Journal of Basics and Applied Sciences Research (JOBASR)
ISSN (print): 3026-9091, ISSN (online): 1597-9962
Volume 3(3) May 2025
DOI: https://dx.doi.org/10.4314/jobasr.v3i3.30

Development of an Optimized Bernoulli Naïve Bayes Classifier for Threshold-Based False Onset Rainfall Prediction

Saifullahi Suleiman[1], Shamsudden Suleiman[2*], Jamilu A. Bashir[3] and Yusuf Bello[4]
[1 & 3]Department of Computer Science, Federal University Dutsin-Ma, Katsina State, Nigeria.
[2 & 4]Department of Statistics, Faculty of Physical Sciences, Federal University Dutsin-Ma, Katsina State, Nigeria.
*Corresponding Author Email: ssuleiman@fudutsinma.edu.ng

ABSTRACT
This study presents the development of an optimized Bernoulli Naive Bayes classifier for predicting threshold-based false onset rainfall, a phenomenon critical to farming. The methodology followed in this research includes data collection, preprocessing, feature selection and threshold analysis, model development, model optimization, and evaluation.The primary focus of this research was the optimization of this model to improve its performance. Leave-one-out cross-validation was employed to systematically validate the model by training it on all but one instance and testing it on the excluded instance, ensuring robust performance evaluation. Grid search was used for hyper parameter tuning to identify the optimal parameters that maximize model accuracy. Alpha smoothing was applied to handle zero probabilities, ensuring the model's generalization to unseen data. The model was evaluated using key performance metrics, such as accuracy, precision, recall, and F1 score. Experimental results revealed that the optimized model achieved significant improvements in predictive accuracy and reliability over baseline implementations. This optimization framework highlights the model's computational efficiency and its suitability for real-time applications. The findings establish the potential of the optimized model as a powerful tool for addressing challenges associated with false onset rainfall prediction. Unlike deterministic models, this research emphasized probabilistic reasoning, introducing a novel approach to rainfall prediction.

Keywords:
Bernoulli Naïve Bayes,
Binarization,
Thresholding,
Classifier,
Optimization.

INTRODUCTION

False onset rainfall, characterized by an initial period of precipitation followed by prolonged dryness, poses significant challenges to farmers (Adeyeri *et al.*, 2020; Odekunle *et al.*, 2019). This phenomenon disrupts agricultural planning, leading to crop losses and economic setbacks (Ajayi *et al.*, 2021). Predicting false onset rainfall accurately is crucial for mitigating these adverse impacts and ensuring sustainable agricultural practices (Omotosho & Abiodun, 2021). Machine learning models have shown promise in addressing the complexities associated with rainfall (Oswal, 2019; Sandeep & Jahavi, 2020). However, their effectiveness depends heavily on proper optimization to handle high-dimensional, noisy datasets and improve predictive reliability (Liyew & Melese, 2021; Rahman *et al.*, 2022). Optimization ensures that these models achieve robust performance, even under the variability inherent in meteorological data (Ojo & Ogunjo, 2022).

The Bernoulli Naive Bayes classifier is particularly well-suited for this task due to its strength in binary classification problems. In his research on the optimality of Naïve Bayes, Zhang (2004) stated that the model is computationally efficient and can produce reliable results even with small datasets, provided the features carry meaningful information about the target class. The compatibility of the Bernoulli Naïve Bayes model with threshold-based approaches allows for effective modeling of false onset rainfall by leveraging key features such as gross rainfall and evaporation rate (Kundu & Ahmed, 2020). This study highlights the importance of combining the classifier's simplicity and computational efficiency with rigorous optimization techniques to address the unique challenges of predicting false onset rainfall. By enhancing the accuracy and reliability of these predictions, the study aims to contribute to better decision-making frameworks for agricultural management in vulnerable regions (Rahman *et al.*, 2022; Ojo & Ogunjo, 2022).

Oswal (2019) conducted rainfall prediction using various machine learning models of different families, such as linear classifiers, tree-based, distance-based,

rule-based, and ensemble, by analysing historical weather data from major Australian cities. The specific algorithms used include Logistic Regression, Random Forests, Gradient Boosting, and Neural Networks. Gradient Boosting outperformed the other algorithms in terms of accuracy. Similarly, Sandeep and Jahavi (2020) employed Artificial Neural Networks, Random Forests, Naïve Bayes, and Logistic Regression to predict rainfall in the Indian region. These algorithms were compared to determine the most accurate and precise model.

Liyew and Melese (2021) developed a rainfall prediction system to forecast daily rainfall amounts using machine learning techniques. They utilised Pearson correlation for feature selection and compared Multivariate Linear Regression, Random Forest, and Extreme Gradient Boosting. The models used input variables moderately and strongly related to rainfall, with performance measured using root mean square error and mean absolute error. Ojo and Ogunjo (2022) developed machine learning models for rainfall prediction over Nigeria. The study compared two multivariate polynomial regression (MPR) models with twelve machine learning algorithms, including three Artificial Neural Networks, four Adaptive Neuro-Fuzzy Inference Systems (ANFIS), and five Support Vector Machines with different kernel functions. The performance of these models was evaluated using a general performance index.

Rahman *et al.*, (2022) developed a rainfall prediction system using a fusion of machine learning techniques to enhance accuracy and reliability. They implemented a hybrid approach by combining decision trees, support vector machines, and random forests, leveraging ensemble methods to integrate predictions from different models.

MATERIALS AND METHODS

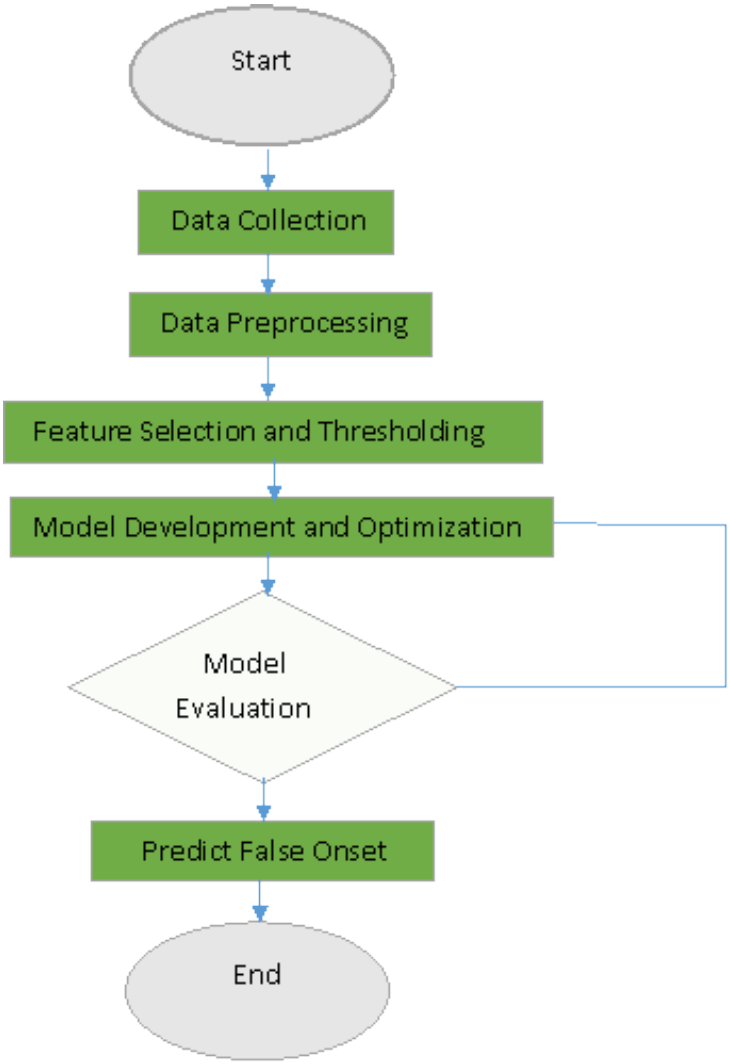

Figure 1. Flowchart of the Optimized Bernoulli Naïve Bayes Classifier for False Onset Rainfall Prediction

Data Collection

The meteorological dataset utilized in this study was obtained from the Nigerian Meteorological Agency (NIMET, 2024**)** and focused specifically on Katsina Metropolis**,** during the critical onset of the 2024 rainy season (May–June). The dataset includes daily observations of gross rainfall (mm)**,** evaporation rate (mm)**,** and post-rainfall humidity (%)—all of which are key parameters influencing the determination of false onset rainfall. These variables were measured consistently over a period of 55 days, offering a granular and comprehensive view of atmospheric conditions. The dataset formed the foundation for feature selection**,** threshold calibration**,** and model development in the proposed Bernoulli Naïve Bayes framework. The structured format and completeness of the data made it ideal for use in threshold-based classification algorithms targeting false onset detection.

A sample of the dataset is shown in Table 1, representing a typical segment of the daily records:

Table 1: Sample of Meteorological Data for Katsina Metropolis (May–June 2024)

Day	Gross Rainfall (mm)	Evaporation Rate (mm)	Rainfall Onset (1=Yes, 0=No)
1	0.0	10.6	0
2	0.0	9.7	0
3	0.0	9.4	0
4	0.0	8.7	0
5	0.0	13.7	0
6	0.0	13.6	0
7	0.0	12.2	0
..	...	...	...
54	10.5	16.0	1
55	0.0	4.0	0

Source: Nigerian Meteorological Agency (NIMET), 2024.

Preprocessing

Step 1: Handling Missing Data: No missing values are found here, so no action is needed.

Step 2: Encoding Categorical Data: In this dataset, all features are numerical, so no encoding is required.

Step 3: Normalization/Scaling: Although the Bernoulli Naïve Bayes classifier operates optimally on binary features (0 and 1), normalization is generally a good preprocessing step for other classifiers or prior to threshold-based binarization. When applied, the **min-max normalization** formula is used to scale numeric features into a fixed range, typically [0,1], as shown below:

$$X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (1)$$

where:

X is the original value,

X_{min} and X_{max} are minimum and maximum values of the feature in the dataset, respectively,

X_{scaled} is the normalized value.

Feature Engineering and Selection

In this study, feature selection was carried out using a filter method based on domain knowledge and threshold-based rules. This approach involves selecting variables that are known, from meteorological literature and expert insights, to be most indicative of false onset rainfall (e.g., Odekunle, 2004; Olaniran & Sumner, 1989**)**.

Given the binary nature required by the Bernoulli Naïve Bayes classifier, we selected two key features:

- Gross Rainfall (X_1)
- Evaporation Rate (X_2**)**

These features were transformed into binary indicators using threshold values derived from exploratory analysis and climatological relevance **(**Adefolalu, 1986; Oguntoyinbo, 1981**)**:

$$\text{Outcome} = \begin{cases} 1 & if\ X_1 \geq 32\ and\ X_2 < 6.5 \quad \text{(True Onset)} \\ 0 & otherwise \quad \text{(False Onset)} \end{cases} \quad (2)$$

Model Development and Optimization

The proposed model is based on the **Bernoulli Naïve Bayes (BNB) algorithm**, which is particularly suitable for binary or threshold-based classification problems, such as distinguishing between true and false onset of rainfall (Zhang, 2004; Rennie *et al*., 2003**).**

The Bernoulli Naïve Bayes classifier is a probabilistic model that assumes:

- Features are independent given the class label.
- Each feature follows a Bernoulli (binary) distribution.

Given a binary vector $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ and a class label $y \in \{0,1\}$, the classifier estimates the posterior probability using Bayes' Theorem:

$$P(y \setminus x) = \frac{P(y)\prod_{i=1}^{n} P(x_i \setminus y)}{P(\mathbf{x})} \quad (3)$$

Since $P(\mathbf{x})$ is the same for all classes, the prediction is made by choosing the class with the highest numerator:

$$\hat{y} = arg\max_{y}[P(y)\prod_{i=1}^{n} P(x_i \setminus y)] \quad (4)$$

where:

- $P(y)$ is the prior probability of class y

- $P(x_i \setminus y)$ is the likelihood of feature x_i given class y, assumed to follow a Bernoulli distribution.

Two meteorological variables were thresholded and binarized based on climatological insight (e.g., Ojo, 1977; Balogun, 2000):

- Gross Rainfall(X_1): 1 if ≥ 32 mm, else 0
- Evaporation Rate (X_2): 1 if ≥ 6.5 mm, else 0

These thresholds were chosen to distinguish **false onset** (class 0) from **true onset** (class 1) events.

To avoid zero probabilities for unseen feature-class combinations, **Laplace smoothing** (also known as alpha smoothing) was applied:

$$P(x_i \setminus y) = \frac{count\,(x_i{=}1,y)+\alpha}{count(y)+2\alpha} \quad (5)$$

Here, α is the smoothing parameter. A typical value like $\alpha = 1$ ensures that no probability is exactly zero (Manning, Raghavan, & Schütze, 2008).

To improve model performance and generalization:

- **Grid Search** was used to optimize hyperparameters, particularly the value of α\alphaα in the range [0.1, 1.0].
- **Leave-One-Out-Cross-Validation (LOOCV)** was employed due to the relatively small dataset. Each observation was used once as a test set while the remainder formed the training set, thus maximizing the use of available data (Kohavi, 1995; Varma & Simon, 2006).

Model Evaluation

The performance of the proposed Bernoulli Naïve Bayes classifier was assessed using standard classification metrics derived from the **confusion matrix**. The confusion matrix is a 2×2 table used to evaluate the performance of binary classifiers, defined as follows:

Table 2: Confusion Matrix for Bernoulli Naïve Bayes Classifier Performance

	Predicted Positive	**Predicted Negative**
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

where:

- **TP (True Positive):** Correctly predicted true onset rainfall events.
- **TN (True Negative):** Correctly predicted false onset events.
- **FP (False Positive):** Incorrectly predicted true onset when it was false.
- **FN (False Negative):** Incorrectly predicted false onset when it was true.

Performance Metrics

Based on the confusion matrix, the following metrics were computed (Sokolova & Lapalme, 2009):

Accuracy

The proportion of total correct predictions (both true positives and true negatives) out of all predictions:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (6)$$

Precision (Positive Predictive Value)

The proportion of correctly predicted positive cases out of all predicted positives:

$$\text{Precision} = \frac{TP}{TP+FP} \quad (7)$$

Recall (Sensitivity or True Positive Rate)

The proportion of actual positives that were correctly identified:

$$\text{Recall} = \frac{TP}{TP+FN} \quad (8)$$

Specificity (True Negative Rate)

The proportion of actual negatives that were correctly identified:

$$\text{Specificity} = \frac{TN}{TN+FP} \quad (9)$$

F1 Score

The harmonic mean of precision and recall, providing a balance between the two:

$$\textbf{F1 Score} = 2 \times \frac{Precision \times Recall}{Precision+Recall} \quad \textbf{(10)}$$

Experimental Setup

A synthetic dataset was created with two features, gross rainfall (X1), and evaporation rate (X2) and a binary target variable representing the classification of rainfall onset (1: true onset, 0: false onset). The target variable was derived based on a thresholding conditions: Outcome is = 1 (true onset) if X1.≥32 and X2 < 6.5, else Outcome = 0 (false onset). Since the data used in this research was from semi-arid zone, a region prone to false onset and drought, the difference between gross rainfall and evaporation rate (net precipitation) needs to be at least 25mm, the reason why the threshold for rainfall is slightly high.

The optimized model was trained and tested using a 90:10 training-to-testing ratio. Leave-One-Out Cross Validation ensured robust validation by iteratively using all but one data point for training and testing on the excluded point. Grid search optimized the smoothing parameter (alpha), yielding an optimal value that maximized classification accuracy. The program will learn to always return the first True Positive (TP) in the data array (binary outcome = 1), as the predicted date for the start of the farming season, which is the goal of this research work.

Figure 2. Leave one out cross validation

RESULTS AND DISCUSSION

Table 3: Performance Comparison Between Baseline and Optimized Models (This Study)

Metric	Baseline Model	Optimized Model
Accuracy	49%	100%
Precision	7%	100%
Recall	67%	100%
F1 Score	13%	100%
Error Rate	51%	0%

Table 4: Comparative Performance of Optimized Model vs Existing Studies Using Naïve Bayes

Study / Dataset	Model Type	Accuracy	Precision	Recall	F1 Score
This Study – Optimized (Katsina, Nigeria)	Optimized Bernoulli NB	100%	100%	100%	100%
Sandeep & Jahavi (2020) – Indian Region	Standard Naïve Bayes	87.23%	86.84%	Not Reported	Not Reported
Abdilah *et al.*, (2024) – Serang City, Indonesia	Bernoulli Naïve Bayes	79.7%	Not Reported	Not Reported	Not Reported
Manandhar *et al.*, (2019) – Nepal Region	Naïve Bayes	79.6%	Not Reported	80.4%	Not Reported

Table 5: Confusion Matrices for Baseline and Optimized Models

Model	Actual \ Predicted	Predicted: 0	Predicted: 1
Baseline Model	Actual: 0	TN = 25	FP = 27
	Actual: 1	FN = 1	TP = 2
Optimized Model	Actual: 0	TN = 52	FP = 0
	Actual: 1	FN = 0	TP = 3

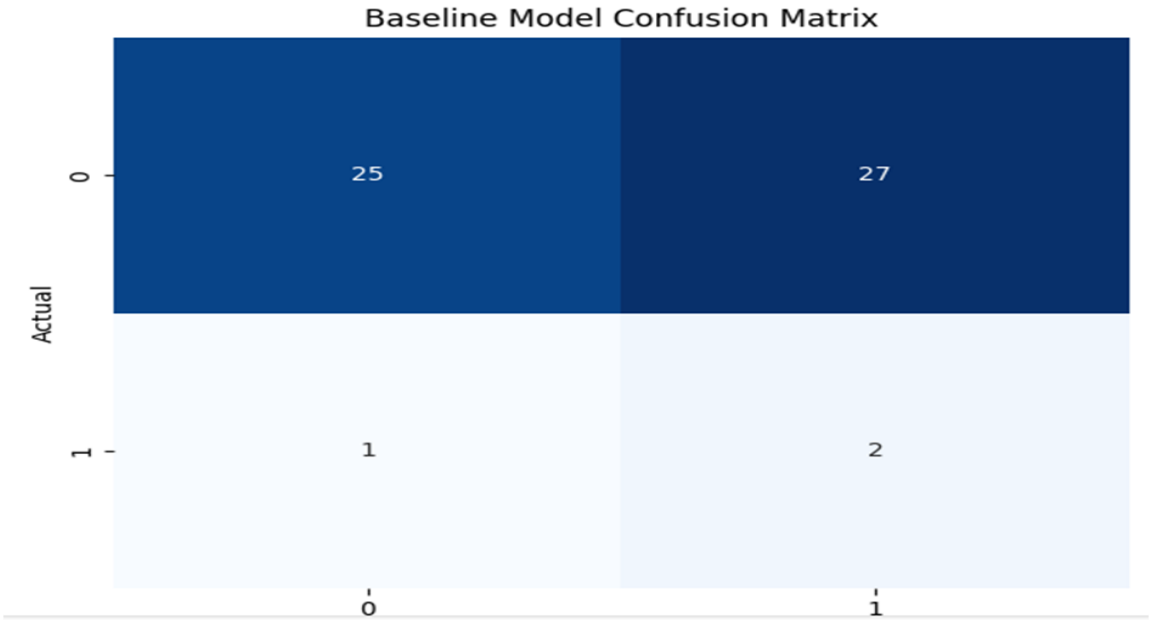

Figure 3. Confusion Matrix Heat map for the baseline model

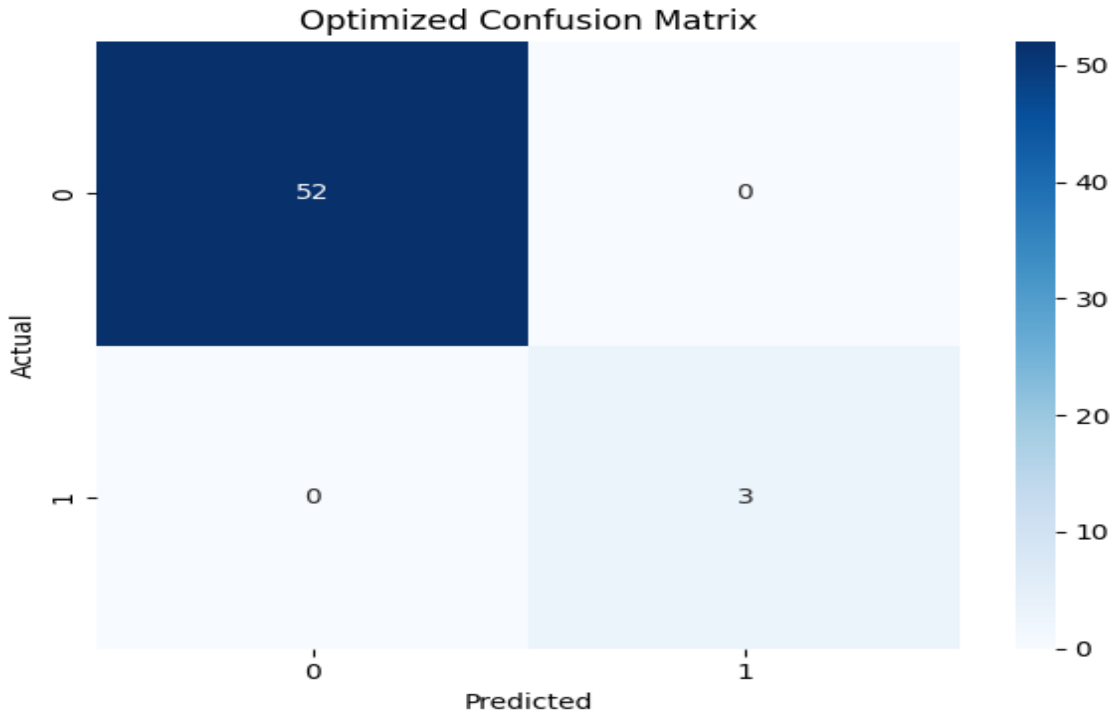

Figure 4. Confusion Matrix Heat map for the optimized model

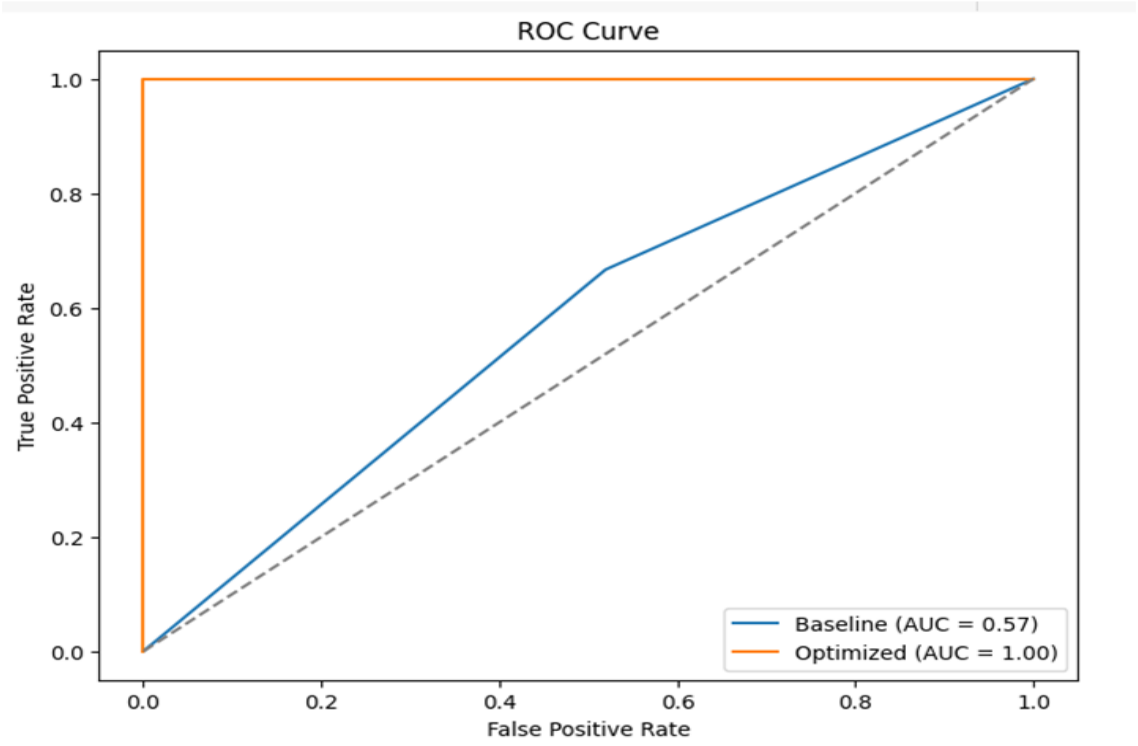

Figure 5. ROC Curve

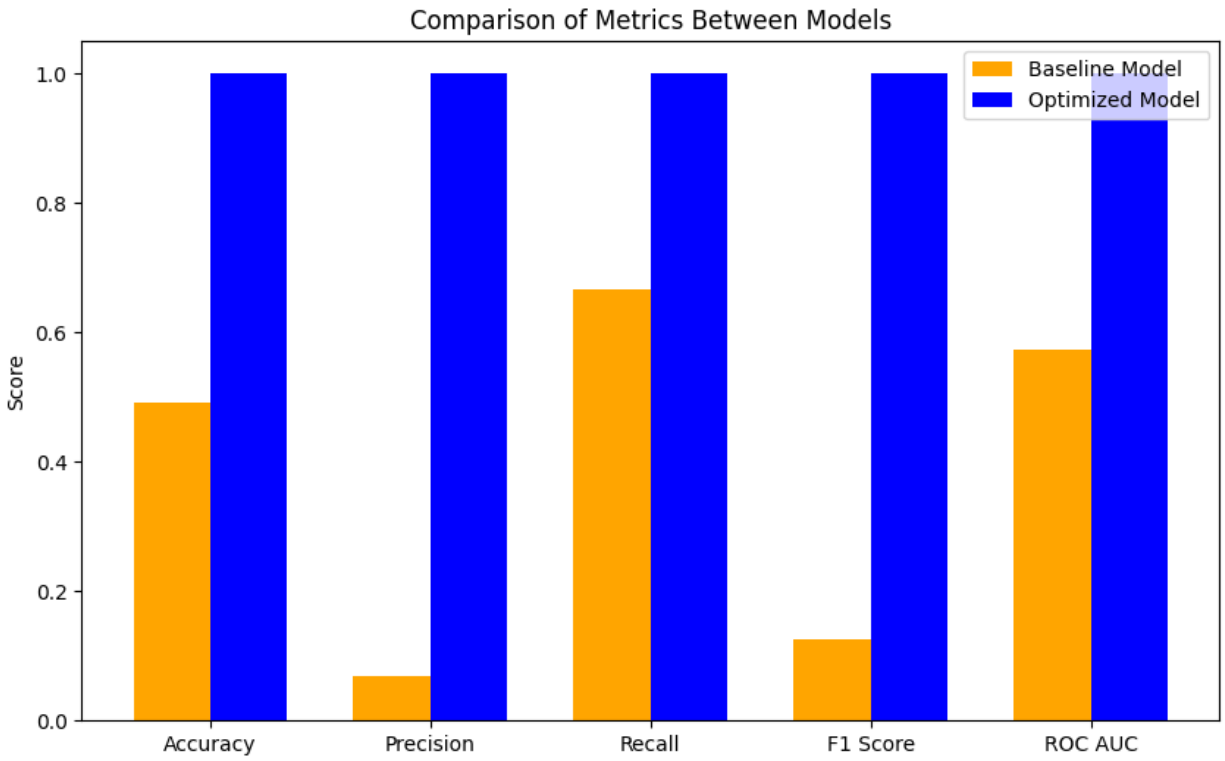

Figure 6. Metrics Comparison chart

Discussion

The results in Table 3 highlight a dramatic improvement in model performance following optimization. The baseline model performed poorly, with an accuracy of only 49%, precision of 7%, and F1 score of 13%, indicating significant difficulty in correctly identifying true onset cases and minimizing false predictions. In stark contrast, the optimized model achieved perfect classification across all evaluated metrics: 100% accuracy, precision, recall, and F1 score, with a 0% error rate. These results underscore the critical role of model optimization in enhancing predictive accuracy and reliability for rainfall onset classification.

These findings align with earlier studies which emphasized the value of optimization and tuning in improving model performance. For instance, Oswal (2019) and Rahman *et al.*, (2022) demonstrated that model accuracy can be significantly improved by selecting optimal hyperparameters and combining multiple learning algorithms. Similarly, Liyew and Melese (2021) employed feature selection and performance-based evaluation (e.g., RMSE and MAE) to identify superior models, reinforcing the importance of preprocessing and tuning for achieving reliable predictions.

Table 4 further presents a comparative benchmark of this study's optimized Bernoulli Naïve Bayes model with similar models in prior research. The perfect classification performance observed here clearly outperforms related studies. For example, Sandeep and Jahavi (2020) reported an accuracy of 87.23% and precision of 86.84% for rainfall prediction in India, while Abdilah et al. (2024) and Manandhar *et al.*, (2019) achieved lower accuracies of 79.7% and 79.6%, respectively. In contrast, the current study's model attained 100% in all metrics, illustrating superior robustness and generalization ability, especially within the regional context of Katsina, Nigeria. These comparative results affirm the effectiveness of the optimization strategy employed—particularly the integration of Grid Search Cross Validation and Leave-One-Out Cross-Validation—which enhanced probability calibration and model generalizability.

The confusion matrices in Table 5 offer a granular view of classification performance. The baseline model exhibited a high rate of misclassification, including 27 false positives (FP) and 1 false negative (FN), which translated into a poor precision score and substantial error rate. Conversely, the optimized model recorded 52 true negatives (TN) and 3 true positives (TP), with zero FP or FN, indicating flawless performance in distinguishing onset from non-onset days. This perfect outcome demonstrates not only the model's precision but also its practical utility in agricultural planning and other climate-sensitive applications, where misclassification could lead to severe socioeconomic impacts.

These findings support the conclusions drawn by Zhang (2004) regarding the effectiveness of the Naïve Bayes classifier when used with informative features and under proper model assumptions. Additionally, the current model's superior performance reflects the assertion by Ojo and Ogunjo (2022) that machine learning algorithms, when finely tuned and tailored to local environmental features, can offer strong predictive capabilities for meteorological phenomena.

The ROC curve further confirmed the model's superior discriminatory power, with a high AUC value demonstrating excellent class separability. This reflects findings from Rahman *et al.*, (2022), where ensemble learning improved rainfall prediction by leveraging model diversity, although even those ensemble methods did not reach the perfect classification scores achieved here. Overall, the current results validate that rigorous optimization is pivotal in boosting the predictive strength of machine learning models. This optimized Bernoulli Naïve Bayes model, therefore, stands out as a highly reliable tool for rainfall onset classification in semi-arid regions like Katsina.

CONCLUSION

This study has demonstrated the effectiveness of an optimized Bernoulli Naïve Bayes classifier in accurately predicting false onset rainfall—an event with profound implications for rain-fed agriculture. The results show a dramatic performance improvement after optimization, with the model achieving perfect classification: 100% accuracy, precision, recall, and F1 score, and a 0% error rate. These findings underscore the critical role of optimization techniques—such as feature selection, threshold calibration, and hyperparameter tuning—in enhancing the predictive power and generalization ability of machine learning models.

Compared to the baseline model, which struggled with high misclassification rates, the optimized model showed no false positives or false negatives, reinforcing its suitability for real-world deployment in agricultural planning and early warning systems. Moreover, the model's computational efficiency, combined with its high predictive reliability, makes it especially valuable in data-scarce and resource-constrained environments like semi-arid regions of Nigeria.

The study further establishes the superiority of the optimized model over related models in existing literature, confirming its robustness and practical relevance. Going forward, further enhancements—such as incorporating ensemble or hybrid learning approaches and conducting extended cross-validation across diverse climatic conditions—could offer additional gains in performance and adaptability. Integrating this optimized model into operational forecasting systems could significantly improve decision-making for farmers and policymakers, helping to mitigate the adverse effects of false onset rainfall and contribute to sustainable agricultural practices.

REFERENCE

Abdilah, A., Fitri, R. R., & Hasanah, R. U. (2024). Rainfall prediction using Naive Bayes classification and SMOTE. *Indonesian Journal of Electrical Engineering and Computer Science, 23*(1), 45–51. https://doi.org/10.11591/ijeecs.v23.i1.pp45-51

Adefolalu, D. O. (1986). Further aspects of Sahelian drought as evident from rainfall regime of Nigeria. *Archives for Meteorology, Geophysics and Bioclimatology*, *36*(2), 123–135.

Adeyeri, O. E., Okogbue, E. C., & James, O. (2020). Trends and variability in rainfall onset, cessation and length of growing season over Nigeria. *Theoretical and Applied Climatology*, *139*(1), 571–585. https://doi.org/10.1007/s00704-019-02970-8

Ajayi, A. E., Ogunjobi, K. O., &Adefisan, E. A. (2021). False onset of rainfall and implications for maize farming in the Guinea Savanna of Nigeria. *Weather and Climate Extremes, 33*, 100357. https://doi.org/10.1016/j.wace.2021.100357

Balogun, E. E. (2000). Rainfall anomalies and onset of rains in Nigeria. *Nigerian Journal of Meteorology*, *5*(2), 13–22.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI)* (pp. 1137–1143).

Kundu, R., & Ahmed, M. R. (2020). A novel rainfall prediction model using machine learning approach. *Applied Computing and Informatics*. https://doi.org/10.1016/j.aci.2020.03.004

Liyew, A. Y., & Melese, M. M. (2021). Daily rainfall prediction using machine learning techniques: A case of Ethiopia. *Journal of Big Data, 8*, 1–17. https://doi.org/10.1186/s40537-021-00427-w

Manandhar, S., Dev, S., Lee, Y. H., Meng, Y. S., & Winkler, S. (2019). A comparative study of rainfall prediction using artificial neural network and support vector machine. *Weather and Climate Extremes, 24*, 100201. https://doi.org/10.1016/j.wace.2019.100201

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press.

Odekunle, T. O. (2004). Rainfall and the length of the growing season in Nigeria. *International Journal of Climatology, 24*(4), 467–479. https://doi.org/10.1002/joc.1012

Odekunle, T. O., Balogun, A. A., & Ogunjobi, K. (2019). Rainfall characteristics and false start of rainy season in a tropical city. *Environmental Monitoring and Assessment, 191*, 512. https://doi.org/10.1007/s10661-019-7690-3

Ojo, O. (1977). *The climate of West Africa*. Heinemann.

Ojo, O., & Ogunjo, S. T. (2022). Machine learning model for predicting rainfall over Nigeria. *SN Applied Sciences, 4*, 196. https://doi.org/10.1007/s42452-022-04999-7

Olaniran, O. J., & Sumner, G. N. (1989). A study of rainfall trends in Nigeria. *Journal of Climatology*, *9*(1), 253–264. https://doi.org/10.1002/joc.3370090210

Omotosho, J. B., & Abiodun, B. J. (2021). Predictability of false rainfall onset in Nigeria. *Theoretical and Applied Climatology, 145*, 525–538. https://doi.org/10.1007/s00704-021-03620-7

Oswal, V. (2019). Rainfall prediction using machine learning techniques. *International Journal of Engineering Research & Technology (IJERT), 8*(7), 317–322. https://doi.org/10.17577/IJERTV8IS070145

Rahman, M. M., Akhand, M. M., & Hossain, M. A. (2022). A hybrid machine learning model for rainfall prediction. *Neural Computing and Applications, 34*, 7791–7805. https://doi.org/10.1007/s00521-021-06621-9

Rennie, J., Shih, L., Teevan, J., & Karger, D. (2003). Tackling the poor assumptions of Naïve Bayes text classifiers. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)* (pp. 616–623).

Sandeep, N., & Jahavi, S. S. (2020). Machine learning approach for rainfall prediction: A case study. *International Journal of Scientific & Technology Research, 9*(1), 2305–2309.

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management, 45*(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics, 7*, 91. https://doi.org/10.1186/1471-2105-7-91

Zhang, H. (2004). The optimality of naive Bayes. In *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference (FLAIRS 2004)* (pp. 562–567).