International Journal on Advances in Engineering Technology and Science
Volume: 5, Issue: 2, 2024, e-ISSN: 2455-3131
DOI: 10.5281/zenodo.12578772

Predicting the appropriate crop based on the climatic situations on the historic data by using Random Forest machine learning algorithms

Harendra Singh[1], Medhavi Bhargava[2]

School of Engineering & Technology, Sanjeev Agrawal Global Educational University, Bhopal

[1]harendra.s@sageuniversity.edu.in, [2]medhavi.b@sageuniversity.edu.in

***Abstract* – Agriculture plays an important role in Indian economy. But now-a-days, agriculture in India is undergoing a structural change leading to a crisis situation. The only remedy to the crisis is to do all that is possible to make agriculture a profitable enterprise and attract the farmers to continue the crop production activities. As an effort towards this direction, this research paper would help the farmers in making appropriate decisions regarding the cultivations with the help of machine learning. This paper focuses on predicting the appropriate crop based on the climatic situations and the yield of the crop based on the historic data by using Random Forest machine learning algorithms. this paper proposes an idea to predict the crop and yield of the crop based on the climatic conditions and historic data related to the crop. The farmer will check the production of the crop as per the acre, before cultivating onto the field. The quantity of grains required by the population in a given year is heavily influenced by population growth and weather changes.**

Keywords - Machine learning, cultivation, decisions, web Application

I. INTRODUCTİON

In machine learning project we have developed an Android app which take the data from the machine learning algorithm for the data that is been trained by the model we have used random forest algorithm because it was giving much more accurate result than any other algorithm that we used. The only remedy to the crisis is to do all that is possible to make agriculture a profitable enterprise and attract the farmers to continue the crop production activities. In the past farmers used to predict their yield from previous year yield experiences. Thus, for this kind of data analytics in crop prediction, there are different techniques or algorithms, and with the help of those algorithms, we can predict crop yield. Nowadays, modern people don't have awareness about the cultivation of the crops at the right time and at the right place. By analyzing all the issues and problems like weather, temperature, and several factors, there is no proper solution and technologies to overcome the situation faced. Accurate information about history of crop yield is an important thing for making decisions related to agricultural risk management. Therefore, this paper proposes an idea to predict the crop and yield of the crop based on the climatic conditions and historic data related to the crop. The farmer will check the production of the crop as per the acre, before cultivating onto the field. The quantity of grains required by the population in a given year is heavily influenced by population growth and weather changes. Because of the sudden change in weather conditions, grains are occasionally damaged and hence are not sent to market, increasing market demand. It's also difficult to anticipate the weather at times. As a result, the following chapter will explain how quantity and price predictions are made.

II. LITERATURE REVIEW

Predicting the yield of the crop using a machine learning algorithm. This focus on predicting the yield of the crop based on the existing data by using the Random Forest algorithm [1].

Machine learning approach for forecasting crop yield based on parameters of climate. It is used to produce the most influencing climatic parameter on the crop yields of selected crops in selected districts and state [2].

Analysis of Crop Yield Prediction by making Use of Data Mining Methods. The main aim is to create a user-friendly interface for farmers, which gives the analysis of crop production based on the available data. For maximizing the crop productivity various Data mining techniques were used to predict the crop yield [3].

Random Forests for Global and Regional Crop Yield Predictions. The generated outputs show that RF is an effective and different machine-learning method for crop yield predictions at regional and global scales for its high accuracy [4].

Crop Prediction using Machine Learning This research work helps the beginner farmer in such a way to guide them for sowing the reasonable crops by deploying machine learning. The various environmental factors like soil, pressure, weather, crop type to predict the maximized profitable crop to grow. It mainly focuses on the algorithms used to predict crop yield, crop cost predictions [5].

Majority of Asian population has influenced their culture, diets and economic condition by the most common agricultural food. For example, more than 50% of the Indian population has rice as the primary source of

nutrition. To achieve the target of large production, productive lands, human settlement, intensive cultivation process of rice, high quality fertilizer, high quality food seed, moderate artificial climate are some of the required facts. But the farmers may not have exact awareness of some of the manmade factors such as precise use of fertilizer, proper measure to prevent and cure the disease. Many diseases considered as minor, may become serious in many rice growing areas [1]. The bibliography is listing the significant contributions over the world and India. With this reference list, significant achievements and importance is state in the following sections.

Smart Farming [2] is not only focused on the analysis of data acquired through various sources like historical, time series, geographical and location dependent or atomization instruments, but implementation of advanced technology, such as drone or robots to name it a real smart system. Smart systems should possess the abilities to record the data and analyze it to make decision, which is more accurate and precise as compared to human expertise. Smart farming employs IoT kind of hardware, electronic interface and cloud storage to capture the data and all these things can be user friendly handled by mobile app kind of simple application to which every single person has access these days, without strong acquaintance of technical knowledge. It works in a manner that the data is organized, accessible all the time and on every aspect of field operation that can be monitored remotely from anywhere in the world.

Smart Farming [2,3] is a budding notion in which the activity to manage the farms is carried out using up to date Information and Communication Technologies. Various hardware devices and software computational techniques can be used to amplify the magnitude and quality of production, increase safety, automatic controlling of environment, field related resources, remote monitoring, hazard avoidance. At the other side, it also optimizes the human labor, soil nutritional element, it lessens dependence on human expertise and approximate measures. It also lessens the dependence on climatic condition which hamper the crop yield. Smart farming or agriculture comprises of Sensors for detecting soil moisture, water, sunlight amount, humidity and temperature in weather conditions, and other supporting and ambient factors. It also comprises historical data analysis where seasonal changes affected crop yield, time series analysis, data analysis where data is contributed to the government agencies for reporting purpose. It also comprises of satellite imagery-based applications, rainfall measures, floods and disaster-based measures and such relational data is to be assessed.

Smart agriculture [4,5] research aims to provide a decision-making support system or framework for overall agricultural management to address the issues of population growth and relative crop yield, climate change and relative crop yield, technological gain, from planting and watering of crops to remote monitoring of health and harvesting of the plants. Deep learning (DL)[5] incorporates a up to date procedure for image processing and big data analysis with huge potential. Deep learning is a recent tool these days in the agricultural domain. It has already been successfully applied to other domains. The chapter analyses the specific employed models for image data taken from the authenticate sources, the performance of iterations, the employed software app and the likelihood of real-time application to evaluate the crop disease, prototype is rice crop. Deep learning offers high precision outperforming other image processing techniques.

III. DISCRIPTION

Figure 1 – Wheat: Major wheat growing states in India are Uttar Pradesh. Wheat is the second most important crop in India next to rice. This food grain of the country is actually the staple food of the people of north-western India. A huge portion of the total cropped area in the country is under the production of wheat crop. It is also said that as a food, wheat is more nutritive as compared to the other cereals. The gluten present in wheat determines its chapati making quality. Hard varieties of wheat are richer in gluten. In India, generally hard varieties of wheat are grown as most of the wheat grown in the country is consumed in the form of chapatis.

Figure 2 – Barley: Barley is primarily a cereal grain popularly known as jau in India. It is the fourth most important cereal crop after rice, wheat and maize. It's converted into malt to use for various food preparations. Barley is commonly used in breads, soups, stews, and health products, though it is primarily grown as animal fodder and as a source of malt for alcoholic beverages, especially beer.

Figure 3 – Millet: Millets are coarse grains like Ragi, Bajra and Jowar. They are highly nutritious and are generally used by rural people. They can be grown in areas of low rainfall and low to medium fertile soils. Jowar, bajra and ragi are the important millets grown in India. They have high nutritional value.

Figure 4 – Rice: Rice is the third-largest crop production, after sugarcane and maize. The main producers of rice are the nations of China, India, Indonesia, Bangladesh, and Vietnam. Rice is a staple crop. ... Not only is rice a key source of food but it is also good source of income for many smallholder farmers. Rice is the most important food crop of the developing world and the staple food of more than half of the world's population. Rich in nutrients and vitamins and minerals, it an excellent source of complex carbohydrates.

Figure 5 – Soyabean oilseed: The soybean, soy bean, or soya bean (Glycine max) is a species of legume native to East Asia ... Soybeans contain a small but significant 2S storage protein. The soybean is economically the most important bean in the world, providing vegetable protein for millions of people and ingredients for hundreds of chemical products. Soybeans are high in protein and a decent source of both carbs and fat. They are a rich source of various vitamins, minerals, and beneficial plant compounds, such as isoflavones.

Fig.1- Wheat

Fig.2- Barley

Fig.3- Millet

Fig.4- Rice

Fig.5- Soyabean

IV. PROPOSED MODEL

The easiest way to the work done in this is based on following ML models:

Bagging: Random forest, for every response y, there are n inputs. Prediction is done by binary tree, at each node, a test to input is applied, mean squared error determines the value and left or right subbranch. Eventually leaf made is prediction node. Feature values are preferred to be categorical. Leaf node is class label. Prediction score of each node is averaged.

Random forest – It's a supervised machine learning algorithm that's commonly used to solve classification and regression problems. It constructs decision trees from various samples, using the majority vote for classification and the average

for regression. Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

The greater number of trees in the forest leads to higher accuracy and prevents the problem of over fitting.

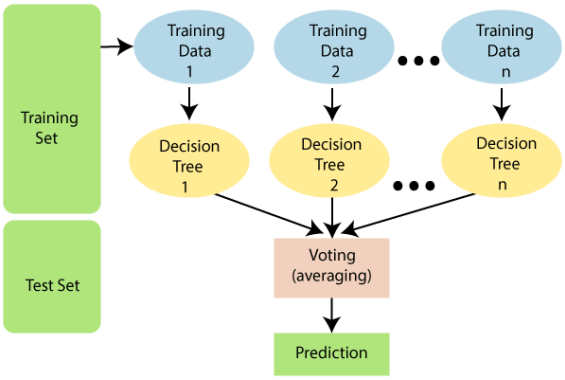

Fig.6- Proposed Architecture

One of the most essential characteristics of the Random Forest Algorithm is that it can handle data sets with both continuous and categorical variables, as in regression and classification. For classification difficulties, it produces superior results.

Seaborn – Seaborn is a Python package based on matplotlib that is open-source. It's used for exploratory data analysis and data visualization. With data frames and the Pandas library, Seaborn is a breeze to use. The graphs that are created can also be readily altered.

2.Matplotlib - Matplotlib is a python visualization toolkit with a low-level graph plotting library. John D. Hunter is the creator of Matplotlib. Matplotlib is free and open-source software that we can utilize. For platform compatibility, Matplotlib is largely written in Python, with a few segments written in C, Objective-C, and JavaScript.

3.NumPy - NumPy is a Python library that allows you to do things with numbers. NumPy is a Python library for working with arrays. The acronym NumPy stands for "Numerical Python."

4.Pandas - Pandas is a valuable data analysis library. It has the ability to manipulate and analyze data. Pandas provide data structures that are both strong and simple to use, as well as the ability to quickly perform operations on them.

5.Sklearn - Scikit-learn is a free Python machine learning library. Support vector machines, random forests, and k-neighbors are among the algorithms included.

Historical data about the crop and the climate at the district level was needed to implement the system. This data has been gathered from the government website www.data.gov.in which includes State, District, Season, Crop, Area and Production. The data about the climate conditions suitable for the particular crops has been collected from the Kaggle which includes Temperature, Population, Precipitation, windspeed, msp.

1. Month: In our dataset we take first row as a month from 1-Jan-00 to 1-Feb-21.
2. Population: This number was obtained from the World Development Indicators and measures the annual change of agriculture production vs. the production from previous years. Population Growth.

3. Temperature: High temperature, even for short period, affects crop growth. We calculate temperature in degree-Celsius.
4. Max-temperature: The largest or highest temperature in that month.
5. Min-temp: The lowest record of the temperature in that month.
6. Precipitation: It is form of liquid or solid water particles that fall from the atmosphere and reach the surface of the Earth.
7. Windspeed: is a fundamental atmospheric quantity caused by air moving from high to low pressure, usually due to changes.
8. MSP (Minimum support Price): It is a minimum price guarantee that acts as a safety net or insurance for farmers when they sell particular crops.

	month	population	temp	maxtemp	mintemp	precipitation	windspeed	msp	dayofprec	production	price
0	1-Jan-00	1056575549	17.3	22.2	7.8	0.0	5.9	610	0	76369.00	4077.33
1	1-Feb-00	1056575550	20.2	24.8	11.0	0.0	7.9	610	0	76369.00	4281.47
2	1-Mar-00	1056575551	23.8	32.0	14.0	0.0	9.0	610	0	76369.00	4201.94
3	1-Apr-00	1056575552	32.2	39.4	25.0	0.6	9.1	610	1	76369.00	4120.83
4	1-May-00	1056575553	32.0	39.1	25.1	0.0	9.5	610	0	76369.00	4357.91
...	...	...	...	...	...	...	...	...	...	...	...
249	1-Oct-20	1380004385	27.9	34.9	20.9	0.0	2.4	1975	0	107592.00	14947.04
250	1-Nov-20	1380004385	20.8	28.2	13.4	21.2	2.5	1975	2	107592.00	15651.92
251	1-Dec-20	1380004385	16.5	23.4	9.6	0.0	3.2	1975	0	107592.00	16006.59
252	1-Jan-21	1393409038	15.0	21.0	9.1	1.0	3.6	2035	1	111895.68	17395.20
253	1-Feb-21	1393409038	20.5	28.4	12.5	5.0	3.9	2035	2	111895.68	17522.70

254 rows × 11 columns

Fig.7- Dataset for wheat

	month	population	temp	maxtemp	mintemp	precipitation	windspeed	msp	dayofprec	production	price
0	1-Jan-00	1056575549	15.6	24.0	7.0	0.0	3.8	445	0	1447	3184.95
1	1-Feb-00	1056575550	20.3	24.2	12.6	0.0	7.5	445	0	1447	3199.44
2	1-Mar-00	1056575551	25.2	31.9	16.5	13.9	5.0	445	1	1447	3242.99
3	1-Apr-00	1056575552	31.1	39.9	23.2	2.0	6.9	445	1	1447	3358.02
4	1-May-00	1056575553	34.3	40.3	27.2	41.6	10.9	445	6	1447	3441.13
...	...	...	...	...	...	...	...	...	...	...	...
249	1-Oct-20	1380004385	28.3	34.9	21.7	0.5	3.4	2150	1	1687	8596.36
250	1-Nov-20	1380004385	21.1	27.9	14.4	14.0	5.1	2150	2	1687	9856.65
251	1-Dec-20	1380004385	18.1	25.6	10.6	0.0	2.3	2150	0	1687	9874.65
252	1-Jan-21	1393409038	16.4	23.0	9.9	9.0	5.8	2645	2	1700	10569.60
253	1-Feb-21	1393409038	21.7	29.4	13.9	0.0	6.4	2645	0	1700	10956.58

254 rows × 11 columns

Fig.8- Dataset for Barley

	month	population	temp	maxtemp	mintemp	precipitation	windspeed	msp	dayofprec	production	price
0	1-Jan-00	1056575549	15.1	25.7	11.0	1.1	10.0	775	3.0	5250	11.59
1	1-Feb-00	1056575550	19.0	26.0	13.2	14.0	10.8	775	1.0	5250	12.16
2	1-Mar-00	1056575551	25.0	32.8	18.9	12.1	13.5	775	3.0	5250	11.69
3	1-Apr-00	1056575552	32.1	40.2	21.9	0.8	15.3	775	3.0	5250	13.68
4	1-May-00	1056575553	31.2	37.2	26.7	1.0	16.7	775	1.0	5250	13.72
...	...	...	...	...	...	...	...	...	...	...	...
249	1-Oct-20	1380004385	26.6	33.4	19.9	24.0	8.1	3800	2.0	10500	23.04
250	1-Nov-20	1380004385	21.9	29.6	14.2	0.2	7.4	3800	1.0	10500	24.18
251	1-Dec-20	1380004385	18.6	25.7	11.5	6.0	7.6	3800	3.0	10500	23.93
252	1-Jan-21	1393409038	18.7	25.5	11.9	16.0	8.5	4050	2.0	10000	31.10
253	1-Feb-21	1393409038	21.6	30.3	12.9	0.0	8.7	4050	0.0	10000	33.20

254 rows × 11 columns

Fig.9- Dataset for Soyabean Oilseed

Exploratory Data Analysis: It refers to the critical process of performing initial investigations on data so as to discover patterns to spot anomalies to test hypotheses and to check assumptions with the help of summary statistics and graphical representations.

Data Cleaning: It is the process of preparing data for analysis by removing or modifying data that is incorrect, incomplete, irrelevant, duplicated, or improperly formatted.

Encoding: It is a required pre-processing step when working with categorical data for machine learning algorithms.

Feature Scaling: It is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units.

Data Partitioning: The Entire dataset is partitioned into 2 parts: for example, say, 75% of the dataset is used for training the model and 25% of the data is set aside to test the model.

V. RESULTS

After training the data with machine learning model,

i) Predicted value of yield and actual value shows an accuracy of approximately 97%.

ii) Predicted value of price and actual value shows an accuracy of approximately 95%.

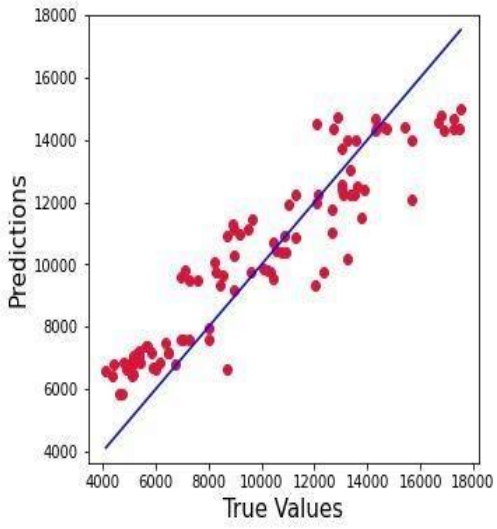

Fig.10- Prediction v/s Actual Values (Crop Price for wheat)

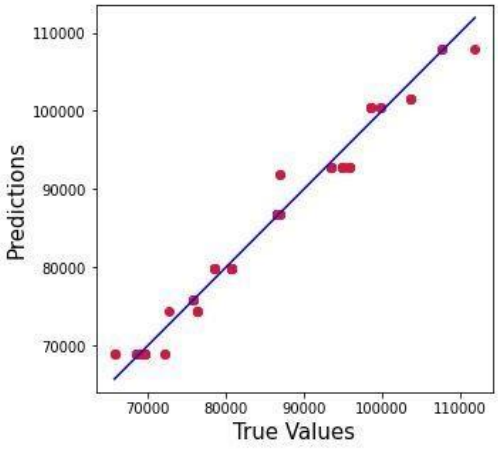

Fig.11- Prediction v/s Actual Values (Crop Production for wheat)

VI. CONCLUSİON & FUTURE WORK

Crop and yield of the crop prediction using intelligent machine learning techniques may improve the crop planning decisions. Our model allows the farmer to anticipate the best yield in terms of quality and quantity. The farmers also get the Information about desired or estimated price for their yield. They can get the information of Quantity and price for their grains with climate change according to his needs. With the help of our model the government can keep a check if there is any kind of black-marketing of grains.

For example: If both the population and production increase every year and if there is a sudden high depend of a grain for the same population. By this the government can trace if there is black-marketing.

Increase the number of grains and will include the pulses and fruits as well. Use of advance machine learning and deep learning to get accurate results. Increasing the dataset by taking the soil profile and judging it in terms of nutrients and quality of soil.

Time series prediction using neural networks is a critical tool for understanding global agricultural commodity futures pricing and, more importantly, for lowering uncertainty and risk in agricultural markets.

The government can use our approach to see if there is any form of grain black-marketing going on. Consider the following scenario: If the population and production both grow every year, and if there is a dramatic increase in the reliance on a particular grain for the same population. The government can use this to see whether there is any black-marketing going on.

Increase the number of grains consumed, as well as pulses and fruits. To achieve accurate results, advanced machine learning and deep learning are used. Increasing the dataset by evaluating the soil profile in terms of nutrients and soil quality.

Increase the number of grains and will include the pulses and fruits as well. Use of advance machine learning and deep learning to get accurate results. Increasing the dataset by taking the soil profile and judging it in terms of nutrients and quality of soil.

REFERENCES

[1] Agila N, Senthil Kumar P, "An Efficient Crop Identification Using Deep Learning", Int Journal of Scientific & Technology Research, vol 9, no 01, pp.2805-2808, January 2020

[2] Santos, Luís & Neves Dos Santos, Filipe & Moura Oliveira, Paulo & Shinde, Pranjali, "Deep Learning Applications in Agriculture: A Short Review" 10.1007/978-3-030-35990-4_vol.12, no(5) 2019.

[3] Yan Guo, et al.," Plant Disease Identification Based on Deep Learning Algorithm in Smart Farming", Hindawi Discrete Dynamics in Nature and Society Volume 2020, Article ID 2479172, 11 pages https://doi.org/10.1155/2020/2479172

[4] M. El-Helly, S. El-Beltagy, and A. Rafea, "Image analysis-based interface for diagnostic expert systems," in Proceedings of the Winter International Symposium on Information and Communication Technologies, pp. 1–6, Trinity College Dublin, Cancun, Mexico, January 2004.

[5] Tellaeche, X. P. Burgos-Artizzu, G. Pajares, and A. Ribeiro, "A vision-based method for weeds identification through the Bayesian decision theory," Pattern Recognition, vol. 41, no. 2, pp. 521–530, 2008.

[6] P. S. Landge, S. A. Patil, D. S. Khot, O. D. Otari, and U. G. Malavkar, "Automatic detection and classification of plant disease through image processing," International Journal of Advanced Research in Computer Science and Software Engineering, vol. 3, no. 7, pp. 798–801, 2013