Machine Learning Engineer Nanodegree

Capstone Proposal

Hajir Almahdi Mohammed Abukhamadah
Feb 24, 2020

# Predicting Crops Yield: A Machine Learning Approach

## Domain Background

Yield prediction is a very important issue in agricultural. Any farmer is interested in knowing how much yield he is about to expect. In the past, yield prediction was performed by considering farmer's experience on field and crop. The yield prediction is a major issue that remains to be solved based on available data.

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

## Problem Statement

Information can be converted into knowledge about historical patterns and future trends, the goal of this project is to predict crops yield from topmost consumed crops worldwide.

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.
The basic ingredients that sustain humans are 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. In this project I will **predict crops yield** worldwide for 10 most consumed crops. It is a regression problem.

## Datasets and Inputs

In this project, machine learning methods are applied to predict crop yield using publicly available data from FAO and World Data Bank.

These corps include:

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

Input Data fields:

The final dataframe will have: Item collected, country, yield, rain, pesticides and temperature values:

- crops yield of ten most consumed crops around the world was downloaded from FAO website. The collected data includes; country, item, year starting from 1961 to 2016 and yield value for these years.

- The climatic factors include rainfall and temperature. They're abiotic components, including pesticides and soil, of the environmental factors that influence plant growth and development.

- Rain has a dramatic effect on agriculture, for this project rain fall per year information was gathered from World Data Bank.

- Pesticides used for each item and country was also collected from FAO database.

- Average Temperature for each country was collected from World Bank Data.

## Solution Statement

The solution is **predictions of yield of crops** worldwide based on collected data. First, I will clean the collected data, merge the dataframes together based on common columns and explore relations between the different variables for correlation. Then I will perform normalization to establish a common scale for all the features, transform any categorical data such as country and crops name to numerical form.

For training I will split the data 70% training to 30% testing and apply different machine learning algorithms to compare which delivers best results.

The total size of the dataset is expected to be 200MB, the files that contain the data has been collected, and final datafame is expected to have: item (crop) country, year, yield value, average rainfall, pesticides and average temperature.

Models that will be used: For this project, we'll compare between the following models:

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

## Benchmark Model

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).

## Evaluation Metrics

**Function: sklearn.metrics.r2_score**

This function computes subset accuracy: the set of labels predicted for a sample must *exactly* match the corresponding set of labels in y_true. For this problem, a model that will preferably achieve an accuracy over 90%.

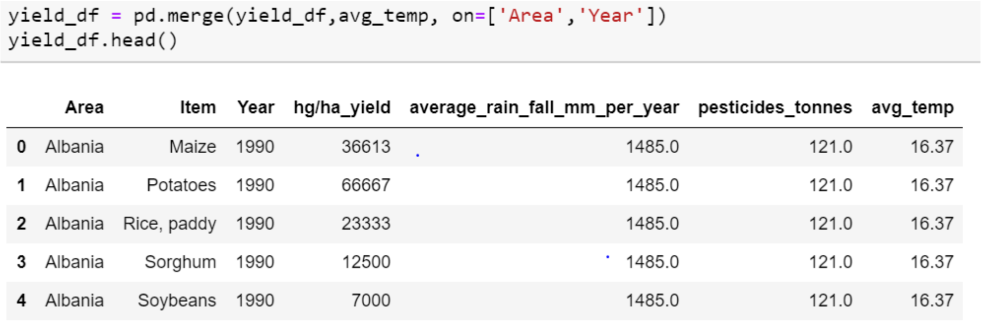
The evaluation metric will be the R^2 (coefficient of determination) regression score function, that will represents the proportion of the variance for items (crops) in the regression model.

## Project Design

After cleaning and exploring the relationship between the features, the final dataframe that contains all the features that will be used for the prediction process can be seen below in the screenshots:

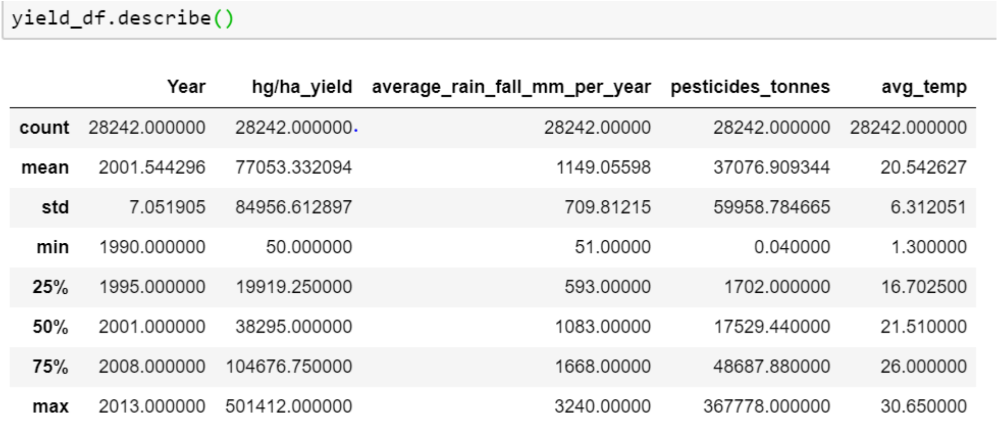
- Area: country of production.
- Item: type of crop.

- Year: year of production.
- Average_rain_fall_mm_per_year: Average amount of rain recorded that year.
- Hg/ha_yield: country's yearly production of the crop that year.
- Pesticides_tonnes: Amount of pesticides used on the crop that year.
- Avg_temp: Average temperature recorded for that year.



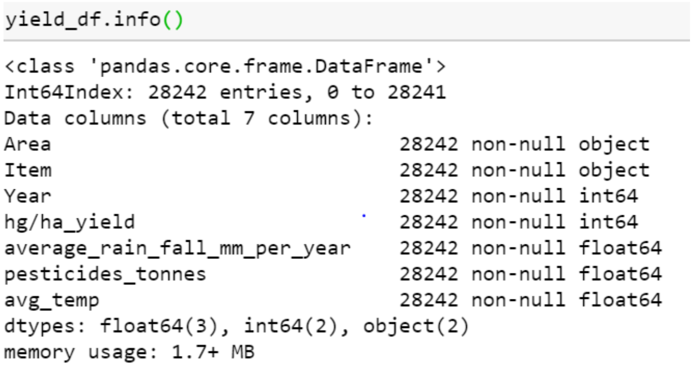
```
yield_df = pd.merge(yield_df,avg_temp, on=['Area','Year'])
yield_df.head()
```

| | 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 |



```
yield_df.describe()
```

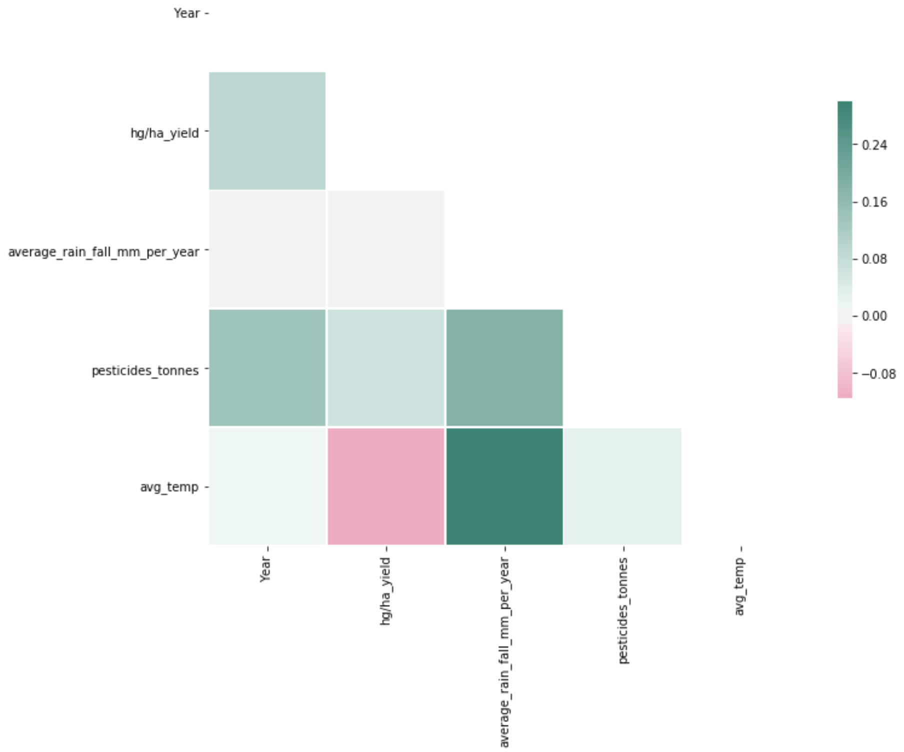
| | Year | hg/ha_yield | average_rain_fall_mm_per_year | pesticides_tonnes | avg_temp |
|---|---|---|---|---|---|
| **count** | 28242.000000 | 28242.000000 | 28242.00000 | 28242.000000 | 28242.000000 |
| **mean** | 2001.544296 | 77053.332094 | 1149.05598 | 37076.909344 | 20.542627 |
| **std** | 7.051905 | 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 |



```
yield_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 28242 entries, 0 to 28241
Data columns (total 7 columns):
Area                             28242 non-null object
Item                             28242 non-null object
Year                             28242 non-null int64
hg/ha_yield                      28242 non-null int64
average_rain_fall_mm_per_year    28242 non-null float64
pesticides_tonnes                28242 non-null float64
avg_temp                         28242 non-null float64
dtypes: float64(3), int64(2), object(2)
memory usage: 1.7+ MB
```

The final dataframe was obtained by joining four different dataframes, from FAO and World Data Bank to collect all needed features. Then after cleaning and transforming the data into standardized form, I've merged them together in the final dataframe yield_df.
To understand relationship between these parameters, I calculated correlation between all the features and illustrated them with diverging color heatmap



To train models, I plan to choose 4 different models to compare, because this is a regression problem. Relying on r2 score to decide which one performs best on dataset. With r2 score I will be able to evaluate the model for each item individually and visualize the final results for each crop (maze, potatoes, etc) yield predictions.

Libraries used in this project:
Pandas, Numpy, sklearn, Seaborn,matplotlib, OneHotEncoder, sklearn.preprocessing (MinMaxScaler), Models: RandomForestRegressor, GradientBoostingRegressor, svm, DecisionTreeRegressor.

## Reference

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