Crops Waste Management by Crop Yield Prediction using ML

By:
Raghav Aggarwal
11710116
KM060B36

Github link- https://github.com/Raghav-16/Crop-Waste-Management

Project Overview

Training machines to learn and produce models for future predictions are widely used in the today's time. Agriculture plays a critical role in the global economy. With the continuing increase in the human population understanding worldwide crop yield to minimize the crop waste is central to addressing food security challenges and reducing the impacts of climate change.

Crop waste management is an important agricultural problem and it can be reduce by predicting the crop yield that's going to be produced. The Agricultural yield primarily depends on conditions such as rain, temperature, pesticides and the history of crop yield. It is an important thing for making decisions related to agricultural risk management and future predictions.

Food varies greatly around the globe, the main ingredients that sustain human life are pretty similar such as corn, wheat, rice and other simple crops. In this project the prediction of top most consumed yields all around the world is established by applying machine learning techniques.

Those crops are:

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

Introduction

In the project, machine learning methods are applied to predict crop yield using publicly available datasets which will be further useful for maintaining the crop waste. Four regression algorithms are used and comparison of which will render the best results to achieve most accurate yield crops predictions.
Regression analysis is one of the predictive modeling techniques which investigate the relationship between a **dependent** (target) and **independent variables** (predictor).

Libraries used are:

- **numpy**- used for multi-dimensional arrays and matrices and high-level mathematical functions.
- **pandas**- for making data frames and reading of the data files.
- **sklearn**- for importing the necessary regressor models.
- **seaborn**- for making of the heatmap and checking correlations.
- **matplotlib**- for plotting the graphs and charts.
- **pydotplus**- used with graphviz for making decision tree chart.

Datasets

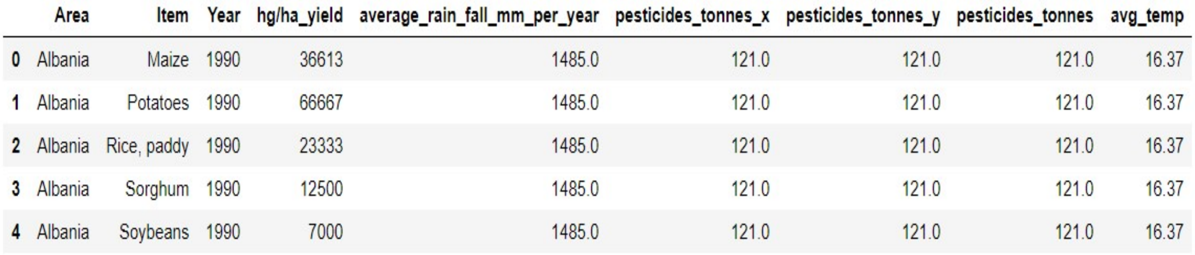
Data sets used are available from Food and Agriculture Organization (FAO) of the United States and World Data Bank.

- http://www.fao.org/home/en/
- https://data.worldbank.org/

The various data taken from these sources are saved in as .csv files, namely;

1. yield.csv for saving the amount of yield produced over the years.
2. rainfall.csv for the data related to precipitation.
3. temp.csv for temperature differences over the time duration.
4. pesticides.csv for the amount of pesticides being used.

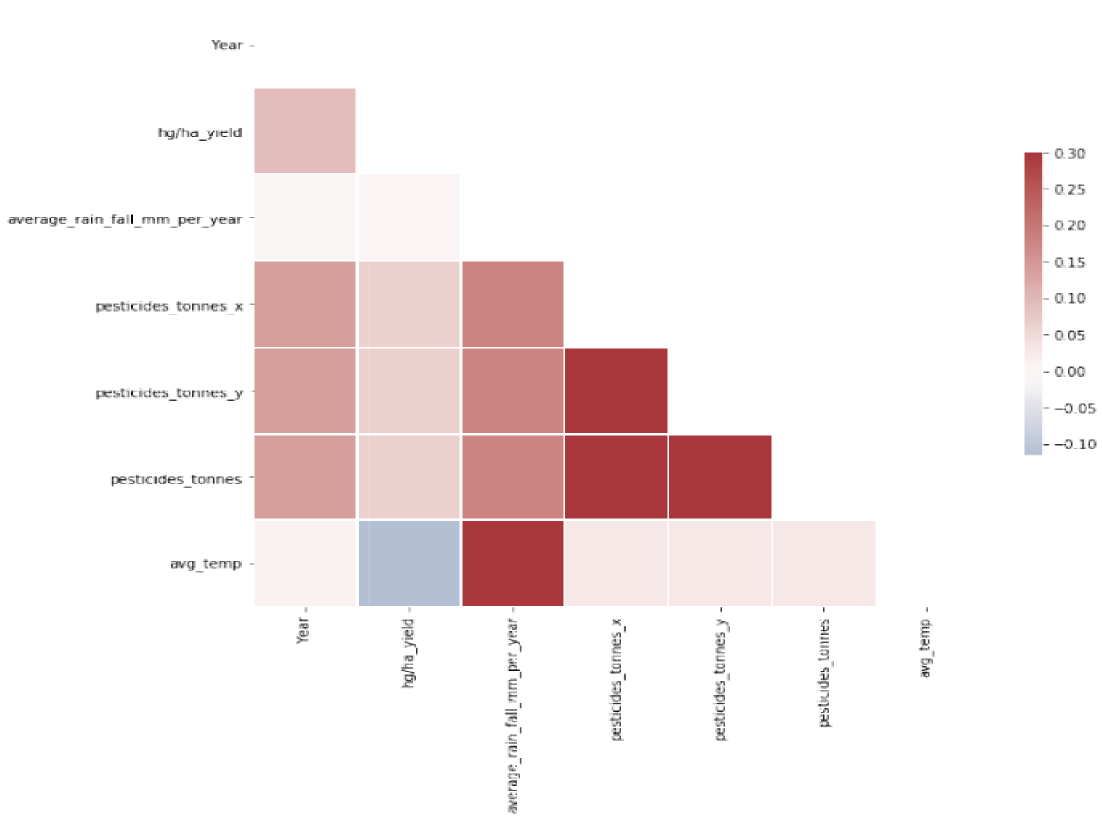
All of the data available are then combined to form one single file that is being saved as yield_df.csv.
The final yield_df.csv file looks like below.



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

Correlations

To check the correlations between the items in the dataset, the best way is to create the heat map. It is quite evident from the below heat map that all variables are independent and there is clearly no correlation between them.



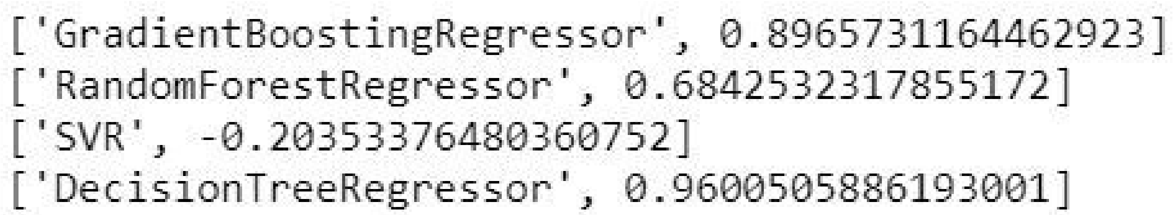
Model Comparison & Selection

Different models and techniques are used to solve the problem and find the most suitable model that will neither over fit nor under fit.

The following Regression models are used by comparing their **Rooted Square Value**:

- Gradient Boosting Regressor
- Random Forest Regressor
- Support Vector Machine
- Decision Tree Regressor

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



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

Results showed that the **Decision Tree Regressor** was the one with the highest R^2 score of **96%**. So, it was used as the Model for the further predictions.

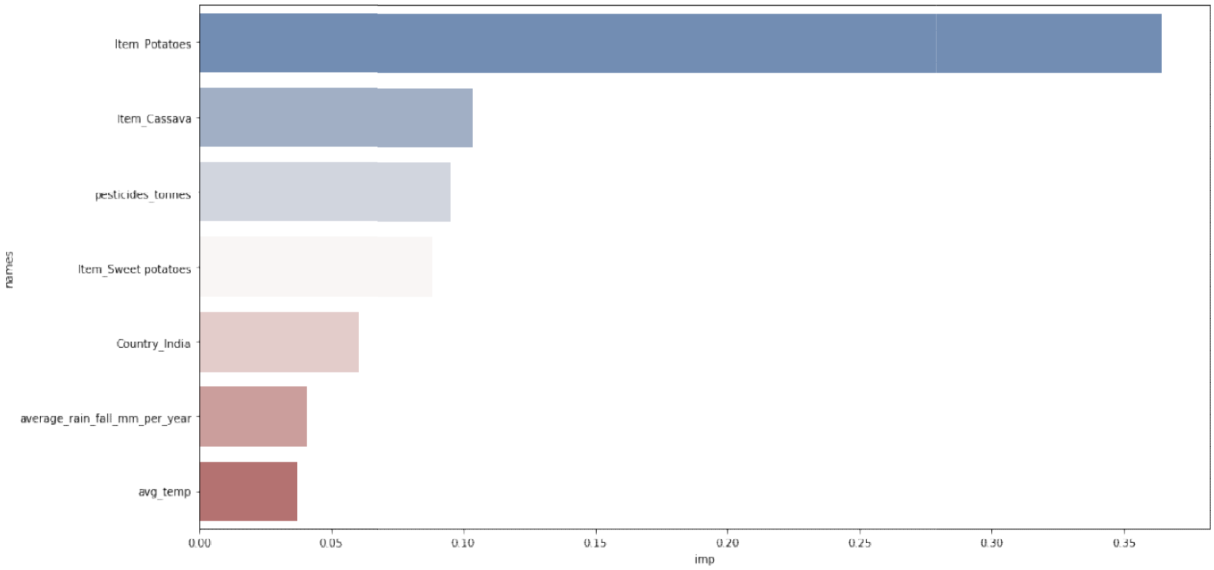
Decision Trees

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. 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 is 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.

Model Results

The most common **interpretation** of **r-squared** is how well the regression model fits the observed data. For example, an **r-squared** of 70% reveals that 70% of the data fit the regression model. Normally, a higher **r-squared** indicates a better fit for the model. From the obtained results, it's clear that the model is fitting 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:

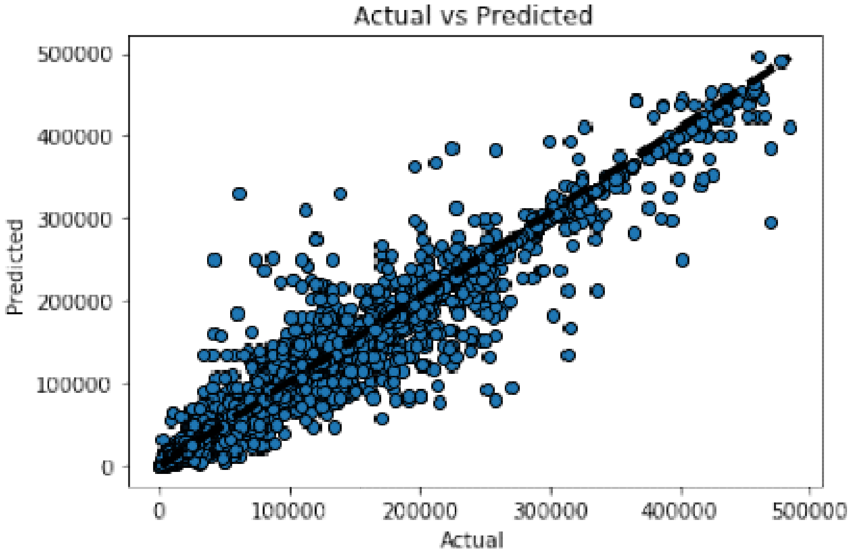


The crop being potatoes has the highest importance in the decision making for the model, where it's the highest crops in the dataset. Cassava too, then as expected we see the effect of pesticide, where it's the third most important feature, and then if the crop is sweet potatoes, we see some of the highest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

Conclusion

The file *model_depth_5*.pdf contains the image of the decision tree upside down with root at the top, in the case the root is potatoes as it's the top feature. The feature importance is quite clear and relationships are easily understandable.

Since encoding the categorical items, the answer must either 0 or 1, that is its 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 one 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.

Improvements

By adding more features and variables, like climatic conditions and its predicted data over the years; wind flow and its velocity over the region and the environmental pollution data, the economical situation of a given country and others will surely enhance the model's over all predictions.

References

1. https://github.com/hajir-almahdi/Machine-Learning-Capstone-Project
2. https://towardsdatascience.com/the-mathematics-of-decision-trees-random-forest-and-feature-importance-in-scikit-learn-and-spark-f2861df67e3
3. http://www.fao.org/home/en/
4. https://data.worldbank.org/
5. https://chrisalbon.com/machine_learning/trees_and_forests/decision_tree_regression/