Using Machine Learning Techniques, a New Method for Predicting Crop Yield

Rashi Tanwar[1], Dr. Kamal Malik[2], Dr. Yogesh Chhabra[3]
[1]Computer Science Department, CT University Ludhiana, India Email: tanwar5390@gmail
[2]Computer Sciences Department, CT University Ludhiana, India Email: kamal17203@ctuniversity.in
[3]Computer Sciences Department, CT University Ludhiana, India Email: dryogeshchhabra@ctuniversity.in

Abstract
Agribusiness is regarded as an essential industry worldwide since there are so many problems to handle while evaluating crops based on environmental factors. This is putting the agricultural nations to the test. Using the most recent innovations, many businesses are reducing manual labor by using IOT-based services and mechanical technology. The majority of the time, such tactics are useful for situations involving reduced physical labor, but not to the extent that was anticipated. In this work, the highest yield is predicted using the most recent machine learning (ML) innovations and KNN grouping computation. Crop production is expected to depend on topsoil and climate parameters. In general, factual models were used to guide the yield analysis and village regeneration projections. However, these quantifiable models have become questionable due to the drastically changing global environment. It becomes acceptable that we turn to alternative, less complicated techniques going forward.
Keyword -Machine Learning, IOT, Crop Yield, ANN, KNN, Classification, Clustering, Random forest, SVM

1. Introduction

Some of the nations that are dependent on agriculture are in India. Agriculture in India is dependent on imports and market prices. Agribusiness is important to the Indian economy. The crop yield has been challenging, and the economic situation has fallen dramatically [7,5,22]. India produces rice, wheat, pulses, and other crops. India is increasing harvest productivity to serve its people, which is dependent on agriculture [13]. The use of AI computations may be the best way to predict elusive traits.
This refers to achieving goals with contemporary technologies. Numerous algorithms calculate crop production. Among these, K-Nearest Neighbor (K-NN) is one that may be quickly put into practice on a low budget. We must determine the frequency of recurring similarities among neighbors and anticipate the outcome using the K-nearest neighbors (KNN) obtained based on prior outcomes.

2. Related Work

2.1. Prediction of efficient crop yield using machine learning algorithms:

To get a cost-effective prediction of crop output, we frequently use descriptive studies in the sector of the agricultural industry for sugar cane farming. You may enjoy three records inside this article: the soil history, the rainfall observations, and the yield record. They merged the dataset's information, used a variety of monitoring models to encourage calculable actual costs, and obtained accurate results. They employed algorithms like K-Neighbor, Support Vector Machine, and Least Square Support Vector Machine to get the desired result [2,6,9,14]. They examined the accuracy results from the aforementioned methods and also found the algorithm's mistakes. The suggested algorithm should be more accurate and have a low rate of mistakes. They assigned the crop yield a rating of "low," "medium," or "high." The paper includes an idea for applying the idea of descriptive analysis to the world of agriculture. Information on how the analysis of the data will be used to create data sets for sugar crops is provided in the study article. The soil record, precipitation record, and yield record are three separate records. These datasets include many criteria that aid in determining the condition of the crops and categorizing the information by using monitored instructions on a dataset from the agricultural sector. Both categorization and regression may be carried out using this approach. While the regression stage determines the specific value of the yield, the classification stage divides the data into three categories (low, medium, and high). We frequently use the three primary supervised learning algorithms, KNN, SVM, and LS-SVM, to train and construct a model [1,9,12,16,23]. This position is not domain-specific. Create a system for a certain industry, such as medical, product comparison, retail, etc. The records must be consistent, but we usually just need to pass them through this system. This analysis will advance to a higher degree. We will construct an agricultural production and marketing recommendation system for farmers that enables producers to make a choice regarding which crops to plant throughout the season in order to increase their profit. This method is effective with organised knowledge sets [3].

2.2 Prediction of Agricultural Crops using the KNN Algorithm

Due to the weather, rainfall, type of soil, and various other influencing elements, agriculture is subject to uncertainty.
Crop forecasting is dubious due to the enormous datasets that make it challenging for farmers to draw conclusions based on them without an accurate system. To fulfil the requirements of the populace, output rates must rise as the population does. These issues can be resolved using techniques for data extraction and machine learning. This system uses real-time agricultural information to estimate the crop, such as soil type, rainfall, and humidity [4,6,9,14]. Mangalore, Kodagu, Kasaragod, and a few more districts in the state of Karnataka were used as samples for this survey. They gathered historical information on the district's agricultural practices and level of production, used it to determine the preferred crop using a well-known algorithm, and discovered a productive method of harvesting. The data gathered from numerous sites is sorted into sets in a well-structured way. Numerous algorithms calculate crop production. Among these, K-Nearest Neighbor (K-NN) is one that may be quickly put into practice on a low budget. We must determine the frequency of recurrent similarities among neighbors and anticipate the outcome using the k-closest neighbors (KNN) determined based on prior results [1].
Since it was initially presented, the K-Nearest Neighbor rule (KNN) has become one of the most well-known supervised learning methods for pattern categorization. The complete training set is simply retained throughout acquiring by this rule, and each query is given a class based on the training set's k-Nearest Neighbors' majority label.

3. Previous Algorithm:

3.1 Machine learning and Data Mining techniques Crop prediction using ANN, SVM, Random forest

- Artificial Neural Networks (ANN)

ANN is a commonly used computational model for machine processing and pattern identification that is based on the structure of the brain. Additionally, agricultural yield forecasting has made extensive use of ANN. This research examines neural network parameters, such as ANN, because they have an impact on predictive performance. The neural network may include more than two layers (input and output). The estimate's exactness is based on the number of layers. A multilayer perceptron and certain lower backpropagation rules are frequently used in this topological method of machine learning. Farmers may view meteorological and environmental conditions for specific crops, which is helpful [10].

- Support Vector Machine

SVM bases its classification into one of two categories on the idea of decision levels that establish boundaries. SVM considers each attribute's own data point [16].

Drawbacks: A mechanism for validation can be introduced.

- Random forest

This algorithm does both regression and classification by creating a large number of decision trees during training. Overfitting is less of an issue with random forests. Pruning is unnecessary. Uses the best forecasters from a group of predictors that were randomly selected at every node to segregate each node.

Drawbacks: The dataset can be expanded to include weather and soil features to increase the generalization of the meta-models.

- Polynomial Regression

Regression analysis is done using polynomial regression. It explains the connection between both variables. It calculates as an nth-degree polynomial in an independent variable and outputs the outcome. mathematical formula for polynomial regression: n is the independent variable, and a + b (degree) is the dependent variable. The model has attribute variables and is non-linear. Additionally, exponential variables are present. Polynomial regression is similar to linear regression. Different varieties of polynomial regression exist. Linear regression with multiple variables: h(x) = b0 + b1x1 + b2x2 + b3x3 +... + bnxn The value (hx) in linear regression is dependent on numerous variables. The value (hx) of a linear regression depends on the strength of the individual variables. A specific variety of linear regression is known as polynomial regression [4,15] in Table .1.

Table 1 Dataset Description with features

Description	Variable
The Production of crop in specific year in kilogram per hectares.	Production
Area of agriculture plants regions in hectares	Area
Plants like are canut, rice, banana ,coconut,etc..	Crop
Data is collected from all the 640 districts and 5924 sub districts in the states of India	District Name
The Data was Collected from 29 States in India	State Name
Seasons like kharif, whole year, rabi etc, crop cultivation season	Season
The data was taken from 1997 onwards	Crop_Year

3.2 Crop prediction using MLR, Density-based clustering technique

- Multiple Linear Regression

The term "multiple regression model" refers to a regression that uses many predictor variables. Modeling the linear relationship between a dependent variable and one or more independent variables is performed using the multiple linear regression (MLR) approach [8, 10]. The independent variables are known as predictors, while the dependent variable is occasionally referred to as predicted [9, 14]. The multiple linear regression (MLR) method is primarily employed in data prediction and is built on least squares. If the crop yield prediction model employs multiple linear regression (MLR) approaches, the crop yield will be predicted, and the predictors will be the year, rainfall, area of sowing, yield, and fertilisers (nitrogen, phosphorous, and potassium).

Drawbacks: Since it is not already present, plant disease detection can be added to the system.

4. The Proposed Method With Algorithm:

Farmers must gather prior output data and analyse it using the current conditions in order to estimate the crop. The statistics obtained using knowledge of the KNN technique are precise. It is challenging to enhance the rate of production without any prior experience in crop forecasting or with data. This method takes into account inputs such as the cultivation area, state, district, and season. If a consumer provides extra parameters, the data may be changed. The system will forecast the appropriate crop according to the input data. The application will forecast the outcome if the qualities are converted into values to test the algorithm's ability to predict crop.

4.1 KNN:

Numerous algorithms calculate crop production. Among these, K-Nearest Neighbor (K-NN) is one that may be quickly put into practise on a low budget. We must determine the frequency of recurring similarities among neighbours and anticipate the outcome using the K-nearest neighbours (KNN) obtained based on prior outcomes.

- **KNN – Algorithm**

Machine learning methods include the k-nearest neighbour algorithm. In this, we take the value as an observation and take into account each of the features to forecast the harvest. The distance between two points is defined as their similarity.The algorithm's flow is shown in the following order:
1. Include the data in the collection
2. Provide the K value.
3. Each data point for Measure the separation between the most recent data and historical data.
4. After calculating the distances, sort the sorted collection of distances.
5. From the sorted collection, select K entries with a high frequency.6. Consider the specifics of the selected K values with a frequency range.
7. Determine the mean of the regression for the K characteristics.
8. Use the K attribute mode for classification.

- **Prediction of Crop Yield through K-NN**

Thus, we implemented a system that forecasts the ideal crop based on input parameters using the KNN algorithm. KNN is utilised to forecast the previously unknown parameter.
Parametersseasons, place, and region were used as entry parameters in this approach to forecast what was appropriate for the given circumstances. Farmers can benefit from the records' predictions by having accurate produce. A suitable crop can be determined by gathering data from the user. Following the collection of information from the previous year, create a CSV file. Import all necessary libraries, including Sklearn, NumPy, and Pandas. There are five variables that need to be entered. District name, crop year, season, crop, and area are the variables. And the attributes should be converted into values from string based on their efficiency. Distance is given by:- $D(a, b) = \sqrt{(a1 - b1)2 + (a2 - b2)2}$. Where a1,a2 are the previous data attribute values and b1, b2 are current data attributes which values should be calculated.

- **KNN has a few further Applications**

A supervised technique called K-Nearest Neighbor Rule (KNN) uses previous data to forecast the unknown variables. The system's dataset can be modified in accordance with the needs of the customer for their agricultural features. A low-cost, accurately personalised system is created using KNN to assist farmers in selecting the best crop and boosting output. KNN is simple to use, and all the necessary information is easily accessible online. Collecting proper prior data is crucial to the system's construction in order to ensure the effectiveness of the present data.

5.Results:

5.1. Dataset Collection:

Agricultural parameters are what we should use to create a prediction model. State, district, cultivable area, weather, and statistics from the prior year are some examples of input variables. Farmers can benefit from the records' predictions by having accurate produce. The user's data might be collected, along with. As a result, a suitable crop is provided. Prepare a CSV file following the collection of information from the prior year in Table.2. Import all necessary libraries, including Sklearn, NumPy, and Pandas[20-25]. There are five variables that need to be entered. District name, crop year, season, crop, and area are the variables and based on how effectively they may be used, the properties should be translated into values using strings.

Table 2 Database description after Preprocessing

Description	Variable
Highest production in a particular year.	Production
Area in hectares.	Area
Crop cultivation seasons.	Season

Suitable crops.	Crop
Data of calculated production from a previous year.	Crop Year
Data is collected from all districts of the state.	District Name
Data of a state in India.	State Name

Thus, we assemble data from various sources and create a dataset. The annual crop report is computed. The proposed model is validated using data derived from data. Data is gathered from an Indian state's district in Table.3

Table 3 Synthetic dataset on Punjab Region

Productionof the previous year	Area to cultivate	Crop suggested	Crop Year	Season	District	State
5430000	162348	Maize	2021	rabi	Patiala	Punjab
241	820	Rice	2021	rabi	Patiala	Punjab
612	190	Sugarcane	2021	rabi	Patiala	Punjab
432	142	Othersgram	2021	Kharif	Patiala	Punjab
4	123	Wheet	2021	Kharif	Patiala	Punjab
2230	1532	Redgram	2021	Kharif	Patiala	Punjab

Table 4 Evaluation measures of Rice

Accuracy (%)	Testing (%)	Training (%)	Method
90.7	30	70	RF
73.3	30	70	SVM
96.4	30	70	ANN

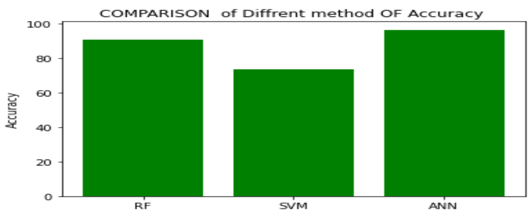

Figure.1: comparison of different method of accuracy

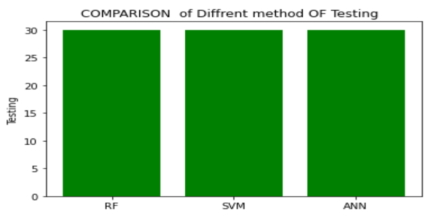

Figure .2 : comparison of different method of Testing

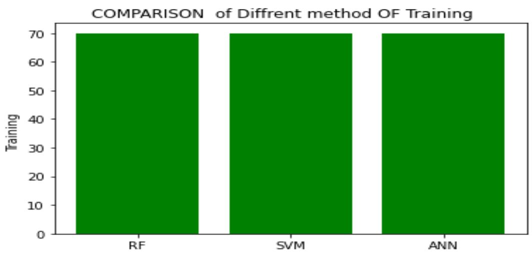

Figure.3 comparison of different method of Training

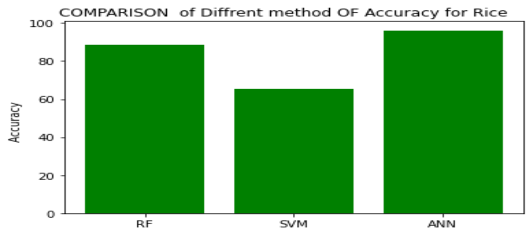

Figure. 4 comparison of different method of Accuracy for Rice

Table 5 Evaluation measures of potato

Accuracy (%)	Testing (%)	Training (%)	Method
88.7	30	70	RF
65.3	30	70	SVM
96.1	30	70	ANN

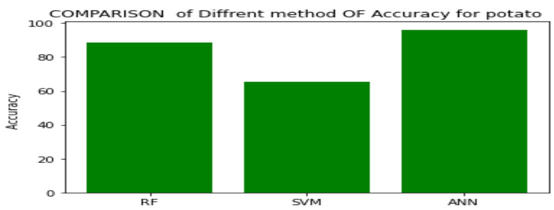

Figure. 5 comparison of different method of Accuracy for potato

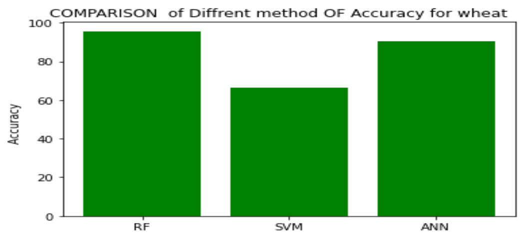

Figure.6 comparison of different method of Accuracy for wheat

Table 6 Evaluation measures of wheat

Accuracy (%)	Testing (%)	Training (%)	Method
95.6	30	70	RF
66.4	30	70	SVM
90.2	30	70	ANN

The technique is designed to assist farmers in increasing production and forecasting crops that will meet their needs. In this system, we employed the KNN algorithm to forecast the harvest depending on the data reports from before Here, we used variables such as the state, district, cultivable area, season, and data from the prior year. Farmers can benefit from the records' predictions by having accurate produce. A suitable crop can be determined by gathering data from the user. Prepare a CSV file following the collection of data from the prior year. Import all necessary libraries, including Sklearn, NumPy, and Pandas. There are five variables that need to be entered. District name, crop year, season, crop, and area are the variables. And, depending on how effectively they are used, the attributes should be converted from strings to values. The algorithm's overall accuracy and the ideal crop will determine the outcome. Thus, we received a 77% accuracy rating, and the crop that was forecasted was based on the consumer's input in Table 4,5,6 and Figure 1,2,3,4,5,6, and 7.

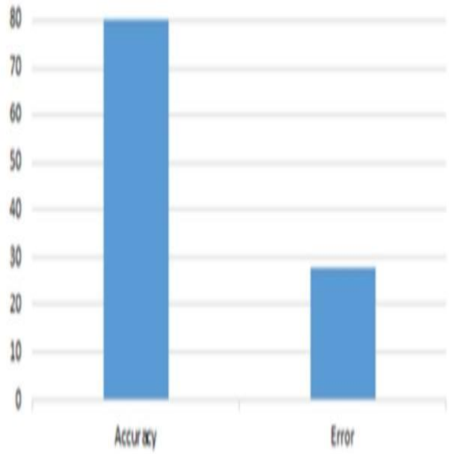

Figure 7 Overall accuracy and error rate of

This application's estimate, which has an accuracy rate of up to 80 and a minimal error rate of 30, will increase the production's accuracy. Six parameters were taken into account, and they were included in the method. Real-time data is gathered from many portals and organised into a dataset to get precise results. Farmers can boost their output by including seasonal changes.

6. Conclusion:

Hence, we draw the conclusion that the customer's agricultural parameters influence the crop prediction. The consumer needs dataset is accessible. So, using the KNN algorithm, predicting an appropriate crop is made simple. Here, we used variables such as the state, district, cultivable area, season, and data from the prior year. Farmers can benefit from the data's predictions by having accurate produce. An appropriate crop can be determined by gathering information from the customer. There are five variables that need to be entered. District name, crop year, season, crop, and area are the variables. And depending on how effectively they may be used, the attributes should be translated into values using strings. The algorithm's accuracy rate and the ideal crop will determine the outcome. Here, the crop prediction system is constructed in an inefficient manner by relying solely on the highly accurate KNN algorithm and a suitable crop for the prediction.

7. **References:**

[1] B. Rama, P. Praveen, H. Sinha and T. Choudhury, "A study on causal rule discovery with PC algorithm," 2017 International Conference on Infocom Technologies and Unmanned Systems (Trends and Future Directions) (ICTUS), Dubai, 2017, pp. 616-621.doi: 10.1109/ICTUS.2017.8286083.

[2] B.Manjula Josephine, K.RuthRamya, K.V.S.N Rama Rao, SwarnaKuchibhotla, P. VenkataBala Kishore, S. Rahamathulla

[3] Chand, R., Raju, S. S. (2008). Instability in AndhraPradesh Agriculture - A Disaggregate Analysis,Agricultural Economics Research Review, Vol. 21(2),PP: 283-288.

[4] D Ramesh and B Vishnu Vardhan, "ANALYSIS OF CROP YIELD PREDICTION USING DATA MINING TECHNIQUES", International Journal of Research in Engineering and Technology eISSN: 2319-1163 | pISSN: 23217308

[5] M S Ramaiah Institute of Technology, Bangalore Vol. 5, Special Issue 2, October 2016.

[6] M. Naveen Kumar1 2, Dr. M. BalakrishnanResearch Scholar, R&D Centre, BharathiarUniversityPrincipal Scientist, National Academy of Agricultural Research Management, ICAR, Rajendra Nagar, Hyderabad, India.

[7] Miss.SnehalS.Dahikar1 ,Dr.Sandeep V.Rode2 PG Student (EXTC), Dept. Of EXTC, Sipna College of Engineering, Amravati, Maharashtra, India1 Dr. Dept. Of EXTC, Sipna College of Engineering, Amravati, Maharashtra, India, January 2014

[8] N.Gandhi and L.J. Armstrong, "Applying data mining techniques to predict yield of rice in Humid Subtropical Climatic Zone of India", Proceedings of the 10th INDIACom-2016, 3rd 2016 IEEE International Conference on Computing for Sustainable Global Development, New Delhi, India, 16th to 18th March 2016.

[9] N.L. Chourasiya, P. Modi , N. Shaikh3 , D. Khandagale, S. Pawar (Department of Computer Engineering, MES College of Engineering/S.P.Pune University, India)

[10] P. Praveen, B. Rama and T. Sampath Kumar, (2017), An efficient clustering algorithm of minimum Spanning Tree Third International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB pp 131-135.

[11] P.Praveen, B.Rama, "An Efficient Smart Search Using R Tree on Spatial Data",Journal of Advanced Research in Dynamical and Control Systems, Issue 4,ISSN:1943-023x.

[12] Peterson, L. E. (2009). K-Nearest Neighbor,Scholarpedia, Vol. 4(2), PP: 1883.

[13] Praveen P., Shaik M.A., Kumar T.S., Choudhury T. (2021) Smart Farming: Securing Farmers Using Block Chain Technology and IOT. In: Choudhury T., Khanna A., Toe T.T., Khurana M.GiaNhu N. (eds) Blockchain Applications in IoT Ecosystem. EAI/Springer Innovations in Communication and Computing. Springer, Cham. https://doi.org/10.1007/978-3-030-65691-1_15

[14] Praveen., P and Ch. JayanthBabu. "Big Data Clustering: Applying Conventional Data Mining Techniques in Big Data Environment." (2019).Innovations in Computer Science and Engineering, Lecture Notes in Networks and Systems 74, ISSN 2367-3370, https://doi.org/10.1007/978-981-13-7082-3_58 Springer Singapore.

[15] R Ravi Kumar MBabu Reddy P Praveen " Text Classification Performance Analysis on Machine Learning" International Journal of Advanced Science and Technology, ISSN: 2005-4238,Vol. 28, No. 20, (2019), pp. 691 – 697.

[16] Singh, K. K., Reddy, D. R., Kaushik, S., Rathore, L. S.,Hansen, J., Sreenivas, G. (2007). Application of seasonalclimate forecasts for sustainable agricultural productionin Telangana subdivision of Andhra Pradesh, India,Climate Prediction and Agriculture - Springer, Berlin,Heidelberg, ISBN: 978- 3-540-44650-7_12, PP: 111-127.

[17] Singha, A. K., Pathak, N., Sharma, N., Tiwari, P. K., & Joel, J. P. C. (2023). COVID-19 Disease Classification Model Using Deep Dense Convolutional Neural Networks. In Emerging Technologies in Data Mining and Information Security (pp. 671-682). Springer, Singapore.

[18] Singha, A. K., Pathak, N., Sharma, N., Tiwari, P. K., & Joel, J. P. C. (2023). Forecasting COVID-19 Confirmed Cases in China Using an Optimization Method. In Emerging Technologies in Data Mining and Information Security (pp. 683-695). Springer, Singapore.

[19] Singha, A.K. and Zubair, S., 2022. Machine Learning for Hypothesis Space and Inductive Bias: A Review. AIJR Abstracts, p.70.

[20] Singha, A.K., PathakS, N., Sharma, N., Urooj, S., Zubair, S., and Larguech, S., An Efficient Integrated Optimize Method Based on Adaptive Meta Optimizer . Intelligent Automation & Soft Computing,(Accepted)

[21] Singha, A.K., PathakS, N., Sharma, N., Urooj, S., Zubair, S., and AreejMalibari. Design of ANN Based Non-Linear Network Using Interconnection ofParallel Processor. Computer Systems Science and Engineering Accepted)

[22] Thayakaran Selvanayagam1 , Suganya S2 , Puvipavan Palendrarajah3 MithunParesith Manogarathash4 , Anjalie Gamage5 , Dharshana Kasthurirathna6 Faculty of Computing, Sri Lanka Institute of Information Technology (SLIIT) Malabe, Sri Lanka, October – 2019.

[23] Vakulabharanam, V. (2004). Agricultural Growth andIrrigation in Telangana: A Review of Evidence.Economic and Political Weekly, Vol. 39(13), PP: 1421-1426.

[24] Zubair, S., Singha, A. K., Pathak, N., Sharma, N., Urooj, S., & Larguech, S. R. (2023). Performance Enhancement of Adaptive Neural Networks Based on Learning Rate. CMC-COMPUTERS MATERIALS & CONTINUA, 74(1), 2005-2019.

**The authors whose names are listed in title page certify that they have no affiliations with or involvement in any organization or entity with any financial interest (such as honoraria; educational grants; participation in speakers' bureaus; membership, employment, consultancies, stock ownership, or other equity interest; and expert testimony or patent-licensing arrangements), or non-financial interest (such as personal or professional relationships, affiliations, knowledge or beliefs) in the subject matter or materials discussed in this manuscript.