A Minor Project Report

On

**Crop yeild prediction using machine learning**

SUBMITTED IN PARTIAL FULFILLMENT FOR THE AWARD OF DEGREE OF

**Bachelor of Technology**

**IN**

**Electronics and Communication Engineering**



**Submitted By:** **Under the Guidance Of**

**naman Sahni (9916102170) Dr. BAJRANG BANSAL**

**Shubham Soni (9916102118)**

**Swapnil srivastav (9916102200)**

**DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING**

**JAYPEE INSTITUTE OF INFORMATION TECHNOLOGY, NOIDA (U.P.)**

**MAY, 2019**

**CERTIFICATE**

This is to certify that the minor project report entitled, “**CROP YEILD PREDICTION USING MACHINE LEARNING**” submitted by **NAMAN SAHNI, SHUBHAM SONI AND SWAPNIL SRIVASTAV** in partial fulfillment of the requirements for the award of Bachelor of Technology Degree in **Electronics and Communication Engineering** of the Jaypee Institute of Information Technology, Noida is an authentic work carried out by them under my supervision and guidance. The matter embodied in this report is original and has not been submitted for the award of any other degree.

**Signature of Supervisor:**

**Name of the Supervisor: Dr. Bajrang Bansal**

**ECE Department,**

**JIIT, Sec-128,**

**Noida-201304**

**Dated:**

**DECLARATION**

We hereby declare that this written submission represents our own ideas in our own words and where other ideas and words have been included have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified and idea/data/fact/source in our submission

**Place:**

**Date:**

Name: **NAMAM SHANI**

Enrollment: **9916102170**

Name: **SHUBHAM SONI**

Enrollment: **9916102118**

Name: **SWAPNIL SRIVASTAV**

Enrollment: **9916102200**

**ABSTARCT**

Wheat is one of the key cereal crops grown worldwide, providing the primary caloric and nutritional source for millions of people around the world. In order to ensure food security and sound, actionable mitigation strategies and policies for management of food shortages, timely and accurate estimates of global crop production are essential. The first step of this project is to develop and evaluate a regression-based model for forecasting wheat production. we have analyzed and used different types of machine learning regression algorithms to predict the yield based on dataset we have collected. The dataset contains different one dependent variable and different independent variable. We have used Polynomial regression, Decision tree regression and Random forest regression to predict the output. We considered dataset of two years and merged them to train and test for our model. After training and testing the dataset to the model we have compared the accuracy of each model individually.

**ACKNOWLEDGEMENTS**

Success is an effort bounded activity that involves cooperation of all. We hereby take the opportunity to express our profound sense of gratitude and reverence to all those who have helped and encouraged us towards successful completion of the project. It has been a great learning experience. We would like to thank our mentor Dr. Bajrang Bansal for his valuable support and guidance, constant encouragement and the opportunity provided to us to complete the project under his supervision. We would also like to express our gratitude to great almighty, our parents and our friends for their concerned support and encouraging words without which this project would not have been the way it is.

**Table of Contents**

Title page**………………………………………………………………………………………**i

Certificate**……………………………………………………………………………………..**ii

Declaration**……………………………………………………………………………………**iii

Abstract**…………………………………………………………………………………….....**iv

Acknowledgements**……………………………………………………………………………**v

**CHAPTER 1: INTRODUCTION………………………………………………………….4**

* 1. General……………………………………………………………………………………4
  2. Motivation ………………………………………………………………………………. .4
  3. Aim & Objective…………………………………………………………………………..4
  4. Introduction to Machine learning………………………………………………………….5
     1. Supervised Machine Learning Algorithms……………………………………………5
     2. Unsupervised Machine Learning Algorithms…………………………………………5
     3. Semi supervised Machine Learning Algorithms……………………………………...5
     4. Reinforcement Machine Learning Algorithms………………………………………..6

**CHAPTER 2: LITERATURE SURVEY………………………………………………….7**   **CHAPTER3: METHOLOGY……………………………………………………………...8**

**CHAPTER 4: PRE PROCESSING…………………………………………………………9** 4.1Dataset Normalization…………………………………………………………………….94.2 Feature Extraction and Correlation……………………………………………………….9

**CHAPTER 5: POST PROCESSING…………………………………………...…………11**

* 1. Model Selection………………………………………………………………………….11
  2. Regression Techniques…………………………………………………………………..11
     1. Polynomial Regression………………………………………………………………11
     2. Decision Tree Regression……………………………………………………………12
     3. Random forest Regression………………………………………………………….13

**CHAPTER 6: TRAINIG AND TESTING OF MODEL…………………………………14**

* 1. Training and Testing of Polynomial Regression Model…………………………………14
  2. Training and Testing of Decision Tree Model…………………………………………...14
  3. Training and Testing of Random Forest Model………………………………………….15

**CHAPTER 7: RESULT AND ANALYSIS………………………………………………..16**

* 1. Accuracy of Polynomial Regression……………………………………………………..16
  2. Accuracy of Decision Tree Regression…………………………………………………..16
  3. Accuracy of Random Forest Regression…………………………………………………17

**CHAPTER 8: CONCLUTION AND FUTURE SCOPE…………………………………18**

**REFRENCES……………………………………………………………………………….19**

**APPENDIX…. ..…………………………………………………………………………….20**

**List of Figures**

Fig 3.1 Block Diagram. Of Methodology**……………………………………………………..**5

Fig 4.1Feature correlation**……………………………………………………………………..**7

Fig 4.2 Dataset in describe from**………………………………………………………………**7

**CHAPTER - 1**

INTRODUCTION

* 1. **General**

The driving force of this project is to predict the production of yield. Accurate and timely estimation of is important for a variety of reasons. On the macroeconomics level they allow societies to understand the food and fiber supply. Yield information gives a farmer a baseline of what is typically expected to be produced and thus can be used to best establish risk, insurance premiums or the value of input costs.

* 1. **Motivation**

Considering the growing volumes and varieties of data available, the need for computational processing is becoming vital to provide deep-rooted information which is cheap and easily available. With the help of A.I. and Machine Learning,[1] it’s possible to automate models which can analyse bigger, complex data to return more fast and accurate results. Organizations are finding profitable opportunities to grow their business by identifying the precise models to avoid unknown risks. The use of algorithms to build a model is helping organizations to bridge the gap between their products and users with better decisions and least human intervention.

Most industries with enormous volumes of data have recognized the value of Machine Learning. By garnering insights from this data, often in real-time, organizations have become more efficient in their workflow and gaining an advantage over other competitors.

* 1. **Aim & Objective – Problem Description**

Objectives of this project is determining the crop yield of an area by taking in the dataset with some features which are important or related to crop production such as temperature, moisture, rainfall, and production of the crop in previous years. Regression models are used to predict a continuous value. It is a supervised technique. While training and building a regression model, it is these coefficients which are learned and fitted to training

data. The aim of training is to find a best fit line such that cost function is minimized. The output function helps in measuring the error. During training process, we try to minimize the error between actual and predicted values and thus minimizing error function. Python is used for this project by using Jupyter platform for simulation.

* 1. **Introduction to Machine Learning**

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves. The process of learning begins with observations or data, such as examples, direct experience, or instruction, in order to look for patterns in data and make better decisions in the future based on the examples that we provide. The primaryaim is to allow the computers learn automatically without human intervention or assistance and adjust actions accordingly.

**1.4.1 Supervised Machine Learning Algorithm**

It can apply what has been learned in the past to new data using labelled examples to predict future events. Starting from the analysis of a known training dataset, the learning algorithm produces an inferred function to make predictions about the output values. The system is able to provide targets for any new input after sufficient training. The learning algorithm can also compare its output with the correct, intended output and find errors in order to modify the model accordingly [2].

* + 1. **Unsupervised Machine Learning Algorithm**

Algorithmsare used when the information used to train is neither classified nor labelled. Unsupervised learning studies how systems can infer a function to describe a hidden structure from unlabelled data. The system doesn’t figure out the right output, but it explores the data and can draw inferences from datasets to describe hidden structures from unlabelleddata [3].

* + 1. **Semi-Supervised Machine Learning Algorithm**

Fall somewhere in between supervised and unsupervised learning, since they use both labelled and unlabelled data for training typically a small amount of labelled data and a large amount of unlabelled data. The systems that use this method are able to considerably improve learning accuracy. Usually, semi-supervised learning [4] is chosen when the acquired labelled data requires skilled and relevant resources in order to train or learn from it. Otherwise, acquiring unlabelled data generally doesn’t require additional resources.

* + 1. **Reinforcement Machine Learning Algorithm**

It is a learning method that interacts with its environment by producing actions and discovers errors or rewards. Trial and error search and delayed reward are the most relevant characteristics of reinforcement learning. This method allows machines and software agents to automatically determine the ideal behaviour within a specific context in order to maximize its performance. Simple reward feedback is required for the agent to learn which action is best; this is known as the reinforcement signal.[4]

**CHAPTER-2**

LITRATURE SURVEY

Many researches have been done on crop yield prediction in the past. In this chapter we will provide insight of some of the research work which helps us to understand the topic in more depth.

**Ramesh, D., and VishnuVardhan, B.,Agrarian et.al [2015]** discussed a several subdivision in India is facing rigorous problem to make the most of the crop productivity. More than 60 out of a hundred the crop still depends on monsoon rainfall. Current growths in Information Technology for agriculture field have developed an interesting research area to forecast the crop yield. The problematic of yield prediction is a major problem that remains to be solved based on accessible data. Data mining methods are the better selections for this purpose. Different Data Mining methods are used and evaluated in agriculture for approximating the upcoming year's crop production. This paper presents a brief analysis of crop yield prediction using Multiple Linear Regression (MLR) method and Density based clustering technique for the particular region i.e. East Godavari district of Andhra Pradesh in India. In this paper an effort is made in command to know the region precise crop yield analysis and it is processed by applying both Multiple Linear Regression method and Density-based clustering method. These models were experimented in respect of all the districts of Andhra Pradesh, then the procedure of evaluation is passed out with only East Godavari district of Andhra Pradesh in India [5].

**Veenadhari, S., Bharat Misra, D Singh et.al [2011]** discussed that the data mining extraction of unseen predictive information from huge records, is a powerful new technology with great potential to help companies focus on the most significant data in their Data warehouses. Data mining tools predict upcoming trends and performance and growth, allowing businesses to make proactive, knowledge driven decisions. Though these methods are plausible, theoretically well created, and perform well on extra or less artificial test data sets, they depend on their skill to make sense of real-world data. This article gave us a detail project that is smearing a range of machine learning plans to problems in agriculture and horticulture. They briefly surveyed some of the techniques emerging from machine learning study, define a software workbench for testing with a variability of methods on real-world data sets, and a learning of dairy herd management in that culling rules were inferred from a medium-sized record of herd information. They also defined a range of machine learning plans to problems in agriculture and horticulture. There is a rising number of applications of data mining methods in agriculture and a rising amount of data that are presently available from several resources. This is relatively a novel research field and it is expected to grow in the upcoming. There is a lot of effort to be done on this emerging and interesting study field. The multidisciplinary method of integrating computer science with agriculture will help in predicting managing agricultural crops effectively [6].

**Ramesh A. Medar and Vijay. S. Rajpurohit et.al [2014]** presented a Precision agriculture (PA) and information technology (IT) are closely interwoven. The former frequently refers to the application of nowadays’ technology to agriculture. Due to the use of sensors and GPS technology, in today’s agriculture several data are collected. Creation use of those data via IT often leads to dramatic improvements in efficiency. For this purpose, the challenge is to change these raw data into useful data. This paper deals with suitable modeling methods for those agricultural data where the objective is to uncover the surviving patterns. In specific, the use of feed-forward back propagation neural networks will be evaluated and suitable parameters will be projected. In consequence, yield prediction is allowed based on cheaply obtainable site data. In this prediction, economic or environmental optimization of, e.g., fertilization can be passed out. Due to the rapidly advancing technology in the last few decades, ever more of our everyday life has been changed by information technology. Data access, once cumbersome and slow, has been turned into “data at your fingertips” at high speed. Technological breakthroughs have been made in industry and services as well as in agriculture [7].

**CHAPTER - 3**

METHODOLOGY

To achieve our objective following methodology will be used:

Data Preprocessing

Data normalization and Exploration

DATA PREPROCESSING

MODEL SELECTION

APPLY SUPERVISED LEARNING ALGORITHM ON SELECTED ALORITHM

SHOW THE ACCURACY

OF PREDICTION

*Fig – 3.1 Block diagram of Methodology*

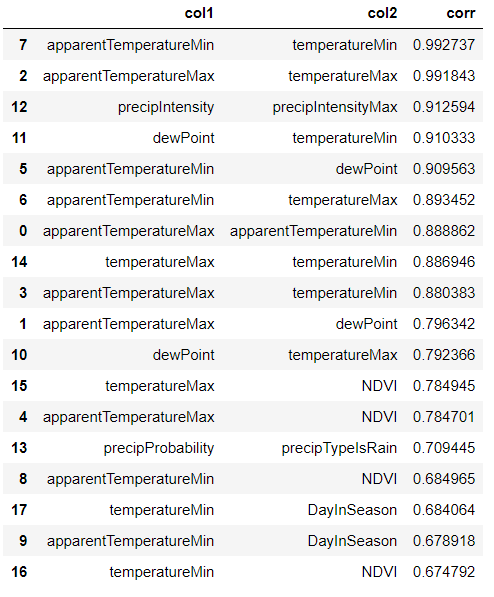
**CHAPTER -4**

PRE PROCESSING

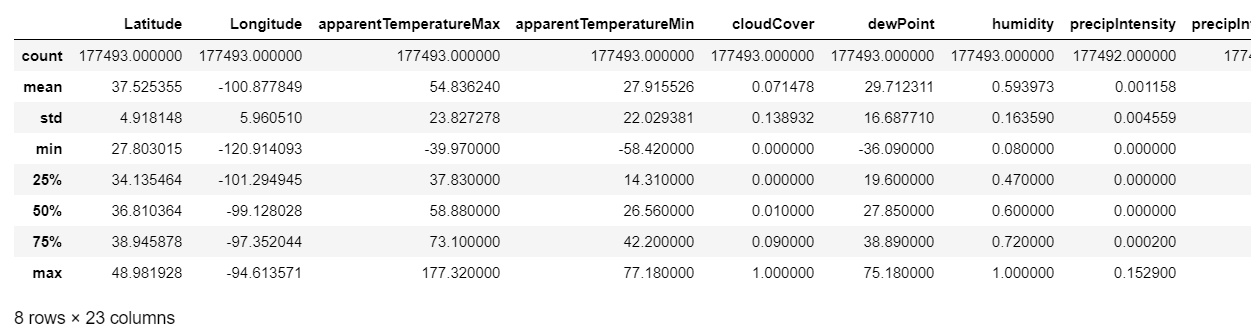
* 1. **Dataset Normalization**

Data pre-processing plays a crucial role. One of the first steps concerns the normalization of the data. This step is very important when dealing with parameters of different units and scales. Therefore, all parameters should have the same scale for a fair comparison between them. Two methods are usually well known for rescaling data [8]. Normalization, which scales all numeric variables in the range [0,1].

In machine learning, pattern recognition and in image processing, feature extraction starts from an initial set of measured data and builds derived values (features) intended to be informative and non-redundant, facilitating the subsequent learning and generalization steps, and in some cases leading to better human interpretations. Feature extraction is a dimensionality reduction process, where an initial set of raw variables is reduced to more manageable groups (features) for processing, while still accurately and completely describing the original data set. When the input data to an algorithm is too large to be processed and it is suspected to be redundant (e.g. the same measurement in both feet and meters, or the repetitiveness of images presented as pixels), then it can be transformed into a reduced set of features (also named a feature vector). Determining a subset of the initial features is called feature selection. The selected features are expected to contain the relevant information from the input data, so that the desired task can be performed by using this reduced representation instead of the complete initial data[7].



*Fig – 4.1 Feature correlation*



*Fig – 4.2 Dataset in describe from [3].*

**CHAPTER – 5**

POST PROCESSING

* 1. **Model Selection**

Model selection is the process of choosing between different machine learning approaches - e.g. Random forest regression, polynomial regression, etc - or choosing between different hyperparameters or sets of features for the same machine learning approach - e.g. deciding between the polynomial degrees/complexities for linear regression.

* 1. **Regression Models Used**
     1. **Polynomial Regression**

In [statistics](https://en.wikipedia.org/wiki/Statistics), polynomial regression is a form of [regression analysis](https://en.wikipedia.org/wiki/Regression_analysis) in which the relationship between the [independent variable](https://en.wikipedia.org/wiki/Independent_variable) *x* and the [dependent variable](https://en.wikipedia.org/wiki/Dependent_variable) *y* is modelled as an *n*th degree [polynomial](https://en.wikipedia.org/wiki/Polynomial) in *x*. Polynomial regression fits a nonlinear relationship between the value of *x* and the corresponding [conditional mean](https://en.wikipedia.org/wiki/Conditional_expectation) of *y*, denoted E(*y* |*x*), and has been used to describe nonlinear phenomena such as the growth rate of tissues, the distribution of carbon isotopes in lake sediments and the progression of disease epidemics. Although polynomialregression fits a nonlinear model to the data, as a [statistical estimation](https://en.wikipedia.org/wiki/Estimation_theory) problem it is linear, in the sense that the regression function is linear in the unknown [parameters](https://en.wikipedia.org/wiki/Parameter) that are estimated from the [data](https://en.wikipedia.org/wiki/Data). For this reason, polynomial regression is considered to be a special case of [multiple linear regression](https://en.wikipedia.org/wiki/Multiple_linear_regression).

*Y=**θ0+θ1x+θ2x2+….. (1)*

Where *Y*  is dependent variable, *θ0* is slope constant, x is independent variable and *θ*n is slope coefficients.

* + 1. **Decision Tree Regression**

A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements.

A decision tree is a flowchart-like structure in which each internal node represents a “test” on an attribute, each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes). The paths from root to leaf represent classification rules.

Tree based learning algorithms are considered to be one of the best and mostly used supervised learning methods. Tree based methods empower predictive models with high accuracy, stability and ease of interpretation. Unlike linear models, they map non-linear relationships quite well. They are adaptable at solving any kind of problem at hand (classification or

regression).

* + 1. **Random Forest Regression**

The random forest model is a type of additive model that makes predictions by combining decisions from a sequence of base models. It is an ensemble approach where we take into account the predictions[8] of several decision regression trees. Can also be mathematically expressed as:

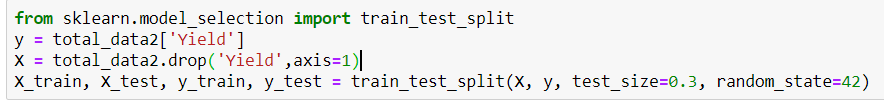
*g(x)=f0(x)+f1(x)+f2(x)+... (2)*

where the final model ***g*** is the sum of simple base models fi. Here, each base classifier is a simple decision tree. This broad technique of using multiple models to obtain better predictive performance is called model ensembling. In random forests, all the base models are constructed independently using a different subsample of the data. Random Forest prevents overfitting (which is common in decision trees) by creating random subsets of the features and building smaller trees using these subsets.

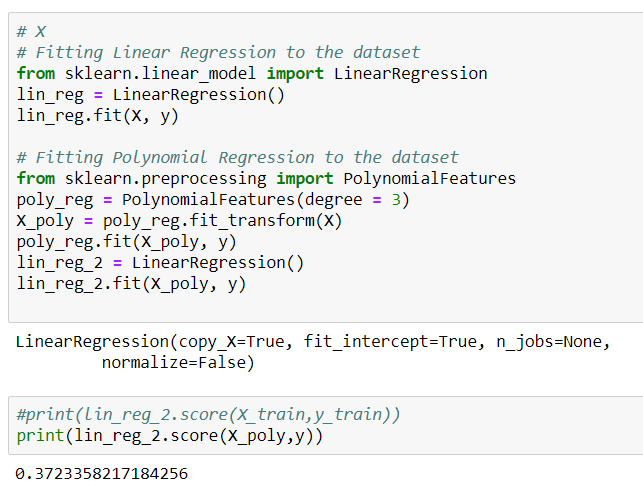
**CHAPTER** - **6**

TRAINING AND TESTING THE MODEL

We have trained our all regression models to fit the data with test size of 30 percent and remaining 70 Percent dataset for training the model [10]. As we increase the size of training dataset the model has been increased for both regression models.

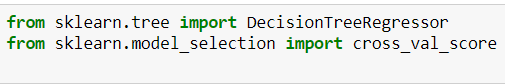


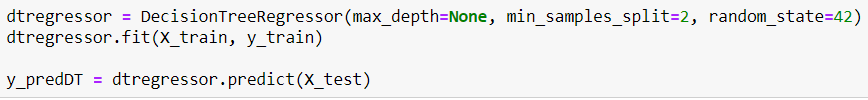
* 1. **Training and Testing of polynomial Regression Model**

We have trained the model with our dataset in polynomial regression model.

* 1. **Decision Tree Regressor Model**

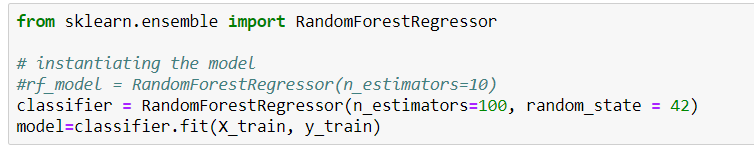
We have trained the model with our dataset in Decision Tree Regression model.





* 1. **Training and Testing of Random Forest Regression**

We have trained the model with our dataset for Random Forest Regression model.



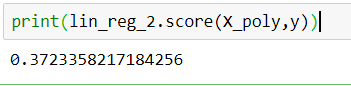
**CHAPTER – 7**

RESULTS AND ANANLYSIS

After successful training and the testing of dataset we moved further for finding the accuracy the model. The accuracy of the model shows that how we have predicted the yield of the crop in compare to original data the more the accuracy of the model the it is near to the original yield value.

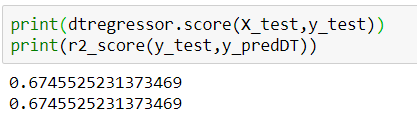
**7.2 Accuracy of Polynomial Regression**

The following figure shows the accuracy of Polynomial Regression Model. Where we have achieved testing accuracy of just 37 percent.



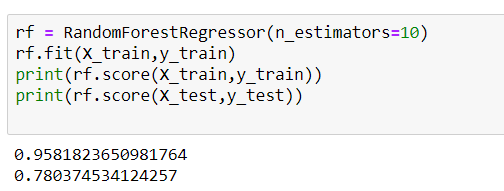
**7.3 Accuracy of Decision Tree Regression**

The following figure shows the accuracy of Decision Tree Regression Model. Where we have achieved testing accuracy of 67 percent.

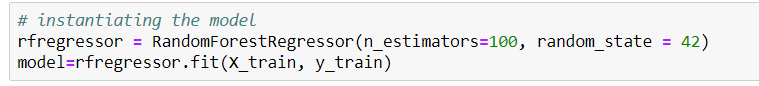


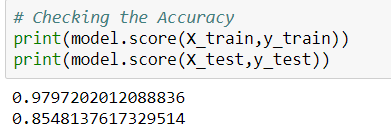
**7.4 Accuracy of Random Forest Regression**

The following figure shows the accuracy of the Random Forest Regression model. Where we have achieved training and testing accuracy of 95 and 78 percent respectively. This figure shows the accuracy with n\_estimators = 10.



Accuracy with n\_estimators =100, We can see the increase in accuracy of training and testing of 97 and 85 percent respectively with increase in time complexity.





**CHAPTER - 8**

CONCLUSIONS AND FUTURE SCOPE

**8.1 Conclusion**

This project presents a Machine learning framework for the task of crop yield prediction. It allows for real time forecasting throughout the year and is applicable world-wide, especially for developing countries where field surveys are hard to conduct. We developed models to predict wheat yield and examined a variety of confounding factors that affect model performance. The model provides us with the state-of-the-art prediction accuracy and will have great impact in sustainable agriculture and food security.

**8.2 Future scope**

Some of the future scopes that can be included in our project are:

1. A web-based application that can be made for our work in future.

2. We can improve our system that can deal with dataset with similar correlation.

3. We can work on deep learning methods to improve our model.

**REFERENCES**

[1] Kuwata, K. and Shibasaki, R. (2015), Estimating crop yields with deep learning and remotely sensed data, *Proceedings* *of 2015 IEEE International Geoscience and Remote Sensing Symposium*, 26-31 July, Milan, Italy, pp. 858-861.

[2] Shweta Taneja, Rashmi Arora, Savneet Kaur, “Mining of Soil Data Using supervised Learning Technique”, International Journal of Applied Engineering Research, ISSN 0973-4562 Vol. 7 No.11, 2012.

[3] https://aerialintel.blob.core.windows.net/recruiting/datasets/wheat-2013-supervised.csv

[4] Dr. Rakesh Poonia1 , Sonia Bhargava “Prediction of Crops Methodology using Data Mining Techniques”, International Journal of Innovative Research in Science, Engineering and Technology, Vol. 6, Issue 10, October 2017.

[5] D Ramesh, B Vishnu Vardhan, “Analysis of Crop Yield Prediction using Data Mining Techniques”, International Journal of Research in Engineering and Technology (IJRET),Vol.4, 2015.

[6] Veenadhari, S., Bharat Misra, D Singh, “Data mining Techniques for Predicting Crop Productivity – A review article”, IJCST, International Journal of Computer Science and technology march 2011.

[7] Ramesh A. Medar and Vijay. S. Rajpurohit “A Survey of data mining techniques for crop yield prediction”, IJARCSMS, Volume 2, Issue 9, September 2014 pg. 59-64.

[8] Leemans V, M F Destain, "A Real Time Grading Method of Apples Based on Features Extracted from Defects", J. Jood Eng., 2004, pages: 83-89.

[9] DakshayiniPatil, Dr. M .S Shirdhonkar, “Rice Crop Yield Prediction using Data Mining Techniques: An Overview”, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 7, Issue 5, ISSN: 2277 128X ,2017.

[10] Rajshekhar Borate., “Applying Data Mining Techniques to Predict Annual Yield of Major Crops and Recommend Planting Different Crops in Different Districts in India”, International Journal of Novel Research and Software Engineering,Vol. 3, Issue 1, pp:, April 2016.

**APPENDIX**

import numpy as np

import matplotlib.pyplot as plt

import pandas as pd

import matplotlib.pyplot as plt

from sklearn.metrics import r2\_score

dataset2013= pd.read\_csv('wheat-2013-supervised.csv')

dataset2014 = pd.read\_csv('wheat-2014-supervised.csv')

#merging the datas to create a single dataset

dataset = pd.concat([dataset2013, dataset2014],axis=0) print(dataset2013.shape)

print(dataset2014.shape)

print(dataset.shape)

dataset.head()

dataset.describe()

total\_data1 = dataset.drop(['CountyName','State','Latitude','Longitude','Date'],axis=1)

total\_data2 = total\_data1.dropna(axis = 0)

print(total\_data2['Yield'])

from sklearn.model\_selection import train\_test\_split

y = total\_data2['Yield']

X = total\_data2.drop(['Yield'],axis=1)

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.3, random\_state=42)

from sklearn import preprocessing

columns\_to\_scale = total\_data2.columns.tolist()

columns\_to\_scale = [x for x in columns\_to\_scale if x != 'Yield']

print(columns\_to\_scale)

std\_scaler = preprocessing.StandardScaler().fit(X\_train[columns\_to\_scale])

minmax\_scaler = preprocessing.MinMaxScaler().fit(X\_train[columns\_to\_scale])

X\_train[columns\_to\_scale] = std\_scaler.transform(X\_train[columns\_to\_scale])

X\_test[columns\_to\_scale] = std\_scaler.transform(X\_test[columns\_to\_scale])

X.astype(float),y.astype(float)

from sklearn.linear\_model import LinearRegression

lin\_reg = LinearRegression(normalize=True)

lin\_reg.fit(X, y)

from sklearn.preprocessing import PolynomialFeatures

poly\_reg = PolynomialFeatures(degree = 3)

X\_poly = poly\_reg.fit\_transform(X)

poly\_reg.fit(X\_poly, y)

lin\_reg\_2 = LinearRegression(normalize=True)

lin\_reg\_2.fit(X\_poly, y)

y\_predPR = lin\_reg\_2.predict(X\_poly)

print(lin\_reg\_2.score(X\_poly,y))

print(r2\_score(y,y\_predPR))

from sklearn.tree import DecisionTreeRegressor

dtregressor = DecisionTreeRegressor(max\_depth=None, min\_samples\_split=2, random\_state=42)

dtregressor.fit(X\_train, y\_train)

y\_predDT = dtregressor.predict(X\_test)

print(y\_predDT)

print(dtregressor.score(X\_test,y\_test))

print(r2\_score(y\_test,y\_predDT))

from sklearn.ensemble import RandomForestRegressor

rfregressor = RandomForestRegressor(n\_estimators=100, random\_state = 42)

model=rfregressor.fit(X\_train, y\_train)

y\_pred = model.predict(X\_test)

print(y\_pred)

total\_data2['y\_pred']=np.NAN

total\_data2.iloc[(len(total\_data2)-len(y\_pred)):,-1:] = y\_pred

total\_data3 = total\_data2.dropna

print(y\_pred)

print(model.score(X\_train,y\_train))

print(model.score(X\_test,y\_test))

imp = list(zip(X,model.feature\_importances\_))

imp=sorted(imp,key=lambda x:x[1])

print(imp)