International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 6s,2025
https://www.theaspd.com/ijes.php

Enhancing Agricultural Yield: A Unified Stacking Ensemble Method for Crop Recommendations using Soil Properties and Weather Attributes

[1]Punith Kumar, [2]Champa H N
[1,2]Department of Computer Science and Engineering, University Visvesvaraya College of Engineering, Bengaluru, India.
* Corresponding author's Email: [1]punithkumar87@gmail.com, [2]champahn@yahoo.co.in

Abstract:
Agriculture plays a remarkably pivotal role in propelling India's economic growth. In the Indian context, the agricultural sector generates more employment opportunities compared to any other domain. It accounts for half of the entire workforce in our nation. However, owing to their limited understanding and lack of awareness about diverse soil constituents and environmental elements, farmers often make incorrect choices regarding the crops they cultivate. This misjudgment significantly hampers agricultural productivity. The creation of a systematic approach that offers scientific guidance to farmers in predicting optimal crops for cultivation based on various factors influencing overall production becomes crucial to surmount this challenge. Accurate crop forecasting holds paramount importance for enhancing agricultural techniques, mitigating risks, ensuring food security, and promoting sustainable farming practices. Throughout the agricultural value chain, such forecasting enables well-informed decision-making and efficient allocation of resources by furnishing insightful insights into crop yields and production. In this endeavor, a dataset incorporating soil parameters like Nitrogen (N), Phosphorous (P), Potassium (K), and soil pH, as well as weather parameters encompassing temperature, rainfall, and humidity for 22 different crops, was employed. As an enhancement derived synthetic data from original data set to bring in novelty in producing larger dataset. Advanced ensemble stacking machine learning techniques were applied to recommend crops with remarkable accuracy and efficiency. Hence, this approach holds the potential to grant farmers greater flexibility and versatility in their agricultural pursuits.
Keywords: *Ensemble stacking, Soil properties and climatic dataset, Crop prediction, Precision agriculture*

INTRODUCTION

Agriculture holds a pivotal role in driving India's economic landscape. Nevertheless, recent times have witnessed the adverse consequences of industrialization and extensive pesticide usage, leading to detrimental effects on soil health. Conventional agricultural practices are proving inadequate to enhance overall productivity. A significant hurdle faced by farmers is the dearth of insights into the most appropriate crops, taking into account their distinct soil prerequisites and prevailing climatic conditions. This deficiency ultimately impacts their productivity. Navigating the complexities of selecting suitable farming techniques and crop choices can be a daunting task for farmers. Frequently, they grapple with determining the optimal crops that yield the maximum output, considering both geographical attributes and financial considerations. Within the realm of agriculture, the ultimate aspiration is to attain the utmost crop yield while minimizing the associated production costs. Addressing these pressing challenges and equipping farmers with enhanced information and decision-making tools can substantially enhance agricultural efficiency and sustainability.Crop prediction models represent advanced tools that amalgamate a variety of data origins, mathematical algorithms, and machine learning methodologies to prognosticate crop yields and production outcomes for specific geographical regions or agricultural domains. These models have become indispensable within contemporary agriculture due to their adeptness in furnishing invaluable insights into crop performance, streamlining the allocation of resources, and curbing risks associated with volatile elements such as weather conditions, pests, and diseases. The foundation of crop prediction models rests on the fusion of diverse datasets, which

encompass historical crop yield records, weather trends, soil attributes, remote sensing images, and agricultural practices. These datasets are sourced from an array of outlets, spanning government entities, research establishments, and repositories maintained by farmers. The acquired data is subsequently subjected to preprocessing and rigorous analysis to unearth underlying patterns and correlations between crop yields and the myriad factors that exert influence. Machine learning techniques, encompassing regression, decision trees, neural networks, and ensemble methodologies, are harnessed to derive insights from historical data and craft predictive frameworks. These frameworks excel in discerning intricate interdependencies among diverse variables, leveraging this comprehension to furnish precise prognostications concerning forthcoming crop yields. A key advantage stemming from crop prediction models lies in their substantial contribution to enhancing agricultural strategizing. Armed with these models, farmers can make judicious choices about optimal planting times and crop selections, calibrate irrigation and fertilization methodologies, and optimize the allocation of resources based on the anticipated crop outputs. Moreover, these models serve as a bulwark against risks by empowering farmers to preemptively address potential hazards such as severe weather occurrences, pest infestations, and diseases.Additionally, crop prediction models assume a pivotal function in safeguarding food security by furnishing policymakers with valuable insights for strategic food supply planning and management. By preemptively identifying potential shortages or excesses in food production, governmental bodies can enact timely measures to stabilize food costs, streamline trade activities, and judiciously allocate available resources. To encapsulate, crop prediction models emerge as indispensable instruments that have brought about a paradigm shift in contemporary agriculture. Through harnessing data, algorithms, and machine learning, these models propel the advancement of sustainable farming methods, financial assurance for farmers, and heightened food security. As the agricultural domain grapples with ongoing challenges, the escalating significance of crop prediction models in molding the trajectory of worldwide food production is undeniable.Crop prediction using the stacking method involves combining the predictions of multiple individual classifiers to make a final prediction. This work create a more complex stacking ensemble where the output of the initial stacking with random forest and naive bayes classifiers is fed into another Random Forest classifier and used appropriate evaluation metrics to assess the performance of the model. The following sections of the paper are organised as follows: section 2 explores related work. The main methodology of the model is demonstrated in section 3. The results and performance analysis are provided in section 4, at the last conclusions are outlined in section 5.

RELATED WORK

Bharath Kumar R et al. [1] introduced a system that employs data mining methodologies to forecast appropriate crops predicated on soil examination outcomes. This predictive framework holds significance not solely for cultivators but also for agricultural institutions, facilitating well-informed choices regarding the optimal crop varieties to cultivate at precise junctures. The primary aims encompass streamlining the process of crop selection, curtailing the manual workload, and proficiently storing and overseeing extensive reservoirs of agricultural data. Through harnessing data mining techniques, the system meticulously evaluates soil attributes, nutrient compositions, and other pertinent factors to offer tailored recommendations for the most fitting crops within specific locales or agricultural domains. Dasari et al.[2] proposed a technique that takes into account three pivotal parameters: soil characteristics, soil types, and crop yield data. Through an in-depth examination of these parameters, the methodology strives to proffer appropriate crop suggestions for farmers to cultivate. Precision agriculture principles are harnessed to curtail the cultivation of unsuitable crops, leading to an amplification in overall productivity. This method confers numerous merits, including refined input and output management efficiency and an enhancement in the farming decision-making processes. The approach employs an ensemble model that integrates majority voting strategies, encompassing machine learning algorithms like random tree, CHAID (Chi-square automatic interaction detection), K-Nearest Neighbor, and Naive Bayes as learners. By amalgamating the outcomes of these algorithms, the recommendation system adeptly furnishes remarkably accurate and efficient crop recommendations, meticulously aligned with the specific soil

parameters.S.Pudumalar et al.[3] devised a framework rooted in the fusion of distinct machine learning (ML) techniques, engineered to anticipate crops well-suited for harvesting. This construct hinges on three pivotal inputs: soil nutrient data, yield data, and soil type data, which harmoniously culminate in precise crop suggestions. Employing an ensemble model methodology, this system adeptly harnesses a spectrum of ML algorithms, encompassing Random tree, CHAID, and SVM. Synthesizing the outcomes of these algorithms empowers the system to present the most credible and probable crop recommendations, all guided by the entered input data. Rikhsit K. et al.[4] presented a study with the goal of assessing different machine learning (ML) technologies to mitigate the risk of erroneous crop selection and attain the highest level of prediction accuracy. For this purpose, the model adopts k-fold cross-validation methodologies, where the value of k is set to 5. In this specific instance, the dataset is partitioned into five subsets. The model is trained on (k-1) subsets and then validated on the remaining subset. This procedure is repeated five times, with each subset serving as the validation set once. The culmination of performance metrics from these iterations culminates in an evaluation of the model's comprehensive efficacy. Subsequent to the comprehensive tests, the model discerns that the Random Forest (RF) machine learning technique delivers the most exceptional performance, underscoring its efficacy in predicting suitable crops grounded in soil attributes, yield statistics, and soil type characteristics. Following suit, the support-vector regression (SVR) technique, coupled with the radial basis function (RBF) kernel, manifests as the runner-up in crop prediction effectiveness.Dr. J. N. Kumar, et al.[5] identified a system that capitalizes on past agricultural records to extract valuable intelligence and generate forecasts concerning yield production. Through an examination of historical data, the system becomes adept at approximating the anticipated yield for impending harvests. This proficiency holds immense significance for farmers as it equips them with vital insights to streamline their agricultural undertakings with precision. The system's predictive prowess extends beyond mere quantitative projections, encompassing qualitative forecasts as well. This expanded scope empowers farmers to preconceive the caliber of their harvests. By factoring in both qualitative and quantitative facets, farmers are empowered to make judicious determinations pertaining to crop administration, storage strategies, and approaches to market their produce.N K Cauvery et al [6] crafted a crop proposal framework that introduces a distinctive approach, amalgamating wisdom from numerous origins through a knowledge-acquisition process. This procedure entails assimilating insights and forecasts derived from a multitude of machine learning (ML) strategies, uniting them into a cohesive representation. By fusing the outputs of myriad ML approaches, the crop proposal framework strives to deliver precise and accurate suggestions regarding crops that are well-suited for cultivation. This amalgamation of diverse ML techniques serves to harness the distinctive merits of each individual model while countering potential limitations, thereby amplifying the precision of predictions. The resultant composite representation stemming from this amalgamative methodology bears immense significance, as it comprehensively accounts for an array of variables and data sources. This broad scope empowers the system to dispense counsel on the most fitting crops for cultivation. The escalated precision attained through this iterative process ensures that farmers are furnished with dependable and well-informed crop recommendations, consequently optimizing their agricultural methods and augmenting productivity. Ajay lokhande et al [7] relies upon a crop dataset encompassing factors such as temperature, rainfall, pH, and humidity pertinent to distinct crops. Employing a diverse array of machine learning techniques, the study endeavors to furnish crop recommendations characterized by both precision and efficacy. The outcome reveals an impressive average accuracy of 99.09% achieved through the utilization of the random forest classifier.Rajak R. K et al [8] leverages data sourced from soil testing laboratories and constructs an ensemble model employing the majority voting strategy. This ensemble model incorporates support vector machine (SVM) and artificial neural network (ANN) as learners to provide crop recommendations tailored to site-specific parameters, demonstrating exceptional accuracy and efficiency. The work [9] Comprise a theoretical and conceptual foundation, the system is based on an integrated approach involving the amalgamation of various components. This framework encompasses the utilization of Arduino microcontrollers for gathering environmental data, machine learning methodologies including Naïve Bayes (Multinomial) and Support Vector Machine (SVM), as well as an unsupervised machine

learning algorithm, K-Means Clustering. Additionally, it incorporates Natural Language Processing, specifically sentiment analysis, an integral facet of Artificial Intelligence. The overarching goal is to recommend the optimal crop for a designated area, accounting for site-specific parameters, all while maintaining an impressive level of accuracy and efficiency.The work in [10] Introduces an approach grounded in IDCSO (Enhanced Distribution-based Chicken Swarm Optimization) coupled with WLSTM (Weight-based Long Short-Term Memory) for the anticipation and counsel of suitable crops. This technique takes into account three critical parameters: the attributes of the soil, the classifications of the soil, and the compilation of crop yield data. Notably, experimental results elucidate that the proposed method, IDCSO-WLSTM, exhibits superior performance compared to its precursor, evident across metrics such as precision, recall, and execution time. A model that uses a semiparametric variant of a deep neural network [11] and Bayesian multi-modeling of deep neuralnets [12] prove instrumental in envisioning the ramifications of climate change on the agricultural domain, yielding crop yield prognostications that are not only more precise and trustworthy but also surpass the standalone capabilities of the 3DCNN (3D Convolutional Neural Network) and ConvLSTM (Convolutional Long Short-Term Memory) networks. This is achieved while effectively factoring in the inherent uncertainties associated with the models. A system for suggesting crops pertaining to a limited selection is initially subjected to preprocessing, and subsequently, the utilization of ensembling techniques plays a pivotal role in effectively categorizing these chosen crops. The individual base learners used in the ensemble model [13] is Random Forest provide the accuracy of 91.99%. Crop recommendations are provided by considering factors such as soil quality, climate conditions, humidity, and other pertinent variables, all aimed at bolstering agricultural output.To anticipate soil fertility and propose optimal crop choices, a quartet of classifiers, namely Artificial Neural Network, Decision Tree, Random Forest, and K-Nearest Neighbors, have been employed. Among these, Random Forest (RF) stands out as the most efficacious algorithm, boasting an impressive accuracy of 98.63% for the crop dataset and 92.61% for the soil dataset, as evidenced by [14]. The work conducted by [15] and [16] introduces an inventive recommendation framework that harnesses Artificial Neural Networks (ANN) to offer tailored crop suggestions. The ANN demonstrates an impressive overall accuracy of 96%, as indicated by [16], while the Decision Tree model achieves an accuracy of 91.5%. On a similar note, [17] utilizes cosine similarity measurements to identify comparable users based on geographic location and employs fuzzy logic to forecast rice crop yields during the Kharif season in the state of Odisha, India. This method is realized through the application of the Mamdani Fuzzy Inference model. The outcomes underscore its capability to provide advanced insights into crop choices prior to the sowing of seeds.In [18], a forward-looking analysis is presented, aims at evaluating the most suitable crop choices for specific weather conditions. The paper also proposes a hybrid recommender system that integrates CBR - Case-Based Reasoning, effectively boosting the system's success rate. In the works of [19], [22], and [23], an array of machine learning algorithms are explored within the context of precision agriculture. Additionally, [20] introduces a lucrative crop recommendation mechanism through the use of an optimized multilayer perceptron regressor. The utilization of sensor data for crop recommendation through machine learning is covered in [21], while a voting classifier-based crop recommendation is discussed in [24]. These contributions collectively offer insightful solutions to the intricacies of crop prediction challenges.In a broader context, these methodologies are strategically developed to elevate agricultural methodologies, furnishing farmers with well-informed suggestions for crop choices that harmonize with their specific soil characteristics and yield-related information. By embracing the precision agriculture paradigm, resource distribution is optimized, wastage is curtailed, and overall productivity is elevated, thereby fostering sustainable and streamlined farming approaches. These systems are dedicated to refining the procedure of crop selection, promoting optimal productivity, and contributing significantly to the enduring growth of the agricultural domain. With the competence to forecast appropriate crops based on soil insights, these systems possess the potential to bring about a transformative shift in contemporary farming methodologies, ultimately amplifying agricultural efficiency.

International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 6s,2025
https://www.theaspd.com/ijes.php

METHOD

A. *Traditonal machine learning algorithms for crop prediction.*

1) Decision Trees: In the realm of crop prediction tasks, decision trees emerge as a widely employed and efficacious machine learning methodology. Their popularity stems from their interpretability, comprehensibility, and versatility in handling diverse data types, whether numeric or categorical. The process commences with the utilization of the training dataset to educate the decision tree model. Through a recursive mechanism, the algorithm dissects the data based on its attributes, ultimately crafting a tree-like structure. The objective is to discern optimal features and thresholds that yield maximal information gain or reduction in Gini impurity at each division. Once the decision tree is honed through training, its predictive capabilities come into play while analyzing the test dataset. Navigating the tree, the model adheres to the established rules at each node, progressing until it arrives at a leaf node, indicative of the anticipated crop classification. However, it's noteworthy that decision trees can encounter limitations, such as susceptibility to minor fluctuations in data and a predisposition to overfitting. To circumvent these challenges, meticulous preprocessing of data and adept model fine-tuning are imperative pre-training steps. In the context of crop dataset predictions, decision trees have demonstrated an impressive accuracy of 99.00% [7].

2) Support Vector Machines: Support Vector Machines (SVMs) emerge as notably potent when confronted with high-dimensional data, exhibiting prowess in navigating both linear and non-linear connections between attributes and target outcomes. One of their salient attributes is sensitivity to feature scaling, whereby it becomes pivotal to normalize the attributes to a uniform range, often accomplished through techniques such as Min-Max scaling or Standardization. In instances where the interplay between attributes and crop categories deviates from linear, SVMs exhibit their prowess by employing the "kernel trick." By opting for diverse kernels like polynomial, radial basis function (RBF), or sigmoid, the model adapts to the intrinsic data characteristics. This adaptive approach leads to a commendable 96.08% accuracy [7], positioning SVMs as more precise than the decision tree algorithm. Notably effective in instances where intricate, non-linear relationships are at play, SVMs excel in scenarios where feature scaling and the judicious choice of kernels play pivotal roles in optimizing predictive accuracy.

3) Logistic Regression: Deployed extensively in binary classification tasks, this algorithm finds its utility in situations where the objective is to prognosticate categorical outcomes that possess two classes, such as "Yes" or "No," or "Crop A" versus "Crop B." While Logistic Regression demonstrates less susceptibility to the impacts of feature scaling in comparison to some other algorithms, adopting feature scaling can occasionally amplify convergence and contribute to the regularization of the model. At its core, Logistic Regression deduces the coefficients of the features, thus delineating an optimal decision boundary capable of segregating the two classes of crop types. Subsequently, this boundary facilitates the classification of data points into distinct crop categories based on acquired insights. For this model, an accuracy of 95.22\% is attained [7]. While surpassing the precision of the decision tree algorithm, it registers a marginally lower accuracy than SVM.

4) Random Forest: Random Forest stands as a remarkably adept choice for tasks involving crop prediction, thanks to its capacity to manage both classification and regression problems, all while affording resilience and mitigating overfitting concerns. This algorithm is emblematic of ensemble learning, constructing numerous decision trees through bootstrap samples derived from the dataset, accompanied by the utilization of random subsets of features. The outcome is determined by each decision tree's vote, where the class with the majority of votes claims victory. Beyond its predictive prowess, Random Forests yield insights into feature importance by assigning scores to each attribute, revealing their respective contributions to crop prediction. By scrutinizing these importance scores, an enhanced understanding of the variables shaping crop forecasting can be gleaned. The model amalgamates the predictions from each individual decision tree to yield the ultimate forecast for each data point. Notably, the accuracy of the random forest algorithm achieves the zenith amongst its counterparts, registering an

impressive 99.09% precision [7].

5) Naïve Bayes Classifier: Naive Bayes operates on a foundation of probability, thus necessitating the conversion of features into discrete variables. When confronted with continuous features, recourse to methods like discretization or binning is required. Through Bayes theorem, Naive Bayes calculates the likelihood of each class, taking into account the provided feature values. The "naive" assumption presumes the conditional independence of all features given the class, streamlining the computation process. By calculating conditional probabilities for each class based on feature values, the model then designates the class with the highest likelihood as the anticipated crop type. While it might not capture intricate interplays between features, Naive Bayes frequently astonishes with its adeptness in diverse classification tasks, including crop prediction. Impressively, this model attains a 99.04% accuracy [7] for the crop dataset.

B. Ensemble stacking machine learning model for crop prediction

As delineated in the preceding section, the findings strongly underscore that random forest and naive bayes stand out as the most adept algorithms for crop prediction in terms of accuracy. Consequently, this novel model adopts these two algorithms as foundational classifiers to construct an ensemble approach. Within this study, the initial stacking ensemble melds the predictions originating from Random Forest and Naive Bayes classifiers, resulting in the formulation of a novel feature set. Subsequently, this freshly minted feature set is inputted into a Random Forest classifier, serving as the meta learner. The final iteration of this process involves the amalgamation of the initial stacking ensemble's output with the original input features, which in turn facilitates the training of the ultimate Random Forest classifier. Figure 1 outlines the system architecture of the proposed stacking ensemble.

1) Dataset Collection: The dataset utilized in this study is sourced from multiple platforms and is characterized as raw data, implying the potential presence of a notable number of inaccuracies and uncertainties. The dataset employed in this research is an amalgamation of two more intricate datasets: the soil content dataset and the climatic condition dataset. The soil content dataset encompasses details concerning the ratios of Nitrogen (N), Phosphorous (P), and Potassium (K) present in the soil, alongside the pH level of the soil. Conversely, the climatic condition dataset encompasses information regarding parameters like rainfall, humidity, and temperature. The culmination of these datasets results in a composite dataset comprising approximately 2200 rows and 8 columns. Each row corresponds to a distinct observation or data point, while each column represents a distinct variable or attribute. For a visual representation, refer to Figure 2, which illustrates a segment of the crop dataset.

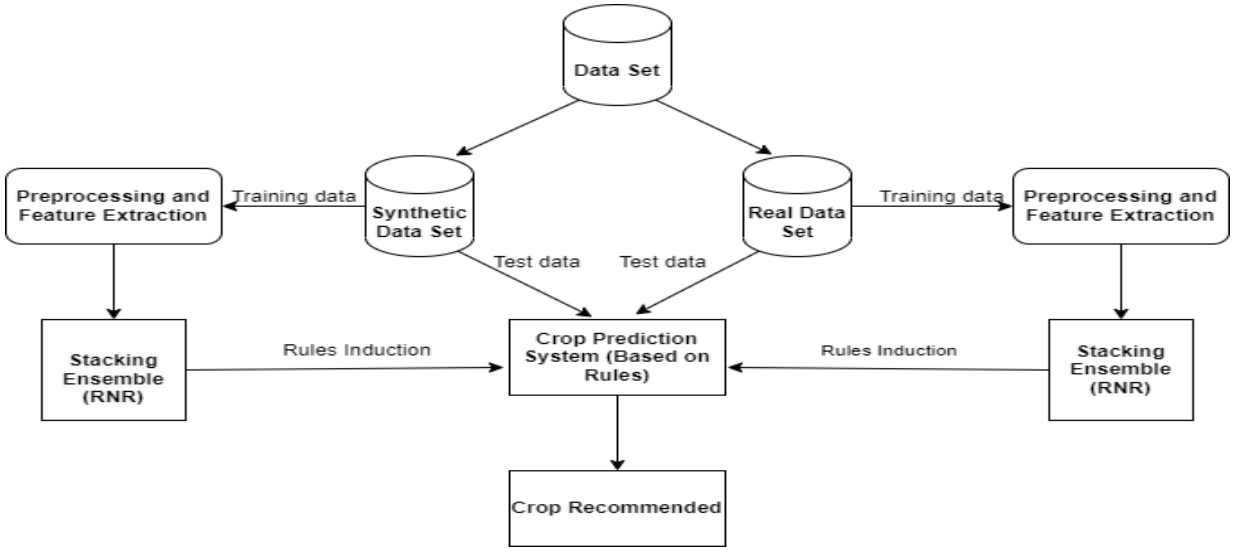

Fig 1: .System Architecture of Stacking Ensemble

International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 6s,2025
https://www.theaspd.com/ijes.php

	N	P	K	temperat	humidity	ph	rainfall	label
2	35	128	205	21.07273	93.56586	6.041054	107.8737	1
3	29	128	198	22.44075	92.70785	5.685062	121.4977	1
4	2	143	196	22.71271	90.45262	5.669489	109.8853	1
5	34	140	198	21.70417	93.44006	5.751707	115.1781	1
6	29	144	204	22.43325	92.48668	5.800449	119.1025	1
7	32	141	203	21.25941	92.84416	5.821348	109.0658	1
8	13	144	197	22.92157	94.89613	6.280223	105.6942	1
9	25	143	198	22.81213	91.51862	6.027314	107.8552	1
10	9	137	200	21.12152	90.68788	5.636687	102.8017	1
11	6	144	198	21.11479	90.31529	5.559364	104.5087	1
12	37	126	196	23.59997	90.97598	5.596449	107.1728	1
13	2	120	203	23.12653	94.71203	5.893493	108.6212	1
14	11	143	197	22.98459	93.32045	5.875719	122.1952	1
15	10	141	201	22.12659	90.97818	6.386021	104.5412	1
16	24	142	202	22.5378	91.48136	5.71082	101.8475	1
17	23	138	195	22.49095	91.70293	5.795986	124.3915	1
18	18	125	204	22.35548	94.47812	6.046674	116.7366	1
19	13	121	196	22.20701	93.50574	6.443383	120.1594	1
20	26	122	202	22.44517	94.73764	5.617227	107.1843	1
21	28	123	202	22.76643	92.12439	6.442289	120.436	1
22	26	121	201	22.19109	90.02575	6.162034	112.3127	1
23	21	137	196	23.61192	91.70294	5.812782	123.5901	1

Fig 2: .Snapshot of the crop dataset

Given the raw nature of the data, which could potentially encompass inaccuracies and uncertainties, it undergoes preprocessing and cleansing before being employed for analysis or modeling purposes. These procedural stages play a pivotal role in upholding the precision and dependability of any conclusions or forecasts derived from the data. Refer to Table 1 for an overview of the dataset's attributes.

Table 1: Dataset details

Lable	Number of rows
Apple	100
Banana	100
Black gram	100
Chickpea	100
Coconut	100
Coffee	100
Cotton	100
Grapes	100
Jute	100
Kidney beans	100
Lentil	100
Maize	100
Mango	100
Moth beans	100
Mung beans	100
Muskmelon	100
Orange	100
Papaya	100
Pigeon peas	100
Pomegranate	100
Rice	100
Watermelon	100

2) Crop prediction model for real data: Crop prediction using the stacking method involves combining the predictions of multiple individual classifiers to make a final prediction.

In the current study, the process followed these steps:

a) Data Preparation: Prepared labeled dataset with input features and target variables (crop types).

b) Splitting the Data: dataset is split into a training set and a validation set.

c) Training the Base Classifiers:

- Trained Random Forest classifier on the training set. This will be the first base classifier in the initial stacking ensemble.
- Trained Naive Bayes classifier on the training set. This will be the second base classifier in the initial stacking ensemble.

d) Generating Predictions from Base Classifiers: Made predictions on the validation set using both the Random Forest and Naive Bayes classifiers.

e) Creating the Initial Stacking Ensemble:

- Combined the predictions from the base classifiers (Random Forest and Naive Bayes) as input features.
- Used the target variables from the validation set to train a new Random Forest classifier, which acts as the meta learner in the initial stacking ensemble.

f) Generating Predictions from the Initial Stacking Ensemble: Made predictions on the validation set using the initial stacking ensemble.

g) Training the Final Random Forest Classifier:

- Combined the predictions from the initial stacking ensemble with the original input features.
- Trained a Random Forest classifier on the combined features, using the target variables from the validation set.

h) Final Predictions: Used the final Random Forest classifier to make predictions on new, unseen data.

i) Evaluation: Evaluated the performance of the final stacking ensemble on a separate test set.

In this approach, an initial stacking ensemble is used to combine predictions from two classifiers: Random Forest and Naive Bayes. These predictions are then used to create a new feature set, which is fed into a Random Forest classifier as the meta learner. The output of the initial stacking ensemble, along with the original input features, is used to train the final Random Forest classifier. By combining the predictions of multiple classifiers and using them as additional features, the final Random Forest classifier can make more informed and accurate predictions. This stacking ensemble technique aims to leverage the strengths of different classifiers and improve the overall performance of the final model. It allows the model to capture complex patterns and relationships in the data, leading to better predictive capabilities and increased accuracy in the predictions.

3) Crop prediction model for synthesized data: The research expands upon previous work [25] by training a stacking ensemble model on synthetic data, generated from an original dataset using the gretel.ai data generation tool. The synthetic dataset comprises 11,000 rows, a fivefold increase compared to the initial dataset, which contained 2,200 rows. Remarkably, the stacking ensemble model achieved an accuracy of 99.86% when evaluated on the synthetic data. This study sheds light on the potential of utilizing synthetic data to augment machine learning models, particularly in enhancing the performance of stacking ensemble techniques. In recent years, the utilization of synthetic data has gained significant attention in various fields, including machine learning and data science. Synthetic data, generated through advanced algorithms and tools, offers a valuable resource for augmenting datasets and enhancing the robustness of machine learning models. In this study, extended previous research by training a stacking ensemble model on synthetic data, generated from an original dataset using the gretel.ai data generation tool. This innovative approach aims to leverage synthetic data to improve the accuracy and performance of stacking ensemble models in predictive analytics tasks.

The methodology involves several key steps. Firstly, we obtain an original dataset comprising 2,200 rows of real-world data. Subsequently, we utilize the gretel.ai data generation tool to generate synthetic data based on the characteristics and patterns observed in the original dataset. The generated synthetic data consists of 11,000 rows, providing a significantly larger dataset for training the stacking ensemble model. Next, we employ the stacking ensemble method, a powerful technique that combines the predictions of multiple base models to improve overall performance. The stacking ensemble model is trained on both the original and synthetic datasets to evaluate its effectiveness in predictive modeling tasks.

RESULS AND DISCUSSIONS

A. Details of the training process:

1) Partitioning the dataset: Partitioning a dataset is a crucial step in machine learning and data analysis. It involves dividing the dataset into separate subsets to facilitate model training, evaluation, and testing. The most common partitioning technique, train-test split method is used. In this method, the dataset is divided into two subsets: a training set and a testing (or validation) set. The model is trained on the training set and evaluated on the testing set to estimate its generalization performance. The split ratio is 80:20 where 80% of the data is used for training, and the remaining percentage 20% is used for testing. Table 2 and Table 3 shows the partitioned sets of real data and synthetic data respectively.

Table 2: Partitioned real dataset

	Number of rows	
Lable	Train	Test
Apple	80	20
Banana	80	20
Black gram	80	20
Chickpea	80	20
Coconut	80	20
Coffee	80	20
Cotton	80	20
Grapes	80	20
Jute	80	20
Kidney beans	80	20
Lentil	80	20
Maize	80	20
Mango	80	20
Moth beans	80	20
Mung beans	80	20
Muskmelon	80	20
Orange	80	20
Papaya	80	20
Pigeon peas	80	20
Pomegranate	80	20
Rice	80	20
Watermelon	80	20
Total	1760	440

2) Training and testing process:

The train-test split data set contains the separate features and target variables for train and test data. Further the train data is split into train and validation sets in the ratio 80:20. This data set is trained and validated using random forest classifier and naive bayes algorithm separately and then make predictions on the train and test data. Further created a new dataset by stacking the predictions of the individual models for train and test data and trained a Random Forest model on the stacked dataset for train and test data respectively. The predictions on the train and test data of the stacked data set is recorded and calculated the evaluation metrics such as average, precision, recall and F1 score. Each model is trained and evaluated for 50 epochs.

Table 3: Partitioned synthetic dataset

	Number of rows	
Lable	Train	Test
Apple	400	100
Banana	400	100
Black gram	400	100
Chickpea	400	100
Coconut	400	100
Coffee	400	100
Cotton	400	100
Grapes	400	100
Jute	400	100
Kidney beans	400	100
Lentil	400	100
Maize	400	100
Mango	400	100
Moth beans	400	100
Mung beans	400	100
Muskmelon	400	100
Orange	400	100
Papaya	400	100
Pigeon peas	400	100
Pomegranate	400	100
Rice	400	100
Watermelon	400	100
Total	8800	2200

3) Discussions on evaluation metrics:

Ajay lokhande et al [7] trained and tested the dataset using random forest, SVM, logistic regression, decision tree, Naive bayes and achieved 99.09%, 96.08%, 95.22%, 99.00%, 99.04% accuracy's respectively. The authors concluded that random forest algorithm is best suited for crop prediction with 99.09% accuracy since there is very less margin of error in prediction. In this work, trained a Random Forest classifier and Naive Bayes classifier on the training set independently. Random Forest is an ensemble learning method that combines multiple decision trees to make predictions. Naive Bayes classifier is a probabilistic classifier based on Bayes' theorem and assumes independence between features. Later generated predictions on the validation set using both the Random Forest and Naive Bayes classifiers. These predictions will serve as inputs for the next step. After that, build a meta learner (the stacking model) that takes the predictions from these base classifiers as inputs and learns to make the final prediction. In this case, another classifier such as Logistic Regression, Decision Tree, or even another

Random Forest can be used as the meta learner. Since the accuracy with random forest was more in [7], trained the random forest based meta learner on the validation set using the predictions from the base classifiers as input features and the actual crop types as the target variable. Once the meta learner is trained, used it to make predictions on new, unseen data (test data).

Evaluated the performance of the final stacking ensemble on a separate test set of both real data and synthetic data. This stacking ensemble is tested on real test set for different performance attributes such as accuracy, precision, recall and F1 Score and obtained score of 99.54%, 99.54%, 99.53%, 99.52% respectively. Then it is tested on synthetic test set for accuracy, precision, recall and F1 Score and obtained score of 99.86%, 99.86%, 99.85%, 99.86% respectively. The results of our experiments demonstrate the efficacy of training the stacking ensemble model on synthetic data. The model achieves an impressive accuracy of 99.86% when evaluated on the synthetic dataset, showcasing its ability to generalize well to unseen data. This significant improvement in accuracy highlights the potential of synthetic data in enhancing the performance of machine learning models, particularly in ensemble techniques like stacking. Moreover, the larger dataset size provided by the synthetic data enables the model to capture a broader range of patterns and relationships, contributing to its superior performance. Table 4 and Table 5 summarizes the performance metrics for real and synthetic data respectively.

Table 4: Performance metrics for real data

Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**
Logistic Regression	95.22%	94.54%	94.53%	93.52%
SVM	96.08%	96.04%	94.53%	93.52%
Decision Tree	99.00%	99.00%	98.53%	98.42%
Naive Bayes	99.04%	99.11%	99.03%	99.03%
Random Forest	99.09%	99.19%	99.09%	99.09%
RNR	**99.54%**	**99.54%**	**99.53%**	**99.52%**

Table 5: Performance metrics for synthetic data

Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**
Logistic Regression	96.77%	96.89%	96.77%	96.78%
SVM	98.31%	98.43%	98.31%	98.31%
Decision Tree	99.54%	99.55%	99.54%	99.54%
Naive Bayes	99.40%	99.41%	99.40%	99.40%
Random Forest	99.59%	99.63%	99.63%	98.62%
RNR	**99.86%**	**99.86%**	**99.85%**	**99.86%**

The accuracy's achieved by the proposed stacking ensemble on real data and synthetic data are represented graphically in figure 3 and figure 4 respectively. In figure 5, a comparison of most recent researches on crop prediction for the real data may be seen. In figure 6, a comparison of different algorithms for the synthetic data is shown. On the crop data set, the proposed work which is RNR (Random forest + Naive Bayes =+ Random Forest) stacking ensemble produced an accuracy of 99.54% on real data set and 99.86% on synthetic data set.

International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 6s,2025
https://www.theaspd.com/ijes.php

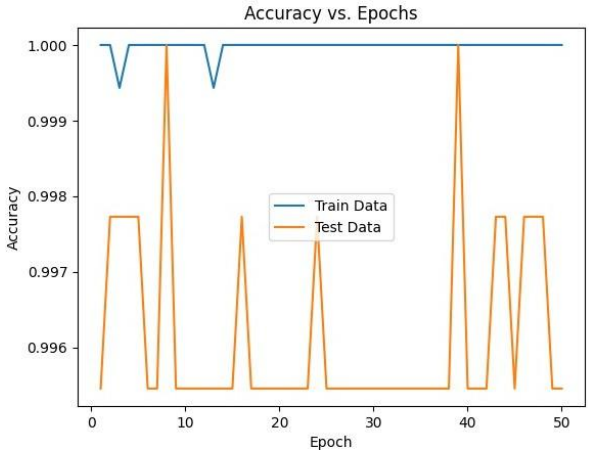

Fig 3: Accuracy of stacking ensemble model on real data set

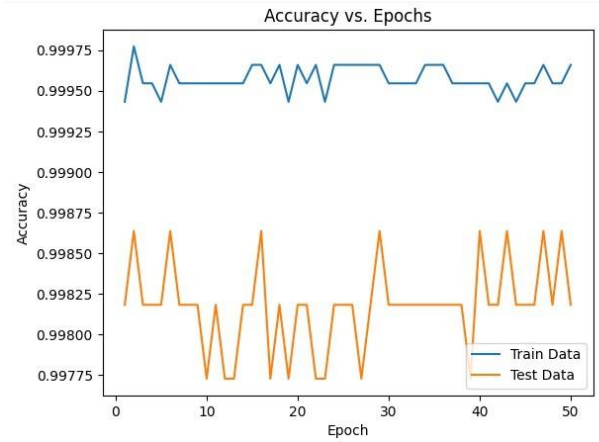

Fig 4: Accuracy of stacking ensemble model on synthetic data set

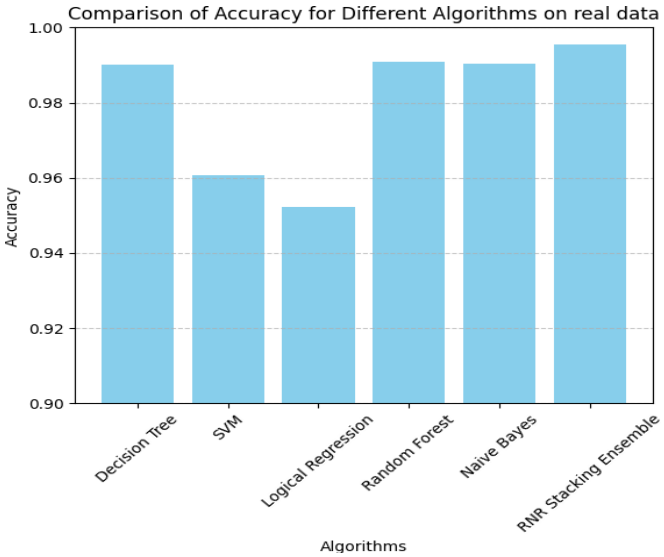

Fig 5: RNR accuracy compared with the other ML models on real data

International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 6s,2025
https://www.theaspd.com/ijes.php

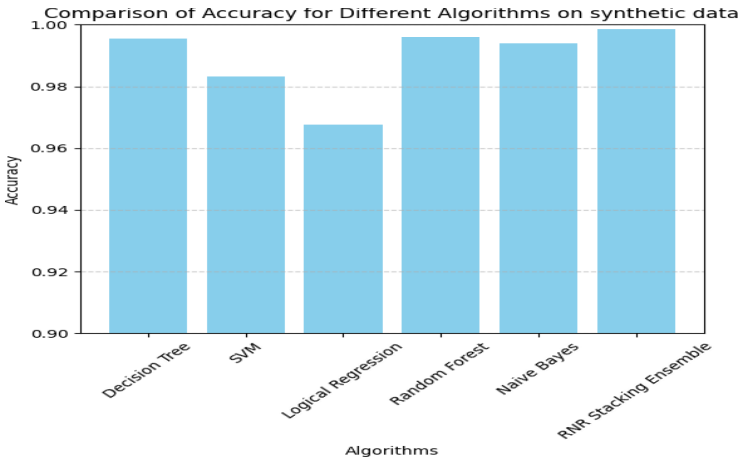

Fig 6: RNR accuracy compared with the other ML models on synthetic data

CONCLUSIONS

This work aims to develop a model that could accurately predict the type of crop that is most likely to be grown in a given area. To achieve this, utilized a dataset of historical crop data along with relevant features such as soil type, climate conditions and soil parameters. This work involves analysis of training and evaluating various machine learning algorithms, including Random Forest and Naive Bayes, using the stacking method. Through thorough evaluation and cross-validation found that, the final stacking ensemble demonstrated superior performance compared to individual classifiers. The model achieved an accuracy of 99.54% on the test set of original data and 99.86% on the test set of derived synthetic data, demonstrating its effectiveness in predicting crop types.During the feature importance analysis, discovered that certain factors played a crucial role in predicting crop types. Soil type, humidity, temperature, and rainfall were identified as the most influential features in this model. Understanding the significance of these features can offer valuable insights for farmers and policymakers in making informed decisions related to crop planning and land management. The successful development of an accurate crop prediction model has numerous practical applications. Agriculture stakeholders, including farmers, land planners, and policymakers, can benefit from using this model to make more informed decisions regarding crop selection, resource allocation, and risk management. By leveraging data-driven predictions, farmers can optimize their agricultural practices, leading to higher yields and more sustainable farming practices. While this model demonstrated strong predictive performance, there is always room for improvement. Future work may involve incorporating additional data sources, such as satellite imagery and advanced climate models, to enhance the accuracy and robustness of the predictions. Additionally, continuous updates to the dataset and retraining the model can help maintain its relevance over time.This study demonstrates the effectiveness of training a stacking ensemble model on synthetic data generated from an original dataset using the gretel.ai data generation tool. The significant improvement in accuracy achieved by the model on the synthetic dataset underscores the potential of synthetic data in enhancing machine learning models' performance. By leveraging synthetic data, researchers and practitioners can augment dataset sizes, improve model robustness, and facilitate more accurate predictions in various predictive analytics tasks. This study contributes to advancing the understanding and application of synthetic data in machine learning and underscores its importance in enhancing model performance and generalization capabilities. In summary, the crop prediction model utilizing the stacking method with Random Forest and Naive Bayes has proven to be a valuable tool for predicting crop types based on soil and environmental factors. This highlights the potential of machine learning and data analytics in revolutionizing agricultural practices, leading to more sustainable and efficient farming systems.

REFERENCES

[1] Bharath Kumar R, Balakrishna K, Bency Celso A, Siddesha M, Sushmitha R, "Crop Recommendation System for Precision Agriculture," International Journal of Computer Sciences and Engineering, Vol.7, Issue.5, pp.1277-1282, 2019.
[2] Dasari Anantha Reddy, Dadore Bhagyashri, Watekar Aarti"Crop Recommendation System to Maximize Crop Yield in Ramtek region using Machine Learning" International Journal of Scientific Research in Science and Technology 485-489. 10.32628/IJSRST196172.
[3] S.Pudumalar, E.Ramanujam, R.Harine Rajashree, C.Kavya, T.Kiruthika, J.Nisha. ``Crop Recommendation System for Precision Agriculture" 8th International Conference on Advanced Computing (ICoAC) IEEE, 2017.
[4] Rikhsit K. Solanki, D Bein, J. A. Vasko, N. Rale. "Prediction of Crop Cultivation" 9th Annual Computing and communication Workshop and Conference (CCWC) IEEE, 2019.
[5] Dr. J. N. Kumar, V. Spandana, V.S. Vaishnavi, ``Supervised Machine learning Approach for Crop Yield Prediction in Agricultural Sector" 5th International Conference on Communication and Electronics Systems (ICCES) June 2020.
[6] N. K. Cauvey, Nidhi H. Kulkarni, Prof. B. Sagar, ``Improving Crop Productivity Through A Crop Recommendation System Using Ensembling Technique" 3rd International Conference on Computational Systems and Information Technology for Sustainable Solutions (CSITSS), pp. 114-119 IEEE, 2018.
[7] Ajay Lokhande, Prof. Manish Dixit, "Crop Recommendation System Using Machine Learning" International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056, p-ISSN: 2395-0072, Volume: 09, Issue: 05, May 2022.
[8] Rajak, R. K., Pawar, A., Pendke, M., Shinde, P., Rathod, S., Devare, A. "Crop recommendation system to maximize crop yield using machine learning technique" International Research Journal of Engineering and Technology pp. 950-953 e-ISSN: 2395-0056, p-ISSN: 2395-0072, Volume: 04 Issue: 12, Dec 2017.
[9] D. Balakrishnan, Anumula Praneeth Kumar, Kristipati Sai Kiran Reddy, R. Ravindra Kumar, K. Aadith, Sudarsi Madhan, "Agricultural Crop Recommendation System", 2023 3rd International Conference on Intelligent Technologies (CONIT), pp.1-5, 2023.
[10] S. Kiruthika, D. Karthika, "IOT-BASED professional crop recommendation system using a weight-based long-term memory approach" Measurement: Sensors, Volume 27,2023,100722,ISSN 2665-9174, https://doi.org/10.1016/j.measen.2023.100722.
[11] Andrew Crane-Droesch "Machine learning methods for crop yield prediction and climate change impact assessment in agriculture" Environ. Res. Lett. 13, 114003, 2018, DOI 10.1088/1748-9326/aae159
[12] Peyman Abbaszadeh, Keyhan Gavahi, Atieh Alipour, Proloy Deb, Hamid Moradkhani, Bayesian Multi-modeling of Deep Neural Nets for Probabilistic Crop Yield Prediction, Agricultural and Forest Meteorology, Volume 314, 2022, 108773, ISSN 0168-1923, https://doi.org/10.1016/j.agrformet.2021.108773.
[13] G. Buvaanyaa, S. Radhimeenakshi, "Crop Recommendation System Using Random Forest Algorithm", 2023 2nd International Journal Of Research Culture Society (IJRCS)", Volume - 7, Issue - 3, March - 2023, DOIs:10.2017/IJRCS/202303014
[14] A. Jhansi Swetha, G. Kalyani, B. Kirananjali, "Advanced Soil Fertility Analysis and Crop Recommendation using Machine Learning", 2023 7th International Conference on Trends in Electronics and Informatics (ICOEI), pp.1035-1039, 2023.
[15] Banavlikar, T., Mahir, A., Budukh, M., & Dhodapkar, S. (2018). Crop recommendation system using Neural Networks. International Research Journal of Engineering and Technology (IRJET), 5(5), 1475-1480.
[16] Madhuri, J., and M. Indiramma. "Artificial neural networks based integrated crop recommendation system using soil and climatic parameters." Indian Journal of Science and Technology 14, no. 19 (2021): 1587-1597.
[17] Kuanr, Madhusree, B. Kesari Rath, and S. Nandan Mohanty. "Crop recommender system for the farmers using mamdani fuzzy inference model." International Journal of Engineering & Technology 7, no. 4.15 (2018): 277-280.
[18] Kamatchi, S. Bangaru, and R. Parvathi. "Improvement of crop production using recommender system by weather forecasts." Procedia Computer Science 165 (2019): 724-732.
[19] Lakshmi, N., M. Priya, Shetty Sahana, and C. R. Manjunath. "Crop recommendation system for precision agriculture." International Journal for Research in Applied Science and Engineering Technology 6, no. 5 (2018): 1132-1136.
[20] Janrao, Surekha, and Deven Shah. "Return on investment framework for profitable crop recommendation system by using optimized multilayer perceptron regressor." IAES International Journal of Artificial Intelligence 11, no. 3 (2022): 969.
[21] Shingade, Sachin Dattatraya, and Rohini Prashant Mudhalwadkar. "Sensor information-based crop recommendation system using machine learning for the fertile regions of Maharashtra." Concurrency and Computation: Practice and Experience (2023): e7774.
[22] Pruthviraj, G. C. Akshatha, K. Aditya Shastry, Nagaraj, and Nikhil. "Crop and fertilizer recommendation system based on soil classification." In Recent Advances in Artificial Intelligence and Data Engineering: Select Proceedings of AIDE 2020, pp. 29-40. Springer Singapore, 2022.
[23] Akshatha, K. R., and K. S. Shreedhara. "Implementation of machine learning algorithms for crop recommendation using precision agriculture." International Journal of Research in Engineering, Science and Management (IJRESM) 1, no. 6 (2018): 58-60.
[24] Bandi, Raswitha, M. Sai Surya Likhit, S. Rajavardhan Reddy, Sathwik Raj Bodla, and Vempati Sai Venkat. "Voting Classifier-based Crop Recommendation." SN Computer Science 4, no. 5 (2023): 516.
[25] Punith Kumar and H. N. Champa, "Crop Disease Identification in Tomato Leaf using Deep Learning," 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT), Delhi, India, 2023, pp. 1-7, doi: 10.1109/ICCCNT56998.2023.10306681.