Crop Production Analysis and Yield Prediction

Authors: Pareekshit Reddy Gaddam, Sree Krishna Suresh, Vinitha Joyce Marathi

GitHub Link

1. Summary

In agriculture, information on the quantity of crops that are produced is known as yield data. It incorporates the yield of crops per hectare of land. Depending on the crop, this can be expressed in tons, bushels, or any other unit of measurement. It crucial since it may assist farmers in making informed decisions about many aspects of farming, such as seed selection, pest management, irrigation scheduling and can be utilized to identify the advantages and disadvantages of various farming techniques and formulate recommendations. Using yield statistics, governments, non-governmental organizations, and other stakeholders may track their progress toward global objectives. But predicting the crop output is exceedingly difficult as it depends on so many variables, including crop genotype, environmental conditions, management strategies, and their interactions, as well as soil, climatic, and crop data [1]. Since, the atmospheric conditions and the terrain, impact the yield, using remote sensing data to predict yield is being currently explored across world. This is because the solution is large scale, continuous, multispectral, long term and cost effective and long-term.

The main objective of this project is to develop an integrated solution which allows a user to access various kinds of yield data predictions and its corresponding explorations along with interactive visualizations including crop disease prediction. We have also developed a website and an interface which provides the user to explore yield variation across years for various crops and countries and visualize the forecasted yield. This implementation uses Time Series Forecasting models such as ARIMA [1], Multivariate LSTM [2] and Image based deep Neural Networks to predict yield. This report focuses on the extensions from phase one involving time series modelling, data analysis and further explores soybean yield prediction using remote sensing data for Illinois, USA.

2. Datasets Overview

As this project is multifaceted, it cannot be addressed using just one dataset. Hence data acquisition is a key step for this project. We acquired the yield (hg/ht), average temperature (°C), pesticide (tonnes), agricultural land area (1000 ha), fertilizer (tonnes) data from FAO (Food and Agriculture Organization) database. This consists of values corresponding to 245 different regions across the world and 300 different crops items, starting from the year 1961 to 2020. The rainfall (mm) data was acquired form World Data Bank.

The remote sensing data was acquired from Earth Engine's public data archive which includes more than forty years of historical imagery and scientific datasets. Primary archives were the Moderate Resolution Imaging Spectroradiometer (MODIS) from which tera surface reflectance,

land surface temperature, land cover type and Daymet V4 were acquired to get daily surface weather and climatological summaries. Also, to get the county boundaries we used TIGER archive by US Census Counties 2018. Further, the corresponding county wise yield data of soybean, state ANSI (American National Standards Institute) code and county ANSI code was taken from the repository of United States Department of Agriculture National. For this project, the focus is on data from year 2002 to 2016.

3. Methods

We preprocessed the data and performed feature engineering to extract the important features in the data (The features that contributed the most to the prediction). For yield, data was unclean and had unnecessary columns and rows, only necessary columns and rows were selected. For example, we dropped the rows which represented whole continents or groups of regions as we only wanted data specific to countries. Also, we imputed the data with zero for data not available. Further, the unique items(crops) were grouped, and their names were modified for ease of use. Next the top 33 crops were identified according to their yield value and their global usage. For the same respective countries were filtered which resulted in 207 countries in total.

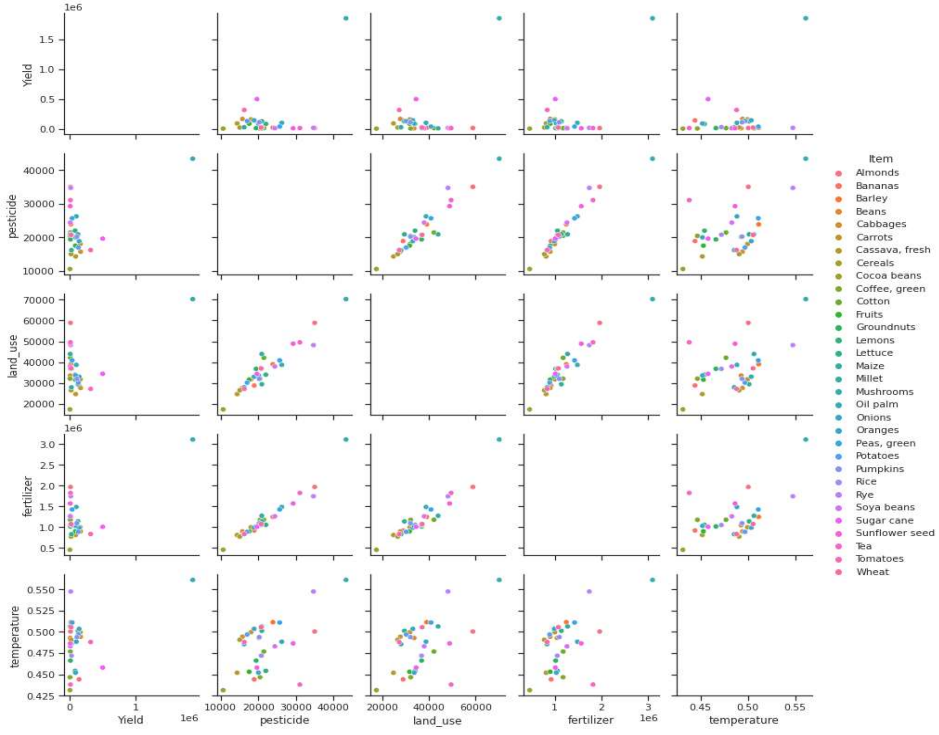

Figure 1: Pair plot to understand the relationship between the variables

The data contains features highly varying in magnitudes, units, and range, example yield of certain crops are very high when compared to others. Hence the features with high magnitudes will weigh in a lot more in the distance calculations than features with low magnitudes. To suppress this effect, we need to bring all features to the same level of magnitude. This was be achieved by scaling using MinMaxScaler.

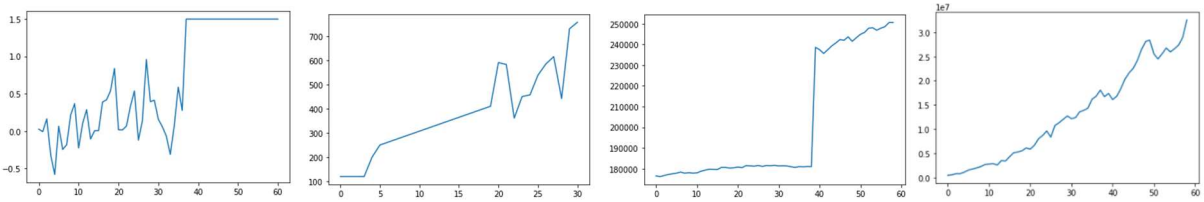

Figure 2: Shows the variation of temperature, pesticide use, land use and fertilizer use with respect to yield for a unique country and a crop across 60 years from 1961 to 2020

3.2. Data Modelling

Statistical Analysis and Yield Prediction

For improving upon existing models developed during phase one we implemented time series model to predict yield. First, we performed univariate time series using ARIMA and next we performed multivariate forecasting with LSTM using 6 different features and measured the root mean square error through a rolling based cross validation for both these models from year 2002 to 2020 and predicted the yield for 2021.

Remote Sensing Data and Yield Prediction

To address this task, we chose Illinois as it has highest production yield of Soybean, the state is divided into 102 counties [Figure 3]. The major portion of the work corresponded with acquiring relevant data in more quantity and further cleaning the data and preprocessing before the modelling.

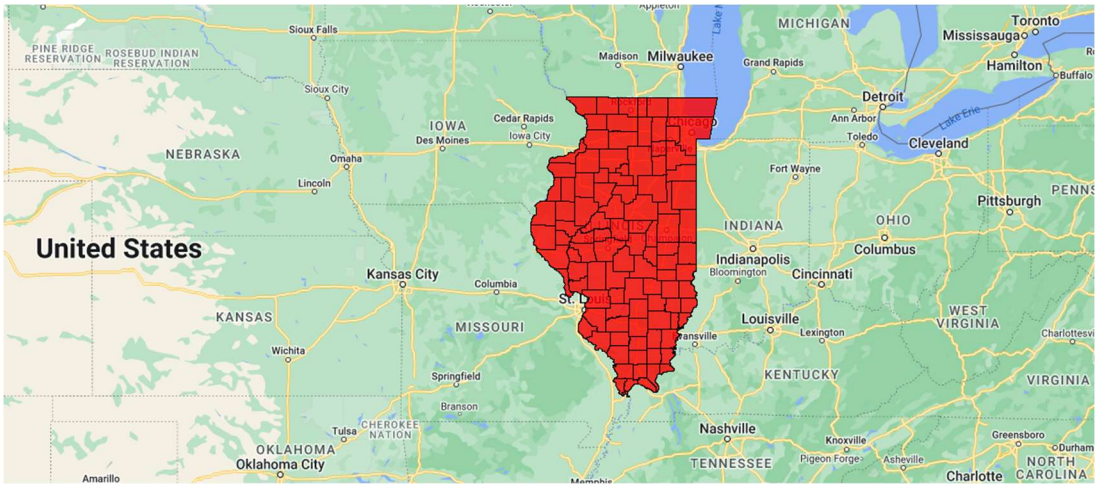

Figure 3: The area for study along with county borders (Illinois)

The data was collected in four stages, for this we used Google Earth Engines (GEE) library for python. We used MODIS by NASA because it offers improved radiometric calibration and superior spectral and spatial resolution. First surface spectral reflectance data was gathered because it reflects the state of crop growth. Next, we got the land surface temperature data followed by griddled estimates of daily weather data and finally land cover data to use as mask. In the first stage we defined a rectangular geometry area to concentrate on covering the area of interest. In GEE the data will be available as Image Collection objects, which has lot of meta data information

including the resolution, number of bands for each image in the collection. Also, the data is stored in raster format (Rasters are spatial data models that define space as an array of equally sized cells). In the second stage, instead of accessing individual images and exporting it to the working directory, GEE provides an option of batch export the image collection as tiff files to drive, where we can filter the dataset by date and sort it in ascending order of date. In the third stage, as we require the data to be concentrated county wise, we used the TIGER dataset which is a feature collection and offers shape files to get the geometry of individual counties of specific states and clipped the generated image collection to be restricted to the shape. In stage four as the problem was time series and the data needed to be relatable with previous year records, we used a mapping function algorithm to encode the temporal information in a cumulative way where, image at timestamp t+1 has the cumulative information from timestamp t. Also, the same mapping is used to reduce the image to selected bands as displayed in table 1.

Dataset	Image Collection ID	Resolution (m)	Yearly Data Frequency	Number of Bands	Units
Terra Surface Reflectance	MOD09A1	500m	46	7	nm
Terra Land Surface Temperature and Emissivity	MOD11A2	1000m	46	2	Kelvin
Daily Surface Weather and Climatological Summaries	DAYMET_V4	1000m	365	2	mm, Pa
Land Cover Type	MCD12Q1	500m	1	1	%

Table1: Data collected from Google Earth Engine details

As part of preprocessing the generated data after loading the tiff file into the working directory, we used rasterio [c] package to extract the raster data from tiff format. Next, we converted the data into NumPy array for easier matrix calculations. This array had dimension equal to

*((bands*number of images * number of years) * height * width)*

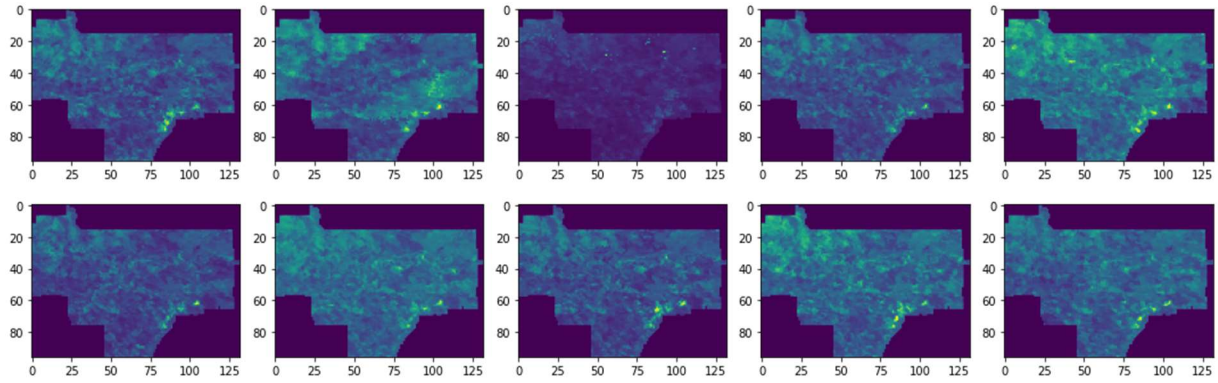

Figure 4: Surface reflectance image for a county from year 2002 to 2021

The array was processed to get the individual images across years as 3D images with the 3rd dimension as the spectral bands. Now to identify only the crop lands from the gathered images we used the land cover type data as a mask (The International Geosphere–Biosphere Programme (IGBP) defines ecosystems surface classifications) over the existing images. We also use the USDA data to get the corresponding yield values to the images generated. As, the DayMet V4 data is collected for each day in a year to concatenate with other dataset images we take the images in intervals of 8 days. Finally, the 3 different image arrays are combined to make a 3D array and are normalized. Final image dimension is scaled to 64*64 matrix with bands as 3rd dimension.

Deep Learning Model Architecture for Remote Sensing Data

We have built a CNN – LSTM architecture which consists of 2 convolutional neural networks and LSTM networks. CNN is used to extract the relevant features from the image and LSTM is used to handle the long-time lags between the images. The CNN has two convolutional layers in the is beginning with 32 filters and 64 filters respectively. These layers are followed by batch normalization and max-pooling layer. The output from this layer is inputted to LSTM with 256 neurons which are followed by a dense layer. The Relu activation function is used to induce non-linearity in the model. The model is run for 20 epochs with 32 as batch size with ADAM optimizer.

4. Results

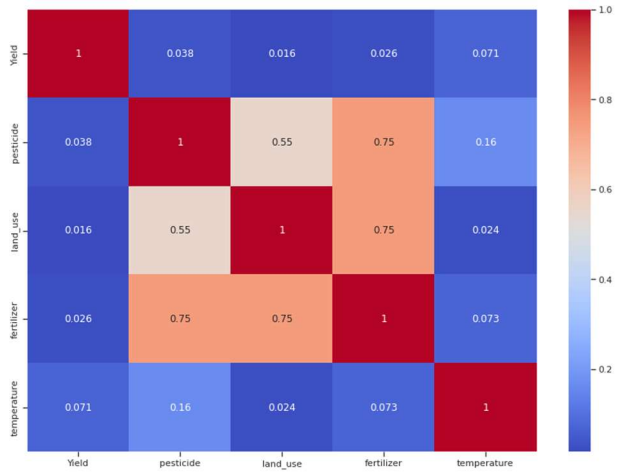

Figure 5: Heat Map showing the relationship between the statistical features generated

Metric	Random Forest		LightGBM		XGBoost		ARIMA	LSTM
	Train	Test	Train	Test	Train	Test		
RMSE	0.211	0.202	0.226	0.249	0.0006	0.051	0.0035	0.001565

Table 2: Shows the RMSE of the different machine learning models used for yield prediction

From Table 2 we can see that the LSTM model has outperformed the other regression models like Random Forest, LightGBM models, andXGBoost. Also, LSTM is performed with multivariate data while ARIMA is just trained on yield value. Likewise, the other models were trained by grouping data based on crops across countries. This might have also contributed to the little higher RMSE.

We have created an interactive world map and time series chart in tableau to visualize the crop yield across countries with drop downs to select the crop, country, and a slider to select the year. We had access to the data only until 2020. The values of yield for the year 2021 you see in the image below are predicted using Time Series Forecasting with LSTM.

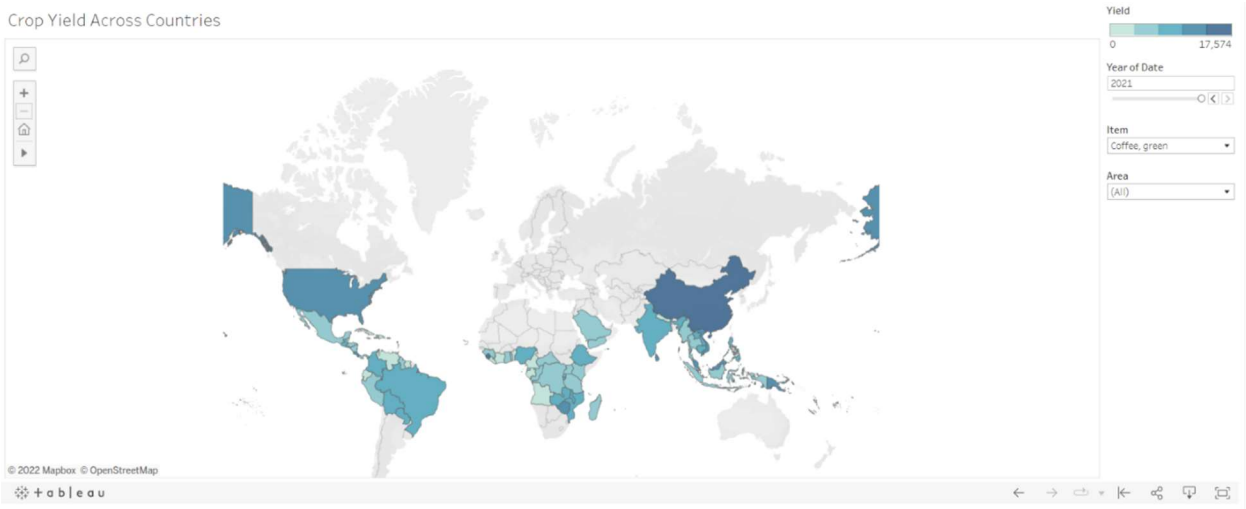

Figure 6: Snapped Image of the crop yield visualization across world

Metric\Year	2011	2012	2013	2014	2015	2016
RMSE	314.54	341.22	302.87	323.13	338.94	306.76

Table 3: Deep Learning Model RMSE Scores for remote sensing data

The Table 3 shows result from the proposed model architecture. This model has advantage of yield prediction in each year except for the year 2012 considering the massive variations in yield data, which can be attributed to the severe weather conditions that year.

5. Discussion

Looking at the model performances on the statistical data we can conclude that deep learning techniques like LSTMs that focus on sequential learning are well suited for performing future predictions data over other machine learning techniques. Also, multivariate regression offers much more better results as it encompasses information from other features corresponding to the yield

output. Since we have different rate of productions of different crops due to its usage across the worlds and while doing statistical analysis one must be cautious to make informed decisions.

The results from time series analysis using remote sensing data with CNN – LSTM based architecture shows even more better results. We have an added advantage over this model generated because the remote sensing images are available for even remote parts, shallow populated or regions with poor infrastructure of the world which may not have good documentation of agriculture data from the past. So, our model can be extended to include different regions from the world and further fine-tuned to the area of interest. As this information can be hosted on the internet anyone with access can utilize the deep learning model to study their land for future yield production and can give promising results on real-time data. These approaches can be used in other fields like weather forecasting, temperature forecasting, and so on.

The project goals have been met and the successful implementation of remote sensing model has expanded the future scope for improvements and feature additions. For the future work forecasting can be done on all the states in the corn or soybean belt instead of just single state. We can integrate stationary data that contains soil properties and feed it into the deep learning model parallelly to increase the performance of the model and overall prediction score [7].

6. Statement of contributions

Pareekshit Reddy Gaddam – Literature Survey and data pre-processing
Sree Krishna Suresh – Data sourcing, Machine Learning and Modelling
Vinitha Joyce Marathi – Exploratory Data Analysis and Website creation

All three group members worked on every aspect of the project interchangeably, above mentioned areas are specific to where a particular group member put higher efforts for the corresponding tasks mentioned.

7. References

[1] K. Akhand, M. Nizamuddin, L. Roytman, and F. Kogan, “Using RemoteSensing Satellite Data and Artificial Neural Network for predictionof Potato yield in Bangladesh,” in Remote Sensing and Modeling ofEcosystems for Sustainability XIII, ser. Proceedings of SPIE, Gao, Wand Chang, NB, Ed., vol. 9975. SPIE, 2016, Conference on RemoteSensing and Modeling of Ecosystems for Sustainability XIII, San Diego,CA, AUG 31, 2016.K. Akhand, M. Nizamuddin, L. Roytman, and F. Kogan, “Using RemoteSensing Satellite Data and Artificial Neural Network for predictionof Potato yield in Bangladesh,” in Remote Sensing and Modeling ofEcosystems for Sustainability XIII, ser. Proceedings of SPIE, Gao, Wand Chang, NB, Ed., vol. 9975. SPIE, 2016, Conference on RemoteSensing and Modeling of Ecosystems for Sustainability XIII, San Diego,CA, AUG 31, 2016.
[2] Khaki Saeed, Wang Lizhi, Crop Yield Prediction Using Deep Neural Networks, Frontiers in Plant Science, 10, 2019, https://www.frontiersin.org/articles/10.3389/fpls.2019.00621, 10.3389/fpls.2019.00621, 1664-462X
[3] Khaki Saeed, Wang Lizhi, Archontoulis Sotirios V., A CNN-RNN Framework for Crop Yield Prediction, Frontiers in Plant Science, 10, 2020, https://www.frontiersin.org/articles/10.3389/fpls.2019.01750, 10.3389/fpls.2019.01750, 1664-462X

[4] Ansarifar, J., Wang, L. & Archontoulis, S.V. An interaction regression model for crop yield prediction. *Sci Rep* 11, 17754 (2021). https://doi.org/10.1038/s41598-021-97221-7
[5] Shahhosseini, M., Hu, G., Huber, I. *et al.* Coupling machine learning and crop modeling improves crop yield prediction in the US Corn Belt. *Sci Rep* 11, 1606 (2021). https://doi.org/10.1038/s41598-020-80820-1
[6] R. J, V. K. G. Kalaiselvi, A. Sheela, D. S. D and J. G, "Crop Yield Prediction Using Machine Learning Algorithm," *2021 4th International Conference on Computing and Communications Technologies (ICCCT)*, 2021, pp. 611-616, doi: 10.1109/ICCCT53315.2021.9711853.
[7] Sun, Jie & Lai, Zulong & Di, Liping & Sun, Ziheng & Tao, Jianbin & Shen, Yonglin. (2020). Multilevel Deep Learning Network for County-Level Corn Yield Estimation in the U.S. Corn Belt. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. PP. 1-1. 10.1109/JSTARS.2020.3019046.
[8] Thomas van Klompenburg, Ayalew Kassahun, Cagatay Catal, Crop yield rediction using machine learning: A systematic literature review,Computers and Electronics in Agriculture,Volume 177, 2020,105709,ISSN 0168699,https://doi.org/10.1016/j.compag.2020.105709.

8. Appendix

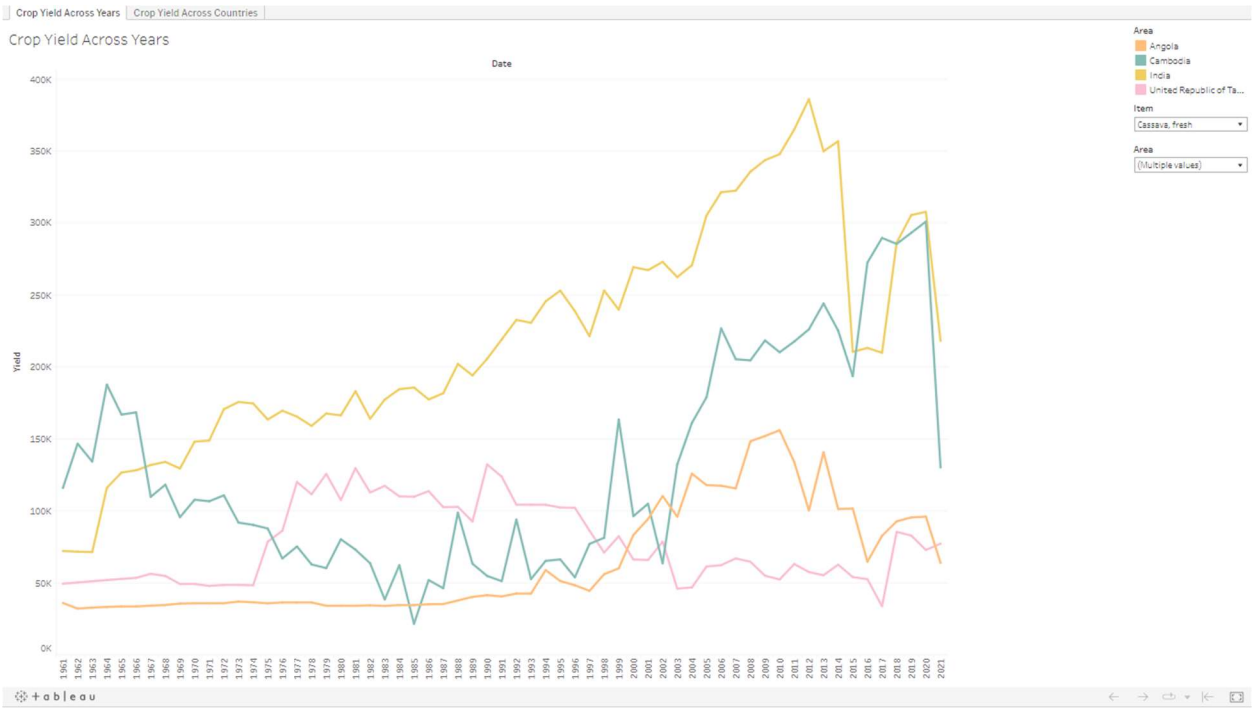

Figure 7: Yield across years with options to select countries and the crop

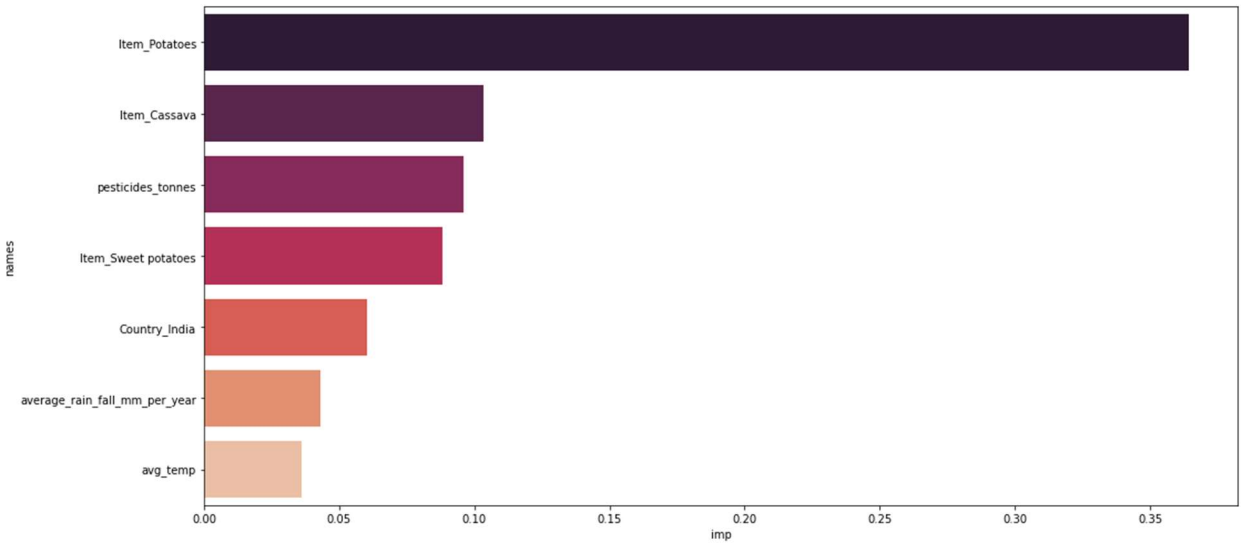

Figure 8: Feature importance generated using random forest