CROP YIELD PREDICTION BASED ON ENSEMBLE MODEL USING HISTORICAL DATA

DEEPAK.G, DEEPIKA.J, DHARSHINI.M, DR.B.VANATHI

[1] *Student, Department of Computer Science and Engineering, SRM Valliammai Engineering College , Chennai, India*
[2] *Student, Department of Computer Scien*ce and Engineering, SRM Valliammai Engineering College , Chennai, India
[3] *Student, Department of Computer Scien*ce and Engineering, SRM Valliammai Engineering College , Chennai, India
[4] *Professor, Department of Computer Scien*ce and Engineering, SRM Valliammai Engineering College , Chennai, India

ABSTRACT

India is a country where agriculture and agriculture related industries are the dominant source of living for the people. Agriculture is in poor condition since before comparing previous years. The vital reason for this is without a well organized pattern about farming and proper instructions to the farmers. Agriculture is an extensive source of economy of the country. Apart from this India also ail from natural disasters like flood and drought which forfeit the crops. One's strategy should be spot on while harvesting the crops, factors like season, soil moisture and weather condition should be well planned and also when to harvest the crop to get the maximum yield. In this project study, we provide an effective recommendation system for the farmers using machine learning techniques. In the proposed system we use ensemble based hybrid model to make the classifier strong. The input dataset is been obtained from the public repository. The input historical dataset in provided to machine learning platforms for data processing. The clustering of data is performed using Guassian Mixture Model clustering and for classification of data we use KNN and XGBboost techniques to define the conditions and solutions. This would provide effective recommendation for the farmers in defining the best crop which can be farmed to obtain huge profitable yield analyzing the environmental factors

Keyword : - *Machine learning (ML), Gaussian Mixture Model(GMM), XGBoost(XGB),Ensemble Model*

1. INTRODUCTION

Agriculture is the backbone of the Indian economy. In India, agricultural yield primarily depends on weather conditions. Rice cultivation mainly depends on rainfall. Timely advice to predict the future crop productivity and an analysis is to be made in order to help the farmers to maximize the crop production of crops. Yield prediction is an important agricultural problem. In the past farmers used to predict their yield from previous year yield experiences. Thus, for this kind of data analytics in crop prediction, there are different techniques or algorithms, and with the help of those algorithms we can predict crop yield. Random forest algorithm is used. Using all these algorithms and with the help of inter-relation between them, there are growing range of applications and the role of Big data analytics techniques in agriculture.

Since the creation of new innovative technologies and techniques the agriculture field is slowly degrading. Due to these, abundant invention people are concentrated on cultivating artificial products that are hybrid products where there leads to an unhealthy life. Nowadays, modern people don't have awareness about the cultivation of the crops at the right time and at the right place. Because of these cultivating techniques the seasonal climatic conditions are also being changed against the fundamental assets like soil, water and air which lead to insecurity of food.

By analysing all these issues and problems like weather, temperature and several factors, there is no proper solution and technologies to overcome the situation faced by us. In India, there are several ways to increase the economic growth in the field of agriculture. There are multiple ways to increase and improve the crop yield and the quality of the crops. Data mining is also useful for predicting crop yield production.

2. LITERATURE SURVEY

2.1 Predicting yield of the crop using machine learning algorithm

Predicting yield of the crop using machine learning algorithm. This paper focuses on predicting the yield of the crop based on the existing data by using Random Forest algorithm. Real data of Tamil Nadu were used for building the models and the models were tested with samples. Random Forest Algorithm can be used for accurate crop yield prediction.

2.2 Random forests for global and regional crop yield prediction

Random forests for global and regional crop yield prediction. Our generated outputs show that RF is an effective and adaptable machine-learning method for crop yield predictions at regional and global scales for its high accuracy and precision, ease of use, and utility in data analysis. Random Forest is the most efficient strategy and it outperforms multiple linear regression (MLR).

2.3 Crop production Ensemble Machine Learning model for prediction

In this paper, KNN and XGBoost are the proposed ensemble model used to project the crop production over a time period. Implementation done using this KNN and XGBoost will boost the efficiency.

2.4 Machine learning approach for forecasting crop yield based on climatic parameters

Machine learning approach for forecasting crop yield based on parameters of climate. In the current research a software tool named Crop Advisor has been developed as a user friendly web page for predicting the influence of climatic parameters on the crop yields.C4.5 algorithm is used to produce the most influencing climatic parameter on the crop yields of selected crops in selected districts of Madhya Pradesh. The paper is implemented using Decision Tree.

2.5 Prediction On Crop Cultivation

Presently, soil analysis and interpretation of soil test results is paper based. This in one way or another has contributed to poor interpretation of soil test results which has resulted into poor recommendation of crops, soil amendments and fertilizers to farmers thus leading to poor crop yields, micro-nutrient deficiencies in soil and excessive or less application of fertilizers. Formulae to Match Crops with Soil, Fertilizer Recommendation.

3 EXISTING SYSTEM

In existing system, the farming land properties keep on changing as per the cultivation and environmental changes. Hence machine learning based historical data analysis of that farm and crop production is essential to recommend the farmer which crop has to be cultivated in the specific farm based on the previous data. Lack of this system only lead to financial and time loss for the farmers costing their life’s in the exixting system K means algorithm is used which has more disadvantage and decision tree algorithm is used.

3.1 DISADVANTAGE OF EXISITING SYSTEM

a)Because we use K-means algorithm:

i.Difficult to predict K-Value.

ii.With global cluster, it didn't work well.

iii.Different initial partitions can result in different final clusters.

iv.It does not work well with clusters (in the original data) of Different size and Different density

b)Because we use decision tree algorithm:

i.A small change in the data can cause a large change in the structure of the decision tree causing instability.

ii.For a Decision tree sometimes calculation can go far more complex compared to other algorithms.

iii.Decision tree often involves higher time to train the model.

iv.Decision tree training is relatively expensive as the complexity and time has taken are more.

v.The Decision Tree algorithm is inadequate for applying regression and predicting continuous values.

4 PROPOSED SYSTEM

In the proposed system, we provide effective farmer recommendation system analyzing the previous historical data. For processing the huge data we need machine learning tools for analyzing and predicting the solutions. In this project study, we provide an effective recommendation system for the farmers using machine learning techniques. In the proposed system we use ensemble based hybrid model to make the classifier strong. The input dataset is been obtained from the public repository. The input historical dataset in provided to machine learning platforms for data processing. The clustering of data is performed using Guassian mixture and for classification of data we use knn and XGBoost to define the conditions and solutions, So that machine itself can match the condition for each clustered data and provide the respective solution for it. This would provide effective recommendation for the farmers in defining the best crop which can be farmed to obtain huge profitable yield analyzing the environmental factors.

4.1 ADVANTAGES OF PROPOSED SYSTEM

a. Because we use Gaussian Mixture Model:

i.It is a probabilistic method for obtaining a fuzzy classification of the observations. The probability of belonging to each cluster is calculated and a classification is usually achieved by assigning each observation to the most likely cluster. These probabilities can also be used to interpret suspected classifications.

ii. Mixture modeling is very flexible.

b. Because we use XGBoost:

i. Regularization

ii. Parallel Processing

iii. Handling Missing Values

iv. Cross Validation

v. Effective Tree Pruning

4.2 PROPOSED ARCHITECTURE

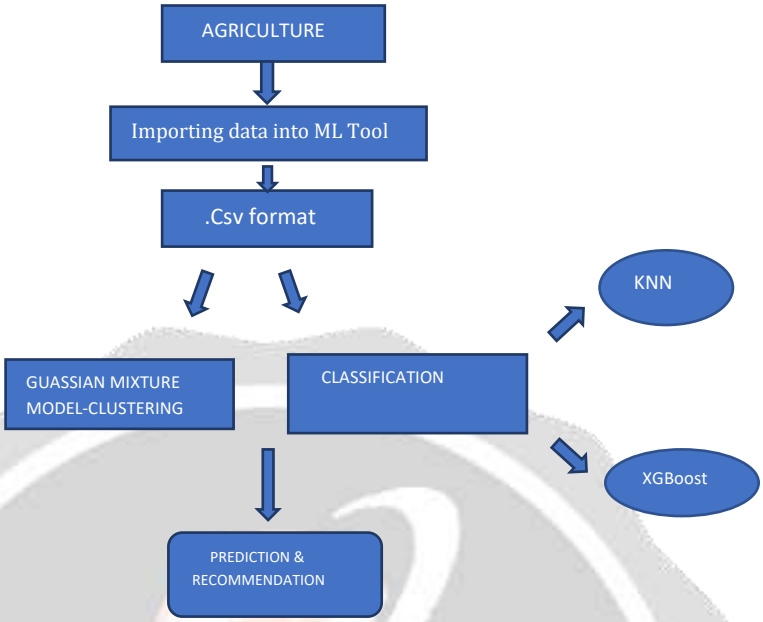

Fig: 1 Architecture of the proposed system

5 METHODOLOGY

5.1 Dataset Acquisition

The climate data obtained from indianwaterportal.org and the crop production data obtained from faostat3.fao.org is taken into account for the study. The climate data contains various variables which are responsible for the rainfall for a specific region and the quantum of crop production for that region is taken into account for this work. The historical climate data of the CSV type with a record for every line of text alike to the data for a given month of an year. The crop production data of the CSV type with a record for each line of text belongs to the data for a specific month of an year.

In the preprocessing stage, though there are many measured parameters available in the raw climate dataset, the less relevant features responsible for the study are ignored and the important features are only taken into account

5.2 Clustering

Clustering is one of the most common exploratory data analysis technique used to get an intuition about the structure of the data.

Gaussian mixture model is best for clustering it overcomes the disadvantage of k mean algorithm guassian model expands its cluster as elliptical to obtain best clustering it is a probabilistic model which datapoints are generated with unknown parameters of finite number of distribution. The below figure 8.1 shows the diagram of the difference between k-means and Gaussian Mixture model.

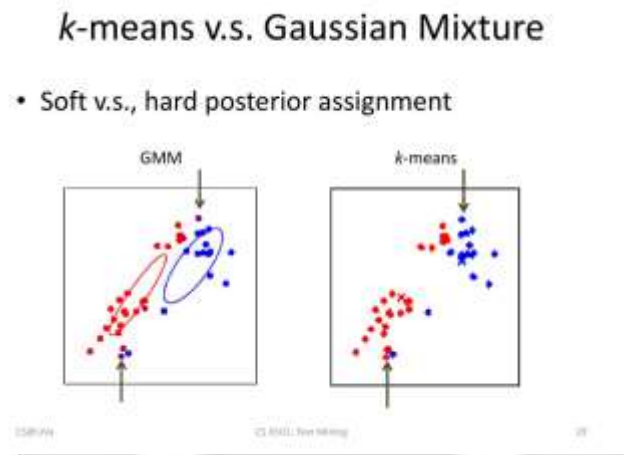

Fig 2 K-Means vs Gaussian Mixture algorithm

5.3 Machine Learning Algorithms

Decision Tree is a supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches. The decisions or the test are performed on the basis of features of the given crop dataset.

The KNN algorithm is a benchmark classifier, which is often used for more complex classification. Despite its simplicity, KNN acts as a more powerful classifiers and is used in most of the applications today. KNN was leveraged in the year 2006 for the assignment of genes based on their expression profiles, but now it is used in many fields for classification and provides very good results. In k-NN classification, the output is a class with similar instances. The input given by us will be classified by a majority vote of its neighbors, with the given k-value being assigned to the class, most common among its k nearest neighbors. If k = 1, then the object is simply assigned to the class of that single nearest neighbor. The k-NN algorithm is among the simplest of all machine learning algorithms. KNN algorithms use a data and classify new data points based on a similarity measures

5.4 Ensemble Model

Combining with many other learning algorithms, the meta algorithm is called as AdaBoost algorithm. This would improve the performance of classification. AdaBoost uses the nested operator and it has a sub process. The sub-processor is used to generate a better model. The ensemble model creates more than one classifier and generates a better model. The accuracy of classification is expanded by creating more than one classifier by the ensemble model. The ensemble model leads to decision making by combining the results of their classification techniques.

5.5 Prediction

Forecasting of crop production is done by using the time series data set precisely than the existing models. By using AdaBoost technique, ensemble models such as KNN and Regression techniques are developed. Importance of crop prediction is highly needed for agriculture and economy.

6 CONCLUSION

In this proposed system, we provide effective farmer recommendation system analyzing the previous historical data. The clustering of data is performed using Guassian Mixture Model- clustering and for classification of data we use KNN and XGBoost techniques to define the conditions and solutions, So that machine itself can match the condition for each clustered data and provide the respective solution for it. This would provide effective recommendation for the farmers in defining the best crop which can be farmed to obtain huge profitable yield analyzing the environmental factors.

7 FUTURE SCOPE

For the Future scope, new algorithm can be added for better performance and User interface can be added so that the use of the system will be easy. Instead of only historical data the live weather conditions will be given as data so that we get best efficient recommendation output.

8 REFERENCES

[1] M. Kalimuthu; P. Vaishnavi; M. Kishore," Crop Prediction using Machine Learning", 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)

[2] Y. Jeevan Nagendra Kumar; V. Spandana; V.S. Vaishnavi; K. Neha; V.G.R.R. Devi," Supervised Machine learning Approach for Crop Yield Prediction in Agriculture Sector", 2020 5th International Conference on Communication and Electronics Systems (ICCES)

[3] S. Veenadhari; Bharat Misra; CD Singh,"Machine learning approach for forecasting crop yield based on climatic parameters",2014 International Conference on Computer Communication and Informatics

[4] Neha Rale; Raxitkumar Solanki; Doina Bein; James Andro-Vasko; Wolfgang Bein,"Prediction of Crop Cultivation",2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC)

[5] Toshichika Iizumia , Yonghee Shinb , Wonsik Kima , Moosup KimbX , Jaewon Choib, "Global crop yield forecasting using seasonal climate information from a multi-model ensemble", International journal on climate services, 2018.

[6] David B. Lobell and Gregory P. Asner, "Comparison of Earth Observing- 1 ALI and Landsat ETM+ for Crop identification and Yield prediction in Mexico", international journal on geo science an remote sensing, June 2003

[7] P. S. VijayabaskarX, Sreemathi.RX, Keertanaa.E,"Crop prediction using predictive analysis", international conference on computation of power energy information and communication, 2017

[8] Yogesh gandge, Sandhya, "A Study on Various Data Mining Techniques for Crop Yield Prediction", international conference on Electrical, Electronics, Communication, Computer and Optimization Technique, 2017

[9] Jin Chaun, Qin Qiming, Zhu Lin, Nan Peng, Abduwasit Ghulam, "TVDI based Crop Yield Prediction model for Stressed Surface", international conference on Autonous Region, 2007.

[10] Joon-Goo Lee, Haedong Lee, Aekyung moon, "Segmentation method of COI for Monitoring and prediction of the crop growth", international conference on embedded system research section, 2014.