Crop Yield Prediction Using Big Data Techniques

M.Sc. Agri Analytics

Big Data Analytics

Submitted by

Group No: 2

Submission date

22nd March 2024

Submitted to

Mr. Kapil Oberai

Indian Institute of Remote Sensing

Individual Contribution in project

Sr.no	Name	Student ID	Contribution
1	Gusai Nidhi	202319015	Create AutoML Regressor, Linear Regression Model in BigQueryML and Report and PPT Creation
2	Nayakapara Drashti	202319019	Create Random Forest Regressor and Boosted Tree Regressor L1 Model in BigQueryML, Dashboard Creation
3	Prajapati Divya	202319024	Project Idea and Data Finding, Dashboard Creation and Report and PPT Creation and Finalization

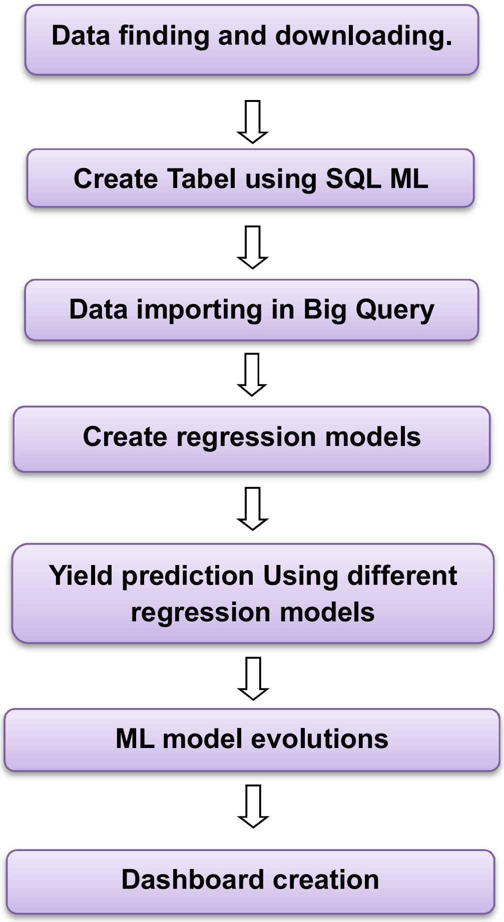
Introduction

The agriculture industry is changing in recent years due to technological and data analytics breakthroughs, which are enabling data-driven problem solving of real-time issues. Crop yield prediction is one of the key areas where advanced technologies are applied. Based on yield estimates, farmers, the national food security department, policy makers, and stockholders can make educated judgments. It also oversees the best use of resources and the reduction of production-related risks in agriculture.

Objective

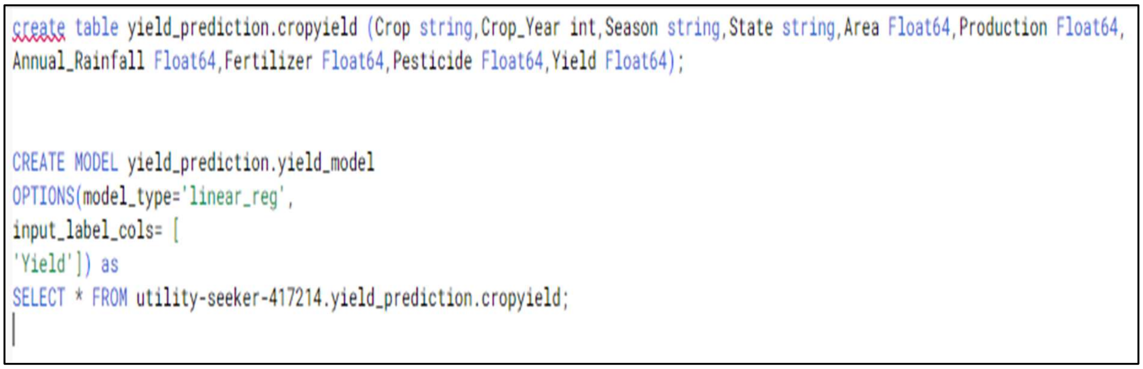
- To get exposures to big data techniques.
- To explore Big Query ML of google cloud for yield prediction.
- To create insightful dashboard for crop yield analysis using Google locker studio.

Flow chart of crop yield prediction


Data finding and downloading.
Create Tabel using SQL ML
Data importing in Big Query
Create regression models
Yield prediction Using different regression models
ML model evolutions
Dashboard creation

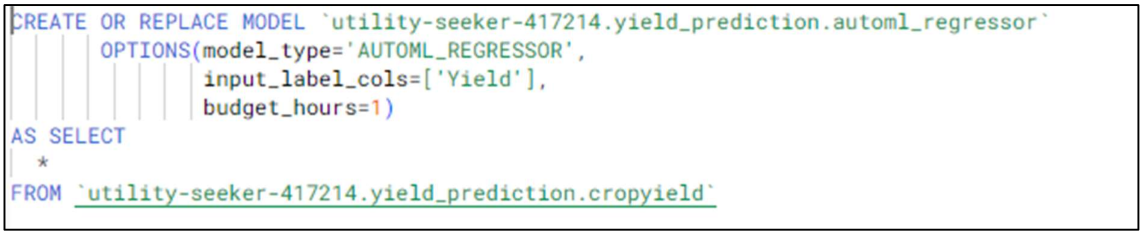
Machine learning models used in project.

Linear Regression: A linear relationship between independent factors and a continuous dependent variable is fitted using the statistical technique of linear regression to forecast. It forecasts numerical results that fall inside a given range and are not endless or undefinable, which avoids problems like overflow or inaccurate forecasts.



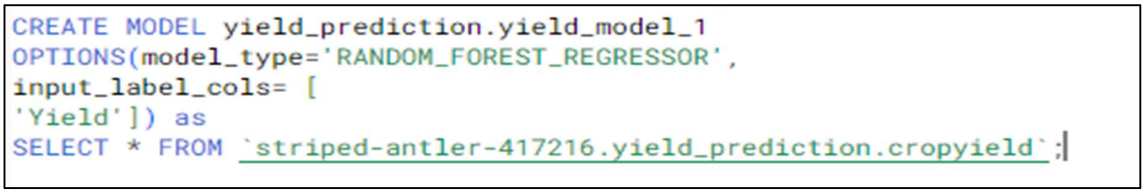
```
create table yield_prediction.cropyield (Crop string,Crop_Year int,Season string,State string,Area Float64,Production Float64,
Annual_Rainfall Float64,Fertilizer Float64,Pesticide Float64,Yield Float64);


CREATE MODEL yield_prediction.yield_model
OPTIONS(model_type='linear_reg',
input_label_cols= [
'Yield']) as
SELECT * FROM utility-seeker-417214.yield_prediction.cropyield;
```

AutoML Regressor: The outcome variable that is being predicted is usually represented by a label column with numeric data type in the regression model. Regression analysis requires mathematical processes, which are made easier by numerical data. The model can learn and predict continuous numerical values by include this column, which enables it to produce precise forecasts based on the input features.



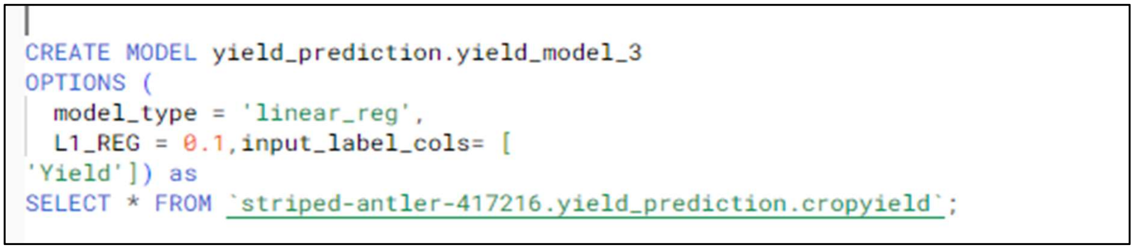
```
CREATE OR REPLACE MODEL `utility-seeker-417214.yield_prediction.automl_regressor`
      OPTIONS(model_type='AUTOML_REGRESSOR',
              input_label_cols=['Yield'],
              budget_hours=1)
AS SELECT
  *
FROM `utility-seeker-417214.yield_prediction.cropyield`
```

Random Forest Regressor: A machine learning approach called Random Forest builds several decision trees using various data subsets. By averaging these trees' results, it improves the accuracy of its predictions. It reduces overfitting problems by merging several predictions, creating a strong model that can produce precise classifications or predictions.



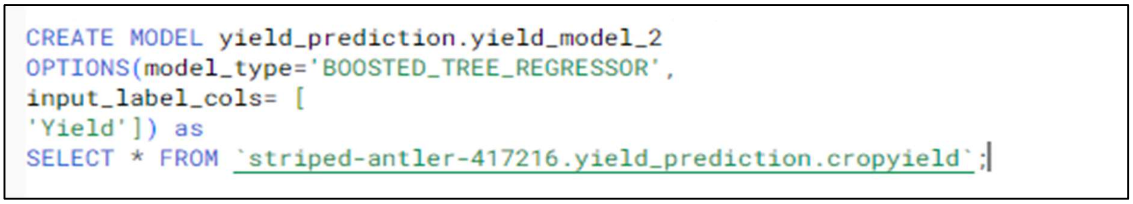
```
CREATE MODEL yield_prediction.yield_model_1
OPTIONS(model_type='RANDOM_FOREST_REGRESSOR',
input_label_cols= [
'Yield']) as
SELECT * FROM `striped-antler-417216.yield_prediction.cropyield`;
```

Linear Regression with L1 Regularization: In addition to the conventional least squares goal, Lasso regression, also known as linear regression with L1 regularization, aims to minimize the sum of the absolute values of the coefficients. By promoting sparse solutions, this regularization method significantly lowers model complexity and may even enhance generalization performance, which helps in feature selection.



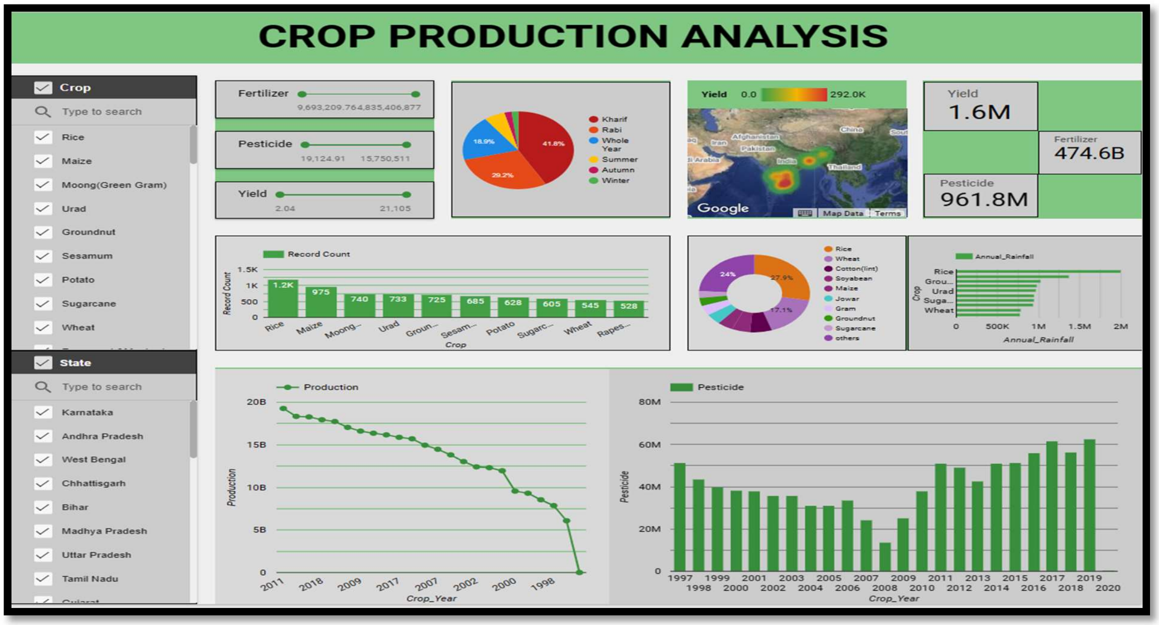
```
CREATE MODEL yield_prediction.yield_model_3
OPTIONS (
  model_type = 'linear_reg',
  L1_REG = 0.1,input_label_cols= [
'Yield']) as
SELECT * FROM `striped-antler-417216.yield_prediction.cropyield`;
```

Bosted Tree Regressor: BigQuery uses Boosted Tree Regression for predictive modeling tasks involving huge datasets. When solving regression issues, it is employed to forecast a continuous numerical value. Boosted Trees are a good choice for applications such as risk assessment, demand prediction, and sales forecasting since they combine numerous weak learners repeatedly and provide excellent predictive accuracy.



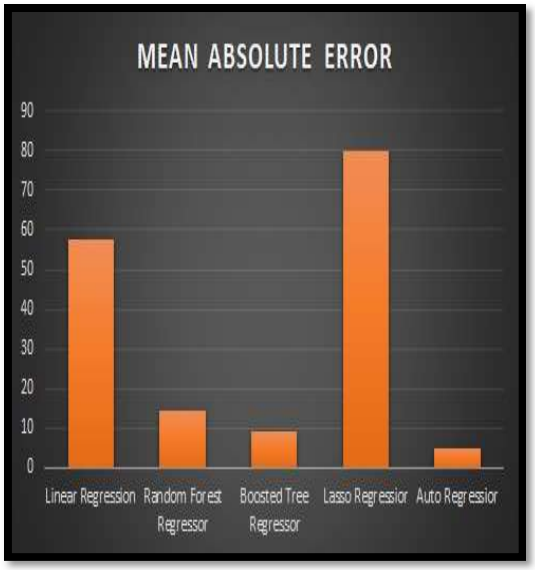
```
CREATE MODEL yield_prediction.yield_model_2
OPTIONS(model_type='BOOSTED_TREE_REGRESSOR',
input_label_cols= [
'Yield']) as
SELECT * FROM `striped-antler-417216.yield_prediction.cropyield`;
```

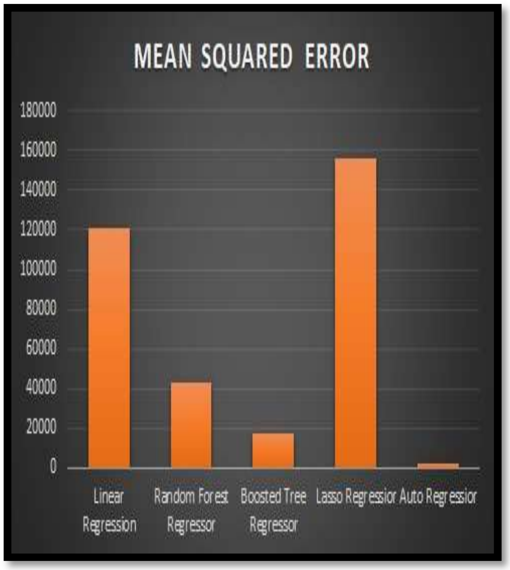
DASHBOARD

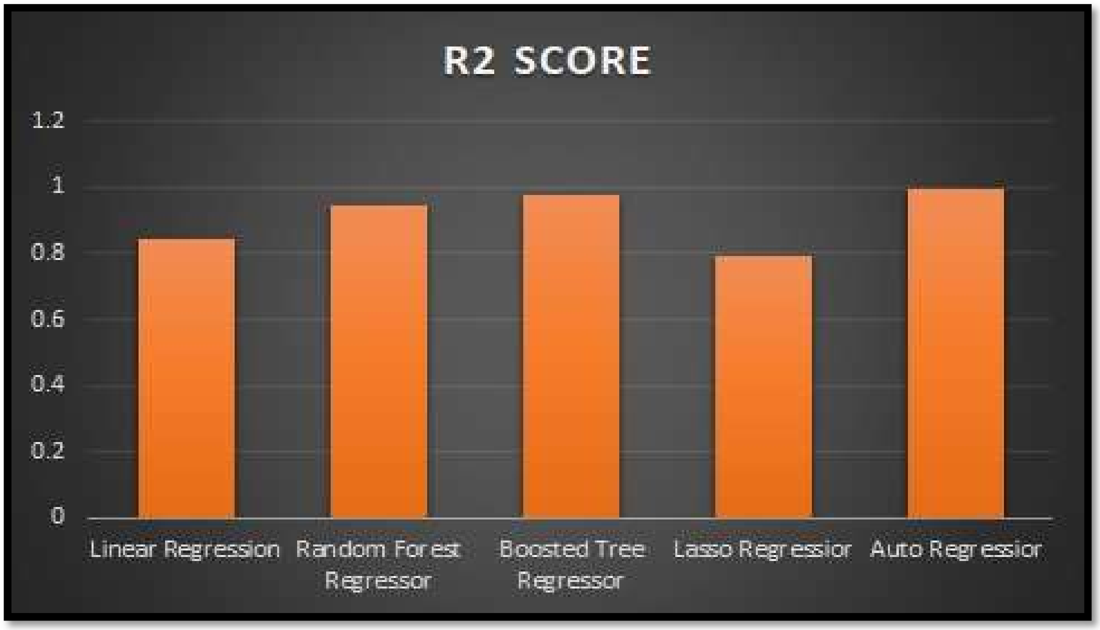


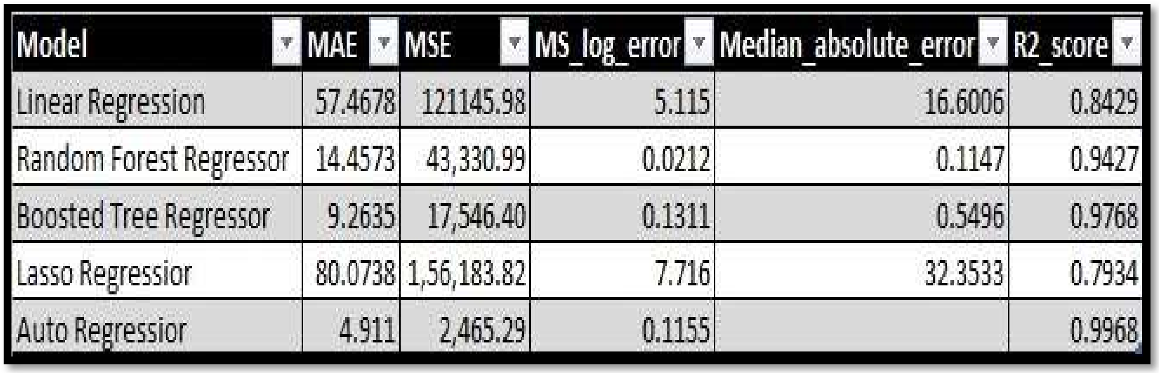
https://lookerstudio.google.com/reporting/b0cc6f82-425b-451c-b012-0826b2bc2a1e

EVALUATION



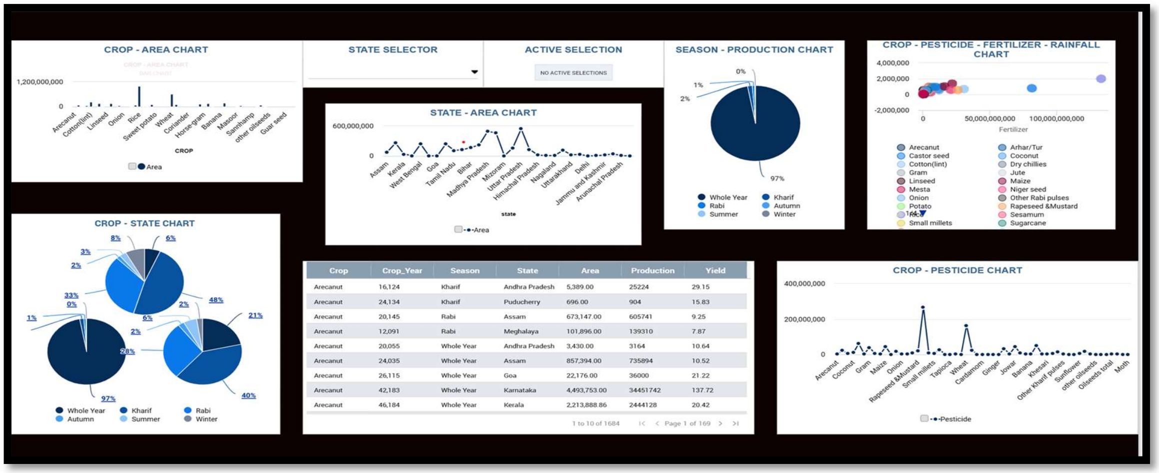






Model	MAE	MSE	MS_log_error	Median_absolute_error	R2_score
Linear Regression	57.4678	121145.98	5.115	16.6006	0.8429
Random Forest Regressor	14.4573	43,330.99	0.0212	0.1147	0.9427
Boosted Tree Regressor	9.2635	17,546.40	0.1311	0.5496	0.9768
Lasso Regressior	80.0738	1,56,183.82	7.716	32.3533	0.7934
Auto Regressior	4.911	2,465.29	0.1155		0.9968

DASHBORAD IN KNOWAGE SOFTWARE



Crop	Crop_Year	Season	State	Area	Production	Yield
Arecanut	16,124	Kharif	Andhra Pradesh	5,389.00	25224	29.15
Arecanut	24,134	Kharif	Puducherry	696.00	904	15.83
Arecanut	20,145	Rabi	Assam	673,147.00	605741	9.25
Arecanut	12,091	Rabi	Meghalaya	101,896.00	139310	7.87
Arecanut	20,055	Whole Year	Andhra Pradesh	3,430.00	3164	10.64
Arecanut	24,035	Whole Year	Assam	857,394.00	735894	10.52
Arecanut	26,115	Whole Year	Goa	22,176.00	36000	21.22
Arecanut	42,183	Whole Year	Karnataka	4,493,753.00	34451742	137.72
Arecanut	46,184	Whole Year	Kerala	2,213,888.86	2444128	20.42

1 to 10 of 1684 |< < Page 1 of 169 > >|

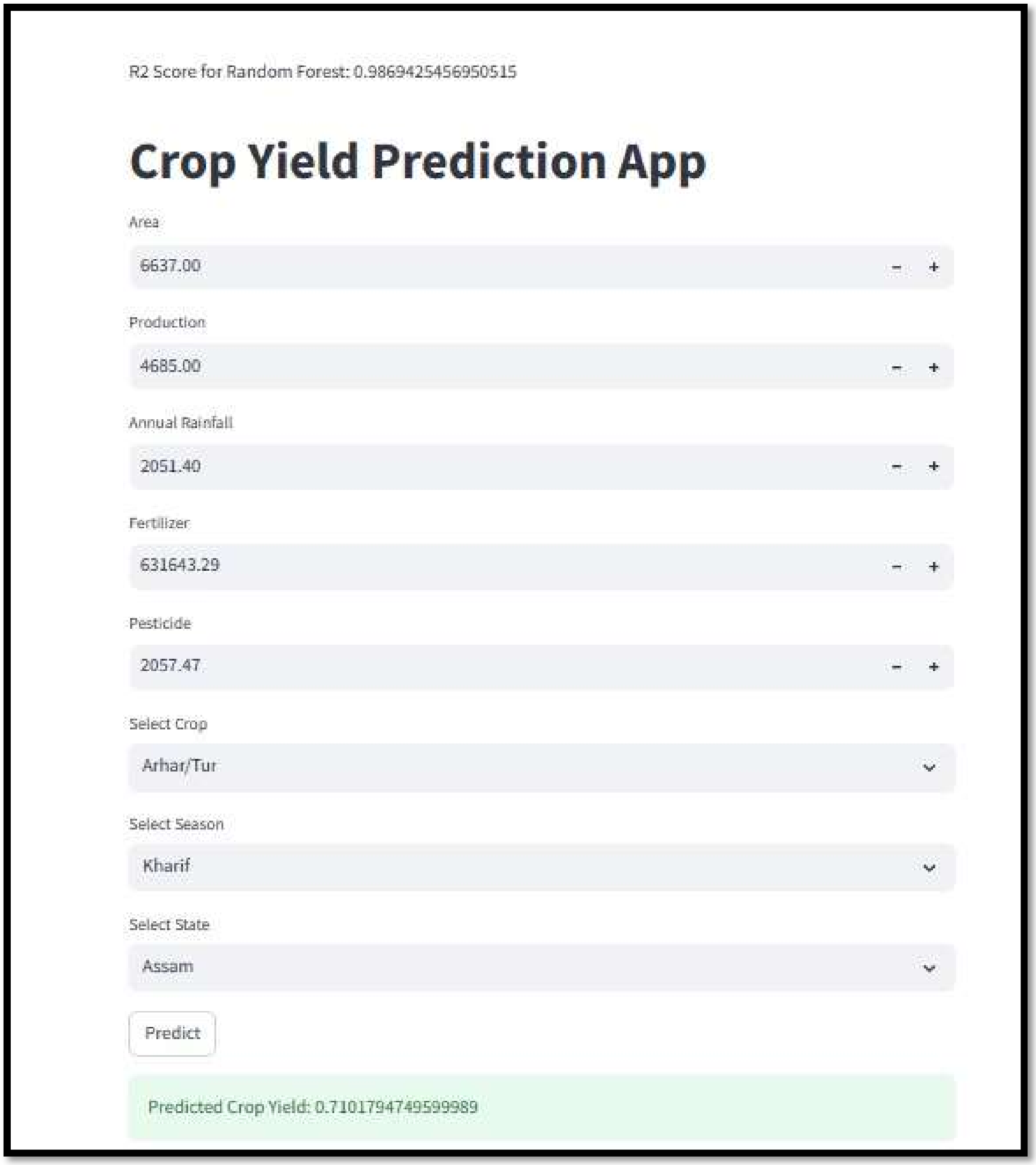
https://demo.knowage-suite.com/knowage-vue/document-browser/document-composite/dashboard

Comparison of Knowage with Looker Studio

	Knowage	Looker Studio
Download and Installation	Downloading and installation of Software is a complicated and time consuming process.	Looker typically offers a user-friendly interface where users can drag and drop data fields to create visualizations and assemble dashboards without needing extensive coding or technical expertise.
Data integration and connectivity	Uploading data is a time consuming prosess.	In comparison to knowage, uploading the data doesn't take too long.
Visualization Capabilities	Less types of charts are available in Knowage compared to looker studio.	Looker offers robust visualization options, customizable dashboards, and dynamic reports. LookML (Looker Modeling Language) allows for deep customization and rich visualization features.
Reference	Videos aren't very informative, and documentation isn't very good either.	The user interface is simple to use and has decent documentation.

CONCLUSION: With a R2 score of 0.9768, the Boosted Tree Regressor has excellent prediction accuracy. The Lasso Regressor, on the other hand, has an accuracy of 0.7934 and works less well. This disparity highlights how well boosting techniques capture intricate patterns in the data when compared to more straightforward linear regression methods such as Lasso.

MODEL DEPLOY USING STREAMLIT LIBRARY



R2 Score for Random Forest: 0.9869425456950515

Crop Yield Prediction App

Area

6637.00

Production

4685.00

Annual Rainfall

2051.40

Fertilizer

631643.29

Pesticide

2057.47

Select Crop

Arhar/Tur

Select Season

Kharif

Select State

Assam

Predicted Crop Yield: 0.7101794749599989

Query Link:

https://console.cloud.google.com/bigquery?project=striped-antler-417216&ws=!1m4!1m3!8m2!1s1016104865611!2s4a440e0864e64507a449fa6dd8905c43

https://console.cloud.google.com/bigquery?sq=39333987166:62917e895754419699bbda349c55b5bd

Dataset Link:

https://www.kaggle.com/datasets/akshatgupta7/crop-yield-in-indian-states-dataset