National College of Ireland

Predicting Wheat Yield Through Ensemble Machine Learning Approach

MSc Research Project
Data Analytics

Ashok Saravanan Sundarrajan
Student ID: x21204764

School of Computing
National College of Ireland

Supervisor: Prashanth Nayak

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Ashok Saravanan Sundarrajan
Student ID:	x21204764
Programme:	Data Analytics
Year:	2023
Module:	MSc Research Project
Supervisor:	Prashanth Nayak
Submission Due Date:	14/08/2023
Project Title:	Predicting Wheat Yield Through Ensemble Machine Learning Approach
Word Count:	7872
Page Count:	21

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Ashok Saravanan Sundarrajan
Date:	18th September 2023

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Predicting Wheat Yield Through Ensemble Machine Learning Approach

Ashok Saravanan Sundarrajan
x21204764

Abstract

Agricultural crop yield productivity plays a significant role in ensuring we meet surging food demand for the growing population in a country, especially in developing countries like India where food crops are grown at high temperatures. Wheat which is a staple food crop in India, plays a major role in the nation's food security and economic stability. Wheat being a majorly consumed crop in India, this study uses a sophisticated ensemble approach coming from different models like Random Forest(RF), Support Vector Machine(SVM), and Decision Tree(DT) to predict the yield and compare the performance of this approach with that of models individually. The crop data used in this study is categorized based on state, district and has seasonal yields ranging from 1997 to 2020. Preliminary results indicate that the ensemble approach used in this study surpasses the individual models in terms of both stability and accuracy over each iteration. this paper discusses the model's performance over ensemble approaches including voting average and stacked ensembling. the voting average has an RMSE of 0.33, however, the Stacking ensemble method achieves the lowest Root Mean Square Error (RMSE) of 0.28 after hyper parameter tuning compared to the voting average method, which is significant for a wheat yield prediction. This innovative approach holds promise in enhancing wheat yield predictions, facilitating informed decision-making for farmers and policymakers, and playing a crucial role in addressing the demand for food crop yield and its sustainability.

1 Introduction

In the rapidly evolving world, the escalating demand for food, driven primarily by population growth, especially in developing countries, underscores the critical role of agriculture. As the backbone of many economies, agriculture's capacity to meet this surging demand is contingent on the reliability and timeliness of crop production. However, the inherent uncertainties in demanding commodities such as wheat provide considerable hurdles for farmers, sometimes resulting in sub-optimal outcomes that struggle to keep up with rising food demand. Wheat crop yield estimation, therefore, emerges as an important part of agriculture, offering insights into crop growth, market pricing, and harvest data. Accurate and accurate production projections can pave the way for more sustainable crops, allowing farmers to maximize yields while lowering costs and increasing profitability.

This study aims to address the challenges in Wheat yield prediction by proposing an innovative ensemble approach that integrates the strengths of Random Forest, Support Vector Machine and Decision Tree machine learning models with a focus on wheat, a

staple food crop in India. The objective is to investigate whether this ensemble approach can enhance the accuracy and sustainability of wheat yield predictions, thereby contributing to food security and economic stability. While numerous studies have explored crop yield prediction using traditional methods such as regression analysis and models like Random Forest, Support Vector Machine, and Artificial Neural Networks (ANN), their effectiveness is often limited by the complex and dynamic nature of agriculture. Moreover, the focus on specific crops in many studies restricts their applicability across different crops.

The research highlights importance of accurate crop yield prediction extends beyond agriculture, influencing aspects such as crop management, food security, and market price determination. It provides valuable insights that guide farmers in making informed decisions about fertilization, irrigation, and harvesting, considering various factors like climatic conditions. It also aids policymakers and food security analysts in making decisions about wheat crop distribution and food security measures.

Research Question. The above research problem motivates the following research question: Can predicting wheat yield using hybrid or Ensemble Model combining Support Vector Machine, Random Forest and Decision Tree models provide significantly greater accuracy than each model individually?

Following this Introduction section, Section 2 discusses the Related Works which are relevant to this study by researchers and experts, Section 3 comprises the Methods and approaches carried out during this research, Section **??** depicts the Design Specification of this research, Section 5, discusses the Implementation, Section 6 describes the Evaluation and Discussion followed by Section 7 which discusses on Conclusion and Future Work of this study.

2 Related Work

A wide range of farming and computational studies have focused on forecasting wheat yield using ensemble machine learning strategies. These strategies outperform existing techniques in terms of predicting capability. To construct reliable models for crop yield data, this segment sheds light on the work carried out by specialists who utilize an assortment of machine learning and artificial intelligence tools. The findings, comparisons, and innovation in this research project are based on the knowledge gathered in this section. There is a huge, yet mutually supportive approaches, from conventional regression based models to deep learning techniques that are very much advanced and used in predicting outcomes of various problems. Each of the approaches has its own strengths and weakness to it. The chance to bring together these diverse approaches into a single, ensemble machine learning model is an exciting prospect in the field of crop yield prediction. As we delve deeper into this field, the rich pool of knowledge from past studies will serve as an invaluable resource to steer our research direction are derived from the studies discussed in this section.

2.1 Traditional Approaches to Crop Yield Prediction

This study by Lobell et al. (2011) examines the impact of climate change on global crop production from 1980 to 2008, primarily focusing on maize, wheat, soybeans, and rice. They argue that the global production of maize and wheat has declined due to climate trends. This paper cleverly uses weather models to see how climate shifts might change

how much crops we can grow. But, it could be better in some ways. For one, they didn't look at the United States when studying heat patterns. That might mess up the findings because the US grows a lot of the world's crops. Also, the time they looked at (1980-2008) might not show the full effects of climate change on crop amounts over a long time. So, even though this paper adds good stuff to what we know about climate change and crops, we have to keep these issues in mind when we think about what they found.

Ray et al. (2015) paper offers a comprehensive study on how climate variability impacts global crop yield variability. While the analysis is detailed, the authors rely primarily on time-series data for their analyses which doesn't necessarily capture the nuances of local agricultural practices or policy changes that could also impact yields. This study leans a lot on stats, which are great for spotting links, but that doesn't mean one thing causes the other. Like, they didn't think about other big stuff that can change how much crop you get. This includes how good the soil is, bugs eating the crops, or how using stuff to make crops grow better can affect the outcome. The authors do admit that climate's impact on yield is location-specific, however, a more detailed examination of these location-based differences might have enriched the paper. Despite these limitations, this paper contributes meaningfully to our understanding of how climate variability might affect crop yields globally and could serve as a foundational reference for future research in this area.

Iizumi et al. (2014) in their study has developed an approach which has most valuable findings. They show there is an increase in yield instability in the Southern Hemisphere, indicating areas of potential future food insecurity. Despite this, the paper doesn't deeply analyze the specific causes of yield instability or consider how socio-economic factors might contribute. The paper concludes that the rise in yield instability could be due to recent climate change. However, it acknowledges that understanding of how climate change impacts crop yields is limited. This indicates a need for future research to close this knowledge gap which paves way for more sophisticated approaches later in this area.

2.2 Machine Learning For Crop Yield Prediction

This paper (Sellam and Poovammal; 2016) certainly puts forward a helpful regression analysis approach for predicting crop yield based on factors like Annual Rainfall, Area under Cultivation, and Food Price Index. However, it's worth noting that the study is limited by its use of only three factors over a relatively short period of 10 years. It also lacks a rigorous model validation or testing, and does not address the potential interplay among the variables considered. The R2 value of 0.7, although decent, still leaves room for error and uncertainty. Furthermore, the fact that the research only considers one crop (rice) limits its generalizability. In future studies, it would be beneficial to consider a wider range of influencing factors, extend the duration of data collection, include more crops, and perform robust validation tests. It's clear, however, that the research provides a good starting point for more complex models

This research (Maimaitijiang et al.; 2020) showcases a valuable exploration into UAV-based multimodal data fusion for soybean yield prediction using deep learning techniques. The study's high point is its integration of different data types and demonstration that a Deep Neural Network (DNN) outperforms other methods, like Partial Least Squares Regression, Support Vector Regression, and Random Forest Regression. However, the paper does have some limitations. The focus solely on soybeans limits its scope and generalizability to other crops. Also, the study's success heavily relies on having a signifi-

cant number of input features, which might not always be feasible in real-world situations. Lastly, although the results are promising, the paper suggests more testing on various crops and in different environmental conditions, implying the work is in an early stage. Overall, the study holds potential and encourages further research, but it should be interpreted with care due to its limitations

The author Prasad et al. (2021) an interesting application of the Random Forest (RF) machine learning algorithm for predicting cotton yield in Maharashtra, India. The research made a promising effort to harness satellite-derived variables and historical crop yield data to inform the model. However, the paper could benefit from a more rigorous examination of the methodological choices. comparing the RF model's performance with traditional linear regression models was a good step. However, including comparisons with other machine learning techniques, like Support Vector Machines or Neural Networks, would provide a more comprehensive view of the RF model's advantages. In essence, while the study shows promise in crop yield prediction, further investigation is required to solidify its findings and potentially improve the model's robustness and accuracy.

2.3 Remote Sensing and Meteorological data based Approaches

This study (Zhou et al.; 2022) presents an novel approach in predicting wheat yield in the regions of China using integrated climate, remote sensing data, and machine learning techniques. The methodology applied and the comparison of three machine learning models (Random Forest, Support Vector Machine, and LASSO) provide a comprehensive approach to understanding yield prediction. However, the study also presents some areas for improvement. First, the use of historical data from 2002 to 2010 might not reflect the current climatic conditions, considering the rapidly changing climate. More recent data could have improved the relevance of the study. Additionally, the use of county-level data may mask micro-level variations in yield prediction and management practices. However, the study is not without limitations. there is a limited exploration of why SVM outperformed other models. However, the limitations identified need to be considered for future research to improve the reliability and applicability of such models.

The main objective of this study (Shah et al.; 2021) is to provide a significant contribution to the field of agricultural planning, demonstrating the power of machine learning in predicting crop yield with 90 percent accuracy. By blending meteorological and remote sensing data with machine learning techniques like XGBoost and Gradent Boost, the authors show the potential for making agricultural decision-making more data-driven. The experiment's specific focus on predicting rice yields in Tamil Nadu, India, reflects a practical application of their methodology. However, the study does not thoroughly detail the feature engineering and outlier correction methods applied, which would be crucial for replicating the study. The reliance on prehistoric data for validation is another limitation, as future climatic conditions may not mirror the past, especially in light of accelerating climate change. Nonetheless, this work represents an important step forward in the utilization of machine learning and remote sensing for agricultural yield predictions.

An insightful approach to wheat yield prediction, utilizing online proximal soil sensing, satellite imagery, and machine learning models has been proposed by (Pantazi et al.; 2016). By using Supervised Kohonen Networks (SKN), Counter-Propagation Artificial Neural Networks (CP-ANN), and XY-Fused Networks (XY-F), it assesses the potential for predicting yield based on soil parameters and vegetation indices. The paper triumphs

in achieving a high accuracy in low yield prediction (91.3 percent), demonstrating the applicability of such an approach. However, it appears to lack thorough explanation of the selected machine learning models and the reasoning behind their selection. It is also not clear how this model would perform under different cropping systems or climatic conditions. While NDVI was found to be more correlated with yield, the paper could also explore the impact of other remote sensing indices. Despite these limitations, the paper sets an important groundwork for future research in precision agriculture.

2.4 Ensemble Methods for Crop Yield Prediction

An intriguing approach to predicting crop yield using an ensemble of Decision Tree and AdaBoost regressors, proving effective with a commendable 95.7 percent accuracy has been deeply discussed in (Keerthana et al.; 2021). However, it leaves room for critique. The paper would have benefited from a deeper exploration of why these specific machine learning algorithms were chosen for the ensemble model. Moreover, it overlooks the variability in regional farming practices and local environmental factors which could affect crop yield. There is also an unspoken assumption that past weather patterns will predict future ones, a concept increasingly challenged in the era of climate change. Lastly, the usage of national-level data may mask localized variations crucial to crop yield. Despite these, the paper provides a solid foundation for further refinement of ensemble machine learning models in agriculture.

This paper (Cao et al.; 2022) does a commendable job exploring a Machine Learning-Dynamical Hybrid model for sub seasonal-to-seasonal winter wheat yield prediction in Northern China. However, it leaves a few areas unattended. The selection of machine learning algorithms and the exclusive focus on the northern China region raise questions about the universal applicability of the model. Additionally, while the approach stands out in outperforming the observational climate data model, the significance of its superiority and the robustness of the model aren't well-discussed. Lastly, the lack of a more exhaustive cross-validation with other hybrid models or a comparative study with other regions leaves some room for skepticism. Nevertheless, the paper provides valuable insights into a novel yield prediction method that could potentially aid farmers and policy-makers.

Heremans et al. (2015) to make early predictions for winter wheat yield in the Huaibei Plain, China using ensemble tree machine learning methods, mainly Boosted Regression Trees and Random Forests, along with remote sensing and meteorological data. The paper deserves praise for its rigorous and comprehensive approach, and the results achieved do indicate a promising potential for these machine learning methods. However, its reliance on just 12 years of data for a highly variable phenomenon like crop yield raises concerns about the robustness of the models. The study's variable selection process also seems overly complex and could potentially overlook significant interactions. Lastly, while the results are promising, the paper lacks a clear comparative analysis with traditional crop yield prediction methods, leaving readers in doubt about the real-world advantages of employing such complex models. It would be beneficial if the authors considered these aspects in their future work.

The research (Prodhan et al.; 2022) delves into a crucial issue, predicting drought impact on crop yield over South Asia, using ensemble machine learning (EML). The use of SPEI drought index and CMIP6 climate models is commendable, providing robust predictions. However, there are limitations. The study focuses only on the magnitude,

intensity, and duration of future drought, missing out on the spatial distribution aspect. The EML approach, combining RF and GBM, while being innovative, isn't compared with other potential ensemble models, leaving a gap in evaluating its relative performance. A more extensive analysis could have added depth. Additionally, the authors' claim about helping policy agencies feels unsubstantiated as the study doesn't elaborate on how the data can be translated into actionable policies. Future work should consider addressing these points for a more well-rounded research outcome.

A significant contribution to agricultural studies with a novel machine learning approach combining Random Forest (RF) and Support Vector Machine (SVM) for enhanced crop classification and area estimation has been done by Löw et al. (2012) . The integration of RF feature importance for SVM feature selection is an innovative idea. However, the authors fail to justify their choice of RapidEye data as the sole satellite data source. They also do not examine the applicability of their model across different types of irrigation systems, which might limit its generalizability. The decision to only present improvements in accuracy, while neglecting other potentially useful metrics, weakens the study's comprehensiveness. Furthermore, although they claim increased user's and producer's accuracy, there is little elaboration on what this would mean in a real-world application, leaving the reader to speculate about its practical implications.

3 Methodology

The research methodology employed in this study is a blend of various machine learning techniques, namely Support Vector Machines (SVM), Decision Trees (DT), and Random Forest (RF). This research essentially gather some insights from the data acquired, analyse the different patterns, then provide a meaningful outcome. These techniques were chosen due to their proven effectiveness in handling high-dimensional data, robustness against over fitting, and capability to capture complex relationships in the data. Firstly, In Data Acquisition - the dataset selection details are discussed in this step, followed by data pre-processing required for the analysis, feature extraction, and transformation phases - the dataset gathered was in a text format, techniques like Label encoding and Time series based cross validation are implemented are discussed and in the Model and Evaluation stages – different machine learning and ensemble models are implemented and evaluated on the wheat among the different crops is discussed. Finally, the results are acquired by comparing with individual methods and evaluating the models using metrics.

3.1 Data Acquisition

This dataset that we intend to use [1] is a combination of agricultural crop data comprising of 345,336 records and it also has eight distinct fields. the data provides an in depth view of crop yields during different seasons in consecutive years from 1997 to 2022 across multiple states and districts in India. Also, the crop has what amount of yield in a year with a given area of land which will be useful information, particularly when developing ensemble model which is core objective of this project. since the record in this dataset has an yield information for particular year, it is safe to say that data is time-series in nature. this characteristics will essentially help us uncover various patterns in the crop over different seasons and find the pattern changes over time, providing valuable insights

[1] `https://data.world/thatzprem/agriculture-india`

State	District	Crop	Crop_Year	Season	Area	Production	Yield
Andhra Pradesh	ADILABAD	Wheat	1997.0	Rabi	3600.0	2000.0	0.56
Assam	DHEMAJI	Wheat	1997.0	Rabi	3500.0	4463.0	1.28
Haryana	KARNAL	Wheat	1997.0	Rabi	163000.0	577000.0	3.54
Assam	DHUBRI	Wheat	1997.0	Rabi	18270.0	25101.0	1.37
Haryana	KAITHAL	Wheat	1997.0	Rabi	160000.0	570000.0	3.56
...	...	...	...	...	...	...	...
Chhattisgarh	SURAJPUR	Wheat	2019.0	Rabi	6538.0	8861.0	1.36
Jammu and Kashmir	KATHUA	Wheat	2019.0	Rabi	43130.0	104473.0	2.42
Haryana	PALWAL	Wheat	2019.0	Rabi	100700.0	487100.0	4.84
Assam	UDALGURI	Wheat	2019.0	Rabi	379.0	483.0	1.27
Jharkhand	WEST SINGHBHUM	Wheat	2019.0	Rabi	1678.0	2084.0	1.24

Figure 1: Overview of dataset

for our research beyond our trivial analysis. The following figure illustrates the first and last few records in our dataset.

3.2 Data Pre-processing

In data pre-processing step, the dataset is subject to a sequences of transformations to ensure its readiness for the subsequent stages in building our ensemble model. Firstly the the columns names of dataset are not properly formatted so we have formatted it as per our requirement. Then the dataset is checked for the null values and it was found out to be present in our dataset. Their NA values are then dropped as we no need them for our further analysis as it may introduce bias to our dataset. Removing the missing values and redundant values are crucial as we are trying to predict the accurate yield of crop in our research. now that we have dataset which is more cleaned and we are further extracting the data we want for building our model.

In the next step, we further clean the data by removing the outliers. the outliers are definitely present in our dataset and we have to remove them. we are using Inter Quartile Range(IQR)[2] methods to detect the outliers and remove it. this will give us a clear idea of how we are going to use the data. the outliers are shown using the bar chart as illustrated in the figure 2. Area, Production, and Yield are numerical variables so we plot it using a box plot. the box plot shows that there are several outliers present in the Area and Production of crops. In IQR method, we see Q1 is the 25th percentile of the data and Q2 is the 75th percentile of data. the IQR is Q3 - Q1. the Lower Outlier Threshold(LOT) is denoted by formula Q1 - 1.5 x IQR and Upped Outlier Threshold(UOT) is denoted by Q3 + 1.5 x IQR. then the cleaned data points are denoted by the formula,

$$\text{Data Cleaned of Outliers} = \text{Data points where } (\text{value} > \text{LOT} \,\&\&\, \text{value} < \text{UOT}) \quad (1)$$

using this approach, we then are extracting the wheat data from the cleaned data and perform a series of transformation steps.

[2] https://en.wikipedia.org/wiki/Interquartile_range

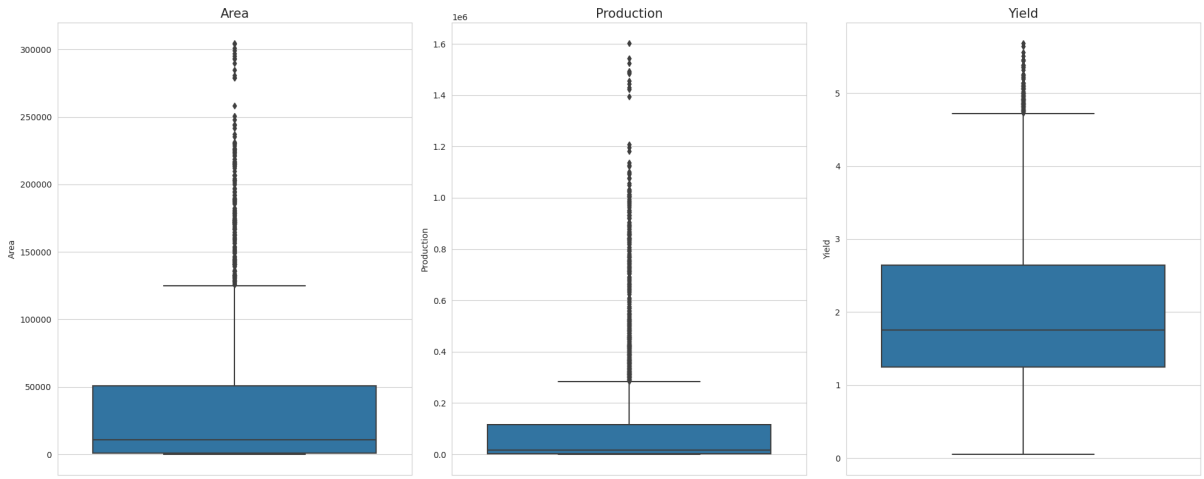

Figure 2: Outliers in data

3.3 Exploratory Data Analysis

In this stage, the crop data that we have been meticulously analysed and visualized using different visualization techniques to attain insights and meaningful information about the different crop variants and patterns of the crop and correlations. The visualization was mainly done to understand the distribution of yield for each crop variant, to analyze the crop variant column data distribution and the yield distribution based on the area, and to analyze the highest number of crop yield instances for each season using Python visualization libraries such as pyplot, matplotlib and seaborn. The number of unique crops in the particular season, their patterns and variation, and the distribution of crops, and also their correlation columns is also taken into account.

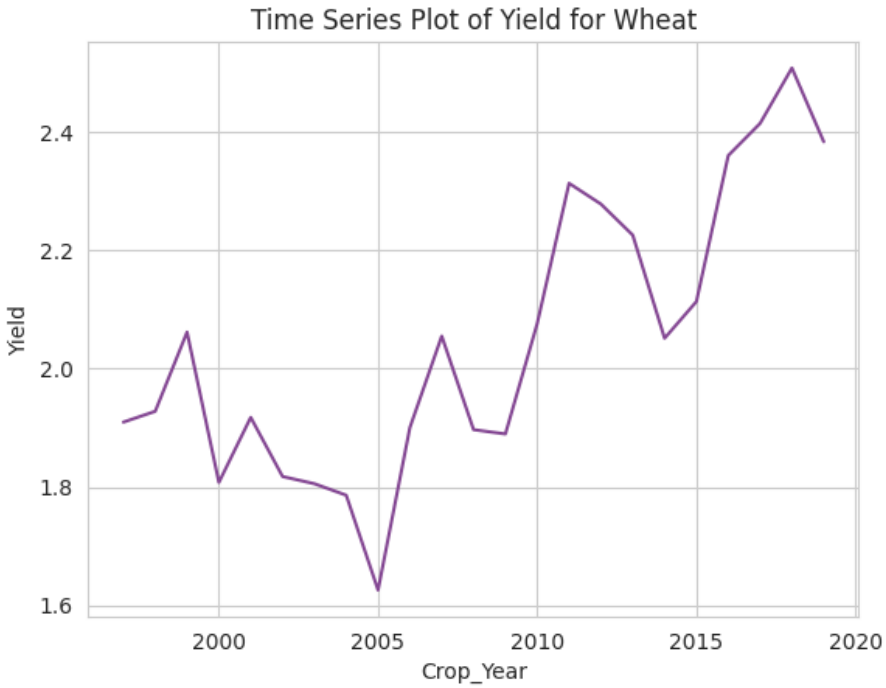

Figure 3: Distribution of Wheat

In the Figure 3, we have visualized the distribution of wheat over time series plot which represent how wheat is giving yield over the given years. It is obvious that wheat has lowest yield due to drought that happened in 2005 and highest yield around the end of 2019. Since then the overall distribution of wheat has fallen.

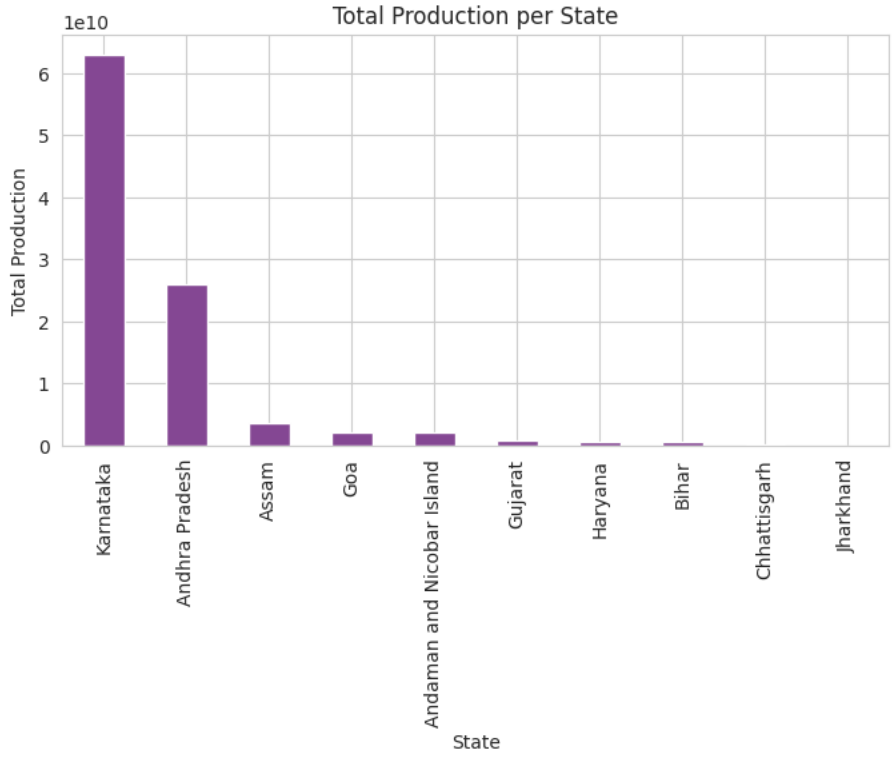

Figure 4: Total Production Per State

In the Figure 4, total wheat production in each state has been clearly visualized. the plot shows that there are 10 different states and each has different wheat production value which represent how wheat is giving production in different states . from the distribution we can conclude that there are lower production of wheat in Jharkhand and Higher production in the state of Karnataka. Andra Pradesh state has significant yield next to Karnataka while other states produce significantly less crops.

Total wheat yield in top 10 states different states has been shown in the Figure 5. the figure shows that there are 10 different states and each has different wheat production value which represent how wheat is giving production in different states . It is evident from the distribution that that there are less number of wheat yield in Madhya pradesh and Chandigarh is the only state which leads in production of wheat. Other areas like Punjab and Haryana are next top producers of wheat. Higher yield means also it has higher land area and production is more. As a result the wheat consumption is more. However, there are lot of other areas still face lot of demand for wheat yield and its highly dependent on various factors like temperature, region, climate, soil and rainfall etc. In such areas where wheat production is less, our ensemble model can greatly improve the efficiency of the yield and ensure the crop sustainability.

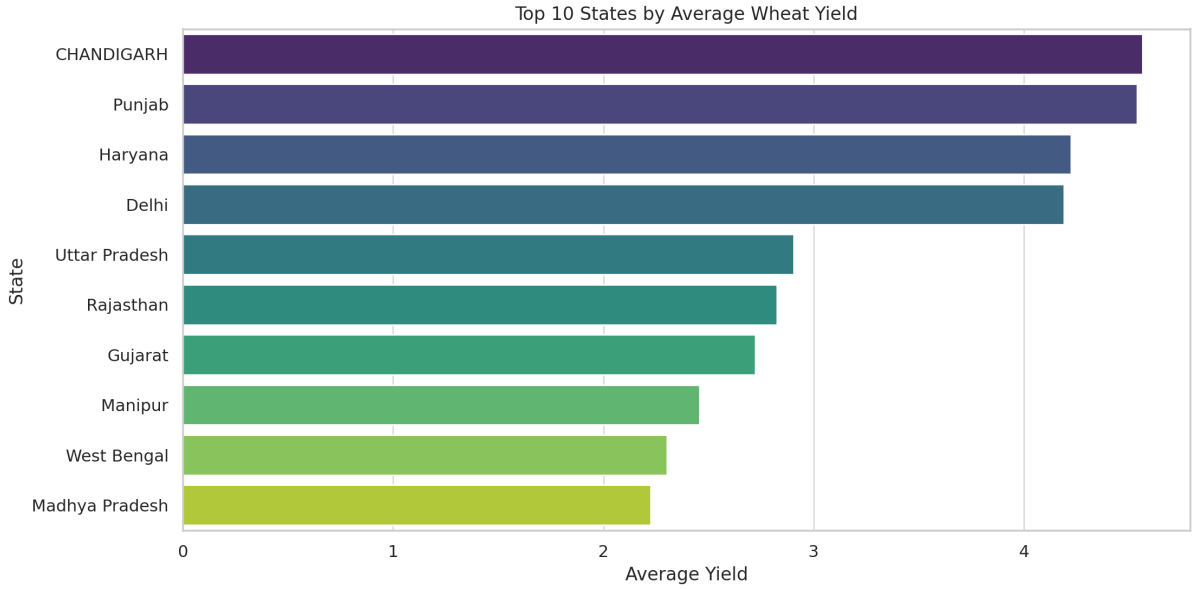

Figure 5: Top 10 States by Average Wheat Yield

3.4 Data Transformation

The data chosen are in time series in nature and it is crucial for us to understand the pattern of the crop and its trends over the season so that we can use that to build ensemble model and compare it to individual models to fully comprehend the performance and perform successful prediction of yield. In data transformation stage, we filter out the wheat from different crops that we have. for data sampling we choose data from 1997 to 2015 for training and 2015 to 2020 for test data which will be used for individual models input to predict the outcome. here yield variable is the independent variable and remaining variables are dependent variable. In this research to understand the crop nature we have undergone series of data transformation techniques and extracted the necessary features out of it which are discussed below.

3.4.1 Label Encoding

The data set we have chosen has to be given as input to our model. This input is essential for a model to predict the outcome better. Model has to successfully predict without any bias with the data. 'State', 'District', 'Season', and 'Crop' columns in our data are categorical variable which is good fit for label encoding. The Label encoder method in scikit-learn package helps in converting the categorical variables into numerical values. By transforming these values into numerical values, we can make the model effectively predict the results. one main advantage of considering the label encoding in this research is that it doesn't unnecessarily create different data dimensions which is trivial in other methods like one hot encoding. Most machine learning algorithms can't work directly with categorical data represented as text, and require numerical values instead. even though some machine learning models like decision trees and random forests can handle categorical values, most other algorithms require numerical inputs.

3.5 Modelling

The primary objective of this research is to forecast the yield of the crop variants, particularly wheat using different machine learning models and build a hybrid model using those models and compare it with the models individually. The Random Forest(RF), Support Vector Machine(SVM), and Decision Tree(DT) were the models that is being used to build an ensemble model. We have done timeseries based cross validation for 5 different time splits to see the models performance and got a significant results compared to that of individual models. Hyper parameter tuning is done to ensure best fit parameters that will yield models best performance. Finally we also built hybrid models combing these models to predict the final outcome and compare the results.

3.5.1 Random Forest vs Support Vector Machine vs Decision Tree

Random Forest(RF) is more robust and user-friendly machine learning model that generates good outcome even without hyper-parameter tuning. Initially we split the dataset into training and test dataset based on the year that we want to predict. The data from 1997 to 2015 is considered for training the model and then from 2015 to 2020 we consider the test data set. With this data we train the model and in our first iteration we have got the RMSE value of 0.55 which is good mean error for a model for such dataset which is timeseries in nature.

Next we try running the same dataset with Support Vector Machine(SVM) and interpret the results. This time we get RMSE of 0.52 which is slightly better error margin than RF but not great results. The reason we choose SVM for this is that it supports both linear and non-linear regression and be more suitable for dataset with time series in nature.

In the third iteration, we slightly take different approach by considering Decision Tree(DT) because it is simplest model to solve the problem and predict results and also provides solid baseline model to compare with other advanced models. This time we have got only RMSE of 0.35 which signifies the model outperforms other two models. This model might be a great model to predict the outcome for single crop and also majorly used model.

3.5.2 Ensemble Models

In addition to the individual models built, the primary objective of this research is to study the efficiency of the models that are combined and to see how it performs the result. This outcome is then compared with outcomes of the models that were built individually to see the performance of our hybrid model. In the Figure 6 it is clearly described that the label encoded training set is run on three different models to get the individual predictions and then those out comes are combined and passed on to ensemble model. hyper parameter tuning is done to ensure the best fit parameters for the model to perform at its best. This ensemble model takes in two different methods, Voting Average method which takes in average of the models outcomes and select the best outcome wherein we got average of 0.33 which is more or less similar to Decision Tree outcome. Next we try Stacking Generalization method which take the outcome of model and use different regression model to predict the results. Here we used Linear Regressor model as estimator for the approach and we got an RMSE of 0.28 which is significant result and outperformed all other models.

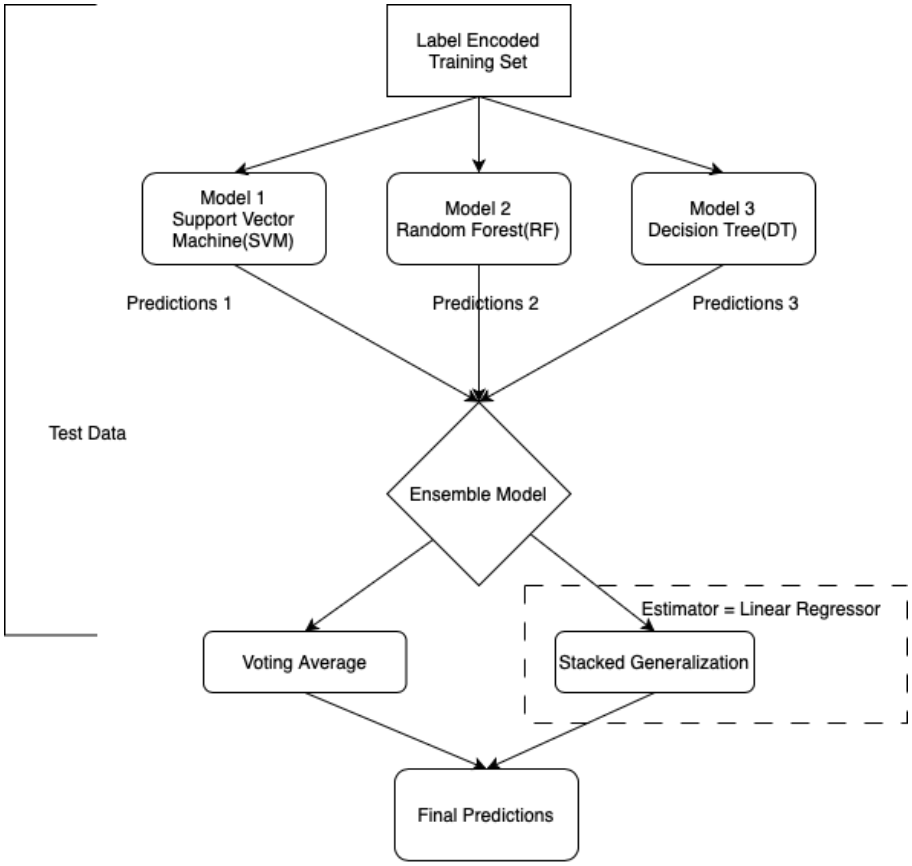

Figure 6: Ensemble Model - RF, SVM, DT

3.6 Evaluation

In this research, we have used a hybrid model and standalone models to do a comparative analysis. The Evaluation metric that we have used for our analysis is Root Mean Square Error(RMSE). The RMSE value is one of the important metrics for evaluating regression problems. Since our model has data that is time-series in nature we considered RMSE as an effective metric for predicting scores. It is simply measuring the error of model in predicting quantitative data. The RMSE score is evaluated as given by formula below,

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

where

- y_i is the actual value for the ith observation.
- $\hat{y}_i$ is the predicted value for the ith observation.
- n is the total number of observations.

In this study, the RMSE for standalone models is 0.55 for Random Forest and 0.52 for Support Vector Machine and 0.35 for Decision Tree which denotes Decision Tree is the best-performing model on our data with the lowest error. In the time series-based cross-validation approach we tried splitting the data based on 5 splits and accumulated

the RMSE and took the average of that RMSE which is 0.30 for the Random Forest, 0.41 for SVM, and 0.37 for the Decision Tree. In this iteration Decision Tree performed badly compared to the other two models. when we compare this with the RMSE of two methods of our ensemble model which is voting average and stacked generalization it has RMSE of 0.33 and 0.28 which has outperformed our standalone model. The stacked generalization method used Linear Regression as the estimator model through and we are able to achieve this level of the lower error bound for our ensemble model.

Alongside RMSE, another critical metric used to evaluate the performance of the models was the Mean Absolute Error (MAE). MAE offers a direct interpretation of how well a model is performing, as it represents the average absolute difference between the observed actual outcomes and the predictions made by the model. Lower values of MAE indicate better predictive accuracy. The mathematical representation of MAE is given by:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{3}$$

where

- y_i is the actual value for the ith observation.
- $\hat{y}_i$ is the predicted value for the ith observation.
- n is the total number of observations.

For the standalone models in our study, the MAE values were as follows: Random Forest had an MAE of 0.42, Support Vector Machine had an MAE of 0.31, and Decision Tree achieved the best result with an MAE of 0.17. This indicates that, on average, the Decision Tree model's predictions were closest to the actual values. In comparison, the ensemble models' MAE scores were lower, with the Stacked Generalization method achieving the lowest MAE of 0.10, demonstrating its superior predictive capability.

Another important metric in our research is calculation of R^2 Score, also known as the coefficient of determination. This metric provides an indication of the goodness of fit of a set of predictions to the actual outcomes. An R^2 score of 1 indicates that the model's predictions perfectly match the actual outcomes, whereas a score of 0 indicates that the model does no better than simply predicting the mean of the actual outcomes for all observations. The R^2 score can be mathematically represented as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}i)^2}{\sum i = 1^n(y_i - \bar{y})^2} \tag{4}$$

where:

- y_i is the actual value for the ith observation.
- $\hat{y}_i$ is the predicted value for the ith observation.
- $\bar{y}$ is the mean value of the observed data.
- n is the total number of observations.

In our study, the standalone models exhibited the following R^2 scores: Random Forest achieved a score of 0.73, Support Vector Machine had a score of 0.79, and the Decision Tree model outperformed the other two with a score of 0.91. In the ensemble methods, the Stacked Generalization approach recorded an outstanding R^2 score of 0.97, indicating that it accounted for 97% of the variability in the data, further underlining its efficacy in predicting wheat yield.

4 Design Specification

The different stages that are carried out in this research are represented in sequential order. Initially, the data is stored in a data layer which is an access point and a single source of truth for the data transaction and its used for pre-processing. The data is loaded and has been used from the data sources like the data world website. The data is injected into Google Colab for pre-processing steps and transformation. Then we perform exploratory data analysis. we do modeling for the label-encoded dataset. For this model, we do data sampling by dividing the dataset based on the crop year. For the training dataset, we have used data from 1997 to 2015 and for the test dataset, we use 2015 to 2022. As part of modeling, we have done model building with a standalone model with and without hyperparameter tuning and we have also time-done time series-based cross-validation for these models. Finally, we build an ensemble model is built using Linear Regressor as an estimator for the stacked generalization method. We use RMSE as an evaluation metric for comparative analysis. the design specification that's used is given below.

IDE	Google Colab, Jupyter Notebook
Programming Language	Python v3.9.1
Modules	Matplotlib, Pandas, Numpy, Scikit-learn
Computation	GPU
Number of GPU	1
GPU Type	Tesla K80 GPU-12GB

Table 1: Design Specification

5 Implementation

For this research we have taken data from data world website which is one of the finest data repository. We have used Google Colab as our development environment as mentioned in Initially we load the data and import required libraries and packages. Next the data is cleaned for NA values which is not required for our analysis. Libraries like Pandas are more useful in pre-processing and transformation of data. We have done the data splitting based on the crop year for train and test data. The programming language that is being used here is Python which is great for data analysis and has wide range of data libraries that supports our research.

The entire model building phase and implementation are performed based on the transformed data. In the transformation phase we have used Label encoding method from scikit-learn to transform the data into set of numerical values for the categorical variables

like State, District, Crop and Season. This will help the model to predict the results accurately which will be primary objective of this research. The novelty of this approach lies in the ensemble model which takes in three different models(RF,SVM,DT) and also timeseries based cross validation is done to ensure the performance of the standalone models, in past studies which combines the different models are done but they have their own limitations which are discussed in related works.

5.1 RF vs SVM vs DT with and without Hyper Parameters

Firstly, RF model was built using the default parameters and evaulated using RMSE to see how it performs on our data and by injecting the best parameters max depth = None, nestimators = 100, we tune the model to give the best performance. the test set is then evaluated and predicted using the regression evaluation metrics.

Secondly we have implemented SVM with its default parameters and we fine tune the models parameters. it was given criterion = poisson, 'max depth = None, min samples leaf = 1, min samples split = 5. this is the best parameters that is found for SVM and we have used GridSearchCV method to implement hyper parameter tuning to find best fit for this model.

Lastly we built Decision Tree using default parameters that has been trained on our data to see how it performed. For this model we have tuned parameters like C = 1000.0, kernel = rbf. this is best parameters in this setting that will produce better results for the model than the model with default parameters.

5.2 Time-series based Cross Validation

In this stage, we have done cross validation using Time Series method. the time series based cross validation chosen over k-fold cross validation because the data is Time Series in nature. Hence the order of our data is much more important to consider and we can see how our model performance over each splits and this will help us understand how the model performance evolve over time. we have used TimeSeriesSplit method using scikit-learn package having n splits parameter as 5. we have done this for all the three models and calculated RMSE for each splits for different model. finally we take average RMSE value for each model and do comparative analysis.

5.3 Ensemble Model - Voting Average

In the voting average method, we have RF models parameters to be n estimators = 100, max-depth = None, min-samples-split = 2. for SVM we have considered C = 10, kernel = 'rbf'. finally for the decision tree model we have chosen max depth = None and min-samples-split = 2. VotingRegressor method from scikit-learn library is used to combine all the three models with the above chosen parameter. At last we calculate square root of mean squared error to get the RMSE.

5.4 Ensemble Model - Stacked Generalization

The Stacked Generalization method uses meta regressor model which combined with our Ensemble model to produce the desired outcome. here we have used Linear Regression model as the meta regressor model. its good for our choice of problem because of time

series nature of data. the scikit-learn library provides StackedRegressor which takes in parameters like estimators which is the array of tuples with the individual models, final estimator as LinearRegressor meta model. Finally we train model with training data and then predict it with test data to get the desired results.

6 Evaluation

The main goal of this research is to forecast the wheat yield in India using an ensemble machine learning model and compare it with the performance of the standalone machine learning models like Random Forest(RF),Support Vector Machine(SVM), Decision Tree(DT). The data spanned from 1997 to 2020 and was categorized by state, district, and season. The models were trained using data up to 2015, and their performance was tested on data from 2016 to 2020.

The results indicated that the ensemble approach outperformed individual models, providing more stability and accuracy over iterations. When comparing the two ensemble methods, the Stacking Ensemble approach, after hyperparameter tuning, achieved the lowest Root Mean Square Error (RMSE) . This result was superior to the Voting Average method, which had an RMSE higher than that. the different results comparison of the models are given in a table below.

Model Measures	**Random Forest**	**SVM**	**Decision Tree**	**Voting Average**	**Stacked Generalization**
RMSE (No Tuning)	0.577	0.515	0.324	–	–
RMSE (With Tuning)	0.205	0.443	0.915	0.285	0.204
RMSE (Cross Validation)	0.487	0.441	0.282	–	–
R2 Score (No Tuning)	0.730	0.785	0.915	–	–
R2 Score (With Tuning)	0.966	0.841	0.325	0.934	0.966
MAE (No Tuning)	0.419	0.309	0.172	–	–
MAE (With Tuning)	0.102	0.232	0.669	0.150	0.101

6.1 Evaluation of Random Forest

From the results table we can understand that Random Forest model, with the default parameters the RMSE was 0.577 and MAE of 0.419 which is quite higher and hence we reject the model as the value appears far from 0. with the hyper parameter tuning setup we were able to achieve RMSE of 0.205 and MAE of 0.102. As far as the R2 score is

concerned we get 0.730 and 0.966 after tuning. there has been slight improvement with different parameters but it doesn't guarantee any performance for any longer run.

6.2 Evaluation of Support Vector Machine

The results of SVM showed that Initially without any hyper parameters the RMSE was 0.515 which is more or less same as RF and MAE was 0.309 which is also higher which shows the models error increased with default parameters. with parameter tuning we were able to achieve RMSE of 0.443 and MAE of 0.232 and r2 score of 0.966 and after tuning 0.841 which is even compared to previous model performance.

6.3 Evaluation of Decision Tree

The Decision Tree has a RMSE before Tuning was 0.324 which better than other two models but 0.915 after tuning which is higher. MAE has been the lowest for this model which was 0.172 and 0.669 which is also higher compared to RF and SVM. It has one of the highest r2 score of 0.915 and 0.325 with tuning which signifies these models are not performing good with the data and not greatly produce results.

6.4 Evaluation of Time Series Cross Validation

Time series cross-validation was used to assess the predictive accuracy of the three models (Random Forest, SVM, and Decision Tree) across different periods. It is essential for models dealing with temporal data, as it provides a robust measure of how well the model can forecast future data points. The cross-validation was performed over five splits, each representing a different time period for training and testing the model. It's important to note that for each split, the training set contains all data up to a certain point in time, and the test set contains data from after that point. the results are illustrated in the Figure 7 .

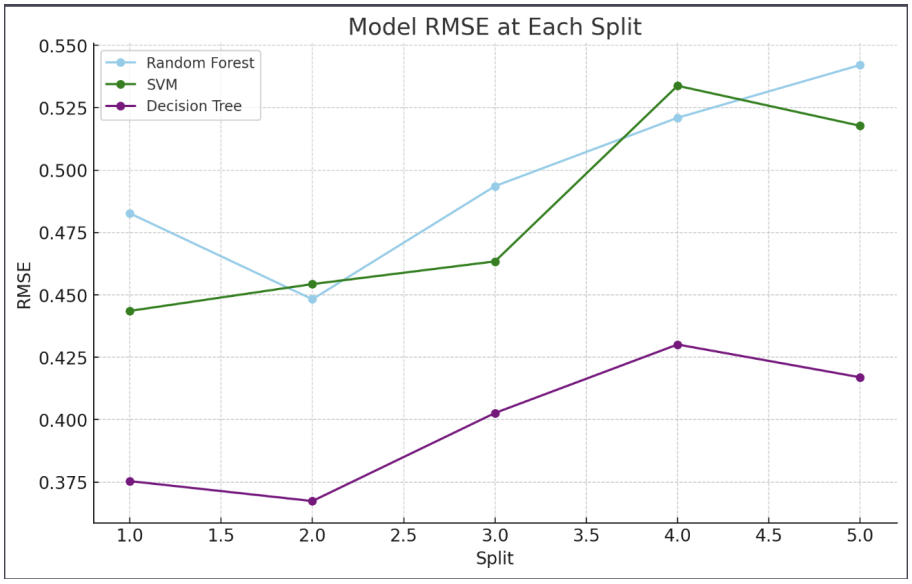

Figure 7: Results of Time Series Cross Validation

The RMSE of the Random Forest model ranged from 0.416 to 0.579 across the five splits. The highest error was found in the most recent split (train: 1997-2016, test: 2016-2020), potentially indicating a decrease in the model's performance over time. However, the average RMSE across all splits was 0.487, a relatively moderate value.

The SVM model's RMSE varied from 0.392 to 0.518 across the splits, with the highest error again observed in the final split (train: 1997-2016, test: 2016-2020). The average RMSE of the SVM model was 0.441, which is slightly lower than that of the Random Forest model, suggesting slightly better performance on average.

The Decision Tree model had the lowest RMSE values among the three models, ranging from 0.202 to 0.358. Interestingly, its highest error was not in the final split but in the first (train: 1997-2001, test: 2001-2005). The average RMSE of the Decision Tree model was 0.282, the lowest among the three models, indicating its superior performance in this validation process.

Overall, all three models showed reasonable performance with RMSE values less than 0.6. However, the Decision Tree model outperformed the Random Forest and SVM models in this time series cross-validation, demonstrating the lowest average RMSE. This is consistent with the results obtained in the earlier model evaluations.

6.5 Evaluation of Ensemble Model

The ensemble model that we used in this research combines three different models Random Forest, Support Vector Machine and Decision Tree. the meta model that has been used in this is the Linear Regression which is good fit for our use case. in this we have used two different approach to evaluate the ensemble models, the Voting Average and Stacked Generalization and the results are interpreted as follows.

6.5.1 Voting Average

The Voting Average method, aggregates the predictions from multiple models. It yielded an RMSE of 0.286, an R2 score of 0.934, and an MAE of 0.150. These metrics suggest that the Voting Average ensemble performed admirably in predicting wheat yield over standalone models that is used in this research. The high R2 score indicates that a substantial portion of the variance in the wheat yield data was accounted for by this model. Additionally, the relatively low RMSE and MAE values demonstrate that the model's predictions were typically close to the actual values.

6.5.2 Stacked Generalization

In the Stacked Generalization method, the performance of the model was siginificantly greater than other models, achieving an RMSE of 0.204, an R2 score of 0.966, and an MAE of 0.101. These improved results indicate a higher level of accuracy and performance. The high R2 score suggests that the model accounted for a significant proportion of the variance in the wheat yield, while the lower RMSE and MAE values indicate that the model's predictions were more accurate, on average.

6.6 Discussion

The study have evaluated on many approaches and presented the results in the Section 6. the further discussions about the results are given below.

6.6.1 Summary of Findings

The research aimed to predict wheat yield in India using machine learning models: Random Forest (RF), Support Vector Machine (SVM), and Decision Tree (DT). Each model was trained on data from 1997 to 2015 and tested on data from 2016 to 2020. The individual models, without hyperparameter tuning, delivered varying RMSE results. RF achieved an RMSE of 0.577, SVM achieved an RMSE of 0.515, and DT yielded the best result with an RMSE of 0.324. These results were mirrored in the R2 scores and Mean Absolute Error (MAE) values, with the Decision Tree model showing superior performance. Hyperparameter tuning considerably improved the performance of the Random Forest model, reducing its RMSE to 0.205, while the performance of the SVM and Decision Tree models deteriorated, with RMSEs of 0.443 and 0.915, respectively. The improvement in the Random Forest model's performance was also reflected in the R2 scores and MAE values, where it surpassed the other two models. The time series-based cross-validation results provided further evidence of the Decision Tree model's superiority in the absence of hyperparameter tuning. The Decision Tree model consistently yielded the lowest RMSE across all splits, followed by SVM and RF. The study then investigated ensemble methods, specifically Voting Average and Stacked Generalization. The Stacked Generalization method achieved the best results with an RMSE of 0.204, an R2 score of 0.966, and an MAE of 0.101.

6.6.2 Comparison with Previous Studies

By comparing with previous study like Cao et al. (2022), this study utilized machine learning models, which are capable of capturing non-linear relationships and interactions among variables that conventional models may miss. Compared to previous studies that employed individual models like Löw et al. (2012), this research incorporated an ensemble approach, combining different models' strengths to improve prediction accuracy and stability. This is a significant step forward, as it leverages the collective power of multiple models to minimize errors and enhance predictive performance.

Furthermore, this study uniquely employed time-series cross-validation, a robust method for assessing model performance over time, which is critical for yield prediction. This approach provides a more realistic estimation of model performance compared to traditional cross-validation techniques used in previous studies.

6.6.3 Theoretical Implications

The research confirms the theoretical proposition that ensemble methods can outperform individual models in machine learning tasks. Particularly, the study reveals that Stacked Generalization, when optimized through hyperparameter tuning, delivers superior performance in predicting wheat yield. These findings contribute significantly to the existing body of literature and a research in crop yield predictions and food crop sustainability in future.

6.6.4 Practical Implications

The study's practical implications are significant. Accurate wheat yield prediction can profoundly impact farmers' decision-making processes, influencing planting schedules,

crop management practices, and resource allocation. Policymakers can use these predictions to strategize food security measures, manage agricultural subsidies, and plan economic policies.

7 Conclusion and Future Work

In conclusion, this study offers a significant contribution to the field of agriculture in crop yield forecasting by employing ensemble machine learning techniques for crop yield prediction. Despite the limitations, the study's findings provide valuable insights for farmers, policymakers, and researchers, highlighting the potential of machine learning in enhancing agricultural productivity and food security. The study also opens up new avenues for future research to further improve the accuracy and applicability of machine learning models in agricultural. Despite the promising results, the study has limitations. The models were trained on historical data, which may not account for future changes in climate, agricultural practices, or socio-economic factors. Also, the models were developed exclusively for predicting wheat yield in India, and their performance may vary when applied to other crops or regions.

Future research should consider integrating more diverse data, such as climate variables or socio-economic indicators, into the models. Additionally, researchers could investigate these techniques' applicability to other crops and regions. Further exploration of ensemble methods and advanced machine learning models could lead to more robust and accurate crop yield prediction models.

References

Cao, J., Wang, H., Li, J., Tian, Q. and Niyogi, D. (2022). Improving the forecasting of winter wheat yields in northern china with machine learning–dynamical hybrid subseasonal-to-seasonal ensemble prediction, *Remote Sensing* **14**(7): 1707.

Heremans, S., Dong, Q., Zhang, B., Bydekerke, L. and Van Orshoven, J. (2015). Potential of ensemble tree methods for early-season prediction of winter wheat yield from short time series of remotely sensed normalized difference vegetation index and in situ meteorological data, *Journal of Applied Remote Sensing* **9**(1): 097095–097095.

Iizumi, T., Yokozawa, M., Sakurai, G., Travasso, M. I., Romanenkov, V., Oettli, P., Newby, T., Ishigooka, Y. and Furuya, J. (2014). Historical changes in global yields: major cereal and legume crops from 1982 to 2006, *Global ecology and biogeography* **23**(3): 346–357.

Keerthana, M., Meghana, K., Pravallika, S. and Kavitha, M. (2021). An ensemble algorithm for crop yield prediction, *2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)*, IEEE, pp. 963–970.

Lobell, D. B., Schlenker, W. and Costa-Roberts, J. (2011). Climate trends and global crop production since 1980, *Science* **333**(6042): 616–620.

Löw, F., Schorcht, G., Michel, U., Dech, S. and Conrad, C. (2012). Per-field crop classification in irrigated agricultural regions in middle asia using random forest and support

vector machine ensemble, *Earth Resources and Environmental Remote Sensing/GIS Applications III*, Vol. 8538, SPIE, pp. 187–197.

Maimaitijiang, M., Sagan, V., Sidike, P., Hartling, S., Esposito, F. and Fritschi, F. B. (2020). Soybean yield prediction from uav using multimodal data fusion and deep learning, *Remote sensing of environment* **237**: 111599.

Pantazi, X. E., Moshou, D., Alexandridis, T., Whetton, R. L. and Mouazen, A. M. (2016). Wheat yield prediction using machine learning and advanced sensing techniques, *Computers and electronics in agriculture* **121**: 57–65.

Prasad, N., Patel, N. and Danodia, A. (2021). Crop yield prediction in cotton for regional level using random forest approach, *spatial information research* **29**: 195–206.

Prodhan, F. A., Zhang, J., Sharma, T. P. P., Nanzad, L., Zhang, D., Seka, A. M., Ahmed, N., Hasan, S. S., Hoque, M. Z. and Mohana, H. P. (2022). Projection of future drought and its impact on simulated crop yield over south asia using ensemble machine learning approach, *Science of The Total Environment* **807**: 151029.

Ray, D. K., Gerber, J. S., MacDonald, G. K. and West, P. C. (2015). Climate variation explains a third of global crop yield variability, *Nature communications* **6**(1): 5989.

Sellam, V. and Poovammal, E. (2016). Prediction of crop yield using regression analysis, *Indian Journal of Science and Technology* **9**(38): 1–5.

Shah, A., Agarwal, R. and Baranidharan, B. (2021). Crop yield prediction using remote sensing and meteorological data, *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*, IEEE, pp. 952–960.

Zhou, W., Liu, Y., Ata-Ul-Karim, S. T., Ge, Q., Li, X. and Xiao, J. (2022). Integrating climate and satellite remote sensing data for predicting county-level wheat yield in china using machine learning methods, *International Journal of Applied Earth Observation and Geoinformation* **111**: 102861.