

# Data-driven Prediction of Crop Yield over Germany

A Thesis

Submitted to

Indian Institute of Science Education and Research Pune  
in partial fulfillment of the requirements for the  
BS-MS Dual Degree Programme



Submitted by:

Prantik Pramanick

Indian Institute of Science Education and Research, Pune

Under the supervision of:

Dr. Manmeet Singh (IITM, Pune)

Co-Supervisor: Dr. Amit Srivastava (The University of Bonn)

TAC Member: Dr. Amit Apte (IISER Pune)



# Certificate

This is to certify that this dissertation entitled **Data-driven Prediction of Crop Yield over Germany** towards the partial fulfilment of the BS-MS dual degree programme at the Indian Institute of Science Education and Research, Pune represents study/work carried out by Prantik Pramanick at Indian Institute of Tropical Metereology, Pune under the supervision of Dr. Manmeet Singh, Scientist, Indian Institute of Tropical Metereology, Pune, during the academic year 2022-2023.

*Manmeet Singh*

**Manmeet Singh**  
Centre for Climate Change  
Research  
IITM Pune

**Committee:**

**Supervisor:** Dr. Manmeet Singh (IITM Pune)

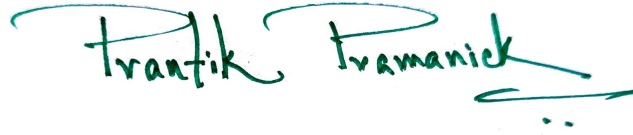
**Co-Supervisor:** Dr. Amit Srivastava (The University of Bonn)

**TAC Member:** Dr. Amit Apte (IISER Pune)



## Declaration

I hereby declare that the matter embodied in the report entitled **Data-driven Prediction of Crop Yield over Germany**, are the results of the work carried out by me at the Indian Institute of Tropical Metereology, Pune, under the supervision of Dr. Manmeet Singh (IITM, Pune), and the same has not been submitted elsewhere for any other degree.



Prantik Pramanick

This Thesis is Dedicated to  
My grandfather, late Paresh Nath Pramanick

## Academic Acknowledgements

March 2022, just after we returned to IISER after COVID-break, a heavy swing smashed the confidence of submitting the Thesis project by June 23. Respected Professor Amit Apte, in parallel of guiding me for semester project, connected me with Dr. Manmeet Singh. Out of all the options from Dr. Manmeet Singh, I got connected with Dr. Amit Srivastava from the University of Bonn. Together offered me this MS project.

I carry extreme gratitude for continuous inspiration and careful guidance.

As soon I started with the project work, a horrific shock was waiting for me from my family. "Bhai" (My Grandfather), identified with stage four liver cancer and was waiting to count his last breathe. Prof. Srivastava stood by me to support in those days and asked me to spend my days with my beloved one.

After almost two months, when I finally started working, Manmeet Sir offered me a cozy office room to work daily on the beautiful campus of IITM with 24×7 access to work without any disturbance.

Till now, including every moment I have received the extended help and support Manmeet Sir and Amit Srivastava sir, have provided every help I've asked for on this project.

I am extremely grateful to Prof Amit Apte for his insightful, constructive feedbacks continuous encouragement, who gave me the first opportunity for a semester project at IISER Pune, which helped me to stay motivated and focused throughout my project.

I would like to thank all my TAC teachers for their unwavering support and encouragement, which gave me the confidence to complete the work done so far. Their profound knowledge, guidance and mentorship have been instrumental in shaping my academic trajectory.

I would like to extend my heartfelt respect for some of the teachers from my school life, without whom, I would not have been successful as today. Their contribution and dedication helped in shaping me to what I am today. Firstly, my late Grandfather Mr. Paresh Nath Pramanick, who sailed me from my childhood in Mathematics and Science. Dipankar-Sir and Asit-Sir from my 11-12th class prepared me to get into this esteemed organization.

Certainly, I'm not able to mention all the names of my teachers. However I am sure their love, dedication and care sailed me here.



## Personal Acknowledgements

Hmm... writing a personal acknowledgment is something that I am extremely bad at. However, I'll try my best.

Firstly, of course, I thank my parents, Mr. Partha Pratim Pramanick and Mrs. Nandini Pramanick, for raising me into the man that I am today. I thank them for their unwavering support, encouragement, and love throughout my academic journey.

Other than them, I must mention my aunt, Mrs. Piyali Chatterjee, and my uncle, Anik Chatterjee, who played a big role in my admission and helped me in every way possible to settle down here. It has always been their house only, which has been my second home during any of the breaks in my academic calendar in the last five years (or should I say 3.5 years?)

I also thank all my other family members, whom I couldn't mention in this short page.

Now, we move on to hostel mates...

Being a really introverted person, my list of friends is small for me, so it is rather easy to write this part.

The first acknowledgment would, of course, be to my actual roommate, Lubdhak Mondal. From helping me write my first-ever semester project mail to helping me find online resources, being an excellent Google searcher as he is, there isn't probably a single thing in academics where this guy hasn't helped me in the five years of our IISER life.

And second will, of course, be my official roommate, Dipayan (Ishita) Pal (Physics major). From helping when I'm stuck in coding to finding small spelling mistakes, which probably even I wouldn't have looked for, this guy probably spent more time with my thesis in the last few days than he did with his own thesis.

I was on the verge of taking a semester extension when these two members of our Three Musketeers group helped me in every way that one can help their friend.

And finally, to my past roommate, Sahil Mulewar, I won't thank you for that vicious conspiracy you had against me last year!

Thanks to Deepesh Khushwani, who helped me a lot with LaTeX, as we say in the Bengali proverb, I didn't know the "L" of LaTeX.

On this point, many thanks to Sagnik-da for this wonderful thesis template. Without this, the thesis writing part probably would have taken double the time it did!

Finally, special mentions to Ayan-da, Saikat, Saket too for making my hostel life as beautiful as home. Thank you for the *sweet* memories.

# Abstract

This research aims to create a system that predicts agricultural yields using an all-encompassing system that integrates Numerical Weather Prediction (NWP) with machine learning. We want to know if combining NWP and ML models improves crop production forecasts over either NWP or ML models alone.

We collect historical crop yield data, weather and soil parameters data. Using the meteorological parameters and yield data, we will train an ML model to predict crop production and find out which model gives the best prediction. Finally, we will combine the ML and NWP models to improve forecast accuracy and reliability.

The system is assessed using multiple metrics. The results will show if the technology can predict crop yields for different crops and regions. For agricultural decision-making, the method may identify the most important meteorological and soil parameters that affect crop yields. Farmers, policymakers, and agricultural stakeholders may benefit from accurate crop output forecasts using the proposed method.

Ultimately, the goal is to contribute to NWP and ML model agricultural production prediction research. By combining these two methods, we can improve crop production estimates, helping farmers make better decisions and improve food security.

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Evolution of Crop Yield Prediction Techniques . . . . .	2
1.2	Numerical Weather Prediction(NWP) . . . . .	3
1.3	Creating an End-to-End Crop Yield Prediction Model . . . . .	4
<b>2</b>	<b>An Overview of the Existing Literature</b>	<b>7</b>
2.1	Winter Wheat Yield Prediction using CNN From Environmental and Phenological Data . . . . .	7
2.2	Predicting Corn and Soybean Yields Simultaneously using Deep Transfer Learning . . . . .	9
2.3	Revisiting Deep Learning Models for Tabular Data . . . . .	10
<b>3</b>	<b>Data Preparation</b>	<b>12</b>
3.1	Data Acquisition . . . . .	12
3.1.1	Weather Data[6] . . . . .	12
3.1.2	Soil Data[8] . . . . .	14
3.1.3	Crop Yield Data[10] . . . . .	15
3.2	Data Analysis . . . . .	15
3.2.1	Distribution of Weather Parameters . . . . .	17

3.2.2	Distribution of Soil Parameters . . . . .	19
3.2.3	Distribution of Crop Yield Values . . . . .	21
3.2.4	Correlation Between The Independent Parameters and The Yield Values . . . . .	25
3.3	Data Pre-processing . . . . .	31
<b>4</b>	<b>Methodology</b>	<b>33</b>
4.1	Model Overview . . . . .	34
4.1.1	Linear Regression . . . . .	34
4.1.2	Decission Tree Regressor . . . . .	35
4.1.3	Random Forest Regression . . . . .	36
4.1.4	Gradient Boosting Regressor . . . . .	37
4.1.5	Extreme Gradient Boosting(XGBoost) . . . . .	39
4.2	Model Evaluation: Assessing Prediction Accuracy . . . . .	40
<b>5</b>	<b>Results</b>	<b>42</b>
5.1	Results of Various Models . . . . .	42
5.1.1	Linear Regression . . . . .	43
5.1.2	Decision Tree Regressor . . . . .	47
5.1.3	Random Forest Regressor . . . . .	51
5.1.4	Gradient Boosting . . . . .	55
5.1.5	Extreme Gradient Boosting(XGBoost) . . . . .	59
5.2	Discussions . . . . .	64
<b>6</b>	<b>Further Work</b>	<b>68</b>



# List of Figures

3.1	Distribution of annual Precipitation(mm) and Minimum Temperature( $^{\circ}\text{C}$ ) . . . . .	17
3.2	Distribution of Minimum Temperature( $^{\circ}\text{C}$ ) and Radiation( $\text{KJ.mol}^{-1}$ )	17
3.3	Distribution of relative humidity(%) and Windspeed( $\text{ms}^{-1}$ ) . . . . .	18
3.4	Distribution of Wilting Point( $\text{cm}^3\text{cm}^{-3}$ ) and Field Capacity( $\text{cm}^3\text{cm}^{-3}$ )	19
3.5	Distribution of Saturation Point( $\text{cm}^3\text{cm}^{-3}$ ) and Bulk Density( $\text{g}^3\text{cm}^{-3}$ )	20
3.6	Distribution of Yield Values of Winter Barley and Spring Barley . .	21
3.7	Distribution of Yield Values of Winter Wheat and Oats . . . . .	22
3.8	Distribution of Yield Values of Sugarbeet and Silaze Maize . . . . .	22
3.9	Distribution of Yield Values of Winter Rape . . . . .	23
3.10	Correlation between different parameters and Winter Wheat yield values . . . . .	25
3.11	Correlation between different parameters and Winter Barley yield values . . . . .	26
3.12	Correlation between different parameters and Spring Barley yield values . . . . .	27
3.13	Correlation between different parameters and Oats yield values . . .	28
3.14	Correlation between different parameters and Sugarbeet yield values	29
3.15	Correlation between different parameters and Winter Rape yield values . . . . .	30

3.16	Correlation between different parameters and Silaze Maize yield values	31
5.1	Actual vs Predicted yield values for Winter Wheat by Linear Regression . . . . .	43
5.2	Actual vs Predicted yield values for Winter Barley by Linear Regression . . . . .	44
5.3	Actual vs Predicted yield values for Spring Barley by Linear Regression . . . . .	44
5.4	Actual vs Predicted yield values for Oats by Linear Regression . . .	45
5.5	Actual vs Predicted yield values for Sugarbeet by Linear Regression	45
5.6	Actual vs Predicted yield values for Winter Rape by Linear Regression	46
5.7	Actual vs Predicted yield values for Silaze Maize by Linear Regression	46
5.8	Actual vs Predicted yield values for Winter Wheat by Decision Tree Regressor . . . . .	47
5.9	Actual vs Predicted yield values for Winter Barley by Decision Tree Regressor . . . . .	48
5.10	Actual vs Predicted yield values for Spring Barley by Decision Tree Regressor . . . . .	48
5.11	Actual vs Predicted yield values for Oats by Decision Tree Regressor	49
5.12	Actual vs Predicted yield values for Sugarbeet by Decision Tree Regressor . . . . .	49
5.13	Actual vs Predicted yield values for Winter Rape by Decision Tree Regressor . . . . .	50
5.14	Actual vs Predicted yield values for Winter Rape by Decision Tree Regressor . . . . .	50
5.15	Actual vs Predicted yield values for Winter Wheat by Random Forest Regressor . . . . .	51
5.16	Actual vs Predicted yield values for Winter Barley by Random Forest Regressor . . . . .	52



5.17	Actual vs Predicted yield values for Spring Barley by Random Forest Regressor . . . . .	52
5.18	Actual vs Predicted yield values for Oats by Random Forest Regressor	53
5.19	Actual vs Predicted yield values for Sugarbeet by Random Forest Regressor . . . . .	53
5.20	Actual vs Predicted yield values for Winter Rape by Random Forest Regressor . . . . .	54
5.21	Actual vs Predicted yield values for Silaze Maize by Random Forest Regressor . . . . .	54
5.22	Actual vs Predicted yield values for Winter Wheat by Gradient Boosting . . . . .	55
5.23	Actual vs Predicted yield values for Winter Barley by Gradient Boosting . . . . .	56
5.24	Actual vs Predicted yield values for Spring Barley by Gradient Boosting . . . . .	56
5.25	Actual vs Predicted yield values for Oats by Gradient Boosting . .	57
5.26	Actual vs Predicted yield values for Sugarbeet Gradient Boosting . .	57
5.27	Actual vs Predicted yield values for Winter Rape by Gradient Boosting . . . . .	58
5.28	Actual vs Predicted yield values for Silaze Maize by Gradient Boosting	58
5.29	Actual vs Predicted yield values for Winter Wheat by XGBoost model	59
5.30	Actual vs Predicted yield values for Winter Barley by XGBoost model	60
5.31	Actual vs Predicted yield values for Spring Barley by XGBoost model	60
5.32	Actual vs Predicted yield values for Oats by XGBoost model . . . .	61
5.33	Actual vs Predicted yield values for Sugarbeet by XGBoost model .	61
5.34	Actual vs Predicted yield values for Winter Rape by XGBoost model	62
5.35	Actual vs Predicted yield values for Silaze Maize by XGBoost model	62

5.36	Comparison between the RMSE Percentage between actual and predicted yields by different models . . . . .	64
5.37	Comparison between the Percentage errors in the prediction of different models . . . . .	65
5.38	Comparison between the $R^2$ values for $y=x$ fit between actual and predicted yields by different models . . . . .	66

# List of Tables

3.1	Summary statistics of the independent variables in the study . . . .	16
3.2	Summary statistics of yield values of the crops. The unit of yield is tons.ha <sup>-1</sup> . . . . .	16
5.1	Accuracy of the Predictions by Linear Regression . . . . .	47
5.2	Accuracy of the Predictions by Decision Tree Regressor Model . . .	51
5.3	Accuracy of the Predictions by Decision Tree Regressor Model . . .	55
5.4	Accuracy of the Predictions by Decision Tree Regressor Model . . .	59
5.5	Accuracy of the Predictions by XGBoost Model . . . . .	63

# Chapter 1

## Introduction

Agriculture has been a seminal discovery for progress of humanity. It has what has led humanity to go from inconsistent animal hunting to consistent form of food gathering which enabled human beings to settle into civilisation, the basis of which forms the modern world as we know today.

It is of the utmost importance that the eight billion people who presently call Earth, home can continue to do so for an indefinite period of time.

The agricultural practises that humanity have perfected over the course of history are the cornerstone upon which our modern civilisation is built. Because there are currently seven billion people who call Earth their home, it is of the utmost importance that they are able to do so for an indefinite amount of time. Affluence, access to natural resources, respect for cultural traditions, and the application of sound farming practices are all factors that have contributed to the development of many new areas of research and technological advancements. It has had a significant impact on our culture as well as our way of life, and it is engrained in a wide variety of activities, ranging from religious observances to communal celebrations.

The agricultural business stands to gain much from this initiative on crop yield prediction utilizing machine learning algorithms and meteorological data. By precisely estimating crop yields, farmers can make better judgments about planting and harvesting dates, use of resources such as water and fertilizer, and market demand. This will raise efficiency and productivity, cut expenses, and improve profitability for farmers. Furthermore, reliable crop yield forecasts can assist governments and policymakers in determining food security and agricultural policy decisions. Overall, the effective implementation of this project has the potential to have a large positive influence on the agricultural sector, resulting in greater

food production and supply, increased sustainability, and economic growth.

## 1.1 Evolution of Crop Yield Prediction Techniques

The process of forecasting crop yields is an essential part of agriculture because it assists farmers in making decisions regarding when to plant, harvest, and sell their crops. The ability to make accurate predictions of yields may be of assistance to farmers in maximizing yields, reducing waste, and increasing earnings. The ability to accurately forecast crop yields has evolved over the course of history, progressing from relatively straightforward empirical methods to more intricate models that make use of recent advancements in data science and machine learning. In this article, we will investigate how predictions of crop production have evolved over the course of time.

In the past, crop yield forecasting relied on empirical methods, which took into account a variety of factors such as historical yield data, the state of the soil, prevailing weather patterns, and the assessment of industry professionals. Farmers would use straightforward models to determine yield based on factors such as the type of soil, temperature, and amount of rainfall. These models were frequently inaccurate because they were unable to take into account the complex relationships that exist between the many factors that influence crop yield. As a direct result of this, farmers frequently relied on methods that were inefficient, time-consuming, and trial-and-error based in order to increase their crop yields.

The prediction of crop production began to make significant strides forward in the 1970s after the introduction of computer modeling. Researchers created simulation models that used mathematical algorithms to forecast crop yields based on variables including weather, soil properties, and plant physiology. These models were used to create crop yield predictions. These models, which were more accurate than empirical techniques, allowed farmers to select their crops with greater knowledge, which resulted in increased productivity. To successfully apply these models, however, required a significant amount of knowledge and complexity on the part of the user.

The estimation of crop yield has made significant strides forward since the 1990s, coinciding with the advent of remote sensing technologies such as satellite photography. Scientists were able to collect a substantial amount of data on crop growth, soil moisture, and weather patterns with the help of remote sensing technology. After that, this information was put to use in the process of developing more accurate models for yield prediction. Researchers were also able to monitor crops in

real time thanks to remote sensing, which enabled farmers to respond quickly to changes in the weather as well as other factors that impact crop productivity.

Early in the 21st century, advances in machine learning and data science began to have a positive impact on the accuracy of crop yield predictions. Machine learning algorithms were used to analyze massive amounts of data in order to create predictive models, which were then used to identify trends that human experts would not have spotted. These models might take into account a wide range of information, like weather patterns, soil properties, plant genetics, and data from remote sensing, which would enable them to make more accurate projections of agricultural yields. In addition, farmers were able to modify their projections using machine learning algorithms to account for the specific aspects of their farming practices and the varieties of crops they grew.

Agricultural yield prediction is currently undergoing significant change as a result of the introduction of novel technologies such as blockchain and the internet of things (IoT). These technologies make it possible for farmers to collect and share data on crop growth, weather patterns, and other variables in real time, which enables the creation of yield prediction models that are more accurate. Additionally, thanks to blockchain technology, farmers can now exchange data in a way that is both secure and open. This fosters an atmosphere of confidence between farmers and the various other participants in the agricultural supply chain.

Finally, yield prediction in agriculture has seen significant progress over the course of time, transitioning from relatively straightforward empirical methods to more intricate models that make use of recent advancements in data science and machine learning. The advancement of agricultural output prediction has been fueled by technological advancements such as remote sensing, machine learning, and blockchain. Researchers and farmers can now anticipate yields with greater precision and individualization than ever before thanks to the real-time collection and analysis of massive amounts of information that is made possible by these technologies. It is possible that in the future, as technology improves, we will be able to anticipate further advancements in agricultural production prediction. These advancements will help farmers maximize their yields and earnings.

## 1.2 Numerical Weather Prediction(NWP)

Numerical Weather Prediction, also known as NWP, is a technique that is utilised in the realm of science to forecast forthcoming weather conditions by making use of mathematical models that are carried out on computers. In order to accurately

simulate the chemistry, physics, and motions that occur in the atmosphere, NWP models employ a wide variety of intricate procedures. These models are based on a set of initial conditions, which may take into consideration observations of the current weather or data from other sources. These conditions could also be taken into account by these models. These models are able to produce reliable forecasts of weather patterns for periods of time ranging from a few days out to many weeks out. These patterns are influenced by a variety of factors, including temperature, wind speed, barometric pressure, and rainfall.

The data that are used in NWP models come from a wide variety of sources, such as observations made on the ground, data from satellites, and data collected by a wide variety of meteorological equipment. Some of these sources include observations made on the ground, data from satellites, and data collected by various types of meteorological equipment. Because of this, the image that these models paint regarding the current status of the atmosphere is one that is not just accurate but also comprehensive. This is because of the fact that these models take into account all of the relevant factors. After that, these data are incorporated into intricate mathematical models that make use of physics, thermodynamics, and fluid dynamics in order to replicate the behaviour of the environment over the course of time. These models are developed in order to predict how the environment will behave in the future. The creation of these models is accomplished through an activity that is known as data mining.

The United States National Weather Service (NWS) has become such an important part of the field of meteorology that it is now considered an essential tool. It is presently used as a primary source of information for forecasting weather, emergency management, aviation, agriculture, and other industries that are influenced in some way by the weather.

## **1.3 Creating an End-to-End Crop Yield Prediction Model**

In recent years, satellite data have been widely available in a variety of geographical, temporal, and spectral resolutions. As a result, it has become feasible to estimate agricultural production on a wide range of scales and in a wide range of locations. Because of how easily accessible Earth observation (EO) data is, agricultural mapping on a large scale has become more efficient. With EO data, which offers a unique method for recording crop information across large areas with rapid updates, it is possible to generate maps of agricultural productivity and yield. These maps may then be shared with others. In spite of the need for

a high spatial resolution in order to make accurate yield predictions, the use of unmanned aerial vehicles (UAV) has been actively promoted for the purpose of data collection. It is not possible to correctly quantify the yield over large regions without the assistance of a significant crew, despite the fact that unmanned aerial vehicle platforms have demonstrated improved image-capturing capabilities. In addition, technological improvements have led to an increase in both the accuracy and accessibility of yield forecasts made with the use of machine learning and statistical techniques. Historically, methods such as random forests, linear regression, and ensemble analysis have been utilised in the process of estimating agricultural production. Deep learning methods, on the other hand, are now in the driver's seat when it comes to estimates of agricultural production. Recent studies have utilised multi-layer perceptrons to anticipate yield in wheat, maize, and strawberry crops. This was accomplished by combining data on observable phenotypic traits with data on environmental traits. Also, there is a developing corpus of research that integrates yield prediction from UAV photos with convolutional neural networks. This research is still in its early stages. Recent research has revealed that the accuracy of utilising deep neural networks to estimate agricultural yields has altered, and this is true independent of the data collection technique that was utilised.

The implementation of numerical weather prediction (NWP), remote sensing datasets such as satellites and radars, and powerful machine learning algorithms have resulted in the creation of the ideal combination for sustainable agriculture. This is the case because NWP allows for more accurate forecasting of weather conditions. During the course of this study, a significant number of direct and indirect agricultural datasets were analysed, the level of predictability that each of these datasets possessed was assessed, and NWP and AI were combined in order to generate agricultural forecasts. Artificial intelligence (AI) algorithms and data from the NWP will be utilised as part of the process of developing an end-to-end crop yield prediction model. These will be used to make predictions for a wide range of crop production factors, such as parameters relating to the weather, parameters relating to the soil, and other factors.

So, in short, the following steps are a rough outline of this work.

1. **Collecting data:** We first gather all available information, including past weather records, crop yields, and details on soil types, farming methods, and land usage. The ML model will be trained on this data, and its predictions will be checked against this data.
2. **Pre-Processing data:** This step involves cleaning, aggregating, and translating the data into an analysis-ready format. Features like scaling and normalization may be engineered in as part of this process.



3. **Feature selection:** Feature selection is choosing the most important meteorological factors (such as temperature, precipitation, humidity, and wind) that will have an impact on harvest success. Both statistical techniques and ML algorithms may be used for feature selection.
4. **Modeling:** A machine learning model to forecast crop production is created, using the information that we have collected. Any ML method that is deemed enough for the job at hand may serve as the basis for the ML model.
5. **Integration:** Now, we combine the ML and NWP models together. High-resolution weather predictions from the NWP model may be fed into the ML model. The ML model may then use this data to provide more precise yield projections.
6. **Cross Validation:** Cross-validation or hold-out validation should be used to assess the efficacy of the integrated model. This will aid in figuring out how well the model predicts reality.
7. **Refinement:** Finally, based on the evaluation's findings, we iteratively tweak the model's parameters, feature selection, and data preparation methods to achieve optimal performance.

In conclusion, observe and compare prediction results generated by different Machine Learning models. The final goal is to enhance agricultural production prediction by combining NWP and ML models and capitalizing on their respective capabilities. The complicated correlations between weather factors and crop production may be learned by ML models, and NWP models can produce high-resolution weather predictions.

# Chapter 2

## An Overview of the Existing Literature

### 2.1 Winter Wheat Yield Prediction using CNN From Environmental and Phenological Data

In recent years, there has been a growing interest in making predictions about agricultural yields by using models that utilize machine learning. It is critical to have an accurate prediction of yield in order to maximize the effectiveness of crop management practices, enhance food security, and reduce the negative effects of climate change. In this paper[1], the authors have used convolutional neural networks (CNNs) to predict winter wheat yield from environmental and phenological data.

The previous research that has been done on the application of machine learning models to the forecasting of crop yield is first discussed in this paper. The authors highlight the limitations of traditional statistical models, which are unable to capture the complex non-linear interactions between the various environmental and agronomic factors that affect crop growth and yield. These models assume linear relationships between the predictor variables and the yield, and they highlight the fact that these linear relationships are assumed to exist. Previous studies have primarily concentrated on predicting crop yield based on remotely sensed data, such as satellite images, while relatively little attention has been paid to using on-site environmental and phenological data. For example, the authors note that previous research has primarily focused on predicting crop yield based on remotely sensed data.

The authors suggest the use of convolutional neural networks (CNNs), a type of deep learning model that is particularly well-suited for the analysis of spatial data such as images, in order to circumvent these limitations. The authors contend that CNNs are capable of learning complex characteristics from environmental and phenological data that are predictive of crop growth and yield. Previous research[2] has demonstrated that CNNs can perform better than traditional models of machine learning when it comes to tasks such as object recognition and image classification.

The dataset that was utilized in the research, which was comprised of three years' worth (2012-2014) of on-site environmental and phenological data collected from winter wheat fields located in Germany. The information gathered consists of a variety of variables, including temperature, precipitation, plant height, and soil moisture, in addition to details regarding the development stage of the crop. The dataset is not very large, consisting of only 63 samples, which makes it difficult to train and evaluate the CNN model.

The structure of the CNN model that was utilized in the research. Following the three convolutional layers are the three fully connected layers. The total number of layers in the model is nine. The model is not particularly complicated, containing a total of only 161,394 parameters; as a result, it is less likely to suffer from the problem of overfitting to the limited dataset. The training and evaluation procedures are also described. These procedures include dividing the dataset into training and testing sets and utilizing mean absolute error (MAE) and coefficient of determination (R squared) as performance metrics.

The CNN model achieved an MAE of 0.24 and an R-squared of 0.60 on the testing set. This indicates that the model is able to accurately predict winter wheat yield based on environmental and phenological data. In addition to this, a sensitivity analysis is carried out in order to determine which variables are most crucial for the yield prediction. According to the findings of the analysis, temperature, precipitation, and plant height are the three most important variables, which is in line with the findings of previous studies on the prediction of winter wheat yield.

CNNs have the potential to be a useful tool for predicting the yield of winter wheat based on phenological and environmental data. The limited size of the dataset presents a challenge for training and evaluating the model, and they suggest that future research should concentrate on collecting larger datasets in order to further improve the accuracy of the model. Using CNNs for crop yield prediction opens up new opportunities for integrating data from different sources, such as remote sensing and on-site data, in order to improve the accuracy of the predictions. This is something that can be done to improve the accuracy of the predictions.

The research emphasizes the significance of gathering data of a high quality and employing sophisticated machine learning algorithms in order to draw conclusions from the data. According to the findings of the study, convolutional neural networks (CNNs) have the potential to outperform conventional machine learning models for the task of predicting crop yields, in particular when analyzing spatial data such as images. The small size of the dataset, which was used in the study, is a limitation, as it may not generalize well to other geographical locations or types of crops. To evaluate the generalizability and scalability of the proposed approach, therefore, the focus of future research should be on recreating the study using datasets that are both larger and more diverse.

## **2.2 Predicting Corn and Soybean Yields Simultaneously using Deep Transfer Learning**

A novel approach for predicting the yield of both corn and soybeans simultaneously using remote sensing data and deep transfer learning is put forth in the paper[3] Accurate yield prediction may assist farmers and agricultural managers in optimizing crop output, increasing resource efficiency, and increasing revenues, making this a significant issue in the area of agriculture.

The first section of the report is a review of earlier studies that used remote sensing data to estimate crop yields. Other techniques, including support vector regression, neural networks, and linear regression, have been employed in earlier research to forecast agricultural yields using remote sensing data. Yet, since they do not adequately account for the intricate connections between environmental variables and crop development, these techniques often suffer from low accuracy and generalizability.

A novel approach is suggested based on deep transfer learning that starts with a pre-trained convolutional neural network (CNN) and trains a fresh CNN for yield prediction in order to overcome these drawbacks. This method is superior to typical machine learning techniques in a number of ways, including improved performance on small datasets, quicker training periods, and the capacity to use previously trained networks' knowledge.

The data utilized in the research, which included yield data gathered from corn and soybean fields in Illinois, USA, and remote sensing data[4] (in the form of NDVI, or Normalized Difference Vegetation Index), are then described. The information was gathered from 2010 to 2017 throughout a number of growth seasons.

The authors describe how they trained their CNN model for the growing simultaneous prediction of maize and soybean yields using a transfer learning methodology. In particular, they started their yield prediction CNN using a pre-trained CNN that was first trained on the ImageNet dataset for image classification. They then adjusted the model using yield information from the corn and soybean fields as well as remote sensing data.

According to the scientists, their model was successful in correctly predicting maize and soybean yields, with correlation values of 0.84 and 0.83, respectively. Moreover, they compare their model to other machine learning methods, such as support vector regression and random forests, and discover that in terms of prediction accuracy, their deep transfer learning method beats these models.

The authors wrap off by going through some possible uses of their approach for farmers and agricultural management. They contend that the capacity to forecast agricultural yields using data from remote sensing may assist guide choices about irrigation, fertilization, and pest control, thereby enhancing farmers' productivity and profitability.

## 2.3 Revisiting Deep Learning Models for Tabular Data

In-depth analysis of various deep learning models for tabular data analysis is provided in Gorishniy et al. [5] a comprehensive review paper. A critical evaluation of several deep learning techniques, including feedforward neural networks, convolutional neural networks, recurrent neural networks, and attention-based models, is provided. The pros and cons of each model are discussed in detail and information is given about how they work and how they are built. Additionally, the problems that arise when training these models on tabular data, such as data preprocessing, feature engineering, and tuning hyperparameters, are talked about.

A thorough literature review of recent research on deep learning models for tabular data analysis is provided in the paper. A number of promising uses for these models, such as credit risk prediction, medical diagnosis, and image recognition, are discussed. Furthermore, the performance metrics like accuracy, precision, recall, and F1 score that are used to judge how accurate and useful these models are, are talked about.

One of the strongest points of this paper is the thorough discussion of the difficulties involved in training deep learning models on tabular data. The significance

of data pre-processing and feature engineering is stressed by the authors in order to get the best performance out of these models. Additionally, the difficulties of tuning hyperparameters are discussed and advice is given on how to choose the right hyperparameters for different deep learning models.

Overall, this is an informative paper that can help researchers and practitioners who want to use deep learning to analyze tabular data. The paper is a good resource for understanding the current state of the art in this field because of its thorough literature review and critical analysis of different deep learning techniques.

# Chapter 3

## Data Preparation

In this chapter, we discuss the data acquisition process and describe the datasets we shall be working on.

### 3.1 Data Acquisition

In this project, we have the data of yield value, weather, and soil characteristics from 1999 to 2019 across all 271 counties in Germany. The parameter sets are as follows:

#### 3.1.1 Weather Data[\[6\]](#)

For 21 years, the Deutscher Wetterdienst (DWD) supplied daily information on wind speed, radiation, precipitation, and temperature (minimum and maximum) (spanning 1999 to 2019). The approach by Zhao et al[\[2\]](#) was used to interpolate this daily data to a 1 km grid scale, and weekly values at the level of NUTS3 (Nomenclature of Territorial Units for Statistics) were then aggregated for use as input in ML models. For the purpose of aggregating to the NUTS3 level, the agricultural land-use ratio for each grid cell was utilized to produce area-weighted average values. The calculations were made using 250 m26-resolution land-use data from CORINE Land Cover 2006[\[7\]](#). We used meteorological parameters such as wind speed, maximum and lowest temperatures, relative humidity, precipitation, and solar radiation in our analysis.

We averaged and aggregated weekly feature values to reduce the sample size of the daily weather data. It is thought that the level of precision and granularity included in everyday data would make it difficult to make discoveries. By focusing on weekly data, we were able to considerably decrease the number of model parameters (from 365 to 52). Daily meteorological data is often preprocessed and downsampled to a weekly level for use in yield prediction studies.

- **Precipitation-n:** This number represents the total quantity of rain that fell during the n-th week of the year. For instance, *Precipitation-4* denotes the total amount of precipitation during the fourth week of the year.

Sufficient precipitation is vital for plant growth since it provides the water required for crop growth and development. Inadequate precipitation can cause water stress, limiting plant development, yield, and quality.

- **Temp-min-n:** The n-th week's lowest temperature for the given year is represented by this value. The term *Temp-min-4* refers to the year's fourth week's lowest temperature.
- **Temp-max-n:** This is the highest temperature that was recorded during the n-th week of the given year. The term *Temp-max-4* refers to the year's fourth-week maximum temperature.

Temperature influences crop development by influencing the rate of photosynthesis, respiration, and other physiological activities. While excessive temperature circumstances can cause heat or cold stress, decreasing crop development and yield, optimal temperature conditions stimulate plant growth and yield.

- **Radiation-n:** This stands for the radiation during the year's n-th week. *Radiation-4* signifies radiation that occurred during the fourth week of the year, for instance.

Solar radiation supplies the energy required for photosynthesis, making it critical for crop growth and development. Inadequate or excessive radiation levels can impact plant development, yield, and quality.

- **RelHumCalc-n:** Use this tool to determine the relative humidity for the nth week of the given year. For instance, *RelHumCalc-4* depicts the relative humidity during the fourth week of the year.

The rate of transpiration, or the process by which plants lose water through their leaves, is affected by relative humidity. High relative humidity can decrease transpiration rates, causing waterlogging and other problems, whereas low relative humidity can increase transpiration rates, causing water stress.



- **Windspeed-n:** This is the typical wind speed for the n-th week of the specified year. As an example, *windspeed-4* represents the wind speed during the fourth week of the year.

Wind has the potential to influence crop development and yield through influencing transpiration rates, modifying the microclimate, and causing physical damage to the plants. High winds can cause water stress, whilst low wind speeds can cause air stagnation around the plants, limiting their growth.

### 3.1.2 Soil Data[8]

Based on the main kinds of soil, soil data were generated and compiled at the DWD grid level (soil categories corresponding to the agricultural land-use categories as per CORINE Land Cover 2006)[7]. The source of the soil data is a soil reconnaissance map of Germany at a scale of 1,000,000 that distinguishes between BÜK1000N (BGR) land usage[9] The parameters of the soil are:

- **Wilting Point(LL):** The wilting point is the lower limit of accessible soil water, at which point plants can no longer draw water from the soil. When a plant doesn't have access to water, it reaches the point of permanent wilting. The volumetric (%) crop available water at the permanent wilting Point is represented by Predictor LL in the dataset.

Water stress, diminished plant development, and yield can result from soil water that is at or below the wilting point.

- **Field Capacity(DUL):** The amount of soil moisture or water content that remains in the soil after excess water has been drained and the rate of downward movement has been greatly slowed down is known as field capacity. The dataset's predictor values represent the volumetric (%) crop available water at the Field Capacity.

Soil water content at field capacity offers necessary moisture to plants without causing waterlogging or soil saturation.

- **Saturation Point(SAT):** The saturation point (SAT) represents the soil's maximum water-holding capacity, which occurs when all pore spaces in the soil are filled with water. Therefore, soil water is no longer available to plants at saturation, and excess water can cause waterlogging, restricted oxygen availability, and other problems. The dataset's predictor values represent the volumetric (%) crop available water at the saturation point.

- **Bulk Density(BD):** Soil compaction is measured by bulk density (BD). It is calculated by dividing the soil's dry weight by volume. The dataset's bulk density estimates are available up to 1.3 meters of soil depth.

High bulk density can restrict root growth and water infiltration, resulting in water stress and decreased plant growth and yield.

To summarize, these soil factors are important in determining the amount of available soil water and nutrients for plant growth, as well as soil physical features that affect plant root growth and nutrient uptake. Knowing and regulating these characteristics can assist optimize soil water and nutrient availability, resulting in increased crop growth and yield.

### 3.1.3 Crop Yield Data[10]

In Germany, winter wheat crop production data for 271 counties at the subnational NUTS328 level were examined from 1999 to 2019. For the research, NUTS3 agricultural yield statistics were extracted from the regional database of Germany.

The crops, whose yield values are available are: *Winter Wheat*, *Winter Barley*, *Spring Barley*, *Oats*, *Sugarbeet*, *Winter Rape*, *Silage Maize*.

## 3.2 Data Analysis

We combined all of the yield, weather, and soil parameter data into a single data file, which will serve as our primary data file for the activities that will come next.

The integration of weekly meteorological, and soil variables over 271 counties in Germany from 1999 to 2019 resulted in a total of 5692 cases and 274 column characteristics.

Variable	count	mean	std	min	25%	50%	75%	max
Precipitation (mm)	5691	14.3	3.07	6.79	12.06	14.06	16.21	29.33
TempMin (°C)	5691	4.44	1.1	1.46	3.62	4.41	5.18	8.29
TempMax (°C)	5691	12.99	1.18	9.45	12.13	12.88	13.73	17.12
Radiation (KJm <sup>-2</sup> )	5691	10630.14	684.02	9082.78	10118.25	10548.04	11096.27	13140.65
RelHum (%)	5691	0.73	0.02	0.68	0.71	0.73	0.74	0.79
Windspeed (ms <sup>-1</sup> )	5691	2.56	0.1	2.35	2.51	2.58	2.61	2.72
LL (cm <sup>3</sup> cm <sup>-3</sup> )	5691	0.14	0.02	0.1	0.14	0.15	0.15	0.28
DUL (cm <sup>3</sup> cm <sup>-3</sup> )	5691	0.27	0.03	0.16	0.26	0.28	0.28	0.42
SAT (cm <sup>3</sup> cm <sup>-3</sup> )	5691	0.43	0.02	0.38	0.42	0.43	0.43	0.53
BD (g.cm <sup>-3</sup> )	5691	2.18	0.64	0.49	1.73	2.16	2.5	4.39

Table 3.1: Summary statistics of the independent variables in the study

Crop Name	Wwheat	Wbarley	Sbarley	oats	sugarbeet	Wrape	Smaize
Total Number of locations	271	271	271	271	271	271	271
Year Range	1999-2019	1999-2019	1999-2019	1999-2019	1999-2019	1999-2019	1999-2019
Count	5509	5371	4970	4734	4184	5210	5163
Mean Yield	5.55	4.2	4.01	15.841	3.09	31.05	32.46
Standard Deviation of Yield	1.001	0.79	0.792	2.955	0.688	5.55	2.728
Minimum Yield	0.53	0.91	0.66	3.78	0.92	2.34	28.19
Maximum Yield	10.04	7.84	7.39	26.50	14.79	58.10	38.91

Table 3.2: Summary statistics of yield values of the crops. The unit of yield is tons.ha<sup>-1</sup>

The above table describes the various sataistics of the yield parameters and yield values.

Next, the distributions plot are presented which consists of a histogram of the distribution overlaid with a kernel density estimate (KDE) curve.

### 3.2.1 Distribution of Weather Parameters

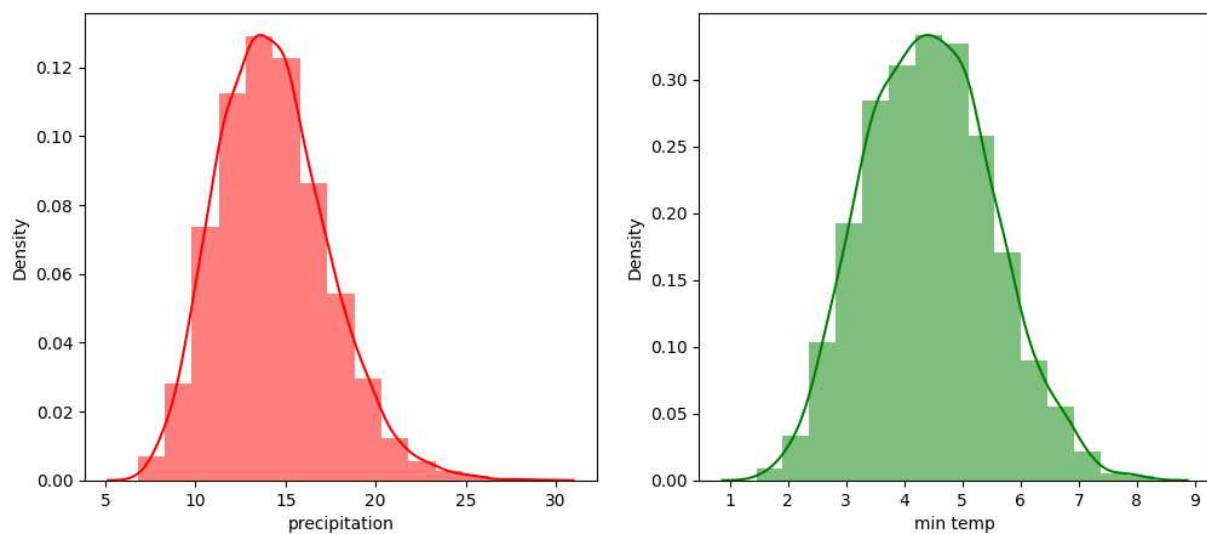


Figure 3.1: Distribution of annual Precipitation(mm) and Minimum Temperature(°C)

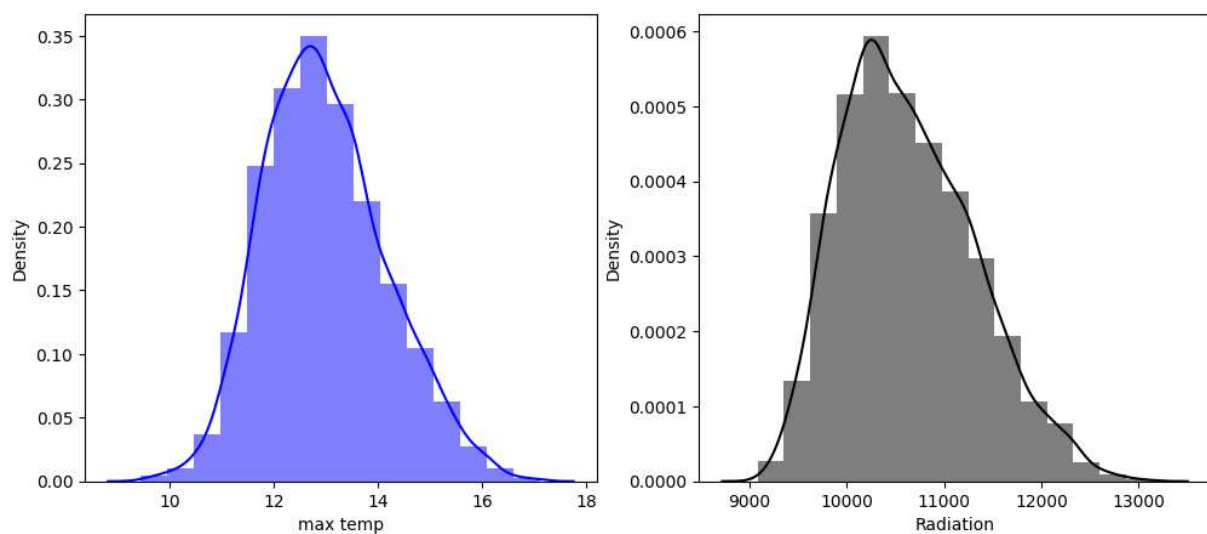


Figure 3.2: Distribution of Minimum Temperature(°C) and Radiation(KJ.mol<sup>-1</sup>)

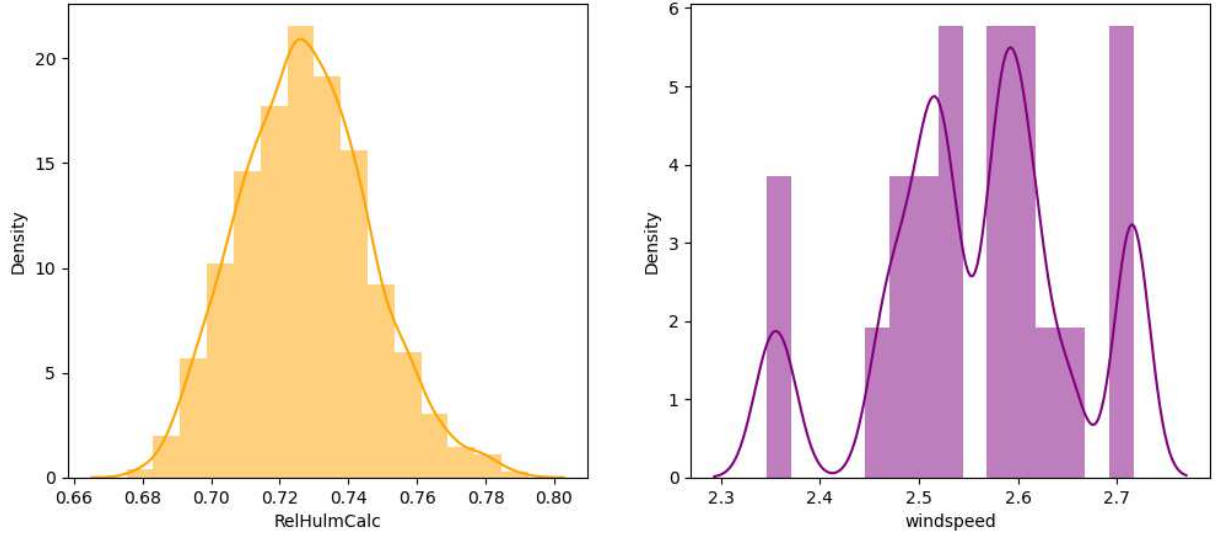


Figure 3.3: Distribution of relative humidity(%) and Windspeed(ms<sup>-1</sup>)

- **Precipitation:** The way the rainfall values are spread out shows that most observations are between 12 and 16 mm, which could mean that this is the best range for crop growth. If the amount of rain falls below this range, crops may experience water stress, which will lower yield. On the other hand, if precipitation levels are too high, crops may be damaged by waterlogging or flooding, which can also lower yield.
- **Minimum Temperature:** Temperature is an important factor for crop growth because it affects how quickly plants grow, make food, and take in nutrients. Most of the observations for the lowest temperatures fall between 4.41 and 5.18 °C, which could mean that this is the best temperature range for crop growth. If the minimum temperatures fall below this range, crops could be damaged by frost, which would result in a lower yield. On the other hand, if the minimum temperatures are too high, crops may experience heat stress, which can also lower yield.
- **maximum Temperature:** Most of the observations for the highest temperatures fall between 12.88 and 13.73°C, which could mean that this is the best temperature range for crop growth. If the highest temperatures fall below this range, crops may grow more slowly, have less photosynthesis, and take in fewer nutrients, which can result in a lower yield. On the other hand, if the maximum temperatures are too high, crops may experience heat stress, which can also lower yield.
- **Radiation:** With a mean of 10630.14K Jm<sup>-2</sup>, the radiation level is quite high. This suggests that there is enough solar energy for crop photosynthesis

and growth. However, it's important to remember that too much radiation can also cause heat stress and damage to crops.

- **Relative Humidity:** The relative humidity (expressed by RelHum) in this dataset has a mean of 0.73, indicating that the air is typically moist. While some humidity is required for crop growth, excessive humidity increases the danger of fungal diseases and pest infestations. It is crucial to note, however, that the effect of relative humidity on agricultural yield varies depending on the crop and stage of growth. Some crops, for example, may require higher humidity levels during specific stages of development, whilst others may be more susceptible to disease in high humidity circumstances.
- **Windspeed:** With a mean of  $2.56ms^{-1}$ , the wind speed is relatively low. While some wind is necessary for crop pollination and disease prevention, too much wind can hurt crops physically. So, the relatively low wind speed seen in this dataset may help crops grow.

### 3.2.2 Distribution of Soil Parameters

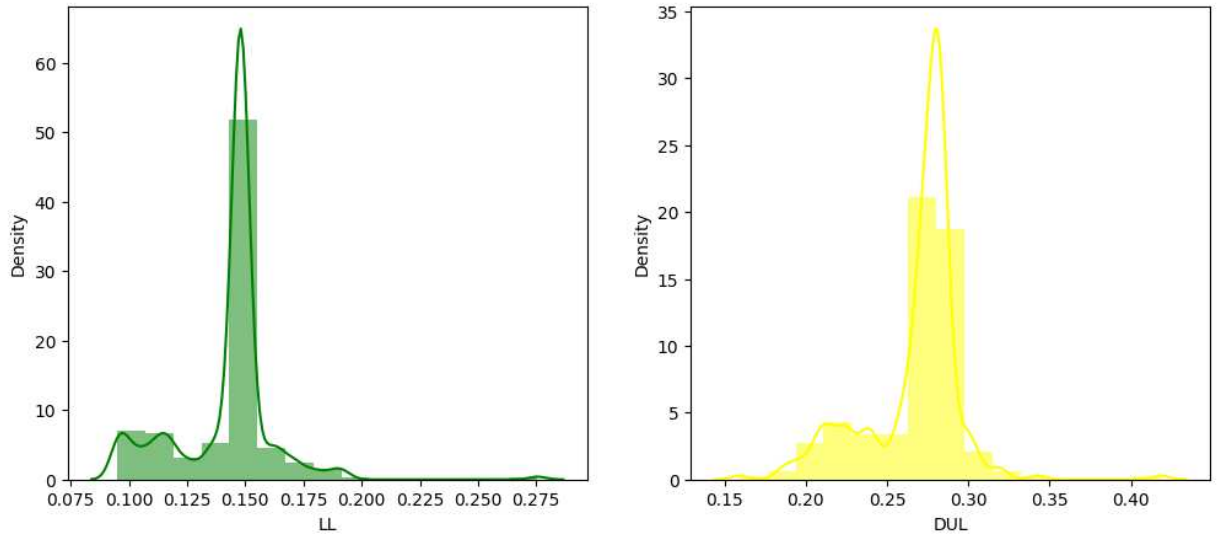


Figure 3.4: Distribution of Wilting Point( $cm^3cm^{-3}$ ) and Field Capacity( $cm^3cm^{-3}$ )

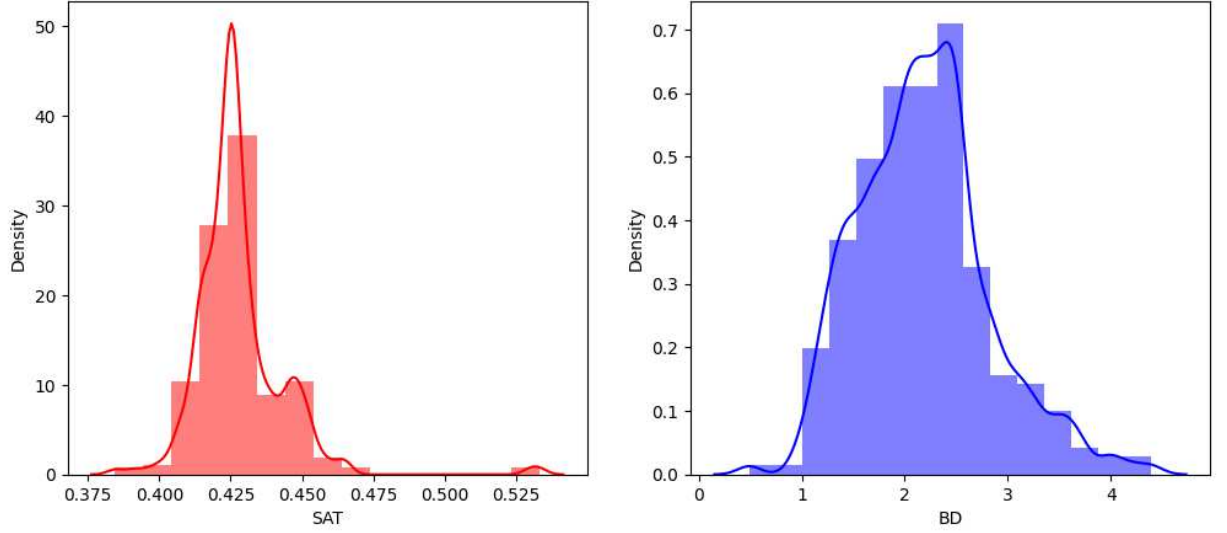


Figure 3.5: Distribution of Saturation Point( $\text{cm}^3 \text{cm}^{-3}$ ) and Bulk Density( $\text{g}^3 \text{cm}^{-3}$ )

Observing the plots, we can comment that:

- **LL:** The distribution of LL values in this dataset has a mean of 0.14 and a relatively narrow range from 0.1 to 0.28. This indicates that the soils tend to have a relatively low water-holding capacity at the permanent wilting point. Soils with low LL values can be more prone to drought stress and may require more frequent irrigation or rainfall to support crop growth. This can impact crop yield by reducing the amount of water available to plants and ultimately limiting their growth and productivity.
- **DUL:** The distribution of DUL values in this dataset has a mean of 0.27 and a range from 0.16 to 0.42. Soils with higher DUL values have a greater water-holding capacity at field capacity, which can be beneficial for crops. However, if the soil has poor drainage, excess water can accumulate and limit the amount of oxygen available to plant roots, which can also impact crop yield. Additionally, soils with high DUL values may be more prone to leaching of nutrients and can require more frequent fertilizer applications to maintain optimal soil fertility.
- **SAT:** The distribution of SAT values in this dataset has a mean of 0.43 and a relatively narrow range from 0.38 to 0.53. Soils with high SAT values have a high water-holding capacity at saturation, which can be beneficial for crops during periods of drought. However, excessive saturation can also lead to waterlogging and reduced oxygen availability to plant roots, which can

negatively impact crop growth and yield. Additionally, soils with high SAT values can be more prone to the leaching of nutrients and may require more frequent fertilizer applications to maintain optimal soil fertility.

- **BD:** The distribution of BD values in this dataset has a mean of 2.18 and a relatively wide range from 0.49 to 4.39. Soils with high BD values tend to be more compacted, which can limit root growth and reduce the availability of water and nutrients to plants. This can negatively impact crop yield by reducing the overall productivity of the plants. Additionally, soils with high BD values can be more prone to erosion and can require additional soil management practices to maintain soil health and fertility.

### 3.2.3 Distribution of Crop Yield Values

1

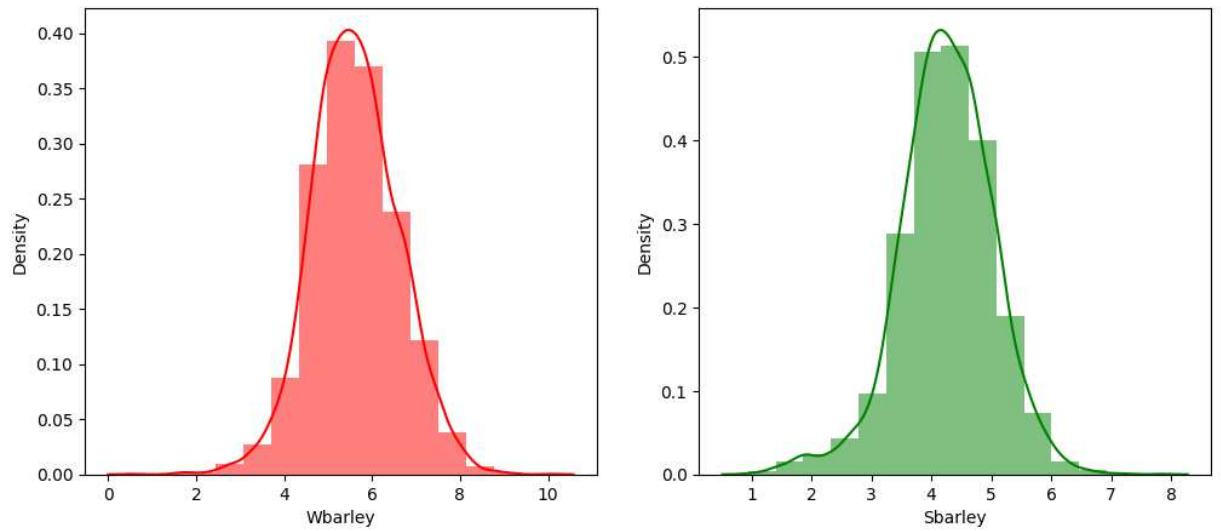


Figure 3.6: Distribution of Yield Values of Winter Barley and Spring Barley

---

<sup>1</sup>All Yield units are in tons.ha<sup>-1</sup>



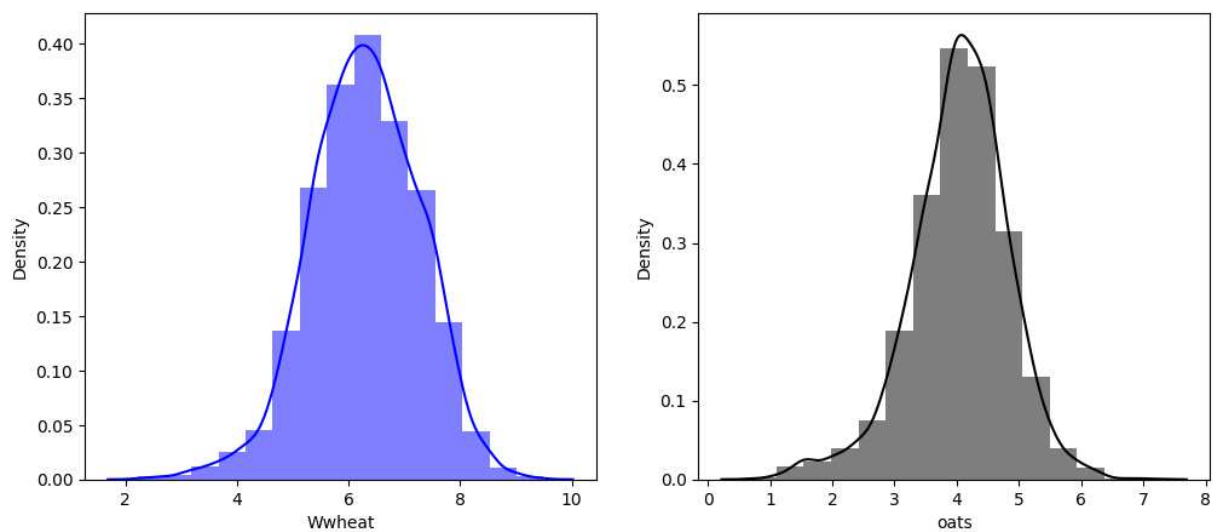


Figure 3.7: Distribution of Yield Values of Winter Wheat and Oats

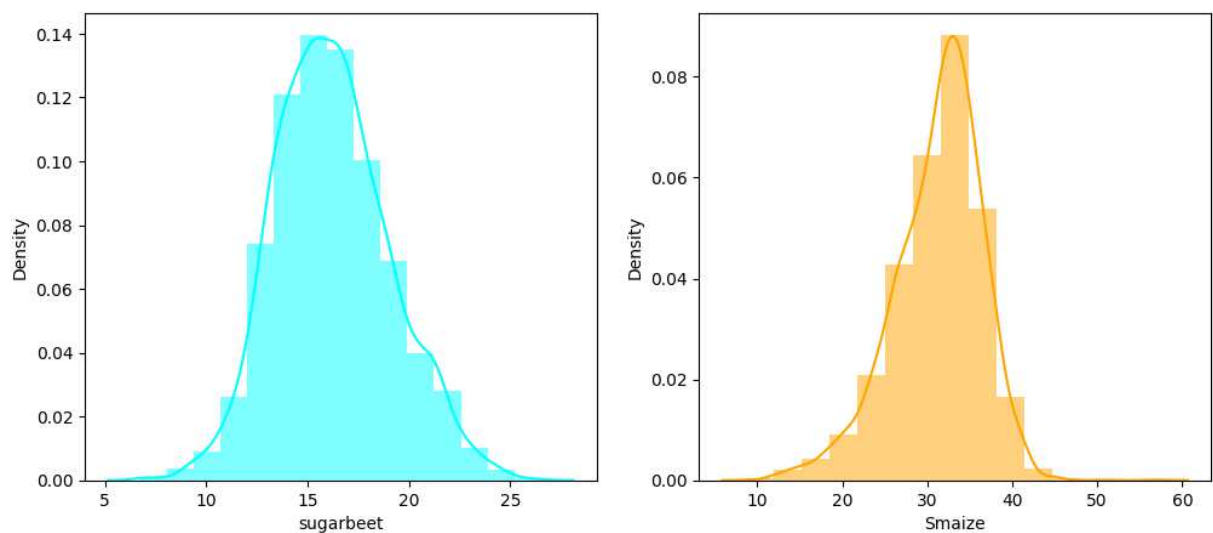


Figure 3.8: Distribution of Yield Values of Sugarbeet and Silaze Maize

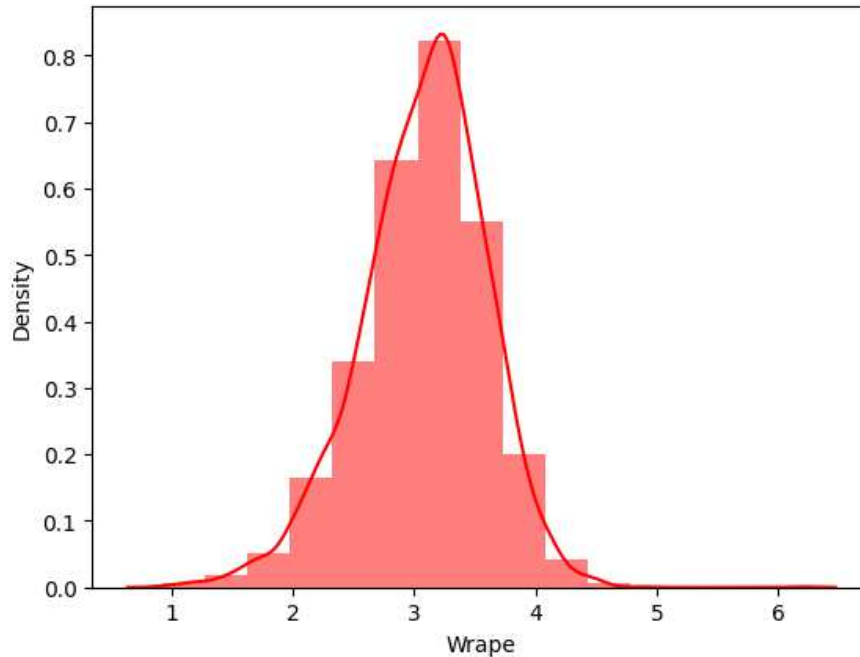


Figure 3.9: Distribution of Yield Values of Winter Rape

After looking at the distributions of the yield values, we can roughly interpret that:

- **Winter Wheat:** The yield distribution for Winter Wheat appears to be relatively normal with a narrow spread, which suggests that most of the yields are concentrated around the mean. This could indicate that wheat yields are relatively consistent across the locations and years in which they were observed.
- **Spring Barley:** The yield distribution for winter barley is also relatively normal, with a slightly wider spread than that of wheat. This suggests that the yields for winter barley are slightly more variable than those of wheat.
- **Spring Barley:** The yield distribution for spring barley is similar to that of winter barley, with a slightly wider spread than that of wheat. This suggests that the yields for spring barley are also slightly more variable than those of wheat.
- **Oats:** The yield distribution for oats is skewed to the right, which suggests that there are a few locations or years in which the yields were much higher than the mean. This could be due to favorable weather or soil conditions, or improved agricultural practices.

- **Sugarbeet:** The yield distribution for sugarbeet is also skewed to the right, which suggests that there are a few locations or years in which the yields were much higher than the mean. This could be due to factors such as improved pest management practices or more favorable weather conditions.
- **Winter Rape:** The yield distribution for rape is also skewed to the right, which suggests that there are a few locations or years in which the yields were much higher than the mean. This could be due to factors such as more favorable weather conditions, improved pest management practices, or the use of higher-yielding varieties.
- **Silaze Maize:** The yield distribution for Silaze Maize is relatively normal, with a narrow spread similar to that of wheat. This suggests that maize yields are relatively consistent across the locations and years in which they were observed.

### 3.2.4 Correlation Between The Independent Parameters and The Yield Values



Figure 3.10: Correlation between different parameters and Winter Wheat yield values

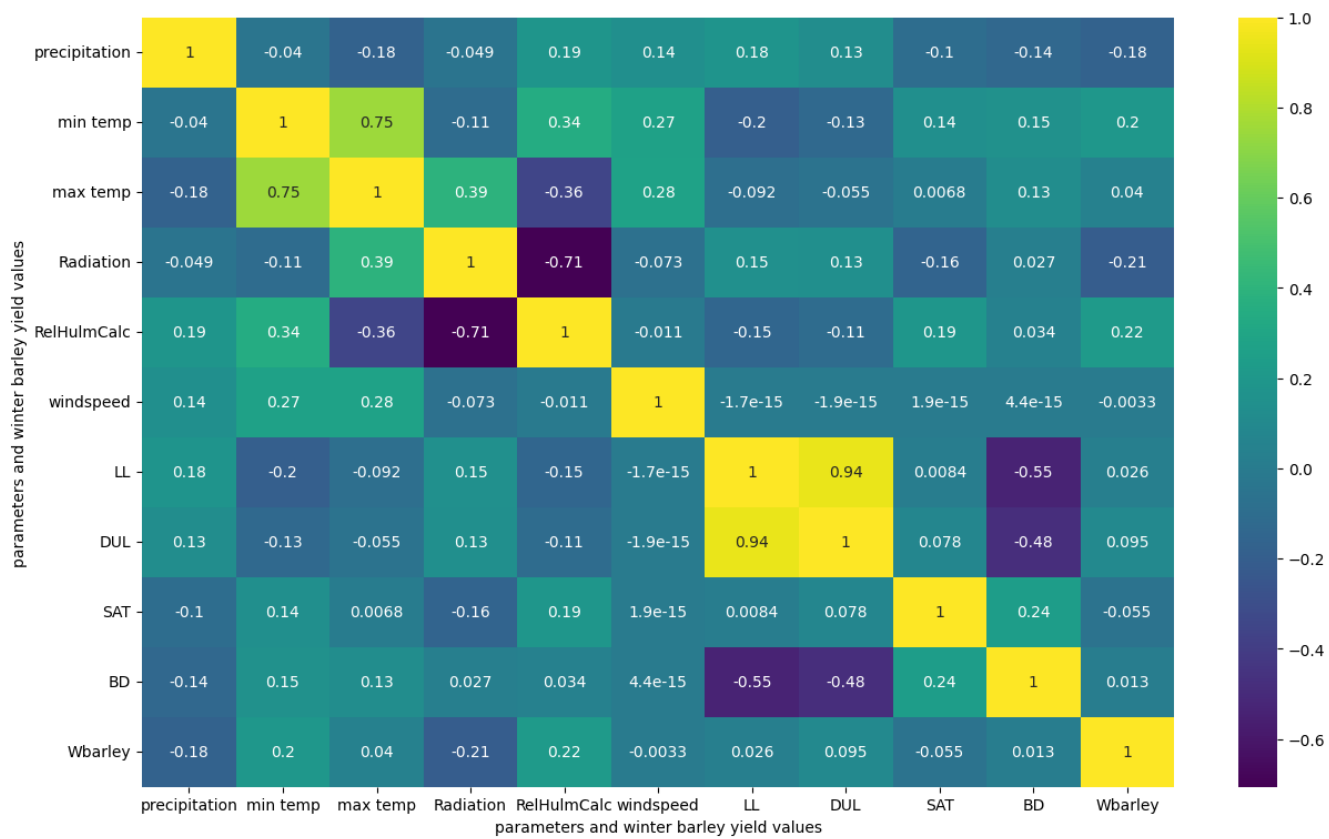


Figure 3.11: Correlation between different parameters and Winter Barley yield values

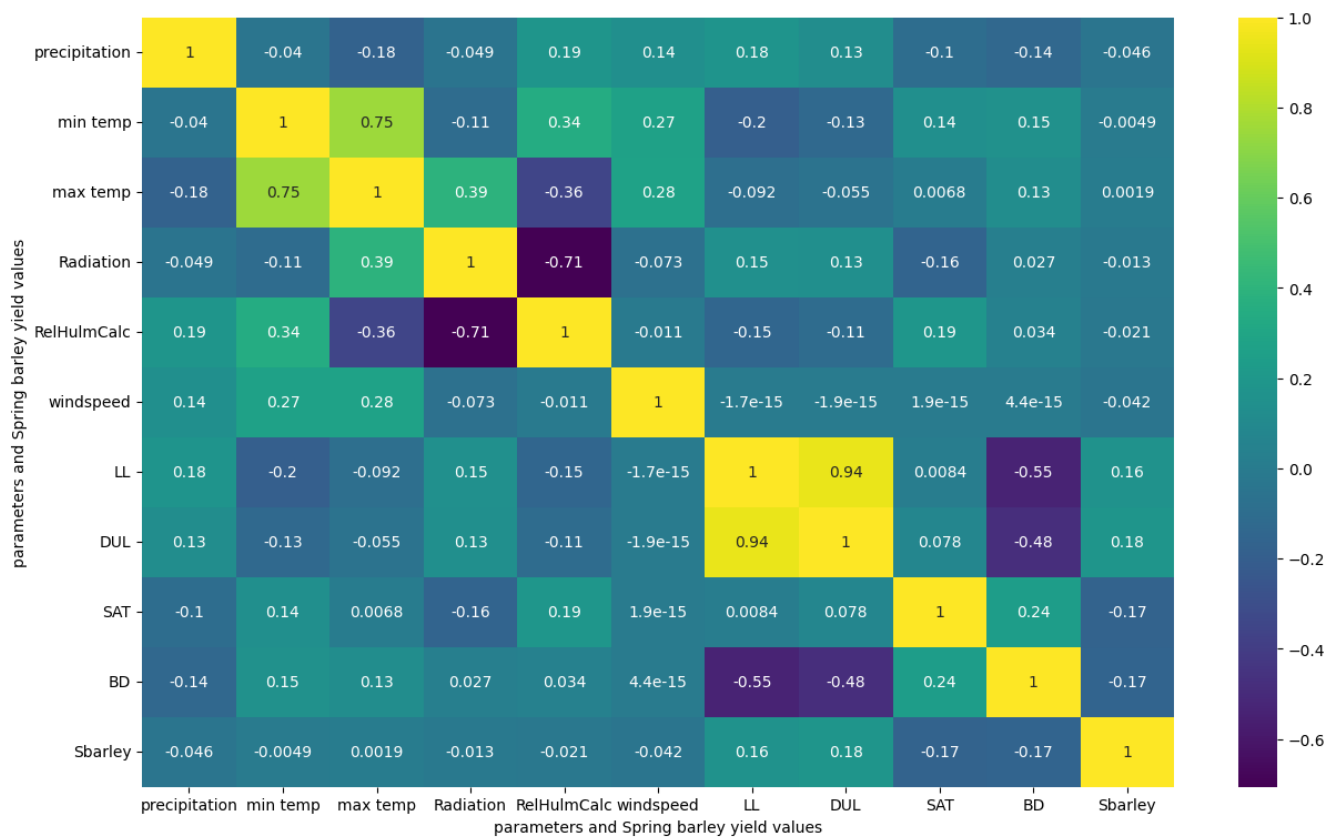


Figure 3.12: Correlation between different parameters and Spring Barley yield values

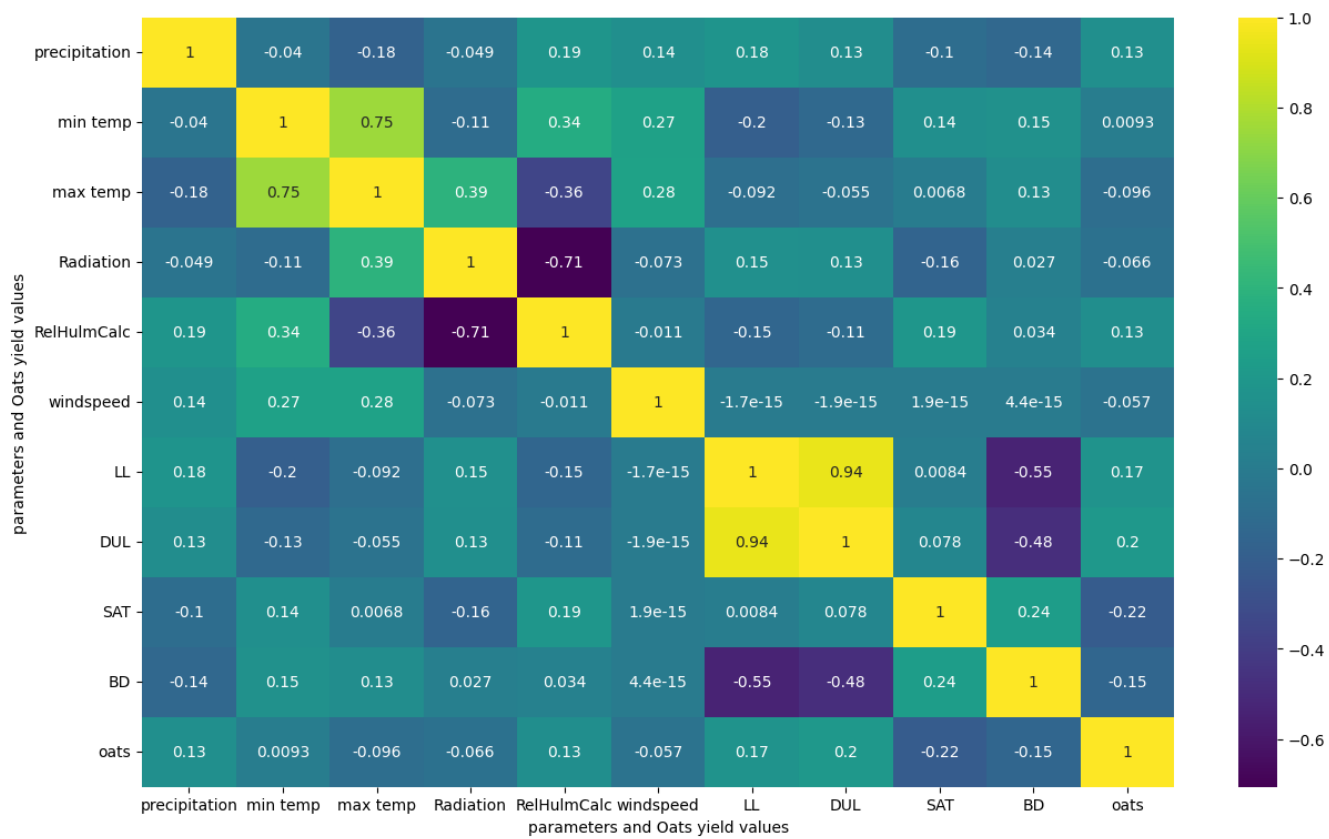


Figure 3.13: Correlation between different parameters and Oats yield values

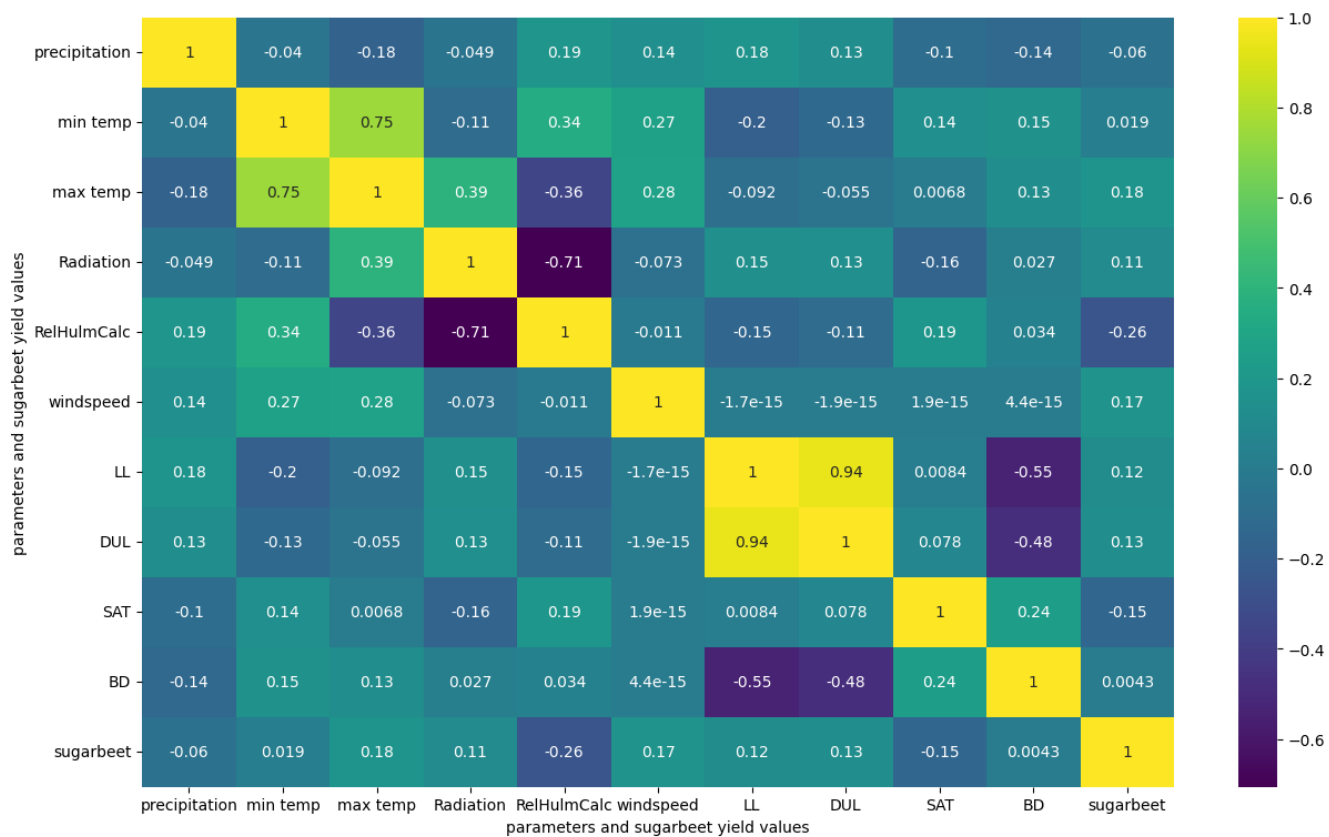


Figure 3.14: Correlation between different parameters and Sugarbeet yield values



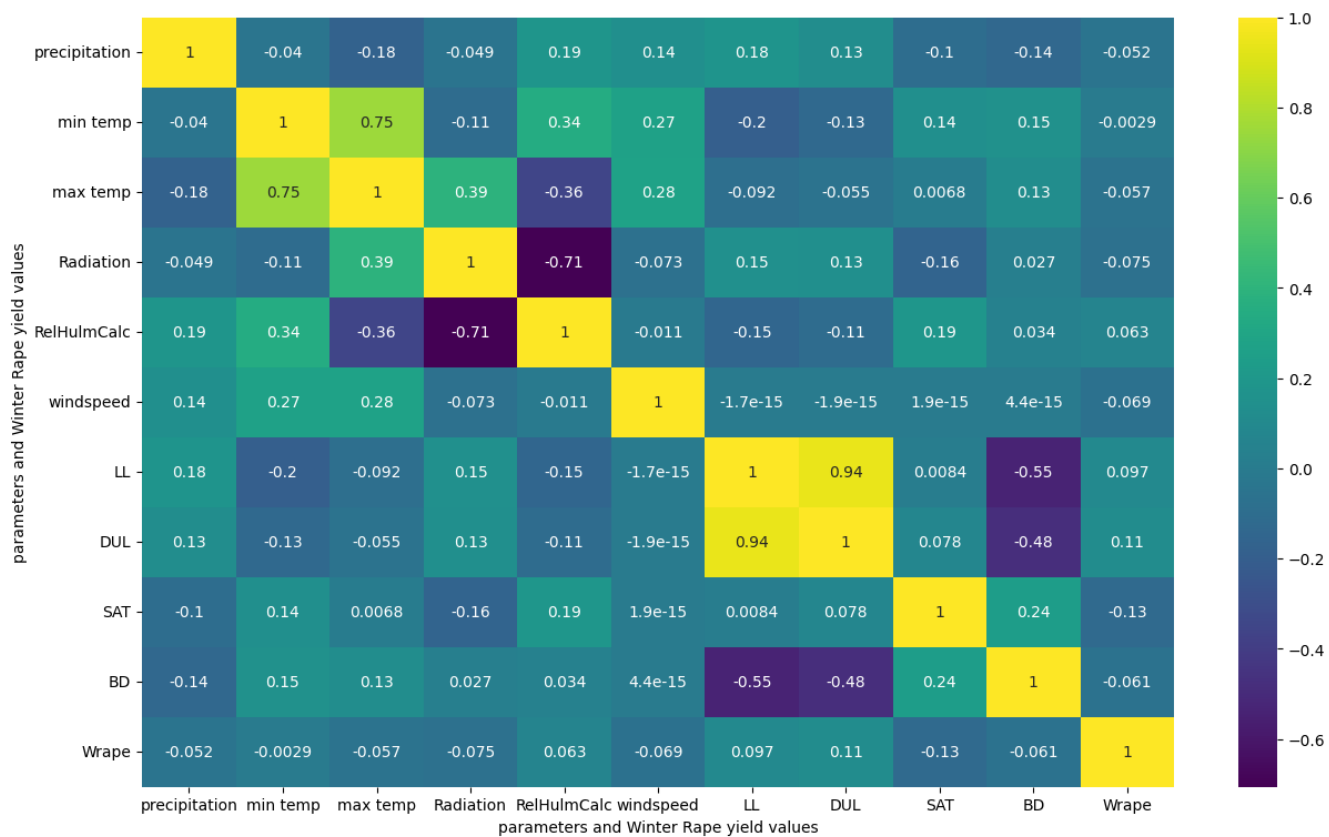


Figure 3.15: Correlation between different parameters and Winter Rape yield values

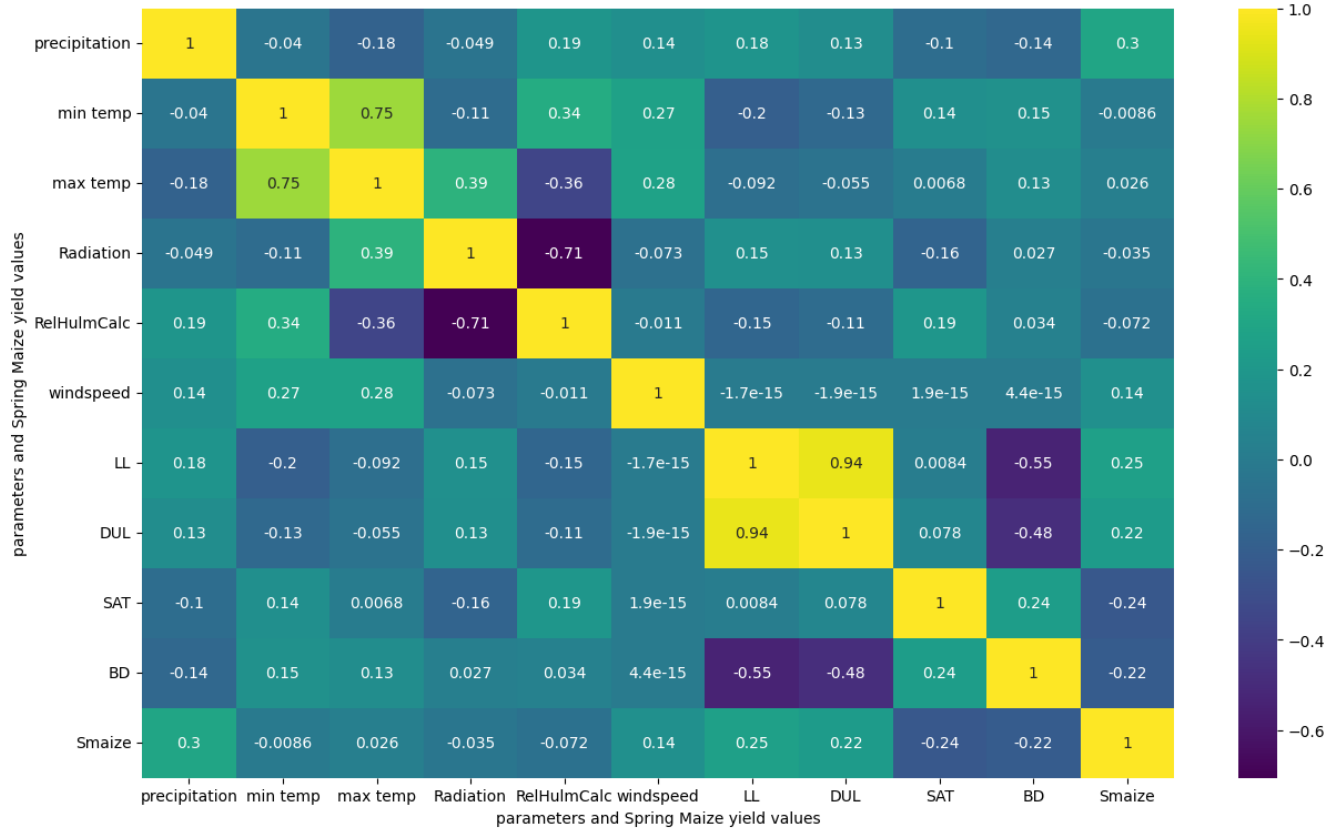


Figure 3.16: Correlation between different parameters and Silage Maize yield values

### 3.3 Data Pre-processing

At first, we remove the categorial columns such as the name of the Kreises, etc. Then we also remove the columns which are not needed for our analysis. These unrequired columns are 'Years', 'code kreise' etc. (This was already done for the heatmaps shown above, but it is part of data pre-processing so it's being mentioned here).

Then we divide the dataset into seven sub-datasets; one for each crop. So, each of the sub-datasets consists of the yield values of that particular crop and all the parameters of interest.

The next step is to remove the unavailable empty cells of the dataframe columns<sup>2</sup> because if there are missing values, the correlation matrix may be incomplete

<sup>2</sup>Pandas dropna function of python has been used

or biased, and may not provide an accurate representation of the relationships between variables.

For training the Machine Learning models, We have split the dataset into a training-testing ratio of 80:20, which means that 80% of the data is used for training the model and 20% of the data is used for testing the model.

# Chapter 4

## Methodology

NWP models provide weather predictions by simulating the behavior of the atmosphere using physics-based models. These models provide massive amounts of data that may be utilized to forecast crop production. Unfortunately, this data may be noisy and difficult to deal with, making precise crop output estimates problematic.

Machine learning may be used to preprocess and evaluate NWP data as well as to create models that estimate crop production based on meteorological data. To discover the complicated correlations between weather patterns and agricultural yields, machine learning techniques such as decision trees, random forests, and neural networks may be trained on historical crop yield data and matching NWP data.

A crop yield prediction model that takes into consideration the intricate relationships between weather patterns and crop development may be constructed by integrating NWP with machine learning.

## 4.1 Model Overview

In this section, we provide an overview of the machine learning models that have been used.

### 4.1.1 Linear Regression

Linear regression is a well-liked and extensively used statistical technique for simulating the connection between a dependent variable (often represented by  $y$ ) and one or more independent variables (usually denoted as  $x$ ). Finding the straight line (or hyperplane, in the case of several independent variables) that best reflects the connection between the variables is the aim of linear regression.

The Linear Regression model can be used to predict crop yield based on the given weather and soil parameters in the context of your training dataset, where  $x$  is the weather and soil parameters and  $y$  is the crop yield value.

The following is the general equation for a Linear Regression model with one predictor variable:

$$y = \beta_0 + \beta_1 x \quad (4.1)$$

where  $\beta_0$  and  $\beta_1$  are the intercept and slope coefficients, respectively. These coefficients are estimated using the training data to fit a line that best describes the relationship between the predictor variable ( $x$ ) and the response variable ( $y$ ).

The goal of Linear Regression is to reduce the sum of squared errors between the predicted and actual values of  $y$ . This is accomplished by minimizing the cost function:

$$J(\beta_0, \beta_1) = \frac{1}{2m} \sum_{i=1}^m (h_{\beta}(x^{(i)}) - y^{(i)})^2 \quad (4.2)$$

where  $h_{\beta}(x)$  is the predicted value of  $y$  given  $x$  and the estimated coefficients  $\beta_0$  and  $\beta_1$ , and  $m$  is the number of training examples.

Linear regression makes the assumption that the relationship between the dependent variable and the independent variable(s) is linear. Linear regression may

not be the optimal model for the data if this assumption is not true. Outliers and multicollinearity may also be sensitive to linear regression (when two or more independent variables are highly correlated).

For modeling the connection between variables and producing predictions, linear regression is a simple yet effective technique. Finance, economics, social sciences, and engineering are just a few of the numerous domains where it is commonly employed.

### 4.1.2 Decision Tree Regressor

The Decision Tree Regressor model of scikit learn works by recursively partitioning the input space into regions that are associated with different output labels. The tree model makes a binary decision at each internal node by comparing the value of a feature to a threshold. The leaves of the tree correspond to the output labels.

The training algorithm for the Decision Tree Regressor model can be summarized as follows:

1. Finding the best feature  $j$  and best threshold  $t_j$  to splitting the training data into two subsets,  $S_{left}$  and  $S_{right}$ , by minimizing the sum of squared errors (SSE) of the target values in each subset:

$$SSE_{split} = \sum_{i \in S_{left}} (y_i - \bar{y}_{left})^2 + \sum_{i \in S_{right}} (y_i - \bar{y}_{right})^2, \quad (4.3)$$

where  $y_i$  is the target value of the  $i$ -th sample, and  $\bar{y}_{left}$  and  $\bar{y}_{right}$  are the mean target values in  $S_{left}$  and  $S_{right}$ , respectively.

2. Recursively applying step 1 to the subsets  $S_{left}$  and  $S_{right}$  until a stopping criterion is met, such as reaching the maximum depth of the tree or having a minimum number of samples in a leaf node.
3. The tree can be used to predict the target value of a new sample by traversing the tree from the root node to a leaf node based on the feature values of the sample. The predicted target value for the sample is the mean target value of the training samples in the corresponding leaf node.

The Decision Tree Regressor model is a powerful and flexible algorithm that can capture complex nonlinear relationships between the input features and the target values. However, it can be prone to overfitting if the tree is too deep or the

training data has noise or outliers. Regularization techniques, such as setting a maximum depth or minimum number of samples in a leaf node, can help prevent overfitting.

### 4.1.3 Random Forest Regression

The Random Forest Regressor is an ensemble learning method that constructs a forest of decision trees and combines their predictions in order to produce more accurate predictions. The algorithm's major stages are as follows:

1. **Random Sampling:** Random Sampling: At each split in the tree, a random subset of features is selected. This is known as random subspace method or feature bagging. This can be represented as:

Let  $m$  be the total number of features and  $m'$  be the number of features to be used for a particular split. Then, the probability of selecting a particular feature  $i$  for a split can be given by:

$$P(i) = \frac{1}{m'} \quad (4.4)$$

2. **Bootstrap Aggregating (Bagging):** The random forest model is trained on multiple bootstrap samples from the training set. This is known as bagging. Each sample has the same size as the original training set but with replacement. The resulting bootstrap samples are used to train individual decision trees.
3. **Decision Trees:** The individual decision trees in the random forest are grown using the CART algorithm. Each tree is grown as follows:
  - (a) **Splitting:** The feature with the best split is chosen from a random subset of features.
  - (b) **Gini Index:** The Gini index is used to determine the best split. The Gini index is a measure of impurity and is calculated as follows:

$$Gini(D) = \sum_{k=1}^K p_k(1 - p_k) \quad (4.5)$$

where  $D$  is the dataset,  $K$  is the number of classes, and  $p_k$  is the proportion of samples belonging to class  $k$  in dataset  $D$ .

- (c) **Recursive Partitioning:** The decision tree is recursively partitioned into smaller subsets using the selected feature and threshold value until the leaf node is reached.
- 4. **Ensemble Learning:** The final prediction is a weighted average of the predictions of all decision trees in the random forest. The weights are proportional to the accuracy of each decision tree. This can be represented as:

$$\hat{y} = \frac{1}{T} \sum_{t=1}^T f_t(x) \quad (4.6)$$

where  $\hat{y}$  is the predicted output,  $T$  is the total number of trees in the forest, and  $f_t(x)$  is the predicted output of the  $t^{th}$  decision tree.

Overall, the Random Forest Regressor technique is a strong and adaptable machine learning model that is frequently utilized for regression issues across a broad range of fields. It is a popular option for many applications because to its capacity to handle complicated relationships and nonlinearities in the data, as well as its resilience to noisy and missing data.

#### 4.1.4 Gradient Boosting Regressor

The gradient boosting algorithm is a machine learning technique that combines multiple weak models into a single strong model. It works by iteratively adding weak models to the ensemble, where each model is fit to the negative gradient of the loss function with respect to the ensemble's predictions. The algorithm can be described with the following equations:

Let  $y$  be the true output,  $f_0$  be the initial model,  $f_t$  be the  $t$ -th model in the ensemble, and  $r_t$  be the negative gradient of the loss function with respect to the ensemble's predictions.

$$y = f_0(x) + \sum_{t=1}^T f_t(x) \quad (4.7)$$

At each iteration, the algorithm fits a weak model  $h_t$  to the negative gradient  $r_t$  of the loss function with respect to the ensemble's predictions:



$$h_t = \operatorname{argmin}_h \sum_{i=1}^n L(y_i, f_{t-1}(x_i) + h(x_i)) \quad (4.8)$$

where  $L$  is the loss function, and  $n$  is the number of samples in the dataset.

The weak model  $h_t$  is then added to the ensemble by computing its weight  $\gamma_t$  using a line search:

$$\gamma_t = \operatorname{argmin}_\gamma \sum_{i=1}^n L(y_i, f_{t-1}(x_i) + \gamma h_t(x_i)) \quad (4.9)$$

The final model is the sum of all weak models weighted by their respective coefficients:

$$f_T(x) = \sum_{t=1}^T \gamma_t h_t(x) \quad (4.10)$$

where  $T$  is the total number of weak models.

The algorithm can be further modified to add regularization terms to prevent overfittings, such as  $L_1$  and  $L_2$  regularization:

$$L_1 = \sum_{i=1}^n L(y_i, f_{t-1}(x_i) + h(x_i)) + \lambda |\gamma| \quad (4.11)$$

$$L_2 = \sum_{i=1}^n L(y_i, f_{t-1}(x_i) + h(x_i)) + \frac{\lambda}{2} \gamma^2 \quad (4.12)$$

where  $\lambda$  is the regularization parameter, and  $|\gamma|$  and  $\gamma^2$  are the L1 and L2 norms of the weight  $\gamma$ , respectively.

By gradually adding weak learners that enhance the predictions, the gradient boosting algorithm iteratively attempts to reduce the loss function. Each weak learner attempts to forecast the residuals of the previous iteration, and the step size (learning rate) determines how much the predictions are updated at each iteration. The final prediction is a weighted average of all weak learners' predictions, with the weights corresponding to the step sizes (learning rates) of each weak learner.

Overall, the Gradient Boosting Regressor technique is a potent and adaptable machine learning model that is often employed to address regression issues. It is a well-liked option for several applications due to its capacity to manage complicated relationships and nonlinearities in the data as well as its tolerance to noisy and missing data.

#### 4.1.5 Extreme Gradient Boosting(XGBoost)

XGBoost is an abbreviation that stands for "extreme gradient boosting," which is a faster and more performant variant of gradient boosted decision trees. Its origins may be traced back to a research conducted in 2016 by Tianqi Chen. The XGBoost method is a decision tree ensemble; it is constructed sequentially from a number of decision trees, with each tree attempting to enhance the performance of the tree before it. With XGBoost, the process of training each tree is parallelized, which significantly enhances the rate at which the training is finished. The XGBoost method has found broad use in agricultural output forecasting. The algorithm can be described with the following equations:

1. **Initialization:** Let  $x_i$  and  $y_i$  be the  $i^{th}$  input feature vector and target value, respectively. Let  $f_0$  be the initial prediction that is set to the mean value of the target values. The objective function of XGBoost can be defined as:

$$Obj(f_t) = \sum_{i=1}^n l(y_i, f_{t-1}(x_i) + h_t(x_i)) + \Omega(h_t) \quad (4.13)$$

where  $f_t$  is the predicted output at iteration  $t$ ,  $l(y_i, f_{t-1}(x_i) + h_t(x_i))$  is the loss function that measures the difference between the predicted output and the actual target value, and  $\Omega(h_t)$  is the regularization term that penalizes complex models.

2. **Gradient Descent:** The gradient of the objective function with respect to the predicted output  $f_t$  is calculated as:

$$g_{i,t} = \frac{\partial l(y_i, f_{t-1}(x_i))}{\partial f_{t-1}(x_i)} \quad (4.14)$$

The gradient of the objective function with respect to the base learner function  $h_t$  is calculated as:

$$h_{i,t} = \frac{\partial^2 l(y_i, f_{t-1}(x_i))}{\partial f_{t-1}(x_i)^2} \quad (4.15)$$

3. **Tree Boosting:** XGBoost uses decision trees as base learners. Each decision tree is grown using the gradient information as follows:
  - (a) Split Finding: The decision tree is recursively partitioned into smaller subsets using the best split that minimizes the loss function.
  - (b) Weighted Quantile Sketch: To find the best split efficiently, XGBoost uses a weighted quantile sketch algorithm that approximates the distribution of feature values.
  - (c) Regularization: XGBoost applies  $L_1$  and  $L_2$  regularization on the weights of the decision tree to prevent overfitting.
4. **Ensemble Learning:** The final prediction is the weighted sum of the predictions of all decision trees in the XGBoost model. The weights are proportional to the accuracy of each decision tree. This can be represented as:

$$\hat{y} = \sum_{t=1}^T \alpha_t f_t(x) \quad (4.16)$$

where  $\hat{y}$  is the predicted output,  $T$  is the total number of decision trees in the XGBoost model,  $\alpha_t$  is the weight of the  $t^{th}$  decision tree, and  $f_t(x)$  is the predicted output of the  $t^{th}$  decision tree.

Overall, XGBoost is a powerful algorithm that can be used to create highly accurate models for predicting crop yields. By updating the predictions using the residuals and fitting new models to the residuals, XGBoost is able to create a highly accurate model that can be used by farmers and agricultural companies to make decisions about planting and harvesting crops. The feature importance metric can also be used to identify which weather variables have the greatest impact on crop yield, providing valuable insights into the relationship between weather and crop growth.

## 4.2 Model Evaluation: Assessing Prediction Accuracy

we applied Machine Learning models one-by-one on each of the crops, and the following has been observed:

- **Root Mean Square Error (RMSE):** We have calculated the RMSE of the errors in the predicted yield values.

$$\sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (4.17)$$

As the scale of the RMSE will be different for the different crops according to their yield range, we normalize the RMSE by dividing it with the maximum value of the actual yield and taking the percentage, i. e.

$$\frac{RMSE}{Maximum\ Value\ of\ the\ actual\ yield\ value} \times 100\% \quad (4.18)$$

where,  $y_i$ ,  $\hat{y}_i$  and  $n$  denote actual yield value, predicted yield value, and the number of total data points, respectively.

- **Percentage Error:** The percentage error between actual and predicted value has been calculated by:

$$\frac{|actual\ yield - predicted\ yield|}{actual\ yield} \times 100\% \quad (4.19)$$

The coefficient of determination, or r-squared ( $R^2$ ), is a statistical measure that represents the proportion of the variance in the dependent variable that is explained by the independent variables.

The formula for r-squared is:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \quad (4.20)$$

where  $SS_{res} = \sum_{i=1}^n (y_i - \hat{y}_i)^2$  is the sum of squares of the residuals (i.e., the differences between the predicted values and the actual values) and  $SS_{tot} = \sum_{i=1}^n (y_i - \bar{y})^2$  is the total sum of squares (i.e., the sum of squares of the differences between the actual values and the mean of the dependent variable).<sup>1</sup>

A higher value of  $R^2$  represents higher accuracy in yield prediction with the values ranging from 0 to 1.

---

<sup>1</sup> $y_i$ ,  $\hat{y}_i$  and  $\bar{y}$  are the  $i$ th observed value of the dependent variable, predicted value of the dependent variable, and mean of the dependent variable respectively

# Chapter 5

## Results

Now, in this section, we present the results of the Machine Learning models one-by-one. The plots and the error analysis for the different Models are attached next.

### 5.1 Results of Various Models

For each of the models, seven scatter plots are shown, which represent the results for the seven crops. In each of the scatter plots (in all the following subsections), the x and y axis represent the actual and predicted yield values respectively, for a certain crop, resulting by the use of the algorithm mentioned in the plot title. and the red line represents the  $y=x$  straight line, i. e. points upon the line imply absolute perfection in the yield prediction.<sup>1</sup>

After the plots, there is a table added, which shows the RMSE Percentages, the percentage errors and the correlation coefficients that we get from the predictions. After that we analyze the results we get from different models.

---

<sup>1</sup>the 'polyfit' function in python has been used to fit the  $y = x$  line

### 5.1.1 Linear Regression

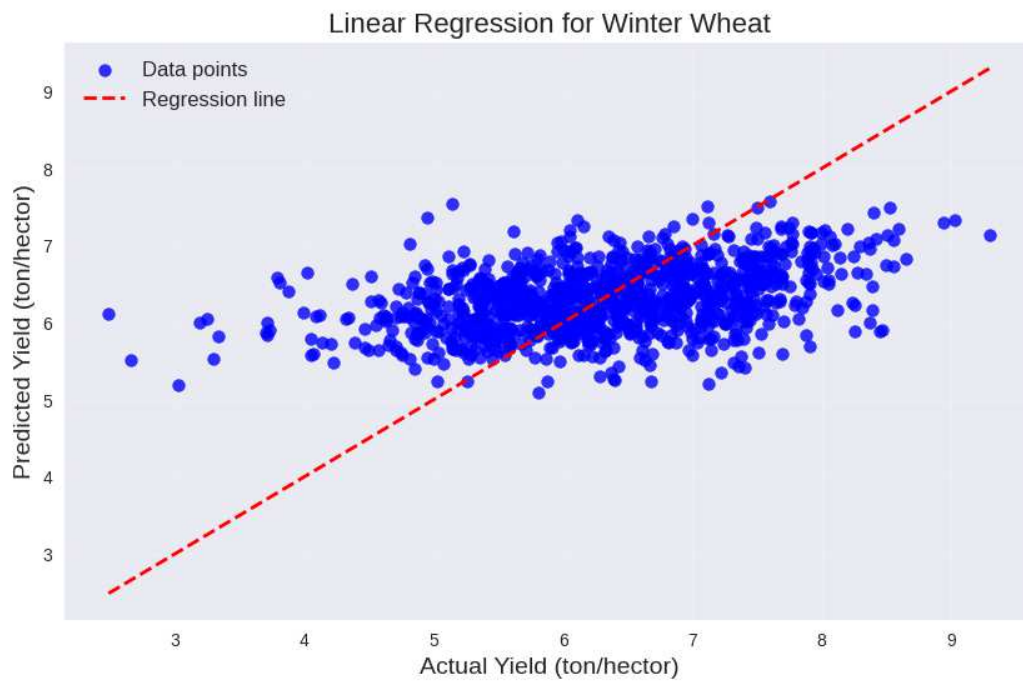


Figure 5.1: Actual vs Predicted yield values for Winter Wheat by Linear Regression

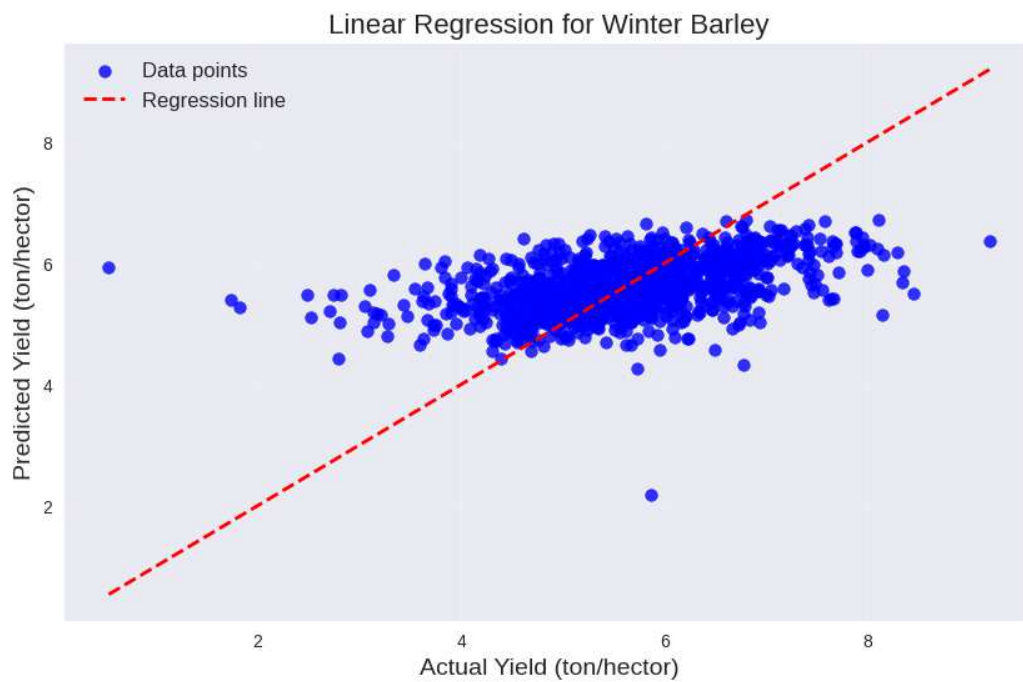


Figure 5.2: Actual vs Predicted yield values for Winter Barley by Linear Regression

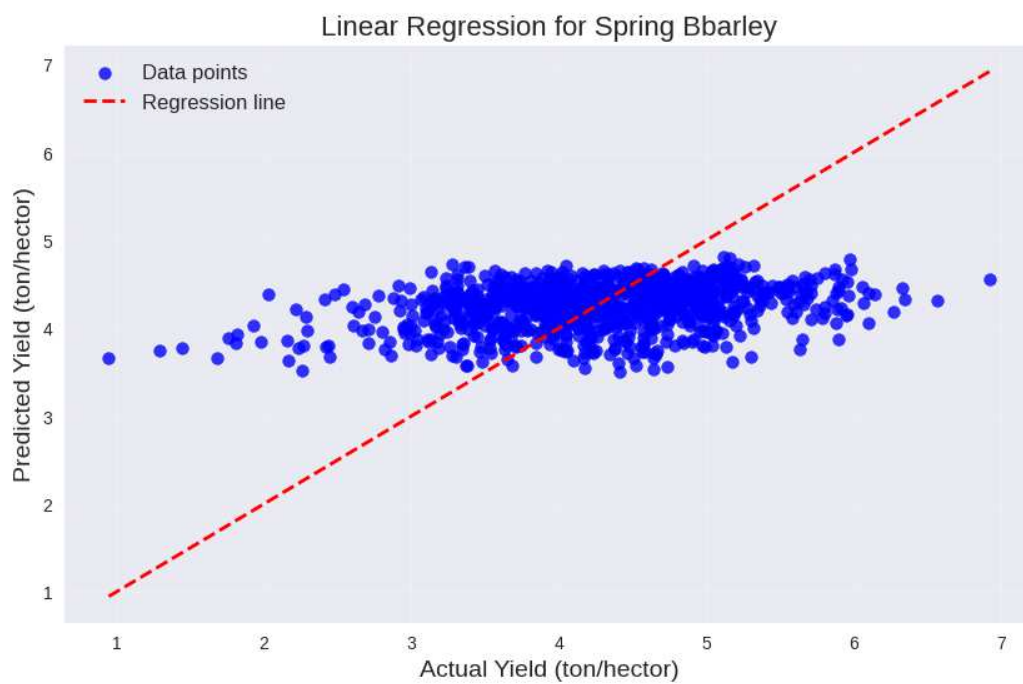


Figure 5.3: Actual vs Predicted yield values for Spring Barley by Linear Regression

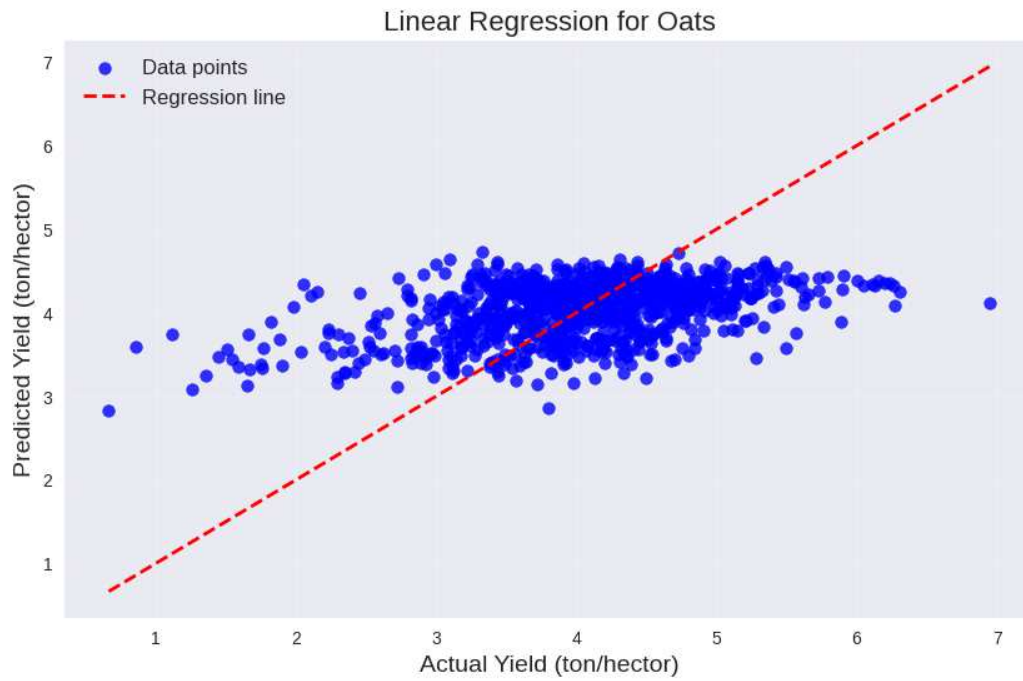


Figure 5.4: Actual vs Predicted yield values for Oats by Linear Regression

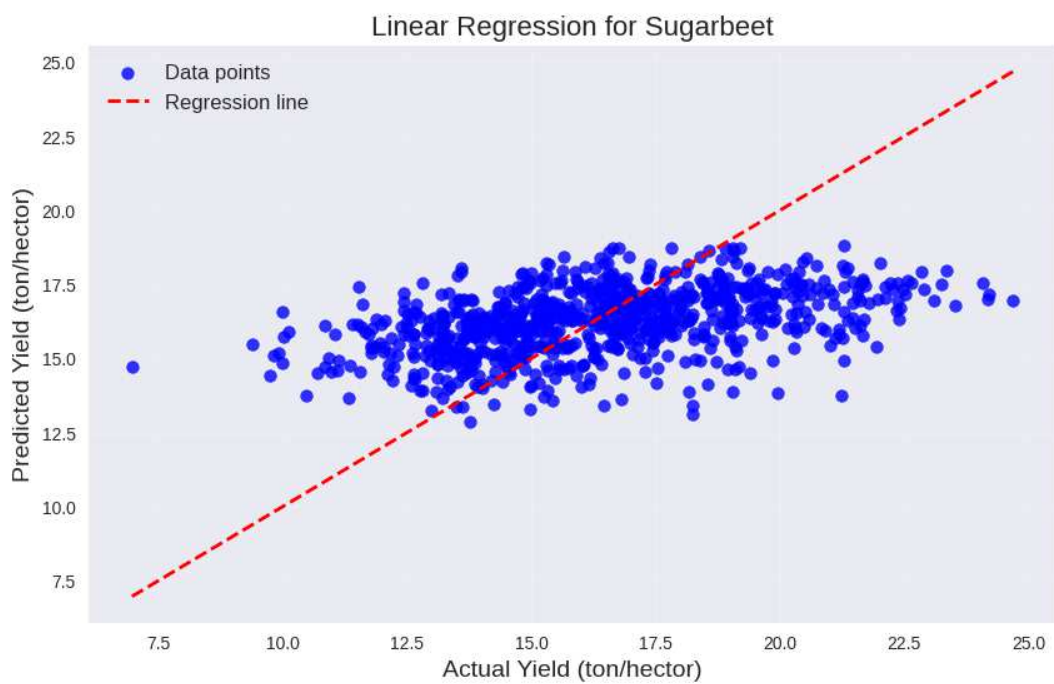


Figure 5.5: Actual vs Predicted yield values for Sugarbeet by Linear Regression



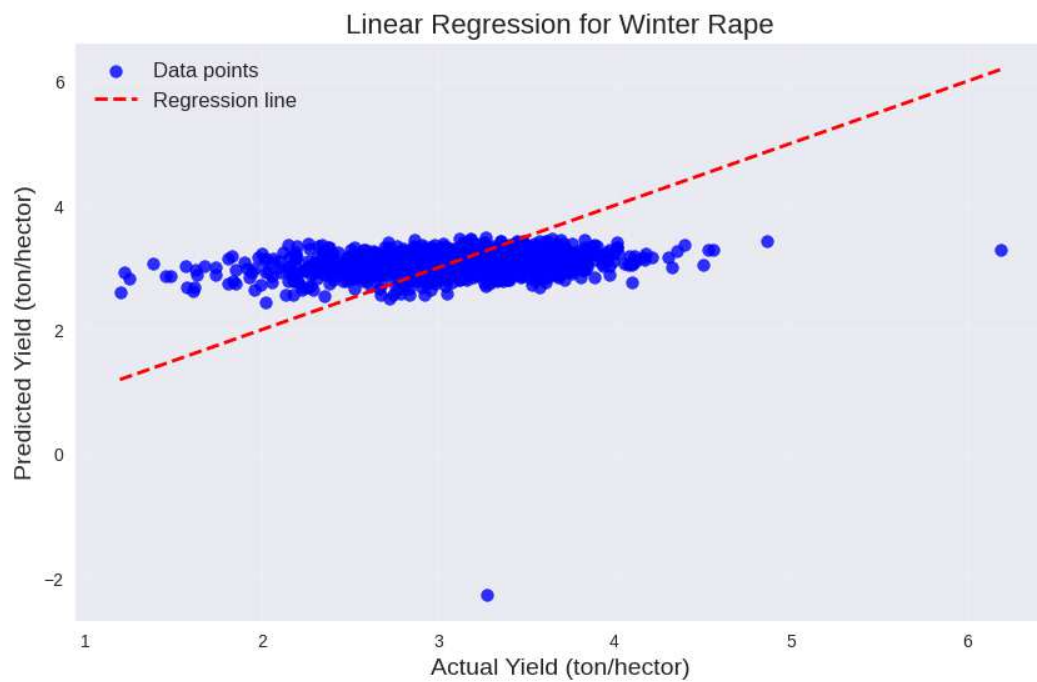


Figure 5.6: Actual vs Predicted yield values for Winter Rape by Linear Regression

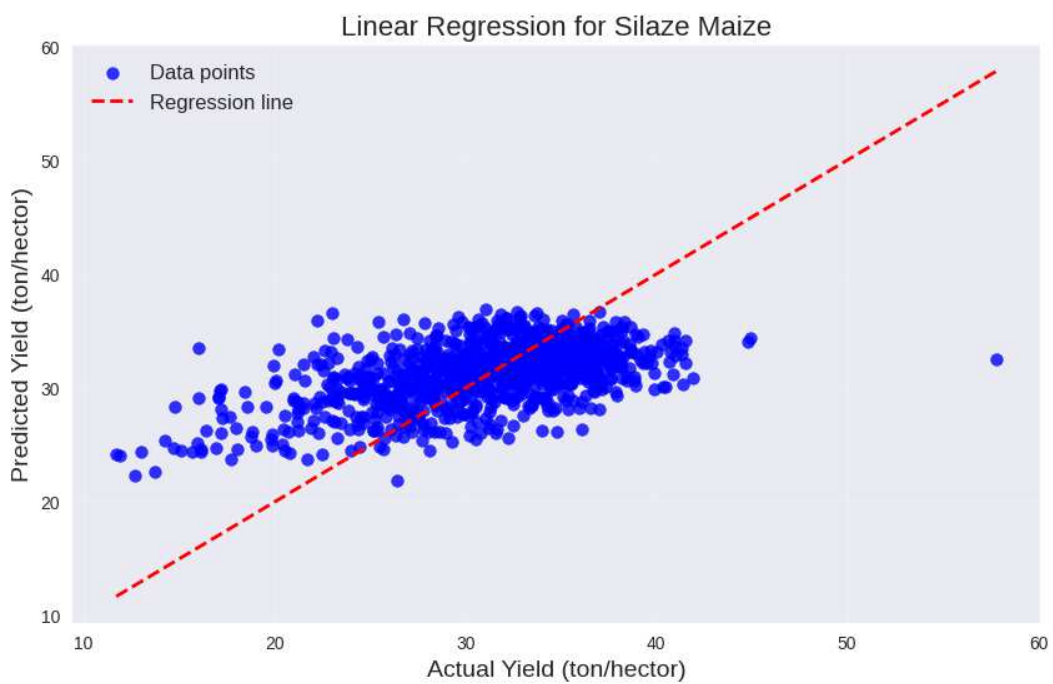


Figure 5.7: Actual vs Predicted yield values for Silaze Maize by Linear Regression

Crop Name	RMSE	Percentage Error	$R^2$ value for $y=x$ fit
Winter Wheat	9.54	12.23	0.16
Winter Barley	8.90	14.09	0.22
Spring Barley	9.76	15.82	0.09
Oats	10.11	16.56	0.19
Sugarbeet	9.59	12.83	0.19
Winter Rape	8.48	13.88	0.01
Spring Maize	7.89	12.87	0.28

Table 5.1: Accuracy of the Predictions by Linear Regression

### 5.1.2 Decision Tree Regressor

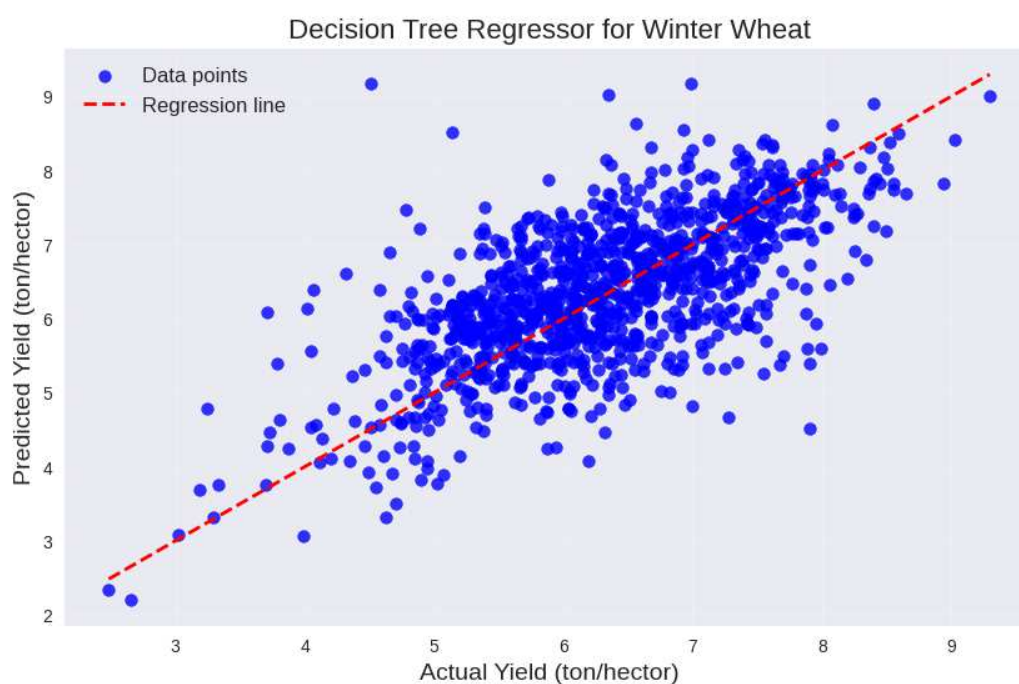


Figure 5.8: Actual vs Predicted yield values for Winter Wheat by Decision Tree Regressor

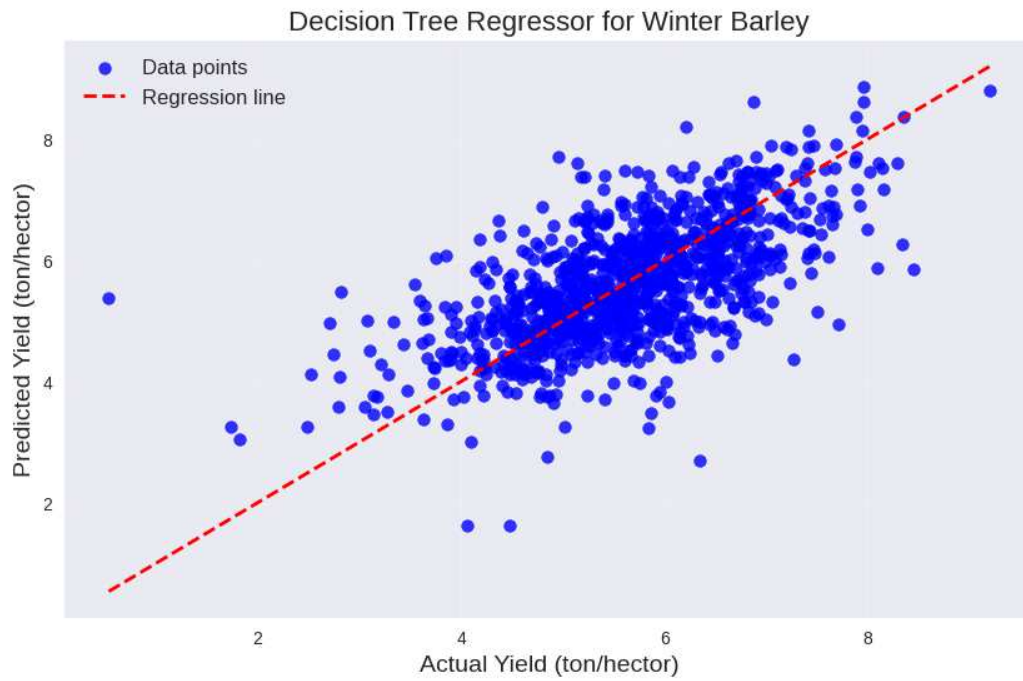


Figure 5.9: Actual vs Predicted yield values for Winter Barley by Decision Tree Regressor

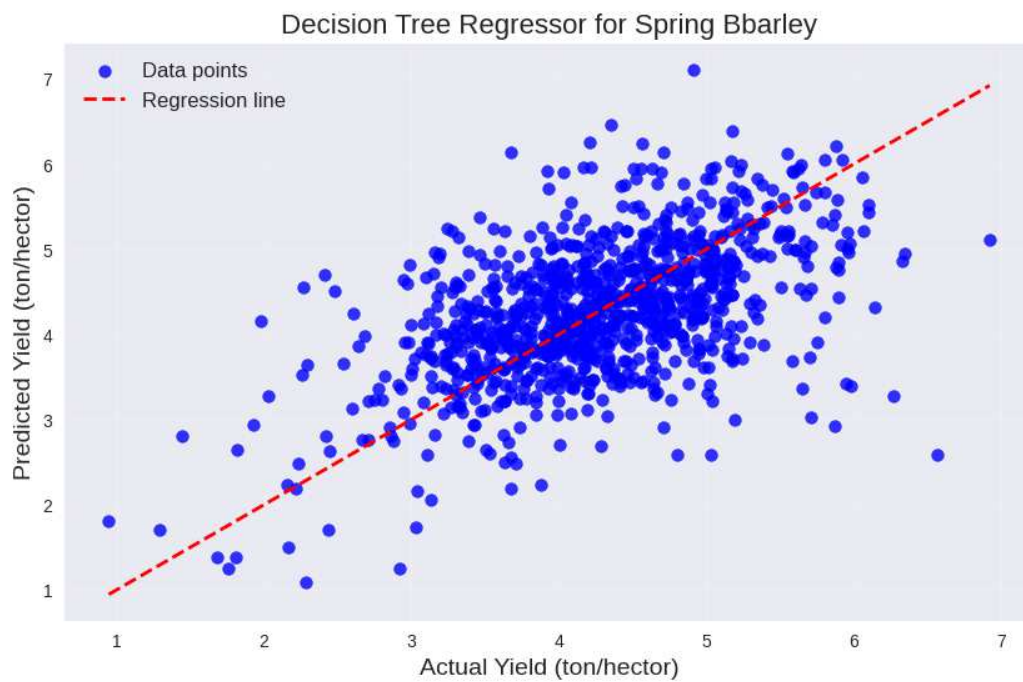


Figure 5.10: Actual vs Predicted yield values for Spring Barley by Decision Tree Regressor

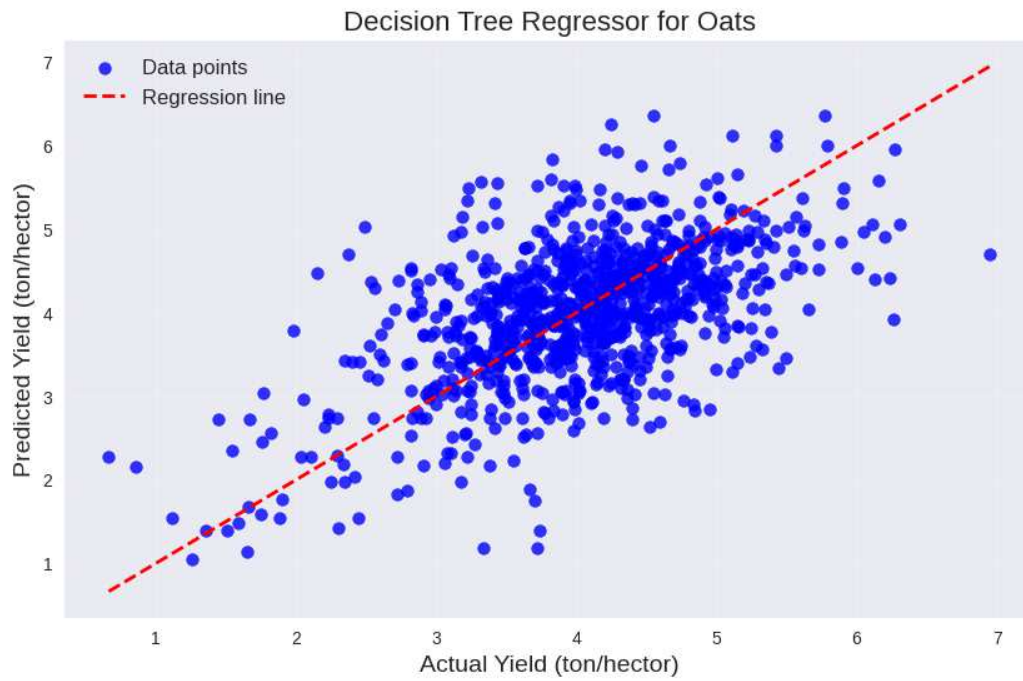


Figure 5.11: Actual vs Predicted yield values for Oats by Decision Tree Regressor

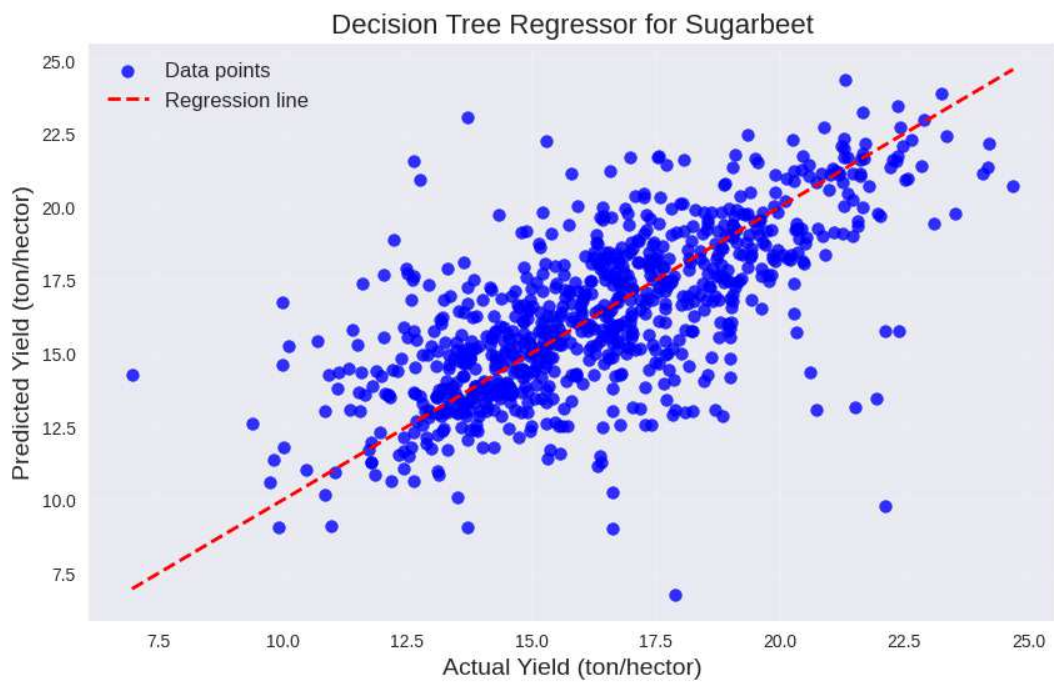


Figure 5.12: Actual vs Predicted yield values for Sugarbeet by Decision Tree Regressor

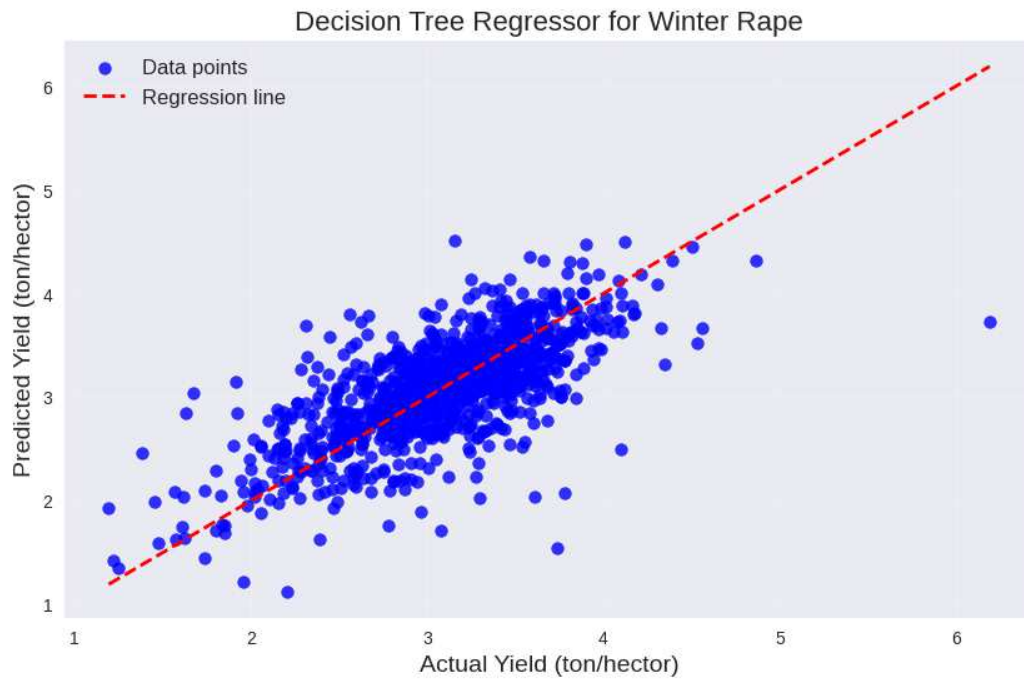


Figure 5.13: Actual vs Predicted yield values for Winter Rape by Decision Tree Regressor

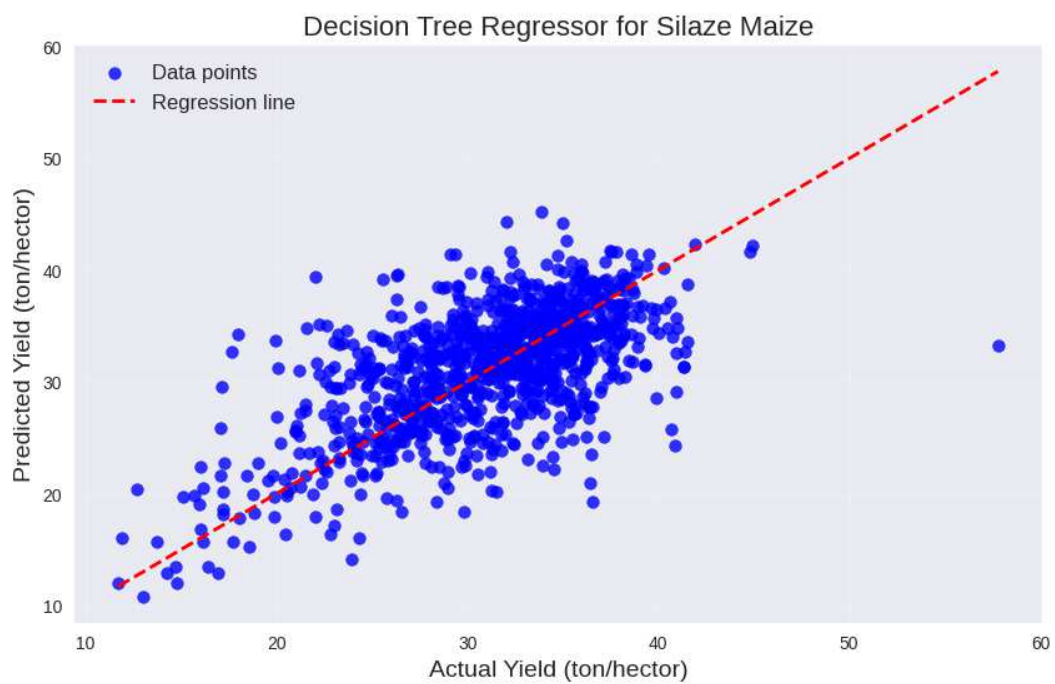


Figure 5.14: Actual vs Predicted yield values for Winter Rape by Decision Tree Regressor



Crop Name	RMSE	Percentage Error	$R^2$ value for $y=x$ fit
Winter Wheat	8.52	10.15	0.33
Winter Barley	8.44	12.75	0.30
Spring Barley	10.03	14.56	0.04
Oats	10.87	16.27	0.07
Sugarbeet	8.20	9.90	0.40
Winter Rape	6.43	10.07	0.42
Spring Maize	8.28	12.14	0.20

Table 5.2: Accuracy of the Predictions by Decision Tree Regressor Model

### 5.1.3 Random Forest Regressor

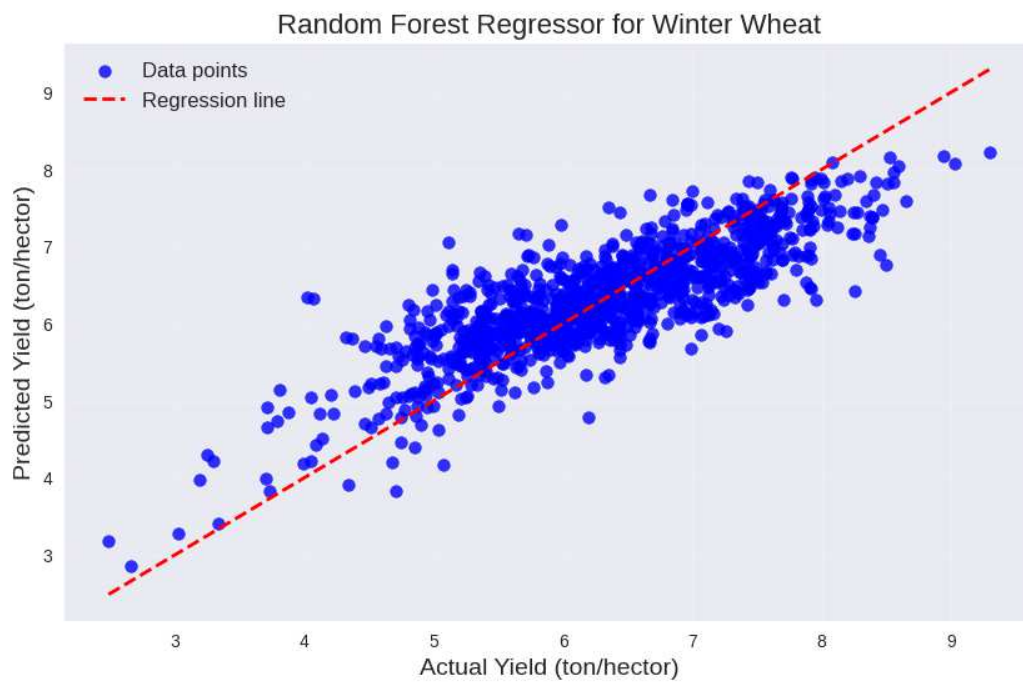


Figure 5.15: Actual vs Predicted yield values for Winter Wheat by Random Forest Regressor

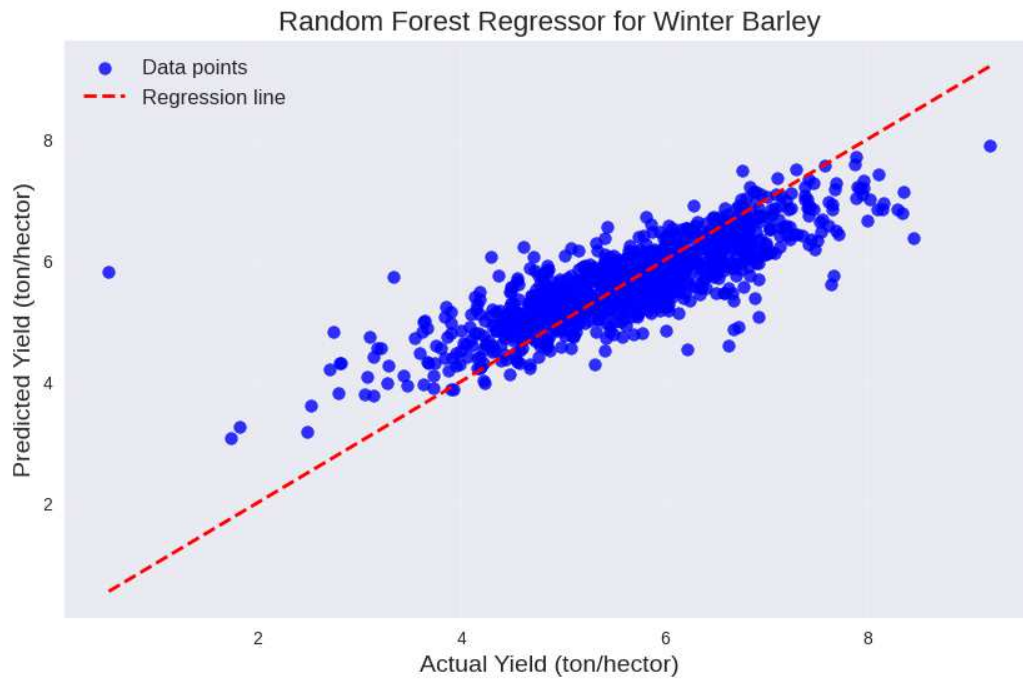


Figure 5.16: Actual vs Predicted yield values for Winter Barley by Random Forest Regressor

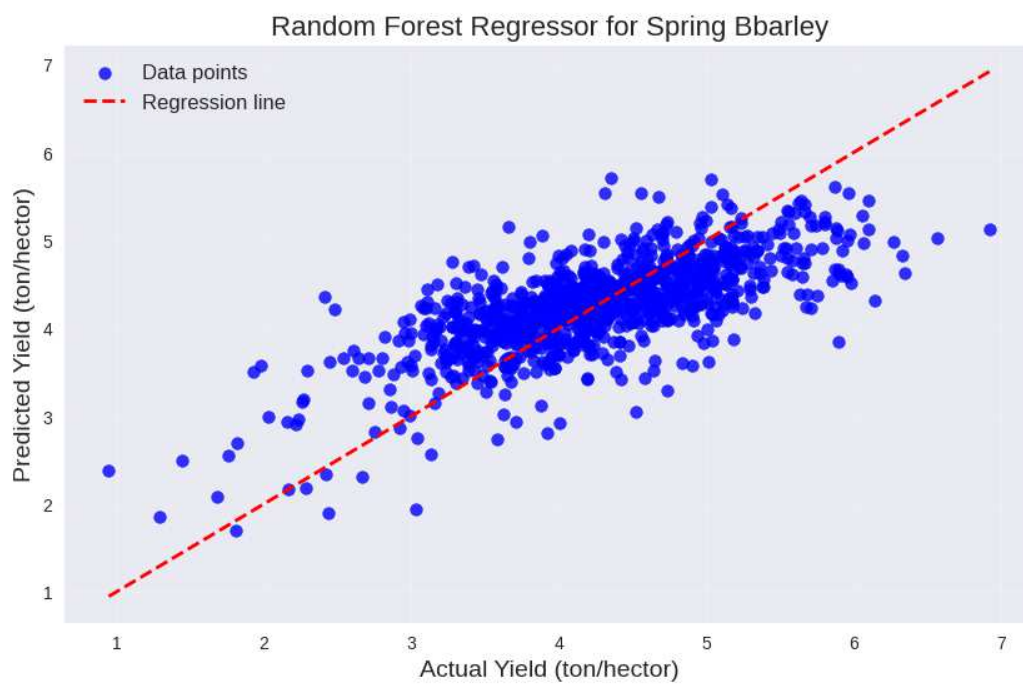


Figure 5.17: Actual vs Predicted yield values for Spring Barley by Random Forest Regressor

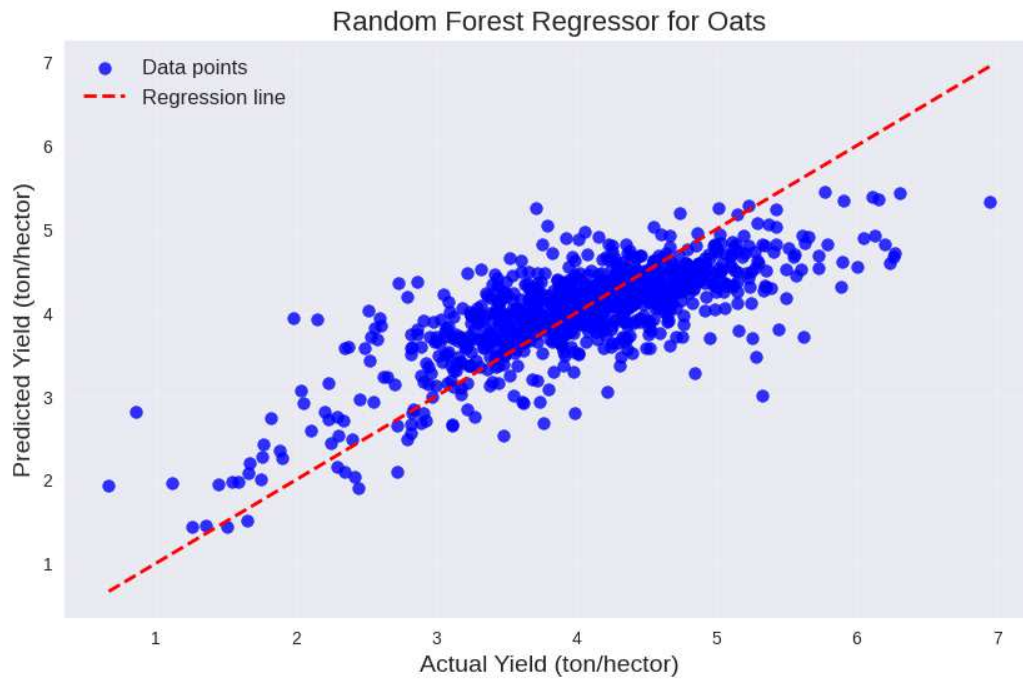


Figure 5.18: Actual vs Predicted yield values for Oats by Random Forest Regressor

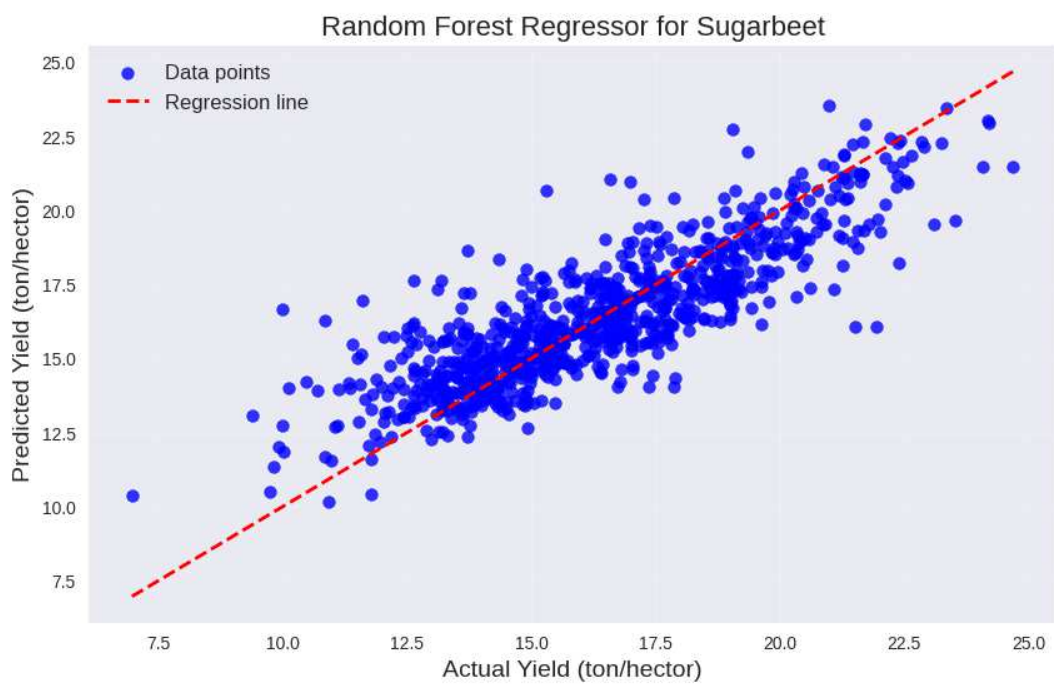


Figure 5.19: Actual vs Predicted yield values for Sugarbeet by Random Forest Regressor



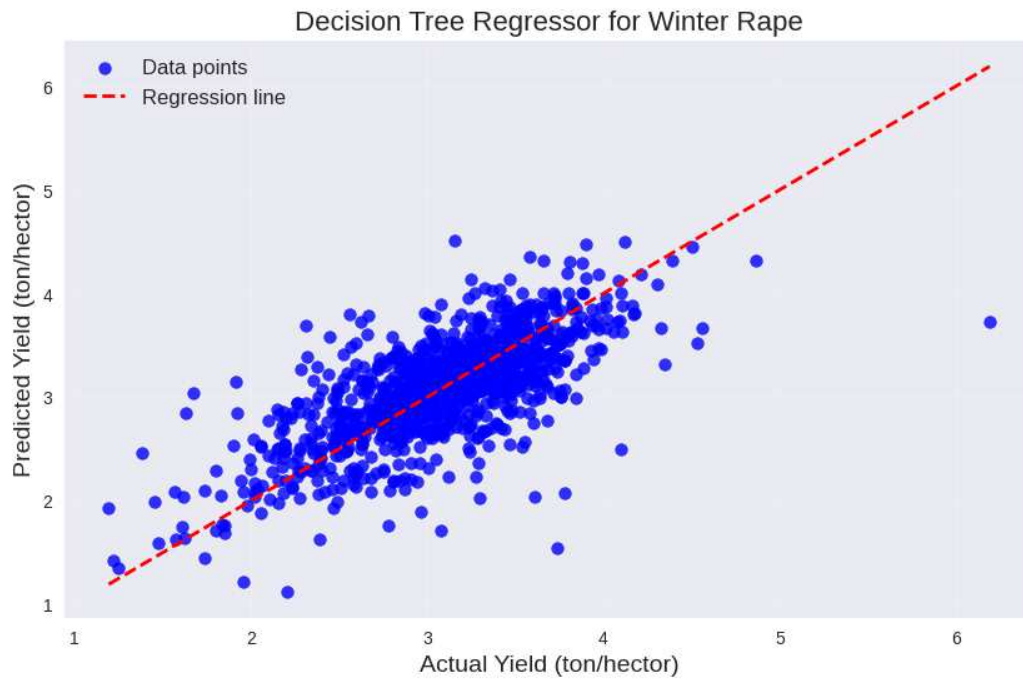


Figure 5.20: Actual vs Predicted yield values for Winter Rape by Random Forest Regressor

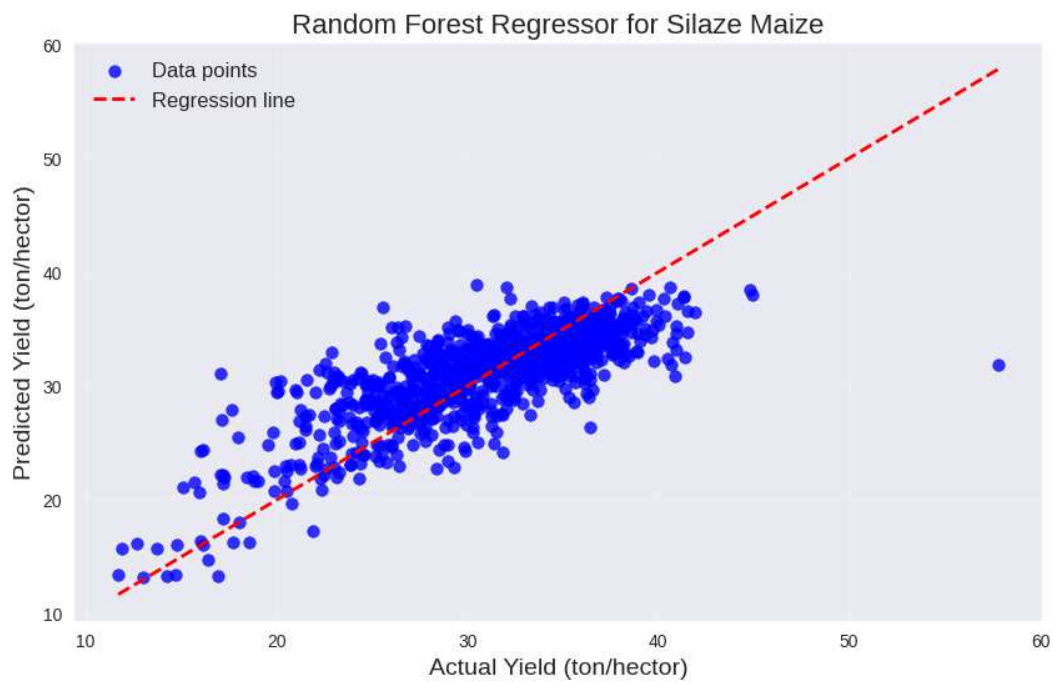


Figure 5.21: Actual vs Predicted yield values for Silaze Maize by Random Forest Regressor

Crop Name	RMSE	Percentage Error	$R^2$ value for $y=x$ fit
Winter Wheat	5.91	7.29	0.68
Winter Barley	5.76	8.92	0.67
Spring Barley	7.13	11.07	0.51
Oats	7.35	11.02	0.57
Sugarbeet	5.69	7.36	0.71
Winter Rape	4.94	8.03	0.66
Spring Maize	5.86	8.76	0.60

Table 5.3: Accuracy of the Predictions by Decision Tree Regressor Model

#### 5.1.4 Gradient Boosting

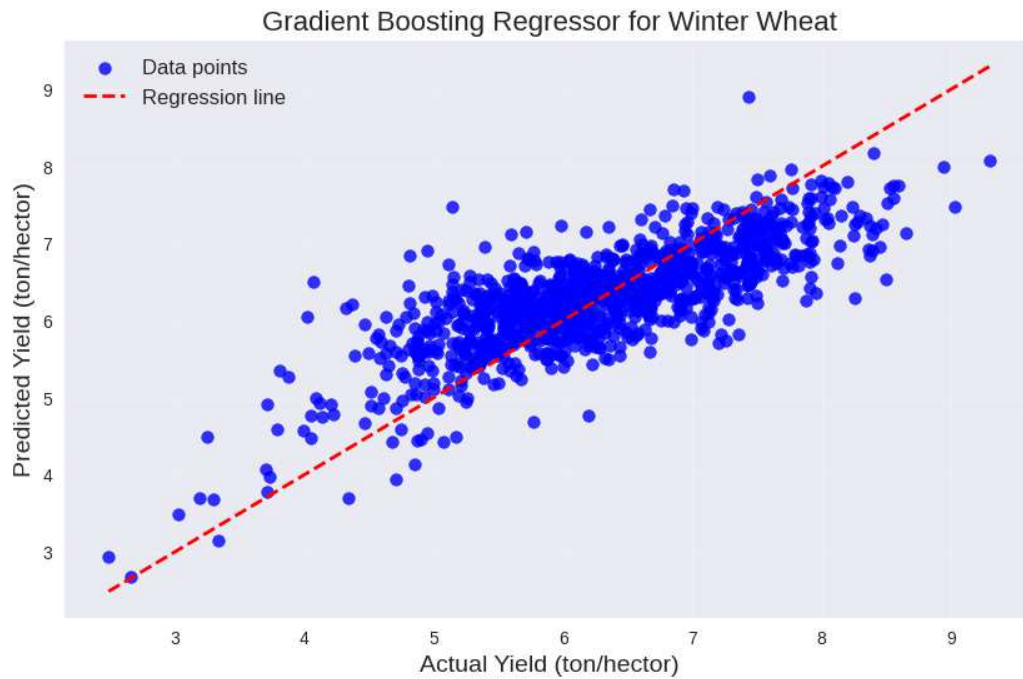


Figure 5.22: Actual vs Predicted yield values for Winter Wheat by Gradient Boosting

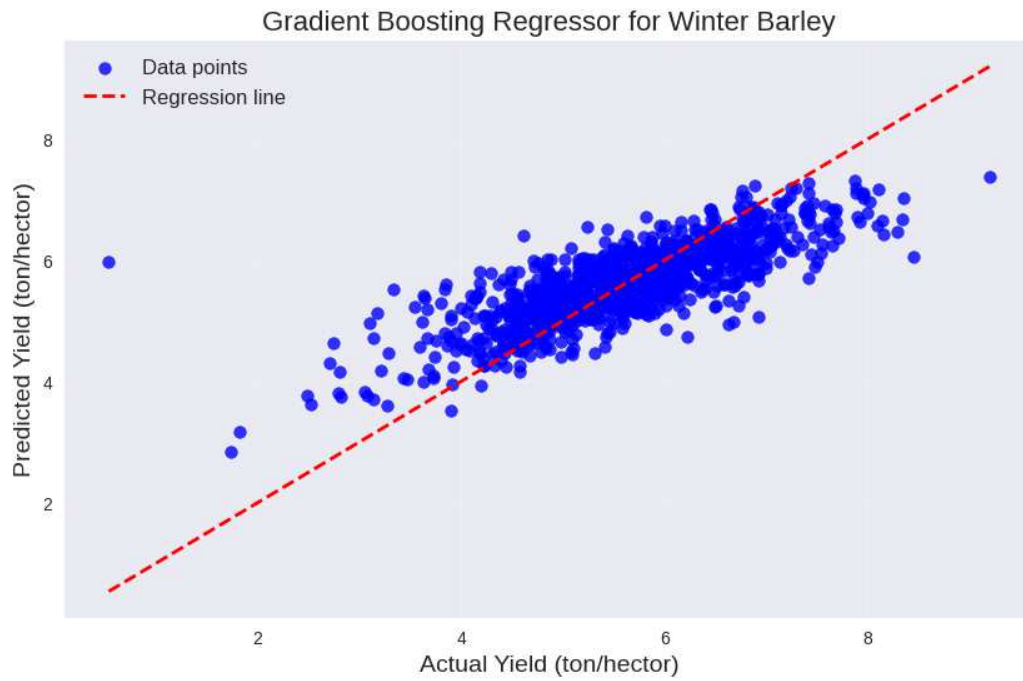


Figure 5.23: Actual vs Predicted yield values for Winter Barley by Gradient Boosting

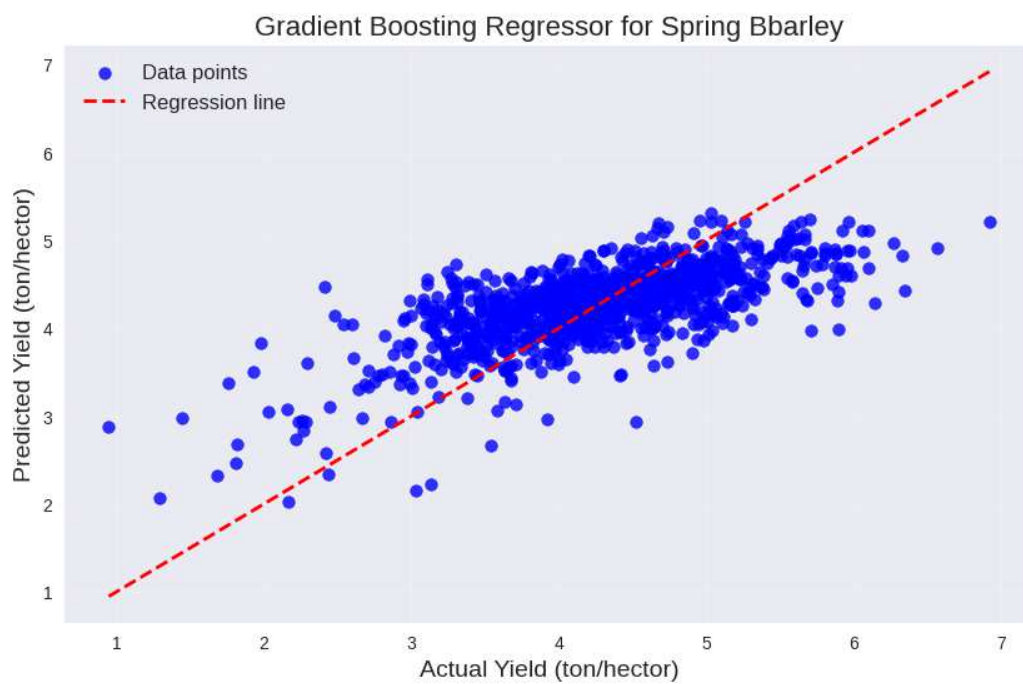


Figure 5.24: Actual vs Predicted yield values for Spring Barley by Gradient Boosting

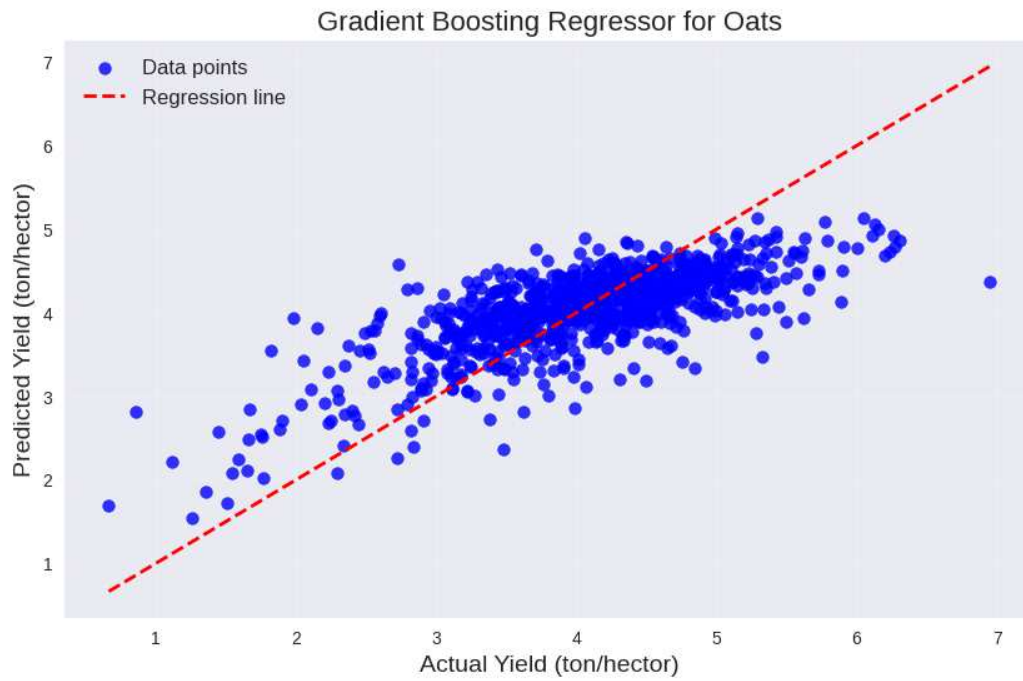


Figure 5.25: Actual vs Predicted yield values for Oats by Gradient Boosting

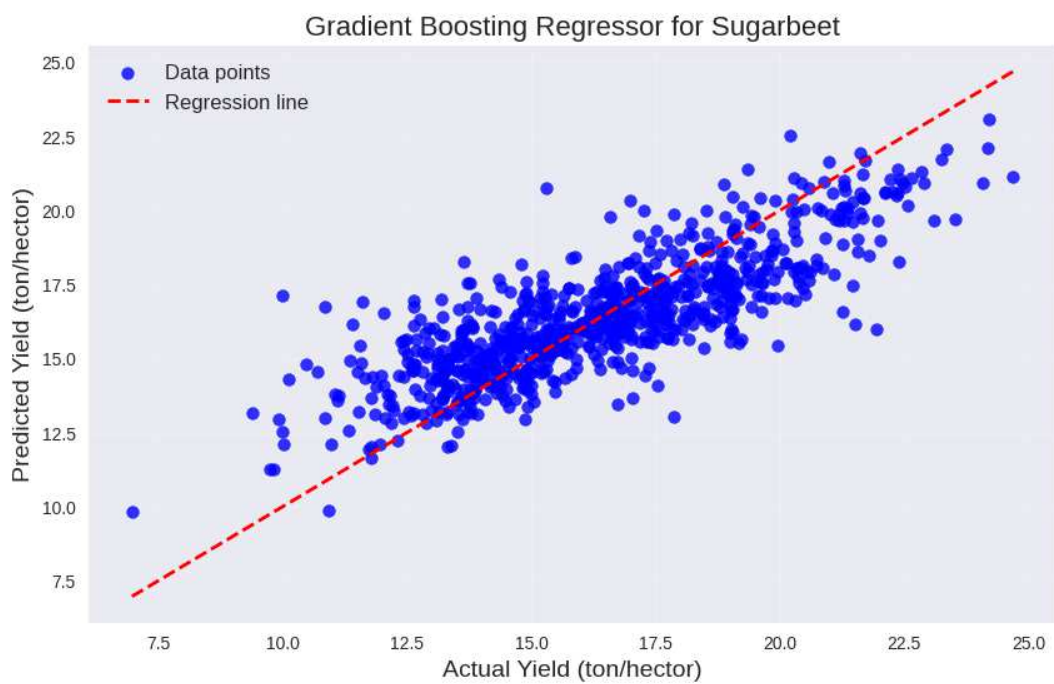


Figure 5.26: Actual vs Predicted yield values for Sugarbeet Gradient Boosting

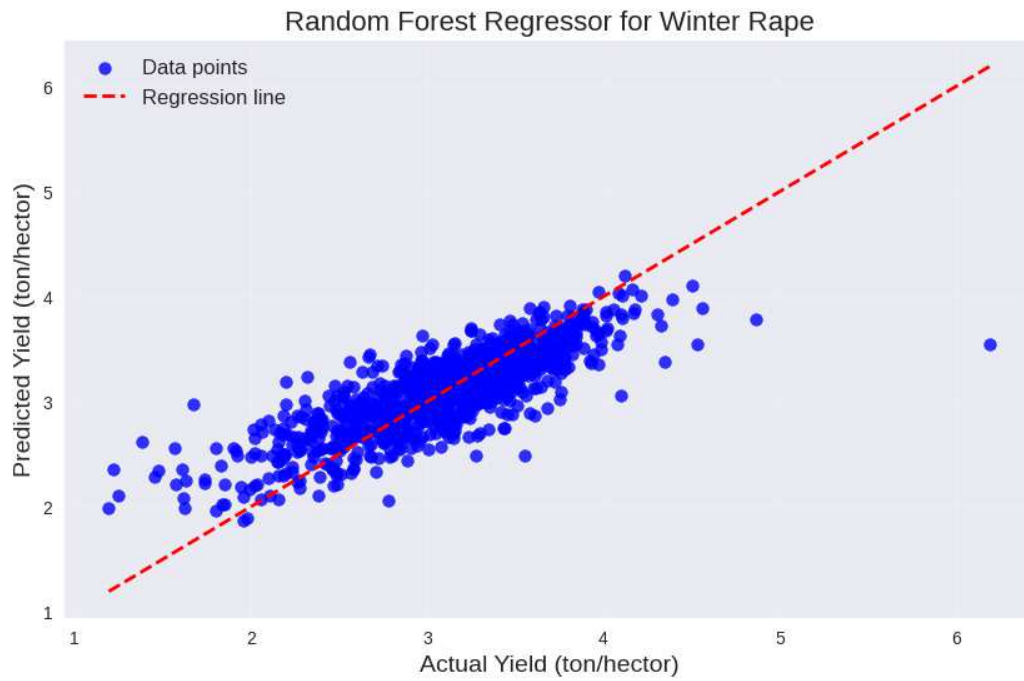


Figure 5.27: Actual vs Predicted yield values for Winter Rape by Gradient Boosting

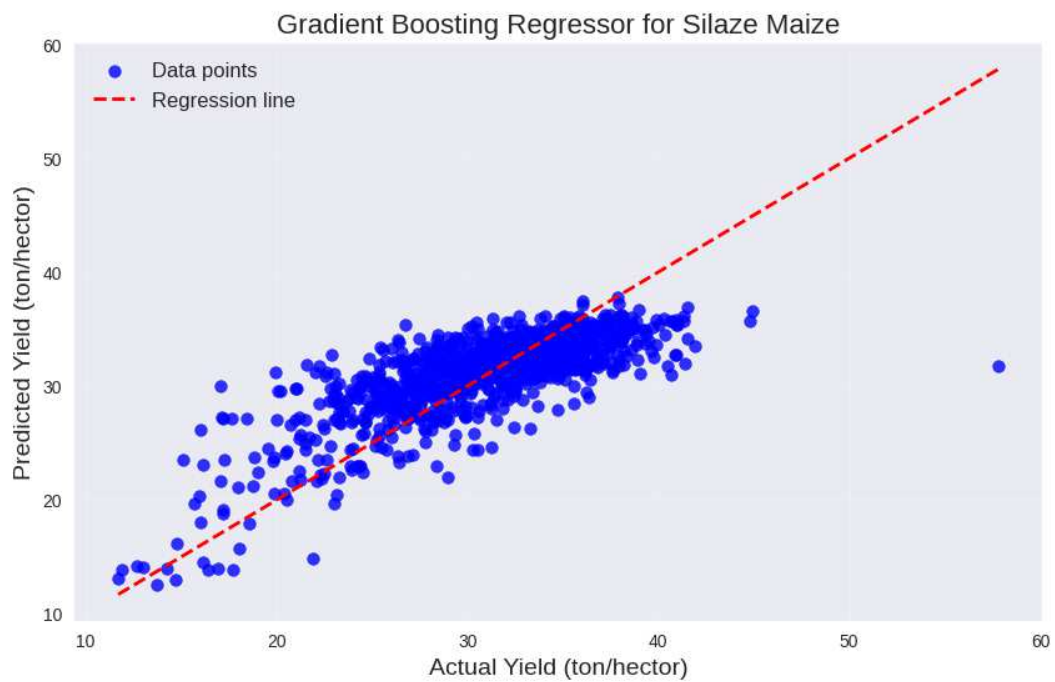


Figure 5.28: Actual vs Predicted yield values for Silaze Maize by Gradient Boosting

Crop Name	RMSE	Percentage Error	$R^2$ value for $y=x$ fit
Winter Wheat	6.56	8.09	0.60
Winter Barley	6.25	9.72	0.61
Spring Barley	7.49	11.78	0.47
Oats	7.75	12.03	0.52
Sugarbeet	6.22	8.06	0.66
Winter Rape	5.46	8.90	0.58
Spring Maize	5.92	9.07	0.59

Table 5.4: Accuracy of the Predictions by Decision Tree Regressor Model

### 5.1.5 Extreme Gradient Boosting(XGBoost)

The Extreme Gradient Boosting (XGBoost) algorithm is a type of gradient boosting algorithm that uses a more regularized model and a more efficient system implementation to improve performance.

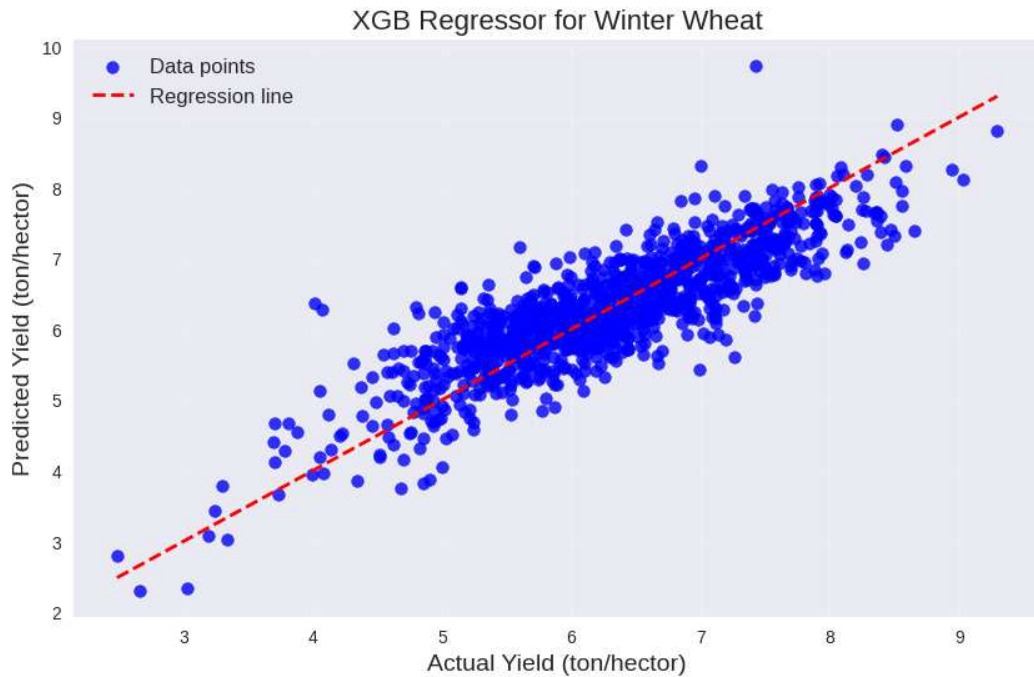


Figure 5.29: Actual vs Predicted yield values for Winter Wheat by XGBoost model



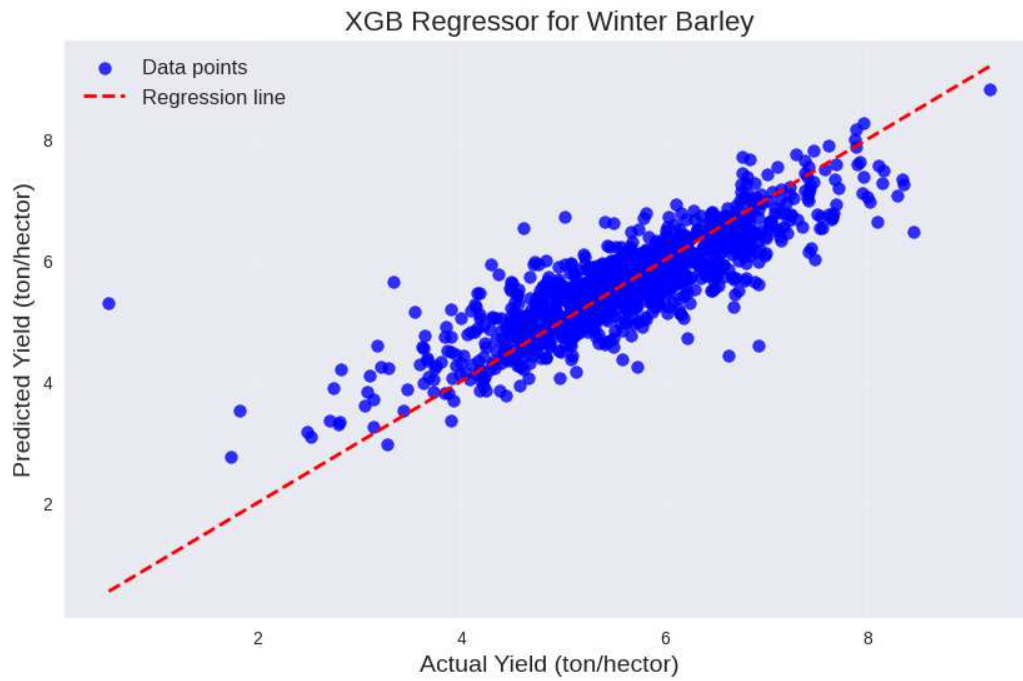


Figure 5.30: Actual vs Predicted yield values for Winter Barley by XGBoost model

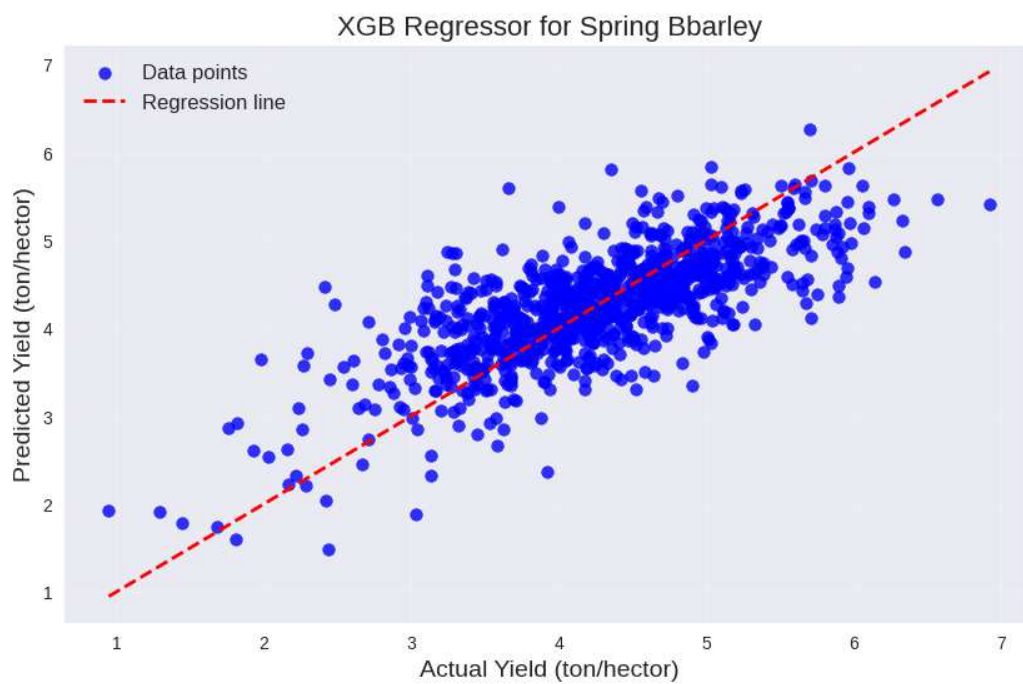


Figure 5.31: Actual vs Predicted yield values for Spring Barley by XGBoost model

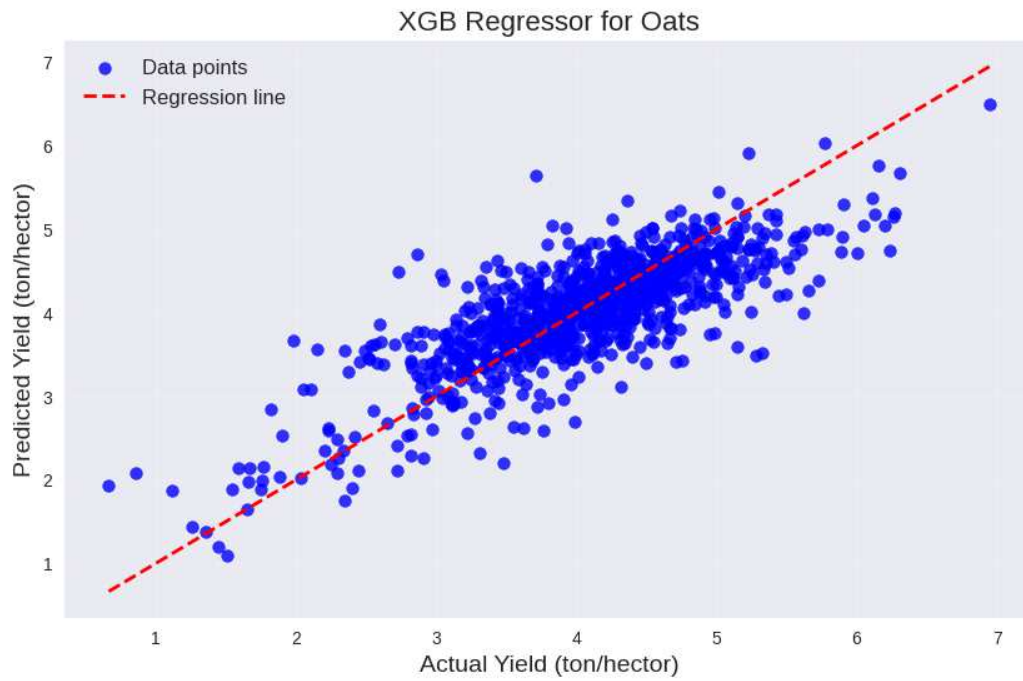


Figure 5.32: Actual vs Predicted yield values for Oats by XGBoost model

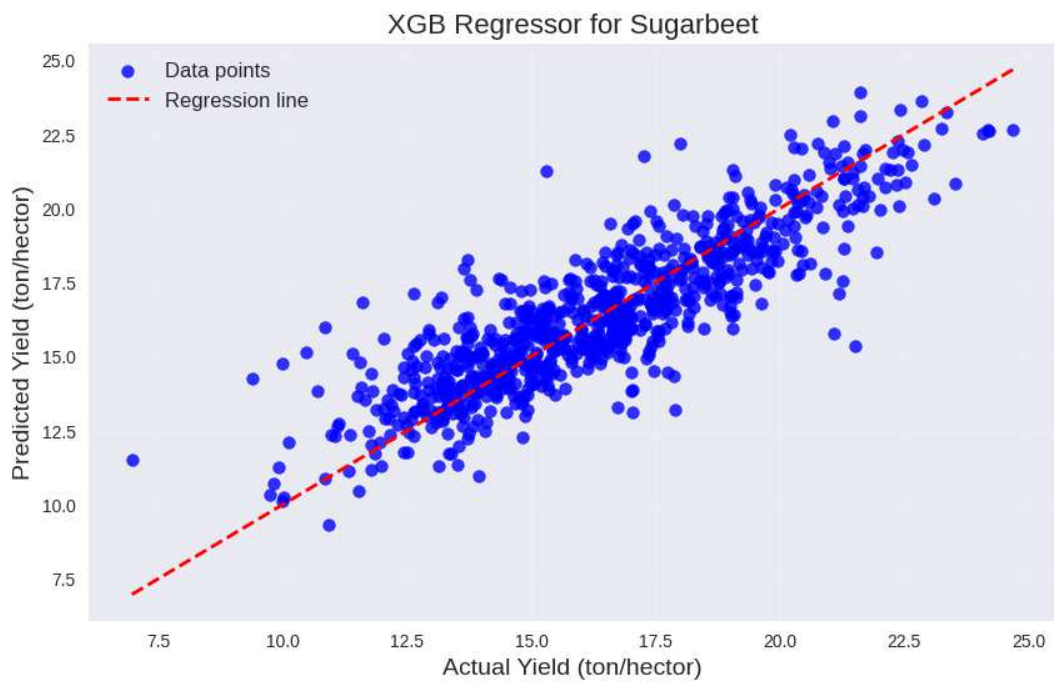


Figure 5.33: Actual vs Predicted yield values for Sugarbeet by XGBoost model



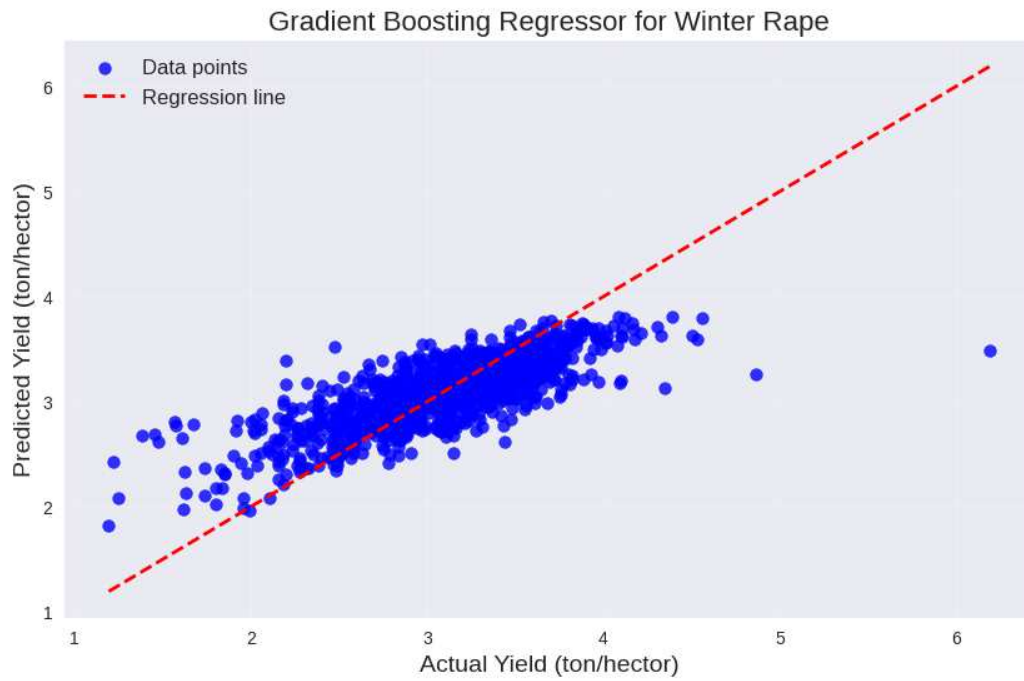


Figure 5.34: Actual vs Predicted yield values for Winter Rape by XGBoost model

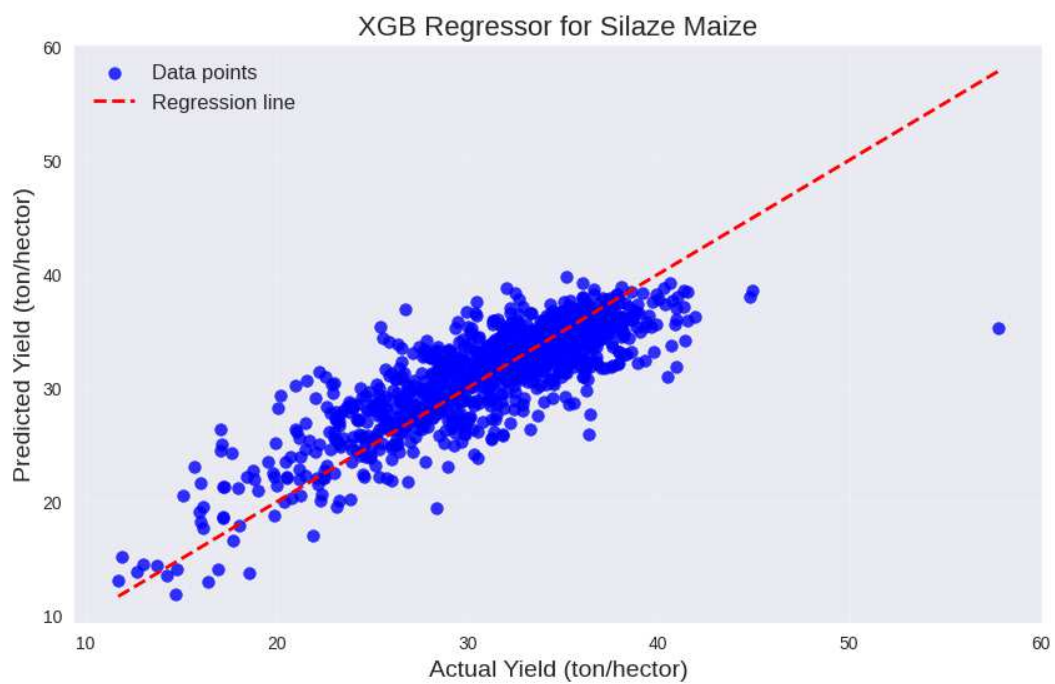


Figure 5.35: Actual vs Predicted yield values for Silaze Maize by XGBoost model

<b>Crop Name</b>	<b>RMSE</b>	<b>Percentage Error</b>	<b>R<sup>2</sup> value for y=x fit</b>
<b>Winter Wheat</b>	6.56	8.09	0.60
<b>Winter Barley</b>	6.25	9.72	0.61
<b>Spring Barley</b>	7.49	11.78	0.47
<b>Oats</b>	7.75	12.03	0.52
<b>Sugarbeet</b>	6.22	8.06	0.66
<b>Winter Rape</b>	5.46	8.90	0.58
<b>Spring Maize</b>	5.92	9.07	0.59

Table 5.5: Accuracy of the Predictions by XGBoost Model

## 5.2 Discussions

When we compare the percentage errors and the  $R^2$  values for  $y=x$  fit side by side, we observe the following:

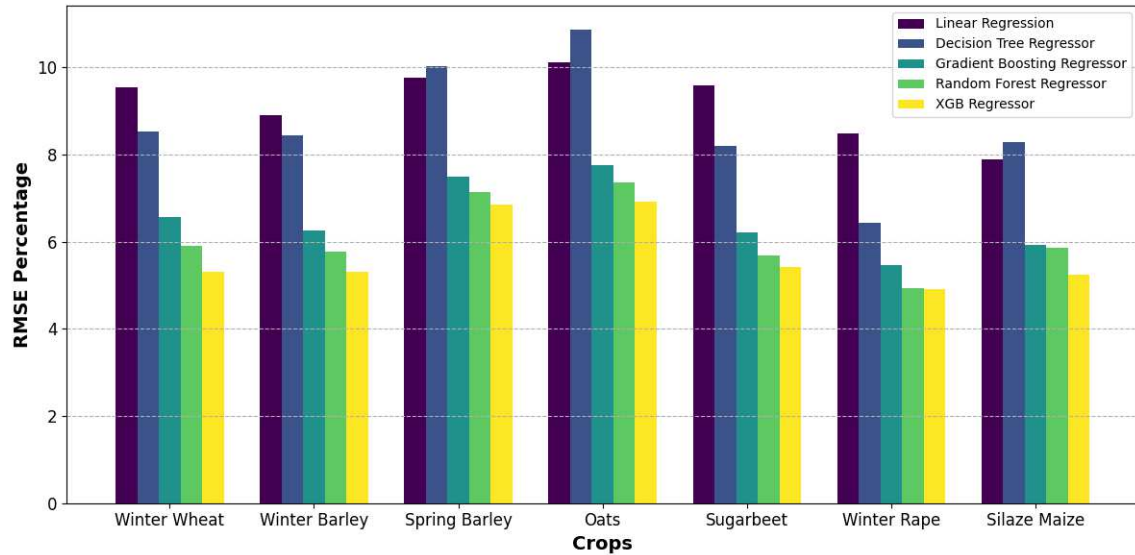


Figure 5.36: Comparison between the RMSE Percentage between actual and predicted yields by different models

The RMSE percentage barplot is used to evaluate the regression models' accuracy. The root-mean-square error (RMSE) percentage, which is reported as a percentage of the actual value, is a measurement of the average difference between the values that were anticipated and those that were actually achieved. A lower RMSE percentage denotes more accurate model predictions.

The bar plot shows how the RMSE percentage varies significantly across the different crops and models. For instance, the XGBoost regressor had the lowest RMSE percentage for the majority of the crops, indicating that it was the most accurate model for forecasting the yield of these crops. On the other hand, the performance of the Decision Tree Regressor was consistently poor across the majority of the crops, as evidenced by the fact that it had the highest RMSE percentage. This demonstrates that decision tree regressors might not be the best choice for predicting agricultural output.

It is possible for the performance of the models to be affected by a variety of factors, such as the type of data and the amount of data used for training, the particular methods that were used, and the model hyperparameters. For the vast majority

of the crops, for example, the random forest regressor and the XGBoost regressor consistently outperformed the control group regressor, demonstrating that they are reliable models that can accommodate a wide range of data and conditions.

The RMSE is a helpful indicator for determining the efficacy of regression models; however, it has significant drawbacks that need to be taken into consideration. Because it does not specify the nature of the error, the root mean square error (RMSE percentage), for instance, treats positive and negative errors equally. The presence of outliers in the data may also have an effect on the RMSE percentage, which may cause the results to be skewed. Therefore, when evaluating the efficacy of different regression models, it is essential to take into account a variety of measures as well as the specific context of the problem.

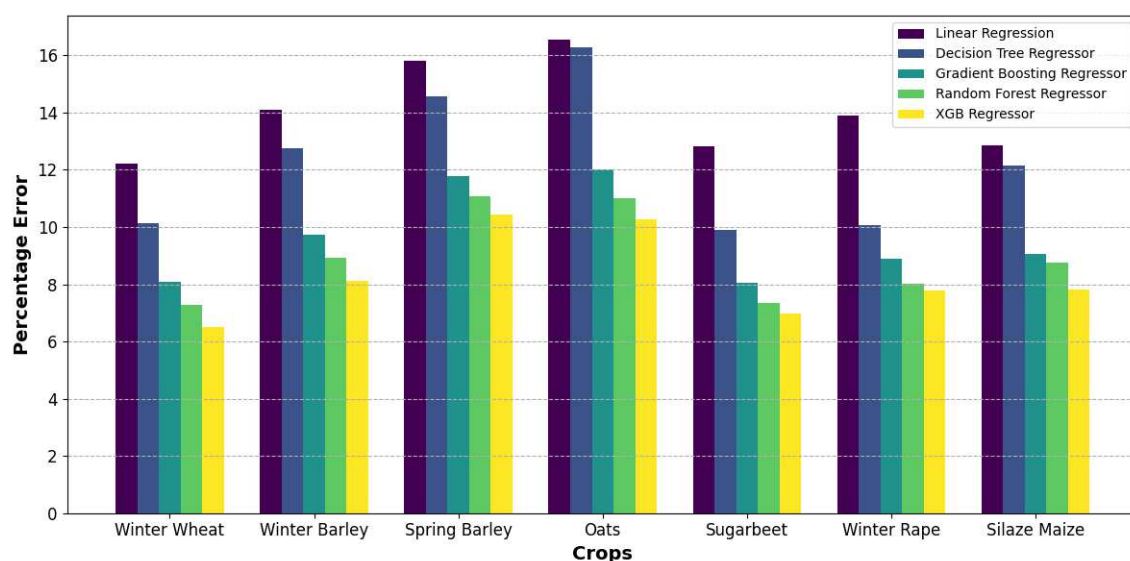


Figure 5.37: Comparison between the Percentage errors in the prediction of different models

The percentage error displays the difference between the actual crop production and the anticipated crop yield.

Similar to the RMSE percentage, a smaller percentage error indicates that the model is more accurate in predicting crop output, whereas a larger percentage error indicates that the model is less reliable.

When we examine the bar plot, we can see that the Decision Tree Regressor and Linear Regression models frequently have higher percentage errors than the Random Forest Regressor and XGBoost Regressor models.

However, the Random Forest Regressor and XGBoost Regressor models do not

perform as well as the Gradient Boosting Regressor.

Overall, the percentage error column gives us a good idea of how accurately each model forecasts crop production. However, as it was also for the RMSE error, the accuracy of a model may be significantly influenced by a number of other variables, such as the quality of the data, the features that are selected, and the hyperparameters that are used in the model.

Now let's look at the plot where a higher value indicates more accuracy.

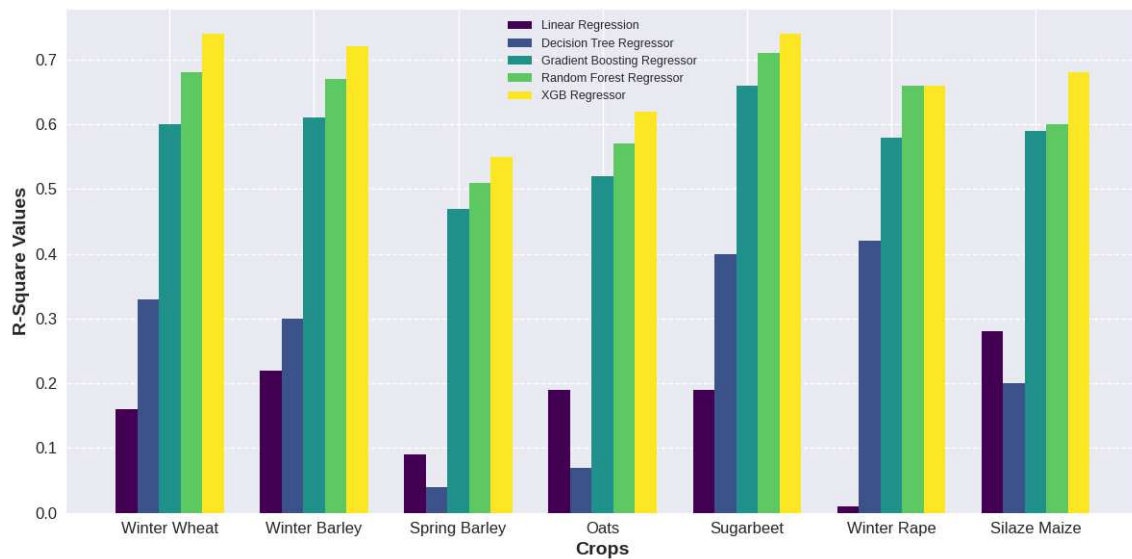


Figure 5.38: Comparison between the  $R^2$  values for  $y=x$  fit between actual and predicted yields by different models

The correlation coefficient is represented by the R-squared value obtained from the plots of the actual versus predicted crop values and the line  $y=x$ . When the R-squared number is close to 1, it indicates that the model can explain the majority of the variance in crop production. This suggests that the predicted values are more accurate. When the R-squared value is closer to 0, on the other hand, this indicates that the regression model performs poorly and does not explain a significant amount of the variance in crop yield.

Therefore, in our plots, an R-squared value that is greater than zero for a particular crop indicates that the matching regression model fits the data well and can forecast crop yield with greater accuracy, whereas an R-squared value that is less than zero indicates that the model's performance is below average.

When we look at the R-squared values in the table, we can see that the Gradient

Boosting Regressor and Random Forest Regressor models perform better than the other models in general, with R-squared values ranging from 0.51 to 0.74. Other models have R-squared values that fall somewhere in the middle of these two extremes. This demonstrates that these models have the potential to explain a sizeable percentage of the variability that is observed in crop yield statistics.

The Decision Tree Regressor and Linear Regression models, on the other hand, frequently perform poorly, with R-squared values ranging from 0.01 to 0.40. This demonstrates that these models are not as capable as the other models when it comes to recognizing the fundamental patterns that can be found in the data.

Due to the fact that the R-squared number only provides insight into the linear relationship that exists between the variables, it is not appropriate to rely solely on this statistic when evaluating the efficacy of the model. In addition to the R-squared value, it is important to take into account other performance metrics for the model, such as the Root Mean Squared Error (RMSE) and the percentage error. However, the values of R-squared that are presented in the table can be used to perform a preliminary analysis of how well the models match the crop yield data.

According on the RMSE percentage and Percentage Error columns, the Random Forest Regressor and XGBoostRegressor models seem to perform better than the other models. For the majority of crops, these models consistently provide lower values in both of these columns.

The higher performance of Random Forest Regressor and XGBoostRegressor may be ascribed to the fact that both models are ensemble approaches, which means they are generated utilizing numerous decision trees. This enables them to capture complicated non-linear correlations between input characteristics and output variables, as well as manage outliers and noise in the data.

Yet, because to their inability to successfully handle non-linear connections and outliers, the Decision Tree Regressor and Linear Regression models may have done worse. Linear Regression, in instance, presupposes a linear connection between the input characteristics and the output variable, which may not necessarily be the case in real-world circumstances. Gradient Boosting Regressor works well, but not as consistently as Random Forest and XGBoostRegressor, potentially because to overfitting on the training data.

It is crucial to remember, however, that the performance of these models will vary based on the dataset and task at hand. When choosing on the optimal model to utilize for a specific job, it is usually suggested to test numerous models and assess their performance on a holdout validation set.

# Chapter 6

## Further Work

As mentioned in the introduction, the further work plan is to incorporate NWP models.

As were the steps mentioned in the introduction, we have progressed till the modeling step, i. e. applying machine learning models to forecast crop production. The next step is to integrate the NWP and ML models together to get more accurate prediction of the crop yield.

We also haven't used the crop Phenology data yet, as it was available for only one crop (Winter Wheat). We hope to get access to the phenology data of thee other crops too. The phenology parameters are expected to increases the prediction results significantly as we will have the knowledge of the quickly the crop grows and quickly it gets ready for harvesting according to different weather and soil conditions.

# References

1. Srivastava, A. K. *et al.* Winter wheat yield prediction using convolutional neural networks from environmental and phenological data. *Scientific Reports* 2022 12:1 **12**, 1–14. ISSN: 2045-2322. <https://www.nature.com/articles/s41598-022-06249-w> (1 Feb. 2022).
2. Wang, Y., Zhang, Z., Feng, L., Du, Q. & Runge, T. Combining Multi-Source Data and Machine Learning Approaches to Predict Winter Wheat Yield in the Conterminous United States. *Remote Sensing* **12**, 1232. ISSN: 2072-4292 (8 Apr. 2020).
3. Khaki, S., Pham, H. & Wang, L. Simultaneous corn and soybean yield prediction from remote sensing data using deep transfer learning. **11**. <https://www.nature.com/articles/s41598-021-89779-z> (2021).
4. *USDA Quick Stats* United States Department of Agriculture, Accessed on April 10, 2023. <https://quickstats.nass.usda.gov/>.
5. Revisiting Deep Learning Models for Tabular Data. <http://arxiv.org/abs/2106.11959> (June 2021).
6. Deutscher Wetterdienst. *Phenology Annual Reports* Accessed on April 10, 2023. [https://www.dwd.de/DE/klimaumwelt/klimaueberwachung/phaenologie/daten\\_deutschland/jahresmelder/jahresmelder\\_node.html](https://www.dwd.de/DE/klimaumwelt/klimaueberwachung/phaenologie/daten_deutschland/jahresmelder/jahresmelder_node.html).
7. Copernicus Land Monitoring Service. *CORINE Land Cover 2006* <https://land.copernicus.eu/pan-european/corine-land-cover/clc-2006>.
8. Bundesanstalt für Geowissenschaften und Rohstoffe. *BUEK1000 Soil Mapping Project* Accessed on April 10, 2023. [https://www.bgr.bund.de/EN/Themen/Boden/Projekte/Informationsgrundlagen\\_abgeschlossen/BUEK1000/BUEK1000\\_en.html](https://www.bgr.bund.de/EN/Themen/Boden/Projekte/Informationsgrundlagen_abgeschlossen/BUEK1000/BUEK1000_en.html).
9. Webber, H. *et al.* No perfect storm for crop yield failure in Germany. *Environmental Research Letters* **15**. ISSN: 17489326 (10 Oct. 2020).
10. Statistische Ämter des Bundes und der Länder. *Genesis Online* Accessed on April 10, 2023. <http://www.regionalstatistik.de/genesis/online>.



## Document Details

<b>Title</b>	MS_Thesis_Prantik_final.pdf
<b>File Name</b>	MS_Thesis_Prantik_final.pdf
<b>Document ID</b>	f35484d0faf8404ba0d4b538eefb267b
<b>Fingerprint</b>	59822f3840be5a2dfa9688baa0561c37
<b>Status</b>	Completed

## Document History

<b>Document Created</b>	Document Created by Manmeet Singh (manmeet.cat@tropmet.res.in) Fingerprint: 1182042abc89252a21d8f67cff3dd524	May 13 2023 05:35AM UTC
<b>Document Signed</b>	Document Signed by Manmeet Singh (manmeet.cat@tropmet.res.in) IP: 122.169.16.206 	May 13 2023 05:35AM UTC
<b>Document Completed</b>	This document has been completed. Fingerprint: 59822f3840be5a2dfa9688baa0561c37	May 13 2023 05:36AM UTC