
| Applying Machine Learning for culling less productive cows | The computer-generated rules outperformed the expert-derived rules. They gave the correct disease top ranking just over 97% of the time, compared to just under 72% for the expert derived rules | WEKA | 1994 |
| Bayesian Classification for Rice Paddy distributions | Interpreting paddy distributions using multitemporal imageries together with cadastre GIS by Bayesian posteriori probability classifier | Bayesian Classification | 2002 |
| Induce a classification system capable of sorting mushrooms into quality grades | The average accuracy of the models was compared favorably with that of the human inspectors and the level of agreement with the human experts was, on average, acceptable | WEKA | 2000 |
| Effect of Pesticides on Humans | Icon based technique which uses features in cartoon-like human faces, each representing variables in order to depict multivariate data | Chernoff faces COF Clustering Tool | 2010 |
| Geospatial Data Mining Techniques | Application of computational characteristic to the needs of agriculture data, as they are uncertain and fundamentally seasonal so use of data mining techniques be helpful in some aspect of agriculture | Knowledge Discovery from Databases OLAP - Online Analytical Processing | 2013 |

**WEKA TOOL**

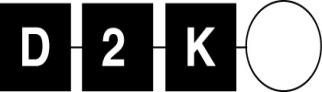
A popular suite of machine learning software written in Java was developed at the University of Waikato, New Zealand. Free software is licensed under the GNU General Public Licence. Contains a collection of visualization tools and algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access to these functions. The original version was primarily designed as a tool for analyzing data from agricultural domains.

**ADVANTAGES OF WEKA INCLUDE**

* Free availability under the GNU General public license.
* Runs on almost any modern computing platform.
* A comprehensive collection of data pre-processing and modeling techniques.
* Ease of use due to its graphical user interfaces.
* Weka supports several standard data mining tasks, more specifically, data pre-processing, clustering, classification regression, visualization,and feature selection.

**COMMERCIAL TOOLS AND PROGRAMMED**

Oracle Data Miner[http://www.oracle.com](http://www.oracle.com/)

Data ToKnowledge

[http://alg.ncsa.uiuc.edu](http://alg.ncsa.uiuc.edu/)

SAS

<http://www.sas.com/>



Clementine<http://spss.com/clemetine/>

IntelligentMiner

<http://www-306.ibm.com/software>

**APPLICATION OF DATA MINING IN AGRICULRURAL**

1. **Segregation of fruits and vegetables based on water level content**

On a normal basis, fruits and vegetables are classified based on their size and color into different categories which determines its cost. However, these are external factors and they don’t really contribute to determine the quality of fruits and vegetables. It mainly depends on the water level content in the fruit, i.e. the amount of water present in the fruit to its weight. Water level in a particular fruit or vegetable can affect its course of life to a certain extent, also abnormal water content can deteriorate the quality of neighboring fruits and vegetables in packaged boxes. Data mining can help us resolve this problem where images of fruits and vegetables are captured at the packaging line, these images are further processed to generate a good guess of the quality of product. (Mirjankar and Hiremath, 2016)

Moreover, records of variety of specimens help to generate a more accurate prediction of the quality of fruits and vegetables. These images can be fed to VGG 19 model, it is a 19 layer deep Convolutional layer used for large scale image recognition. The image of 224 X 224 RGB image is fed to VGG19 model and at the output layer we use ‘Softmax’ as an activation function that would give quality rating for the input image in the range of 10 output labels. The training of the model is required with the images of fruits taken with labels provided to them as the rating out 10 by a human.

1. **Using data mining to maximize yield depending on the quality of the soil**

Cultivation of a particular crop on any land which doesn’t meet the minimalistic requirements of the crop would generate yield of lower quality and less revenue for the farmer. Determination of quality of the soil is pre-requisite in agriculture. This gives an analysis of proportions of nutrients and minerals present in the soil. Quality of the soil depends on factors like alkalinity, salinity, moisture contents etc. Data mining is used to study various natures of soil. Soil data analysts suggest the type of crop to be grown and harvested depending upon the fertility of soil which would generate maximum yield. Data mining provides a large set of data for different varieties of soil which can help to predict several traits for cultivation depending on the season and climatic conditions. Implementation of data mining techniques with a wider range of statistical and analytical data and improvises accuracy in extracting information and can also automate results for generic cases (Patel and Patel, 2014). Data mining can also be used to study cross cultivation where in different crops which can be grown simultaneously which would bring in more revenue than single crop cultivation utilizing resources to the best possible extent without affecting the fertility of soil. The scope of data mining is large and its scope can be seen on the soil analysis as follows

1. Crops can be adopted by sensing and detecting soil which can be done by Artificial Neural Networks.
2. Previously unknown patterns of soil can be discovered
3. Traits and behavior of soils can be predicted on the basis of climate conditions and ingredients.
4. Testing of soil fertility can be done by statistical methods (Palepu and Muley, 2017).
5. **Using data mining to achieve accurateness in agriculture using data mining techniques**

Excessive use of pesticides hampers the overall agricultural productivity and in order to tackle this issue there is a need to minimize the use of pesticides in agriculture. Data mining can be used to design automated systems to detect weeds growing in fields (Tellaeche et al.,2007). This uses image processing mechanism and it primarily depended on aspect ratios, shapes and surface area. Later on, images of area under cultivation are being processed to find weed patches using specific algorithms (Yang, 2003). Color density in the images is used to represent the density of growth of crops in a particular area where the irregular growth of crops was represented by a different color.

1. **Analyzing performance of chicken by neural network models**

Neural network is a set of artificial neurons which is functional replica of biological neurons interconnected to form different layers of neurons. The first neural layer is said to be as input layer and the last layer as output layer, and the layers between them is called as hidden layers. The neural networks containing multiple hidden layers is called as Deep Neural Network. This setup of neurons can be used in many prediction and classification problems by training the neural network with previous records. With slight hyper-parameter tuning these type of trained networks can achieve human level accuracy or more but eventually less then 100% accuracy. This analysis of artificial neural networks on feeding Efficiency and weight gain from set of data proposed that concentration of dietary protein is more significant than threonine concentration. A study was proposed that diet containing 0.73% threonine and 18.69% protein may lead to producing weight gain, while on the other hand the efficiency can be achieved from the above data with 0.2 or 0.3 percent standard deviation.

1. **Optimizing the use of pesticides by data mining**

Excessive use of pesticides can harm the farmer in multiple ways . In agriculture, crop yield forecast is a very important problem. Agriculture researchers conducted a recent study which have shown that in order to maximize the crops yield,the pesticides are over used which is extremely dangerous for the environment. Also, excessive pesticide usage may lead to immunity in pests, which ultimately makes them more harmful to crops and less susceptible to reduction. As a result the overuse of pesticides is creating a health hazard and is imposing a financial burden on farmers and their family. With the help of clustering, one of the method of data mining which can cluster the features by giving interesting patterns of farmer practices and thus provide useful information which will highlight the detrimental effect of excessive pesticide use (Paulson, 2015). Advanced recent concept of spatial correlation has thoroughly impacted the yield prediction in a positive way. For development of forecasting and forewarning models of plant diseases the Artificial Neural Networks plays a very important role. Independent Component Analysis is used for the extraction of independent sources as it is a signal processing technique. It is a numerical and statistical technique for finding hidden characteristics that subtend signals, measurements and various sources (Bhagawati, 2016). A generative model is proposed for the obtained or observed data..It helps to identify patterns of weather data for optimization of pesticides usage method like integration of agricultural data is more often used as it includes pest scouting, pest usage, and meteorological recording.

1. **Prediction of problematic wine fermentations**

Wine is usually produced all around the world. The process of fermenting the wine is very important, because it can implement the productivity of wine-related industries and also the quality of wine. If we could predict how the fermentation is going to be at the early stages of the process, we could interfere with the process in order to guarantee a regular and smooth fermentation. Fermentations are studied by using different techniques, such as, for example, the k-means algorithm, and atechnique for classification based on the concept of biclustering (Paulson, 2015).

1. **Detection of diseases from sounds issued by animals**

The detection of animal's diseases in farms can impact positively the productivity of the farm, since sick animals can cause contaminations and spread contagious diseases. Moreover, the early detection of the diseases can allow the farmer to cure the animal as soon as the disease appears. Sounds made by pigs can be analyzed for the detection of diseases. In particular, their coughs can be studied, because they indicate their sickness content. A computational system is under development which is able to monitor pig sounds by microphones installed in the farm, and which is also able to discriminate among the different sounds that can be detected (Paulson, 2015).

1. **Optimizing pesticide use by data mining**

Recently conducted studies by agriculture researchers showed that attempts of cotton crop yield maximization through pro-pesticide state policies have led to a dangerously high pesticide use. These studies have reported a negative correlation between pesticide use and crop yield procedure. Hence excessive use of pesticides is harming the farmers with adverse financial situations, environmental and social impacts. By data mining the cotton Pest Scouting data was shown that how pesticide use can be optimized (reduced). Clustering of data revealed interesting patterns of farmer practices along with pesticide use dynamics and hence help identify the reasons for this pesticide abuse (Paulson, 2015).

1. **Sorting apples by water cores**

Before reaching the market, apples are checked and the ones showing some defects are removed. However, there are also unseen defects, that can spoil the apple flavor and look. An example of unseenable defect is the watercore. This is an apple disorder that can affect the longevity of the fruit. Apples with few or mild watercores are sweeter, but those apples with moderate to severe degree of watercore cannot be stored for long length of time. In addition, a few fruits with severe watercore could spoil a whole batch of apples. Due this reason, a computational system is undergoing study which takes Xray photographs of the fruit while they run on conveyor belts, and which is also able to analyse (by data mining techniques) the taken pictures and estimate the possibility that the fruit contains watercores (Paulson, 2015).

|  |  |
| --- | --- |
| **Methodology** | **Applications** |
| K-means | Forecasts air pollution, soil classification in combination with GPS |
| k-nearest Neighbour | Simulating daily precipitations and other weather |
| Support Vector Machine | Analysis of different possible change of the weather scenario. |
| Decision Tree Analysis | Prediction of soil depth |
| Unsupervised Clustering | Generate cluster and determine any existence of pattern. |
| WEKA Tool | Classification system for sorting and grading mushrooms. |
| Artificial neural network | Discriminating between good and bad fruits. |

**Fig 7: Data mining methodologies and its uses in agriculture**

**CURRENT ISSUE OF DATA MINING IN SMART AGRICULTURE**

There are still some challenges in data mining applications that need to be overcome.

* **More generalization:** Data mining techniques are not specified for the particular field or data.
* **Require special knowledge:** A major issue that is attached with the data mining is the results obtained the applications need special knowledge or experience to identify and utilize the given data volume.
* **Test against the small size of data:** Most of the existing systems considered only part of the dataset or small size of the dataset to determine the result. It is limiting the efficiency of crop yield prediction. Therefore, there is a vacancy for approaches that are increasing the efficiency of crop yield prediction by considering the entire data set.
* **High Computational Cost:** It is to address the agricultural problems presents set of the data mining techniques such as clustering, classification, k-Nearest Neighbor, k-Means. The problem of the k-Means algorithm is there can be a choice of various parameters ‘k’. Also, it takes the cost of computational. The WEKA (Waikato Environment for Knowledge Analysis) system has not been supported to handle multiple algorithms at the same time. It needs more time and a high computational cost.
* **Limited resources in India:** Most of the farmers have a small area of land. It is difficult to use applications, machines produced for the large agricultural field. Less coverage of the Internet limits the usage of IT application. Less availability of the infrastructure facilities.

**CONCLUSION**

* Agriculture is the most important application area mainly in the developing countries like India. Data mining not only gives us information about particular task but there are chances where the obtained information can be changing according to the modern technologies emerging widely. Use of information technology in agriculture can change the condition of decision making and farmers can yield in better way. Data mining plays a crucial role for decision making on several issues related to agriculture field. Additionally, user-focused access to hidden patterns in data is provided by data mining. For a variety of purposes, including problem prediction, disease detection, pesticide optimisation, and other fields, agricultural institutions apply data mining applications. Therefore, we may conclude that data mining has benefited the agricultural industry.

**REFERENCE**

Annonymous (2023b) Data Mining: Process, Techniques and Major Issues In Data Analysis, <https://www.softwaretestinghelp.com/data-mining/> > accessed on 18 July, 2023.

Annonymous, (2023a).Goals of data mining. Available at <https://people.cs.pitt.edu/~chang/156/21mining.html> > accessed on 18 July, 2023.

Annonymous, (2023c).Sources of data for mining. Available at [https://www.geeksforgeeks.org/types-of-sources-of-data-in-data-mining**/**](https://www.geeksforgeeks.org/types-of-sources-of-data-in-data-mining/) **>** accessed on 18 July, 2023.

Bagal, Y. V.; Pednekar, S. V.; Pandey, A. R. and Dhamdhere, T. B. (2020). Data Mining in Agriculture-A Novel Approach. *International Journal of Engineering Research and Technology,* **9**(8): 213-215

Bhagawati, K.; Sen, A.; Shukla, K. K. and Bhagawati, R. (2016). Application and scope of data mining in agriculture. *International Journal of Advanced Engineering Research and Science,* **3**(7): 236783.

Han, J., Kamber, M. and Pei, J. (2011). Data reduction. *Data Mining, Concepts and Techniques, 3rd ed.; The Morgan Kaufmann Series*, 99-110.

Kodeeshwari, R. S. and Ilakkiya, K. T. (2017). Different types of data mining techniques used in agriculture-a survey. *International Journal of Advanced Engineering Research and Science*, **4**(6): 17-23.

Majumdar, J., Naraseeyappa, S. and Ankalaki, S. (2017). Analysis of agriculture data using data mining techniques: application of big data. *Journal of Big data*, *4*(1):20.

Mirjankar, N. and Hiremath, S. (2016). Application of data mining in agriculture field. *International Journal of Computer Engineering and Applications*, **10(**6).

Palepu, R. B. and Muley, R. R. (2017). An analysis of agricultural soils by using data mining techniques. *International Journal of Engineering Science Computer*, ***7***(10).

Patel, H. and Patel, D. (2014). A brief survey of data mining techniques applied to agricultural data. *International Journal of Computer Applications,***95**(9).

Paulson, S. (2015). A Survey on Data Mining Techniques in Agriculture International Journal of Engineering Research and Technology, **3**(30).

Rajeswari, V. And Arunesh, K. (2016).Analyzing Soil Data Using Data Mining Classification Techniques. *Indian Journal Of Science And Technology,*9(19): 1-4.

Ramesh, D. and Vardhan, B. V. (2013). Data mining techniques and applications to agricultural yield data. *International journal of advanced research in computer and communication engineering*, ***2***(9), 3477-3480.

Sindhu, S. and Sindhu, D. (2017). Role of data mining techniques in agriculture improvement. **6**: 654-663.

Tellaeche, A.; BurgosArtizzu, X. P.; Pajares, G. and Ribeiro, A. (2007). A vision-based hybrid classifier for weeds detection in precision agriculture through the Bayesian and Fuzzy k-Means paradigms. *Innovations in hybrid intelligent systems*, 72-79.

Yang, C. C.; Prasher, S. O.; Landry, J. A. and Ramaswamy, H. S. (2003). Development of an image processing system and a fuzzy algorithm for site-specific herbicide applications. *Precision agriculture*, 4: 5-18.