

Integration of Probabilistic Multi-Class Labeling and Adaptive K-Means Clustering with KNN Classification: Application to Weather Data

Husni Lubis¹, Ihsan Lubis², Herlina Harahap³, Tommy⁴, Rosyidah Siregar⁵


^{1,2}Department of Information System, Universitas Harapan Medan, Sumatera Utara, Indonesia

^{3,4,5}Department of Information Technology, Universitas Harapan Medan, Sumatera Utara, Indonesia

ABSTRACT

complex datasets. Despite their broad applications across fields such as pattern recognition, market segmentation, anomaly detection, and weather prediction, these techniques face significant limitations. Clustering methods like K-Means assume known cluster numbers and data distributions, while classification approaches such as K-Nearest Neighbors (KNN) rely heavily on the quality of labeled data. These challenges are particularly pronounced in the context of dynamic weather data, which exhibits high variability and complexity. This research addresses these limitations by integrating probabilistic multi-class labeling with an adaptive K-Means clustering approach. Probabilistic labeling allows data points to belong to multiple classes, reflecting the nuanced nature of overlapping weather conditions. Adaptive K-Means dynamically determines the optimal number of clusters, overcoming traditional constraints. By combining these methods with KNN classification, the proposed approach enhances the accuracy of weather classification. KNN leverages cluster centroids and class probabilities to provide more precise predictions. This approach provides a robust foundation for further research and optimization of adaptive methods applicable to other complex data types. Ultimately, the proposed model contributes significantly to advancing data analysis methods, particularly for dynamic and multi-class datasets like weather data.

Keyword : Clustering; Classification; Probabilistic Labeling; Adaptive K-Means; Weather.

 This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Corresponding Author:

Husni Lubis,
Department of Information System
Universitas Harapan Medan Sumatera Utara
Jl. HM Joni No 70 C Medan, 20217, Indonesia.
Email : husni.lubis82@gmail.com

Article history:

Received Aug 13, 2024
Revised Aug 17, 2024
Accepted Aug 20, 2024

1. INTRODUCTION

Clustering and classification technologies have become fundamental pillars in data analysis, enabling researchers and practitioners to identify hidden patterns within complex datasets. Clustering, in particular, serves to group similar data into clusters without requiring predefined labels. On the other hand, classification uses models trained on labeled data to categorize new data. These techniques have wide-ranging applications across various fields such as pattern recognition (Ahmed, Seraj, & Islam, 2020), market segmentation (Tabianan, Velu, & Ravi, 2022), anomaly detection (Wang, Ying, & Yang, 2020), and weather prediction (Kareem, Abdulazeez, & Hasan, 2021).

Despite the extensive use of clustering and classification, each approach has significant limitations. Clustering, particularly K-Means, tends to work under the assumption that the number of clusters and the distribution of data are known beforehand, which is often unrealistic in real-world scenarios (Ikotun, Almutari, & Ezugwu, 2021). Meanwhile, classification techniques like K-Nearest Neighbors (KNN) heavily rely on the quality and representation of labeled data (Zhang, 2021). These shortcomings often lead to inaccurate analysis results, especially in the context of highly dynamic data like weather data (Kareem, Abdulazeez, & Hasan, 2021).

Weather data, with its highly fluctuating and often unstructured nature, requires a more adaptive approach to analysis (Shofura, Suryani, Salma, & Harini, 2021). Changing weather systems not only affect weather predictions themselves but also impact various sectors such as agriculture (Ben Ayed & Hanana, 2021), transportation (Pang, Zhao, Yan, & Liu, 2021), and public health (Ajina, Jaya, Bhat, & Saxena, 2023). Therefore, it is essential to develop more sophisticated methods for analyzing weather data that can capture the complexity and variability of these phenomena.

One of the primary challenges in analyzing weather data is the need to handle multi-class data (Purwandari, Sigalingging, Cenggoro, & Pardamean, 2021), where a single set of weather data may belong to multiple categories such as rainy, cloudy, and clear. Traditional approaches in classification often fail to capture the nuances of data with multiple classes, frequently leading to less accurate or even misleading results (Cho, Yoo, Im, & Cha, 2020).

The solution proposed in this research is the integration of probabilistic multi-class labeling with an adaptive K-Means clustering approach (Sinaga & Yang, 2020). Probabilistic labeling allows each data point to have affiliations with more than one class, providing a more realistic representation of overlapping weather conditions. Additionally, adaptive K-Means is proposed to address the limitations in determining the number of clusters dynamically based on the variation in the existing data. By combining probabilistic multi-class labeling and KNN in classification following the clustering process, this approach is expected to improve the accuracy of weather classification. KNN is used to predict weather categories based on the centroids of the resulting clusters, considering the probability of each class. This enables the model to provide more informed prediction results, reflecting the complexity and overlap in the data.

This integration not only offers more accurate results but also enhances the model's flexibility in dealing with varying weather data. With this approach, researchers and practitioners can more easily adapt their analysis models to the frequent changes in weather patterns, making them more reliable for data-driven decision-making.

Furthermore, the proposed method can be implemented in various practical applications such as early warning systems, energy demand prediction, and natural resource management. The development of this model also opens opportunities for further research into optimizing other adaptive methods that can be applied to other types of data with similar characteristics to weather data. Ultimately, this research aims to make a significant contribution to the development of more robust data analysis methods, particularly in the context of weather data. By combining the strengths of probabilistic multi-class labeling, adaptive clustering, and KNN classification, this solution is expected to address existing challenges and provide a solid foundation for further advancements in the future.

2. RESEARCH METHODS

2.1 Adaptive K-Means Clustering

Adaptive K-Means clustering is an extension of the traditional K-Means algorithm designed to address the limitations of determining the number of clusters in advance. In traditional K-Means, the number of clusters (K) is a critical parameter that must be specified before the algorithm runs. However, in many real-world applications, especially with complex and dynamic data like weather data, it is difficult to predetermine the optimal number of clusters. This can lead to either over-clustering, where too many small clusters are formed, or under-clustering, where distinct groups are merged into a single cluster.

Adaptive K-Means clustering addresses this issue by dynamically adjusting the number of clusters based on the characteristics of the data. Instead of starting with a fixed K, the algorithm begins with an initial estimate and iteratively adjusts K as it analyzes the data. This approach allows the algorithm to better capture the underlying structure of the data, leading to more accurate and meaningful clustering results.

The adaptability of this method is particularly beneficial in scenarios where data distributions are non-uniform or when the data set contains natural variations over time, such as in weather data. By allowing the number of clusters to evolve, adaptive K-Means can provide a more nuanced grouping that reflects the real patterns present in the data.

The adaptive K-Means clustering algorithm can be described as follows:

1. Initialization:

- a. Start with an initial estimate for the number of clusters K_0 .
- b. Randomly select K_0 initial centroids from the dataset.

2. Assignment Step:

- a. For each data point x_i , calculate the distance to all centroids using Euclidean distance using the following equation :

$$d_{(x,y)} = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (1)$$

- b. Assign each data point to the nearest centroid, forming initial clusters.

3. Update Step:

- a. Recalculate the centroid of each cluster by averaging the data points assigned to it.
- b. Record the classes of all data points within each cluster.
- c. Calculate the probability of each class within the cluster by dividing the number of data points belonging to a particular class by the total number of data points in the cluster.

$$P(C_j|k) = \frac{n_{C_j,k}}{n_k} \quad (2)$$

Where $P(C_j|k)$ is the probability of class C_j within cluster k . $n_{C_j,k}$ is the number of data points in cluster k that belong to class C_j . n_k is the total number of data points in cluster k .

- d. Assign a multi-class label to each cluster, with each class having an associated probability. This label reflects the likelihood that a given cluster belongs to each of the recorded classes.
4. **Convergence Check.** Repeat the Assignment, Update, and Adaptation steps until the algorithm converges, i.e., until the centroids no longer change significantly, or the cluster configuration stabilizes.
5. **Finalization.** The algorithm terminates when an optimal number of clusters is reached, resulting in a final set of clusters that best represent the data's structure.

22 KNN Classification

In traditional KNN, a data point is classified by looking at the majority class among its k nearest neighbors (Wang, Han, Li, Zhang, & Cheng, 2021). However, when integrated with adaptive K-Means clustering, KNN classification can be enhanced by considering the probabilistic labels of the clusters (Kusy & Kowalski, 2022). Instead of merely counting the nearest neighbors, the classification decision incorporates the likelihood of each class as indicated by the cluster labels (Huang, Xu, Chen, & Guo, 2023).

This probabilistic approach allows the KNN classifier to provide more nuanced predictions, especially in cases where the data exhibits overlapping class distributions. By taking into account the probability distributions within the clusters, the classifier can deliver more accurate results, particularly in complex datasets like weather data where multiple weather conditions may occur simultaneously. The KNN classification process, when used in conjunction with adaptive K-Means clustering with multi-class labeling, is described as follows:

1. A set of k centroids from the final clusters produced by the adaptive K-Means algorithm, each labeled with multi-class probabilities $P(C_j|k)$.
2. A new data point x_{new} that needs to be classified.
3. Compute the distance between x_{new} and each of the k centroids. The distance can be calculated using any suitable distance metric, such as Euclidean distance.
4. Identify the k nearest centroids to x_{new} based on the calculated distances.
5. For each of the k nearest centroids, consider their associated class probabilities $P(C_j|k)$.
6. Aggregate the probabilities of each class across the k nearest centroids to determine the overall probability of each class for x_{new} . This can be done by summing the probabilities for each class across the nearest centroids.

$$P(C_j|x_{new}) = \frac{1}{k} \sum_{i=1}^k P(C_j|Centroid_i) \quad (3)$$

7. Assign x_{new} to the class with the highest aggregated probability $P(C_j|x_{new})$.
8. Alternatively, if probabilistic outputs are desired, the final classification can be expressed as a distribution over the possible classes, reflecting the calculated probabilities.
9. Output the predicted class, or the probabilistic distribution over all possible classes.

The research begins with the collection of weather data from a publicly available dataset on Kaggle, accessible at [<https://www.kaggle.com/code/devsubhash/weather-analysis-dashboard-using-pywedgeda/input>]. The dataset, named "weather.csv," contains the following columns: Formatted Data, Summary, Precip Type, Temperature (C), Apparent Temperature (C), Humidity, Wind Speed (km/h), Wind Bearing (degrees), Visibility (km), Cloud Cover, and Pressure (millibars). The class column for this research is the "Summary," which categorizes the weather conditions.

Before proceeding with clustering and classification, the data undergoes a preprocessing step to prepare it for analysis. This step is crucial to ensure that the data is in a suitable format for the adaptive K-Means clustering algorithm.

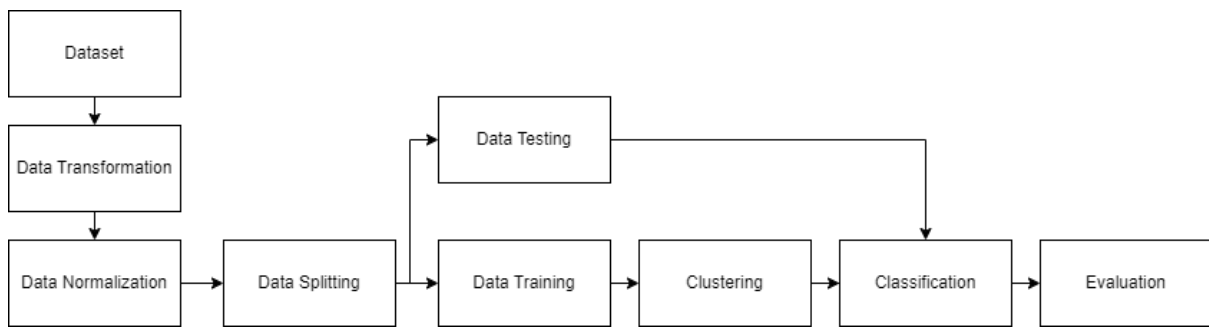


Fig 1. Clustering and Classification Stages

2.3 Date Transformation

The "Formatted Data" column, which contains datetime information in the format "2006-04-01 00:00:00.000 +0200," is transformed into a numerical attribute called `date_numeric`. This transformation is performed using the following equation :

$$datenumeric = (month \times 10000) + (day \times 100) + hour. \quad (4)$$

This equation converts the date into a normalized numeric value by multiplying the month by 10,000, the day by 100, and adding the hour. This transformation allows the date information to be incorporated into the clustering process in a meaningful way.

2.4 Data Normalization

After the date transformation, the dataset is normalized to ensure that all features are on a similar scale. Normalization is essential in clustering algorithms like K-Means to prevent features with larger numerical ranges from dominating the clustering process.

The normalization can be performed using the min-max scaling method, where each feature x is scaled to a range of $[0, 1]$ using the formula:

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (5)$$

Where x_{min} and x_{max} are the minimum and maximum values of the feature, respectively.

2.5 Clustering

With the preprocessed data, the next step involves clustering using the adaptive K-Means clustering algorithm as discussed earlier. This algorithm is particularly suited for this research because it dynamically adjusts the number of clusters based on the data's characteristics, providing more accurate and meaningful groupings.

The adaptive K-Means algorithm starts with an initial estimate of the number of clusters and proceeds through iterative assignment and update steps. The clusters are labeled with multi-class probabilistic labels, reflecting the likelihood of each class (e.g., weather condition) within the cluster.

The clustering configuration is evaluated at each iteration. If the clustering is found to be suboptimal (e.g., poor separation or overlap between clusters), the number of clusters K is adjusted. This could involve splitting clusters with high variance or merging clusters that are too close together. The process continues until the algorithm converges, resulting in a final set of clusters, each labeled with a multi-class probability distribution over the "Summary" classes.

2.6 Classification

For a new data point, the distance to each cluster centroid is calculated using Euclidean distance. The class probabilities for the new data point are aggregated across its k nearest centroids. The new data point is assigned to the class with the highest aggregated probability, or a probabilistic classification is produced, indicating the likelihood of each class.

3. RESULTS AND DISCUSSION

The dataset utilized in this study comprises 97,453 weather records, which were partitioned into training and testing datasets using an 80:20 ratio. Consequently, the data was divided as follows:

- a. **Training Dataset:** 77,962 records (80% of the total data)

b. **Testing Dataset:** 19,491 records (20% of the total data)

Table 1. Weather Datasets

Formatted Date	Summary	Precip Type	Temperature (C)	Apparent Temperature (C)	Humidity	Wind Speed (km/h)	Wind Bearing (degrees)	Visibility (km)	Pressure (millibars)
2006-04-01 00:00:00.000 +0200	Partly Cloudy	rain	9.472222	7.388889	0.89	14.1197	251	15.8263	1015.13
2006-04-01 01:00:00.000 +0200	Partly Cloudy	rain	9.355556	7.227778	0.86	14.2646	259	15.8263	1015.63
2006-04-01 02:00:00.000 +0200	Mostly Cloudy	rain	9.377778	9.377778	0.89	3.9284	204	14.9569	1015.94
2006-04-01 03:00:00.000 +0200	Partly Cloudy	rain	8.288889	5.944444	0.83	14.1036	269	15.8263	1016.41
2006-04-01 04:00:00.000 +0200	Mostly Cloudy	rain	8.755556	6.977778	0.83	11.0446	259	15.8263	1016.51
2006-04-01 05:00:00.000 +0200	Partly Cloudy	rain	9.222222	7.111111	0.85	13.9587	258	14.9569	1016.66
2006-04-01 06:00:00.000 +0200	Partly Cloudy	rain	7.733333	5.522222	0.95	12.3648	259	9.982	1016.72
...
2016-09-09 23:00:00.000 +0200	Partly Cloudy	rain	20.43889	20.43889	0.61	5.8765	39	15.5204	1016.16

The initial dataset underwent a crucial transformation process to convert date and time information into numerical features suitable for analysis. Specifically, the transformation applied a function to convert the date and time columns into a numeric format, facilitating their integration into the machine learning models.

The transformation equation used was designed to encode the date and time data as a continuous numeric attribute. This approach involved extracting key components from the date and time, such as year, month, day, and time of day, and converting these components into a single numerical value. This value represents the temporal aspect of each record in a format that allows for easier manipulation and analysis by the subsequent clustering and classification algorithms.

Table 2. Transformed Dataset

Summary	Temperature C	Apparent Temperature C	Humidity	Wind Speed (km/h)	Wind Bearing degrees	Visibility (km)	Pressure millibars	Date Numeric
Clear	12.73	12.73	0.81	0.35	310	16.1	1019.75	90203
Clear	12.2	12.2	0.81	8.05	290	16.1	1019.55	90204
Clear	11.93	11.93	0.84	0.31	317	15.18	1019.47	90205
Partly Cloudy	11.14	11.14	0.86	0	0	16.1	1019.73	90206
Partly Cloudy	14.82	14.82	0.79	0	0	16.1	1019.74	90207
Partly Cloudy	18.46	18.46	0.68	3.2	316	15.18	1019.75	90208
Partly Cloudy	22.67	22.67	0.52	0.14	40	16.1	1019.73	90209
...
Partly Cloudy	20.44	20.44	0.61	5.88	39	15.52	1016.16	90923

Following the transformation of date and time data into numerical features, the dataset was divided into training and testing subsets. The training dataset, consisting of 77,962 records, was subjected to clustering to identify multi-label clusters. This process involved applying the K-Means clustering algorithm to the training data, which resulted in distinct clusters with multiple labels corresponding to the weather classes present in the data.

The multi-label clusters derived from the K-Means algorithm were then utilized to classify the testing dataset. This classification step involved assigning each record in the testing dataset to one of the identified clusters based on its feature similarity. The labels associated with these clusters were used to predict the weather classes for the test records.

The clustering process commenced with an initial setup of $k = 27$ clusters. The initial cluster labels were determined based on randomly selected centroids from the dataset. This initial random selection of centroids served as the starting point for the K-Means clustering algorithm. As the clustering process progressed, the labels of the clusters evolved to reflect the proportions of the weather classes within each cluster. The K-Means algorithm iteratively updated the cluster centroids and re-assigned data points based on the nearest centroid. Consequently, the labels of the clusters were adjusted to represent the dominant class proportions of their members.

Table 3. Cluster Labels

Cluster	Labels	Probabilities
1	[Clear, Foggy, Mostly Cloudy, Overcast, Partly Cloudy]	[0.15206929740134745, 0.2564966313763234, 0.16987487969201154, 0.24205967276227142, 0.1794995187680462]

2	[Clear, Drizzle, Foggy, Humid and Mostly Cloudy, Humid and Partly Cloudy, Mostly Cloudy, Overcast, Partly Cloudy, Rain]	[0.1792086889061288, 0.001034393586759762, 0.026377036462373934, 0.002068787173519524, 0.0002585983966899405, 0.2904059994828032, 0.13369537108869925, 0.36669252650633566, 0.0002585983966899405]
3	[Clear, Dry and Partly Cloudy, Mostly Cloudy, Overcast, Partly Cloudy]	[0.1687648860156516, 0.0003402517863218782, 0.270840421912215, 0.025859135760462743, 0.5341953045253488]
4	[Clear, Dry, Dry and Mostly Cloudy, Dry and Partly Cloudy, Foggy, Mostly Cloudy, Overcast, Partly Cloudy]	[0.055865921787709494, 0.0026289845547157412, 0.001643115346697338, 0.008544199802826159, 0.00032862306933946765, 0.29181728557344727, 0.015445284258954979, 0.6237265856063096]
5	[Breezy, Breezy and Foggy, Breezy and Mostly Cloudy, Breezy and Overcast, Breezy and Partly Cloudy, Clear, Foggy, Humid and Mostly Cloudy, Humid and Overcast, Mostly Cloudy, Overcast, Partly Cloudy]	[0.00048355899419729207, 0.0038684719535783366, 0.0009671179883945841, 0.0024177949709864605, 0.00048355899419729207, 0.02562862669245648, 0.0024177949709864605, 0.00048355899419729207, 0.00048355899419729207, 0.4410058027079304, 0.14941972920696325, 0.3728239845261122]
6	[Clear, Foggy, Mostly Cloudy, Overcast, Partly Cloudy]	[0.19123505976095617, 0.00099601593625498, 0.28834661354581675, 0.11354581673306773, 0.40587649402390436]
7	[Clear, Drizzle, Foggy, Light Rain, Mostly Cloudy, Overcast, Partly Cloudy, Rain]	[0.15118577075098813, 0.0004940711462450593, 0.17045454545454544, 0.00024703557312252963, 0.23789525691699603, 0.21195652173913043, 0.22702569169960474, 0.0007411067193675889]
8	[Breezy, Breezy and Mostly Cloudy, Breezy and Overcast, Breezy and Partly Cloudy, Clear, Dry, Dry and Mostly Cloudy, Dry and Partly Cloudy, Mostly Cloudy, Overcast, Partly Cloudy]	[0.004203888596952181, 0.024567280848687884, 0.0027917364600781687, 0.03740926856504746, 0.005780346820809248, 0.05862646566164154, 0.00446677833612507, 0.0016750418760469012,

		0.010050251256281407, 0.3316582914572864, 0.038525963149078725, 0.4868788386376326]
9	[Clear, Drizzle, Foggy, Light Rain, Mostly Cloudy, Overcast, Partly Cloudy]	[0.10441880101322826, 0.0014072614691809737, 0.06501547987616099, 0.0025330706445257528, 0.27694905713481566, 0.3216999718547706, 0.22769490571348155]
10	[Breezy, Breezy and Mostly Cloudy, Breezy and Overcast, Breezy and Partly Cloudy, Clear, Dangerously Windy and Partly Cloudy, Mostly Cloudy, Overcast, Partly Cloudy, Windy, Windy and Foggy, Windy and Mostly Cloudy, Windy and Overcast, Windy and Partly Cloudy]	[0.021231422505307854, 0.2611464968152866, 0.20063694267515925, 0.2070063694267516, 0.006369426751592357, 0.0010615711252653928, 0.07749469214437367, 0.02653927813163482, 0.0732484076433121, 0.005307855626326964, 0.0021231422505307855, 0.027600849256900213, 0.03397027600849257, 0.05626326963906582]
11	[Breezy, Breezy and Foggy, Breezy and Mostly Cloudy, Breezy and Overcast, Breezy and Partly Cloudy, Clear, Foggy, Light Rain, Mostly Cloudy, Overcast, Partly Cloudy, Windy and Foggy, Windy and Mostly Cloudy, Windy and Overcast]	[0.0031628887717448603, 0.00790722192936215, 0.025830258302583026, 0.05851344227727991, 0.00790722192936215, 0.045861887190300474, 0.03900896151818661, 0.002635740643120717, 0.2878228782287823, 0.40168687401159725, 0.11755403268318397, 0.0005271481286241434, 0.0005271481286241434, 0.0010542962572482868]
12	[Clear, Drizzle, Foggy, Humid and Mostly Cloudy, Humid and Overcast, Humid and Partly Cloudy, Light Rain, Mostly Cloudy, Overcast, Partly Cloudy]	[0.07466666666666667, 0.0007619047619047619, 0.0053333333333333333, 0.004190476190476191, 0.00038095238095238096, 0.0015238095238095239, 0.00038095238095238096, 0.3798095238095238, 0.22895238095238096, 0.304]
13	[Clear, Foggy, Mostly Cloudy, Overcast, Partly Cloudy]	[0.14592703648175911, 0.0004997501249375312, 0.30184907546226886,

		0.09745127436281859, 0.4542728635682159]
14	[Clear, Foggy, Mostly Cloudy, Overcast, Partly Cloudy]	[0.12112754674853475, 0.32291375941948086, 0.18336589450181412, 0.21378732905386547, 0.15880547027630476]
15	[Breezy, Breezy and Mostly Cloudy, Breezy and Overcast, Breezy and Partly Cloudy, Clear, Foggy, Humid and Mostly Cloudy, Light Rain, Mostly Cloudy, Overcast, Partly Cloudy]	[0.0005428881650380022, 0.025515743756786103, 0.016286644951140065, 0.006514657980456026, 0.0494028230184582, 0.002171552660152009, 0.0010857763300760044, 0.0005428881650380022, 0.032263694706949, 0.25234693611875914, 0.0400394062565651]
16	[Breezy, Breezy and Mostly Cloudy, Breezy and Overcast, Breezy and Partly Cloudy, Clear, Foggy, Mostly Cloudy, Overcast, Partly Cloudy]	[0.015215015215015215, 0.03318603318603318, 0.004794004794004794, 0.01894801894801895, 0.3094133094133094, 0.000733000733000733, 0.14111214111214112, 0.20542620542620543, 0.27605627605627605]
17	[Clear, Drizzle, Foggy, Light Rain, Mostly Cloudy, Overcast, Partly Cloudy, Rain]	[0.11909535388721868, 0.001534926563383401, 0.09183250368845623, 0.010820307464320457, 0.3092138876603641, 0.31047201789022743, 0.1578066702522094, 0.0007843333205791854]
18	[Clear, Foggy, Mostly Cloudy, Overcast, Partly Cloudy]	[0.16489324423561866, 0.005679084346874828, 0.26325982692825865, 0.13485990689172658, 0.3912089375975213]
19	[Clear, Foggy, Mostly Cloudy, Overcast, Partly Cloudy]	[0.12256984713603215, 0.013471691547597077, 0.3053738357881145, 0.22105661423088335, 0.3375280112973729]
20	[Breezy, Breezy and Mostly Cloudy, Breezy and Overcast, Breezy and Partly Cloudy, Clear, Foggy, Humid and Mostly Cloudy, Light Rain, Mostly Cloudy, Overcast, Partly Cloudy]	[0.004372136091796369, 0.005285038255639321, 0.007975428249509283, 0.001831485847821485, 0.29458746776346746, 0.011929431637524277, 0.002274572059184877, 0.03374954960706692,

		0.3331741286461517, 0.2900421512811483, 0.02090361974161306]
21	[Clear, Foggy, Mostly Cloudy, Overcast, Partly Cloudy]	[0.1466731581515851, 0.0008120434603717498, 0.2758954237989037, 0.09096501325821496, 0.4866543613309244]
22	[Clear, Drizzle, Foggy, Light Rain, Mostly Cloudy, Overcast, Partly Cloudy, Rain]	[0.0816919499283216, 0.0009043665818638245, 0.08720482008696551, 0.009075402477842383, 0.312776213951258, 0.2845631231422931, 0.2127847969440906, 0.013798298826334146]
23	[Clear, Foggy, Mostly Cloudy, Overcast, Partly Cloudy]	[0.1404374249497024, 0.009762575874350686, 0.23341094496067218, 0.13938914734155162, 0.4779999078737221]
24	[Clear, Drizzle, Foggy, Light Rain, Mostly Cloudy, Overcast, Partly Cloudy, Rain]	[0.1358826497243763, 0.003191831825487416, 0.07847374283973402, 0.009018046775826056, 0.2809573487289298, 0.3361649641213483, 0.152048926084365, 0.004262961204083629]
25	[Clear, Foggy, Mostly Cloudy, Overcast, Partly Cloudy]	[0.12769394553220637, 0.002632839235007051, 0.2712822772957642, 0.1405439310632365, 0.4588470068737868]
26	[Breezy, Breezy and Mostly Cloudy, Breezy and Overcast, Breezy and Partly Cloudy, Clear, Foggy, Mostly Cloudy, Overcast, Partly Cloudy]	[0.027547809735978744, 0.01006822269362103, 0.009712020882216032, 0.002175575166279118, 0.2056474801056945, 0.005562650290411953, 0.2479605946498852, 0.24817734476201576, 0.24458533115987862]
27	[Clear, Foggy, Mostly Cloudy, Overcast, Partly Cloudy]	[0.1751488633319697, 0.001929253273073371, 0.30682967903758356, 0.11592737706873184, 0.4001648272886415]

Once the testing dataset was classified using the multi-label clusters, the results were thoroughly analyzed. The performance of the classification was evaluated by comparing the predicted weather

classes with the actual classes in the testing dataset. Metrics such as accuracy, precision, recall, and F1-score were calculated to assess the effectiveness of the classification approach.

Table 4. Classification Results

Label	Precision	Recall	F1-Score	Support
Breezy	1	1	1	13
Breezy and Dry	0	0	0	1
Breezy and Foggy	1	0.9091	0.9524	11
Breezy and Mostly Cloudy	0.9902	1	0.9951	101
Breezy and Overcast	1	1	1	119
Breezy and Partly Cloudy	1	1	1	81
Clear	1	1	1	2184
Dangerously Windy and Partly Cloudy	0	0	0	0
Drizzle	1	0.6667	0.8	3
Dry	1	1	1	5
Dry and Mostly Cloudy	0	0	0	0
Dry and Partly Cloudy	1	1	1	15
Foggy	1	0.9993	0.9997	1446
Humid and Mostly Cloudy	1	1	1	7
Humid and Overcast	1	1	1	2
Humid and Partly Cloudy	1	1	1	5
Light Rain	1	0.7778	0.875	9
Mostly Cloudy	0.9996	1	0.9998	5707
Overcast	1	1	1	3347
Partly Cloudy	0.9994	1	0.9997	6395
Rain	1	1	1	2
Windy	1	1	1	3
Windy and Dry	0	0	0	1
Windy and Foggy	1	1	1	1
Windy and Mostly Cloudy	1	1	1	8
Windy and Overcast	1	1	1	11
Windy and Partly Cloudy	1	1	1	14
micro avg	0.9996	0.9996	0.9996	19491
macro avg	0.8515	0.8279	0.8378	19491
weighted avg	0.9995	0.9996	0.9996	19491
Summary	0.8613	0.8393	0.8486	77964

The model evaluation results reveal an exceptional performance in terms of accuracy and other evaluation metrics. The model achieved an impressive accuracy of **0.9996**, indicating its ability to classify data with almost no errors. Precision, reflecting the model's accuracy in identifying positive classes, is also exceptionally high at **0.9995**, demonstrating that the model consistently makes correct predictions for positive classes.

Recall, which measures how well the model identifies all instances of the positive class, reached **0.9996**, indicating that nearly all positive data is correctly identified. The F1 Score, combining precision and recall into a single metric, also achieved a near-perfect value of **0.9996**, reflecting an excellent balance between accuracy and coverage.

Table 5. Classification Scores

Variable	Value
Accuracy	0.9996

Precision	0.9995
Recall	0.9996
F1 Score	0.9996

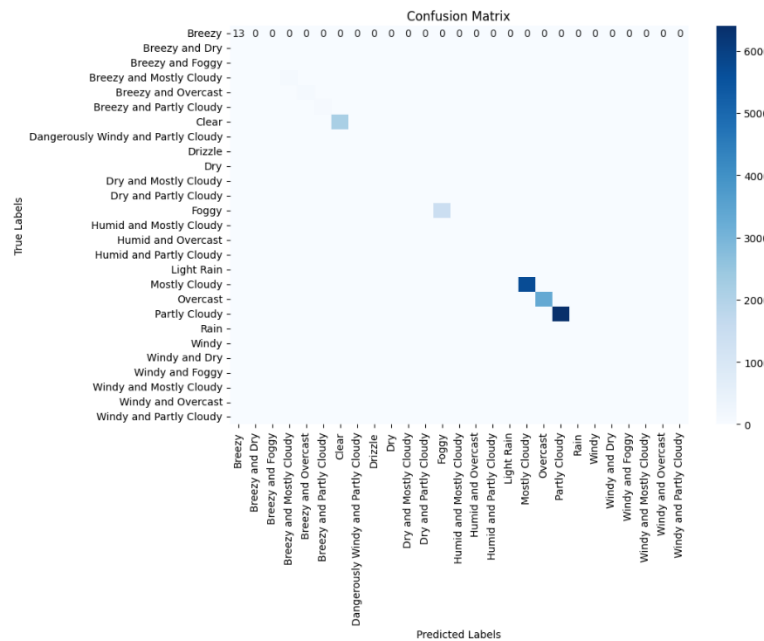


Fig 2. Confusion Matrix

In summary, the evaluation metrics underscore the model's outstanding performance across all critical aspects of classification. With an accuracy of **0.9996**, it demonstrates near-perfect precision and recall, reflecting its robust ability to correctly classify instances and minimize both false positives and false negatives. The near-perfect F1 Score of **0.9996** further highlights the model's effectiveness in balancing precision and recall.

These results not only validate the model's reliability but also underscore its potential for practical applications where high accuracy and precision are essential. The exceptional performance metrics suggest that the model is well-suited for complex classification tasks, offering valuable insights and accurate predictions in its domain of application. This level of performance provides a strong foundation for future deployments and further refinement to enhance its capabilities even further.

4. CONCLUSION

The results underscore the remarkable effectiveness of the model in weather data classification, with an impressive accuracy of 0.9996. This exceptional accuracy highlights the model's capability to classify nearly all instances correctly with minimal errors. Precision, at 0.9995, signifies the model's reliability in correctly identifying positive classes, ensuring accurate predictions. The recall score of 0.9996 reflects the model's proficiency in identifying almost all relevant instances of positive classes, while the F1 Score, also at 0.9996, demonstrates an excellent balance between precision and recall, showcasing the model's comprehensive performance. A key factor contributing to this success is the integration of K-Nearest Neighbors (K-NN) classification within the adaptive K-means clustering framework. The K-NN algorithm enhances the classification process by evaluating the proximity of data points to their neighbors, which aids in accurate label prediction based on the majority class among the nearest neighbors. The adaptive K-means clustering, coupled with K-NN classification, further refines the model's performance. The clustering algorithm begins with an initial number of clusters, assigns data points to the nearest centroids, and iteratively updates centroids while recording class probabilities within each cluster. This approach ensures that each cluster is labeled with a multi-class label reflecting the likelihood of each class being present. Together, the adaptive K-means clustering and K-NN classification techniques enable a detailed and nuanced analysis of weather data, resulting in highly accurate and balanced performance metrics. The model's ability to effectively integrate these advanced methods demonstrates

its superior capability in handling complex data structures and achieving exceptional results in weather data classification.

REFERENCES

- Ahmed, M., Seraj, R., & Islam, S. (2020). The k-means algorithm: A comprehensive survey and performance evaluation. *Electronics*, 9(8), 1295.
- Ajina, A., Jaya, C., Bhat, D., & Saxena, K. (2023). Prediction of weather forecasting using artificial neural networks. *Journal of applied research and technology*, 21(2), 205-211.
- Ben Ayed, R., & Hanana, M. (2021). Artificial intelligence to improve the food and agriculture sector. *Journal of Food Quality*, 2021(1), 5584754.
- Cho, D., Yoo, C., Im, J., & Cha, D. (2020). Comparative assessment of various machine learning-based bias correction methods for numerical weather prediction model forecasts of extreme air temperatures in urban areas. *Earth and Space Science*, 7(4), e2019EA000740.
- Huang, A., Xu, R., Chen, Y., & Guo, M. (2023). Research on multi-label user classification of social media based on ML-KNN algorithm. *Technological Forecasting and Social Change*, 188, 122271.
- Ikotun, A., Almutari, M., & Ezugwu, A. (2021). K-means-based nature-inspired metaheuristic algorithms for automatic data clustering problems: Recent advances and future directions. *Applied Sciences*, 11(23), 11246.
- Kareem, F., Abdulazeez, A., & Hasan, D. (2021). Predicting weather forecasting state based on data mining classification algorithms. *Asian Journal of Research in Computer Science*, 9(3), 13-24.
- Kusy, M., & Kowalski, P. (2022). Architecture reduction of a probabilistic neural network by merging k-means and k-nearest neighbour algorithms. *Applied Soft Computing*, 128, 109387.
- Pang, Y., Zhao, X., Yan, H., & Liu, Y. (2021). Data-driven trajectory prediction with weather uncertainties: A Bayesian deep learning approach. *Transportation Research Part C: Emerging Technologies*, 130, 103326.
- Purwandari, K., Sigalingging, J., Cenggoro, T., & Pardamean, B. (2021). Multi-class weather forecasting from twitter using machine learning approaches. *Procedia Computer Science*, 179, 47-54.
- Shofura, S., Suryani, S., Salma, L., & Harini, S. (2021). The Effect of Number of Factors and Data on Monthly Weather Classification Performance Using Artificial Neural Networks. *International Journal on Information and Communication Technology (IJoICT)*, 7(2), 23-35.
- Sinaga, K., & Yang, M. (2020). Unsupervised K-means clustering algorithm. *IEEE access*, 8, 80716-80727.
- Tabianan, K., Velu, S., & Ravi, V. (2022). K-means clustering approach for intelligent customer segmentation using customer purchase behavior data. *Sustainability*, 14(12), 7243.
- Wang, B., Ying, S., & Yang, Z. (2020). A Log-Based Anomaly Detection Method with Efficient Neighbor Searching and Automatic K Neighbor Selection. *Scientific Programming*, 2020(1), 4365356.
- Wang, L., Han, M., Li, X., Zhang, N., & Cheng, H. (2021). Review of classification methods on unbalanced data sets. *Ieee Access*, 9, 64606-64628.
- Zhang, S. (2021). Challenges in KNN classification. *IEEE Transactions on Knowledge and Data Engineering*, 34(10), 4663-4675.