# A
# Summer Internship Report
# On
# “Gujarat Crop Production Analytics”

(CE346 – Summer Internship - I)

**Prepared by**

Vansh Desai-19CE018

**Under the Supervision of**

Prof. Mayuri Popat

**Submitted to**

Charotar University of Science & Technology (CHARUSAT)
for the Partial Fulfillment of the Requirements for the
Degree of Bachelor of Technology (B.Tech.)
for Semester 5

**Submitted at**

Accredited with Grade A by NAAC
Accredited with Grade A by KCG

**U & P U. PATEL DEPARTMENT OF COMPUTER ENGINEERING**
**Chandubhai S. Patel Institute of Technology (CSPIT)**
**Faculty of Technology & Engineering (FTE), CHARUSAT**
**At: Changa, Dist: Anand, Pin: 388421.**
**July 2021**

अमृतं तु विद्या CHARUSAT
CHAROTAR UNIVERSITY OF SCIENCE AND TECHNOLOGY

**Accredited with Grade A by NAAC**
**Accredited with Grade A by KCG**

# CERTIFICATE

This is to certify that the report entitled "**Gujarat Crop Production Analytics**" is a bonafied work carried out by **19CE018** under the guidance and supervision of **Prof. Mayuri Popat**/ **Mr. Manthan Bhikhadiya** for the subject **Summer Internship – I (CE346)** of 5th Semester of Bachelor of Technology in **Computer Engineering** at Chandubhai S. Patel Institute of Technology (CSPIT), Faculty of Technology & Engineering (FTE) – CHARUSAT, Gujarat.

To the best of my knowledge and belief, this work embodies the work of candidate himself has duly been completed, and fulfills the requirement of the ordinance relating to the B.Tech. Degree of the University and is up to the standard in respect of content, presentation and language for being referred by the examiner(s).

Under the supervision of,

| | |
|---|---|
| Mpopat | M.V. Bhikadiya |
| Prof. Mayuri Popat<br>Assistant professor<br>U & P U. Patel Dept. of Computer Engineering<br>CSPIT, FTE, CHARUSAT, Changa, Gujarat | Mr. Manthan Bhikhadiya<br>ML team head<br>Krsikx India llp |

Dr. Ritesh Patel
Head - U & P U. Patel Department of Computer Engineering,
CSPIT, FTE, CHARUSAT, Changa, Gujarat.

**Chandubhai S. Patel Institute of Technology (CSPIT)**

**Faculty of Technology & Engineering (FTE), CHARUSAT**

At: Changa, Ta. Petlad, Dist. Anand, Pin: 388421. Gujarat.

KrsikX
CULTIVATING IDEAS FOR CULTIVATORS

**Krsikx India LLP**
Tel :-+91 9924720999 | +91 75673 21109
Email :- krsikx@gmail.com
URL :- www.krsikx.com
LLPIN :- AAT-7820

Date: 15th July 2021

# CERTIFICATE OF COMPLETION

This is to certify that **Vansh Desai**, a student of Chandubhai S Patel Institute of Technology, CHARUSAT has undergone his internship in the role of **Machine Learning Intern** with **KrsikX India** from **04/06/2021 to 11/07/2021**.

He has worked on the project titled **Gujarat Crop Production Analytics**. This project was aimed to calculate and analyze production of crop for different scenarios. During his internship period he has worked in different algorithms. He exhibited excellent skills in statistical analysis.

He has demonstrated his skills with self-motivation to learn new skills. His performance exceeded our expectations, and he was able to complete the target on time. He is well organized and is able to work independently and effectively multi task to ensure that assignments are looked after and completed in a professional and timely manner.

We wish him all the best for his upcoming future.

D. K. Zala

**Divyarajsinh Zala**

**Founder & CEO**

**KRSIKX INDIA LLP.**

K.Baldaniya

**Krishna Baldaniya**

**Co – Founder & CTO**

**KRSIKX INDIA LLP.**

# Table of Contents

# Acknowledgement

I, the developer of the machine learning model " Gujarat Crop Production Analytics ", am very happy and promised to present the mission of the project. The development of this project has provided me with ample opportunities to think, implement and interact with various aspects of management skills and machine learning.

Every successful job done by a person is inseparable from the continued encouragement, kindness and support of the people around him. I take this opportunity to thank many people for their valuable time, full support and cooperation in the development of this project.

I would like to express my deep gratitude to prof Aayushi Chaudhary faculty of our CE department. Thanks to Mayuri popat. It is because of her that she motivates me to work hard and adopt new technologies.

I would also like to thank my mentors Manthan Bhikhadia and Vanshita Agarwal, for their guidance during the model development phase. they will help me whenever I fall into the concept of machine learning.

I am very grateful to Mr. Krishna Baldania who gave me the opportunity to become a member of the company.

I sincerely thank everyone at Krsikx for helping me with the project in some way.

# Abstract

Machine learning (ML) methods are used in many areas, from assessing customer behavior in supermarkets to predicting customer phone use. Machine learning has also been used in agriculture for many years. Crop production prediction is one of the most challenging problems in precision agriculture.

So far, many models have been proposed and verified. This problem requires the use of multiple data sets, because crop production depend on many different factors, such as climate, climate, soil, fertilizer use, and seed varieties. This shows that crop production forecasting is not a trivial task, on the contrary, it consists of several complex steps. Today, crop production prediction models can reasonably estimate actual production, but better production prediction is still needed.

Our dataset was taken from Kaggle and it has no null value. In the first look dataset is looking very easy and simple but the work is complicated. Whole dataset about Gujarat district and weather. Our main aim is to process data-set and get good-fitted regression line. At the end we talked about feature tuning and cross validation.

# List of Figures

# INTRODUCTION

## 1.1 Introduction to Topic

Machine learning is a branch of artificial intelligence (AI) that focuses on learning and is a practical method that can provide better performance predictions based on multiple characteristics. Machine learning (ML) can determine patterns and correlations, and discover insights from data sets. The model must be trained using a data set, where the results are expressed based on past experience. The predictive model is built using multiple features; therefore, the model parameters are determined using historical data during the training phase. For the test phase, some of the historical data that is not used for training is used for performance evaluation purposes.

ML models can be descriptive or predictive, depending on the research question and the research question. Descriptive models are used to gain insight into the collected data and explain what happened, while predictive models are used to predict the future. When seeking to build high-performance predictive models, ML research contains different challenges. Choosing the right algorithm to solve the problem at hand is crucial, and the underlying algorithm and platform must also be able to handle large amounts of data.

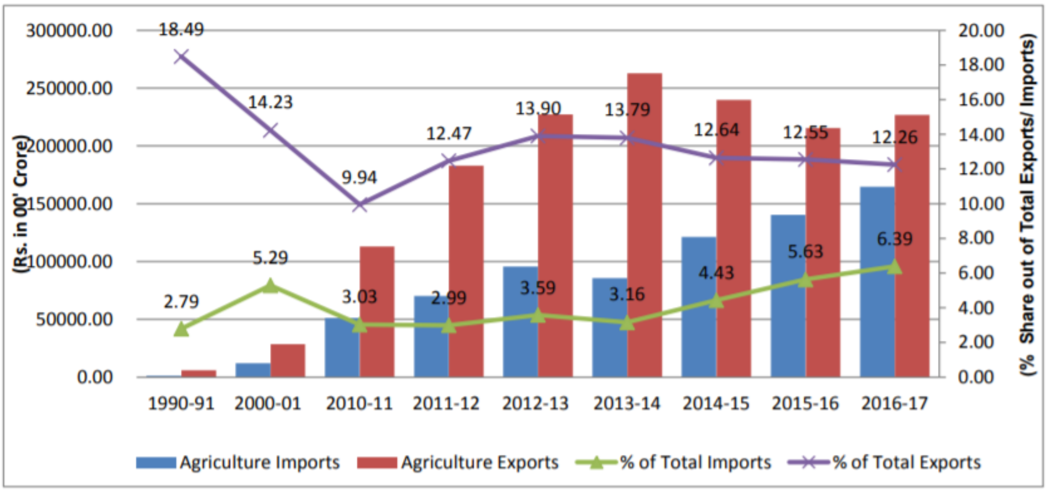
Crop yield prediction is an essential task for the decision-makers at national and regional levels (e.g Directorate of Economics and Statistics) for rapid decision-making. An accurate crop production prediction model can help farmers to decide on what to grow and when to grow. There are different approaches to crop production prediction.

Crop yield prediction is extremely challenging due to its dependence on multiple factors such as crop genotype, environmental factors, management practices, and their interactions Environments, changing spatially and temporally, have huge effects on year-to-year and location-to-location variations in crop yield.

## 1.2 Motivation

The state of Gujarat is described as hot and semi-arid, and 49% of its total area is built up land (9.6 million hectares). In the monsoon the flooded area only represents 32% of the total area of impacted areas. The remaining areas face severe water shortages. The quality of the groundwater is bad. The alluvial soil is alkaline / saline-alkaline, it lacks nitrogen, phosphorus and zinc. The editing intensity is as low as 118%. The state's total grain output is 5.26 million tons. The key yields are peanuts, cotton, pearl millet, corn, sorghum, castor, gram and mustard.

A farmer must have a good knowledge of the soil type, the biotic factors governing the soil and a thorough knowledge about the traditional agricultural practices to gain maximum crop production. Such practice may include harrowing and ploughing using inputs such as fertilizers, insecticides and herbicides. Government watch and predict the harvesting outcome’s of particular area. Government able to provide resources to farmer accordingly that. Gujarat stands first in area and production but productivity wise Tamil Nadu state stands first in the country.



*Source:* Directorate General of Commercial Intelligence & Statistics, D/o Commerce, Kolkata.

**Figure 1.1: Trends in Agricultural Imports/ Exports and their percentage share in Total Imports**

As you can see in Figure 1.1, we are trying to decrease imports and increasing exports in India. If we increase our production by 40% we can double this ratio.

## 1.3 Problem Statement

We have dataset from Kaggle which contain information of State (Gujarat), City, Year, Season, Crop, Area - In Hectare, average temperature (In degree Celsius), Cloud Cover (In Percentage), maximum temperature (In degree Celsius), Precipitation (In mmvap), Pressure (In hPa (Hectopascal)), Rainfall (In mm), Wet Day Freq (In a number of days), minimum temperature (In degree Celsius), Production (In Tone).

We have to extract that information and fit the regression curve properly. We have to check different model with different algorithm.

## 1.4 Objectives

- Build a model with clean data
- Model must have good feature and easily implementable
- Model must properly and easily compatible with flask
- Model should not overfitted
- All statistical parameter analyze in model evaluation
- Good cross validation and r2 score

# 2.Literature Review

Machine learning is an important decision support tool for predicting crop yield, including supporting decisions about what crops to plant and what to do during the growing season of the crop. Various machine learning algorithms have been applied to support crop yield forecast research. In this research, we performed a systematic literature review (SLR) to extract and synthesize the algorithms and characteristics that have been used in crop yield prediction research. Based on our search criteria, we retrieved ten related studies from six electronic databases, of which two studies were selected for further analysis of the inclusion and exclusion criteria.

We carefully research these selected studies, discuss the methods and characteristics used, and provide recommendations for future research. According to our analysis, the most used features are temperature, rainfall, and soil types, and the most used algorithms in these models are artificial neural networks. After observing the analysis of 2 articles based on machine learning, we performed additional searches in the electronic database to identify research based on deep learning, we got 1 article based on deep learning and extracted the applied deep learning algorithm. Based on this additional analysis, convolutional neural networks (CNN) are the most widely used deep learning algorithms in these studies, and the other most widely used deep learning algorithms are short-term memory (LSTM) and deep neural networks (DNN).

An SLR study is expected to be replicable, which means that all the steps taken need to be explained clearly, and the results should be transparent for other researchers. The critical factors for a successful SLR study are objectivity and transparency (Kitchenham et al., 2007). As its name indicates, an SLR needs to be systematic and cover all the literature published so far. This study presents all the available literature published so far on the application of machine learning in crop yield prediction problem. In this study, we present our empirical results and responses to the research questions defined as part of this review article.

# 3. Proposed Model/Architecture

## 3.1 ML Flow Diagram

Before work on model we have to work on design of flow. Design of flow varies on data in this data.There are both type of categorical and numerical variable in the dataset. In figure 3.1 we describe our basic flow of diagram.

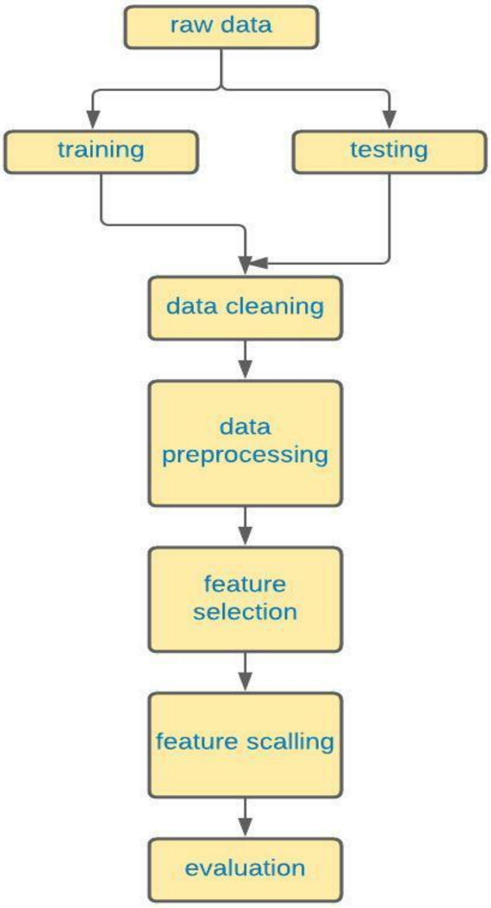


**Figure 3.1 Flow Diagram**

Our project is divided in six parts:

1. Distribution of data in train and test
2. Data preprocessing
3. Feature selection
4. Feature scaling
5. Algorithms
6. Evaluation of models

## 3.2 Distribution of data in train and test

Why we should do training and testing before cleaning and pre-processing there is question in mind? Answer is that if you can do training and testing after pre-processing statistically, we build relationship between both data it is called data-leakage. Data leakage become constraint when we deploy model in real world. We make that distribution with the help of train_test_split module of skit-learn library. Our dataset has total 8436 rows and 15 columns. We divided train and test in 60:40 ratio.

## 3.3 Data preprocessing

Data preprocessing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model.

A real-world data generally contains noises, missing values, and maybe in an unusable format which cannot be directly used for machine learning models. Data preprocessing is required tasks for cleaning the data and making it suitable for a machine learning model which also increases the accuracy and efficiency of a machine learning model.

It involves below steps:

### 3.3.1 Getting the dataset

To create a machine learning model, the first thing we required is a dataset as a machine learning model completely works on data. The collected data for a particular problem in a proper format is known as the dataset. Our dataset is in CSV file. We use panda library for import the dataset

### 3.3.2 Importing libraries

In order to perform data preprocessing using Python, we need to import some predefined Python libraries. These libraries are used to perform some specific jobs. There are three specific libraries that we will use for data preprocessing. We use library like pandas, NumPy ,Seaboarn,Sikit-learn etc.

### 3.3.3 Encoding categorical data

Since machine learning model completely works on mathematics and numbers, but if our dataset would have a categorical variable, then it may create trouble while building the model. So it is necessary to encode these categorical variables into numbers.

We have total 3 categorical variable. we will convert the country variables into categorical data. So to do this, we will use LabelEncoder() class from preprocessing library.

## 3.4 Feature selection

Feature Selection is one of the core concepts in machine learning which hugely impacts the performance of your model. The data features that you use to train your machine learning models have a huge influence on the performance you can achieve.

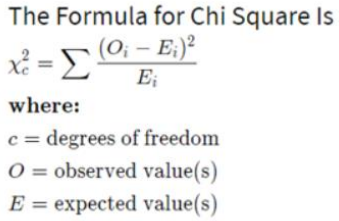
Feature Selection is the process where you automatically or manually select those features which contribute most to your prediction variable or output in which you are interested in. We use two feature selection techniques for feature selection

1. Univariate Selection
2. Correlation with Heatmap

### 3.4.1 Univariate selection

Univariate feature selection works by selecting the best features based on univariate statistical tests. We compare each feature to the target variable, to see whether there is any statistically significant relationship between them. It is also called analysis of variance (ANOVA). When we analyze the relationship between one feature and the target variable, we ignore the other features. That is why it is called 'univariate'. Each feature has its test score.

We use chi$^2$ test for univariate selection. A chi-square test is used in statistics to test the independence of two events. Given the data of two variables, we can get observed count O and expected count E. Chi-Square measures how expected count E and observed count O deviates each other.



The Formula for Chi Square Is

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

**where:**

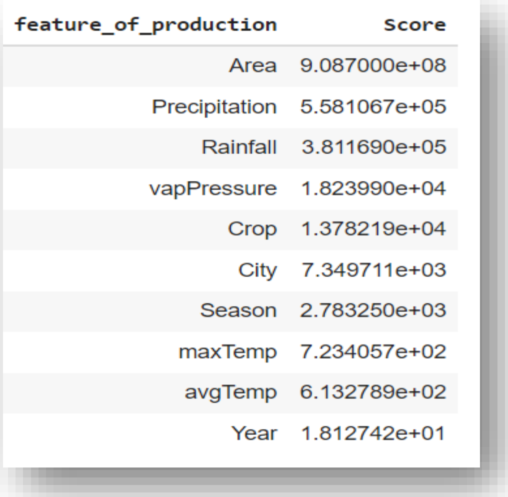
$c$ = degrees of freedom

$O$ = observed value(s)

$E$ = expected value(s)

**Figure 3.2: Formula of chi$^2$ test**

The scikit-learn library provides the SelectKBest class that can be used with a suite of different statistical tests to select a specific number of features. We use that class with chi$^2$ test. In figure 3.3 you can see our result of chi$^2$ test



| feature_of_production | Score |
|---|---|
| Area | 9.087000e+08 |
| Precipitation | 5.581067e+05 |
| Rainfall | 3.811690e+05 |
| vapPressure | 1.823990e+04 |
| Crop | 1.378219e+04 |
| City | 7.349711e+03 |
| Season | 2.783250e+03 |
| maxTemp | 7.234057e+02 |
| avgTemp | 6.132789e+02 |
| Year | 1.812742e+01 |

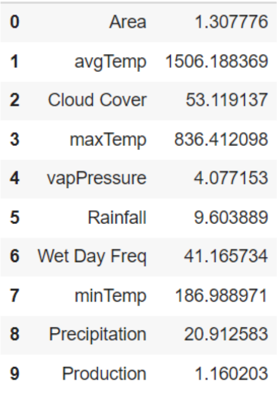
**Figure 3.3: Result of chi$^2$ test**

We can see in figure 3.3 Area has major score so it is very important feature the importance of feature is decided with chi$^2$ test.

### 3.4.2 Correlation with heatmap

Correlation explains how one or more variables are related to each other. These variables can be input data features which have been used to forecast our target variable. Correlation, statistical technique which determines how one variables moves/changes in relation with the other variable. It gives us the idea about the degree of the relationship of the two variables. It's a bi-variate analysis measure which describes the association between different variables. In most of the business it's useful to express one subject in terms of its relationship with others. Correlation helps us to eliminate dependent variable. We use two method variance inflation factor (VIF) and heat-map for determine correlation.

Variance Inflation Factor (VIF) is used to detect the presence of multicollinearity. Variance inflation factors (VIF) measure how much the variance of the estimated regression coefficients are inflated as compared to when the predictor variables are not linearly related.



| | | |
|---|---|---|
| 0 | Area | 1.307776 |
| 1 | avgTemp | 1506.188369 |
| 2 | Cloud Cover | 53.119137 |
| 3 | maxTemp | 836.412098 |
| 4 | vapPressure | 4.077153 |
| 5 | Rainfall | 9.603889 |
| 6 | Wet Day Freq | 41.165734 |
| 7 | minTemp | 186.988971 |
| 8 | Precipitation | 20.912583 |
| 9 | Production | 1.160203 |

**Figure 3.4: VIF score**

VIF is greater than 10 means multi-collinearity will be detected. In figure 3.4 you can see that rainfall, area ,vap pressure, production is independent variable.

Pearson’s heat-map is another method to find multi collinearity. In figure 3.5 more dark part means more multi collinearity.

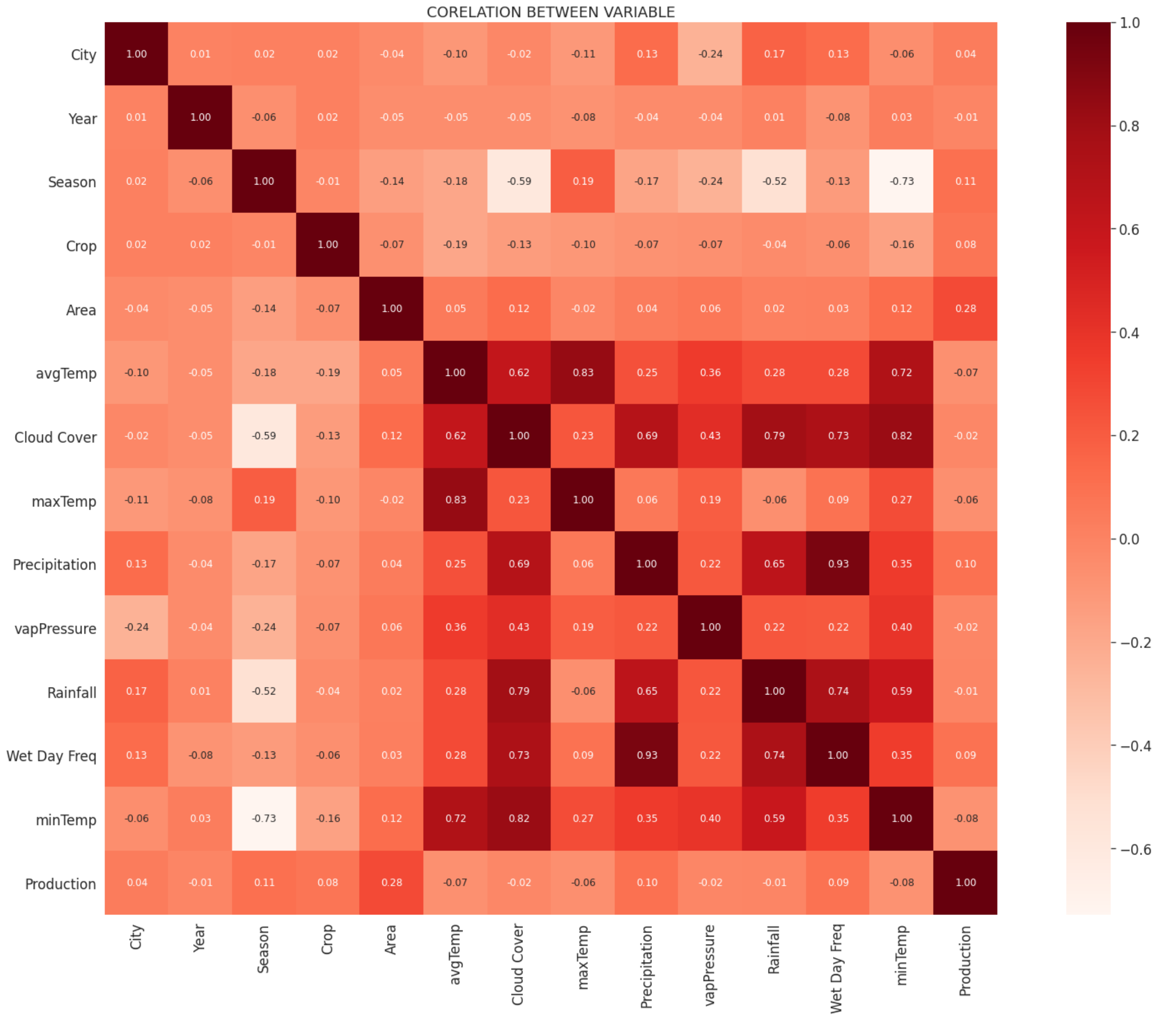


**Figure 2.5 Heat map**

From VIF and heat map we can conclude that temperature variables has large correlation. Rainfall is better than precipitation. So our features are 'Area', 'Rainfall’, ‘Crop', 'vapPressure', 'City’, ‘Season'.

## 3.5 Data scaling

Generally we talked about scaling in data-preprocessing part but in this data-set scaling is played an important role. We try to scale both dependent and independent variable and then use inverse transform function.

We use three different types of scalers robust, standard, min_max and normalize scaler but in standard scaler we get good accuracy.

### 3.5.1 Standard scaler

Standard Scaler follows Standard Normal Distribution (SND). Therefore, it makes mean = 0 and scales the data to unit variance. Standardization scales each input variable separately by subtracting the mean (called centering) and dividing by the standard deviation to shift the distribution to have a mean of zero and a standard deviation of one.

## 3.6 Algorithms

In Machine Learning, we use various kinds of algorithms to allow machines to learn the relationships within the data provided and make predictions based on patterns or rules identified from the dataset. So, regression is a machine learning technique where the model predicts the output as a continuous numerical value.

The algorithms we are going to cover are:

1. Linear Regression
2. Decision Tree
3. Support Vector Regression
4. Adaboost Regression
5. Random Forest

### 3.6.1 Linear Regression

Linear Regression is an ML algorithm used for supervised learning. Linear regression performs the task to predict a dependent variable(target) based on the given independent variable(s). So, this regression technique finds out a linear relationship between a dependent variable and the other given independent variables. Hence, the name of this algorithm is Linear Regression.

### 3.6.2 Decision Tree

The decision tree models can be applied to all those data which contains numerical features and categorical features. Decision trees are good at capturing non-linear interaction between the features and the target variable. Decision trees somewhat match human-level thinking so it’s very intuitive to understand the data.

### 3.6.3 Support Vector Regression

You must have heard about SVM i.e., Support Vector Machine. SVR also uses the same idea of SVM but here it tries to predict the real values. This algorithm uses hyperplanes to segregate the data. In case this separation is not possible then it uses kernel trick where the dimension is increased and then the data points become separable by a hyperplane.

### 3.6.4 Adaboost Regression

The AdaBoost algorithm involves using very short (one-level) decision trees as weak learners that are added sequentially to the ensemble. Each subsequent model attempts to correct the predictions made by the model before it in the sequence. This is achieved by weighing the training dataset to put more focus on training examples on which prior models made prediction errors.

### 3.6.5 Random Forest Regression

Random Forests are an ensemble(combination) of decision trees. It is a Supervised Learning algorithm used for classification and regression. The input data is passed through multiple decision trees. It executes by constructing a different number of decision trees at training time and outputting the class that is the mode of the classes (for classification) or mean prediction (for regression) of the individual trees.

## 3.7 Evaluation of models

Model evaluation is very important in data science. It helps you to understand the performance of your model and makes it easy to present your model to other people. There are many different evaluation metrics out there but only some of them are suitable to be used for regression. This section will cover the different metrics for the regression model and the difference between them. Hopefully, after you read this post, you are clear on which metrics to apply to your future regression model.

There are 3 main metrics for model evaluation in regression:

1. R Square/Adjusted R Square

2.Root Mean Square Error (RMSE)

3. Mean Absolute Error (MAE)

4. Cross-validation score

# 4.Implementation Environment

Our implementation environment is google Colab. Colab notebooks allow you to combine executable code and rich text in a single document, along with images, HTML, LaTeX and more. When you create your own Colab notebooks, they are stored in your Google Drive account. You can easily share your Colab notebooks with co-workers or friends, allowing them to comment on your notebooks or even edit them.

# 5.Experimental Results

## 5.1 Analysis of Random Forest Regressor

MAE :0.06103549853853965

RMSE :0.268340435184688

r2: 0.9279934108448923

cross-validation score: 0.776587805887862

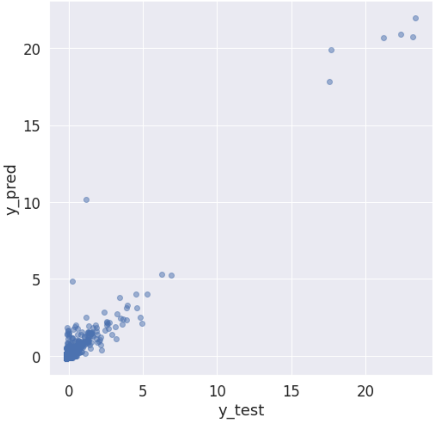


**Figure 5.1:test vs predication of random forest**

## 5.2 Analysis of Decision tree Regressor

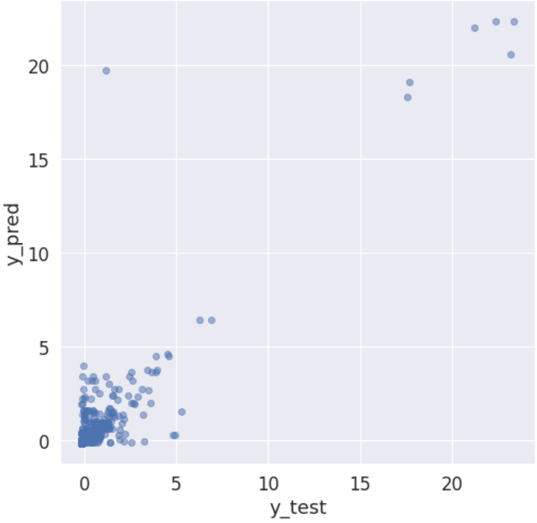
MAE 0.07842691684281601

RMSE 0.44391249969474994

r2 0.8029416926147586

cross-validation score: 0.6266165058383556

**Figure 5.2:test vs prediction of decision tree**



## 5.3 Analysis of support vector Regressor

MAE 0.18901486156474287

RMSE 0.9798147365995549

r2 0.03996308194234488

cross-validation score: 0.33596569535247267

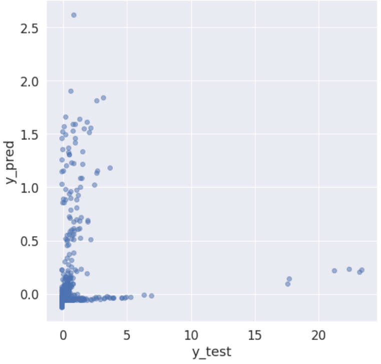


**Figure 5.3:test vs prediction of support vector**

## 5.4 Analysis of AdaBoostRegressor

MAE 0.39743375569485 89

RMSE 0.4543718674563228

r2 0.7935462060642539

cross-validation score: 0.5847625522973583

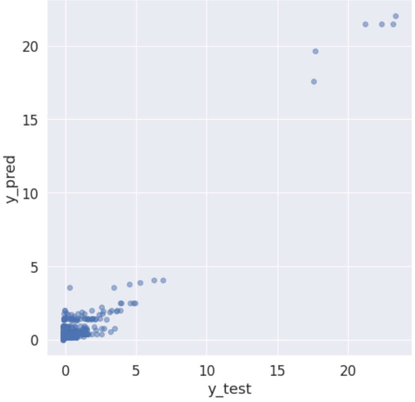


**Figure 5.2: test vs prediction of AdaBoostRegressor**

# 6. Limitations and Future Enhancements

Weather data has a major role in affecting the Agriculture data of crop yield. For crop yield prediction, Random Forest for prediction model gives better Results in training as well as validation for the data then other three techniques which are Artificial Neural network and two regression techniques multiple linear regression(MLR) and Regression tree. Apart from that, if using larger dataset of previous years can also be affect the accuracy of algorithms and can get better accuracy or less RMSE value in future.

We can also build inverse scaling function. Artificial Nural Network can open the doors of production prediction.

# Conclusion

Machine learning in agricultural is good domain for research. Regression has some relation with scaling techniques. Ensemble techniques work better on the dataset. Flask is good tool for implement that model. regression technique is not giving better result. We have to train nural network

# References

**Paper:**

Vinita Shah, Prachi Shah,” Groundnut Crop Yield Prediction Using Machine Learning Techniques”,2018

**Web:**
https://towardsdatascience.com/feature-selection-techniques-in-machine-learning-with-python-f24e7da3f36e

https://medium.com/analytics-vidhya/what-is-correlation-4fe0c6fbed47

https://www.analyticsvidhya.com/blog/2021/05/5-regression-algorithms-you-should-know-introductory-guide/

https://www.javatpoint.com/data-preprocessing-machine-learning