Factors affecting Crop Production in Central and Southern Asia

Ananya Upadhyay | upadha3@rpi.edu

Rensselaer Polytechnic Institute,, Troy, NY, United States,

Abstract

This poster addresses the critical issue of food security in Central and Southern Asia, a region of 12 countries(India, Pakistan, Sri Lanka, Maldives, Bangladesh, Tajikistan, Turkmenistan, Uzbekistan, Kyrgyz Republic, Kazakhstan, Nepal) heavily reliant on wheat, rice, and potatoes as staple crops. Fluctuating yields due to complex interactions of climate, operational, and economic factors threaten food stability. My research aimed to identify key factors influencing yield variations and develop a robust model to predict yields for these vital crops. Employing four comprehensive datasets and advanced machine learning models, particularly Random Forest, I achieved exceptional accuracy in yield forecasting. This empowers policymakers to proactively manage resources and ensure food availability, while enabling farmers to optimize practices and maximize productivity. This also aligns with the UN's Zero Hunger goal, paving the way for a food-secure future in the region.

My research was motivated by a profound understanding of the critical role of yield prediction in ensuring food security. This will help unraveling the secrets of crop yields in Central and Southern Asia, I pave the way for a future where food security is not a distant aspiration but a tangible reality for all.

Data

The data varies across the years 1961 to 2021 and consists of the following:
- World Development Index data of the economical agricultural factors from the region of Central and Southern Asia.
- Crop and Livestock Products from Food and Agriculture Organization for the three staple crops(Rice, Wheat and Potato) in those regions.
- Observed Temperature Data (in Degree C)
- Observed Precipitation Data (in mm)

Methods

1.Multiple Linear Regression Model (MLR):
In my yield project, MLR uncovers the linear relationship between crop yields and multiple factors, providing insights into how various variables collectively influence production.

2.Decision Tree Model:
Decision Tree Model segments data based on critical factors, capturing non-linear dynamics and interactions to understand decision-making in agriculture.

3.Random Forest Model:
Tailored to my yield project, the Random Forest Model, an ensemble of decision trees, excels in predicting crop yields by leveraging diverse data subsets and random feature selections.

4.Polynomial Regression Model:
Polynomial Regression extends beyond linear relationships, accommodating non-linear terms to understand nuanced interactions between variables influencing crop yields.

Workflow and Results

Common workflow for the Project

Data Collection And Merging	• From various online resources consisting Enviornmental and Econmic features of three staple crops in the Central and South Asian Regions
Data Exploration and Preparation	• Explore the dataset to get descriptive results • Find out data features and outliers, then filter them.
Model Development and Evaluation	• Build the Models • Evaluate them based on factors such as RMSE and R-Squared Values, etc.

1. Data Collection and Merging

R language (R)

```
wdi.data<-read.csv('WDI Data.csv')
climate.pcpt.data<-read.csv('climatePcpt.csv')
potato<-read.csv('yield_potato.csv')
rice<-read.csv('yield_rice.csv')
wheat<-read.csv('yield_wheat.csv')
data1<-left_join(wdi.data,climate.pcpt.data,by=c('Country_Code','Country_Name','Year'))
data2<-left_join(potato,rice,by=c('Country_Name','Year'))
data2<-left_join(data2,wheat,by=c('Country_Name','Year'))
data<-left_join(data1,data2,by=c('Country_Name','Year'))
```

The data was collected and merged as mentioned in the above R code.

2. Data Exploration and Preparation

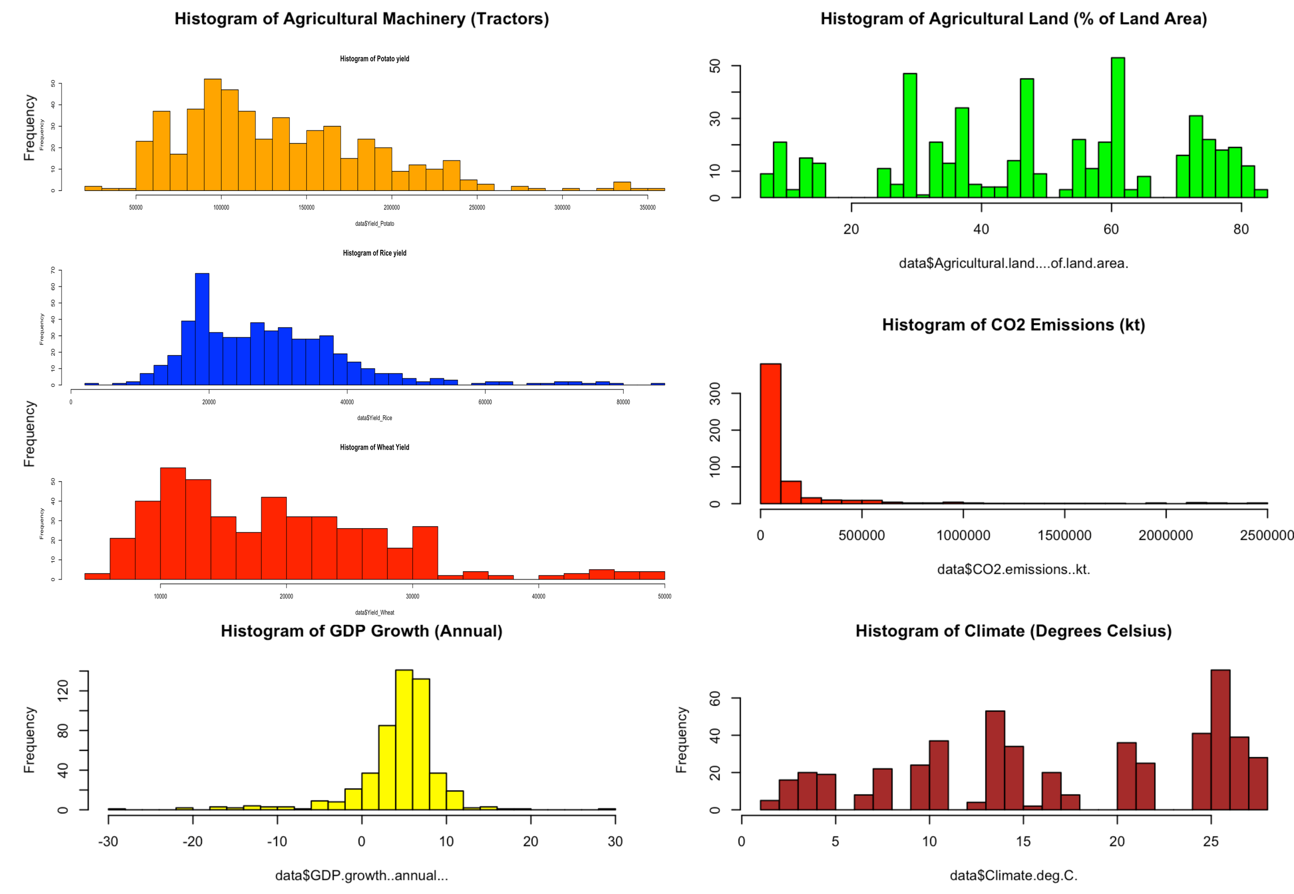
Several methods are used to inspect the dataset.

A. Explore the dataset:

Three commonly used exploratory plots are scatter, box, and histograms.

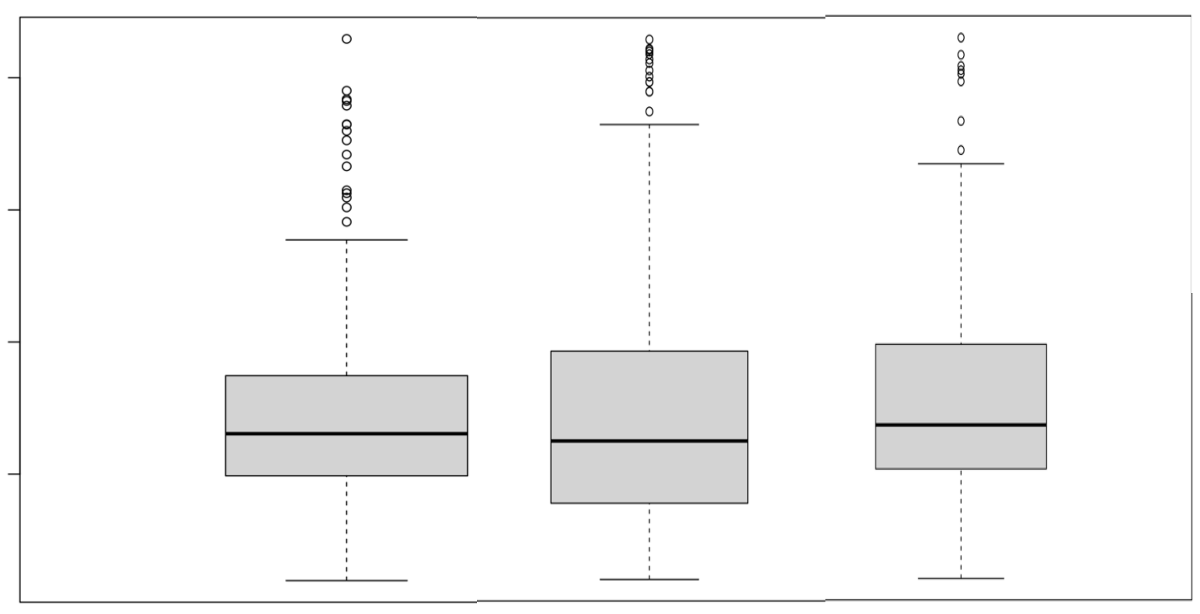
i. Histograms(**R:**)

```
hist(data$Yield_Potato, main = "Histogram of Potato yield", col = "
hist(data$Yield_Rice, main = "Histogram of Rice yield", col = "blue
hist(data$Yield_Wheat, main= "Histogram of Wheat Yield", col="red",
```



ii. Box plot **R**:

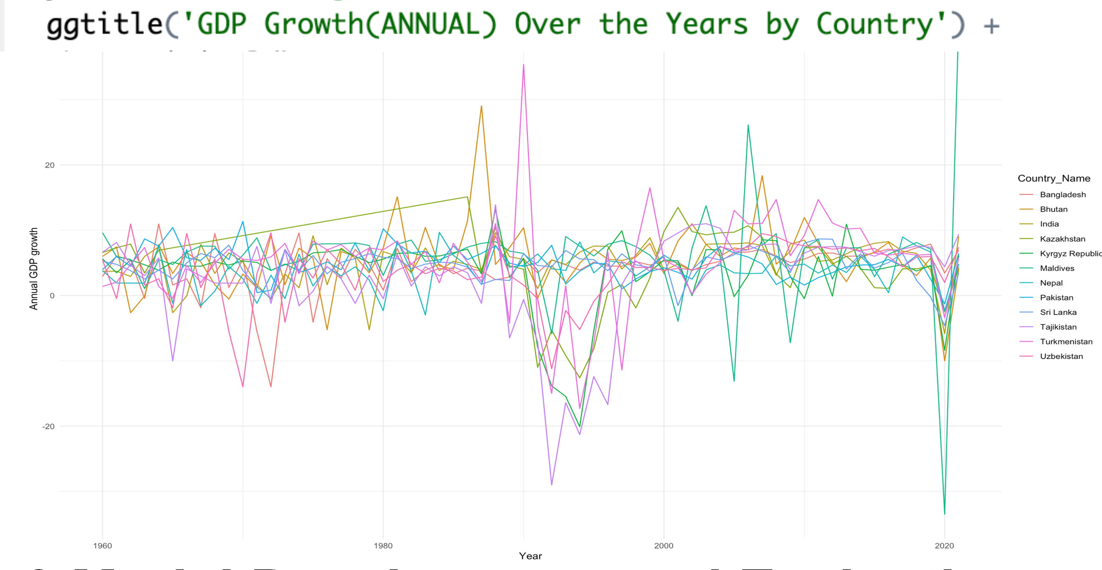
```
boxplot(numeric_data[[var]], main = paste("Boxplot of'
```



iii. Line Plot

Plotted the attributes over the years for each country.

```
ggplot(data |, aes(x = Year, y = GDP.growth..annual...
                  , group = Country_Name, color = Country_Name)) +
  geom_line() +
  xlab('Year') +
  ylab('Annual GDP growth') +
  ggtitle('GDP Growth(ANNUAL) Over the Years by Country') +
```



3.Model Development and Evaluation

1. Multiple Linear Regression Model (MLR):

MLR	Rice	Potato	Wheat
RMSE	6444.398	23622.58	4401.186
RSQUARE	0.65081	0.7790014	0.7582456

Potato has the highest R-squared value, indicating strong explanatory power, Wheat stands out with the lowest RMSE, suggesting more accurate predictions.

2.Decision Tree Model:

Decision Tree	Rice	Potato	Wheat
RMSE	5691.204	25895.34	4651.095
RSQUARE	0.6312062	0.7363685	0.6522271

Decision Tree model shows varying performance across different crops, with improvements in some cases (Rice and Wheat) and degradation in others (Potato) compared to the MLR model.

3. Random Forest Model:

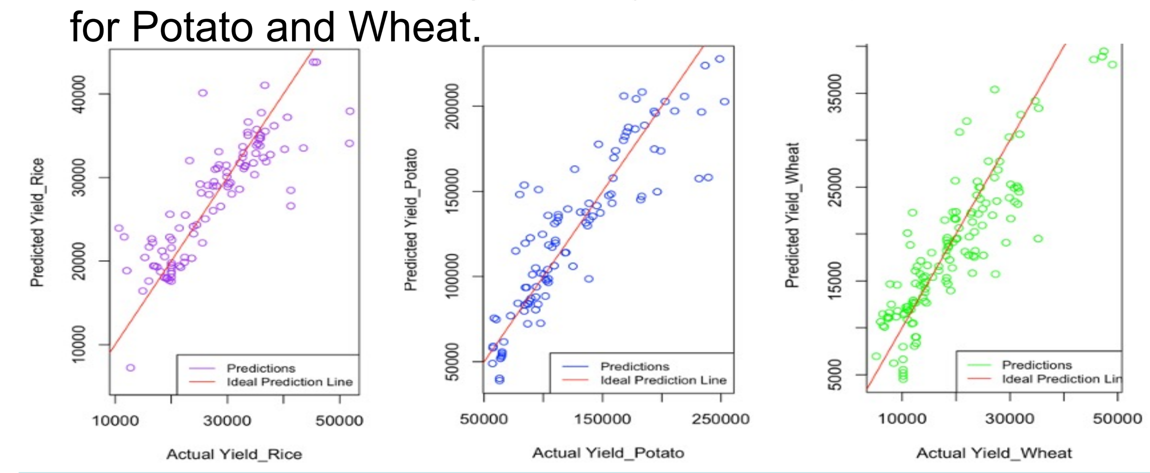
Random Forest	Rice	Potato	Wheat
RMSE	4672.501	18394.53	3175.575
RSQUARE	0.7514155	0.8669755	0.8378824

Random Forest model outperforms both the Multiple Linear Regression and Decision Tree models across all three crops (Rice, Potato, and Wheat) in terms of both RMSE and R-squared. Gives a better predictive accuracy.

4.Polynomial Regression Model:

Polynomial Regression	Rice	Potato	Wheat
RMSE	5018.033	23494.38	4462.389
RSQUARE	0.6764281	0.7813935	0.7514752

Polynomial Regression model shows mixed results across different crops. It appears to improve the performance for Rice but does not significantly enhance predictive accuracy for Potato and Wheat.



Conclusion

In conclusion, this study aimed to address food security in Central and Southern Asia by employing advanced machine learning models to predict crop yields. While the Multiple Linear Regression model demonstrated moderate accuracy, the Random Forest Regressor emerged as the most effective, consistently outperforming other models across wheat, rice, and potatoes. This model not only enhanced predictive accuracy but also provided valuable insights for policymakers and farmers to manage resources proactively, contributing to the region's food security and aligning with the UN's Zero Hunger initiative.

Glossary:

R – A program to process data and perform statistical analysis

UN- United Nations is an intergovernmental organization whose stated purposes are to achieve international cooperation.

RMSE- Root Mean Square Error, average magnitude of prediction errors.

Rsquared- quantifies the proportion of variance in the dependent variable explained by the independent variable(s), varies between 0 to 1.

Resources:

https://sdgs.un.org/goals
https://www.worldbank.org/en/home
https://www.fao.org/faostat/en/

Related Work:

-Crane-Droesch, A. (2018). Machine learning methods for crop yield prediction and climate change impact assessment in agriculture. Environmental Research Letters, 13(11), 114003

-Bowman, M. S., & Zilberman, D. (2013). Economic factors affecting diversified farming systems. Ecology and society, 18(1).

-Nigam, A., Garg, S., Agrawal, A., & Agrawal, P. (2019, November). Crop yield prediction using machine learning algorithms. In 2019 Fifth International Conference on Image Information Processing (ICIIP) (pp. 125-130). IEEE

-Everingham, Y., Sexton, J., Skocaj, D., & Inman-Bamber, G. (2016). Accurate prediction of sugarcane yield using a random forest algorithm. Agronomy for sustainable development, 36(2), 27