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

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, Clustering and Classification to achieve the solution.

II. Background

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. 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.

D. 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.

E. 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.

III. Methodology

Before you begin to format your paper, first write and save the content as a separate text file. Complete all content and organizational editing before formatting. Please note sections A-D below for more information on proofreading, spelling and grammar.

Keep your text and graphic files separate until after the text has been formatted and styled. Do not use hard tabs, and limit use of hard returns to only one return at the end of a paragraph. Do not add any kind of pagination anywhere in the paper. Do not number text heads-the template will do that for you.

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.

IV. Results And Discussion

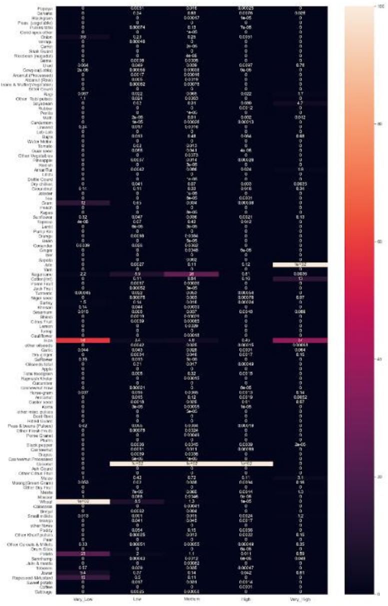

Figure 1

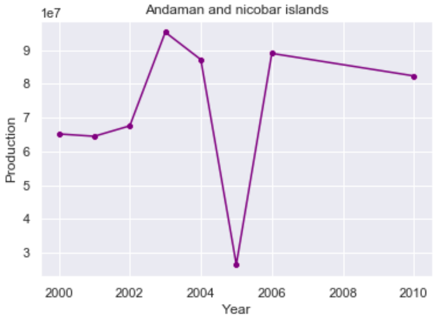

Figure 2

The heatmap (Figure 1) is plotted between Rain-Crop-Production. X-axis has discretized rain classes and Y-axis is the crops. From a quick analysis it can be 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.

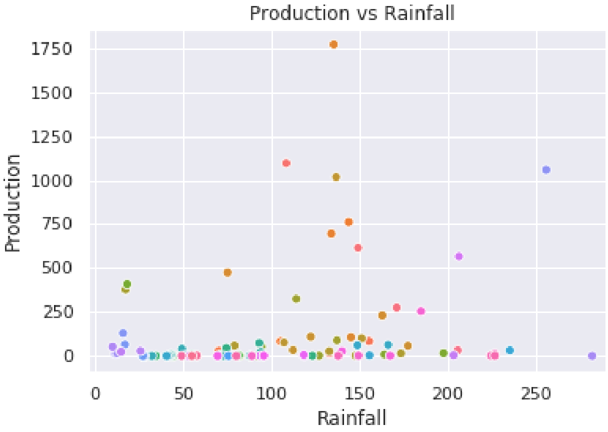

Figure 3

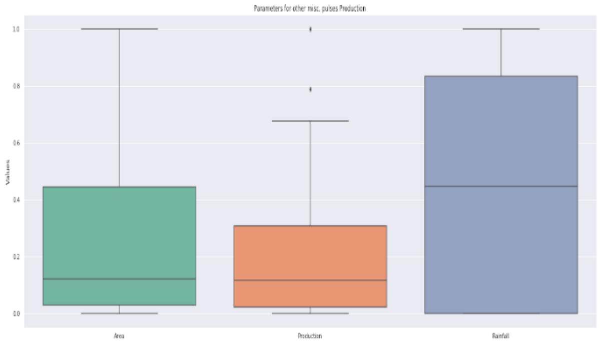

Figure 4

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.

REFERENCES

[1] 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.

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

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