Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 13 Issue 9, September 2024

Crop Prediction System From the Weather Data using Machine Learning

Bhargav Shivbhakta
Dept. of Electronics and Computer Engineering
PES Modern College of Engineering
Pune, India

Anusha Karve
Dept. of Electronics and Computer Engineering
PES Modern College of Engineering
Pune, India

Aryan Humnabadkar
Dept. of Electronics and Computer Engineering
PES Modern College of Engineering
Pune, India

Mrs. S. V. Hon
Dept. of Electronics And Computer Engineering
PES Modern College Of Engineering
Pune, India

***Abstract*— Precise crop prediction is essential for maximizing agricultural efficiency and guaranteeing food stability. This work investigates the implementation of a crop prediction system that utilizes meteorological and soil data and employs machine learning algorithms. The crop recommendation dataset obtained from the Kaggle contains nitrogen, phosphorous, potassium, temperature, humidity, pH, and rainfall. Twenty different crops were considered for the label. For this approach, the SVM, KNN, DT, and GB algorithms were trained on different sets of crops, with 20 classes for proper training and validation. Data preparation also included cleaning and normalizing the data set, dealing with missing values, and encoding category attributes. We evaluated the algorithms on F1-score, recall score, accuracy, and precision. Results showed that Gradient Boosting and KNN (K=3) are the best estimation algorithms for crop prediction. The data will be useful for the farmers and agricultural strategists by providing them with decision-making tools based on weather conditions and soil characteristics, helping to improve crop management and productivity. Additional research will be conducted to scale up the dataset, incorporate other related environmental factors, and improve the model's accuracy for practical use.**

***Keywords*—Crop prediction, machine learning, weather data,**

I. INTRODUCTION

Food, basic resources, and employment opportunities are all provided by agriculture in a variety of economies across the world. Agriculture plays a crucial part in providing these necessities. In this regard, an urgent need is to set up a vibrant management mechanism for agricultural processes to ensure economic growth and food security and attain sustainable development. In recent years, the agriculture sector has faced numerous challenges, from population expansion to land, water deficit, and soil erosion, besides the increasing intensity of climate change.

These challenges gave rise to the need for new approaches and effective solutions to these constraints, ultimately leading to higher output per unit area. Using historical weather data, sophisticated Machine Learning (ML) systems can accurately predict agricultural yield. This article will examine the types and levels of harvests as we attempt to explain the differences between crop forecasting from one season to another. Using exact harvest anticipating, ranchers can make informed decisions regarding edit choice, the best establishing time, and asset portion. It also helps legislators and agricultural strategists develop policies that make using resources and having food available easier. Standard ways of evaluating plant creation energetically depend upon human expertise and evident data, which are leaned to tendency and subjectivity. Contrastively to its counterpart, artificial intelligence is a process that requires careful data analysis in order for it to identify patterns and make accurate predictions.

With advancements in machine learning (ML), it is possible to handle tough nonlinearity and complex connections within agricultural data with strong tools. So here, Machine (ML) comes into the picture, which uses Large Datasets to find the pattern and help us get very accurate predictions. This work is on a crop prediction system that forecasts the best suitable crop to cultivate in such a climate. This will use data on the weather as well as ML algorithms.

This study used the Kaggle dataset for its analysis, which included various climatic and soil parameters such as temperature, humidity, pH value, and water availability by season. As such, these attributes should be considered as potential indicators of whether or not a particular crop may grow well in some regions. The overall goal of this study is to use different types of ML models to categorize crops, given the features provided. The approaches chosen for this experiment are Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Tree (DT), and Gradient Boosting. Because of their specificity and benefits, such models can be used in various datasets with prediction tasks. Grounded in previous work, this study compares the performance of different crop forecasting systems to determine an optimal type.

Multiple contributions of this research will help disciplines like agricultural data science. In the first half of this work, we examine several machine-learning models for crop prediction and discuss some positive or negative items regarding each strategy. Second, it discovers the best model to predict crops due to weather and soil data, helping farmers and agricultural planners. Third, the study provides concrete and real-time forecasts of how ML might enhance agricultural yield and sustainability.

These are significant implications that come from this research. These could help farmers optimize their resources,

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 13 Issue 9, September 2024

reduce waste, and increase crop yields by identifying which crops would grow well under which conditions. This may enhance meal safety and economic results in agricultural communities. In addition, applying machine learning to agriculture might evoke some creativity rather than deploying traditional crop management practices in bigger numbers, and more frequently, new instruments or methods may reveal themselves.

Nonetheless, there are some limitations and troubles to take into account. One big challenge is good data and enough of it. In order for ML models to be able to make predictions, they need to learn from robust and high-quality datasets. The performance of the models might have been constrained by low or missing agricultural data in some areas. Additionally, detecting masks may result in lower generalization and more unfit models across regions due to different edaphic-climatic conditions. Validation is key to ensuring that the models are robust and practical for almost all datasets.

II. LITERATURE SURVEY

This study must implement the best crop watering system that Muangprathub et al. referred to, which builds upon a wireless sensor network, which consequently analyses monitored data to estimate the water and moisture level of their surrounding environment [1] and temperature. Thorat et al. [2] Monitor environmental parameters and detect leaf diseases. IoT cannot, under any stretch of feasibility, suggest the best crops to grow without weather forecast factors. ML is a data analysis technique that uses weather data to model new yet accurate analytical models [3], which helps predict or anticipate future conditions and brings forth an effective forecasting solution. Furthermore, for a given problem to which an ML model applies, since it is intelligent enough, recent or previous data can be analyzed with these algorithms, and suggestions could become available.

Patil et al. [4] proposed the use of semi-supervised learning with a Q-learning algorithm for identifying key parameters like pH, MC (moisture content), T (temperature), and RH(humidity). It also helps in the analysis and prediction part based on soil characteristics like pH, phosphate (P2O5), and potassium level to give you better insights into the fertilizer quantity required when using. Girish ([5]) tried various ML techniques, i.e., Linear Regression, Support Vector Machine (SVM), K Nearest Neighbor (KNN) technique, and Decision Tree algorithm to forecast the output variables using a predictive analysis. Farmers use this data to choose crops that fit their historical rainfall patterns and market prices. The work done by Nischitha et al. [6] has suggested a method that would help to find the best crop suitable for that land by good extent and exactness. However, this method is fully based on the nature and structure of the soil and depends upon environmental factors like temperature, humidity, pH value, and rain. Both SVM and Decision Trees are two standard ML approaches for predictive analysis.

According to the study's author, farmers may utilize data mining and ML techniques to forecast agricultural productivity and make informed decisions about crop selection [7]. Data mining is discovering a new pattern within a large amount of data. This pattern is used to compute agricultural output and information that assists farmers in selecting crops based on several easily available attributes. The author used linear regression to make predictions about the data.

According to the author of [8], several factors, including temperature, land productivity, water quantity and quality, seasons, and produce prices, affect agricultural forecasting. ML can accurately estimate agricultural outputs with location, weather information, and season. It assists farmers in cultivating crops most appropriate for the regions in which they operate. The author assessed the forecasting system using historical data and ML methods, including SVM, RF, and Iterative Dichotomize 3. The SVM algorithm yields the most precise outcomes.

In [9], the researchers utilized the meta-learning approach to forecast the values of various significant commodities over some time. This mix of a long short-term model and a self-organized map is used to train crop prices and agricultural production databases. For instance, this combination can be used in adaptive crop price prediction based on ML. The trials' findings demonstrate room for significant improvement in the degree of cross-correlation entropy and precision of the current agricultural price prediction systems.

III. PROPOSED SYSTEM

The block diagram of the proposed system is shown in Fig.3.1.

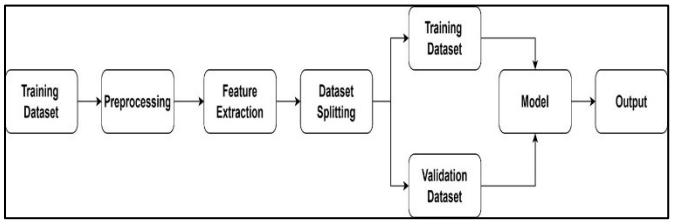

Fig. 1. Block diagram of the proposed system

A. Input Dataset

This information is intended primarily for agricultural planning and crop yield forecasting. By analyzing the relationships between environmental factors (temperature, humidity, and rainfall), soil parameters (Nitrogen, phosphorous, and potassium), and the type of crop grown, data scientists and agricultural experts can develop models and algorithms to predict crop yields accurately. This predictive capability can contribute to better resource management, informed decision-making for planting and harvesting, and the development of sustainable agricultural practices. The dataset comprises the following columns:

- Nitrogen (N): Ratio of the Nitrogen in the soil
- Phosphorous (P): Amount of Phosphorous in the soil
- Potassium (K): Amount of Potassium in the soil
- Temperature: The region's average temperature expressed in degrees Celsius, reflecting the local climate.
- Humidity: The percentage representation of the relative humidity level, which indicates the relative moisture content of the air.
- pH: The soil's pH value indicates how acidic or alkaline it is.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 13 Issue 9, September 2024

- Rainfall: The amount of rainfall in mm.
- Label: The crop label that describes the kind of crop being grown. The dataset contains information on various crops, such as rice, maize, chickpea, kidney beans, pigeon peas, moth beans, mungbean, black gram, lentil, etc.

B. Preprocessing

ML project must include data preparation, cleaning, and preparing the data for analysis. Data preprocessing in this study included encoding categorical features, standardizing the data, and addressing missing values. In order to get accurate predictions, ML models must acquire high-quality input data, which can only be achieved through proper preprocessing.

C. Dataset Splitting

The dataset is divided into different subsets to facilitate training and testing. Training is assigned 80% of the data, while testing is reserved 20%. This component ensures that a substantial percentage of the dataset is utilized for training the ML algorithms while a separate subset is used for evaluation. This allows for an unbiased evaluation of the model's performance.

D. Training and classification

The trained ML algorithms have the recovered information as the input. Different classification methods, like DT, GB, KNN, and SVM, are applied to the data, discovering patterns or correlations in order. The main task of these algorithms is to recommend crops based on weather and soil data.

a. Support Vector Machines (SVM): Gradient Boosting is a powerful, supervised ML technique for regression and classification problems. With all its parameters, SVMs can tackle the most complex problems with data containing many dimensions. SVM efficiently does this by discovering the hyperplane that defines a decision boundary in a multi-dimensional input space and separating distinct classes. A hyperplane can be considered a discriminant that demarcates the feature space into solid partitions corresponding to classes. The idea of SVM is to find the hyperplane with some margin, i.e., a minimal distance between parallel hyperplanes and data points from each class. Support vectors are the data points that are positioned closest to the hyperplane.

- Hyperplane: A hyperplane is a straight line that separates the two classes for binary classification problems.
- Margin: It is the distance between the support vectors and the hyperplane. That is to say, as we know, in SVM, the hyperplane separates two classes and gets the maximum margin.
- The nearest ones to the plane are classified as support vectors, affecting the determination of a division line. Well, the orientation and position of the hyperplane are decided most by these sites, where a specific category lies or not.
- SVM uses the kernel trick technique to transform input data into higher-dimension feature space. It can use a kernel function to calculate the dot products between data points in higher-dimensional space without transforming the features explicitly. SVM is good at classifying concerning a linear boundary, but it can also use this same trick for nonlinear decision boundaries.
- SVM can handle non-linearly separable data through a soft margin, allowing some misclassification. A slack variable is given to do so, permitting some data points even on the wrong side of the hyperplane or within the margin.
- Training and Classification: You will require an input feature-tagged dataset with the corresponding target labels to train an SVM model. The SVM approach employs various optimization techniques to find the optimal hyperplane, which maximizes this margin. After the model has been trained, it is used to help us classify new data items by predicting their class labels.

SVMs also have several advantages: they can deal with high dimensional input data, are very good at handling complicated decision boundaries, and avoid overfitting essentially. SVMs are particularly cost-intensive in computation with high dimensional data, and sometimes, it is complex to choose the correct kernel function along with hyperparameters. Selecting these settings wisely and configuring them based on the use case for optimal performance is important. Furthermore, SVMs can be scaled to solve multi-class Classification problems using methods like one-vs-all or One-vs-One and binary classification. In addition to classification problems, SVMs can be used for regression tasks to predict a real value instead of a categorical class.

b. KNN*:* This is a common method in supervised ML for classification and regression tasks. K-nearest neighbors (k-NN) is a non-parametric method with no explicit assumption for the data distribution. This breaks what KNN operates down to in the simplest form: group or predict new data points on a majority vote (or average) of its k-nearest neighbors from all other training examples in feature space. k of KNN (No of neighbors to be taken into account for prediction is user-defined parameter)

- Training Phase: In this phase, the K-nearest neighbors(KNN) Model saves the p–p-dimensional feature vectors and corresponding class labels or target values from the labeled training set.
- When an unseen data point in a novel requires prediction or classifying, the algorithm calculates the distance between this new unannotated point and all other points from the training dataset. It commonly uses Euclidean distance to determine the similarity between data points, but it can also work with other metrics such as Manhattan or Minkowski distance.
- Select Neighbours: Determine the k nearest neighbor to the new data point based on distances (calculated above). Out of them, k data points from the training dataset closest to the new point will be chosen.
- The new data point is assigned to a category either using the average technique (for regression) or the majority vote method (classification). This is achieved by checking from k nearest neighbors which class label occurs most frequently. We can predict the target value of a new data point for regression problems by taking the average over all k (which can be tuned) closest neighbors.
- Prediction: now that the learning process is over, a data set we have not seen before {xi} inbounded in X needs to be submitted for evaluation (also known as prophesying), which would result in an associated class label c or predictor estimate y.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 13 Issue 9, September 2024

KNN may be effective for some dataset types and is simple to comprehend and apply. However, with high-dimensional data, its performance may suffer, and it might be sensitive to irrelevant or noisy aspects. Having a large amount of training data is crucial for making accurate predictions.

c. The decision tree is a widely used supervised machine-learning technique for regression and classification tasks. The result is a structure that resembles a flowchart. In this structure, each internal node represents a feature, each branch offers an alternative based on that information, and each leaf node represents the outcome or forecast. The decision tree method recursively divides the incoming data according to the features to produce a structure like a tree. The objective is to partition the data to optimize the target variable's homogeneity or purity within each partition. There are several algorithms and criteria for constructing decision trees, but the " CART " algorithm is the most commonly used one (Classification and Regression Trees). Below is a summary of the general operation of the decision tree algorithm:

- Feature Selection: The algorithm chooses the most useful feature to partition the data. It assesses several attributes according to standards like knowledge gain and Gini impurity. The objective is to identify the characteristic that most effectively separates the groups or minimizes the variability in the target variable.
- Partitioning: After identifying the most favorable characteristic, the dataset is divided into smaller groups according to the potential values of that characteristic. The splitting procedure is applied recursively to each subset, which results in each subset becoming a child node of the current node.
- The recursive splitting method continues until a halting criterion is met. A minimal number of samples in a leaf node, a maximum depth, or the point at which further splitting does not appreciably enhance performance could all be considered examples of this criterion.
- Formation of Leaf Nodes: Once the halting condition is satisfied, the decision tree generates its final nodes, also known as leaf nodes. Every terminal node indicates a forecasted category or numerical value in the context of regression.
- Prediction: The input is propagated through the decision tree, following the path of features and decisions until it reaches a leaf node to make predictions on fresh, unknown data. The decision tree's output is the prediction at that leaf node.

These issues are commonly addressed by ensemble approaches such as Random Forests, which incorporate many decision trees to improve prediction performance and generalization, or by boosting algorithms like Gradient Boosting Machines.

d. Gradient Boosting*:* Boosting is a powerful ML technique in which a single or more weak learners, usually decision trees, can be aggregated into a strong prediction model. New models are fitted on the residuals of older ones by an ensemble method to give an additive model. In general, the algorithm of gradient boosting can be expressed as:

- Train a weak (base) model: A simple or base-generalized ML model is first trained from the training data. Weak learners are often composed of decision trees with shallow depth.
- Residuals: These are the errors of the base model; they can be calculated by subtracting the predicted value from the actual target.
- Model: The residuals over time are used as input to train a new model (often the decision tree again). The underlying idea is to train this model to predict the discrepancies and identify some patterns of information that had been beyond driving models' comprehension.
- Update the model: The new model is added to your ensemble by blending it with current models (Training). To achieve this, models can be weighted based on performance or their contribution.
- Iterative process: Steps 2-4 are done iteratively, with each new model focusing on residuals of models in the previous ensemble. The ensemble is given more additional models, and the weights of all those models are trained to minimize joint prediction errors.
- Migration Transition of Flocks: The flocks forecast each ensemble model, and the summation gives the final forecast. The contribution of each model is determined by a weighting factor that considers the performance and speed at which it learns; this determines how many future models use its influence to develop.

IV. RESULTS AND ANALYSIS

This section displays the performance metrics of suggested solutions in various testing environments. Various test cases are used to check the functioning of each application. The ML algorithm included the system with GB, KNN, DT, and SVM algorithms of historical data to identify the crop. Results of the F1-score, Recall, accuracy, and precision for several ML approaches are presented in TABLE I.

TABLE I. PERFORMANCE OF ML ALGORITHM FOR CROP SUGGESTION SYSTEM

Algorithm		Precision	Recall	F1-Score	Accuracy
SVM	Linear	0.98	0.97	0.97	0.98
	Rbf	0.99	0.98	0.98	0.98
	Polynomial	0.94	0.92	0.92	0.92
KNN	K=3	0.99	0.99	0.99	0.99
	K=5	0.98	0.98	0.98	0.98
	K=7	0.98	0.98	0.98	0.98
DT		0.98	0.97	0.97	0.97
GB		1.00	1.00	1.00	1.00

The performance metrics used are Precision, Recall, and F1-score for the classifiers chosen. The SVM method was tested on the Linear, RBF, and Polynomial kernels. The linear SVM model had a Precision of 0.98, Recall of 0.97, F1-Score of 0.97, and accuracy of 0.98. The RBF kernel exhibited superior

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 13 Issue 9, September 2024

performance to the Linear kernel, achieving a Precision of 0.99, Recall of 0.98, F1-Score of 0.98, and an identical accuracy of 0.98. The Polynomial kernel exhibited worse performance, achieving a Precision of 0.94, Recall of 0.92, F1-Score of 0.92, and accuracy of 0.92. The KNN algorithm was evaluated using several values of K, specifically 3, 5, and 7.

The achievement of the GB with a Perfect Score in all metrics (Precision, Recall, F1-Score, and Accuracy= 1.00) shows that considering it into the Crop Suggestion System improves performance among the algorithms studied for this work. They rigorously experimented and evaluated the project, which featured them with good results, as seen from how different algorithms performed on various performance metrics. The talk starts by giving an overview of the project's goals and methods, focusing on the significance of weather and soil parameters when robust models can be used for crop prediction. This series has stressed the importance of preprocessing and data availability to determine how your ML model will perform.

V. CONCLUSION

This paper introduces a Crop Prediction System that has been created and assessed. The system uses ML algorithms to categorize crops by leveraging meteorological and soil data. The system underwent training and testing using a Kaggle dataset that included important variables such as temperature, humidity, pH, water availability, and season. Various ML algorithms, including SVM, KNN, DT, and GB, were applied to the dataset and evaluated using Precision, Recall, F1-Score, and Accuracy metrics. The results indicate that GB outperformed all other algorithms, achieving perfect scores across all metrics. KNN with K=3 also demonstrated exceptional performance, closely following GB. The findings underscore the potential of ML in enhancing agricultural productivity by providing accurate crop predictions based on environmental conditions. The comparative analysis of different algorithms reveals that ensemble methods like Gradient Boosting are particularly effective for this task. This research contributes to the field by identifying the most suitable algorithms for crop prediction, thereby aiding farmers and agricultural planners in making informed decisions that optimize resource allocation and crop yield.

While the current study has achieved promising results, there are several avenues for future research to improve the Crop Prediction System further. Expanding the dataset to include additional environmental variables such as wind speed, precipitation, and soil type could enhance the model's predictive accuracy. Incorporating remote sensing data and satellite imagery could provide more comprehensive insights into crop health and growth conditions, enabling more precise predictions. Moreover, developing hybrid models that combine multiple ML techniques' strengths could improve performance. For instance, integrating deep learning models with traditional ML algorithms may help capture complex patterns in the data more effectively.

Additionally, real-time data processing capabilities could make the system more responsive to changing environmental conditions, allowing for dynamic crop management recommendations. Another important area of future research is the model's generalization across different geographical regions. Validating and refining the model with data from diverse climates and soil types will ensure its applicability on a global scale. Finally, exploring the integration of this system into existing agricultural management platforms could provide farmers with actionable insights, making it easier to apply these predictions in practical settings. The potential for ML to transform agriculture and enhance food security can be fully realized by continuing to refine and expand this research.

REFERENCES

[1] Muangprathub, J., Boonnam, N., Kajornkasirata, S., Lekbangpong, N., Wanichsombat, A. and Nillaor, P. (2019) IoT and Agriculture Data Analysis for Smart Farm. Computers and Electronics in Agriculture, 156, 467-474. https://doi.org/10.1016/j.compag.2018.12.011

[2] Thorat, A., Kumari, S. and Valakunde, N.D. (2017) An IoT Based Smart Solution for Leaf Disease Detection. 2017 International Conference on Big Data, IoT and Data Science (BID), Pune, 20-22 December 2017, 193-198. https://doi.org/10.1109/BID.2017.8336597

[3] Katarya, R., Raturi, A., Mehndiratta, A. and Thapper, A. (2020) Impact of Machine Learning Techniques in Precision Agriculture. 2020 3rd International Conference on Emerging Technologies in Computer Engineering: Machine Learning and Internet of Things (ICETCE), Jaipur, 7-8 February 2020, 1-6 https://doi.org/10.1109/BID.2017.8336597

[4] Patil, B., Maske, R., Nakhate, R., Nikam, R. and Javheri, P.S. (2020) Soil Fertility Detection and Plant Prediction Using IoT and Machine Learning Approach. International Journal of Innovative Research in Science, Engineering and Technology, 9, 3329-3334.

[5] Girish, L., Gangadhar, S., Bharath, TR, Balaji, KS and Abhishek, K.T. (2018) Crop Yield and Rainfall Prediction in Tumakuru District Using Machine Learning. National Conference on Technology for Rural Development, 61-65.

[6] Nischitha, K. and Dhanush, V. (2020) Crop Prediction Using Machine Learning Approaches. International Journal of Engineering Research and Technology, 9, 23-26.

[7] Ms. Fathima, Ms. Sowmya K, Ms. Sunita Barker, Dr. Sanjeev Kulkarni(2020), "Analysis of Crop yield Prediction using Data Mining Technique" International Research Journal of Engineering and Technology (IRJET), Volume: 07 Issue: 05.

[8] A. Gonzalez-Sanchez, J. Frausto-Solis, and W. Ojeda-Bustamante(2014): "Predictive ability of machine learning methods for massive crop yield prediction," Spanish Journal of Agricultural Research, vol. 12, no. 2, pp. 313–328.

[9] D. K, R. M, S. V, P. N, and I. A. Jayaraj(2021)," Meta-Learning Based Adaptive Crop Price Prediction for Agriculture Application," in 2021 IEEE, 5th International Conference on Electronics, Communication, and Aerospace Technology (ICECA), pp. 396-402, DOI: 10.1109/ICECA52323.2021.9675891.