Predicting CO_2 emission per crop-type using Machine learning and Neural Network algorithms

Nisha Chauhan
STUDENT NUMBER: 2033808

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN DATA SCIENCE & SOCIETY
DEPARTMENT OF COGNITIVE SCIENCE & ARTIFICIAL INTELLIGENCE
SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
TILBURG UNIVERSITY

Thesis committee:

prof. Dr. Eric Postma
prof. Dr. Sharon Ong

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science & Artificial Intelligence
Tilburg, The Netherlands
May 2020

Contents

1 Introduction **1**

2 Related Work **2**
2.1 CO_2 emanating sources in agriculture . . . 3
2.2 Related studies . . . 3
2.3 Algorithms and Features . . . 5

3 Experimental Setup **6**
3.1 Data setup . . . 6
3.2 Models . . . 10
3.3 Experiment framework . . . 11
3.4 Evaluation metrics: . . . 11

4 Results **12**
4.1 Experiments . . . 12
4.2 Carbon emission profile . . . 15

5 Discussion **17**

6 Conclusion **18**

Appendices **22**

A Figures **22**

B Tables **23**

List of Figures

1 Raw data Density distribution . . . 8
2 Raw data distribution using boxplot . . . 8
3 Normalized data distribution . . . 9
4 Cross-correlation matrix of Normalized dataset . . . 10
5 Experiment Framework . . . 11
6 Experiment 1: Multi Linear regression model results . . . 13
7 Experiment 1: RFR model results . . . 14
8 Experiment 1: MLP model results . . . 15
9 Total CO_2 emission from residue management practice in year 2017 . . . 15
10 Total CO_2 emission per crop in China from residue management practice in year 2017 . . . 16
11 Total CO_2 emission per crop in Netherlands from residue management practice in year 2017 . . . 16
12 Density pair plot of the raw dataset . . . 22
13 Density pair plot of Log10 Normalized data . . . 23
14 Density plot of Normalized (Log1p) data . . . 24
15 Percentage of total emission per crop type in year 2017 . . . 24

Preface

This thesis is written to fulfill the graduation requirement of Masters of Data science and society at Tilburg University, Netherlands. This thesis is an original work by Nisha Chauhan, and it was conducted under the supervision of prof. Dr. Eric Postma in Department of Cognitive science . Artificial Intelligence, University of Tilburg.

The driving force behind this research is my personal interest to ultimately preserve the nature by addressing a key global problem of Climate Change due to the carbon footprint of Agricultural activities. Considering rising pressure on Agriculture industry due to the facts that agriculture practices are not standard globally, the demand of food is growing due to increasing population. Therefore my approach is to use the historical emission data and predict the CO_2 emission per crop type using neural networks and machine learning algorithms. The results of this study has the potential to contribute in creating innovative solutions to grow a given crop with maximum efficiency and minimum CO_2 emission.

Through out this study I felt so aligned with what Albert Einstein once said "The more I learn, the more I realize how much I don't know". The curiosity to learn was one of the driving force during this research. I would like to take this opportunity to express my humble gratitude to my supervisor Dr. Eric Postma for his time, valuable input and committed support throughout the period of study and especially during research period. Finally, I would like to thank my family and friends for their encouragement and unfailing support.

Nisha Chauhan
Tilburg University
May 2020

Predicting CO_2 emission per crop-type using Machine learning and Neural Network algorithms

Nisha Chauhan

As climate change is emerging as a global challenge, agriculture has secured its position as the third-largest anthropogenic greenhouse gases (GHG) emission contributor. In the past few decades, a substantial number of researches has been done to predict GHG emission and the nature of contributing factors in several sectors including agriculture. In the agricultural sector the majority of the research focuses on forecasting overall GHG emission. However, no research has been done with a focus on profiling CO_2 emission per field crop using Artificial neural networks (ANN) or machine learning(ML) algorithms. This research intends to fill this knowledge gap and presents the use of the ANN and ML algorithm for establishing CO_2 emission profile per crop-type. For the training and testing model, CO_2 emission data of 12 crops of 200 countries were selected from the Food and Agriculture Organization of the United Nation's (FAO) data center (FAOSTAT). Input features were constituted of categorical and continuous variables. The models used for predicting CO_2 emissions from crop residue and burning crop residue were Multilayer perceptron (MLP), Random forest regressor (RFR), and Multilinear regressor (MLR). The results show that all three models were successful in forecasting CO_2 emission per crop type. In addition to that, with the given data, Random forest regressor (RFR) performed better (RMSE 0.032) than MLP (RMSE 0.69) and MLR (RMSE 0.671). The paper also discusses the future application of the results and suggests how the prediction can be made more robust.

1. Introduction

To address the climate change, the United Nations Framework Convention on Climate Change's (UNFCCC) 2015 Paris agreement focuses on keeping global average temperature rise well below 2°C (Rogelj et al. 2016). To achieve this target, the inclusion of agricultural emission is crucial as it is responsible for approximately 11% of global anthropogenic greenhouse gas (GHG) emissions (Ge and Friedrich 2020).

The agricultural-based emissions on the national level are about 35% in developing countries and about 12% in developed countries. With the rapid socio-economic growth in developing nations, their respective contribution is estimated to continue rising in the near future (Richards, Wollenberg, and Buglion-Gluck 2015; Wollenberg et al. 2016). In addition to that, the agricultural sector is already under pressure to produce more food to eradicate food insecurity for roughly 2 billion people (FAO 2009; Smith, Martino, and Cai 2007). The problem will be further intensifies with the population growth as the world population is expected to grow by 2.3 billion by 2050. The additional global demand for food and fodder will put more pressure on the agricultural sector and

eventually cause higher levels of GHG emission (Food and Organization; Balafoutis et al. 2017).

Interestingly, in contrast with other sectors, agriculture has a dual role where it acts as both source and sink (carbon sequestration) of GHG (McCarl and Schneider 2000). This potential of being a sink can be used to achieve carbon-neutral agriculture by minimizing GHG emission and maximizing carbon storage (Lehner and Rosenberg 2017). However, since agriculture involves a complex living biological system, it is hard to measure GHG emissions accurately (McMahon 2019). In addition to that, agricultural born greenhouse gas fluxes are complex and heterogeneous in nature, which gives rise to the challenge of accurately predicting agricultural born emissions. Having accurate information on GHG emission has a direct relation to managing it in agriculture (Olander et al. 2013). By keeping in mind the current trends in global emission, the contributing factors, and how their effect on GHG emissions is likely to grow in the future, it is crucial to further identify, define and quantify the contributing entities and processes used in the agricultural practices. Out of several options of mitigation, the prominent ones are improving agronomic practices, residue management, and fertilizer use (Smith, Martino, and Cai 2007). This is why in this research the focus is on predicting CO_2 emission caused by residue management practices (Smith, Martino, and Cai 2007).

The main aim of the proposed research is to estimate CO_2 emission arising from a particular crop-type by applying artificial neural network (ANN) and machine learning (ML) algorithms. The research addresses the question "How to develop a residue born CO_2 emission profile per crop-type using neural networks and ML algorithms and by using input features such as crop production, census data available in time series?". The prediction of GHG per crop-type is a novel approach with little to no work done on the topic.

The results of the study have the potential to suggest mitigation strategies, site-specific solutions, reforming existing policies as well as forming new ones to manage the CO_2 or GHG emission rate in agronomy. For instance, if CO_2 emission per crop is known, strategies like precision farming can be applied as a spot-on solution to manage high CO_2 emanating crops. Along with that similar models can be applied to further identify GHGs other than CO_2.

The rest of the study is organized as follows: Section 2 describes related work on different methods used for forecasting GHG emission and understanding the important features and their relationship. Section 3 provides a detailed view of the experimental setup including data, pre-processing, and algorithms. Following that, section 4 shows the results obtained from the models. Section 5 discusses the results and implications while Section 6 concludes the study.

2. Related Work

The related work is arranged in a manner where in the first section, main GHG emanating sources in agriculter are discussed briefly followed by the second section analyzing past studies conducted to predict GHG emissions. It is worth noting that literature on predicting GHG emission in agriculture is highly limited especially with crop-based agriculture. Therefore, the literature review not only includes the directly related literature but also a few papers about GHG emission prediction from sectors other than agriculture to understand if there are some interesting trends or techniques that may be of interest to agriculture emission or predictions. Subsequently, their methods and results will also be critically analyzed and discussed. Finally, a brief introduction of the methodology and features adopted for the study is discussed.

2.1 CO_2 emanating sources in agriculture

The agricultural sector mainly produces carbon dioxide (CO_2), methane (CH_4), and nitrous oxide (N_2O) GHGs (McCarl and Schneider 2000). CO_2 emission arises from the direct combustion of residue together with soil respiration and also residual biomass decomposition. In addition to that, farm operations (e.g., energy usage, land use) and inputs (e.g., fertilizers, pesticides) also cause CO_2 emissions. CH_4 is mainly caused by the enteric fermentation of ruminants in livestock farming, fertilizer, and manure management. N_2O is produced mostly as the result of microbial activities to transform nitrogen in soils and by application of chemical fertilizers (Balafoutis et al. 2017).

2.2 Related studies

Among various sectors, agriculture is the third major contributor to the global anthropogenic greenhouse gas (GHG) emissions (about 11-15%), preceded by heat and electricity (31%), and transportation (15%) (Ge and Friedrich 2020; Balafoutis et al. 2017). Managing GHG emission is crucial for agriculture as it impacts its growth and sustainability negatively (Murad et al. 2010). To manage the GHG emissions it is highly imperative to have accurate information on GHG emission in agriculture (Olander et al. 2013).

To understand the relationship between climate change and agriculture growth, a linear regression model was used by Murad et al. (2010). The study concluded that the agriculture production index and per capita CO_2 emissions share bidirectional positive and significant relationships, which indicates that an increase in agricultural production will cause inflation in per capita CO_2 emissions. This study had a limitation of limited data, which might have influenced how significantly the agriculture growth may impact climate change.

Subsequently, few studies focus on forecasting total CO_2 emission for a particular region and given crop, especially rice or wheat. Farag et al. (2013) conducted a study to calculate the total carbon footprint caused by rice fields in Egypt, by aggregating the emission data from rice cultivation, mechanical operations such as irrigation pumping, tillage, harvesting, nitrogen fertilization and burning rice residue. The individual emissions were calculated using mathematical equations and finally, the total carbon footprint was calculated by taking into account all GHGs associated with farming rice paddy (Farag et al. 2013). The calculation of carbon footprint heavily relied on emission factors, which were extracted from limited literature. Therefore, the accuracy of extrapolations might be affected as well. Although it would be interesting to explore if the calculations of the study could be made more robust, and if the model can be expanded to all countries and crops worldwide instead of just rice crop per country.

A similar study was carried out in New Zealand with wheat fields, ANN and a multilinear regression model (MLR) (Safa et al. 2016). The study initially considered 16 features, which were reduced to 8 by employing PCA. The features considered were wheat area harvested, irrigation applications, machinery condition, insecticides, and pesticides. The result of the study shows that ANN predicted CO_2 emission better than the MLR model with ±11% margin error under different conditions. Another interesting study was done by Khoshnevisan et al. (2013), where the total CO_2 emission from the wheat field was carried based on input energy. The input energy includes direct and indirect emissions caused by machinery, fuels, fertilizers, pesticides, and electricity. The CO_2 emission per feature was calculated by taking the multiplication of the input feature with the GHG coefficient. The used ANN model performed well with the mean

squared error (MSE) of .064 (Khoshnevisan et al. 2013). Above mentioned studies acted as a useful foundation for deciding the possible approaches for creating carbon profile per crop type. Although they were different in terms of features, methods, focus areas such as crop, and country, but the underlying concept was similar where the intention was to predict the emission for a selected crop in a country. This research builds on the top of these underlying ideas and targets to expand the existing research to create the carbon emission profile per crop in each country.

In 2016, Tongwane et al. (2016) conducted a study to forecast GHG emissions from field crops and investigated the contribution by each field crop using crop management practices such as the application of fertilizers, lime, crop residue management. The calculations were done by Agriculture and Land Use National Greenhouse Gas Inventory (ALU) Software, which is based on the Intergovernmental Panel on Climate Change (IPCC) Guidelines for National GHG catalog. As a result of the study, Tongwane et al. (2016) were able to predict the GHG emissions by crop management practices and by crop commodity for South Africa. This research is similar to the proposed thesis to quite some extent, which is why a critical analysis of the literature was required. The close analysis of the study done by Tongwane et al. (2016) reveals few limitations. The first limitation of the study is that the ALU software works with IPCC compliant data only, therefore it might not work with unstructured data or the data that is not IPCC compliant. Secondly, for making future predictions, it uses the "mitigation" function, which treats the past year as the baseline instead of extracting the patterns from the whole data (Milne et al. 2013). In addition to that, the tool covers only emission source features associated with the agriculture and land use (Milne et al. 2013), which means it is not possible to include more features such as demographic, economic, and energy consumption in the model, which can be useful in order to have more accurate information (Farag et al. 2013; Alston and Pardey 2014; Smith, Martino, and Cai 2007). Thirdly, the researchers assumed that the selected year was an average representative of other years and that there are no changes in crop management practices (Tongwane et al. 2016). Furthermore, the study also lacks in stating the prediction accuracy and publishes the predicted values. Consequentially it is hard to understand the accuracy of the predictions and reliability and scalability of the model. The limitations of the Tongwane et al. (2016) study has partially fueled the motivation to use neural networks and ML algorithms to create the CO_2 profile.

Another interesting study focused on predicting GHG from manure management practices was done by Lim, Moon, and Kim (2007). They used an approach called PWA, which is a combination of Principal component analysis, Weight portioning method, and Artificial neural network to forecast ammonia gas emission from field-applied manure and to assess the relative significance of ammonia emission factors. The complex nonlinear effects between the ammonia emission variables such as soil and manure states, climate, and agronomic factors were predicted with high accuracy (MSE=0.8333) by the PWA technique. The model was also able to identify the most significant contributing factors (wind speed, soil pH value, average air temperature, and manure pH value) in emission from field-applied manure.

As mentioned earlier, the literature review is limited due to relatively small number of titles on the topic of predicting GHG emissions. To overcome this limitation, the literature review was extended to sectors other than agriculture to learn more about the different prediction techniques employed and features used for forecasting the GHG emissions. Meng and Niu (2011) used the logistic function to improve the CO_2 emissions prediction from fossil fuel combustion from industries such as agriculture, construction, transport, storage, wholesale and retail trades, hotels, and catering services in China.

The study concluded that logistic regression was more suitable for the prediction as compared to the linear regression model. A similar study was carried out by Liu et al. (2012), for predicting CO_2 emissions in China, using time series input features like gross domestic product (GDP), consumer price index (CPI), population, subprime mortgage, European debt crisis. The neural network proved to be promising in predicting the emissions with the relative error between -1% to 1%.

Adaptive Network-based Fuzzy Inference System (ANFIS) was also used to forecast the annual emission of greenhouse gases (GHG) using features such as deforestation indexes, energy consumption, economical, and population growth index. The Neural-Fuzzy System proved to be useful in forecasting GHG and produced consistent results (Rodrigues et al. 2009). A similar study was done in Russia by Samoilov and Nakhutin (2009), to predict total anthropogenic GHG emission in the country. They employed a statistical method (linear extrapolation method) and selected macroeconomics and microeconomics features to forecast total GHG emissions. Intending to create a more robust method, Radojević et al. (2013) used an artificial neural network (ANN) to forecast CO_2 emission, with the dataset having missing value. As input features, sources of energy, gross domestic product (GDP), gross energy consumption, and energy intensity were selected. The results showed that the relative prediction error was less than 10% and a conclusion was made that ANN can be useful to simulate development scenarios and their environmental impact. These extended literature review was useful in understanding the input features that affect the agriculture born emission such as GDP, CPI, and energy usage.

As the review reveals, a considerable number of studies have been performed by using various methods to forecast the overall GHG emissions. However, no study was found that focuses on profiling CO_2 emission for the field crop per country using ANN or ML, which motivates the aim of this research. The objective is to develop CO_2 emission profiles per crop-type using residue management emission data. The dataset consists of a six-decade data of crop production, agriculture activities, and census data.

2.3 Algorithms and Features

ANN is slowly making its mark in agriculture with recent applications, which mostly focuses on plant classification, weed, pest identification and control, yield prediction, disaster monitoring, fruits counting, robotic harvesting, and animal behavior classification (Zhu et al. 2018; Karmokar et al. 2015). For most of the applications, the strong capability of image processing of ANN models such as CNN, GAN is used (Hanson et al. 2017; Zhu et al. 2018). The main reason behind the popularity of ML and ANN approaches is that these approaches are well capable of recognizing the hidden patterns in the high-dimensional data and that they are able to solve linear and nonlinear functions (Hassoun et al. 1995). This study attempts to use the prediction capabilities of ML and ANN, which is also supported by the literature (Lim, Moon, and Kim 2007; Radojević et al. 2013; Behrang et al. 2011; Safa et al. 2016).

To be able to create CO_2 profile per crop-type, the agriculture specific features should also be available per crop type. After exploring the data available, yearly production features per field-crop per country such as area harvested, yield, production quantity features were selected. The selection is also supported by the literature which points out that production data is related to the residue produced by the crop (Scarlat, Martinov, and Dallemand 2010). Literature also suggests that the Gross Domestic Product (GDP), population, per capita income, and employed technology also affects the agricultural born emission (Radojević et al. 2013). To find if a country and its

annual population conceals the information of GDP, per capita income in some way, population per country was also selected as a feature. The Features and employed model is discussed in more detail in the "Experimental setup" section.

3. Experimental Setup

This section mainly discuss the data processing, selected models, Experiment framework, and evaluation metrics used in the study.

3.1 Data setup

This section firstly focuses on describing the source of data and selected features and labels. Successive subsections concentrate on the observations obtained by exploring data and their analysis. Subsequently, data cleaning and pre-processing were described.

3.1.1 Data source and data exploration. The primary source of the data used as part of this research is an open-source data provided by the Food and Agriculture Organization of the United Nations (FAO). FAOSTAT is FAO's data center that provides food and agriculture-related data for about 245 countries and duration ranging from 1961 to 2017. It is worth highlighting that the provided data by FAOSTAT is a combination of official reported data of countries, imputed data for unreported commodities or countries, and computed data for some elements from the existing available data. The features and labels were extracted from the raw datasets. The details of the raw dataset are as follows:

Production dataset: The features extracted from the production dataset are the record of crops grown in 245 countries for years 1961-2017, with their statistics such as area-harvested (hectare), production quantity (hectogram per hectare), and yield (tonnes) of around 6 decades (1961-2017). "Area harvested" value refers only to the area from which there was an actual harvest collected and excluded the area that was sown but the crop was not harvested due to any reason such as damage or failure. Area harvested is recorded as many times as crop is harvested, this can be more than once in case of successive cropping and only once in case of successive gathering from the same standing crop during the year. "Production quantity" is simply the amount of crop produced per year. "Yield" is the ratio of production amount per unit of harvested area for given crop products. The yield data is obtained by dividing harvested production per unit of harvested area for crop-products and not recording.

Population dataset: Total annual population per country was selected as a feature.

Crop residue and Burning crop residue dataset: The crop residue dataset provides information about direct and indirect GHG emissions from crop residue left on agricultural fields. The burning crop residue dataset holds information of GHG produced by on-site combustion of crop residue. The CO_2 emission(gigagram) per crop per country was extracted as label.

The final dataset was constructed using two categorical variables: country and crops, and five continuous variables: "Area harvested", "Production quantity", "Yield", "Annual population", and "Years". To have singular value as the label, arithmetic sum of residue emission and burning residue emission was taken which includes the CO_2 emissions produced from crop residues burnt on-site and deposition and leaching processes of the leftover on-field residue. Python library pandas was used to work with raw datasets (McKinney et al. 2010).

3.1.2 Data Pre-processing. The first observation while exploring the dataset was the presence of null labels. The "Crop residue" and the "Burning crop residue dataset" provide GHG emission for only 12 crops, for the reason that not all crops produce residue in reasonable amounts that can be burnt or where the leftover can produce CO_2 in significant amounts, for example, apples, almonds, coriander. The 12 crops for which CO_2 emission data is present are: Barley, Maize, Millet, Potatoes, Rice, paddy, Wheat, Dry Beans, Oats, Rye, Sorghum, Soybeans, and Sugar cane. As the first step in data cleaning, all the data points with null labels were discarded, which means all the crops except the 12 crops mentioned above were discarded. In addition to that, about 1.5% of records were found having a null value in Area harvested, Production quantity, and Yield features. The missing values were treated using the complete case analysis as the proportion of missing variables was less than 5% (Pigott 2001).

The Table 1 summarizes the descriptive statistics of the raw data, where it is evident that although the data is scattered across the wide range, mostly the data is clustered to a certain interval. The statistics from the table show that the features and labels have high standard deviation and the mean is also influenced by the extremely large values. The distribution of the data was also investigated using scatter plots by employing python packages seaborn and matplotlib (Hunter 2007; Waskom et al. 2017). The Figure 1 shows that the data distribution was highly skewed. In general terms, skewness is a measure of asymmetry in a given distribution. The distribution of raw data is positively skewed as the tail extends out to the right and leptokurtic. Table 2 shows the calculated coefficients for skewness and kurtosis. In general, values closer to 0 indicates, approximate normal distribution (Lewis et al. 2005).

Table 1: Descriptive statistics of the raw data set

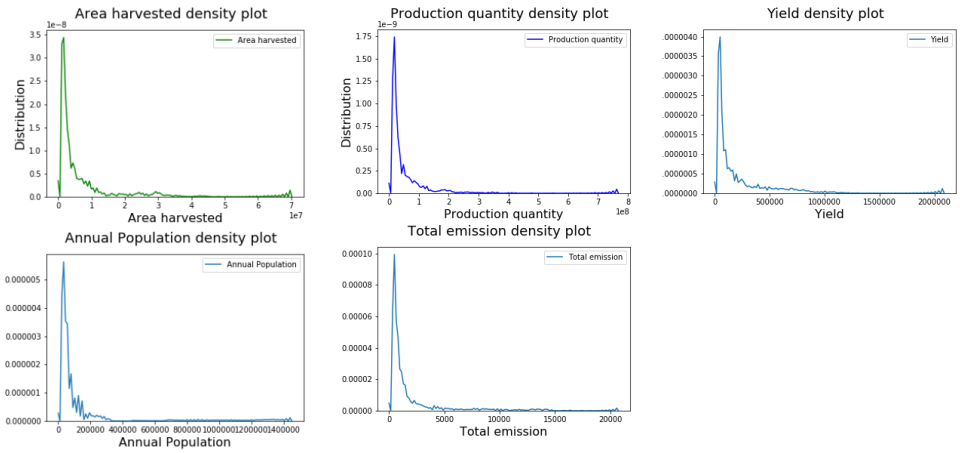
	Area harvested	Production quantity	Yield	Annual Population	Total emission
count	66584.0	66584.0	66584.0	66584.0	66584.0
mean	782179.8	3253076.9	81957.5	55908.6	172.0
std	3431116.1	18863772.4	174639.4	183014.4	953.5
min	0.0	0.0	0.0	4.4	0.0
25%	4100.0	8728.5	10263.0	4059.7	0.5
50%	40000.0	97399.0	20000.0	10434.5	5.1
75%	225654.8	802767.5	50331.2	31411.1	35.3
max	70205008.0	768594154.0	2088710.0	1452625.2	20742.2

The table shows the descriptive statistics of the raw dataset. The data is spread across wide range, but most of the values are concentrated to a definite range. The extreme values causing the mean to shift towards the right.

The data distributions were further investigated to understand the reason for the outliers values. Figure 2 shows that Area harvested and Production quantity is having quite some outliers. Few random cases of detected outliers were investigated to find if the extreme values were outliers or not. It was found that the selected outlier values were not outliers, but were affected by external factors such as recurrent droughts, flood and cyclones, climate change, political conflicts, investment or disinvestment in agricultural research, evolving agro-management practices. (FAO 2016; Qureshi and Afghan 2005; Aye and Edoja 2017). Following the affecting historic trends and concluding the cases are outliers or not, is extremely difficult and time-consuming.

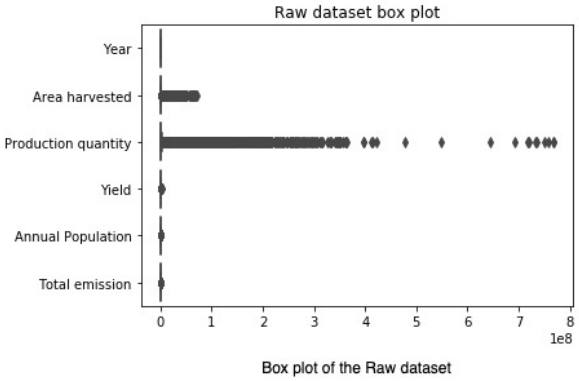
Therefore, to address the skewness and outliers, Logarithmic (Log10) Normalisation was selected (Lewis et al. 2005). Logarithmic transformation is a powerful and popular method to reduce the skewness by dampening the effect of larger values. Before applying the Log transformation, the zero values present in features needed to be handled first. On closer observation, an anomaly in the data records was found, e.g.

Figure 1: Raw data Density distribution



The figure shows the density distribution of the raw dataset. The distribution is positively skewed as the tail is extended towards the right.

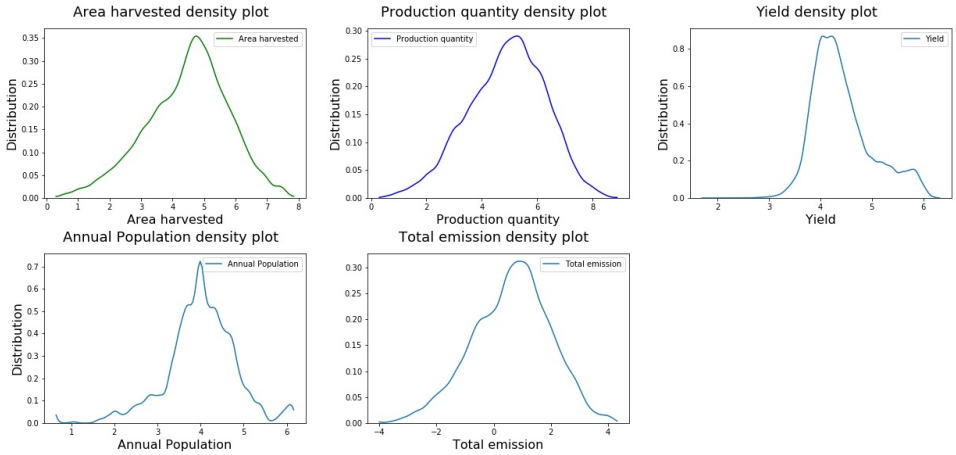
Figure 2: Raw data distribution using boxplot



Note: The figure shows the density distribution of the raw dataset. The distribution is positively skewed as the tail is extended towards the right.

it is not possible that if the area harvested for a crop is zero, the production quantity and yield are non-zero. In simple words, if farming or harvesting of a crop is not done, production quantity and yield are bound to be 0. For simplicity, the records with zero value (.07% of the dataset) were discarded. Figure 3 shows that the distribution of data became relatively normalized after applying the log10 normalization technique. Table 2 shows the coef. of skewness before and after applying log normalization and Table 3 shows the descriptive statistics of Normalised data. It is clear from the table that the mean is relatively balanced and the standard deviation is also reduced. This happens because the Log10 transformation has dampened the effect of extremely large value, consequently, their effect on mean was also reduced.

Figure 3: Normalized data distribution



The figure shows the distribution of raw dataset after applying Log10 Normalisation became relatively normalized.

Table 2: Data Skewness and Kurtosis coef.

Feature/Label	Coef. of skew	Coef. of skew after Normalization	Kurtosis. Coef. Of Raw data	Kurtosis. Coef. after normalization
Area harvested	8.636	-0.380	94.509	0.044
Production quantity	16.954	-0.294	452.409	-0.150
Yield	3.682	0.807	15.242	0.314
Annual Population	5.654	-0.205	32.968	1.001
Total emission	10.747	-0.257	135.501	-0.055

The table shows the Skewness and Kurtosis coef. of the dataset before and after applying the Log10 Normalization. Post normalization the coef. were close to zero, which implies approximately normal distribution.

Table 3: Descriptive statistics of log10 Normalized data

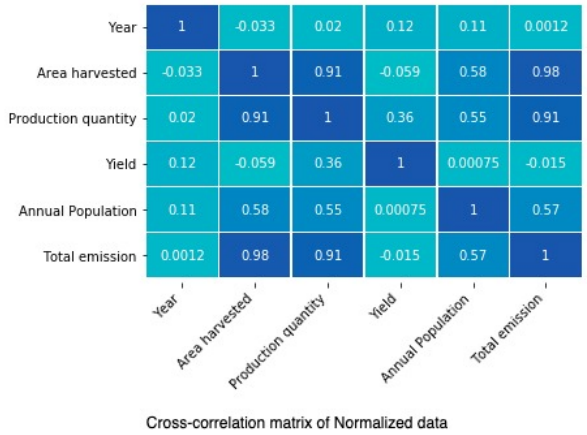
	Area harvested	Production quantity	Yield	Annual Population	Total emission
count	66539.000	66539.000	66539.000	66539.000	66539.000
mean	4.464	4.888	4.424	4.023	0.615
std	1.322	1.413	0.584	0.786	1.369
min	0.000	0.000	1.699	0.646	-4.000
25%	3.615	3.944	4.011	3.608	-0.307
50%	4.602	4.990	4.301	4.018	0.710
75%	5.354	5.905	4.702	4.497	1.548
max	7.846	8.886	6.320	6.162	4.317

The table shows the descriptive statistics of the Log10 Normalized dataset.

As secondary approach, instead of deleting the records log1p normalization was also tried, and to some extent the feature distribution was normalized, but the label data distribution moved far away from normalised distribution (Appendix A, Figure: 14), which is why this approach was rejected.

Finally, the Pearson correlations of normalized data were also investigated (Figure: 4). The three evident strong correlation coef.(r) found were: (1) Area harvested and production quantity (r=0.91), (2) Area harvested and Total emission (r= 0.98) and (3) production quantity and Total emission (r=0.91).

Figure 4: Cross-correlation matrix of Normalized dataset



The figure shows the Pearson correlation of the Log10 Normalized dataset. Area harvested and Production quantity shows strong correlation with Total emission.

The strong correlations were further probed to find out if a causal relationship exists between the variables or it is a coincidental correlation. The "Area harvested" and "Production quantity" are directly related to the amount of residue produced (Boateng, Obeng, and Mensah 2017) and which eventually directly impacts residue born emission (Murad et al. 2010). This direct relation explains that strong correlations are causal relations. To avoid the strong impact of "Area harvested" and "Production quantity" on to-be predicted label, these features were deleted from the dataset.

The data exploration and pre-processing was done by using python packages such as seaborn (Waskom et al. 2017), matplotlib (Hunter 2007), numpy (Oliphant 2006), pandas (McKinney et al. 2010), math (Van Rossum 2020), and Scipy (Virtanen et al. 2020).

3.2 Models

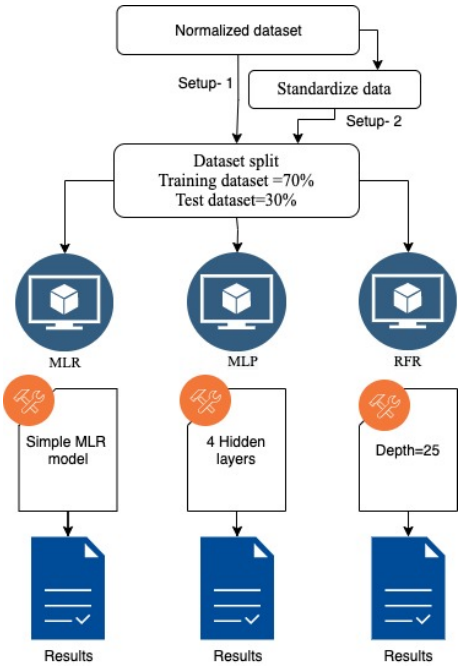
As discussed in related work, Artificial neural network (ANN) and machine learning(ML) models are quite popular for prediction tasks including GHG emissions forecasting (Lim, Moon, and Kim 2007; Radojević et al. 2013; Behrang et al. 2011; Safa et al. 2016). For this research MLP, which is the most common artificial neural network was selected as one of the models (Safa et al. 2016). In addition to that, MLR (Safa et al. 2016) and Random forest regressor (RFR) algorithms are also selected.

To build MLP, several number of trials led to the selection of four hidden layers, the ReLU activation function, and the "adam" optimizer. Similar to MLP, after trying different combinations, the optimized RFR model was obtained with a depth of 25 and 200 trees in the model. Finally, the MLR was also built. Due to the absence of any baseline from the related work, the MLR performance will be treated as baseline performance and the overall performance of the model will be accessed against the baseline model. All the models are built using an open source library scikit-learn (Pedregosa et al. 2011).

3.3 Experiment framework

The overall experiment framework is shown in Figure 5. The experiment (Experiment 1) was divided into two sub experiments (Experiment_1a and Experiment_1b), where each model was trained and tested with setup-1 and setup-2. In setup-1, the normalized data was used for training and testing the models. In setup-2, the normalized data was first standardized (Z-score standardization) and then was used for training and testing the models. The data was randomly split into train and test set in a proportion of 70% and 30% respectively. To create the emission profile, a subset of data with given year is kept aside to be used as unseen examples to forecast the CO_2 emission with best performing model. This subset will not be used in training or validation of the selected model, this means that model will see these samples first time. Using the result of these unseen sample the profile will be created, which will be discussed in "Result" section.

Figure 5: Experiment Framework



The figure shows the overall experiment framework consisting of three models and two experiment setups.

3.4 Evaluation metrics:

The model results are analyzed using Mean squared error (MSE), Root mean squared error (RMSE), coefficient of determination (R^2) metrics. The MSE and RMSE are commonly used metrics in forecasting tasks to verify the model performance. The underlying assumption of MSE and RMSE is that the distribution of the data is normal. Furthermore, the RMSE represents errors in the same units as predicted and recorded data. The smaller the RMSE is, the more accurate the model is in forecasting the values. R^2 is the proportion of the variance in the label that is explained by the feature(s) and it is scaled between 0 and 1. Ideally, closer the value of R^2 to 1, better the model fits the

data. R^2 is easy to interpret and provides a quick overview of how well the model is able to explain the variance in the dataset. However, it does not indicate to what extent the predictions deviate from the actual values in the dataset. The RMSE and MSE provides more information about the deviations by quantifying the errors. The R^2 is a relative measure of fit and RMSE is an absolute measure of fit. In addition to that, the MSE and RMSE are sensitive to outliers, and as shown in Table 1 and 3, the given dataset has some extremely small and some extremly large values. Which is why for the given dataset RMSE and MSE are suited more. Moreover, the related studies also show the use these metrics for evaluation of forecasting task.

4. Results

This section primarily discusses the results obtained from the proposed models. Later, the subsection Carbon emission profile subsection examine the created CO_2 emission profile and compare it among countries such as China, Netherlands, and USA.

4.1 Experiments

In this section the models were trained on 70% randomly selected samples of data and the rest 30% samples were used for validation. As explained in "Models" section, each model was trained and tested with setup-1 and setup-2. The achieved results are shown in Table 4.

Table 4: An overview of R^2, MSE,and RMSE, obtained in Experiment_1a and Experiment_1b

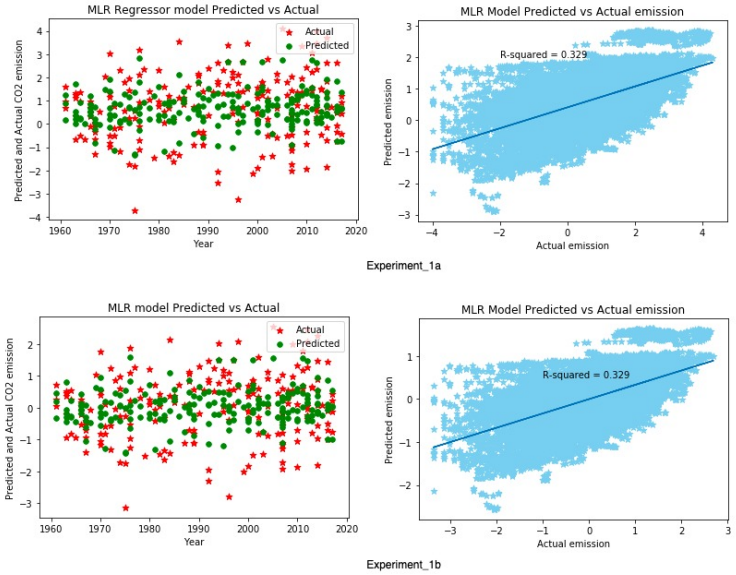
	Experiment_1a			Experiment_1b		
Evaluation metrics	MLR	RFR	MLP	MLR	RFR	MLP
R^2	0.329	**0.97**	0.33	0.329	**0.968**	0.31
MSE	1.121	0.24	1.124	0.819	0.18	0.831
RMSE	1.258	**0.057**	1.264	0.671	**0.032**	0.69

4.1.1 Multi Linear Regressor. The MLR determines the linear relationship between features and label. The equation (1) shows the linear equation to predict CO_2 emission using input features such as Country, Crop, Year, Yield, and Population. Where β_0 is the intercept and $\beta_1, \beta_2, \beta_3 ...$ are the regression coefficients.

$$\textit{Total emission} = \beta_0 + \beta_1 \textit{Country} + \beta_2 \textit{Crop} + \beta_3 \textit{Year} + \beta_4 \textit{Yield} + \beta_5 \textit{Population} + \epsilon \quad (1)$$

In Experiment_1a, the MLR model was able to be fitted to the CO_2 emission data and was accounted for 33.8% of the variance in training data and 32.9% of variance in test data. The upper part of the (Figure 6) shows the results obtained on Experiment_1a. The scatter plot shows the predicted values and actual values of CO_2 emission per year. For simplicity of understanding, the scatter plot only shows first 200 recorded and predicted values from the test set. As it can be seen in the scatter plot, the model is able to predict relatively accurately for the values around the mean. The right side of the Figure shows the model fitness to the data, where it can be seen that due to outlier values the model is accounted for 32.9% of variance in the test data.

Figure 6: Experiment 1: Multi Linear regression model results

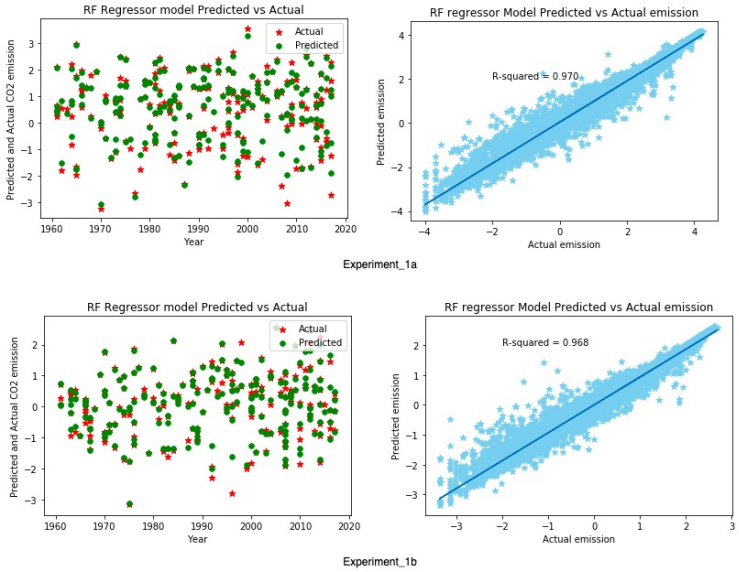


The left side of the figure shows the MLR model's predicted values and recorded values on validation dataset and the right side of the figure shows how well the model was able to explain the variance in validation dataset.

To improve the accuracy, Z-score standardization was applied to the data (setup 2) in Experiment_1b. As presented in Table 4, the RMSE value improved from 1.258 to 0.671. In addition to that, it is interesting to notice that the R^2 the model remained the same. Due to absence of related research, in this study the MLR model is treated as a baseline model and as MLR model is performing better in Experiment_1b, the result metrics are treated as baseline metrics.

4.1.2 Random Forest Regressor. The RFR model was also trained and validated with and without standardized data. In the Experiment_1a, the model is able to fit pretty well to the CO_2 emission data and is accounted for approximately 97% variance in the test data. The RMSE value for the setup 1 is 0.057. As shown in the scatter plot of the Figure 7: Experiment_1a, the predicted values are close to the actual values. Similarly, in Experiment_1b the RMSE obtained is .032 and model is accounted for 96.8% variance in the test data. On closely obsevering the scatter plot of the Figure 7: Experiment_1b and comparing it with the scatter plot of Experiment_1a, it is clear that there are less prediction errors. Overall the standardization of the data shows improvement in RMSE from 0.057 to .032 from Experiment_1a and Experiment_1b respectively. As comparing the results of Experiment_1b with the baseline results, RFR is clearly showing better performance with RMSE (.032) as comared to the MLR baseline RMSE (0.671). As RFR model is performing significantly better than the MLR model, it is interesting to explore if the model is over-fitting to the noise of the data. The RMSE on Train and test set for Experiment_1b is 0.005 and 0.032 respectively. The model do not show any signs of over-fitting. Appendix B Table 5 shows the complete results on training and test data for Experiment_1a and Experiment_1b.

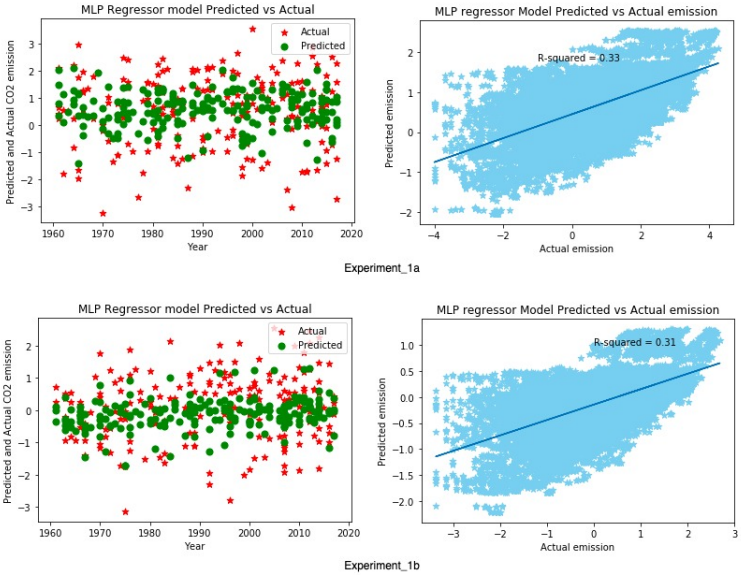
Figure 7: Experiment 1: RFR model results



The left side of the figure shows the RFR model's predicted values and recorded values on validation dataset and the right side of the figure shows how well the model was able to explain the variance in validation dataset.

4.1.3 Multi Layer Perceptron. The results of MLP are similar to MLR. In Experiment_1a, the RMSE value for MLP model is 1.264, which is slightly more than the MLR Experiment_1a RMSE(1.258). The MLP model is able to fit the CO_2 emission data and is accounted for 33% of the variance in the validation set. In Experiment_1b, the model RMSE improved from 1.264 to 0.69, with relative fit of the model being approximately the same 33%. The Figure 8 shows the predicted and recorded values plotted as scatter plot and the right side plot shows how well the model is able to explain the variance in the data. The MLP model shows relatively same characteristics of the MLR model. It is interesting to notice that the MLP model is performing very close to the the baseline. On comparing with the baseline (RMSE =0.671), the MLP model is performing slightly worse with RMSE (0.69). The Figure 8 illustrates also shows that the model is not able to predict the values correctly which are far from the mean.

Figure 8: Experiment 1: MLP model results

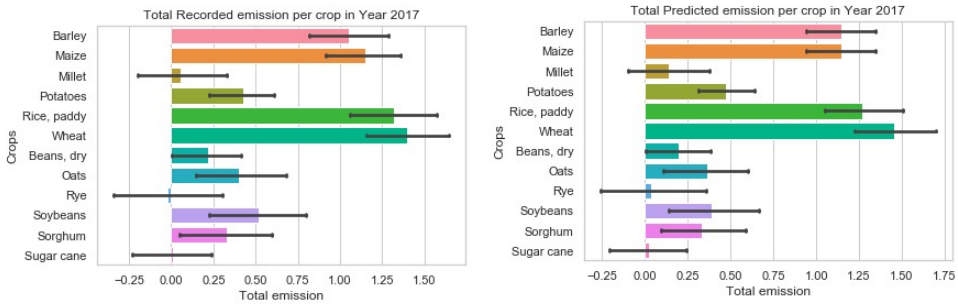


The left side of the figure shows the MLP model's predicted values and recorded values on validation dataset and the right side of the figure shows how well the model was able to explain the variance in validation dataset.

4.2 Carbon emission profile

The data of 189 countries for the year 2017 is fed to the RFR model to make predictions. It is worth highlighting that the model is neither trained nor testing on this subset of the data set and it is an unseen sample for the model. For simplicity of Experiment_1a setup is used for this task. Figure 9 show the total recorded and predicted emission per crop. The shown results are the aggregation of emission produced by the particular crop for all the countries present in the subset. The figure illustrates that wheat, paddy rice, and maize are the top three CO_2 emanating contributors. It is intriguing to notice that barley, wheat, paddy rice and soyabean are evidently different from the recorded emission values. For the comparison purpose, four major agro-based countries selected

Figure 9: Total CO_2 emission from residue management practice in year 2017



The left side of the figure shows the recorded CO_2 emission from residue management practices per crop for 189 countries in year 2017. Similarly, the right side of the figure shows the predicted CO_2 emission for the same year

are: China, Netherlands, and USA. Figure 10 shows the CO_2 emission profile per field

crop type of China for year 2017. The figure shows the actual as well as the predicted value of CO_2 emission. As China profile from the figure illustrates that major CO_2 emanating crops were maize, paddy rice and wheat. As per the predicted value, the maize was responsible for producing 16175.9 gigagram of CO_2, however the observed emission was 15920.8 gigagram. This difference is significant and points to the need of improvement in the prediction accuracy.

Figure 10: Total CO_2 emission per crop in China from residue management practice in year 2017

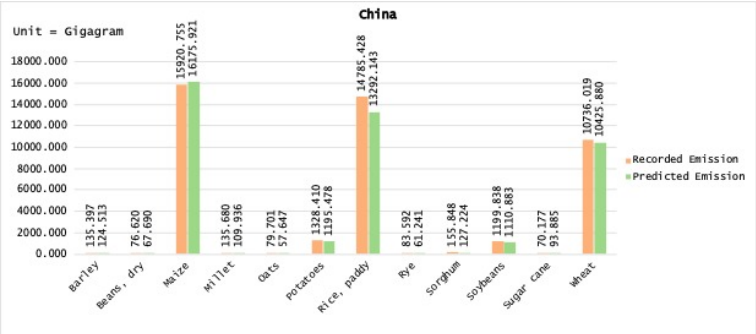
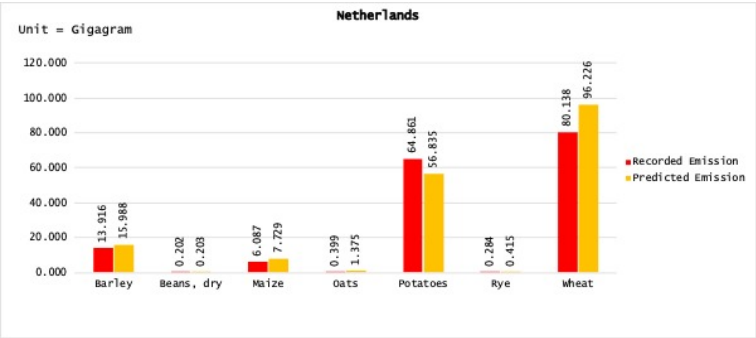


Figure 11 shows the CO_2 emission profile of Netherlands for year 2017. On contrary to China profile, Netherlands highest CO_2 producing crops were potatoes and wheat. This comparison highlights the reason why one framework do not fit all and to achieve carbon neutral farming, every country need to focus on different crops. Another interesting area to explore is that the emission caused by wheat in Netherland is considerably lower than the China. One of the obvious answer can be that China's area harvest (24510490 hectares) is larger than the Netherlands (115923 hectares). However, it is worth exploring if the difference is also due to dissimilar technology or residue management practice used. These comparative connections can be helpful to adopt or change the new technology or residue management practice.

Figure 11: Total CO_2 emission per crop in Netherlands from residue management practice in year 2017



5. Discussion

The MLR which was the baseline model, produced a considerable RMSE (0.662) on the standardized validation test set. MLP comparatively performed slightly worse than baseline with RMSE (0.69). One of the possible reasons may be that MLP requires more data to achieve better accuracy, as it is powered by high dimensional data. The result of this study shows that, RFR clearly performed better (R^2 =.968, RMSE=.032) as compared to the MLP (R^2 =.31, RMSE=.69) and the MLR (R^2 =.329, RMSE=.671). In addition to that, the RFR model does not show any signs of over-fitting. Comparison of the results of the models from Experiment_1a and Experiment_1b, suggest that the models perform better with standardized data (setup-2).

Although all of the three models were able to predict the CO_2 emission per crop-type per country, the current performance is not qualified yet for the practical implementation. Even though for the given task RFR performed very well, however it should be used cautiously as it tends to overfit the data. In addition to that, generally, RFR can work with limited data and gives good results. But after a certain limit, adding more samples do not improve the accuracy. In contrary, MLP benefits from the massive amount of data and can achieve improved accuracy with high dimensional data. MLP is also more robust to work with incomplete datasets (Radojević et al. 2013). However, if limited data is a constraint, MLP might not be the best choice. The MLR is more suitable for exploring the linear relationship of features and labels. The model is very sensitive to outliers. Another MLP underlying assumption is the absence of multicollinearity among the features, which is far away from reality. In general, the MLR is a good starting point to analyze the relationship within the data, however not quite suitable for complex practical applications, due to over-simplification of the problem by the underlying assumption about the data.

As extensively discussed in "Carbon emission profile " subsection of the "Results" section, the RFR model was selected to create CO_2 emission profile per crop type. This paper fills the knowledge gap of creating CO_2 emission profile per crop type using ML and ANN. It further refines the CO_2 emission contribution, caused by crop residye management practices per crop-type for a given country. This study can act as guidance research that can be extended and expanded. Extension in terms of more data can be used to capture the trends directly and indirectly related to farming, to make the predictions more reliable and expanding in respect of creating a complete GHG emission profile (CO_2, CH_4, and N_2O) per crop type. In addition to that, the models proposed in this research are robust and are capable of handling noisy data to a certain extent. The preferable model will depend on the goal to be achieved.

Three sections of society can directly benefit from the result of this research, farmers, government or policymakers, and the farming innovation industry. This research can be used as a baseline to build a system that can track CO_2 emission profile per crop type. It will be especially helpful for farmers, as they will know in order to reduce the carbon emission which crops should be the focal point. Once the focal point is known, direct and indirect solutions can be employed to achieve carbon-neutral farming such as precision farming, using biomass to produce electricity. With CO_2 emission information in hand, the farmers can decide the extent to which the emission reduction is possible. The farming-innovation industry can use this research to build a system that can provide a real-time carbon profile per crop, which will help the industry in researching new ways to solve the problem and coming up with the long term as well as spot-on solutions to reduce CO2 emission. The government can harness the benefits from this research as a guide to form a more efficient national level agricultural borne

smart and emission mitigation policy in terms of rebate, tax, investment, and subsidy to promote the reduction of CO2 footprint.

Similar to every research, this research also had a few limitations, which are discussed briefly below. First of all, due to extremely limited related literature and the absence of baseline from related work, makes it challenging to critically analyze the results of this research. Furthermore, the FAOSTAT emission data were computed at Tier 1 by following IPCC 2006 Guidelines for National GHG Inventories and their prediction accuracy is unknown (FAO 2016). This implies that the inaccuracies in the data computation were inherently included in the model as well and might have impacted the model performance. Additionally, it is recommended that future studies take into account of diverse features selection and feature engineering methods.

6. Conclusion

The results of this research can be perceived as a first effort to propose ML and ANN that are appropriate for estimating CO_2 emissions from residue management practices by considering production and census data. This proposed approach can also be replicated to create CH_4 and N_2O emission profile per crop-type per country. So that eventually, a complete GHG emission profile can be created per crop-type for a given country. The results of this research can benefit the society, where especially the farmers can benefit by providing them the CO_2 emission profile per crop so that emission management strategies can be applied. The scientific community can use the outcome of this research to further build a more accurate and robust model which not only captures the directly affecting factor but also the indirect factors. For the policymakers, the information and results from this research will help them to judge the best regulatory directives and policies to implement and control the CO2 emission.

References

Alston, Julian M and Philip G Pardey. 2014. Agriculture in the global economy. *Journal of Economic Perspectives*, 28(1):121–46.

Aye, Goodness C and Prosper Ebruvwiyo Edoja. 2017. Effect of economic growth on co2 emission in developing countries: Evidence from a dynamic panel threshold model. *Cogent Economics & Finance*, 5(1):1379239.

Balafoutis, Athanasios, Bert Beck, Spyros Fountas, Jurgen Vangeyte, Tamme Van der Wal, Iria Soto, Manuel Gómez-Barbero, Andrew Barnes, and Vera Eory. 2017. Precision agriculture technologies positively contributing to ghg emissions mitigation, farm productivity and economics. *Sustainability*, 9(8):1339.

Behrang, MA, E Assareh, MR Assari, and A Ghanbarzadeh. 2011. Using bees algorithm and artificial neural network to forecast world carbon dioxide emission. *Energy Sources, Part A: Recovery, Utilization, and Environmental Effects*, 33(19):1747–1759.

Boateng, Kofi K, George Y Obeng, and Ebenezer Mensah. 2017. Rice cultivation and greenhouse gas emissions: a review and conceptual framework with reference to ghana. *Agriculture*, 7(1):7.

FAO. 2016. Mainstreaming ecosystem services and biodiversity into agricultural production and management in east africa.

FAO, October. 2009. How to feed the world 2050 high-level expert forum. *Available at: fao. org/fileadmin/templates/wsfs/docs/Issues_papers/HLEF2050_Global_Agriculture. pdf [accessed on 10 November 2014]*.

Farag, AA, HA Radwan, MAA Abdrabbo, MAM Heggi, and BA McCarl. 2013. Carbon footprint for paddy rice production in egypt. *Nature and Science*, 11(12):36–45.

(Food, FAO and Agriculture Organization). 2009. Global agriculture towards 2050. In *High Level Expert Forum-How Feed World*, volume 2050, pages 1–4.

Ge, Mengpin and Johannes Friedrich. 2020. 4 charts explain greenhouse gas emissions by countries and sectors.

Hanson, AMGJ, MG Joel, A Joy, and J Francis. 2017. Plant leaf disease detection using deep learning and convolutional neural network. *International Journal of Engineering Science*, 5324.

Hassoun, Mohamad H et al. 1995. *Fundamentals of artificial neural networks*. MIT press.

Hunter, John D. 2007. Matplotlib: A 2d graphics environment. *Computing in science & engineering*, 9(3):90–95.

Karmokar, Bikash Chandra, Mohammad Samawat Ullah, Md Kibria Siddiquee, and Kazi Md Rokibul Alam. 2015. Tea leaf diseases recognition using neural network ensemble. *International Journal of Computer Applications*, 114(17).

Khoshnevisan, Benyamin, Shahin Rafiee, Mahmoud Omid, Marziye Yousefi, and Mehran Movahedi. 2013. Modeling of energy consumption and ghg (greenhouse gas) emissions in wheat production in esfahan province of iran using artificial neural networks. *Energy*, 52:333–338.

Lehner, Peter and Nathan A Rosenberg. 2017. Legal pathways to carbon-neutral agriculture. *Envtl. L. Rep. News & Analysis*, 47:10845.

Lewis, Scott M, Trenton A Jerde, Charidimos Tzagarakis, Pavlos Gourtzelidis, Maria-Alexandra Georgopoulos, Nikolaos Tsekos, Bagrat Amirikian, Seong-Gi Kim, Kâmil Uğurbil, and Apostolos P Georgopoulos. 2005. Logarithmic transformation for high-field bold fmri data. *Experimental brain research*, 165(4):447–453.

Lim, Youngil, Young-Sil Moon, and Tae-Wan Kim. 2007. Artificial neural network approach for prediction of ammonia emission from field-applied manure and relative significance assessment of ammonia emission factors. *European journal of agronomy*, 26(4):425–434.

Liu, Peng, Guoxing Zhang, Xutao Zhang, and Sujie Cheng. 2012. Carbon emissions modeling of china using neural network. In *2012 Fifth International Joint Conference on Computational Sciences and Optimization*, pages 679–682, IEEE.

McCarl, Bruce A and Uwe A Schneider. 2000. Us agriculture's role in a greenhouse gas emission mitigation world: An economic perspective. *Review of Agricultural economics*, 22(1):134–159.

McKinney, Wes et al. 2010. Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, volume 445, pages 51–56, Austin, TX.

McMahon, Jeff. 2019. Why agriculture's greenhouse gas emissions are almost always underestimated.

Meng, Ming and Dongxiao Niu. 2011. Modeling co2 emissions from fossil fuel combustion using the logistic equation. *Energy*, 36(5):3355–3359.

Milne, Eleanor, Henry Neufeldt, Todd Rosenstock, Mike Smalligan, Carlos Eduardo Cerri, Daniella Malin, Mark Easter, Martial Bernoux, Stephen Ogle, Felipe Casarim, et al. 2013. Methods for the quantification of ghg emissions at the landscape level for developing countries in smallholder contexts. *Environmental Research Letters*, 8(1):015019.

Murad, Wahid, Rafiqul Islam Molla, Mazlin Bin Mokhtar, and Abdur Raquib. 2010. Climate change and agricultural growth: an examination of the link in malaysia. *International Journal of Climate Change Strategies and Management*.

Olander, Lydia, Eva Wollenberg, Francesco Tubiello, and Martin Herold. 2013. Advancing agricultural greenhouse gas quantification. *Environmental Research Letters*, 8(1):011002.

Oliphant, Travis E. 2006. *A guide to NumPy*, volume 1. Trelgol Publishing USA.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830.

Pigott, Therese D. 2001. A review of methods for missing data. *Educational research and evaluation*, 7(4):353–383.

Qureshi, M Asghar and Shahid Afghan. 2005. Sugarcane cultivation in pakistan. *Sugar Book Pub. Pakistan Society of Sugar Technologist*.

Radojević, D, V Pocajt, I Popović, A Perić-Grujić, and M Ristić. 2013. Forecasting of greenhouse gas emissions in serbia using artificial neural networks. *Energy Sources, Part A: Recovery, Utilization, and Environmental Effects*, 35(8):733–740.

Richards, Meryl, Eva K Wollenberg, and S Buglion-Gluck. 2015. Agriculture's contribution to national emissions.

Rodrigues, Joaquim Augusto Pinto, Luiz Biondi Neto, Pedro Henrique Gouvêa Coelho, and João Carlos Correia Baptista Soares de Mello. 2009. Estimating greenhouse gas emissions using computational intelligence. In *ICEIS (2)*, pages 248–250.

Rogelj, Joeri, Michel den Elzen, Niklas Höhne, Taryn Fransen, Hanna Fekete, Harald Winkler, Roberto Schaeffer, Fu Sha, Keywan Riahi, Malte Meinshausen, and et al. 2016. Paris agreement climate proposals need a boost to keep warming well below 2°c.

Safa, Majeed, Mohammadali Nejat, Peter L Nuthall, and Bruce J Greig. 2016. Predicting co emissions from farm inputs in wheat production using artificial neural networks and linear regression models-case study in canterbury, new zealand.

Samoilov, IA and AI Nakhutin. 2009. Estimation and medium-term forecasting of anthropogenic carbon dioxide and methane emission in russia with statistical methods. *Russian Meteorology and Hydrology*, 34(6):348–353.

Scarlat, Nicolae, Milan Martinov, and Jean-François Dallemand. 2010. Assessment of the availability of agricultural crop residues in the european union: potential and limitations for bioenergy use. *Waste management*, 30(10):1889–1897.

Smith, Pete, Z Martino, and D Cai. 2007. 'agriculture', in climate change 2007: mitigation.

Tongwane, Mphethe, Thandile Mdlambuzi, Mokhele Moeletsi, Mitsuru Tsubo, Vuyo Mliswa, and Lunga Grootboom. 2016. Greenhouse gas emissions from different crop production and management practices in south africa. *Environmental Development*, 19:23–35.

Van Rossum, Guido. 2020. *The Python Library Reference, release 3.8.2*. Python Software Foundation.

Virtanen, Pauli, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, CJ Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake Vand erPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1. 0 Contributors. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*.

Waskom, Michael, Olga Botvinnik, Drew O'Kane, Paul Hobson, Saulius Lukauskas, David C Gemperline, Tom Augspurger, Yaroslav Halchenko, John B. Cole, Jordi Warmenhoven, Julian de Ruiter, Cameron Pye, Stephan Hoyer, Jake Vanderplas, Santi Villalba, Gero Kunter, Eric Quintero, Pete Bachant, Marcel Martin, Kyle Meyer, Alistair Miles, Yoav Ram, Tal Yarkoni, Mike Lee Williams, Constantine Evans, Clark Fitzgerald, Brian, Chris Fonnesbeck, Antony Lee, and Adel Qalieh. 2017. mwaskom/seaborn: v0.8.1 (september 2017).

Wollenberg, Eva, Meryl Richards, Pete Smith, Petr Havlík, Michael Obersteiner, Francesco N Tubiello, Martin Herold, Pierre Gerber, Sarah Carter, Andrew Reisinger, et al. 2016. Reducing emissions from agriculture to meet the 2 c target. *Global change biology*, 22(12):3859–3864.

Zhu, Nanyang, Xu Liu, Ziqian Liu, Kai Hu, Yingkuan Wang, Jinglu Tan, Min Huang, Qibing Zhu, Xunsheng Ji, Yongnian Jiang, et al. 2018. Deep learning for smart agriculture: Concepts, tools, applications, and opportunities. *International Journal of Agricultural and Biological Engineering*, 11(4):32–44.

Appendices

A. Figures

Figure 12: Density pair plot of the raw dataset

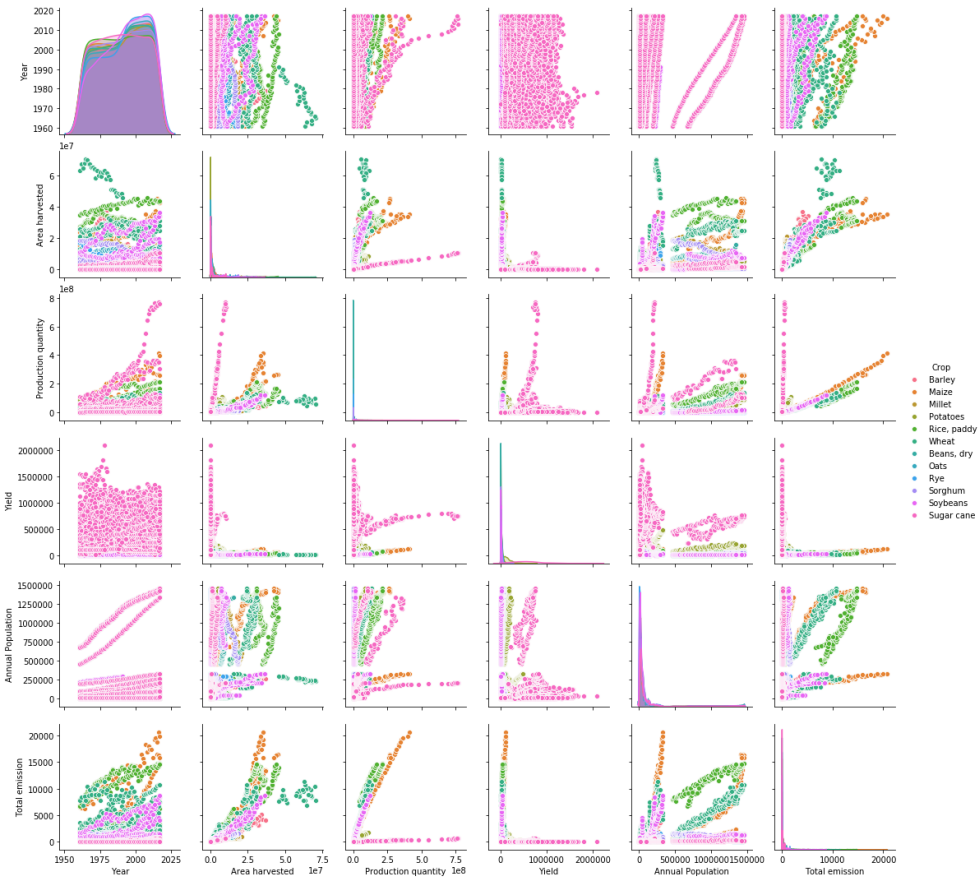
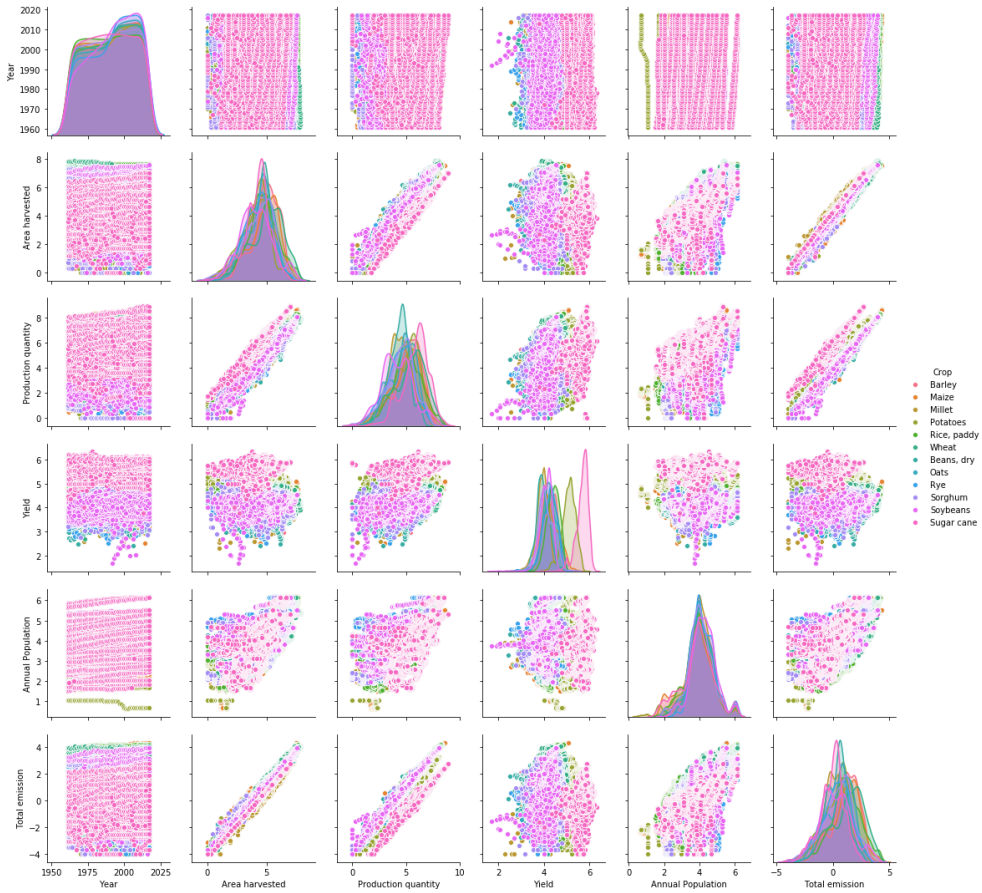


Figure 13: Density pair plot of Log10 Normalized data



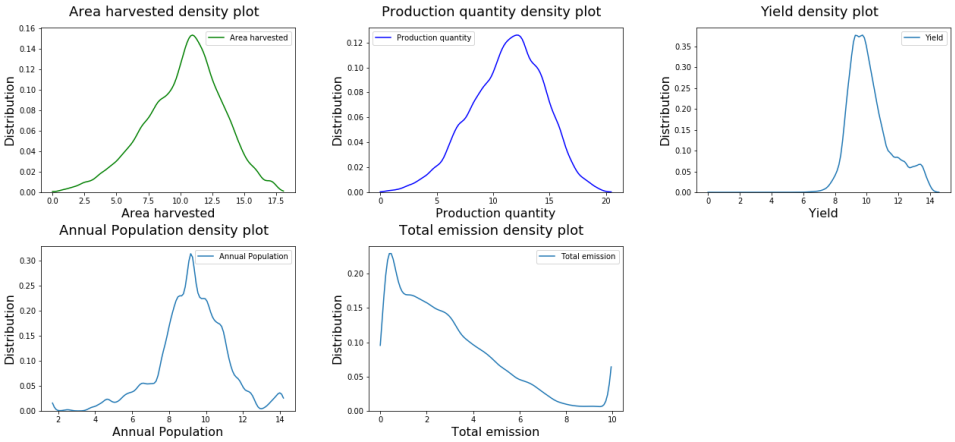
B. Tables

Table 5: Experiment 1 results

	Experiment_1a						Experiment_1b					
	MLR		RFR		MLP		MLR		RFR		MLP	
Evaluation metrics	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train
R-squared	0.329	0.338	0.970	0.995	0.33	0.332	0.329	0.338	0.968	0.995	0.31	0.317
MSE	1.121	1.114	0.24	0.09	1.124	1.118	0.819	0.814	0.18	0.067	0.831	0.826
RMSE	1.258	1.241	0.057	0.009	1.264	1.249	0.671	0.662	0.032	0.005	0.69	0.683

The table shows the results obtained from the models on the validation and training set in Experiment_1a and Experiment_1b.

Figure 14: Density plot of Normalized (Log1p) data



The figure shows the distribution of the dataset after applying Log1p normalization. To certain extent, the distribution became normalized for all features, except "Total emission"

Figure 15: Percentage of total emission per crop type in year 2017

