International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 10 Issue 5 ǁ 2022 ǁ PP. 11-15

Application of Open Source software (WEKA) in Data Mining related to Agriculture

Mrs.C.Kalaivani,
Assistant Professor, Department of Computer Applications,
Chevalier T.Thomas Elizabeth College for women, Chennai - 11

Ms. R. Rashmika
II BCA, Chevalier T. Thomas Elizabeth College for women, Chennai - 11

Abstract
Large amounts of data are now used by agricultural groups. In this avalanche of agricultural data, the processing and retrieval of critical data is essential. Agriculture is the main focus of research. Agriculture is the main focus of research. Agriculture is the primary source of income in India for a wide range of people and communities. Agricultural enterprises have a lot of raw data and a lot of it. It is required to collect and arrange them in a logical manner, and their integration allows for the establishment of an agricultural information system. Data mining allows agricultural enterprises to foresee trends in customer conditions or behaviour in addition to crop data. This is performed by examining data from a variety of angles and uncovering connections and interconnections among seemingly unrelated facts. The purpose of this report is to get the researcher acquainted with the open-source software WEKA. WEKA is an open-source Java-based software suite that may be used for data preparation, classification, clustering, association rule mining, and visualization. This study also demonstrates what data mining is, as well as its methodologies, methods, and applications in agriculture, utilizing WEKA, an open-source software which can be used by anybody at free of cost, downloadable from the internet.

Keywords: *data mining, agriculture, open source, WEKA, clustering, classification, dataset, data mining*

Date of Submission: 12-05-2022 Date of acceptance: 26-05-2022

I. INTRODUCTION

Agriculture is India's main occupation, and the country's rural economy is entirely based on it. Farming accounts for over 70% of all primary and secondary businesses. As a result, many farmers are attempting to improve their farming operations through the use of new technologies and tactics. On the other hand, most people are blissfully unaware of these technologies and the importance of producing crops at the appropriate time and location. Specifically, in this case identifying a variety of elements that have an impact on productivity in a specific situation is more important. Crop yield and adaptability can lead to improved crop quality and production in terms of higher productivity and economic growth. Crop development is a complicated process in data mining that necessitates the use of agricultural input parameters. Crop production estimates obtained from agricultural data after data mining are used by businesses such as seed, fertiliser, agrochemical, and agricultural machinery to plan production and marketing activities [3]. Farmers and government can benefit from the following two criteria which will be decided after data mining during decision making:

- It aids farmers in risk management by providing previous crop yield records together with forecasts.
- It aids the government in the formulation of crop insurance and supply chain policies.

Data mining is a method for generating new meaningful information from previously collected information from massive databases. Data mining assists businesses in gaining a better understanding of future patterns and features, allowing them to make more educated decisions. Some of the important tools in data mining are tracking patterns, classification, association, outlier detection, clustering, prediction and regression which will be useful to generate important patterns and decisions based on them [7].

The WEKA (Waikato Environment for Knowledge Analysis) is in charge of data mining tool implementations. Under the GNU General Public License, WEKA is free and open-source software that contains machine learning algorithms for data mining tasks.[4] The University of Waikato in New Zealand is responsible for the development of the system. Data pre-processing, classification, clustering, and association rule extraction are the components included in WEKA.

II. DATA MINING IN THE FIELD OF AGRICULTURE

Agriculture data mining is currently a prominent topic in academics. It comprises data mining techniques for agricultural applications. Today's technology may provide a plethora of data regarding agricultural activities, which can then be analysed to find important truths. Precision agriculture is a term that is similar but not identical to precision farming. Information technology has found widespread use in a range of human activities, including agriculture, during the contemporary age. Because of the development and deployment of new information technologies that allow for worldwide networking, agriculture has been nicknamed "IT agriculture." Information technologies are gradually contributing in the development of a systematic approach to agricultural problem-solving. The ability to compile accurate reports, such as those on the use of protective equipment, the number of work hours spent by the machine on a certain crop, or the number of seasonal workers hired, is made possible by having access to the correct information. Agriculture is the study or practice of cultivating land, which includes increasing soil quality over time in order to produce harvests for food, fleece, and other items. As a result of rising urbanization and industrialization, the amount of land under cultivation has dropped considerably over time, and the agriculture business has been badly damaged by population control and climate change. Till now only a few farmers have advanced the usage of progressive techniques, apparatuses, and cultivation approaches. The agricultural industry is highly reliant on data. Agriculture is influenced by cultivation, irrigation, rainwater collecting, fertilizers, climate, soil, pesticides, weeds, and other factors. Companies in the agricultural sector use data mining to estimate production in order to plan and implement supply chain strategies. Predicting crop yields is crucial in agriculture.

The huge quantity of records generated due to those tactics holds a huge ability for enhancing the performance of related sectors. Data mining also can discover hidden facts which could assist farm managers make higher judgments. Data mining is split into types: descriptive records mining and predictive records mining, typically called Knowledge Discovery in Data (KDD).[7] In predictive records mining, there are values to expect in the future, while descriptive records mining jobs give an explanation for the elements of the records in a goal records collection. On the other hand, predictive analysis has a broader range of applications than descriptive analysis. Clustering, classification, affiliation rule mining, regression, and marketplace length estimation are all records mining methods.

III. DATA MINING TECHNIQUES IN AGRICULTURE -WEKA TOOLS

The two most basic forms of data mining algorithms are classification and clustering algorithms. Unknown samples are classified using classification techniques based on information provided by a set of identifiable samples. Because it is meant to teach the classification approach how to perform classification in general, this set is frequently referred to as a training set. Neural networks and support vector machines, for example, require training sets to fine-tune their parameters in order to solve a given classification task. To put it differently these two classifications techniques, use a training set to learn how to categorise unknown samples or samples whose classification is unknown. Because it employs the training set every time a classification is needed, the k closest neighbour classification method, for example, has no learning phase. As a result, the k nearest neighbour classifier has earned the moniker "lazy classifier."

3.1 K-means clustering using WEKA:

The technique of arranging a collection of abstract things into groups is known as clustering. Remember that a group of data components might be considered a single entity. Cluster analysis separates the data set into groups and names them based on data similarity.

K-means Clustering is an unsupervised learning method that is simple to implement. The records items (n') are separated into 'k' clusters, with every commentary being allotted to the cluster with the nearest mean. It defines a total of 'k' sets, one for each cluster k n, with the point serving as the centre of a one-dimensional or two-dimensional figure. The clusters are separated by a significant amount of distance. After that, the information is categorised into usable data sets and linked to the closest collection. The first stage is more difficult to finish if no data is waiting; in this situation, an early grouping is conducted. The barycenters of the clusters from the previous stage must be recalculated for the 'k' new set. The identical data set points and the nearest new sets are connected together after these 'k' new sets are created. After that, a loop is created. The 'k' sets progress in a step-by-step way until the loop produces no more modifications.[6]

K-means are used to evaluate soil fertility. The weighted K-means clustering set of rules may be used to estimate soil fertility. The set of rules makes use of AHP to calculate the load of soil nutrient characteristics. The K-means clustering technique changed into then utilised. Finally, the wise clustering set of rules may be advanced through locating the satisfactory class primarily based totally on operational performance and accuracy. Tests found out that the weighted K-means clustering set of rules has drastically better accuracy and operational performance than the un-weighted clustering set of rules whilst in comparison to the conventional K -means clustering set of rules; a complete assessment of the adjustments in soil vitamins after precision fertilisation the use of the set of rules. The soil fertility degree has significantly advanced after years of chronic precision fertilising. The findings suggest that the advanced clustering set of rules is a beneficial device for figuring out general soil fertility. [8]

The following are the steps to be followed to implement K-means Clustering in WEKA

Step 1: Import the appropriate dataset into the Weka Explorer in the pre-processing interface; we'll use the iris. arff dataset.

Step 2: Select the selection button from the explorer's 'cluster' tab to perform clustering. A dropdown list of available clustering algorithms appears as a result of this phase, and the simple-k-means algorithm is chosen.

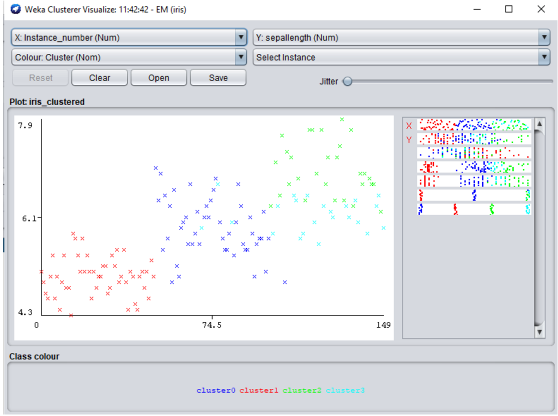
Step 3: Then, to the right of the pick icon, tap the text button to bring up the popup window seen in the screenshots. We enter three for the number of clusters and leave the seed value blank in this window. The seed value is used to generate a random number that is used to internally assign cluster instances to one another.

Step 4: One of the choices has been chosen. We need to make sure they're in the 'cluster mode' panel before we run the clustering technique. The choice to use a training set is selected, after which the 'start' button is pressed. The screenshots below show the process and the resulting window.

Step 5: The centroid of every cluster, in addition to statistics at the range and percentage of times assigned to every cluster, are displayed withinside the end result window. Each cluster centroid is represented through an average vector. This cluster may be used to explain a cluster.

Step 6: Visualizing the traits of each cluster is another approach to grasp them. Right-click the result set on the result to accomplish this. Using the list column to choose to visualise cluster assignments.

Figure 1*(K-means clustering)*



Weka Clusterer Visualize: 11:42:42 - EM (iris)

X: Instance_number (Num)

Y: sepallength (Num)

Colour: Cluster (Nom)

Select Instance

Plot: iris_clustered

Source: (GreeksforGreeks, 2021)

3.2 Decision Tree Analysis using WEKA:

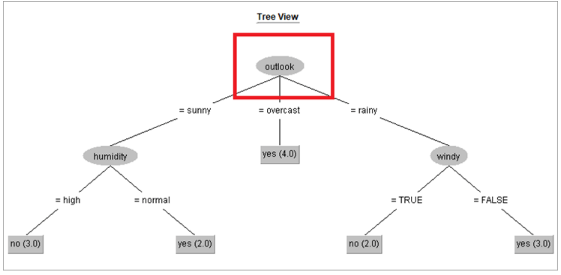
A decision tree is a supervised learning technique used in data mining for classification and regression. It is a tree that assists us in making decisions. As a tree structure, the choice tree creates class or regression fashions. It divides a record set into smaller subsets while also developing a choice tree at the same time. A tree containing choice nodes and leaf nodes is the very final tree. A choice node has at least one branch. The leaf nodes represent a class or option. We can't achieve further separation on leaf nodes, which are the tree's uppermost choice nodes and correspond to the fine predictor known as the foundation node.[1][5]

Each express and numerical record can be addressed by decision bushes. By inputting average temperature, humidity, and pressure, it can be used to predict events such as fog, rain, and thunder. Which farmers and individuals from all walks of life may utilize to make informed judgments. Weather forecasting is the usage of technology and era to count on atmospheric situations for a sure region and time. Weather forecasts are created with the aid of accumulating quantitative records at the cutting-edge situation of the ecosystem at a certain region and the usage of meteorology to estimate how the ecosystem will change. Farmers rely on weather forecasts to decide what work to do on any particular day. For example, drying hay is only feasible in dry weather but on the other hand, prolonged periods of dryness can ruin cotton, wheat, and corn crops.

Weather has a tremendous effect on agricultural production. it has a significant impact on a crop's growth, development, and yields, as well as the occurrence of pests and diseases, water requirements, and fertiliser requirements. Despite meticulous micro-scale agronomic making plans to suit the neighbourhood climate, plants are subjected to plenty of climate fluctuations from 12 months to 12 months. Abnormal climate happens extra often in nearly all years, places, and seasons. Pest management is necessary to protect farms and crops from insects. Weather forecasts assist farmers in determining when pesticides and herbicides should be applied to avoid crop loss. WEKA affords a number of choice tree category techniques. J48 is a choice tree-primarily based totally categorization algorithm. The choice tree may be proven using the Classify tab.[9] If the

choice tree becomes too crowded, tree pruning may be executed from the Preprocess tab with the aid of removing non-vital attributes and restarting the category process.
The following are the steps to be followed to implement decision tree in WEKA
Step 1: Import the appropriate dataset into the Weka Explorer in the pre-processing interface; we'll use the weather.nominal.arff file for classification.
Step 2: To categorise unclassified data, go to the "Classify" tab. Select the "Choose" option. Choose "trees -> J48" from this menu.
Step 3: select the Start Button. The output of the classifier will be displayed on the right-hand side.
Step 4: Right-click on the end result and choose "visualise the tree" from the menu.

Figure 2 (Decision Tree)



source: (softwaretestinghelp, 2022)

IV. Conclusion

Agriculture is India's primary occupation, and its rural economy is entirely dependent on it. Agriculture data mining is a hot issue in academia right now. Data mining strategies are implemented for agricultural applications. Data mining also can find hidden statistics which can assist farm managers make higher decisions. In agricultural organisations, the usage of statistics mining strategies promotes instances for making right judgments and so gaining aggressive advantage. Agricultural establishments use statistics mining equipment to make higher judgments in a whole lot of sectors, which includes trouble prediction, disorder diagnosis, pesticide optimization, and so on. As a result, we might also additionally argue that statistics mining has benefited agriculture. In this paper we've mentioned Data mining withinside the area of Agriculture and the usage of open supply device referred to as WEKA. We have also discussed the certain data mining techniques available in WEKA namely K-means clustering and Decision tree. The method of arranging a set of summary matters into agencies is called clustering. Remember that a set of statistics additives is probably taken into consideration as an unmarried entity. Cluster evaluation separates the statistics set into agencies and names them primarily based totally on statistics similarity. K-way Clustering is an unmanaged studying approach that is straightforward to implement. A decision tree is a supervised studying method utilized in statistics mining for category and regression. As system studying generation evolves and matures, studying algorithms need to be delivered to the computers of those who paint with statistics and apprehend the software area from which it comes. It's critical that algorithms go away from the lab and attain those who can use them. WEKA is a full-size breakthrough withinside the implementation of system studying withinside the workplace.

References

[1]. Figure 1(K-means clustering). (2021, May 4). [Image]. https://www.geeksforgeeks.org/k-means-clustering-using-weka/. https://media.geeksforgeeks.org/wp-content/uploads/20210518114512/5.PNG
[2]. Figure 2 (Decision Tree). (2022, May 5). [Image]. https://www.softwaretestinghelp.com/weka-datasets/. https://www.softwaretestinghelp.com/wp-content/qa/uploads/2020/07/Image-19-Decision-tree.png
[3]. Anjali, J. (2019, October 4). Decision Tree Analysis. The Investor Book. Retrieved 7 May 2022, from https://theinvestorsbook.com/decision-tree-analysis.html
[4]. L. (2017, December 22). The 7 Most Important Data Mining Techniques. Data Science Central. Retrieved 15 May 2022, from https://www.datasciencecentral.com/the-7-most-important-data-mining-techniques/
[5]. Majumdar, J. (2017, July 5). Analysis of agriculture data using data mining techniques: application of big data - Journal of Big Data. SpringerOpen. Retrieved 15 May 2022, from https://journalofbigdata.springeropen.com/articles/10.1186/s40537-017-0077-4
[6]. Wikipedia contributors. (2022, February 21). Data mining in agriculture. Wikipedia. Retrieved 15 May 2022, from https://en.wikipedia.org/wiki/Data_mining_in_agriculture
[7]. Mohyuddin, U. (2021, December 1). Data Mining in WEKA. Baeldung on Computer Science. Retrieved 15 May 2022, from https://www.baeldung.com/cs/weka-data-mining
[8]. Mohan, A. (2021, December 7). Decision Tree Algorithm with Hands-On Example - DataDrivenInvestor. Medium. Retrieved 15 May 2022, from https://medium.datadriveninvestor.com/decision-tree-algorithm-with-hands-on-example-e6c2afb40d38
[9]. Imad Dabbura. (2018, September 17). K-means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks. Towards Data Science. Retrieved 7 May 2022, from https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a
[10]. Web Reference - https://www.techopedia.com/definition/25827/knowledge-discovery-in-databases-kdd
[11]. Abuzir, Dr. Yousef; Al-Quds Open University - https://dspace.qou.edu/bitstream/194/948/2/1441-5206-1-RV.pdf

[12]. Comparative Analysis of Classification Algorithms on Different Datasets using WEKA, Rohit Arora M.Tech. CSE Deptt. Hindu College of Engineering Sonepat, Haryana, India and Suman Asstt. Prof. CSE Deptt. Hindu College of Engineering Sonepat, Haryana, India https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.258.9202&rep=rep1&type=pdf