Impact of Historical Weather Data on Crop Selection and Fertilizer Recommendation using Machine Learning Stacking Approach: Indian Cities

MSc Research Project
Data Analytics

Vikraman Kannappan Sankar
Student ID: x23104881

School of Computing
National College of Ireland

Supervisor: Prof. Jorge Basilio

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Vikraman Kannappan Sankar		
Student ID:	x23104881		
Programme:	MSc Data Analytics	**Year:**	2023 - 24
Module:	MSc Research Project		
Supervisor:	Prof. Jorge Basilio		
Submission Due Date:	12/08/2024		
Project Title:	Impact of Historical Weather Data on Crop Selection and Fertilizer Recommendation using Machine Learning Stacking Approach: Indian Cities		
Word Count:	7382	**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: Vikraman Kannappan Sankar

Date: 12th August 2024

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):	

Impact of Historical Weather Data on Crop Selection and Fertilizer Recommendation using Machine Learning Stacking Approach: Indian Cities

Vikraman Kannappan Sankar
x23104881

Abstract

The rise in the global population has increased the demand for the food production, this possesses a big challenge to the agriculture sector. Traditional farming methods are not being sufficient due to the unpredictable weather patterns and the environmental changes. Farmers due to the drastic environmental condition and weather changes, they choose wrong crop and fertilizer for cultivation which will significantly have negative impact on the crop productivity. This obstacle can be overcome with a data-driven approach, in this study we will be leveraging the historical weather data to create a recommendation system using machine learning stacking technique. Using stacking technique will enhance the performance of the model by integration of multiple model such as random forest classifier, decision tree classifier, K-nearest neighbors and MLP classifier. This study also compares the performance of the standalone and the stacking model with the evaluation metrics of Accuracy, Precision, Recall and F1 Score.

Keywords: Stacking Technique**,** Weather Data**,** Crop Recommendation**,** Fertilizer Recommendation, Machine Learning

1 Introduction

1.1 Background and Motivation

The rapid increases in the world population, the United Nation projects the population will reach up to 9.8 billion by the year 2050 and it will increase further to 11.2 billion in 2100. This possesses a big threat to the agriculture industry in terms of the rapid increase in the need of food production [1]. There are several problems occurring in the agriculture field, one of the commonly occurring problem is that the conversion of the farmable land to housing development area this was observed by (Saputra et al., 2022) research done on the Indonesia's population. This possesses the urgent of the need of finding new ideas to improve the agricultural productivity to ensure the food security. The traditional farming methods are failing due to changes in the weather patterns and environmental changes.

The Indian farmers rely on the monsoon season for the cultivating of the rice and the wheat now the faces the inconsistent in the yield due to the environmental shifts, similar to that in

[1] https://www.un.org/en/desa/world-population-projected-reach-98-billion-2050-and-112-billion-2100

United States article published by the Washington post 2024 reported that the Midwest region experienced intense storms and floods [2]. Countries started to take some measure to resolve the population problem, China introduced a policy known as the one child policy to control the population in the country, it was introduced in the year 1979 and was the great success to reduce the population [3]. Governments all around the world recognize this issue and are taking measures to support the farmers, in India government offering welfare schemes with the low interest bank loans [4]. This impact in the agriculture is mainly due to the global warming, in India the climate model projects the significant increase in the global surface air temperature by 4.0-5.8 °C in the coming decades, this is mainly a big threat to the agricultural sector which will face the more challenges. With the India's population is expected to hit 1.4 billion by 2025, it is necessary to prepare for the impacts (Chauhan et al., 2014). This challenge possesses the needs to develop methods and techniques that can increase the agricultural productivity in the face of the climate change.

This proposed study aims to solve the gap between traditional farming methods and the modern technology by developing a crop and fertilizer recommendation system by leveraging the historical weather data, fertilizer usage and the crop production data. Using the machine learning stacking of random forest classifier, decision tree classifier, K-nearest neighbors, MLP classifier models to optimize the crop and the fertilizer selection. Dataset for this study is sourced from the open platforms Kaggle, which includes the crop production statistics – India data [5], city-specific historical weather data [6], and records of fertilizers data [7]. In this study we will merge all the dataset to single dataset, in order for creating the recommendation system which will focus on the Indian Cities Crop and Fertilizer recommendation. (Elbasi et al., 2023) demonstrated that the use of the machine learning algorithms is optimizing the crop production and reducing the waste with the high accuracy, these approaches show the potential of the machine learning to enhance the agricultural productivity through the data driven decision making approach, in this proposed study we will implement a machine learning based recommendation system using the stacking technique to improve the crop and the fertilizer recommendations.

1.2 Research Question and Objectives

The proposed research question for this study:

"How can stacking of machine learning models, using the historical weather data to enhance performance of both crop selection and fertilizer recommendations of India Cities?"

[2] https://www.washingtonpost.com/weather/2024/02/26/storm-front-snow-fires-midwest-plains-rockies-snow/

[3] https://www.britannica.com/topic/one-child-policy/Consequences-of-Chinas-one-child-policy

[4] https://timesofindia.indiatimes.com/india/empowering-indias-farmers-list-of-schemes-for-welfare-of-farmers-in-india/articleshow/107854121.cms

[5] https://www.kaggle.com/datasets/nikhilmahajan29/crop-production-statistics-india

[6] https://www.kaggle.com/datasets/hiteshsoneji/historical-weather-data-for-indian-cities

[7] https://www.kaggle.com/datasets/gdabhishek/fertilizer-prediction- Kaggle

The objective of the study is to develop a recommendation system using the historical weather data of the Indian Cities, incorporating machine learning stacking of random forest classifier, decision tree classifier, K-nearest neighbors and MLP classifier models to optimize the crop and fertilizer selection. By integrating the multiple datasets including the crop production statistics, historical weather pattern data and fertilizer records, the study aims to enhance the performance of the model. Comparing the performance of the single standalone and stacking model by evaluating accuracy, F1, recall and precision metrics. Data-driven approach provide the farmers with the valuable insights that will enabling them to make the informed decision that will improve the agriculture efficiency and the resilience to environmental changes. Ultimate goal is to bridge the gap between the traditional farming practices and the modern technology advancement to ensure the sustainable food production in the face the growing global population.

1.3 Document Structure

The structure of the paper is as follows, introduction part will be covered in Section 1, Section 2 will contain relevant research of integration of the machine learning algorithm and breakdown section will contain the methods used in crop recommendation system. Section 3 will cover about the methodology which will contain the data preprocessing and feature engineering. The design specification of the model will be covered in the Section 4. Implementation of the work will be covered in Section 5. Evaluation of the implemented model under different case studies will be covered in Section 6. Finally, Section 7 will contain the conclusion and the future works.

2 Related Work

The development of crop selection and fertilizer recommendation model is a difficult aspect of the agricultural management as it depends on the various factors like the weather pattern, crop types, soil characteristics and environmental conditions. Building a recommendation system can help the farmers to make the informed decision of crop and fertilizer usage which will help increases the crop productivity. In this section reviews about the previous literature and highlighting key contribution in the field of the agricultural data mining and machine learning.

2.1 Integration of Machine Learning in Agricultural

Machine learning model have become the game changer in the many sectors, in agriculture it has influenced by providing the opportunity of more data analysis and the predictive modelling which enhance decision making process in the farming practices. ML algorithms are had been applied to various agricultural task that include crop yield prediction and assessment on soil quality (Elbasi et al., 2023). IoT sensors was used to gather the weather data and it was used to understand about the weather patterns, similar way IoT was used to gather the information about the soil properties and the pest outbreaks for the prediction of the crop growth this precision farming technique was used by (Cravero et al., 2022) to optimize irrigation and the pest control in order of improving the crop yield and the quality. Implementation of this type of technology have several challenges that will require large datasets in order of training the ML models which is bit difficult to obtain in some agricultural settings. (Maharana et al., 2022)

Despite of all the drawbacks the integration of the ML techniques in the agriculture that provide the great promise in increasing the efficiency and the productivity. The integration of (ML) into agriculture that has significantly transformed the farming practice, the research by (Banerjee & Chand Mondal, 2023) creation of an effective crop prediction system using different machine learning algorithms, the parameters used for the analysis temperature, rainfall and the soil type. Study compared different deep learning algorithms such as the Convolutional Neural Networks (CNNs), Long-Short Term Memory (LSTM) this algorithm used Lasso, Kernel Ridge and Enet to improve the prediction accuracy. Disadvantage in this study, since all the deep learning model required a large amount of data to process, more complexity in training this model which will require high computing power and operational cost.

2.2 Historical Weather Data and Its Impact on Agriculture

The historical weather data significantly impacts the agriculture industry as per the study done by (Gupta et al., 2021), highlighted the importance collection and the analysis of extensive weather data which include the temperature, rainfall and the humidity and the wind speed. This study used the Big Data Analytics used to identify the patterns and predict the short-term and the long-term climatic changes. (Schwalbert et al., 2020) study integrate the weather data such as the temperature and the perception in the ML models for the improved soybean yield prediction in southern Brazil. Combining the satellite images with the weather data enhances the model accuracy by reduction in the mean absolute error (MAE) and the root mean square error (RMSE), this analysis highlighted the importance of the using the weather data in the agriculture field.

(Shook et al., 2021) highlighted the importance of the historical weather data in the prediction of the crop yield by the incorporating the genotype information using the deep learning models. The challenges of the accounting the climate variable caused the need to incorporate stacking number of models to provide the insights during the growing season. (Medar et al., 2019) for the forecasting the sugarcane yield in the Karnataka, India used the supervised ML techniques this model incorporates the same Long Term Time Series (LTTS) the forecasting weather and the soil attributes combination of the normalized the difference vegetation index. The most importance factors used for crop prediction are soil, moisture, windspeed, temperature and the other environmental factors. (Laxmi & Kumar, 2011) study showed the impact of the weather data used Multilayer Perceptron (MLP) model that uses the non-linear elements to arrange the successive layers, the information flow in uni-direction from the input to output through the hidden layers, this model achieved the forecast that closely matched the crop yield. This demonstrated the MLP achieved accurate prediction using the weather data.

2.3 Machine Learning-Based Crop Recommendation Systems

(Gosai et al., 2021) created a crop-based recommendation system using the machine learning, to recommend the suitable crops based on the various environmental and the soil factors. This study incorporated multiple algorithms such as the Decision Tree, K-Nearest Neighbour (K-NN), Random Forest and Neural Networks to improve the accuracy. Factors such the soil type, pH, temperature are used to create recommendation crops. New approach of creation of the dual sub-system of the crop suitability and the rainfall prediction to achieve high accuracy in crop recommendation to optimizing the crop selection and improve yield and resource efficiency. (Kumar et al., 2023) created a recommendation using the ensemble stacking technique, the idea of the approach was to combine the multiple classifiers through the stacking ensemble method. Random Forest, Decision Tree, KNN and Naive Bayes was used for the prediction and the creation of the new features to train the meta learner which is the Random Forest Classifier creating the initial stacking layer. Final model is the combination of the prediction from the ensemble and input features. This method shows the potential strength of combining different models to increase the better predictivity of the model, evaluation metrics like recall and accuracy was used.

Research done by (Priyadharshini et al., 2021) of creation of the Intelligent Crop Recommendation system similar to the previous studies most important features like the environmental parameters, soil type, climate and the seasonal patterns for making the accurate prediction. This all study shows that above mentioned Features are very important to create the accurate recommendation system. (Oswal et al., 2021) study used fuzzy logic to address the challenge of the feature selection in the imbalanced datasets. (Suruliandi et al., 2021) used the Recursive Feature Elimination (RFE) technique was used to improve performance of the crop prediction model. This study suggested that the further research have to incorporate the weather data to improve the accuracy.

2.4 Limitations and Takeaway from the Literature

The literature on the machine learning crop recommendation systems reveals the limitation and the key takeaways. The primary limitation seen in the literature is that handling the non-linear relationship between the different agricultural variables includes the soil, moisture and the temperature which are more important for the accurate crop prediction in the agriculture field. In addition to it, often the studies required the large datasets that may contain the inaccuracies, studies showed that fuzzy logic and Recursive Feature Elimination methods can be used for the feature selection. Most of the studies did not incorporate the dynamic weather data for the creation of the model, which is the most essential features in improving the prediction accuracy. Positive side is that the various studies used the potential of integration multiple machine learning models such as the Random Forest, Decision Tree, KNN, Naive Bayes and MLP classifier and the stacking methods to improve the prediction.

3 Methodology

The proposed solution flowchart shows in Figure 1, first initial step of the design which involves the all the Dataset are pushed to the Jupiter Notebook for Data Pre-Processing which involves Data Cleaning and other preprocessing step including merging of the dataset, after merging of the dataset the final dataset is achieved. The next Step is splitting the data to test and train data. Train data is used creation of the machine learning model once the first step base model is created and then next step the meta model created to recommend the crop and the fertilizer and the test data is used to evaluate the performance.

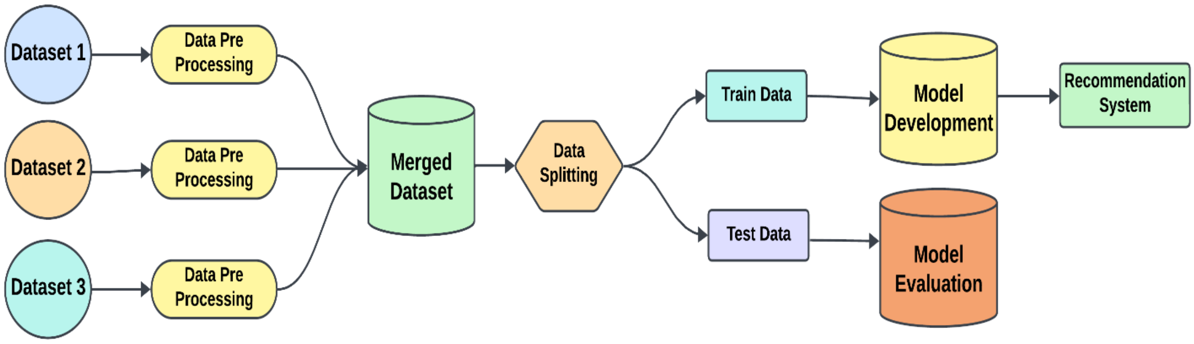

Figure 1: Proposed Solution Flowchart

3.1 Data Sourcing

3.1.1 Crop Production Dataset

Data set for this study is sourced from the open-source platform Kaggle, total three types dataset which include Crop Production Statistics – India, Historical Weather Data for Indian Cities and Fertilizer information. The Crop Production Statistics – India which contains the crop production of India which are categorized by the State and the district and the cover the four seasons major crop seasons, namely kharif, rabbi, summer, and autumn from the year 1997 to 2020, there are total 345337 records and 8 columns. The Dataset is shown in Figure 2, the data includes such as State, District, Crop, Crop_Year (crop harvested year), Season (growing season of the crop), Area (area under the cultivation of the crop), Production (production output of the crop) and Yield (yield of the crop cultivated per unit area).

	State	District	Crop	Crop_Year	Season	Area	Production	Yield
0	Andaman and Nicobar Island	NICOBARS	Arecanut	2007	Rabi	1626.4	2277.0	1.40
1	Andaman and Nicobar Island	NICOBARS	Arecanut	2008	Autumn	4147.0	3060.0	0.74
2	Andaman and Nicobar Island	NICOBARS	Arecanut	2009	Autumn	4153.0	3120.0	0.75
3	Andaman and Nicobar Island	NICOBARS	Arecanut	2009	Summer	4153.0	2080.0	0.50
4	Andaman and Nicobar Island	NICOBARS	Arecanut	2006	Whole Year	896.0	478.0	0.53

Figure 2: Crop Production Statistics Dataset

3.1.2 Historical Weather Dataset

Second data set is about the Historical Weather Data of the Indian Cities Bengaluru, Bombay, Delhi, Hyderabad, Jaipur, Kanpur, Nagpur and Pune all the data includes 25 columns. Most important attributes such the date_time (timestamp of the weather data which include date and time), maxtempC and minctempC (Maximum and Minimum temperature), sunHour (number of hours the sunshine recorded for the specific time duration), uvIndex (uv Index indicating the level of ultraviolet radiation) , humidity (humidity observed during the time period), tempC (temperature during the time period), windirDegree (wind direction), windspeedKmph (windspeedKmph), moonset, moonrise along with the City names is show in the Figure 3.

	date_time	maxtempC	mintempC	sunHour	pressure	uvIndex	sunHour	moonset	moonrise	humidity	tempC	winddirDegree	windspeedKmph	City
0	2009-01-01 00:00:00	27	12	11.6	1014	5	11.6	10:03 PM	09:58 AM	91	14	109	8	Bengaluru
1	2009-01-01 01:00:00	27	12	11.6	1014	5	11.6	10:03 PM	09:58 AM	93	14	85	6	Bengaluru
2	2009-01-01 02:00:00	27	12	11.6	1014	5	11.6	10:03 PM	09:58 AM	94	13	61	4	Bengaluru
3	2009-01-01 03:00:00	27	12	11.6	1014	5	11.6	10:03 PM	09:58 AM	96	12	37	3	Bengaluru
4	2009-01-01 04:00:00	27	12	11.6	1015	5	11.6	10:03 PM	09:58 AM	88	14	45	3	Bengaluru

Figure 3: Historical Weather Data of the Indian Cities

3.1.3 Fertilizer Dataset

Third data set is about the fertilizer information and contains the 9 columns it includes the temperature, Humidity, Moisture, Soil Type, Crop Type, Nitrogen, Potassium, Phosphorous and the Fertilizer Name show in figure 4.

	Temparature	Humidity	Moisture	Soil Type	Crop Type	Nitrogen	Potassium	Phosphorous	Fertilizer Name
0	26	52	38	Sandy	Maize	37	0	0	Urea
1	29	52	45	Loamy	Sugarcane	12	0	36	DAP
2	34	65	62	Black	Cotton	7	9	30	14-35-14
3	32	62	34	Red	Tobacco	22	0	20	28-28
4	28	54	46	Clayey	Paddy	35	0	0	Urea

Figure 4: Fertilizer Dataset

3.2 Data Pre-Processing

Data Pre-Processing is the important step in the creation of the Machine Learning Model, this involves few steps to process the data (Figure 5). The First step of the process is the preparation of the Data, the next step is the Data Cleaning which will basically include the handling of the missing values, correcting the inconsistencies in the data and removing of the duplicates. Next is the data transformation is done converting data types and the normalizing the numerical values and encoding of the categorical variables. Data reduction is done by selecting the required features by focusing on the essential features. Last step is the Feature Engineering is applied to create new insightful feature from the exiting data.

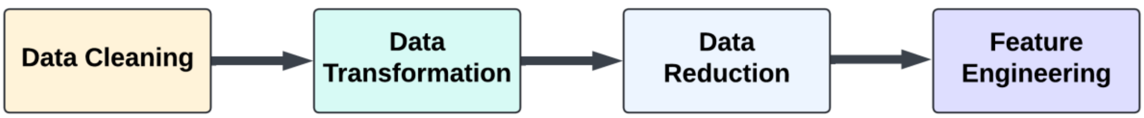

Figure 5: Data Pre-Processing Flowchart

3.3 Data Cleaning

3.3.1 Handling the Missing Values Crop Production Data

In the data cleaning process, the Crop Production Statistics dataset is read form the CSV file using the pandas libraries. The dataset is converted pandas Data Frame, when analysing the data set it was found that there are few missing values were identified in the Crop and the Production columns. In Crop column 9 rows and in Production column 4948 rows were missing.

Cluster-Based Interpolation: As per the (Kaiser, 2014) study the missing values can be dropped if the volume of the missing data is less, in our case the crop column has 9 rows of the missing data which is less when compared to main data set, so missing crop values were dropped. For the production column there is a greater number of missing values, in this work we adopted a method similar to the one proposed by (Shi et al., 2020), which utilized the clustering algorithm to handle the missing values and to provide the improved mean. The first step of the process of cluster interpolation (Figure 6), is that selection of features from the cleaned dataset. The selected features Crop_Year, Area, Yield are scaled and K-Means cluster is applied, elbow method was used to find the optimal number of clusters (k=3). Next step is that performing interpolation with each cluster to fill the missing values. Last Step is dropping the cluster column after verification of the missing values.

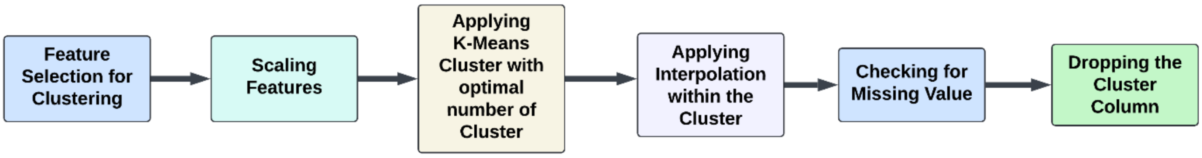

Figure 6: Cluster Interpolation

Data Standardization: Column names were stripped for removing the extra spaces using the pandas, in order to avoid the problems in the data manipulation. Next the data types are converted, the categorical columns converted to category type and the numerical columns to numeric types for ensuring the data is standardized through the process.

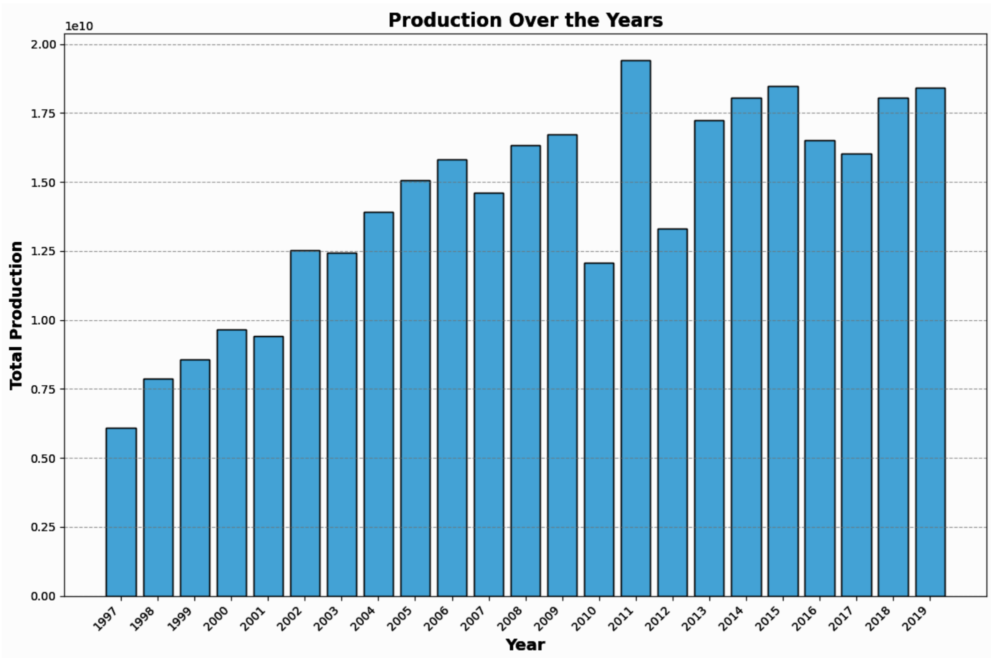

Figure 7: Crop Production Over the Years

Figure 7 show the Crop production over the years from 1997 to 2019, with the x-axis show the year of the crop production and y-axis show the crop production in billion. Graph showing the upward trend in the production, this show that the demand of the food production gradually increased in years. In the year 2007 there faced a dip in the crop production this could have been varying due to the climatic conditions or economic factors affecting agricultural production output. After 2013, production is stable which having slight minor variation over the years. This stability shows that improvements done in the agriculture practice or the favourable climate condition favoured in sustaining the high productivity.

3.3.2 Combining and Cleaning Weather Data

The historical weather data set pre-processing, the weather data is collected from multiple cities, and the dataset is stored as csv. Using the pandas library the data is imported, first step of the process all the files are joined as a single pandas data frame.

Combining Data Frames: A tuple with all the file path of weather of 7 major Indian city Bengaluru, Bombay, Delhi, Hyderabad, Jaipur, Kanpur, Nagpur by iterating over the list each CSV file into pandas Data Frame. Next is the appending the file using City column to indicate the data's origin and collect this Data Frame into a list. Then the individual data frames are concatenate into a single combined Data Frame.

Data Cleaning: Data cleaning process is processed for the combined data frame, first step of the process checking if there are any missing values in the dataset. No missing values found in the combined data set. Similarly, no duplicate entries of the data set were seen, column are converted to appropriate data type, checking the unique value of the city column to ensure that the all the city is having consistent city names.

Temperature Analysis: The box plot Figure 8 represents the maximum temperature across the Indian cities which include Bengaluru, Bombay, Delhi, Hyderabad, Jaipur, Kanpur, and Nagpur.

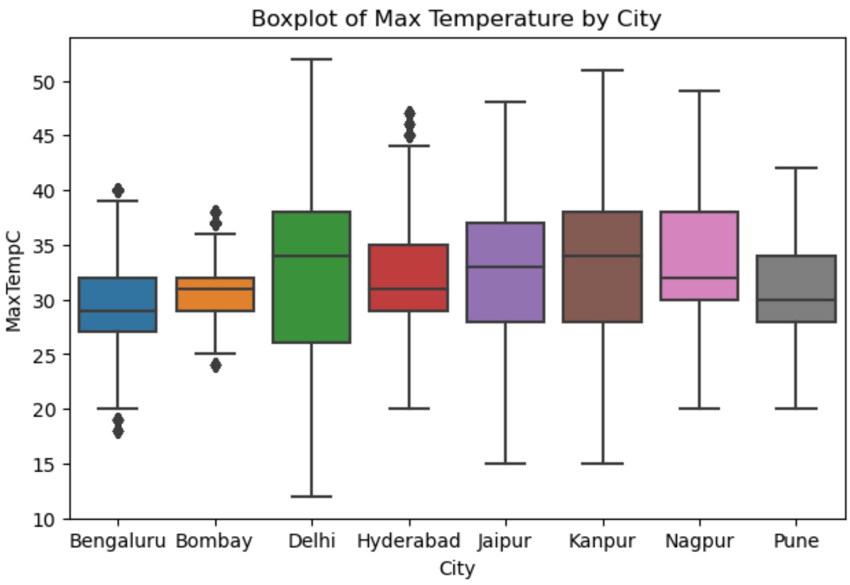

Figure 8: Maximum Temperature of the City

Analysis the outcome of the chart Bengaluru showing a moderate temperature range with the median having around 29°C which is constantly maintaining a moderate temperature is crucial for the good crop yield. The increase in the maximum temperature will affect the crops whereas the rise is in minimum temperature is beneficial but not much effective for good crop yield (Birthal et al., 2014). Maximum Temperature and the Minimum Temperature becomes the crucial factors for the good crop yield, Bombay having median around 31°C and indicating an occasional deviation for low temperature, this show Bombay having stable temperature pattern with the mild variations. Delhi being the widest box with the reflecting high temperature close to 40°C which extreme high temperature which will be affecting the crop production in large number, this large number in the temperature show that there is frequent deviation in the temperature from extreme hot to cool. Comparing to Delhi the Jaipur, Kanpur, and Nagpur other cities experiencing the lower similar temperature that ranges from 35-40°C this shows that it is more consistent temperature across the cities. This Chart shows that there are few outliers present in the data, this outlier well be solved in the next steps of the methodology.

3.3.3 Formatting Fertilizer Data

Fertilizer data set, using the pandas data frame the csv file is imported and converted to pandas data frame, in the initial analysis it was found that there is no missing values and duplicate values found in the data.

Cleaning and Formatting: To ensure the consistency the column name consisted of white space was stripped down and the numeric columns was converted to appropriate data types. The 'Soil Type' and 'Crop Type' columns were capitalized to ensure the data is in the correct format.

Nutrient Content Analysis: The Figure 9 show the nutrient content of the different fertilizers which special contain nitrogen, potassium and phosphorus. Urea which contains the high amount of the nitrogen and which lacks in the potassium and phosphorus, DAP (Diammonium Phosphate) is rich in the phosphorus with the moderate amount of nitrogen and no potassium.

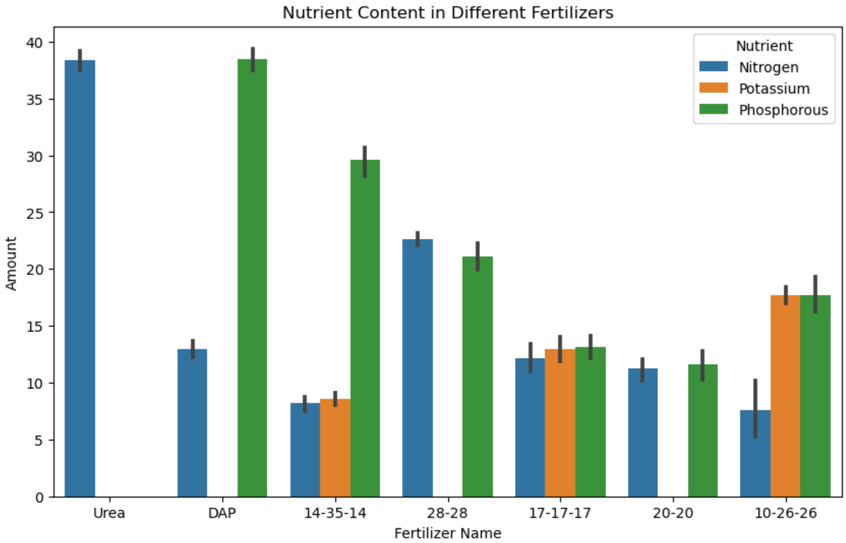

Figure 9: Nutrient Content of Different Fertilizers

All other fertilizer '14-35-14' contained equal range of Nitrogen and little higher range of Potassium. When analyzing all the fertilizer concentration of the nutrient content in all the fertilizer the nitrogen is less than the potassium and phosphorous this up trend is seen showing the important of phosphorous and potassium.

3.4 Data Transformation

Data transformation is the important part of the study as it involves the conversion of the raw data into a format suitable for performing the analysis. This process follows the several steps that aim to enhance the quality of the data making sure the data is consistent.

Label Correction: The primary step in the data transformation is the correction of the data labels to ensure the structure is uniform. In the crop production dataset, crop name consists of inconsistencies as the varying names of the same crop example 'Cotton(lint)' instead of normal 'Cotton' this can lead to confusion during the merging of the dataset. For performing this task Pandas library used, allowing to replacement the values within the Data Frame, crop Name Values ('Cotton(lint) replaced with Cottan', 'Small millets replaced to Millets', 'Oilseeds total and other oilseeds replaced to Oil seeds', 'Other Kharif pulses, Other Rabi pulses, and Peas & beans (Pulses) were standardized under Pulses' similar way the district name is changed to 'Bengaluru urban to Bengaluru', 'Mumbai to Bombay', 'Delhi_Total to Delhi' and 'Kanpur Dehat updated to Kanpur'.

Extracting and Grouping Date Column: Historical weather data set the 'date_time' column used to extract the year, month features and then the data is grouped for creating the seasonal categories. Initial stage of the process the two new columns was created as year and month, Next essential step of the data transformation is the rounding the numerical value to standardize the data, for example in the weather data rounding the columns 'MaxTempC', 'MinTempC' and 'Humidity' this helps in keeping the data standardized.

Converting Data Types: Converting the data type is also an essential part in the data transformation, ensuring all the column in the dataset has appropriate data type. Example the 'Year' column is converted to integers data type and 'Temperature' values to float for accurate analysis, the Pandas and NumPy libraries was used for data type conversion.

3.5 Data Reduction

Data reduction is important step in order to prepare our data for the analysis. The process begins with the handling of the missing values we used the interpolation technique to fill up the missing values, and reducing not important factors in order to improve the quality of the data. Once the data is merged the irrelevant columns such as 'Humidity_x', 'Temperature', and the 'Crop Type' are dropped as this will create the data redundance.

3.6 Feature Engineering

Feature engineering involves the transformation of the data into meaningful feature, in the weather data the datetime column using the pandas data is changed to time function format, from it year and month is extract as new columns. Next step a function was created in order to label the season based on the months. Label season function was used to categorize each row to specific season based on the created month column. The Categorized season included winter, kharif, rabi, autumn, whole year and summer, Winter season during from December to February, Summer from March to June, Autum from October to November, Kharif from June to October and Rabi from November to March. Once the labelling of the season, the new feature season is created and it then grouped by the year, city and season, this grouping of the data is the crucial step it is important to identifying the season trends. Finally, the grouped seasonal data is joined into a single data frame.

3.7 Merging Dataset

Once all the three data set is pre-processed next step is the merging of the dataset, which involves identifying the common columns in the data set. The common columns in the crop production and the weather data are the 'Year', 'City', and the 'Season' this are the key columns that will used to merge the data. Pandas library merge function was used to merge the crop production and the weather data after merging the combined data contains the insights that will link the crop yield and the weather patterns. Next Step of the process is merging the fertilizer data with the already combined data, the fertilizer data is integrated with the combined dataset using the 'Crop Type' as the main joining key column for this also the merge function from pandas is used to create the final combined dataset. Joining the fertilizer data will allow analysis get more information about the fertilizer impact on the crop production under the different climatic conditions.

Outlier Analysis: Outlier was identified when analysing the merged dataset, statistical methods Interquartile Range (IQR) was used to solve the outlier. This IQR method was used by (Schwertman et al., 2004) this study showed that properly treating of the outliers is important for improving the model accuracy. In this study the outliers are defined the value beyond the 1.5 times of the IQR from the first and the third quartiles.

3.8 Exploratory Data Analysis

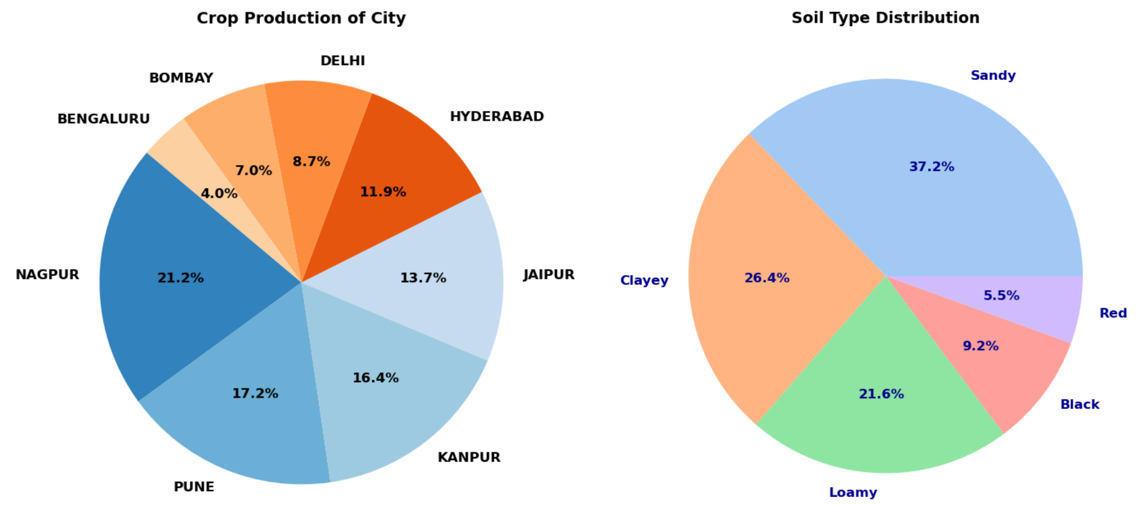

Figure 10: Crop Production of City and Soil Type Distribution

In this section we will discuss about the exploratory data analysis of the merged dataset, Figure 10 shows the distribution of the agricultural crop production of the 8 cities, Nagpur standing out from the list with high percentage this show that high level of agricultural activity is performed in this city when compared it to another city. City like Bengaluru and Bombay show less percentage shows that this city have the less emphasis on the agriculture. Kanpur and the Hyderabad also having the significant count which is less than the Nagpur which show that area of improvement needed in this city. This analysis reveals that more metro city like Bengaluru and Bombay must increases the agriculture productivity. Soil Type Distribution from the Figure 10 shows that Clayey, Sandy and Loamy are more preferred type of soil for the cultivation of the crop in the India Cities. EDA is helpful in identifying the importance of the features used for building up of the model.

4 Design Specification

The design specification involves of sourcing the data from Kaggle, the data include the crop production statistics, historical weather data of the major Indian city and the fertilizer information. Data cleaning process involved in the correction of the missing values using the cluster interpolation method and dropping the not required columns for analysis, converting the numerical values to appropriate data type and standardizing and normalizing the features making sure the consistency of the data. Label encoding was used to converting the categorical variable to numerical. Merging of the dataset crop production data with weather data based on common columns and another merger with fertilizer data. Visualization tools like Matplotlib and Seaborn was used to understand the data links and trends, library such as Pandas, NumPy and scikit-learn was used for performing the data manipulation techniques. This section includes the details of the methods and the algorithms to be applied and the performance metrics, the execution phase cover about the selection of the best model and applying the training data to train the model and evaluate the matrices of the model.

4.1 Modelling Technique

- Random Forest Classifier (RF): Random Forest classifier is the robust learning technique that use multiple decision trees to enhance the accuracy of the model. (Geetha et al., 2020) used the Random Forest for the efficient crop prediction due to their ability of handling large dataset and variable. It effectively reduces the risk of overfitting and improve the model generalization and making this an ideal choice for the agricultural applications.

- Decision Tree Classifier (DT): Decision Tress classifier is a powerful supervised learning algorithm that is used for the classification task. Splitting the dataset into subset based on the significant features forming a tree-like structure with each node representing the decision rule and branch represents the outcomes. (Dhanapal et al., 2021) used this technique to predict the crop price using the historical rainfall and wholesale price index data (WPI).

- K-nearest neighbors (KNN): It is a simple effective machine learning method used for the classification and the regression task. KNN is used where the decision boundary is not well defined and the data is noisy. (Wickramasinghe et al., 2019) study used it for prediction of fertilizer recommendation, due to its ability to handle multiclass classification.

- MLP Classifier (MLP): Multilayer Perceptron (MLP) classifier is a powerful neural network classifier, which consists of multiple layers of neurons including an input layer and one or more hidden layers and an output layer, this will allow the model to learn the complex pattern in the data. (Garg & Alam, 2023) used the MLP classifier for the effective crop prediction and it demonstrated the ability to predict the crop in high accuracy. Additional advantage is that MLP can be tuned through the hyperparameter optimization to improve the model performance.

Stacking technique involves the training of the multiple base model (random forest classifier, decision tree classifier, K-nearest neighbors and MLP classifier) and the combining the predictions using the meta modal. Base models provide the perspectives on the dataset and the output are used as the input for the meta-model. Meta models are the decision tree, random forest, KNN and MLP used to understand the best combination of the base model's prediction to improve overall performance. In our study different meta models are evaluated to determine best combination of the accurate crop and the fertilizer recommendations.

4.2 Evaluation Technique

- Accuracy: Accuracy is the critical metrics that will indicate the percentage of the correct predictions of the model, it is calculated by dividing the number of accurate forecasts by the total number of forecasts and it displayed in percentage.

- Precision and recall: Both the metrics are essential in the classification tasks. Precision is used to measure the true positive of the predictions among all the positive predictions. Recall indicates the percentage of the actual positive samples that correctly identified by the model. This metrics is used for the both binary and multi class classification.

- F1score: F1 Score is the important metric that has the balance precision and recall this both are calculated as harmonic mean of the two. This is useful when the class distribution is uneven.

5 Implementation

Developing a machine learning model will involves several processes and executed procedures. To ensure the model's proper functionality in the real-world application will require a through approach at each stage of the development and the implementation.

5.1 Tools Used

Range of tools and libraries are required to develop and implement the machine learning model. In this study Python programming language was used for coding, data manipulation and preprocessing of the data was handled using the Pandas and NumPy libraries and the Seaborn and Matplotlib used for the data visualization. Model building and evaluation was done using

Scikit-learn specific models was used like train_test_split for the data splitting and various classifiers like RandomForest, DecisionTree, KNeighbors and MLP. Additional to it Pickle was used to for saving and loading the trained model.

5.2 Feature Selection

Feature selection is the important step in the development of the crop and fertilizer recommendation system. It is important to select the correct feature for the model to improve the model performance and the stability. (Suruliandi et al., 2021) used the Recursive Feature Elimination (RFE) technique to improve the performance of the crop prediction model. In this study RFE is used or the feature selection, RFE was employed with the Random Forest Classifier to identify the most important features of the prediction task. Selected features included ‘City’, ‘Season’, ‘MaxTempC’, ‘MinTempC’, ‘Pressure’, ‘TempC’, ‘WindSpeedKmph’, ‘Humidity’, ‘Moisture’ and ‘Soil Type’ this feature was chosen since the contributed more in the prediction in the accuracy of the models.

5.3 Hyper parameter Tuning

The selection of the optimal hyperparameter value is important for a machine learning model. Hyperparameter tuning is critical for optimizing the model performance, (Schratz et al., 2019) study demonstrated the importance of the parameter tuning to maintain the consistency of the classifier performance by reducing biases. In our study the parameters are set before the training process the following hyperparameter are used.

- Random Forest Classifier: Grid Search CV Parameters: *n_estimators, max_depth, min_samples_split, min_samples_leaf, bootstrap.*
- Decision Tree Classifier: Grid Search CV Parameters: *max_depth, min_samples_split.*
- K-Nearest Neighbors (KNN): Grid Search CV Parameters: *n_neighbors, weights, metric.*
- Neural Network (MLP Classifier): Grid Search CV Parameters: *hidden_layer_sizes, max_iter, activation.*

Table 1: Best Parameters for Tuning

Model	Random Forest (RF)	Decision Tree (DT)	K-Nearest Neighbors (KNN)	Multi-Layer Perceptron (MLP)
Best Parameters	*{'max_depth': 20, 'n_estimators': 200}*	*{'max_depth': 20, 'min_samples_split': 10}*	*{'n_neighbors': 3, 'weights': 'distance'}*	*{'activation': 'tanh', 'hidden_layer_sizes': (50, 50), 'max_iter': 1000}*

Table 1 show the values of the best parameters, this was achieved by using the GridSeachCV method, by testing different combination of the definer parameter grid with having 5-fold cross-validation and the accuracy. The hyperparameters values can significantly impact the model effectiveness the tuning is the important in the classification task.

Splitting Data: Last step before modelling of the data is to split the dataset into training and testing sets, the dataset was split it into 70% of the data for training and 30% of the data for testing was split using the Sklearn libraries.

6 Evaluation

In the process of the Machine Learning pipeline, evaluation is the important phases to evaluate the model effectiveness and there by ensuring the it functioning as per the model design. The selection of the appropriate model assessment metrics and proper evaluation of the model's performance is essential.

6.1 Case Study 1: Random Forest Model as the Meta Model

In the first case study the Random Forest Model was used as the meta model and the base model included random forest classifier, decision tree classifier, K-nearest neighbors and MLP classifier. With the default parameters the model achieved accuracy of 84.47% (Table 2) with the hyperparameter tuning with the best parameters 'max_depth: 20', 'n_estimators': 200' the model showed slight improved performance to 84.96%. The F1 score, precision and recall only had the slight difference in the numbers showing that the tuning did not significantly increase the performance of the model. Fertilizer model performance was stronger compared to the crop it showed 94.05% to 96.47% after the tuning. Precision, Recall showed in the value after tuning, increase in the value after the tuning increase and the F1 score to 0.95 showed that the model there is balance in between the precision and the recall. Fertilizer Recommendation model showed the effectiveness of the usage of the hyperparameter tuning model.

Table 2: Evaluation Metrics for Different Meta Models

Meta Model	Recommendation	Accuracy	Accuracy (Tuned)	Precision	Precision (Tuned)	Recall	Recall (Tuned)	F1 Score	F1 Score (Tuned)
Random Forest	Crop	84.47%	84.96%	0.86	0.88	0.26	0.27	0.24	0.24
	Fertilizer	94.05%	96.47%	0.93	0.95	0.92	0.94	0.93	0.95
Decision Tree	Crop	82.48%	84.85%	0.86	0.90	0.24	0.28	0.22	0.25
	Fertilizer	94.68%	96.66%	0.94	0.95	0.95	0.95	0.94	0.95
K-Nearest Neighbors	Crop	88.36%	89.56%	0.62	0.62	0.44	0.48	0.45	0.49
	Fertilizer	94.77%	96.60%	0.94	0.95	0.94	0.95	0.94	0.95
Multi-Layer Perceptron	Crop	86.85%	87.14%	0.64	0.61	0.33	0.35	0.33	0.35
	Fertilizer	94.62%	96.77%	0.94	0.95	0.94	0.95	0.94	0.95

The individual model of the Random Forest Classifier showed the comparatively less accuracy of 75.04 % crop and 66.91 % fertilizer recommendation (Table 3), the crop had the recall and the F1 score around the same level and the precision of 0.96, similar case was seen in the fertilizer recall and the F1 having the same value of (0.51,0.53) and having the precision of 0.84. Model having the higher precision and the recall show that it is more accurate when prediction the positive prediction, thought the model is missing to predict the lot of true positives.

Table 3: Evaluation Metrics for Individual Models

Model	Recommendation	Accuracy	Recall	Precision	F1 Score
Random Forest	Crop	75.04%	0.15	0.96	0.14
	Fertilizer	66.91%	0.51	0.84	0.53
Decision Tree	Crop	75.15%	0.18	0.89	0.17
	Fertilizer	63.46%	0.65	0.69	0.62
K-Nearest Neighbors	Crop	85.19%	0.36	0.49	0.38
	Fertilizer	90.44%	0.89	0.89	0.89
Multi-Layer Perceptron	Crop	88.00%	0.41	0.59	0.41
	Fertilizer	94.60%	0.94	0.94	0.94

6.2 Case Study 2: Decision Tree Model as the Meta Model

Decision Tree model was used as the meta model in the second case, particularly in the fertilizer recommendation it achieved the accuracy of 96.77 % also maintaining a high precision and the recall, on other side crop recommendation was lower with the accuracy 87.14% and the F1 of 0.35. The individual model Decision Tree it had 75.15% for the crop and the 63.16% for the fertilize, while there was balance in the F1 and the precision for the fertilizer it was lower for the crop. Hyper tuning of the DT showed that there was little increase in the accuracy of the crop models recommendation model, and it had dip in the fertilizer recommendation model the parameters used for are training are ‘max_depth: 20’, ‘min_samples_split: 10’. The meta model show that integrated the predictions of the other models and enhance the performance with increase in the accuracy. In the overall view the decision trees showed better performance when using but the standalone model had more true positive value when compared it to the meta model of the decision tree (Figure 11), the standalone model had ROC curve are of the 0.73 and the 0.27 with the stacked meta model, for the fertilizer recommendation and the individual model should the reasonable performance in the fertilizer recommendation it missed to predict the number of true positives.

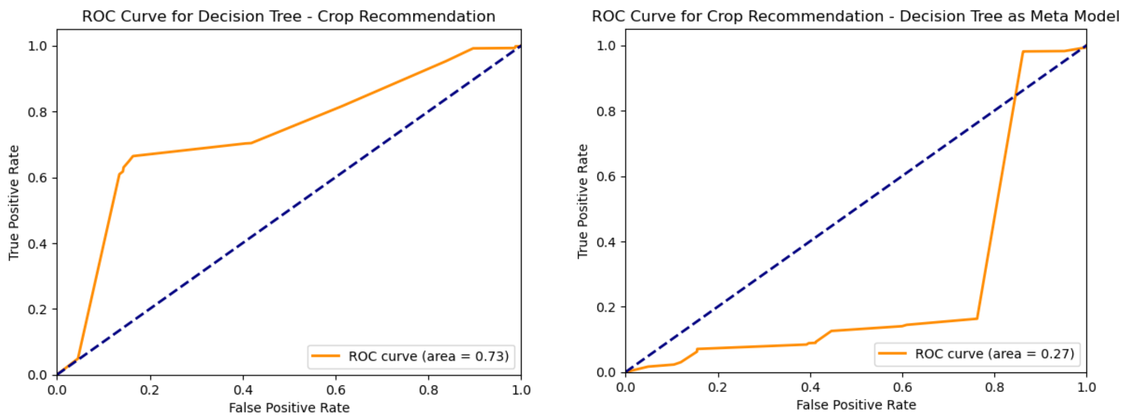

Figure 11: ROC Curve of Decision Tree of Standalone and Stacking Modelling

6.3 Case Study 3: K-Nearest Neighbors as the Meta Model

Using the K-Nearest Neighbours as the meta model showed an improvement when comparing it with the standalone and other meta models. The crop recommendation that had an increase in the accuracy of 88.36% and the tuned model showed 89.56% a dip in the precision to 0.62, the recall of the model improved to 0.44 and the tunned recall had good increase to 0.48, F1 score of the model only showed a slight increase in the score. The fertilizer recommendation should only a slight increase in the accuracy when comparing with other meta models. Precision value remained stable and the recall had a slight increase. Overall, the KNN metamodel should the improvement in the recall and the F1score for the crop production and high stable precision in the fertilizer recommendations. But the model predicted a greater number of false positive in the crop and fertilizer recommendation, after the hyper tuning the model showed little increase in the prediction of true positive.

6.4 Case Study 4: Multilayer Perceptron as the Meta Model

Certain changes were observed in the results of MLP model, the individual model accuracy increased to 88% which was higher than all the individual model Random Forest, Decision Tree and KNN. Stacked meta model of MLP crop recommendation accuracy was little less when comparing it to the KNN meta model. The hyper tuning of the model showed slight drop in the value in the precision. Tunned F1 Score and Recall of the model increased to 0.35. The Fertilizer recommendation accuracy which remains at around the 94% for the individual and the stacked model and the precision was consistent. Result indicates that the MLP Model as meta model achieved the balanced performance with the improvement in the metrics after the hyperparameter tuning.

6.5 Discussion

This section is the review of the four recommendation system case studies that was implemented using the Random Forest, Decision Tree, K-Nearest Neighbors and Multilayer Perceptron as Metamodel and Standalone Model. In first the case the Random Forest model showed a balanced performance with good accuracy but had lot of high number of the false positives particularly in the fertilizer recommendation. Standalone Model had accuracy of 75.04% but the precision value was too high 95.6% indicated that model had few correct number predictions. Decision Tree that had better performance in the standalone mode in the crop recommendation of 75.14%, the precision of the fertilizer model was lower comparing to that of Random Forest, the KNN after tuning the accuracy of the model improved but the precision and the recall metrics showed a large number of false positives particularly in the crop recommendation. MLP model had the high standalone model of accuracy of the 88% the stacked model performance was lower than KNN, recall improved in the hyperparameter tuning showing that the model balanced improvement across the metrics.

In all the observation across all the models it was noted that there was high number of false positives, particular in the fertilizer recommendation of the all model. This show that the models were good in the identifying the correct class, often it misclassified the wrong class as well. Hyperparameter tuning showed the marginal improvement in the precision and in the recall but did not reduce a lot of the false positive values. This show that more advanced technique is needed and the real time data gathering system needs to be used. Methods like Adaptive Learner can be adapted to this domain to reduce the False Positive (Pietraszek & Tanner, 2005). This technique showed good results in the classifying the alerts with the High accuracy and can potentially improve the performance of the crop and fertilizer recommendation systems by minimizing the false positives rate through adaptive learning technique.

7 Conclusion and Future Work

The experiments case results demonstrated that the stacking model provided better performance in terms of the accuracy compared to the individual models. But it had high false positive rates this show that area for improvement. Key findings include that stacking models after the hyperparameter tuning there was marginal improvement in the precision and the recall but it did not significantly reduce the false positive. MLP model achieved the high standalone accuracy, but did not perform well in the stacked meta model when compared to KNN. The implementation of this system in real-world scenarios has some risk due to the high false positive that may lead to incorrect recommendation will potentially affect the crop productivity and the fertilizers usage. Future research should focus on using the data augmentation techniques to increase the diversity and the volume of the training data and to develop system that integrate the real-time weather data information. To reduce the false positive methods like bagging and boosting can be used to reduce the false negative (Yap et al., 2014). Furthermore, the Hybrid recommendation system ((Paradarami et al., 2017) can be incorporated using artificial neural networks and different combination of models can be used to minimize the misclassification using the deep learning approaches. By applying this technique, it is possible to achieve a better result of the recommendation systems.

References

Banerjee, S., & Chand Mondal, A. (n.d.). A Region-Wise Weather Data-Based Crop Recommendation System Using Different Machine Learning Algorithms. International Journal of Intelligent Systems and Applications in Engineering, 2023(Vol. 2023, Issue 3).

Birthal, P. S., Khan, T., Negi, D. S., & Agarwal, S. (2014). Impact of Climate Change on Yields of Major Food Crops in India: Implications for Food Security. Agricultural Economics Research Review, 27(2), 145. https://doi.org/10.5958/0974-0279.2014.00019.6

Chauhan, B. S., Prabhjyot-Kaur, Mahajan, G., Randhawa, R. K., Singh, H., & Kang, M. S. (2014). Global Warming and Its Possible Impact on Agriculture in India. In D. L. Sparks (Ed.), Advances in Agronomy (Vol. 123, pp. 65–121). Academic Press. https://doi.org/10.1016/B978-0-12-420225-2.00002-9

Cravero, A., Pardo, S., Sepúlveda, S., & Muñoz, L. (2022). Challenges to Use Machine Learning in Agricultural Big Data: A Systematic Literature Review. Agronomy, 12(3). https://doi.org/10.3390/agronomy12030748

Dhanapal, R., AjanRaj, A., Balavinayagapragathish, S., & Balaji, J. (2021). Crop price prediction using supervised machine learning algorithms. Journal of Physics: Conference Series, 1916(1), 12042.

Elbasi, E., Zaki, C., Topcu, A. E., Abdelbaki, W., Zreikat, A. I., Cina, E., Shdefat, A., & Saker, L. (2023). Crop Prediction Model Using Machine Learning Algorithms. Applied Sciences, 13(16). https://doi.org/10.3390/app13169288

Gosai, D., Raval, C., Nayak, R., Jayswal, H., & Patel, A. (2021). Crop Recommendation System using Machine Learning. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 558–569. https://doi.org/10.32628/CSEIT2173129

Gupta, R., Sharma, A. K., Garg, O., Modi, K., Kasim, S., Baharum, Z., Mahdin, H., & Mostafa, S. A. (2021). WB-CPI: Weather-Based Crop Prediction in India Using Big Data Analytics. IEEE Access, 9, 137869–137885. https://doi.org/10.1109/ACCESS.2021.3117247

Kaiser, J. (2014). Dealing with Missing Values in Data. Journal of Systems Integration, 5(1), 42–51.

Kumar, P., Varun Prabhu, H., & Champa, H. N. (2023). Crop Recommendation using Ensemble Stacking Machine Learning approach. 2023 IEEE 3rd Mysore Sub Section International Conference, MysuruCon 2023. https://doi.org/10.1109/MysuruCon59703.2023.10396866

Laxmi, R. R., & Kumar, A. (2011). Weather-based forecasting model for crops yield using neural network approach. Statistics and Applications, (Vol. 9).

Maharana, K., Mondal, S., & Nemade, B. (2022). A review: Data pre-processing and data augmentation techniques. Global Transitions Proceedings, 3(1), 91–99. https://doi.org/10.1016/j.gltp.2022.04.020

Medar, R. A., Rajpurohit, V. S., & Ambekar, A. M. (2019). Sugarcane Crop Yield Forecasting Model Using Supervised Machine Learning. International Journal of Intelligent Systems and Applications, 11(8), 11–20. https://doi.org/10.5815/ijisa.2019.08.02

Oswal, M., Mahajan, K., Pagare, S., Kasa, V., Malhotra, J., & Sarode, S. (2021). Feature-Based Analytical Crop Recommendation System. 2021 IEEE Pune Section International Conference, PuneCon 2021. https://doi.org/10.1109/PuneCon52575.2021.9686530

Paradarami, T. K., Bastian, N. D., & Wightman, J. L. (2017). A hybrid recommender system using artificial neural networks. Expert Systems with Applications, 83, 300–313. https://doi.org/10.1016/j.eswa.2017.04.046

Pietraszek, T., & Tanner, A. (2005). Data mining and machine learning: Towards reducing false positives in intrusion detection. Information Security Technical Report, 10(3), 169–183. https://doi.org/10.1016/j.istr.2005.07.001

Priyadharshini, A., Chakraborty, S., Kumar, A., & Pooniwala, O. R. (2021). Intelligent Crop Recommendation System using Machine Learning. Proceedings of the 5th International Conference on Computing Methodologies and Communication, ICCMC 2021, 843–848. https://doi.org/10.1109/ICCMC51019.2021.9418375

Saputra, R. A., Tisnanta, H. S., Sumarja, F. X., & Triono, A. (2022). Agricultural land conversion for housing development and sustainable food agricultural land. www.techniumscience.com

Schratz, P., Muenchow, J., Iturritxa, E., Richter, J., & Brenning, A. (2019). Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data. Ecological Modelling, 406, 109–120. https://doi.org/10.1016/j.ecolmodel.2019.06.002

Schwalbert, R. A., Amado, T., Corassa, G., Pott, L. P., Prasad, P. V. V., & Ciampitti, I. A. (2020). Satellite-based soybean yield forecast: Integrating machine learning and weather data for improving crop yield prediction in southern Brazil. Agricultural and Forest Meteorology, 284. https://doi.org/10.1016/j.agrformet.2019.107886

Schwertman, N. C., Owens, M. A., & Adnan, R. (2004). A simple more general boxplot method for identifying outliers. Computational Statistics and Data Analysis, 47(1), 165–174. https://doi.org/10.1016/j.csda.2003.10.012

Sharma, P., Dadheech, P., & Senthil, A. V. S. K. (2023). AI-Enabled Crop Recommendation System Based on Soil and Weather Patterns. In Artificial Intelligence Tools and Technologies for Smart Farming and Agriculture Practices (pp. 184–199). IGI Global. https://doi.org/10.4018/978-1-6684-8516-3.ch010

Shi, H., Wang, P., Yang, X., & Yu, H. (2020). An Improved Mean Imputation Clustering Algorithm for Incomplete Data. Neural Processing Letters, 54. https://doi.org/10.1007/s11063-020-10298-5

Shook, J., Gangopadhyay, T., Wu, L., Ganapathysubramanian, B., Sarkar, S., & Singh, A. K. (2021). Crop yield prediction integrating genotype and weather variables using deep learning. PLoS ONE, 16(6). https://doi.org/10.1371/journal.pone.0252402

Suruliandi, A., Mariammal, G., & Raja, S. P. (2021). Crop prediction based on soil and environmental characteristics using feature selection techniques. Mathematical and Computer Modelling of Dynamical Systems, 27(1), 117–140. https://doi.org/10.1080/13873954.2021.1882505

Wickramasinghe, C. P., Lakshitha, P. L. N., Hemapriya, H. P. H. S., Jayakody, A., & Ranasinghe, P. G. N. S. (2019). Smart Crop and Fertilizer Prediction System. 2019 International Conference on Advancements in Computing (ICAC), 487–492. https://doi.org/10.1109/ICAC49085.2019.9103422

Yap, B. W., Rani, K. A., Rahman, H. A. A., Fong, S., Khairudin, Z., & Abdullah, N. N. (2014). An application of oversampling, undersampling, bagging and boosting in handling imbalanced datasets. https://link.springer.com/chapter/10.1007/978-981-4585-18-7_2