

# Identifying And Imparting A Suitable Data Mining Methods For Precision Agricultural Practices

<sup>1</sup>K.Vijesh , <sup>2</sup>Dr. R. Nagaraj

<sup>1</sup>Research Scholar, Department of Computer Science, Kamadhenu Arts and Science College, Sathyamangalam, TN, India.

<sup>1</sup>Associate Professor, Department of Computer Science, Kamadhenu Arts and Science College, Sathyamangalam, TN, India.

DOI: 10.47750/pnr.2022.13.S10.13

## Abstract

Numerous ongoing projects and a growing global population continuously put pressure on the agricultural sector. As much as 70% more production efficiency is needed for all types of crops from the same land resources in order to meet demand without negatively impacting the environment. At the same time, one-third of our food is wasted. Other variables, such as global warming, more frequent storms, water shortages, and problems with soil and water stream contamination, which can result in waste and decreased output, might make it more difficult for farmers to fulfill the growing demand for crops for human and animal use. We evaluate a technology to help manage these problems. We examine a technology to help the agricultural sector, which will benefit both farmers and the general public by producing high-quality, safe food that satisfies both present and future demand, in order to control these difficulties. Smart farming, the Internet of Things, big data, high-performance computing, and cloud computing are all combined in the solution to provide precision agricultural services innovation that address real-world issues. The creating a solution with a data-driven applications will be facilitated by storing agriculture data on supercomputers. In soil and plant research, supercomputers store information from multiple sources, including sensors on the ground. It is possible to collect valuable operational information about soil quality, crop nutrients, forecasts of crop yields, the best harvesting time, gas emissions, and soil contamination through satellite and drone imagery. Through machine learning and artificial intelligence, the supercomputer also extracts crucial data and converts it into apps that are sent through different routes, with direct delivery to farmers' cell phones being the most convenient. In this paper we will be discussing about best practices we carry forward to implement a precision agriculture practice in India.

**Key Word:** Smart Farming, (IOT)Internet of Things, , High-performance computing and cloud computing technologies, Big data, Machine Learning, Precision Agriculture practice.

## Introduction:

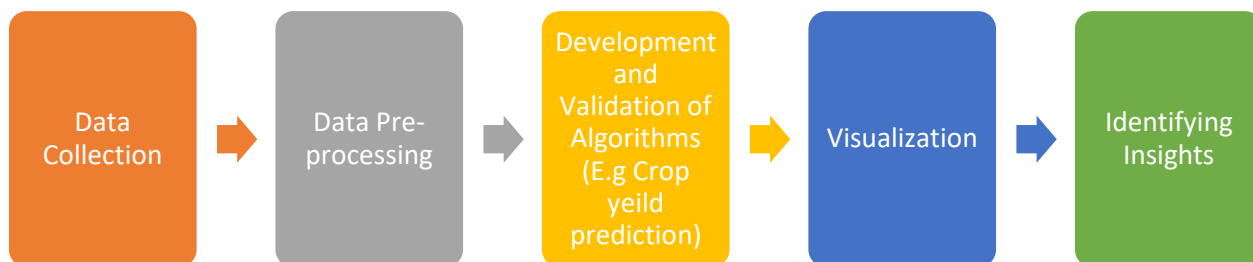
Farmers can make informed decisions using machine learning and artificial intelligence applications. When monitoring fertilizer application, herbicide spraying, or crop harvesting, data visualization can provide insight into the potential of autonomous vehicles. People are alerted when a problem may exist by automated information.

- There is a lack of nutrients in certain areas of their land assets (or)
- Symptoms of an outbreak of a disease (or)
- Detects harvesting readiness of crops and provides alerts.

A farmer cannot make informed decisions without the right data. Various methods are used to collect the data: machine learning, last year's data, and other automated AI applications. With tools like Tableau or Power BI, a farmer can quantify their data to make well informed decisions.

## Data collection approach:

The data is entered from a variety of sources, which goes like this:



In order to help farmers make informed decisions, we collect data. Pre-processing and purification of data, analysis of data, structuring and removing poor bits of data, algorithm development and validation, visualization, interpretation, and development of performance business models.

Using precision farming technologies, data can be visualized and analysed to surface trends and segment tasks best performed by autonomous vehicles. Both common and vulnerable crops and fruits can also be modeled using weather forecasts, satellite images, and drone imagery.

Improved crop quality and production will be achieved by integrating local weather and imaging in a more detailed manner. With the world's population reaching 10 billion by 2050, farmers are increasingly open to using technology that allows them to meet this demanding need in a safe and efficient manner while improving quality of life and the environment. In the past, farmers have used traditional farming practices to meet the demands of their communities; however, with the world's population expected to reach 10 billion by 2050, they are becoming increasingly open to new technologies that will assist them in meeting these demands in an efficient and effective manner.

### Stages of data mining process:

Existing data mining approaches can be classified into 10 stages right from Data acquisition, Data Cleaning, Data preparation, Data Transformation, Data Analysis, modeling, classification/clustering, forecasting and report

generation.

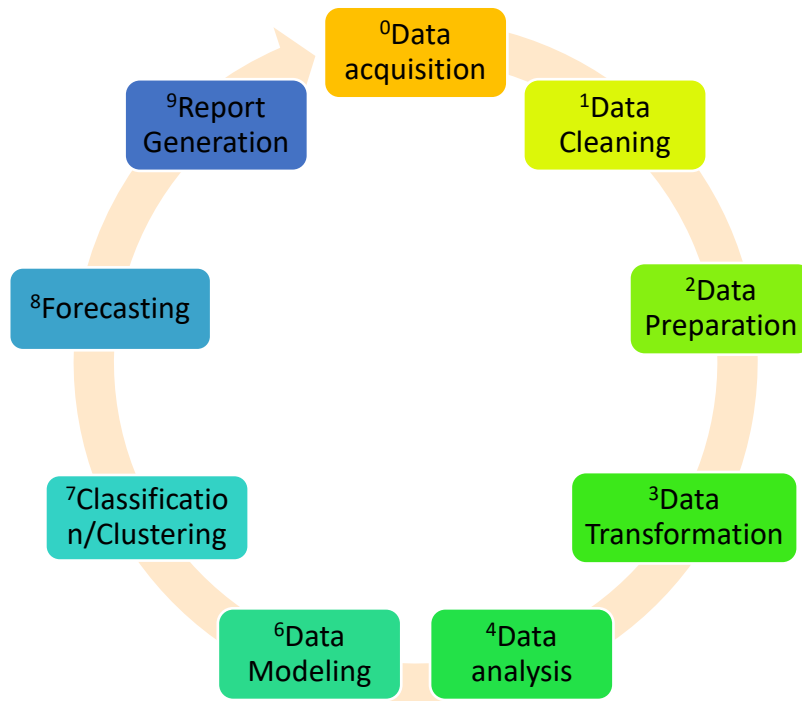


Figure (1)

With algorithms imparting proper data mining methods, we can generate meaningful reports on agriculture related data, from scanning for frequent patterns, detecting associations and correlations, classifications, clustering, and detecting outliers. In this paper, we present an analysis of possible solutions that can be implemented to enable people who want to start farming in open land to have more opportunities. Another idea is to educate and improve existing traditional farmers to start practicing precision agriculture.

### Analysis of best method for informed decision making:

This title depends on choosing a best algorithm to implement best practices for the farmers to take informed decisions. Let us analyze the classification and clustering algorithmic approaches,

### Classification algorithm:

In classification algorithms, input information is weighed so that positives and negatives are produced. There are several algorithms available that allow us to learn more about K-Nearest Neighbors and Decision Trees. These algorithms (K-Nearest Neighbors & Decision Tree) is mainly used to forecast the output of categorical data, and it is first aimed at determining the category of a given dataset.

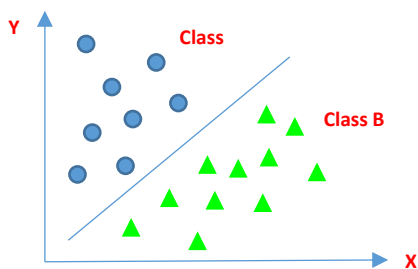


Fig.2 (Sample KNN – K-Nearest Neighbours classification output)

It's part of classification analysis the output data used to retrieve important and relevant information about data, and metadata.

### K-Nearest Neighbours algorithm and its role:

The KNN algorithm is an easy-to-use supervised machine learning algorithm used for regression and classification. An unlabeled input data may be used to train a function that provides relevant outputs when given an unlabeled input.

### KNN: our first choice for understanding:

Imagine the computer as a child. We are its supervisor, and we want it to learn what a dog looks like. We will show the youngster a variety of photographs, some of which will be of dogs and the remainder might be of anything (cats, dogs, etc).

To teach a machine how to identify dogs, we show it examples of images of dogs. If it gets the correct answer, we say yes and if not we say no. **Classification problems** can be solved using supervised machine learning algorithms

For a classification task, you can only be one of two outcomes. For instance, **"I like Strawberry on Ice Cream"** or **"I do not like Strawberry on Ice Cream."** Furthermore, the difference between recognizing a dog and cat is also the same.

Age	Likes Strawberry on Ice Cream
42	1
65	0
50	0
76	1
96	1
50	1
91	0

Table (1) distinct values likes (1) dislikes (0)

As you can see in the image above, classification data might look something like this. In this case, we have a predictor (or set of predictors) and a label. According to the picture, someone might like Strawberries (1) or dislike them (0) on their Ice Cream depending on their age (the predictor).

In general, classification methods produce integer results (labels) such as 1, -1, or 0. Statistical analysis in this situation is merely illustrative. It would be ineffective to expose them to mathematical methods. Take time to consider this. What is the distinction between "likes Strawberry" and "hates Strawberry"? Exactly. We shouldn't add their numerical equivalents since we can't add them.

### K-Nearest Neighbours - a typical example:

Close proximity is indicative of similar objects. In other words, it would be like a group of people with similar interests getting together to hang out. As a result, comparable objects are located close to one another.

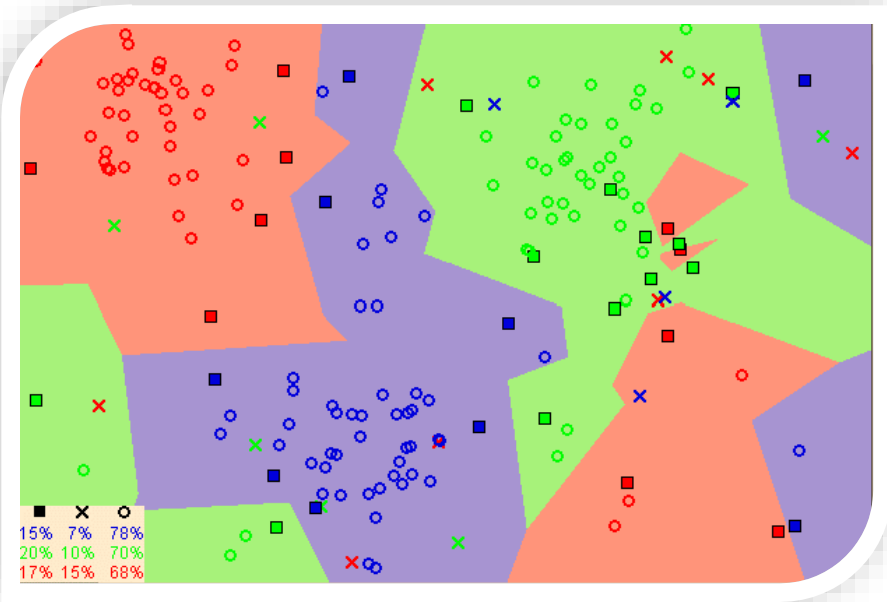


Fig (3) – Sample classification for identifying similarities

The above figure depicts the comparable data points are usually adjacent to one other. Consequently, the KNN algorithm is designed based on the assumption that this assumption is true enough that the algorithm will be beneficial. Similar to the math we learn as children, such as calculating distances between points on a graph, KNN uses the concept of similarity (also known as distance, proximity, or closeness).

There are numerous methods for determining distance, and one method may be preferred depending on the task at hand. However, the straight-line distance (also known as the Euclidean distance) is a popular and well-known option.

### The KNN Algorithm

1. Initiate the first step by loading the data.
2. Your chosen number of neighbors should be used as the initial value for K.
3. According to the data, for each example, Utilize the current example and the query example to calculate the distance between them.  $\text{distance} = \sqrt{a^2 + b^2}$ 
  - 3.2 Organize an ordered collection by adding the distance and example index.
4. Sort distances and indexes according to smallest to largest (in ascending order).
5. Select the first K entries from the sorted collection.
6. Get the labels for K entries.
7. In the event that there was a regression, calculate the mean of the K labels.
8. If the K labels are classified, return their mode.

### How to select the right set of values for K:

A prediction is produced by running the KNN algorithm multiple times with different K values. In the prediction process, we select the K that will give us the best prediction, without affecting how the prediction performs with new data..

Here are some things to know,

1. When K equals one, KNN is unreliable. When there are more red points in the neighborhood than green ones, but only one green neighbor is the query point, it still predicts the point to be green.
2. It is more likely that the forecast will be accurate when K is high enough. If the K value is high enough, there will be mistakes in the output.
3. In the case of a tie, we tend to make K an odd number with a tiebreaker.

### Advantages

1. Implementing the algorithm is simple and straightforward.
2. No model is required, no parameters need to be tuned, no additional assumptions need to be made.
3. Regression and classification are both possible with artificial intelligence.

### Disadvantages

1. The performance of the algorithm will decrease as the number of independent variables increases.

### Decision Tree algorithm:

Classification and regression are done using decision trees, which replicate the rules determined from data. In this algorithm, decision rules are derived from attributes in the data and used to train a model. Piecewise approximations are illustrated by the tree.

### Some advantages of decision trees are:

1. Interpretable and easily understood. It is possible to visualize trees.
2. Data preparation is minimal, but missing values are not handled by the module.
3. In machine learning, the quantity of data consumed determines the amount of data consumed.
4. A number of different datasets can be processed in Scikit-Learn, including numerical and categorical data. Categorical data, however, are not supported for the time being.
5. Multi-output difficulties could be managed by programs using this algorithm.
6. Validating a model can be accomplished using a statistical test. The model's dependability can be accounted for in this way.

### The disadvantages of decision trees include:

1. Keeping leaf nodes and depths shallow will prevent overfitting in decision-tree learners.
2. Despite their usefulness in generating predictions and classifications, decision trees can be unstable for certain datasets. A decision tree can, however, be incorporated into a larger ensemble in cases of complex problems in order to produce stable results.

### Enhanced Hybrid-Nearest Neighbour algorithm (HNN)

The new HNN is the combination of K-Nearest Neighbours and Decision tree algorithm. Due to the disadvantages found in KNN, this algorithm is impractical when dealing with large data sets. The outcome value of KNN is further brought to decision tree algorithmic approach and iteration will be done at this level which is an extensive method varies from KNN and large volumes of data can be iterated in this model.

By using a set of if-then-else decision rules, the improved version will provide better results by applying decision trees to data. As the tree gets deeper, the decision rules and model fit become more complex.

### References:

1. Ester, M, Kriegel, H P, Sander, J, & Xiaowei, Xu. "A density-based algorithm for discovering clusters in large spatial databases with noise." United States.

2. Satoru Hayasaka, "What is Clustering and How Does it Work?", 2021, Source: URL -<https://www.knime.com/blog/what-is-clustering-how-does-it-work>
3. L. Bottou and Y. Bengio. "Convergence properties of the kmeans algorithm. In Advances in Neural Information Processing Systems". 1995. Google Scholar
4. Aravind CR, "Exploring Clustering Algorithms: Explanation and Use Cases", 2022, Open web article, Source: URL – <https://neptune.ai/blog/clustering-algorithms>
5. Open web article, Source: URL – [https://developers.google.com/machine-learning/glossary#classification\\_model](https://developers.google.com/machine-learning/glossary#classification_model)
6. J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra. "Efficient projections onto the  $l_1$ -ball for learning in high dimensions." In ICML '08: Proceedings of the 25th international conference on Machine learning, 2008. Google Scholar Digital Library
7. George Seif, "The 5 Clustering Algorithms Data Scientists Need to Know", 2018, Open web article, Source: URL - <https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68>
8. C. Elkan. "Using the triangle inequality to accelerate k-means". In ICML '03: Proceedings of the 20th international conference on Machine learning, 2003. Google Scholar
9. D. D. Lewis, Y. Yang, T. G. Rose, and F. Li. Rcv1: "A new benchmark collection for text categorization research". J. Mach. Learn. Res., 5, 2004. Google Scholar Digital Library
10. D. Witten and R. Tibshirani. "A framework for feature selection in clustering. To Appear: Journal of the American Statistical Association", 2010. Google Scholar
11. Jason Brownlee, "Python Machine Learning", 2020 Open web article, Source: URL - <https://machinelearningmastery.com/clustering-algorithms-with-python/>
12. X. Wu and V. Kumar. "The Top Ten Algorithms in Data Mining". Chapman & Hall/CRC, 2009.
13. Nagesh Singh Chauhan, "KDNuggets – Building and optimizing decision tree classifiers with knowledge of decision trees", 2022 Open web article, Source: URL - <https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html>
14. Arya Talathi, "Decision Tree Algorithm", 2019, Open web article, Source: URL - <https://medium.com/nerd-for-tech/decision-tree-algorithm-3b5e8b0aed2a>
15. Open web article, Source: URL – <https://scikit-learn.org/stable/modules/tree.html#classification> (L. Breiman, J. Friedman, R. Olshen, and C. Stone. Classification and Regression Trees. Wadsworth, Belmont, CA, 1984.)