*American Journal of Data Mining and Knowledge Discovery*
2024, Vol. 9, No. 1, pp. 1-19
https://doi.org/10.11648/j.ajdmkd.20240901.11

Research Article

# Exploring a Novel Approach of K-mean Gradient Boosting Algorithm with PCA for Drought Prediction

**Babatunde Isaiah Ayinla*, Rasheedat Aderonke Abdulsalam**

Department of Computer Science, University of Ibadan, Ibadan, Nigeria

## Abstract

Drought poses a significant threat to essential resources like food, land, and public health. Machine Learning (ML) has emerged as a powerful tool in weather forecasting, leveraging algorithms to predict weather phenomena with remarkable accuracy. ML models excel in navigating complex atmospheric systems, including those affected by climate change, offering precision beyond traditional forecasting methods. However, predicting drought remains challenging due to its uneven distribution and varying degrees. To tackle this challenge, an exploration of a novel approach of combining K-means++ clustering and Gradient Boosting Algorithm (KGBA) with Principal Component Analysis (PCA) for dimensionality reduction was carried out. Using a dataset spanning from 2000 to July 2016, comprising 2,756,796 US Drought Monitor records, the study developed and evaluated the KGBA model's effectiveness in drought prediction. The results demonstrated the superiority of high precision and recall rates, particularly in forecasting extreme and exceptional drought periods. Specifically, KGBA attained precision accuracies of 33% and 74%, along with recall rates of 72% and 77% for predicting extreme and exceptional drought periods, respectively. The model had an overall accuracy of 46% in predicting all the multiple classes of droughts. A performance that is slightly better than other ensemble methods that had the closest performance. These findings underscore the potential of KGBA in enhancing the predictive capabilities for drought mitigation efforts, as it outperformed other models such as Gradient Boosting, Random Forest, Bayes Naive, and K-Nearest Neighbor.

## Keywords

K-means++, Gradient Boosting, Drought, Principal Component Analysis, Machine Learning and Climate Change

## 1. Background of the Study

Climate refers to the long-term patterns and variations in key atmospheric factors like temperature, precipitation, and wind. Essentially, climate represents a comprehensive summary of weather conditions over time. Drought, on the other hand, is a notable environmental occurrence that carries substantial consequences for agriculture, water resources, and natural habitats. Precise forecasting of drought events can play a crucial role in proactive planning and implementing strategies to alleviate the adverse effects of droughts.

According to the National Oceanic and Atmospheric Administration, US Commerce 2021 report. The world experienced coronavirus pandemic shutdowns in 2020, yet there was a surge in the level of anthropogenic greenhouse gases; that is, $CO_2$ rose to about 40% from 25% since 1953, and methanes also increased in its percentage (NOAA Science Report, 2021). Climate change is beyond rising in tempera-

*Corresponding author: ayinlab@gmail.com (Babatunde Isaiah Ayinla)

**Received:** 11 June 2024; **Accepted:** 8 July 2024; **Published:** 23 July 2024

tures. However, Global temperatures continue to rise to about 1.980 F (1.10C) between 1901 and 2020. The effect of Climate change is also reflected in the rise of sea level from 1.7 mm/year, mostly in the twentieth century, to 3.2 mm/year. The shifting weather patterns brought about by climate change are causing both droughts and floods, adversely impacting essentials for human and animal survival [1].

The adverse effects of climate change are obvious in various aspects of our society. However, drought can be devastating, where food production could be grossly affected, and human health could be highly degraded. In a similar vein, flooding can lead to the spread of diseases that can wipe out a complete community and cause damage to social amenities and ecosystems. Climate change (CC) manifests its negative impact across various aspects of our world. It is also relevant that climate change differs from neighbourhoods or individuals to others.

The world has been facing the challenges of adaptation and strategies to survive in the midst of climate change for decades. One of the key reasons for the failure of countries to tackle climate change vulnerability is the lack of appropriate steps in risk management. Mortuza–observed that the government of countries like Bangladesh tends to prioritize response and recovery efforts over monitoring, preparedness, and mitigation strategies [2]. Therefore, there is a need for accurate drought projection tools for the sustainable management of all aspects affected by CC, such as agriculture, health resources, and more. Given the unpredictable and diverse characteristics of drought, which can vary in intensity and occurrence across different locations, there's a critical need to develop swift, reliable, and accurate prediction models. These models are essential for quantifying the risks associated with drought events and for better understanding their potential impacts [2].

Traditional drought prediction systems frequently rely on meteorological indicators such as temperature and precipitation. These approaches, however, may fail to capture the intricate linkages and patterns associated with drought [26]. In recent years, machine learning algorithms have shown promise in improving drought forecast accuracy by incorporating new data sources and recognizing non-linear connections. Khan reported the utilization of various machine learning (ML) models, including support vector machine (SVM), artificial neural networks (ANN), and Extreme Learning Machine (ELM) [3]. However, the most widely used model in the domain of drought prediction has been Support Vector Machine (SVM) along with other potent algorithms [4-7]. The algorithm is combined with SPEI to predict drought over Pakistan, Palmer Drought Severity Index (PDSI) to predict drought over Turkey. Some researchers have examined the performance of ANN in predicting the Standardized Precipitation Index (SPI) over Iran.

## 1.1. Principal Component Analysis (PCA)

This is a sophisticated statistical method for reducing the dimensionality of large data sets. It is widely used in data science. The approach has its strength in managing complex and highly dimensional datasets without significant loss of information. This methodology has taken a strong position in various fields, from finance to genomics. These areas are known for massive amounts of data that require interpretation. It highlights different patterns in the dataset as well as any similarities among them, essentially converting the original features into new, uncorrelated features called principal components. The components are orthogonal axes of maximum variance that stand as datasets in a reduced-dimensional space [8].

For instance, consider a data matrix, X, with column-wise zero empirical mean, a situation where each of the columns of the sample data has been shifted to zero. Each of the *n* rows stands as the repetition of the experiment, where each of the p columns generates a new feature.

In mathematical terms, the transformation involves a collection of weight vectors or coefficients, each of size p dimensions, which map every row vector of X is transformed into a new set of principal components. scores, denoted as t. These coefficients are structured to maximize the variance inherited from X across the individual variables of t in the dataset. Furthermore, each coefficient vector w is treated as a unit vector, typically leading to a reduction in the dimensionality of the observations.

When the First component is observed to maximize variance, the first weight vector $w_{(1)}$ must satisfy. In other word, expressing this in matrix form yields:

$$x_1 = \lambda_{11}f_1 + \lambda_{21}f_2 + \ldots\ldots \lambda_{1m}f_m + e_1 \quad (1)$$

$$x_2 = \lambda_{21}f_1 + \lambda_{22}f_2 + \ldots\ldots \lambda_{2m}f_m + e_2$$

$$\ldots$$

$$\ldots$$

$$X_p = \lambda_{p1}f_1 + \lambda_{p2}f_2 + \ldots\ldots \lambda_{pm}f_m + e_p$$

where $\lambda_{jk}$, j = 1, 2,...,p; k = 1, 2,...,m are constants called the factor loadings, and $e_j$, j = 1, 2,...,p are error terms, sometimes called specific factors (because $e_j$ is 'specific' to $x_j$ , whereas the $f_k$ are 'common' to several $x_j$).

Since $w_{(1)}$ has been defined as a unit vector, it also satisfies the equation.

The Rayleigh quotient is the quantity to be maximized. The maximum possible value of the quotient, indicated by the largest eigenvalue of the matrix, is a standard result for positive semidefinite matrices like $X^TX$. This maximum value is achieved when *w* corresponds to the eigenvector.

The $w_{(1)}$ is the first principal component of a data vector $x_{(i)}$, that is

$$\text{score } t_{1(i)} = x_{(i)} \cdot w_{(1)} \quad (2)$$

represented as the transformed co-ordinates, or could the seen

as the corresponding vector in the original variables, $\{x_{(i)} \cdot w_{(1)}\} w_{(1)}$.

The other components of the PCA can be deduced as:

The $k$-th component is the subtraction of first $k-1$ principal components from matrix X in (1):

Then, the weight vector is subsequently found, which is the extraction from the maximum variance of the new data matrix.

The above result in (2) extends to the remaining eigenvectors of $X^TX$, with the maximum values for the quantity in brackets given by their corresponding eigenvalues. Hence, the weight vectors are eigenvectors of $X^TX$.

The $k$-th principal component of a data vector $x_{(i)}$ is defined as the

$$\text{score } t_{k(i)} = x_{(i)} \cdot w_{(k)} \quad (3)$$

in the transformed coordinates, or as the corresponding vector in the space of the original variables,

$\{x_{(i)} \cdot w_{(k)}\} w_{(k)}$, where $w_{(k)}$ is the $k$th eigenvector of $X^TX$.

The entire principal components decomposition of X can be defined as follow:

W represents a $p$-by-$p$ matrix of weights, where each column corresponds to the eigenvectors of $X^TX$. The transpose of W, often referred to as the whitening or sphering transformation, is calculated by scaling the columns of the W matrix by the square root of the corresponding eigenvalues. This process, essentially multiplying the eigenvectors by their respective variances, results in loadings in PCA or Factor analysis, as described by Sidak in 2023 [9].

This can further be mathematically represented as follows:

$$PCA\_1 = a\_11 * x\_1 + a\_12 * x\_2 + ... + a\_1p * x\_p \quad (4)$$

$$PCA\_2 = a\_21 * x\_1 + a\_22 * x\_2 + ... + a\_2p * x\_p$$

$$...$$

$$PCA\_k = a\_k1 * x\_1 + a\_k2 * x\_2 + ... + a\_kp * x\_p$$

where a_ij is the loading or weight of variable x_j on principal component PCA_i, and x_j is the jth variable in the data matrix X. The principal components are ordered such that the first component PCA_1 captures the most significant variation in the data, the second component PCA_2 captures the second most significant variation, and so on as earlier described. The number of principal components used in the analysis, k, determines the reduced dimensionality of the dataset [8, 9].

The PCA is useful in three major areas of model building, especially during the preprocessing stage. The areas are data reduction, an approach that simplifies model building in machine learning and statistical analysis by reducing the number of variables under consideration. Secondly, data analysis exploration; It uncovered hidden patterns in the preliminary stages of data analysis and thirdly, multivariate Analysis that deal with observations of multiple interrelated features.

## 1.2. Machine Learning Models

According to Mokhtar et al, machine learning models are mathematical representations of sets of data where predictions can be made for decision-making [10]. The models are built from training machine learning algorithms to learn from historical datasets, which are either labelled or unlabelled. Once the training is done, a generalized prediction can be made from unseen datasets. These models have revolutionalised several domains of interest, including security, health, finance, and the like. The model has the ability to uncover insights, including patterns and irregularities in data, in a more sophisticated way than the traditional statistical models [11]. Furthermore, Machine Learning Models (MLMs) exhibit portability, robustness, and flexibility, enabling them to perform effectively across diverse tasks, including assessing patient risk levels, making diagnoses, and predicting outcomes [12]. However, most MLMs are black boxes, and explainability and interpretability are concerns [13, 14].

### 1.2.1. Gradient Boosting Algorithm (GBA)

Gradient boosting is a versatile ensemble machine learning method suitable for regression and classification tasks. It combines the predictions of numerous weak learners, typically decision trees, sequentially. The primary aim is to enhance predictive accuracy by optimizing the model's weights. These weights are determined by the errors of previous iterations, gradually reducing errors to refine the final model accuracy. Employing an arbitrary differentiable loss function, the model is systematically constructed stage by stage, akin to other boosting algorithms.

Similarly, the gradient boosting algorithm originated from Leo Breiman's idea that boosting can be conceptualized as an optimization algorithm on an appropriate cost function [15]. In the years 2001 and 2002, Jerome H Friedman developed a regression version of the gradient boosting algorithm. The algorithm uses the approach proposed by Mason et al. [16-18].

There are basically three family members of GBA: XGBoost, LightGBM, and CatBoost, with the aim of achieving better accuracy and speed optimization as the focus. Extreme Gradient Boosting Algorithm (XGBA) is known for its scalability, efficiency and reliability among the machine learning algorithms. However, LightGBM is extremely fast in model training with the use of selective samplings of high-gradient records. In a similar way, CatBoost places a premium on the accuracy prediction of the model by modifying the computation of gradients [19].

The Gradient tree boosting algorithms are based on the derivation outlined by Friedman et al. with minor enhancements made to the regularized objective function, which have proven to be beneficial in practical applications [20].

The Gradient Tree Boosting ensemble model can be illus-

trated with Eq. (5). This involves functions being treated as parameters, a process that conventional optimization techniques in Euclidean space cannot handle. Essentially, the model is trained in an additive manner, allowing for optimization through the inclusion of these functions as parameters. Assume $\hat{y}$(t) as the value of prediction of the *i*-th tuple at the looping *t*-th. To minimize the objective, $f_t$ needs to be added.

$$L(\emptyset) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \quad (5)$$

where $\Omega(f) = \gamma T + \frac{1}{2}\lambda || \omega ||^2$

In this expression, *l* symbolizes a differentiable convex loss function, which evaluates the difference between the predicted value $\hat{y}_i$and the actual target value prediction $\hat{y}_i$. The subsequent term Ω penalizes the complexity of the model, particularly focusing on the regression tree functions. Furthermore, the regularization term is incorporated to refine the final learned weights, effectively addressing the issue of overfitting.

$$L^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) \quad (6)$$

The equation in (6) implies that the $f_t$ that most improves it is greedily added.

According to Friedman et al., 2000, the second-order approximation is speedily applied to optimize the objective as presented in (5).

$$L^{(t)} \simeq \sum_{i=1}^{n} [l(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t) \quad (7)$$

where gi = ∂yˆ(t−1)l(yi , yˆi (t−1)) and hi = ∂ 2 yˆ(t−1)l(yi , yˆi (t−1)), the first and second-order gradient statistics are computed on the loss function. After removing the constant terms, the simplified objective at step *t* is obtained.

L˜(t) = Xn i=1 [gift(xi) + 1 2 hif 2 t (xi)] + Ω(ft) (3) Define Ij = {i|q(xi) = j} as the instance of leaf j. We can rewrite (7) by expanding Ω as

$$\tilde{L}^{(t)} = \sum_{i=1}^{n} [g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t) \quad (8)$$

The final prediction for a given group of samples is the sum of all the predictions from each tree as follows

$$\tilde{L}^{(t)} \simeq \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)\right] + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2 \quad (9)$$

$$\sum_{j=1}^{T} [\left(\sum_{i \in I_j} g_i\right) w_j + \frac{1}{2}\left(\sum_{i \in I_j} h_i + \lambda\right) \omega_j^2] + \gamma T$$

The optimal weight $w_j$* for a given structure *q*(*x*) can be calculated using (9).

$$\omega_j^i = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda} \quad (10)$$

The corresponding optimal value can be determined by the following calculation:

$$\tilde{L}^{(t)}(q) = -\frac{1}{2}\sum_{j=1}^{T} \frac{\left(\sum_{i \in I_j} g_i\right)^2}{\sum_{i \in I_j} h_i + \lambda} + \gamma T \quad (11)$$

The equation in (11) can serve as a scoring function to evaluate the quality of a tree structure *q*. This score functions similarly to the impurity score used in decision trees but is applicable to a broader range of objective functions.

### 1.2.2. Random Forest (RF)

The RF is a special model that is based on multiple decision trees that are called forest with controlled variance [13, 21]. This model, called Random Forest, has been potency to work with both continuous and discrete datasets. The classifiers are known as regression and classification models, respectively. A random forest regression is a bootstrap ensemble. It works with random binary trees that made use of a subset of the datasets through bootstrapping, from a random subset of the training dataset isolated to build the models [10].

## 2. Literature Review

In the study "Applying Machine Learning for Threshold Selection in Drought Early Warning System" by Luo et al. 2022, the correlation between NDVI readings and drought categories is investigated across a 34-year timeframe in two distinct climate zones in Australia. The research aims to establish NDVI threshold values for different drought categories through a threshold selection approach. While the model provides valuable insights into drought severity and lays the groundwork for future drought classification models, additional efforts are necessary to enhance the accuracy of the model.

With a one-month lead time, artificial neural networks are used in "Applying machine learning for drought prediction using data from a large ensemble of climate simulations" [22]. This method predicts the onset of drought in two European domains. The paper addresses the application of explainable AI methods to acquire insights into the outcomes using data from a model of a large ensemble. Consequently, the models give chances to examine the impact of input variables on drought formation and serve as a foundation for the creation of future drought prediction models, but weak prediction accuracies are noted [23].

According to Jiang and Luo's article "An Evaluation of Machine Learning and Deep Learning Models for Drought Prediction Using Weather Data" [11]. An experiment was conducted to analyze different AI models both machine and deep learning models to predict drought using dataset from United State. It was reported that no one model performs best

for all the assessment criteria, due to the imbalanced nature of drought events, special attention was given to developing models capable of accurately predicting drought occurrences.

The study "The global k-means clustering algorithm" by Likas, Vlassis, and Verbeek in 2011, suggested changes to the technique to lessen the computing effort without significantly reducing solution quality [24]. The evaluation took into account both the solution quality and computational complexity. It was discovered that the global k-means algorithm was quite effective. But there is a significant issue that needs more research, and that has to do with the potential creation of theoretical underpinnings for the method's underlying presumptions [24].

Tri et al. in the study titled "Application of Meteorological and Hydrological Drought Indices to Establish Drought Classification Maps of the Ba River Basin in Vietnam,". The aimed of this study was to generate maps illustrating the lack of discharge in the Ba River basin of Vietnam. They employed various indices, such as the Standardized Precipitation Index (SPI), the Drought Index (I), and the Ped Index (Ped), in conjunction with the Soil and Water Assessment Tool (SWAT) model and the hydrological drought index (KDrought), to create these maps [25].

The hydrological drought index for the study area was derived by utilizing the simulation outcomes from the SWAT model. The impacts of the drought on both the spatial and temporal dimensions of the study area were assessed through drought classification maps generated from the calculated drought index (KDrought). While there were limited calibrations and validations conducted on the SWAT model, the study identified a correlation between the moisture regime and drought occurrences in the Central region.

## 3. Research Methodology

This study was performed using the US drought dataset, which contains different drought levels by state in the US from 2000 to 2016. The size of the data was 18.28 MB. Similarly, the total records in the dataset were 19,300,680, and 16,543,884 out of the records had null values. Two million, Seven hundred and fifty-six, Seven hundred and ninety-six (2,756,796) were left as viable for the experiment. The dataset was obtained from data. World, it is a dataset containing different drought levels by state in the United States (US) [27]. The structure of the dataset is described in Tables 1 and 2, and the stages of the experiment can be viewed diagrammatically in Figure 1.

Figure 1 shows the extreme skewness of the entire dataset, favouring more free drought seasons. The no drought, I mean class 0 dataset, was approximately 60% of the whole dataset, signifying 1,652,230 out of 2,756,796 rows. The other classes, such as abnormal dry (class 1), moderate drought (class 2), severe drought (class 3), and extreme drought (class 4), had 17% (466,944), 11% (295,331), 7% (196,802) and 4% (106,265) of tuples respectively. The last class of the drought, that is, exceptional drought (class 5) had extremely lowest representation of 1% (39,224) observations.

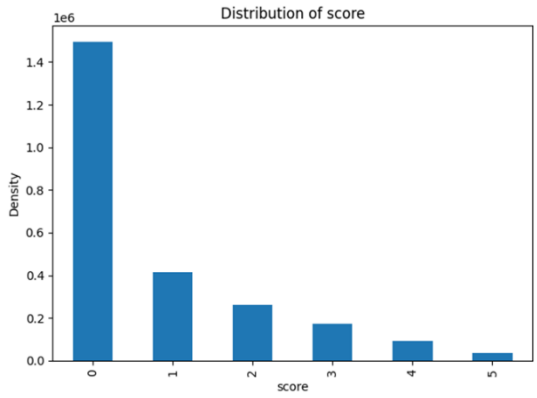



***Figure 1.*** *The Data Distribition according to the target variable.*

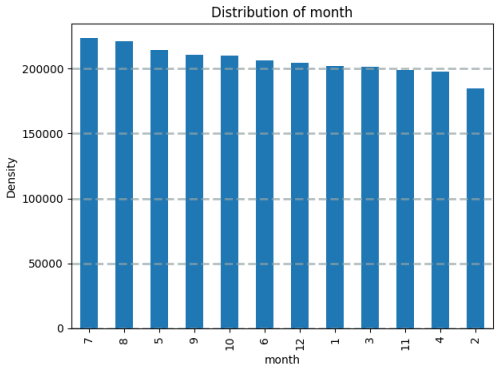



***Figure 2.*** *The Data Distribition according to the months of the year.*

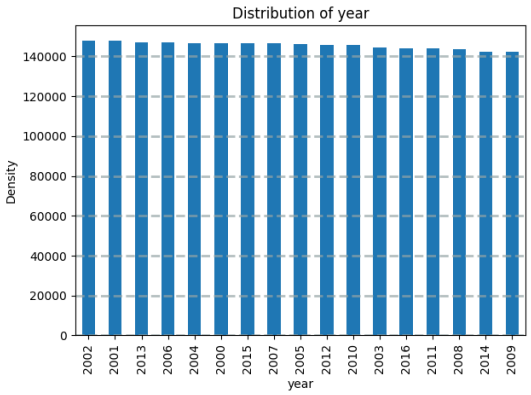



***Figure 3.*** *The Data distribution according to the years.*

Further exploration of the dataset, as revealed in Figure 2, examined the distribution of the dataset according to the months of the years. The figure shows that the months of July and August were prevalent in the dataset, while observations from February and April were least represented. However, the state of the climate change looks similar bimonthly.

Figure 3, reveals that non of the year had less that 140, 000

records in the whole dataset, although 2009 and 2014 had the lowest representations.

*Table 1. The Attributes of the drought's tuples; Independent Variables.*

| S/N | Attributes | Description |
|---|---|---|
| 1. | WS10M_MIN | Minimum Wind Speed at 10 Meters (m/s) |
| 2. | QV2M | Specific Humidity at 2 Meters (g/kg) |
| 3. | T2M_RANGE | Temperature Range at 2 Meters (C) |
| 4. | WS10M | Wind Speed at 10 Meters (m/s) |
| 5. | T2M | Temperature at 2 Meters (C) |
| 6. | WS50M_MIN | Minimum Wind Speed at 50 Meters (m/s) |
| 7. | T2M_MAX | Maximum Temperature at 2 Meters (C) |
| 8. | WS50M | Wind Speed at 50 Meters (m/s) |
| 9. | TS | Earth Skin Temperature (C) |
| 10. | WS50M_RANGE | Wind Speed Range at 50 Meters (m/s) |
| 11. | WS50M_MAX | Maximum Wind Speed at 50 Meters (m/s) |
| 12. | WS10M_MAX | Maximum Wind Speed at 10 Meters (m/s) |
| 13. | WS10M_RANGE | Wind Speed Range at 10 Meters (m/s) |
| 14. | PS | Surface Pressure (kPa) |
| 15. | T2MDEW | Dew/Frost Point at 2 Meters (C) |
| 16. | T2M_MIN | Minimum Temperature at 2 Meters (C) |
| 17. | T2MWET | Wet Bulb Temperature at 2 Meters (C) |
| 18. | PRECTOT | Precipitation (mm day-1) |

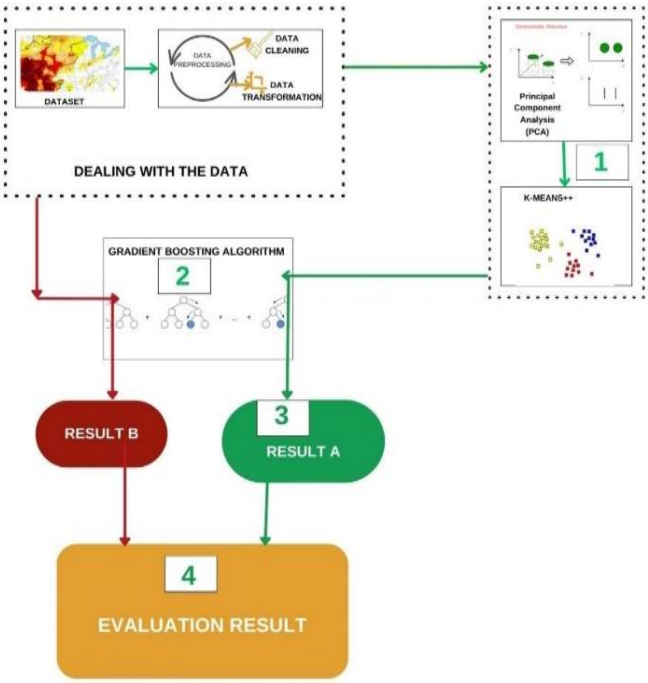



*Figure 4. Generic Architecture of KGBA Model.*

The Gradient Boosting Algorithm (GBA) is a popular technique in supervised machine learning used for regression and classification tasks, as depicted in Figure 4. It works by combining multiple weak individual models to create a stronger final model. This process involves iteratively adjusting the weak hypothesis or learning algorithm to enhance its predictive power.

Among the various frameworks available for Gradient Boosting, LightGBM (LGBM) stands out due to its speed, memory efficiency, tunability, and flexibility. LightGBM utilizes tree-based learning algorithms and is designed to be distributed and efficient, offering advantages over traditional boosting methods. Another popular boosting method, XGBoost, is also notable for its high performance and can serve as a benchmark for comparison [27].

KMeans Gradient Boosting Algorithm (KGBA)

Step 1: Clean the Data

*cleaned_data = clean(data)*

Step 2: Dimensionality Reduction using Principal Component Analysis (PCA)

*reduced_data = reduce_dimensionality_using_PCA(cleaned_data)*

Step 3: Kmeans++

*num_clusters = choose_num_clusters(reduced_data)*

*centroids = initialize_centroids_using_kmeans_pp(reduced_data, num_clusters)*

# Iterate until convergence

*while True:*

# Assign data to closest centroids

*cluster_assignments = assign_data_to_closest_centroids(reduced_data, centroids)*

# Update centroids

*new_centroids = update_centroids(cluster_assignments, reduced_data)*

# Check for convergence

*if centroids == new_centroids:*

break;

*centroids = new_centroids*

Step 4: Gradient Boosting

*X_train, X_test, y_train, y_test = split_data(cluster_assignments, cleaned_data)*

## 3.1. Data Preprocessing

As per the United States Environmental Protection Agency's Climate Change Indicators report on Drought in 2023, this indicator assesses the drought status across U.S. territories. Various indices have been developed to gauge the severity of drought conditions, incorporating factors like precipitation, soil moisture, stream flow, vegetation health, among others. However, the Palmer Drought Severity Index stands out as the most commonly utilized metric, deriving from measurements of precipitation and temperature collected at weather stations. The algorithm 1 provided serves as a preliminary outline of the algorithm employed in this study. It delineates the sequential steps involved and elucidates how various processes interact with each other.

There were 18 independent variables in the dataset with 5 targets as dependent variables as shown in Tables 1 and 2, respectively. The preprocessing aspect is an important step due to the unstructured nature of the dataset. It is imperative to check the data for any errors, noise or missing values for a reliable outcome. Therefore, all uncompleted and duplicated records were removed, and the data were transformed to make it coherent and easy to read by machine learning algorithms. The machine learning algorithms such as K-means++, Principal Component and Gradient Boosting algorithms do not work with string data. Hence, a data encoding function of scikit library of Python was employed to translate the data into a numerical dataset.

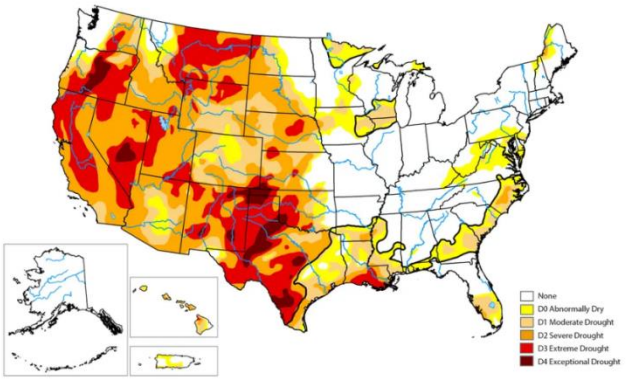



***Figure 5.*** *US Drought Map [28].*

***Table 2.*** *The Map Key Description.*

| S/N | Label | Description |
|---|---|---|
| a. | None | Drought Absent |
| b. | D0 | Abnormally dry |
| c. | D1 | Moderate drought |
| d. | D2 | Severe drought |
| e. | D3 | Extreme drought |
| f. | D4 | Exceptional drought |

Table 2 and Figure 5 show the Drought areas and the major keys description of the map. The U.S. Drought Monitor is jointly produced by the National Drought Mitigation Center at the University of Nebraska-Lincoln, the United States Department of Agriculture, and the National Oceanic and Atmospheric Administration. Map courtesy of NDMC.

*Dependent and Independent Variables Analysis*

The Univariate Analysis as shown in Figures 6 and 7 described each of the independent variables to determine the level of skewness and noise in the dataset.

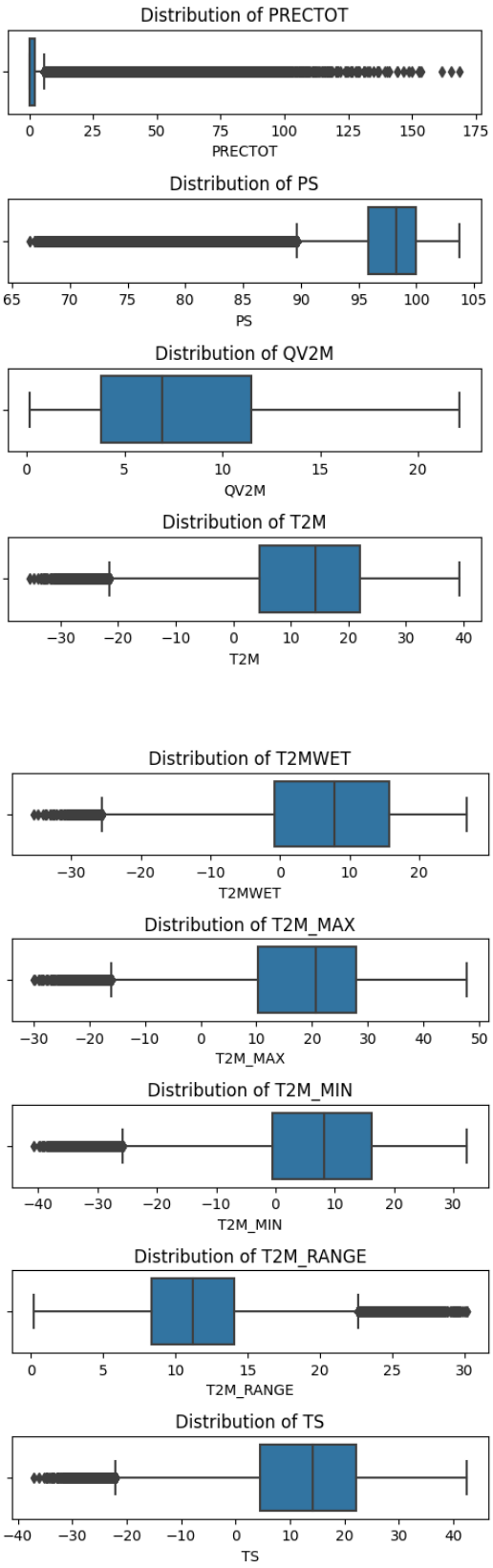



***Figure 6.*** *The univariant Features' Outlier Analysis.*

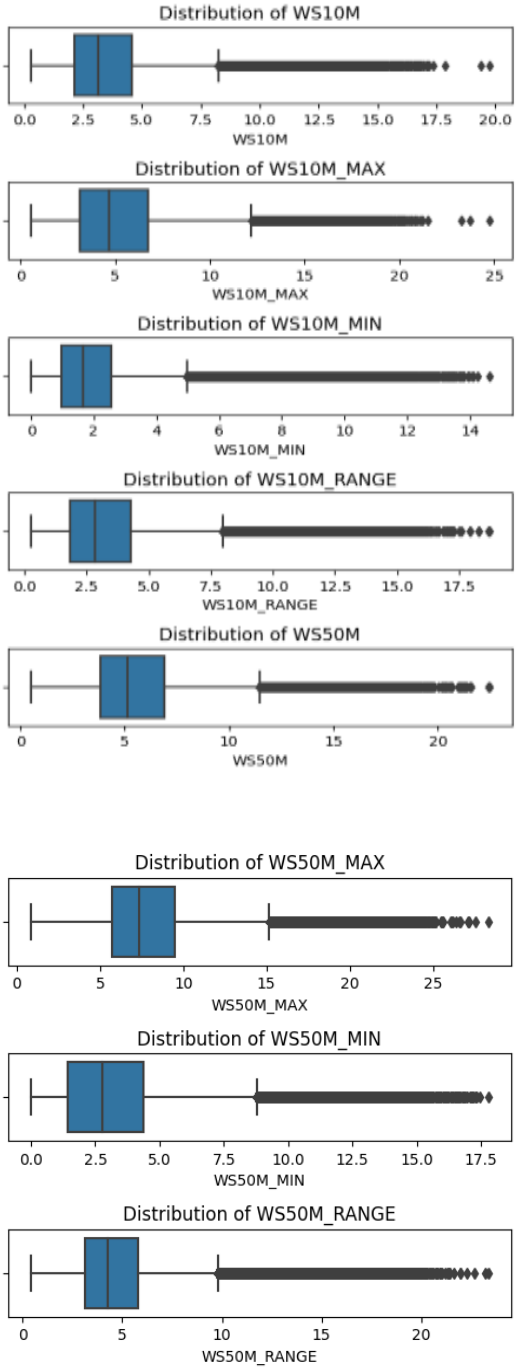



***Figure 7.*** *The Univariant Features' Outlier Analysis.*

Any data point that significantly deviates from the typical range of values within a dataset is known as outlier. These outliers can represent exceptional cases that fall well outside the norm for individual samples or entire populations. They are distinct data points that stand out due to their extreme values. Similarly, Figures 3 and 4 illustrate the status of each independent variable, serving the purpose of identifying and removing irrelevant sets or features. The dataset initially contained 2,756,796 records, among which 65,933 (2.39%) were identified as outliers for Precipitation (mm day-1) (PRECTOT). Similarly, surface pressure (kPa) (PS) exhibited

73,197 outlier records, approximately 2.65%. Conversely, Specific Humidity at 2 Meters (g/kg) (QV2M) showed exceptional consistency with only one outlier, which was extremely impressive. Temperature at 2 Meters (C) (T2M) presented 4,531 (0.16%) outlier records, indicating some variability. Dew/Frost Point at 2 Meters (C) (T2MDEW) demonstrated a fair representation, with 2,023 (~0.07%) outlier records identified. Wet Bulb Temperature at 2 Meters (C) (T2MWET) displayed 1,814 (0.06%) outlier records, a marginal proportion compared to the dataset's volume. Maximum Temperature at 2 Meters (C) (T2M_MAX) showed 3,384 (~0.12%) outlier records, indicating a fair representation as well. In contrast, the Minimum Temperature at 2 Meters (C) (T2M_MIN) exhibited 6,944 (~0.25%) outliers. Temperature Range at 2 meters (C) (T2M_RANGE) indicated 3,628 (~0.13%) standard outliers. Earth Skin Temperature (C) (TS) contained 4,762 (~0.17) outlier values, which were negligible in the context of the entire dataset.

Upon examination of Wind Speed at 10 Meters (m/s) (WS10M), 29,954 (~1.08%) records were identified as outliers. Maximum Wind Speed at 10 Meters (m/s) (WS10M_MAX) recorded 23,387 (~0.84%) outliers, while Minimum Wind Speed at 10 Meters (m/s) (WS10M_MIN) exhibited a relatively high outlier count at 39,901, representing approximately 1.4% of the total records. Minimum Wind Speed at 10 Meters (m/s) (WS10M_RANGE) displayed 35,979 outliers, approximately 1.3% of the total records.

For Wind Speed at 50 Meters (m/s) (WS50M), 23,090 outliers were identified, constituting approximately 0.8% of the total records. Maximum Wind Speed at 50 Meters (m/s) (WS50M_MAX) exhibited 25,985 outliers, representing approximately 0.94% of the dataset. Minimum Wind Speed at 50 Meters (m/s) (WS50M_MIN) displayed 19,569 outliers, constituting approximately 0.70% of the total records.

Notably, the Wind Speed Range at 50 Meters (m/s) (WS50M_RANGE) contained 33,808 noisy data points, representing approximately 1.22% of the total records, marking it as one of the highest outlier counts in the dataset.

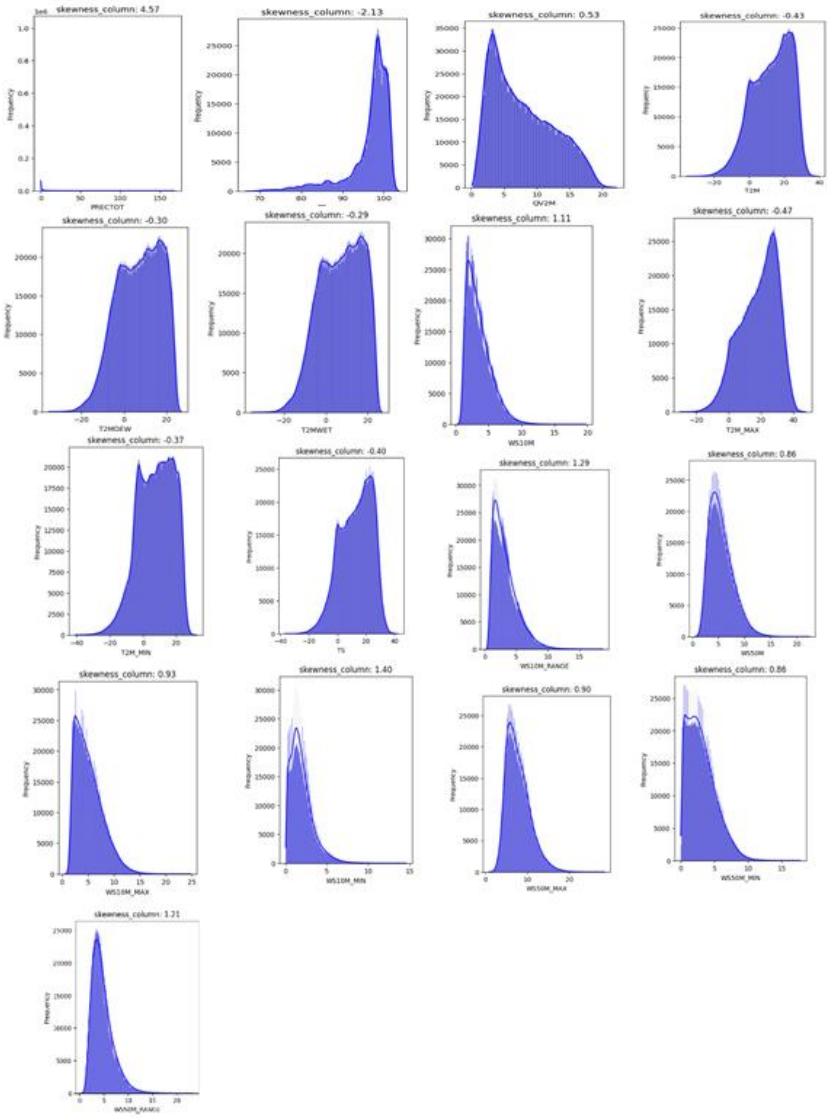


***Figure 8.*** *The Skewed Analysis of the Dependent Variables.*

Figure 8 shows a classical view of each of the variable's skewness after the elimination of noise and null values.

Applying Bulmer, 1979 skewness rules to Figure 8, the variables PRECTOT (4.56), WS10M (1.11), WS10M_MIN (1.40), WS10M_MAX (1.28), and WS50M_RANGE (1.2) were highly skewed positively, and this can impact the accuracy of the model's prediction. Similarly, QV2M, T2M_RANGE, WS10M_RANGE, WS50M, WS50M_MAX, and WS50M_MIN were moderately positively skewed with various degrees as follows: 0.52, 0.09, 0.93, 0.86, 0.89, 0.86, respectively. Conversely, the following independent variable can be classified to be approximately symmetric: fips (-0.07), T2M (-0.42), T2MDEW (-0.30), T2MWET (-0.28), T2M_MAX (-0.46), T2M_MIN (-0.36), TS (-0.39). The only feature that was moderately negatively skewed was PS (-2.13).

***Table 3.*** *Independent Variables' Kurtosis Analysis.*

| S/N | Feature | Kurtosis values | S/N | Feature | Kurtosis values |
|---|---|---|---|---|---|
| 1. | fips | -1.10 | 13. | WS10M_MAX | 0.70 |
| 2. | PRECTOT | 33.30 | 14. | WS10M_MIN | 3.15 |
| 3. | PS | 4.81 | 15. | WS10M_RANGE | 2.08 |
| 4. | QV2M | -0.78 | 16. | WS50M | 0.81 |
| 5. | T2M | 0.55 | 17. | WS50M_MAX | 0.98 |
| 6. | T2MDEW | -0.73 | 18. | WS50M_MIN | 0.59 |
| 7. | T2MWET | -0.75 | 19. | WS50M_RANGE | 2.20 |
| 8. | T2M_MAX | -0.50 | 20. | Score | 1.38 |
| 9. | T2M_MIN | -0.44 | 21. | Year | 1.20 |
| 10. | T2M_RANGE | -0.31 | 22. | month | 1.20 |
| 11. | TS | -0.53 | 23. | day | 1.19 |
| 12. | WS10M | 1.41 | | | |

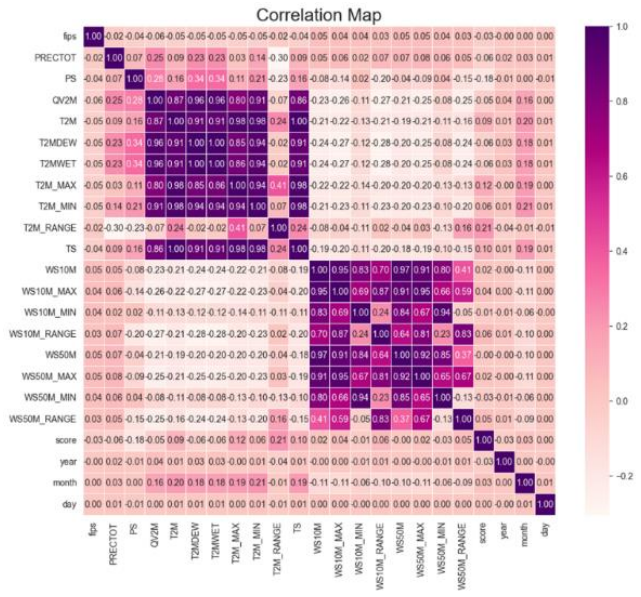



***Figure 9.*** *Correlation Plot For Feature Selection.*

Further investigation into the data from Table 3 revealed a state where the distribution of classes or categories within the dataset is highly skewed, with one or more classes being significantly more prevalent than others. The Kurtosis values of the data show how tailed to the right or left the data is, which gives more information on outliers in the dataset.

Figure 9 shows the visualization of the correlation plot; this plot shows are how closely related correlated are the variables; the closer to 1, the darker the shade, the more correlated the features are. As illustrated in Figure 9, the features to be selected for the experiments were plotted against each other to calculate the degree of mathematical relationship known as correlation. This aspect of preprocessing is crucial to building a good model that can be generalized easily. All correlation coefficients of 1.0 to 0.8 indicate that the correlation was very strongly positive, such as Specific Humidity at 2 Meters (g/kg) (QV2M), Temperature at 2 Meters (C) (T2M), Dew/Frost Point at 2 Meters (C) (T2MDEW), Maximum Temperature at 2 Meters (C) (T2M_MAX), Wet Bulb Temperature at 2 Meters (C) (T2MWET), Minimum Temperature at 2 Meters (C) (T2M_MIN), Temperature Range at 2 meters (C)

(T2M_RANGE) and Earth Skin Temperature (C) (TS).

A similar very strong correlation was also found among the following features: Wind Speed at 10 Meters (m/s) (WS10M), Maximum Wind Speed at 10 Meters (m/s) (WS10M_MAX), Minimum Wind Speed at 10 Meters (m/s) (WS10M_MIN), Minimum Wind Speed at 10 Meters (m/s) (WS10M_RANGE), Wind Speed at 50 Meters (m/s) (WS50M), Maximum Wind Speed at 50 Meters (m/s) (WS50M_MAX), Minimum Wind Speed at 50 Meters (m/s) (WS50M_MIN) and Wind Speed Range at 50 Meters (m/s) (WS50M_RANGE).

However, there were some with weak positive correlations, such as PRECTOT and QV2M, T2MDEW, T2M, T2MWET, and T2M_MIN. However, most of the temperature features were negatively and weakly correlated with the wind features.

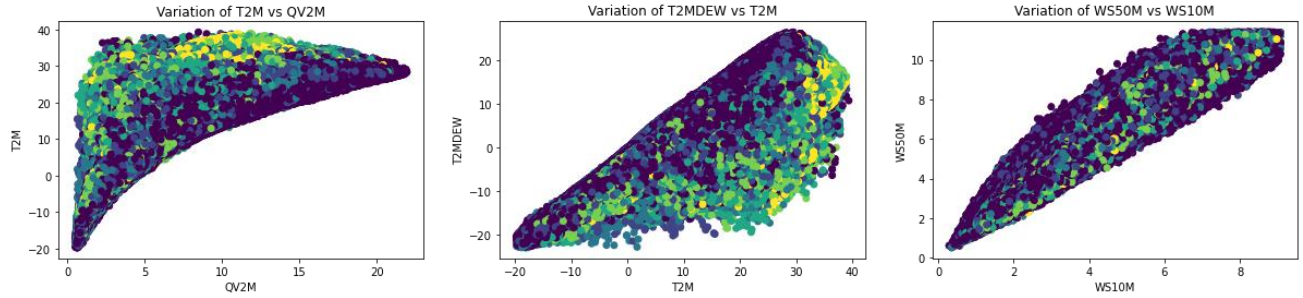



***Figure 10.*** *The Bi-variance Correlation between some Dependent Variables.*

Attributes QV2M, T2M, T2MDEW, T2MWET, T2M_MAX, T2M_MIN and TS have shown strong positive correlation. Similary, WS10M, WS10M_MAX and WS10M_MIN have shown a strong positive correlation. Likewise, WS50M, WS50M_MAX and WS50M_MIN show strong positive correlation. However, the scatter plots above show significant variance between the data points despite the strong positive correlation. Hence, we'll retain all these variables and try other feature selection methods.

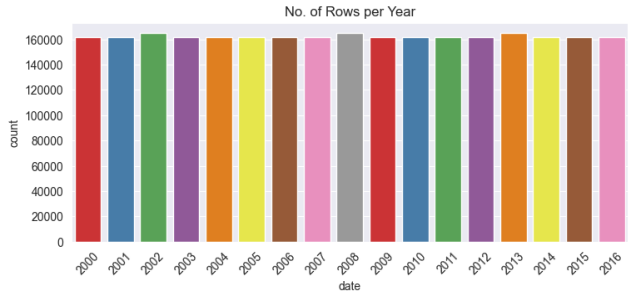



***Figure 11.*** *An Overview of Numbers of Tuples Across 17 Year.*

## 3.2. Model Building

Following essential data exploration and preprocessing steps, the dataset underwent systematic division into three distinct subsets to facilitate the development and evaluation of the machine learning model. The training set assumed a pivotal role, comprising data dedicated solely to training the model, enabling it to grasp patterns and relationships within the dataset. Concurrently, the validation set emerged as a crucial tool for fine-tuning hyperparameters and evaluating the model's performance during training, thereby mitigating overfitting risks. The test set remained isolated throughout model development and was exclusively reserved for final evaluation. It served as a litmus test for the model's generalization capabilities, offering insights into its performance on previously unseen data. To ensure the model's training on past data and evaluation on future data, the division of training, validation, and test sets occurred over time using the date feature, as seen in Figures 10 and 11.

Additional analysis was conducted on the dataset, where models were constructed using the ensemble algorithm, specifically the Extreme Gradient Boosting Algorithm, following the procedures outlined in Algorithm 1. Furthermore, a hybridized model combining K-means++ and the Gradient Boosting Algorithm(KGBA) was developed with a Principal Component Analysis function for dimensionality reduction, as detailed in Figure 11. The Performance Metrics were calculated as an average percentage of test options (Training set and Testing set) used to run the selected ensemble algorithm (Gradient Boosting Algorithm).

One very important thing in K-means++ is choosing the number of clusters to use, which is the K. In this study, the elbow method was employed to help determine the proper number of clusters for K-means++ clustering. Figure 12 shows the plot for the elbow method, and this was done by plotting the distortion against the number of clusters. By visually inspecting the plot in Figure 12 and identifying the point at which additional clusters bring diminishing returns in terms of reducing distortion. The distortion measures how evenly distributed the data points are inside each cluster. Better grouping is indicated by lower distortion. The plot assists in determining the appropriate number of clusters by finding the point of inflection or "elbow" in the plot when the distortion improvement is greatly reduced. As indicated

in Figure 12, three points with cluster numbers 3, 4 and 5 were seen as the point of inflection, so this research work tested all three points to see which had a better performance and cluster number 5 was concluded to have the best performance and was chosen as the K, which is the K needed for the implementation of the K-means++.

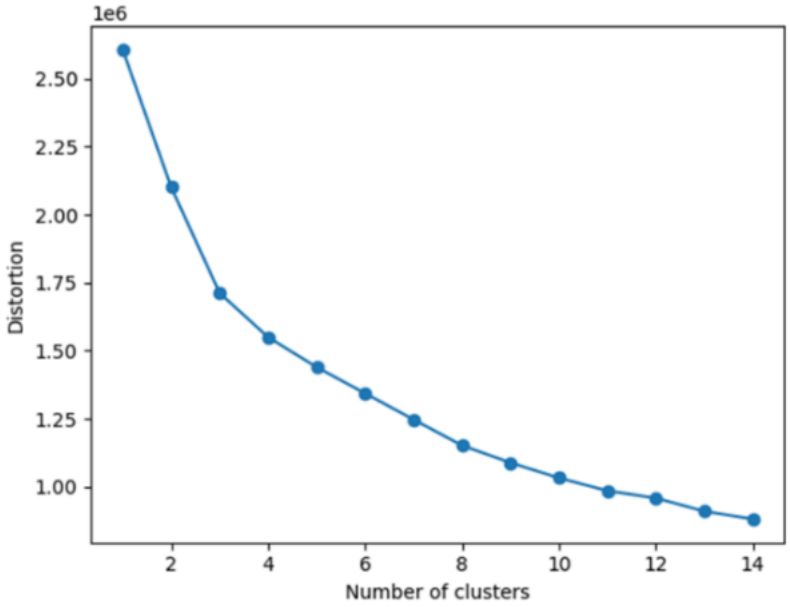



***Figure 12.*** *The Elbow Method To Determine Number of Clusters.*

The scaled feature matrix X_scaled is subjected to Principal Component Analysis (PCA) and the data is transformed into the new lower-dimensional space denoted by X_pca using the fit_transform() method after computing the principal components using the PCA() function from scikit-learn.

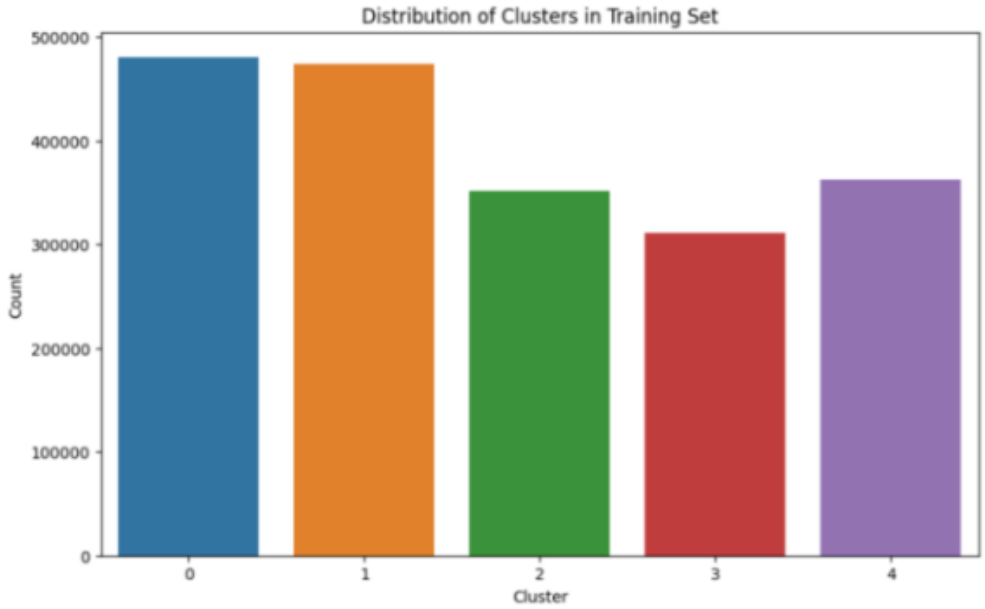



***Figure 13.*** *Training Dataset Distribution After PCA and K++Mean According To Each Class.*

Figure 13 illustrates the distributions of the dataset for model training. The label class 0,1,2,3,4 had about 490,000, 490,000, 350,000, 280,000, and 360,000 records, respectively.

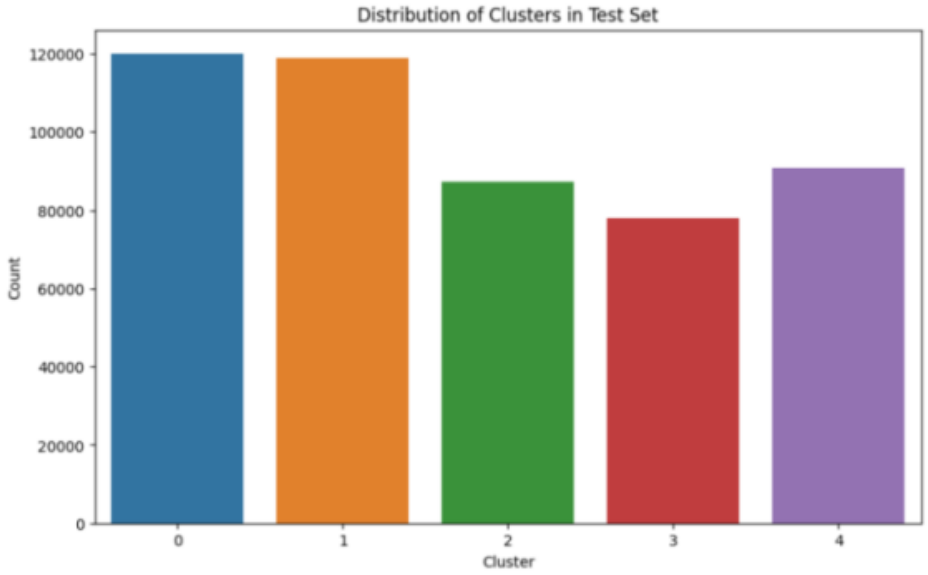



***Figure 14.*** *Test Dataset Distribution After PCA and K++Mean According To Each Class.*

The distributions of the dataset during the testing is depicted in Figure 14. The class 0,1,2,3,4 of the label had about 120,000, 120,000, 90,000, 79,000 and 95,000 records, respectively.

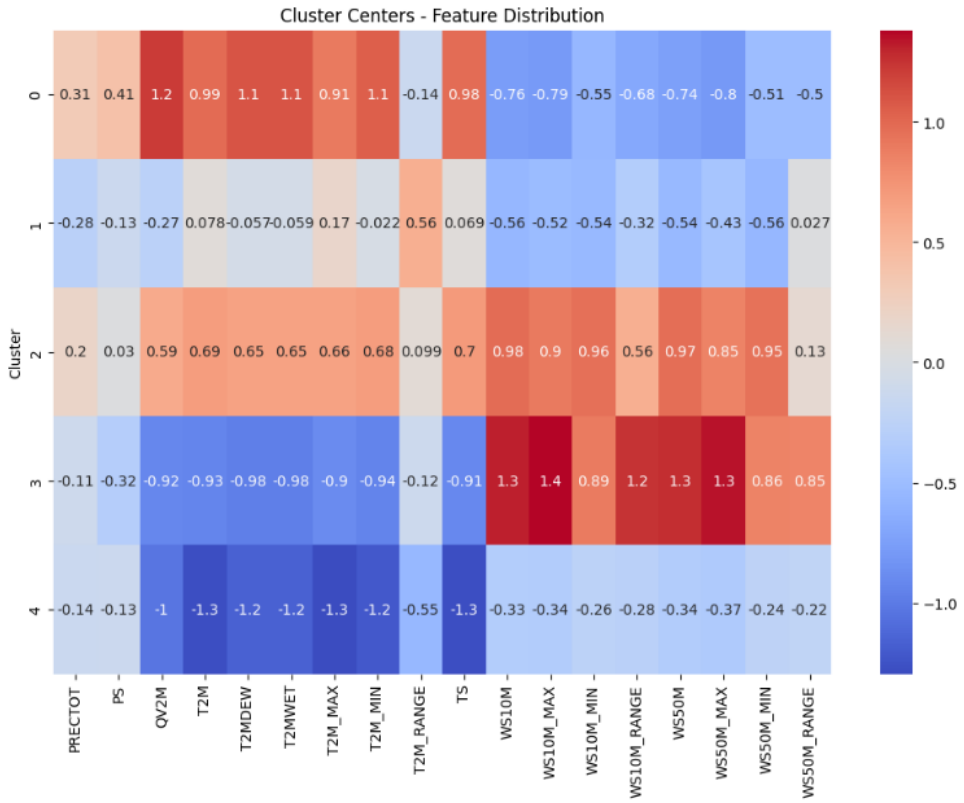



***Figure 15.*** *Data Clusters During PCA.*

The modified data X_pca, which contains the principal components derived from PCA, is subjected to K-Means++ clustering, and the data were clustered using the K-Means++ algorithm into n_clusters (in this case, 5 separate clusters) as depicted in Figure 15. The fit_predict() method is used to compute the clusters for the data once the KMeans++ function from scikit-learn is used to generate a K-Means++ object. The clusters array contains the assigned clusters as a result. A new

feature is then added to the DataFrame X named "kmeans_scaled" that holds the cluster designations for each data point after applying K-Means++. Afterwards, the machine learning model uses this new feature as an additional input.

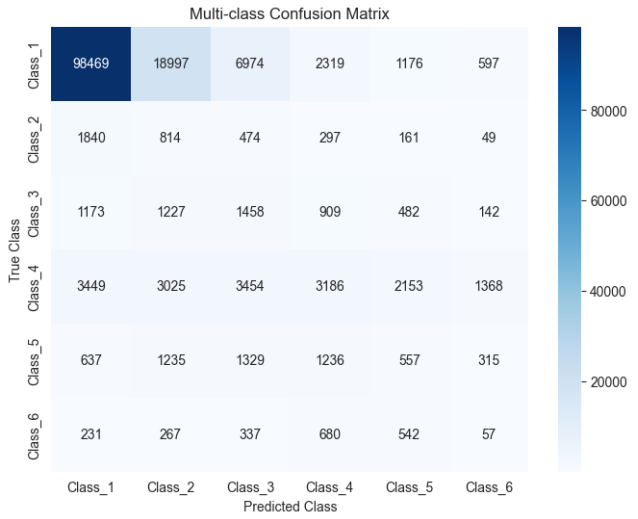



***Figure 16.*** *Confusion matrix for K-Means++ Gradient Boosting Algorithm (KGBA).*

Data exploration and analysis were performed on the dataset, which includes dimensionality reduction using Principal Component Analysis and clustering using K-means++ as shown in Figures 13, 14, and 15. The implementation was carried out on the virtual machine of kaggle.com running Graphics Processing Unit (GPU) with Python Scikit-learn libraries. Figure 16 shows the confusion matrix for the model developed in this study, K-means++ Gradient Boosting Algorithm. This was done to assess the model's performance on each class independently, providing insights into how well it handles minority classes. This confusion matrix was used to generate a classification report in Table 4 to gain more insight into the performance of the model.

The explorations began with the first phase, which was the data preprocessing stage. This stage of data transformation allows the machine to easily interpret and use the data. In the second phase, Principal Component Analysis (PCA) was used to reduce dimensionality. At the end of the operation the following features were considered viable for building the model: PRECTOT, PS, QV2M, T2M, T2MDEW, T2MWET, T2M_MAX, T2M_MIN, T2M_RANGE, TS, WS10M, WS10M_MAX, WS10M_MIN, WS10M_RANGE, WS50M, WS50M_MAX, WS50M_MIN, WS50M_RANGE. It reduced the number of features while preserving the crucial information in the data by focusing on the most significant variances, as illustrated in Figures 10 & 11. To improve the performance of the model, the K-means++ was used to separate the points into different clusters. In the third phase, the clustered points were added as a new feature and the Gradient Boosting Algorithm (GBA) was used to build the prediction model with a weather dataset. The machine learning standard metrics were used for the evaluation to know how the new model K-Means++ Gradient Boosting Algorithm (KMGBA) fares.

## 4. Result and Discussion

The process of data exploration was completed in section 3.0, during which the preprocessing activities were strictly carried out. The models were trained using different algorithms, among which are Gradient Boosting (GB), K-Nearest Neighbour(KNN), Random Forest (RF) and Naïve Bayes, and the new algorithm, which is the combination of K-Nearest Neighbour (KNN++) and Gradient Boosting (KGBA). The 5 levels of drought (D0: Abnormal dry, D1: Moderate drought, D2: Severe drought, D3: Extreme drought and D4: Exceptional drought) in the dataset were predicted, and standards metrics were used, such as Recall, Precision, F1Score and accuracy to compare the performance of the models. The outcome of the models can be viewed in Tables 5-8. Although emphasis was laid on the recall due to the nature of the dataset.

***Table 4.*** *KXGBA Classification Report for All the Classes.*

| | precision | recall | f1-score | Support |
|---|---|---|---|---|
| Class 0 | 0.57 | 0.50 | 0.53 | 10904 |
| Class 1 | 0.56 | 0.23 | 0.33 | 10737 |
| Class 2 | 0.69 | 0.16 | 0.26 | 10880 |
| Class 3 | 0.28 | 0.34 | 0.31 | 10932 |
| Class 4 | 0.33 | 0.72 | 0.45 | 10890 |
| Class 5 | 0.75 | 0.78 | 0.76 | 10803 |
| Accuracy | | | 0.46 | 65146 |
| macro avg | 0.53 | 0.46 | 0.44 | 65146 |
| weighted avg | 0.53 | 0.46 | 0.44 | 65146 |

Table 4 presents the performance results of the K-means++ Gradient Boosting Algorithm. The algorithm achieved a precision of approximately 57%, indicating that out of the 10,904 instances predicted as the "no drought" (Class 0), roughly 6,215 were accurate positive predictions of all instances. However, the remaining 43% of the records were erroneously classified as positive.

The recall value for this algorithm was lower, standing at around 50%. This implies that the algorithm correctly identified only half of the actual "no drought" instances in the dataset.

When assessing the records classified as Abnormal drought (Class 1), approximately 56% of the 10,737 instances were

accurately identified as positive (true positives) out of all instances predicted as positive, encompassing both true positives and false positives. This corresponds to 6,010 instances within the class. However, the recall prediction was less promising, indicating only a 23% accurate prediction of actual positive instances of Abnormal drought.

The precision value for the Moderate drought (Class 2) was notably high, standing at 69%. This indicates that out of all instances predicted as positive by the model, 7,507 were indeed positive out of the total 10,880 instances. However, the recall value for this class is strikingly low compared to the precision prediction. This suggests that only 16% of the actual moderate drought instances were correctly identified as moderate drought this may be due to the skewness of the dataset.

For the Severe drought (Class 3), the model achieved a precision of 28%, indicating that approximately 28% of the instances predicted as positive were accurate predictions out of the total instances. Conversely, the recall value was recorded at 34%, suggesting that only 34% of the actual severe drought instances were correctly identified by the model, totaling 3,717 of the actual instances out that were actual severe drought.

Similarly, for the Extreme drought (Class 4), the precision prediction value was 33%, indicating that around 33% of the total instances predicted as positive were accurate predictions, with a false positive prediction rate of 67%. However, the recall value was notably higher, standing at 72%, indicating that the model accurately predicted the presence of extreme drought in approximately 7,840 cases.

The exceptional drought (Class 5) was predicted with 75% precision, an indication of 8,102 postive instances correctly predicted from 10,803 instances. The 25% of the total records were erroneously classified as positive. In similar way, the recall value for the same exceptional drought was high. The 78% of instances were accurately predicted of the actual exceptional drought cases. It signifies that the model wrongly classified 22% of this class.

The overall accuracy of the model in the prediction of this multiple classes is 46% because the model failed to perform well in some of the classes probably because of low percentage of representation.

***Table 5.*** *GBA Classification Report For All the Classes.*

| Class Name | precision | recall | f1-score | support |
|---|---|---|---|---|
| Class 0 | 0.56 | 0.49 | 0.52 | 10904 |
| Class 1 | 0.53 | 0.23 | 0.32 | 10737 |
| Class 2 | 0.68 | 0.15 | 0.25 | 10880 |
| Class 3 | 0.28 | 0.35 | 0.31 | 10932 |
| Class 4 | 0.33 | 0.72 | 0.45 | 10890 |
| Class 5 | 0.74 | 0.77 | 0.76 | 10803 |
| accuracy | | | 0.45 | 65146 |
| macro avg | 0.52 | 0.45 | 0.43 | 65146 |
| weighted avg | 0.52 | 0.45 | 0.43 | 65146 |

Table 5 presents the classification report depicting the performance metrics for Drought Prediction utilizing the Traditional Gradient Boosting Algorithm. The model's effectiveness in predicting drought situations across different classes is assessed based on precision, recall, and F1-scores.

The performance metrics for this model are as follows: Class 0 achieved a precision of 56%, a recall of 49%, and an F1-score of 52%, with a support of 10,904 instances. For Class 1, the precision was 53%, the recall was 23%, and the F1-score was 32%, with a support of 10,737 instances. Class 2 exhibited a precision of 68%, a recall of 15%, and an F1-score of 25%, with a support of 10,880 instances. Class 3 had a precision of 28%, a recall of 35%, and an F1-score of 31%, with a support of 10,932 instances. Class 4 displayed a precision of 33%, a recall of 72%, and an F1-score of 45%, with a support of 10,890 instances. Class 5 demonstrated a precision of 74%, a recall of 77%, and an F1-score of 76%, with a support of 10,803 instances. The overall accuracy of the model was 45%, calculated over a total of 65,146 instances.

***Table 6.*** *Random Forest Model's Performance across all the classes of Drought.*

| Class Name | precision | Recall | f1-score | Support |
|---|---|---|---|---|
| Class 0 | 0.6095 | 0.6039 | 0.6067 | 10904 |
| Class 1 | 0.4020 | 0.3868 | 0.3943 | 10737 |
| Class 2 | 0.2748 | 0.2707 | 0.2727 | 10880 |
| Class 3 | 0.1716 | 0.1701 | 0.1709 | 10932 |
| Class 4 | 0.1565 | 0.1680 | 0.1620 | 10890 |
| Class 5 | 0.8606 | 0.8578 | 0.8592 | 10803 |
| accuracy | | | 0.4089 | 65146 |
| macro avg | 0.4125 | 0.4095 | 0.4110 | 65146 |
| weighted avg | 0.4118 | 0.4089 | 0.4103 | 65146 |

Table 6, presents the classification report showcasing the performance metrics for Drought Prediction with Random Forest model. The classification model's performance metrics are detailed as follows: For Class 0, a precision of 60%, recall of 60%, and F1-score of 60% were achieved, with a support of 10,904 instances. Class 1 exhibited a precision of 40%, recall of 38%, and F1-score of 39%, with a support of 10,737 in-

stances. Class 2 demonstrated a precision of 27%, recall of 27%, and F1-score of 27%, with a support of 10,880 instances. For Class 3, precision was 17%, recall was 17%, and F1-score was 17%, with a support of 10,932 instances. Class 4 displayed a precision of 15%, recall of 16%, and F1-score of 16%, with a support of 10,890 instances. Finally, Class 5 showed a precision of 86%, recall of 85%, and F1-score of 85%, with a support of 10,803 instances. The overall accuracy of the model was 40%, calculated over a total of 65,146 instances.

***Table 7.*** *KNN Classification Report For All the Classes.*

| Class Name | precision | recall | f1-score | support |
|---|---|---|---|---|
| Class 0 | 0.3554 | 0.4534 | 0.3985 | 10904 |
| Class 1 | 0.2651 | 0.2493 | 0.2570 | 10737 |
| Class 2 | 0.1807 | 0.1833 | 0.1820 | 10880 |
| Class 3 | 0.1472 | 0.1452 | 0.1462 | 10932 |
| Class 4 | 0.2031 | 0.1811 | 0.1915 | 10890 |
| Class 5 | 0.6855 | 0.6098 | 0.6455 | 10803 |
| accuracy | | | 0.3033 | 65146 |
| macro avg | 0.3062 | 0.3037 | 0.3034 | 65146 |
| weighted avg | 0.3057 | 0.3033 | 0.3030 | 65146 |

Table 7 provides the classification report detailing the performance metrics for Drought Prediction using the K-Nearest Neighbor (KNN) algorithm. The model's ability to predict drought situations across different classes is evaluated based on precision, recall, and F1-scores.

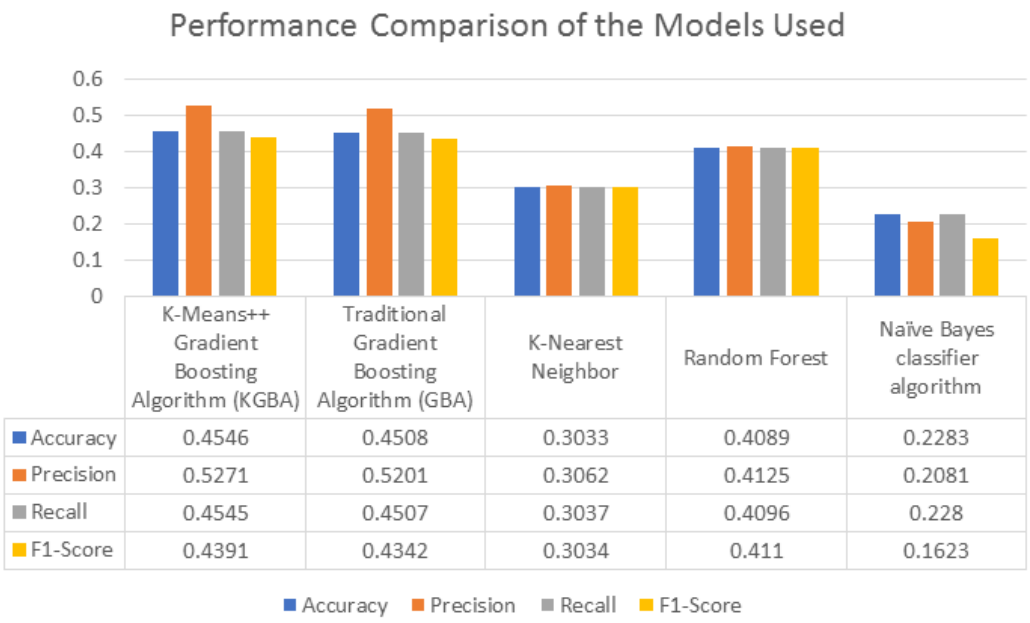
The model's performance are summarized as follows: For Class 0, a precision of 35%, recall of 45%, and F1-score of 39% were achieved, with a support of 10,904 instances. Class 1 exhibited a precision of 26%, recall of 24%, and F1-score of 25%, with a support of 10,737 instances. Class 2 demonstrated a precision of 18%, recall of 18%, and F1-score of 18%, with a support of 10,880 instances. For Class 3, precision was 14%, recall was 14%, and F1-score was 14%, with a support of 10,932 instances. Class 4 displayed a precision of 20%, recall of 18%, and F1-score of 19%, with a support of 10,890 instances. Finally, Class 5 showed a precision of 68%, recall of 60%, and F1-score of 64%, with a support of 10,803 instances. The overall accuracy of the model was 30%, calculated over a total of 65,146 instances.

Table 8 is the classification report which shows the performance metrics for Drought Prediction using the Naïve Bayes Classifier. Table 8 presents the classification report detailing the performance metrics for Drought Prediction. The model's ability to predict drought situations across diverse classes is assessed through precision, recall, and F1-scores.

The results are as follows: For Class 0, a precision of 0.2647, recall of 0.1045, and F1-score of 0.1499 were achieved, with a support of 10,904 instances. Class 1 exhibited a precision of 0.1584, recall of 0.0392, and F1-score of 0.0629, with a support of 10,737 instances. Class 2 demonstrated a precision of 0.1171, recall of 0.0342, and F1-score of 0.0529, with a support of 10,880 instances. For Class 3, precision was 0.0639, recall was 0.0369, and F1-score was 0.0467, with a support of 10,932 instances. Class 4 displayed a precision of 0.2375, recall of 0.9366, and F1-score of 0.3789, with a support of 10,890 instances. Finally, Class 5 showed a precision of 0.4070, recall of 0.2164, and F1-score of 0.2826, with a support of 10,803 instances. The overall accuracy of the model was 0.2283, calculated over a total of 65,146 instances.

***Table 8.*** *Naive Bayes Classifier Classification Report For All the Classes.*

| Class Name | precision | recall | f1-score | support |
|---|---|---|---|---|
| Class 0 | 0.2647 | 0.1045 | 0.1499 | 10904 |
| Class 1 | 0.1584 | 0.0392 | 0.0629 | 10737 |
| Class 2 | 0.1171 | 0.0342 | 0.0529 | 10880 |
| Class 3 | 0.0639 | 0.0369 | 0.0467 | 10932 |
| Class 4 | 0.2375 | 0.9366 | 0.3789 | 10890 |
| Class 5 | 0.4070 | 0.2164 | 0.2826 | 10803 |
| accuracy | | | 0.2283 | 65146 |
| macro avg | 0.2081 | 0.2280 | 0.1623 | 65146 |
| weighted avg | 0.2079 | 0.2283 | 0.1623 | 65146 |




| | K-Means++ Gradient Boosting Algorithm (KGBA) | Traditional Gradient Boosting Algorithm (GBA) | K-Nearest Neighbor | Random Forest | Naïve Bayes classifier algorithm |
|---|---|---|---|---|---|
| Accuracy | 0.4546 | 0.4508 | 0.3033 | 0.4089 | 0.2283 |
| Precision | 0.5271 | 0.5201 | 0.3062 | 0.4125 | 0.2081 |
| Recall | 0.4545 | 0.4507 | 0.3037 | 0.4096 | 0.228 |
| F1-Score | 0.4391 | 0.4342 | 0.3034 | 0.411 | 0.1623 |

***Figure 17.*** *Performance Comparison of the model used.*

The performance of various machine learning algorithms for prediction of drought can be compared based on their accuracy, precision, recall, and F1-score metrics. The comparative analysis of the algorithms mentioned can simply be summaried further from Figure 17 as follows:

Accuracy: K-Means++ Gradient Boosting Algorithm (KGBA) achieved the highest accuracy of approximately 46% across all the classes of drought levels, followed closely by Traditional Gradient Boosting Algorithm (GBA) and K-Nearest Neighbor algorithm, both with an accuracy of 45%. Random Forest algorithm had an accuracy of 41%, while Naïve Bayes classifier algorithm exhibited the lowest accuracy of 23%.

Precision: KGBA also outperformed other algorithms in terms of precision, with a precision value of ~53%. GBA followed closely with a precision of ~52%, while Random Forest algorithm had a precision of ~41%. K-Nearest Neighbor algorithm exhibited lower precision of 31%, and Naïve Bayes classifier algorithm had the lowest precision of 21%.

Recall: KGBA and GBA had similar recall values, both around ~45%. Random Forest algorithm and K-Nearest Neighbor algorithm also exhibited comparable recall values, approximately ~41% and ~30% respectively. Naïve Bayes classifier algorithm had the lowest recall value of ~23%.

F1-score: KGBA had the highest F1-score of ~44%, followed closely by GBA with a score of ~43%. Random Forest algorithm and K-Nearest Neighbor algorithm had similar F1-scores, around ~41% and 30% respectively. Naïve Bayes classifier algorithm had the lowest F1-score of ~16%.

In summary, comparing the performance of these machine learning algorithms, it's evident that K-Means++ Gradient Boosting Algorithm (KGBA) achieved the highest accuracy, precision, recall, and F1-score among the algorithms considered. Following closely behind KGBA was the Traditional Gradient Boosting Algorithm (GBA), which exhibited similar performance across all metrics. Random Forest algorithm showed moderate performance, while K-Nearest Neighbor algorithm demonstrated relatively lower precision and recall values. Notably, Naïve Bayes classifier algorithm exhibited the lowest performance across all metrics, with significantly lower accuracy, precision, recall, and F1-score values compared to the other algorithms. In summary, KGBA and GBA performed the best overall, followed by Random Forest, K-Nearest Neighbor, and finally, Naïve Bayes classifier algorithm.

KGBA demonstrated the best overall performance across all metrics, followed closely by GBA. Random Forest algorithm performed moderately well, while K-Nearest Neighbor algorithm showed relatively lower performance. Naïve Bayes classifier algorithm exhibited the lowest performance among the algorithms evaluated.

## 5. Conclusion

In this study, the K-means++ Gradient Boosting Algorithm (KGBA) was introduced to enhance the performance of the Gradient Boosting Algorithm (GBA) in classifying drought-prone areas. Through comparisons with standard models like Random Forest and Naïve Bayes classifiers, the study evaluated KGBA's efficacy in drought prediction using criteria such as accuracy, precision, recall, and F1-Score. Results showed KGBA's slight advantage over GBA in forecasting drought likelihood based on historical data, highlighting its potential for improving predictive capabilities in drought mitigation.

By integrating K-means++ into GBA, KGBA aimed to better identify and categorize drought-prone regions, offering a more precise model for drought prediction in line with Jiang and Luo as suggested in the article. Ensemble models like

KGBA and Random Forest performed notably well, particularly in predicting extreme drought seasons [11]. Incorporating advanced techniques such as K-means++ and Principal Component Analysis during feature engineering could further enhance prediction accuracy. These findings support previous research by Likas et al., endorsing the effectiveness of the K-means algorithm in clustering for both solution quality and computational complexity considerations [24].

## 6. Recommendation

Further research can be conducted to validate the KGBA model's usefulness in multiple circumstances by applying it to other geographical regions and datasets. Furthermore, the model can be improved by taking into account additional environmental and socioeconomic elements that influence drought occurrence, allowing for a more complete knowledge and prediction of drought events.

Overall, the KGB model represents an encouraging step forward in drought prediction, with potential benefits for resource management, agriculture, and water conservation activities.

## Abbreviations

| | |
|---|---|
| CC | Climate Change |
| MLM | Machine Learning Model |
| RF | Random Forest |
| GBA | Gradient Boosting Algorithm |
| KGBA | K-means++ Clustering and Gradient Boosting Algorithm |
| XGBA | Extreme Gradient Boosting Algorithm |
| LightGBA | Light Gradient Boosting Algorithm |
| PCA | Principal Component Analysis |
| ML | Machine Learning |
| SVM | Support Vector Machine |
| ANN | Artificial Neural Networks |
| ELM | Extreme Learning Machine (ELM) |
| SPI | Standardized Precipitation |
| I | Drought Index |
| Ped | Ped Index |
| SWAT | Soil and Water Assessment Tool |
| NDMC | National Drought Mitigation |

## Author Contributions

**Babatunde Isaiah Ayinla:** Conceptualization, Methodology, Project administration, Supervision, Writing – review & editing

**Rasheedat Aderonke Abdulsalam:** Data curation, Investigation, Software, Visualization, Writing – original draft

## Conflicts of Interest

The authors declare no conflicts of interest.

## References

[1] *NOS Science Report 2021* https://oceanservice.noaa.gov/about/nos-science-report/2021/ accessed 02 May 2023.

[2] Mortuza, M. R., Moges, E., Demissie, Y., & Li, H. Y. (2019). Historical and future drought in Bangladesh using copula-based bivariate regional frequency analysis. *Theoretical and Applied Climatology*, *135*(3–4), 855–871. https://doi.org/10.1007/s00704-018-2407-7

[3] Khan, N., Sachindra, D. A., Shahid, S., Ahmed, K., Shiru, M. S., & Nawaz, N. (2020). Prediction of droughts over Pakistan using machine learning algorithms. *Advances in Water Resources*, *139*. https://doi.org/10.1016/j.advwatres.2020.103562

[4] Barua, S., Ng, A. W. M., & Perera, B. J. C. (2012). Artificial Neural Network–Based Drought Forecasting Using a Nonlinear Aggregated Drought Index. *Journal of Hydrologic Engineering*, *17*(12), 1408–1413. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000574

[5] Ghimire, S., Deo, R. C., Downs, N. J., & Raj, N. (2019). Global solar radiation prediction by ANN integrated with European Centre for medium range weather forecast fields in solar rich cities of Queensland Australia. *Journal of Cleaner Production*, *216*, 288–310. https://doi.org/10.1016/J.JCLEPRO.2019.01.158

[6] Xiang, B., Lin, S. J., Zhao, M., Johnson, N. C., Yang, X., & Jiang, X. (2019). Subseasonal Week 3–5 Surface Air Temperature Prediction During Boreal Wintertime in a GFDL Model. *Geophysical Research Letters*, *46*(1), 416–425. https://doi.org/10.1029/2018GL081314

[7] Yang, T., Zhou, X., Yu, Z., Krysanova, V., & Wang, B. (2015). Drought projection based on a hybrid drought index using Artificial Neural Networks. *Hydrological Processes*, *29*(11), 2635–2648. https://doi.org/10.1002/HYP.10394

[8] Jolliffe, I. T. (2002). Principal component analysis for special types of data (pp. 338-372). *Springer*, New York. https://doi.org/10.1007/b98835

[9] Sidak, K. (2023, December). *Overview of Principal Component Analysis (PCA)*. https://codefinity.com/blog/Overview-of-Principal-Component-Analysis-(PCA) accessed 02 May 2023.

[10] Mokhtar, A., Jalali, M., He, H., Al-Ansari, N., Elbeltagi, A., Alsafadi, K., Abdo, H. G., Sammen, S. S., Gyasi-Agyei, Y., & Rodrigo-Comino, J. (2021). Estimation of SPEI Meteorological Drought Using Machine Learning Algorithms. *IEEE Access*, *9*, 65503–65523. https://doi.org/10.1109/ACCESS.2021.3074305

[11] Jiang, W., & Luo, J. (2021). *An Evaluation of Machine Learning and Deep Learning Models for Drought Prediction using Weather Data*. https://doi.org/10.3233/JIFS-212748

[12] Gan, T. Y., Ito, M., Hülsmann, S., Qin, X., Lu, X. X., Liong, S. Y., Rutschman, P., Disse, M., & Koivusalo, H. (2016). Possible climate change/variability and human impacts, vulnerability of drought-prone regions, water resources and capacity building for Africa. *Hydrological Sciences Journal*, *61*(7), 1209–1226. https://doi.org/10.1080/02626667.2015.1057143

[13] Ayinla, B., & Akinola, S. O. (2021). An Improved Collaborative Pruning Using Ant Colony Optimization and Pessimistic Technique of C5.0 Decision Tree Algorithm. *Article in International Journal of Computer Science and Information Security*. https://doi.org/10.5281/zenodo.4427699

[14] Zhong, R., Chen, X., Lai, C., Wang, Z., Lian, Y., Yu, H., & Wu, X. (2019). Drought monitoring utility of satellite-based precipitation products across mainland China. *Journal of Hydrology*, *568*, 343–359. https://doi.org/10.1016/J.JHYDROL.2018.10.072

[15] Breiman, L. (1997). *ARCING THE EDGE*.

[16] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

[17] Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics and Data Analysis*, *38*(4), 367–378. https://doi.org/10.1016/S0167-9473(01)00065-2

[18] Mason, L., Bartlett, P., Baxter, J., & Frean, M. (2000). Boosting Algorithm as Gradient Descent. *Advances in Neural Information Processing Systems*, 512–518.

[19] Bent ́ejac, C., Csörgő, A., & Mart ́nez-Mu ̃noz, G. (2021). A Comparative Analysis of XGBoost. *Artificial Intelligence Review*, *54*, 1937–1967. https://doi.org/10.1007/s10462-020-09896-5

[20] Friedman, J., Hastie, T., & Tibshirani, R. (2000). ADDITIVE LOGISTIC REGRESSION: A STATISTICAL VIEW OF BOOSTING. In *The Annals of Statistics* (Vol. 28, Issue 2).

[21] Breiman, L. (2001). Random forests. *Kluwer Academic Publishers*, Netherlands *45*(1), 5–32.

[22] Luo, H., Bhardwaj, J., Choy, S., & Kuleshov, Y. (2022). Applying Machine Learning for Threshold Selection in Drought Early Warning System. *Climate*, *10*(7). https://doi.org/10.3390/cli10070097

[23] Felsche, E., & Ludwig, R. (n.d.). *Applying machine learning for drought prediction using data from a large ensemble of climate simulations*. https://doi.org/10.5194/nhess-2021-110

[24] Likas, A., Vlassis, N., & Verbeek, J. (n.d.). *The global k-means clustering algorithm The global k-means clustering algorithm. [Technical*. https://hal.inria.fr/inria-00321515

[25] Tri, D. Q., Dat, T. T., & Truong, D. D. (2019). Application of meteorological and hydrological drought indices to establish drought classification maps of the Ba River basin in Vietnam. *Hydrology*, *6*(2). https://doi.org/10.3390/hydrology6020049

[26] Christoph, M. (2021, July 23). *Predict Droughts using Weather & Soil Data*. https://www.kaggle.com/datasets/cdminix/us-drought-meteorological-data accessed 18 May 2023

[27] Nitin. (2020, April 22). *LightGBM Binary Classification, Multi-Class Classification, Regression using Python*. Https://Nitin9809.Medium.Com/Lightgbm-Binary-Classification-Multi-Class-Classification-Regression-Using-Python-4f22032b36a2 accessed 18 May 2023

[28] Amber, T., & US, D. M. (2021). *amberthomas/us-drought-monitor | Workspace | data. world*. https://data.world/amberthomas/us-drought-monitor/workspace/project-summary?agentid=amberthomas&datasetid=us-drought-monitor accessed 20 May 2023

## Biography

**Babatunde Isaiah Ayinla** is a distinguished Computer Science lecturer at the University of Ibadan, Nigeria, where he bagged both his M.Sc. and Ph.D. His academic journey includes a fellowship at the College of Charleston, USA, in 2002. Specialising in cybersecurity, machine learning, and data science, Dr. Ayinla contributes significantly to the field through teaching and research. He imparts essential programming skills to students and supervises Master's dissertations, shaping future computer scientists. His expertise in developing robust cybersecurity systems, creating intelligent machine learning algorithms, and extracting insights from complex datasets makes him a valuable asset in the evolving landscape of computer science. Dr. Ayinla is currently en route to the Federal University of Lavras, Brazil, to pursue postdoctoral research in the Department of Automatic and System Engineering. His academic achievements and practical experience position him as an influential figure in advancing computer science knowledge and applications.

**Rasheedat Aderonke Abdulsalam** hails from Saki, Oyo State, Nigeria. She has an MSc. in Computer Science from the prestigious University of Ibadan and a BSc. in Computer Science from Alhikmah University. With her background in computer science, she has built a career and is currently working as a Technical Writer at FlowCentral Technologies and as a Program Analyst at the MIS Unit, Federal College of Education (Special), Oyo. As a Technical Writer, Abdulsalam excels in creating detailed documentation and user guides that make technical materials understandable to any audience. At the MIS Center, she uses her analytical skills to create and manage information systems that improve the institution's operational efficiency. She is a member of the Nigeria Computer Society and the Computer Professionals Registration Council of Nigeria. Her dedication to these organizations is a reflection of her passion to stay abreast of technological advancements and contribute to its development.