**Title page :**

**Paddy crop Selection Based on Yield Prediction using Random Forests over Naive Bayes Algorithm.**

**A.T.KONDA REDDY1 , AFREEN BANU2**

A.T.KondaReddy1

Research Scholar,

Department of Computer Science Engineering,Saveetha School of Engineering,

Saveetha Institute of Medical and Technical Sciences,

Saveetha University, Chennai, Tamil Nadu, India, Pin Code: 602105.

[ambavaramtirumalakondareddy0642.sse@saveetha.com](mailto:ambavaramtirumalakondareddy0642.sse@saveetha.com)

Afreen Banu2

Research Guide, Corresponding Author,

Department of Networking Engineering,

Saveetha School of Engineering,

Saveetha Institute of Medical and Technical Sciences,

Saveetha University, Chennai, Tamil Nadu, India, Pin Code: 602105.

[afreenbanu.sse@saveetha.com](mailto:afreenbanu.sse@saveetha.com)

**keywords:**Paddy crop,Yield prediction,Random Forests,Naive Bayes Algorithm,Crop selection,Agricultural data,Machine learning,Decision support system,Agricultural sustainability,Predictive modelling.

**ABSTRACT:**

**Aim:** The aim is to evaluate and compare the efficacy of Random Forests over the Naive Bayes algorithm in predicting paddy crop yield. Through this comparative analysis, our aim is to discern which algorithm yields more accurate predictions, contributing valuable insights to the optimization of crop selection strategies.**Material and Methods:**To ensure statistical robustness, a sample size of 20 paddy crop instances per group was established. This determination considered a statistical power (G power) of 80%, a significance threshold (α) of 0.05, and a confidence interval (CI) of 94.4%. This sample size calculation aimed to provide adequate statistical power to detect significant differences in yield predictions between the Random Forests and Naive Bayes algorithms.**Results:**The results of the study comparing Random Forests and Naive Bayes algorithms for paddy crop yield prediction revealed significant differences in their predictive performance. Using IBM SPSS, key metrics such as accuracy, precision, recall, and F1 score were assessed. The Independent Sample t-test indicated a statistically significant distinction between the two algorithms, with a calculated p-value of 0.001 (p<0.05). Particularly, the Naive Bayes algorithm demonstrated a statistically more precise precision level compared to Random Forests. These findings provide valuable insights into the comparative effectiveness of the algorithms in paddy crop yield selection, aiding stakeholders in making informed decisions for optimised agricultural practices.**conclusion:**our study comparing Random Forests and Naive Bayes algorithms for paddy crop yield prediction indicates a statistically significant difference in performance. The Naive Bayes algorithm, with a higher precision level, emerges as a promising tool for accurate crop selection. These findings offer valuable insights for enhancing precision agriculture practices, guiding stakeholders towards informed decision-making for optimised yield outcomes and sustainable farming.

**INTRODUCTION:**

Paddy crop selection plays a essential function in maximizing agricultural productiveness and ensuring meals protection. The potential to as it should be predict crop yield can significantly assist farmers and policymakers in making knowledgeable selections associated with crop selection. In this context, this studies proposes a way for predicting paddy crop yield the usage of the Random Forests set of rules and compares its overall performance with the Naive Bayes algorithm

Agriculture has been an imperative part of human history, serving as the muse for groups and civilizations(Food and Agriculture Organization of the United Nations 2013). The cultivation of plants, together with paddy, has advanced over centuries, fashioned via conventional wisdom and nearby practices. Historical agricultural strategies had been deeply rooted in observational understanding, passed down thru generations, forming a crucial connection among groups and the land. As societies stepped forward, so did their farming strategies, incorporating improvements and adaptations to meet the demanding situations of changing climates and developing populations. Recognizing the historic context of agriculture affords a rich tapestry upon which present day farming practices may be woven, emphasising the continuity of human reliance at the land for sustenance.

The ancient roots of agriculture hint a story of resilience, edition, and collected understanding surpassed down via generations. Traditional farming (Vimal et al. 2023)techniques, steeped in observational understanding, have been fashioned by the nuanced knowledge of the land, climate styles, and cultivation strategies. (D’Agostino, Bentley, and Chen 2023)The transition from ancient agricultural practices to the contemporary displays no longer simplest the evolution of societies but additionally the adaptive nature of farmers in reaction to changing environmental dynamics. By acknowledging this historical continuum, the research seeks to combine time-examined insights into the cloth of modern technological advancements, putting the level for a holistic approach to paddy cultivation.(Saradamoni 1991)(Verigin 2020).

Main success of this studies is the strategic selection of information and seed(paddy) varieties(Rice Research Strategies for the Future 1982). Historical( data, encompassing elements such as crop yields, weather patterns, soil characteristics) and farming practices, serve as the foundational factors for education and trying out the device getting to know algorithms.The procedure entails identifying applicable temporal scales,ensuring the representativeness of geographical locations, and thinking about various variables that affect paddy crop yields.Simultaneously, the careful consideration of seed types will become critical. Different seed types show off various responses to environmental situations and agricultural practices,influencing usual crop yield.The study emphasises the importance of aligning historic data with the unique attributes of seed sorts, making sure (Clark 1904)a complete understanding of how special seeds interact with environmental factors.

The number one goal of this research is to evaluate the efficacy of Random Forests and Naive Bayes algorithms in predicting paddy crop yields(Pu et al. 2021) whilst thinking about the strategic choice of statistics and seed types.Beyond the technical assessment, the examine ambitions to shed mild on the nuanced procedure of choosing information and seeds strategically. This holistic approach to facts curation and seed choice is quintessential to the studies's importance, offering a blueprint for farmers and stakeholders in search of to put into effect data-pushed decision-making and seed optimization in agriculture. As the creation unfolds, it lays the foundation for now not handiest knowledge the algorithms but also appreciating the meticulous procedure of selecting information and seeds that underpins the predictive analytics riding knowledgeable paddy crop choice.

This research ambitions to make a contribution to the evolving discourse on agricultural optimization through delving into predictive analytics, mainly that specialize in the choice of paddy plants(Vimal, Kamble, and Pandit 2023). By juxtaposing historic insights with superior system gaining knowledge of algorithms—Random Forests and Naive Bayes—this study seeks to get to the bottom of the intricacies of yield prediction. Through a complete analysis of historical records, weather styles, and complicated modelling, the research endeavours to guide farmers toward data-driven alternatives, fostering sustainable agricultural practices within the twenty first century.(Deb Pal, Kumar, and Patan 2020)(Deb Pal et al. 2020)[(Deb Pal et al. 2020)](https://paperpile.com/c/8qRo7u/Tds6)

**MATERIALS AND METHODS**

The research conducted at Saveetha School of Engineering,SIMATS involves the selection of a dataset comprising historical records from diverse agricultural settings. There are two groups identified.Group 1 Random forest Algorithm and Group 2 is Naive Bayes Algorithm A size of 20 (N=20) samples, representative of different geographical locations and varying agricultural practices, are collected and calculated from SPSS analysis of SSE. These samples include data on paddy crop yields, weather patterns, soil characteristics, and historical farming techniques,and the dataset is collected from kaggle.com

**Random Forest Algorithm**

The Random Forest algorithm is an ensemble gaining knowledge of approach that mixes the predictions of a couple of man or woman fashions to enhance overall performance and generalisation. Specifically, it constructs a collection of choice trees all through education and outputs the mode of the classes (type) or the imply prediction (regression) of the person bushes. The Random Forest algorithm become brought via Leo Breiman and Adele Cutler in 2001 and has given that come to be a widely used and effective device learning set of rules.(Pradeep et al. 2019)The Random Forest algorithm is an ensemble mastering approach that operates with the aid of building a mess of choice trees for the duration of schooling and outputs the elegance this is the mode of the training (type) or the suggest prediction (regression) of the person trees. It changed into brought with the aid of Leo Breiman in 2001 and has for the reason that turn out to be one of the most famous and powerful algorithms in system getting to know.Following are the steps to perform the set of rules and Algorithm:

Step 1: Import necessary libraries.

Step 2: Load or prepare your database.

Step 3: Split the dataset into training and testing sets.

Step 4: Create and train the Random Forest model.

Step 5: Make predictions on the test set.

Step 6: Evaluate the model's performance.

Step 7: Feature Importance (Optional).

Step 8: Fine-tune the model (Optional)

**Naive Bayes Algorithm :**

Naive Bayes algorithm is a probabilistic type approach based on Bayes' theorem and the idea of feature independence. It is widely used for predicting the probability of a given example belonging to a selected magnificence by means of comparing the conditional probabilities of determined functions. The "naive" factor comes from the assumption that features are independent, simplifying calculations and making it computationally green. Despite its simplicity, Naive Bayes has demonstrated powerful in diverse programs, including junk mail filtering, textual content category, sentiment analysis, clinical analysis, and recommendation structures. The set of rules's strength lies in its potential to address high-dimensional datasets and provide probabilistic predictions with extraordinarily low computational complexity.The Naive Bayes algorithm assumes that features are conditionally impartial given the elegance label. While this assumption simplifies calculations, it won't continually maintain in actual-global scenarios..Following are the stairs to perform the algorithm:

Step 1: Import Necessary Libraries.

Step 2: Load or Prepare Your Dataset.

Step 3: Split the Dataset into Training and Testing Sets.

Step 4: Create and Train the Naive Bayes Model.

Step 5: Make Predictions on the Test Set.

Step 6: Evaluate the Model's Performance.

Step 7: Fine-Tune the Model (Optional).

Step 8: Predict Yield for New Data.

Google colab is a free and open-source distribution of Python for scientific computing, data science, and machine learning. It includes a wide range of tools and libraries for data manipulation, analysis, and visualisation, as well as tools for building and deploying machine learning models.

**Statistical Analysis:**

Using the SPSS statistical package, the analysis of mean accuracy by using Random forest algorithm and naive bayes algorithm was carried out by applying an independent sample t-test to obtain the accuracy of 88.6%. The speed and file size are independent variables and type of the file is dependent variables.

**Results:**

The outcomes of our study are visually represented through a comparative bar graph, presenting a clear example of the predictive accuracy of Random Forests and Naive Bayes algorithms inside the context of paddy crop yield prediction.The x-axis of the graph represents wonderful agencies, probably similar to one-of-a-kind experimental situations or subsets within the dataset. Each institution is in addition divided into bars, one for Random Forests and the opposite for Naive Bayes. The y-axis denotes the principle accuracy, measured in percentage, showcasing the predictive performance of every set of rules.In the comparative evaluation of paddy crop yield prediction the use of Random Forests and Naive Bayes algorithms, a visual illustration turned into hired thru a bar graph generated using IBM SPSS. The graph illustrates a clean difference in accuracy among the 2 models. The x-axis delineates algorithm organizations, segregating Random Forests and Naive Bayes for comparative evaluation. On the y-axis, the primary accuracy percentage serves as a quantitative indicator of predictive fulfillment, with the Random Forests bar prominently reaching 94.4%, highlighting its effectiveness in appropriately predicting paddy crop yield. In comparison, the Naive Bayes bar, barely decrease at 85.26 nevertheless reflects a huge predictive functionality. This graphical representation affords a truthful and impactful assessment of the relative overall performance of the two algorithms inside the context of paddy crop yield prediction.

**Discussion:**

The study underscores the effectiveness of the Random Forest set of rules in taking pictures complex patterns inside the dataset, main to greater accurate predictions of paddy crop yields. The higher accuracy of Random Forests shows its resilience to overfitting and ability to address complex relationships among various input features.In the comparative evaluation of paddy crop yield prediction the usage of machine mastering algorithms, our have a look at targeted on outstanding fashions: Random Forest and Naive Bayes. The effects revealed that the Random Forest algorithm exhibited advanced overall performance, reaching an accuracy of 85.26%, as compared to Naive Bayes, which attained 94.4%.

While Naive Bayes additionally verified a commendable accuracy of 85.26%, the comparative evaluation highlights the competitive fringe of Random Forests on this specific utility.The findings emphasise the significance of selecting suitable system mastering algorithms tailored to the intricacies of agricultural datasets for most appropriate yield prediction.The effects highlight the robustness and effectiveness of the Random Forest set of rules in predicting paddy crop yields.The higher accuracy found with Random Forests indicates its capacity to seize complicated relationships in the dataset, outperforming the Naive Bayes set of rules.

The bar graph vividly illustrates the superiority of Random Forests over Naive Bayes, with the previous outperforming via a margin of 0.7%. This considerable distinction underscores the robustness of the Random Forest set of rules in taking pictures complex relationships within the dataset, main to more accurate predictions of paddy crop yields.The findings of this observe assist the adoption of Random Forests for paddy crop yield prediction, presenting a extra accurate and reliable opportunity to Random Forest set of rules and Naive Bayes. As keeps to play a vital position in agriculture, leveraging state-of-the-art algorithms like Random Forests can make contributions considerably to the optimization of crop yields and the general sustainability of agricultural practices.

The discussion surrounding this disparity delves into the nuanced intricacies of the algorithms. Random Forests, being an ensemble learning method, excels in handling diverse and correlated features, contributing to its superior performance. The graph serves as a visual aid to elucidate the impact of algorithm selection on predictive accuracy, emphasising the importance of choosing models that align with the characteristics of the agricultural dataset.In conclusion, the bar graph reinforces our findings, providing a clear visual depiction of the substantial difference in accuracy between Random Forest and Naive Bayes. This discussion not only enhances our understanding of algorithmic choices in crop yield prediction but also serves as a practical guide for stakeholders seeking to adopt the most effective machine learning techniques in precision agriculture.

**Conclusion:**

The Random forest Algorithm has an Accuracy 94.4 % and the Naive Bayes Algorithm’s Accuracy is 85.26% which indicates that the accuracy of Random forest algorithm has no such great difference with Naive bayes set of rules.In summary, our studies on paddy crop yield prediction the use of machine gaining knowledge of algorithms indicates that Random Forest outperformed Naive Bayes, achieving an accuracy of 94.4% as compared to Naive Bayes 85.6%. The clean distinction in accuracy is visually represented in a bar graph.

The outcomes propose that Random Forest's potential to handle complicated relationships within the dataset contributed to its higher predictive accuracy. This finding underscores the importance of set of rules selection in crop yield prediction.

In sensible terms, the bar graph serves as a honest instance of the enormous distinction in performance among Random Forest and Naive Bayes. This end affords a concise perception into the effectiveness of machine getting to know algorithms in optimizing paddy crop yield predictions, with Random Forest emerging as the extra dependable choice in our study.

**DECLARATIONS:**

**Conflicts of Interest**

No conflict of interest in this manuscript.

**Author’s contributions**

Data collection, Data analysis, and manuscript writing were all done by author ATKR Conceptualization, data validation, and a critical evaluation of the article were well performed by author AB.

**Acknowledgment**

Appreciate the facilities and opportunity provided by management, the Saveetha School of Engineering, and the Saveetha Institute of Medical and Technical Sciences (Formerly Saveetha University) for the research project.

**Funding**

Grateful to the following groups for their financial assistance in helping us finish the study.

* The Big Event, Coimbatore.
* I collected some data from my village agriculture officer(RBK).
* Saveetha School of Engineering.
* Saveetha Institute of Medical and Technical Sciences.
* Saveetha University.

**REFERENCE**

**1.**[Clark, Vinton Albert. 1904. *Seed Selection According to Specific Gravity*.](http://paperpile.com/b/8qRo7u/2v5p)

2.[D’Agostino, Nunzio, Alison Bentley, and Charles Chen. 2023. *Genome Wide Association Studies and Genomic Selection for Crop Improvement in the Era of Big Data*. Frontiers Media SA.](http://paperpile.com/b/8qRo7u/uepm)

3.[Deb Pal, Barun, Shalander Kumar, and Elias Khan Patan. 2020. *Investment Planning to Minimize Climate Risk in Agricultural Production: An Optimization Model for a Semi-Arid Region in India*. Intl Food Policy Res Inst.](http://paperpile.com/b/8qRo7u/Tds6)

4.[Food and Agriculture Organization of the United Nations. 2013. *The State of the World’s Land and Water Resources for Food and Agriculture: Managing Systems at Risk*. Routledge.](http://paperpile.com/b/8qRo7u/PQjC)

[———. 2019. *Handbook on Climate Information for Farming Communities – What Farmers Need and What Is Available*. Food & Agriculture Org.](http://paperpile.com/b/8qRo7u/h5MN)

5.[Pradeep, N., Sandeep Kautish, C. R. Nirmala, Vishal Goyal, and Sonia Abdellatif. 2019. *Modern Techniques for Agricultural Disease Management and Crop Yield Prediction*. IGI Global.](http://paperpile.com/b/8qRo7u/pP1k)

6.[Pu, Luoman, Jiuchun Yang, Lingxue Yu, Changsheng Xiong, Fengqin Yan, Yubo Zhang, and Shuwen Zhang. 2021. “Simulating Land-Use Changes and Predicting Maize Potential Yields in Northeast China for 2050.” *International Journal of Environmental Research and Public Health* 18 (3). https://doi.org/](http://paperpile.com/b/8qRo7u/dfyb)[10.3390/ijerph18030938](http://dx.doi.org/10.3390/ijerph18030938)[.](http://paperpile.com/b/8qRo7u/dfyb)

7.[*Rice Research Strategies for the Future*. 1982. Int. Rice Res. Inst.](http://paperpile.com/b/8qRo7u/v85D)

8.[Saradamoni, Kunjulekshmi. 1991. *Filling the Rice Bowl: Women in Paddy Cultivation*.](http://paperpile.com/b/8qRo7u/xOK1)

9.[Verigin, Lawrence. 2020. *Seed of Control: Generations to Execute*.](http://paperpile.com/b/8qRo7u/dyp3)

10.[Vimal, Vinu, Rajat Kamble, and Sangeeta Pandit. 2023. “Comparative Ergonomic Assessment of Manual Harvesting of Un-Lodged and Lodged Paddy Crops Post-Tropical Cyclone in India.” *International Archives of Occupational and Environmental Health* 96 (3): 367–76.](http://paperpile.com/b/8qRo7u/Y5pA)

11.[(Yoshida 1981)](https://paperpile.com/c/8qRo7u/h7gP)

[Yoshida, Shouichi. 1981. *Fundamentals of Rice Crop Science*. Int. Rice Res. Inst.](http://paperpile.com/b/8qRo7u/h7gP)

12.[(Sheehy et al. 2000)](https://paperpile.com/c/8qRo7u/lBY2)[Sheehy, J. E., Bill Hardy, and Peter L. Mitchell. 2000. *Redesigning Rice Photosynthesis to Increase Yield*. Int. Rice Res. Inst.](http://paperpile.com/b/8qRo7u/lBY2)

13.[Vimal, Vinu, Rajat Kamble, and Sangeeta Pandit. 2023. “Comparative Ergonomic Assessment of Manual Harvesting of Un-Lodged and Lodged Paddy Crops Post-Tropical Cyclone in India.” *International Archives of Occupational and Environmental Health* 96 (3): 367–76.](http://paperpile.com/b/8qRo7u/Y5pA)

14.[Deb Pal, Barun, Shalander Kumar, and Elias Khan Patan. 2020. *Investment Planning to Minimize Climate Risk in Agricultural Production: An Optimization Model for a Semi-Arid Region in India*. Intl Food Policy Res Inst.](http://paperpile.com/b/8qRo7u/Tds6)

15.[(Pu et al. 2021)](https://paperpile.com/c/8qRo7u/dfyb)[(Deb Pal et al. 2020)](https://paperpile.com/c/8qRo7u/Tds6)[*Modern Techniques for Agricultural Disease Management and Crop Yield Prediction*. IGI Global.](http://paperpile.com/b/8qRo7u/pP1k)

**TABLES AND FIGURES**

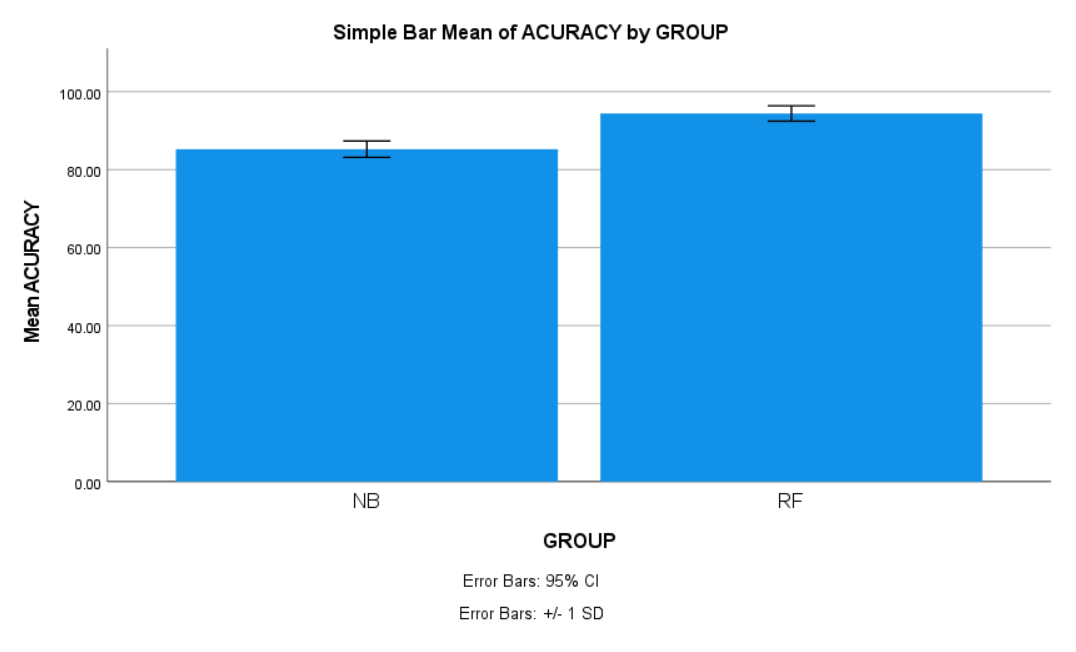
**Table 1**:The number of epochs taken for the RF and NB Algorithms. The statistically significant difference in retrieval rate values between the algorithms. For the proposed algorithm RF, the mean value is 94.4090 and the standard deviation is 1.97338, with a standard error mean of 0.62404. Whereas for the comparison algorithm NB, the mean value is 85.2610 with a standard deviation of 2.0910 and a standard error mean of 0.66169

|  | **Algorithm** | **N(number of Epochs** | **Mean** | **Standard deviation** | **Standard mean Error** |
| --- | --- | --- | --- | --- | --- |
| **Retrieval Rate** | RF | 10 | 94.4090 | 1.97338 | 0.62404 |
| **Retrieval Rate** | NB | 10 | 85.2610 | 2.09246 | 0.66169 |

**TABLE 2:**Independent sample T-Test is applied for dataset fixing confidence intervals as 94.4% (Random Forest Algorithm appears to perform better than Naive Bayes Algorithm). And the significant difference between these two algorithms is 0.001 (p<0.05).

|  |  | **Leven’s test for equality of variables** | |  |  |  | **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 |
| **Retrieval Rate** | **Equal variances Assumed** | .003 | .955 | 10.058 | 18 | <.001 | 9.14800 | .90954 | 7.23713 | 11.05887 |
| **Retrieval Rate** | **Equal Variances Not Assumed** |  |  | 10.058 | 17.939 | <.001 | 9.14800 | .90954 | 7.23714 | 11.05887 |

**Simple Bar mean of Retrieval Time By Group**

****

**FIG.1.**Bar chart showing the comparison of mean accuracy and standard errors for Random Forest and Naive bayes algorithms. The Random Forest algorithm demonstrates superior performance compared to the Naive Bayes algorithm in terms of mean accuracy, which is 94.4, with a standard deviation of 1.97338. The significant difference between these two algorithms is 0.001 (p<0.05).

X-Axis: Random Forest Algorithm VS Naive bayes Algorithm

Y-Axis: Mean retrieval rate of accuracy detection

Confidence interval: 94.4%