Proceedings of the 2nd International Conference on Inventive Communication and Computational Technologies (ICICCT 2018)
IEEE Xplore Compliant - Part Number: CFP18BAC-ART; ISBN:978-1-5386-1974-2

Comparative Analysis of Data Mining Models for Crop Yield by Using Rainfall and Soil Attributes

Kunal Teeda[1] Nandini Vallabhaneni[2] Dr.T.Sridevi[3]

***Abstract*— Till the seventies of the last century, Indian agriculture was in a poor condition. The agrarian economy was largely consumption-oriented and there were poor irrigation facilities and simple agricultural implements. Agricultural yield was very low and dependency on nature was very high. The food grains were not enough to feed the population. With a view to augment the yield, the Indian government had no option but to introduce Green Revolution. The Green Revolution was a movement towards excessive mechanisation of agriculture. The agriculturists were motivated and assisted to undertake the technology-based farming. Irrigation facilities were developed. However, the results of Green Revolution were not uniform all over the country. Neither has there has been uniform impact on all kinds of crops nor has there been uniform impact on all the regions and all categories of farmers. Even today the farmer falls prey to the risks unleashed by the nature. Success or failure of rain fed vegetation depends upon the sample and amounts of rainfall. But, other factors like temperature, photoperiod and grid additionally notably influence crop boom and yield. The analysis of climate performs a key role in planning better farming structures to enhance and stabilise yields, and to design appropriate crop breeding strategies. With the use of technology, it has also become possible to minimise the risks involved in agriculture to which the early farmers were awfully exposed. There are in particular two procedures to predict rainfall. Empirical technique and dynamical method. In our method we use the empirical technique that is based on evaluation of historical information of the rainfall and its dating to a spread of atmospheric variables over different components of the nation. The most broadly used empirical approaches used for weather prediction are regression, artificial neural network, fuzzy logic and institution approach of statistics dealing with. We use data mining techniques such as clustering and classification techniques for rainfall prediction.**

I. INTRODUCTION

Agriculture in India has a full-size history. nowadays, India is ranked 2nd worldwide in farm output. Agriculture and allied sectors like forestry and fisheries accounted for 16.6 percent of the GDP 2009, about 50 percent of the overall workforce. The monetary contribution of agriculture to India's GDP is regularly declining with the united states' large-primarily based economic boom. Agriculture is a form of an enterprise with a chance.The production of plants relies on different factors like on climatic, geographical, organic, political and financial elements. Accurate statistics about the character of an ancient yield of the crop is important modelling input, which is useful to farmers and authorities organisation for decision-making technique in establishing right policies associated with subsequent manufacturing. The advances in computing and information storage have provided largely at the maximum of information. The project has been to extract expertise from this uncooked statistics, statistics mining that may bridge the understanding of the facts to the crop yield estimation. This task aimed to statistics mining strategies and follow them to the various variables consisting inside the database to set up if significant relationships may be discovered and the usage of fuzzy common sense to discover the circumstance of crops on a diverse situation of rainfalls. Bangladeshi student proposes Data mining techniques to predict annual yield of major crops and recommend planting different crops in different districts in Bangladesh[5]. They considered the effects of biotic(pH, soil salinity), environmental(weather), and area of production as factors towards crop production in Bangladesh. Taking these factors into consideration as datasets for various districts, they applied clustering techniques to divide regions; and then they apply suitable classification techniques to obtain crop yield predictions. In the research paper by David H White and S Mark Howden[9], they focus on the climates determinants of crop productivity.They considered how the climate envelopes different crops based on temperature, moisture and light influence the distribution of cropping and other land uses around the world. They also discuss how these and other climatic variables influence the growth and yield of crops. Adaption strategies are also discussed that helps a lot to assist the crop producers to cope with the rising global temperatures and carbon dioxide (CO2) levels, along with the often reduced rainfall, soil moisture and water availability.

II. RELATED WORKS

[1] In this paper, the author Dr. D. Ashok Kumar states that the purpose of the examination at it is to observe the best techniques to extract new understanding and information from present soil profile data contained within ISRIC-WISE soil statistics set. Numerous records mining techniques like Support Vector Machines, K nearest Neighbours, Bayesian Networks. Also various optimisation strategies like Ant colony optimisation, Particle Swarm Optimisation. [2] This review article written by Dr. Bharath Misra focuses mostly on various attributes to be taken in consideration while applying data mining techniques in the field of agriculture and also concludes that The multidisciplinary approach of integrating computer science with agriculture will help in forecasting/ managing agricultural crops effectively. [3] Sally Jo Cunningham emphasises on the usage of data mining techniques and its process model to derive innovative applications in the field of agriculture. He also visualises the applications of data mining, the goal might be to use a model predictively, to

provide automated classification of new instances. He finally concludes that the usage of Weka tool and its comprehensive suite of facilities for applying data mining techniques to large data sets. [4] Jayanta Basak, the author of this paper discusses about The central problem in weather and climate modelling, which is to predict the future states of the atmospheric system. Also the atmospheric correlation is being discussed in this paper. The author has provided techniques for determining the strongest independent components in the multidimensional data set. [5] In this paper published as part of the CRPIT conference, the author Ahsan Abdullah has taken the attribute pesticide into consideration and devised a correlation matrix between pesticide usage and yield. The input data is being subjected to RNR(Recursive Noise Removal) Framework and Similarity Matrix Corresponding to Input Data is been drawn. [6] Several information mining techniques utilised in the examination of agriculture. The author has discussed few of the techniques here. Different changes to weather have been analysed by the usage of SVM. Additionally, K-Means technology is used to forward the pollutants in the atmosphere. K Means technique is likewise used to classify the soil and plants. It's far observed that efficient methods may be advanced and analysed using the best information, the information which is collected from Kolhapur district to clear up complex agricultural problems by using information mining techniques. [7] The author proposes the methodology as follows, Take an input Training Dataset such as Mushroom or Soyabean. Apply PSO-SVM Feature Selection Algorithm for the selection of important Features from the Dataset on the basis of which classification can be done accurately. Apply Fuzzy Decision Tree for the Generation of Decision Tree and generate rules using Fuzzy decision Tree.Store the Generated rules. This methodology provides high Accuracy as compared to the existing methodology and less Error Rate and High Positive Rate. [8] Berhanu Borena in this paper has discussed the crop productivity of Ethiopia in this paper. He has taken various attributes that come into consideration. Also he has provided detailed description about handling large data sets and how to pre-process it.This study recommends for policy makers to make proactive decisions in identifying which factors are the most important to increase productivity. From the result, we can conclude three major things. The first one is out of all attributes used, fertiliser use has the highest predictive power. The second one is, out of the three algorithms tested, J48 has shown more predictive power. Of course all of the three algorithms have shown almost the same efficiency. The third conclusion is that the data may not have efficient predictable power as only one year.

III. PROPOSED WORK

In our paper, we propose to discuss various models and how they perform for the given datasets. In the result analysis we would discuss the results obtained by the most accurate models.

A. Artificial Neural Networks

Artificial Neural Network is one of the most used technique for prediction models, ANN is based on the structure and features of Neural Networks, the imitation of human brain. In this the primary computational units are called as neurons, these neurons are connected together in layers, where the data is passed in as the input the network is trained throughout with special equations called as the activation functions. The applications of neural network is widely used for agricultural practices.As soon as the neural network is skilled it is able to expect the crop yield in comparable patterns, despite the fact that the prior data consists of a few mistakes. Even if the statistics are complex, multivariate, nonlinear this community offers the correct effects and also without any of underlying concepts the relationship between them and the output is extracted. The process of classification by an ANN is done by these following steps.

1) Initialize the input data, classify into test and train dataset.
2) Run a sample from the training set, by means of giving its characteristic values as input.
3) Propagate the input throughout the network with variable weights, with a suitable activation function.
4) Minimize the network error by using back propagation process.
5) The summation of weights and activation functions location unit implemented at each node of hidden and output layers until the output is generated.
6) Compare the output with the predicted output from training set.
7) If the expected output or the accuracy is achieved then, the model will be predicting the required features for the crop predicting model.

The benefit of ANN system over the alternative system as in, is that it is able to version the rainfall. Also it predicts the pest assault incidence for one week in advance. Information mining tools are beginning to expose cost in studying massive statistic units from complicated structures and supplying notable records. Artificial neural network (ANN) is an appealing opportunity for building a information-discovery environment for a crop production system. The model we use for these agriculture data is a sequential model and the most used activation function for this kind of datasets is linear activation function and sigmoid. The output also varies based on the type of activation function we choose so it is essential to train your model with proper activation function.

B. Support Vector Machines Rainfall prediction Model

Support vector machine is a supervised prediction model which comes with a classification problem. In this SVM the data is represented in a planar model where it can predict whether a brand new example falls into the same category or alternate categories. Every set in SVM is categorised into a training model. In the rainfall prediction data set the attributes like temperature average, the diurnal range is taken, and then a planar model is applied. This model is similar to

regression model where the best fit line is drawn between the categorised training sets. When there are two attributes to be predicted these are mapped between categorised groups on a two-dimensional plane. When it comes to a three-dimensional SVM technique, then n attributes can be taken into a single category and the prediction is applied to the model. Based on the principle of scale compatibility model the classifier is then changed into a frame. The prediction model in SVM is also changed based on the parameters and the error resources that we consider for training the process. The advantages of SVM technique is that it is easy to implement when compared to the other prediction models. It has good reasonable baseline performance. To perform better, the data can be vectorised and then prediction can be applied. Dimensional reduction is also one of the important features that are used for training the model.

C. Bayesian network

A Bayesian network is a probabilistic graphical model which can be used for statistical analysis of the attributes for a given dataset. In this model, the attributes are represented in charts which are directed node by node. The nodes represent the probabilistic function, and the edges represent the conditional dependencies of the attributes. The statistically fit functions can be calculated outdrawn from the given graphical models where the predictions can be made.

This approach is particularly useful for ecological modelling because anticipated patterns may emerge at a variety of scales, constraining a multiplicity of model forms. In this work, we didn't train a Bayesian network but a simple graph is represented how we can future use it for a prediction model. This approach explicitly offers with the uncertainty of facts and relationships and may consist of both qualitative and quantitative variable. The disadvantages of the Bayesian network is that it cannot be applied to the large datasets and few functions take more computation time for the prediction. Few of the functions which can be used on the agricultural data are sigma functions and cross-corpora functions.

D. Clustering Model

Clustering is a technique where we categorise attributes which are correlated into various subsets, such that the intersection of any two subsets results in a null set. This comes under an unsupervised problem where the data is unlabeled and not continuous, This clustering technique is widely used for grouping the data based on the categories. For this present problem of analysis of the agricultural data, this gave us one of the best results as the correlation of attributes between each is broadly recognized. So when it comes to attributes like average temperature, precipitation we can group this using the clustering techniques so that the prediction model will be more accurate than the supervised training. Based on the predicted attributes the clustering techniques can be applied to two types. Firstly two-dimensional clustering where we take in two characteristics as input and categorise them into clusters and three-dimensional clusters where we can take n attributes and classify them. Since our data has few attributes and more accuracy is expected we choose a two dimension clustering model. When it comes to training, the rainfall data set the time complexity is increased as the grouping take more computation power than the normal linear model. Experimented techniques on the dataset using clustering gave better results gave us more accuracy than other models but the time taken by the algorithm is more when compared.

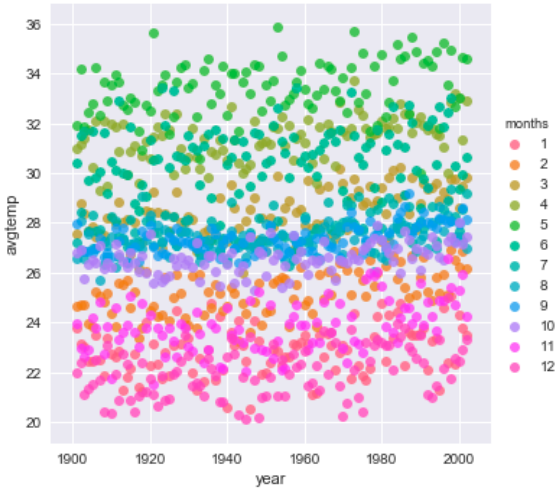

Fig. 1. Clustering predictions with respect to year and temparature.

Further, the bi-clustering techniques can be applied where the data is represented in a tabular or a matrix format. Once the data is represented in a matrix format, the similar rows and columns are categorised into bi-clusters then the prediction model is applied.

E. Conventional Techniques for prediction using Decision trees

A decision tree is used to visualise the data in a tree-like pattern which flows/increases with several conditions and terminates at the particular point. The termination if a decision point can be a prediction or an analysis. Decision trees are one of the most commonly used techniques in data mining. These can be applied to several fields like agricultural data or any network applications such as e-commerce websites. This decision tress basically starts with a root on the top of the tree and stops with some conclusion at the end. The branches and internal sub-tress are called as the nodes of the tress, which are several conditions and mathematic functions. When these decision trees come to rainfall prediction or agriculture analysis, we can easily represent the feature importance and the relations between several attributes in the dataset.

F. ARIMA Prediction Model

Agriculture mostly depends on the weather, so in this case, we see a lot of variations between agricultural attributes

concerning the time. So for the highest accuracy of the prediction, it is important to consider time as one the main important feature. This ARIMA(Autoregressive integrated moving average model) is mostly used for prediction which involves the time and forecasting parameters. It is a nonlinear transformation method which is used for the logging and deflating predictions. To apply the ARIMA model for the best accurate training we need to follow the below steps. Firstly, we need to convert all the time attributes into a time series model. Then the attributes depending on the time should be normalised. Next, we calculate the autocorrelation factor (ACF) and partial autocorrelation factor (PACF) from the normalised data.

$$\frac{\text{Covariance}(x_t, x_{t-h})}{\text{Std.Dev.}(x_t)\text{Std.Dev.}(x_{t-h})} = \frac{\text{Covariance}(x_t, x_{t-h})}{\text{Variance}(x_t)} \quad (1)$$

$$\rho_1 = \frac{\text{Cov}(x_t, x_{t+1})}{\text{Var}(x_t)} = \frac{\phi_1 \text{Var}(x_t)}{\text{Var}(x_t)} = \phi_1 \quad (2)$$

ACF, PACF formula.

These two factors will give us the time patterns and the correlation between the variables. Once the elements are known we fit our inputs into the ARIMA model and then the forecasting of the desired attributes like rainfall, climate moisture control is made. We plot the ACF and PACF factors of the ARIMA model so that no more extraction is left out.

IV. RESULT ANALYSIS

A. Grouping of attributes based on K- Nearest Neighbours

In this KNN the classification model is created by the training dataset. The class labels are then classified based on the correlated and associated dataset. The correlation between the data is calculated by a mathematical formula called as Euclidean distance. Mathematically the euclidean distance is represented by the formula. The accuracy levels of KNN are similar to the clustering technique as both the algorithms the data is classified based on the correlation of the attributes.

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \quad (3)$$

N nearest neighbours: 2
Training data accuracy: 90.92607733462055
Testing data accuracy : 88.73142013546567

N nearest neighbours: 11

Training data accuracy: 0.9412973104699223
Testing data accuracy : 0.881218513239697

TABLE I
ACCURACY METRICS BASED ON N NEIGHBORS

train on 1	0.8521386022239681
test on 1	0.7903567159622974
train on 2	0.9092607733462055
test on 2	0.8873142013546567
train on 3	0.9200017636367522
test on 3	0.887083916875027
train on 4	0.9186770428293242
test on 4	0.8884225911339733
train on 5	0.9299339066248916
test on 5	0.8845623211826944
train on 6	0.9322563517755024
test on 6	0.8836090297219555
train on 7	0.936967836206562
test on 7	0.8848134277850955
train on 8	0.9382930319975269
test on 8	0.8835048035075506
train on 9	0.9399501469463843
test on 9	0.8827601790846294
train on 10	0.9394392712091164
test on 10	0.8836072947393012
train on 11	0.9412973104699223
test on 11	0.881218513239697
train on 12	0.9416958755604256

B. Multivariate Linear Regression

Linear regression is one of the most used techniques for predicting a series. This linear regression comes under supervised learning where the data is arranged continuously pattern. We choose linear regression as our first step for the model as our data is unlabeled. The features of our crop yield prediction include average temperature, precipitation, diurnal temperature range, potential evapotranspiration concerning year. All this data is arranged in heat map where the index is the month, and the column is the year. Initially, the model we took for this prediction is a simple line equation mathematically, y=mx+c.

Regression Line:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\epsilon}_i \quad (4)$$

Slope:

$$\hat{\beta}_1 = \frac{\sum(X_i \bar{X})(Y_i \bar{Y})}{\sum(X_i \bar{X})^2} \quad (5)$$

Intercept:

$$\hat{\beta}_0 = \bar{Y}\hat{\beta}_1 \bar{X} \quad (6)$$

The process for prediction first includes cleaning of the given dataset. First, clean the data by removing/replacing all 'NaN'/missing values. To apply a linear model to a dataset the data set is split into two more data frames Training dataset and testing dataset. The training dataset contains all the input features given above, and the test data set is divided to find the accuracy of the model. We use other metrics like Standard Deviation, Mean square error for detecting the accuracy of the model.

After visualising the data, we see the correlation between few attributes like average temperature, and we found that linear regression can be applied to prediction model rather than any other techniques. The inputs we took our data of past 100 years with all these attributes. We used the above procedure and divided the data into train and test sub-datasets and then fit into the linear model. The accuracy we found was 68 percent, and the results were pretty decent when compared to other linear models. All this computation was made using

sklearn a python library for the rainfall prediction. The prediction using the linear model are tabulated below

TABLE II
TEST DATA SET PREDICTION USING LINEAR REGRESSION

[1989]	[23.18732596]
[1990]	[23.19603139]
[1991]	[23.20473681]
[1992]	[23.21344223]
[1993]	[23.22214766]
[1994]	[23.23085308]
[1995]	[23.23955851]
[1996]	[23.24826393]
[1997]	[23.25696936]
[1998]	[23.26567478]
[1999]	[23.27438021]
[2000]	[23.28308563]
[2001]	[23.29179105]
[2002]	[23.30049648]

KNN is based on feature similarity; it depends on the reach of how approximately the training set determines to classify the given data at a point. The nearest neighbour parameter is defined as the number of training samples described in the closest new point for the prediction.It is the user-defined constant based on the local density of points. For this agricultural data, we trained our prediction model by changing the nth nearest neighbours parameters from 1 to 11 due to the number of months present in our data. We found the results to be varying for each nth neighbours. The accuracy of the test and train data are tabulated below. These increase continuously until 11 nearest neighbours and then reached a saturation point.

TABLE III
JANUARY AVERAGE TEMPERATURE PREDICTION

[2018]	[23.43978327]
[2019]	[23.44848869]
[2020]	[23.45719412]
[2021]	[23.46589954]
[2022]	[23.47460497]
[2023]	[23.48331039]
[2024]	[23.49201581]
[2025]	[23.50072124]
[2026]	[23.50942666]
[2027]	[23.51813209]
[2028]	[23.52683751]
[2029]	[23.53554294]

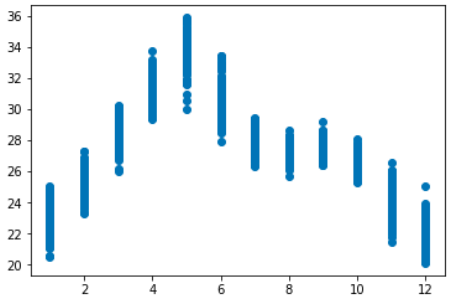

Fig. 2. Image predicted average temperature for month. (X axis - temparatue ranges, Y axis - month)

V. CONCLUSIONS

In this work, we applied several prediction models for soil behaviour and rainfall analysis; we found KNN classification model to be the best as the training accuracy is higher when compared to other models. The linear model underperformed as the data is non-continuous. There are very few applications for models like Bayesian Networks and Decision tress as they are only confined to limited attribute prediction. We can achieve more accuracy for our proposed model by using particle swarm, ant colony optimization.

REFERENCES

[1] [1]Dr. D. Ashok Kumar, N. Kannathasan, A Survey on Data Mining and Pattern Recognition Techniques for Soil Data Mining
[2] S.Veenadhari, Dr. Bharat Misra, Dr. CD Singh, Data mining Techniques for Predicting Crop Productivity A review article, International Journal of Computer Science and Technology IJCST Vol. 2, Issue 1, March 2011.
[3] Developing innovative applications in agriculture using data mining, Sally Jo Cunningham, Geoffrey Holmes
[4] Jayanta Basak, Anant Sudarshan, Deepak Trivedi, MS Santhanam,Weather data mining using independent component analysis,Journal of Machine Learning Research,Volume 5, 2004
[5] Ahsan Abdullah, Stephen Brobst, Ijaz Pervaiz, Learning Dynamics of Pesticide Abuse through Data Mining.
[6] Perpetua Noronha, Divya .J, Shruthi .B.S, Comparative Study of Data Mining Techniques in Crop Yield Prediction.
[7] Raorane A.A., Kulkarni R.V, Data Mining: An effective tool for yield estimation in the agricultural sector.
[8] Zekarias Diriba, Berhanu Borena,Application of Data Mining Techniques for Crop Productivity Prediction.
[9] White, David, Howden, S. Mark,Climate Change: Significance for Agriculture and Forestry,Springer,May 1994.
[10] Ms.Kalpana.R, Dr.Shanthi. N , Dr.Arumugam.S, A Survey on Data Mining Techniques in Agriculture.
[11] Searching for Activation Functions ,Prajit Ramachandran, Barret Zoph, Quoc V. Le