Crop Suggestions using Data Mining Approaches

Ameya Salankar
Department of Computer Science and Information Systems
BITS Pilani, Hyderabad Campus
Hyderabad, India
f20170182@hyderabad.bits-pilani.ac.in

Atharva Sune
Department of Computer Science and Information Systems
BITS Pilani, Hyderabad Campus
Hyderabad, India
f20170183@hyderabad.bits-pilani.ac.in

Prakhar Suryavansh
Department of Computer Science and Information Systems
BITS Pilani, Hyderabad Campus
Hyderabad, India
f20171017@hyderabad.bits-pilani.ac.in

Harsh Kumar
Department of Computer Science and Information Systems
BITS Pilani, Hyderabad Campus
Hyderabad, India
f20171584@hyderabad.bits-pilani.ac.in

***Abstract*— We propose a solution using Association Rule Mining and Clustering where, given the conditions, a farmer will be suggested the suitable crop(s) that he can cultivate based on the past years' data.**

Keywords—Data Mining, Association Rule Mining, Clustering, Agriculture

I. Introduction

Data Mining is the extraction of hidden useful information from huge databases and is a groundbreaking new innovation with extraordinary potential to assist organizations with concentrating on the most significant data in their information stockrooms. Agriculture constitutes the single biggest part of India's GDP, contributing about 25% of the aggregate and almost 60% of Indian populace relies upon this occupation. Data Mining techniques can prove to be very useful in helping the farmers in growing the suitable crops.

Many times farmers need to know the optimal crop to be cultivated, given the specific conditions of rainfall, temperatures and region. They would benefit largely if they have a rough idea on the choices of crops to grow which could maximize their production.

A lot of solutions to the above problem based on neural networks and data mining techniques have been proposed by researchers before but they target yield prediction from the given data.

In this paper, we aim to suggest a set of crops, based on a predefined set of parameters, that will maximize the production of the crop. The proposed framework uses Data Mining techniques like Association Rule Mining and Clustering to achieve the solution.

II. Literature Review

There have been various research studies that lay emphasis on the significance of utilizing data mining as a strengthening device in changing enormous volumes of agricultural data into meaningful data.

A research survey by Dr. D. Ashok Kumar and N. Kannathasan focuses on different data mining techniques we can use in agriculture. The research survey conducted by them reveals that a comparison of various data mining techniques can be used to develop an efficient algorithm [1].

Data mining techniques in agriculture are often used to study soil attributes. For instance, the K-Means approach is utilized for classifying soils, also using GPS-based advancements [3] to classify soils and plants and SVM to classify crops [2].

A study shows the impact of climatic factors on major kharif and rabi crops in Bhopal District of Madhya Pradesh State [4]. The study indicated that the yield of soybean crop was for the most part impacted by Relative humidity followed by Rainfall and Temperature. The yield of paddy crop was for the most part affected by Rainfall followed by Relative humidity and Evaporation. For Wheat crop the study showed that the profitability is for the most part impacted by Temperature followed by Relative humidity and Rainfall. The studies clearly show that rainfall is one of the major factor affecting crop productivity.

The three major factors affecting crop yield are Season, Soil Type and Rainfall.

III. Techniques

A. Normalization

In many data mining tasks, variables that have vastly different scales from each other which may cause problems e.g. height in inches vs cms or weight in pounds vs kilos Expressing an attribute in smaller units will lead to a larger range for that attribute, and thus tend to give such an attribute greater effect. To avoid dependence on the choice of measurement units, the data should be normalized or standardized.

In **z-score normalization** (or zero-mean normalization), the values for an attribute, A, are normalized based on the mean (i.e., average) and standard deviation of A.

$$v_i = \frac{v_i - \mu}{\sigma}$$

B. Binarization

Binarization maps a continuous or categorical attribute into one or more binary variables. It is typically used for association analysis. It often converts a continuous attribute to a categorical attribute and then converts the categorical attribute to a set of binary attributes. This was done here because association analysis needs asymmetric binary attributes.

C. Data Visualisation

1) Box Plots

Box plots are a convenient way of showing interval data through the use of its quartiles. It shows different characteristics of the data - the minimum, the first quartile (i.e. data at the 25th percentile), the second

quartile (data at the 50th percentile), the third quartile (the data at the 75th percentile) and the maximum. In addition to this, the outliers are also shown.

Box plots are useful because they indicate the spread of the data visually, around the median of the data. They can handle large data easily. Also, outliers can be visually identified with them.

2) Heat Maps

A heat map (or heatmap) is a graphical representation of data where the individual values contained in a matrix are represented as colors.

Heat maps are used when there is a need to show data which is dependent on two independent variables as a color coded image plot. The values represented by each cell of a heatmap are represented by colors, and intensities. Generally, the darker the color, the lower is the value. But there are various variations to this.

A careful observation of heat maps can reveal outliers and data points that don't follow the general trend.

3) Line Plots

A line chart or line plot or line graph or curve chart is a type of chart which displays information as a series of data points called 'markers' connected by straight line segments.

A line chart is very helpful in displaying temporal (time-dependent) data.

When plotting temporal data, the X-axis represents time in seconds, minutes, years, etc. and Y-axis represents the attribute/value being evaluated over time.

D. Clustering

1) K-Means Clustering

Typically, unsupervised algorithms make inferences from datasets using only input vectors without referring to known, or labelled, outcomes. The K-means algorithm tries to group similar data points together and discover underlying patterns. K-means looks for a fixed number (k) of clusters in a dataset. The K-means algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the clusters as small as possible. After each iteration, the centroids of each cluster is updated. The process stops when there is very less change in the centroids for successive iterations. The input 'k' is user defined. But it is generally calculated by a method known as the 'Elbow Method'. Here, the average sum of squared distance of each point to its cluster centre is plot against the number of clusters. The point at which a sudden change in the average distance is observes is called the 'elbow point'. The number of clusters corresponding to this elbow point is taken as 'k'.

2) DBSCAN

Density-based spatial clustering of applications with noise (DBSCAN) is a data clustering algorithm. It is a density-based clustering algorithm. Given a set of points, it puts all the densely packed points in the same cluster and marks the points which lie alone as noise (outliers). It is one of the most popular clustering algorithms.

It defines two parameters – ε, the radius of a neighborhood with respect to some point and MinPts, the minimum number of points that should be in epsilon neighborhood of a point for it to be classified as core point. There are 3 types of points: core points, density reachable, and outliers. A point p is a core point if at least minPts points are within distance epsilon of it (including p). A point q is directly reachable from p if point q is within distance epsilon from core point p. A point q is reachable from point p, if there exists a chain of points from p to q, all of which are directly reachable from one another, i.e: p->p+1->p+2…. ->q. All points not reachable from any other point are outliers or noise points. If p is a core point, then a cluster is formed together with all points (core or non-core) that are reachable from it.

E. Association Rule Mining

The aim of Association Rule Mining is to find out rules which give the items appearing most frequently together in dataset. The main way by which this is achieved is by finding how many times every possible set of items appear in the dataset. But this task is computationally extensive. Hence, two algorithms have been developed to tackle this issue.

1) Apriori

Apriori is an algorithm for frequent itemset mining and association rule learning. It works by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database. The frequent itemsets determined by Apriori can be used to determine association rules which highlight general trends in the database. It is faster than the brute force method because once an itemset is found to be infrequent, all its supersets are discarded from the list of frequent itemsets.

2) FP Growth

FP Growth is another algorithm for frequent itemset mining. FP here stands for Frequent Pattern. This method was developed to handle extremely large databases. In this method, the data is first preprocessed. Infrequent items are discarded entirely. Each transaction of the dataset is then sorted according to the frequency of the items. All the transactions in the dataset are stored in a tree. This tree is also known as an FP-Tree.

This tree can be then mined for frequent itemsets. Items are considered in the increasing order of the frequencies. For each item, a new conditional tree is constructed by starting from every occurrence of that item in the tree and reaching to the root from that node. The support count of each item in the conditional tree is updated. The items which don't satisfy the minimum support requirement are pruned. This process is performed recursively. Growth of the itemset stops if there is only the root node in the conditional tree of an itemset. All the frequent itemsets are thus generated.

IV. METHODOLOGY

A. Dataset

The dataset used was the crop production dataset provided by the Ministry of Agriculture and Farmers Welfare, Department of Agriculture, Govt. of India. The dataset consists of crop production statistics in each season and district of India. This dataset was merged with a dataset

containing state-wise, month-wise rainfall of India from 1997 to 2015, provided by IMD, Pune.

B. Data Cleaning and Preprocessing

Data cleaning was performed on the dataset. Missing values were handled by ignoring some tuples and using mean to fill the remaining NaN values.

Data Preprocessing was performed using the following techniques

- **Numerosity Reduction**: Stratified sampling was used to reduce the data size to 70 percent since the data was very large.
- **Feature Construction**: A new feature named '**yield**' (crop production per unit area) was added to the dataset. This was done because it combined the information of two attributes - the crop production and the area under production, in a useful way. It increased insight gained.
- **Normalization:** Z-score normalization was used to normalize *rainfall* and *yield* to bring them to the same scale.
- **Discretization:** *Rainfall* and *Yield* fields were discretized into five categories - *very_low*, *low*, *medium*, *high* and *very_high*. This was done in order to do cluster analysis and classification.
- **Binarization**: In order to do association analysis Rainfall and Yield fields were mapped to five binary variables corresponding to the five categories. A binary variable represents the presence or absence of a record in that category.

C. Data Visualisation

Box Plots, Heat Maps, Line charts were plotted from the data and studied.

- A **heatmap** is plotted with the seasons, Kharif, Rabi, Autumn, Spring, Summer, Whole Year, on the X-axis and crops on Y-axis. (Season-Crop Production). The cells of the matrix represent the normalized production values of the crops in the respective season.
- After preprocessing the data, the rainfall values have been discretized into 5 categories *Very Low*, *Low*, *Medium*, *High*, *Very High* and another **heatmap** was plotted with crops on Y-axis and cells having the value of production (normalized)
- **Line charts** have been plotted for every state .The line charts are plotted for net production of state over time. X-axis denotes years and Y-axis denotes the net production in those years
- A **scatter plot** was made between average rainfall and production and the crops. The rainfall values were averaged over districts and time.
- **Box Plots** were plot for each crop. The variables taken into consideration were area, production and rainfall added over all the years and all the districts. The fields were further normalized to ease the visualization. The plots help us to know the median values required of rainfall and area for each of the crops. They also show the median values of production for each crop.

D. Clustering

1) K-Means Clustering

K-Means Clustering algorithm was applied to divide the data points into clusters. When a new input is given by the user, it is assigned one of the clusters and then the suitable crops from that cluster is given as the output. The inputs asked from the user are *Season*, *Soil Type* and *Rainfall*. By using the elbow method and after some trial and error, dividing into 12 clusters was found to be optimal. The inputs from the user were used as the attributes for the K-Means algorithm. The data was first binarized so that it could be fed to the K-Means Clustering algorithm. 12 random points were chosen as the initial mean for each cluster. Distance of each point from these 12 mean points was calculated and the point was assigned to the cluster with the minimum distance. After each iteration, the mean was updated and the process repeated with the new mean. The iterations stopped when there was no change in points in each cluster.

2) DBSCAN

The data was preprocessed by converting the categorical labels via one hot encoding. Then the data was passed to the algorithm, which first found the core points and the neighbors of all the points. Then it made clusters using only the core points (basic clusters). Then for each cluster, it found all the reachable points. All the points that remained unclassified after the formation of the clusters were termed as NOISE/OUTLIERS.

E. Association Rule Mining

In the preprocessing stage, the rainfall and yield columns of the dataset were discretized. The dataset needed further preprocessing for Association Rule Mining. It was further processed by appending the column name to each data cell of the column. Thus, for e.g., *very_low* in the *Rainfall* column became *very_low_rainfall* and the same was done for the *Yield* column. Thus, the dataset now consisted of five columns – *Rainfall*, *Soil*, *Season*, *Crops* and *Yield*. Each row of the dataset was treated as a transaction and each individual element in the row as an item of the transaction.

The data was then gone through once and the support count of each item in the dataset counted. The items below a support of 50 were discarded. This corresponds to 0.03% of the dataset being discarded. Each of the transactions in the dataset was sorted according to their support counts.

Using this dataset, two Association Rule Mining methods were applied.

1) Apriori Algorithm

Apriori algorithm was applied on the dataset to find frequent itemsets and then interesting rules from those datasets. The rules will predict the crops and their yields given other factors.

2) FP Growth Algorithm

An FP Tree was constructed on using the transactions from the dataset. This FP Tree was subsequently mined for the frequent patterns.

The frequent patterns of length five were extracted from the set of mined frequent patterns from both of these

algorithms separately. This was done because the number of columns in our original dataset was five and all the columns were needed to predict the crops from the set of given conditions.

Using the extracted frequent patterns, rules were generated. The rules had the antecedent as a three-tuple of *Season*, *Soil* and *Rainfall*. The consequent was a two-tuple of *Crop* and *Yield*. Confidence of each of these rules was calculated.

$$Confidence = \frac{Support(Rule)}{Support(Antecedent)}$$

A threshold confidence of 0.02 was set. All the rules having confidence below this value were discarded. This corresponds to 30% of the rules that were discarded.

The crops could now be predicted using these rules. The antecedent can be taken as an input from the user. All the rules with the same antecedent as the input are marked. They are then sorted according to the confidence values and returned. Only rules with *Yield* as *very_high*, *high* or *medium* can be considered.

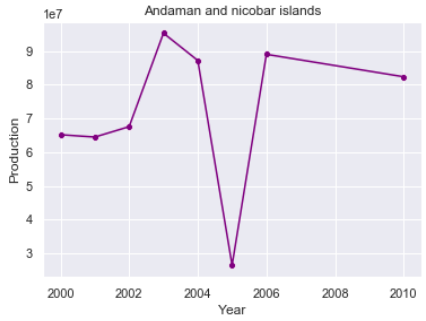

Figure 1

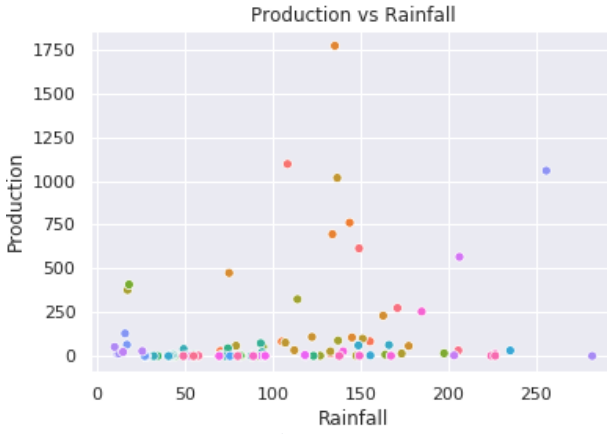

Figure 2

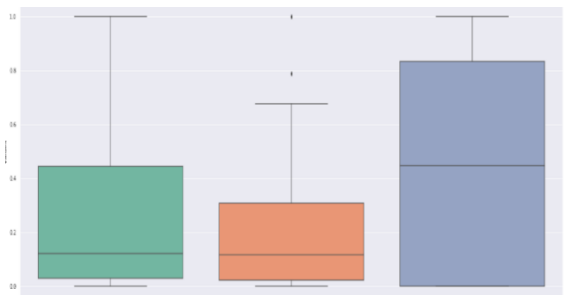

Figure 3

V. Results And Discussion

The heatmap (Figure 1) is plotted between Rain-Crop-Production. X-axis has discretized rain classes and Y-axis is the crops. It was observed that rainfall alone is not enough to determine the production of crops. It can be seen from the column of *Very_Low* rainfall. Though most crops follow the trend with having good production values in *Medium* to *High* rainfall environments and almost no production in *Very_Low* and *Very_High* rainfall. But some crops do have a good amount of production in these conditions as well.

The line chart (Figure 2) is one of the 33 line charts that are obtained for each state. The line charts depict the change in net production of the state over the years.

A cursory glance on the production vs rainfall scatter (Figure 3) plot shows us the clusters of crops which require a similar amount of rainfall. There appear to be some crops which have a larger production than others who had similar rainfall. This plot gives us an indication of the closeness of some crops to others.

Box plot of crops such as Figure 4 (of misc. pulses) shows us the variance in the parameters over different places and years. This shows us the median rainfall required to grow this crop and the median area that people have used to grow in the past. This gives us a fair idea of the variance of the values we could accept for this crop.

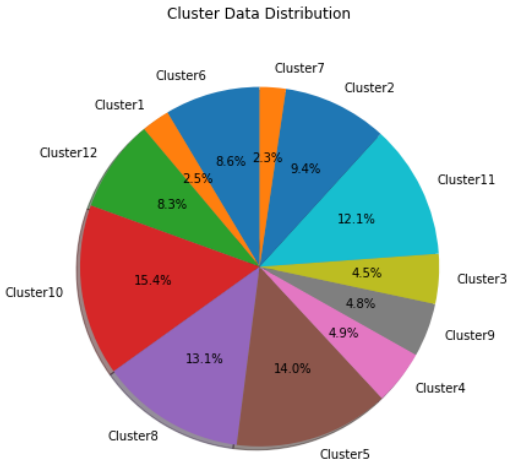

Figure 4

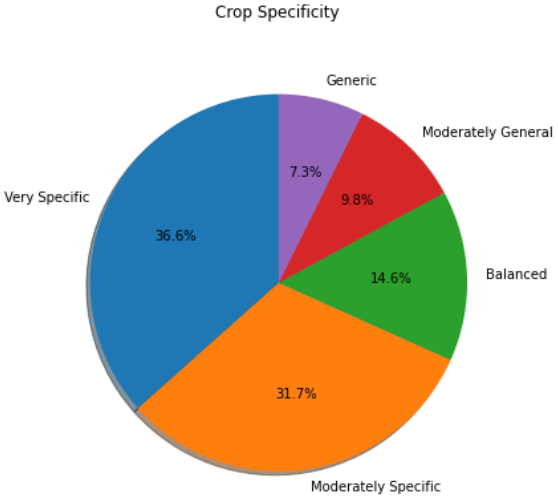

Figure 5

In K-Means Clustering, after applying the elbow method and after some trial and error, 12 clusters are found to be optimal. Dividing the dataset into these 12 clusters, it is found that crops that are grown in similar type of conditions tend to fall in the same cluster. Figure 4 shows the relative sizes of the cluster. Figure 5 shows the specificity of each crop. A crop with low specificity appears in majority of the clusters (i.e. > 9 clusters). Thus, this crop is likely to grow with less regard to the given conditions. On the other hand, a crop with a high specificity is very specific and belongs to a less number of clusters.

In DBSCAN, For the given dataset the algorithm was observed to take a large amount of time. The dataset contained 200,000 distinct data points, and since neighbors for all needed to be found, distance for every point from every other point needed to be calculated. The machine implementing the algorithm took approximately 3 days to find the basic clusters consisting only of the core points. Thus the use of DBSCAN on this dataset was abandoned.

In Association Rule Mining, using the mentioned values of support and confidence (i.e. 50 and 0.02 respectively), 7551 different frequent itemsets were generated. As many as 782 distinct frequent itemsets of length five were identified. Thus, 782 distinct rules were generated. The minimum confidence threshold could be changed at will depending on the output. It was further found that the returned sets were satisfactory when the antecedent contained values which were highly frequent in the dataset and were generally empty otherwise. One reason for this could be that the itemsets from which the rules were generated were discarded because they did not satisfy the minimum support threshold.

VI. Conclusion

In Clustering, crops that are grown in similar type of conditions fall in the same cluster. This was used to suggest the crops which give high yield when a condition is given as input. Also, the use of DBSCAN for a large dataset, such as the one used here, was found inefficient.

In Association Rule Mining, since each of the data cell was considered an item of a transaction, it was possible to form rules which could suggest the crop (with their yield values) for the given input conditions. The set of the returned crops from these rules represent the crops which have a higher probability of producing a high yield in the given conditions based on the past data. Hence, the farmers can plant any of the crops depending on the cost and availability, from the set, singularly or using farming techniques like crop rotation.

Moreover, the results obtained from these two different techniques can be used as potential fallbacks in case one of the technique fails to give a satisfactory result set.

It is hoped that this will benefit the farmers immensely.

VII. Future Scope

There is much scope ahead regarding this area.

- More parameters can be used while applying the techniques mentioned here. For e.g. *Humidity*, *Region*, *Temperature*, *pH of Soil*.
- Prediction of *Yield* classes can be done when specific conditions and a *Crop* is given.
- The algorithms could be modified to incorporate real time data to obtain updated models.

References

[1] D, Ashok Kumar & Kannathasan, N. (2011). A Survey on Data Mining and Pattern Recognition Techniques for Soil Data Mining. International Journal of Computer Science Issues. 8.

[2] Mathur, Ajay & Foody, Giles. (2008). Crop classification by support vector machine with intelligently selected training data for an operational application. International Journal of Remote Sensing - INT J REMOTE SENS. 29. 2227-2240. 10.1080/01431160701395203.

[3] Chougule, Archana & Jha, Vijay & Mukhopadhyay, Debajyoti. (2019). Crop Suitability and Fertilizers Recommendation Using Data Mining Techniques. 10.1007/978-981-13-0224-4_19.

[4] Mall, Rajesh & Sonkar, Geetika & Sharma, Narendra & Singh, Nidhi. (2016). Impacts of Climate Change on Agriculture Sector in Madhya Pradesh. 10.13140/RG.2.1.3010.0247.

[5] Dr. Rahul G. Thakkar, Dr. Manish Kayasth, Hardik Desai, "Rule Based and Association Rule Mining On Agriculture Dataset," International Journal of Innovative Research in Computer and Communication Engineering, Vol. 2, Issue 11, November 2014.

[6] Tan, Pang-Ning & others, Introduction to Data Mining, Pearson Education, 2006.

[7] Han J & Kamber M, Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers, Second Edition, 2006.

[8] Rakesh Agrawal and Ramakrishnan Srikant Fast algorithms for mining association rules. Proceedings of the 20th International Conference on Very Large Data Bases, VLDB, pages 487-499, Santiago, Chile, September 1994.

[9] Han (2000). Mining Frequent Patterns Without Candidate Generation. Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data. SIGMOD '00. pp. 1–12.

[10] Data-Mining-CP repository on GitHub, Data Mining Course Project https://github.com/AtharvaSune/Data-Mining-CP