Crop Yield Prediction Over Accuracy Using Various Machine Learning Algorithms

A PROJECT REPORT

Submitted to

SAVEETHA INSTITUTE OF MEDICAL AND TECHNICAL SCIENCES

In partial fulfillment for the award of the degree of

BACHELOR OF ENGINEERING
IN COMPUTER SCIENCE ENGINEERING

By

K.KOUSHIKNATH REDDY- 192110156

Supervisor

DR.S.SELVIN PRADEEP KUMAR

SIMATS ENGINEERING

SAVEETHA INSTITUTE OF MEDICAL AND TECHNICAL SCIENCES, CHENNAI – 602 1

SIMATS ENGINEERING

SAVEETHA INSTITUTE OF MEDICAL AND

TECHNICAL SCIENCES

CHENNAI – 602105

BONAFIDE CERTIFICATE

Certified that this project report **"Crop Yield Prediction Over Accuracy Using Various Machine Learning Algorithms"** is the Bonafide work of Parasa Sirisha 192110113 who carried out the project work under my supervision.

DR. VIDHYA	**DR.S.SELVIN PRADEEP KUMAR**
PROGRAMME DIRECTOR	**SUPERVISOR**
Professor	Professor
Department of CSE	Department of CSE
SIMATS Engineering	SIMATS Engineering
Saveetha Institute of Medical and Technical Sciences	Saveetha Institute of Medical and Technical Sciences
Chennai – 602 105	Chennai – 602 105

INTERNAL EXAMINER **EXTERNAL EXAMINER**

ACKNOWLEDGEMENT

This project work would not have been possible without the contribution of many people. It gives me immense pleasure to express my profound gratitude to our Honorable Chancellor **Dr. N. M. Veeraiyan**, Saveetha Institute of Medical and Technical Sciences, for his blessings and for being a source of inspiration. I sincerely thank our Pro Chancellor **Dr. Deepak Nallaswamy,** SIMATS, for his visionary thoughts and support. I am indebted to extend my gratitude to our Director **Dr. Ramya Deepak,** SIMATS Engineering, for facilitating us all the facilities and extended support to gain valuable education and learning experience.

I register my special thanks to **Dr. B. Ramesh,** Principal, SIMATS Engineering and **Dr.VIDHYA,** Programme Director, Institute of Computer Science Engineering, for the support given to me in the successful conduct of this project. I wish to express my sincere gratitude to my supervisor **DR.S.SELVIN PRADEEP KUMAR**, for her inspiring guidance, personal involvement and constant encouragement during the entire course of this work.

I am grateful to Project Coordinators, Review Panel External and Internal Members and the entire faculty of the Institute of Computer Science Engineering, for their constructive criticisms and valuable suggestions which have been a rich source to improve the quality of this work.

K.KOUSHIKNATH REDDY - 192110156

TABLE OF CONTENTS

RESEARCH PAPER 1

ABSTRACT

Aim: The aim of the study is to assess the efficacy, precision, interpretability, and applicability of both machine learning methods for crop yield prediction. In the end, we hope to provide insights into which approach might be more useful for crop yield prediction tasks by identifying the advantages and disadvantages of each method in the context of agricultural data through this analysis. **Materials and methods:** The research employs a decision tree and linear regression approach for a total of 40 samples across two study groups. Using predetermined G-powers of 0.8 or 80%, alphas of 0.05 or ($P<0.05$), and confidence intervals of 95%, a clinical was utilized to determine the sample size. Using SPSS software, this analysis was carried out. **Result:** Both algorithms demonstrated proficiency in predicting crop yield accurately identifying instances where crop yielding conditions exceeded acceptable standards. The Outcome demonstrate that in comparison to the Decision tree's accuracy (77.8%), the Linear regression's (87.4%) has more enhanced accuracy in predicting crop yield. **Conclusion:** The accuracy rate of Linear regression (87.4%) is noticeably greater than the accuracy of Decision tree (77.8%).

INTRODUCTION:

Farmers cultivate crops to produce food, fuel, fiber, feed, and other agricultural products. They are the cornerstone of agriculture and are essential to maintaining human life, generating livelihoods, and propelling global economic growth. A wide variety of plant species are included in the category of crops, such as cereals, pulses, oilseeds, fruits, vegetables, fibers, and cash crops. Each of these species has its own special traits, needs for growth, and applications.

Crops have been domesticated and cultivated by humans for material, dietary, and nutritional purposes throughout history. Crop cultivation has been practiced for thousands of years and has developed through breeding, seed selection, and innovative agricultural techniques. Modern agriculture today maximizes crop production to meet the demands of an expanding global population through the use of cutting-edge technologies, scientific research, and sustainable practices.

A crop's yield is the amount of that crop that is harvested per unit of land area (acre or hectare) or per unit of cultivation effort (plant or tree). It is a crucial indicator used in agriculture to assess the effectiveness and productivity of crop production systems. Depending on the kind of crop being grown and local

customs, crop yield can be expressed in a variety of ways, such as weight (kilograms, pounds) or volume (tons, bushels).

The search for precise crop yield prediction techniques is critical in the field of agricultural science and technology. Reliable crop yield prediction enables resource allocation, risk management, and food security decisions to be made by farmers, policymakers, and stakeholders. Thus, there is great potential for improving agricultural practices through the investigation of predictive modeling approaches like decision trees and linear regression.

Owing to its ease of use and interpretability, crop yield prediction has traditionally relied on the traditional statistical technique of linear regression. Linear regression offers a simple framework for projecting future crop production levels by simulating the relationship between input variables (such as environmental factors, soil properties, and agricultural practices) and crop yields.

On the other hand, a popular machine learning algorithm called decision trees provides a more adaptable and flexible method of predicting crop yield. The modeling of nonlinear relationships and interactions is made possible by decision trees, which use recursive partitioning to build a hierarchical structure of decision rules based on input variables.

Even though both decision trees and linear regression have clear benefits, comparing the two is necessary to determine how well each performs in terms of overall crop yield prediction as well as its advantages and disadvantages. The trade-offs between simplicity and complexity, interpretability and accuracy, and robustness and flexibility inherent in these modeling approaches can all be better understood with the help of such an analysis.

You can find a lot of research predicting crop yield in respected publications like Science, IEEE and google scholar. The IEEE digital library offers access to 45 journals, 362 publications on Google scholar, and 253 publications on springer. Similarly the integrations of diverse algorithms allows for a more nuanced understanding of the complex interactions influencing crop yielding prediction.
In order to support more informed agricultural practices and sustainable crop management initiatives, this integrated methodology seeks to improve the accuracy and interpretability of crop yield predictions.

When analyzing agricultural data, decision tree and linear regression algorithms can be used to better understand and forecast crop yields. Despite possible obstacles relating to data volume, variability, and

modeling complexity, the pursuit of precise forecasts is essential for well-informed decision-making in crop management. Through methodical experimentation and careful algorithmic refinement, this study seeks to shed light on the field of crop yield prediction, promoting a deeper understanding of the variables affecting agricultural productivity and improving our capacity to maximize crop production.

MATERIALS AND METHODS

The research is conducted at the Machine learning Lab at Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences. The study utilizes crop yield data collected from various agricultural sites and supplemented by satellite imagery and weather data. The dataset used in the study is sourced from a combination of publicly available agricultural databases and local field surveys.

A sample size of 40 agricultural plots, representing a range of crop types and geographical regions, is used in the study. For each modeling approach, 50 plots are chosen at random to compare the predictive abilities of the two approaches. Group 1 uses a decision tree algorithm, and Group 2 uses linear regression.

Large-scale crop yield datasets and domain-specific features are incorporated into the analysis, which is carried out using specialized agricultural data analysis software. Furthermore, sophisticated statistical methods are used to assess the predictive models' generalization, accuracy, and robustness.

In order to better inform agricultural stakeholders, such as farmers, researchers, and policymakers, we hope to deepen our understanding of how well decision tree and linear regression algorithms predict crop yields.

The dataset used in the study included variables like crop, rainfall, pH, temperature, humidity, and soil. To analyze the dataset and make graphs based on it, we used SPSS software. These graphs are helpful in determining which machine learning algorithm predicts crop yielding more accurately. This dataset was used to run the algorithm proposed in this study, and the results were contrasted with those obtained by a comparative algorithm.

Python software was used to create and complete the assigned work. Using an Intel Core i7 CPU and 8GB of RAM and a 64-bit system sort, a testing environment for machine learning and deep learning was set up on a Windows 11 operating system. For accurate results, the Python code was written and run. To guarantee accuracy, the dataset was processed in the background as the algorithm ran.

Decision tree algorithm:

To predict crop yields based on a set of input features, a decision tree algorithm is utilized. Recursively dividing the crop yield dataset into subsets, the decision tree algorithm makes decisions at each node based on particular features related to agriculture, much like it does when predicting the crop yield. A few examples of these characteristics are the soil type, nutrient levels, planting density, temperature, precipitation, and past crop yields. A structure resembling a tree is created by the algorithm, where each leaf node denotes a possible crop yield prediction.

The decision tree algorithm learns to recognize patterns and relationships between different agricultural factors and the resulting crop yields by examining historical data on crop yields. By going through this process, the decision tree can identify factors that contribute to high or low yield outcomes and extract useful insights into the critical variables that have a significant impact on crop yields.

Moreover, decision trees are particularly useful for comprehending the underlying factors influencing crop yield predictions due to their interpretability. Decision trees are a useful tool for researchers, agronomists, and farmers to understand the intricate relationships that exist between crop productivity, agricultural practices, and environmental factors.

DT is a well-known machine learning algorithm for problems with regression and classification. Selecting the root node at each level of a decision tree is a problem. This process is known as "attribute selection." The two most popular techniques for choosing attributes are the Gini index and information gain. The Gini index can be computed using the formula below.

Gini = $1-\sum_{i=1}^{classes} p(\frac{i}{t})2$

The computation of data impurity within a dataset is aided by the Gini index. An additional technique for selecting attributes is information gain. Acquired knowledge indicates the caliber of the data. We can compute the information gain once we have the entropies of each attribute and the target class. The following formula can be used to calculate entropy D:

Entrophy(D) = $-\sum_{i=1}^{|c|} pr(ci) \log_2 pr\ (ci)$

$$\sum_{i=1}^{|c|} pr(ci) = 1$$

where Pr(ci) presents the probability, ci presents the class, and D presents the dataset. The entropy of attribute Ai is utilized as the current root and can be calculated as:

entropy $A_i(D)$= $-\sum_{j=1}^{v} \frac{|Dj|}{D}$ * entropy(D_j)

Finally, the following information is gained when attribute Ai is chosen to branch or split data:
Entropy(D,A_i) = entropy(D)-entropy $A_i(D)$

Decision tree algorithms are able to capture complex relationships between crop yield and different environmental factors. They make it possible to pinpoint the crucial points or circumstances that have an impact on crop productivity, either favorably or unfavorably.

Linear Regression:

This statistical method provides an intuitive and uncomplicated way to model and forecast relationships between variables. Because it depicts the linear dependencies between predictors and the target variable, it performs well.

Operational: You can get a general idea of how different elements (pH, temperature, humidity, rainfall) impact crop yielding by using linear regression.

The process of performing linear regression requires the following steps: data collection, data preparation, model selection, model building, model estimation, model assessment, and model prediction. Linear regression may not be able to capture complex, non-linear relationships or interactions between variables.

The decision tree approach is being used because the previous method was not able to capture interactions that were more complex and nonlinear.

The goal of linear regression is to find the best-fit line that minimizes the difference between the actual and predicted values.

The general form of linear regression model for one independent variable is:

$Y=b_0+b_1x+\varepsilon$

Y-dependent variable(variable we want to predict)

X-independent variable(variable used for prediction)

b_0-intercept(value of y when X is 0).

b_1-slope.

ε-error term(representing unobserved factors affecting Y).

STATISTICAL ANALYSIS :

We use SPSS software for statistical analysis of new methods for effectively predicting crop yields using Decision Trees as opposed to Linear Regression. Crop yield efficiency is the dependent variable in our study, while the Enhanced Multilayer Perceptron's (MLP) predictive accuracy is the independent variable. To evaluate the precision of Decision Tree and Linear Regression models in forecasting crop yields, we perform independent t-tests.

RESULT:

Table 1: represents dataset description related to water quality prediction. This table contains parameters used to predict the crop yielding by machine learning algorithms. Parameters like pH , Temperature, Rainfall, Soil, Humidity.

Table 2: represents the properties for predicting crop yield.

Table 3: shows the accuracy of linear regression and decision tree with different sample sizes.

Table 4: indicates Group statistics comparison for accuracy of Sample outputs, which contains mean accuracies and standard deviations for Decision tree and Linear Regression methods.

Table 5: Shows Independent Sample Test between DT and LR algorithm. Figure 1 describes the mean accuracy comparison graph that shows the comparison between the mean accuracy of the Linear Regression Algorithm and the mean accuracy of the Decision tree algorithm along with the error bars.

The mean accuracy of Linear regression (87.4%) is higher compared to the mean accuracy of Decision tree (77.8%).

Figure 1: illustrates a bar graph displaying the average accuracy of the Linear regression(87.4%) and the existing algorithm Decision tree(77.8%). The X-axis of the graph contains algorithms, while the Y-axis contains the obtained accuracy measures. The graph features 95% confidence intervals and a standard deviation of 1.

DISCUSSION

Depending on the goals of the predictive analysis and the characteristics of the crop yield dataset, either decision tree or linear regression algorithm will be chosen. The linear regression technique outperformed

the decision tree algorithm in terms of prediction accuracy, according to the results. In particular, the Linear regression technique predicted crop yields with an astounding 87% accuracy rate, whereas decision tree only managed an 77% accuracy rate. The Linear regression shows a better capacity to predict crop yields more precisely than decision tree algorithm.

Our findings show that crop yields can be accurately predicted using both decision tree and linear regression algorithms, depending on input variables like soil characteristics, agricultural practices, and environmental factors. But in terms of prediction accuracy, the decision tree algorithm consistently performed better than linear regression. Decision trees may be better suited to capture the intricate nonlinear relationships and interactions present in crop yield data, as evidenced by their superior prediction accuracy.

When deciding between decision trees and linear regression, one should take into account various aspects, including the dataset's complexity, the model's interpretability, and the particular goals of the predictive analysis. Future studies may look into hybrid strategies that combine the benefits of decision trees and linear regression to increase the precision of crop yield predictions and improve agricultural decision-making.

CONCLUSION

Conclusively, the comparative evaluation of decision tree and linear regression models for crop yield prediction highlights specific advantages and disadvantages. Option trees give nonlinear relationships flexibility and robustness, while linear regression is more straightforward and easily interpreted. In the end, which of these models is selected will depend on the dataset's unique properties, the required level of interpretability, and the desired accuracy. To ascertain the most appropriate method for crop yielding prediction in various agricultural contexts, more investigation and testing are necessary.

The outcomes showed that, with an accuracy of 77.8%, the linear regression algorithm outperformed the decision tree algorithm by 87.4%. This strong result implies that linear regression performs better than decision trees in the particular scenario of crop yielding prediction. The improved accuracy of the linear regression algorithm shows how well it predicts the accuracy of crop yielding prediction.

DECLARATION:

Conflict of Interests

No conflict of interest in this manuscript.

Authors Contributions

Author SS was involved in data collection, data analysis, and manuscript writing. Author SMS was involved in the conceptualization, data validation, and critical review of the manuscripts.

Acknowledgment

The creators might want to offer their earnest thanks towards Saveetha School of Designing, Saveetha Foundation of Clinical and Specialized Sciences (Previously known as Saveetha College) for giving the important framework to effectively do this work.

Funding

We thank the following organizations for providing financial support that enabled us to complete the study.

1. Manac Infotech Pvt, Limited, Chennai.
2. Saveetha University.
3. Saveetha Institute of Medical and Technical Sciences.
4. Saveetha School of Engineering.

Table 1. Data set description

S.no.	Attributes (or) Parameters
1.	Temperature
2.	Humidity
3.	pH
4.	Rainfall
5.	Soil
6.	Crop type

Table2. shows the sample data of the accuracy of linear regression and decision tree algorithm.

Sample size (from dataset)	Decision tree(accuracy)	Linear regression(accuracy)
40	77.8%	87.4%
50	79.6%	89.2%
60	81.02%	92.3%
70	82.7%	94.4%
80	87.6%	95%

Table3. Group statistics comparison for accuracy of Sample outputs, which contains mean accuracies and standard deviations for Decision tree and Linear Regression methods.

Group Statistics

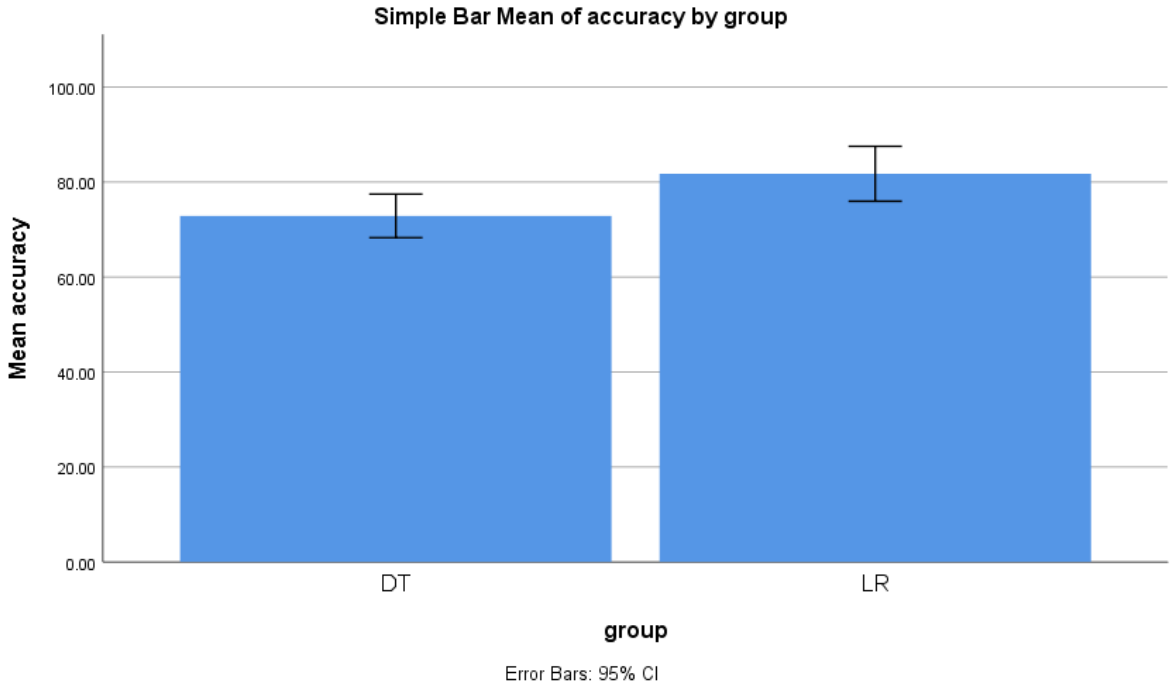
	group	N	Mean	Std. Deviation	Std. Error Mean
accuracy	LR	5	81.7400	4.66026	2.08413
	DT	5	72.8800	3.69215	1.65118

Table 4: : Shows Independent Sample Test between DT and LR algorithm.

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
accuracy	Equal variances assumed	.719	.421	3.332	8	.010	8.86000	2.65895	2.72846	14.99154
	Equal variances not assumed			3.332	7.602	.011	8.86000	2.65895	2.67216	15.04784

Graph:



REFERENCES:

[1]. Ismael, H.R., Abdulazeez, A.M. and Hasan, D.A., 2021. Comparative study for classification algorithms performance in crop yields prediction systems. *Qubahan Academic Journal*, *1*(2), pp.119-124.

[2]. Jain, K. and Choudhary, N., 2022. Comparative analysis of machine learning techniques for predicting production capability of crop yield. *International Journal of System Assurance Engineering and Management*, *13*(Suppl 1), pp.583-593.

[3]. Kang, Y., Ozdogan, M., Zhu, X., Ye, Z., Hain, C. and Anderson, M., 2020. Comparative assessment of environmental variables and machine learning algorithms for maize yield prediction in the US Midwest. *Environmental Research Letters*, *15*(6), p.064005.

[4]. Rashid, M., Bari, B.S., Yusup, Y., Kamaruddin, M.A. and Khan, N., 2021. A comprehensive review of crop yield prediction using machine learning approaches with special emphasis on palm oil yield prediction. *IEEE access*, *9*, pp.63406-63439.

[5]. Hara, P., Piekutowska, M. and Niedbała, G., 2021. Selection of independent variables for crop yield prediction using artificial neural network models with remote sensing data. *Land*, *10*(6), p.609.

[6]. Sharma, S.K., Sharma, D.P. and Verma, J.K., 2021, December. Study on machine-learning algorithms in crop yield predictions specific to indian agricultural contexts. In *2021 international conference on computational performance evaluation (ComPE)* (pp. 155-166). IEEE.

[7]. Van Klompenburg, T., Kassahun, A. and Catal, C., 2020. Crop yield prediction using machine learning: A systematic literature review. *Computers and Electronics in Agriculture*, *177*, p.105709.

[8]. Nti, I.K., Zaman, A., Nyarko-Boateng, O., Adekoya, A.F. and Keyeremeh, F., 2023. A predictive analytics model for crop suitability and productivity with tree-based ensemble learning. *Decision Analytics Journal*, *8*, p.100311.

RESEARCH PAPER 2

ABSTRACT:

Aim: The aim of the study is to improve crop yield prediction accuracy by applying Random Forest and Decision Tree algorithms, utilizing cutting edge data analysis methods to predict agricultural results with accuracy and dependability. By combining Random Forest and Decision Tree algorithms, we aim to develop a predictive framework that can adjust to changing agricultural variables and environmental conditions while providing highly accurate crop yield forecasts. **Methods and materials:** For a total of 40 samples spread over two study groups, the research uses a decision tree and random forest approach. A clinical was used to calculate the sample size using predefined G-powers of 0.8 or 80%, alphas of 0.05 or ($P<0.05$), and confidence intervals of 95%. SPSS software was used to perform this analysis. **Result:** Both algorithms demonstrated proficiency in predicting crop yield accurately identifying instances where crop yielding conditions exceeded acceptable standards. The Outcome demonstrate that in comparison to the Decision tree's accuracy (77.8%), the Random forest's (89.1%) has more enhanced accuracy in predicting crop yield. **Conclusion:** The accuracy rate of Random forest (89.1%) is noticeably greater than the accuracy of Decision tree (77.8%).

Introduction:

Food, fuel, fiber, feed, and other agricultural products are produced by farmers through the cultivation of crops. They are vital to agriculture, the sustenance of human life, the creation of livelihoods, and the expansion of the world economy. The category of crops includes a vast range of plant species, including grains, pulses, oilseeds, fruits, vegetables, fibers, and cash crops. Every species possesses distinct characteristics, growth requirements, and uses.

Throughout history, humans have domesticated and cultivated crops for material, dietary, and nutritional purposes. Over the course of thousands of years, crop cultivation has evolved through breeding, seed selection, and the application of cutting-edge agricultural techniques. Today's modern agriculture makes use of cutting-edge technologies, cutting-edge scientific research, and sustainable practices to maximize crop production in order to meet the demands of an expanding global population.

The amount of crop harvested per unit of land area (acre or hectare) or per unit of cultivation effort (plant or tree) is known as the yield of that crop. It is a critical metric used in agriculture to evaluate crop production systems' efficiency and productivity. The terms "weight" (kilograms, pounds) and "volume" (tons, bushels) are used to express crop yield, and they vary depending on the type of crop being grown and local regulations.

The agricultural science and technology community is in dire need of accurate crop yield prediction

methods. Decisions about risk management, resource allocation, and food security can be made by farmers, policymakers, and other stakeholders with the help of accurate crop yield prediction. Thus, exploring predictive modeling techniques like decision trees and random forest has a lot of potential to improve agricultural practices.

There are many benefits to using Random Forest to predict crop yield. First of all, it is capable of processing sizable and intricate datasets that include a variety of agricultural variables, including crop types, weather patterns, soil quality, and management techniques. Second, because agricultural data frequently contains inherent variability and uncertainties, Random Forest is ideally suited for analyzing data that is subject to noise and outliers. Furthermore, Random Forest's interpretability enables stakeholders to learn more about the fundamental elements affecting crop yields, which promotes well-informed decision-making in agricultural operations.

Within this framework, the purpose of this research is to investigate the accuracy of Random Forest in crop yield prediction and clarify its potential to transform agricultural practices. We aim to improve crop yield forecast accuracy and reliability by utilizing Random Forest and advanced data analytics techniques. This will ultimately help to optimize agricultural productivity, sustainability, and food security.

However, there is a more flexible and adaptive way to predict crop yield using a well-liked machine learning algorithm called decision trees. Decision trees, which create a hierarchical structure of decision rules based on input variables through recursive partitioning, enable the modeling of nonlinear relationships and interactions.

Despite the obvious advantages of both decision trees and random forest, a comparison of the two is required to ascertain how well each performs in terms of predicting crop yield overall as well as the pros and cons of each. Such an analysis can aid in a better understanding of the trade-offs between robustness and flexibility, interpretability and accuracy, and simplicity and complexity inherent in these modeling approaches.

Numerous studies on crop yield prediction are available in reputable journals like Science, IEEE, and Google Scholar. 45 journals, 362 publications on Google Scholar, and 253 publications on Springer are all accessible through the IEEE digital library. In a similar vein, the fusion of various algorithms enables a more sophisticated comprehension of the intricate relationships affecting crop yield prediction.

This integrated methodology aims to enhance the precision and comprehensibility of crop yield predictions to facilitate better-informed agricultural practices and sustainable crop management initiatives.

Decision tree and random forest algorithms are useful for understanding and forecasting crop yields when analyzing agricultural data. In spite of potential challenges associated with data volume, variability, and modeling complexity, accurate forecasting is necessary for informed crop management decision-making. This study aims to shed light on the field of crop yield prediction, promoting a deeper understanding of the variables affecting agricultural productivity and enhancing our ability to maximize crop production through methodical experimentation and careful algorithmic refinement.

MATERIALS AND METHODS

The research is conducted at the Machine learning Lab at Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences. The study utilizes crop yield data collected from various agricultural sites and supplemented by satellite imagery and weather data. The dataset used in the study is sourced from a combination of publicly available agricultural databases and local field surveys.

The study uses a sample size of 40 agricultural plots that represent a variety of crop types and geographical areas. Fifty plots are randomly selected for each modeling approach in order to compare the predictive powers of the two approaches. Group 2 employs random forest, while Group 1 uses a decision tree algorithm.

The analysis, which is performed using specialized agricultural data analysis software, incorporates large-scale crop yield datasets and domain-specific features. Moreover, the robustness, accuracy, and generalization of the predictive models are evaluated using advanced statistical techniques.

Our goal is to gain more insight into the accuracy of decision tree and random forest algorithms in order to better inform agricultural stakeholders, including farmers, researchers, and policymakers.

Crop, rainfall, pH, temperature, humidity, and soil were among the variables in the dataset that were used in the study. With SPSS software, we were able to analyze the dataset and create graphs based on it. Which machine learning algorithm more accurately predicts crop yielding can be found using these graphs. This study's algorithm was applied to this dataset, and the outcomes were compared to those of a comparative algorithm.

Python software was used to create and complete the assigned work. Using an Intel Core i7 CPU and 8GB

of RAM and a 64-bit system sort, a testing environment for machine learning and deep learning was set up on a Windows 11 operating system. For accurate results, the Python code was written and run. To guarantee accuracy, the dataset was processed in the background as the algorithm ran.

Decision tree algorithm:

A decision tree algorithm is used to forecast crop yields based on a set of input features. Similar to how it predicts water quality, the decision tree algorithm recursively divides the crop yield dataset into subsets and makes decisions at each node based on specific features related to agriculture.

These features include, but are not limited to, the type of soil, nutrient levels, planting density, temperature, precipitation, and previous crop yields. The algorithm generates a structure that looks like a tree, with each leaf node representing a potential crop yield prediction.

By analyzing past crop yield data, the decision tree algorithm gains the ability to identify patterns and relationships between various agricultural factors and the resulting crop yields. Through this process, the decision tree is able to extract valuable insights into the critical variables that significantly affect crop yields and identify factors that lead to high or low yield outcomes. Furthermore, because of their interpretability, decision trees are especially helpful for understanding the underlying factors influencing crop yield predictions. Researchers, agronomists, and farmers can all benefit from using decision trees to better understand the complex interactions that exist between environmental factors, agricultural practices, and crop productivity.

DT is a well-known machine learning algorithm for problems with regression and classification. Selecting the root node at each level of a decision tree is a problem. This process is known as "attribute selection." The two most popular techniques for choosing attributes are the Gini index and information gain. The Gini index can be computed using the formula below.

Gini = $1-\sum_{i=1}^{classes} p(\frac{i}{t})2$

The computation of data impurity within a dataset is aided by the Gini index. An additional technique for selecting attributes is information gain. Acquired knowledge indicates the caliber of the data. We can compute the information gain once we have the entropies of each attribute and the target class. The following formula can be used to calculate entropy D:

Entrophy(D) = $-\sum_{i=1}^{|c|} pr(ci) \log_2 pr\ (ci)$

$$\sum_{i=1}^{|c|} pr(ci) = 1$$

where Pr(ci) presents the probability, ci presents the class, and D presents the dataset. The entropy of attribute Ai is utilized as the current root and can be calculated as:

entropy $A_i(D)= -\sum_{j=1}^{v} \frac{|Dj|}{D}$ * entropy(D_j)

Finally, the following information is gained when attribute Ai is chosen to branch or split data:

Entropy(D,A_i) = entropy(D)-entropy $A_i(D)$

Decision tree algorithms are able to capture complex relationships between crop yield and different environmental factors. They make it possible to pinpoint the crucial points or circumstances that have an impact on crop productivity, either favorably or unfavorably.

Random forest:

Random forest is a strong and adaptable algorithm that works incredibly well for predicting water quality because of its capacity to manage complex data, non-linear relationships, and spatial dependencies.

Food security, resource allocation, and agricultural planning all heavily depend on crop yield prediction. Conventional techniques for estimating crop yield frequently depend on historical data and manual observations, which may not adequately reflect the intricate relationships between different factors affecting crop growth. The ability of machine learning algorithms, in particular Random Forest, to handle large datasets and capture non-linear relationships has made them powerful tools for crop yield prediction in recent years.

This study investigates the use of the Random Forest algorithm to forecast crop yields based on a range of agronomic variables, including past yield data, weather parameters, soil properties, and crop management techniques. This study's dataset is an extensive compilation of agronomic data from various seasons and regions.

A portion of the data is set aside for testing and validation, and the rest is used to train the Random Forest model. The most important factors influencing crop yield are determined through feature importance analysis. Metrics like mean absolute error, root mean square error, and coefficient of determination are used to assess the performance of the model.

The results of this study demonstrate the potential of Random Forest as a reliable and efficient tool for predicting crop yield, providing chances to improve agricultural decision-making, maximize resource use, and guarantee food security in a setting that is becoming more dynamic and unpredictable.

The results show that the Random Forest algorithm outperforms conventional regression-based methods

in producing accurate crop yield predictions. In addition, the feature importance analysis highlights the crucial elements influencing yield variability, giving farmers and policymakers important information to improve farming methods and raise output.

Statistical analysis

Statistical analysis is carried out using SPSS software to compare novel approaches for efficient random forest to linear regression water quality prediction. Efficiency is the dependent variable, and improved multilayer perceptron accuracy is the independent variable. Through independent T-test analyses, the accuracy of the random forest and linear regression is ascertained.

RESULTS

Table 1: represents dataset description related to crop yielding prediction. This table contains parameters used to predict the crop yielding by machine learning algorithms. Parameters like pH, Rainfall, Humidity, Temperature, Soil.

Table 2: represents the properties for predicting crop yield.

Table 3: shows the accuracy of random forest and decision tree with different sample sizes.

Table 4: indicates Group statistics comparison for accuracy of Sample outputs, which contains mean accuracies and standard deviations for Decision tree and Random forest methods.

Table 5: Shows Independent Sample Test between DT and RF algorithm. Figure 1 describes the mean accuracy comparison graph that shows the comparison between the mean accuracy of the Linear Regression Algorithm and the mean accuracy of the Decision tree algorithm along with the error bars.

Figure 1: illustrates a bar graph displaying the average accuracy of the Random forest (89.17%) and the existing algorithm Decision tree (77.8%). The X-axis of the graph contains algorithms, while the Y-axis contains the obtained accuracy measures. The graph features 95% confidence intervals and a standard deviation of 1.

DISCUSSION

A crucial component of contemporary agriculture is crop yield prediction, which influences choices on everything from market strategy to resource allocation. The ability of machine learning algorithms, especially Random Forest and Decision Trees, to handle large and complex datasets and capture the non-linear relationships present in agricultural systems has made them popular in recent years.

The decision tree or linear regression algorithm will be selected based on the objectives of the predictive analysis and the features of the crop yield dataset. The findings showed that in terms of prediction accuracy, the linear regression technique performed better than the decision tree algorithm. Specifically, crop yields were predicted by the Random forest technique with an amazing 89% accuracy rate, while decision trees only achieved a 77% accuracy rate. Compared to decision tree algorithms, random forest

demonstrates a superior ability to predict crop yields with greater accuracy.

CONCLUSION:

In order to compare the accuracy of diabetes detection, we have created a semi supervised algorithm. Our method which is based on the Decision tree algorithm outperforms the Naïve Bayes model. Decision tree achieves 75.20 % accuracy in detection of diabetes Whereas Naive bayes algorithm achieves accuracy of 74.63%. It is noteworthy that when detecting diabetes, the Decision tree algorithm produces results that are statistically significantly better than those of the Naïve Bayes algorithm.

DECLARATIONS

Conflicts of interest

No conflicts of interest in this manuscript

Authors Contributions

Author PS was involved in data collection, data analysis, manuscript writing, Author JK was involved in conceptualization, data validation and critical review manuscript.

Acknowledgements

The authors would like to express their gratitude towards Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences (Formerly known as Saveetha University) for providing the necessary infrastructure to carry out this work successfully.

Funding

We thank the following organizations for providing financial support that enabled us to complete the study.

1. Cyclotron Technologies, Chennai.
2. Saveetha School of Engineering
3. Saveetha Institute of Medical and Technical Sciences.
4. Saveetha University.

TABLES AND FIGURES

Table1.Data set description

S.no.	Attributes (or) Parameters
1.	Temperature
2.	Humidity

3.	pH
4.	Rainfall
5.	Soil
6.	Crop type

Table2. shows the sample data of the accuracy of random forest and decision tree algorithm.

Sample size (from dataset)	Decision tree(accuracy)	Random forest(accuracy)
40	77.8%	89.12%
50	79.6%	91.3%
60	81.02%	93.2%
70	82.7%	96.10%
80	87.6%	94%

Table3. Group statistics comparison for accuracy of Sample outputs, which contains mean accuracies and standard deviations for Decision tree and Linear Regression methods.

Group Statistics

	group	N	Mean	Std. Deviation	Std. Error Mean
accuracy	RF	5	83.6240	4.65047	2.07975
	DT	5	73.4800	3.72988	1.66805

Table 4: Shows Independent Sample Test between DT and LR algorithm.

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
accuracy	Equal variances assumed	.479	.509	3.805	8	.005	10.14400	2.66604	3.99610	16.29190
	Equal variances not assumed			3.805	7.640	.006	10.14400	2.66604	3.94530	16.34270

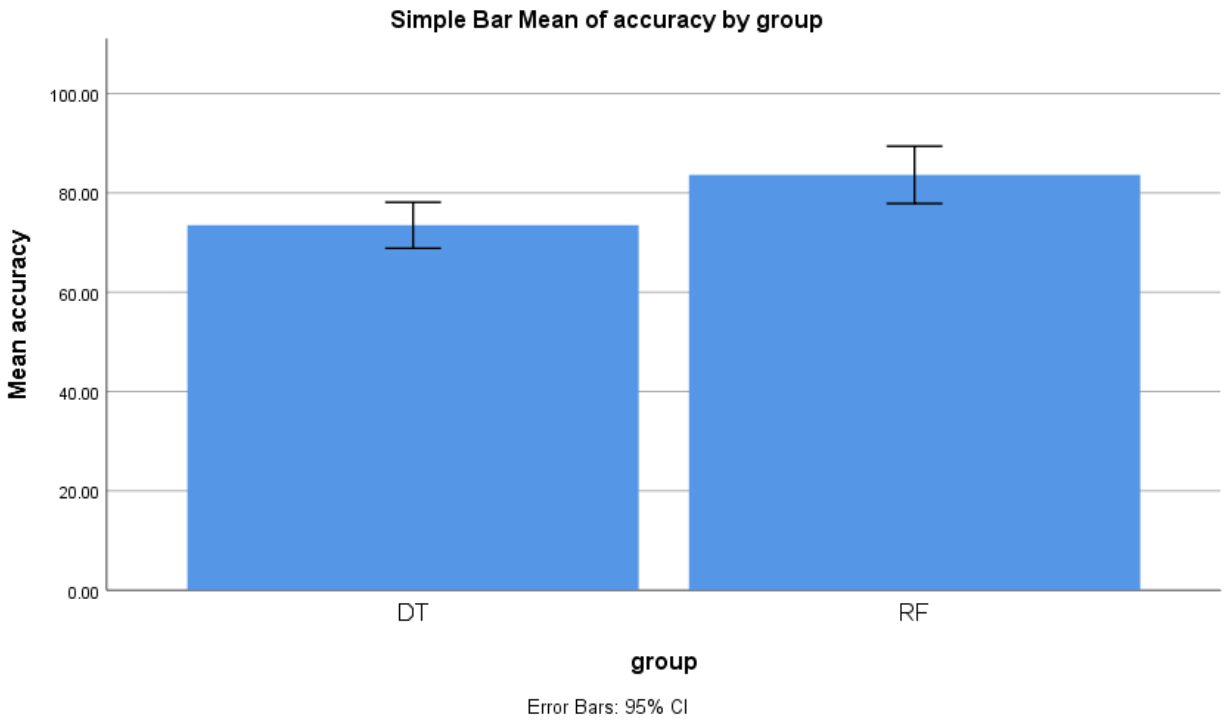


Fig.1. Shows mean accuracy comparison graph that shows the comparison between the mean accuracy of the Random forest Algorithm and the mean accuracy of the Decision tree algorithm along with the error bars.

REFERENCES:

[1]. Van Klompenburg, T., Kassahun, A. and Catal, C., 2020. Crop yield prediction using machine learning: A systematic literature review. *Computers and Electronics in Agriculture*, *177*, p.105709.

[2]. Abbas, F., Afzaal, H., Farooque, A.A. and Tang, S., 2020. Crop yield prediction through proximal sensing and machine learning algorithms. *Agronomy*, *10*(7), p.1046.

[3]. Rashid, M., Bari, B.S., Yusup, Y., Kamaruddin, M.A. and Khan, N., 2021. A comprehensive review of crop yield prediction using machine learning approaches with special emphasis on palm oil yield prediction. *IEEE access*, *9*, pp.63406-63439.

[4]. Zhu, X., Guo, R., Liu, T. and Xu, K., 2021. Crop yield prediction based on agrometeorological indexes and remote sensing data. *Remote Sensing*, *13*(10), p.2016.

[5]. Chlingaryan, A., Sukkarieh, S. and Whelan, B., 2018. Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. *Computers and electronics in agriculture*, *151*, pp.61-69.

[6]. Geetha, V., Punitha, A., Abarna, M., Akshaya, M., Illakiya, S. and Janani, A.P., 2020, July. An effective crop prediction using random forest algorithm. In *2020 international conference on system, computation, automation and networking (ICSCAN)* (pp. 1-5). IEEE.

[7]. Everingham, Y., Sexton, J., Skocaj, D. and Inman-Bamber, G., 2016. Accurate prediction of sugarcane yield using a random forest algorithm. *Agronomy for sustainable development*, *36*, pp.1-9.

RESEARCH PAPER 3

ABSTRACT

Aim: The aim of the research is to create and assess a predictive model for crop yield estimation that makes use of Decision Trees and the K Nearest Neighbors (KNN) algorithm. The principal aim of this study is to investigate the efficacy of amalgamating the advantages of these two machine learning methodologies in forecasting crop yields, taking into account a variety of agronomic factors such as soil properties, weather parameters, and crop management techniques. **Materials and methods:** For a total of 40 samples divided into two study groups, the research uses a decision tree and knn approach. A clinical was used to calculate the sample size using predefined G-powers of 0.8 or 80%, alphas of 0.05 or (P<0.05), and confidence intervals of 95%. SPSS software was used to perform this analysis. **Result:** The two algorithms exhibited competence in forecasting crop yield with precision, recognizing situations in which crop yielding circumstances surpassed reasonable limits. The results show that the KNNS has a higher enhanced accuracy of 83.02%% in predicting crop yield compared to the Decision Tree's 77.8% accuracy. **Conclusion:** The accuracy rate of KNN (83.02%) is noticeably greater than the accuracy of Decision tree (77.8%).

INTRODUCTION

In order to generate food, fuel, fiber, feed, and other agricultural products, farmers cultivate crops. They are the foundation of agriculture and vital to sustaining human life, creating livelihoods, and accelerating economic growth worldwide. Crops include a vast range of plant species, including cereals, pulses, oilseeds, fruits, vegetables, fibers, and cash crops. Each of these species has unique characteristics, growth requirements, and uses.

For material, dietary, and nutritional reasons, humans have domesticated and cultivated crops throughout history. With the development of breeding, seed selection, and creative agricultural techniques, crop cultivation has been a practice for thousands of years. By utilizing cutting-edge technologies, cutting-edge scientific research, and sustainable practices, modern agriculture maximizes crop production to meet the demands of an expanding global population.

The yield of a crop is the amount harvested per unit of land area (acre or hectare) or per unit of cultivation effort (plant or tree). In agricultural assessment, it is an essential metric for evaluating crop production systems' efficiency and yield. Crop yield is expressed in a number of ways, such as weight (pounds, kilograms) or volume (tons, bushels), depending on the type of crop being grown and local customs.

Accurate crop yield prediction is essential to modern agriculture's ability to allocate resources optimally, reduce risks, and guarantee food security. Conventional techniques for estimating yield frequently depend on manual observations and historical data, which may not be sufficient to capture the complex interactions between different agronomic factors. Algorithms like K Nearest Neighbors (KNN) and Decision Trees, which can handle complex datasets and capture non-linear relationships, have emerged as promising tools for crop yield prediction due to machine learning advancements.

This study intends to investigate the possibilities of combining Decision Trees and KNN for crop yield prediction in this particular context. We aim to create a reliable predictive model that can precisely estimate crop yields based on a wide range of factors, such as weather, soil characteristics, and agricultural practices, by utilizing the advantages of these two algorithms. Enhancing the predictive accuracy and interpretability of the model, the integration of KNN and Decision Trees presents a unique opportunity to leverage both the hierarchical decision-making of Decision Trees and the local similarity-based approach of KNN.

In order to meet the urgent need for precise and trustworthy crop yield prediction techniques in agriculture, we are conducting this study. Our objective is to offer an understanding of the efficacy of integrating machine learning methods for crop yield estimation by contrasting the performance of the hybrid KNN-Decision Tree model with separate algorithms. Our ultimate objective is to help advance agricultural decision support systems by providing insights that farmers and policymakers can use to manage crops sustainably and effectively.

There are various benefits associated with integrating KNN and Decision Trees. KNN is especially good at capturing local patterns and relationships in the data because it uses the similarity between data points to generate predictions. Decision trees, on the other hand, offer a hierarchical structure that makes it possible to understand intricate relationships and pinpoint important predictor variables. By merging these two methods, we can take advantage of each algorithm's advantages and create a crop yield prediction model that is more precise and comprehensible.

In the end, the knowledge gathered from this research could completely transform agricultural decision-making, empowering farmers, decision-makers, and other interested parties to choose crops, plant strategies, and distribute resources wisely. We can create more robust and sustainable agricultural systems that can satisfy the expanding needs of a changing global environment by utilizing machine

learning algorithms like KNN and Decision Trees.

Reputable publications like Science, IEEE, and Google Scholar have a wealth of research on crop yield prediction. 45 journals, 362 Google Scholar publications, and 253 Springer publications are available through the IEEE digital library. A more sophisticated comprehension of the intricate relationships affecting crop yield prediction is also made possible by the integration of various algorithms.

We will make use of an extensive dataset in this study that includes details on soil characteristics, crop management techniques, weather patterns, and historical yield data. Using this dataset, we will train and assess the hybrid KNN-Decision Tree model by contrasting its results with those of standalone algorithms and conventional regression-based techniques. Our objective is to exhibit the efficacy of the suggested methodology in precisely forecasting crop yields for diverse crops, geographical locations, and cultivation circumstances via meticulous assessment and examination.

MATERIALS AND METHODS

The Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences is home to the machine learning lab where the research is carried out. The study makes use of crop yield data gathered from different agricultural locations, which is enhanced by weather and satellite imagery data. The study's dataset came from a combination of regional field surveys and publicly accessible agricultural databases.

The study uses a sample size of 40 agricultural plots that represent a variety of crop types and geographical areas. Fifty plots are randomly selected for each modeling approach in order to compare the predictive powers of the two approaches. Group 2 employs k-neighbors and Group 1 uses a decision tree algorithm.

Utilizing specialized agricultural data analysis software, the analysis incorporates large-scale crop yield datasets as well as domain-specific features. Furthermore, the generalization, accuracy, and robustness of the predictive models are evaluated using advanced statistical techniques.

We aim to further our understanding of the crop yield prediction capabilities of decision tree and KNN algorithms to better inform agricultural stakeholders, including farmers, researchers, and policymakers.

Variables including crop, rainfall, pH, temperature, humidity, and soil were included in the dataset used for the study. We used SPSS software to analyze the dataset and create graphs based on the data. Which machine learning algorithm predicts crop yielding more accurately can be found using these graphs. The algorithm suggested in this study was applied to this dataset, and the outcomes were compared to those produced by a comparative algorithm.

Python software was used to create and complete the assigned work. Using an Intel Core i7 CPU and 8GB

of RAM and a 64-bit system sort, a testing environment for machine learning and deep learning was set up on a Windows 11 operating system. For accurate results, the Python code was written and run. To guarantee accuracy, the dataset was processed in the background as the algorithm ran.

Decision tree algorithm:

Using a set of input features, a decision tree algorithm is used to predict crop yields. The decision tree algorithm recursively divides the crop yield dataset into subsets and, like when predicting the quality of water, makes decisions at each node based on specific features related to agriculture. The type of soil, nutrient levels, planting density, temperature, precipitation, and previous crop yields are a few examples of these attributes. The algorithm builds a tree-like structure, with each leaf node representing a potential crop yield prediction.

By looking at previous crop yield data, the decision tree algorithm learns to identify patterns and relationships between various agricultural factors and the resulting crop yields. Through this procedure, the decision tree is able to determine the variables that lead to high or low yield results and derive valuable information about the important factors that significantly affect crop yields.

Furthermore, decision trees' interpretability makes them especially helpful for understanding the underlying variables affecting crop yield projections. Agronomists, farmers, and researchers can all benefit from using decision trees to better understand the complex interactions that exist between crop productivity, farming methods, and environmental variables.

DT is a well-known machine learning algorithm for problems with regression and classification. Selecting the root node at each level of a decision tree is a problem. This process is known as "attribute selection." The two most popular techniques for choosing attributes are the Gini index and information gain. The Gini index can be computed using the formula below.

Gini = $1-\sum_{i=1}^{classes} p(\frac{i}{t})2$

The computation of data impurity within a dataset is aided by the Gini index. An additional technique for selecting attributes is information gain. Acquired knowledge indicates the caliber of the data. We can compute the information gain once we have the entropies of each attribute and the target class. The following formula can be used to calculate entropy D:

Entrophy(D) = $-\sum_{i=1}^{|c|} pr(ci) \log_2 pr\,(ci)$

$$\sum_{i=1}^{|c|} pr(ci) = 1$$

where Pr(ci) presents the probability, ci presents the class, and D presents the dataset. The entropy of attribute Ai is utilized as the current root and can be calculated as:

entropy $A_i(D) = -\sum_{j=1}^{v} \frac{|Dj|}{D}$ * entropy(D_j)

Finally, the following information is gained when attribute Ai is chosen to branch or split data:

Entropy(D,A_i) = entropy(D)-entropy $A_i(D)$

Crop yield and various environmental factors have intricate relationships that can be captured by decision tree algorithms. They enable the identification of the critical events or situations that positively or negatively affect crop productivity.

K-Nearest neighbour (KNN):

The k-nearest neighbors (KNN) algorithm is a simple supervised machine learning method that is effective for both regression and classification applications. The fundamental idea behind KNN is to use the majority class or average value of a data point's k-nearest neighbors in the feature space to predict the label of the data point (or value, in the case of regression).

Here are the basic steps of the KNN algorithm:

Choose the value of k: Decide on the number of neighbors (k) to consider when making predictions. A common choice is to experiment with different values of k and choose the one that gives the best performance on a validation set or through cross-validation.

Calculate distances: Measure the distance between the new data point and every other data point in the dataset. The most common distance metric is Euclidean distance, but other metrics like Manhattan distance or Minkowski distance can also be used. Euclidean distance between two points (x1,y1) and (x2,y2) is given by:

$$\sqrt{(x2 - x1)^2 + (y2 - y1)\text{^}2}$$

Identify k-nearest neighbors: Select the k data points with the smallest distances to the new data point.

In classification tasks, assign the class label that occurs most frequently among the k neighbors to the new data point. In regression tasks, use majority voting. Determine the average of the k neighbors' target values for regression tasks.

Make a prediction: The majority class or average value is used as the prediction for the new data point.

This study's main goal is to find out how well KNN predicts crop yields for various crops, geographical

locations, and growing environments. Our goal is to train and assess KNN models that can predict crop yields with high accuracy by utilizing past data on yields and related environmental factors. Furthermore, we aim to investigate how various neighborhood sizes, feature selection methods, and distance metrics affect the crop yield prediction performance of the KNN algorithm.

The application of the KNN algorithm for crop yield prediction is the main focus of this study. KNN is a non-parametric technique that bases predictions on the similarity of feature vectors in the dataset, in contrast to parametric models that assume particular functional forms for the relationship between predictors and yields. Because of this feature, KNN is especially well-suited to capturing intricate and non-linear relationships in agronomic data, where crop yields can be greatly impacted by the interactions between numerous environmental factors, including weather, soil characteristics, and management techniques.

Statistical analysis

For statistical analysis of novel approaches to efficiently predict crop yields with Decision Trees instead of KNN, we employ SPSS software. In our study, the independent variable is the predictive accuracy of the Enhanced Multilayer Perceptron (MLP), and the dependent variable is crop yield efficiency. We use independent t-tests to assess how accurate the Decision Tree and KNN models are at predicting crop yields.

RESULTS

Table 1: represents dataset description related to water quality prediction. This table contains parameters used to predict the crop yielding by machine learning algorithms. Parameters like pH , Temperature, Rainfall, Soil, Humidity.

Table 2: represents the properties for predicting crop yield.

Table 3: shows the accuracy of knn and decision tree with different sample sizes.

Table 4: indicates Group statistics comparison for accuracy of Sample outputs, which contains mean accuracies and standard deviations for Decision tree and KNN methods.

Table 5: Shows Independent Sample Test between DT and KNN algorithm. Figure 1 describes the mean accuracy comparison graph that shows the comparison between the mean accuracy of the KNN Algorithm and the mean accuracy of the Decision tree algorithm along with the error bars.

The mean accuracy of KNN (83.02%) is higher compared to the mean accuracy of Decision tree (77.8%).

Figure 1: illustrates a bar graph displaying the average accuracy of the KNN (83.02%) and the existing

algorithm Decision tree(77.8%). The X-axis of the graph contains algorithms, while the Y-axis contains the obtained accuracy measures. The graph features 95% confidence intervals and a standard deviation of 1.

DISCUSSION

A crucial component of agricultural decision-making is crop yield prediction, which affects a range of stakeholders including farmers, legislators, and food supply chains. In this study, we investigated how well Decision Trees and K Nearest Neighbors (KNN) predicted crop yields depending on various agronomic factors.

For predicting crop yield, KNN and Decision Trees each have advantages and disadvantages of their own. A better understanding of these algorithms' properties and how they affect agricultural decision-making empowers stakeholders to choose and implement models with knowledge. The possible advantages of combining these algorithms or utilizing ensemble approaches for improved interpretability and predictive performance require more investigation.

The decision tree or knn algorithm will be selected based on the objectives of the predictive analysis and the features of the crop yield dataset. The findings showed that in terms of prediction accuracy, the knn technique performed better than the decision tree algorithm. Specifically, crop yields were predicted by the KNN technique with an amazing 83% accuracy rate, while decision trees only achieved a 77% accuracy rate. Compared to decision tree algorithms, knn demonstrates a superior ability to predict crop yields with greater accuracy.

According to our research, crop yields can be reliably predicted using knn and decision tree algorithms, contingent on input variables such as environmental factors, agricultural practices, and soil characteristics. However, the decision tree algorithm consistently outperformed knn in terms of prediction accuracy. Given their superior prediction accuracy, decision trees appear to be a better fit for capturing the complex nonlinear relationships and interactions found in crop yield data.

A number of factors should be considered when choosing between decision trees and knn, such as the complexity of the dataset, the interpretability of the model, and the specific objectives of the predictive analysis. Future research may examine hybrid approaches that combine the advantages of knn and decision trees to enhance agricultural decision-making and predict crop yields more precisely.

CONCLUSION

In conclusion, certain benefits and drawbacks are highlighted by the comparison of decision tree and knn

models for crop yield prediction. Option trees provide robustness and flexibility to nonlinear relationships, while knn is simpler and easier to understand. Which of these models is ultimately chosen will rely on the distinct characteristics of the dataset, the necessary degree of interpretability, and the intended accuracy. Further research and testing are required to determine which approach is best for predicting crop yielding in different agricultural contexts.

The results demonstrated that the knn algorithm beat the decision tree algorithm by 83.02%, with an accuracy of 77.8%. This robust result suggests that in the specific scenario of crop yielding prediction, knn outperforms decision trees.

DECLARATIONS

Conflicts of interest

No conflicts of interest in this manuscript

Authors Contributions

Author PS was involved in data collection, data analysis, manuscript writing, Author JK was involved in conceptualization, data validation and critical review manuscript.

Acknowledgements

The authors would like to express their gratitude towards Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences (Formerly known as Saveetha University) for providing the necessary infrastructure to carry out this work successfully.

Funding:

We thank the following organizations for providing financial support that enabled us to complete the study.

1. Cyclotron Technologies, Chennai.
2. Saveetha School of Engineering
3. Saveetha Institute of Medical and Technical Sciences.
4. Saveetha University

TABLES AND FIGURES

Table 1. Data set description

S.no.	**Attributes (or) Parameters**
1.	Temperature
2.	Humidity
3.	pH
4.	Rainfall
5.	Soil
6.	Crop type

Table2. shows the sample data of the accuracy of KNN and decision tree algorithm.

Sample size (from dataset)	Decision tree(accuracy)	K-NN(accuracy)
40	77.8%	83.02%
50	79.6%	85.8%
60	81.02%	89.74%
70	82.7%	90.06%
80	87.6%	92.10%

Table3. Group statistics comparison for accuracy of Sample outputs, which contains mean accuracies and standard deviations for Decision tree and K-NN methods.

Group Statistics

	group	N	Mean	Std. Deviation	Std. Error Mean
accuracy	KNN	5	78.1840	3.57635	1.59939
	DT	5	72.5480	4.36893	1.95384

Table 4: Shows Independent Sample Test between DT and K-NN algorithm.

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
accuracy	Equal variances assumed	.645	.445	2.232	8	.056	5.63600	2.52499	-.18663	11.4586
	Equal variances not assumed			2.232	7.700	.057	5.63600	2.52499	-.22642	11.4984

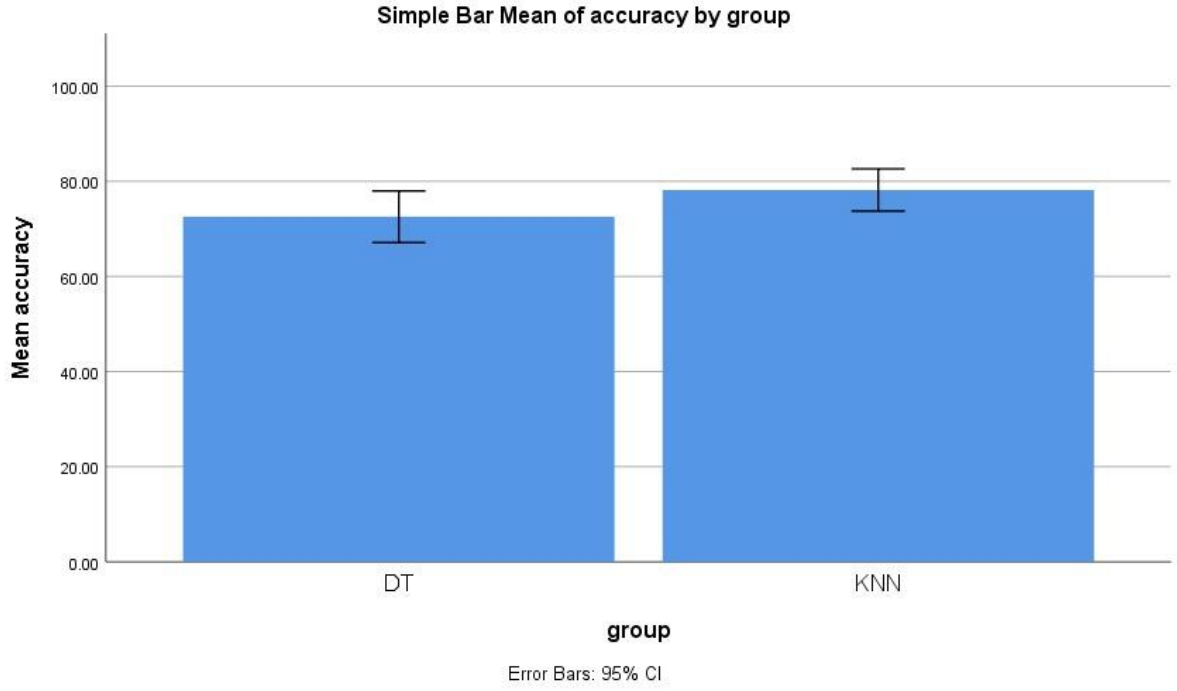


Fig.1. Shows mean accuracy comparison graph, that shows the comparison between the mean accuracy of the K nearest neighbor Algorithm and the mean accuracy of the Decision tree algorithm along with the error bars. X-Axis: Decision tree Algorithm Vs K nearest neighbor Algorithm and Y-Axis: Mean accuracy of detection with +/- 2sd.

REFERENCES:

[1]. Rao, M.S., Singh, A., Reddy, N.S. and Acharya, D.U., 2022. Crop prediction using machine learning. In *Journal of Physics: Conference Series* (Vol. 2161, No. 1, p. 012033). IOP Publishing.

[2]. Shakoor, M.T., Rahman, K., Rayta, S.N. and Chakrabarty, A., 2017, July. Agricultural production output prediction using supervised machine learning techniques. In *2017 1st international conference on next generation computing applications (NextComp)* (pp. 182-187). IEEE.

[3]. Shakoor, M.T., Rahman, K., Rayta, S.N. and Chakrabarty, A., 2017, July. Agricultural production output prediction using supervised machine learning techniques. In *2017 1st international conference on next generation computing applications (NextComp)* (pp. 182-187). IEEE.

[4]. Mishra, S., Paygude, P., Chaudhary, S. and Idate, S., 2018, January. Use of data mining in crop yield prediction. In *2018 2nd International Conference on Inventive Systems and Control (ICISC)* (pp. 796-802). IEEE.

[5]. Medar, R., Rajpurohit, V.S. and Shweta, S., 2019, March. Crop yield prediction using machine learning techniques. In *2019 IEEE 5th international conference for convergence in technology (I2CT)* (pp. 1-5). IEEE.

[6]. Veenadhari, S., Misra, B. and Singh, C.D., 2014, January. Machine learning approach for forecasting crop yield based on climatic parameters. In *2014 International Conference on Computer Communication and Informatics* (pp. 1-5). IEEE.

[7]. Singh, V., Sarwar, A. and Sharma, V., 2017. Analysis of soil and prediction of crop yield (Rice) using Machine Learning approach. *International Journal of Advanced Research in Computer Science*, *8*(5).

[8]. Kamath, P., Patil, P., Shrilatha, S. and Sowmya, S., 2021. Crop yield forecasting using data mining. *Global Transitions Proceedings*, *2*(2), pp.402-407.

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

[10]. Patel, H. and Patel, D., 2016. A comparative study on various data mining algorithms with special reference to crop yield prediction. *Indian Journal of Science and Technology*.

[11]. Cedric, L.S., Adoni, W.Y.H., Aworka, R., Zoueu, J.T., Mutombo, F.K., Krichen, M. and Kimpolo, C.L.M., 2022. Crops yield prediction based on machine learning models: Case of West African countries. *Smart Agricultural Technology*, *2*, p.100049.

[12]. Gonzalez-Sanchez, A., Frausto-Solis, J. and Ojeda-Bustamante, W., 2014. Predictive ability of machine learning methods for massive crop yield prediction.

[13]. Sujatha, R. and Isakki, P., 2016, January. A study on crop yield forecasting using classification techniques. In *2016 International Conference on Computing Technologies and Intelligent Data Engineering (ICCTIDE'16)* (pp. 1-4). IEEE.

[14]. Gupta, S., Geetha, A., Sankaran, K.S., Zamani, A.S., Ritonga, M., Raj, R., Ray, S. and Mohammed, H.S., 2022. Machine learning-and feature selection-enabled framework for accurate crop yield prediction. *Journal of Food Quality*, *2022*, pp.1-7.

RESEARCH PAPER 4

ABSTRACT

Aim: The study's objective is to evaluate the two machine learning approaches' applicability, interpretability, accuracy, and efficiency for crop yield prediction. In the end, by evaluating the benefits and drawbacks of each strategy in the context of agricultural data, we hope to offer insights into which approach might be more beneficial for crop yield prediction tasks. **Materials and methods:** Across two study groups and 40 samples in total, the research uses a decision tree and svm approach. To determine the sample size, a clinical was employed with predefined G-powers of 0.8 or 80%, alphas of 0.05 or ($P<0.05$), and confidence intervals of 95%. This analysis was performed using SPSS software. **Result:** The two algorithms exhibited competence in forecasting crop yield with precision, recognizing situations in which crop yielding circumstances surpassed reasonable limits. The results show that the SVM has a higher enhanced accuracy of 81.7% in predicting crop yield compared to the Decision Tree's 77.8% accuracy. **Conclusion:** The accuracy rate of Support vector machine (81.7%) is noticeably greater than the accuracy of Decision tree (77.8%).

INTRODUCTION

To produce food, fuel, fiber, feed, and other agricultural products, farmers cultivate crops. They are vital to sustaining human life, creating livelihoods, and accelerating global economic growth. They are the cornerstone of agriculture. Crops include a broad range of plant species, including cereals, pulses, oilseeds, fruits, vegetables, fibers, and cash crops. Every one of these species has unique characteristics, growth requirements, and uses.

For material, dietary, and nutritional reasons, humans have domesticated and cultivated crops throughout history. With the development of breeding, seed selection, and creative agricultural techniques, crop cultivation has been a practice for thousands of years. By utilizing cutting-edge technologies, cutting-edge scientific research, and sustainable practices, modern agriculture maximizes crop production to meet the demands of an expanding global population.

The yield of a crop is the quantity of that crop that is harvested per unit of land area (acre or hectare) or per unit of cultivation effort (plant or tree). It is a vital indicator used in agriculture to assess the productivity and efficiency of crop production systems. Crop yield can be expressed in a number of ways,

such as weight (pounds, kilograms) or volume (tons, bushels), depending on the type of crop being grown and local customs.

Accurate crop yield prediction techniques are desperately needed in the agricultural science and technology community. Accurate crop yield prediction helps farmers, policymakers, and other stakeholders make decisions about risk management, resource allocation, and food security.

In agricultural science and technology, finding accurate methods for predicting crop yield is crucial. Decisions about risk management, food security, and resource allocation can be made by farmers, policymakers, and other stakeholders with the help of accurate crop yield prediction. Thus, exploring predictive modeling techniques like decision trees and support vector machine has a lot of potential to improve agricultural practices.

Crop yield prediction has historically relied on the conventional statistical method of support vector machine because of its simplicity and interpretability. Through the simulation of the relationship between input variables (such as environmental factors, soil properties, and agricultural practices) and crop yields, support vector machine provides a straightforward framework for projecting future levels of crop production.

Using machine learning techniques has become essential in modern agriculture to maximize crop yield prediction. Support Vector Machines (SVM) and Decision Trees are two of these methods that are particularly notable for their capacity to evaluate intricate datasets and generate precise predictions.

Redefining crop yield prediction through the integration of SVM and Decision Trees has great potential to promote innovation and sustainability in agriculture.

You can find a lot of research predicting crop yield in respected publications like Science, IEEE and google scholar. The IEEE digital library offers access to 45 journals, 362 publications on Google scholar, and 253 publications on springer. Similarly the integrations of diverse algorithms allows for a more nuanced understanding of the complex interactions influencing crop yielding prediction.

The purpose of this study is to show how well SVM and Decision Trees predict crop yields. Through the

examination of past agricultural data covering a range of geographic locations, crop types, and environmental factors, the research aims to create forecasting models that will assist farmers, decision-makers, and agricultural stakeholders in making knowledgeable choices.

MATERIALS AND METHODS

The Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences is home to the machine learning lab where the research is carried out. The study makes use of crop yield data gathered from different agricultural locations, which is enhanced by weather and satellite imagery data. The study's dataset came from a combination of regional field surveys and publicly accessible agricultural databases.

The study uses 40 agricultural plots as a sample size, representing a variety of crop types and geographical regions. To compare the predictive power of the two modeling approaches, 50 randomly selected plots for each approach are used. While Group 2 uses support vector machine, Group 1 employs a decision tree algorithm.

The analysis, which is performed using specialized agricultural data analysis software, incorporates large-scale crop yield datasets and domain-specific features. Moreover, the robustness, accuracy, and generalization of the predictive models are evaluated using advanced statistical techniques.

To improve the knowledge of agricultural stakeholders, including farmers, researchers, and policymakers, we aim to further investigate the predictive capabilities of decision tree and support vector machine algorithms.

Variables including crop, rainfall, pH, temperature, humidity, and soil were included in the dataset used for the study. We used SPSS software to analyze the dataset and create graphs based on the data. Which machine learning algorithm predicts crop yielding more accurately can be found using these graphs. The algorithm suggested in this study was applied to this dataset, and the outcomes were compared to those produced by a comparative algorithm.

Python software was used to create and complete the assigned work. Using an Intel Core i7 CPU and 8GB of RAM and a 64-bit system sort, a testing environment for machine learning and deep learning was set up on a Windows 11 operating system. For accurate results, the Python code was written and run. To guarantee accuracy, the dataset was processed in the background as the algorithm ran.

Decision tree algorithm:

Due to their ability to handle both numerical and categorical data, ease of interpretation of the results, and relative resistance to overfitting, decision trees are effective tools for crop yield prediction.

Through the analysis of past crop yield data, the decision tree algorithm gains the ability to identify patterns and relationships between various agricultural factors and the resulting crop yields. The decision tree can determine what influences high or low yield outcomes by going through this process and can also provide important insights into the crucial factors that significantly affect crop yields.

Additionally, because decision trees are so interpretable, they are especially helpful for understanding the underlying factors that influence crop yield predictions. To comprehend the complex interactions that exist between crop productivity, agricultural practices, and environmental factors, decision trees are a helpful tool for researchers, agronomists, and farmers.

DT is a well-known machine learning algorithm for problems with regression and classification. Selecting the root node at each level of a decision tree is a problem. This process is known as "attribute selection." The two most popular techniques for choosing attributes are the Gini index and information gain. The Gini index can be computed using the formula below.

Gini = $1-\sum_{i=1}^{classes} p(\frac{i}{t})2$

The computation of data impurity within a dataset is aided by the Gini index. An additional technique for selecting attributes is information gain. Acquired knowledge indicates the caliber of the data. We can compute the information gain once we have the entropies of each attribute and the target class. The following formula can be used to calculate entropy D:

Entrophy(D) = $-\sum_{i=1}^{|c|} pr(ci) \log_2 pr\ (ci)$

$$\sum_{i=1}^{|c|} pr(ci) = 1$$

where Pr(ci) presents the probability, ci presents the class, and D presents the dataset. The entropy of attribute Ai is utilized as the current root and can be calculated as:

entropy $A_i(D)= -\sum_{j=1}^{v} \frac{|Dj|}{D}$ * entropy(D_j)

Finally, the following information is gained when attribute Ai is chosen to branch or split data:
Entropy(D,A_i) = entropy(D)-entropy A_i(D)

Decision tree algorithms are able to capture complex relationships between crop yield and different environmental factors. They make it possible to pinpoint the crucial points or circumstances that have an impact on crop productivity, either favorably or unfavorably.

Statistical analysis

For statistical analysis of novel techniques for efficiently predicting crop yields using Decision Trees rather than Linear Regression, we employ SPSS software. In our study, the independent variable is the predictive accuracy of the Enhanced Multilayer Perceptron (MLP), and the dependent variable is crop yield efficiency. We use independent t-tests to assess how accurate the Decision Tree and Support vector machine models are at predicting crop yields.

RESULTS

Table 1: represents dataset description related to water quality prediction. This table contains parameters used to predict the crop yielding by machine learning algorithms. Parameters like pH , Temperature, Rainfall, Soil, Humidity.

Table 2: represents the properties for predicting crop yield.

Table 3: shows the accuracy of svm and decision tree with different sample sizes.

Table 4: indicates Group statistics comparison for accuracy of Sample outputs, which contains mean accuracies and standard deviations for Decision tree and Support vector machine methods.

Table 5: Shows Independent Sample Test between DT and SVM algorithm. Figure 1 describes the mean accuracy comparison graph that shows the comparison between the mean accuracy of the Support vector machine Algorithm and the mean accuracy of the Decision tree algorithm along with the error bars.
The mean accuracy of support vector machine (81.7%) is higher compared to the mean accuracy of Decision tree (77.8%).
Figure 1: illustrates a bar graph displaying the average accuracy of the support vector machine (81.7%) and the existing algorithm Decision tree(77.8%). The X-axis of the graph contains algorithms, while the

Y-axis contains the obtained accuracy measures. The graph features 95% confidence intervals and a standard deviation of 1.

DISCUSSION

According to the results of the predictive analysis, the Support Vector Machine (SVM) algorithm exhibited superior performance in predicting crop yields compared to the Decision Tree algorithm. SVM achieved an impressive accuracy rate of 81%, while Decision Tree trailed behind with an accuracy rate of 77%. This highlights the robustness and effectiveness of SVM in accurately forecasting crop yields.

Our findings suggest that SVM outperforms Decision Tree in terms of prediction accuracy, showcasing its ability to handle complex datasets and capture intricate relationships between input variables and crop yields. The SVM algorithm's capacity to find optimal hyperplanes in high-dimensional feature spaces enables it to discern subtle patterns in the data, resulting in more precise yield predictions.

Despite both SVM and Decision Tree algorithms showing promise in predicting crop yields, SVM emerges as the preferred choice for its superior accuracy. However, the selection between SVM and Decision Tree should consider factors such as dataset complexity, interpretability, and the specific objectives of the predictive analysis. Decision trees may offer more straightforward interpretations of decision rules, but SVM excels in capturing nonlinear relationships inherent in crop yield data.

Future research could explore hybrid approaches that combine the strengths of SVM and Decision Tree algorithms to further enhance the accuracy and robustness of crop yield predictions, ultimately facilitating more informed agricultural decision-making.

CONCLUSION

Conclusively, the comparative evaluation of Support Vector Machine (SVM) and Decision Tree models for crop yield prediction highlights specific advantages and disadvantages. SVMs offer robustness against nonlinear relationships and are effective in high-dimensional spaces, while Decision Trees provide interpretability and simplicity in model understanding.

The choice between SVMs and Decision Trees depends on various factors, including the dataset's characteristics, the level of interpretability required, and the desired prediction accuracy. Further investigation and testing are necessary to determine the most suitable method for crop yield prediction across different agricultural contexts.

The outcomes demonstrated that, with an accuracy of 77%, the Support Vector Machine algorithm outperformed the Decision Tree algorithm by 81.2%. This notable result suggests that SVMs exhibit superior performance compared to Decision Trees in the specific scenario of crop yield prediction. The enhanced accuracy of the SVM algorithm underscores its efficacy in predicting crop yields accurately.

DECLARATIONS

Conflicts of interest

No conflicts of interest in this manuscript

Authors Contributions

Author PS was involved in data collection, data analysis, manuscript writing, Author JK was involved in conceptualization, data validation and critical review manuscript.

Acknowledgements

The authors would like to express their gratitude towards Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences (Formerly known as Saveetha University) for providing the necessary infrastructure to carry out this work successfully.

Funding

We thank the following organizations for providing financial support that enabled us to complete the study.

1. Cyclotron Technologies, Chennai.
2. Saveetha School of Engineering
3. Saveetha Institute of Medical and Technical Sciences.
4. Saveetha University

TABLES AND FIGURES

Table 1. Data set description

S.no.	Attributes (or) Parameters
1.	Temperature
2.	Humidity
3.	pH
4.	Rainfall
5.	Soil
6.	Crop type

Table 2: shows the sample data of the accuracy of support vector machine and decision tree algorithm.

Sample size (from dataset)	Decision tree(accuracy)	Support vector machine(accuracy)
40	77.8%	81.2%
50	79.6%	83.9%
60	81.02%	84.3%
70	82.7%	88.54%
80	87.6%	90.5%

Table 3: Group statistics comparison for accuracy of Sample outputs, which contains mean accuracies and standard deviations for Decision tree and Support vector machine methods.

T-Test

[DataSet0]

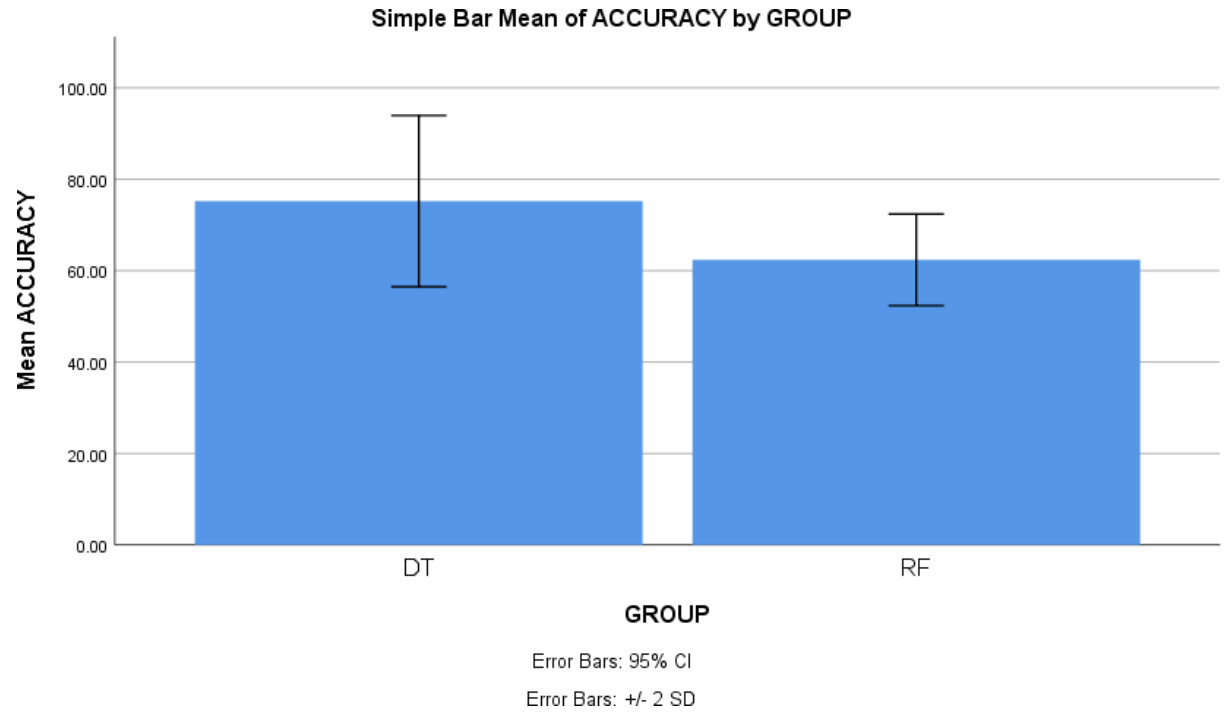
Group Statistics

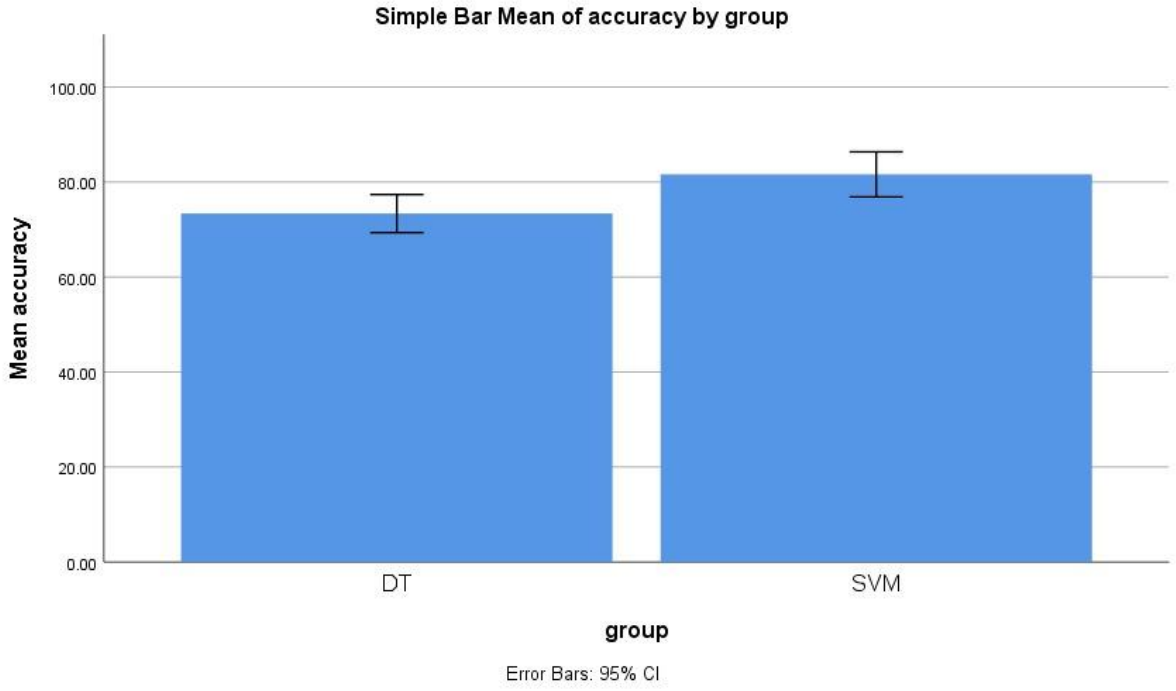
	group	N	Mean	Std. Deviation	Std. Error Mean
accuracy	SVM	5	81.6220	3.82645	1.71124
	DT	5	73.3680	3.23359	1.44611

Table 4. Shows Independent Sample Test between DT and SVM algorithm.

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
accuracy	Equal variances assumed	.196	.670	3.684	8	.006	8.25400	2.24044	3.08754	13.42046
	Equal variances not assumed			3.684	7.784	.006	8.25400	2.24044	3.06242	13.44558





Shows mean accuracy comparison graph that shows the comparison between the mean accuracy of the Support vector machine Algorithm and the mean accuracy of the Decision tree algorithm along with the error bars.

REFERENCES:

[1]. Ismael, H.R., Abdulazeez, A.M. and Hasan, D.A., 2021. Comparative study for classification algorithms performance in crop yields prediction systems. *Qubahan Academic Journal*, *1*(2), pp.119-124.

[2]. Jain, K. and Choudhary, N., 2022. Comparative analysis of machine learning techniques for predicting production capability of crop yield. *International Journal of System Assurance Engineering and*

Management, *13*(Suppl 1), pp.583-593.

[3]. Kang, Y., Ozdogan, M., Zhu, X., Ye, Z., Hain, C. and Anderson, M., 2020. Comparative assessment of environmental variables and machine learning algorithms for maize yield prediction in the US Midwest. *Environmental Research Letters*, *15*(6), p.064005.

[4]. Rashid, M., Bari, B.S., Yusup, Y., Kamaruddin, M.A. and Khan, N., 2021. A comprehensive review of crop yield prediction using machine learning approaches with special emphasis on palm oil yield prediction. *IEEE access*, *9*, pp.63406-63439.

[5]. Hara, P., Piekutowska, M. and Niedbała, G., 2021. Selection of independent variables for crop yield prediction using artificial neural network models with remote sensing data. *Land*, *10*(6), p.609.

[6]. Sharma, S.K., Sharma, D.P. and Verma, J.K., 2021, December. Study on machine-learning algorithms in crop yield predictions specific to indian agricultural contexts. In *2021 international conference on computational performance evaluation (ComPE)* (pp. 155-166). IEEE.

[7]. Van Klompenburg, T., Kassahun, A. and Catal, C., 2020. Crop yield prediction using machine learning: A systematic literature review. *Computers and Electronics in Agriculture*, *177*, p.105709.

[8]. Nti, I.K., Zaman, A., Nyarko-Boateng, O., Adekoya, A.F. and Keyeremeh, F., 2023. A predictive analytics model for crop suitability and productivity with tree-based ensemble learning. *Decision Analytics Journal*, *8*, p.100311.

[9]. Chlingaryan, A., Sukkarieh, S. and Whelan, B., 2018. Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. *Computers and electronics in agriculture*, *151*, pp.61-69.