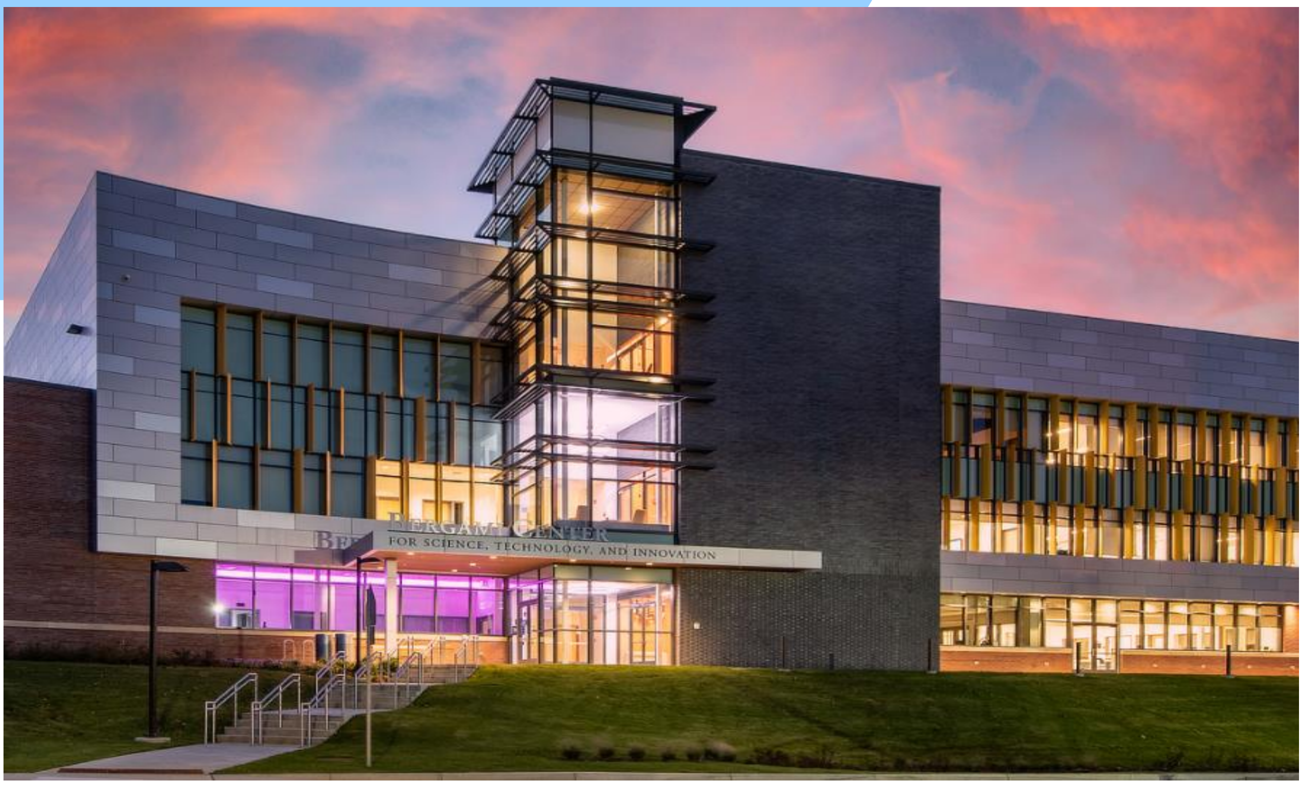
University of New Haven
TAGLIATELA COLLEGE OF ENGINEERING
Electrical & Computer Engineering and Computer Science

DISTRIBUTED AND SCALABLE DATA ENGINEERING

TECHNICAL REPORT



Team Members:
Name 1 Sireesha Chimbili
Name 2 Sushma Mandati
Name 3 Poojitha Mandapati
Name 4 Mustak Ahamed Yadiki

CONTENTS

ABSTRACT:

The science of training machines to learn and produce models for future predictions is widely used, and not for nothing. Agriculture plays a critical role in the global economy. With the continuing expansion of the human population understanding worldwide crop yield is central to addressing food security challenges and reducing the impacts of climate change. Crop yield prediction is an important agricultural problem. The Agricultural yield primarily depends on weather conditions (rain, temperature, etc), pesticides and accurate information about history of crop yield is an important thing for making decisions related to agricultural risk management and future predictions. Yield prediction benefits the farmers in reducing their losses and to get best prices for their crops. In our current times, owing to unforeseeable climate change, farmers are unable to achieve a reasonable amount of crop production.

In order to feed the World's growing population, it is important to integrate new and innovative technologies and resources in the agricultural sector. This Study Focuses on training machine learning models to predict the crop production of the world's most popular crops grown. Factors such as Rainfall, Temperature and Pesticide Input are considered in predicting the crop yield. We compare the accuracy of regression models such as Decision Tree Regressor, Gradient Boosting Regressor, Random Forest Regressor.

INTRODUCTION:

Agriculture is one of the most significant factors in the growth of the developing countries such as India where the agricultural ecosystem contributes to about 17-18% of the country's GDP. Agriculture and related industries employ more than 70% of the nation's population and thus is a key source of survival for many. Agriculture also plays a crucial role in the global economy. With the continued expansion of human population awareness of global crop yields is essential to resolving food security issues and reducing the effects of climate change. Crop yield forecasting is an important agricultural problem. Policy makers depend on accurate predictions to pass legislations on import and export policies to strengthen national food security. Farmers also benefit from accurate predictions by making informed strategic management and financial decisions.

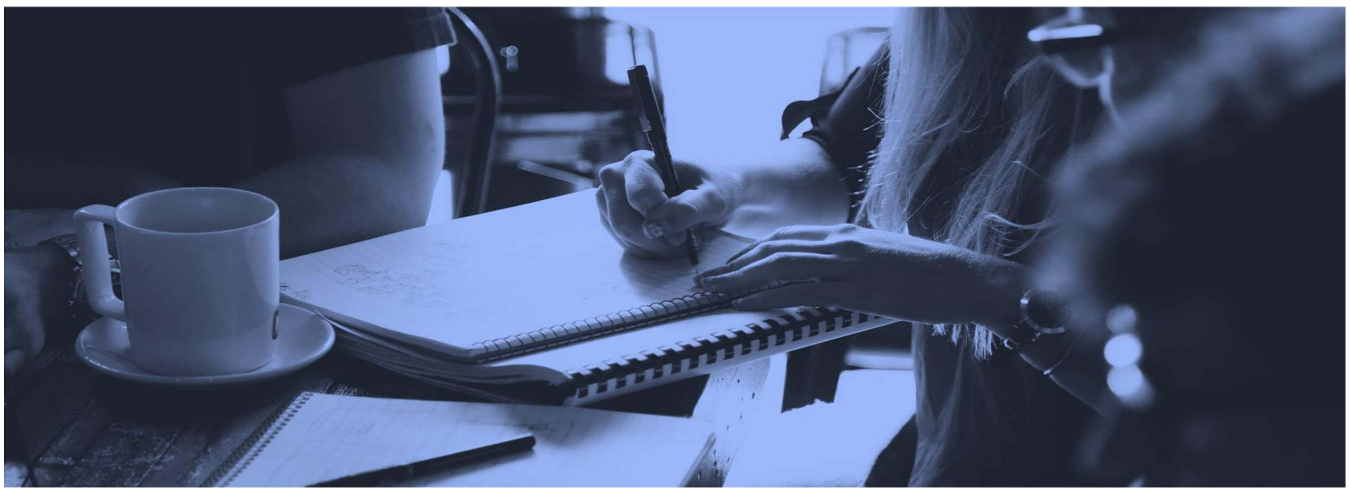
Agricultural yield depends primarily on weather conditions such as rain, temperature, etc. and environmental conditions such as Soil Quality, pesticides etc. Accurate knowledge on the history of crop yields is critical for decision-making on agricultural risk management and future predictions. Although cuisine varies greatly across the world, the essential ingredients that support humans are very similar. The World consumes a lot of maize, wheat, rice and other basic crops. In this study, machine learning approaches are used to forecast the 10 most consumed crops using publicly accessible data from the Food and Agriculture Organization (FAO) and the World Data Bank.The goal is to analyze the impact of the features (rainfall, Temp…) on the crop productivity and identify the most productive crops and predict the future productivity based on the history of data using the Regressor models like Linear Regression, Decision Tree Regressor, Random Forest Regressor etc…

Executive Summary:

- Our main objective is to develop a model to predict crop yield prediction based on the inputs provided by the user like Area, Crop, the impacted features.
- We also succeeded in analyzing which features impact the most to judge the yield of the crop.
- We have selected only a few regions/countries as we have a large group of data. So, we have selected the top 10 Regions where agriculture is the main occupation.
- We have done the pre-processing of the data and trained the model using the cleaned data.
- Used Data Visualization from the pandas, sklearn, numpy libraries to analyze the impact of the features on the productivity and the highest produced crops and predict the future productivity.
- Split the data to training and testing with 75% of the training data and 25% of the data to testing data.
- Different regressor algorithms have been used like Gradient Boosting Regressor, Xboost Regressor, Decision Tree Regressor, Support Vector Machines(SVM), Random Forest Regressor to train the data.
- We have used AWS Resources like Amazon Sage maker, S3 bucket, Sagemaker Estimator, EC2, Flask App, Visual Studio Code, Python Advanced Libraries to work on the project.

Highlights of the project:

- Crop Yield Dataset has been taken from kaggle.
- The data like Pesticides, Yield data is collected from Food and Agriculture Organization(FAO).
- The other relevant data like Weather data like Avg. Temperature and Rainfall is collected from World Data Bank.
- CRISP Methodology has been used for the analysis and to design the model as it is satisfying the business challenge.
- For modeling we have used regression models like SVM, Gradient booster regression, random forest regression and decision tree regression to build the model and R2 score model to evaluate the model.
- The efficiency of the model is calculated using Mean Squared Error/Mean Averaged Error to find how the model works on the test data and compared the predicted output with the actual outputs.

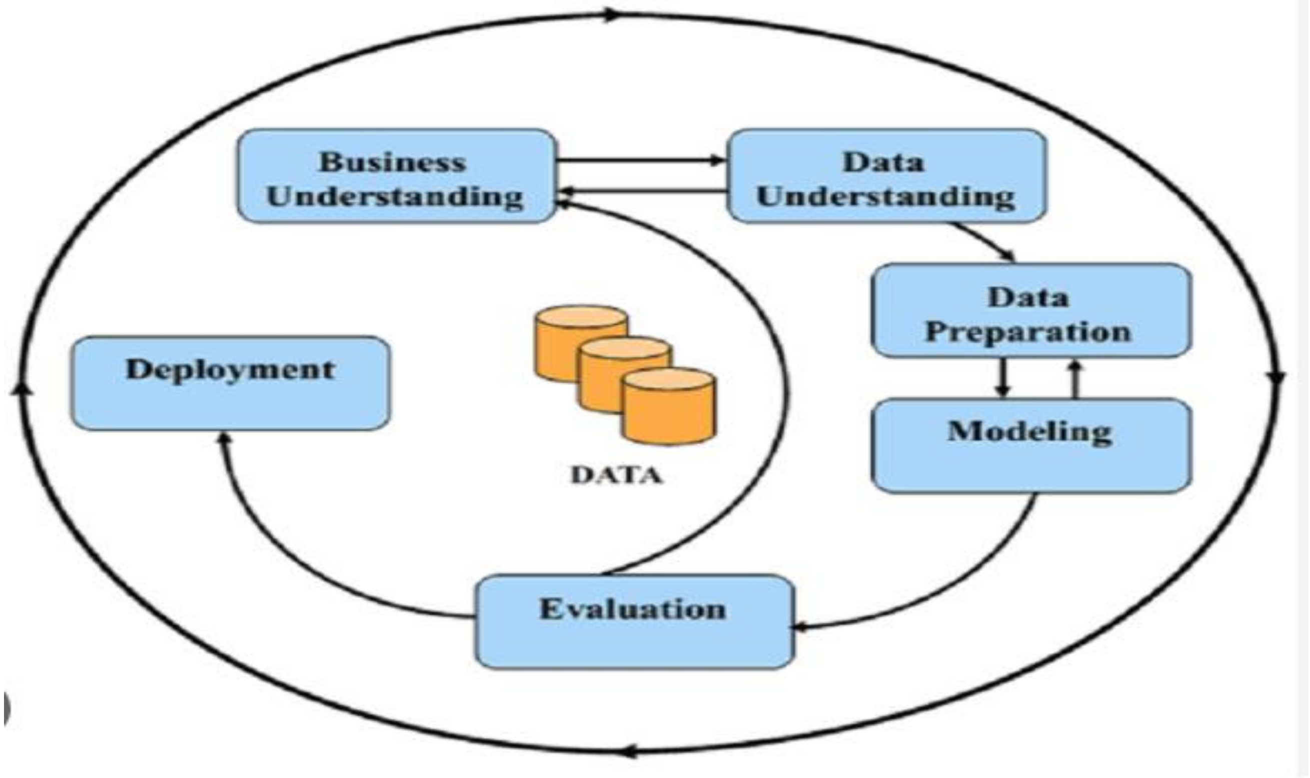


Methodology:

The project involved CRISP methodology, which includes:

- Business understanding
- Data understanding
- Data preparation
- Modelling
- Evaluation
- Deployment

This is the CRISP methodology, where it involves the understanding of the challenge and the data then, building the model using the data through which it is evaluated as if it is satisfying the business challenge and making it deployed.



Business Understanding

We have studied the dataset and the dataset contains the data of 10 crops over 101 countries with different features like rainfall, Avg. temperature, pesticides, year, Area, Crop, yield, Domain code, pesticide Name etc.

Data Understanding:

We understood all the attributes and their spread among the yield of the crop and for this, we have used different data visualizations like Correlation(using heat map to find the correlations between the features) and selected the features that have the most impact on the target field. Using Box plot to find the outliers of the data and removed the rows.

Data Preparation:

We have prepared the data like replacing the null values to appropriate values, Encoded the data using One hot encoder, Label Encoder to convert the categorical features of type objects to the numerical values (int) which is the acceptable format for the models. Scaled the data using the sklearn preprocessing library to bring all the features to the same level of magnitude.

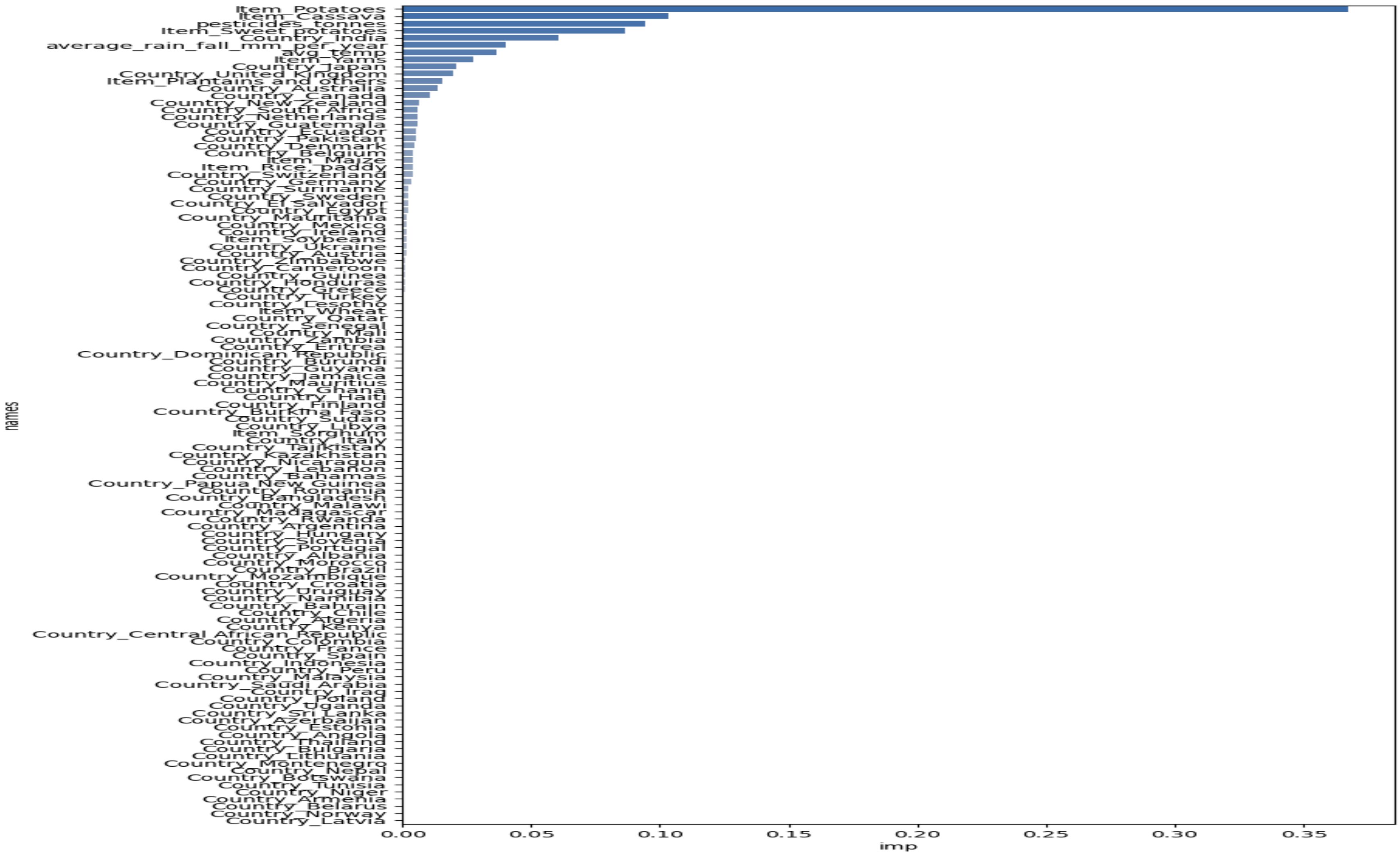


Fig: Showing the importance of the features and the crops given the probability of the yield

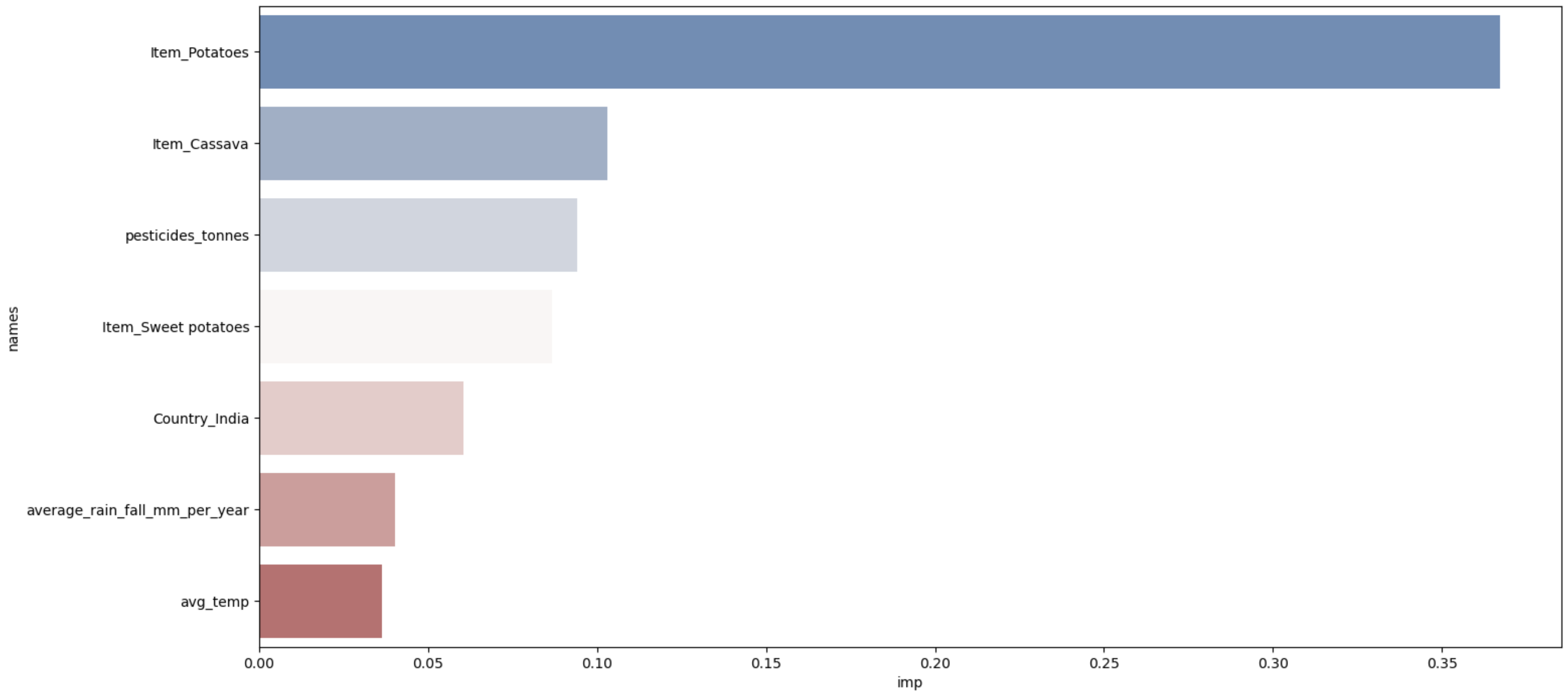


Fig: The most 7 impacted features on the Yield

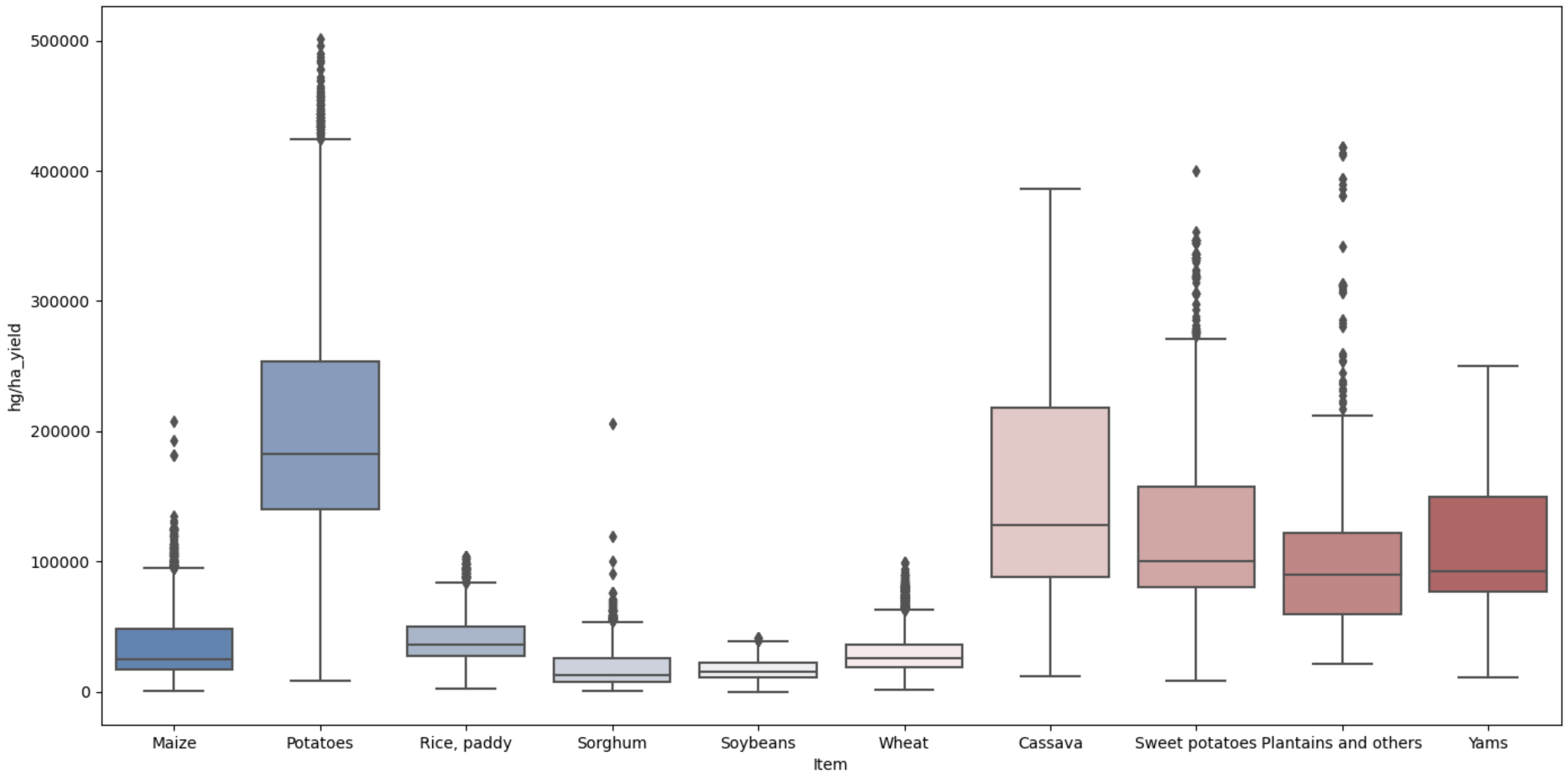


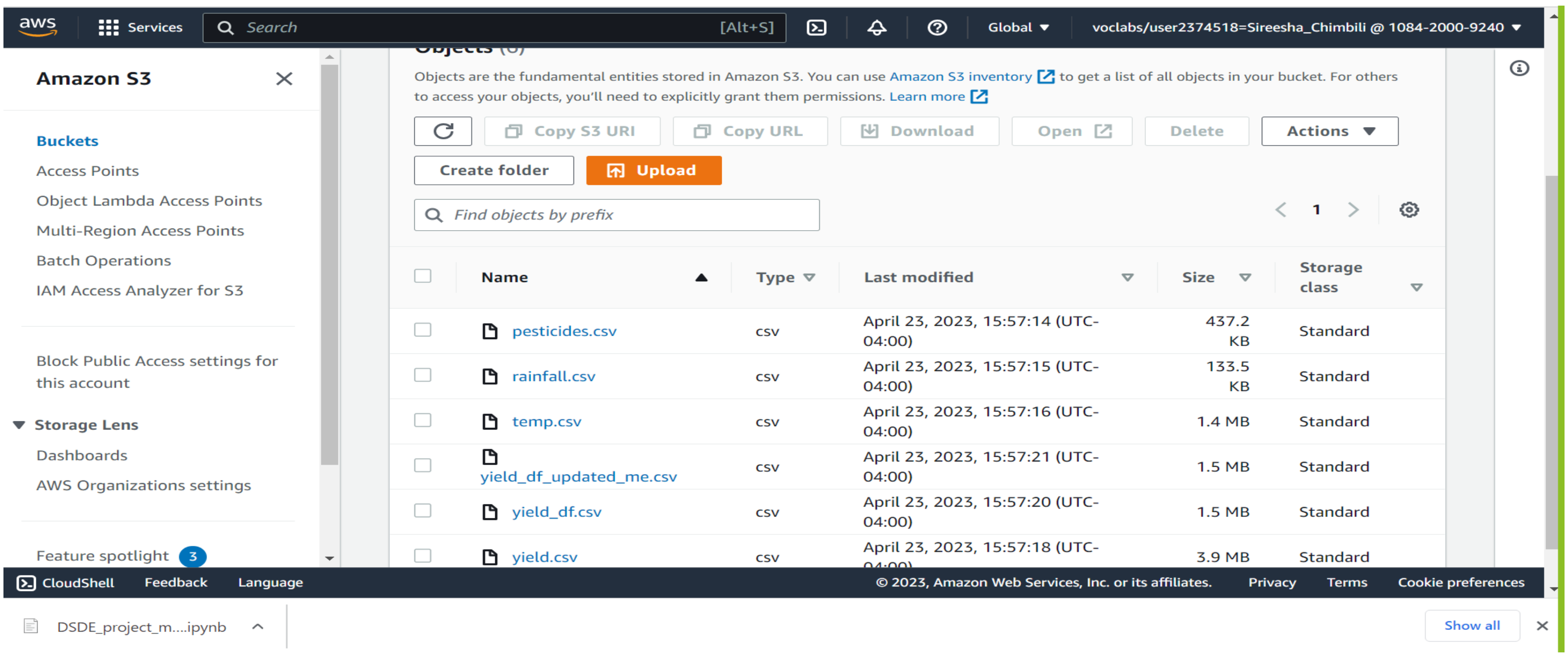
Fig: Box plot showing the yield spread of the Crops.

Modeling:

We have used Decision Tree Regressor to build the model and our target is to predict the Yield of the crop based on the inputs provided by the user.

For the Modeling we have targeted to use AWS resources to design the app and to work on the project, We have used Amazon sagemaker to create the notebook instance in which we have imported all the required libraries of sagemaker, python to model the project.

We have created an S3 bucket to access the data. Using Boto3, we have uploaded the data to the S3 bucket from the Kaggle and from there, we retrieved the data and cleaned all the data as we had 5 different data sets each representing a feature of our concentration.



Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more

	Name	Type	Last modified	Size	Storage class
□	pesticides.csv	csv	April 23, 2023, 15:57:14 (UTC-04:00)	437.2 KB	Standard
□	rainfall.csv	csv	April 23, 2023, 15:57:15 (UTC-04:00)	133.5 KB	Standard
□	temp.csv	csv	April 23, 2023, 15:57:16 (UTC-04:00)	1.4 MB	Standard
□	yield_df_updated_me.csv	csv	April 23, 2023, 15:57:21 (UTC-04:00)	1.5 MB	Standard
□	yield_df.csv	csv	April 23, 2023, 15:57:20 (UTC-04:00)	1.5 MB	Standard
□	yield.csv	csv	April 23, 2023, 15:57:18 (UTC-04:00)	3.9 MB	Standard

We have cleaned each dataset individually and worked on processing the data and removed the unnecessary features for the area of implementation.

We have merged all the datasets to a single datset which containing all the necessary features for building the model.

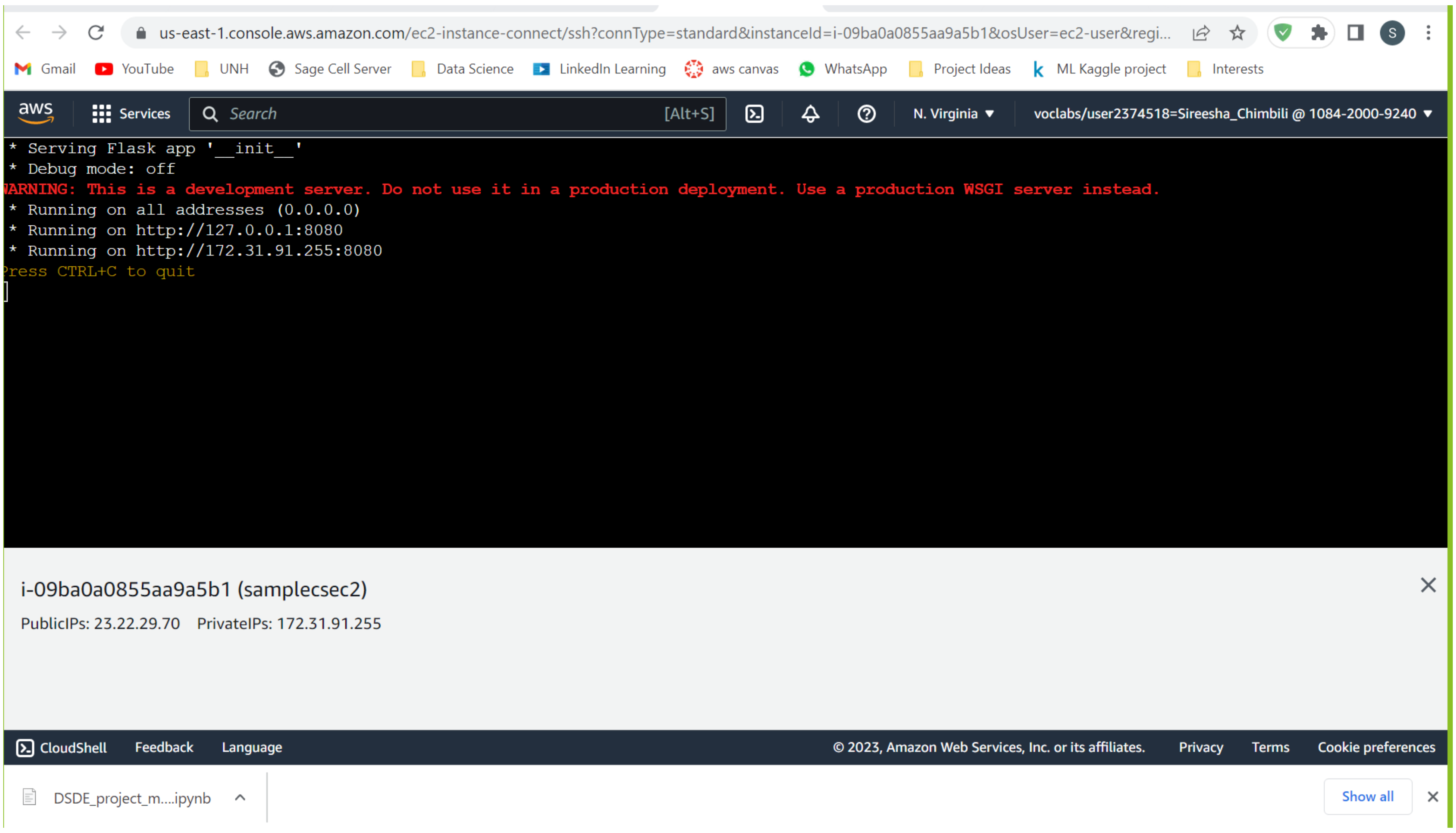
Then using Correlation, we have identified the correlation between the features. We have removed some of the features which are highly correlated to each other and the selected features have a good correlation with respect to the target field.

Using Boxplot, the data which is not appropriate is been removed and we saw that data is not scaled as it contained different representations and units as per each individual data set, so, we have scaled the data using the preprocessing library.

The data has had some features representing in the object type which is not acceptable format. So, we have encoded the data using one hot encoder.

The data is been trained using the sage maker estimator of Xgboost and decision tree and performed well on the test data too.

The code is first tested by deploying it in the local server using the flask app and then it's been deployed in the EC2 instance to run on all web host addresses and allowed the port of 8080 in both http and https web addresses.


us-east-1.console.aws.amazon.com/ec2-instance-connect/ssh?connType=standard&instanceId=i-09ba0a0855aa9a5b1&osUser=ec2-user&regi...
Gmail
YouTube
UNH
Sage Cell Server
Data Science
LinkedIn Learning
aws canvas
WhatsApp
Project Ideas
ML Kaggle project
Interests
aws
Services
Search
[Alt+S]
N. Virginia
voclabs/user2374518=Sireesha_Chimbili @ 1084-2000-9240

```
 * Serving Flask app '__init__'
 * Debug mode: off
ARNING: This is a development server. Do not use it in a production deployment. Use a production WSGI server instead.
 * Running on all addresses (0.0.0.0)
 * Running on http://127.0.0.1:8080
 * Running on http://172.31.91.255:8080
ress CTRL+C to quit
```

i-09ba0a0855aa9a5b1 (samplecsec2)

PublicIPs: 23.22.29.70 PrivateIPs: 172.31.91.255

Evaluation:

The accuracy score has been used to evaluate the model. 75% of the data has been used for training and 25% is to test the model and we have achieved 95.64% accuracy and it can be improved by training the model with more data.

Deployment:

- We dumped our model as a pickle file and built a flask app and then deployed the flask app in the EC2 instance to run on all addresses.
- The Deployed model can be accessed through http://ec2-54-211-144-106.compute-1.amazonaws.com:8080/ or http://127.0.0.1:5000

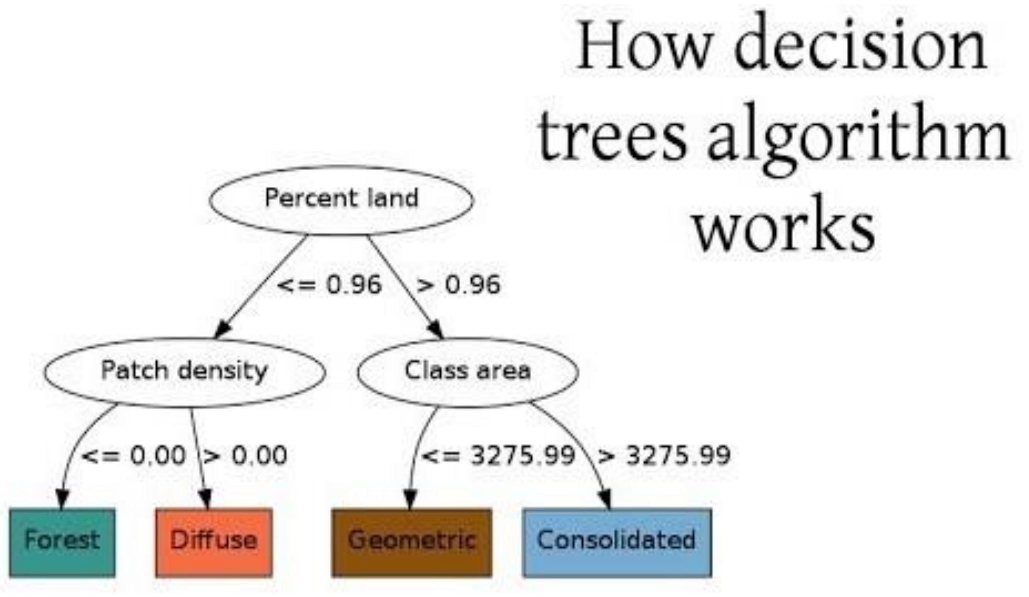
Models used:

- Decision Tree Regressor

Why have we used Decision Tree Regressor algorithm?

Decision Tree regressor model is a method commonly used in data mining applications. The aim of the model is to predict the value of a dependent variable based on several independent variables. The Decision tree iteratively makes decisions on the value of a particular independent variable and continually classifies the dependent variable to make prediction easier. Each internal node of the tree asks a simple question about the value of a certain input feature.

Working of Decision Tree Regressor algorithm:



The following steps explain the working Decision Tree Regressor Algorithm:

Step 1: This algorithm will construct a decision tree for every training data.

Step 2: If the land percent is >= 0.96 patch density is calculated otherwise class area is calculated.

Step 3: if patch density is <=0.00 land is considered as forest if not considered as diffuse.

Step 4: if the class area is <=2375.99 area is considered as Geometric if not consolidated.

Results:

Here while applying these techniques to dataset, we also evaluate methods by following metrics:

• Classification Accuracy: it is a measure of correct predictions in percentage

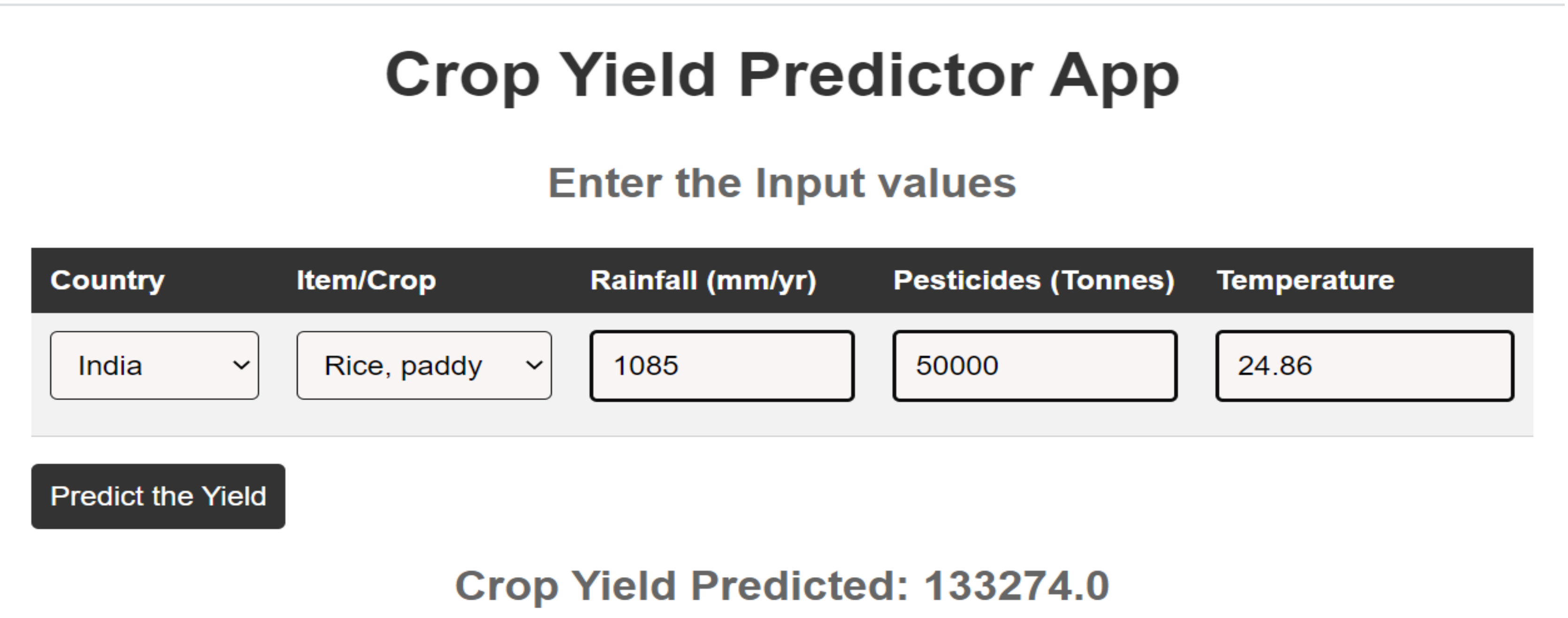
• Confusion Matrix: Tabular format to describe performance of model o True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN)

• Precision: it measures how often correct positive value is predicted.

• AUC Score: it is area under ROC (Receiver Operating Characteristic) curve in percentage.

Model	Accuracy (%)
Decision Tree Regressor	95.64%
Random Forest Regressor	68.54%
Gradient Boosting Algorithm	89.76%
SVM	-19.89%

Here are few snapshots of working of the app using the model designed.

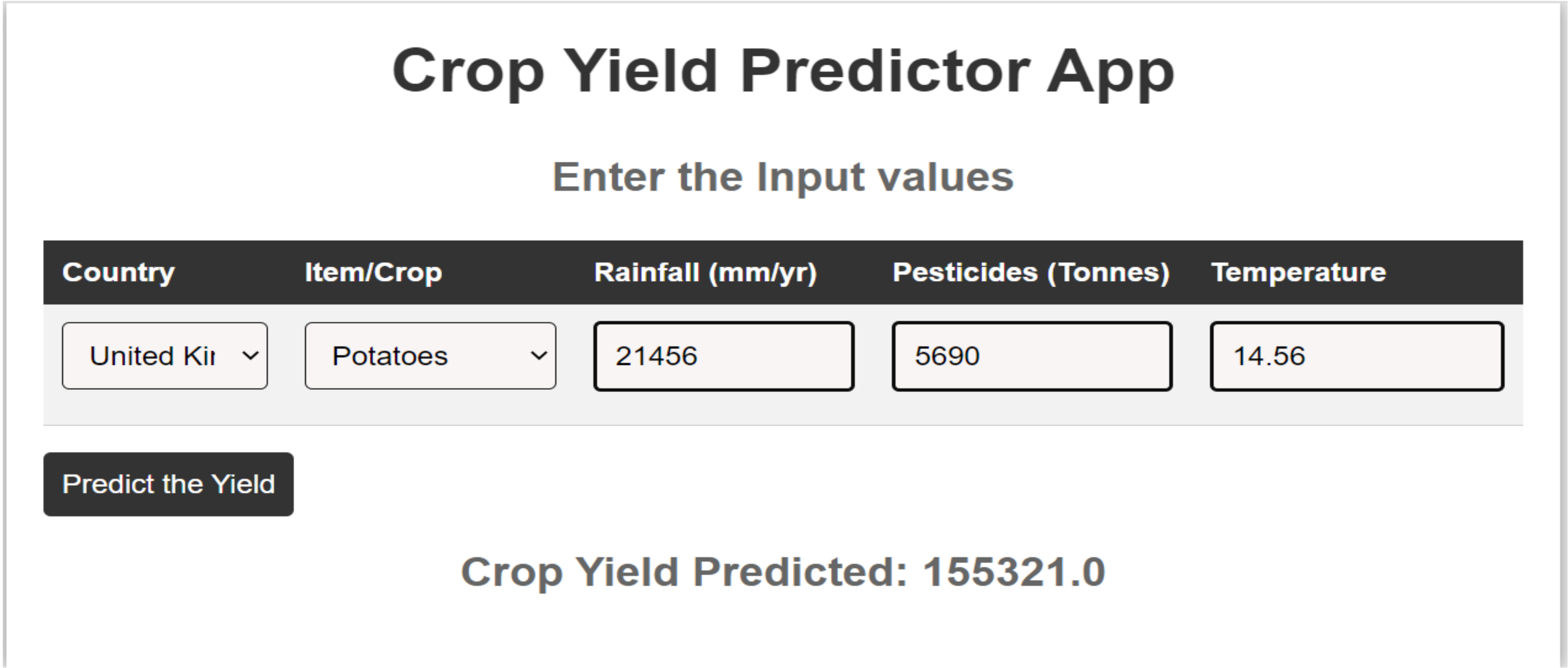


Crop Yield Predictor App

Enter the Input values

Country	Item/Crop	Rainfall (mm/yr)	Pesticides (Tonnes)	Temperature
India	Rice, paddy	1085	50000	24.86

Crop Yield Predicted: 133274.0



Crop Yield Predictor App

Enter the Input values

Country	Item/Crop	Rainfall (mm/yr)	Pesticides (Tonnes)	Temperature
United Ki	Potatoes	21456	5690	14.56

Crop Yield Predicted: 155321.0

Conclusion:

A model is build to predict the Yield of the crop based on the several factors that impact the crop. The Yield depends on many factor which are some controlled by the human and some are not. If we can understand what crops suits the temperature or the soil condition of the area/country, the average rainfall per year or the amount of pesticides used in terms of tons on the particular cop, or the average temperature of the area, it can help the farmers in deciding the best suited crop for their conditions and We can analyze which crop is produced on a large scale in terms of areas based on the conditions specified.

This is the motivation for us to study this specific instance on finding the best suited crop/ identifying the yield of the crop. So, we have built a model where the model is trained with the data received by the food and weather organizations based on the attributes and the model can predict the avg yield that can be gained based on the conditions/inputs provided by the user based on their area and crop.

Contributions/References:

- Y. J. N. Kumar, V. Spandana, V. S. Vaishnavi, K. Neha and V. G. R. R. Devi, "Supervised Machine learning Approach for Crop Yield Prediction in Agriculture Sector", *2020 5th International Conference on Communication and Electronics Systems (ICCES)*, pp. 736-741, 2020, June.
- A. Patil, S. Kokate, P. Patil, V. Panpatil and R. Sapkal, "Crop Prediction using Machine Learning Algorithms", *International Journal of Advancements in Engineering & Technology*, vol. 1, no. 1, pp. 1-8, 2020.
- E. Khosla, R. Dharavath, and R. Priya, "Crop yield prediction using aggregated rainfall-based modular artificial neural networks and support vector regression," Environment, Development and Sustainability, pp. 1-22, 2019.
- S. Khaki and L. Wang, "Crop yield prediction using deep neural networks", *Frontiers in plant science*, vol. 10, pp. 621, 2019.
- A. Chandgude, N. Harpale, D. Jadhav, P. Pawar and S. M. Patil, "A Review on Machine Learning Algorithm Used For Crop Monitoring System in Agriculture", *International Research Journal of Engineering and Technology (IRJET)*, vol. 5, no. 04, pp. 1470, 2018.
- S. Sunder, “India economic survey 2018: Farmers gain as agriculture mechanisation speeds up, but more r&D needed,” The Financial Express, 2018 .
- D. Ramesh and B. V. Vardhan, "Analysis of crop yield prediction using data mining techniques," International Journal of research in engineering and technology, vol. 4, no. 1, pp. 47-473, 2015.