Predicting Crops Yield : A Machine Learning Approach

By: Hajir Almahdi Mohammed Abukhamadah

I. Definition

Project Overview

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.

Cuisine varies greatly around the world, the basic ingredients that sustain humans are pretty similar. We eat a lot of corn, wheat, rice and other simple crops. In this project the prediction of top 10 most consumed yields all over the world is established by applying machine learning techniques.

These corps include:

- Cassava
- Maize
- Plantains and others
- Potatoes
- Rice, paddy
- Sorghum
- Soybeans
- Sweet potatoes
- Wheat
- Yams

Problem Statement

In the project, machine learning methods are applied to predict crop yield using publicly available data from FAO and World Data Bank. The application of four regression algorithms and comparison of which will render the best results to achieve most accurate yield crops predictions.

Regression analysis is a form of predictive modelling technique which investigates the relationship between a **dependent** (target) and **independent variable (s)** (predictor).

Regression models used for this project:

- Gradient Boosting Regressor
- Random Forest Regressor
- SVM
- Decision Tree Regressor

Metrics

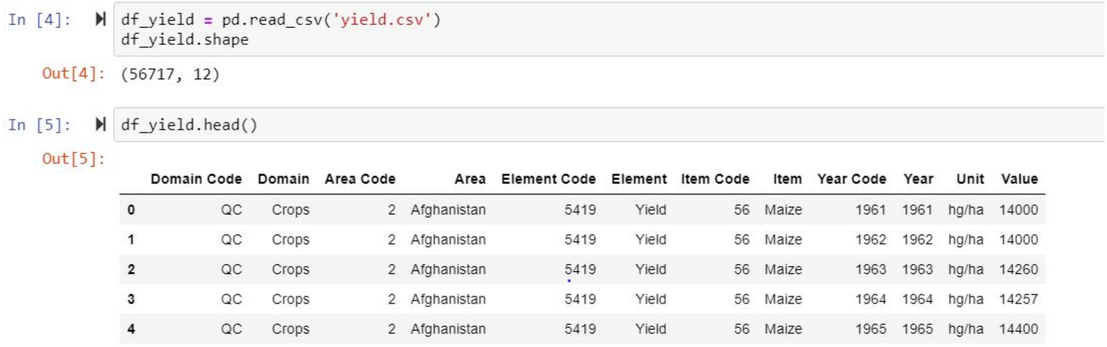
sklearn.metrics.r2_score function computes **Coefficient of determination R squared,** is the proportion of the variance in the dependent variable that is predictable from the independent variable, where it is a statistical measure between 0 and 1 which calculates how similar a regression line is to the data it's fitted to. If it's a 1, the model 100% predicts the data variance; if it's a 0, the model predicts none of the variance.

The evaluation metric will be the R squared score function, that will represent the proportion of the variance for items (crops) in the regression model.

II. Analysis

Gathering & Cleaning Data

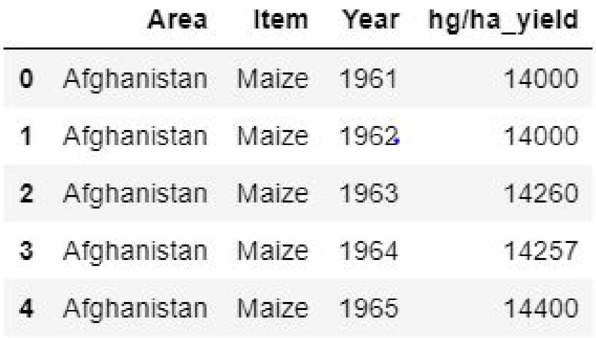
Data collection is the process of gathering and measuring information on variables of interest. FAOSTAT provides access to over 3 million time-series and cross sectional data relating to food and agriculture. The FAO data can be found in csv files. FAOSTAT contains data for 200 countries and more than 200 primary products and inputs in its core data set. It offers national and international statistics on food and agriculture. The first thing to get is the crops yield for each country.



```
In [4]:  df_yield = pd.read_csv('yield.csv')
         df_yield.shape

Out[4]:  (56717, 12)

In [5]:  df_yield.head()

Out[5]:
```

	Domain Code	Domain	Area Code	Area	Element Code	Element	Item Code	Item	Year Code	Year	Unit	Value
0	QC	Crops	2	Afghanistan	5419	Yield	56	Maize	1961	1961	hg/ha	14000
1	QC	Crops	2	Afghanistan	5419	Yield	56	Maize	1962	1962	hg/ha	14000
2	QC	Crops	2	Afghanistan	5419	Yield	56	Maize	1963	1963	hg/ha	14260
3	QC	Crops	2	Afghanistan	5419	Yield	56	Maize	1964	1964	hg/ha	14257
4	QC	Crops	2	Afghanistan	5419	Yield	56	Maize	1965	1965	hg/ha	14400

Now the data looks clean and organized, but dropping some of the columns such as Area Code, Domain, Item Code, etc, that won't be of any use to the analysis. Also, renaming **Value** to **hg/ha_yield** to make it easier to recognise that this is our crops yields production value. The end result is a four columns dataframe that contains: country, item, year and crops yield corresponds to them.

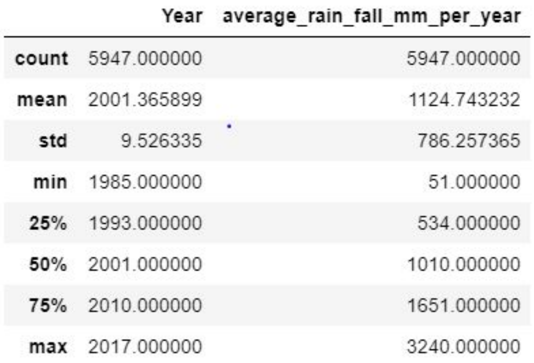


	Area	Item	Year	hg/ha_yield
0	Afghanistan	Maize	1961	14000
1	Afghanistan	Maize	1962	14000
2	Afghanistan	Maize	1963	14260
3	Afghanistan	Maize	1964	14257
4	Afghanistan	Maize	1965	14400

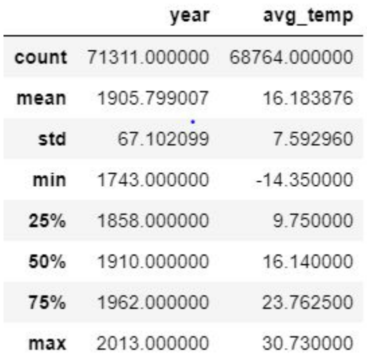
Using describe() function, few things come clear about the dataframe, where it starts at 1961 and ends at 2016, this is all the available data up to date from FAO.

Climatic factors include humidity, sunlight and factors involving the climate. Environmental factors refers to soil conditions. In this model two climate and one environmental factors are selected, rain and temperature and pesticides that influence plant growth and development.

Rain has a dramatic effect on agriculture. For this project rainfall per year information was gathered from the World Data Bank in addition to average temperature for each country.

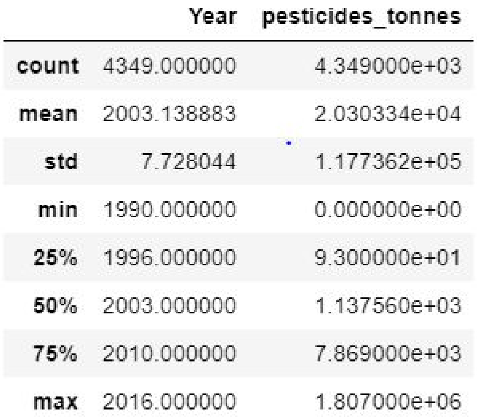


	Year	average_rain_fall_mm_per_year
count	5947.000000	5947.000000
mean	2001.365899	1124.743232
std	9.526335	786.257365
min	1985.000000	51.000000
25%	1993.000000	534.000000
50%	2001.000000	1010.000000
75%	2010.000000	1651.000000
max	2017.000000	3240.000000



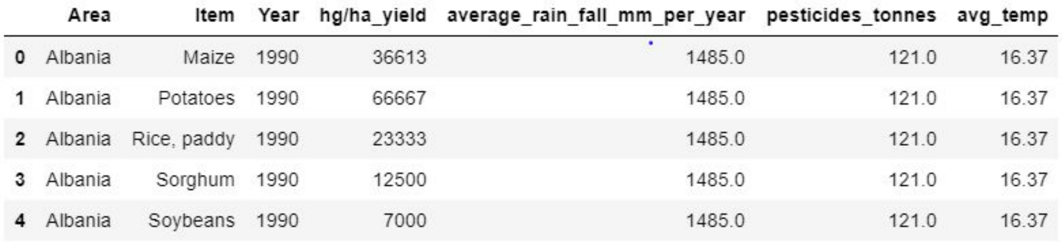
	year	avg_temp
count	71311.000000	68764.000000
mean	1905.799007	16.183876
std	67.102099	7.592960
min	1743.000000	-14.350000
25%	1858.000000	9.750000
50%	1910.000000	16.140000
75%	1962.000000	23.762500
max	2013.000000	30.730000

The final dataframe for average rainfall includes; country, year and average rainfall per year. The dataframe starts from 1985 to 2017, on other hand, the average temperature dataframe includes country, year and average recorded temperature. The temperature dataframe starts at 1743 and ends at 2013. The variation in years will compromise the collected data a bit where having to unite a year range to not include any null values.



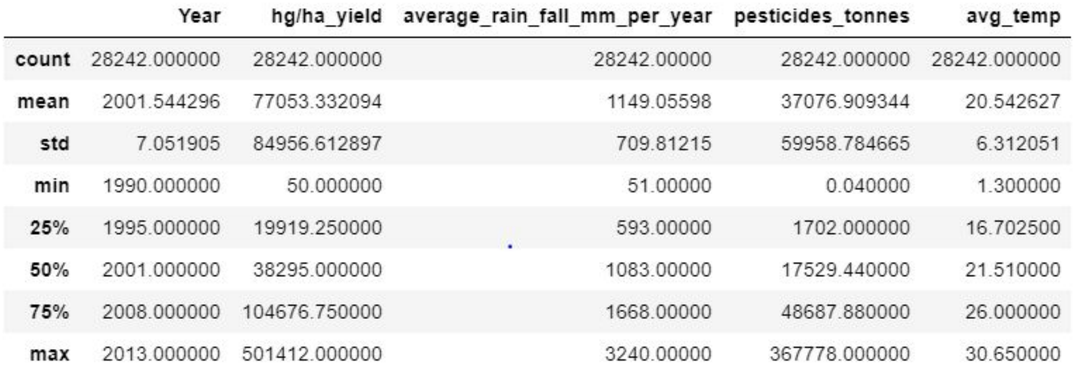
	Year	pesticides_tonnes
count	4349.000000	4.349000e+03
mean	2003.138883	2.030334e+04
std	7.728044	1.177362e+05
min	1990.000000	0.000000e+00
25%	1996.000000	9.300000e+01
50%	2003.000000	1.137560e+03
75%	2010.000000	7.869000e+03
max	2016.000000	1.807000e+06

Data for pesticides was collected from FAO, it’s noted that it starts in 1990 and ends in 2016. Merging these dataframes together, its expected that the year range will start from 1990 and ends in 2013, that is 23 years worth of data.



	Area	Item	Year	hg/ha_yield	average_rain_fall_mm_per_year	pesticides_tonnes	avg_temp
0	Albania	Maize	1990	36613	1485.0	121.0	16.37
1	Albania	Potatoes	1990	66667	1485.0	121.0	16.37
2	Albania	Rice, paddy	1990	23333	1485.0	121.0	16.37
3	Albania	Sorghum	1990	12500	1485.0	121.0	16.37
4	Albania	Soybeans	1990	7000	1485.0	121.0	16.37

Attached above the final dataframe with selected features for the application of model.

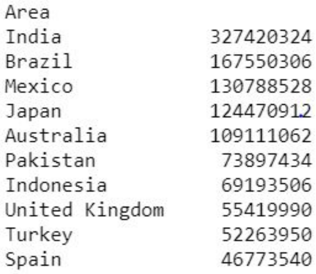


	Year	hg/ha_yield	average_rain_fall_mm_per_year	pesticides_tonnes	avg_temp
count	28242.000000	28242.000000	28242.00000	28242.000000	28242.000000
mean	2001.544296	77053.332094	1149.05598	37076.909344	20.542627
std	7.051905	84956.612897	709.81215	59958.784665	6.312051
min	1990.000000	50.000000	51.00000	0.040000	1.300000
25%	1995.000000	19919.250000	593.00000	1702.000000	16.702500
50%	2001.000000	38295.000000	1083.00000	17529.440000	21.510000
75%	2008.000000	104676.750000	1668.00000	48687.880000	26.000000
max	2013.000000	501412.000000	3240.00000	367778.000000	30.650000

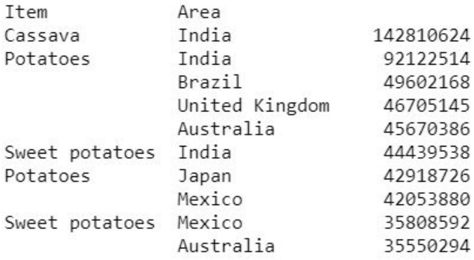
As expected, the dataframe starts in 1990 and ends in 2013. Making sure there are no empty entities, I can move to the next step. On a final note, the high variance in the values for each column, later on I'll account for that will scaling.

Data Exploration

First of all, knowing how many countries there are in the dataframe in addition to what are the highest crops yield countries are relevant to understand. There are 101 countries in the dataframe, with India having the highest crop yield.



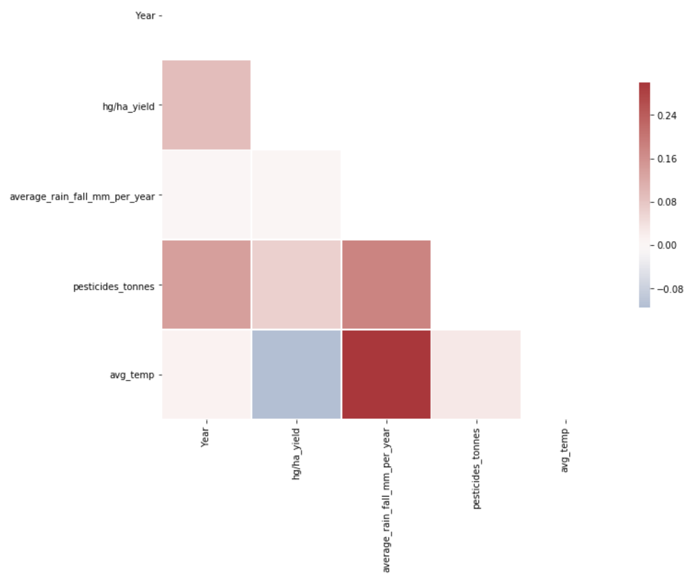
```
Area
India              327420324
Brazil             167550306
Mexico             130788528
Japan              124470912
Australia          109111062
Pakistan            73897434
Indonesia           69193506
United Kingdom      55419990
Turkey              52263950
Spain               46773540
```



```
Item            Area
Cassava         India             142810624
Potatoes        India              92122514
                Brazil             49602168
                United Kingdom     46705145
                Australia          45670386
Sweet potatoes  India              44439538
Potatoes        Japan              42918726
                Mexico             42053880
Sweet potatoes  Mexico             35808592
                Australia          35550294
```

Grouped by the item (crop), India is the highest for production of cassava and potatoes. Potatoes seem to be the dominant crop in the dataset, being the highest in 4 countries.
The final dataframe starts from 1990 and ends in 2013, that's 23 years worth of data for 101 countries.

Now, exploring the relationships between the columns of the dataframe, a good way to quickly check correlations among columns is by visualizing the correlation matrix as a heatmap.



It is evident from the heatmap above that all of the variables are independent from each, with no correlations.

Benchmark

Default SciKit-Learn Gradient Boosting Regressor and Random Forest Regressor will be used as benchmarks. Several models will then be explored to improve over the benchmark including Decision Tree Regressor and Support-Vector Machines (SVM).

III. Methodology

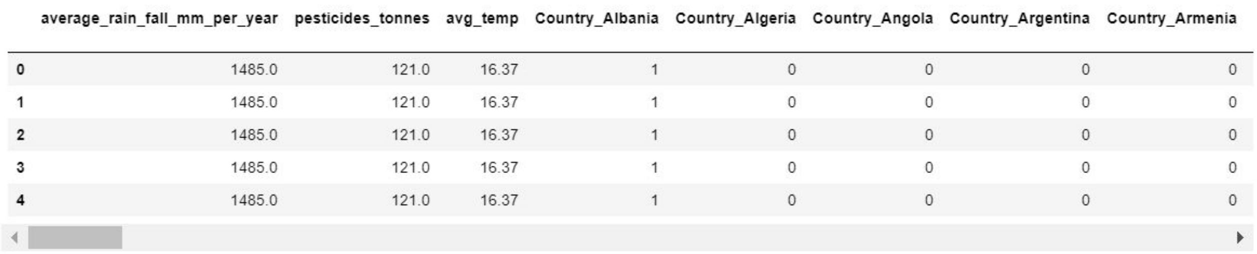
Data Preprocessing

Data Preprocessing is a technique that is used to convert the raw data into a clean data set. In other words, whenever the data is gathered from different sources it is collected in raw format which is not feasible for the analysis.

In the final dataframe there are two categorical columns in the dataframe, categorical data are variables that contain label values rather than numeric values. The number of possible values is often limited to a fixed set, like in this case, items and countries values. Many machine learning algorithms cannot operate on label data directly. They require all input variables and output variables to be numeric.

This means that categorical data must be converted to a numerical form. One hot encoding is a process by which categorical variables are converted into a form that could be provided to ML algorithms to do a better job in prediction. For that purpose, One-Hot Encoding will be used to convert these two columns to one-hot numeric array.

The categorical value represents the numerical value of the entry in the dataset. This encoding will create a binary column for each category and returns a matrix with the results.



	average_rain_fall_mm_per_year	pesticides_tonnes	avg_temp	Country_Albania	Country_Algeria	Country_Angola	Country_Argentina	Country_Armenia
0	1485.0	121.0	16.37	1	0	0	0	0
1	1485.0	121.0	16.37	1	0	0	0	0
2	1485.0	121.0	16.37	1	0	0	0	0
3	1485.0	121.0	16.37	1	0	0	0	0
4	1485.0	121.0	16.37	1	0	0	0	0

The features of the dataframe will look like the above with 115 columns. Taking a look at the dataset above, it contains features highly varying in magnitudes, units and range. 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 magnitudes. This can be achieved by scaling with MinMaxScaler.

The final step on data preprocessing is the training and testing data. The dataset will be split into two datasets, the training dataset and test dataset. The data usually tend to be split inequality because training the model usually requires as much data-points as possible.The common splits are 70/30 or 80/20 for train/test.

The training dataset is the initial dataset used to train ML algorithms to learn and produce right predictions. (70% of dataset is training dataset)

The test dataset, however, is used to assess how well ML algorithm is trained with the training dataset. You can’t simply reuse the training dataset in the testing stage because ML algorithm will already “know” the expected output, which defeats the purpose of testing the algorithm. (30% of dataset is testing dataset).

Model Comparison & Selection

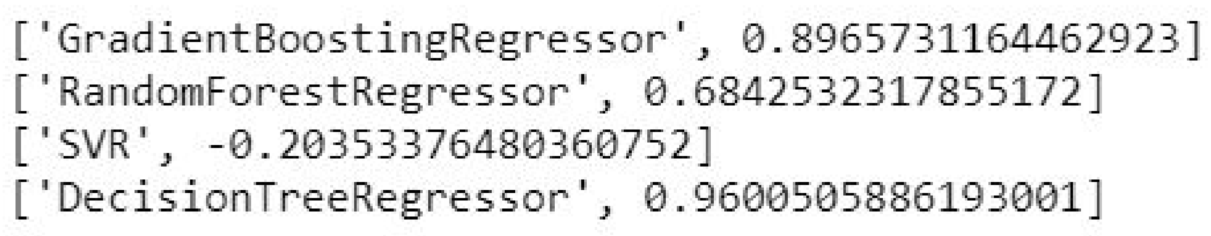
Before deciding on an algorithm to use, first we need to evaluate, compare and choose the best one that fits this specific dataset.

Usually, when working on a machine learning problem with a given dataset, we try different models and techniques to solve an optimization problem and fit the most suitable model, that will neither overfit nor underfit the model.

For this project, we'll compare between the following models through their Rooted Square Value:

- Gradient Boosting Regressor
- Random Forest Regressor
- SVM
- Decision Tree Regressor

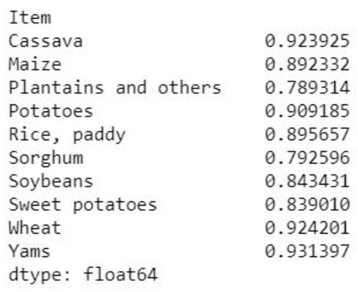
The evaluation metric is set based on **R^2 (coefficient of determination)** regression score function, that will represent the proportion of the variance for items (crops) in the regression model. **R^2** score shows how well terms (data points) fit a curve or line.



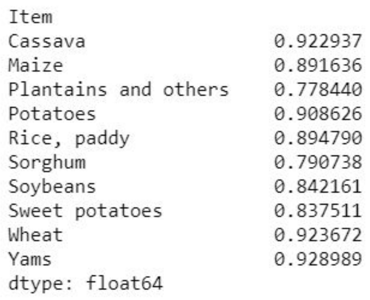
```
['GradientBoostingRegressor', 0.8965731164462923]
['RandomForestRegressor', 0.6842532317855172]
['SVR', -0.20353376480360752]
['DecisionTreeRegressor', 0.9600505886193001]
```

From results viewed above, **Decision Tree Regressor** has the highest R^2 score 0f **96%**, **GradientBoostingRegressor** comes second.

I'll also calculate **Adjusted R^2** , where it also indicates how well terms fit a curve or line, but adjusts for the number of terms in a model. If you add more and more useless variables to a model, adjusted r-squared will decrease. If you add more useful variables, adjusted r-squared will increase. Adjusted R2 will always be less than or equal to R2.



```
Item
Cassava                   0.923925
Maize                     0.892332
Plantains and others      0.789314
Potatoes                  0.909185
Rice, paddy               0.895657
Sorghum                   0.792596
Soybeans                  0.843431
Sweet potatoes            0.839010
Wheat                     0.924201
Yams                      0.931397
dtype: float64
```



```
Item
Cassava                   0.922937
Maize                     0.891636
Plantains and others      0.778440
Potatoes                  0.908626
Rice, paddy               0.894790
Sorghum                   0.790738
Soybeans                  0.842161
Sweet potatoes            0.837511
Wheat                     0.923672
Yams                      0.928989
dtype: float64
```

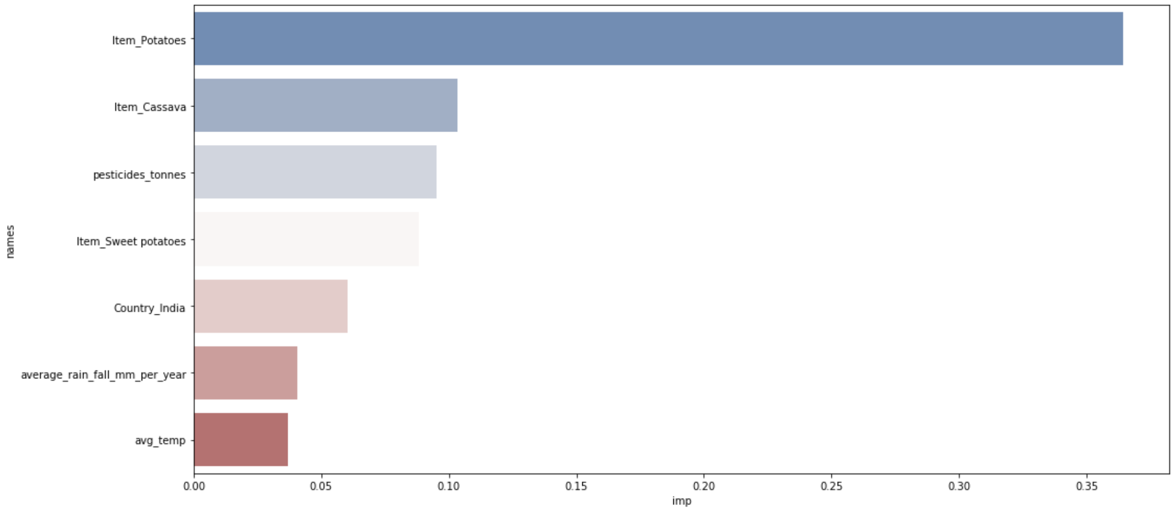
R^2 and adjusted R^2 results respectively for each item.

IV. Results

Model Results & Conclusions

The most common **interpretation** of **r-squared** is how well the regression model fits the observed data. For **example**, an **r-squared** of 60% reveals that 60% of the data fit the regression model. Generally, a higher **r-squared** indicates a better fit for the model. From the obtained results, it’s clear that the model fits the data to a very good measure of 96%.

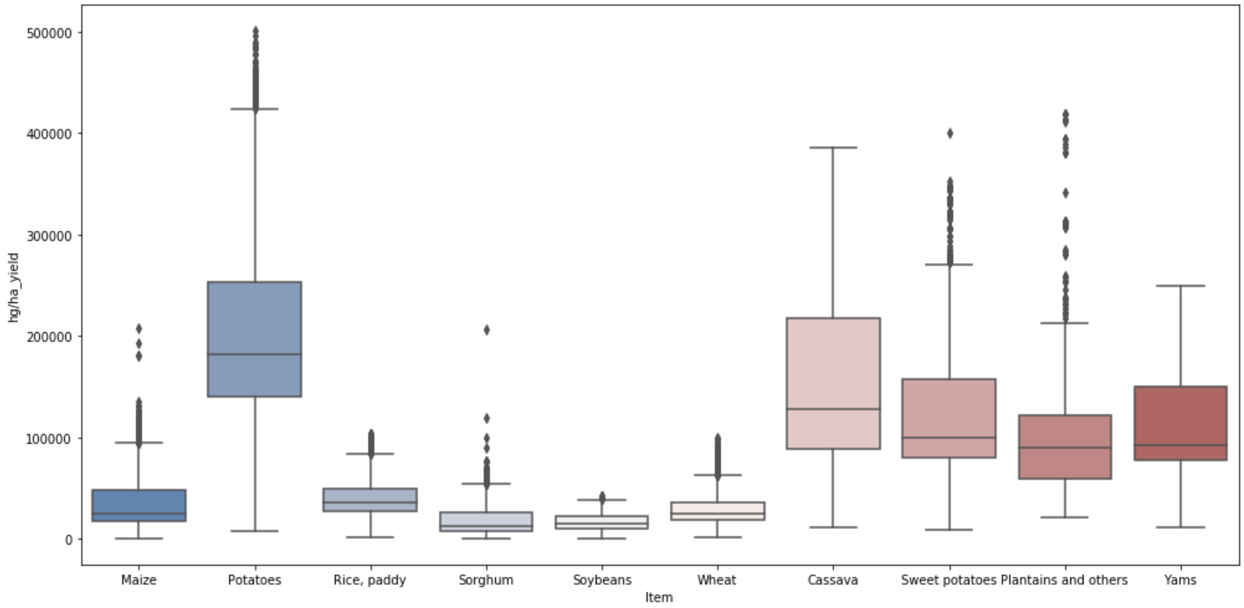
Feature importance is calculated as the decrease in node impurity weighted by the probability of reaching that node. The node probability can be calculated by the number of samples that reach the node, divided by the total number of samples. The higher the value the more **important** the **feature**. Getting the 7 top features importance for the model:



The crop being potatoes has the highest importance in the decision making for the model, where it's the highest crops in the dataset. Cassava too, then as expected we see the effect of pesticides, where its the third most important feature, and then if the crop is sweet potatoes, we see some of the highest crops in features importance in dataset.

If the crop is grown in India, it makes sense since India has the largest crops sum in the dataset. Then comes rainfall and temperature. The first assumption about these features were correct, where they all significantly impact the expected crops yield in the model.

The boxplot below shows the yield for each item. Potatoes are the highest, Cassava, sweet potatoes and Yams.



V. Conclusion

The file *model_depth_5*.pdf attached contains the drawing of the decision tree upside down with root at the top, in this case the root is item potatoes as it’s the top feature. The feature importance is clear and relations can be viewed easily.

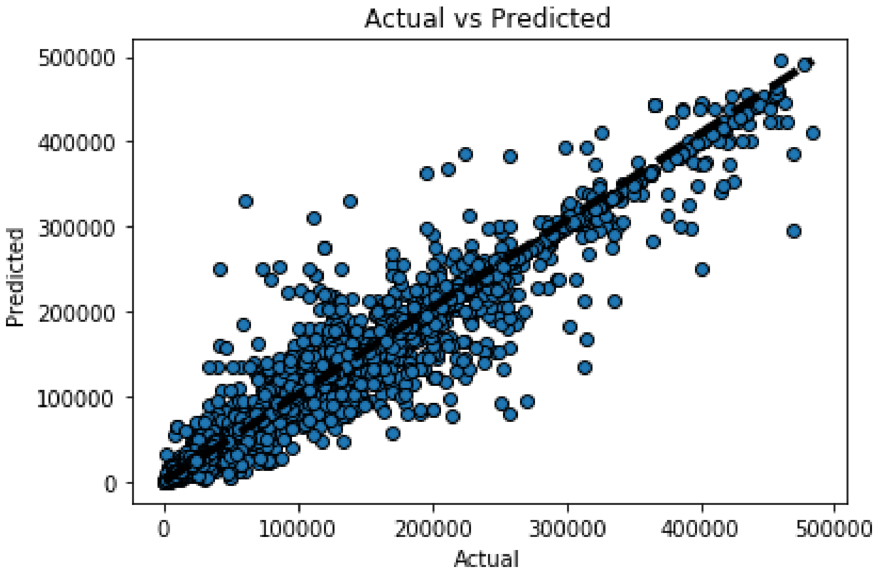
Decision Tree algorithm has become one of the most used machine learning algorithms both in competitions like Kaggle as well as in business environments. Decision Tree can be used both in classification and regression problems.

A decision tree typically starts with a single node, which branches into possible outcomes. Each of those outcomes leads to additional nodes, which branch off into other possibilities. A decision node, represented by a square, shows a decision to be made, and an end node shows the final outcome of a decision path. A node represents a single input variable (X) and a split point on that variable, assuming the variable is numeric. The leaf nodes (also called terminal nodes) of the tree contain an output variable (y) which is used to make a prediction.

The decision tree is arriving at an estimate by asking a series of questions to the data, each question narrowing our possible values until the model gets confident enough to make a single prediction. The order of the question as well as their content are being determined by the model. In addition, the questions asked are all in a True/False form. Decision trees regression uses mean squared error (MSE) to decide to split a node in two or more sub-nodes.

The root node is item potato, where its most important feature in the model. The model asks if it's potato then based on that it follows the branch if it's true or false. the algorithm first will pick a value, and split the data into two subset. For each subset, it will calculate the MSE separately. The tree chooses the value with results in the smallest MSE value up until it reaches a leaf node.

Since encoding the categorical items, the answer is either 0 or 1, it's either yes or no. Then the two internal nodes at the depth of one, if the true branch is followed, "*Is the item cassava*?<= 0.5". The other node, will ask "*pesticides_tonnes* <= *0.005*", following the decision tree to a deeper level and so on.



The figure above shows the goodness of the fit with the predictions visualized as a line. It can be seen that R Square score is excellent. This means that we have found a good fitting model to predict the crops yield value for a certain country. Adding more features, like climate data; wind and pollution, the economical situation of a given country and so on will probably enhance the model's predictions.

References

https://towardsdatascience.com/the-mathematics-of-decision-trees-random-forest-and-feature-importance-in-scikit-learn-and-spark-f2861df67e3
https://gdcoder.com/decision-tree-regressor-explained-in-depth/
http://www.fao.org/home/en/
https://data.worldbank.org/
https://chrisalbon.com/machine_learning/trees_and_forests/decision_tree_regression/