क्षत्रपति(kṣetrapati): Farm Maestro

Mannat Doultani
Computer Engineering
VESIT
Mumbai

Nishit Ahuja
Computer Engineering
VESIT
Mumbai

Chirag Rajavat
Computer Engineering
VESIT
Mumbai

Mitesh Methani
Computer Engineering
VESIT
Mumbai

Saurabh Sahu
Computer Engineering
VESIT
Mumbai

ABSTRACT
This paper proposes a near perfect recommendation, early warning systems and other climate information services which will help farmers and government agencies plan farming activities and programming system for the farmer in term of crop planning and finance management. The system will help them to achieve better sustainability of crops from climatic changes while providing them assistive plan for greater yields.

Keywords
RANSAC, K-Means, Mean-Shift, ARIMA

1. INTRODUCTION
This system provides a one stop solution which integrates various essential services for agriculture and forms a single package which will be able to meet all of the necessary requirements of decision making in agriculture. The system covers all major aspects of agriculture to develop the solution namely climatic conditions, soil nutrients, fertilizers used and economic value of yield.

2. LITERATURE SURVEY
The literature survey provides that no recent work endow the farmer with the appropriate decision of crop cultivation throughout the life cycle of crop. There are many solution which are provided for different aspects essential for agricultural development which address different issues but there is no one stop solution addressing all the problems providing an integrated approach. The paper “Economic impacts of climate change on agriculture” by Peng Zhang, Junjie Zhang and Minpeng Chen states relationship between temperature and precipitation on crops yield of wheat, maize and rice in China [2]. A regression model was fitted to predict the yield of these crops. The paper “Review on Data Mining Techniques for Fertilizer Recommendation” by Jignasha M. Jethva, Nikhil Gondaliya and Vinita Shah compares results of various data mining techniques for fertilizers recommendation [5] .This paper provides an analysis of data mining techniques used in agriculture soil data and also compares the J48, Naïve Bayes, JRip, K-Means Classifier algorithm.

3. DATASETS
Dataset for temperature and precipitation is taken from indiawaterportal.org. The website works primarily through partnerships with non-profit organizations, CSR divisions of multinational corporations and the media to provide resources, working papers, reports, data and discussions on water [12]. The dataset consists of monthly average of temperature and precipitation of Maharashtra state district wise from the year 1901 to 2002. The average of these parameters season wise, which is July to October for Kharif crops, October to March for Rabi crops, March to June for summer crops is then calculated. Crop yield dataset is taken from data.gov.in [15]. The datasets consists of cultivated area and production of a crop yearly district wise throughout India. The dataset for NPK values required by the crop is taken from the “Four Decades of STCR Research - Crop Wise Recommendations” report by the Indian Council of Agricultural Research (ICAR). The dataset attributes are: state, district, crop name, season, available NPK, required NPK and yield. The fertilizer dataset is taken from megapib.nic.in. The website contains information modules on natural resources, Soil crop Suitability, Agro-meteorological Services, Food Processing and Fertilizers. The dataset for fertilizers has attributes fertilizer name, NPK percentage and its price. Crop market prices in various markets of Maharashtra are taken from agmarknet.gov.in [14]. The website provides weekly price trend report for important markets in respect of major agricultural commodities. The dataset consists of crop name, market location, date, and price as attributes. For testing, farmer details such as farmer name, location, field area and soil NPK values is taken from soil health card issued by Ministry of Agriculture and Farmers Welfare, Government of India. The dataset of soil health card can be found at soilhealth.dac.gov.in [9].

4. MODULAR DESCRIPTION

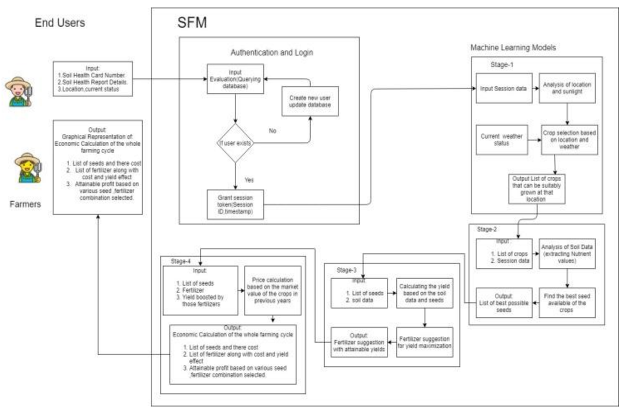

Figure 1: Block Diagram of the system

The system can be divided into four modules. Each module consists of different parameters and machine learning model best fitted on these parameters. The first module covers the effect of climatic conditions on crop yield using the parameters location, temperature and precipitation. The random sample consensus (RANSAC) algorithm is used. The algorithm predicts the yield of every crop and returns a list of most suitable crops based on the yield. The RANSAC algorithm takes random sample called the consensus set from the dataset and the model is fitted. Then it calculates the sum of squared differences (SSD) between the inlier for each point and includes a point as inlier if the SSD is within the threshold else the point is considered as outlier. This step is repeated until enough points are in the consensus set. The model has some change in all iterations. This new model is considered if it contains more number of points in consensus set than the previous one. Else the new step model is not considered and previous model is used for next iteration. In this way, RANSAC removes all the exceptions or so called outliers points from the model. The RANSAC model for each crop is created. Then for each input, the yield of each crop is predicted and the list of crops sorted by their corresponding predicted yield is returned.

This list of suitable crops is then given as input to the next module. The module also takes soil nutrients - N (Nitrogen), P (Phosphorus) and K (Potassium) as input parameters. K-means clustering algorithm is applied to cluster crops based on their NPK values. The nutrients values suitable for various crops are clustered. K-means algorithm initializes k random points as centroids. Then, the Euclidean distance between NPK values of a crop and each centroid is calculated. The crop is then put to a cluster from which its distance is smallest. After clustering each crop once, the centroids are now shifted to a point which is average of all the points which belongs in that cluster. This step is repeated until the difference between new centroid position and its previous position is greater than the given threshold. The sum of distance of all points to its cluster is called the score of K-means model. The score is calculated for values of k from 2 till the number of crops in dataset. The optimal k value is found to be 196 as at this point, the score is found to be smallest. The input NPK values taken from the soil health card is then clustered. This model then returns the target yield and required NPK nutrients to achieve that target.

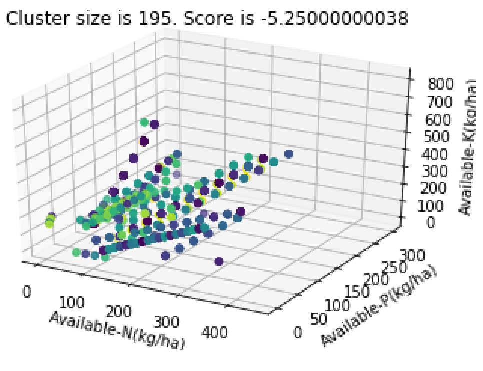

Figure 2: K-means clustering on crop NPK values dataset with k=195.

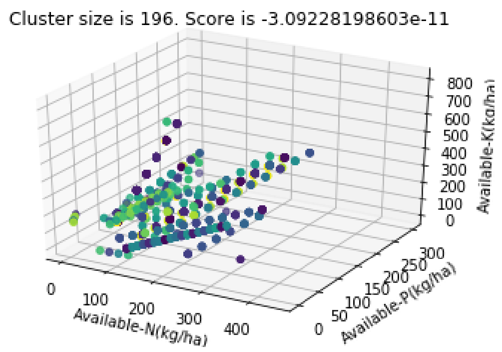

Figure 3: K-means clustering on the crop NPK values dataset with k=196. The score was found to be minimum at this point

The third module is fertilizer suggestion module. It takes required NPK values for the crop and suggests the best fertilizers set along with its cost price. The Mean-Shift clustering algorithm is used. Fertilizers are first clustered based on their percentage NPK values. The bandwidth parameter, which determines the distances between the clusters to be in a cluster, is chosen as 0.1 to get the total clusters as 43.From the required amount of nutrients required for the crop for a particular yield, the percentage amount of each required nutrient in a 100kg bag of fertilizer is found. Then, the nearest cluster where the required NPK values belong is found. If the ratio exactly satisfies the requirement, the system recommends only this fertilizer. Else, if it does not exactly satisfies, remaining amount is calculated. This amount is satisfied with the only-one-nutrient fertilizers. Then, overall price of the fertilizer set needed is calculated.

The fourth module deals with the economic aspect of farming. It analyzes the previous years' prices of crops in all markets to predict the price for this year. Autoregressive Integrated Moving Average (ARIMA) model is applied. The algorithm calculates the prices of all the crops which are predicted suitable for growing from the first two modules. The date is considered as index and price of crop as column. The algorithm takes date as the dependent variable and has seasonality trends in the data. The stationarity of time series is checked. A time series is stationary if its statistical properties such as mean or variance do not change with time. We use rolling statistics and dicky fuller test to check stationarity. Rolling statistics is moving average of last 12 months at any instant t. The dickey fuller test considers the null hypothesis that 'the time series is not at all stationary', then it calculates the test statistics and critical values. If the test statistics is less than critical values then the time series is stationary and reject the null hypothesis. The log of crop prices is taken to perform transformation to remove the trend in the time series. For eliminating the seasonality, the differencing technique is used. In this, the difference between prices of two consecutive dates is taken. It can be observed that time series has now improved in its stationarity. The ARIMA model consists of 3 parameters p, d and q which define the model. The values of p, d and q are chosen and then fit the ARIMA model to predict the prices. This module then computes the total selling price of each suggested crop.

A final report is generated which provides complete assistive plan to the end user. The report, based on results of every module, enlists most suitable crop to be chosen, predicted yield along with required NPK to achieve the target, amount of fertilizers required along with its cost and total selling price on yield.

5. RESULTS

The dataset shows that there is a relation between precipitation and temperature on crop yield. Following graph shows the relationship:

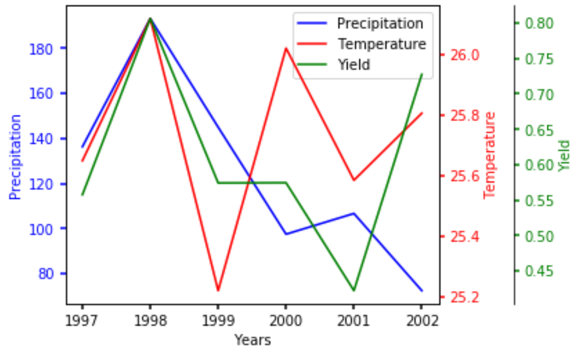

Figure 4: Relation between Temperature, Precipitation and crop yield for the crop 'Bajra' in the district Ahmednagar (Maharashtra)

The RANSAC model was compared with other regression models such as Linear Regression and Polynomial regression with degree 3,4 and 5. The RANSAC algorithm was chosen as it clearly has the lowest mean square error values for the crops when compared with other models. This lower rate of error can be justified by the removing of outlying tuples for fitting the model by ransac, while all other models consider such tuples.

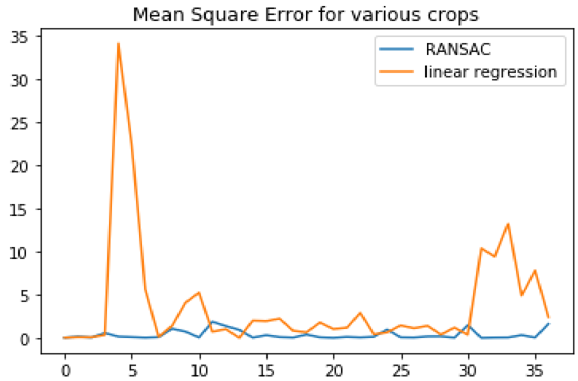

Figure 5: Comparison of Mean Square Error values of RANSAC and Linear Regression.

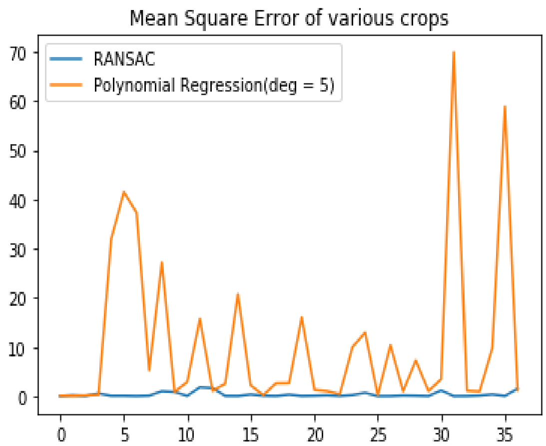

Figure 6: Comparison of Mean Square Error values of Polynomial Regression with degree 5 and RANSAC. Clearly, the RANSAC performs better in case of all the crops.

For the second module, K-means Clustering was compared with Agglomerative Clustering algorithm and MeanShift clustering methods.In K-means, the number of clusters were given which needs to be formed as input whereas in agglomerative clustering it forms clusters on its own in the form of hierarchy and the number of clusters is decided by the line which cuts the hierarchy. In MeanShift too, the number of clusters are decided on its own and bandwidth as given as the input. After giving the required inputs the score of the clusters was calculated and it was seen that the k-means score was the best out of all others.For the fertilizer suggestion module, different approaches like taking ratios, percentages for suggesting fertilizers were compared and then the number of clusters was calculated through MeanShift and the score was found. Then in the third approach the sum of all requirements was taken and then the proportions that would be present in a 100kg bag of fertilizer were found and the clusters were predicted for the given input. It was seen that the fertilizer clusters predicted by this method had the lowest amount of investment required than the above methods. This in turn reduced the amount of investment required by the farmer.

In the crop price prediction module, three different models were compared- Autoregressive (AR) model, Moving Average (MA) model and the Autoregressive Integrated Moving average (ARIMA) model. The RSS (Residual Sum of Squares) value was calculated and the AR and MA model have almost similar RSS value but when both the models are integrated RSS there is a significant difference found. The below graph is the plot of the predicted and original price from the test dataset. The average mean square error between the original and predicted values of crop is 14.36%.

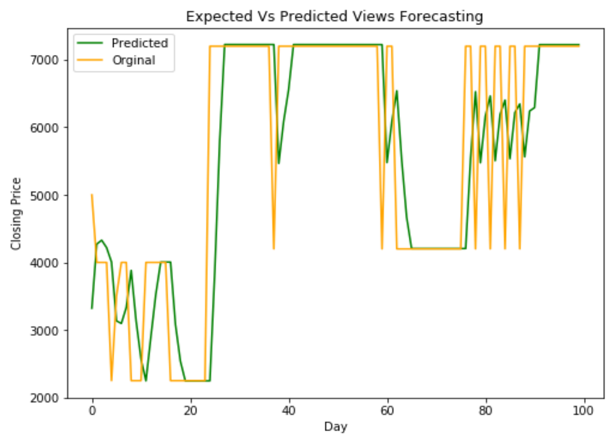

Figure 7: Comparison of Original crop prices with predicted crop prices in the ARIMA module.

6. CONCLUSION

Farming remains an important field in the nation and plays a significant role in driving nation's' economy. The conventional process of selection and decision making taken by farmers are majorly driven by lack of knowledge of clear idea of various parameters and how they affect the crop yield. This leads to drop in crop yield figures. Hence, a system which assists a farmer is necessary. Ksetrapati aims to assist a farmer in taking accurate decisions which lead to higher yield of the crops. The technique of machine learning is being used widely for finding relations between parameters that affect the output parameter and giving suggestions which are highly reliable. This projects could help farmer to reduce cost, improve and maximize crop yield, from various suggestions choose correct fertilizers and market which would provide maximum selling price for the crop.

7. REFERENCES

[1] Haedong Lee and Aekyung Moon, 2014 , “Development ofYield Prediction System Based on Real-time Agricultural meteorological Information”, International Conference on Advanced Communications Technology

[2] Peng Zhang, Junjie Zhang and Minpeng Chen, 2017, “Economic impacts of climate change on agriculture”, Journal of Environmental Economics and Management

[3] Mr. Ambarish G. Mohapatra, Dr. Bright Keswani and Dr. Saroj Kumar Lenka, 2017, “Soil N-P-K prediction using location and crop specific random forest classification technique in precision agriculture”, International Journal of Advanced Research in Computer Science

[4] Jay Gholap, Anurag Ingole, Jayesh Gohil, Shailesh Gargade and Vahida Attar, “Soil Data Analysis Using Classification Techniques and Soil Attribute Prediction”, College of Engineering, Pune, Maharashtra

[5] Jignasha M. Jethva, Nikhil Gondaliya and Vinita Shah , 2018, “Review on Data Mining Techniques for Fertilizer Recommendation”, International Journal of Scientific Research in Computer Science

[6] Navneet and Nasib Singh Gill, 2014, “Algorithm for Producing Compact Decision Trees for Enhancing Classification Accuracy Navneet and Nasib Singh Gill, 2014, “Algorithm for Producing Compact Decision Trees for Enhancing Classification Accuracy in Fertilizer Recommendation of Soil”, International Journal of Scientific Research in Computer Science

[7] Anna X. Wang, Caelin Tran, Nikhil Desai, David Lobell and Stefano Ermon, 2018, “Deep Transfer Learning for Crop Yield Prediction with Remote Sensing Data”, ACM SIGCAS Conference on Computing and Sustainable Societies (COMPASS 2018)

[8] Nikita Gandhi, Leisa j. Armstrong, Owaiz Petkar and Amiya Kumar Tripathy, 2016, “Rice Crop Yield Prediction in India using Support Vector Machines”, Joint Conference on Computer Science and Software Engineering (JCSSE)

[9] Soil Health Card: https://soilhealth.dac.gov.in/

[10] Department of Agriculture, Government of Maharashtra: http://krishi.maharashtra.gov.in/1250/Taluka-Level

[11] National agriculture market, Government of India: http://enam.gov.in/NAM/home/dashboard4.html#

[12] Indiawaterportal:http://www.indiawaterportal.org/met_data/

[13] Fertilizer NPK values details and dataset: http://agritech.tnau.ac.in/agriculture/agri_nutrientmgt_nutrientcontent.html

[14] Cropprices:http://agmarknet.gov.in/PriceAndArrivals/DatewiseCommodityReport.aspx.

[15] Maharashtra taluka wise yield dataset: http://krishi.maharashtra.gov.in/1250/Taluka-Level