



Water Level Prediction in Water Shed Management Utilizing Machine Learning

Dr. K. Balasubramanian^{1*}, K. Shobiya²

^{1*}Associate Professor, CSE, E.G.S. Pillay Engineering College, Nagapattinam, India.

²P.G Scholar E.G.S Pillay Engineering College Nagapattinam, India.

Email: ²shobiyakrishnan@gmail.com

Corresponding Email: ^{1*}bala@egspec.org

Received: 05 June 2021

Accepted: 20 August 2021

Published: 18 September 2021

Abstract: Due to uneven rainfall, nowadays the amount of rain to be showered in a month is getting showered in few days. The massive wastage of water occurs due to irregular heavy rainfall and water released from dams. To avoid this, the proposal suggests an idea to develop a watershed and to predict the water level measurement by Bayesian classification, clustering, and optimization techniques. Artificial Neural Network is one of the previous techniques used to predict water level which gives approximate result only. To overcome the disadvantage, this proposal suggests an idea to develop the watershed by using different machine learning techniques. The level of water that can be stored is calculated using Bayes Network which will classify the data into labels according to the condition of the capacity of the minimum and maximum storage level of the watershed. The standardized data considered for the classification are normalized using the z-score normalization. Classification will represent the result by means of the instances that are correctly classified. The output of the classified data is fed into clustering algorithm where the labels are grouped into different clusters. The K-Mean algorithm is utilized for clustering which iteratively assign data point to one of the k group according to the given attribute. The clustered output gives the result of how many instances are correctly clustered. The clustered output will be refined for further process such that the data will be extracted as ordered dataset of year wise and month wise data.

For the extracted data gradient descent algorithm is applied for reducing the error and predicting the amount of water stored in watershed for upcoming years by means of calculating the actual and prediction value. Later the result will be visualized in the form of graph. The obtained output is considered as an input for posterior probability that uses J48 algorithm which gives the result of probability of event happened after all the evidence is taken for consideration and gives the accurate result. The above methodology provides high performance and efficient result.

Keywords: Bayes Network, K-Means, Gradient Descent, J48 Algorithm.



1. INTRODUCTION

Water scarcity is the major issue in today's world. Due to uneven rainfall the amount of water to be showered in months is getting showered in days. The massive wastage of water occurs due to uneven rainfall and water released from dams because of this issue each district, town and village suffer by water scarcity. We are pushed to a situation where we must save water in every possible way. To save water and to avoid wastage of water this proposal suggests an idea to develop a watershed in town situated in Thanjavur district which has three water resources such as, and to predict the water level measurement in watershed. A watershed is an area of land that stores rainwater into one location such as a stream, lake, or wetland. These water sheds are used for supplying the drinking water, water for agriculture and provides habitation to various plants and animals.

Artificial neural networks back propagation method is one of the previous techniques used to predict the water level. It gives less accurate and approximate results. To overcome this disadvantage and to increase the accuracy, this proposal uses machine learning techniques such as Bayesian network classification, k mean clustering, gradient descent, and posterior probability to find the monthly quantity of water that can be stored in the watershed which benefits the 106 taluks in Kumbakonam town. The overall process has three major components such as representation, evaluation, and optimization.

1.1 Data preprocessing

Past 15 years dam outlet readings and rainfall readings at a particular region is taken as input data set. The preprocessing of data is done by standardization and normalization. The rainfall readings in millimeter standard and dam outflow readings in TMC standard are standardized into cubic cm standard. The standardized readings are normalized by z-score normalization. Normalization is done to make the standardized data into a particular range.

1.2 Representation

Bayesian network classification is used for representing the conditional dependencies of a set of data. It is a probabilistic graphical model. It is represented by directed acyclic graph. It is used for finding the probabilistic relationship between data where events are represented as nodes and edges represent probabilistic relation between events. If the nodes are not connected it represents the conditional independency of variables. Since the proposal being a prediction model, Bayesian network is most suitable algorithm.

K means clustering is used to cluster the data. This proposal uses large data set, so to process the data quickly and efficiently fast algorithm like K mean is used. K clusters are partitioned from number of observations in which each cluster contains observation.

1.3 Optimization

Gradient descent is used for the optimization. It is computationally efficient which generates stable convergence and stable error gradient. It reduces prediction error and improves prediction accuracy. Another name of gradient descent. Machine learning problems such as linear regression can be solved by Gradient descent. The relationship between two or more independent variables and one continuous variable can be explained by multiple linear regression. It is used to forecast values for future years.



1.4 Evaluation

Posterior probability can be obtained by updating the prior probability using Bayes' theorem. The posterior probability of a random event is captured after the evidence or background is considered into account. The posterior probability is the probability of event B occurring in which the event A had occurred already. The posterior probability is directly proportional to the product of likelihood and prior probability. By using these techniques more accurate and optimized results can be obtained.

Over 95 TMC to 110 TMC water is released from dam but not even a 4 TMC is reaching the surrounding towns which is stated in many research and records. This proposal gives a gist to save the water in the form of watershed which gives good level of storage of water and this watershed development appears to help in terms of groundwater recharge which benefits the farmer as well as for domestic purpose.

2. PREVIOUS WORK

2.1 Artificial Neural network

It is one of the methods used in predicting rainfall which gives only the approximate results. To predict exact results existing techniques can be enhanced by the addition of some techniques like Regression, clustering, and optimization process. ANN is the collection of connected nodes called artificial neurons. Neurons may be implemented in multiple layers, where each layer process signal and performs some specific task. This processed information is then passed to next node as a signal (weight), which leads to desired output. ANN automatically learns from its training set which leads to its popular usage. ANN performance can be improved by number of hidden layers (up to 2) and number of nodes implemented at each layer.

In previous work the system uses back propagation method to handle local minima phenomenon that happen in watershed graphs. This system also process data using previous work method without addition of runoff drainage rate. The predictions result from both models then compared by its accuracy and precision rate. Model scheme for previous work method can be seen on Figure 1.1.

The optimum prediction for each method is generated from learning rate, hidden node, and additional momentum rate if necessary. Proposed research manages training and testing with various conditions to analyze impact of variable changes to the data processing. Basically, the amount of hidden node should be between input node and output node. Optimum variables (learning rate, hidden node, and momentum) for water level prediction.

3. PRESENT WORK

To overcome this disadvantage of the existing system the proposed system in Fig. 3.1 uses Bayesian network which is the representation method that used to find probabilistic relationship between nodes (states of the world). Gradient descent is used to reduce the prediction error and improves prediction accuracy. Posterior probability is used to make the predicted water level more accurate.

The acquired by the past rainfall readings and dam outlet readings are standardized by Z-score normalization. The relationship between is calculated by correlations. Bayesian Network classification is used for finding probability relationship and data are clustered using

K-Means clustering. Gradient Descent reduces prediction error and improves prediction accuracy. Posterior probability is applied to make the result more accurate.

3.1 Architecture Diagram

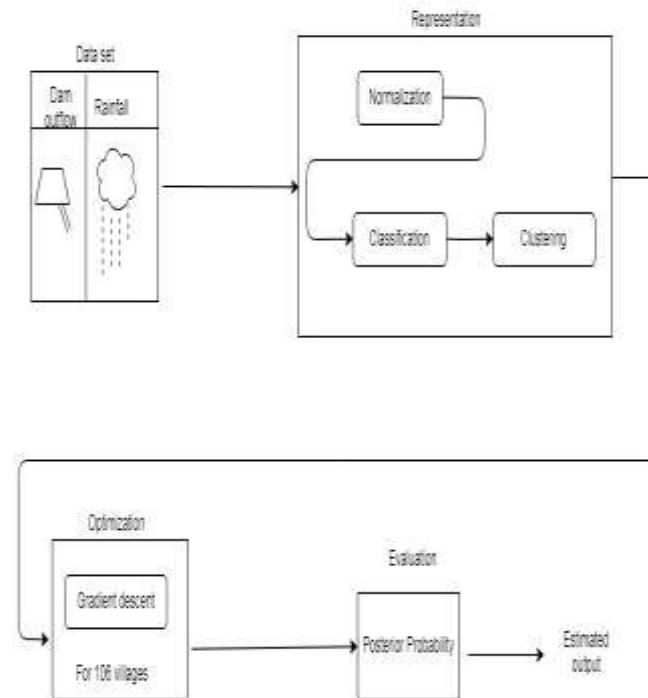


Fig. 3.1 Architecture diagram

4. DESIGN AND IMPLEMENTATION

4.1 Standardization

The original rainfall data is in millimeter. The Dam reading is in cusec. Then both data must be converted into common unit. It must be converted into cm^3 .

4.1.1 Dam data conversion (TMC to cm^3)

$\text{TMC} \rightarrow \text{m}^3 \rightarrow 1 \rightarrow \text{cm}^3$

$$1 \text{ TMC} = 28316846.592 * 1000 * 1000 \text{ cm}^3 \quad (4.1)$$

Where $1 \text{ TMC} = 28316846.592 \text{ m}^3$

$$1 \text{ m}^3 = 1000 \text{ l}$$

$$1 \text{ l} = 1000 \text{ cm}^3$$

4.1.2 Rain data conversion (mm to cm^3):

$\text{mm} \rightarrow \text{cm} \rightarrow \text{cm}^3$

$$1 \text{ mm} = 0.1 * (\text{cm}) \quad (4.2)$$

Where $1 \text{ mm} = 0.1 \text{ cm}$

$$0.1 \text{ cm} = (0.1)^3 \text{ cm}^3$$

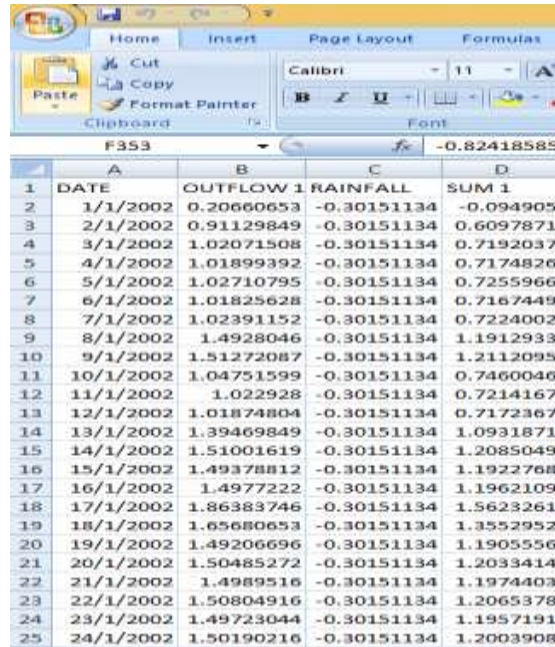
4.1.3. Normalization using Z-Score:

$$V^i = \frac{V - \bar{A}}{\sigma A} \quad (4.3)$$

Where \bar{A} is the Mean.

σ_A is the square root of the variance.

$$\text{Variance} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \quad (4.4)$$



	A	B	C	D
1	DATE	OUTFLOW 1	RAINFALL	SUM 1
2	1/1/2002	0.20660653	-0.30151134	-0.094905
3	2/1/2002	0.91129849	-0.30151134	0.6097871
4	3/1/2002	1.02071508	-0.30151134	0.7192037
5	4/1/2002	1.01899392	-0.30151134	0.7174826
6	5/1/2002	1.02710795	-0.30151134	0.7255966
7	6/1/2002	1.01825628	-0.30151134	0.7167449
8	7/1/2002	1.02391152	-0.30151134	0.7224002
9	8/1/2002	1.4928046	-0.30151134	1.1912933
10	9/1/2002	1.51272087	-0.30151134	1.2112095
11	10/1/2002	1.04751599	-0.30151134	0.7460046
12	11/1/2002	1.022928	-0.30151134	0.7214167
13	12/1/2002	1.01874804	-0.30151134	0.7172367
14	13/1/2002	1.39469849	-0.30151134	1.0931871
15	14/1/2002	1.51001619	-0.30151134	1.2085049
16	15/1/2002	1.49378812	-0.30151134	1.1922768
17	16/1/2002	1.4977222	-0.30151134	1.1962109
18	17/1/2002	1.86383746	-0.30151134	1.5623261
19	18/1/2002	1.65680653	-0.30151134	1.3552952
20	19/1/2002	1.49206696	-0.30151134	1.1905556
21	20/1/2002	1.50485272	-0.30151134	1.2033414
22	21/1/2002	1.4989516	-0.30151134	1.1974403
23	22/1/2002	1.50804916	-0.30151134	1.2065378
24	23/1/2002	1.49723044	-0.30151134	1.1957191
25	24/1/2002	1.50190216	-0.30151134	1.2003908

Fig. 4.1 Normalized values of dam outflow and rainfall

4.2 Classification using Bayes

Classification was applied for year wise data. Here Bayesian Network is used. It is used to find probabilistic relation between events. Bayesian network is used when outcome is uncertain.

4.2.1 Steps

xlsx → csv (comma separated value) file

csv → .arff (Attribute Relation File Format) file.

4.2.2 Based on Water storing capacity of a watershed

Min: 2.19 TMC

Max: 95.66 TMC

4.2.3 Data classified into two types

Value (Dam outflow + rain data) >=

-0.70711 (Z- score value) → H (4.5)

Value (Dam outflow + rain data) <

-0.70711 (Z- score value) → L (4.6)

4.2.4 For Dam Outflow 2 Timings are used

capacity1 → 8.00 AM

capacity2 → 4.00 PM

4.2.5 Result obtained for Bayes net

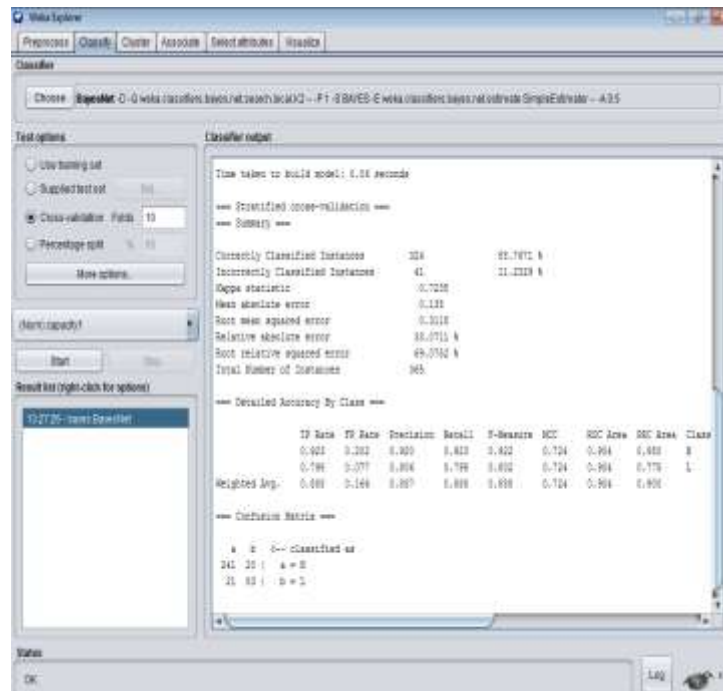


Fig. 4.2 Bayesian Network gives accuracy of classification.

The obtained result shows the accuracy of classification in percentage values

Total Number of instances: 364

Correctly Classified: 324 i.e.,89%

Incorrectly Classified: 41 i.e.,11%

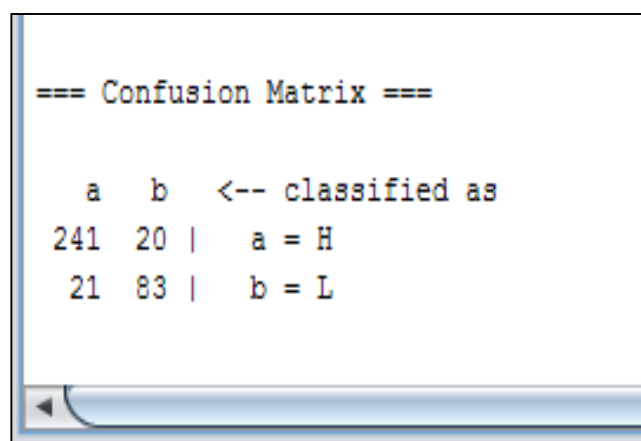


Fig.4.3 Confusion Matrix in Bayes net.

Here:

- 241 represents correctly classified H label.
- 20 represents incorrectly classified H label as L label.
- 83 represents correctly classified L label.

- 21 represents incorrectly classified L label as H label.

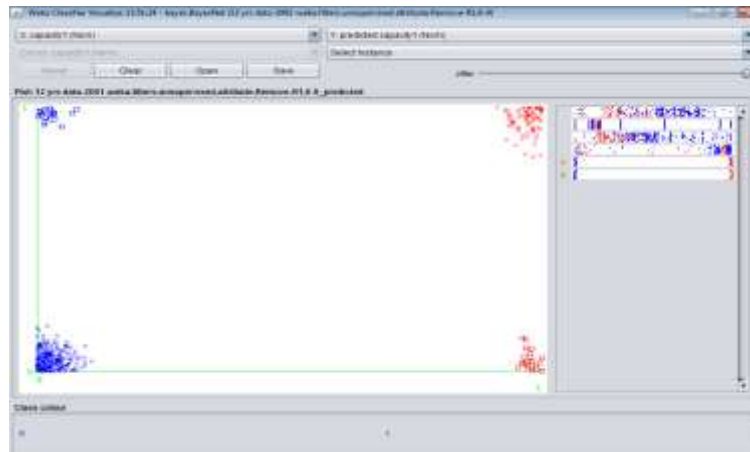


Fig. 4.4 Classification Visualization window

Indicators:

H →blue color

L →red color.

x →correctly classified.

box →incorrectly classified.

The below mentioned Table 1 shows the accuracy percentage of capacity 1 and capacity 2 of year wise data.

Table I Classification Accuracy Table in %

YEAR	CAPACITY 1	CAPACITY 2
2001	88.7671	100
2002	100	99.726
2004	99.7268	100
2005	99.1098	99.7033
2006	100	94.2466
2007	100	100
2008	100	100
2009	99.726	100
2010	100	99.726
2012	100	100
2013	100	99.6711
2014	99.726	100
2015	100	100
2016	98	99.22
2017	97	100
2018	99.54	98
2019	100	100
2020	98.32	96.33

4.3 Clustering using K-means

4.3.1 Steps

xlsx → csv (comma separated value) file

csv → .arff (Attribute Relation File Format) file.

4.3.2 Finding the range for “H” label

capacity 1:

Min of “H”: 6.816772

Max of “H”: -2.77594

Average: 2.020414

capacity 2:

Min of “H”: 6.997528

Max of “H”: -0.70674

Average: 3.145394

4.3.3 Data for “H” label is divided into 2 groups as

Capacity1:

value \geq 2.020414 → HH

value $<$ 2.020414 → HL

Capacity2:

value \geq 3.145394 → HH

value $<$ 3.145394 → HL

3 groups of data such as HH (High), HL (Medium), L (Low) is used in clustering. Clustering is done for Whole data.

4.3.4 Result obtained for K-means

Simple K-Means algorithm is used for clustering.

Number of clusters=3.

Here class label is ignored in attributes list.

Number of iterations: 8

Here initial cluster point for 3 clusters is shown.

Final cluster point for data set is shown in the below cluster information.

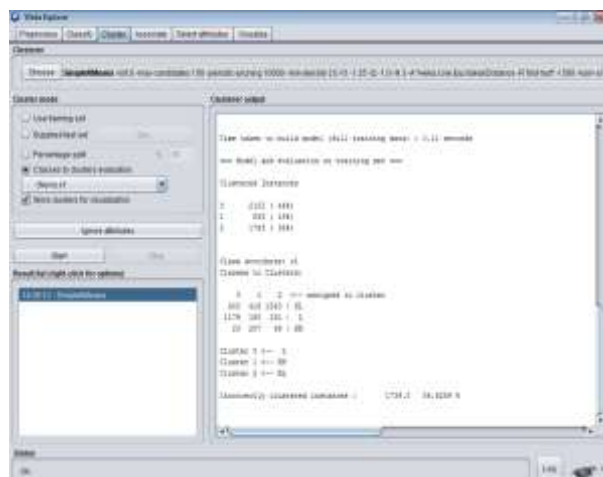


Fig. 4.5 Clustered Information

Cluster 0 → 2102 data (both correct and incorrect data)

Similarly, for cluster 1 and 2.

Cluster 0 → L

Cluster 1 → HH

Cluster 2 → HL

Out of 4748 instances 1739 instances i.e., 37% are incorrectly clustered

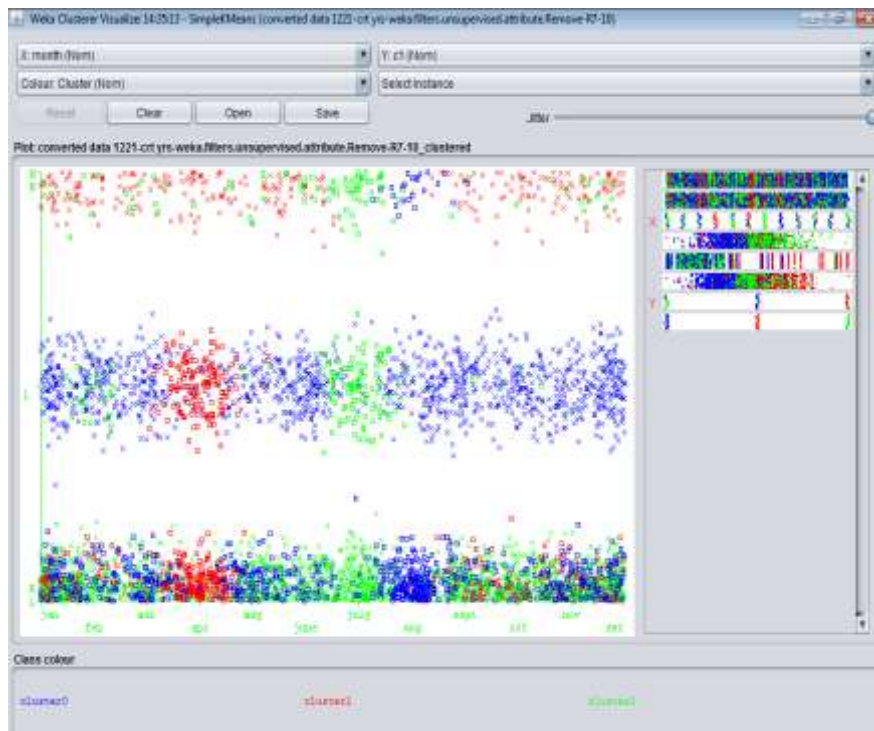


Fig. 4.6 Cluster visualization

The Fig 4.6 shows the visualization of cluster where the blue indicates cluster 0 i.e., L, red indicates cluster 1 i.e., HH and green indicates cluster 2 i.e., HL

In visualization window click on any point. This displays instance information. The Fig. 4.7 shows the instance information for the instance 1252 and month-June, amount of water storage will be low “L” and weka correctly clustered it as “cluster 0” (L).

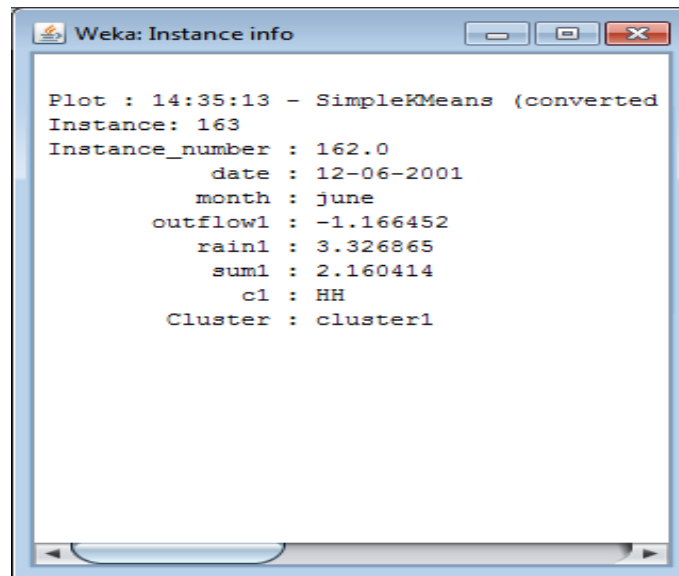


Fig. 4.7 Instance information window

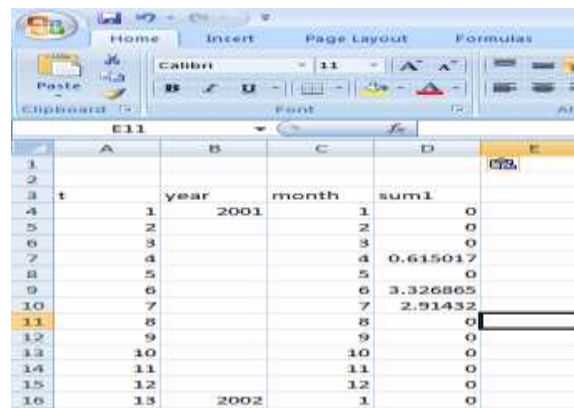
4.4 Gradient Descent

4.4.1 Steps

For each month, based on cluster value (HH, HL, L) average of sum value is taken.

For capacity 1 → Cluster0(L), Cluster1(HH), Cluster2(HL).

For capacity 2 → Cluster0(HL), Cluster1(HH), Cluster2(L).



	A	B	C	D	E
1					
2					
3	t	year	month	sum1	
4		1	2001	1	0
5		2		2	0
6		3		3	0
7		4		4	0.615017
8		5		5	0
9		6		6	3.326865
10		7		7	2.91432
11		8		8	0
12		9		9	0
13		10		10	0
14		11		11	0
15		12		12	0
16		13	2002	1	0

Fig 4.8 Actual data for gradient descent

The Figure 4.8 shows the actual data used for calculating gradient descent. By following various steps, we can forecast the capacity of water to be stored in exact numbers.

4.4.2 Forecasting calculation

4.4.2.1 mavg, Cmavg values are calculated

$$\text{mavg} = \text{AVERAGE} (D4:D15) \quad (4.7)$$

where E14, E15 represent sum values.

Average of sum value for first 12 values. For subsequent values of mavg, move one cell down then take average of sum value.

$$C_{mavg} = \text{AVERAGE}(E14:E15) \quad (4.8)$$

where E14, E15 represent mavg values.

Average of 1st 2 values of mavg. Similarly, by moving one cell down, take average for 2 values of mavg.



Fig. 4.9 Graph for actual and calculated values

The Fig. 4.9 shows the graph representation of sum and C_{mavg} values. Here the blue line indicates the actual value, and the red line indicates the predicted values. The actual and predicted values have large difference so to minimize that both intercept and slope values are calculated.

4.4.2.2 StIt, St, Deseasonalize, It values are calculated

St, It, forecast values are calculated for 2016, 2017, 2018, 2019, 2020 also.

$$StIt = D14/F14 \quad (4.9)$$

where D14 represents sum.

F14 represents C_{mavg}

$$St = \text{AVERAGEIF}(C14:C157, U25, G14:G157) \quad (4.10)$$

where C14:C157 represents month value.

U25 represents criteria for month value

G14:G157 represents StIt value.

$$\text{Deseasonalize} = D4/H4 \quad (4.11)$$

where D4 represents sum value.

H4 is the St value.

4.4.2.3 Linear regression analysis in Excel

To find the intercept values, linear regression is calculated through data analysis in excel. It will provide the summary output of linear regression the Fig. 4.10 shows the output for linear regression.

The coefficient of intercept and t values is taken for further calculation.

4.4.2.4 Linear regression analysis in Excel

$$It = \$C\$225 + \$C\$226 * A4 \quad (4.12)$$

where t is series number.

C225 is intercept value. C226 represents t.

$$\text{Forecast} = St * It \quad (4.13)$$

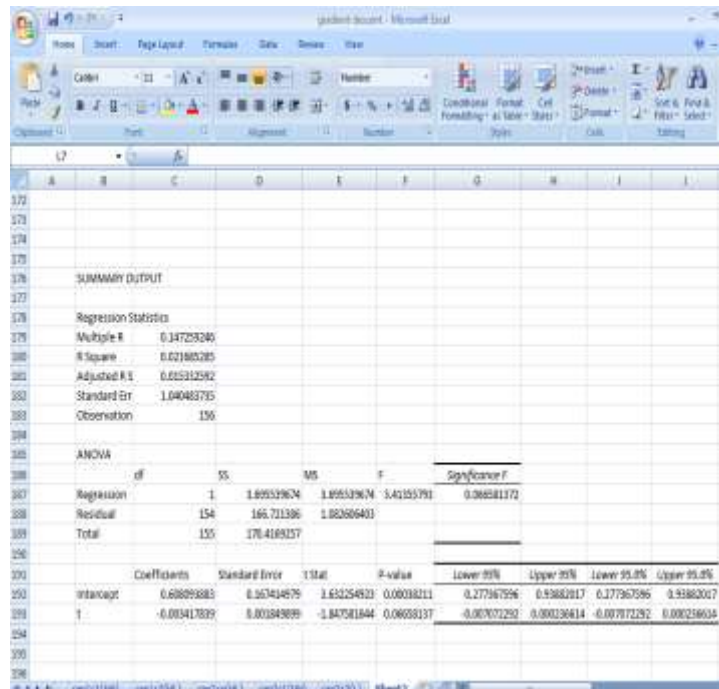


Fig. 4.10 Linear regression output

The Fig. 4.11 shows the graph of forecast values and the green line indicates the forecast values.



Fig. 4.11 Graph for Actual, Calculated and Forecast values.

The output values are not in normalized values so conversion must be done using the NORMINV function.



$$\text{value}(\text{cm}^3) = \text{NORMINV}(K4, Q4, R4) \quad (4.13)$$

where

K4 is the forecast value.

Q4 is mean value for normalized original data.

R4 is standard deviation for normalized original data.

4.5 Posterior Probability

The calculated forecast values are classified based on cluster label (HH, HL, L) using J48 algorithm for each month. The input data consists of month, label and forecast values for each capacity separately which is shown in the Fig. 4.12.

No.	1: year	2: month no	3: month	4: cluster	5: forecast
	Numeric	Numeric	Nominal	Nominal	Numeric
1	200...	1.0	jan	L	-1.7
2		2.0	feb	L	-1.3
3		3.0	mar	L	-1.44
4		4.0	apr	L	-1.08
5		5.0	may	L	-1.18
6		6.0	jun	L	-1.1
7		7.0	jul	L	-0.76
8		8.0	aug	L	-1.27
9		9.0	sep	L	-1.41
10		10.0	oct	L	-1.24
11		11.0	nov	L	-0.7
12		12.0	dec	L	-1.32
13	200...	1.0	jan	L	-1.68
14		2.0	feb	L	-1.29
15		3.0	mar	L	-1.43
16		4.0	apr	L	-1.07
17		5.0	may	L	-1.17
18		6.0	jun	L	-1.09
19		7.0	jul	L	-0.75
20		8.0	aug	L	-1.25
21		9.0	sep	L	-1.4
22		10.0	oct	L	-1.23
23		11.0	nov	L	-0.69
24		12.0	dec	L	-1.31
25	200...	1.0	jan	L	-1.66
26		2.0	feb	L	-1.27
27		3.0	mar	L	-1.41
28		4.0	apr	L	-1.06
29		5.0	may	L	-1.16
30		6.0	jun	L	-1.08
31		7.0	jul	L	-0.74
32		8.0	aug	L	-1.24
33		9.0	sep	L	-1.38
34		10.0	oct	L	-1.21
35		11.0	nov	L	-0.68
36		12.0	dec	L	-1.3
37	200...	1.0	jan	L	-1.64

Fig. 4.12 Input ARFF file for J48

4.5.1 Result obtained for J48

The J48 algorithm is used for classification, here number of instances for leaf node is 2. i.e., M2. Confidence factor is 0.5 (for more accuracy). Total number of instances used her is 612. All these details were obtained through the J48 output window that is shown in the Fig. 4.13.

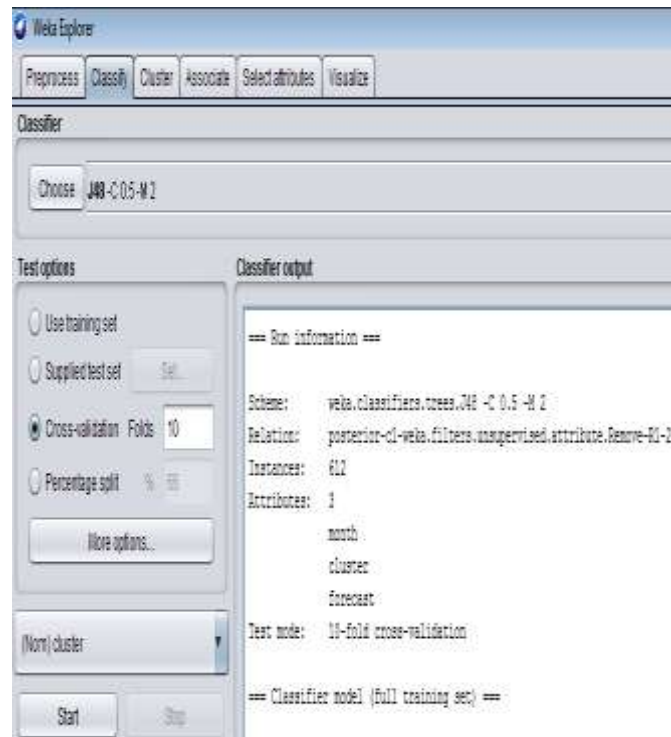


Fig. 4.13 Output window for J48

The J48 accuracy is shown in the Fig. 4.14 where the number of leaves obtained is 32. Here correctly classified instances were 550 in percentage it is 90% and incorrectly classified instances were 62 in percentage it is 10%. The confusion matrix of J48 show the details in term of HH, HL and L labels which is shown in the Fig. 4.15.

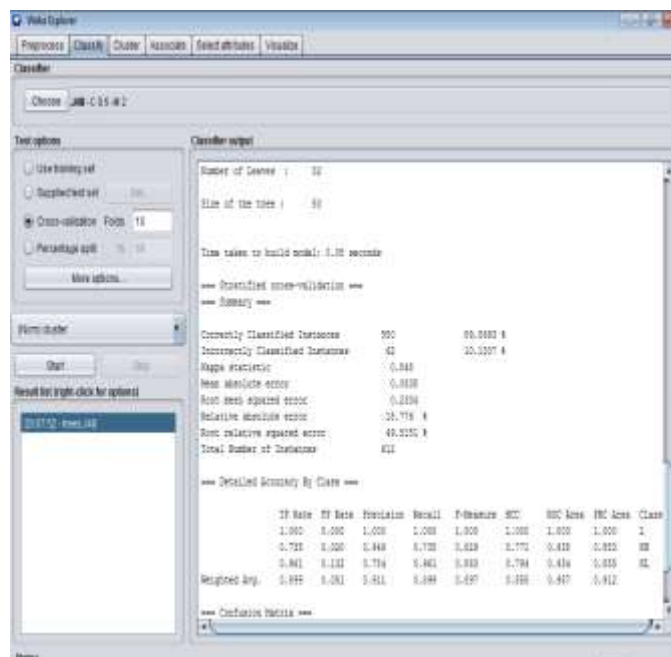


Fig. 4.14 J48 Accuracy

The classifier visualization is shown in the Fig. 4.17 where the X-axis denotes forecast values and Y-axis denotes cluster values. The accuracy percentage of the capacity is also shown in the Table. 4.2.

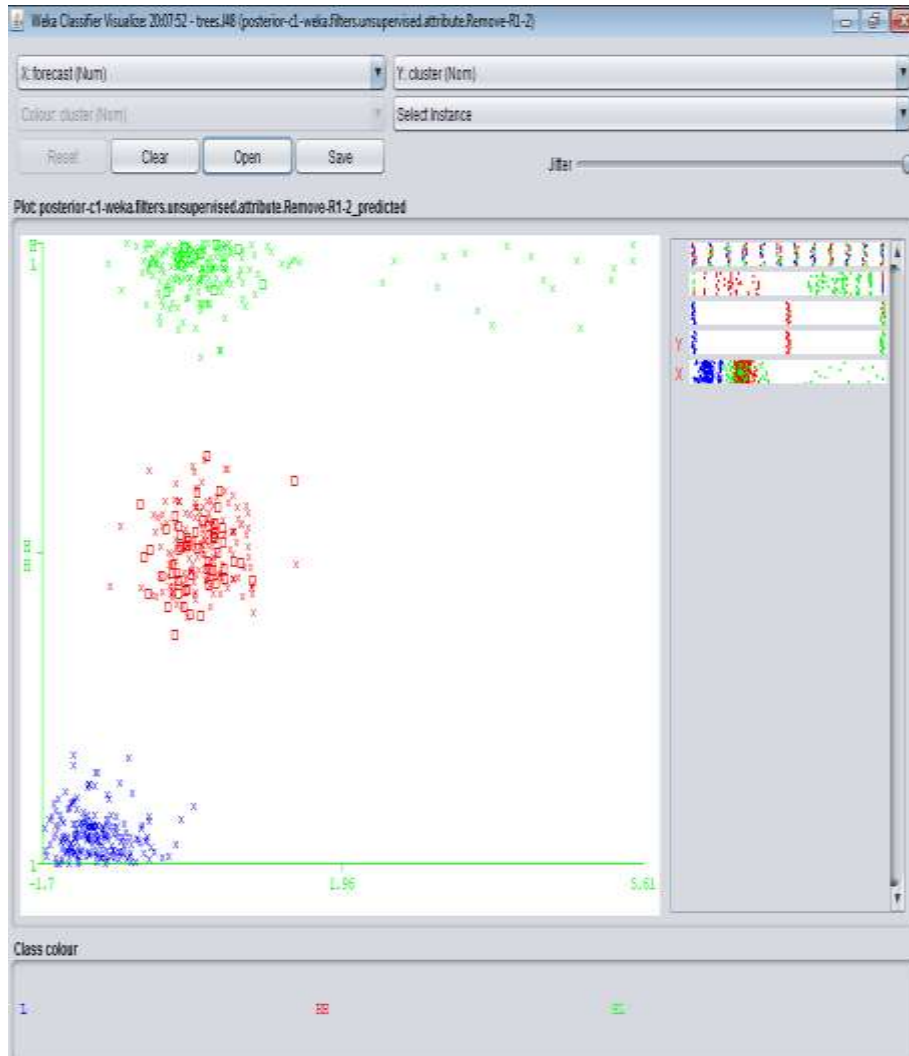


Fig. 4.17 Visualization window for J48

Table II J48 classification accuracy table in %

capacity	Accuracy in %
Capacity 1	90
Capacity 2	80.2288

The J48 algorithm will classify the label and gives the results as predicted clustered labels it is shown in the Fig.4.18, we can find the difference between the calculated classified labels and weka classified labels.



5. CONCLUSION

Applying above specified data techniques the accuracy of system is increased certainly when compared to previous techniques applied through ANN technique. Where ANN leads to desired output but there are some uncertainties like it provide results only up to approximate level and do not give accurate results. By using data mining techniques, values are forecasted here will be useful for predictive analysis and will gives the accurate result in numbers and percentage.

6. REFERENCES

1. Azfina Putri Anindita, Pujo Laksono, I Gusti Bagus Baskara Nugraha ,(2016), Dam Water Level Prediction System UtilizingArtificial Neural Network Back Propagation, Case Study: Ciliwung Watershed, Katulampa Dam ,proceedings of “2016 International Conference on ICT For Smart Society Surabaya,”
2. Punithavathi.J, Baskaran,R, (2011), Land use and land cover using remote sensing and GIS techniques - A case study of Thanjavur District, Tamil Nadu, India, -proceedings of “International Journal of Current Research “ Vol. 3, Issue, 10, pp.237-244.
3. J.Nittin Johnson, S.Govindaradjane, T.Sundararajan,(2013),Impact of Watershed Management on the Groundwater and Irrigation Potential: A Case Study, proceedings of “International Journal of Engineering and Innovative Technology (IJEIT) “Volume 2, Issue 8.
4. J.Nittin Johnson, S.Govindaradjane, T.Sundararajan,(2013),Impact of Watershed Management on the Groundwater and Irrigation Potential: A Case Study, proceedings of “International Journal of Engineering and Innovative Technology (IJEIT) “Volume 2, Issue 8.
5. Parneet Kaur, Manpreet Singh, Gurpreet Singh Josan,(2015),Classification and prediction based data mining algorithms to predict slow learners in education sector.
6. Dr. Sudhir B. Jagtap, Dr. Kodge B. G.(2013), Census Data Mining and Data Analysis using WEKA, proceedings of “XVIII International Conference on Water Distribution Systems Analysis”.
7. Raziye Farmani, Konstantinos Kakoudakis, Kourosh Behzadian and David Butler,(2017), Pipe Failure Prediction in Water Distribution Systems Considering
8. Dr. Sudhir B. Jagtap, Dr. Kodge B. G.(2013), Census Data Mining and Data Analysis using WEKA, proceedings of “XVIII International Conference on Water Distribution Systems Analysis”.



ARFF-Viewer - C:\Users\TASL\Desktop\step7PosteriorProbability\visop1.arff

File Edit View

visop1.arff

Relation: posterior-c1-weka.filters.unsupervised.attribute.Remove-R1-2_predicted

No.	1: month Nominal	2: prediction margin Numeric	3: predicted cluster Nominal	4: cluster Nominal	5: forecast Numeric
1	sep	-0.545455	HL	HH	0.14
2	apr	1.0	HH	HH	-0.02
3	may	1.0	HH	HH	0.48
4	mar	1.0	HH	HH	0.21
5	apr	1.0	HH	HH	0.34
6	sep	-0.545455	HL	HH	0.17
7	feb	1.0	HH	HH	0.03
8	aug	1.0	HH	HH	0.06
9	jul	1.0	HH	HH	0.39
10	apr	1.0	HH	HH	0.02
11	nov	1.0	HH	HH	0.05
12	jan	-0.6	HL	HH	0.19
13	jan	1.0	HH	HH	0.35
14	oct	-0.5	HL	HH	0.34
15	apr	1.0	HH	HH	0.3
16	jun	1.0	HH	HH	-0.02
17	jun	1.0	HH	HH	0.24
18	feb	1.0	HH	HH	-0.03
19	sep	1.0	HH	HH	0.05
20	apr	1.0	HH	HH	0.05
21	jul	1.0	HH	HH	-0.07
22	feb	1.0	L	L	-1.3
23	jan	1.0	L	L	-1.68
24	oct	1.0	L	L	-1.21
25	may	1.0	L	L	-1.08
26	jul	1.0	L	L	-0.76
27	feb	1.0	L	L	-1.2
28	apr	1.0	L	L	-0.98
29	jun	1.0	L	L	-0.98
30	jun	1.0	L	L	-1.02
31	sep	1.0	L	L	-1.38
32	may	1.0	L	L	-1.12
33	dec	1.0	L	L	-1.23
34	aug	1.0	L	L	-1.09
35	jun	1.0	L	L	-1.01
36	jul	1.0	L	L	-0.65
37	dec	1.0	L	L	-1.2

Fig. 4.18 output ARFF file for J48