PredYield - Agricultural Yield Predictor

Abstract — **PredYield is a solution for farmers facing Indian Agriculture Challenges due to uncertain weather conditions, inappropriate farming, and irrigation. Indian economy is largely dependent on agricultural activities. Due to variation in climate, wrong agricultural methods, and other prices influencing parameters, farmers face massive losses due to failure of crops and uncertain price fluctuations, which lead to rising suicidal deaths of farmers. Traditionally yield prediction was performed with guesswork based on farmer's experience on-field and crop. This system can be of great advantage for farmers. The proposed system focuses on forecasting the crop's yield and price using historical data considering various parameters affecting the agricultural output. Knowing prices and yield productions, farmers will gain a huge boost in their decision-making. It will also allow farmers to have a brief understanding of the nature of yield production with easy-to-understand visualized charts based on historical data.**

Keywords— Price prediction, yield production, decision-making

I. INTRODUCTION

Our country's economy is mainly based on agriculture. It has been seen that the prediction was performed based on the farmer's experience on-field and crop. Due to variation in climatic change, wrong methods, and other prices influencing parameter farmers face massive loss such as failure of crops and uncertainties in the price fluctuation. In agriculture, yield and price prediction is a critical topic. Every farmer wants to know how much produce they should expect.

Machine learning plays a crucial role in crop yield prediction. It acts as a decision support tool for predicting what crops can be grown in the area. Different machine learning algorithms can be used to predict crop yield.

Robotic Process Automation will be used to collect agricultural data. Data can be gathered by scraping various websites and using APIs. Robotic Process Automation is mimicking human actions to perform a sequence of steps that lead to meaningful activity, without any human intervention. Uipath is globally available free software for RPA[14][15]. It can perform various types of Automation. It can be hosted in cloud environments or virtual terminals.

II. LITERATURE SURVEY

Rohith, Vishnu, Kishore, Deeban Chakkaravarthi[1]:- In this paper, the emphasis is on machine learning techniques to predict the Price of the Crop using the support vector regression Algorithm. Using the Regression technique as a method to predict the price of the crop. Thereby solving the crop price prediction problem to ensure the guaranteed benefits to the farmers.

Sunil Kumar, Vivek Kumar, R. K. Sharma[2]:- This paper used a support vector machine to predict the rice yield. Different experiments of SVM training were done using a one-to-one multi-classification process, k-fold cross-validation, and a polynomial kernel function. A 4-fold cross-validation approach is used to find the best prediction accuracy for the 4-year relative average rise.

S Brunda, Nimish, Chiranthan, Arbaaz Khan [3]:- In this paper, the data consists of 330 different crops grown across India, with prices calculated as the average of prices available in different locations. The Crop prices make up the dataset. For the prediction model, three main factors were focused i.e. item name, prices, and date of production. For optimizing they used Decision Tree Regressor. In addition, the model's hyperparameter was tuned to improve price prediction accuracy. But using a random forest regressor they attained the same results even without tuning the hyperparameters. Thus Random Forest Regressor is an effective tool for prediction.

M. Kalimuthu, P. Vaishnavi, and M. Kishore[4]:- In this paper, the authors have done the crop prediction process by consolidating the preceding data and the present data of a particular month to prove the accuracy of climatic data. The system followed the Naive Bayes classifier, the supervised learning algorithm consisting of the four levels to calculate and predict the crop for the suitable climate in phenomena such as Data collection, Preprocessing Step, Feature Extraction, and Data Prediction. DHT11 was used for implementation.

Raissa Uskenbayeva, Zhyldyz Kalpeyeva, and Ryskhan Satybaldiyeva[5]:- In this paper, the author's used high-level analysis to identify potentially robotized processes. This followed by a selection of pilot processes, the calculation of anticipated performance metrics, the comparison of RPA vendors, and the purchase of licenses are all tasks that must be completed. The next step is to develop and test the selected processes of the previous stage. The main objective here is to transfer the robots to a productive environment, scaling, and the formation of a center of competence within the company.

Piotr Andruszkiewicz and Henryk Rybinski[6]:- In this paper, the process started by gathering structured and reliable data from the digital libraries for addressing the issue of not always up-to-date data in digital libraries. URLs were filtered using an SVM classifier to identify potentially useful pages. Using the machine learning algorithm of crowdsourcing data they extracted information. For Data Acquisition, they used a crawling method to gather web pages from the Internet. After this, they gathered data from unstructured sources. They used SVM and CRF models for extracting necessary information about a conference from its website.

Y. Yang, L.T. Wilson, J. Wang[7]:- This paper provides an overview of the processes and designs that led to the development of an automated climatic data scraping, filtering, and exploring system, and the integration of climatic database with cropping systems applications. The method of constructing climate data is divided into five main software modules. First, it processes the web pages for the requested data. Second, it fetches the data made accessible by data sources in response to a request then decompresses and saves the data to a centralized SQL Server database after parsing the contents of the original data file. Data Filter aids in the discovery of additional data quality control, as well as the detection of outliers and the estimation of missing data. Data Explorer is a web-based user interface that provides users with access to consolidated and filtered climatic data through dropdown lists and map-based navigation.

From the above mentioned papers, we studied there are various techniques such as machine learning, data mining, and IoT are proposed for the same. Existing system does not take into account the agricultural economy's demand.

III. PROPOSED SYSTEM

A. PROBLEM SCENARIO

An Indian farmer faces Agriculture Challenges due to uncertain weather conditions, inefficient storage, improper warehousing, and inappropriate farming and irrigation[10][11]. Traditionally the prediction of crops is based on the experience of the farmers. Farmers are increasingly vulnerable to unpredictable, more frequent, severe, and longer-lasting drought events, which have an impact on yields, production, and livelihoods. This is due to a combination of climatic and non-climatic factors, such as extreme weather and climate variability, as well as soil type or management. As a result, farmers have been conducting their prediction activities in increasingly unpredictable and challenging conditions like irrigation pattern, and fertilizers, etc.

B. Technologies

a. Robotic Process Automation

Robotic Process Automation is the process of automating the workflow with the help of robots/software to reduce the involvement of humans. With UiPath, one can perform various types of automation such as Excel Automation, Web Automation, PDF Automation, Web Scraping, and so on. Extracting data and storing it into files and databases from various websites can be done using Web Automation[14][17].

b. Supervised Machine learning

Supervised learning is the machine that learns from our data when we specify a target variable. Based on the data that is used to train and test the model. Next, is to split the data into a training dataset and test dataset where model gets trained using training dataset whereas the test dataset acts as new data for predicting results by which we can see the accuracy of our model. After training the model based on the input patterns algorithm generates the expected output which helps in predicting the correct output.

Support vector machine uses data points that are outside the training set. Gatherings of information focuses that could draw a straight line with the end goal that one class is on one side of the line and the other class is on the opposite side of the line. The line that separates the dataset is called a hyperplane. Its objective is to find a plane with a maximum margin, that is, the maximum distance between data points of the data set. The points nearest to the separating hyperplane are called support vectors. Maximizing the margin distance provides some reinforcement so that new data points can be classified more accurately[16].

Decision Tree Regressor is a flowchart-like model that contains conditional control statements, involving choices and their plausible results. The yield identifies with the naming of unpredicted information. In the tree representation, the leaf nodes represent class labels, and the internal nodes correspond to the attributes. A decision tree can be used to solve problems with Boolean functions as well as discrete attributes.

Gradient Boosting Regressor is used for regression analysis. It calculates the difference between the current prediction and the known correct target value. It is called residual. After training a weak model it maps features to the residual. The residual anticipated by a frail model is added to the common model information taking the model towards the right yield. For improvising the model repetition of this step is required.

Random Forest Regressor is an entity technique capable of performing both regression and classification tasks with the utilization of multiple decision trees. It focuses on combining multiple decision trees to determine the final output instead of relying on individual decision trees. It has multiple decision trees as base learning models. For every model, it randomly performs row sampling and feature sampling from the dataset.

C. *Proposed features*

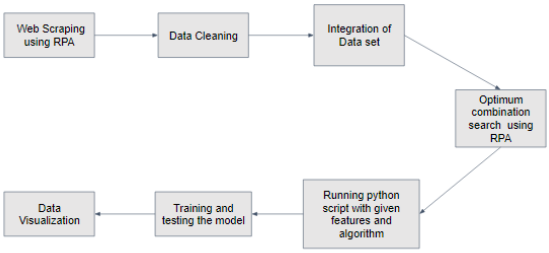

Fig. 1. shows the flowchart for the model.

The proposed system is a centralized system that can be used by farmers. It consists of the following features:

a. Data Collection

All of the datasets used in the study were obtained from openly accessible records from across India[9][10][12][13]. Data is scrapped using RPA from various sources. The data set includes crop yields in India from the year 1990 to the year 2015.

b. Data Pre-procesing

Data preprocessing includes data cleaning, data integration, data transformation, and data reduction. The data which is collected contains some unnecessary columns. After dropping the columns, the next phase is to integrate all the acquired data from scrapping with the given database. Merging of different features like temperature, rainfall, etc. is done by performing a different pre-processing task on each dataset.

	State Name	Dist Name	Year	Crop	Area (1000 ha)	Production (1000 tons)	Yield (kg per ha
0	Chhattisgarh	Durg	1966	RICE	548.0	185.0	338.0
1	Chhattisgarh	Durg	1967	RICE	547.0	409.0	748.0
2	Chhattisgarh	Durg	1968	RICE	556.3	468.0	841.0
3	Chhattisgarh	Durg	1969	RICE	563.4	400.8	711.0
4	Chhattisgarh	Durg	1970	RICE	571.6	473.6	829.0

Fig. 2. Shows merging of different features such as temperature, rainfall, etc. by performing pre-processing task on each dataset.

c. Data Modelling

For creating a model, different machine learning algorithms will be used. Algorithms such as Gradient booster regressor, Decision tree, Random forest tree and Support vector regressor are used in the project. The algorithm that gives the accurate and real value will be further used for different crops for predicting the yield. For better results model will be divided into 70:30 ratio for training and testing purpose.

```
['GradientBoostingRegressor', 0.8935799021468619]
['RandomForestRegressor', 0.685140599475108]
['SVR', -0.1901665990814032]
['DecisionTreeRegressor', 0.9623417470727424]
```

Fig. 3. represents running of different algorithms to check the accuracy.

Item	
Cassava	0.949523
Maize	0.880277
Plantains and others	0.666172
Potatoes	0.918130
Rice, paddy	0.871750
Sorghum	0.831521
Soybeans	0.816393
Sweet potatoes	0.856023
Wheat	0.940815
Yams	0.911715

Fig. 4. Shows the use of the decision tree regressor algorithm for the prediction of the crops.

d. Data Visualization

The yield and price will be represented graphically using data visualization. It creates different charts and graphs so that users can easily understand the insights extracted from the data.

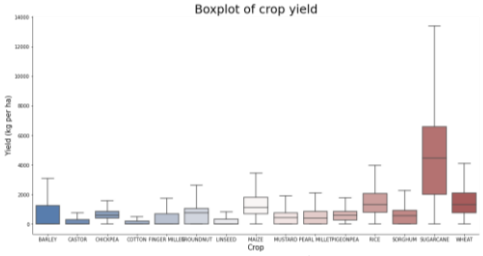

Fig. 5. Shows representation of yield in boxplot.

Box plots visualize 5 kinds of information for a dataset. The middle line reflects the median value, that is, there are 50% values on either side of the median. Within the box itself, we have 50% of the data, 25% above the median and 25% below it. The outliers, if any are plotted beyond the whiskers. The given boxplot in figure 5 shows the dispersion of the crop yield. The maximum and minimum values at the end of whiskers giving a visual understanding of the spread of the yields in kg per hectare. The plot depicts that Sugarcane has the highest yield and spread with the median value roughly around 4000 kg/ha.

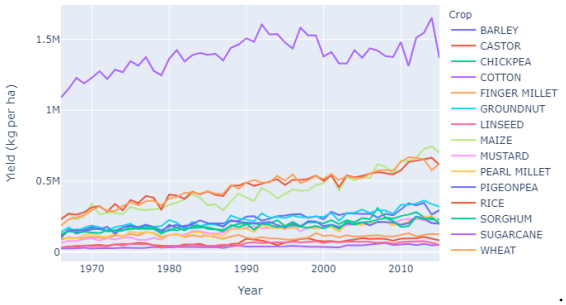

Fig. 6. The plot displays the overall yield of a crop over the years. Each line shows the rise and fall trends for that particular crop. It can see observed that sugarcane has the highest yield when compared to others.

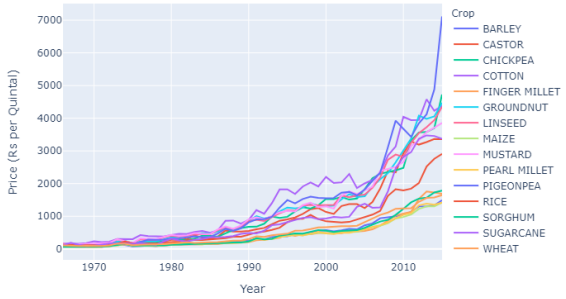

Fig. 7. Shows the graph that describes the rising trend in the prices of crops per quintal over the years, where each color visualize a crop.

e. Predicting prices and quantity of crops

This feature will help the user to determine what will be quantity produced in the coming year and how the prices will change based on the availability and the need of the people. With the price prediction, they can choose to invest towards the production of that crop or even cut down some of the investment of resources towards it.

f. Sort data according to the need

The user can change the duration, region, and the crops for which the data is being visualized which will give him the ease of comparing multiple crops in a region or the same crop in multiple regions and get a broader view on the topic.

IV. SYSTEM FLOW

All the datasets used in the research are sourced from openly accessible records from all over India. Various datasets used from the past 20 to 50 years will be scrapped. From the vast initial dataset, only a limited number of important factors which have the highest impact on agricultural yield are selected for the research. The parameters selected for the study are Area, Crop name, Year of production, Crop Production, Rainfall (in mm), Fertilizers, Temperature, Irrigated area, etc. Data pre-processing techniques, the second step, involves cleaning and converting raw data into an understandable format. After pre-processing techniques, the next phase is to integrate all the acquired data from scrapping with the given database. The next phase involves finding the best set of features that is to be used for optimal prediction using Machine Learning models. For creating a model different machine learning algorithms will be processed on the dataset. After which training and testing of the model will take place in a 70:30 ratio. 70% of the dataset will undergo training whereas the remaining 30% will be used for testing purposes. The model will be trained with the given set of features to predict the yield and prices of the crops. Based on the results, data is visualized using different plots and graphs.

V. RESULTS

Models were trained with the parameters like Rainfall, Fertilizers, Area, Temperature, etc. Data had been collected from different open sources such as various Indian government websites. Since data has been collected from a number of sources, we can see a variation in the number of records of each table. The outcomes of various algorithms were reported and are shown in Table I.

Table I. The accuracy for crop yield prediction

Sr. No	***Algorithm***	***Tool used***	***Dataset***	***Prediction Accuracy (%)***
1.	Gradient Booster Regressor	Jupyter notebook, flask	Multiple samples	89.35
2.	Random Forest Regressor	Jupyter notebook, flask	Multiple samples	68.51
3.	Support Vector Regressor	Jupyter notebook, flask	Multiple samples	0.19
4.	Decision Tree Regressor	Jupyter notebook, flask	Multiple samples	96 .23

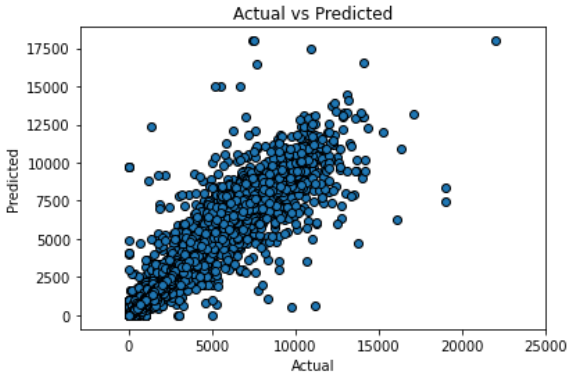

Fig. 8. Shows a graph for the actual vs predicted value.

The above figure 8 shows the goodness of the fit with the predictions visualized as a line. This means that we have found a good fitting model to predict the crop yield value.

VI. CONCLUSION

PredYield will provide a solution for Indian Agriculture Challenges faced by the farmers by predicting the price and yield of the crop. It will help farmers in increasing their profit. It will provide an interface where farmers can interact and examine multiple charts and visualizations. This will help the farmers to make the right decision for the right crop such that the agricultural sector will be developed by innovative ideas. Predicting the price of a given crop with yield in the future is extremely valuable when determining which types of crops to encourage and plant.

REFERENCES

[1] Rohith R, Vishnu R, Kishore A, Deeban Chakkarawarthi "Crop Price Prediction and Forecasting System using Supervised Machine Learning Algorithms".
https://ijarcce.com/wp-content/uploads/2020/03/IJARCCE.2020.9306.pdf

[2] Sunil Kumar, Vivek Kumar, R. K. Sharma: "Rice Yield Forecasting Using Support Vector Machine".
https://www.ijrte.org/wp-content/uploads/papers/v8i4/D7236118419.pdf

[3] S Brunda, Nimish L, Chiranthan S, Arbaaz Khan "Crop Price prediction using Random Forest and Decision Tree Regression ".
https://www.irjet.net/archives/V7/i9/IRJET-V7I938.pdf

[4] M. Kalimuthu, P. Vaishnavi, M. Kishore: "Crop Prediction using Machine Learning" https://ieeexplore.ieee.org/abstract/document/9214190

[5] Raissa Uskenbayeva; Zhyldyz Kalpeyeva; Ryskhan Satybaldiyeva "Applying of RPA in Administrative Processes of Public Administration".
https://ieeexplore.ieee.org/abstract/document/880779

[6] Piotr Andruszkiewicz; Henryk Rybinski " Data Acquisition and Information Extraction for Scientific Knowledge Base Building".
https://ieeexplore.ieee.org/document/8334468

[7] Y. Yang, L.T. Wilson, J. Wang "Development of an automated climatic data scraping, filtering, and display system" https://www.sciencedirect.com/science/article/abs/pii/S0168169909002348

[8] Pedro Martins; Filipe Sa; Francisco Morgado; Carlos Cunha "Using machine learning for cognitive Robotic Process Automation (RPA)".
https://ieeexplore.ieee.org/abstract/document/9140440

[9] World Bank Open Data https://data.worldbank.org/

[10] Food and Agriculture Organization of the United Nation http://www.fao.org/home/en/

[11] Annual Report 2016-17, Department of Agriculture and Cooperation, Ministry of Agriculture & Farmer Welfare, Government of India, available: http://agricoop.nic.in/sites/default/files/Annual_rpt_201617_E.pdf

[12] Open Government Data(OGD) Platform India

[13] TerraClimate- Climatology Lab

[14] How Does RPA Work? | What Is Robotic Process Automation (RPA)? https://www.youtube.com/watch?v=5fn8QcIGJc0

[15] Robotic_Process_Automation https://en.wikipedia.org/wiki/Robotic_process_automation

[16] MachineLearning https://en.wikipedia.org/wiki/Machine_learning

[17] UiPath Academy
https://www.uipath.com/rpa/academy