Survey of Various Machine Learning Algorithms for weather and crop yield prediction

Saamiya Newrekar, Student, Vidyalankar Institute of Technology, Mumbai, India
Kushagra Soni, Student, Vidyalankar Institute of Technology, Mumbai, India
Tanya Shrivastava, Student, Vidyalankar Institute of Technology, Mumbai, India
Prof. Kanchan Dhuri, Assistant Professor, Vidyalankar Institute of Technology, Mumbai, India

Abstract- Agriculture acts as the backbone for any developing country. In India, more than half of the employment opportunities are provided via agriculture. The crucial aspects that are of utmost significance to the farmers during the crop cultivation cycle are weather, yield, price, and soil condition. The aim of this survey paper is to provide a simple, yet informative application that includes all the required information like weather conditions, crop selection, crop scheduling, fertilizer application, best market rates, etc., i.e., providing a scrupulous and comprehensive guide to the farmers. Technology is an asset to today's farmers and by comparing various machine learning models like artificial neural networks, support vector regression, logistic regression, LASSO, random forest regression, k-nearest neighbors, etc., we aim to develop the most accurate and optimized model for predicting rainfall, humidity, temperature and crop yield so as to help the farmers make informed judgements and reduce losses and improve the overall quality of agriculture and farming for the farmers of today.

Index Terms – forecasting, machine learning, prediction, support vector machine, multiple linear regression, artificial neural networks.

I. INTRODUCTION

Agriculture plays a crucial part in a developing nation like India. There are different methods to obtain weather, yield or price predictions, including prediction models, private websites, and government data. But most farmers in India, either do not have the means to access such information or are not technically sound to be able to extract such information. Even though all the required information can be extracted and compiled from multiple sources, there is no unified platform that can provide weather, yield, and price predictions, along with crop scheduling information for a given district and crop, at one place.

Our research aims to tackle this problem at a micro level by using weather data from Indian Meteorological Department and studying very models that have been previously developed for predicting rainfall, humidity, temperature and crop yield and developing the most optimum model for improving the livelihood of farmers and agriculture today.

Weather prediction has always been a major concern for meteorologists because of its non-linearity. Predicting rainfall accurately does not only affect agriculture but can also help prevent the serious damage caused by natural disasters like landslides and floods. Knowing these occurrences in advance can prepare the areas to be evacuated, preventive measures to be taken and financial losses to be minimized. Rainfall prediction proves to be difficult because of the dynamic nature of the atmospheric processes affecting it.

Other factors affecting the weather of a region is temperature and humidity. Temperature is a measure of the degree of hotness or coldness in an area. Humidity is the amount or quantity of water vapor present in the air in a region. Weather affects the agriculture produce greatly. Each crop has an ideal weather condition associated with it and by predicting the weather beforehand we can also predict the crop that would flourish in a given climate- thus maximizing the yield that the farmer can produce.

Traditionally, weather predictions involved using numerical methods and solving the equations for the various atmospheric parameters. This was possible due to the atmosphere being fluid in nature and solving these equations of fluid dynamics using partial differentiation and thermodynamics the next state of the fluid could be predicted. This could hardly be consistent and accurate due to the chaotic nature of the weather. These equations could not only be completely solved but the time needed to solve them was so drastic, it would lead to more errors due to the weather changing drastically in the same time frame. However, due to the recent advancements in the field of technology, we do not need to rely on such time-consuming methods to make predictions. Machine learning is defined as the use and development of computer systems that have the

 DOI : 10.35291/2454-9150.2021.0556

ability to learn and adapt without needing to be explicitly instructed. Machine learning models use algorithms and statistical models to find hidden patterns in the historical dataset and draw inferences from these patterns. Machine learning is a subset of artificial intelligence that would allow the model to become more accurate with time and by fine tuning them without having to explicitly being programmed to do so. Thus, using technology we can build machine learning models to predict the weather in a short amount of time and the model will learn from the wrong predictions to improve accuracy with time and training.

Another module of our research is crop yield prediction. Knowing how many seeds are optimum for a given land area, how much yield is expected from a given sowing season- these factors can help reduce the burden on the farmers with respect to storing, selling, importing/exporting their agricultural produce and also help the farmers by minimizing their finances required for a given batch of produce. Buying fertilizers, insecticides, etc., in excess/dearth and overusing/underusing them on the crop can affect the final harvest. Thus, keeping these parameters in mind we will predict the most optimum crop that should be sown in a given land area, during which month is it best to sow these seeds, how much fertilizer should be used along with all the other factors affecting the yield along with the weather conditions during the sowing and harvesting period.

South India is most affected with climatic changes due to the drastic, chaotic and extreme weather processes occurring in the region. For our research, we have focused on southern states of India- Kerala, Andhra Pradesh, Karnataka and Tamil Nadu. We will build the machine learning model for one state and expand it further for the rest and eventually optimizing it for a nation-wide level.

II. LITERATURE SURVEY

The four core modules of our research are- rainfall, temperature, humidity and crop yield prediction models. For this purpose, we have studied previous works done in all of these models and we have reported their findings below.

In [1] the author compares artificial neural networks (NN) using radial basis function(RBF) and support vector machines(SVM) for rainfall predictions. The paper also discusses using SVMs along with other techniques like PSO. RBF NN and wavelet SVM gave remarkable results. In conclusion, the author remarks that any half breed approach using NN would give, good accurate results for rainfall prediction.

The author in [2] studies and examines artificial neural networks such as feed forward neural network(FFNN) for predicting rainfall dur to its soft computing and adaptability. An accuracy of 93.55% was achieved using FFNN. The performance of the model was evaluated using confusion matrix. Root mean square error(RMSE) of less than 0.5 means that the said model is good and acceptable. The RMSE value for the FFNN was found to be 0.254 which is good.

The research in [3] aims to study the weather conditions to support crop, water and flood management for both the safety of the region during natural disasters and in assisting farmers to minimize their losses due to unforeseen management. The paper focuses on random forest (RF), support vector regression(SVR) and decision trees(DT). The study compares the above models. RF outperforms SVR and DT. Further, each of the model is fine-tuned with respect to the kernel functions. In conclusion, RF is proven to be the best among the three with an adjusted r-squared value of 0.98. For the SVM model, radial basis kernel function tends to give better results than polynomial function SVM.

In [4] compares artificial neural networks(ANN) and LASSO regression models. The author also explains why models like ARIMA are not a good fit for season wise real time rain predictions. After comparing the results, LASSO model is said to be more accurate with an accuracy of 94% compared to ANN with an accuracy of 77%

The author in [5] compares decision trees(DT), logistic regression, k-nearest neighbors(KNN), random forest(RF) for predicting rainfall. The models are evaluated on the basis of accuracy, area under the curve, recall, precision, confusion matrix, stratified k-fold and their F1 score. Logistic regression worked best with undersampled data and KNN worked best with oversampled data. The paper concludes that the input has a very important role to play in deciding the best model for rainfall prediction. Under future scope, they wish to implement deep learning models like multilayer preceptors, neural networks, etc.

In [6] a low cost, effective model is proposed for predicting rainfall for farmer. The model proposed is random forest classification. An accuracy of 87.9% was achieved and the model was evaluated using the confusion matrix. Real time data is used along with historical data. For real time analysis, raspberry pi board is used along with barometric pressure sensor.

The author in [7] compares support vector machine(SVM), random forest(RF), neural networks(NN) algorithms for crop yield predictions. Leave One Out K-fold validation technique is used to tune the parameters in each of these models. The average error on a weekly basis is shown to be 3.14% while the maximum error seen is 9.66% in the prediction of the crop yield.

The author in [8] uses three varying coefficient regression model in predicting the crop yield. From the initial model, the parameters which have a negative or high-penalty or non-significant impact on the outcome are dropped and the model is reduced. The fourth reduced model is the finalized model for this paper in which out of 45 predictors, they have reduced it to 11 predictors using selected aggregate

 DOI : 10.35291/2454-9150.2021.0556

variables. The mean absolute percentage error is seen to be 0.74.

In the paper [9], decision tree algorithm is used for crop prediction. The rules for the model are based on soybean crop in the district of Dewas and the extend of cloud cover in mm keeping rainfall and temperature as affecting parameters. For the soybean crop, an accuracy of 87% was achieved.

Four machine learning models are compared in [10] namely- random forest regression, support vector regression, k-nearest neighbors and L2 linear regression. Random forest regression performed best with its default settings followed by k-nearest neighbors, L2 regression with polynomial features and lastly support vector regression using radial basis function.

In [11] compares the simple linear regression model to k-nearest neighbors(KNN) for crop yield prediction and the linear regression model outperforms the later with an accuracy of 95% as opposed to the 86% of KNN model.

The author [12] analyzes the multiple linear regression model along with density based clustering technique for crop prediction for Godavari district in Andhra Pradesh. The results achieved using the multiple linear regression model ranges between -14% to +13% for a 40-year period.

The paper [13] discusses the use of a multidimensional model and OLAP for cropping systems, fertilizer consumption and simplifying the process for farmers. It also is effective in using weather and meteorological data for analyzing its effect on agriculture.

The author [14] compares various models like k-nearest neighbors(KNN), k-means clustering, artificial neural networks, multiple linear regression(MLR) and support vector machines on four parameters- year, rainfall, area of sowing, and production. Out of these MLR performed the best with an accuracy of 98%, k-means following closely behind with an accuracy of 96%.

In the paper [15], PAM(Partition Around Medoids), CLARA(Cloud Learning Autonomous Reacting Algorithms), DBSCAN(Density based spatial clustering of applications with noise) and MLR models are compared on parameters like year, district, area in hectares, average temperature, soil type, soil pH value, minimum rainfall required and minimum temperature required. Out of all the models, DBSCAN and CLARA perform approximately at the same level and PAM gives the worst of the clustering results.

ARIMA and LSTM are compared in the paper [16] for humidity prediction. ARIMA is a better algorithm for predicting humidity because LSTM due to its very complex neural network is more accustomed for time series analysis.

FP growth algorithm is discussed in the paper [17]. The algorithm proving to be effect for classifying weather parameters like humidity. FP growth is an effective, scalable technique used in data mining used for storing information about frequent patterns.

In [18] uses k-means clustering model for predicting humidity. Clustering being a good technique to train the model to find hidden trends in the data. The paper affirms the use-case of k-means algorithm. It can perform very well in many situations where predictions like that of humidity are required, despite being a simple algorithm.

To summarize our survey, we developed a better understanding of which model performs better under which circumstances, for example- oversampled and undersampled data yields different results because the input to the model is varied. We understood how the non-linearity of climatic parameters affect the model in question and SVM and DT are not a good fit for weather predictions. And hence, models such as MLR, RF regression and ANN are appropriate and have proven to yield results with great accuracy for weather prediction. For crop yield prediction, SVM has consistently proven to be accurate and a good fit. All in all, the models to be finalized will be decided after fine tuning the hyperparameters and comparing all the models in question with one another.

III. PROBLEM STATEMENT

To develop a simple, easy to navigate, multilingual application for farmers that would predict the rainfall, temperature and humidity in a given region, in a given period of time, along with predicting the crop yield in a given acre of land so as to optimize the agriculture industry by using a digitalized approach. The application would include insurance policies along with other essential information that could assist the farmers by avoiding losses and opportunities and improving their livelihood.

IV. PROPOSED METHODOLOGY

Machine learning by definition is used to train a model on a lot of data and used these trained models to make predictions on unseen, new data. Hence, to make sure the model proposed works effectively, many steps are to be followed so as to avoid cases like overfitting or underfitting, and to improve the overall accuracy of the said model.

The methodology that we have proposed consists of the following steps-

1. Data Collection
2. Data Preprocessing
3. Developing the Machine Learning Model
4. Performance Evaluation
5. Application Interface

A. Data Collection

South India is known for facing calamities due to rainfall and extreme weather conditions. The dataset has been collected from Indian Meteorological Department(IMD) and world

 DOI : 10.35291/2454-9150.2021.0556

weather online [19]. We have targeted the states of South India for our study, namely- Andhra Pradesh, Tamil Nadu, Kerala and Karnataka. The dataset for each of the states consists of approximately 90,000 rows and 25 columns. Each column is a feature describing the dataset.

B. Data Preprocessing

The real-world data scraped from the web is incomplete, inconsistent, noisy and consists of null values which affect the accuracy of the prediction model. To make the dataset uniform, and to improve the performance of the model, the data needs to be standardized and cleaned using various data preprocessing [20] techniques. Data preprocessing includes- data cleaning, data transformation and data reduction.

Data cleaning is used to first detect and then remove corrupt or inaccurate data from the dataset. Missing or null values are computed using the mean/median/mode of the dataset or the tuples having null values are simply removed.

The data extracted from the internet is inconsistent depending on the various sources or databases used. To deal with this, data transformation is used as inconsistent data can result in faulty predictions. Techniques like standardization and normalization are used to transform a dataset with a wide range of values to a shorter, easily computable range.

Large datasets can be hard to deal with and processing them could be time consuming and require higher computational power. Querying such data or working on machine learning models with a huge dataset can result in undesirable outputs. Thus, the dataset is reduced in size by selecting the most relevant features affecting the output and dropping the ones that don't contribute to the output as much. Techniques like feature extraction, dimension reduction, etc. are used for data reduction.

C. Developing the Machine Learning Models

Our research consists of crop yield prediction model and weather prediction models like rainfall prediction, humidity prediction, and temperature prediction. Studying various papers previously published in these areas, we have decided to use support vector regression [21] for crop yield prediction and multiple linear regression(MLR) and artificial neural networks(ANN) for weather prediction. Both MLR [22] and ANN [23] have previously been used and studied for weather prediction models and have performed well. In certain cases, MLR performed better than ANN and in certain cases, ANN outperformed MLR. The reason for suggesting two models for weather prediction is that many of the cases studies have concluded that each machine learning model performs differently based on the dataset that is inputted in the model and thus we want to compare the performances of these two models on our dataset and see which one yields better results.

D. Performance Evaluation

We propose to evaluated our model using mean square error(MSE) [24] and R-squared [25] techniques. And depending which of the two MLR and ANN models performs better in case of each of the three weather prediction models, we will choose that model for that prediction.

E. Application Interface

The application's front-end is proposed to built using flutter, making it cross-platform so both android and IOS users can make use of it. It will be a simple UI with multilingual support and various tabs for insurance policies, YouTube videos, news, shop for buying and selling agricultural supplies, etc. The profile of the users will include details like the name of the users, phone number, size of land, crop details if any is available. Location feature will be enabled at all times so as to give live weather updates and predictions. There will be a feedback mechanism which will enable the models to learn from more data and give better accuracy for future predictions.

V. CONCLUSION

In conclusion, the papers we have reviewed have been enlightening and fundamental in our research project. Agro being an application which is machine learning intensive is going to be a helping hand to farmers, starting from the states of South India. For weather forecasting models which include rainfall prediction, humidity prediction and temperature prediction we have proposed to implement multiple linear regression(MLR) and artificial neural networks(ANN). To predict the crop yield, we are going to implement support vector regression(SVR). The reasons for choosing these models is the literature survey- confirming their accuracy and relevancy to our output parameters.

REFERENCES

[1] Naveen I. Mohan H. S., 2019, "Atmospheric Weather Prediction Using various machine learning Techniques: A Survey", Proceedings of the Third International Conference on Computing Methodologies and Communication, IEEE Xplore.

[2] A. Kala, Dr. S. Ganesh Vaidyanathan, 2018, "Prediction of Rainfall using Artificial Neural Networks", Proceedings of the International Conference on Invention Research in Computing Applications, IEEE Xplore.

[3] Tharun V. P., Ramya Prakash, S. Renuga Devi, 2018, "Prediction of Rainfall using Data Mining Techniques", Proceedings of the 2nd International Conference on Invention Communication and Computing Technologies, IEEE Xplore.

[4] Kaushik Dutta, Gouthaman P, May 2020, "Rainfall Prediction using Machine Learning and Neural Network", Internation Journal of Recent Technology and Engineering

 DOI : 10.35291/2454-9150.2021.0556

[5] Nikhil Oswal, October 2019, "Predicting Rainfall using Machine Learning Techniques", ResearchGate.

[6] Nitin Singh, Saurabh Chaturvedi, Shamim Akhter, 2019, "Weather Forecasting using Machine Learning Algorithm, Proceedings of the International Conference on Signal Processing and Communication, IEEE Xplore.

[7] Ranjini B. Guruprasad, Kumar Saurav, Sukanya Randhawa, 2019, "Machine Learning Methodologies for Paddy Yield Estimation in India: A Case Study", Proceedings of the IEEE International Geoscience and Remote Sensing Symposium, IEEE Xplore.

[8] E. Manjula, S. Djodiltachoumy, March 2017, "A Model for Prediction of Crop Yield", International Journal of Computational Intelligence and Informatics.

[9] S. Veenadhari, Bharat Misra, CD Singh, October 2014, "Machine learning approach for forecasting crop yield based on climatic parameters", Proceedings of the International Conference on Computer Communication and Informatics. IEEE Xplore.

[10] Neha Rale, Raxitkumar Solanki, Doina Bein, March 2019, "Prediction of Crop Cultivation", Proceedings of the IEEE 9th Annual Computing and Communication Workshop and Conference, IEEE Xplore.

[11] B. Devika, B. Ananthi, December 2018, "ANALYSIS OF CROP YIELD PREDICTION USING DATA MINING TECHNIQUE TO PREDICT ANNUAL YIELD OF MAJOR CROPS" , International Research Journal of Engineering and Technology.

[12] D. Ramesh, B Vishnu Vardhan, January 2015, "ANALYSIS OF CROP YIELD PREDICTION USING DATA MINING TECHNIQUES", International Journal of Research in Engineering and Technology.

[13] Constanta Zoie Radulescu, Marius Radulescu, Adrian Turek Rahoveanu, March 23 - 25, 2009, "A Multidimensional Data Model and OLAP Analysis for Agricultural Production", Proceedings of the 10th WSEAS Int. Conference on Mathematics and Computers In Business And Economics

[14] Ramesh D, Vishnu Vardhan B., 2013, "Data mining techniques and applications to agricultural yield data", International journal of advanced research in computer and communication engineering.

[15] Jharna Majumdar, Sneha Naraseeyappa and Shilpa Ankalaki, July 2017, "Analysis of agriculture data using data mining techniques: application of big data", Journal of Big Data 2017.

[16] ZhiQiang Li, HongXia Zou, Bin Qi, 2019, "Application of ARIMA and LSTM in Relative Humidity Prediction", IEEE 19th International Conference on Communication Technology

[17] Christy Kunjumon, Sreelekshmi S Nair, Deepa Rajan S, 2018, "Survey on Weather Forecasting Using Data Mining", . IEEE Conference on Emerging Devices and Smart Systems

[18] Badhiye S. S. , Dr. Chatur P. N., Wakode B. V., 2012, "Temperature and Humidity Data Analysis for Future Value Prediction using Clustering Technique: An Approach", International Journal of Emerging Technology and Advanced Engineering

[19] https://www.worldweatheronline.com/

[20] https://www.analyticsvidhya.com/blog/2021/08/data-preprocessing-in-data-mining-a-hands-on-guide/

[21] https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html

[22] https://www.analyticsvidhya.com/blog/2021/05/multiple-linear-regression-using-python-and-scikit-learn/

[23] https://en.wikipedia.org/wiki/Artificial_neural_network

[24] https://link.springer.com/referenceworkentry/10.1007%2F978-0-387-30164-8_528

[25] https://www.investopedia.com/terms/r/r-squared.asp

 DOI : 10.35291/2454-9150.2021.0556