RECENT DEVELOPMENT AND FUTURISTIC TRENDS OF DATA MINING IN AGRICULTURE

Abstarct

Data mining is pivotal in revolutionizing agriculture by extracting valuable insights from vast and diverse datasets. Key data mining methods, including classification, clustering, and regression, are discussed in relation to critical tasks such as crop yield prediction, disease detection, and precision agriculture. This chapter delves into the applications of data mining techniques in agriculture, emphasizing their potential to enhance decision-making, optimize resource usage, and elevate overall crop productivity. Leveraging diverse data sources like satellite imagery, weather data, and sensor networks, the study focuses on gathering information about soil conditions, crop health, and environmental factors. The integration of machine learning algorithms facilitates the identification of patterns, trends, and correlations within these datasets, providing actionable intelligence for farmers and researchers. The chapter also addresses challenges associated with handling agricultural data, such as issues related to data quality and scalability. Emphasis is placed on the role of data mining in promoting sustainable farming practices, mitigating environmental impact, and tackling global food security challenges.

Keywords: Data mining, Methods of data mining, Applications of data mining

Authors

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

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

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

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

I. 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.

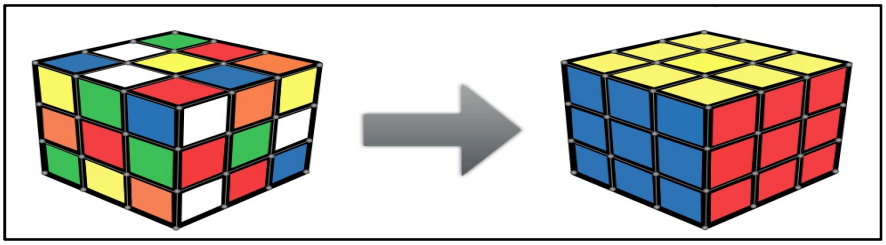
There is an essential desire to improve the the management of agricultural resources in light of the rising population and the corresponding requirement for rising agricultural productivity. Agriculture is the backbone of the Indian economy since it directly supports about two-thirds of the population. 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. In addition, uniform input application in traditional agricultural field management not only ignores the idea of spatial and temporal heterogeneity within a crop field but also has negative effects on the environment and farm income. Researchers, producersand farmers all around the world have argued for the necessity of site-specific management or precision agriculture. Precision agriculture is based on cutting-edge information technology that can offer rapid and affordable techniques to discover spatial heterogeneity within crop fields. Additionally, recent technological advancements in remote sensing have made them useful tools for site-specific management in crop production and protection(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. Agricultural businesses will be able to forecast trends concerning consumer conditions or behaviour thanks to data mining techniques, which will supply information about crops. It examines the data from many angles and assists in identifying connections between what at first glance may seem to be unconnected facts. Researchers should take into account the computing requirements of agricultural data and how data mining techniques might be employed as a tool for knowledge management in agriculture. Transaction management, information retrieval and data analysis are greatly facilitated by the ability to create data warehouses to store farm data. (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.The objective of data mining is to divide data into groups, each of which stands for a feature that the data may have(Han *et al*.,2011).



Data mining is a process for knowledge discovery that examines potentially interesting and undiscovered patterns in a vast amount of datasets by condensing the relevant information from several different perspectives.

The practice of obtaining valuable information from huge data sets is known as data mining. Agriculture-related data mining is a relatively new area of study. A very significant issue in agriculture is yield prediction. Any farmer is curious to know how much of a crop he can expect. In the past, farmer experience with a certain field and crop was taken into account when predicting production. Take into account that historical data with matching yield projections is accessible for a period of time in the past. The training data for every data mining technique must be gathered from some point in the past, and the gathered data is employed in terms of training that must be utilized to learn how to classify.This needs to be taken advantage of to understand how to categorize predictions of future yield (Ramesh and Vardhan, 2013).

For the management of crops and pests, wasteland management and soil categorization, data mining becomes prominent in the field of agriculture. In order to forecast the significant associations and give association rules for various soil types in agriculture, a range of association strategies from data mining were evaluated. Similar to how crop forecast, disease detection, and pesticide optimization are previously studied using various data mining techniques (Rajeswari, and Arunesh, 2016).

A variety of data sets are mined for knowledge that can be used to make better decisions in time-sensitive circumstances. Data mining is the standard term describing knowledge discovery in databases (KDD). The primary objective of data mining is to extract knowledge from the dataset and transform it into human intelligible form for further usage. We are able to create and produce relevant and knowledgeable data by using data mining.

In several fields, data mining techniques are employed to raise the quality and importance of valuable data. In each specialized field, it is essential.Data mining is a broad and valuable idea, particularly in the realm of agriculture. 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.

The two categories of data mining tasks are as follows: 1) Descriptivedata mining and 2) Predictive data mining.While predictive data mining is used to forecast the direct values based on patterns identified from previous results, descriptive data mining activities characterize the overall qualities of the data in the database. Prediction is the process of determining unknown or future values of other important variables by using some variables or database fields. Predictive data mining methods are employed the majority of the time in data mining techniques. Utilizing predictive data mining, it is possible to estimate future crop production, weather, the need for pesticides and fertilizers, as well as the amount of money that will be made (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 datamining.

II. 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 improve user experience and engagement by giving users personalized recommendations based on their preferences and previous interactions.

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.

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

The data mining process typically involves several key steps:

1. **Data Collection:**Gathering appropriate information from numerous sources, including databases, spreadsheets, web logs, social media or other data repositories, is the initial step.

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:**A format for data that can be analyzed is created. This could involve scaling or normalizing numerical values, encoding categorical variables or developing new features to improve data mining.

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.

IV. IMPORTANCE OF DATA MINING

1. **Precision Agriculture:** Data mining facilitates precision agriculture, where farmers can make informed decisions at a highly granular level. By examining data from a variety of sources, including agricultural production information, weather forecasts, satellite photography, and soil sensors. Farmers may maximize their field-specific irrigation, fertilization, and pest control tactics. Crop yields are increased and resource waste is reduced with this focused strategy.

2. **Crop Disease Detection**: Data on crop health and environmental circumstances can be analyzed in both the past and the present using data mining techniques. Data mining aids in the early diagnosis of illnesses, pests, and nutrient deficits by spotting patterns and anomalies. This enables prompt action and reduces crop losses.

3. **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.

4. **Market Analysis**: By using data mining to examine price patterns, market trends and consumer preferences, farmers are able to make well-informed choices about the crops they should plant and how best to meet consumer demand.

5. **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.

6. **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.

7. **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.

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

1. **Data Warehouse:** A data warehouse is a centralized repository of structured and organized data that is designed to support business intelligence (BI) activities, reporting, and data analysis. It is a critical component of an organization's data management strategy and serves as a foundation for making data-driven decisions.E.g.Use in data mining, business decision-making, etc.

2. **Transactional Database**: A transactional database, also known as an OLTP (Online Transaction Processing) database, is a type of database system designed to efficiently manage and support the day-to-day operations of an organization. These operations typically involve numerous short and frequent database transactions, such as adding, updating, or deleting records. E.g. Object databases, distributed systems, banking applications, etc.

3. **Multimedia Databases**: Multimedia databases are a type of database designed to store and manage multimedia data, which includes various forms of non-textual data such as images, audio, video, and even 3D models. These databases are used in applications and systems where multimedia content plays a significant role, such as digital libraries, content management systems, media streaming platforms, and multimedia search engines. Application: Online music databases, video on demand, news on demand, and digital libraries.

4. **Spatial Database**: Keep geographic data on hand, stores information in the form of coordinates, topology, lines, polygons, and other shapes.Application: Maps, Global positioning, etc.

5. **Time-series Databases**: Time series databases provide stock exchange data and user-logged activity, as seen in item a.handles a numeric array that is indexed by a time, date, etc.Analyses in real time are necessary. This sort of information is gathered over time, including stock prices, weather information, and website visitor statistics.Application: eXtremeDB, Graphite, InfluxDB, etc.

6. **Multimedia Databases**: Multimedia databases are specialized databases designed to store and manage multimedia data, which includes various forms of non-textual content such as images, audio, video, animations, and 3D models. These databases are used in applications and systems where multimedia content plays a central role, such as digital libraries, content management systems, media streaming platforms, and multimedia search engines.

7. **WWW**: The World Wide Web (WWW), commonly known as the web, is a system of interconnected documents and resources that are linked together and accessible via the internet. It's a fundamental part of the internet and has revolutionized how people access and share information, communicate, and conduct business.Application: Online shopping, Job search, Research, studying, etc.

8. **Structured Data**: This kind of information is arranged in a particular format, such a database table or spreadsheet. Data on transactions, clients and inventories are a few examples.

9. **Semi-Structured Data**: Examples of this form of data include email messages, XML and JSON files, which have some organization but not as much as structured data.

10. **Unstructured Data** : This kind of data can be in any format, including text, photos, audio, and video. Customer reviews, news stories, and social media posts are a few examples.

11. **External Data** : External data refers to information and datasets that originate from sources outside of an organization or system. This data is typically acquired, integrated, and utilized by an organization for various purposes, such as analysis, decision-making, reporting, and enhancing existing datasets. External data can come from a wide range of sources, both digital and non-digital, and it often complements an organization's internal data resources.

12. **Streaming Data**: This kind of data, like sensor data, social media feeds and log files are continuously produced.

13. **Relational Data**: A relational database houses this kind of data, which may be accessed using SQL queries.

14. **NoSQL Data**: NoSQL (Not Only SQL) is a category of database management systems that diverge from traditional relational database management systems (RDBMS). These databases are designed to handle various types of unstructured, semi-structured, or structured data and can provide more flexibility and scalability for certain use cases.

15. **Cloud Data**: AWS, Azure and GCP are just a few examples of cloud computing platforms where this kind of data is stored and handled.

16. **Big Data**: Big data technologies like Hadoop and Spark can be used to store and process this type of data because of its enormous volume, high velocity, and high variety.

VI. 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.

2. **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.

3. **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.

4. **Association Rule Mining**: This method locates intriguing connections or correlations between various dataset elements. For instance, it can aid in the discovery of patterns such as "Customers who buy product A are likely to buy product B."

5. **Anomaly Detection**: This approach focuses on finding outliers and unusual patterns in the data that considerably depart from the norm. To detect fraud, diagnose problems or recognize unusual activity in cybersecurity, anomaly detection is essential.

6. **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.

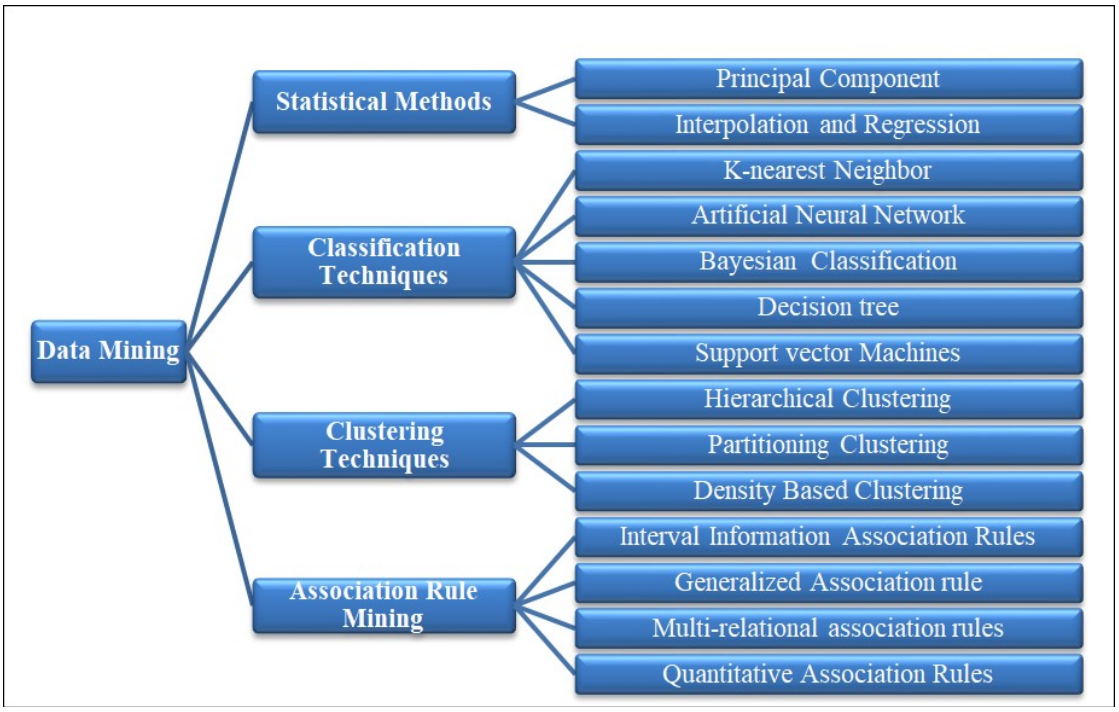


Figure 1: Data Mining Techniques

VII. ALGORITHMS

1. **Supervised Learning:** A supervised learning algorithm examines the practice data and generates an inferred function that may be applied to the mapping of new cases. The algorithm will be able to accurately determine the class labels for instances that are not yet visible in an ideal environment.This necessitates that the learning algorithm generalize in a "reasonable" manner from the training data to hypothetical situations.
   Eg.:Regression models, k-Nearest-Neighbor, Decisiontrees, Neuralnetworks

2. **Unsupervised Learning:** Inferring a function to characterize hidden structures from unlabeled data is a machine learning task. Unlabeled examples are used to teach students because there are no error or reward signals to assess potential solutions. Unsupervised learning can now be distinguished from supervised learning and reinforcement learning.
   Eg.:K-meansclustering, Self-organizedmap, Association rule mining

VIII. CLASSIFICATION TECHNIQUES

It is a machine learning-based classification method.Every time a set of data is categorized, one of a predetermined set of groups is used. Software that can learn how data objects are categorized into groups is developed for categorization. The scalability, speed, predictive accuracy, robustness and interpretability of the categorization and prediction model outputs are assessed(Kodeeshwari and Ilakkiya, 2017).

Rule-Based Classifiers, Bayesian Networks (BN), Decision Trees(DT), Nearest Neighbor (NN), Artificial Neural Networks (ANN), Support Vector Machines (SVM), Rough Sets, Fuzzy Logic and Genetic Algorithms are some of the classification methods used to find knowledge.

1. **Decision Tree**: The most reliable categorization method used in data mining is the decision tree. It is a flowchart with a tree-like organization. Every internal node in this diagram denotes a conditional test and each branch denotes the result of the test (whether it is true or false). A decision tree's leaf nodes each contain a class label. The data can be divided into many classes following the decision tree. It would foretell which category the newly added data point would fall into in the decision tree that was generated. Lines both horizontally and vertically define its forecast boundaries.

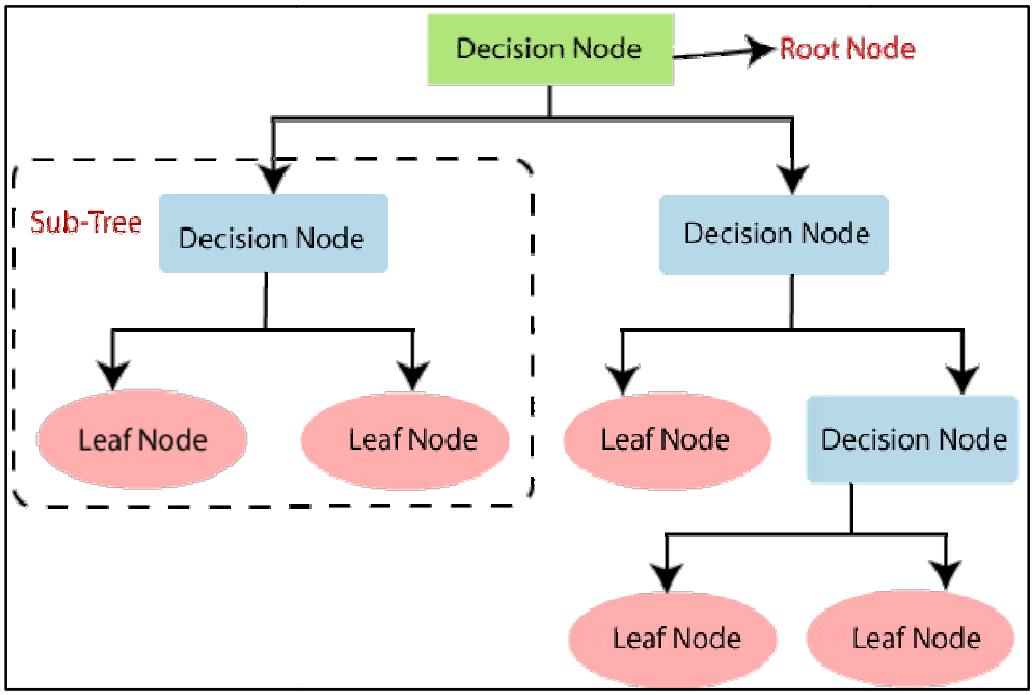


Figure 2: Decision Tree

2. **Naive Bayes:** The naive Bayes method assumes that each characteristic is distinct from the others and that each feature contributes equally to the final result. The notion that all features are equally important is another presumption made by this algorithm. In the modern world, it can be used for a variety of things, including document classification and spam filtering. The estimate of the necessary parameters only needs a little amount of training data when using naive bayes.Other sophisticated and advanced classifiers are substantially slower than a basic bayes classifier. Because it assumes all features are of equal value, which is untrue in the majority of real-world circumstances, the naive bayes classifier is known for being bad at estimating.

$$P(A/B) = \frac{P(B/A)P(A)}{P(B)}$$

3. **Support Vector Machine:** Support Vector Machine (SVM) is a popular machine learning method primarily used for classification tasks. Here is an overview of the steps involved in implementing an SVM:

- **Data Collection and Preprocessing:** Gather and prepare your dataset. Ensure it's labeled and contains the features (input variables) and the corresponding target variable (class labels). Perform data preprocessing, which may include handling missing values, feature scaling, and feature engineering.

- **Feature Selection (Optional):** Depending on the dataset, you may need to select relevant features or perform dimensionality reduction to improve the SVM's performance.

- **Split Data into Training and Testing Sets:** Divide your dataset into two parts: a training set and a testing set. The training set is used to train the SVM model, and the testing set is used to evaluate its performance.

- **Choose the SVM Kernel:** Select the appropriate kernel for your problem. Common choices include: Linear Kernel: Used for linearly separable data. Polynomial Kernel: Suitable for data with some degree of non-linearity. Radial Basis Function (RBF) Kernel: Appropriate for highly non-linear data. Sigmoid Kernel: Rarely used in practice, mainly for special cases.

- **Model Training:** Train the SVM model on the training data using the chosen kernel. The training process involves finding the hyperplane that best separates the data points while maximizing the margin. This is typically done through optimization techniques.

- **Hyperparameter Tuning:** SVM has hyperparameters like C (regularization parameter) and kernel-specific parameters (e.g., degree for polynomial kernel, gamma for RBF kernel). Perform hyperparameter tuning using techniques like cross-validation to find the best combination of hyperparameters for your specific problem.

- **Model Evaluation:** Use the testing dataset to evaluate the SVM's performance. Common evaluation metrics for classification tasks include accuracy, precision, recall, F1-score, and ROC curves. Visualize the decision boundary to understand how the SVM classifies data points.

- **Model Deployment (if needed):** If the model performs well, deploy it in a real-world application. This may involve integrating it into a larger system or creating an API for predictions.

- **Model Maintenance:** Continuously monitor and update your SVM model as new data becomes available or as the problem evolves. Retraining the model periodically may be necessary.

- **Handling Imbalanced Data(if applicable):** If your dataset has imbalanced class distribution, consider techniques like oversampling, undersampling, or using class-weighted SVM to address this issue.

- **Interpretation:** Understand the importance of different features and how the SVM makes predictions. Some SVM variants, like Support Vector Classification (SVC), provide coefficients that can be used for feature importance analysis.

- **Regularization and Cross-Validation:** Regularization is an essential aspect of SVM. It helps control overfitting. Cross-validation is used to ensure that your model's performance estimates are robust and do not overfit a specific dataset.

- **Scaling and Normalization:** Properly scale and normalize your data to ensure that SVM works effectively. Different kernels might have different requirements regarding data scaling.

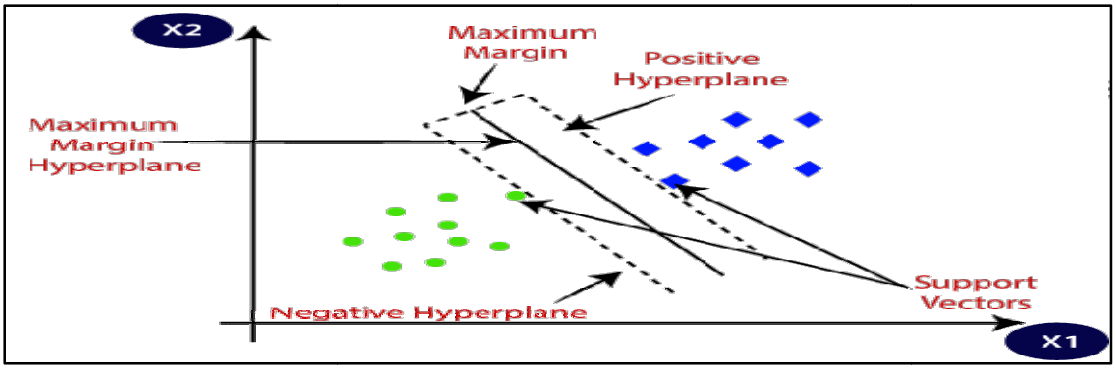


Figure 3: Support Vector Machine

Remember that SVM is a versatile algorithm with various configurations and parameters, so the specific steps you follow may vary based on your dataset and problem. The choice of kernel and hyperparameters often requires experimentation and tuning to achieve the best results.

4. **K-Nearest Neighbour:** A non-linear classifier is part of the k-nearest neighbor algorithm. By determining the classes of its k closest neighbors, it can predict the class of a new test data point. Using the Euclidean distance, you've chosen the k closest neighbors of a test data point. You must count the number of data points present in each category in the k nearest neighbors and then allocate the new data point to the category with the greatest number of neighbors. Because it requires a lot of resources to find the value of k, the algorithm is highly expensive. Its computational cost is further increased by the requirement to calculate the distance between each occurrence and each training sample.

5. **Artificial Neural Network:** Artificial neural networks are also used in various hydrological applications, such as forecasting daily water stress and flow, to assist in the prediction of rainfall.

Good production efficiency and good product quality are requirements for modern agriculture. This holds for both the raising of crops and cattle. Advanced methods of data analysis, especially those developed from artificial intelligence methods are increasingly used to satisfy these criteria. One of the most well-liked tools of this kind is artificial neural networks (ANN). They are frequently employed in the resolution of numerous

classification and prediction challenges. They have been utilized for a while in the broadly defined field of agriculture. They could be included in decision-support and precision agricultural systems. One of the key alternatives to traditional mathematical models is artificial neural networks, which can take the place of traditional modeling techniques. The range of artificial neural network applications is fairly broad. These tools have been used for a very long time by researchers from all over the world to support agricultural production, increasing its efficiency and supplying the highest-quality products conceivable.The goal of this Special Issue was to present excellent research and review papers on the use of various artificial neural network types in solving pertinent tasks and issues related to what is commonly referred to as agriculture.

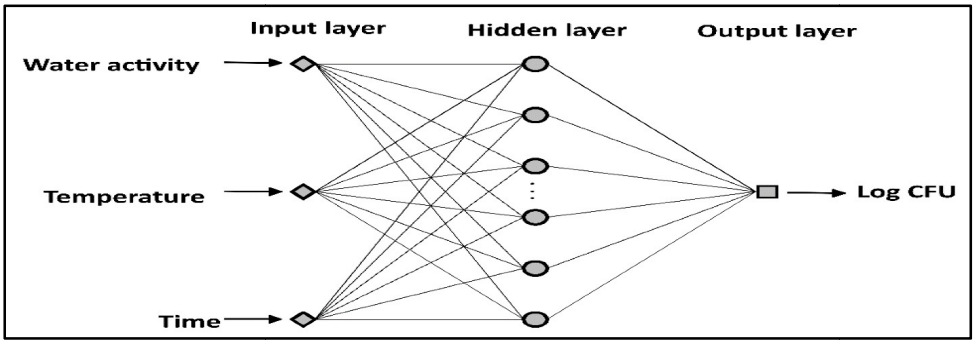


Figure 4: Artificial Neural Network

IX. CLUSTERING TECHNIQUES (Kodeeshwari andIlakkiya, 2017)

A data mining technique called clustering is used to divide the collection of data objects into various clusters or sub-classes. Clustering is a type of unsupervised machine learning technique used to group similar data points based on some similarity or distance metric. There are various clustering techniques, each with its own strengths, weaknesses, and ideal use cases. Here are some of the most commonly used clustering techniques.

A data mining approach called clustering groups similar and dissimilar instances together as well as those that belong to different groups based on the data instance. Subsets of the data instance are created. Because it connects with instances where similarities and ranges concur, the clustering technique is utilized to identify various pieces of information. This method does not require any prior data knowledge. Clustering is a method of unsupervised learning that separates unlabeled data records and is used for data processing. A cluster center, which houses all the clusters, is the result of clustering. High quality clusters will be produced by a well-defined clustering algorithm. i) Inter class[similarity low] and ii) Intra class[similarity high]

1. K-Means Clustering

- **Method:** Divides data into a pre-defined number (k) of clusters.
- **Strengths:** Simple, computationally efficient, and works well for globular-shaped clusters.
- **Weaknesses:** Requires the number of clusters to be specified in advance, sensitive to

initial cluster centers, and doesn't work well for non-linear or irregularly shaped clusters.

2. **Hierarchical Clustering**

- **Method:** Builds a hierarchy of clusters by successively merging or splitting existing clusters.
- **Strengths:** Doesn't require specifying the number of clusters beforehand, can be visualized as a dendrogram, and is useful for identifying nested clusters.
- **Weaknesses:** Can be computationally intensive, and the choice of linkage method (single, complete, average, etc.) can impact results.

3. **DBSCAN (Density-Based Spatial Clustering of Applications with Noise)**

- **Method:** Forms clusters based on the density of data points. It identifies core points, border points, and noise points.
- **Strengths:** Can find clusters of arbitrary shapes, doesn't require specifying the number of clusters, and is robust to outliers.
- **Weaknesses:** Sensitive to the choice of distance metric and density parameters.

4. **Agglomerative Clustering**

- **Method:** Starts with individual data points as clusters and iteratively merges the closest clusters until a stopping criterion is met.
- **Strengths:** Hierarchical and doesn't require specifying the number of clusters beforehand.
- **Weaknesses:** Can be computationally expensive for large datasets, and choice of linkage method matters.

5. **Gaussian Mixture Models (GMM)**

- **Method:** Models data as a mixture of multiple Gaussian distributions. Data points are assigned probabilities of belonging to each cluster.
- **Strengths:** More flexible than K-Means, can model elliptical clusters, and provides probabilistic cluster assignments.
- **Weaknesses:** Sensitive to initialization and may converge to local optima.

6. **Spectral Clustering**

- **Method:** Applies graph theory to cluster data by representing it as a graph and performing spectral decomposition.
- **Strengths:** Can discover non-convex clusters, is less sensitive to initialization, and can handle data with complex structures.
- **Weaknesses:** Can be computationally intensive for large datasets, and you need to specify the number of clusters.

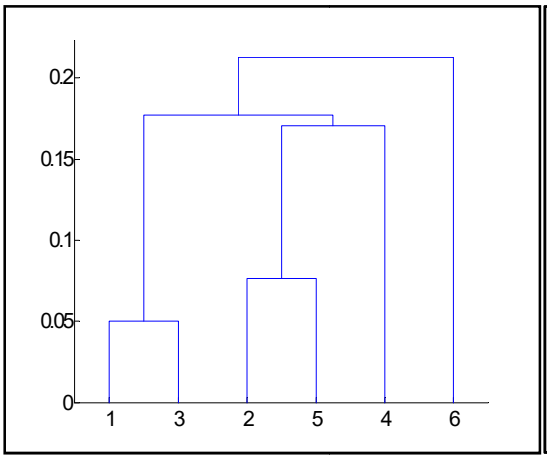
7. Mean Shift

- **Method:** Shifts data points iteratively towards regions of higher data point density, identifying cluster centers.
- **Strengths:** Automatically detects the number of clusters and is robust to irregularly shaped clusters.
- **Weaknesses:** Can be sensitive to bandwidth parameter, and the runtime can be high for large datasets.

8. Affinity Propagation:

- **Method:** Uses message-passing between data points to find clusters. It automatically determines the number of clusters.
- **Strengths:** Can discover exemplars within clusters and doesn't require specifying the number of clusters.
- **Weaknesses:** Can be sensitive to the similarity metric and may result in a large number of clusters.

Choosing the right clustering technique depends on your data, the shape of clusters, the number of clusters, and the computational resources available. It's often a good practice to try multiple techniques and evaluate their performance using appropriate metrics before selecting the one that best suits your problem.



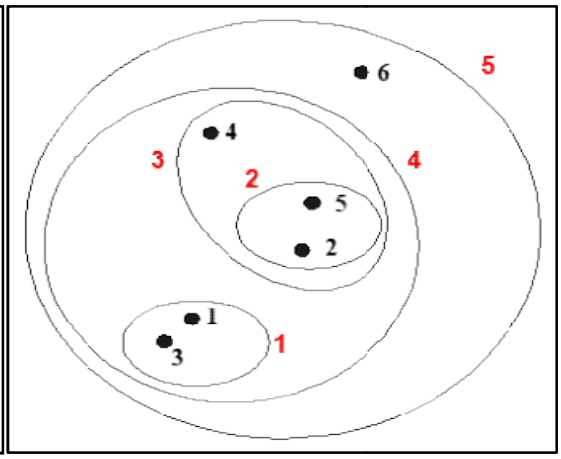


Figure 5: Agglomerative Methods (Bottom-Up) **Figure 6:** Divisive Methods (Top-Down)

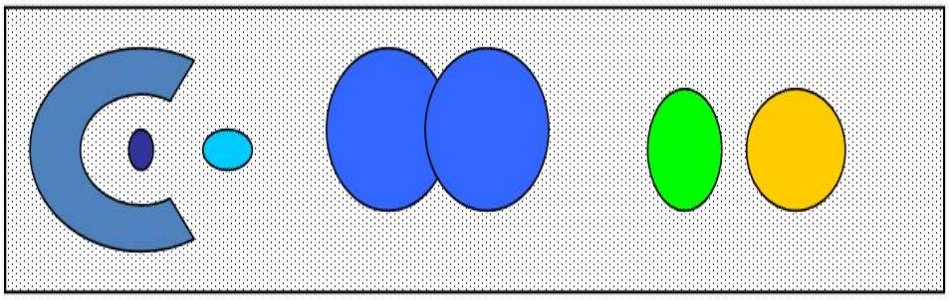


Figure 7: Density Base Clustering

X. ASSOCIATION RULE MINING(Kodeeshwari and Ilakkiya, 2017)

Agrawal, Imielinski, and Swami created the data mining method known as association rule mining in 1993. This is one of the well-organized data mining techniques for looking for hidden or desired patterns in data. Finding relationships between different items in the relational database is the major goal of this approach. In a dataset with more absolute selections of elements, association rules are utilized to locate patterns and identify the elements that occur repeatedly.

Association rule mining in agriculture involves applying data mining techniques to discover interesting relationships or associations between different agricultural factors or events. These associations can provide valuable insights for optimizing agricultural practices, improving crop yields, reducing costs, and enhancing overall farm management. Here's a step-by-step method for using association rule mining in agriculture:

1. **Data Collection**: Gather relevant agricultural data, which may include information on crops, soil types, weather conditions, fertilizer usage, pest and disease occurrences, irrigation practices, and more. Ensure that your data is well-structured, clean, and suitable for association rule mining.

2. **Data Preprocessing:** Clean and preprocess the data to handle missing values, outliers, and inconsistencies. Encode categorical variables into numerical formats if necessary. Normalize or scale numerical variables to ensure they have similar ranges.

3. **Data Exploration**: Perform exploratory data analysis (EDA) to understand the dataset better. This includes summary statistics, data visualizations, and identifying potential relationships.

4. **Association Rule Mining Algorithm**: Choose an association rule mining algorithm. The Apriori algorithm is a widely used method for discovering association rules, but other algorithms like FP-Growth can also be effective.

5. **Set Parameters**: Define parameters for the algorithm, including the minimum support and minimum confidence thresholds. The minimum support threshold specifies the minimum frequency (occurrence) of an itemset or association in the dataset. The minimum confidence threshold sets the level of confidence required for an association rule to be considered interesting.

6. **Generate Frequent Itemsets**:Apply the chosen association rule mining algorithm to the preprocessed dataset to generate frequent itemsets. Frequent itemsets are sets of items (e.g., agricultural practices or conditions) that meet the minimum support threshold.

7. **Generate Association Rules**: From the frequent itemsets, generate association rules that meet the minimum confidence threshold. Association rules typically have the form "if {antecedent} then {consequent}" and indicate the relationships between different agricultural factors.

8. **Evaluate and Interpret Rules**:Evaluate the generated association rules to identify meaningful and actionable insights.Pay attention to the support, confidence, and lift of the rules.Support indicates the frequency of occurrence of the rule's antecedent and consequent. Confidence measures how often the rule is correct.Lift quantifies how much more likely the consequent is given the antecedent compared to random chance.

9. **Interpretation and Actionable Insights**: Interpret the discovered association rules in the context of agriculture.Identify patterns or relationships that can lead to practical recommendations for optimizing farming practices, crop management, pest control, resource allocation and more.

10. **Implementation and Monitoring:** Implement the insights gained from association rule mining in actual agricultural practices.Continuously monitor and assess the impact of these recommendations on crop yields, resource usage, and other relevant metrics.

11. **Iterate and Refine:** As new data becomes available, periodically re-run the association rule mining process to identify evolving patterns and refine recommendations accordingly.

By applying association rule mining in agriculture, farmers and agricultural researchers can make data-driven decisions that lead to improved crop production, reduced resource wastage, and more efficient farm management practices. It can also aid in sustainable agriculture and help address environmental concerns by optimizing resource usage.

The significance of a rule like this is that transactions that contain A also contain B within the database. As an example, the association expression between the items A and B are of the form A => B. When using association rules, we need take into account two crucial basic rule metrics like support and confidence. These two metrics have a threshold value that can be used to extract intriguing patterns from a large body of data. Of course, support has real value, and confidence has probability value. Uninteresting rules are those that do not meet the cutoff value. A simple, certain, and useful association rule is required in order to gauge how fascinating it is. The rule basic measure support and confidence between the items A and B is defined as

$$\text{Confidence}\ (\text{A} \Rightarrow B) = \frac{\text{Transactions containing both A \& } B}{\text{Transactions containing A}}$$

$$\text{Support}\ (\text{A} \Rightarrow B) = \frac{\text{Transactions containing both A \& } B}{\text{Transactions containing both A \& } B}$$

When the confidence value in association rule mining is 100%, the examined data is always right; such rules are referred to as precise. The minimal support threshold and minimum confidence threshold are features of strong association rules.

XI. REGRESSION(Kodeeshwari andIlakkiya, 2017)

Regression is a data mining function that is used to predict a numeric or continuous value. A predictive modeling technique called regression analysis shows the relationship between the independent variable (X) and dependent variable (Y). Dependent variables are those that have been predicted and independent variables are those that are predicted and utilized to forecast the values of dependent variables. Regression is a key tool for data analysis and data modeling.

The fluctuations in one variable that follow variations in another variable are likewise represented by regression. One crucial aspect of regression is that it thoroughly explains the connections between the variables. Correlation is a measure of how strongly the variables are related. Frequently, classification problems with quantitative class labels are used to describe regression tasks.

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

$$Linear\ regression\ equation\ (y) = a + bx + \varepsilon$$
$$Non-linear\ regression\ equation\ (y) = f(X, \beta) + \varepsilon$$

Where, a = Intercept
b = Slop
ε = Error
y = Dependent Variable
X = Independent Variable

Table 1: Comparison of Data Mining Technique

Differentiator	Classification	Clustering	Association Rules	Regression
Methods	Predictive method	Descriptive method	Descriptive method	Predictivemethod
Usage	Used to predict the instance class from a pre- labeledinstance.	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	*Hierarchical clustering *Partition clustering *Density clustering	*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

Table 2 : Various Tools Used for Data Mining

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

Commercial Tools And Programmed

- Oracle Data Minerhttp://www.oracle.com
- Data ToKnowledgehttp://alg.ncsa.uiuc.edu
- SAShttp://www.sas.com/
- Clementinehttp://spss.com/clemetine/
- IntelligentMinerhttp://www-306.ibm.com/software

XII. APPLICATION OF DATA MININGIN AGRICULTURE

1. **The Process of Segregating Fruits and Vegetables Based on their Water Content Levels has been Enhanced:** Normal classification of fruits and vegetables into different price ranges is based on their size and color. These, however, are outside variables, and they have little to do with how well fruits and vegetables are grown. The amount of water in the fruit, or how much water is present relative to the fruit's weight, is what determines its key characteristics. A fruit or vegetable's water content can have a limited impact on how long it lasts, and abnormal water content can degrade the quality of nearby fruits and vegetables in boxed produce. Images of fruits and vegetables are taken at the packing line

and then processed to create a digital representation that can be used to tackle this problem(Mirjankarand Hiremath, 2016).

Additionally, records of a range of specimens aid in producing a more precise estimate of the fruit and vegetable quality. These photos can be used to feed the VGG 19 model, a deep convolutional layer with 19 layers that is utilized for large-scale image recognition. The output layer of the VGG19 model receives the 224 x 224 RGB picture, and "Softmax" is used as the activation function to provide a quality rating for the input image in the range of 10 output labels. It is necessary to train the model using pictures of fruits that have labels attached and have been rated out of 10 by a human.

2. **Leveraging Data Mining Techniques to Optimize Crop Yield based on Soil Quality:** Growing a certain crop on any area that doesn't fulfill the crop's minimal requirements will result in lower-quality yields and less money for the farmer. Agriculture requires that the soil's quality be determined. In this, the proportions of nutrients and minerals found in the soil are analyzed. Alkalinity, salinity, moisture content and other variables all affect the soil's quality. The many types of soil are studied *via* data mining. Depending on the soil's fertility and the desired yield, soil data analyzers recommend the type of crop to be planted and harvested. To anticipate various qualities for agriculture depending on the season and climatic conditions, data mining provides a wide amount of data for various soil variations. Data mining techniques are being used with a larger variety of statistical and analytical data, which improves the accuracy of information extraction and also has the potential to automate results for general scenarios (Patel and Patel, 2014). Data mining can also be used to research cross-cultivation, which is the simultaneous cultivation of many crops. This practice would increase revenue over cultivation of a single crop while making the most use of available resources without compromising soil fertility. The following are some examples of how data mining's broad application may be seen in soil analysis.

   - Artificial Neural Networks can be employed to sense and detect soil conditions to recommend suitable crops for adoption
   - Newly unidentified soil patterns can be uncovered
   - The characteristics and behavior of soils can be forecasted based on the prevailing climate conditions and soil composition.
   - Soil fertility testing can be enhanced through the application of statistical methods (Palepu and Muley, 2017).

3. **Leveraging Data Mining Techniques in agriculture to Enhance Accuracy and Precision in Decision-Making:** Reduced pesticide use in agriculture is necessary to address the problem of excessive pesticide use, which lowers overall agricultural production. According to Tellaeche et al. (2007), data mining can be utilized to create automated systems to find weeds growing in fields. This makes use of an image processing system and relied heavily on surface area, aspect ratios, and forms. Later, using particular algorithms, photographs of the region being farmed are processed to identify weed patches (Yang, 2003). In the photographs, color density is used to depict the density of crop growth in a specific location, while a different color was used to depict uneven crop development.

4. **Evaluating the Performance Of Chicken Using Modified Neural Network Models:** An artificial neural network comprises interconnected artificial neurons that emulate the functionality of biological neural networks. It consists of layers, with the first layer being the input layer, the last layer as the output layer, and the layers in between as hidden layers. Deep neural networks have multiple hidden layers. Through training with historical data, these neural networks can be applied to various prediction and classification tasks. Such trained networks can achieve human-level or higher accuracy but not 100%, typically requiring minor hyperparameter adjustments. In an artificial neural network analysis of feeding effectiveness and weight gain based on a dataset, it was observed that dietary protein concentration holds greater significance than threonine concentration. According to a study, a diet containing 18.69% protein and 0.73% threonine can promote weight gain, although achieving the same level of efficiency with a standard deviation of 0.2 or 0.3 percent is feasible.

5. **Enhancing the Efficiency of Pesticide Utilization Through Data Mining:** Overuse of pesticides can be detrimental to farmers in a variety of ways. Crop yield prediction is a significant issue in agriculture. According to a recent study by agricultural researchers, pesticides are overused, which is quite damaging for the environment, in order to enhance crop productivity. Additionally, overuse of pesticides can render pests immune, which makes them more dangerous to crops and harder to control. Because of this, using pesticides excessively puts farmers' families at financial risk and poses a health risk.With the aid of clustering, one of the data mining techniques, it is possible to group the features by revealing intriguing patterns in farmer behavior and so deliver valuable information that will draw attention to the negative effects of excessive pesticide use (Paulson, 2015).

6. **Prediction of Problematic Wine Fermentations:** Wine production is a global industry, and the fermentation process plays a pivotal role in determining both the productivity of wine-related sectors and the quality of the wine itself. Predicting the progression of fermentation during its early stages could enable proactive intervention, ensuring a consistent and smooth fermentation process. Various techniques, including the k-means algorithm and biclustering-based classification (Paulson, 2015), are employed to study fermentations.

   The latest advancements in understanding spatial correlations have significantly enhanced yield forecasts. Artificial Neural Networks are instrumental in developing models for forecasting and issuing warnings regarding plant diseases. Independent component analysis, a signal processing technique, is utilized to isolate independent sources, revealing hidden traits through numerical and statistical analysis of data from diverse sources (Bhagawati, 2016). This technique employs generative models on observed data.

   To optimize pesticide usage, the detection of weather data patterns is crucial. Agricultural data integration, a technology that combines pest scouting, pesticide application, and climatic monitoring, is increasingly employed for this purpose.

7. **Detection of Diseases from Sounds Issued by Animals:** Since sick animals can contaminate food and spread infectious diseases, finding animal ailments early on can help farms become more productive. Additionally, early disease diagnosis enables the

farmer to treat the animal as soon as the illness manifests. Diseases can be found by listening to the sounds that pigs make. Their coughs in particular can be examined because they reveal the severity of their illnesses. A computational system that can distinguish between the many sounds that can be heard and monitor pig sounds using microphones deployed in the farm is currently being developed (Paulson, 2015).

8. **Optimizing Pesticide use by Data Mining:** Recent agricultural research studies revealed that attempts to maximize cotton crop output through pro-pesticide state laws have resulted in an alarmingly high pesticide use. These studies have found a link between pesticide use and crop production management that is unfavorable. Therefore, overusing pesticides has a negative impact on farmers' livelihoods, the environment, and society. It was demonstrated how pesticide use can be optimized (decreased) by data mining the cotton Pest Scouting data. Intriguing patterns of farmer behavior and pesticide use dynamics were discovered by clustering the data, which helped pinpoint the causes of pesticide abuse (Paulson, 2015).

9. **Sorting Apples by their Water Cores:** Apples are inspected before they are brought to market, and those that have flaws are eliminated. Unseen flaws, on the other hand, have the potential to ruin the apple's appearance and flavor. The watercore is an illustration of an unseenable flaw. This apple condition may shorten the fruit's shelf life. Apples with few or mild watercores have a higher sugar content; nevertheless, apples with moderate to severe watercores cannot be kept for an extended period of time. Additionally, a small number of fruits with a bad water core could ruin an entire batch of apples.Due to this, a computational system is being investigated that can both acquire X-ray images of fruit as it moves along a conveyor belt and analyze those images (using data mining techniques) to determine whether the fruit may contain watercores(Paulson, 2015).

Table 3: Data Mining Methodologies and its Uses In Agriculture

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.

XIII. CURRENT ISSUE OF DATA MINING IN SMART AGRICULTURE

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

1. **More generalization:**Data mining techniques are not specified for the particular field or data.

2. **Require special knowledge:**A significant problem with data mining is that the applications that use the results require specialized knowledge or experience to recognize and make use of the available data volume.

3. **Test against the small size of data:**Most of the systems in use today only took into account a small portion of the dataset while determining the outcome. It is reducing agricultural production prediction's effectiveness. As a result, there is a need for methods that make use of the complete data set to boost crop yield forecast efficiency.

4. **High Computational Cost:**It offers a variety of data mining techniques, including clustering, classification, k-Nearest Neighbor, and k-Means, to handle agricultural problems.The k-Means algorithm's drawback is the possibility of selecting from a range of different 'k' parameters. Additionally, it includes the cost of computation.The simultaneous handling of various algorithms by the WEKA (Waikato Environment for Knowledge Analysis) system has not been supported. It takes longer and requires expensive processing.

5. **Limited resources in India:**The majority of farmers just have a modest plot of land. Applications and equipment created for the expansive agricultural sector are challenging to use. Internet coverage is less widespread, which restricts the use of IT applications. fewer infrastructure facilities are available.

XIV. CONCLUSION

The most significant application area is primarily in underdeveloped nations like India. Data mining not only provides information about a specific task, but there is a potential that the information will change as new, widely used technologies are developed.By using information technology in agriculture, farmers may produce more effectively and change the way decisions are made. Making decisions about a variety of topics pertaining to the agricultural area requires the use of data mining. 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 optimization, and other fields, agricultural institutions apply data mining applications. Therefore, we may conclude that data mining has benefited the agricultural industry.

REFERENCE

[1] Annonymous (2023b) Data Mining: Process, Techniques and Major Issues In Data Analysis, https://www.softwaretestinghelp.com/data-mining/> accessed on 18 July, 2023.
[2] Annonymous, (2023a).Goals of data mining. Available at https://people.cs.pitt.edu/~chang/156/21mining.html> accessed on 18 July, 2023.

[3] Annonymous, (2023c).Sources of data for mining. Available at https://www.geeksforgeeks.org/types-of-sources-of-data-in-data-mining/> accessed on 18 July, 2023.
[4] 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
[5] 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.
[6] Han, J., Kamber, M. and Pei, J. (2011). Data reduction. *Data Mining, Concepts and Techniques, 3rd ed.; The Morgan Kaufmann Series*, 99-110.
[7] 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.
[8] 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.
[9] Mirjankar, N. and Hiremath, S. (2016). Application of data mining in agriculture field. *International Journal of Computer Engineering and Applications*, **10(**6).
[10] 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).
[11] Patel, H. and Patel, D. (2014). A brief survey of data mining techniques applied to agricultural data. *International Journal of Computer Applications,* **95**(9).
[12] Paulson, S. (2015). A Survey on Data Mining Techniques in Agriculture International Journal of EngineeringResearch and Technology, **3**(30).
[13] Rajeswari, V. And Arunesh, K. (2016).Analyzing Soil Data Using Data Mining Classification Techniques. *Indian Journal Of Science And Technology,*9(19): 1-4.
[14] 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.
[15] Sindhu, S. and Sindhu, D. (2017). Role of data mining techniques in agriculture improvement. **6**: 654-663.
[16] 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.
[17] 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.

