A recommendative system for Indian Agriculture using Supervised Learning Techniques

Gowri Srinivasa
Department of Computer Science and Engineering
PES University
Bangalore, India
gsrinivasa@pes.edu

Guruprasad M, Jai Agarwal, Shreyas Nitin Pujari
Department of Computer Science and Engineering
PES University
Bangalore, India
mguru1998@gmail.com, jai.bhageria@gmail.com, pujari.shreyas@gmail.com

***Abstract*—Agriculture is the most important economic sector for a developing country like India, whose major population resides in rural areas. Agriculture forms a major portion of India's GDP, and is a livelihood for major part of the population of India. Precision Agriculture is a technique that involves studying crops and the factors associated with agriculture like weather, soil type, fertilizers among others, in detail supported by a variety of data. Insights from this data can help us predict crop trends and also conserve resources by suggesting correct quantities of resources to be utilised. This could help save a lot of resources and time and an effective advise could be beneficially economical as well. In this project we look to develop a recommendation system to suggest farmers with the correct crops they should plant looking at various factors like land area, weather patterns and crop season. This will help them save time and money and hence maximise their yield and also satisfy consumer demands effectively.**

***Keywords*—Precision, Agriculture, Neural Network, Regression, Supervised, Learning, Time Series Analysis, Decision Tree**

I. Introduction

Agriculture, in simple terms, refers to cultivation and harvesting of crops. It is the Indian economy's most important sector in terms of employment and food production. A majority of the nation's population is dependent on agriculture directly or indirectly. Some of the main crops grown in India include Wheat, Rice, Maize, Sugarcane, Cotton, Coffee, Coconut and Tea. All these crops require varied climatic conditions and crop seasons for optimal growth.

Precision agriculture refers to the methodology of ensuring high yield of crops using minimal resources. This is achieved by applying the suggesting from various decision support systems. This method is very effective compared to classical methods as these systems work based on real-time data and also provide a very high accuracy.

The current agricultural system in India is not very effective. A high yield of crops cannot be generated because most of the times the farmers do not know which crops to plant in which part of the country for particular crop seasons like Kharif and Rabi. We try to improve on that particular aspect by looking at the dataset of the crops, area and production over the years. We also try to study the weather patterns in certain parts of the country and try to estimate as to what the production of a particular crop in a particular area would be, during the various crop seasons. An efficient recommendative system would help the farmers in taking the right decisions and hence maximizing their revenue by optimal utilization of the available land and resources during various parts of the year.

The organization of the rest of this report is as follows: Section II presents a summary of the past related work carried out in this domain, Section III describes the proposed problem statement, Section IV discusses the proposed solution for the problem statement, Section V illustrates the results obtained and the relevant discussions and Section VI describes the conclusions and future work planned to be carried out in order to obtain a better and optimum solution to the proposed problem statement. Section VII describes the role of each team member.

II. Related Work

In the past, work in this domain has been carried out in order to develop predictive and recommendative models for increasing the production of crops with the help of precision agriculture techniques. A summary of the works are presented in this section.

Lakshmi. N et al (2018) [1] in their work have shown how large amounts of data can be mined effectively and can be used to build predictive systems. They have developed a system which considers various factors and predicts appropriate soil types as well as crops that are suited for

those soil types based on weather patterns over the years. For storage of the large volumes of data, Hadoop distributed File system is used. Using the Hadoop Map-reduce framework, weather data for multiple years have been aggregated for 28 locations in order to generate cumulative rainfall. Soil related attributes considered include texture, pH, drainage and permeability. Crops like Paddy, Pulses, Sugarcane and Cotton were also considered. The difference in weather pattern between successive years has been determined by measuring the Euclidean Distance and similar weather patterns have been identified using the Nearest Neighbours technique. The aspect of prediction using various factors and patterns has been incorporated in our work.

Usage of Supervised learning techniques like Random Tree, K-Nearest Neighbour and Naive Bayes Classifiers for recommending crops based on various parameters at farming locations has been demonstrated in their work by S.Pudumalar et al (2016) [2]. They have collected the soil specific attributes of various farmlands in Madurai district to be used as one of the datasets. Another dataset comprises the crop data of the main crops grown in that region. A highly used ensembling technique known as Majority Voting Technique has been used in order to effectively suggest the suitable crop for the characteristics of that particular farmland. An accuracy of 88 percent has been obtained by their system. The usage of Naive Bayes classifier for predicting crops has been incorporated in our system.

John M.Antle et al (2016) [3] have described the usage of Agricultural System models in improving the predictive capabilities of farming parameters. Emphasis has also been laid on the potential of these models in order to aid the development of newer models to improve the quality of information made available to the decision-makers in the Agricultural sector. They have also spoken about the change in approach - from a Supply Driven Approach of model development to a Demand Driven approach which utilises user data, and also the need for the development and usage of products which encourage the usage of outputs of various models. The need for computational research, which is significantly better than classical methods for the purpose of efficient agricultural recommendations has also been stressed upon by the authors. The aspect of usage of computational and a demand-driven approach has been incorporated in our work.

Daniel Rodriguez et al (2018) [4] have highlighted the need for modern and advanced methods to bridge the productivity gap and increase profits of farmers. They have mentioned that by understanding of a crop environnment, seasonal climate forecasts, projected food supply/demand gap and crop genetics, optimum crop designs could be predicted which is highly effective. Also presented in their report is the fact that increasing crop production per unit area of available land is better than incorporating new land. Growing of crops

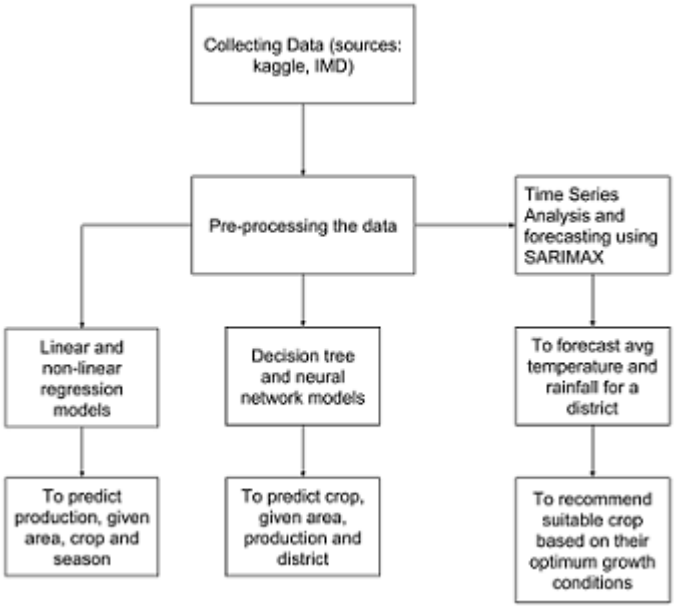

Fig. 1. A block diagram of the proposed system depicting the various components

is affected by soil condition and also rainfall. They have also illustrated the high reliability and accuracy of Australia's POAMA-2 forecasting system compared to the SOI phase system. The usage of APSIM sorghum model for crop model simulation has also been elucidated by the authors. The aspect of using projected food supply/demand gap in order to design and improve the crop recommendation model has been incorporated in our system.

III. Proposed Problem Statement

To develop a crop recommendation system for Indian farmers based on multiple factors like seasonal and climatic conditions dependency of certain crops, and also to determine a relation between area and production of crops, so as to recommend optimum utilisation of available land of farmers using Supervised Learning approaches.

IV. Proposed Solution

The block diagram (Fig. 1) shown above depicts the various stages of our proposed solution which includes data collection, pre-processing and application of various models.

A. Data collection

The main dataset, which contained Area, Production and Season for various crops, district-wise, was collected from Kaggle. The other datasets, which included the average monthly rainfall and average temperatures of certain districts of Tamil Nadu for 102 years and also the optimal growth conditions of few crops, were collected from various online sources.

B. Data pre-processing

First we cleaned the dataset and and trimmed it by removing all the NA values. Basic summary statistics were calculated and it was observed that there were few outlier values, leading to high skewness, hence a 10 percent Trimmed Mean was obtained, which gave a better representation of the data, hence the bottom and top 10 percent of data values sorted by production were removed from the dataset.
In order to handle categorical data, a copy of the dataset was created, wherein the categorical variables were replaced by numbers, based on sequential order. Similar numbers were used to represent similar values.
For the other dataset, the columns were obtained in the required format for Time Series Analysis, that is YYYY-MM-01 format.

C. Usage of Linear Regression, Naive-Bayes classifier and Non-linear Regression

Linear regression method was used to make the predictions. Linear regression is one of the most commonly used predictive modelling techniques. The aim of linear regression is to find a mathematical equation for a continuous response variable Y as a function of one or more X variable(s). So that you can use this regression model to predict the Y when only the X is known. The X here being Area and the Y being Production. The aim of our model is to predict the amount of production (Y) when only the Area(X) is given.
Initially, a scatterplot of production and area was plotted. (Fig. 2).A linear model was then created from the given parameters, wherein the data set was divided into training and testing sets, 70 percent of dataset size being the training and remaining 30 percent being the testing dataset. The classifier was run on the training model and the predictions were made on the testing set. We then plotted the predicted values and the actual values. From the plot, it could be inferred that the relationship between production and the other parameters is non-linear.
The second model tries to predict the crop given the district name, area and production given, For this approach a very generic classifier was used - The Naive Bayes Classifier.

The data was divided into training and testing sets of size. 70 percent size being the training and 30 percent size being the testing dataset. The classifier was run on the training model and the predictions were made on the testing set.

The Naive Bayes classifier takes as input the district name, area and the production in the form of a table and runs it through the model built and gives the appropriate crop as the output. The naive bayes classifier was found not to be a good classification method for the data due to low accuracy, hence we rejected the classifier and move onto non-linear regression. To test for non-linearity, we tried to predict production , using area, crop and season as parameters. We used the LASSO regression technique to predict the values of production by varying the degree of non-linearity from 2 to 6.

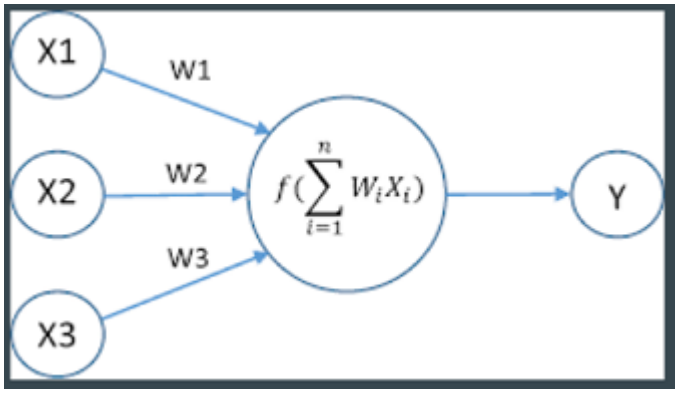

Fig. 2. A block diagram depicting the neural network model used

D. Usage of Decision Tree and Neural Network models

A decision tree is a decision support tool that uses a tree like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. For our model, we used two inputs, Area and Production, in order to generate the output as a crop number, which was then converted to a crop using the encoding technique described above. The crop output predicted is the best possible crop to harvest for the given parameters.

A neural network is a statistical model that uses examples to automatically infer rules for recognizing outputs based on the given inputs. The neural network approach was used because the decision tree approach and regular statistical models weren't accurate enough. Our neural network model has 3 layers: One input layer that has 2 dimensions, The activation function which is relu (rectified linear unit), one hidden layer which is also relu and the output layer, which is a cost function called the softmax function. The use of softmax function in the last layer (output layer) is important because of the presence of categorical data, for which the softmax function is necessary.
We calculated both the training and testing efficiency of the dataset with Adam optimizer and the evaluation metric used was "accuracy".

E. Usage of Time Series Analysis for Forecasting and Knowledge based recommendation

A time series model was built in order to forecast temperature and rainfall. The idea was to obtain the forecasted temperature and rainfall for a particular district, for the upcoming months, which would help us make a knowledge based recommendation on the crop to be planted. This recommendation would help farmers in making the right decisions according to the prevailing weather trends.
SARIMAX model was used to fit the time series model and hence obtain the forecasts. The dataset we worked on was inherently stationary and had a seasonal trend associated with it, which was why we chose SARIMAX as rainfall and temperatures have a obvious seasonal trend associated with them, and X represents exogenous parameters can influence

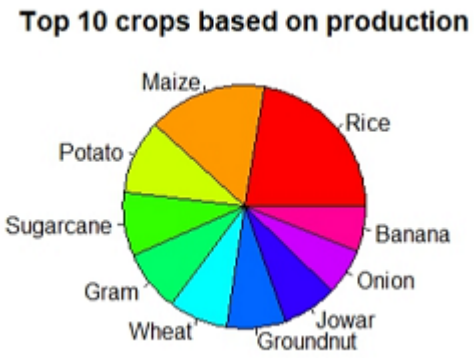

Fig. 3. Pie chart depicting top ten crops by production

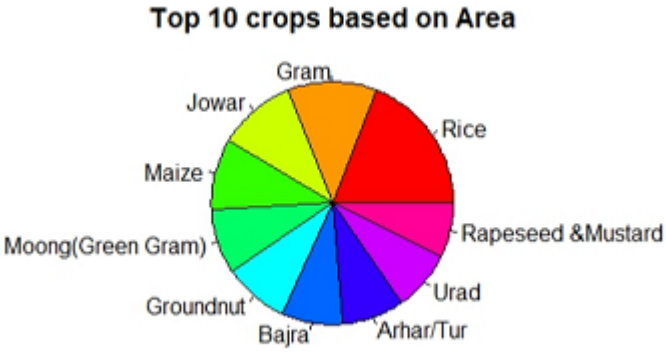

Fig. 4. Pie chart depicting top ten crops by area

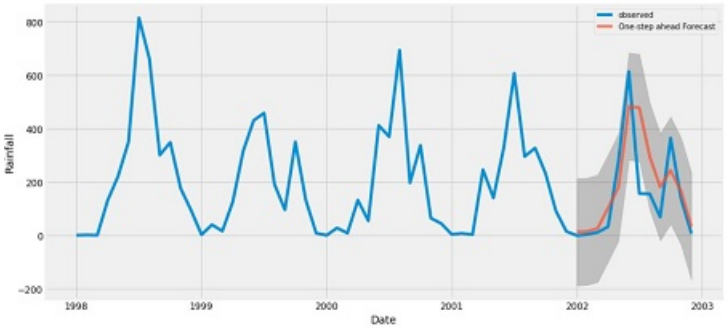

Fig. 5. Graph depicting observed rainfall and one-step ahead forecast

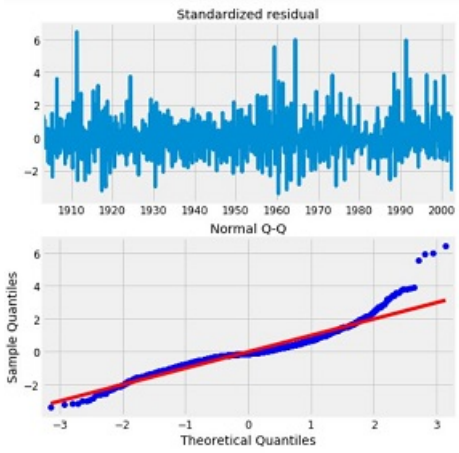

Fig. 6. Graph depicting standardized residual and Q-Q plot for rainfall

weather patterns which we are accounting for by using this model.

The parameters used for evaluating the accuracy is AIC criterion to obtain accurate p,d,q values for model. RMSE score with the observed and predicted values was calculated and also a graph depicting forecasted and observed values was drawn to observe the trends.

Using the forecasted values we calculated a Euclidean distance parametric with available records. A low euclidean distance suggests us the name of the crop which is closest to the forecasted growth conditions.

F. Visualisation

Using the available data, some basic visualisations were carried out, which includle: Pie charts (Fig. 3 and Fig. 4) depicting the top ten crops overall by area and also by production and a map (Fig. 6) showing the districts where farming is carried out.

The time series plots depict the observed rainfall values and the one-step ahead forecasted values for Coimbatore district

V. Results and Discussions

A Q-Q plot (Fig. 10) was generated for the transformed values of Production and Area. This plot nearly resembled a 45 degree straight line, hence we concluded that they are nearly normally distributed.

A scatterplot of log transformed values of Production on the Y axis and Area on the X axis was obtained and the correlation co-efficient was calculated. The obtained correlation co-efficient was low, hence the linear regression model was rejected.

A Naive Bayes classifier was used in order to predict crop based on district name, area and production. The

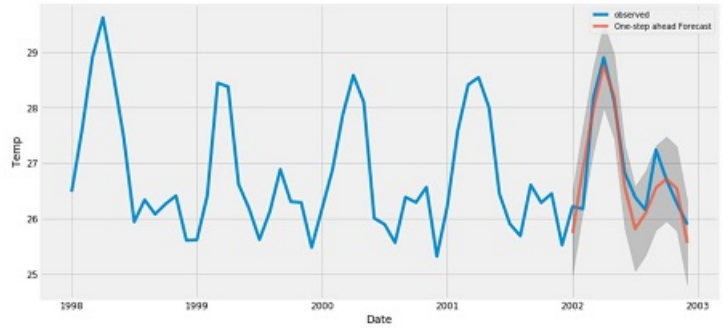

Fig. 7. Graph depicting observed temperature and one-step ahead forecast

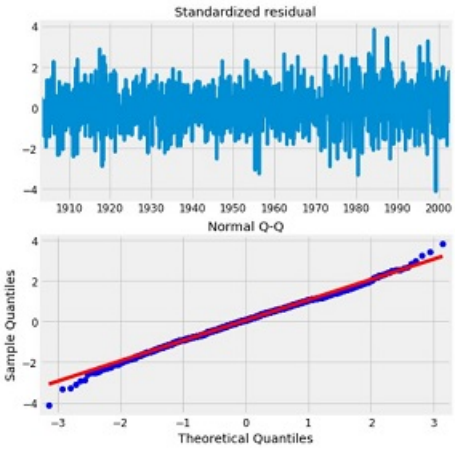

Fig. 8. Graph depicting standardized residual and Q-Q plot for temperature

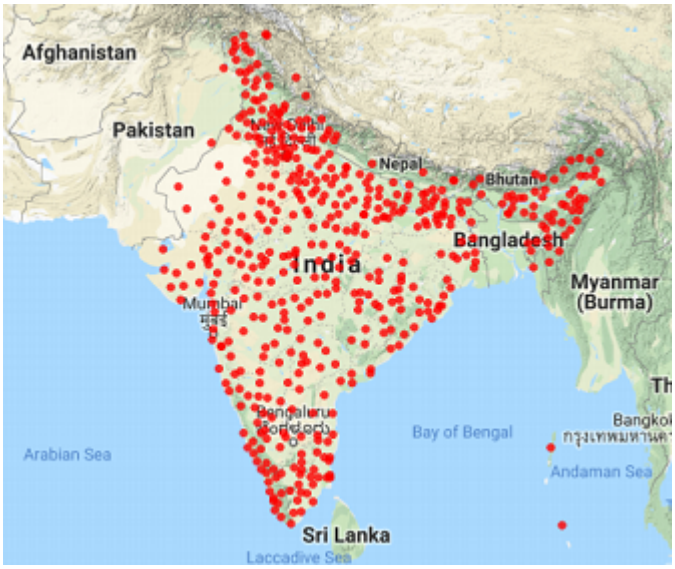

Fig. 9. Map depicting the Indian districts involved in agriculture

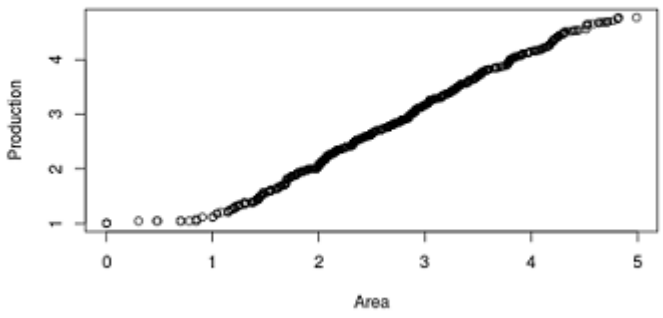

Fig. 10. A Q-Q plot to check if data is normal

accuracy obtained was around 12 percent, which was too low, hence the model was discarded. The non-linear regression model built gave us a test score of 0.39, which meant that it was a better model compared to a linear regression model but still not the best model to represent production as a function of area, season and crop. The decision tree model had a testing efficiency of 40.32 percent, which was mainly due to the limitation of the dataset. The neural network model had a testing efficiency of around 62 percent and a training efficiency of around 41 percent, which was again mainly due to the limitation of the dataset. Hence, we can conclude that a neural network model is better than the decision tree model. The time series and knowledge based models predicted the average temperature and rainfall of Coimbatore district for the month of November 2018 to be around 26.19 degree Celsius and 156.56 cm, which were reasonably accurate. Using this information, the knowledge based recommender system predicted the optimum crop to be Sugarcane.

VI. Conclusions and Future Work

We were able to reasonably achieve the desired results, but due to the constraints on time and the fact that data for agriculture in India was not found readily we could not achieve a good accuracy. The datasets we found were scattered and had only part of the required information. Thus a major part of our project was spent on exploration and cleaning of datasets. The models we have built are working perfectly and can be used on better datasets to achieve the desired results. We have built a neural model that predicts a crop based on a farmers available land area and his expected production. This could help the farmer decide and plan well ahead on what type of crop has to be planted. We further suggest a crop based on the climatic factors in his district. All these crucial decisions can help the farmer save a lot of time and money, enabling him/her to make a reasonable profit on the yield.

We further would have predicted the crop prices in the market by following it's supply chain data, but due to the very scarce data available for this kind of analysis, we had to forego it. The strategy ahead is to use various other factors like soil fertility, water logging characteristics of soil, minerals available and also fertilisers used and cost, to make better crop predictions. This would require us to delve deeper into the agriculture domain, and hence require more time and detailed analysis along with better data.

References

[1] Lakshmi. N, Priya. M, Mrs. Sahana Shetty, Mr. Manjunath C.R (2018), "Crop Recommendation System for Precision Agriculture", International Journal for Research in Applied Science and Engineering Technology (IJRASET)

[2] S.Pudumalar, E.Ramanujam, R.Harine Rajashree, C.Kavya, T.Kiruthika, J.Nisha (2016), "Crop Recommendation System for Precision Agriculture", 2016 IEEE Eighth International Conference on Advanced Computing (ICoAC) pages 33 to 36

[3] John M. Antle, James W. Jones, Cynthia E. Rosenzweig (2016), "Next generation agricultural system data, models and

knowledge products: Introduction", online Journal Report at Elsevier
[4] D. Rodriguez , P. de Voil, D. Hudson, J. N. Brown, P. Hayman, H. Marrou and H. Meinke (2018), "Predicting optimum crop designs using crop models and seasonal climate forecasts", Online Scientific Report
[5] C. T. DE WIT AND F. w. T. PENNING DE VRIES (1985), "Predictive models in agricultural production", Phil. Trans. R. Soc. Land. B 310 pages 309-315
[6] Meonghun Lee, Jeonghwan Hwang, and Hyun Yoe (2013), "Agricultural Production System based on IoT", 2013 IEEE 16th International Conference on Computational Science and Engineering pages 833-837

VII. Role of each member

Guruprasad: Data Collection, Data Preprocessing and implementation of non linear and linear regression models.

Shreyas: Neural network and decision trees and analysis of the dataset to get the top crops of various states by production over the years.

Jai: Time series analysis of rainfall data and recommendation system to predict the suggested crop for a district.