computers

MDPI

Article

Machine Learning-Based Crop Yield Prediction in South India: Performance Analysis of Various Models

Uppugunduri Vijay Nikhil [1], Athiya M. Pandiyan [1], S. P. Raja [1] and Zoran Stamenkovic [2,3,*]

[1] School of Computer Science and Engineering, Vellore Institute of Technology, Vellore 632 014, Tamil Nadu, India; vijaynikhil47@gmail.com (U.V.N.); athiyapandiyan@gmail.com (A.M.P.); avemariaraja@gmail.com (S.P.R.)
[2] Institute for Computational Science, University of Potsdam, An der Bahn 2, 14476 Potsdam, Germany
[3] IHP—Leibniz-Institut für Innovative Mikroelektronik, Im Technologiepark 25, 15236 Frankfurt, Germany
* Correspondence: stamenkovic@uni-potsdam.de or stamenko@ihp-microelectronics.com

Abstract: Agriculture is one of the most important activities that produces crop and food that is crucial for the sustenance of a human being. In the present day, agricultural products and crops are not only used for local demand, but globalization has allowed us to export produce to other countries and import from other countries. India is an agricultural nation and depends a lot on its agricultural activities. Prediction of crop production and yield is a necessary activity that allows farmers to estimate storage, optimize resources, increase efficiency and decrease costs. However, farmers usually predict crops based on the region, soil, weather conditions and the crop itself based on experience and estimates which may not be very accurate especially with the constantly changing and unpredictable climactic conditions of the present day. To solve this problem, we aim to predict the production and yield of various crops such as rice, sorghum, cotton, sugarcane and rabi using Machine Learning (ML) models. We train these models with the weather, soil and crop data to predict future crop production and yields of these crops. We have compiled a dataset of attributes that impact crop production and yield from specific states in India and performed a comprehensive study of the performance of various ML Regression Models in predicting crop production and yield. The results indicated that the Extra Trees Regressor achieved the highest performance among the models examined. It attained a R-Squared score of 0.9615 and showed lowest Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) of 21.06 and 33.99. Following closely behind are the Random Forest Regressor and LGBM Regressor, achieving R-Squared scores of 0.9437 and 0.9398 respectively. Moreover, additional analysis revealed that tree-based models, showing a R-Squared score of 0.9353, demonstrate better performance compared to linear and neighbors-based models, which achieved R-Squared scores of 0.8568 and 0.9002 respectively.

Keywords: crop yield prediction; machine learning

Citation: Nikhil, U.V.; Pandiyan, A.M.; Raja, S.P.; Stamenkovic, Z. Machine Learning-Based Crop Yield Prediction in South India: Performance Analysis of Various Models. *Computers* **2024**, *13*, 137. https://doi.org/10.3390/computers13060137

Academic Editor: M. Ali Akber Dewan

Received: 19 April 2024
Revised: 22 May 2024
Accepted: 23 May 2024
Published: 29 May 2024

1. Introduction

India is an agricultural nation; it relies on agriculture as a major contributor towards its economy. According to the estimates, released by the Ministry of Statistics & Programme Implementation (MoSPI), the Gross Value Added (GVA) of agriculture and allied sectors in 2020–2021 was 20.3%, it was 19% in 2021–2022 and it again came down to 18.3% in 2022–2023 [1]. India being one of the world's agricultural powerhouses, is the world's top producer of several spices, cereal crops such as rice and wheat, fruits and vegetables, along with commercial crops such as tea. Being one of the biggest producers not only produces for its own consumption but is also a key exporter of agricultural goods to several nations, rice and sugar being the major agricultural exports. India's export trade was around Rs 380,000 crore in 2021–2022 (Department of Commerce, Government of India). Majority of India's people live in rural areas, and agriculture is the major source of income for these

people. Moreover, over 50% of the Indian population depend on agriculture. And about 54% of the total workers are agricultural workers [2]. 57% of rural households are engaged in agriculture.

The total food grain production of the country is estimated to be 3296.87 Lakh tons [3]. The major cereal crops are wheat, and rice grown in states such as UP, Punjab, Haryana, West Bengal, Andhra Pradesh. cash crops such as sugarcane are grown in UP and Maharashtra. Tamil Nadu, Kerala, Andhra Pradesh, Rajasthan, Gujarat grow several oil seeds. Fiber crops such as cotton, jute, silk and hemp are also grown in the states of Maharashtra, Gujarat, UP, Kerala. Plantation crops such as tea, coffee, rubber is also grown in the states of Assam, Karnataka, Kerala. Spices such as pepper, ginger, turmeric are grown in Tamil Nadu, Kerala, UP and Andhra Pradesh, this exemplifies how much India is dependent on agriculture and how almost the entire nation is a contributor to producing agricultural crops.

The Indian subcontinent is position in the southern Asia with the Indian Ocean influencing its climate, bordered by the Himalayas in the north. It has a tropical monsoon climate with several parts receiving monsoon showers, but the mean rainfall is very varied, parts of Meghalaya receive the heaviest rainfalls whereas some locations like Ladakh and Thar desert remain dry for most of the year. Regions of Ganga plains and coastal areas receive rainfall during the months of July and August, whereas places like Goa and Hyderabad receive rainfall during June and July. Several coastal areas of Tamil Nadu and West Bengal also face cyclonic rains during the months of October-November. The Indian climate and agriculture are greatly influenced by these rains. India has great annual temperature ranges with the coastal regions having lower moderate temperature, with the desert regions of Rajasthan having extreme high temperatures and the regions of Himalayas having cold temperature. The monsoon winds along with the temperatures and rainfalls greatly affect the agriculture production of India.

India also has a great distribution of soils. Alluvial soil being the most predominant one covers about 40% of the total land area and are present throughout the northern plains and river valleys. Black Soil covers about 15% of the land area and is found in regions of Deccan Plateau and are used to cultivate cotton, pulses and sugarcane. Crops such as wheat, oilseeds and cotton are also grown in red and yellow soils which are found in regions of low rainfall such as Odisha and Chhattisgarh. Peaty soil which is rich in humus is also found in regions of Tamil Nadu, Uttarakhand, Bihar and West Bengal. Laterite soil, Mountain soil, Desert Soil, Saline Soil is also found in parts of India.

These above climactic, soil and other factors greatly influence the agriculture of India and its production capabilities. These factors have a complex relationship that predicts the production of crop and its yield. The changing climate patterns have also made it that much difficult to predict the production of crops year after year.

Crop production and yield is an important parameter that is used to predict if the demand is met by the supply, to ensure if the crops produced meet the consumptive needs of the nation as well as the export needs of the nation. It enables the farmers to predict how much yield their farms will produce, which can be useful in many ways such as to prepare storage, optimize the use of fertilizers and natural resources and to increase efficiency and decrease costs and also allows to choose the crops that give the most yield.

Crop yield prediction is a crucial task that is of great importance in many areas. Crop production depends on several parameters such as the crop itself, the soil, the region, the climate. The parameters behave in a complex fashion that determines the crop production and yield. Traditionally, farmers predict the crop production and yield, by considering the rainfall trends, and the number of crops sown, but the changing climate as well as the intrinsic complexity of the factors that influence the crop production have made the prediction of crop production more difficult and less accurate.

To overcome these limitations, there are other ways in which crop production can be predicted which mainly include linear models, crop models and Machine Learning (ML) models. Linear models aim to predict by assuming the additive nature of parameter under consideration and fall short of effective prediction due this assumption of linearity

of variables. Crop models aim to model by defining and including several explicit ways in which interactions occur through equations, this can be very complicated and expensive to construct along with the necessity of domain-specific knowledge. Crop models may also be computationally slow and resource-intensive.

ML models on the other hand are efficient in agricultural predictions and are not as resource intensive or complicated as crop models. These models are trained on data, and once trained, it can be used to predict crop production and yield. ML models are essentially models that learn over data provided to them by identifying patterns and such, and can use this learned knowledge to predict from unseen data. The availability of huge amounts of data such as climate datasets, soil data, and region data enables the building and training of such ML models to great accuracy, moreover the complex relationships among the weather, region, and soil parameters that influence the production of crops make ML models an optimal choice to tackle this problem statement.

Machine learning is especially useful in agriculture, where a huge amount of data is available and this can be used to make predictions or decisions on plants and animals. It also helps reduce the uncertainty in agricultural trends that are arising due to climatic changes hence providing a more consistent prediction. ML has widespread applications in agriculture such as disease detection, grading, irrigation, weather monitoring, animal welfare monitoring and so on.

In this paper, we have explored several ML models and trained them on the datasets available to predict the crop production and yield. The regions taken into account are the Indian states of Andhra Pradesh, Telangana, Karnataka, Kerala and Tamil Nadu. Figure 1 highlights the areas considered in this research for crop yield prediction.

The crops that have been considered include rice, sorghum, cotton, sugarcane and rabi. Soil data consists of various soil attributes for each of the chosen districts from the chosen states.

The ML models that are used prevalently can be categorized into: regression, clustering, Bayesian models, and Artificial Neural Networks. Clustering Models aim to group data points based on similar characteristics, and hence enable grouping of a dataset into subgroups that have some inherent similarities. Regression and Bayesian models are suitable for predicting various agricultural parameters. We have trained with several ML models such as Linear Regression, Gradient Boosting Regression, Random Forest Regression, K-nearest Neighbors Regression, LGBM Regression and Decision Tree and Bagging Regression. The models have been evaluated using Mean Absolute Error (MAE), Root Relative Squared Error (RRSE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R-squared score.

Prediction of crop production and yield using machine learning models greatly benefit the farmers and to predict the supply to meet the demand in an efficient and optimal way.

Our main contributions are:

- A unique dataset has been compiled by us, encompassing essential factors influencing crop growth such as meteorological factors, soil factors, and agricultural factors for the certain specific states in India as mentioned above.
- A statistical feature analysis was conducted on the influence of specific features related to meteorological, soil, and crop data on crop production and yield. This feature selection analysis helped in identifying the most contributing features to the prediction of crop production and yield. The impact of these features on crop production and yield was examined, focusing on their relationship to the outcome variables.
- A comprehensive analysis has been conducted on the trained ML models, evaluating their performance using selected metrics. The ML models were also compared with each other based on these metrics.

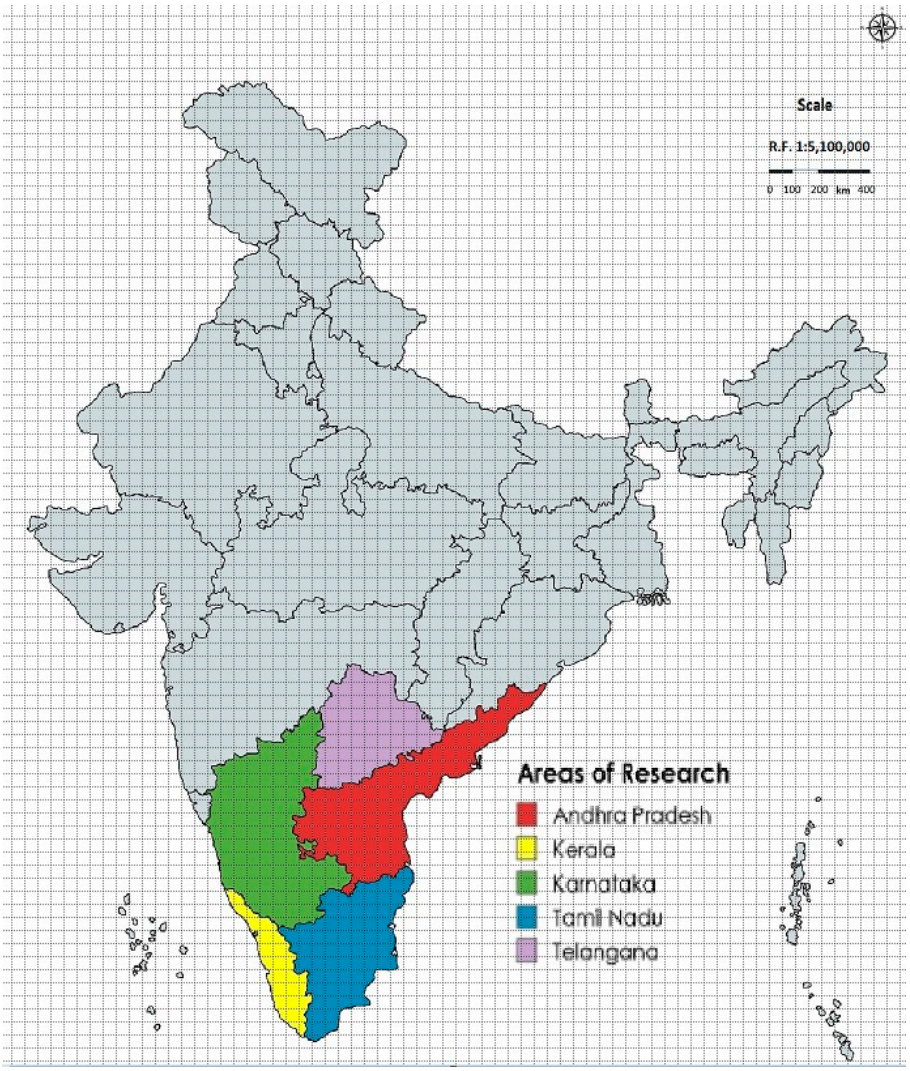

Figure 1. States considered in the research.

2. Literature Review

Anakha Venugopal et al. [4] has proposed a mobile application which predicts the crop's name and also calculates its yield. The dataset used included some meteorological data such as temperature, wind speed and humidity and also the crop production related data. However, the data related to soil was not included in the dataset and the absence of the soil data limited the analysis of soil-related factors. The classification algorithms used in the paper are Logistic Regression, Naïve Bayes and Random Forest. And among these, the highest accuracy of 92.81% was shown by Random Forest model followed by Naïve Bayes with 91.50% accuracy and Logistic Regression with 87.80% accuracy. Thomas van Klompenburg et al. [5] conducted a Systematic Literature Review (SLR) to gather algorithms and attributes used in crop prediction studies. Their analysis revealed that the most used features in these studies are rainfall, temperature and soil type. And also, it showed that the most commonly used algorithm is Artificial Neural Networks (ANN). The paper also discussed the evaluation metrics used in these studies, which revealed that Root Mean Squared Error (RMSE) was the most popular choice. And further, it provided an additional analysis of deep learning-based studies. However, the review doesn't delve into the challenges related to the data quality, feature selection or model interpretability. Sonal Agarwal et al. [6] proposed a hybrid approach for crop yield prediction, combining machine learning and deep learning techniques. The enhanced model provided a better accuracy of 97% compared to the existing model which had an accuracy of 93%. Support Vector

Machine (SVM) was used for machine learning algorithm and for deep learning algorithms, Long Short-Term Memory networks (LSTM) and Recurrent Neural Network (RNN) was used. The study lacks an in-depth exploration of the trade-offs between different hybrid models or their computational requirements.

A.B. Sarr et al. [7] investigate crop yield prediction methods specifically for Senegal. They proposed a study in which they used three machine learning models which are SVM, Random Forest and Neural Network and one multiple linear regression that is Least Absolute Shrinkage and Selection Operator (LASSO) to predict the yield of essential food staple crops in Senegal. They used three combinations of predictors: vegetation data, climate data and combination of both, for training the models. The best performance was shown by models trained with combination of both vegetation and climate data. However, This study overlooked the influence of soil conditions on the prediction of crop yields. S. S Kale et al. [8] aimed to predict crop yield of different crops using neural network regression. Dataset is obtained from Indian government websites for districts in Maharashtra, India. The model predicted with an 82% accuracy by using three layered ANN that uses *Rectified Linear activation function* (RELU) activation function and Adam optimizer. N. Bali et al. [9] explored various machine learning algorithms and techniques used in crop yield prediction, and assesses advanced techniques like deep learning in such estimations and also explores the efficiency of hybridized models. It concluded that factors such as precipitation and temperature were the most influencing factors along with agronomic practices adopted by farmers. ANN and Adaptive Neuro-Fuzzy Inference System (ANFIS), hybridized fuzzy and ANN models showed the best accuracy. In addition to the mentioned studies, various other studies have successfully incorporated neural networks into crop yield prediction, such as those referenced in citations [10–13].

Research conducted by Hames Sherif [14] identified the important factors responsible for staple crops production in semi-arid and desert climates in Africa to predict their yield. Machine learning models used in the research were Multiple Linear Regression (MLR) model and random forest regressor. Metrics such as RMSE and R-squared score were used to compare the performances between the models. It was found that random forest regressor had better accuracy in its predictions compared to the MLR model. However, the run time for training the random forest model was significantly higher than that of MLR model. One limitation of this research paper is that the accuracy of the data obtained from various sources is difficult to assess, as it is largely collected by member countries and includes imputations for missing data whose accuracy is unknown. H.A. Burhan [15] evaluated regression machine learning methods for crop yield prediction of major crops in turkey. The data used to train these models includes pesticide use, meteorological factors and crop yield values. The best R-squared scores were shown by random forest and decision tree regression methods, but support vector regression seemed to show extremely poor performance.

M. Kuradusenge et al. [16] used yields and weather data from a district in Rwanda to predict crop harvest specifically Irish potatoes and Maize. The models used to predict are Polynomial, Random Forest and Support Vector regressors. Among these models, random forest showed the best performance with R-squared score of 0.875 for potatoes and 0.817 for maize. However, the paper did not include other weather-related features such as humidity, wind speed and solar radiation and also soil data for training the models, the predictions made does not incorporate the impact of these factors, despite their actual significance. F. Abbas et al. [17] used four machine learning algorithms, particularly linear regression, elastic net, K-nearest Neighbor (KNN) and Support Vector Regression (SVR) to predict the potato tuber yield from crop and soil data. The best performance among these models is shown by SVR model, while KNN model showed poor performance among them. Furthermore, several other studies have discussed the influence of soil factors on crop yield prediction such as [18,19].

P. Das et al. [20] presented a novel hybrid approach of combining the soft computing algorithm, multivariate adaptive regression spline (MARS) for feature selection with SVR

and ANN models to predict grain yield. The MARS-based hybrid models showed better performance compared to the regular models. Y. Shen et al. [21] proposed an architecture combining long short-term memory neural network and random forest (LSTM-RF) to predict wheat yield using multispectral canopy water stress indices (CWSI) and vegetation indices (VIs) as training data. The combined model, LSTM-RF, showed a better R-squared score of 0.71 than just LSTM model with R-squared score of 0.61. Other research works that utilized LSTM for improved crop yield prediction performance include [22–25].

N. Banu Priya et al. [26] has aimed to predict crop yield production using data mining techniques and machine learning algorithms, to improve the accuracy of the crop production to manage agricultural risk. Random forest regression, decision tree regression and gradient boost regression have been used and have achieved 88% R-squared score with random forest regression. However, a limitation is that the study uses relatively simple factors like the state, district, crop, and season, which may not fully capture the complexity of crop yield variability. The ability of statistical models to predict crop yield production with respect to changes in mean temperature and precipitation was examined by David B. Lobell et al. [27] as simulated by a crop model (CERES Maize). Results suggested that statistical models when compared to crop models represent a useful but imperfect tool in prediction crop production, however it was also observed that they performed better at broader scales concluding that statistical models would still play an important role in predicting the impacts of climate change. One limitation of this study is that these models rely on historical yields and simplified weather measurements, which may not fully capture the complexity of climate impacts on agriculture.

Douglas K. Bolton et al. [28] Used country level data from the USDA (United States Department of Agriculture) from NASA's Moderate Resolution Imaging Spectroradiometer (MODIS) to develop empirical models for predicting soybean and maize in Central United States. It was found that inclusion of phenology data greatly improved the model performance, however crop phenology data may not be readily available especially when taking into account regional crop varieties and it can also complicate the interaction relationships and increase model complexity. The paper provided by Yogesh Gandge et al. [29] aims to predict crop yield to help farmers and government to plan better, and uses data mining techniques to efficiently extract features and data used to predict crop yield and finds that there is still scope for improvement and to use better unified models along with a bigger dataset to predict crop yield at a greater accuracy.

Keerthana Mummaleti et al. [30] analyzed the usage and implementation of ensemble techniques in predicting the crop type from location parameters, by retrieving 7 features from various databases with 28,252 instances. The paper concluded that an ensemble of decision tree regression and Ada boost regression gave the best accuracy, giving a recommendation of which crop should be cultivated in the region based on weather conditions. Although this study uses various models to predict the crop type, it doesn't consider other important climatic factors and soil quality, which can significantly affect the predictions. Shah Ayush et al. [31] suggested the optimal climate factors to maximize crop yield and to predict crop yield, using multivariate polynomial regression, support vector machine regression and random forest models. It uses yield and weather data from United States Department of Agriculture. The paper found that support vector regression obtained the best possible results. One limitation of the paper is that it does not consider other factors such as soil quality, which can also impact yields.

S. Misra Veenadhari et al. [32] predicted the influence of climatic parameters on the crop yields in selected districts of Madhya Pradesh, India but other agricultural parameters were not considered in this paper. A prediction accuracy of 76 to 90% was achieved for the selected crops and districts and an overall prediction accuracy of 82%. The paper's limitation in predicting crop yield using machine learning and climatic parameters lies in its exclusive focus on climate-related factors, neglecting other crucial agro-input variables that influence crop productivity. V. Sellam et al. [33] analyzed the influence of environmental parameters like Area Under Cultivation (AUC), Annual Rainfall (AR), and Food Price

Index (FPI) in crop yield. This has been done using Regression Analysis and achieved an accuracy of 0.7 (R-squared measure) using their linear regression model with least squares fit. The limitation of the paper in its reliance on a single predictive model and data from a single country, suggesting the need for future studies to explore other machine learning algorithms and expand the scope of the research to other regions. P. Mishra et al. [34] used Gradient Boosting Regression to improve the prediction of crop yields for districts in France. The model showed a R-squared score of 0.51 which was significantly better than other models, namely Ada Boosting, KNN, Linear and Random Forests. The limitation of this paper is that it focuses only on predicting maize yields in France and does not consider other crops or regions. Gradient Boosting Regression was also used in other studies [35–37] to improve the accuracy of crop yield prediction.

Leveraging Data Mining techniques, particularly KNN, V. Latha Jothi et al. [38] provided research which focused on using historical data like rainfall, temperature, and groundwater levels to predict future crop production, aiding in analyzing past and predicting future groundwater levels for improved agricultural planning. The limitation of this research paper is the difficulty in estimating the rainfall precisely, which is an important factor for crop yield prediction. Similar research work of using KNN models for crop yield prediction was done in [39–42].

3. Methodology

The steps to predict the crop production yield are:

- Variable identification,
- Data collection,
- Data pre-processing,
- Model training, and
- Model evaluation.

Variable identification, data collection and pre-processing steps are some of the most important steps when it comes to training any ML (Machine Learning) model as the performance of the model depends greatly on the quality, consistency and accuracy of the data it is trained on. Therefore, it is crucial that these steps are done with diligence. Figure 2 shows the overall methodology flow.

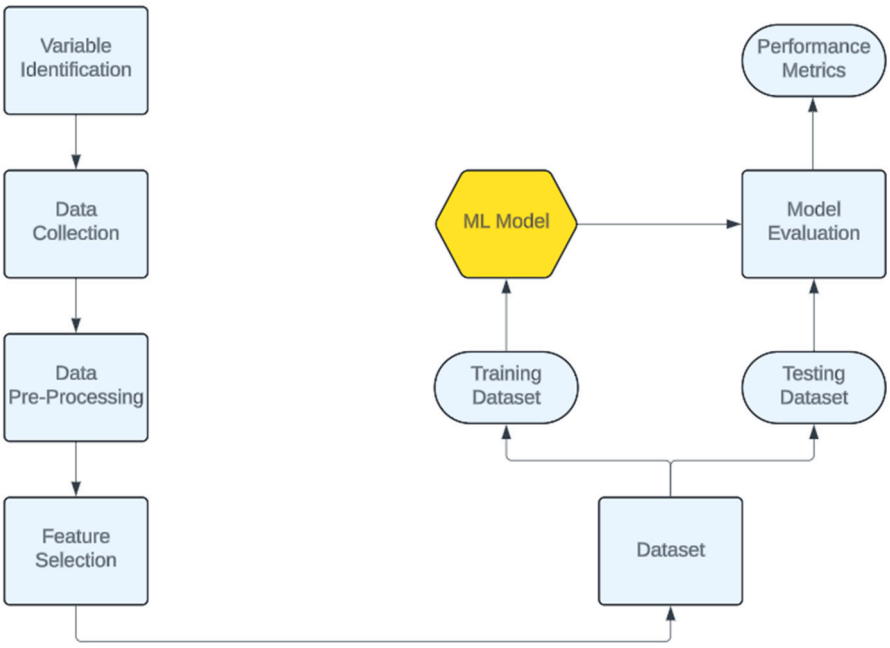

Figure 2. Overall methodology flow.

3.1. Variable Identification

We first identify the variables that are required that will be useful to predict crop yield and production. The most contributing factors that influence crop yield and production are the region under consideration, the meteorological factors, the soil type and soil profile, and the crop itself.

The regions of consideration are identified. We have considered particular districts from the following states in India to predict crop yield and production. The meteorological and soil data attributes are identified.

3.2. Data Collection

Once the required data is identified, we proceeded to identify the data sources to collect the data. The data is from government or other open access data sources.

Data Source

The dataset was built based on the data from official and verified websites, they are:

- data.icrisat.org (http://data.icrisat.org/dld/, accessed on 1 November 2023): The data related to the crop production such as yield, area, crop type etc. was obtained from the ICRISAT website.
- power.larc.nasa.gov: The meteorological data included in the dataset such as temperature, utilized in this study was obtained from this website.
- geoportal.natmo.gov.in: The district wise soil related data in the dataset was obtained from this website.

3.3. Data Pre-Processing

Data pre-processing involves the steps necessary to make the data more usable and in the right structure and format for the ML model. This is done by handling missing values, data discrepancies and inconsistencies. First the independent datasets of meteorological data and agricultural data were merged based on the common district and year attributes which in turn was merged with soil data based on the district attribute. By merging the datasets, a more comprehensive and diverse set of features was created by leveraging the information from multiple sources which is much more usable by the ML model. After merging all the datasets, we have total of 2550 rows (34 districts $\times$ 15 years $\times$ 5 crops). Figure 3 shows the algorithm for merging the datasets.

Input: Meteorological and Crop Dataset, Soil Dataset
Output: Merged Dataset containing all the attributes

- Load the two datasets.
- Choose and verify the attribute that is common in both datasets.
  (District and year are the common attribute)
- Merge or join the two datasets on this common attributes

Figure 3. Algorithm for merging the datasets.

Secondly, the predictor variables that were redundant and could potentially lead to overfitting were dropped. For example, Area and Yield were two predictor variables, the product of which was Production (another existing predictor variable) which is redundant here as it is directly dependent on these two predictor variables. Thus, this variable contributes nothing unique of its own and thus was dropped. This leads to prevention

of the overfitting that would have been caused by an unnecessarily complex model due to the presence of redundant predictor variables. Thirdly, after the above two steps some of the categorical variables must be converted to numeric or variables as the models that are used to train accept numeric values as inputs. Variables like Nitrogen content, Phosphorus content, Soil type, Soil depth, pH was converted using one hot encoding, and Crop Types was converted using label encoding. Finally, the data is normalized, which is the transforming of data into standard ranges. This removes the impacts of different units and scales, increases the speed of model training, helps in giving consistent results and handles outliers and skewed data distributions. We standardized the data by removing the means and scaling it to unit variance. And also, outliers were removed, after which, the range of values for the target variable was between 1 and 1000. This also includes the removal of rows with the target variable as zero, indicating that the specific crop was not cultivated in the corresponding district and year A total of 918 rows were thus excluded. The final dataset, after pre-processing, comprised 38 attributes (21 attributes without one hot encoding) and consisted of 1632 rows. Figure 4 shows the algorithm for encoding and normalization.

Input: Dataset with all the necessary attributes.
Output: Encoded and Normalized Dataset

- The dataset has a number of categorical values which are not suitable as inputs for machine learning models. Therefore, the categorical attributes are encoded.
- Identify all categorical attributes in the dataset.
- Identify the categorical variables for each attribute.
- Add new columns corresponding to each categorical variable and set its value to 1 or 0 depending on whether or not that instance has that categorical value as its attribute (one hot encoding).
- Remove the mean and scale each attribute to unit variance (Normalization)

Figure 4. Algorithm for encoding and normalization.

Feature Selection

The next important step in the pipeline is feature selection. Feature selection is the process of identifying and reducing the dataset to the most important features. This consequently includes removing features that do not affect the output variable or may even interfere and deviate the outputs. This also helps in reducing the dimensionality of the dataset and thus reduces risk of over fitting on the data. Feature Selection can be done in the following major ways: Filter methods, wrapper methods, embedded methods and Dimensionality reduction techniques.

Filter methods essentially access feature relevance by performing statistical tests and select features independently of the ML model. It can be correlation based where the higher correlation features are considered more relevant. It can be chi-squared where the dependence of categorical data with the target variable is evaluated based on the observed versus expected frequencies. It can be information gain which identifies the amount of

information gain a feature contributes towards the target variable. Figure 5 shows the feature selection process.

Figure 5. Feature selection process.

Wrapper methods identify the feature subsets by training different combinations of the dataset on the ML model, thus taking into account the performance of the model and the features influencing the performance. Therefore, here the feature selection is essentially wrapped around the ML model. They are computationally expensive but access feature relevance more accurately. There is Recursive feature elimination (RFE) where we begin with all features and recursively eliminate unimportant features by recursively fitting the data on a smaller subset of features. There is Forward feature selection where we add on the most important features one at a time, backward feature selection which eliminates features based on their contribution to model performance. Embedded methods select the features as a part of the model's training process while considering the performance of the model at the same time. They can be L1 regularization, tree based or neural network based. Tree based methods aim to assign scores of importance to features based on their overall contribution to prediction the output variable whereas neural network-based methods are used for unsupervised feature selection. Dimensionality reduction techniques like PCA aim to reduce the dimensionality of the dataset by transforming the features into a smaller dimension and encompassing most of the information contained in the original dataset. Figure 6 shows the feature selection algorithm.

Input: All features in the dataset.
Output: Extracted features after feature selection

1. Identify k value which is the number of features to be selected.
2. We use ranking based on correlation as the scoring function to determine the most important features.
3. After determining the most important features, we filter them from the dataset.

Figure 6. Feature selection algorithm.

We employed a filter-based feature selection method for its simplicity and computational efficiency. This approach allowed us to identify a subset of the best features by evaluating their relevance to the target variable. To accomplish this, we utilized the f_regression score function, which ranks features based on their capability to explain the variance observed in the target variable. To determine the optimal number of features for selection, we examined the relationship between the number of selected features and the average R2 score across all models as shown in Figure 7. After pre-processing, the dataset contained 38 features available for selection. Through analysis, we discovered that the highest average R2 score was obtained when using a subset of 31 features. Hence, we considered this subset as the optimal number of features for further analysis.

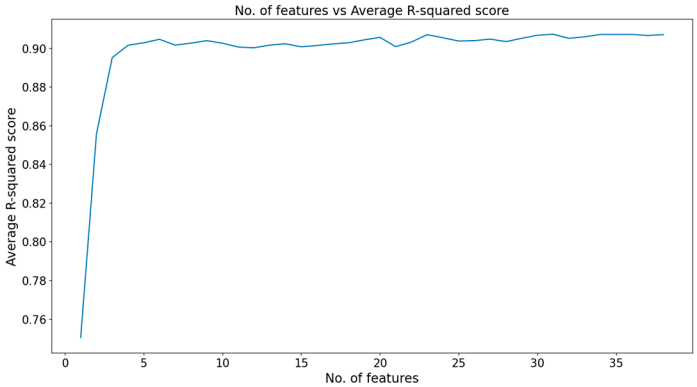

Figure 7. Number of features vs. Average R-squared score.

3.4. Train-Test Split

The next step before training the model on the data is to split the dataset into training and testing datasets. This is done so that the model can be trained on the training dataset, and it can be separately tested on unseen data from the testing dataset. We have chosen 80-20 training-test dataset split for the models.

3.5. Model Training

Now the training data is fed as inputs to the various ML models that we have considered. They include:

3.5.1. Linear Regression

This algorithm aims to fit a straight line by minimizing the deviations of the predicted result from the actual values.

This model is very simple, easy to predict. straightforward and training is quick. It has applications in a wide variety of fields. But this method also has limitations such as assuming that the variables are linearly dependent and have a linear relationship, whereas many real-time datasets have complex non-linear relationships. Moreover, the variables should be independent of each other and the residuals should follow normal distribution and minimal outliers.

Despite these limitations, linear regression performs well in sales predictions, risk assessment and problems where there is an identifiable trend.

$$Y = a + bX + cZ \tag{1}$$

Equation (1) explains linear regression where Y is the target variable which follows a linear dependency with the attributes X and Z.

Ridge Regression is a modification to the linear regression model, in which we try to reduce variance of a linear regression fit by introducing a bias in the minimization equation.

$$\beta_{\text{ridge}} = \text{argmin } | y - X\beta |^2 + \lambda | \beta |^2 \tag{2}$$

Ridge regression generally outperforms linear regression when the dataset is small, because the minimizing of the least squares over fits the data resulting in low variance. It can also be used for discrete data and also with logistic regression. Some of the limitations of this model is that it sometimes shrinks coefficients to zero and trades off variance for bias.

Bayesian Ridge regression is another regression model that aims to find linear regression fits through probability estimates rather than point estimates. Bayesian ridge regression aims to combine Bayesian regression with ridge regression, this done by probabilistically estimating the relationship between attributes and target while also considering uncertainties, making it good for handling multicollinearity. However, it can be computationally expensive. Figure 8 shows the algorithm for linear regression.

Inputs: Training data
Outputs: Trained machine learning (linear regressor) model

1. Initialize model with arbitrary values for the weights and bias. The meteorological, crop and soil data are fed as inputs to the model.
2. Initial predictions of the crop yield are computed from the initialized model.
3. A **loss function** is used to find out how poorly the model performed in its predictions, here mean squared error (MSE) is used as the loss function.
4. Now the weights and bias are updated by using gradient descent optimization algorithm, it computes the negative gradient of the loss function with respect to the weights and bias to minimize the loss function, the step size of the updates is determined by the learning rates.
5. This process is repeated till the line/model that best fits the input data is obtained.

Figure 8. Algorithm for linear regression.

3.5.2. K Neighbors Regression

K Neighbors Regression, unlike the linear regression is a non-parametric model, it uses the average of the neighborhood of X (determined by k value) to estimate Y. For k = 1 it is a perfect fit, but also causes over fitting, whereas on high values of k it performs poorly on both trained and unseen data but an optimum k is identified at the elbow point by cross validation. This model thus performs better than linear regression by allowing for more flexibility through estimations through approximations of neighbors, but it suffers from the problem of dimensionality and is thus unsuitable for datasets with higher number predictor variables.

3.5.3. Decision Tree Regression

Decision tree regression model builds a tree to predict the target variable by splitting data based on features that best predict the target variable (using standard deviation, information gain or entropy) and creating a tree structure with leaf nodes containing the predictions.

This model can handle multiple predictors well but can overfit, to prevent this a minimum number of observations per split is required. It works well with noisy data but is computationally expensive and unstable.

3.5.4. Bagging & Boosting Regressions

Bagging is an ensemble learning technique. Ensemble learning techniques are essentially models which combine multiple regression models, and this is done mainly to reduce over fitting and improve model accuracy. It involves two important processes: bootstrapping and aggregating. Bootstrapping is the division of the dataset and training different models over random samples of it. Aggregating is the combining of the results produced by the different models to produce the final result, this can be done by voting or averaging. Bagging helps with the problem of over fitting as multiple models are trained over different subsets of the data. It is simpler to implement and can also handle complex relationships between variables. Boosting on the other hand sequentially combines multiple weak base

estimators to create a more powerful, robust and accurate model. Figure 9 shows the decision tree regressor flowchart.

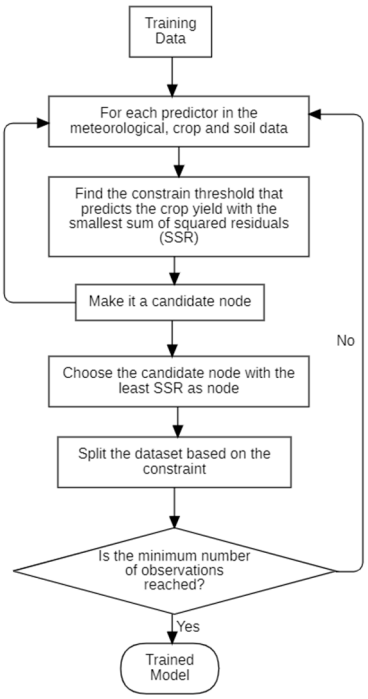

Figure 9. Decision tree regressor flowchart.

3.5.5. Bagging Regression

Uses decision trees as the base estimator and combines their predictions into a final prediction. This is done by bootstrap aggregating. This helps to reduce over fitting and improve the model and enabling it to capture different patterns and relationships in data. The aggregation of predictions is done by averaging the predictions of each base estimator. This technique helps to minimize the impact of outliers and noise in the crop production and prediction dataset. Figure 10 shows the Bagging regressor flowchart.

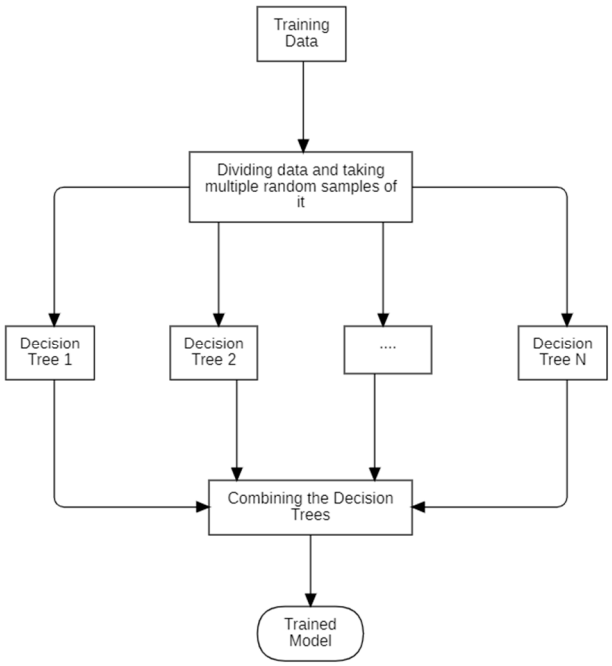

Figure 10. Bagging regressor flowchart.

3.5.6. Random Forest Regression

This is an extension of bagging regression with additional randomization techniques. It performs feature subsampling, optimizes the splits by considering a random subset of features at each node of the decision tree. Random forest regression performs better due to this.

3.5.7. Extra Trees Regression

This is a variation of decision tree-based ensemble models. Additional randomness is introduced during the training when compared to traditional random forests. This model also performs feature subsampling. It also randomly selects threshold for each candidate feature instead of exhaustive searching. The extra randomizations help in further reduction of variance in the crop production prediction. It is robust to noisy data and also performance good on the higher number of features present in this crop dataset. Figure 11 shows the Extra Trees regressor flowchart.

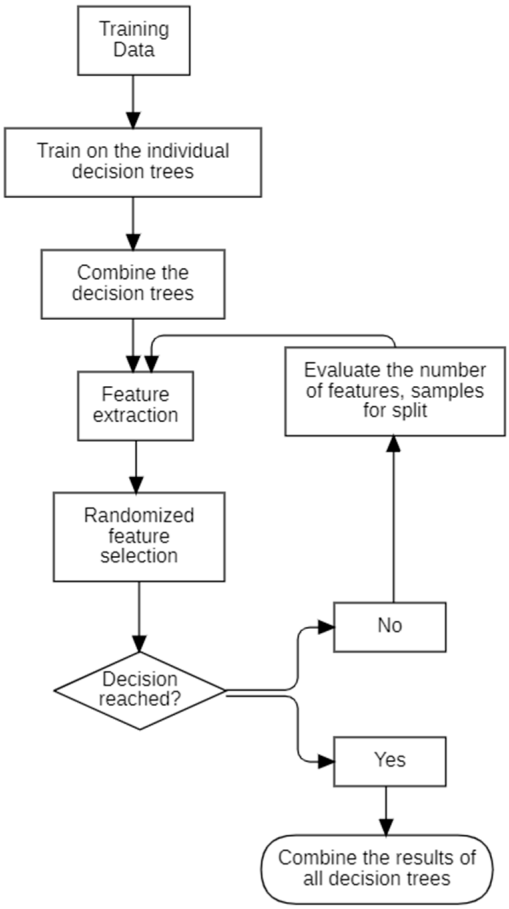

Figure 11. Extra trees regressor flowchart.

3.5.8. Gradient Boosting Regression

This model utilizes gradient boosting to perform regression predictions. It combines multiple weak regression models to create a powerful model. It iteratively builds an ensemble of weak regressors where the loss function to be minimized is the negative gradient of the loss with respect to the previous regressor's predictions. This enables it to capture the complex relationships and patterns that are present in crop and climate data and the factors influencing crop production and yield. Figure 12 shows the algorithm for gradient booting regression.

Inputs: Training data

Outputs: Trained machine learning model

1. Initialize the model
    - Set the crop yield as the initial prediction of the ensemble model.
2. For each weak learner (decision tree) in the ensemble model:
    - Calculate the negative gradient of the loss function by computing the residuals with respect to the model's prediction of the crop yield.
    - Train a weak learner (decision tree) on the residuals by fitting it using the meteorological, crop and soil data as the input features.
    - Update the model's prediction by adding the prediction of the weak learner (decision tree) to it.
    - Calculate the new residuals finding the difference between the decision tree's predictions from the negative gradient values (ensemble model's predictions). Make the new residuals the target variable for the next boosting iteration of the next weak learner (decision tree).

Figure 12. Algorithm for Gradient boosting regression.

3.5.9. Light Gradient Boosting Regression (LGBM)

It is an accurate, efficient and scalable gradient boosting algorithm that is suitable for large scale datasets like crop datasets. It iteratively combines on multiple trained weak learners or regressors in a boosting manner to make better and more accurate predictions. It uses gradient based methods to determine best splits during tree growth. It grows trees leaf wise focusing on leaves with larger gradients, and it can also handle categorical features without encoding.

LGBM is very similar to gradient boosting approach algorithm-wise, but it makes some optimizations to make the training faster and increase performance. Instead of using every unique feature value as a potential split point, it uses a histogram-based approach where it uses bins to discretize the values of the features and uses the histograms to efficiently find the optimal split points.

3.6. Model Evaluation

In model evaluation, the predicted values from the trained models are compared against the actual values to calculate relevant performance metrics. These metrics provide valuable insights for further analysis of the model's performance. Figure 13 shows the model evaluation technique.

Input: Testing Data, Trained ML model
Output: Accuracy/Performance Evaluation of the Model

- We feed the testing data as inputs to the trained model.
- On obtaining the predicted outputs Op from the model, we compare it to the actual outputs.
- Obtain performance metrics from this comparison.

Figure 13. Model Evaluation.

The performance of the trained machine learning models is analyzed based on the below evaluation metrics:

Mean Absolute Error (MAE)

Mean Absolute Error calculates the average absolute difference between the actual and the predicted values. It measures the average magnitude of the errors but does not account for the direction.

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}$$

Mean Squared Error (MSE)

Mean Squared Error calculates the average of squared differences between the predicted and actual values. It measures average squared magnitude of errors and penalizes larger errors more heavily compared to *MAE*.

$$MSE = \frac{\sum_{i=1}^{n} (y_i - x_i)^2}{n}$$

Root Mean Squared Error (RMSE)

Root Mean Squared Error is the square root of *MSE*, which makes it easier to interpret as it is on the same scale as the target variable. *RMSE* measures the standard deviation of the residuals, giving a measure of the average magnitude of errors.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y_i - x_i)^2}{n}}$$

R-squared Score

R-squared is a statistical measure that measures the ability of the independent variables to explain the variation in the target variable. It ranges from 0 to 1. A R^2 of 1 indicates a perfect fit, while 0 shows there is no explanation of the variance.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - x_i)^2}{\sum_{i=1}^{n} \left(\bar{x} - x_i\right)^2}$$

Root Relative Squared Error (RRSE)

Root Relative Squared Error of a model is the square root of ratio of the average of squared differences between its predicted values and actual values to the average of squared differences between predictions of a naïve mean model and the actual values.

$$RRSE = \sqrt{\frac{\frac{1}{n}\sum_{i=1}^{n} (y_i - x_i)^2}{\frac{1}{n}\sum_{i=1}^{n} \left(\bar{x} - x_i\right)^2}}$$

where,

x_i = Actual value
$\bar{x}$ = Mean of the actual values
y_i = Predicted value
n = Number of observations/rows

4. Results and Discussions

4.1. Dataset Description

Data pertaining to the fundamental factors influencing crop prediction, encompassing soil-related factors, meteorological factors, and agricultural factors, has been collected for 34 districts across South India, covering the states of Tamil Nadu, Andhra Pradesh, Telangana, Karnataka, and Kerala, spanning the time period from 2001 to 2015. And the data collected was used to create three primary datasets each containing the attributes related to the discussed key factors. The three primary datasets are soil related dataset, agricultural dataset and meteorological dataset.

4.1.1. Soil Related Dataset

This dataset includes attributes that are associated with the properties and characteristics of the soil that directly affect the crop growth. Table 1 shows a sample of soil dataset. The attributes are:

- Nitrogen (N): The attribute includes values indicating the amount of nitrogen in the soil, classified as low, medium, and high, for the particular district.
- Phosphorous (P) The attribute includes values indicating the amount of phosphorous in the soil, classified as low, medium, and high, for the particular district.
- Potassium (K): The attribute includes values indicating the amount of potassium in the soil, classified as low, medium, and high, for the particular district.
- Soil Type: The attribute 'Soil Type', categorized district-wise, comprises values representing different soil types. The values include red, black, mixed red and black and alluvial soil.
- Soil Depth: The attribute 'Soil depth' refers to the thickness of soil layer at the particular district. This attribute has values categorized as 0 to 25 cm, 50 to 100 cm, 100 to 300 cm and above 300 cm.
- pH: The 'pH' attribute in the soil dataset represents the measurement of the soil's acidity or alkalinity level for the particular district. The values are categorized as Slightly Acidic, Slightly Alkaline, Neutral, Strongly Alkaline and Strongly Acidic.

Table 1. Sample of soil dataset.

District	Nitrogen (N)	Phosphorus (P)	Potassium (K)	Soil Type	Soil Depth	pH
anantapur	low	low	medium	red	100–300	Neutral
chitoor	medium	low	low	red	0–25	Slightly Acidic
guntur	low	low	low	black	above 300	Strongly Alkaline
kadapa	low	low	low	Mixed red and black	100–300	Strongly Alkaline
nellore	medium	low	low	alluvial	above 300	Slightly Alkaline

4.1.2. Meteorological Dataset

This dataset comprises attributes pertaining to the district wise atmospheric conditions and weather conditions information that influence the growth and development of the plants. Table 2 summarizes the attributes of this dataset and Table 3 shows a sample of meteorological dataset.

Table 2. Summary of the attributes in meteorological dataset.

Attributes	Description
PS (Surface Pressure)	Refers to the surface pressure measured in kilopascals (kPa).
TS (Earth Skin Temperature)	Represents to the temperature of the surface of the Earth measured in degrees Celsius (°C).
QV2M (Specific Humidity)	Denotes the specific humidity measured at a height of 2 m above the earth surface expressed in grams per kilogram (g/Kg).
WS2M (Wind Speed)	Indicates the wind speed, the rate at which air is moving horizontally, measured at a height of 2 m above the surface represented in meters per second (m/s).
T2M_MAX (Temperature Maximum)	Represents the maximum temperature recorded at a height of 2 m above the surface measured in degrees Celsius (°C).
T2M_MIN (Temperature Minimum)	Represents the minimum temperature recorded at a height of 2 m above the surface measured in degrees Celsius (°C).
ALLSKY_KT (All Sky Insolation Clearness Index)	Refers to the clearness index of insolation, which is the ratio of the actual solar radiation received on the Earth's surface to the maximum possible solar radiation under clear sky conditions. It is a dimensionless index.
CLOUD_AMT (Cloud Amount)	Indicates the proportion of the sky covered by clouds and is expressed as a percentage (%).
PRECTOTCORR (Precipitation)	Denotes the corrected precipitation measurement, which represents the average amount of precipitation (rainfall) in millimetres per day (mm/day).
ALLSKY_SFC_UVA (All Sky Surface UVA Irradiance)	Refers to the ultraviolet-A (UVA) irradiance received at the Earth's surface under all-sky conditions and it is measured in watts per square meter (W/m^2).
ALLSKY_SFC_UVB (All Sky Surface UVB Irradiance)	Represents the ultraviolet-B (UVB) irradiance received at the Earth's surface under all-sky conditions and the irradiance is measured in watts per square meter (W/m^2).
ALLSKY_SFC_SW_DWN (Sky Surface Shortwave Downward Irradiance)	Refers to the total shortwave downward irradiance received at the Earth's surface under all-sky conditions and the irradiance is measured in kilowatt-hours per square meter per day ($kW\text{-}hr/m^2/day$).
ALLSKY_SFC_PAR_TOT (All Sky Surface PAR Total)	Represents to the total surface photosynthetically active radiation (PAR) in watts per square meter (W/m^2) under all-sky conditions.

Table 3. Sample of meteorological dataset.

District	Year	PS	TS	QV2M	WS2M	T2M_ MAX	T2M_ MIN	ALLSK Y_KT	CLOU D_AMT	PRECTO TCORR	ALLSKY_ SFC_UVA	ALLSKY_ SFC_UVB	ALLSKY_SF C_SW_DWN	ALLSKY_SF C_PAR_TOT
adilabad	2001	96.98	27.66	11.78	2.31	46.4	9.43	0.54	55.90	2.18	12.34	0.33	5.05	95.93
adilabad	2002	97.02	28.35	11.37	2.38	45.99	9.44	0.56	49.80	2.36	12.84	0.35	5.28	100.34
adilabad	2003	97.03	28.01	12.07	2.36	46.61	8.28	0.56	55.35	2.72	12.78	0.34	5.22	99.29
adilabad	2004	97.04	27.95	11.87	2.33	45.23	8.00	0.57	49.77	1.85	13.07	0.35	5.30	101.68
adilabad	2005	97.02	27.23	12.08	2.22	46.11	7.53	0.56	53.54	3.24	12.65	0.34	5.18	98.32

4.1.3. Agricultural Dataset

This dataset includes a range of practices and techniques carried out by farmers to enhance crop production. Table 4 shows a sample of the agriculture dataset. The attributes are:

- Crop Type: This attribute denotes the specific type of crops cultivated in the district for the particular year.
- AREA (1000 ha): This attribute represents the total area of land, measured in thousands of hectares, dedicated to crop cultivation in the district.
- IRRIGATED AREA (1000 ha): This attribute signifies the extent of land, measured in thousands of hectares, that is irrigated for crop cultivation in the district.

- YIELD (Kg per ha): This attribute quantifies the crop yield per hectare of cultivated land in kilograms, reflecting the productivity of the district's agricultural output.
- PRODUCTION (in 1000 tons): This attribute indicates the total crop production, measured in KGs, achieved in the district. It is derived from area and yield attribute. This is the target variable that is to be predicted. And also, given the corresponding Area, yield can be calculated from it.

Table 4. Sample of agricultural dataset.

Year	District	Crop Type	AREA (1000 ha)	YIELD (Kg per ha)	IRRIGATED AREA (1000 ha)
2001	anantapur	RICE	71	2880.56	70.94
2002	anantapur	RICE	40	2150	39.9
2003	anantapur	RICE	28.34	2482.36	28.32
2004	anantapur	RICE	33.58	3176.59	33.5
2005	anantapur	RICE	48.15	2607.68	48.06

4.2. Interpretation of Results

Based on the production data from 2001 to 2015 for districts across South India, including Tamil Nadu, Andhra Pradesh, Telangana, Karnataka, and Kerala, it is evident that rice was the most prominently cultivated crop during that period. This trend is attributed to the region's warmer and wetter climate, which provides optimal conditions for rice cultivation. Additionally, crops such as sugarcane and sorghum have significant production levels. Conversely, crops like rabi had significantly lower production levels, suggesting that their cultivation was notably restricted compared to other crops due to their preference for cold and dry climates. The bar plot showing average production (in 1000 tons) for each crop can be shown in Figure 14.

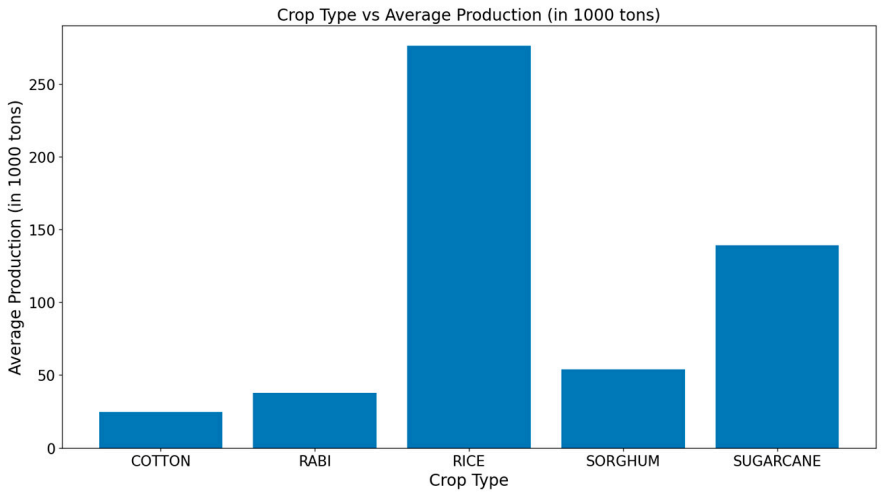

Figure 14. Bar plots showing average production (in 1000 tons) for each crop.

We employed a scatter plot to depict the relationship between irrigated area, measured in thousands of hectares, and production, measured in thousands of tons as shown in Figure 15. Additionally, to observe the underlying trend, we introduced a linear regression line on the plot. The equation of the fitted line was:

$$Production\ (in\ 1000\ tons) = 3.10 \times IRRIGATED\ AREA\ (1000\ ha) + 28.30$$

This line represents the linear relationship between irrigated area and production, where the slope coefficient (3.10) indicates the change in production per unit increase in irrigated area, and the intercept term (28.30) denotes the estimated production when the irrigated area is zero. Moreover, the correlation coefficient (Pearson's r) for this relationship

was calculated to be 0.8931, signifying a strong positive correlation between irrigated area and production.

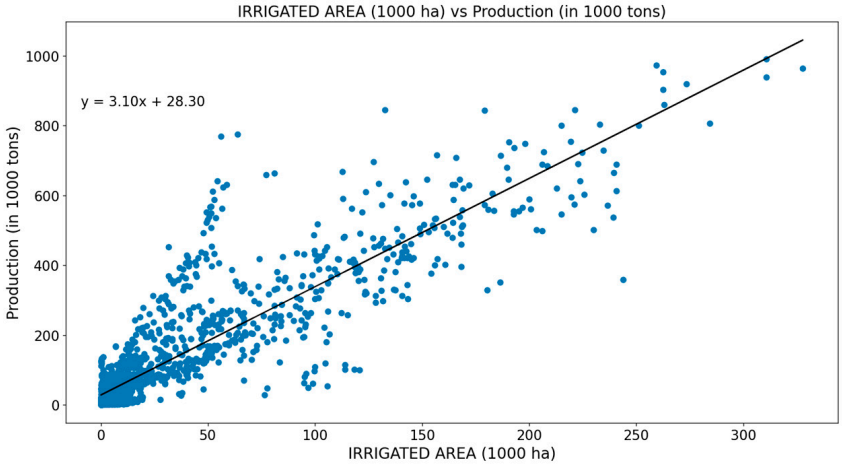

Figure 15. Scatter plot of Irrigated area (1000 ha) vs. Production (in 1000 tons).

In order to examine the influence of meteorological factors on crop production, we specifically focused on rice as our chosen crop type. Various histograms were plotted as shown in Figure 16. The analysis of the histogram depicting the average production based on surface pressure bars revealed that the highest average rice production occurred within the range of 97 to 101 kPa of surface pressure. Increased surface pressure benefits rice crop production by retaining soil moisture and reducing evaporation, which helps regulate temperature. Furthermore, an increase in surface temperature exhibited a positive correlation with average rice production. Regarding specific humidity, the study observed a rise in average rice production with increasing specific humidity until it reached 16 g/Kg, beyond which a decline was observed suggesting that rice crop growth was optimal under moderate humidity levels. The average rice production levels remained relatively consistent until a cloud cover of approximately 60%. However, beyond this threshold, increased cloud cover resulted in a slight decrease of average rice production, possibly due to diminished sunlight availability for photosynthesis, decreased temperatures, heightened humidity fostering moisture stress and fungal diseases, altered microclimates, and reduced solar radiation important for plant metabolism. UVB radiation detrimentally impacts plant physiology by impairing photosynthesis, disrupting nutrient absorption, and stunting overall plant growth. Consequently, these effects contribute to diminished yields and reduced productivity levels. Furthermore, heightened UVB exposure triggers the generation of reactive oxygen species (ROS) within plant cells, exacerbating cellular damage and hindering growth and development.

Taking sugarcane as an example, when analyzing the impact of soil factors on crop yield, it can be observed that sugarcane thrive in areas with slightly acidic soils. This conforms to the fact that sugarcane thrives in slightly acidic soil due to nutrient availability and sugar production benefits within a pH range of 5.5 to 6.5 [43]. Also, it is observed regions with black soil with high water retaining capacity has relatively higher sugarcane production, as shown in the Figure 17.

Table 5 presents the evaluation metrics for various machine learning models trained using the training dataset and then tested on the testing dataset. These training data includes the data of all the crops, where even the crop type is given as an input.

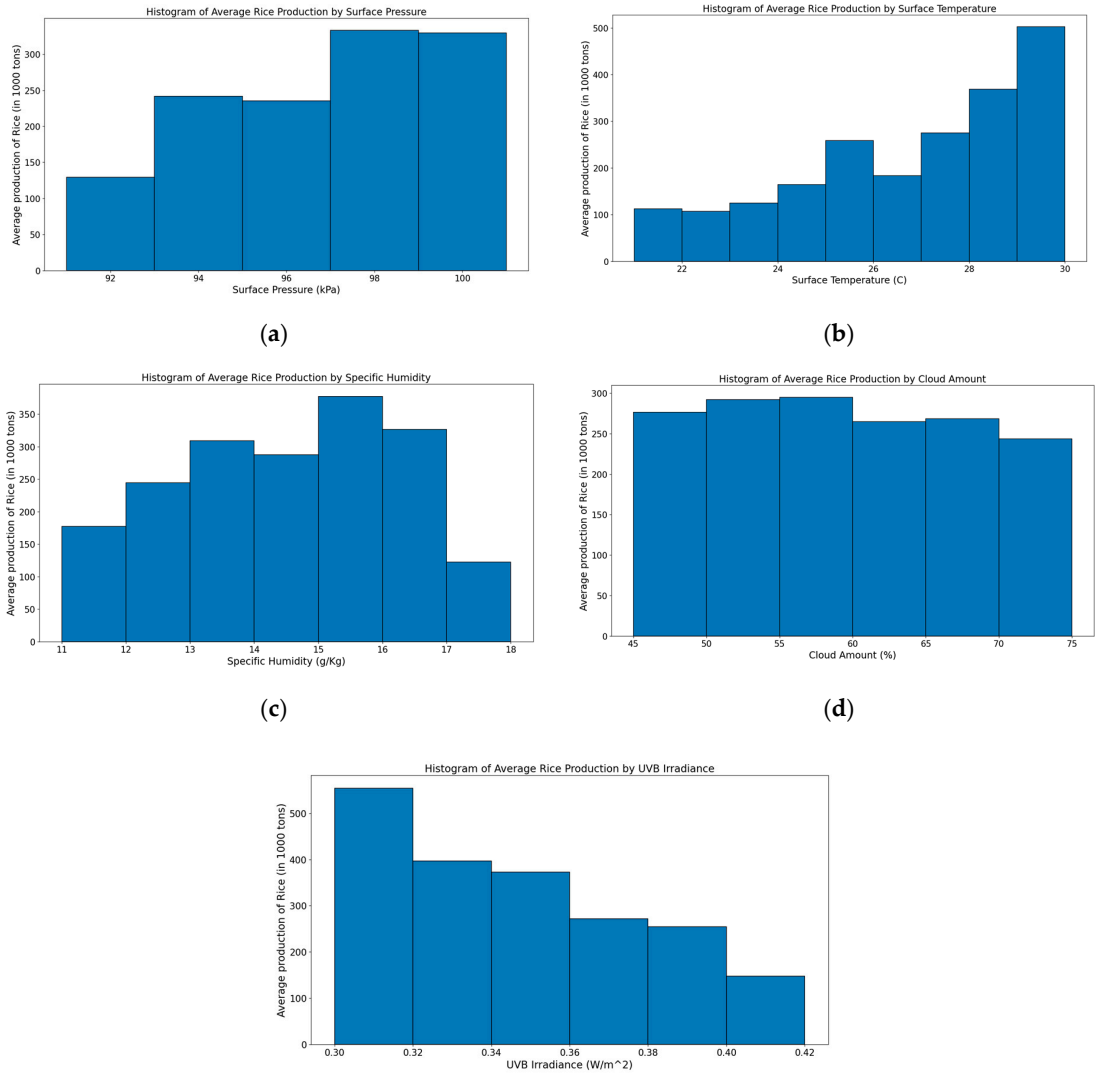

Figure 16. Bar plots showing influence of meteorological factors on rice crop production (**a**) Surface Pressure (kPa) vs. Average production of Rice (in 1000 tons) (**b**) Surface Temperature (°C) vs. Average production of Rice (in 1000 tons) (**c**) Specific Humidity (g/Kg) vs. Average production of Rice (in 1000 tons) (**d**) Cloud Amount (%) vs. Average production of Rice (in 1000 tons) (**e**) UVB Irradiance (W/m^2) vs. Average production of Rice (in 1000 tons).

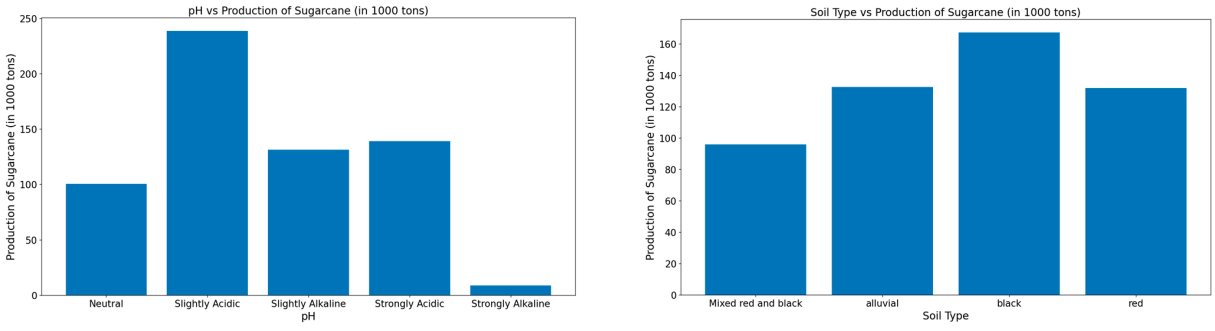

Figure 17. Bar plots showing influence of soil factors on sugarcane crop production (**a**) pH vs. Average production of Rice (in 1000 tons) (**b**) Soil Type vs. Average production of Rice (in 1000 tons).

Table 5. Performance metrics values for each model.

Model	MAE	MSE	RMSE	R-Squared Score	RRSE	Difference = (RMSE − MAE)
Linear Regressor	45.70	4292.94	65.52	0.8570	0.3782	19.82
Gradient Boosting Regressor	30.64	2259.24	47.53	0.9247	0.2744	16.69
Random Forest Regressor	25.30	1690.09	41.11	0.9437	0.2373	16.25
K-nearest Neighbours Regressor	36.13	2996.43	54.74	0.9002	0.3160	18.61
LGBM Regressor	25.85	1808.11	42.52	0.9398	0.2455	16.67
Decision Tree Regressor	31.29	2822.89	53.13	0.9059	0.3067	23.15
Bagging Regressor	25.86	1905.60	43.65	0.9365	0.2520	16.25
Extra Trees Regressor	21.06	1155.31	33.99	0.9615	0.1962	13.06
Bayesian Ridge Regressor	45.64	4303.85	65.60	0.8566	0.3787	19.96
Ridge Regressor	45.67	4298.35	65.56	0.8568	0.3785	19.89

The least MAE can be observed for extra trees regressor indicating this model has the lowest average magnitude of errors in its predictions compared to the other models. And the highest MAE is shown by linear regressor, suggesting the predictions by this model is less accurate compared to other models. While for RMSE values, the least value is still shown by extra trees regressor but the highest value is shown by Bayesian ridge regressor.

$$MAE_{Extra\ Trees} < MAE_{Random\ Forest} \approx MAE_{LGMB} \approx MAE_{Bagging} < MAE_{Gradient\ Boosting} < MAE_{Decision\ Tree} < MAE_{KNN} < MAE_{Bayesian\ Ridge} \approx MAE_{Ridge} \approx MAE_{Linear}$$

$$RMSE_{Extra\ Trees} < RMSE_{Random\ Forest} < RMSE_{LGMB} < RMSE_{Bagging} < RMSE_{Gradient\ Boosting} < RMSE_{Decision\ Tree} < RMSE_{KNN} < RMSE_{Linear} \approx RMSE_{Ridge} \approx RMSE_{Bayesian\ Ridge}$$

Looking at random forest regressor and bagging regressor, the MAE values of these two models are almost the same which indicates similar average performance. But there is a relatively larger difference between their RMSE values, that is, RMSE of bagging regressor is more than that of random forest regressor. This suggests that bagging regressor has larger errors in magnitude compared to random forest regressor, despite their similar average absolute differences. This difference in RMSE values is seen as it penalizes larger errors more than MAE. So, we can say that the predictions for bagging regressor have more uneven distribution of errors as compared to random forest regressor, although having similar average magnitude of errors. This fact can also be observed in Figure 18.

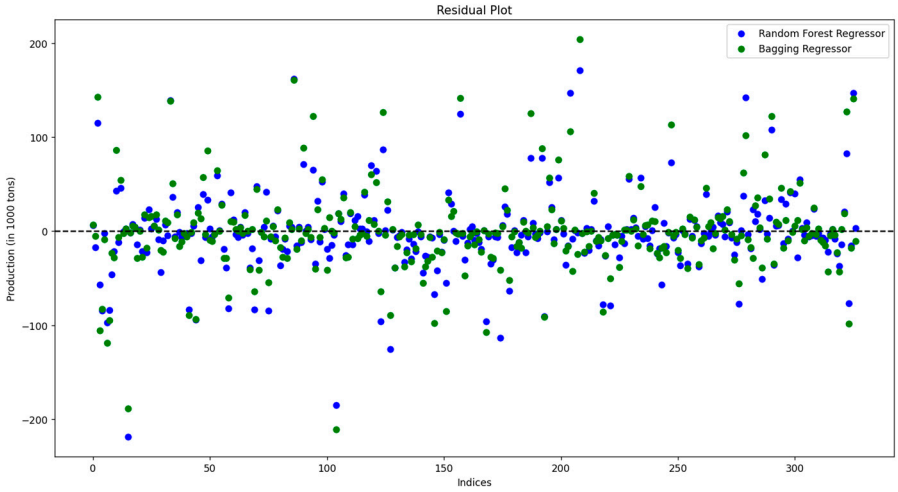

Figure 18. Residual Plots of Random Forest regressor and Bagging regressor.

And also, the difference between the RMSE and MAE value is the most for decision tree regressor, implying it is having the more uneven distribution or higher variance in errors compared to the other models. On the other hand, the lowest difference can be

observed for extra trees regressor, indicating that it has more even distribution or lower variance in errors compared to the other models. As the difference increases, the variance in errors of the model's predictions also increases.

The highest R-squared score is observed for extra trees regressor (0.9615) followed by random forest regressor (0.9437) and LGBM regressor (0.9398). This suggests that these models are better at explaining the variance in the target variable using the independent variable compared to the other models. And Bayesian regressor (0.8566) is having the least R-squared score among the other models indicating it is less effective in capturing the relationship between target variable and the independent variables. Looking at the RRSE values, to compare the models to the baseline model. The baseline model is a simple model which is represented by using the mean value of the dependent variable as a predictor. The RRSE values of all the models ranges between 0.1962 and 0.3787, indicating that the prediction errors made by these models is significantly lower compared to the baseline model.

The models are grouped based on their underlying methodologies and techniques, and then were compared at group level. The above models are grouped in three categories: Linear Models, Tree-Based Models and Neighbors-Based Models as shown in Table 6.

Table 6. Models grouped based on their underlying techniques and methodologies.

Linear Models	Tree-Based Models	Neighbours-Based Models
• Linear Regressor • Bayesian Ridge Regressor • Ridge Regressor	• Gradient Boosting Regressor • Random Forest Regressor • Bagging Regressor • Extra Trees Regressor • LGBM Regressor • Decision Tree Regressor	• K-nearest Neighbours Regressor

The average MAE, MSE, RMSE, R-squared score and RRSE of the models, according to their groups is shown in Table 7.

Table 7. Group wise average performance of models.

Groups	MAE	MSE	RMSE	R-Squared Score	RRSE
Linear Models	45.67	4298.38	65.56	0.8568	0.3784
Tree-Based Models	26.67	1940.20	43.65	0.9353	0.2520
Neighbours-Based Models	36.13	2996.43	54.74	0.9002	0.3160

$$MAE_{Tree-Based} < MAE_{Neighbours-Based} < MAE_{Linear}$$

Here, it can be observed that average MAE values of the tree-based and neighbors-based models are significantly lower compared to that of linear models. The same holds true for RMSE values as well. And the average R-squared scores of linear models is less than that of tree-based and neighbors-based models, suggesting that these are comparatively less effective in explaining the variability in the target variable using the independent variables. The average RRSE values for all the groups are low, suggesting the average prediction errors caused by the model groups are significantly lower compared to the baseline model. Overall, the performance of tree-based and neighbors-based models is better than that of the linear models.

Table 8 displays how well different models performed for each crop. Rice had the highest average R-squared score of 0.9119, showing it was predicted most accurately. Linear models did particularly well for rice, with an average score of about 0.9214. Sugarcane also had good accuracy, with an average score of 0.8472. However, linear models didn't perform as well for sugarcane, with an average score of only 0.7159. For crops like sorghum

and rabi, all types of models did decently, but tree-based models stood out for their good performance. Sorghum had average scores of 0.4350 for linear, 0.4778 for neighbors-based, and 0.8323 for tree-based models. Rabi had scores of 0.5031, 0.6211, and 0.8642 respectively. Cotton's tree-based and neighbors-based models performed well, with scores of 0.7797 and 0.7723. However, linear models for cotton had negative scores, indicating they were worse than simply guessing the mean yield.

Table 8. Performance of the models for each crop.

Crop Type	Model	MAE	MSE	RMSE	R-Squared Score	RRSE
RICE	Linear Regressor	43.76	3178.22	56.38	0.921	0.281
	Gradient Boosting Regressor	44.87	3799.84	61.64	0.9056	0.3073
	Random Forest Regressor	41.75	3206.09	56.62	0.9203	0.2823
	K-nearest Neighbours Regressor	50.02	4558.69	67.52	0.8867	0.3366
	LGBM Regressor	44.86	3753.12	61.26	0.9067	0.3054
	Decision Tree Regressor	48.65	5235.52	72.36	0.8699	0.3607
	Bagging Regressor	41.50	3449.62	58.73	0.9143	0.2928
	Extra Trees Regressor	32.31	1943.77	44.09	0.9517	0.2198
	Bayesian Ridge Regressor	43.60	3150.30	56.13	0.9217	0.2798
	Ridge Regressor	43.63	3163.85	56.25	0.9214	0.2804
SUGARCANE	Linear Regressor	61.96	7581.16	87.07	0.7196	0.5295
	Gradient Boosting Regressor	34.90	2640.61	51.39	0.9023	0.3125
	Random Forest Regressor	31.30	2497.48	49.97	0.9076	0.3039
	K-nearest Neighbours Regressor	39.57	3377.08	58.11	0.8751	0.3534
	LGBM Regressor	26.88	2063.17	45.42	0.9237	0.2762
	Decision Tree Regressor	33.10	3080.69	55.50	0.8861	0.3376
	Bagging Regressor	32.39	2922.90	54.06	0.8919	0.3288
	Extra Trees Regressor	22.74	1685.81	41.06	0.9376	0.2497
	Bayesian Ridge Regressor	62.87	7774.76	88.17	0.7124	0.5362
	Ridge Regressor	62.52	7690.41	87.69	0.7156	0.5333
SORGHUM	Linear Regressor	38.45	2902.37	53.87	0.4285	0.756
	Gradient Boosting Regressor	22.01	1070.03	32.71	0.7893	0.459
	Random Forest Regressor	16.69	721.64	26.86	0.8579	0.377
	K-nearest Neighbours Regressor	36.67	2651.82	51.50	0.4778	0.7226
	LGBM Regressor	16.35	525.66	22.93	0.8965	0.3217
	Decision Tree Regressor	20.49	1166.01	34.15	0.7704	0.4792
	Bagging Regressor	18.93	926.60	30.44	0.8175	0.4271
	Extra Trees Regressor	16.85	700.09	26.46	0.8621	0.3713
	Bayesian Ridge Regressor	38.08	2840.88	53.30	0.4406	0.7479
	Ridge Regressor	38.22	2864.71	53.52	0.4359	0.7511
RABI	Linear Regressor	35.51	3276.84	57.24	0.5012	0.7063
	Gradient Boosting Regressor	22.05	2043.72	45.21	0.6889	0.5578
	Random Forest Regressor	16.40	923.02	30.38	0.8595	0.3748
	K-nearest Neighbours Regressor	28.97	2489.15	49.89	0.6211	0.6155
	LGBM Regressor	14.62	768.85	27.73	0.8830	0.3421
	Decision Tree Regressor	14.29	492.58	22.19	0.9250	0.2738
	Bagging Regressor	14.16	599.75	24.49	0.9087	0.3021
	Extra Trees Regressor	14.02	525.83	22.93	0.9200	0.2829
	Bayesian Ridge Regressor	35.32	3252.62	57.03	0.5049	0.7036
	Ridge Regressor	35.44	3264.14	57.13	0.5031	0.7049

Table 8. *Cont.*

Crop Type	Model	MAE	MSE	RMSE	R-Squared Score	RRSE
COTTON	Linear Regressor	43.51	4045.49	63.60	−2.7917	1.9472
	Gradient Boosting Regressor	16.11	443.91	21.07	0.5839	0.645
	Random Forest Regressor	9.24	183.45	13.54	0.8281	0.4147
	K-nearest Neighbours Regressor	10.06	242.92	15.59	0.7723	0.4772
	LGBM Regressor	10.29	238.09	15.43	0.7768	0.4724
	Decision Tree Regressor	9.36	237.04	15.40	0.7778	0.4714
	Bagging Regressor	9.23	172.37	13.13	0.8384	0.4019
	Extra Trees Regressor	6.80	135.10	11.62	0.8734	0.3558
	Bayesian Ridge Regressor	42.57	3966.81	62.98	−2.7180	1.9282
	Ridge Regressor	42.99	3995.74	63.21	−2.7451	1.9352

5. Conclusions

In conclusion, this research paper has compiled data from various sources to analyze the primary factors influencing crop yield in selected districts. Our findings highlight the importance of soil factors, meteorological conditions, and agricultural practices. Each of these factors was thoroughly investigated by compiling a primary dataset for each category, which was later merged into a comprehensive dataset.

The study also examined the influence of specific features within each factor on crop yield. The original dataset was utilized to train various regression machine learning models, and their performance was compared using metrics such as the R-squared score and RMSE. The Extra Trees Regressor model achieved the highest R-squared score of 0.9615, indicating its good prediction accuracy. Furthermore, the ML models were categorized into distinct groups based on their underlying techniques and methodologies, specifically linear, neighbors-based, and tree-based models.

Analyzing the average performances of these model groups revealed that the tree-based models demonstrated the highest average R-squared score of 0.9353, followed by neighbors-based models with a score of 0.9002, and linear models with a score of 0.8568. Additionally, the study briefly discusses the performance of the models in predicting crop yields for each specific crop, which is presented in a tabulated format.

Overall, this research paper provides valuable insights into the factors influencing crop yield and demonstrates the effectiveness of machine learning models in predicting and understanding agricultural outcomes. The findings contribute to the existing body of knowledge and underscore the significance of considering various factors in optimizing crop production.

Author Contributions: Conceptualization, U.V.N. and A.M.P.; methodology, U.V.N.; software, A.M.P.; validation, U.V.N., A.M.P. and S.P.R.; formal analysis, S.P.R.; investigation, Z.S.; resources, U.V.N. and A.M.P.; data curation, U.V.N. and A.M.P.; writing—original draft preparation, U.V.N.; writing—review and editing, S.P.R.; visualization, S.P.R.; supervision, Z.S.; project administration, Z.S. All authors have read and agreed to the published version of the manuscript.

Funding: The work presented in this paper was supported by the German Federal Ministry for Education and Research in form of the Brandenburg/Bayern Initiative for Integration of Artificial Intelligence Hardware Subjects in University Curriculum (BB-KI Chips), project no. 16DHBKIO20.

Data Availability Statement: Data available in a publicly accessible repository. The original data presented in the study are openly available at http://data.icrisat.org/dld/, http://power.larc.nasa.gov, http://geoportal.natmo.gov.in.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Available online: https://pib.gov.in/PressReleaseIframePage.aspx?PRID=1909213 (accessed on 1 November 2023).
2. Available online: https://pib.gov.in/PressReleasePage.aspx?PRID=1601902 (accessed on 1 November 2023).
3. Available online: https://pib.gov.in/PressReleasePage.aspx?PRID=1968931 (accessed on 1 November 2023).

4. Venugopal, A.; Aparna, S.; Mani, J.; Mathew, R.; Williams, V. Crop Yield Prediction using Machine Learning Algorithms. *Int. J. Eng. Res. Technol. IJERT* **2021**, *9*. Available online: https://ieeexplore.ieee.org/abstract/document/8985951 (accessed on 1 November 2023).
5. van Klompenburg, T.; Kassahun, A.; Catal, C. Crop yield prediction using machine learning: A systematic literature review. *Comput. Electron. Agric.* **2020**, *177*, 105709. [CrossRef]
6. Agarwal, S.; Tarar, S. A Hybrid Approach for Crop Yield Prediction Using Machine Learning and Deep Learning Algorithms. *J. Physics Conf. Ser.* **2021**, *1714*, 012012. [CrossRef]
7. Sarr, A.B.; Sultan, B. Predicting crop yields in Senegal using machine learning methods. *Int. J. Clim.* **2023**, *43*, 1817–1838. [CrossRef]
8. Kale, S.S.; Patil, P.S. A Machine Learning Approach to Predict Crop Yield and Success Rate. In Proceedings of the 2019 IEEE Pune Section International Conference (PuneCon), Pune, India, 18–20 December 2019.
9. Bali, N.; Singla, A. Emerging Trends in Machine Learning to Predict Crop Yield and Study Its Influential Factors: A Survey. *Arch. Comput. Methods Eng.* **2021**, *29*, 95–112. [CrossRef]
10. Khaki, S.; Wang, L. Crop Yield Prediction Using Deep Neural Networks. *Front. Plant Sci.* **2019**, *10*, 621. [CrossRef] [PubMed]
11. Srivastava, A.K.; Safaei, N.; Khaki, S.; Lopez, G.; Zeng, W.; Ewert, F.; Gaiser, T.; Rahimi, J. Winter wheat yield prediction using convolutional neural networks from environmental and phenological data. *Sci. Rep.* **2022**, *12*, 3215. [CrossRef]
12. Gong, L.; Yu, M.; Jiang, S.; Cutsuridis, V.; Pearson, S. Deep Learning Based Prediction on Greenhouse Crop Yield Combined TCN and RNN. *Sensors* **2021**, *21*, 4537. [CrossRef]
13. Sadenova, M.; Beisekenov, N.; Varbanov, P.S.; Pan, T. Application of Machine Learning and Neural Networks to Predict the Yield of Cereals, Legumes, Oilseeds and Forage Crops in Kazakhstan. *Agriculture* **2023**, *13*, 1195. [CrossRef]
14. Sherif, H. Machine Learning in Agriculture: Crop Yield Prediction. Master's Thesis, Rochester Institute of Technology, Rochester, NY, United States, 2022.
15. Burhan, H.A. Crop Yield Prediction by Integrating Meteorological and Pesticides Use Data with Machine Learning Methods: An Application for Major Crops in Turkey. *Ekon. Polit. Ve Finans. Araştırmaları Derg.* **2022**, 1–18. Available online: https://dergipark.org.tr/en/pub/epfad/article/1148948 (accessed on 1 November 2023). [CrossRef]
16. Kuradusenge, M.; Hitimana, E.; Hanyurwimfura, D.; Rukundo, P.; Mtonga, K.; Mukasine, A.; Uwitonze, C.; Ngabonziza, J.; Uwamahoro, A. Crop Yield Prediction Using Machine Learning Models: Case of Irish Potato and Maize. *Agriculture* **2023**, *13*, 225. [CrossRef]
17. Abbas, F.; Afzaal, H.; Farooque, A.A.; Tang, S. Crop Yield Prediction through Proximal Sensing and Machine Learning Algorithms. *Agronomy* **2020**, *10*, 1046. [CrossRef]
18. Chandraprabha, M.; Dhanaraj, R.K. Soil Based Prediction for Crop Yield using Predictive Analytics. In Proceedings of the 2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), Greater Noida, India, 17–18 December 2021; pp. 265–270.
19. Paudel, S.; Nakarmi, R.; Giri, P.; Karki, S.B. Prediction of Crop Yield Based-on Soil Moisture using Machine Learning Algorithms. In Proceedings of the 2022 2nd International Conference on Technological Advancements in Computational Sciences (ICTACS), Tashkent, Uzbekistan, 10–12 October 2022; pp. 491–495.
20. Das, P.; Jha, G.K.; Lama, A.; Parsad, R. Crop Yield Prediction Using Hybrid Machine Learning Approach: A Case Study of Lentil (*Lens culinaris* Medik.). *Agriculture* **2023**, *13*, 596. [CrossRef]
21. Shen, Y.; Mercatoris, B.; Cao, Z.; Kwan, P.; Guo, L.; Yao, H.; Cheng, Q. Improving Wheat Yield Prediction Accuracy Using LSTM-RF Framework Based on UAV Thermal Infrared and Multispectral Imagery. *Agriculture* **2022**, *12*, 892. [CrossRef]
22. Bhimavarapu, U.; Battineni, G.; Chintalapudi, N. Improved Optimization Algorithm in LSTM to Predict Crop Yield. *Computers* **2023**, *12*, 10. [CrossRef]
23. Wang, J.; Si, H.; Gao, Z.; Shi, L. Winter Wheat Yield Prediction Using an LSTM Model from MODIS LAI Products. *Agriculture* **2022**, *12*, 1707. [CrossRef]
24. Di, Y.; Gao, M.; Feng, F.; Li, Q.; Zhang, H. A New Framework for Winter Wheat Yield Prediction Integrating Deep Learning and Bayesian Optimization. *Agronomy* **2022**, *12*, 3194. [CrossRef]
25. Haider, S.A.; Naqvi, S.R.; Akram, T.; Umar, G.A.; Shahzad, A.; Sial, M.R.; Khaliq, S.; Kamran, M. LSTM Neural Network Based Forecasting Model for Wheat Production in Pakistan. *Agronomy* **2019**, *9*, 72. [CrossRef]
26. Banu Priya, N.; Tejasvi, D.; Vaishnavi, P. Crop yield prediction based on Indian agriculture using machine learning. *Int. J. Mod. Agric.* **2021**, *10*, 73–82.
27. Lobell, D.B.; Burke, M.B. On the use of statistical models to predict crop yield responses to climate change. *Agric. For. Meteorol.* **2010**, *150*, 1443–1452. [CrossRef]
28. Bolton, D.K.; Friedl, M.A. Forecasting crop yield using remotely sensed vegetation indices and crop phenology metrics. *Agric. For. Meteorol.* **2013**, *173*, 74–84. [CrossRef]
29. Gandge, Y. A study on various data mining techniques for crop yield prediction. In Proceedings of the 2017 International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEECCOT), Mysuru, India, 15–16 December 2017; IEEE: New York, NY, USA, 2017.

30. Keerthana, M.; Meghana, K.J.M.; Pravallika, S.; Kavitha, M. An ensemble algorithm for crop yield prediction. In Proceedings of the 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), Tirunelveli, India, 4–6 February 2021; IEEE: New York, NY, USA, 2021.
31. Shah, A.; Dubey, A.; Hemnani, V.; Gala, D.; Kalbande, D.R. Smart farming system: Crop yield prediction using regression techniques. In *Proceedings of International Conference on Wireless Communication: ICWiCom 2017*; Springer: Singapore, 2018.
32. Veenadhari, S.; Misra, B.; Singh, C. Machine learning approach for forecasting crop yield based on climatic parameters. In Proceedings of the 2014 International Conference on Computer Communication and Informatics, Coimbatore, India, 3–5 January 2014.
33. Sellam, V.; Poovammal, E. Prediction of Crop Yield using Regression Analysis. *Indian J. Sci. Technol.* **2016**, *9*, 1–5. [CrossRef]
34. Mishra, P.; Khan, R.; Baranidharan, D.B. Crop Yield Prediction using Gradient Boosting Regression. *Int. J. Innov. Technol. Explor. Eng.* **2020**, *9*, 2293–2297. [CrossRef]
35. Lamos-Díaz, H.; Puentes-Garzón, D.E.; Zarate-Caicedo, D.A. Comparison between Machine Learning Models for Yield Forecast in Cocoa Crops in Santander, Colombia. *Rev. Fac. Ing.* **2019**, *29*, e10853. [CrossRef]
36. Pradeep, G.; Rayen, T.D.V.; Pushpalatha, A.; Rani, P.K. Effective Crop Yield Prediction Using Gradient Boosting To Improve Agricultural Outcomes. In Proceedings of the 2023 International Conference on Networking and Communications (ICNWC), Chennai, India, 5–6 April 2023; pp. 1–6.
37. Yasaswy, M.K.; Manimegalai, T.; Somasundaram, J. Crop Yield Prediction in Agriculture Using Gradient Boosting Algorithm Compared with Random Forest. In Proceedings of the 2022 International Conference on Cyber Resilience (ICCR), Dubai, United Arab Emirates, 6–7 October 2022; pp. 1–4.
38. Jothi, V.L.; Neelambigai, A.; Nithish Sabari, S.; Santhosh, K. Crop Yield Prediction Using KNN Model. *Int. J. Eng. Res. Technol. IJERT* **2020**, *8*. [CrossRef]
39. Suresh, A.; Kumar, P.G.; Ramalatha, M. Prediction of major crop yields of Tamilnadu using K-means and Modified KNN. In Proceedings of the 2018 3rd International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 15–16 October 2018; pp. 88–93.
40. Pavani, S.; Augusta Sophy Beulet, P. Prediction of Jowar Crop Yield Using K-Nearest Neighbor and Support Vector Machine Algorithms. In Proceedings of the International Conference on Futuristic Communication and Network Technologies, Niagara Falls, ON, Canada, 9–11 August 2022.
41. Sundari, M.; Rekha, G.; Siva Rama Krishna, V.; Naveen, S.; Bharathi, G. Crop Recommendation System Using K-Nearest Neighbors Algorithm. In Proceedings of the 6th International Conference on Recent Trends in Computing, Chennai Campus, India, 14–15 December 2023; pp. 581–589.
42. Karn, R.K.; Suresh, A. Prediction of Crops Based on a Machine Learning Algorithm. In Proceedings of the 2023 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 23–25 January 2023; pp. 1–8.
43. Cheong, L.R.N.; Kwong, K.F.N.K.; Du Preez, C.C. Effects of sugar cane (*Saccharum hybrid* sp.) cropping on soil acidity and exchangeable base status in Mauritius. *S. Afr. J. Plant Soil* **2009**, *26*, 9–17. [CrossRef]