MATHEMATICAL SCIENCES FOR CLIMATE RESILIENCE INTERNSHIP PROGRAM

(MS4CR-IP)

(AIMS-IITA)

INTERNATIONAL INSTITUTE OF TROPICAL AGRICULTURE, Nairobi

INTERNSHIP REPORT

Assessment of Weather Forecast data in crops yield prediction: Case of ECMWF data

Written by :

Cedric NGAKOU SOTAKOUTSING

supervised:

EDUARDO GARCIA And GHOSH ANIRUDDHA

Contents

1 General Introduction

1.1 Background of the project

Climate change refers to changes beyond the average atmospheric's condition that are caused both by natural factors and anthropogenic factors. Therefore, the agricultural system is influenced by climatic seasonality and elements such as temperature, precipitation, solar radiation, etc. On other hand, it is also affected by hydrology, including groundwater and limited by availability of resources such as water and soil nutrients. The direct consequence of changes in these components is the reduction of crop yields, the nutritional quality of major crops and lowering livestock productivity. Substantial investments in adaptation will be required to maintain current yields and achieve production and food quality increases to meet demand. Therefore, it is becoming necessary, to find short- and long-term solutions to address this problem . the issue requires understanding relationships between agricultural management to compensate the lack of nutrient in the soil and also the behaviour of weather variables for a better adaptation.

Machine learning techniques are used in many disciplines to solve complex problems. In agriculture, crop yield forecast is among the most difficult issues in the field. This relates to the need for quality ground truth data, together with the information about the different biophysical factors affecting yield such as soil information, climate, etc. These information is now more accessible from various sources, but additionally, it is important to know the agronomic practices. To solve this major constrain, in this work we will use legacy agriculture data from experiments put together by the carob project which is an approach to reshape primary agricultural research data from experiments into a standard format ready for research and analysis. Together with historical weather data from **WorldClim** [14], CHIRPS [6], CHIRTS [2] and Weather Forecast data from The European Centre for Medium-Range Weather Forecasts (ECMWF). The main goal of this work is to develop a good methodology to assess weather forecast data in the crops yield forecast and sort out their impact in providing to the farmers useful information for climate change adaptation and crop management.

1.2 Specific Objectives

1. Assessment of the ground truth dataset (Quality control and pre-analysis)
2. Validate a baseline model (ML) for crop yield prediction using historical seasonal weather data
3. Assessment of agro-meteorological forecast product From ECMWF .

1.3 Structure of work

This work is divided mainly into Three sections: The first section is about general introduction describing the problem, the second section is focused on methodology where the different methods and

mathematical formulations are described as well as the materials used. In this section we describe the different datasets used, since the target variable is a numerical value, we select among the different regression models the best one using model selection method metrics and finally apply tuning parameters to choose the best hyper-parameters. The last section describes the results, analysis and a conclusion.

2 Methodology

2.1 Materials

2.1.1 Carob description

The aim of the Carob project is to create reproducible workflows to reshape primary agricultural data from experiments and survey into a standard format, aggregating individual datasets into larger collections that can be used in further research. Carob is a workflow developed to standardize these data, writing an R script for each individual dataset. After standardization, the datasets are combine based on the feature, to have a large base of information for the purpose of analysis.

Figure 1: Carob logo

2.1.2 Legacy experiment and survey Data

Legacy data is generally conceived as data from a previous experiments or data collection efforts. This can be identified and re-used. Typically, this information stored in bespoke formats and structures, making it difficult to access or process[1]. Where accessible, this data is valuable to derive new insights beyond those for which the data was initially collected [**bonilla2021**].

2.1.3 Climatic data Elements

A climatic data element is a measured parameter which helps to specify the climate of a specific location or region, such as precipitation, temperature, wind speed , humidity etc...[3]

Short-range is referred to forecasts for periods on the order of hours to days and up to two weeks ahead.

Sub-seasonal forecast is referred to as extended-range prediction, looking about two weeks to one or two month ahead. It can provide vital information for early warning of weather extremes.

seasonal forecasts refers to a long range prediction, cover a time period of 3 to 9 months.

Climate Forecast System (CFS) models the interactions between Earth's oceans, land, and atmosphere on a global scale.

The Global Forecast System (GFS) is a weather forecast model that generates data for dozens of atmospheric and land-soil variables, including temperatures, winds, precipitation, soil moisture, and atmospheric ozone concentration [7].

2.2 ECMWF Seasonal forecast system 5

ECMWF Forecast are calculates, the evolution of the atmosphere, ocean and land surface starting from an initial state based on observations of the Earth system. Due to limitations on information about the system, the initial state is not perfectly known [9]. This is due to the chaotic property of atmospheric systems. These models are very sensitive to small errors in the initial conditions which limits the ability to forecast daily weather variations beyond 10 to 15 days in the future. However, longer term predictions of the climate in the weeks, months, and years ahead are possible due to a number of known processes in the atmospheric system. The evolution of the atmosphere is slower and consequently, the information from their initial state for longer periods and its evolution can be predicted on long timescales.

Seasonal Forecast is a statistical summary of the daily weather calculated by the forecast model in the months ahead.

2.2.1 The ECMWF Seasonal forecast system

The system consists of an ocean analysis to estimate the initial state of the ocean, a global coupled ocean atmosphere general circulation model to calculate the evolution of the ocean and atmosphere, and a post-processing suite to create forecast products from the raw numerical output[11].

2.2.1.1 Models description

A-Ocean and sea ice model

SEAS5 uses the NEMO (Nucleus for European Modelling of the Ocean) ocean model with some modification regarding the model version, ocean physics and resolution. The ocean model used contains upgrades regarding aspects of ocean-surface wave interaction (Breivik et al. 2015) originally introduced at ECMWF. These aspects include a momentum flux estimated from the dissipation term; the surface boundary condition of the turbulent kinetic energy equation, which now account for the energy flux from breaking waves (Craig and Banner 1994); and the Coriolis-Stokes forcing term is introduced in the momentum equation.

B-Atmospheric model

The atmospheric component of SEAS5 is the ECMWF IFS (Integrated Forecast System). The atmospheric model contain horizontal and vertical resolution, the spectral horizontal resolution used for the main dynamic part of the model.

C-Coupling model

A gaussian method is used for interpolation in both directions, primarily due to the complexity of the 0.25 degree ORCA grid. The gaussian method automatically accounts for the inevitably different coastlines of the atmosphere and ocean models - values at land points are never used in the coupling, since these can be physically very different to conditions over water. The coupling interval is 1 hour, which allows resolution of the diurnal cycle.

2.3 Crop yield forecast using machine leaning model

machine learning can be an important tool for crop yield prediction and for decision making [13] specially on what crops to grow and what to do during the growing season of the crops. Thomas van Klompenburg and al. [13] have carried out a systematic literature review on crops yield prediction using machine leaning. The table below presents a summary of their results.

2.3.1 Short review on Machine Leaning model for crop yield prediction

Title	Algorithm used	reference	year
Data Mining with Neural Networks for Wheat Yield Prediction	Neural networks	reference	2008
Yield Prediction Using Artificial Neural Networks	Neural networks	reference	2011
Predictive ability of machine learning methods for massive crop yield prediction	M5-prime regression tree, k-nearest neighbor, support vector machine		2014
Application of supervised self-organizing models for wheat yield prediction	Neural networks		2014
Yield prediction for precision territorial management in maize using spectral data	Polynomial regression, logistic regression		2015
Maize yield forecasting by linear regression and artificial neural networks in Jilin, China	Neural networks, multiple linear regression		2015
Applying data mining techniques to predict annual yield of major crops and recommend planting different crops in different districts in Bangladesh	Linear regression, neural networks, clustering, k-nearest neighbor		2015

Title	Algorithm used	reference	year
Random forests for global and regional crop yield predictions	Random forest, linear regression		2016
Accurate prediction of sugarcane yield using a random forest algorithm	random forest		2016
Sugarcane Yield Grade Prediction Using Random Forest with Forward Feature Selection and Hyper-parameter Tuning	Random forest		2019
Artificial Neural Networks for Soil Quality and Crop Yield Prediction using Machine Learning	Neural networks		2019
yield prediction using sentinel-based optical and SAR data in Sahibganj district, Jharkhand (India)	Linear Regression		2019
Design of an integrated climatic assessment indicator (ICAI) for wheat production: A case study in Jiangsu Province, China	Random forest, support vector machine		2019

2.3.2 Syntheses of review

The systematic literature review carries out by homas van Klompenburg and al [13]. Shows that,

1. The main features used by most authors for crops yield prediction are:

Temperature, soil type ,Rainfall,Crop information,soil maps, Humidity,PH-value,solar-radiation,fertilizer,soil information .

2. Most used Machine leaning algorithms

   (a) Neural Networks

   (b) Linear regression

   (c) Random forest

   (d) Support vector machine

(e) Gradient boosting Tree

Random forest and Support vector machine are the most important Algorithm used [13].

3. Most used evaluation metric

Rank	Evaluation Metric	Key
1	Root Mean square Error	RMSE
2	R-Square	R^2
3	Mean Absolute Error	MAE
4	Mean square Error	MSE

2.3.3 Overview on mathematical formulation of Machine learning algorithm

2.3.3.1 Random Forest algorithm

Random forests are a combination of tree predictors in such a way that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. The generalization error for forests converges to a limit as the number of trees in the forest increases[4].

Mathematical formulation

Given training vectors $x_i \in R^n, i = 1, 2, ...l$ and a label vector y $\in R^n$, decision tree recursively partitions the feature space such that the samples with the same labels or similar target values are grouped together.

Let the data at node m be represented by $Q_m with h_n$ samples. For each candidate split $\theta = (j, t_m)$, where j is a feature and t_m the threshold, partition the data two subsets $Q_m^{left}(\theta)$ and $Q_m^{right}(\theta)$.

$$Q_m^{left}(\theta) = \{(x, y)|x_j \leqslant t_m\} \tag{1}$$

$$Q_m^{right}(\theta) = Q_m \setminus Q_m^{left}(\theta) \tag{2}$$

The quality of a candidate split of node m is computed using an imputing function or loss function depending on the problem that we have to solve (classification or regression).

$$G(Q_m, \theta) = \frac{n_n^{left}}{n_m} H(Q_m^{left}(\theta) + \frac{n_n^{right}}{n_m} H(Q_m^{right}(\theta)) \tag{3}$$

The next objective is to find the argument θ that will minimize the imputing function.

$$\theta^* = argmin_\theta G(Q_m, \theta) \tag{4}$$

Recurse for subsets $Q_m^{left}(\theta)$ and $Q_m^{right}(\theta)$ until the maximum allowable depth is reached $n_m < min_{samples}$ or $n_m = 1$

If we are dealing with a classification problem, Let's consider the target or outcome taking on 1,2....k-1 for node m. Let

$$P_{mk} = \frac{1}{n_m} \sum_{y \in Q_m} I(y = k) \tag{5}$$

a proportion of class k observations in node m. Common measure of impurity are the following:

$$H(Q_m) = \sum_k P_{mk}(1 - P_{mk}) \tag{6}$$

this is the loss function in classification and we can also get the Log of loss or Entropy as follows:

$$H(Q_m) = -\sum_k P_{mk} log(P_{mk}) \tag{7}$$

If we are dealing with Regression problem, Let's consider that the target is a continuous value, then for the node m, common criteria to minimize as for determining locations for future splits are Mean squared Error(MSE orL_2 error), Poisson deviance as well as Mean Absolute Error(MAE or L_1 error) [5].

MSE and poisson deviance both set the predicted value of terminal node to learned mean value y_m of the node whereas the MAE sets the predicted value of terminal nodes to median $(y)_m$ [5].

Mean Squared Error:

$$H(Q_m) = \frac{1}{n_m} \sum_k (y - \bar{y}_m)^2 \tag{8}$$

Where

$$\bar{y}_m = \frac{1}{n_m} \sum_{y \in Q_m} y \tag{9}$$

Half of Poisson deviance:

$$H(Q_m) = \frac{1}{n_m} \sum_{y \in Q_m} y log(\frac{y}{\bar{y}_m} - y + \bar{y}_m) \tag{10}$$

2.3.3.2 Gradient Tree Boosting algorithm

Gradient boosting is a method standing out for its prediction speed and accuracy, particularly with large and complex datasets.Gradient boost is very helpful in minimizing bias error of the model.

The main idea behind this algorithm is to build models sequentially and these subsequent models try to reduce the error of the previous model. This is done by building a new model on the error or residuals of the previous model [8].

Mathematics formulation

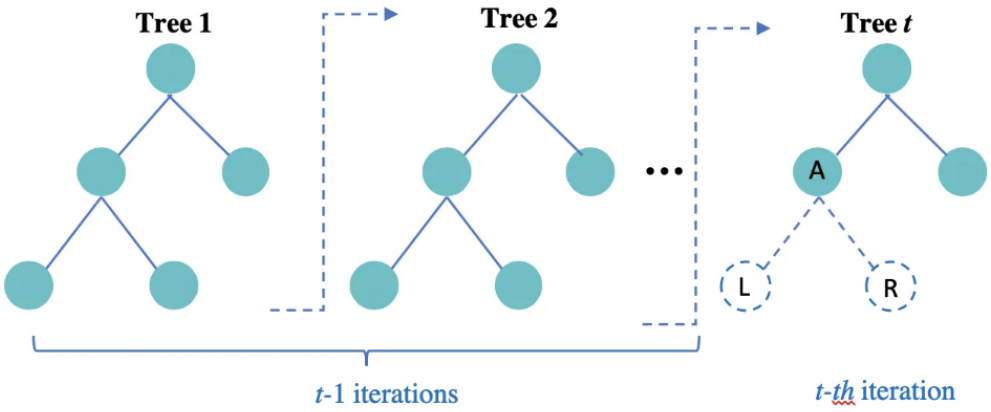

Figure 2: Algorithm of Gradient Tree Boosting

1.) build the base model to predict the observations in the training dataset

The idea is to find the predicted value for which Loss function is minimal.

$$F_0(x) = \underset{\gamma}{argmin} \sum_{i=1}^{n} L(y_i, \gamma) \tag{11}$$

L is Loss function

y_i is observation (target variable)

γ is the predicted value

$$L = \frac{1}{n} \sum_{i=0}^{n} (y_i - \gamma_i)^2 \tag{12}$$

$$\frac{dL}{d\gamma} = -\frac{2}{n} \sum_{i=0}^{n} (y_i - \gamma_i) \tag{13}$$

2.) Calculate the pseudo residuals

Since we are dealing with minimizing the previous error at each iteration, the pseudo residuals will be calculated using the following formula:

$$r_{im} = - \left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)} \right]_{F(x)=F_{m-1}(x)} \tag{14}$$

Where: $F(x_i)$ is the previous model (the output of the previous model) and **m** is a number of decision Tree

this equation will be writing in simple way like following:

$$r_{im} = - \left(Observevalue - predictedvalue \right) \tag{15}$$

3.) Build a model on these pseudo residuals and make prediction

we want to minimize these residuals and eventually improve the model accuracy and prediction. The idea is to use the residual as target to generate the new predictions (which will be the error value). let $H_m(x)$ our DT made on the residuals, we have

$$\gamma_m = \underset{\gamma}{argmin} \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) + \gamma \times H_m(x_i)) \tag{16}$$

Where: γ is the output value at the different leaves of the Tree and m the number of Trees.

2.) Update Model

Based on the logic above, the updated model is:

$$F_m(x) = F_{m-1}(x) + \nu_m H_m(x)) \tag{17}$$

Where: ν_m is the learning rate at tree m.

New prediction = previous prediction + learning rate* The tree made on residuals

2.3.3.3 Light Gradient Boosting Machine algorithm

LightGBM is a gradient-boosting framework based on decision trees to increase the efficiency of the model and reduces memory usage. It uses two new techniques:

1. Gradient-based One Side Sampling(GOSS)
2. Exclusive Feature Bundling (EFB)

It is designed to handle large-scale datasets and performs faster than other popular gradient-boosting frameworks like XGBoost.

Architecture of LightBGM

LightGBM splits the tree leaf-wise as opposed to other boosting algorithms that grow tree level-wise (Figure 3). It chooses the leaf with the maximum delta loss to grow. Since the leaf is fixed, the leaf-wise algorithm has a lower loss compared to the level-wise algorithm. Leaf-wise tree growth might increase the complexity of the model and may lead to overfitting in small datasets.

2.3.3.4 Feature importance

Feature importance is calculated as the decrease in node impurity weighted by the probability of reaching that node[12]. 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.

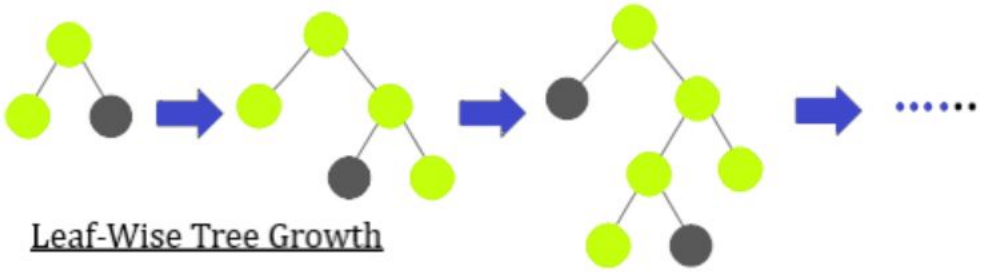

Figure 3: LGBM architecture

Let n_{ij} be the importance of node j regarding feature i,

$$n_{ij} = W_j C_j - W_{left(j)} C_{left(j)} - W_{right(j)} C_{right(j)} \tag{18}$$

W_j= weighted number of samples reaching node j

C_j = impurity value of node j

$W_{left(j)}$= Child node from left split on node j

$W_{right(j)}$=Child node from right split on node j

we can calculates the importance for each feature as follow:

$$FI_i = \frac{\sum_j n_{ij}}{\sum_{K \in all\ node} n_{iK}} \tag{19}$$

FI_i feature importance at node j split on feature i

2.4 Assess Weather Forecast product method

Here is the proposed methodology for assessment of weather forecast data from ECMWF. We first of all validate the baseline machine learning model using historical weather data from worclim,and then use Weather forecast data from ECMWF as a new dataset to test our model and compared the result.

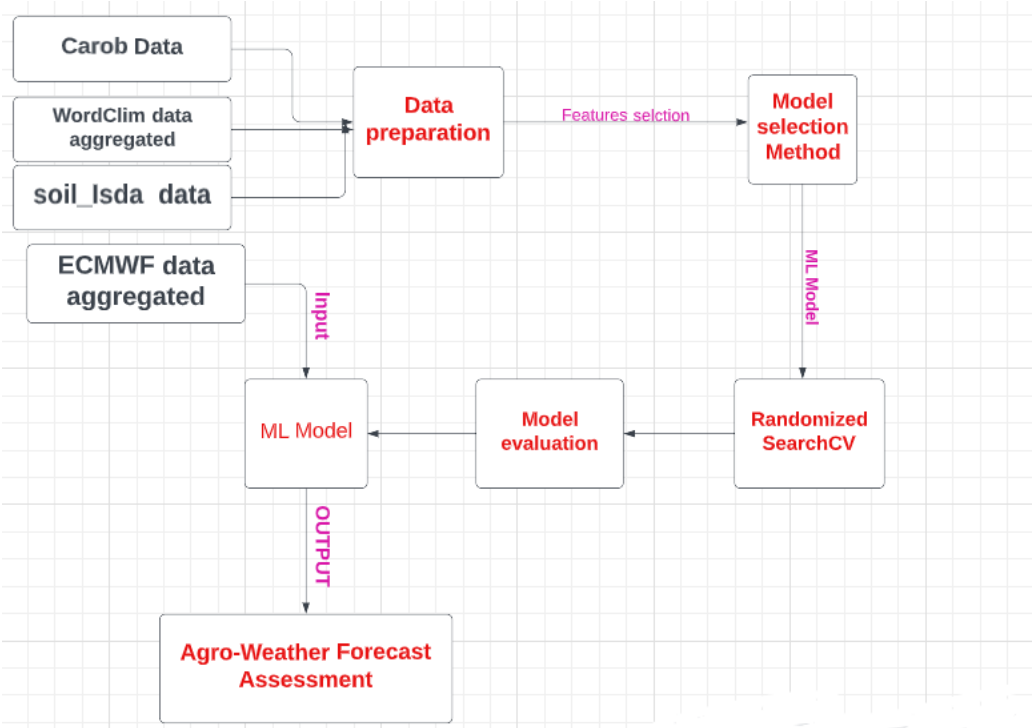

Figure 4: Assessment methodology

2.5 Data preparation and preliminary analysis

2.5.1 Data description

2.5.1.1 Carob data description

Carob is a workflow where we can access Agronomic Ground data sets aggregated into groups. From carob database we can access the Group of data Show in the table below:

Group	Number of Dataset	Records
Fertilizer	45	93144
Maize trials	7	6878
Wheat trials	14	49148
Variety trials	2	432
Lateblight	2	188
Crop cuts	4	5027

The dataset in carob contains many variables selected based on the critical importance in agronomy in general and crop productivity in particular.

2.5.1.2 WorldClim data description

WorldClim is a database of high spatial resolution global weather and climate data. From WordClim database we can Access:

1. Historical climate data from 1970 to 2000
2. Historical monthly weather data From 1960-2018
3. Future climate.The monthly value average over 20years periods(2021-2040,2041-2060,2061-2080,2081-2100) are available in the following spacial resolution: 10 minutes, 5 minutes, 2.5 minute, and 30 second.

Climate products available are **Minimum temperature, Maximum temperature and precipitation**

years	minimum temperature	maximum temperature	precipitation
1960-1969	tmin_1960-1969	tmax_1960-1969	prec_1960-1969
1970-1979	tmin_1970-1979	tmax_1970-1979	prec_1970-1979
1980-1989	tmin_1980-1989	tmax_1980-1989	prec_1980-1989
1990-1999	tmin_1990-1999	tmax_1990-1999	prec_1990-1999
2000-2009	tmin_2000-2009	tmax_2000-2009	prec_2000-2009
2010-2019	tmin_2010-2019	tmax_2010-2019	prec_2010-2019
2020-2021	tmin_2020-2021	tmax_2020-2021	prec_2020-2021

Figure 5: Example of Historical monthly climate product available

2.5.1.3 iSDA soil data description

iSDA soil is an open soil data for Africa where we can access a high spatial resolution soil information service, mapped at 30m spatial resolution for two standard depth intervals $(0-20)$ cm and $(20-50)$ cm.

1. Soil chemical properties such as Aluminium, carbon, Nitrogen, PH,etc...
2. Soil Physical properties and landscape such as (clay content, sand content, slope Angle, etc..)
3. Soil nutrients such as (calcium, Iron,Magnesium, Phosphors,Potassium sulfur,...)
4. Agronomy information such as (cropland,land cover,..)

2.5.1.4 CHIRPS data description

CHIRPS and CHIRTS were created in collaboration with scientists at the USGS Earth Resources Observation and Science (EROS) Center in order to deliver complete, reliable, up-to-date data sets for a number of early warning objectives. From these data, we can access:

1. rainfall from 1981 up to near today
2. maximum temperature from 1983 up to 2016

2.5.2 Quality control of the data

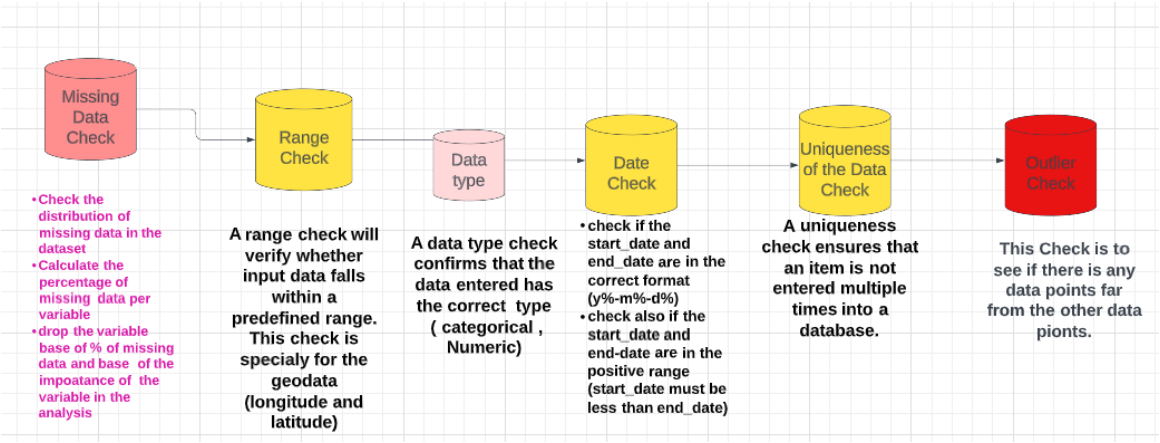

Figure 6: Step of Quality control

Missing data check

After checking the distribution of missing data using heat map visualization and the percentage of missing data per variable, we clean the data. The figure below gives an idea about the statistic.

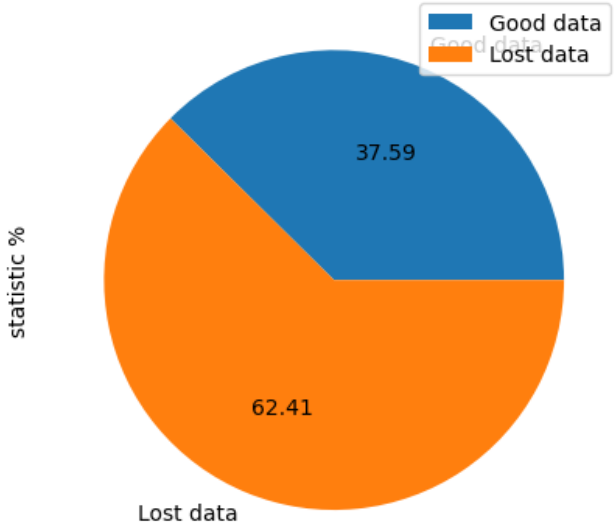

Figure 7: Statistic of carob data after cleaning

From the original dataset of 130204 Rows, 48943 Rows were considered acceptable to be use for this specific analysis. This means after cleaning the data we have about **37%** of the original dataset (The global description of the data will be given at the document annex).

longitude and latitude check

The goal here is to check if all the Geo-point are inside the range (longitude between 90 deg and - 90 deg,latitude between 180 deg and - 180 deg) and also have a global look on map. As you can see on the Figure 8 below some points of our data sets were out of the land and we had to fix the longitude

and latitude coordinates.

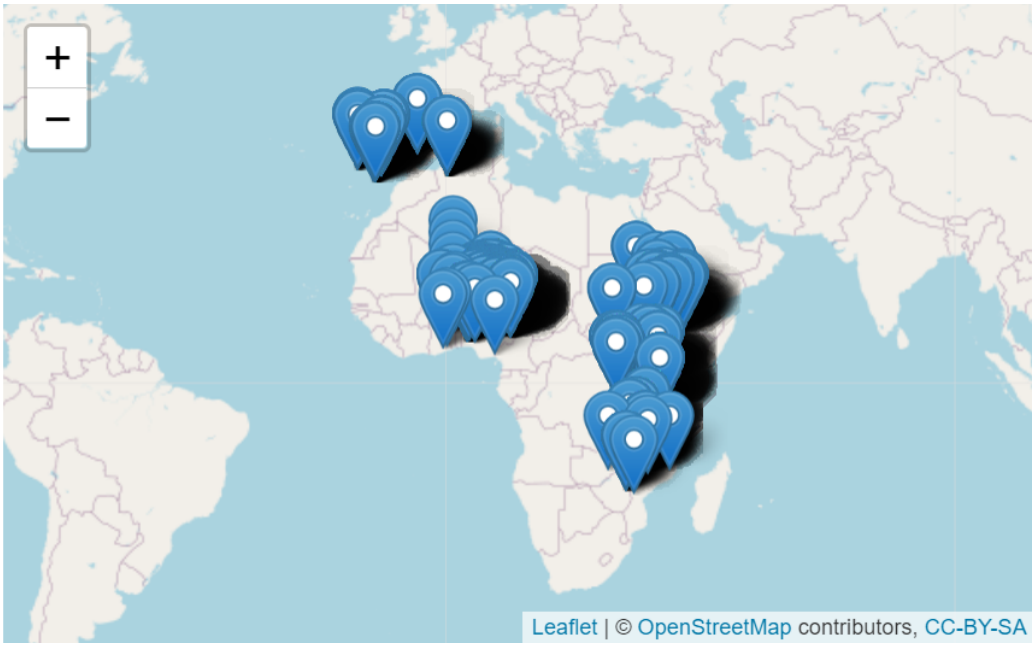

Figure 8: Map of long and lat coordinate

Outlier Check

We find the outliers with the IQR or STD. for the IQR the data point should be less than the inter quantille range from the quantiles and for STD the data point should be less than three times the standard deviation away from the mean.

After applying the above data checks, the final dataset is described below:

Total number of observations: **4820 rows**

Carob variables(ground dataset)	Weather product	Soil variables
N fertilizer	maximum temperature	Soil phosphor
P fertilizer	minimum temperature	soil potassium
K fertilizer	maximum rainfall	soil Zinc
Yield	Average rainfall	Soil sulfur
spacial temporal variables	Total rainfall	

2.5.3 Data exploration and Data visualization

In this part, we will be focusing on understanding the crop data. So we will have a look at each variable and try to understand their meaning and relevance to the problem. In the data set we have categorical variables and numerical variables.

1. Categorical variables

Histogram

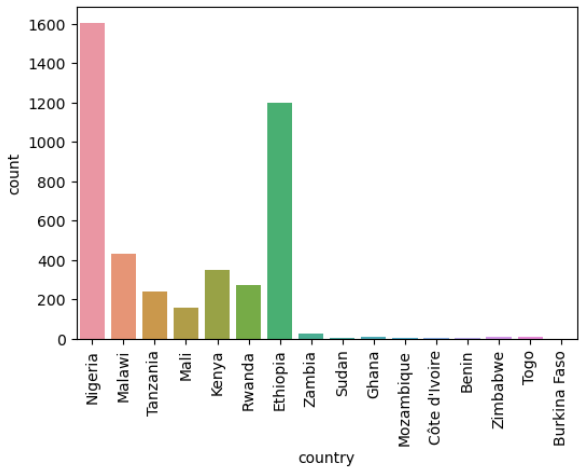

Figure 9: Country Histogram

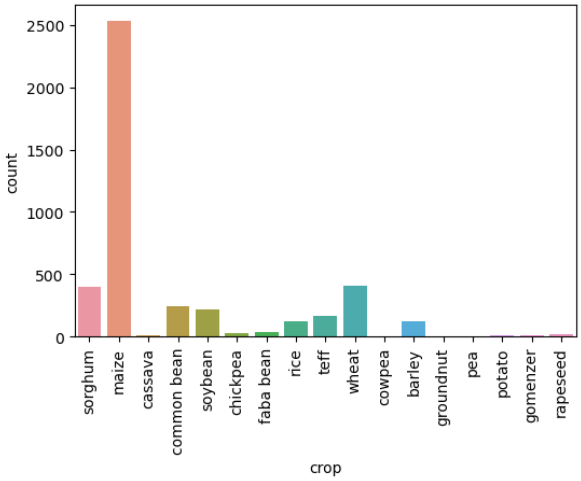

Figure 10: Crop Histogram

As Figure 9 shows, Nigeria and Ethiopia are the most represented countries, followed by Malawi, Kenya and Tanzania. In term of crops, maize has the highest number of records, followed by sorghum and wheat.

2. Numerical variables

Histogram

The aim of this part is to understand the distribution of data points in the dataset.

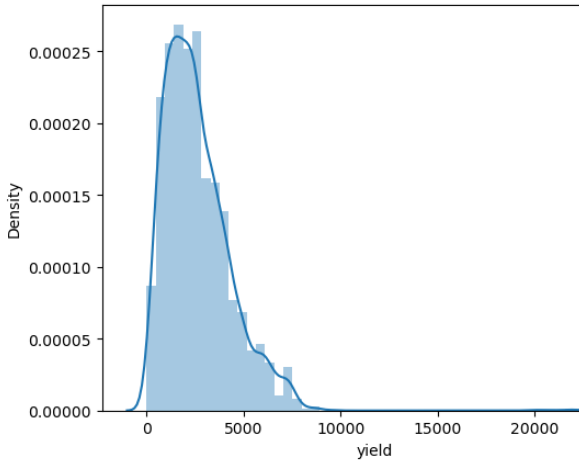

Figure 11: yield distribution

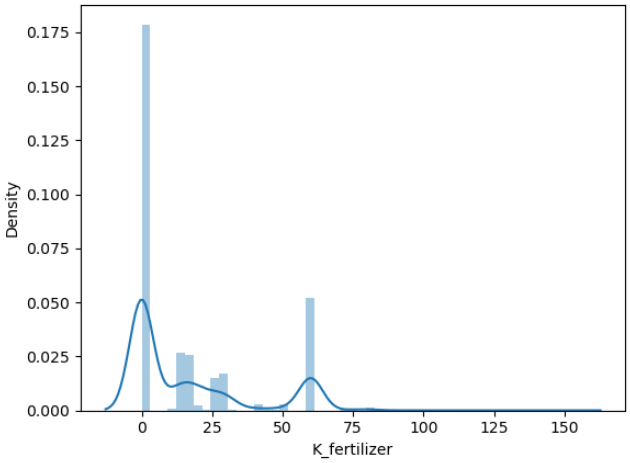

Figure 12: Kfertilizer distribution

yield skewness: 2.515223487745546 ; yield Kurtosis: 19.493078

Kfertilizer skewness:1.2254343259395635 ;K Kurtosis: 0.359684

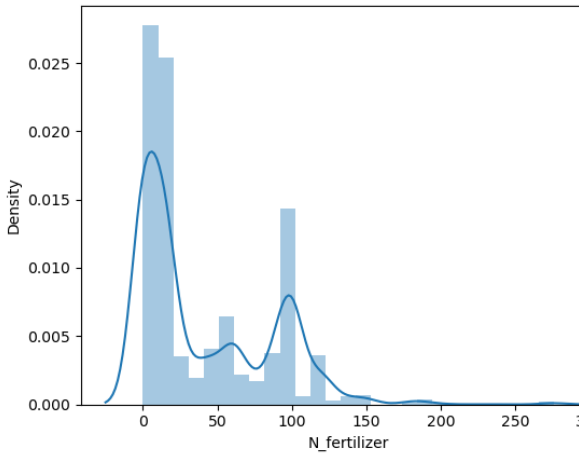

Figure 13: Nfertilier distribution

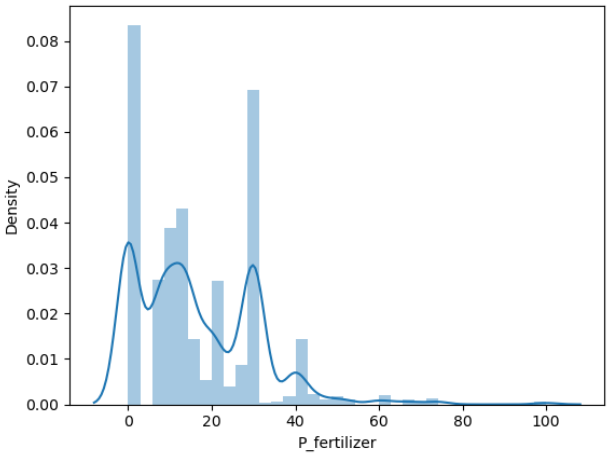

Figure 14: Pfertlizer distribution

Nfertilizer skewness: 1.0180608200049874; Pfertilizer skewness: 1.1163691376649636

N Kurtosis: 0.736094 ; P Kurtosis: 2.669068

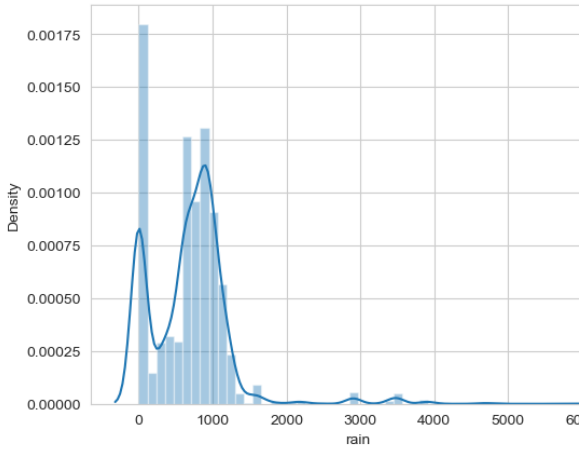

Figure 15: Total rainfall

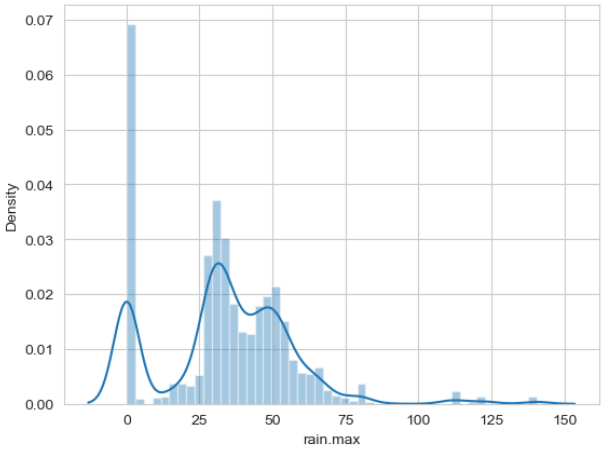

Figure 16: Maximum rainfall

Total rainfall skewness: 2.1661782697395857; Total rainfall Kurtosis: 11.227415

maximum Rainfall skewness: 0.60663313073341 ; maximum Rainfall Kurtosis: 2.047716

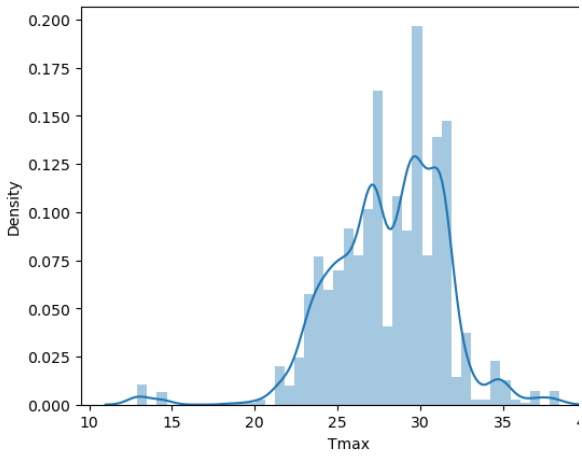

Figure 17: Maximum Temperature

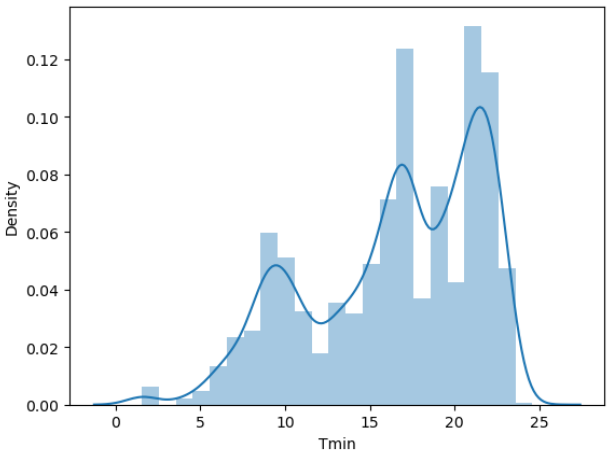

Figure 18: Minimum Temperature

maximum temperature skewness: -0.6932931 maximum temperature Kurtosis: 2.461563 minimum temperature skewness: -0.6288102 minimum temperature Kurtosis: -0.584133

Skewness is a measure of symmetry, or more precisely, the lack of symmetry. A distribution, or data set, is symmetric if it looks the same to the left and right of the center point.

Kurtosis is a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution. That is, data sets with high kurtosis tend to have heavy tails, or outliers. Data sets with low kurtosis tend to have light tails, or lack of outliers.

According to Hair et al. (2010) and Bryne (2010) [10] the data is considered to be normal if skewness is between -2 to 2 and kurtosis is between -7 to 7.

So base on that, yield and total rainfall data points are not well distributed. In the case of yield this can be related to the fact that most of the data include nutrient omission trials (NOTs), with no fertilizer applied, thus resulting in low yields and skewing the data. For the others variables, there is symmetry but a simple data transformation can solve the problem.

3.Relationship between the variables

The aim of this part is to have a good understanding on how the variables (input variables and output variable) are correlated. We will check the correlation in between input variables using correlation matrix, between output variable and each input variable.

Figure 19 show the correlation between the variables. From there we can notice that:

1. Minimum temperature and maximum temperature are highly correlated.
2. N and K fertilizer are most correlated with yield than others variables.
3. Weather variables (minimum and maximum temperature, rainfall) have a very low correlation with yield.

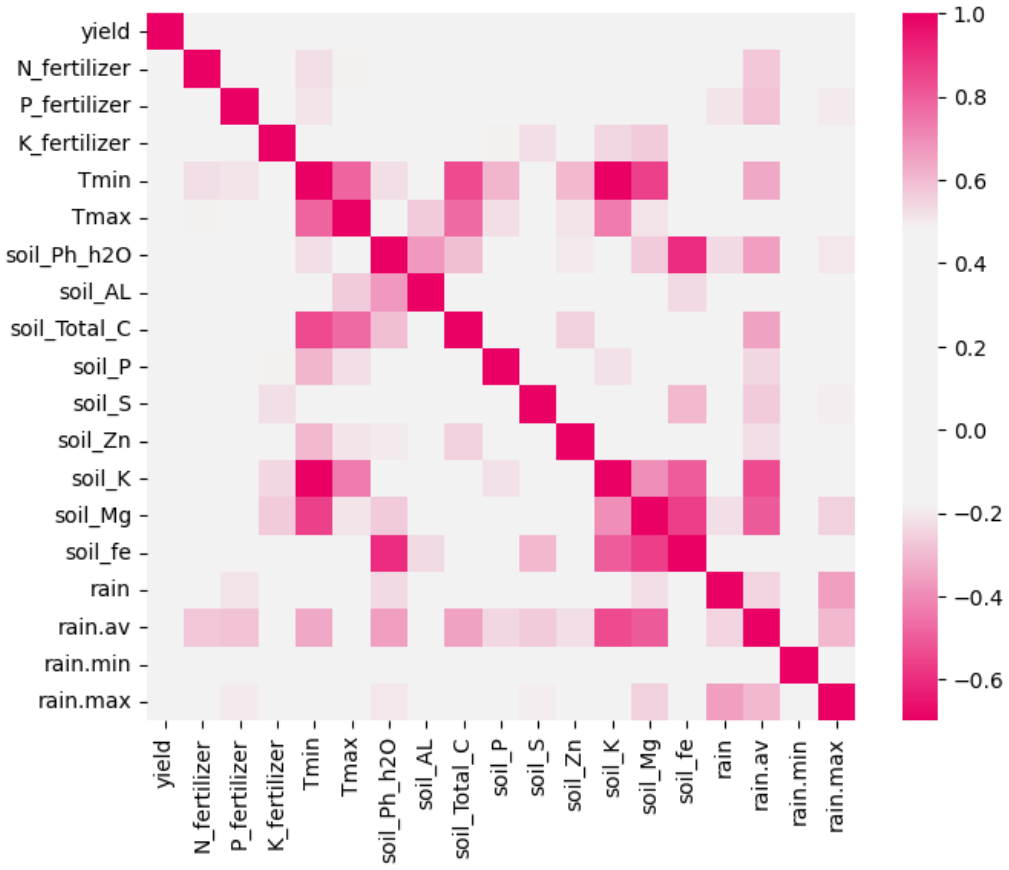

Figure 19: correlation matrix

2.5.4 Data transformation and feature engineering

Data transformation is the process of converting raw data into a a format or structure that would be more suitable for the model or algorithm and also data discovery in general.

in this work, we will transform data into the same scale to allow the algorithm to compare the relative relationship between data points better.

There is three main scaling transformation used in machine leaning: Min Max scaler-Normalization, Standard Scaler - standardization and Robust Scaling.

Min Max scaler-Normalization

MinMaxScaler() is applied when the dataset is not distorted. It normalizes the data into a range between 0 and 1. The mathematical formulation is :

$$X_{new} = \frac{X_i - X_{min}}{X_{max} - X_{min}} \tag{20}$$

Standard Scaler - standardization

We use standardization when the dataset conforms to normal distribution. StandardScaler() converts the numbers into the standard form of mean = 0 and variance = 1. The mathematical formulation is

:

$$X_{new} = \frac{X_i - \bar{X}}{Std(X)} \tag{21}$$

Robust Scaling

RobustScaler() is more suitable for datasets with skewed distributions and outliers because it transforms the data based on median and quantile, specifically. The mathematical formulation is :

$$X_{new} = \frac{X - Median}{inter - quartilerange} \tag{22}$$

We can also notice that all these techniques are applied on numerical variables. For categorical variables, we will use in this work Label encoder function which will transform categorical variables into numeric and Min Max scaler-Normalization for numerical feature transformation.
To make data tidier and more insightful, we created new feature related to fertilizer level and soil nutrients level. These new feature helps the model to learn about the different level of fertilizer application.

2.6 Model selection

The challenge of applying machine learning sometimes belong to how to choose amongst a range of different models ,the best that can be used for the problem. In this work, we perform a model selection among six different Machine learning models: Random Forest (RF),AdaBoost algorithm,Light Gradient Boosting Machine (LGBM), Extreme Gradient Boost (XGB), Decision Tree (DT), Linear Regression (LR) .
To evauluate the performance of the model,we use resampling methods; this means we estimate the performance of the model on out-of-sample data. This is achieved by splitting the training dataset into sub train and test sets, then fitting the model on sub train set and evaluating it on the test set. Resampling method include:

1. Random train/test split
2. Cross-validation (K-fold)

3 Result

3.1 Result1: Validation of baseline Model

In this first part, we will present the results of model selection. The measure use to evaluate the performance of each model is Mean Absolute Error (MAE) and the best model will be the one have a low MAE. After selecting the best models we will present the performance based on three different metrics: Mean Square Error(MSE),Mean Absolute Error (MAE) and R^2 Error.

3.1.1 Model selection result

we decided to subset the dataset by crops, and the result will be for maize since it's the most important crop in the dataset.

3.1.1.1 Mean Absolute Error

N/	Models	Mean Absolute Error (MAE)
01	Ada Boosting	1101.2532
02	Random Forest (RF)	**698.94**
03	Extreme Gradient Boosting (XGB)	**669.867**
04	Light Gradient Boosting Machine (LGBM)	**673.088**
05	Decision Tree (DT)	807.72
06	Linear Regression (LR)	1132.58

Figure 20: Model selection

3.1.1.2 Visualization

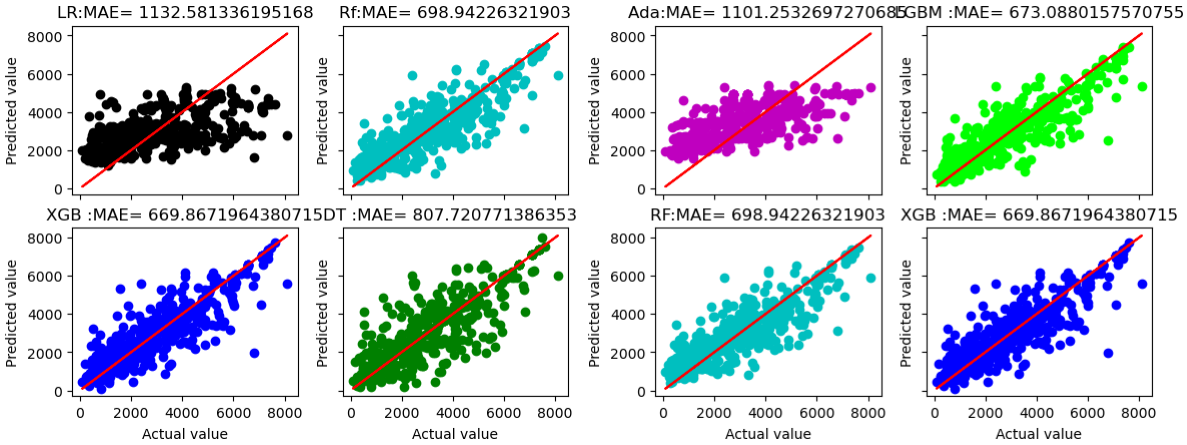

Figure 21: Visualization of models performance during the resampling method

As we can see on Figure 20, we have tested six different models and we can notice that:

1. Extreme Gradient Boosting has a low MAE, follow by Light Gradient boosting,Random Forest and Decision Tree
2. Linear regression is the one that has a large MAE follow by Ada Boost

From Figure 21, we are able to see how each model fits with the data points.

- Random Forest, LGBM,XGB has a good fit of data points than Decision Tree.
- We can also confirm that Ada Boost model and Linear Regression have a bad fit with the data points.

In this analysis , we consider as the best model, the model having the lowest output error and having the best feature leaning reflecting the reality of the problem we are dealing with. That means, for the next step we choose to continue with three models: Random Forest, XGB and LGBM since the their Output Error are in the same range. The final choice will be done after the model assessment in the next part.

3.1.2 Model performance

Models	MSE	MAE	R Square (training set)	R Square (test set)
Random Forest	**871.366**	**611.2082**	0.626526	**0.76393**
XGBoost	886.10413	620.44851	0.6478868	0.75588
LGBM	878.05307	641.014	**0.6600301**	0.7602963

Figure 22: Model selection

These results are obtained after searching for the best hyper parameters of each model using hyper-parameters turned method considering the following elements:

- Random Search Function
- K fold Cross validation With K=3
- 100 different combinations

From the results presented on Figure 22 we notice that:

- All the models have a Good performance in the Test set than the training set based on R^2 Error. That probably means,they was some noise in the training data during the training process.
- Random Forest Shows the overall best performance (Best MSE, best MAE and Best R^2 Error in test set)

At this stage it will be difficult to choose our baseline model, because even if Random forest seem to be the best, all the model Errors are close. The best model will be choose after evaluate how each model is learning from the feature.

3.1.3 Model Assessment

The assessment of the model here refers to Feature Importance. We want to know a score for all the input features for each of the three Models. the scores simply represent the importance of each feature. A higher score means that the specific feature will have a larger effect on the model that is being used to predict yield variable.

3.1.3.1 Feature importance Random Forest

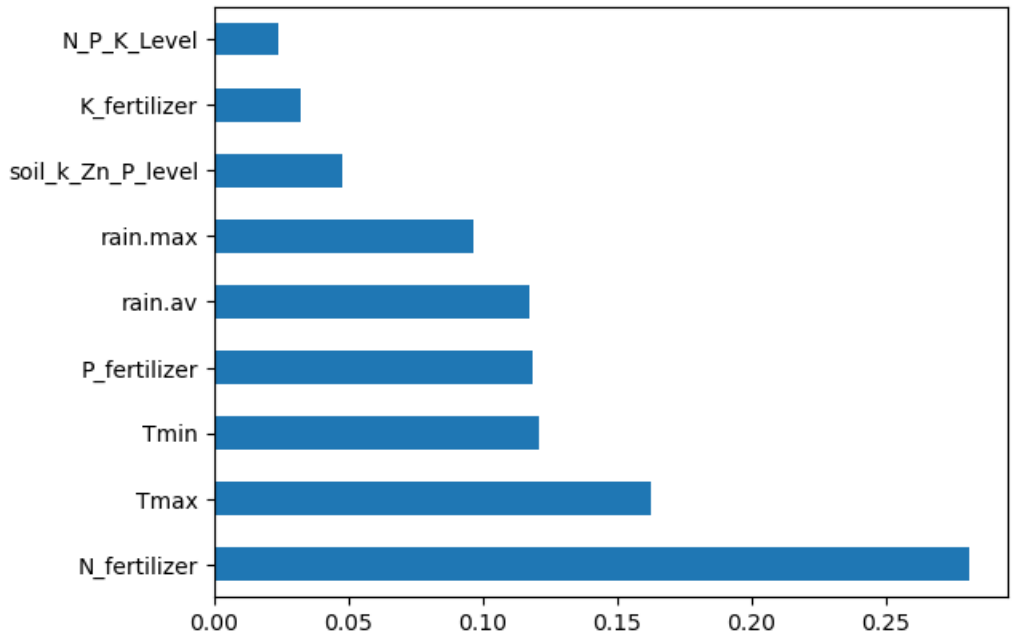

Figure 23: Feature importance RF

From Figure 23, RF model shows that, N fertilizer is the most important variable for yield prediction, followed by Maximum Temperature, minimum temperature,P fertilizer, Average rainfall, maximum rainfall and K fertilizer.

3.1.3.2 Feature importance Light Gradient Boosting Machine

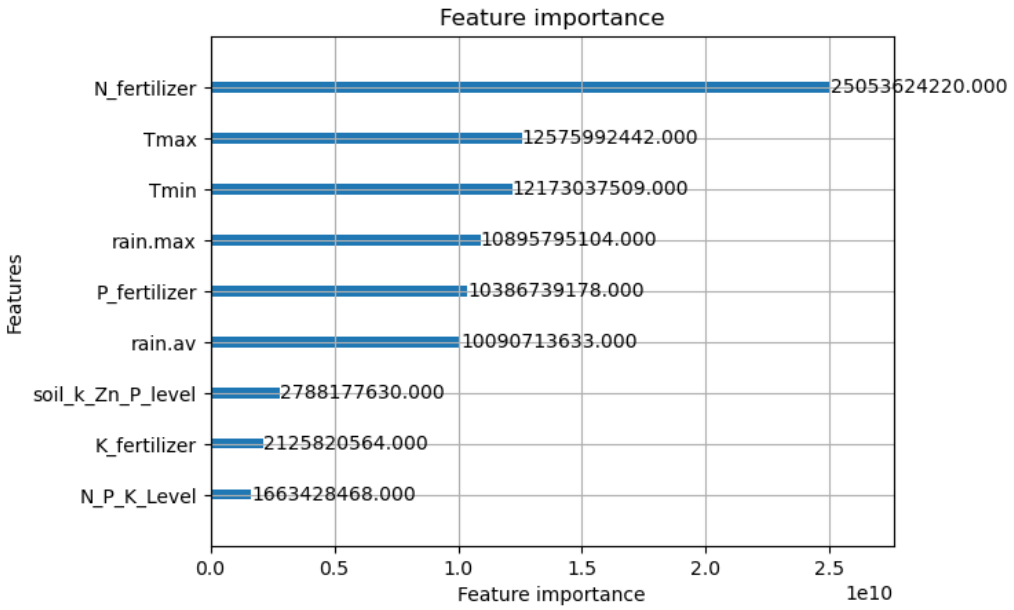

Figure 24: Feature importance LGBM

From Figure 24, LGBM model shows that, N fertilizer is the most important variable for yield

prediction, followed by maximum Temperature, minimum Temperature, maximum rainfall, P fertilizer, Average rainfall, soil nutrient level and K fertilizer.

3.1.3.3 Feature importance Extreme Gradient Boosting

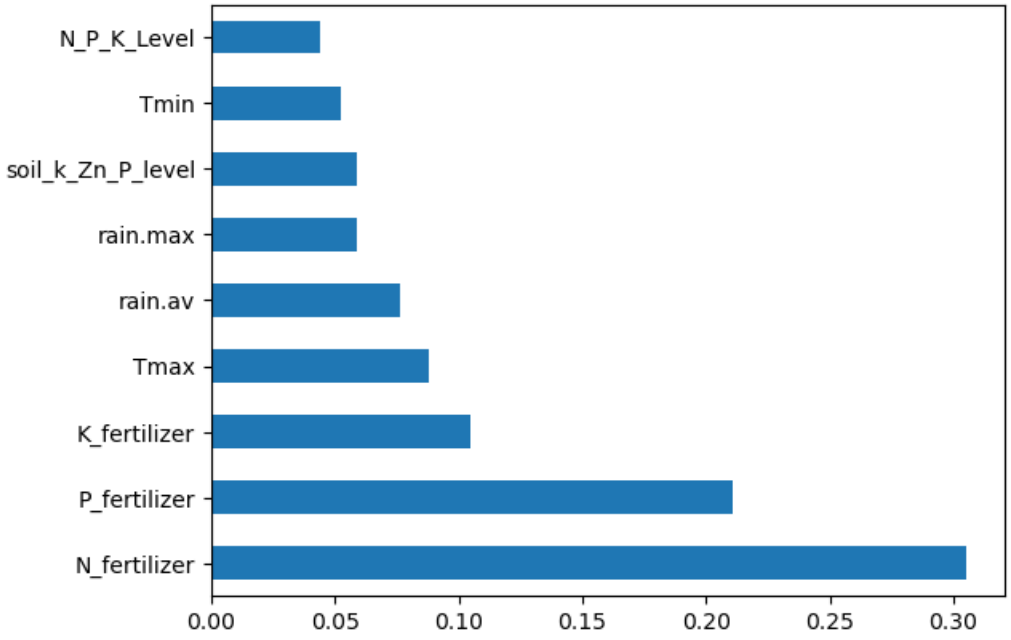

Figure 25: Feature importance XGB

From Figure 25, XGB model shows that, N fertilizer is the most important variable for yield prediction, followed by P fertilizer , K fertilizer and weather product.

3.1.3.4 Synthesis and result analysis

		Models		
		RF	LGBM	XGB
N/	Feature	Rank	Rank	Rank
1	**N fertilizer**	1	1	1
2	P fertilizer	4	5	2
3	K fertilizer	8	8	3
4	**Minimum temperature**	3	3	8
5	**Maximum temperature**	2	2	4
6	**Average rainfall**	5	6	5
7	**Maximum rainfall**	6	4	6
8	N, P, K level	9	9	9
9	Soil (k Zn P) level	7	7	7

Figure 26: Rank of feature per model

The table of the Figure 26, shows the rank of each feature per model. Focusing on Weather features, random Forest and Light Gradient Boosting are almost similar, weather variables are in

the top 6 of the important variables for both model. It shows also that temperature (minimum and maximum) is important than rainfall(average and maximum). On the other hand, Extreme Gradient Boost shows rather an interest on fertilizer (N,P,K) and weather products come in the second plan. So this means that, N,P,K fertilizer are the most important variables for crop yield productivity.
To understand why the model is learning better from fertilizer features specially from N fertilizer, let's have a look on the Figure 27. This figure simply shows that, the dataset we are working on comes from two main countries (probably Nigeria and Ethiopia as we can see on figure 9). looking at the color we notice there is not much change in rainfall data as in N fertilizer. That means:

- The data was collected almost at the same place or at nearby place.
- The dataset is for fertilizer trial, meaning that the output of our different model make sense.

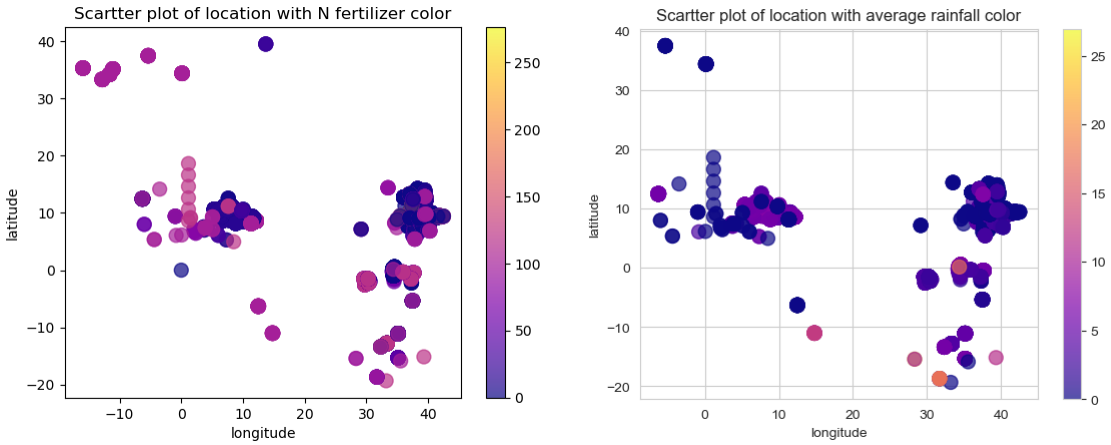

Figure 27: scatter plot of location

Selection of baseline model

We will select the baseline model that will be used to assess agro-weather forecast data based on how the weather variables affect the model accuracy. Considering our dataset without weather variables (**Minimum and Maximum temperature, rainfall**), we notice that:

- A decrease of 32.65% of accuracy on training set and 32.76% on testing set for Random Forest
- A decrease of 34.6% of accuracy on training set and 29.77% on testing set for Light Gradient Boost
- A decrease of 32.81% of accuracy on training set and 27.192% on testing set for Extreme gradient boost.

Based on that result, we will use Random forest as our baseline model for Agro-weather forecast data assessment since it shows the highest impact of historical weather data than others models especially in training set and also the best accuracy.

3.1.3.5 Random forest performance: correlation between predicted data and original data

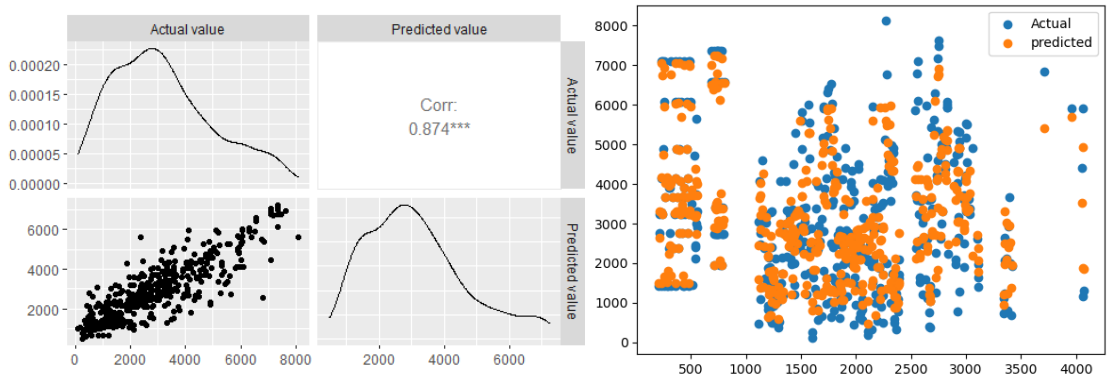

Figure 28: scatter plot and correlation

3.2 Result2: Evaluation of Agro-weather forecast data from ECMWF

3.2.1 Model performance assessment

In this part we comparing the output Error of Random forest when the input data contains historical weather data and when the input data contains weather forecast data from ECMWF.

3.2.1.1 First case: use ECMWF data as input to validate the model

After training and test the model with the historical data,We have tested the model with the same dataset but instead of having historical weather product in the dataset, we have seasonal weather forecast data. the table 1 below shows the obtained result:

Metrics	Weather data	ECMWF data
MSE	574	1119.33
MAE	319.8	806.8
R^2	89%	60%

Table 1: model performance with Weather forecast data and historical weather data.

A) Similarity between yield and predicted yield with observation weather /Forecast data.

Similarity measure	Weather data	ECMWF data
Jacard similarity	0.01055	0
Cosine similarity	0.9876	0.9524
euclidean distance	15422	30040
Pearson correlation	0.946	0.8103

Table 2: Similarity analysis

B) Visualization: correlation between yield and predicted yield

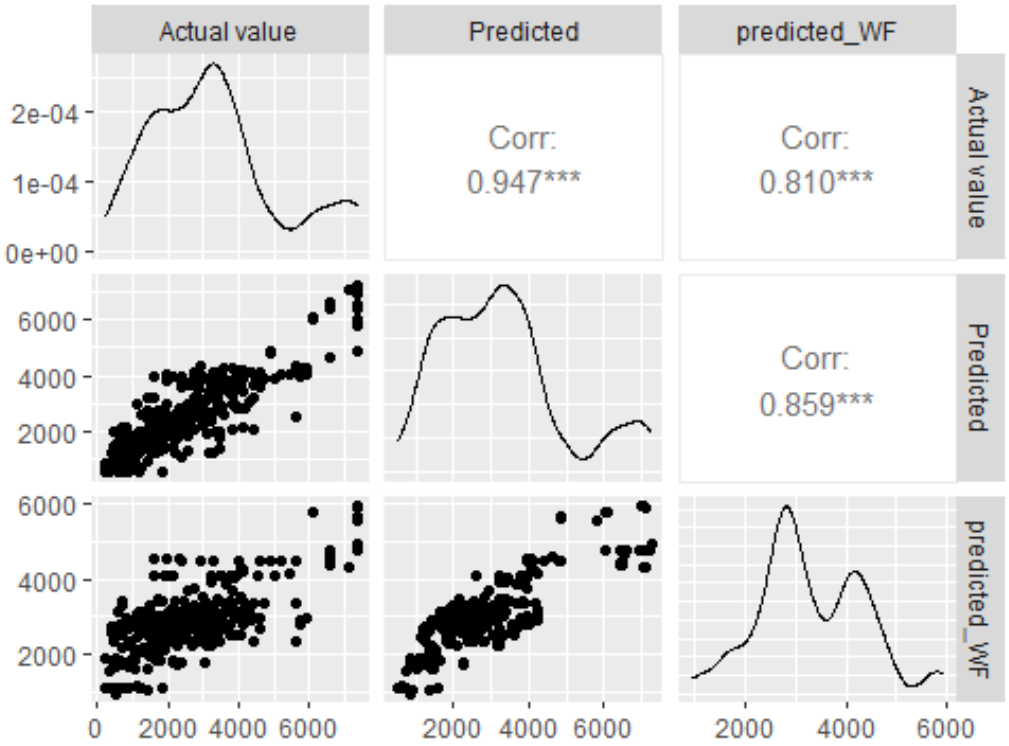

Figure 29: scatter plot and correlation

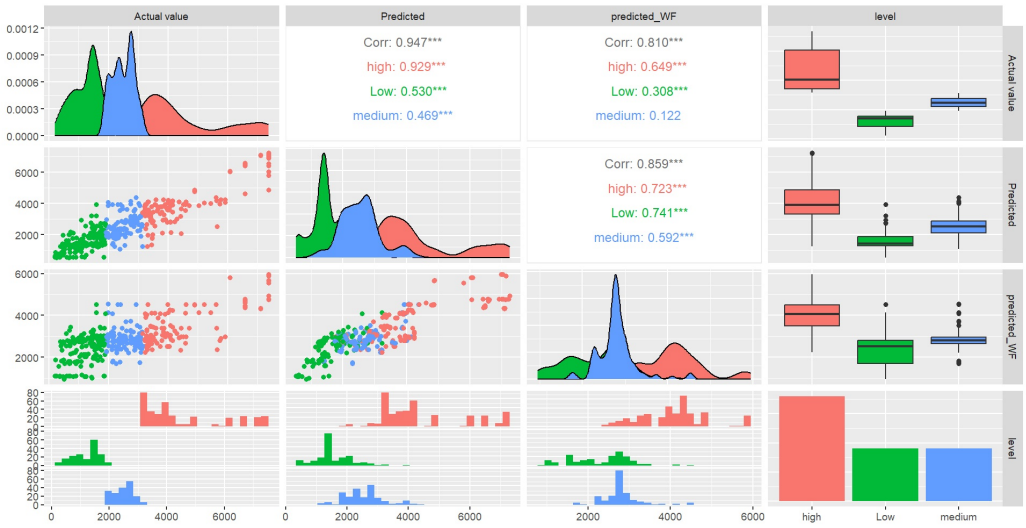

Figure 30: scatter plot correlation with level

3.2.1.2 Second case: use ECMWF data to train and test the model

In this part,we would see how the model performs using weather forecast product instead of historical weather data in the training process. The table below gives the comparative result.

Metrics	Weather data	ECMWF data
MSE	574	676
MAE	319.8	517
R^2	89%	85%

Table 3: model performance with Weather forecast data and historical weather data.

A) Similarity between yield and predicted yield with historical weather /Forecast data.

Similarity measure	Weather data	ECMWF data
Jacard similarity	0.01055	0.0013
Cosine similarity	0.9876	0.9829
euclidean distance	15422	18140
Pearson correlation	0.946	0.925

Table 4: Similarity analysis

B) Visualization: correlation between yield and predicted yield

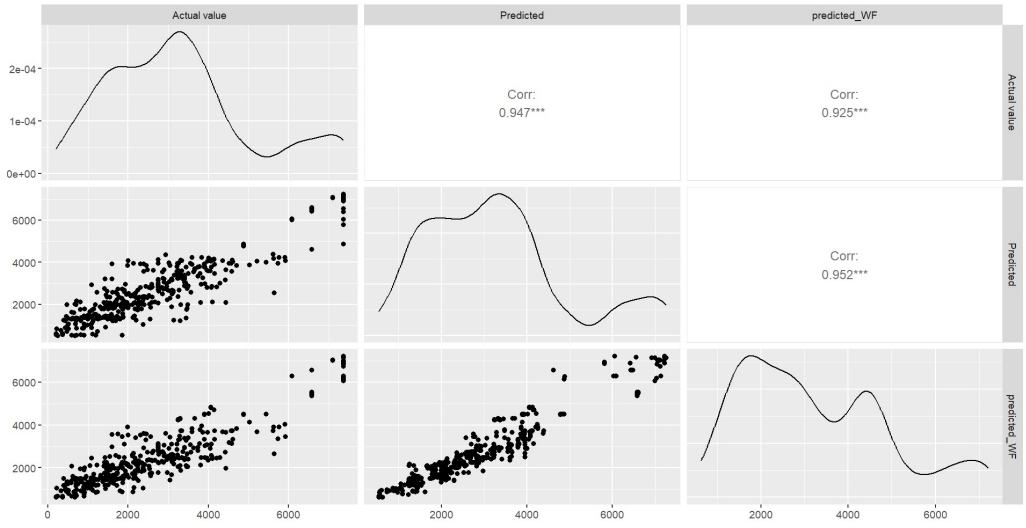

Figure 31: scatter plot correlation with level

3.2.2 Interpretation and Analysis

From Table1 we notice that:

- MSE of the model using weather Forecast data is almost 2 times greater than the MSE of model using historical weather data.
- MAE of the model using weather Forecast data is 2.5 times greater than the MSE of model using weather data.
- From R^2 Error we observe a decrease of 32.58% of the Error when weather Forecast data is used.

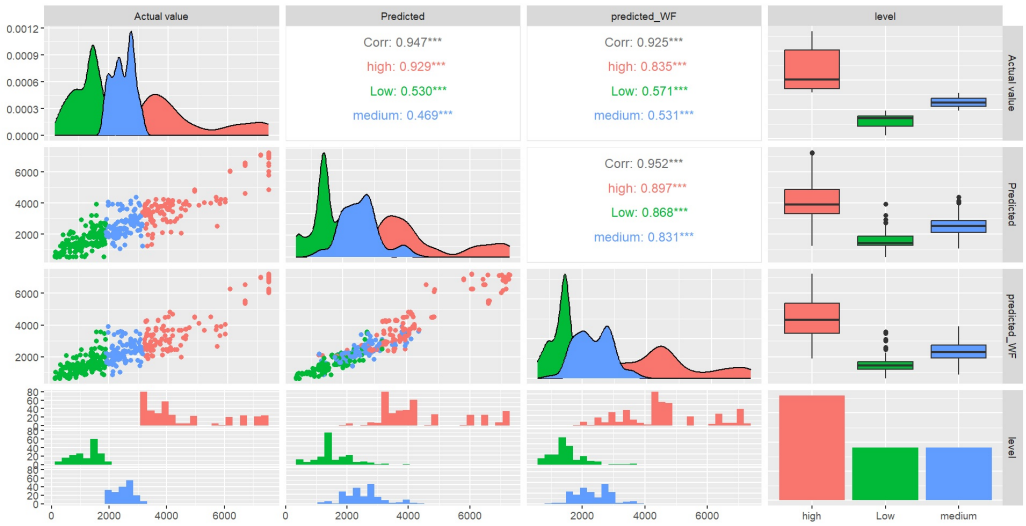

Figure 32: scatter plot correlation with level

Based on this result, there is a considerable gap between **Agro-weather forecast data point** and **Agro-weather data point**. To have more information and better conclusion about the Forecast data in yield prediction, we performed the similarity analysis between the two predicted yields to compare how similar are they with the original yield data. The results presented in Table 2 show that:

- Jacard similarity is equal to 0% for Forecast data and 1% for weather data. This means that 1% of the original yield data have been predicted accurately using weather data against 0% when using weather forecast data.
- Euclidean distance between predicted yield and observed yield using weather data is 2 times less than euclidean distance between predicted yield and observed yield using forecast data. This simply means, the yield value predicted using forecast data is 2 times far from the original yield than the yield value predicted using weather.

In this stage we already have an idea about the accuracy of weather Forecast data compared to original weather data but some close analysis have to be done before concluding. From Figure 29 and 30, we have the visualization of the correlation between the predicted yield and original yield and also the information about the different stage of learning.

- The correlation between predicted yield and observe yield using weather is 0.947, and the correlation between predicted yield and observe yield using weather Forecast is 0.810.
- From Figure 29 we compare the distribution of the data using density plot. Predicted yield has a good similarity with the original yield specially at the end. But for predicted yield using forecast data it is difficult to see the similarity; That means the distribution of the predicted data point is different.

- From Figure 30 we have a close analysis. For both data predicted, certain values are out of the boundary of quantile of the original data as shown in the boxplot.

For this First part of analysis focusing on using Agro-forecast data from ECMWF as test data, we can conclude that:

- ECMWF data is widely different from weather data in terms of MSE,MAE and R^2 error and also in terms of euclidean distance similarity, data point distribution.
- But comparing in terms of correlation between predicted yield and original yield using weather data and predicted yield and original yield using ECMWF data, we observe that the gap is about 13%

We have also performed a comparison when the ECMWF data is used to train and test the model. at this point it's important to mention that we are not much focused on this result since in reality we don't have future data to train the model. But considering we still notice a difference between both results as Table 3, Table 4, Figure 31 and Figure 32 are showing.

4 Conclusion

Climate change is becomes a more and more a worrying subject affecting several fields. Agricultural systems are one of the most affected in the sense that climate change has a direct influence on agricultural climatic elements such as temperature,precipitation and sunshine. All these changes have a direct consequence on agricultural productivity and food security. Therefore, it becomes imperative to find a solution to a better adaptation. It starts by understanding the behaviour of Agroclimatic element in the future; that is why it's important to know how accurate is the weather forecast product. This study was focusing on assessing Agro-weather forecast From ECMWF in crop yield prediction. The idea behind was to estimate the performance of weather forecast data compared to historical weather data both aggregated in the same time period ,and see whether it's possible to use ECMWF data for yield prediction, decision making and crop management. To achieve this objective, Machine learning approach was used to validate a baseline model for crop yield prediction using historical data through the model selection method. Out of six different models,we found that, Random Forest,Extreme Gradient Boosting and Light Gradient Boosting Machine perform better than Multi-Linear Regression Model,Ada Boosting and Decision Tree Model. After performing hyper-parameters tuning and feature importance method, we found that Random Forest has the overall best performance than other models. After training and testing the model with historical data we validate using weather forecast data from ECMWF. We found the decrease of 32.58% of R^2 Error,the MAE 2.5 times greater and the MSE 2 times greater when using weather forecast data. In terms of similarity analysis we notice that, the Euclidean distance between predicted yield and original yield has increased by 50% , Pearson correlation has decreased by 13 % and the other similarity measures such as Jacard similarity,cosine similarity were also completely different and less accurate when using weather forecast data from ECMWF. That simply means,there is a considerable gap between historical weather data and weather forecast data especially in predicting accurately crop yield data. In this work we also experiment the performance of forecast data as a training and testing data. the result was very curios and impressive since we got a better output not far from the output give by historical weather data especially in R^2 and Pearson correlation. It's important to notice that, it was not the best way to go since in reality there is no forecast data to train the model. As future work,we will improve the quality of carob data, test the performance of other forecast products, quantify the gap between these products and historical weather data and finally show the value of implementing the results in agronomy to provide insights for decision-makers into relevant investment areas based on seasonal outlooks in the productive agricultural sector, improving food security.

Reference

[1] Pallab Chatterjee. "What is Legacy Data?" In: *Legacy Data: A Structured Methodology for Device Migration in DSM Technology*. Boston, MA: Springer US, 2003, pp. 5–10. ISBN: 978-1-4615-0241-8. DOI: 10.1007/978-1-4615-0241-8_2. URL: https://doi.org/10.1007/978-1-4615-0241-8_2.

[2] *CHIRTS*. https://www.chc.ucsb.edu/data/chirtsdaily.

[3] *Climate Elements Natural Resources Conservation Service, formerly known as the Soil Conservation Service, is an agency of the United States Department of Agriculture that provides technical assistance to farmers and other private landowners and managers*. https://www.nrcs.usda.gov/wps/portal/wcc/home/climateSupport/fieldOfficeGuide/climaticDataElements.

[4] Adele Cutler, David Cutler, and John Stevens. "Random Forests". In: vol. 45. Jan. 2011, pp. 157–176. ISBN: 978-1-4419-9325-0. DOI: 10.1007/978-1-4419-9326-7_5.

[5] *Decision Trees*. http://scikit-learn.org/stable/modules/tree.html#tree.

[6] Chris Funk et al. "The climate hazards infrared precipitation with stations - A new environmental record for monitoring extremes". In: *Scientific Data* 2 (Dec. 2015), p. 150066. DOI: 10.1038/sdata.2015.66.

[7] *Global Forecast System (GFS)*. https://www.ncei.noaa.gov/products/weather-climate-models/global-forecast.

[8] *Gradient Boost Machine*. https://www.analyticsvidhya.com/blog/2021/09/gradient-boosting-algorithm-a-complete-guide-for-beginners/.

[9] S. J. Johnson et al. "SEAS5: the new ECMWF seasonal forecast system". In: *Geoscientific Model Development* 12.3 (2019), pp. 1087–1117. DOI: 10.5194/gmd-12-1087-2019. URL: https://gmd.copernicus.org/articles/12/1087/2019/.

[10] *Structural equation modeling with AMOS: Basic concepts, applications, and programming*. https://www.researchgate.net/file.PostFileLoader.html?id=551bcf4cd2fd6424088b45e4&assetKey=AS:273770781052931@1442283449007.

[11] *System5-guide*. https://www.ecmwf.int/sites/default/files/medialibrary/2017-10/System5_guide.pdf, note=Version 1.2.

[12] *The Mathematics of Decision Trees, Random Forest and Feature Importance in Scikit-learn and Spark*. https://towardsdatascience.com/the-mathematics-of-decision-trees-random-forest-and-feature-importance-in-scikit-learn-and-spark-f2861df67e3.

[13] Thomas van Klompenburg, Ayalew Kassahun, and Cagatay Catal. "Crop yield prediction using machine learning: A systematic literature review". In: *Computers and Electronics in Agriculture* 177 (2020), p. 105709. ISSN: 0168-1699. DOI: https://doi.org/10.1016/j.compag.2020.105709. URL: https://www.sciencedirect.com/science/article/pii/S0168169920302301.

[14] *worldclim*. https://www.worldclim.org/data/index.htmltext=WorldClim)..