Machine learning techniques on agro meteorological parameters to predict the crop yield in support of building decision support system to farmer’s community

Sowmya B J, Seema S, Daniel Monteiro, Zaid Hossain, Zubin Paymaster and Aditya Tomar

Ramaiah Institute of Technology, MSR Nagar, Bengaluru, Karnataka 560054, India

E-mail: sowmyabj@msrit.edu

Abstract. Over the years, farmer grievances have been rapidly increasing due the barrage of issues that they face on a regular basis. This includes not getting good harvests, shortage of resources, etc. Decision making is an important aspect of farming, as it can improve results tremendously, hence helping farmers. With the rise in technology, and great research in the realm of Data Science, there is an increasing demand for technical solutions for farmers. To solve this, we built a Decision Support System that predicts a crop to be grown based on agrometeorological factors such as rainfall, temperature, and humidity. The system was built using various machine learning algorithms. We compared multiple models, change algorithmic parameters, and decide the best model to be used. This system aims to aid the decision-making process of a farmer, and hence help them gain maximum yield with the knowledge and resources available.

Keywords: Agro meteorological, Decision Support, Humidity, Classification, Decision tree, Manhattan distance

1. INTRODUCTION

Farming is the foundation of the Indian economy, though the business as of now needs more help than some other. India is a nation of over a billion people in populace, out of which, over 70% of the populace lives in the country. Comprising 40% of the nation's workforce, agriculture is a noteworthy industry and a key influencer of the Indian economy. Regardless of this, its commitment to the 2.3 lakh crore rupee economy is only a pitiful 16% of the whole GDP. In India, agriculture needs institutional consideration, support from banks in terms of loans and welfare schemes, and experiences the ill effects of a horde of fiascos like exhausting groundwater levels in country zones, environmental change, dry seasons, floods, uncalled for value fixing approaches of produce, relocation of farmers towards the urban communities looking for better paying occupations, and much more. Agriculture is one division in charge of feeding each person, yet the general population associated with it are the last to be dealt with. In the wake of coming up short establishments, time has without a doubt desired innovation to assume control over the change. With more up to date issues springing up each day in the most inescapable indigenous areas, it's about time that we resort to rising technologies for arrangements. Farmers currently depend on traditional methods of farming, hence rendering their produce ineffective in the vast majority of instances. With the help of technology, farmers will be empowered with the knowledge required to make the right decisions in order to yield
maximum benefit from the resources they have at hand. This is a study of the dependence of various crops on certain agrometeorological factors like humidity and rainfall. Based on the datasets, various classification algorithms will be applied and compared in order to figure out which one is the best suited algorithm for this Decision Support System. Our work also aims to modify algorithmic parameters in classification algorithms and analyze the effect the modification has on the accuracy they achieve. This is an analysis of datasets obtained regarding the dependence of various crops on agrometeorological factors such as humidity, temperature and rainfall. The main aspects of the project converge on the inferences drawn using the different machine learning approaches on these datasets.

From the above we can formulate the problem statement as: "To create an efficient Decision Support System to decide the right crops to be grown based on agrometeorological factors like rainfall, humidity, and temperature by applying various classification algorithms, modifying algorithmic parameters, and comparing their accuracy in order to decide the best suited algorithm, and hence provide a computational measure in order to help farmers gain maximum yield with the resources available to them and provide a computational solution to the agriculture sector."

The objectives of the work are:

- To build a decision support system to decide the right crops to be grown based on agrometeorological factors.
- To compare the accuracy of various classification algorithms.
- To modify certain parameters in the classification algorithms used and study the effects of the modification and its accuracy.
- To identify the best-suited algorithm for this dataset.

The deliverables of the idea can be stated as:

- Analysis of crop requirements.
- Helps farmers make better use of the resources available.
- Helps farmers decide the right crops to grow and get a better yield.
- To find the differences in accuracy of predicting crops based on various classification algorithms.

Agriculture is one sector that is responsible for feeding the person, but it is the last to take care of the people involved in it. Yes, the time has come for technology to take over the transition after institutions have collapsed. For new problems in the most unavoidable indigenous industries happening every day, it is high time we resorted to emerging solutions technologies. One way to solve this problem is to aid the decision-making process of farmers. This can be done by building decision support systems. This can be found using the available data sets that are reliable and genuine. The current scope hence depends on the availability of this data or feasibility collection of such data for a given region.

2. LITERATURE SURVEY

The study, Machine learning approach to crop yield prediction based on climatic parameters [1] has shown the possible use of data mining techniques to predict crop yield based on climatic input parameters. The developed webpage is user-friendly and prediction accuracy in all crops and districts selected in the study showing higher prediction accuracy is above 75 percent. Any consumer can use the user-friendly web page built to predict crop yield by providing climate data from that location. The study, Climate Attribute Analysis to Predict Crops for the Season Using Data Mining [2] addressed how variables such as temperature, photoperiod and grid also have a significant impact on crop boom and yield. In general, there are two methods for forecasting rainfall. Empirical technique and method of dynamics. Regression, artificial neural network, fuzzy logic and systematic approach to dealing with statistics are the most commonly used statistical methods used for weather prediction.
They use data mining techniques like clustering and precipitation forecasting classification techniques. An article, Crop Yield Prediction Article Using Data Mining Techniques [3], analyzes several models of prediction. The Multiple Linear Regression technique is initially applied to existing data. The results thus obtained have been checked and analyzed using the technique of Data Mining, namely clustering based on size. The findings of two methods are compared according to the particular region i.e. in this procedure, Andhra Pradesh district in East Godavari, India. Similar process has been adopted for all Andhra Pradesh districts to boost and authenticate the yield prediction validity that is useful for Andhra Pradesh farmers to forecast a specific crop. A comparison of the crop yield forecast can be made with the entire set of available data in the subsequent work and will be committed to appropriate approaches to improve the effectiveness of the proposed technique. A Research & Survey on Rainfall Prediction and Crop Production Using Data Mining Techniques [4] is another research. We use the K-Means in clustering, is one of the simplest and best predictive work methods. Second, we've been using the fuzzy logic. Eventually, they used NeuroFuzzy's genetic algorithm hybrid process. In that we conclude that, as opposed to the other algorithm, these three models & Optimization of these three K-Means, Fuzzy Logic & NeuroFuzzy with Genetic Algorithm models are useful to determine how excellent and reliable agricultural crop production results are based on prediction of rainfall. A crop yield prediction model [5] relies on a predictive model that can be used to forecast crop yield in the future. This paper presents a brief analysis of crop yield prediction based on data mining technique for the selected area, i.e. Tamil Nadu district in India. The experimental results indicate that the proposed work effectively predicts the production of crop yields. A research paper, Rainfall Prediction using Data Mining Techniques: A Systematic Literature Review [6], discusses a valuable yet daunting role in forecasting rainfall. Data mining techniques have the ability to predict the rainfall by extracting and using the hidden knowledge from past weather data. In the last decade, many researchers have worked to increase the accuracy of rainfall prediction by optimizing and integrating data mining techniques. Various models and techniques are available today for successful prediction of rainfall, but there is still a lack of a concise analysis of literature and a comprehensive mapping study that could reflect current challenges, proposed solutions and the latest trends in this area. It was therefore concluded that upgrades, modifications and integration of data mining methods were essential in order to explore and solve these problems. Another review paper, The Machine Learning Techniques for Rainfall Prediction: A Review [7,8] recognizes that rainfall prediction accuracy is of great importance to countries such as India, whose economies are largely dependent on farming. Statistical techniques do not provide good accuracy for predicting precipitation due to the dynamic nature of the atmosphere. Nonlinearity of precipitation data [10] is a good strategy for Artificial Neural Network. Study work and comparison of various approaches and algorithms that researchers use to forecast rainfall is shown in a tabular form. The aim of this paper [9] is to provide non-experts with easy access to rainfall forecasting techniques and approaches. Many papers focused on predicting crop yield for the next season and predicting the next season's rainfall. All the papers made use of existing algorithms and made classifications/predictions without changing mathematical parameters. Due to this, there is a scope for predicting crops based on agrometeorological data. Further changing the mathematical parameters and verifying whether it’s accuracy can be better than the existing algorithms, which is what we are going to do over the course of our project. Also, the papers did not consider humidity as a factor for growing crops; we are using it as a factor in our classification.

3. DESIGN

The architecture diagram depicts our model as shown in fig 1. Starting with applying classification models to our data set to changing the algorithmic parameters of the best model (KNN here) and finding the best accuracy.
4. IMPLEMENTATION AND RESULTS

4.1. Explanation of Algorithm and how it has been implemented
4.1.1. Random Forest. Random Forest is an algorithm for supervised learning. It produces a forest, as you can see from its name, and makes it somehow random. The forest it creates is an ensemble of Decision tree, many equipped with the system of "bagging." The general idea of the bagging approach is that the overall result is improved by a mixture of learning models.

Simply put, random forest constructs and merges several decision trees to achieve a more accurate and stable prediction. In the diagram above, multiple trees perform classification, a vote is then taken, and the majority class is selected as the final class.

4.1.2. Decision Trees.
Tree like decision graph or template and its potential sequences. The goal is to create a model that predicts the value of a target variable by learning simple decision rules based on data characteristics. The splits in the tree as shown in Fig 2 are made using calculations of entropy, gini and information gain.

In the fig 3, the root node splits into multiple decision nodes, and upon further splitting, and finally into terminal nodes that are classes.
4.1.3. K-Nearest Neighbours. K-Nearest Neighbours is in Machine Learning, it is one of the most simple yet important classification algorithms. It belongs to the supervised learning field and sees heavy use in pattern recognition, data mining and detection of invasion. It is widely available in real-life scenarios as it is non-parametric, meaning; it makes no underlying assumptions about data distribution (as opposed to other algorithms such as GMM, which assume a Gaussian distribution of the data). Some prior data (also called training data) are given to us, which classifies coordinates into groups defined by an attribute.

In the figure 4, there are two classes in the form of a diamond and a square. The diagram shows how they are classified based on a distance measure.

4.2. Information about implementation of Modules

4.2.1. Mean Measures of Crops. The fig 5 graph shows the mean measure of the crops vs rainfall, crops vs temperature, and crops vs humidity.
We conclude that Sugarcane and Rice comparatively require more rainfall than the other crops. Millet and Barley require relatively lower rainfall levels.

![Image](image1.png)

Figure 6. Mean measure of humidity of all the crops

From the fig 6 graph above, we infer that the average humidity of all the crops are quite similar to each other.

![Image](image2.png)

Figure 7. Mean measure of temperature of all the crops

From the fig 7 graph above, we infer that the average temperatures of all the crops are quite similar to each other.

4.2.2. Outlier Analysis. A dataset with 7500 tuples was considered for this project. Since all the values of temperature and humidity lied in the same range for all the crops, there was only a need to check for outliers in the rainfall data. We used the IQR Method to detect outliers as shown in fig 9.
Humidity vs Temperature: We plotted graphs of humidity vs temperature for all the crops as shown in fig 8. The correlation inferred from the graph helped us gain an understanding of the type of regions suitable for growing the respective crops.

4.2.3. Random Forest Implementation. We considered 100 estimators in this model of random forest classification. This means that there are 100 decision trees in the voting system to decide the class the data belongs to. This model is used to make sure that one tree doesn’t make the only decision. It’s always better to have multiple trees. Accuracy of the model = 90.76% and is shown in fig 10.
4.2.4. Decision tree. A decision tree is used to classify agrometeorological data into one of the six crops in the dataset. The tree’s root node is rainfall. From the decision tree, we infer places with rainfall which happens to be >200 is likely to grow sugarcane. The other factors which play a role in deciding the class are temperature and humidity. Accuracy of the model is 90.82% as shown in fig 11.

4.2.5. K-Nearest Neighbours.
The KNN algorithm is a supervised algorithm that takes data like rainfall, temperature, and humidity and classifies it into one of the crops in the dataset and is depicted in fig 12. There are many distance measures that we can use in K-Nearest Neighbours. When we ran KNN against Decision Tree and Random Forest, it provided the best accuracy. By default, the Euclidean distance measure is used. The accuracy of this model turned out to be the best = 92%.

4.3. Modifying K-Nearest Neighbours Parameters
Since KNN was the best model, we decided to change the distance measures into 3 distinct categories. Euclidean distance (power = 2), Manhattan distance (power = 1) and Minkowski distance (power = 5).
Based on these formulas, we have tested the machine learning model. Moreover, the number of neighbors is also a very important factor in deciding the accuracy of the machine learning model. We have tested the values of the accuracy of these models for a value of k ranging from [1 - 300].

Figure 12. Formulas of different distance measures.

Figure 13. Accuracy Plot of Euclidean Distance against values of K from 1-300.

Figure 14. Accuracy Plot of Manhattan Distance against values of K from 1-300.
Figure 15. Accuracy Plot of Minkowski Distance against values of K from 1-300

From this, we finally conclude that we should build a KNN classifier using the Manhattan distance measure, and setting K= 85 for the best accuracy. Hence, we have chosen this for our decision support system. The various measures as shown in the fig 13, 14 and 15, it shows accuracy levels for different distance formula.

5. CONCLUSION

Random Forest and Decision Tree models gave a very similar accuracy, the models were quite deep in terms of the levels in the tree in order to provide a good accuracy. The accuracy of these models was lower than K-Nearest Neighbors. The best model to use is the K-Nearest Neighbors model. The distance to be used is Manhattan Distance, as it provided an accuracy of 92.46 %, which is higher than the other versions of the model. We tested multiple K (neighbor) values, and observed that for Manhattan Distance, the K value of 85 got the best accuracy. For the given dataset, it was observed that crops like Sugarcane and Rice require heavy rainfall, whereas crops like Millet and Barley require low rainfall. Rainfall was the most relevant parameter in this dataset as the mean measures for the temperature and rainfall were very similar among all the crops in the dataset, as observed from the graphs. The decision support system can now be used to successfully predict the right crops based on agrometeorological data like rainfall, humidity, and temperature, and hence aid the decision-making process of farmers, and help them get a better yield. The model can be made better by obtaining the dependencies of crops on more factors, but this requires that the government assigns a body or a set of groups to collect such data. The study presented in this project could be used by government organizations to more effectively help farmers. It could be extended to an application that all these bodies or farmers use to predict what sort of crops can be grown in a particular region. The government of the respective city could invest in accumulating real-time data regarding crop vs factors that affect its growth, and then hire a team which obtains these inferences. This can then be used on a common platform all farmers across the country.

REFERENCES

[1]. Veenadhari S, Misra B, Singh CD, Jan. 03 – 05, 2014. International Conference on Computer and Informatics (ICCCI -2014). Machine learning approach for forecasting crop yield based on climatic parameters.

[2]. Teeda K, Vallabhaneni N, Sridevi T, International Journal of Pure and Applied Mathematics Volume 119 No. 12 2018. Analysis of Weather Attributes to Predict Crops for the Season Using Data Mining

[3]. Ramesh D, Vishnu Vardhan B, IJRET: International Journal of Research in Engineering and, Volume: 04 Issue: 01 | Jan-2015. ANALYSIS OF CROP YIELD PREDICTION USING DATA MINING TECHNIQUES.
[4]. Sundaravalli N, Geetha A, *International Research Journal of Engineering and Technology (IRJET)*. Volume: 03 Issue: 12 | Dec -2016. A Study & Survey on Rainfall Prediction And of Crops Using Data Mining Techniques.

[5]. Manjula E, Djodiltachoumy S, *International Journal of Computational Intelligence and*, Vol. 6: No. 4, March 2017. A Model for Prediction of Crop Yield.

[6]. Aftab S, Ahmad M, Hameed N, Bashir M, Ali I, Nawaz Z, *(IJACSA) International Journal of Computer Science and Applications*, Vol. 9, No. 5, 2018. Rainfall Prediction using Data Mining Techniques: A Systematic Literature Review.

[7]. Parmar A, Mistree K, Sompura M, *2017 International Conference on Innovations in Embedded and Communication Systems (ICIIECS)*. Machine Learning Techniques For Rainfall Prediction: A Review.

[8]. Zingade DS, Buchade O, Mehta N, Ghodekar S, Mehta C, *Special Issue on Recent Trends in Data Engineering* Volume 4, Special Issue 5, Dec.-2017. Crop Prediction System using Machine Learning.

[9]. Kumar A, Kumar S, *International Journal of Computer Science and Mobile Computing*, Vol. 4 Issue. 8, August-2015. Prediction of Production of Crops using K-mean & Fuzzy Logic.

[10]. Moorthy A, Buermann W, and Rajagopal D, June 12, 2012. *The Impact of Climate Change on Crop Yields in India from 1961 to 2010.*