*Project For the Subject*

**INDUSTRIAL ENGINEERING**

**(UME515)**

Submitted by

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
| Gurriday Chawla | (102108211) |

### Group: 2MEE8

**Research Term Project Title:**

**Crop Yield Predicition Using Gaussian Process Regression and Autoregressive Integrated Moving Average.**

Course Instructor

## Dr. Rajnish Mallick



**THAPAR INSTITUTE OF ENGINEERING AND TECHNOLOGY, PATIALA**

Session: January – May 2023

# Contents

|  | **Title** | **Page No.** |
| --- | --- | --- |
|  | Abstract | 1 |
|  | Introduction | 2 |
| 1. | Why use Machine Learning models like GPR for crop yield prediction? | 3 |
| 2. | Methodology | 3 |
| 3. | Framework | 5 |
| 4. | Data Analysis. | 6 |
| 4.1 | Closing Price Visualization Along with 50 and 200- day moving averages. | 6 |
| 4.2 | Predictions from ARIMA. | 7 |
| 4.3 | Gaussian Process Regression. | 8 |

**Abstract**

This research term project studies and examines the forecasting of Crop Yield using past data- previously collected data using statistical and computing tools like GPR (Gaussian Process Regression) and ARIMA (Auto Regressive Integrated Moving Average) to predict crop yield for the next month.

We would like to thank Mr. Rajnish Mallick for his immense support. With his determination and confidence we were able to finish our project while gaining the knowledge about different models of ML used for future prediction.

# Introduction

Crop yield prediction is one of the challenging tasks in agriculture. It plays an essential role in decision making at global, regional, and field levels. The prediction of crop yield is based on soil, meteorological, environmental, and crop parameters. Decision support models are broadly used to extract significant crop features for prediction. Precision agriculture focuses on monitoring (sensing technologies), management information systems, variable rate technologies, and responses to inter- and intravariability in cropping systems. The benefits of precision agriculture involve increasing crop yield and crop quality, while reducing the environmental impact.

Crop yield simulations help to understand the cumulative effects of water and nutrient deficiencies, pests, diseases, the impact of crop yield variability, and other field conditions over the growing season.

The study aims to predict crop yield for the next month using Gaussian Process Regression and ARIMA .

## Why use Machine Learning models like GPR for crop yield prediction?

Machine Learning is a successful dynamic device for foreseeing crop yields, just as for choosing which harvests to plant and what to do about them during the developing season. Since it operates with a large amount of data produced by several variables, the farming system is highly complicated. Methods of machine learning can aid intelligent system decision-making.

By taking into account several variables, machine learning algorithms can help farmers decide which crop to grow in addition to increasing yield. Farmers can benefit from yield estimation because it allows them to minimize crop loss and obtain the best prices for their crops.

**Methodology**

**We have assigned an agricultural csv dataset for Crop yield prediction.** **As some models cannot take string values as inputs, we had to do encoding of those variables.** **This will help the machine understand the columns and take them as inputs. Although label encoding has a few benefits where each value is given a unique integer value based on order which is alphabetical. But it used that integer value as a rank for those values and this relationship among the values affects the model's interpretation of the values.**

**FRAMEWORK**

**The Workflow diagram illustrates the components of the system and how they function. This framework can predict the crop's yield. The**[**Fig. 1**](https://www.sciencedirect.com/science/article/pii/S0965997822002277#fig0001)**depicts a clear image of how data is computed from the agriculture website and pre-processed.**

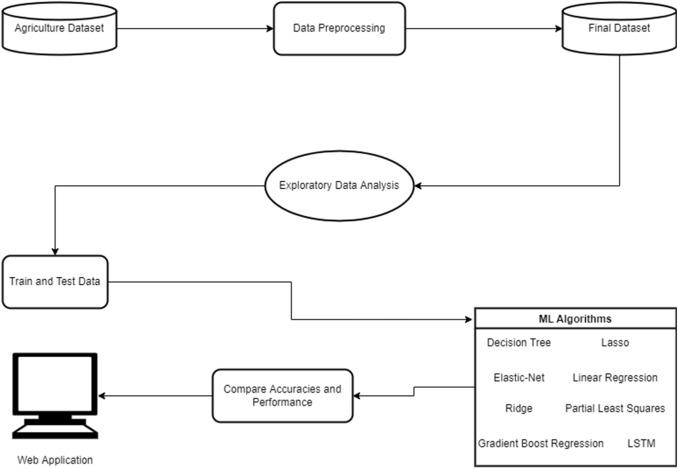


Fig. 1. Work Flow Diagram

The machine learning algorithms we took upon for this prediction were Gaussian Process Regression and autoregressive integrated moving average.

For comparing the efficiency of these regression models, we used [Mean Absolute Error](https://www.sciencedirect.com/topics/engineering/mean-absolute-error), [Root Mean Square Error](https://www.sciencedirect.com/topics/engineering/root-mean-square-error). So, our architecture focuses on the three modules i.e., Data Pre-processing, Training models, and Testing, comparing results.

A picture containing diagram

Description automatically generated

R^2 = coefficient of determination

SS res = sum of squares for residuals

SS tot = total sum of the squares



MAE = Mean Absolute Error

*n* = Number of errors

Σ = Symbol for summation

|xi – x| = Absolute error



p: Polynomial number

Gaussian Progress Regression for yield prediction:

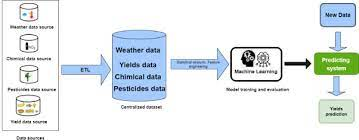
The GP regression is a powerful yet disciplined nonparametric method that has seen wide applications in statistical analysis and machine learning. This modeling approach is probabilistic in nature and yields not only point estimates but entire predictive distributions. This is particularly appealing to one of the primary focuses on the crop yield estimation, which is to obtain reliable predictive yield distribution.

ARIMA Model:

ARIMA, short for ‘Auto Regressive Integrated Moving Average’ is actually a class of models that ‘explains’ a given time series based on its own past values, that is, its own lags and the lagged forecast errors, so that equation can be used to forecast future values

**Any ‘non-seasonal’ time series that exhibits patterns and is not a random white noise can be modeled with ARIMA models.**

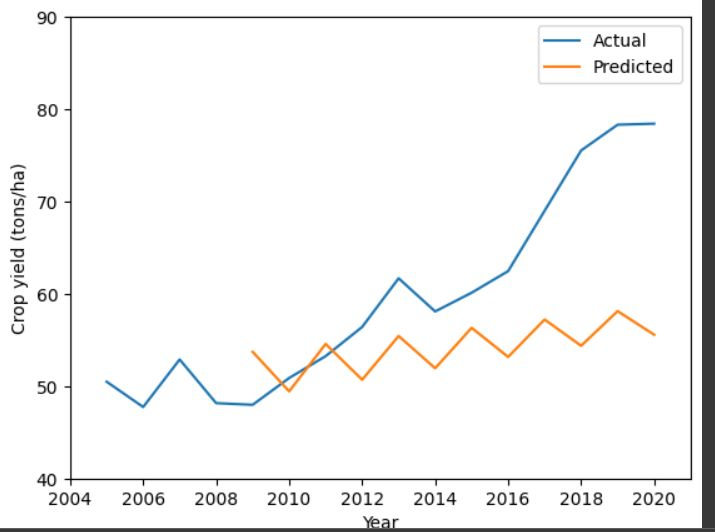
**An ARIMA model is characterized by 3 terms: p, d, q.**

****

## 

# DATA ANALYSIS:

Predictions from Autoregressive integrated moving average (ARIMA).

****

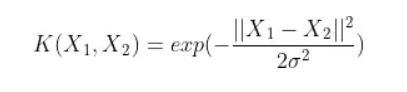
**Fig 4.2:** Crop yield vs Year

We obtained a Root mean squared error of **12.124183423096063**

* 1. **Gaussian Process Regression.**

Kernel Choice:

* + - The RBF kernel function for two points X₁ and X₂ computes the similarity or how close they are to each other.



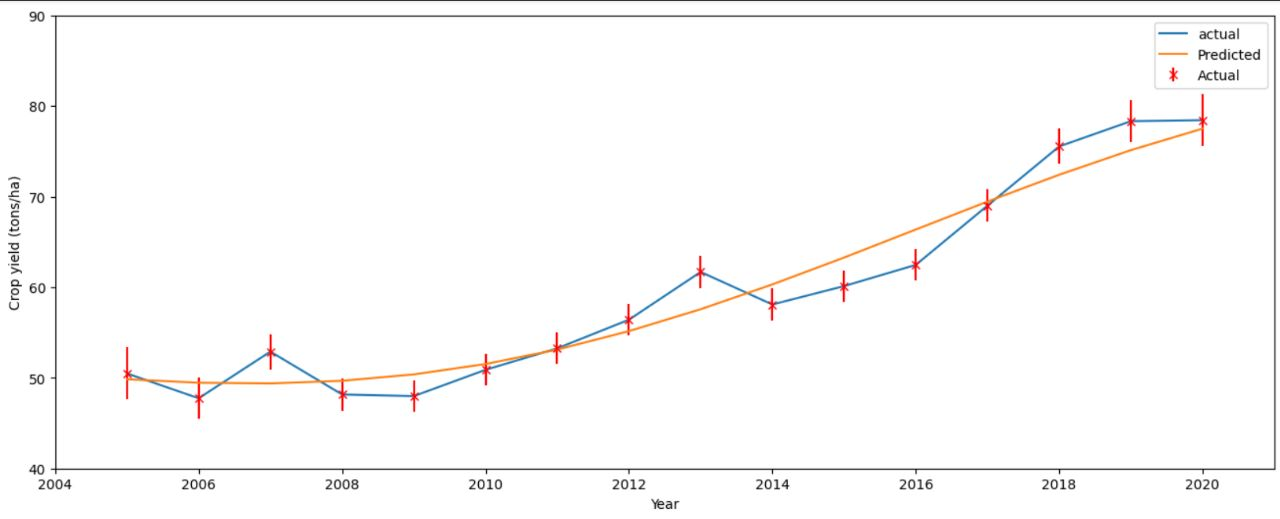
where,  
1. ‘σ’ is the variance and our hyperparameter  
2. ||*X₁ - X₂||*is the Euclidean (L*₂*-norm) Distance between two points X₁ and X₂

White kernel.

The main use-case of this kernel is as part of a sum-kernel where it explains the noise of the signal as independently and identically normally-distributed.

Constant kernel.

used as part of a product-kernel where it scales the magnitude of the other factor (kernel) or as part of a sum-kernel, where it modifies the mean of the Gaussian process.

Fig: crop yield vs Year, using Gaussian Process regression

### We obtained a Root mean squared error of 2.4200917103329895

### Mean Absolute Error: 2.055229991668719

### Mean Square Error: 5.856843886422455

THANK YOU!

## 