Decision Making Using C4.5,KNN,Regression Algorithm On Agricultural Factors

[1]Sheetal Jagtap, [2]Sumedha Kolhe, [3]Bhagyashri Shinde,[4]Pravin Tak

[1]Student, [2]Student, [3]Student,[4]Assistant professor
Computer department,
Zeal college of engineering, pune

Abstract : Agriculture is the best source of earning and living a life in India from histories. Agriculture is the single most important contributor to the Indian economy. India is a country of different kind of season in every season there are different crops grown up which also depends on economic and biological conditions .The future of crop yield is challenging task for every country. Farmers usually plant the cultivation process based on their previous experiences. Due to lack of knowledge of cultivation farmers are unable to take beneficial crop. To help the farmers take decisions that can make their farming more efficient and profitable, we establish an intelligent information prediction analysis on farming. In the system, by using various techniques and terminologies we try to overcome the farmer's problems and help them to analyze the crop selection in different area and weather. Main purpose through this system is to help farmers and decrease the percentage of farmer's suicide deaths. System assisting farmers to make better decisions by providing them real time data processing, and a dynamic data service composition method, to enhance and monitor the agricultural productivity. Thus to help farmers in decision making and in better management of natural resources. The system suggests area based beneficial crop rank before the cultivation process. It indicates the crops that are cost effective for cultivation for a particular area of land. To achieve these results, system considers major crops. Supervised learning algorithm helps to predict the analyzed data.

IndexTerms **- Data Mining, Predictive Analytics, Stepwise Linear Regression Model, C4.5 Decision Tree, K-Nearest Neighbor, Machine Learning Algorithm, Regression Techniques, and Classification Techniques, Agriculture.**

I. INTRODUCTION

Agriculture, the main backbone of India, is the development of plants for food, bio-fuel, medicinal plants and other products used for sustaining and enhancing individual life. The history of agriculture engages thousands of years back, and its growth has been motivated and defined by several atmospheres, cultures and technologies. It is a crucial part of our economy and has always been one of the vital occupations that serve mankind, both in terms of livelihood and employment. Due to the substantial increase in the population, the nutritional status of the poor is growing bad, which must be improved. The major effect of population increase has been prominently shown on the environment, the damage of which is increasing rapidly, which ultimately hinders agricultural production. The vision of meeting world's food demands for the increasing population throughout the world is becoming more important in these recent years.

India is an Agricultural nation. Almost majority of the Indians traditional occupation is Agriculture. Many of them worship their land as the god. It is one of the prominent factors that decide the economy of our country. Table 1 clearly presents the contribution of agriculture to the national income and its share in export for a period of 50 years. An examination of Table 1 makes clear that the share of agriculture in the national income and in the total export is declining consistently. Now agriculture contributes only about one-third to the national income as against 54% in 1950-51. Similarly, the share of agricultural goods in export has declined from 52.5% in 1950-51 to only 16.5% in 1990-91. To increase the agricultural contribution to the national income; the production of crops should be increased.

Due to the lack of use of technology and scientific methods to farming, farmers in most cases do not get the preferred output. In this era of technology, applying scientific methods and automated learning in problem solving has come a long way from being a trend to become a necessity. The process of extracting important and useful information from large sets of data is called Data Mining. Data mining in agriculture applications involves the conceptualization, design, development, estimation and application of modern ways for utilizing the information and communication technologies (ICT) in rural domain including with the major objective on agriculture productivity. Different modelling processes and simulation methods have been implemented for dynamic systems in agriculture. The mined information is typically represented as a model of semantic structure of database, wherein the model may be used on new data for prediction or classification of agricultural data. The major challenge in agriculture is that no specific measures have been taken out with the large sets of agricultural data. The foremost issues in agriculture and modern techniques are associated to the overall crop production. Analyzing these data, the algorithms give the result which predicts the preferred profitable output. Nowadays, plant diseases and land degradation are the most important problems in agriculture. Therefore, the improvement of agricultural is monitored by data mining techniques.

Table 1. India: Position of agriculture in national income and total export (1950-91)

Year	Contribution of agriculture to national income	Share of agriculture to total exports of India
1950-51	54	52.5
1950-51	49	44.0

1950-51	47	37.5
1950-51	36	25.5
1950-51	31	16.5

Source: Food and Agricultural Organization (2000: Press Note)

The objective of the research is to provide a learning agent that can aid in taking decisions to make the farming more efficient and profitable through technology. The research provides a list of profitable crops in a particular area using decision making algorithms. The research focuses on the twelve major crops Rice, Coconut, Sugarcane, Mango, Grapes, Cashew nut, Ragi, Wheat, Cotton, Bajra, Barley, and Jowar for Maharashtra region, stored in database system. The dataset contains details on crops yield per hectare (M.Ton), average of minimum and maximum temperature, rainfall, year range, and region, pest/disease build on crop, crop price data of region. The goal of this research is to help the farmers maximize their profit margin by providing predictions on crops that will give the maximum output in a particular area. Crop selector could be applicable for minimize losses when unfavourable conditions may occur and this selector could be used to maximize crop yield rate when potential exists for favourable growing conditions. Many research intended to agriculture planning is carried out, where the goal is to get an efficient and accurate model for crop yield prediction, crop classification, soil classification, weather prediction, crop disease prediction, classification of crops based on growing phase. A statistical and machine learning both techniques were modelled. Agricultural management specialists need simple and accurate estimation techniques to predict crop yields in the planning process. Over the last few decades, statistical methods have traditionally been used for predictions and classifications. Some of the common traditional statistical techniques used for predictions and classifications are multiple regression, discriminate analysis, logistic regression etc.

Achieving high crop yields is the principle aim of agricultural production. Early detection and management of problems associated with crop yield indicators can help increase yield and subsequent profit. Predictions could be used by crop managers to minimize losses when unfavourable conditions may occur. Additionally, these predictions could be used to maximize crop prediction when potential exists for favourable growing conditions. There are a number of crop yield prediction models which use either statistical or crop simulation models. Over the last decade it has been observed that Artificial Intelligence (AI) techniques provide a more effective approach to predicting crop yield under different cropping scenarios.

Yield prediction is one of the most critical issues faced in the agricultural sector. Farmer's lack of knowledge about harvest glut, uncertainties in the weather conditions and seasonal rainfall policies, depletion of nutrition level of soils, fertilizer availability and cost, pest control, post–harvest loss and other factors leads to decrease in the production of the crops. Crop production rate depends on geography of a region (e.g. hill area, river ground, depth region), weather condition (e.g. temperature, cloud, rainfall, humidity), soil type (e.g. sandy, silty, clay, peaty, saline soil), soil composition (e.g. PH value, nitrogen, phosphate, potassium, organic carbon, calcium, magnesium, sulphur, manganese, copper, iron) and harvesting methods. Different subsets of these influencing parameters are used for different crops by different prediction models.

Price prediction is a very important problem for any farmer as he is the one who should know how much cost he would expect for his crops. In past years, price prediction was done by judging farmer's experience on particular crop and field. Maximizing production rate of crop is an interesting research field to agro-meteorologists which play a significant role in national economic. Agricultural production of next year is affected from crop degree of preceding year, price, consumption patterns, and imported agricultural products. It takes into consideration with the impact of a combination of many different factors.

In agriculture, crop yield is measured as a degree of the yield per unit area of the cultivated land and the seeds produced from the same crop. The reasons for low productivity of Agriculture productivity are Human factor i.e. lack of training and efficiency of farmers, huge population, Traditional methods of cultivation, Problems of soils, Pests, Diseases, feeble cattle, Lack of credit facility, Inadequate Irrigation facilities, Unreliable Monsoon and Improper marketing of crops etc.

Suitable application of curative measures may decrease the yield loss. For the application of these areas, one must have prior knowledge of the time and the harshness of the outbreak of pests and diseases.

Now a day's advanced technology are helping a agriculture in crop yield. As the development of technology in different field increases day by day, so there arise necessity to play a conclusive roll in agriculture by crop and predictive tool.

Crop yield is a unified bio-socio-system comprised of complex interaction among the soil, the air, the water, and the crops grown in it, where a comprehensive model is required which are possible only through classical engineering expertise. To predict the crop future consider different sources meteorological data, agro-meteorological (phonology, yield), soil (water holding capacity), remotely sensed, agricultural statistics.

II. LITERATURE SURVEY

Aditya Shastry et.al. [Ref.2] in their paper, they carried out an experiment on wheat, cotton and maize data sets using Quadratic, Linear, Polynomial, GLM and SLM. By utilizing the best regression model for the survey, the forecast of generation of wheat, maize and cotton is done for chosen years. The results are compared obtained from them. The accuracy is measured by using R^2, RMSE and MPPE. It was concluded that pure quadratic model accurately predicts the wheat yield, Stepwise Linear Regression model accurately predicts the cotton yield and Generalized Linear Regression is used for Maize yield prediction.

Fahad Sheikh [Ref.3] in his paper, various data mining techniques are reviewed and performance comparison between various classification algorithms. Algorithms compared are C4.5, CART, k-means clustering, ANN and MLR. It was found that the performance of C4.5 (J48) with accuracy 88.2 % is better than the Naive Bayes with accuracy 54.8%.

Dewi Sinta et al [Ref.1] in their paper, by using KNN model price prediction of rice crop is done from January to December 2012 in Indonesia. The model performance is based on the value of MAPE, MAE and RMSE. The best model is used to predict has the value of MAPE, MAE and RMSE smaller. They found that Ensemble KNN has better yield prediction than single KNN method.

R. Sujatha et al [Ref.10] in their paper stated that there are various classification methods such as Naïve Bayes, J48, Random Forest, Artificial Neural Network, Decision Tree and Support Vector Machine (SVM) for solving the problem of yield prediction. The paper describes how improving agriculture efficient by prophesying and improves yields by previous agricultural information.

Giritharan Ravichandran et al [Ref.14] in their paper, they take the input as various parameters that decides the productivity, process them based on the algorithm provided by ANN, and predict suitable crop for the land. Also, the paper suggests the some fertilizers. They found that with increase in the number of hidden layers, the performance increases along with the complexity so, hidden layer is chosen by trial and error method.

S. Ruggieri et al [Ref.11] in their paper analytical evaluation of the runtime behavior of the C4.5 algorithm is presented which gives rise to more efficient implementation, with a five times performance gain. The paper stated that the EC4.5 is among the best strategies over C4.5 according to the analytic comparison of the efficiency.

III. PROPOSED WORK

The main goal of the system is to give farmers the overview of crops that they can take in their farms with the estimated crop yield and the market price prediction of that particular crop.

System architecture includes an input module, which is responsible for taking input from the farmers. In that the farmer has to provide Location, Soil type/ Land type, Size of land, Water Source. The farmer is also responsible for interacting with predicted results. After selection the location parameter, the input query block is responsible for subset selection of an attribute from crop details. It gives all possible crops that are to be sown at a given time stamp suitable for given location. The prediction model is used to predict weather, pest/ disease build on crop and market price. Yield rates of these crops are evaluated, if yield rate per day of these crops are fair (within tolerance) then those crops are selected for crop sequences. The main feature of the system is it gives output in statistics form. To this end, the methodological approach that it will follow is composed of these steps:

1) The description of user's farm.

2) The selection of crops suitable.

3) Predictive analytics modelling for predicting total yield and market price of those crops.

Strategic Module consists of 2 blocks:

1. Query Block
2. Predictive Analytics block

Describing the data i.e. summarizing its statistical attributes (such as means and standard deviations), visually reviewing it using charts and graphs, and looking for potentially meaningful links among variables (such as values that often occur together) is the simple first analytical step in data mining.

Data mining takes advantage of advances in the fields of artificial intelligence (AI) and statistics which is being used for pattern recognition and classification of problems [9].

Prediction types using data mining techniques are:

(1) *Classification:* predicting into what category or class a case falls.

(2) *Regression*: predicting what number value a variable will have (if it is a variable that varies with time, it's called 'time series' prediction).

Classification:

Aim of Classification problems is to identify the characteristics that group each case. Understand the existing data and to predict how new instances will behave is the task of classification problem. Data mining builds classification models by examining already classified data (cases) and inductively finding a predictive pattern. These existing cases come from a historical. They may also come from an experiment in which a sample of the entire database is tested in the real world and the results used to create a classifier. Sometimes an expert classifies a sample of the database, and this classification is then used to create the model which will be applied to the entire database [9].

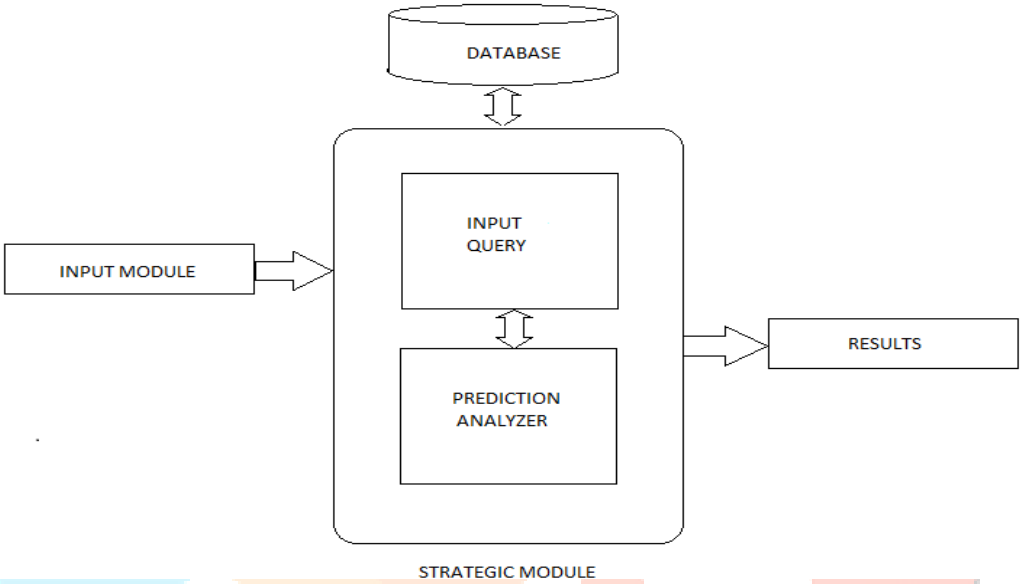

Fig 1: System Architecture

Regression

Using existing values to forecast what other values will be is Regressions. For simple case, regression uses standard statistical techniques such as linear regression. But, many real-world problems are not simply linear projections of previous values. For instance, sales volumes, stock prices, and product failure rates are all very difficult to predict because they may depend on complex interactions of multiple predictor variables. Therefore, more complex techniques may be necessary to forecast future values. The same model types can often be used for both regression and classification. There are varieties of data mining methods including Support Vector Machines (SVM), Artificial Neural Networks (ANN), Naïve Bayesian Classifier, Genetic Algorithm, and K-Nearest Neighbor (KNN).

The system uses predictive analytics to extract knowledge from the existing data used for future planning and also gives idea about trends and outcomes. Predictive analytics is the practice of extracting information from existing data sets in order to determine patterns and predict future outcomes and trends. Here Predictive analytics is used for predicting three terms which will greatly benefit the farmer for deciding the crop he should take [5].

- Weather
- Crop Yield
- Crop Price

Predictive model building uses three algorithms: C4.5 Classification algorithm for Weather Prediction, Generalized Linear Regression Model for Crop Yield Prediction, KNN Regression algorithm for Crop Price Prediction.

IV. ALGORITHMS

C4.5

C4.5 is an algorithm used to generate a decision tree.C4.5 is an extension to ID3 algorithm. The decision tree generated by C4.5 is used for classification and is referred as statistical classifier. C4.5 uses a information gain while generating the decision tree. It also uses a single pass pruning process for reducing the over-fitting. It works on both types of data like continuous and discrete data. C4.5 is a supervised learning algorithm. The C4.5 working is, it developed a classifier in the form of decision tree. To do this, C4.5 is given a set of data representing things that are already classified. The classifier which constructed by C4.5 is a tool in data mining that takes a bunch of data representing things we want to classify and attempts to predict which class the new data

belongs. It also support for binary and n-array outcomes. The additional feature of C4.5 is it also support for tree pruning and missing value handling [11]. Figure 2 shows the working of C4.5 Algorithm.

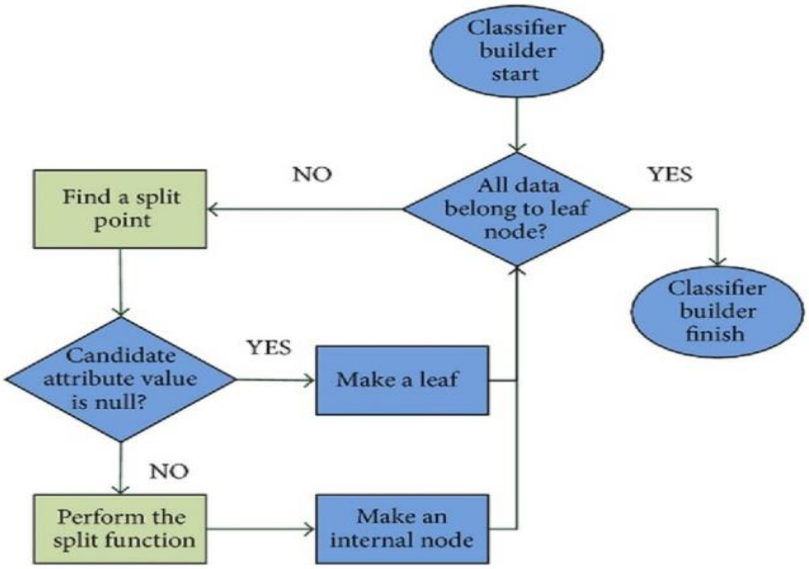

Fig 2: C4.5 Algorithm Flowchart

The objective is to predict if the crop is suitable or not in that weather by using C4.5. C4.5 generated decision tree is used as Classification method. Suppose a dataset contains a list of crops information which includes name, location, crop yield, crop price, etc. These are all called attributes. Now from these entire attribute we want to predict whether the crop is suitable or not in that environment. Using these crop attributes and crop information class, C4.5 developed a decision tree which will predict the class for new crop based on their attributes [11].

C4.5 Tree-Construction Algorithm

The decision tree is constructed by using training data set S, which is set of tuples in database terminology. Each attribute have either discrete or continuous values. A decision tree creates something similar to flowchart to classify new data. Decision tree contain decision node and leaves. The leaf specifies a class value and decision node specifies a test over one of the attribute. For each possible outcome, a child node is present. The path from root to leaf of decision tree is followed by attribute values of case. The class which is present at leaf node is predicted by the decision tree [11].

Tree-Construction Algorithm

The C4.5 constructs the tree using divide and conquer method. Construction of the decision tree is done by selecting the best possible attribute which will be able to split set of samples in effective manner. The attribute which contain highest entropy difference or gain is selected as the splitting criteria for that node. Each penultimate node contains the multiple or last attribute for making the final decision. In C4.5, each node contains a set of cases. Cases have weight to take into account unknown attribute values. At start only the root node is available and associated with whole training set S and with all case weights equal to 1.0 [11].

ALGORITHM C4.5(S)

Input: Dataset S with attribute value

Output: Decision tree or/and set of rules that assigns a class to a new case

1) Tree= {}

2) ComputeClassFrequency(S)

3) If oneClass

return leaf;

create decision Tree D

4) ForEach Attribute A;

ComputeGain (A);

5) D.test=Best attribute according to above Criteria(Step 4);

6) if D.test continue

find threshold;

7) ForEach S' in the splitting of S

8) if S' is Empty

child of D is a leaf

else

9) Child of D=formTree(S')

Attach T Tree to the corresponding branch of Tree;

10) Compute Errors of D;

Return decision Tree D;

Let S be the dataset given to the system. At start the Tree is empty having no node. The weighted frequency of dataset S is computed (step (2)) freq (b_i, S) in S which contain class b_i for i € [1, Dclass].If all cases belong to same class (step (3)), then the node becomes leaf with given class (respectively with most frequent class). If S contain cases belonging to two or more classes (step (4)), then information gain is calculated. For discrete attributes, the information Gain helps to split the cases in S into sets with distinct an attribute values. For continues attributes, the information gain split the set S into subsets, namely, cases with an attribute value which is not greater than a local threshold which is determined during gain [11].

Information Gain

The information gain of an attribute A for a set of cases S is calculated as follows:

$$Gain = Entropy(S) - \sum_{i=1}^{c} \frac{|S_i|}{|S|} \times \text{Entropy}(S_i)$$

Where,

$$Entropy(S) = -\sum_{j=1}^{NClass} \frac{freq(b_j,S)}{|S|} \times \log \frac{freq(b_j,S)}{|S|}$$

The attribute which contain high gain is (step (5)) is selected for next process. The test select the continuous attribute then threshold is calculated (step (6)) as the calculated threshold is below the local threshold. If S_i empty then the child c node is set as leaf, with associated class the most frequently used class. If S_i is not empty, then divide and conquer approach apply recursively on same operation n set and also on unknown parameters. The unknown parameter of selected attribute is replicated in each child with their weights proportional to cases in S with their known value of selected attribute. Finally classification error is calculated as sum of error of child node [11].

REGRESSION

To solve the selection problem there are many regression algorithms are available. The stepwise regression predicts the variables choice for regression models and this process done automatically. In each step variable is added or subtracted from set. It also works on forward selection, backward elimination, and bidirectional elimination. In forward selection, start from no variables in model; test the addition of each model according to the given criteria. This process continues till the adding variable gives statically improved fitting. Backward elimination starts from all candidate variables. It deletes the each variable according to the fit criteria. Delete the variable whose loss gives the most statically significant loss fit[13].The goal of stepwise regression is to maximize the prediction by using minimum number of variables[12].The objective of paper is to predict the suitable and favourable crop for farmer by using the stepwise linear regression. The Fig. shows the actual working of SLM.

Regression Algorithm

Input: Independent variable for crop dataset

Output: Predict crop yield for the dataset

1. Initially all observations are in the root node;
2. Start from the root node ,partition the samples using the recursive procedure;
3. Giving a tree node;
4. Perform stepwise linear regression on samples in current node;
5. Calculate the residual sum square(RSSnode);
6. If (number of samples>predefined minimum node size);
7. For each predictor variable x;
8. Sort x in ascending order ;
9. for each value d in the above sorted list;
10. using d as the threshold value ,partition the samples within current node into two subsets;
11. perform stepwise regression on each subset and calculate the corresponding RSS;
12. calculate the subtotal RSS as the sum of those of the two subsets;
13. end of for loop on threshold value d;

14. find the minimum subtotal RSS achievable from splitting current node on x;
15. end of for loop on predictor variable x;
16. find the minimum RSS of all possible splits of current node(MinRSSsplitting);
17. calculate the improvement from splitting current node using equation (4);
18. if(improvement>predefined minimum improvement);
19. split current node into two new nodes using the variable and threshold value that give the MinRSSsplitting ;
20. for each of the two new nodes ,go to step 3;
21. end of if in step 18;
22. end of if in step 6;
23. end of recursion;

Initially all observations are in a set, the root node of the regression tree. The programme starts from the root node and recursively partitions the samples into subsets until no leaf node can be further split. A node is split into two subsets when a split leads to substantial reduction in the residual sum square (RSS) of estimation. For a given node RSS is calculated using equation (3), and is referred to as RSSnode. Similarly the RSSs of the two subsets after a split, referred to as RSSleft and RSSright, are also calculated using this equation. The total RSS after a split, referred to as RSSsplitting, is the sum of RSSleft and RSSright. The minimum RSSsplitting of all possible splits of a node is referred to as MinRSSsplitting. A node can not be further split when the number of samples in that node is less than a predefined minimum node size or when the improvement from splitting that node is less than a predefined threshold value [13]. The improvement from splitting a node is calculated as:

$$Improvement = RSSnode - MinRSSsplitting\ RSSnode \times 100\%$$

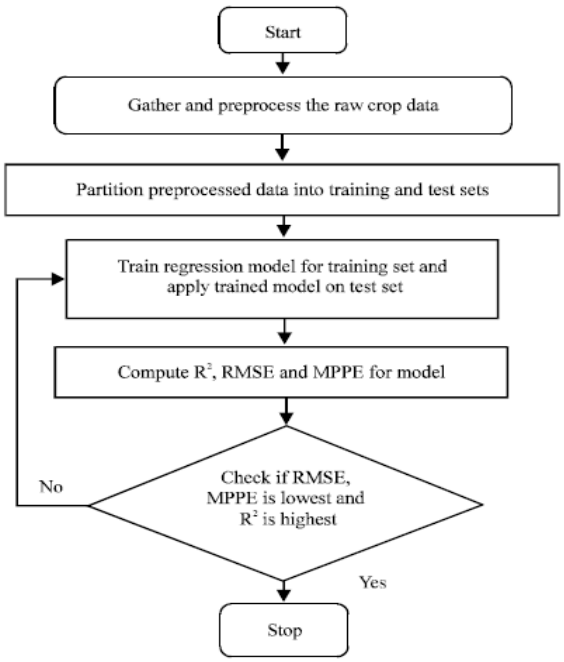

Fig 3: Regression Methodology for crop yield prediction

KNN

K-nearest neighbor is the fundamental and simple classification technique which is suitable for little or no prior knowledge about how data is distributed.

This technique is a machine learning algorithm that has easy implementation. *k*NN is lazy learning which gives the closest *k* records of the training data set that have the highest similarity to the test (i.e. query record) unlike other learning that builds a model or function. An example for the k-nearest neighbor classification is given in figure 4 [8]

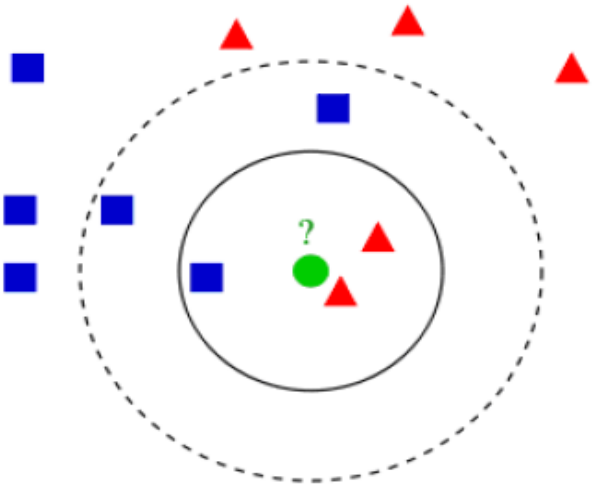

Fig 4: Example of K-nearest Neighbor Classification

Regression problems deal with prediction of the outcome of a dependent variable while set of independent variables are known. Figure 5, shows the relationship between the independent variable *x* and the dependent variable *y* (red curve) by a set of points (green squares). Given the set of

green objects (known as examples) the KNN method is used to predict the outcome of *X* (also known as query point) given the example set (green squares) [9].

Consider an example of the 1-nearest neighbor method. In such case the example set (green squares) is searched and located the one which is closest to the query point *X*. For such

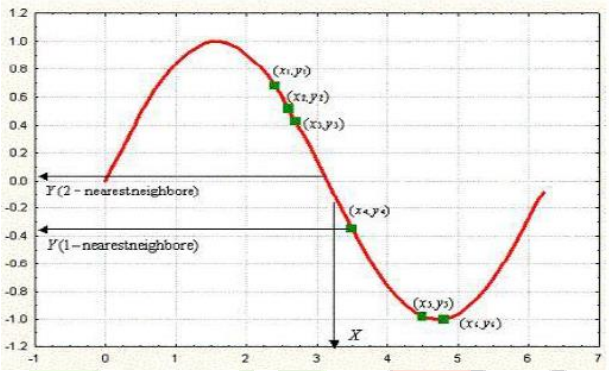

Fig 5: KNN Decision Rule for Regression

particular case, it happens to be *x*4. The outcome of *x*4 (i.e., *y*4) is thus then taken to be the answer for the outcome of *X* (i.e., *Y*). Hence for *1*-nearest neighbor we can write:

$$Y = y_4$$

In next step, the *2*-nearest neighbor method is considered. Here, the first two closest points to *X* are located, which happen to be *y*3 and *y*4. Taking the average of their outcome, the solution for *Y* is then given by:

$$Y = \frac{y_3 + y_4}{2}$$

The average of the outcomes of its *K* nearest neighbors is the outcome *Y* of the query point *X*.

Similarly, assigning the property value for the object to be the average of the values of its K nearest neighbors can be used for regression. It can be useful to determine the contributions of the neighbors, so that the nearer neighbors contribute more to the average than the more distant ones.

K neighbors closest to that point is responsible for the predictions made by KNN. Therefore, to make predictions with KNN, we need to define a metric for measuring the distance between the query point and cases from the examples sample. Euclidean is mostly used measure this distance. Also, Euclidean squared, City-block, and Chebychev can be used [9].

$$D(x,p) = \sqrt{((x-p)\text{^}2)} \quad \text{.. Euclidean}$$

$$D(x,p) = (x-p)\text{^}2 \quad \text{.. Euclidean squared}$$

$$D(x,p) = |x-p| \quad \text{.. City-block}$$

$$D(x,p) = \text{Max}(|x-p|) \quad \text{.. Chebychev}$$

where *x* and *p* are the query point and a case from the examples sample, respectively.

Predictions based on the KNN examples are made after selecting the value of K . For regression, KNN prediction is the average of the K nearest neighbors outcome:

$$y = \frac{1}{K}\sum_{i=1}^{k} y_i$$

where *yi* is the *i*th case of the examples sample and *y* is the prediction (outcome) of the query point [9]. Unlike regression, in classification problems, for determining the class label a majority vote is performed among the selected k records which are later assigned to the query record on which KNN prediction is based [7].

KNN has several many advantages such as simplicity, effectiveness, intuitiveness and competitive classification performance in many domains. It is Robust to noisy training data and is effective if the training data is large [9].

Despite the advantages, KNN has a few limitations such as poor run-time performance when the training set is large. It is very sensitive to irrelevant or redundant features because all features contribute to the similarity and thus to the classification. By careful feature selection or feature weighting, this can be avoided. Two other disadvantages of the method are [9]:

- Distance based learning is not clear which type of distance to use and which attribute to use to produce the best results.
- Computation cost is quite high because we need to compute distance of each query instance to all training samples.

V. CONCLUSION

Defining decisional result data set from existing data set is crucial task especially if that correlates with agriculture. Since data set is somewhat complex and ambiguous in nature. As by considering the motivation behind this work the algorithmic strategies such as C4.5, KNN, SLM are used to define effective result data set so that the farmer will help them to take real time decisions awhile in agriculture.

VI. ACKNOWLEDGMENTS

The authors would like to thank their family, supportive knowledge given reference papers and management of Zeal College of Engineering for their kind support.

REFERENCES

[1] Dewi Sinta , Ensemble K-Nearest Neighbors Method to Predict Rice Price in Indonesia, *Applied Mathematical Sciences, Vol. 8, 2014, no. 160, 7993 – 8005*

[2] Aditya Shastry, H.A. Sanjay and E. Bhanusree, 2017. Prediction of Crop Yield Using Regression Techniques. *International Journal of Soft Computing, 12: 96-102.*

[3] Fahad Sheikh, S. Karthick, D. Malathi, J. S. Sudarsan and C. Arun, , 2016. Analysis of Data Mining Techniques for Weather Prediction. *Indian Journal of Science and Technology, Vol 9(38)*

[4] Kaur, Manpreet & Gulati, Heena & Kundra, Harish. (2014). Data Mining in Agriculture on Crop Price Prediction: Techniques and Applications. *International Journal of Computer Applications. 99. 1-3. 10.5120/17422-8273.*

[5] S. Lamrhari, H. Elghazi, T. Sadiki and A. El Faker, "A profile-based Big data architecture for agricultural context," *2016 International Conference on Electrical and Information Technologies (ICEIT), Tangiers, 2016, pp. 22-27.*

[6] M. R. Bendre, R. C. Thool and V. R. Thool, "Big data in precision agriculture: Weather forecasting for future farming," *2015 1st International Conference on Next Generation Computing Technologies (NGCT), Dehradun, 2015, pp. 744-750.*

[7] Alkhatib K, Najadat H, Hmeidi I, Shatnawi MKA. 2013. Stock Price Prediction Using K-Nearest Neighbor (kNN) Algorithm. *International Journal of Business, Humanities and Technology.* 3 (3): 32 - 44.

[8] Chitra A, Uma S. 2010. An Ensemble Model of Multiple Classifiers for Time Series Prediction. *International Journal of Computer Theory and Engineering*. 2 (3): 454 - 458. http://dx.doi.org/10.7763/ijcte.2010.v2.184

[9] Imandoust, S.B. & Bolandraftar, Mohammad. (2013). Application of K-nearest neighbor (KNN) approach for predicting economic events theoretical background. Int J Eng Res Appl. 3. 605-610.

[10] R. Sujatha and P. Isakki, "A study on crop yield forecasting using classification techniques," *2016 International Conference on Computing Technologies and Intelligent Data Engineering (ICCTIDE'16)*, Kovilpatti, 2016, pp. 1-4.

[11] S. Ruggieri, "Efficient C4.5 [classification algorithm]," in *IEEE Transactions on Knowledge and Data Engineering*, vol. 14, no. 2, pp. 438-444, Mar/Apr 2002.

[12] S. Nagini, T. V. R. Kanth and B. V. Kiranmayee, "Agriculture yield prediction using predictive analytic techniques," *2016 2nd International Conference on Contemporary Computing and Informatics (IC3I)*, Noida, 2016, pp. 783-788.

[13] C. Huang & J. R. G. Townshend (2003) A stepwise regression tree for nonlinear approximation: Applications to estimating subpixel land cover, *International Journal of Remote Sensing, 24:1, 75-90.*

G. Ravichandran and R. S. Koteeshwari, "Agricultural crop predictor and advisor using ANN for smartphones," *2016 International Conference on Emerging Trends in Engineering, Technology and Science (ICETETS)*, Pudukkottai, 2016, pp. 1-6.