International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 04 | April - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

Crop Yield Prediction Using Machine Learning Algorithms

Ajithkumar.A [1] DheenaHarrish.K[2], Jagadeesh.R[3], Mukilraja.V [4], Mrs.P.Premadevi [5]

1,2,3,4B.E, Student Department Computer Science and Engineering, Angel College Of Engineering and Technology,Tiruppur , TamilNadu, India

5Asst.Professor, Department Computer Science and Engineering, Angel College Of Engineering and Technology, Tiruppur,Tamilnadu, India

Abstract— Among worldwide, agriculture has the major responsibility for improving the economic contribution of the nation. However, still most agricultural fieldsare underdeveloped due to the lack of deployment of ecosystem control technologies. Due to theseproblems, the crop production is not improved which affects the agriculture economy. Hence a development of agricultural productivity is enhanced based on the plant yield prediction. To prevent this problem, Agricultural sectors have to predict the crop from given dataset using machine learning techniques.The examination of dataset by coordinated ML techniques. A comparative study between machine learning algorithms had been carried out in order to determine which algorithm is the most accurate in predicting the best crop .In this we are going to predictthe yield if a specific crop is selected else we will predict the yield ofall the crops using the parameters District name, season and year.

Keywords : dataset, Machine Learning- Regression methods, mean absolute error, R2- score

INTRODUCTION

In our research, we found that most of previous papers used climatic factors like rainfall, sunlight. Whenfocused based on soilthey used parameters like soil type, soil PH, crop sensitivity etc., Regarding rice crop prediction they used parameters like soil, sunshine,fertilizer, temperature, paddy rainfall and pest.

As these terms are bit difficult to explain to the farmersbecause hemay not be able to understand these terms so to make this process easy, we are making an application which uses parameters likedistrict name, season and year and it predicts the crop yield this make farmers .

Agriculture is the backbone of every economy. Agriculture is considered as the main and the foremostculture practiced in India. The main goal of agricultural planning is to achieve maximum yield rateof crops by using limited number of land resources.

Many machine learning algorithms can help in improving the production of crop yield rate.

Whenever there is a loss in unfavorable conditions. We can apply crop selecting method and reduce the losses. The maximizing of yield rate helps in improving economy. We have observed that there is an increase inthe suicide rates. So, we want to help the farmers to understandthe importance of prior prediction of crop, to increase their knowledge about quality of soil, to understand location wise weather constraints, in orderto gain intense yield of crop through our technology solution.

A website created as part of the "Smart Farming using Machine Learning" project helps farmers by predicting thecrop that will be grown. This calls for specific conditions including temperature, precipitation, and soil moisture. The suggested systemspecifies the kind of crops a farmer may raise on his property. A suitable dataset that describes the best crop is required for the crop prediction process in order to reduce the likelihood of crop failure. Another thing to keep in mind is that it uses technology for prediction more and does not requirea lot of human resources.

About 70% of Indians work in the agricultural sector,which is why it was important to include it in our analysis of the economy of the nation. Crop yield prediction is a huge problem in theagricultural sector. Crop prediction is the process of figuring out what thefarmer can grow. Building a system that would operate with maximum accuracy and take into account all significant variables that can affect the outcome of the crop prediction is imperative. Numerous studies have been conducted to forecast thecrop that a farmer can grow. Most of the farmers try to know crop yield and whether it meets their expectations. They evaluatethe previous experience of the farmer on a specific crop yield.

INTRODUCTION TO ML.

Agribusiness is partner exchange area that is profiting intensely from the occasion of finder innovation, information science, and AI (ML) methods inside the most recent years. Theseimprovements get back to fulfill ecological and populace pressures round-looked by our general public, any place reports show a necessity for powerful worldwide agribusiness yield increment to create nourishment for a developing populaceon a more sweltering planet. The vast majority of the work tired the area of yield predicting through cubic centimeter utilizes some sort of far-off detecting information over the homestead..

Agribusiness looks to broaden and improve the cropyield and hence the nature of the yields to support human existence. Nonetheless, inside the current time, people will in general require a great deal of like a shotappreciated positions. There are less, and less people worried incrop development. moreover, the consistent increment of human populace makes the development of the yields at the legitimate time and opportune spot evena great deal of crucial, in light of the fact that the environment is dynamic and accordingly the movements from conventional climate designare a ton of continuousthan before make.

The data hole between antiquated ways that of developing and new agrarian advancements might be survived if the PC code might be intended to show the intelligent effect of environment factors,especially the effect of greatest occasions (for example warmth, rainfalls andoverabundance water) happening at totallyextraordinary developing periods of crops.

The temperature change without a doubt influences the local and world food creation, along these lines arranging PC code to show crop forecasts needs new strategy for for temperature change examines, circumstances for temperature change transformation, and policymakers which will restrict the overwhelmingimpacts of climate on food give. The dirt sort will adjustment after some time due to climate and vermin, consequentlycrop the executives should deal with an extravagant amount . temperature change examines, circumstances for temperature change transformation, and policymakers which will restrict the overwhelming impacts of climateon food give. The dirt sort will adjustment after some time due to climate and vermin, consequentlycrop the executives should deal with an extravagant amount of data straihght forwardly or by implication related to each other . it will along these line by thinking about a work on the real world , to allow a brisk appraisal of theeffect of temperature change in agribusiness

Farming ought to adjust to those environment changesand it will do it will do accordingly bycreating models which will in principle enhance the executives rehearses, augment the turns of the it will do accordingly bycreating models which will in principle enhance the executives rehearses, augment the turns ofthe occasional environment changes might bediscovered and recorded in an incredibly convenient way. Later on, by exploitation PC code upheld AI, one will conveniently evaluate the temperature change effect and check achievable circumstances that consolidate found out changes in climatic conditions and water dispersion.

Information processing} is that the way toward dissecting the test information gathered overa sum and changed areas from totally various separate patterns or examples of information of information on info} and switch them into supportive data for clients. The examples, affiliations, or connections among this information will extra be reawakenedinto data that is offered to the client The examples, affiliations, or connections among this information will extrabe reawakened into data that is offered to the the client as recorded examples and future patterns. This data given by AI will encourage ranchers with crop development by foreseeing probabilities of crop misfortunes or stop misfortunes out and out

Yield forecast benefits the ranchers in diminishing their misfortunes and to get best costs for their harvests . Agriculture is most significant occupation drilled in our country. It is the biggest financial area and assumes a significant part in by and large improvement of the country .

significant issue in horticultural arranging. Yield creation rate relies upon topography of an area, climate condition, soil type and gathering techniques. Various subsets ofthese impacting boundaries are utilized for various yields by various expectation models. AI strategy is utilized for expectation of harvest yield utilizing various calculations. AI strategies which are broadly utilized in expectation strategy are relapse tree, irregular timberland, convolution neural organization and K-closest calculation. There is consistently a huge danger factor for the ranchers to choose which specific yield oughtto develop during season, on specific landasset. It relies upon various boundaries, for example, creation rate, cost and diverse government approaches. Independent of the capital put as far as soil, water and nature of seeds of the yield rate creation the harvest maybomb carrying shocking misfortunes to the rancher and his family

In this pre-owned AI approach with created in harvest or plant yield expectation sincehorticulture has diverse information like soil information, crop information, and climate information. Plant development forecast is proposed for checking the plant yield viably through the AI methods.

II RELATED WORK

Virendra Panpatil et al[1] had made huge work for Indian farms by building profitable yield recommendation schema.

The planned schema is used find the bestseason for planting, advancement of plant and Plant fulfillment. They used discrete classifier for generating better correctness for occurrence. The best agreeable area of framework that it can without much of a extent all-around all things be used to check on various yields.

Mayank et al[2] created improvised schema for crop yield using executed AI computationsand with target to give effortless to use User Interface, rise the exactness of crop yield estimate, examining discrete climatic parameters.

Zhihao et al[3] two relapse administered AI strategies SVM, RVM to show viability in soilquality forecast. a shrewd remote gadget for detecting soil dampness and meteorological information.

Sabri Arik et al[4] includes a analysis regarding Soil Fertility , Plant Nutrient by utilizing back spread computation . The outcomes are exact and empowers advancement in soil properties.

It works best when contrasted with conventional techniques. Be that as it may, framework is moderate wasteful and not steady.

Shivnath et al[5] proposed about BPN to assess the test informational collection. BPN utilizes a concealed layer which supports in better execution in foreseeing soil properties. BPN present, is utilized to build up a self- prepared capacity to foresee soil properties with boundaries.

This results in more exactness and executes better compared to the customarily utilized strategies, in any case, at times theframeworkturns be moderate and irregularityis found in the yield.

Raval et al[6] examine regarding the Knowledge Discovery Process and the rudiments of different Data Mining approaches, for example, Association rules, Classification etc.,

Agrawal et al examine regarding different Data Mining devices, for example, Dashboards, Text-Mining instruments. They give an outline about these apparatuses and thedifferent situations where they can be conveyed.

Grajales et al[7] recommended a web app that uses open dataset like verifiable creation,land cover, nearby environment surroundings and coordinates them to give simpleadmittance to the ranchers. The suggested

engineering basically centers around open-source devices for the advancement of the application. The client can choose areafor which the subtleties are accessible at a single tick.

Bendre et al[8] gathers information about GIS, GPS, VRT and RS are controlled utilizing Map Reduce calculation and direct relapse calculation to figure the climateinformation that can be utilized in accuracy agribusiness.

Verma, A. et al[9] used classification techniques for crop prediction like Naïve Bayes, KNN algorithm on soil datasetswhich consists of nutrients of soil like zinc, Phosphorous , copper , pH, iron, Sulphur, manganese, Organic Carbon, nitrogen, andpotassium

Chakrabarty, A. et al[10] made crop prediction in Bangladesh. They considered parameters like soil composition, type of fertilizer, type of soil and its structure, soil consistency, reaction and texture

III PROPOSED SYSTEM

In this proposed work we need to loadthe information, check for null and duplicate values, and then we trim and clean our dataset for examination. We should makesure that werecord our means carefully and legitimize your cleaning preferences. This is the most energizing stage in Applying Machine Learning to any Dataset. It is otherwise called Algorithm choice for Predicting the best outcomes. Usually, Data Scientists use various kinds of ML calculations to the huge data collections. Yet,at significant level every one of those various calculations can be of these two types of gatherings: regulated and solo learning. Managed learning: supervised learning is a schema where the necessary information and needed yield information is also given. Info and yield information are usedfor future data handling. Learning issues can be of two types Regression and Classification issues

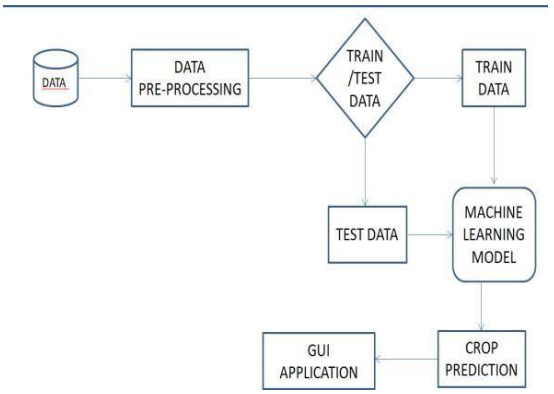


Fig. Architecture of proposed Model

3.1 TRAIN THE DATASET

At First, we import all the required modules. Then we use appropriate cleaning methods to clean our data. Then we split our data according to the required output. We split ourdataset into train and test data using train_test_split technique. The X prefix indicates the element esteems and y prefix indicates target esteems. This splits the dataset into train and test information in the proportion which we mentioned. Here we are using a proportion of 80:20. At this point we epitomize any calculation. Here we fit our preparation information into this calculation so our system can get prepared utilizing this information. Here the training part is finished.

3.1 TEST THE DATASET

To get confidence in our trained model weusesome data of the dataset as test data. Herewe have divided the dataset into train and testdatain 80:20 proportion. We predict the yield values with the trained model by sending the feature values from testing data. We now compare the predicted values with the actual values using metrics such as R2 score, mean squared error etc. these metrics tell us how close we are able to predict the actual data. Mean squared the low the better. R2 score the higher the better.r2 score calculates the

correlation between predicted and actualvalues.

IV MODULES DESCRIPTION

4.1 DATA VALIDATION ANDCLEANING

4.1.1 DATA SET DESCRIPTION

We took our dataset from dataworld website. dataset has more than 250000 data points. As machine learning models work better when wehave more data this dataset having large data is very useful. Dataset has the following attributes state, district, area, production, season, year.

4.1.2 DATA CLEANING

From the available attributes we drop the state attribute as we will use district. As state is dependent on the district, we don't need it for training. We are dropping rows whichhavenegative values for area and production as these are supposed to positive and hence are inconsistent. We also drop any na,nan and empty values before training as these rows could disrupt training. For the target of our training model, we use a derived attribute production Per Area as we are interested in yield per area. Hence, we drop productionand area as these are included in our derived target attribute Production Per Area.

To get confidence in our trained model we use some data of the dataset as test data. Herewe have divided the dataset into train and testdata in 67:33 proportion. We predict the yield values with the trained model by sending the feature values from testing data. We now compare the predicted values with the actual values using metrics such as R2 score, mean squared error etc. these metrics tell us how close we are able to predict the actual data. Mean squared the low the better. R2 score the higher the better.r2 score calculates thecorrelation between predicted and actualvalues.

4.2 DATA NORMALIZATION

4.2.1 PREPROCESSING AND SCALING

To train data with string datatype we use label encoder to convert them to numeric data.Label encoding simply assigns each data point a number starting from .the attributes weare label encoding is district name, season andcrop. Then to scale each data attribute we use two types of scaling techniques which are MinMax transform and normalization. The attributes district name,season, crop is scaled by MinMax transform and production per area is scaled by normalization. MinMax transformis done by subtracting each datapoint from minimum among datapoints then dividing by maximum- minimum. normalization is done by dividing datapoint with standard deviation. We choosenormalization to productionperarea as many data points are low and min max transform is yielding good results.

4.3 RANDOM FOREST

It is a very famous ML algorithm that felicitates in cases of both classification and regression issues. This algorithm works on thenotion of ensemble learning, which workson the principle of merging several classifiersto give the solution for any tough problems and increase the precision and performanceof the applied model. Random Forest is aclassifier orregressor that contains multiple decision treeson several subsets of a given dataset and considers the average to improve the dataset'shigh accuracy in case of classification problems in case of regression problems it helps us to lower the error value. In other words, rather than depending on a single decision tree, this algorithm considers predictions from every tree and predicts the final result based on the predictions. On observation we can say that:

$$No.\ of\ Trees\ in\ forest a\ Accuracy\ of\ the\ model$$

4.3.1 Decision Tree

It is a very famous ML algorithm that is used incases of both classification and regression issues. Itconsists of three nodes. The First node which is the initial node is the root node. The Interior nodes describes the features of the data set where as the branches represent the decision rules. Finally, the leaf node shows the result.

4.4 K-NEAREST NEIGHBOR (KNN)

It is a regulated AI figuring which keeps all events identify with planning data centers in n-dimensional spaces. It separates the closest k number ofcases saved and gives the most broadly perceived class as the assumption and for authentic regarded data it reestablishes the mean of k nearest neighbors. In case of distance weighted nearest neighbor it stacks the responsibility of all of the k neighbors based on their distance using the going with request giving more critical burden to the closest neighbors. Makes assumptions regarding the endorsement set using the entire getting ready set. KNN makes an estimate about another case by means of glancing through the entire setto find the k "closest" events. "Closeness" is settled usinga proximity assessment across all features.

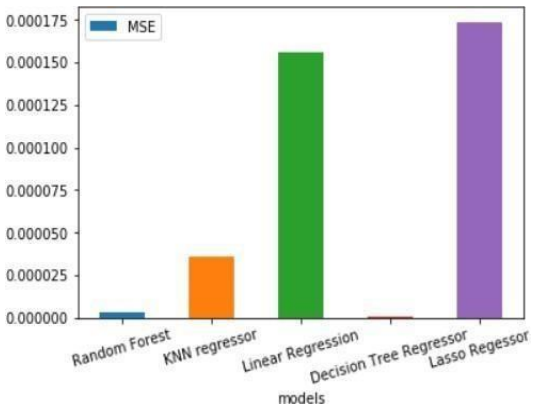
4.6. LASSO REGRESSION

Lasso regression a kind of linear regression thatuses depreciation, i.e., data points are diminished towards a mid-point, as done when finding the mean. This model is particularly useful and suits well for models that show high levels of multicollinearity as is the case in this particular dataset. It executes L1 regularization. Lasso solutions are quadratic programming problems, that use the following formula:$\Sigma(yi - \Sigma xij\beta j)2 + \lambda\Sigma|\beta j|pj=1\ jni=1$. The intensity of the L1 penalty is controlled by a tuning parameter. is essentially the amount of shrinking. If it is zero, we knowthat all features are taken into account, and it is equivalent to linear regression, in which just the residual sum of squares is used to form a predictive model. If it is infinity, it means that no features are taken into account. The bias increases while the variance decreases with an increase in λ, and vice-versa

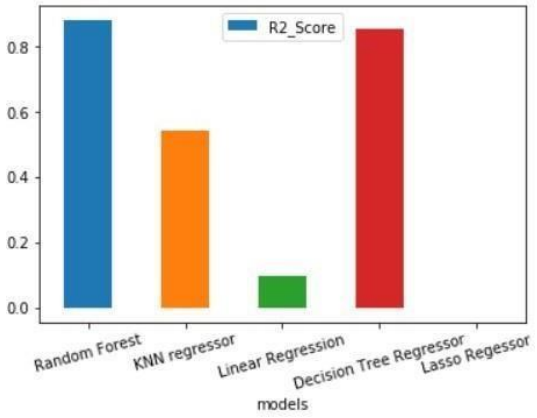
V. RESULT AND DISCUSSION

	MSE	MAE	RMSE	R2_Score	models
0	3.426823e-06	0.001505	0.004558	0.880158	Random Forest
1	3.595377e-05	0.003637	0.008913	0.541729	KNN regressor
2	1.557740e-04	0.006636	0.012498	0.098844	Linear Regression
3	1.761657e-07	0.001490	0.005049	0.852930	Decision Tree Regressor
4	1.737817e-04	0.007141	0.013166	-0.000002	Lasso Regessor

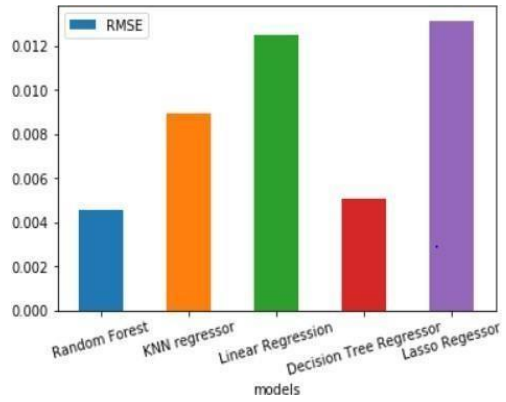
MSE PLOT



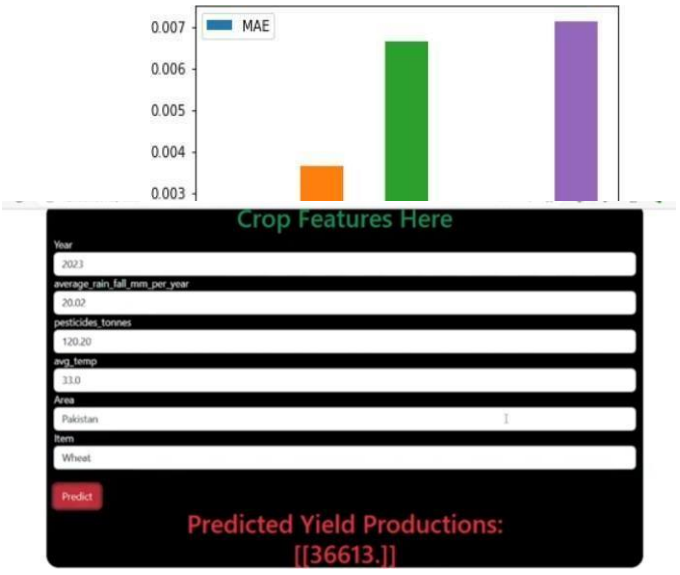
R2_Score plot



RMSE PLOT



MAE PLOT



CONCLUSION

The logical connection started from data cleaning, visualizing the data preprocessing then training and testing the dataset examination finally model design and evaluation. Finally, we anticipate the yieldusing ML algorithms with different results. This brings a bit of the going with encounters about yield gauge. As most prominent sorts ofyields will be covered under this structure, farmer may turn out to be more familiar with about the collect which may never have been created and penetrates down each possible yield, it helps the farmer in unique of which respect create.

It gives the yield if a specific crop is given elseit will give the yield of all crops.

REFERENCES

[1] ARTIFICIAL INTELLIGENCE AND INTERNET OF THINGS FOR SUSTAINABLE FARMING AND SMART AGRICULTURE (2023) Ahmad Ai Alzubid, Kalda Galyna

[2] CROP YIELD PREDICTION USING MACHINE LEARNING APPROACHES WITH SPECIAL EMPHASIS ON PALM OIL YIELD PREDICTION (2021) Mamunur Rashid, Bifta Sama Bari, Yusri Yusup,
Mohamad Anuar

[3] PRECISION FARMING FOR SUSTAINABLE AGRICULTURE (2020) ABHILASH JOSEPH E. ,ABDUL HAKKIMV.M, SAJEENA.S

[4] MACHINE LEARNING APPLICATION FOR PRECISION AGRICULTURE(2021) ABHINAV SHARMA,ARPIT JAIN, PRATEEK GUPTA.