JINAV: Journal of Information and Visualization *ISSN: 2746-1440 (online)*
Vol. 7 No. 1 (2026)
https://doi.org/10.35877/454RI.jinav4788

Rainfall Classification Using Output Statistics Models Based on Classification and Regression Trees with Principal Component Analysis Preprocessing

Zulkifli Rais*, Hardianti Hafid, & Yhegi Rombe Bunga

Department of Statistics, Faculty of Mathematics and Natural Sciences, Universitas Negeri Makassar, Indonesia

Abstract

Makassar City has a varied monsoon rainfall pattern, so rainfall prediction is an important challenge in disaster mitigation and resource management. Data mining techniques such as classification with the Classification and Regression Trees (CART) algorithm can be used to classify rainfall and analyze historical data, but the risk of overfitting high-dimensional data requires dimension reduction such as Principal Component Analysis (PCA). To improve accuracy, the Output Statistics Model (MOS) approach that combines numerical data and observations is also used. The results of dimension reduction using the Principal Component Analysis (PCA) method showed that of the initial seven variables, only three main components were retained because they had eigenvalues greater than 1 and were able to explain the data variance significantly. The decision tree model that was formed resulted in an accuracy rate of 72.34% in training data. Where the model can classify most of the training data into the correct rainfall category. In the data testing, the model was able to achieve an accuracy level of 71.43%, which shows that the model has good generalization ability to new data and does not experience overfitting.

Keywords: principal component analysis, classification and regression trees, rainfall.

Received: 11 October 2025 Revised: 30 March 20246 Published: 30 April 2026

1. Introduction

Data Mining It is a process of collection, the use of historical data to find order, relationship patterns in large datasets. Output from Data Mining It can be used to make decisions in the future (Handoko et al., 2020). In data mining, the data processing process is carried out with various techniques. One of the techniques that is often used in data mining is classification (Wijaya & Triayudi, 2023). Classification is a simple method used to recognize a class or data model that is then used as an approach in predicting a problem. One of the classification algorithms is the algorithm Classification and Regression Trees (CART) (Hasanah et al., 2021).

Classification and Regression Trees (CART) are an algorithm of a decision tree technique known as Decision Trees. CART is referred to as a statistical and nonparametric algorithm that can be described as a response variable with one or more predictor variables. CART aims to obtain an accurate set of data as a characterization of a classification (Pratiwi & Zain, 2014).

Large data requires larger memory and more variables, and too large data can affect computer memory usage. Memory limitations will affect the classification performance to be suboptimal so that more input variables will result in overfitting of the data. To overcome the data dimension that is too large, the data dimension must be reduced so that the classification process can run properly. One way to overcome overfitting is to reduce the data dimension.

Dimension reduction is a technique to reduce multicollinearity in input variables. Dimension reduction works by reducing the number of variables on the Dataset without omitting important information in Dataset aforementioned. One method to reduce dimensions is Main Component Analysis (PCA) (Rumah & Basaruddin, 2011).

* Corresponding author.
E-mail address: zulkifli.rais89@unm.ac.id

Makassar City based on weather conditions and rainfall, is included in the group of areas with a temperate to tropical climate and includes having a monsumal rainfall pattern (Nensi et al., 2016). By classifying rainfall into specific categories, such as low, moderate, or high, decision-makers can more easily understand rainfall patterns and implement strategic measures to mitigate the negative impacts of extreme rainfall (Juliati, 2023). However, accurate rainfall prediction is still a challenge due to the complexity of the factors that affect it. Therefore, an effective approach is needed to classify and predict rainfall based on available historical data. One approach that can be used is Model Output Statistics (MOS), which combines the output of numerical models with observational data to improve prediction accuracy (Ramadan & Septiadi, 2024).

Research related to MOS was conducted by Niswatul et al (2021) conducted research on Output Statistics Model with Principal Component Regression (PCR), Least Square Regression (PLSR), and ridge regression. The results of the study concluded that MOS was able to correct the forecast bias of the NWP by more than 50%. Another study by Purnamawati et al (2022) for the prediction of bicycle users based on the weather using the CART method showed an accuracy of 90%. The next research conducted by Musfiroh et al. (2023) is the Application of Principal Component Analysis (PCA) and Short-Term Memory (LSTM) Methods in predicting Daily Rainfall. In this study, PCA was able to increase accuracy in considering all parameters and choosing the effective one.

2. Literature Review

Classification and Regression Trees (CART) is one of the methods or algorithms of one of the data exploration techniques, namely the decision tree technique. CART aims to obtain an accurate data group as a characteristic of a classification. The resulting tree model depends on the scale of the response variable, if the data response variable is continuous then the resulting tree model is a regression tree (regression trees) while if the response variable has a categorical scale, the resulting tree is a classification tree (Classification Trees) (Pratiwi & Zain, 2014).

The classification tree is a method of recursive partitioning of data in a repetitive and binary manner (binary recursive partitioning), because it always divides the data set into two partitions. Each data restriction is expressed as a node. Selection is carried out on each node until a terminal/end node is obtained. The variable that sorts at the main node is the most important variable in estimating the class from observation. Lewis (2000) in Yusri (2008) calls the root node the parent node, while the fraction of the parent node is called the internal node. The final node is also referred to as the terminal node where there is no election. The depth of the tree is calculated starting from the main node (t_1) is at depth 1, while it is at depth 2, and so on until the final node.

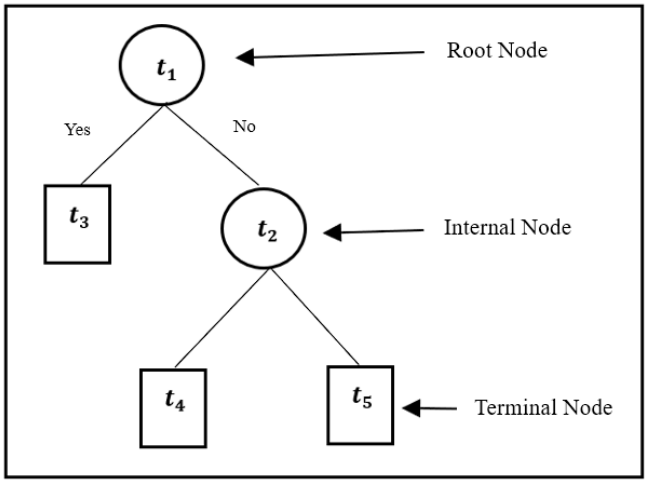

Figure 1. Classification tree structure (Tanjung & Kartiko, 2017)

The application of the CART algorithm method consists of two stages, namely the formation of the classification tree, and the pruning of the classification tree.

2.1. Formation of the Classification Tree

In the preparation of the classification tree, there are three core processes, namely the selection of sorters, the determination of terminal nodes, and the marking of class labels as follows:

a. Selection of Sorters

The selection of each node is carried out to obtain the sorter that has the most homogeneous variable value. Selecting a sorter with a Gini index before it is done, it would be better to first get gain information for each node with the formula used as follows:

$$GI(t) = -\sum_{j}^{n} P(j|t)\log_2 P(j|t) \quad (1)$$

where:

$GI(t) =$ Gain Information on nodes t
$P(j|t)$ = Proportion of classes j on the knot t where j = 1,2, 3..., n, with $P(j|t) = \frac{n_{j(t)}}{n(t)}$
$n_{j(t)}$ = The number of observation processes in class j on Note t
$n(t)$ = The number of observation processes in Note t

A sorting method that can measure the level of class heterogeneity of a node in the classification tree is by using the impurity measure i(t) method. The Gini index on the t-node of the jth class can be written as follows:

$$\begin{aligned} i(t) &= \sum_{j=1}^{k} p(j|t)(1 - p(j|t) \\ &= \sum_{j=1}^{k} p(j|t) - p^2(j|t) \\ &= \sum_{j=1}^{k} p(j|t) - \sum_{j=1}^{k} p^2(j|t) \\ &= 1 - \sum_{j=1} P^2(j|t) \end{aligned} \quad (2)$$

where:

$i(t) =$ gini index

$P(j|t)$ = Proportion of classes j on the knot t where j = 1,2, 3, …, n, with $P(j|t) = \frac{n_{j(t)}}{n(t)}$
$n_{j(t)}$ = The number of observation processes in class j on Note t
$n(t)$ = The number of observation processes in Note t

The attributes obtained from the selection results will build a set of classes called nodes or nodes. Furthermore, to determine the goodness of split criteria which include the assessment of the selection by the s-ordinator at t can also be referred to as a decrease in heterogeneity, the formula is:

$$\Delta i(s,t) = i(t) - P_L i(t_L) - P_R i(t_R) \quad (3)$$

where:

$\Delta i(s,t)$ = value goodness of split
$i(t)$ = Heterogeneity function on nodes t
P_R = Proportion of right-node observations
P_L = Proportion of left-node observations
$i(t_L)$ = heterogeneity function in the left child node
$i(t_R)$ = heterogeneity function in the right child node

For and can be calculated by entering the probability of occurrence of many objects, the formula is as follows:$t_L t_R$

$$P_R = \frac{\text{right node}}{\text{Data Training}} \quad (4)$$

$$P_L = \frac{\text{Left Node}}{\text{Data Training}} \quad (5)$$

The selection that obtained the highest results $\Delta i(s,t)$ was the best sorter. The development of the tree is carried out at the main node t_1 and then selected to be t_2 and so on t_3.

b. Determination of Terminal Nodes

The determination of the node t can be a terminal node or not by selecting, if there is no decrease in heterogeneity in the node. The criteria for determining the node can be a terminal node, namely if there is a node $n > 5$ and the tree formed will stop if it has reached the level specified in the maximum tree criterion.

c. Class Label Marking

Marking a class is a form of identification for each node in a given class. The class tagging process is carried out by the terminal node, non-terminal node, and root node. To perform class label marking based on the maximum number rule is if:

$$p(j_0|t) = \max_j p(j|t) = \max_j \frac{N_j(t)}{n(t)} \quad (6)$$

where:

$p(j_0|t)$ = The proportion of the class on the nodej_0 t
$p(j|t)$ = Proportion of classes j on the knot t
$N_j(t)$ = Number of class observations j at knot t
$N(t)$ = Number of observations on the node t

The symbol is the label of the terminal class of jthe t node that has the largest guessing value of all the t-node classification errors. As for the formation of a classification tree, if there is a discontinuation from the observation of each child node or a minimum limit n, there is an observation in the child node that is similar, and has a maximum tree depth or limit of the number of levels.

2.2. Classification Tree Pruning (Pruning)

Pruning is carried out on less important parts of the tree, so that an optimal classification tree will be obtained. The pruning measure used to obtain a decent tree size is called Cost complexity minimum (Evalina & Sinambela, 2008). In order to have a suitable tree size, it is necessary to take trees that have been pruned based on the size formula Cost Complexity Minimum The following:

$$R_\propto(T) = R(T) + \alpha|\widetilde{T}| \quad (7)$$

where:

$R_\propto(T)$ = Resubtition of a T tree on complexity α
R(T) = Resubtition Estimate (substitute assumption)
α = Cost complexity parameter for adding one node Tree End T
$|\widetilde{T}|$ = The size of the number of terminal nodes of the tree

3. Research methods

3.1. Types of Research

This study uses a quantitative approach because this research focuses on processing and analyzing numerical data, such as meteorological data from BMKG. This research involves measuring variables, applying statistical methods (Model Output Statistics) for bias correction, reducing data dimensions using Principal Component Analysis (PCA), and applying the Classification and Regression Trees (CART) machine learning algorithm to map the relationship between predictor variables and rainfall.

3.2. Data Source

The data in this study is secondary data, namely Makassar City rainfall data sourced from the website of the Meteorology, Climatology, and Geophysics Agency of the Paotere Maritime Meteorological Station. The data is daily data for the period of January 1, 2014 to January 1, 2024.

The results of the PCA reduction in the form of several main components are then used as predictive variables to build tree classifications. Meanwhile, the rainfall response variable will be classified into 5 categories, with the following criteria (Maraun & Widmann, 2017):

Table 3. Classification of Rainfall According to Its Intensity

Rain Classification	Rainfall intensity (mm/day)
Sunny cloudy	Rainfall ≤ 0.1
Light rain	$0.1 <$ Rainfall ≤ 20
Moderate rain	$20 <$ Rainfall ≤ 50
Cloudburst	$50 <$ Rainfall ≤ 100
Heavy rain	Rainfall > 100

4. Results and Discussion

4.1. Dimension Reduction with Principal Component Analysis (PCA)

Main Component Analysis (PCA) is used to find out how many new components are formed to be able to explain the influence of rainfall that occurs in Makassar City. The first is to use a scree plot.

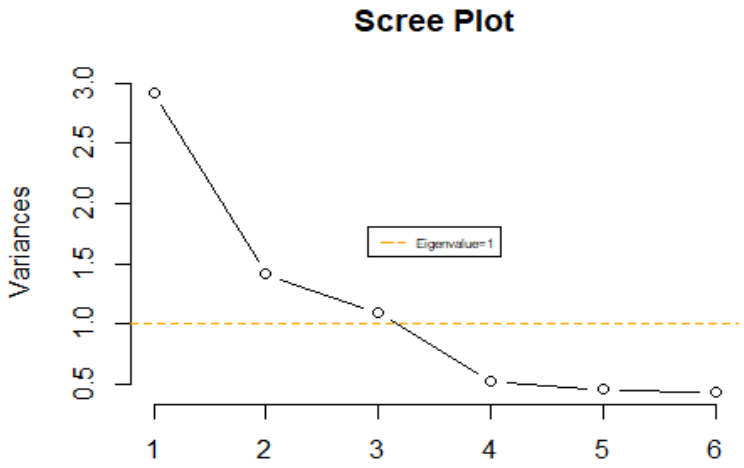

Figure 3. Scree Plot Main components

Based on Figure 3, the components are taken up to the third component in the scree plot because the first three components have an eigenvalue above 1. This criterion suggests that only components with an eigenvalue of > 1 are worth maintaining because each account for a greater variance than one original variable. After the third component, the eigenvalue drops below 1, so the fourth to sixth components are not considered important enough to be included in further analysis. So, 3 main components were taken because only three met the significance criteria based on the eigenvalue.

In addition to using a scree plot, the determination of the number of factors can also be done through more than one eigen. The following are the calculation results for eigenvalues, total proportions of variance, and cumulative proportions.

Table 4. Eigenvalues, total proportions of variance, and cumulative proportions

Nilai Eigen	Proportion of Total Variance (%)	Cumulative (%)
1.71	41.73	41.73
1.19	20.19	61.92
1.05	15.60	77.53
0.72	7.39	84.92
0.67	64.75	91.39
0.66	61.94	97.59
0.41	24.15	100

Based on Table 4, it can be seen that the new components formed are 3, it can be seen from the eigenvalue of >1 there are as many as 3. The first main component obtained was with an eigenvalue of 1.71 with a variance of 41.73%. These three components can explain the total cumulative variance of 77.53%. Next, the eigenvector value is searched

based on the previously obtained eigenvalue. The results obtained indicate that there are 3 main components, then the constituent elements of the three main components are selected based on the largest loading value. The loading value is the correlation identification value of the factor formed with the variable value. The closer the relationship between the factor and the variable, the greater the loading value. The loading values of the three components are presented in Table 5.

Table 5 Loading values

Variable	PC_1	PC_2	PC_3
X_1	-0.2837452	0.0404156	0.7675183
X_2	**-0.4543388**	-0.1597612	0.0328701
X_3	-0.5045168	-0.0851576	0.2858604
X_4	0.4400725	0.1740962	0.3885521
X_5	-0.3974832	-0.2779021	-0.3759548
X_6	0.2754601	-0.6133468	0.1041977
X_7	0.1760477	-0.6941597	0.1579463

Table 5 explains the relationship between the original variable and the new variable formed by Principal Component Analysis. Table 4.2 shows that the largest loading value on each of the x variables is indicated by the bolded values, i.e., and on, and on and on. These variables are then grouped into new main components. The explanation of the relationship between these variables can be seen in Table 6.

Table 6 Loading values

PC	Variable	Variance described
PC_1	X_2 X_3 X_4	41.73%
PC_2	X_6 X_7	20.19%
PC_3	X_1	15.60%

Based on the loading values in Table 6 and the PCA results of the loading values in Table 4.3, the variables that represent each main component are shown. The first main component is PC_1 represented by the minimum temperature variable (X_2), average temperature (X_3) and average humidity (X_4) , for the second main component or is PC_2 represented by the variable by the maximum wind speed (X_6) and average wind speed (X_7), and for the third main component or PC_3 represented by the minimum temperature variable (X_1) . Next, the main component equation will be formed as follows:

$$PC_1 = -0.2837452X_1 - 0.4543388X_2 - 0.5045168X_3 + 0.4400725X_4 - 0.3974832X_5 + 0.2754601X_6 + 0.1760477X_7$$
$$PC_2 = 0.0404156X_1 - 0.1597612X_2 - 0.0851576X_3 + 0.1740962X_4 - 0.2779021X_5 - 0.6133468X_6 - 0.6941597X_7$$
$$PC_3 = 0.7675183X_1 + 0.0328701X_2 + 0.2858604X_3 + 0.3885521X_4 - 0.3759548X_5 + 0.1041977X_6 + 0.1579463X_7$$

Which results in the values of the main components in Table 7

Table 7. Value of the main components

PC_1	PC_2	PC_3
3.66369	-0.36319	1.12354
2.37788	-4.25255	2.05542
.	.	.
.	.	.
.	.	.
s-0.17281	-0.30652	0.67325
1.97917	0.52043	-0.68183

Table 7 is the result of obtaining data that has been reduced in dimension using Principal Component Analysis where from the initial data 7 variables after being reduced to data with 3 new variables. The data is then used to classify using Classification and Regression Trees.

4.2. Analysis of Classification and Regression Trees

In this study, data analysis was used using the CART (Classification and Regression Trees) method, namely:

a. Dividing sample data into two parts, namely training data and testing data

In the process of analyzing data this research uses the division of training data (90%) and testing data (10%) where the total data is 132 sample data, for training data as many as 118 samples and data testing as many as 14 samples.

b. Formation of classification trees

1) The selection of sorters with the Gini index selection rules is further filtered based on the goodness of split criteria

For the selection of sorters, a variety or category is carried out on variables. The selection of the sorter using the Gini index obtained by the prospective sorter is shown in Table 8.

Table 8. Variable categories

Knot		Rain Category					Sum	Total
		Bright Overcast	Rain Light	Rain Keep	Rain Dense	Rain Dense Very		
PC_1	≥ 0.1868	6	8	14	4	2	34	118
	< 0.1868	60	23	1	0	0	84	
PC_2	≥ 1.2858	3	14	7	2	1	27	118
	< 1.2858	63	17	8	2	1	91	
PC_3	≥ -0.5082	48	26	15	4	0	93	118
	< -0.5082	18	5	0	0	2	25	

2) Gain Information

In finding the value of gain information, first look for the probability value of each node to make it easier. As in the first node, equations (4) and (5) are used as follows:

The same step is repeated according to the above work and the results are presented in the Table 9.

Table 9. Node Probability Calculation

Knot	P_L	P_R	Class	$p(j\|t_L)$	$p(j\|t_R)$
1	0.7119	0.2881	Sunny Cloudy	0.7143	0.1765
			Light rain	0.2738	0.2353
			moderate rain	0.0119	0.4118
			cloudburst	0	0.1177
			It rained heavily	0	0.0588
2	0.7712	0.2289	Sunny Cloudy	0.6923	0.1111
			Light rain	0.1868	0.5185
			moderate rain	0.0879	0.2593
			cloudburst	0.0219	0.0741
			It rained heavily	0.0219	0.0370
3	0.2119	0.7881	Sunny Cloudy	0.72	0.5161
			Light rain	0.2	0.2796
			moderate rain	0	0.1613
			cloudburst	0	0.0430
			It rained heavily	0.08	0

Table 9 will be used to detect each attribute that has information based on a certain class, namely the value of gain information. In equation (8) it will be used to get the gain information value for each attribute with the following formula:

$$GI(t) = -\sum_{j}^{n} P(j|t) log_2 P(j|t)$$
$$GI(P_R) = -(0.1765)^2\, log(0.1765) + (-(0.2353)^2 log\ (0.2353)) + (-(0.4118)^2\, log(0.4118)) + (-(0.1177)^2\, log(0.1177)) + (-(0.0588)^2\, log(0.0588))$$
$$= 0.4672$$

The same step is repeated according to the above work and the results are presented in Table 10.

Table 10. Gain Information Calculation

Knot	P_R	P_L	$GI(P_R)$	$GI(P_L)$	Average
1	≥ 0.1868	< 0.1868	0.4672	0.3884	0.4278
2	≥ 1.2858	< 1.2858	0.4519	0.3712	0.4116
3	≥ 0.5082	< 0.5082	0.4745	0.3619	0.4182

Table 10 can be seen that the value of gain information that has the most information is the attribute of node 1 by looking at the average of each attribute above 41%, then the other attributes are not much different in value so that all attributes can be included in the data analysis process.

3) Gini Index

Next, the value of the gini index is obtained using the value in Table 11. For each node, equation (9) is used so that the Gini index value for the first candidate node is obtained as follows:

$$i(t) = \sum_{j=1}^{k} p(j|t)(1 - p(j|t))$$
$$= 1 - (0.7119)^2 - (0.2881)^2 = 0.4102$$

The search for the value of the Gini index is performed for all prospective nodes which can be seen in the Table 11.

Table 11. Gini Index Details

Knot	P_R	P_L	$i(t)$
1	≥ 0.1868	< 0.1868	0.4102
2	≥ 1.2858	< 1.2858	0.3529
3	≥ 0.5082	< 0.5082	0.3339

4) Goodness of split

Then a node candidate process is carried out which will be the parent node or sorter or root node with the criteria of goodness of split. The calculation of the goodness of split on the first node candidate using equation (10) is obtained:

$$i(t_R) = 1 - (0.1765)^2 - (0.2353)^2 - (0.4118)^2 - (0.1177)^2 - (0.0588)^2 = 0.7266$$
$$i(t_L) = 1 - (0.7143\ )^2 - (0.2738)^2 - (0.0119)^2 - (0)^2 - (0)^2 = 0.4147$$
$$\Delta i(s,t) = i(t) - P_L i(t_L) - P_R i(t_R) = 0.4102 - (0.7119)(0.4147) - (0.2881)(0.7266)$$
$$= (-0.0944)$$

The goodness of split value search is performed for all prospective nodes obtained in the Table 12.

Table 12. Goodness of split

Knot	$\Delta i(s,t)$	Criterion Goodness
1	-0.0944	1
2	-0.1625	2
3	-0.2529	3

In Table 12, it can be seen that the goodness of split value that meets as the candidate node with the highest value is the candidate of the 1st node, which is -0.0944, then the candidate of the 1st node will be the root node or parent node, PC_1 . where the 1st node will branch into the left branch is the attribute: PC_1 (< 0.1868) and the right branch of the attribute ($PC_1 \geq 0.1868$) which is shown in the following figure:

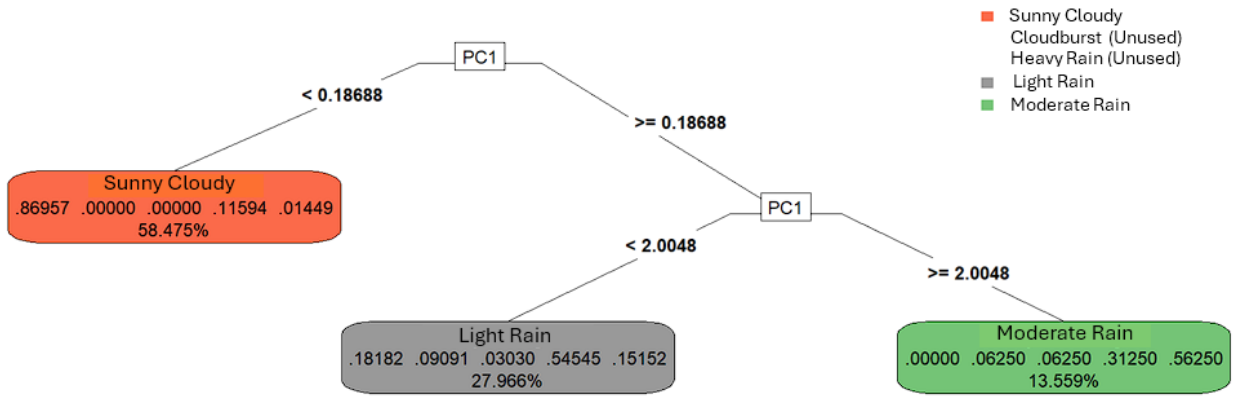

Figure 4. Root node separation process

Based on the structure of the tree, it can be concluded that most of the data falls into the category of sunny and cloudy, followed by light rain and moderate rain. The other two categories, namely heavy rain and heavy rain, are marked as unused because they do not appear in the final result of the decision tree. This suggests that the two categories do not contribute significantly to the separation of data by variable PC_1, so they are not used by the CART algorithm in the formation of this decision tree structure. So that the maximum decision tree can be formed as follows:

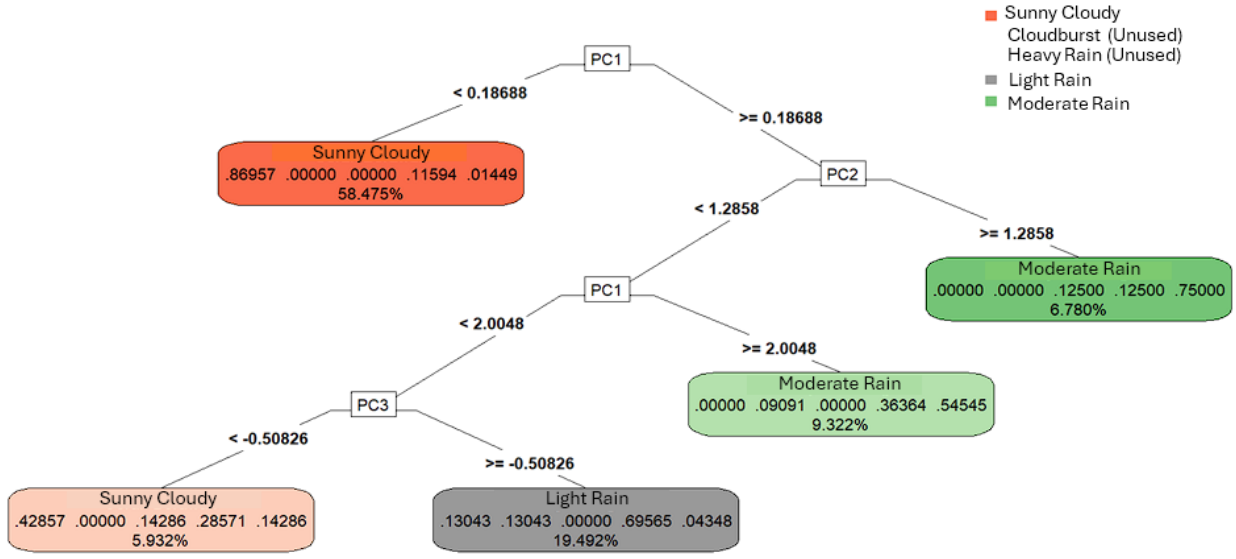

Figure 5. Shape of the Maximum Decision Tree

Based on the structure of this decision tree, it can be seen that the majority of the data falls into the category of sunny and cloudy (58.475%), followed by light rain (19.492%), moderate rain (9.322% and 6.780% knots combined), and the rest are sunny and cloudy from the deepest node (5.932%). The categories of heavy rain and heavy rain are once marked as unused because they do not appear in the tree structure. This indicates that the two categories do not provide sufficient separating power to the data based on the main variables (PC_1, PC_2 , and PC_3), so they are not selected by the CART algorithm in the process of forming the Decision tree

So, in detail the shape of the tree Figure 5 can be summarized in the Table 13.

Table 13. Table of nodes in the shape of the decision tree

Knot	Name	Information
1	PC_1	Non-terminal nodes (root node)
2	$PC_1 < 0.1868$	Terminal node (sunny cloudy)
3	$PC_1 \geq 0{,}1868$	Non-terminal nodes
4	$PC_2 < 1.2858$	Non-terminal nodes
5	$PC_2 \geq 1.2858$	Terminal node (moderate rain)
6	PC_1 $2.0048<$	Non-terminal nodes
7	$PC_1 \geq 2.0048$	Terminal node (moderate rain)
8	$PC_3 \geq -0.5082$	Terminal node (light rain)
9	$PC_3 < -0.5082$	Terminal node (sunny cloudy)

It can be seen in Table 13 that the terminal node or terminal node is the last node in the decision tree that does not undergo further separation because the impurity value is quite low according to the goodness of split criterion. Based on the structure of the decision tree formed, terminal nodes are indicated by nodes 2, 5, 7, 8, and 9 and non-terminal nodes, namely nodes 1, 3, 4, and 6.

5) Class Labeling

For class labelling, it is carried out according to equation (6) where based on the rule of the maximum number of each class in the bound variable or response. Sourced from Figure 5 of the decision tree shape, the class label marking for each node or node, specifically the class label marking on the terminal node is presented in the following table:

Table 14. Table nodes in the shape of the Decision tree

Knot	Name	Class Labels	Percentage
1	$PC_1 < 0.1868$	Sunny cloudy	58.475%
2	$PC_1 \geq 0.1868$, $1.2858 PC_2 \geq$	Moderate rain	6.475%
3	$PC_1 \geq 0.1868$, < 1.2858, $2.0048 PC_2 PC_1 \geq$	Moderate rain	9.322%
4	$PC_1 \geq 0.1868$, < 1.2858, PC_2 2.0048, $< PC_1 \geq PC_3 - 0.5082$	Sunny cloudy	5.932%
5	$PC_1 \geq 0.1868$, < 1.2858, PC_2 2.0048, $PC_1 \geq PC_3 \geq -0.5082$	Light rain	19.492%

From Table 14, it is clear that the label of the cloudy sunny class is at nodes 1 and 4, for the medium rain class label is at nodes 2 and 3 and for the light rain class label is at node 5.

6) Classification Tree Cutting

Pruning trees to get the best tree. Tree branches that can be pruned are those that have a minimum complexity parameter value. So, the value of the complexity parameter is obtained in Table 15.

Table 15. Complexity Parameter Values

N0	CP	Xerror	Nsplit
1	0.326923	1.00000	0
2	0.096154	0.71154	1
3	0.038462	0.75000	2
4	0.019231	0.80769	3
5	0.000001	0.78846	4

It can be seen in Table 15 that the complexity parameter value for the split in order is 0.326923; 0.096154; 0.038462; 0.019231; 0.000001. The minimum complexity parameter value chosen is 0.096154, because it provides the lowest cross-validation error value (error), and produces a tree with 1 split. So that the optimal decision tree structure that has been pruned is obtained as follows:

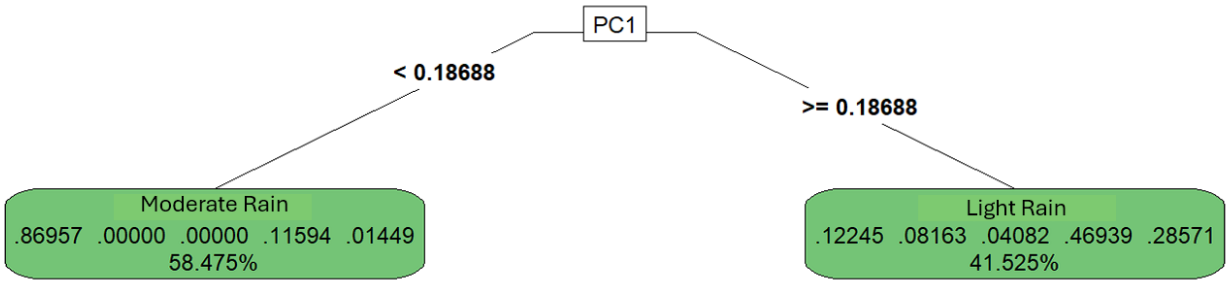

Figure 6. Optimal Decision Tree Shape

Seen in Figure 6, it shows the results of the maximum decision tree from the CART model which uses the main variable, i.e., to separate the data into two classes. The separation is carried out based on the threshold value of 0.18688. If the value PC_1 is less than 0.18688, then the data is classified into the left node with the class label "Sunny Cloudy". At this node, the majority of data comes from sunny and cloudy weather conditions with a percentage of 86.957%, and the total percentage of data entering this node is 58.475%. This indicates that the left node has a high level of purity (impurity).

Meanwhile, if the value PC_1 is greater than or equal to 0.18688, the data is classified into the right node with the class label "Light Rain". At this node, the class distribution was dominated by light rain at 46.939%, followed by sunny cloudy at 12.245% and other classes with smaller proportions. The percentage of total data that goes into this node is 41.525%. Although the right node has a lower level of purity than the left node, it still provides useful information for classification. Overall, this decision tree divides data simply but quite effectively with just one main variable, PC_1.

7) Results of the Classification Decision

The accuracy level of the optimal tree classification results resulting from the training data can be calculated based on Table 16.

Table 16. Classification of Rainfall Training Data on Optimal Trees

Current	Predictions					Accuracy Classification (%)	Error Classification
	Bright Overcast	Rain Light	Rain Keep	Rain Dense	Rain Dense Very		
Sunny Cloudy	60	6	0	0	0	90.91%	6
Light Rain	8	23	0	0	0	74.19%	8
Moderate Rain	1	14	0	0	0	0%	15
Cloudburst	0	4	0	0	0	0%	4
Heavy Rain	0	2	0	0	0	0%	2

Based on Table 16, the classification error of the observation class occurred in all classes, where the moderate rain class, the heavy rain class and the heavy rain class once produced a classification accuracy of 0%. This means that there is not 1 class 5 observation data that is properly classified. The greatest classification accuracy occurred in class 1 (sunny and cloudy) with a percentage of 90.91%.

Using the information in Table 14, the accuracy of the classification of training data can be calculated as follows

$$1 - \text{APER} = \left(1 - \frac{6 + 8 + 15 + 4 + 2}{118}\right) \times 100\% = 70.34\%$$

The results of the calculation of the accuracy of the training data classification were 70.34%. This means that the optimal classification tree is able to classify rainfall observations into category classes correctly by 70.34%.

The optimal classification tree that is formed needs to be validated to know if it is feasible and can be used to classify new data. The accuracy level of the optimal tree classification results resulting from the testing data can be calculated based on Table 17.

Table 17. Classification of Rainfall Testing Data on Optimal Trees

Current	Predictions					Accuracy Classification (%)	Error Classification
	Bright Overcast	Rain Light	Rain Keep	Rain Dense	Rain Dense Very		
Sunny Cloudy	8	0	0	0	0	100%	0
Light Rain	2	2	0	0	0	50%	2
Moderate Rain	0	2	0	0	0	0%	2
Cloudburst	-	-	-	-	-	-	-
Heavy Rain	-	-	-	-	-	-	-

So that the accuracy of classification for data testing can be calculated as follows:

$$1 - \text{APER} = \left(1 - \frac{0 + 2 + 2 + 0 + 0}{14}\right) \times 100\% = 71.43\%$$

The calculation results showed that the accuracy of the classification of testing data was 71.43%. This means that the model has the ability to correctly predict new data as much as 71.43% of all tested data.

5. Conclusions

Based on the analysis carried out, the following conclusions were obtained: (1) the results of dimension reduction using the Principal Component Analysis (PCA) method showed that of the initial seven variables, only three main components ($PC_1, PC_2,$ and PC_3) were retained because they had eigenvalues greater than 1 and were able to explain the data variance significantly, (2) the results of the accuracy of the rainfall forecast classification using the Classification and Regression Trees (CART) method showed quite good performance. The decision tree model that was formed resulted in an accuracy rate of 72.34% in training data. This means that the model is able to classify most of the training data into the correct rainfall category, (3) the performance of the CART model after being validated with data testing showed consistent results. In the data testing, the model was able to achieve an accuracy level of 71.43%, which shows that the model has good generalization ability to new data and does not experience overfitting.

References

Aminuddin, J., (2016). Pengaruh Kecepatan Angin Terhadap Evapotranspirasi Berdasarkan Metode Penman Di Kebun Stroberi Purbalingga. *Elkawnie: Journal of Islamic Science and Technology*, *2*(1), 21–28. www.jurnal.ar-raniry.com/index.php/elkawnie

Azmi. B., Hermawan. A., & A. . D. (2023). Analisis Pengaruh Komposisi Data Training dan Data Testing pada Penggunaan PCA dan Algoritma Decision Tree untuk Klasifikasi Penderita Penyakit Liver. *JTIM : Jurnal Teknologi Informasi Dan Multimedia*, *4*(1), 281–290. https://doi.org/10.35746/jtim.v4i4.298

Azmi, U. (2017). *Prediksi curah hujan melalui Model Output Statistics menggunakan Classification And Regression Trees dengan Pre-Processing Principal Component Analysis* [Skripsi, Institut Teknologi Sepuluh Nopember, 2017]. http://repository.its.ac.id/3294/

Dwirani, F. (2019). Menentukan stasiun hujan dan curah hujan dengan metode polygon thiessen daerah kabupaten lebak. *Jurnal Lingkungan Dan Sumberdaya Alam (JURNALIS)*, *2*(2), 139–146. https://ejournal.lppm-unbaja.ac.id/index.php/jls/article/view/674

Evalina, Y., & Sinambela, S. (2008). *Penerapan Metode Pohon Klasifikasi Dengan Algoritma CART pada Data Status Daerah Kabupaten Di Indonesia*. Skripsi, Tidak Diterbitkan, Institut Pertanian Bogor, 2008.

Firdaus, R. F. (2022). *Prediksi Curah Hujan Menggunakan Metode Long Short Term Memory (Studi Kasus : Kota Bandung)* [Skripsi, Universitas Islam Indonesia]. https://dspace.uii.ac.id/handle/123456789/42746

Handoko, S., Fauziah, F., & Handayani, E. T. E. (2020). Implementasi Data Mining Untuk Menentukan Tingkat Penjualan Paket Data Telkomsel Menggunakan Metode K-Means Clustering. *Jurnal Ilmiah Teknologi Dan Rekayasa*, *25*(1), 76–88. https://doi.org/10.35760/tr.2020.v25i1.2677

Hasanah, M. A., Soim, S., & Handayani, A. S. (2021). Implementasi CRISP-DM Model Menggunakan Metode Decision Tree dengan Algoritma CART untuk Prediksi Curah Hujan Berpotensi Banjir. *Journal of Applied Informatics and Computing*, *5*(2), 103–108. https://doi.org/10.30871/jaic.v5i2.3200

Imah, E. M., & Basaruddin, T. (2011). Klasifikasi Beat Aritmia Pada Sinyal EKG Menggunakan Fuzzy Wavelet Learning Vector Quantization. *Jurnal Ilmu Komputer Dan Informasi*, *4*(1), 1–9.

Juliati, D. (2023). Analisis Karakteristik Curah Hujan Dengan Menggunakan Klasifikasi Schmidr-Fergusson Di Kota Makassar. *Jurnal Sains Dan Pendidikan Fisika (JSPF)*, *2*, 229–235. https://pdfs.semanticscholar.org/0f27/312229fc0a96555b3395beb96aebe38cbe18.pdf

Koentjoro, Y. (2014). Dampak Perubahan Pola Curah Hujan Terhadap Pertumbuhan Tanaman Pangan Di Kabupaten Pasuruan. In *Academia.Edu*. https://www.academia.edu/download/39359102/Tugas_Bu._Kokom_Perubahan_pola_curah_hujan.pdf

Malino, C. R., Arsyad, M., & Palloan, P. (2021). Analisis Parameter Curah Hujan dan Suhu Udara di Kota Makassar Terkait Fenomena Perubahan Iklim. *Jurnal Sains Dan Pendidikan Fisika (JSPF)*, *17*(2), 139–145.

Maraun, D., & Widmann, M. (2017). Model Output Statistics [Skripsi, Institut Teknologi Sepuluh Nopember]. In *Statistical Downscaling and Bias Correction for Climate Research*. https://doi.org/10.1017/9781107588783.013

Maulidani S, S., Ihsan, N., & Sulistyawati. (2015). Analisis Pola Dan Intensitas Curah Hujan Berdasakan Data Observasi Dan Satelit Tropical Rainfall Measuring Missions (TRMM) 3B42 V7 Di Makassar. *Jurnal Sains Dan Pendidikan Fisika (JSPF)*, *11*(1), 98–103.

Mulyana, E. (2002). Pengaruh Dipole Mode Terhadap Curah Hujan Di Indonesia. *Jurnal Sains & Modifikasi Cuaca*, *3*(1), 39–43.

Musfiroh, M., Novitasari, D. C. R., Intan, P. K., & Wisnawa, G. G. (2023). Penerapan Metode Principal Component Analysis (PCA) dan Long Short-Term Memory (LSTM) dalam Memprediksi Prediksi Curah Hujan Harian. *Building of Informatics, Technology and Science (BITS)*, *5*(1), 1–11. https://doi.org/10.47065/bits.v5i1.3114

Nensi, T., Ihsan, N., & Patandean, A. . (2016). hujan bulanan minimum yaitu pada bulan Juni , Juli atau Agustus dan puncak maksimum musim hujan yaitu pada bulan Proses Pengambilan Data . Proses Analisis Data. *Jurnal Sains Dan Pendidikan Fisika.*, *12*(3), 324–329.

Nuraliza, H., Pratiwi, O. N., & Hamami, F. (2022). Analisis Sentimen IMBd Film Review Dataset Menggunakan Support Vector Machine (SVM) dan Seleksi Feature Importance. *Jurnal Mirai Manajemen*, *7*(1), 1–17.

Prakoso, D. (2018). Analisis pengaruh tekanan udara, kelembaban udara dan suhu udara terhadap tingkat curah hujan di kota semarang. *Jurnal Universitas Negeri Semarang*, 1–77. http://lib.unnes.ac.id/id/eprint/36742

Pratiwi, F. E., & Zain, I. (2014). Klasifikasi pengangguran terbuka menggunakan CART (Classification and regression tree) di Provinsi Sulawesi Utara. *Jurnal Sains Dan Seni ITS*, *3*(1), D54–D59. http://www.ejurnal.its.ac.id/index.php/sains_seni/article/view/6129

Ramadhan, I. A., & Septiadi, D. (2024). *The Utilization of Model Output Statistic (MOS) in Improving Weather Prediction Model Accuracy of Integrated Forecasting System (IFS)*. *16*(2).

Safitri, R., & Sutikno. (2012). Model Output Statistics dengan Projection Pursuit Regression untuk Meramalkan Suhu Minimum, Suhu Maksimum, dan Kelembapan. *Jurnal Sains Dan Seni ITS*, *1*(1), 296–301. http://ejurnal.its.ac.id/index.php/sains_seni/article/view/2070

Saleh, A., & Nasari, F. (2018). Penerapan Equal-Width Interval Discretization Dalam Metode Naive Bayes Untuk Meningkatkan Akurasi Prediksi Pemilihan Jurusan Siswa. *Masyarakat Telematika Dan Informasi : Jurnal Penelitian Teknologi Informasi Dan Komunikasi*, *9*(1), 1. https://doi.org/10.17933/mti.v9i1.113

Saputra, K. A., Hardinata, J. T., Lubis, M. R., Andani, S. R., & Saragih, I. S. (2020). Klasifikasi Algoritma C4.5 Dalam Penerapan Tingkat Kepuasan Siswa Terhadap Media Pembelajaran Online. *Kajian Ilmiah Informatika Dan Komputer*, *1*(3), 113–118. https://djournals.com/klik

Sari, Irma Purnamasari, A., & Rinaldi Dikanda, A. (2023). Implementasi Data Mining Dalam Menentukan Pola Penjualan Vitamin Blackmores. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *7*(2), 1265–1269. https://doi.org/10.36040/jati.v7i2.6534

Sudipa, I. G. I., Darmawiguna, I. G. M., Dendi, I. M., & Sanjaya, M. (2024). *Buku ajar data mining*. Jambi : PT. Sonpedia Publishing Indonesia.

Tanjung, R. H., & Kartiko. (2017). Penerapan Metode CART (Classification and Regression Trees) Untuk Menentukan Faktor-faktor Yang Mempengaruhi Pembayaran Kredit Oleh Nasabah (Studi Kasus Bank BRI Unit Aek Tarum-Sumatera Utara). *Jurnal Statistika Industri Dan Komputasi*, *2*(2), 78–83.

Tinungki, G. M., & Sunusi, N. (2018). Penerapan Sparse Principal Component Analysis dalam Menghasilkan Matriks Loading yang Sparse. *Jurnal Matematika Statistika Dan Komputasi*, *15*(2), 44. https://doi.org/10.20956/jmsk.v15i2.5713

Wijaya, Y. F., & Triayudi, A. (2023). Perbandingan Algoritma Klasifikasi Data Mining Pada Prediksi Penyakit Diabetes. *Journal of Computer System and Informatics (JoSYC)*, *5*(1), 165–174. https://doi.org/10.47065/josyc.v5i1.4614