The International Conference on Research Perspectives : IoT in Hybrid Grid Integrated Renewable Energy Sources
In association with International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)

# Prediction of Crop Yield and Cost by Finding Best Accuracy using Machine Learning Approach

**Swathi, Mrs. Soja Rani**
Department of CSE, New Horizon College of Engineering, Bangalore, Karnataka, India

## ABSTRACT
Among around the world, agribusiness has the significant duty regarding improving the financial commitment of the country. Still the most agrarian fields are immature because of the absence of arrangement of biological system control advances. Because of these issues, the yield creation isn't improved which influences the farming economy. Subsequently an improvement of rural profitability is upgraded dependent on the plant yield expectation. To forestall this issue, Agricultural areas need to anticipate the yield from given data set utilizing AI procedures. The outcomes show that the viability of the proposed AI calculation strategy can be contrasted and best exactness with accuracy.

**Keywords :** Dataset, Crop Yield, Machine Learning- Classification Method.

## I. INTRODUCTION

In agricultural nations, cultivating is considered as the significant wellspring of income for some individuals. In current years, the rural development is locked in by a few advancements, conditions, methods and civic establishments. Moreover the usage of data innovation may change the state of dynamic and in this manner ranchers may yield the most ideal way. For dynamic cycle, information mining procedures identified with the horticulture are utilized. Information mining is a cycle of separating the most huge and helpful data from the gigantic measure of data sets. These days, we utilized AI approach with created in harvest or plant yield forecast since agribusiness has distinctive informationlike soil information, crop information, and climate information. Plant development expectation is proposed for checking the plant yield adequately through the AI strategies. It is additionally relevant for the computerized cycle of cultivating is the start of another time that will be reasonable for the ranchers who look for specialists to take proposal about the fitting yield on explicit area of their property and don't have any desire to fail to remember any progression of the development all through the cycle.

## II. CROP YIELDPREDICTION

Crop yield prediction that is a fundamental endeavor for the pioneers at public and typical levels for enthusiastic dynamic. Careful gathering of yields understand model can help farmers with picking what to make and when to make. There are different approaches to manage administer oversee crop yield

assumption. Different advances that can be associated with crop yield forecast are information obtaining, information pre-handling, highlight choice, order and forecast. The harvests that were considered in the model for forecast incorporate coriander, beats, cotton, paddy, sorghum, groundnut, sugarcane, banana and vegetables. Various credits of the dirt were considered in request to anticipate the harvest, which included pH, profundity, disintegration, surface, waste, dater holding also, soil tone. The arranged work presents effective degree affordable crop proposal framework. Utilization of guileless mathematician makes the model horrendously practical as far as calculation.

## III. DATASETS

A mix of models is a dataset and when working with AI procedures we normally need a couple datasets for various purposes. Preparing Dataset: A dataset that we feed into our AI check to set up our model. It could be known as the underwriting dataset.

### A. Preparing the Dataset

The demo dataset is right now gave to AI model dependent on this educational assortment the model is readied. Each new detail involved at the hour of design goes probably as a test instructive assortment. After the movement of testing, model estimate reliant on the inferring it closes dependent on the readiness educational assortments. Satellite Imagery (Remote Sensing Data), has been comprehensively used for predicting crop yield. This dataset is accumulated using the sensors mounted on satellites or planes, which recognize the energy (electromagnetic waves), reflected or diffracted from surface of the earth. Inaccessible distinguishing data has a lot of energy gatherings to bring to the table, anyway basically relatively few of them have been used for crop yield conjecture. In any case, there are a couple of gathering who have made a pass at creating material features using the gatherings which are typically dismissed, and they have been productive with improving results with that. In case of this dataset, by far most only sometimes research the high- demand depictions of the features. Considering these datasets people have used computations like Regression models, Random Forest and Nearest Neighbor, etc

| Variable | Description |
|---|---|
| Crop | Crop name |
| State Name | Indian state name |
| District Name | District name list of each state |
| Cost of Cultivation | Cultivation amount for C2 Scheme |
| Cost of Production | Production amount for Scheme |
| Yield (Quintal/ Hectare) | Yield of crop |
| Crop year | Crop year list |
| District Name | District name for each state |
| Area | Total area of each place |
| Rainfall | Water availability of each crop |
| Average Moistness | Straightforwardly impacts the water relations of plant and by implication influencesleaf development |
| Mean Temperature | Climate of r each particular crop |

**Fig.1. Table shows details of the datasets**

### B. Exploratory Data Analysis

In this piece of the information, you will stack in the data, check for cleanliness and a short time later trim and clean your dataset for examination. Guarantee that you fileyour methods carefully and legitimize your cleaningdecisions.

### C. Training Dataset:

The first line imports iris instructive assortment which is as of now predefined in sklearn module. Iris enlightening assortment is basically a table which contains information about various combinations of iris blooms.

- For model, to import any computation and train_test_split class from sk-learn and numpy module for use in this program.
- This procedure secludes dataset into getting ready and test data discretionarily in extent of point we encapsulate any computation.
- In the accompanying line, we fit our planning data into this estimation so PC can get readied using this data. By and by the readiness part is done.

### D. Testing dataset:

- Now we have measurements of another bloom in a numpy cluster called 'n' and we need to foresee the types of this blossom.
- We do this utilizing the foresee technology which accepts exhibit info which lets active objective incentive as yield.
- Objective worth recovers out to be 0. At last discovery grade which is the respective no. of expectations discovered right complete forecasts made.
- This utilizing the technique that fundamentally analyzes an real estimations of the test set with the anticipated qualities.

## IV. MACHINE LEARNING

Machine learning is concerned with learning the computer programs and improve automatically with the experience. However there is no clear idea about how to make computers learn as we humans do but there are several algorithms that are used for certain type of learning task. From age old year's humans worked under every different categories as time passed machines came into use which was trained with algorithms and spoon fed how to work then complete a given task and now it is the decades of the concept ML where in machines are not trained how to work in real but allowed itself to learn from experience using the input, analyze the data by itself and give the results. The input can be divided into 2 sector training data and testing data. Training data is used to guide the system; the system will learn from these data and will produce the output. Later comes the role of testing data where the trained system is checked for its correctness. Machine learning algorithms also performed in the field called data mining to determine the knowledge from the databases, loan transactions, and financial transactions.

### A. Types of Machine Learning

There are three sector of ML techniques which is based on the input type, input length, time duration given to solve etc.

#### ❖ Supervised Learning

In this technique set of data in the form of examples need to be provided. Each example will be associated with labels; a learning algorithm is to be given to example. It is possible for this technique to predict the label for the examples that never seen before when it is fully trained. Since supervised learning focuses on singular task, so we can feed more and more examples until it performs the task accurately.

Supervised learning can further be divided into two kinds of problems regression problems and classification problems. Regression problems have to find or draw a linear line (in case of linear regression) which classifies that given data maximally correct. As the name suggests classification problem has to classify the given input correctly based on the other inputs whose classification is previously known. A simple example is given input data of which is the suitable day to play tennis in the form of attributes such as temperature, wind, outlook etc., and giving the testing input to check if tennis can be played on a particular given day.

### ❖ Unsupervised Learning

It is too sensitive converse to the supervised learning because we need to provide lot of data to understand its properties, as the data are unlabeled Unsupervised learning is said to be data driven because it is mainly based upon data and its properties. The outcomes are based on data and the way they are formatted. As in supervised learning the input data consists of inputs and also its possible classification (in terms of target attribute) the unsupervised learning does not consist of target attribute, only input data is given and based on the common things/ similarities the system should group the data and provide the output. Example is given the input of number of patients who are having diabetes, attributes such as their age, height, sex, blood pressure etc., are also given, now the system has to learn by itself from these data and when a testing input is given it should predict whether that person has diabetes or not.

### ❖ Reinforcement Learning

Let us consider a simple example of playing checkers game, in this game we make several moves and at the end the result is either win, loss or a draw. Let us assume that in the first game you lost the game so while playing second round you will think which are the moves you took in the first game which led to failure and change your some of the moves accordingly (we cannot say that all the moves made in the first game is wrong, some moves made may be the reason for the failure) and eventually win the game. So Reinforcement Learning is somewhat different from supervised and unsupervised learning as it will learn from mistakes. For support learning, we need a specialist and a climate. To associate the climate and the specialist we need to give two signs: refreshed state and award.

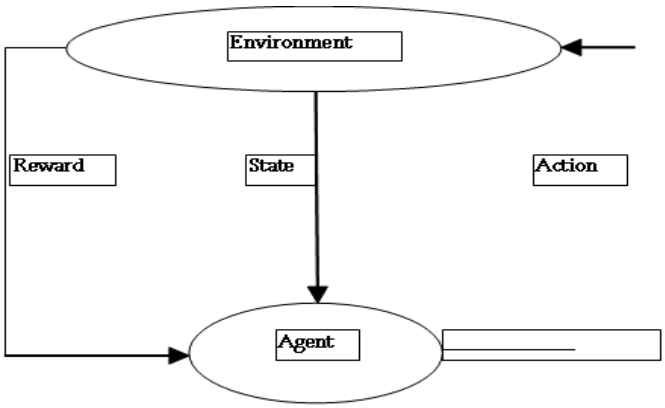


Figure 2: Reinforcement Learning

In the above diagram it is clearly explained how a reinforcement learning works, the environment gives the changes in the environment as input to the learning system called agent, the agent using its previous experience takes an action and depending on its decision the environment either provides rewards or blames the system. This is represented in equation form.

### B. Designing a learning system

- Choosing the preparation experience: this is initial step included where we will pick the preparation information. This preparation information will have an impact on progress or disappointment of the student.

There are several key attributes. The first key attribute is providing a direct or indirect feedback about the choices made. Second is how much the student controls succession of the preparation models.

The third characteristic is the means by which well the circulation of the models over the last framework execution is addressed. That is if the preparation models follow the conveyance like the future test models then the learning is more solid.

- Choosing the target function: here we need to think about what kind of knowledge is gained and used by the performance system. There will be several optimization problems encountered such as scheduling and controlling manufacturing processes where the steps are understood but the strategy for sequencing is not.
- Choosing a Representation for the target function: We need to select a representation that will be crucial and very expressive such that it is almost approximate to the ideal objective capacity. Assuming the portrayal is expressive, we can pick among the elective speculation it can address.
- Choosing the capacity guess calculation: We require a bunch of preparing guides to gain proficiency with the objective capacity to do this we need to initially appraise the preparation esteems and afterward change the loads appropriately so it best fits the preparation models and diminishes the squared mistake between the preparation esteems and the qualities that are anticipated.
- Final Design: It contains four distinct component
  - The performance system: It takes new problem instance as input and output as a trace of solution. It is mainly used to solve given performance task .The performance should increase as the evaluation function becomes more accurate.
  - Critic: takes the input as history and output as set of training examples. Critic is nothing but it corresponds to the training rule.
  - Generalizer: The input is a training examples obtained from the critic and the hypothesis is the output. The hypothesis will be generalized such that it covers all the training examples and the cases beyond the training examples.
  - Experiment Generator: input will be a current hypothesis and the output will be new problem for the performance system in this manner it increases the learning rate of the system

### C. Commonly used Machine Learning algorithms

Some of the most commonly used ML algorithms are

1. **Decision Trees**
2. **Support Vector Machines(SVM)**
3. **Naïve Bayes**
4. **K-Nearest Neighbors (KNN)**
5. **K-Means**
6. **Random Forest**

#### 1. Decision Trees

The most frequently used supervised learning algorithm is decision tree. It works well for both categorical (discrete) as well as for continuous dependent variable. The various distinct groups are made. In simple words the structure is similar to a tree present in data structures. The tree consists of a main node *Root* internal nodes which is nothing but attributes given in input labeled based on some given conditions( test on the attribute) and finally the last nodes are the leaves which is the target concept / possible outputs for a given problem statement, branches are the values obtained after the attribute is tested.

#### 2. Support Vector Machines

It is a support learning algorithm. In this modular the points will be represented in N dimensional space so that examples with different features will be clearly separated by a large gap. The new instances will be then classified to the category in which there is small gap. SVM can easily perform both linear and non-linear classification using kernel trick that is it maps

the inputs to high dimensional output. Nonlinear SVM is nothing but the boundary that the algorithm calculates is not a straight line. SVM is very much suitable in the following cases:

- When there is huge number of training data.
- When the number of zero values are more.
- To solve problems like image classification, genes classification etc.

There are several drawbacks as well:

- Input data has to be labeled.
- Probabilities of estimated data has finite data will not be estimated.
- It is difficult to interpret the parameters of solved model.
- SVM is applicable only for class two tasks. If there is multi class task binary problems have to be applied.

### 3. Naïve Bayes

It is one of the highly sophisticated classification methods very much useful for larged at a set. It is mainly based on Bayes' Theorem. Naive Bayes classifier assumes that each of the feature is independent of the other. The Naive Bayes classifier will consider all properties contribute independently to the probability even if the features are related on each other.

### 4. K-Nearest Neighbors

It is commonly used to solve the classification and regression problems. KNN algorithm will store all the available cases and when a new case is encountered the case will be assigned for class which is pure common among the K nearest neighbor calculated by Distance function. There are various distance functions such as Euclidian, Manhattan and Murkowski used for similarity of function and hamming distance used for particular sector variables.

### 5. K-Means

It is unsupervised centroid based algorithm used for clustering. It mainly aims at forming a k clusters from observations. Clustering is done based on the nearest mean value. In order to pause the clustering in K-means algorithm

- Do not change the centroid of newly formed cluster.
- All the points should remain in the same cluster.
- Maximum iterations has to be reached.

### 6. Random Forest

Random forest is a collection of decision tree mainly used for classification and regression. It is supervised algorithm that doesn't over fit the model, handles the missing value and mainly modeled for categorical value. It is better than decision tree because it contains only the subset of features

## D. Disciplines of machine learning

### ❖ Artificial intelligence

Symbolic representation of concepts are learnt using artificial intelligent. It is also used as approach to improve problem solving. Using prior knowledge along with training details used as guide for learning.

### ❖ Bayesian method

The best hypothesis which is most probable can be determined by using Bayes' theorem. It is important in machine learning because it provides quantitative approach in identifying the evidence supporting alternating hypothesis. It provides basis for learning algorithm as it will directly manipulate probability and provides framework for analyzing the operation of algorithm.

### ❖ Computational complexity theory

The multifaceted design of different learning task are assessed by the computational effort, bungles made, planning models, etc

❖ **Control theory**

To predict the next state of the process and also to optimize the predefined objectives we need to learn the control process Information theory learning to provide minimum description length. For encoding hypothesis with optimal codes and their relationship to optimal training sequences.

❖ **Philosophy**

To determine the best hypothesis that is simple. To generalize the observed data analyzing the justification.

❖ **Psychology and neurobiology**

According to the power law for a very wide range of learning problem response time of people increases with practice. Artificial neural network models of learning are motivated by Neurobiological studies.

❖ **Statistics**

When estimating the accuracy of the hypothesis based on the data given characterization of errors may occur.

**E. Applications of machine learning**

❖ **Recognition of spoken words:**

SPHINX system is one such that recognizes the primitive sounds and words from the speech signal and learns the speaker specific strategies. There are certain methods that are effective for customizing individual speakers, vocabularies, background noise such as methods for hidden Markov models, Neural network learning methods.

❖ **Driving an autonomous vehicle:**

The PC controlled vehicles are prepared to guide accurately when driving on assortment of street types. ALVINN framework is trained to mimic the steering commands of human driving the vehicle and it was successful in driving on public highways at speed up to 70 miles per hour and for a distance of 90 miles.

❖ **Classifying the astronomical structure**

In order to classify the celestial object decision tree learning algorithm have been used by NASA. It will automatically classify objects in the sky survey.

## V. REFERENCES

[1]. Guanghui1,DongryeolRYU2,∗,JIAOCaixia1 and HONG Changqiao1 Estimation of Organic Matter Content in Coastal Soil UsingReflectance Spectroscop Research 2015

[2]. ManashProtimGoswami , BabakMontazer, and UtpalSarma, Member, IEEE Design and Characterization of a Fringing Field CapacitiveSoil MoistureSensor 2018

[3]. Zhiqiang Cheng, JihuaMeng * and Yiming Wang Improving Spring Maize Yield Estimation at Field Scale by Assimilating Time-Series HJ-1 CCD Data into the WOFOST Model Using a New Method with Fast Algorithms2016

[4]. Becker-Reshef, E. Vermote, M. Lindeman A generalized regression-based model for forecasting winter wheat yields in Kansas and Ukraine using MODIS data2010

[5]. Kim, NariLee,Yang-Won Machine Learning Approaches to Corn Yield Estimation Using Satellite Images and Climate Data: A Case of Iowa State 2016

[6]. SudhanshuSekharPanda, GerritHoogenboom, and Joel O.Paz Remote Sensing and Geospatial Technological Applications for Site-specific Management of Fruit and Nut Crops: A Review 2010

[7]. K. Sathishkannan; G.ThilagavathiOnline farming based on embedded systems andwireless sensor networks2013

[8]. B. Abishek; R. Priyatharshini; M. AkashEswar; P. DeepikaPrediction of effective rainfall and crop water needs using data miningtechniques 2017

[9]. Crop Yield Estimation Using Time-Series MODIS Data and the Effects of Cropland Masks inOntario, Canada 2019

[10]. Sheena Angra; SachinAhujaMachine learning and its application2017

[11]. FilippoSciarroneMachine Learning and Learning Analytics: Integrating Data with Learning2018

[12]. Pavan Patil1, Virendra Panpatil2, Prof. ShrikantKokate Crop Prediction System using Machine Learning Algorithms2020

[13]. Md. TahmidShakoor, KarishmaRahman, Chakrabarty.2017."Agricultural Production Output Prediction Using Supervised Machine LearningTechniques".978-1-5386-3831-6/17/$31.00 ©2017 IEEE

[14]. I. Ahmad, U. Saeed, M. Fahad, A. Ullah, M. Habib-ur-Rahman, A. Ahmad, J.Judge Yield forecasting of spring maize using remote sensing and crop modeling in Faisalabad-Punjab

[15]. S. Pudumalar, E. Ramanujam, R. H. Rajashree, C. Kavya, T. Kiruthika and J. Nisha, "Crop recommendation system fo r prpecision agriculture," 2016 Eighth International Conference on Advanced Computing (ICoAC), Chennai, 2017, pp. 32-36. doi: 10.1109/ICoAC.2017.7951740.

[16]. Naive Bayes classifier available at https://en.wikipedia.org/wiki/Naive_Bayes_classifier

[17]. S Brunda1, Nimish L2, Chiranthan S2, Arbaaz Khan2 Cro p Price prediction using Random Forest and Decision Tree Regression2020

[18]. s://medium.com/swlh/random-forest- and-its-implementation-71824ced454f

[19]. Y. Peng, C. Hsu and P. Huang, "Developing crop price forecasting service using open data from Taiwan markets," 2015 Conference on Technologies and Applications of Artificial Intelligence (TAAI)

[20]. Ananya Roy; Prodipto Das; Rajib DasTemperature and humidity monitoring system for storage rooms of industries2017