International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-2, July 2019

# Quality Assessment of Ground Water in Pre and Post-Monsoon using Various Classification Technique

**Aiswarya Vijayakumar, A S Mahesh**

***Abstract*: Quality assessment of water is one of the basic points which have pulled in a lot of thought in the progressing years. Diverse kinds of classification system are most convenient for the examination in this field of study. The present examination investigates the quality of ground water in Agastheeswaram which is located in Tamilnadu. Totally 138 water samples was accumulated in the midst of pre-monsoon (PRM) and post-monsoon (PSM) from the year of 2011 to 2012.The water quality (WQ) evaluation was carried out by assessing chemical parameters for both the seasons. This paper explores various classifier models such as DT, KNN and SVM to achieve prediction of groundwater quality. The classification is done based on the WQI of each sample. A near investigation of characterization systems was done dependent on the confusion matrix, accuracy, f1 score, precision and recall. The outcomes propose that SVM is a better method having high accuracy rate than other models.***

***Index Terms*: Classification Algorithms, Water Quality Index, Support Vector Machine, Decision Tree; K Nearest Neighbors.***

## I. INTRODUCTION

Nowadays, use of groundwater extended at an aggravating amount over the world. Abuse of water has expanded significantly due to irregularity of monsoon, which badly influences surface water resources. In current trends, the number of aquifers is going to exploit at a sustainable level. In the future, most of the irrigated agriculture and drinking water supplies are dependent on groundwater. So, the assessment of water quality is a serious concern for current and future needs. Such huge numbers of geochemical works are done in India and abroad to discover the reasonableness of groundwater. WQI is a helpful tool for quality analysis of water. The similar type of studies was carried out for the analysis of WQ in PSM and PRM. The variation in physical and chemical properties of groundwater in different seasons is the major reason for the change in water quality index in Bhopal region [10]. Quality evaluation and hydro geochemical attributes of groundwater was done on a similar locale.

In this field, classification is a helpful method for announcing the appraisal of water. For a specific point in time, an order will demonstrate to us where the nature of the water condition is great and where it might require making the stride. Classification techniques are intended to perceive the participation of each article to its legitimate class based on a lot of estimations. Classification of information into various classes is a strategy to arrange extensive information for efficient computation. As such, order strategies find numerical connections between a lot of enlightening factors (e.g., Chemical measures) and a qualitative variable. Characterization techniques (additionally called supervised methods) are progressively utilized in a few fields, for example, science, process checking, medical sciences, pharmaceutical science, and social and monetary sciences [4].

## II. RELATED WORKS

In water quality research examines, diverse methods of arrangements have been used, for example, SVM [20], neural network [14], K-NN [25] and classification trees [17].

Classification techniques have been utilized widely in different fields of study. There are a lot of investigations demonstrated that these techniques are useful with sensible precision for water quality research., Recently utilized techniques are LDA and Naive Bayesian Regularization for water quality examinations with nine sampling stations with four properties in Karaj River, Iran [19].

Sakizadeh et.al were used SVM and K-NN classifier models for the analysis of WQ with pollution level readings in Khuzestan Province, Iran. For these 17 parameters from 41 samples were utilized for the evaluation [18].

In comparison with different classification models such as DT, KNN, and SVM on samples that are amassed from Madhya Pradesh, India. The interpretation is carried out by Electrical Conductivity levels. The classification learner application from MATLAB is used for the implementation process [16].

An examination was done to interpret the WQI on samples with 16 parameters that collected from Iran Ministry of Energy, using Artificial Neural Network with Bayesian Regularization. The study that also depicts the sensitivity analysis to show importance of each properties in prediction of WQI.

**Revised Manuscript Received on 30 July 2019.**
***** Correspondence Author
**Aiswarya Vijayakumar***, Department of Computer Science & IT, Amrita School of Arts & Sciences Kochi, Amrita Vishwa Vidyapeetham India
**A S Mahesh**, Department of Computer Science & IT, Amrita School of Arts and Sciences Kochi, Amrita Vishwa Vidyapeetham, India.

## III. PROPOSED METHOD

Currently, there are a lot of classification models were utilized for assessment of groundwater quality with different interpretation strategies the significant concentration of this paper to evaluate the quality of groundwater in PRM and PSM, for that the samples were obtained from coastal area because the large amount of contamination and more seasonal variations in observed values was noticed.

However, the fundamental aim of this paper was to utilize different classification models (K-NN, DT, and SVM) for quality assessment of groundwater in different seasons. The range for WQI differing for each sample in Pre and Post Monsoon, because of variations in observed values of each parameter. The data set for this investigation were collected from the southern part of India. The interpretation depends on the confusion matrix, precision, recall, and accuracy.

## IV. MATERIALS AND METHODS

### *4.1 Study Area*

Totally 138 samples were amassed and assessed for this investigation to obtain the WQ in PRM and PSM.The region that covers for this study is 279.4 km2. The longitude and latitude for this study area is 8˚4′ N to 8˚13′45″ N and 77˚18′45″ E to 77˚35′15″ E.

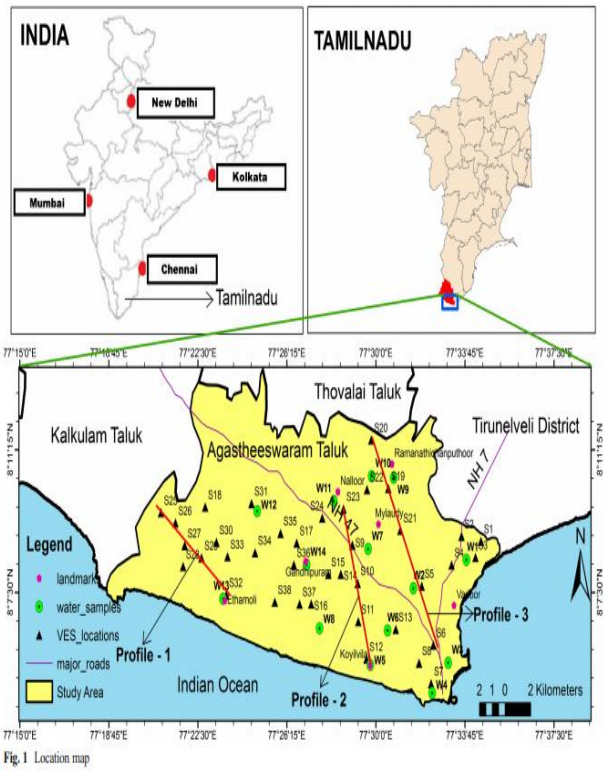


Fig. 1 Location map

**Fig 1: An illustration of study is in Agastheeswaram Taluk**

### *4.2 Workflow Diagram*

The figure demonstrated that the complete process of the research study. Data procuring is the foremost step and then the data is pretreated, i.e. the noisy data will be removed and WQI for each sample gets calculated. Then the data gets separated into training and test set (70% and 30%) respectively further the datasets are implemented based on three classification algorithms. Then the testing of data is carried out using test data. Performance evaluation is done on the basis of confusion matrix, recall, f1 score and precision.

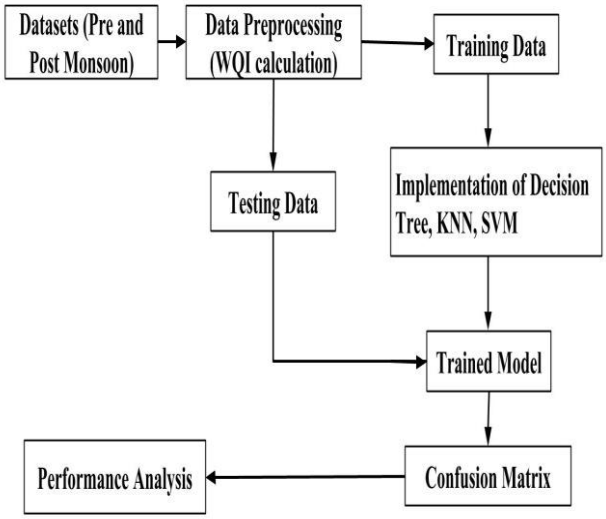


**Fig 2: An Illustration That Depicts Flow Diagram Of This Study**

### *4.3 Water Quality Index Calculation*

The samples were amassed and processed using 7 parameters and Quality Rating (QR) to calculate WQI. For this examination we utilized process recommended by Horton [8] was used, further the permissible and desirable limit of each parameter were taken from Bureau of Indian Standards (BIS).

The succeeding equations confess calculate WQI

a. WQI

$$WQI = \sum \boldsymbol{Q}iWi \; / \sum \boldsymbol{W}i \qquad (1)$$

Where,

Qi = QR of ith WQ parameter.

Wi = UW of ith WQ parameter, UW = Unit Weight

b. Quality rating (Qi)

The Qi is calculated by equation (2)

$$Qi = [(Ui - Uid) / (Si - Uid)] \times 100 \qquad (2)$$

Where,

Ui = Sample location's estimated value of ith WQ parameter

Uid = Pure water ideal value for ith parameter

(Uid for pH = 7 and 0 for all others)

Si = ith WQ parameter of Standard permissible value.

c. Unit weight (Wi)

$$Wi = C / Si \qquad (3)$$

*Retrieval Number: B3709078219/19©BEIESP*
*DOI: 10.35940/ijrte.B3709.078219*
*Journal Website: www.ijrte.org*



*Published By:*
*Blue Eyes Intelligence Engineering*
*& Sciences Publication*

Where,
Si = Standard permissible value of ith WQ parameter.
C = Constant of proportionality
C = [1 / (∑1/ Sn=1, 2… i)] (4)

**Table 1: The Table Illustrates The Calculation Of WQI For Sample S1**

| Sl. No. | Parameters | Observed Values | Standard Values ($S_i$) | Unit Weight ($W_i$) | Quality Rating ($Q_i$) | $W_iQ_i$ |
|---|---|---|---|---|---|---|
| 1 | pH | 7.97 | 8.5 | 0.6525 | 194.00 | 126.57 |
| 2 | TH | 210.16 | 200 | 0.0277 | 105.08 | 2.91 |
| 3 | Ca | 2.004 | 75 | 0.0739 | 2.67 | 0.197 |
| 4 | Mg | 49.856 | 30 | 0.1849 | 166.19 | 30.72 |
| 5 | Cl | 88.75 | 250 | 0.0222 | 35.50 | 0.787 |
| 6 | TDS | 65.28 | 500 | 0.0111 | 13.06 | 0.144 |
| 7 | SO4 | 12.25 | 200 | 0.0277 | 6.13 | 0.169 |
| | | | | ∑Wi = 1.000 | ∑Qi = 522.62 | ∑QiWi = 161.51 |
| Water Quality Index = ∑Qi Wi/ ∑ Wi = 161.51 | | | | | | |

*4.3.1 WQI Status*

The WQI status reveals the quality rating of water based on water quality index values

**Table 2. The Table Shows The Quality Rating With Corresponding WQI Values**

| SL NO | Water Quality Index Values | WQ Rating |
|---|---|---|
| 1 | 0 - 25 | Excellent |
| 2 | 26 – 50 | Good |
| 3 | 51 – 75 | Fair |
| 4 | 76 -100 | Poor |
| 5 | >100 | Unfit |

*4.4 Decision Tree*

Decision tree is a tree structure like flowchart is used to predict the classification of unseen records based on a series of rules. The attributes of classes can be any type of variable from nominal, quantitative, binary, ordinal values. A decision tree produces a series of questions that can be used to recognize the class. The basic DT demonstrated in Fig3

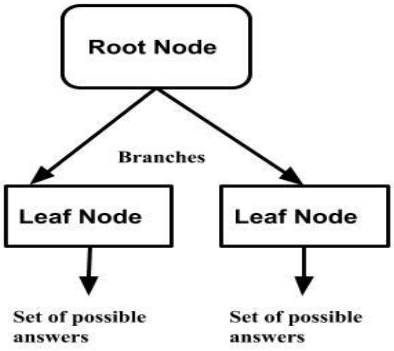


**Fig 3: An Illustration Of Basic Decision Tree**

The decision tree are regularly utilized for decision making through a set of information. The representation of DT starts with a root node on which will be the base of the decisions and each internal node tests an attribute on some property, finally the possible outcome represented in branches. Generally used decision tree algorithms are CART and ID3, which handles various splitting criteria for splitting the node. These algorithms searches for the best attribute that can used for splitting. There are different part criteria dependent on nodes impurity. The primary point of splitting criteria is to lessen the impurity of a node. There are numerous proportions of splitting that can be utilized to decide the most ideal approach to split the records. These splitting measures are characterized as far as the class distribution of the records when split.
Impurity measure comprises:
ENTROPY: Impurity measure of a node
**Entropy(t) : - P(m/n) $\log_2$ P(m/n) ,** *(5)*
where m and n are level of target variable
GINI INDEX : It is another proportion of impurity influence that estimates the divergences between the likelihood dissemination of the objective characteristic qualities.
**Gini Index : 1 - $[P(m/n)]^2$** *(6)*

*4.5 K – Nearest Neighbor*

It is a nonparametric algorithm which classifies new cases based on distance functions. In Knn, for every test information point, we would take a glimpse at the K nearest training information and take the most often occurring classes and dole out that class to the test information. Hence K shows the number of training data point lying in vicinity to the test information point which we are going to use to discover the class.

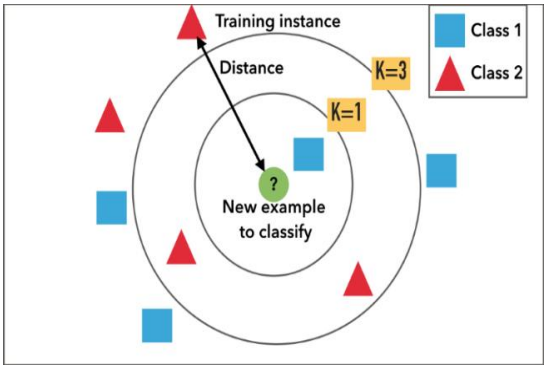


**Fig: 4. The Basic Illustration Of KNN Classification.**

K nearest neighbor measured using the following distance functions

Euclidean $= \sqrt{\sum_{t=1}^{k}(Xt - Yt)^2}$ *(7)*

Manhattan $= \sum_{t=1}^{k}|Xt - Yt|$ *(8)*

Minkowski $= (\sum_{t=1}^{k}(|Xt - Yt|)^q)^{1/q}$ *(9)*

For non-numeric and categorical attribute, the possible distance metric is Hamming

$d_h = \sum_{m=1}^{K}|Xm - Ym|$ *(10)*

if X = Y => 0

if X not equal Y => 1

Where X, Y are two instances of attribute, if both the instances are same then distance is taken as 0.For a given number of nearest neighbors K and an unknown sample point X and a distance metric D, a KNN classifier performs the following two functions

1. It searches through the full dataset and calculate the distance between X and each observation for training. Let assume that the K points in the training data that are adjacent to X and reside on U set. It is noteworthy that K usually to be an odd number as it prevents the tied situation.
2. Then it calculates for each class conditional probability, i.e., the possibility for a points of fraction present in the set U for a given class label. Finally, the unknown sample point X is allocated based on highest probability of class.

### *4.6 Support Vector Machine*

It is a regulated learning model intended to accomplish superior exhibitions in pragmatic applications that make grouping by identifying the ideal hyper plane that helps the edge of detachment inside the two classes. In spite of the fact that SVM builds a complex model and calculations (contains a wide-going class of neural nets, radial, and polynomial classifiers, and so on [7]; it is easy to examine numerically, in light of the fact that it maps the contributions with an immense dimensional element space through some nonlinear mapping in the input space. Through some kernel tricks, necessary calculations can be performed for input space. To minimizing the error the data should be classified correctly with right hyper plane.

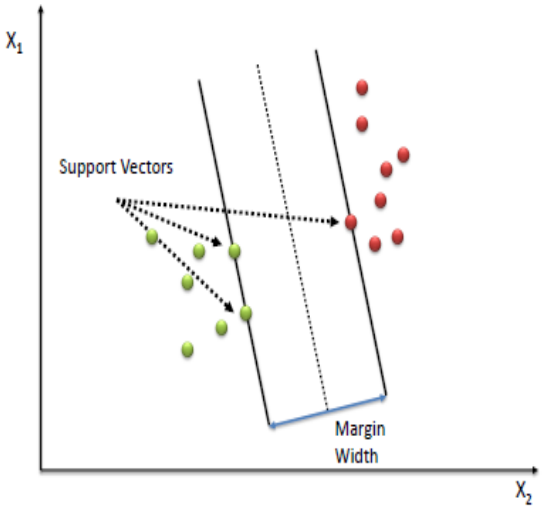


**Fig 5: The Basic Illustration Of Support Vector Machine Classifier**

The model finds the hyperplane that maximizes the margin between the two classes for classification procedure. Support vectors are defining the hyperplane. The margin width(W) should be maximized to get optimal hyperplane. The perfect SVM examination should deliver a hyperplane that totally differentiate the vectors in to two classes which are not overlapping. Sometimes it is difficult to get the actual separation because the result shows that classification of model is not accurate. The easiest approach to isolate two sets of data with one dimensionally, two dimensionally and multi dimensionally. There exist many circumstances in which the nonlinear region can isolate the groups more effectively and Kernel functions are used to handle this situation. Kernel function can delineate information into an alternate space where a hyperplane can't be utilized to do the partition. It implies that non-linear function found out by a direct learning machine in a high dimensional component space while the limit of the framework is constrained by parameter that depends on space dimensionality. This is known as Kernel Trick which implies the kernel function change the information into multi-dimensional component space to make it conceivable to play out the straight detachment.In Support Vector Machines a number of Kernels can be utilized. These comprises of Polynomial, Radial Base Function, Linear and Sigmoid.

**Kernel Functions**

Polynomial K $(X_m, X_n) = (\gamma X_m . X_n + C)^d$ *(12)*

RBF K$(X_m, X_n) = \exp(-\gamma |X_m - X_n|^2)$ *(13)*

Linear K $(X_m, X_n) = X_m . X_n$ *(14)*

Sigmoid K $(X_m, X_n) = \tanh(\gamma |X_m . X_n + C)$ *(15)*

Where K$(X_m, X_n) = \phi(X_m) . \phi(X_n)$

Equation (12 -15) explain the working of kernels which transforms the feature space by multiplying the input data set to higher dimension using ϕ. The kernel functions can be adjusted using Gamma. The commonly used SVM kernel is radial base function, this is because of a direct result of their restricted and limited reactions over the whole scope of the genuine x-axis.

## V. EXPERIMENTS AND DISCUSSIONS

The proposed method had been tested in datasets of both PRM and PSM. Experiments were conducted to check the efficiency of each model.

### *5.1 Performance evaluation*

In machine learning the performance can be assessed using Confusion Matrix. It is an execution estimation for machine learning grouping issue where yield can be at least two classes. The confusion matrix is a table that consist with 4 distinct mixes of anticipated and genuine qualities. That is, it consists of true positives ($T^+$) and negatives($T^-$) and false positives ($F^+$) and negatives ($F^-$). The model assessment criteria have been characterized utilizing components of the confusion matrix. The performance interpretation can be done on the basis of following metrics for a model.

Accuracy

Accuracy is the most natural execution measure and it is just a proportion of accurately anticipated perception to the complete perceptions. It can be calculated using the equation (16).

$$A = (T^+ + T^-)/ (T^+ + T^- + F^+ + F^-) \quad (16)$$

Precision(P)

Accuracy is the proportion of accurately anticipated positive perceptions to the absolute anticipated positive. This can be calculated using equation 17

$$P = T^+ / (T^+ + F^+) \quad (17)$$

Recall(R)

Sensitivity is the proportion of effectively anticipated positive perceptions to the all perceptions in genuine class – yes. Equation 18 shows how to calculate recall

$$R = T^+ / (T^+ + F^-) \quad (18)$$

F1 Score(F1)

Weighted average of Recall and Precision gives f1 score

$$F1 = 2 *(R*P) / (R+P) \quad (19)$$

### 5.2 Result Analysis

The effectiveness of the proposed method discussed in detail in this section

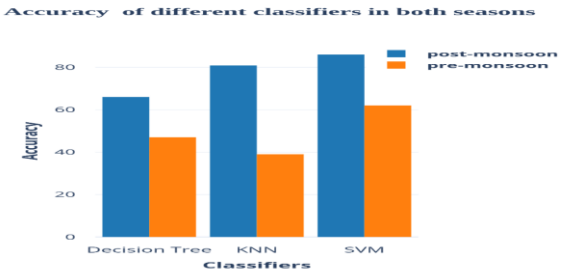


**Fig 6: Shows Different Classifier Models With Respective Accuracies**

The Fig 6 represent the accuracies of different classifiers in PRM and PSM and SVM gives highest accuracy i.e., 86% in post monsoon and 62% for pre monsoon. Whereas KNN gives 80.9% and 39% for both seasons and DT gives least accuracy for both seasons i.e. 66 and 47 percentage.

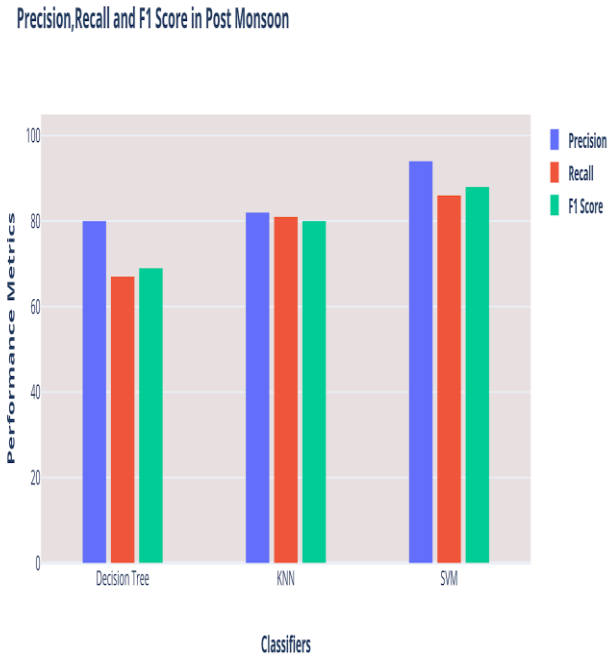


**Fig 7: Graphical Representation Of Different Performance Metrics In Post Monsoon**

The Fig 7 shows recall, f1 score and precision of different models in post monsoon. It explains that SVM gives highest performance rate than other models. The table 3 shows that the percentages of performance for each model.

**Table 3: Performance Metrics values in PSM**

| Data Division | DT | KNN | **SVM** |
|---|---|---|---|
| Precision (%) | 80 | 83 | **94** |
| Recall (%) | 67 | 81 | **86** |
| F1 Score (%) | 69 | 80 | **88** |

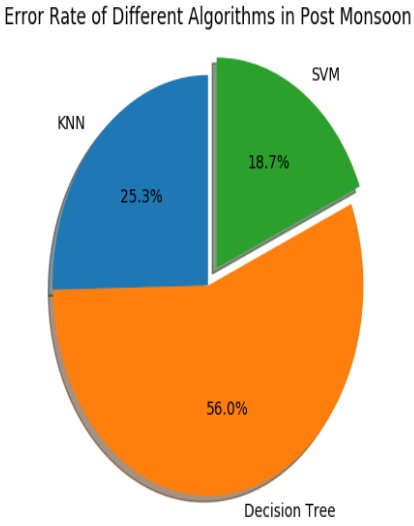


**Fig 8: Overall Error Rate Of Different Models**

Figure 8 reveals that SVM model gives less error than other models in assessment of water quality. That is, it gives good prediction rate in post monsoon.

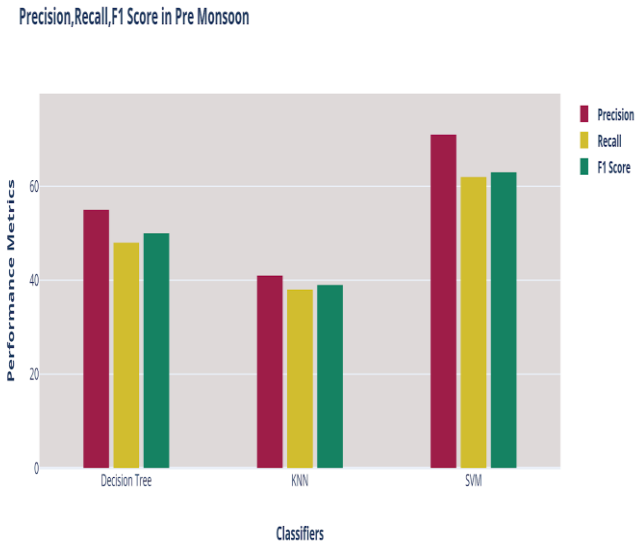


**Fig 9: Graphical representation of different performance metrics in PSM**

The Fig 9 shows that the performance metrics of different models in pre monsoon. It clearly shows that SVM gives highest percentage of performance rate than other models. The table 4 shows that the percentages of performance for each model.

**Table 4: Performance metrics values for different classifiers in Pre-Monsoon**

| Data Division | KNN | DT | **SVM** |
|---|---|---|---|
| Precision (%) | 41 | 55 | **71** |
| Recall (%) | 38 | 48 | **62** |
| F1 Score (%) | 39 | 50 | **63** |

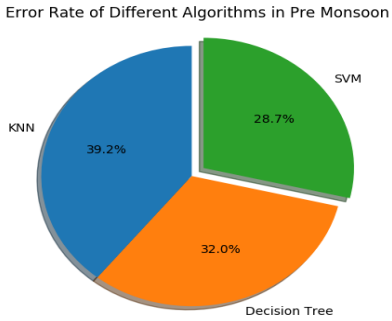


**Fig 10: Overall Error Rate Of Different Models In Pre Monsoon**

According to Figure 10, it depicts that overall error rate for SVM is lower than other classifiers. That is, it gives good assessment rate for pre monsoon.

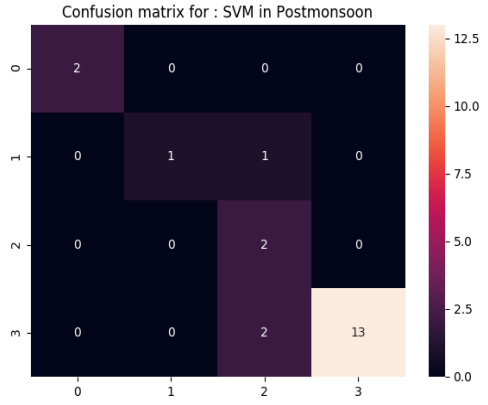


**Fig 11: Illustration CM for SVM in Post-Monsoon**

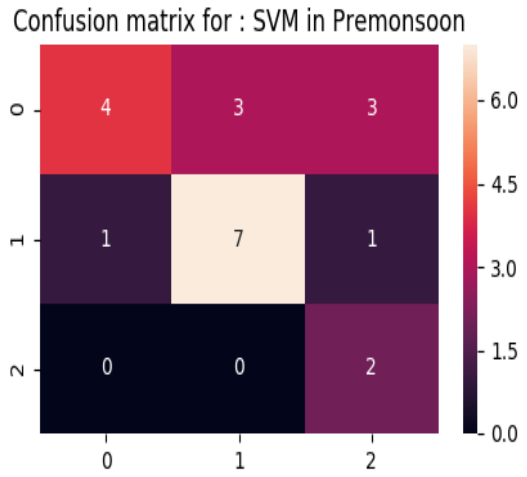


**Fig 12: Illustration CM For SVM In Pre-Monsoon**

In pre and post monsoon 69 genuine dataset where used to assess WQ. For the analysis 70:30 ratio used to train and test respectively. Figure 11 and 12 reveals that quantity of samples where predicted as true and partially predicted classes in both seasons. According to Figure 6 to 12 and table 3 and 4, clearly shows that the effectiveness of Support Vector Machine for analysis of water quality in post monsoon and pre monsoon. This study identifies that SVM will be the better option for quality assessment of groundwater with seasonal variations of water quality index.

## VI. CONCLUSION

Observing the nature of water has turned into a key part as a result of the generally predominant interminable waterborne maladies. Nowadays, Classification techniques have been utilized broadly in different fields of study; in any case, their utilization in hydrogeology is restricted due to unavailability of data. This investigation demonstrated that the utilization of these techniques with sensible exactness in quality of water related research contemplates. This paper proposes effectiveness of different classifiers for quality assessment of groundwater with seasonal variations. The utilized models are KNN, Decision Tree and SVM. The dimension of WQI was the grouping criteria for the groundwater samples gathered from the southern tip of India. The implementation of all models is done in PYTHON-3.6.5. The efficiency of model analyzed on Confusion Matrix, Precision, Recall, F1 and error rate. This study clearly depicts that SVM classifier is the best model for water quality assessment in both PRM and PSM. The performance metrics of SVM gives higher accuracy, recall, precision, f1 score, and lower error rate in both seasons than other classifiers.

## ACKNOWLEDGMENT

The authors gratefully acknowledge Dr Hudson Oliver for providing genuine data to carry out this study.

## REFERENCES

1. Central Ground Water Board, Ministry of Water Resources,Government of India- BIS Standard.
2. Cruz, J. A., &Wishart, D. S. (2006). Applications of machine learning in cancer prediction and prognosis. Cancer informatics, 2, 117693510600200030.
3. Deng, Z., Zhu, X., Cheng, D., Zong, M., & Zhang, S. (2016). Efficient kNN classification algorithm for big data. Neurocomputing, 195, 143-148.
4. Dollar, E. S. J., James, C. S., Rogers, K. H., &Thoms, M. C. (2007). A framework for interdisciplinary understanding of rivers as ecosystems. Geomorphology, 89(1-2), 147-162.
5. Dubey, H. (2013). Efficient and accurate kNN based classification and regression. A Master Thesis Presented to the Center for Data Engineering, International Institute of Information Technology, Hyderabad-500, 32.
6. Guo, G., Wang, H., Bell, D., Bi, Y., & Greer, K. (2003, November). KNN model-based approach in classification. In OTM Confederated International Conferences" On the Move to Meaningful Internet Systems" (pp. 986-996). Springer, Berlin, Heidelberg.
7. Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., &Scholkopf, B. (1998). Support vector machines. IEEE Intelligent Systems and their applications, 13(4), 18-28.
8. Horton, R. K. (1965). An index number system for rating water quality. Journal of Water Pollution Control Federation, 37(3), 300-306.
9. Huang, L., Chen, Y., She, C., Wu, Y., & Zhang, S. (2018). Application of Classifiers in Predicting Problems of Hydropower Engineering. Applied and Computational Mathematics, 7(3), 139-145.
10. Jinwal, A., & Dixit, S. (2008). Pre-and post-monsoon variation in physico-chemical characteristics in groundwater quality of Bhopal “The City of Lakes” India. Asian Journal of Experimental Sciences, 22(3), 311-316
11. Khalil, A., Almasri, M. N., McKee, M., &Kaluarachchi, J. J. (2005). Applicability of statistical learning algorithms in groundwater quality modeling. Water Resources Research, 41(5).
12. Khan, M. M. R., Arif, R. B., Siddique, M., Bakr, A., &Oishe, M. R. (2018). Study and Observation of the Variation of Accuracies of KNN, SVM, LMNN, ENN Algorithms on Eleven Different Datasets from UCI Machine Learning Repository. arXiv preprint arXiv:1809.06186.
13. Li, X., &Guo, Y. (2013, August). Active Learning with Multi-Label SVM Classification. In IJCAI (pp. 1479-1485).
14. Modaresi, F., &Araghinejad, S. (2014). A comparative assessment of support vector machines, probabilistic neural networks, and K-nearest neighbor algorithms for water quality classification. Water resources management, 28(12), 4095-4111.
15. Moore, A. W. (2001). Support vector machines. Tutorial. School of Computer Science of the Carnegie Mellon University. Available at http://www. cs. cmu. edu/~ awm/tutorials.[Accessed August 16, 2009].
16. Prakash, R., Tharun, V. P., & Devi, S. R. (2018, April). A Comparative Study of Various Classification Techniques to Determine Water Quality. In 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT) (pp. 1501-1506). IEEE.

17. Saghebian, S. M., Sattari, M. T., Mirabbasi, R., & Pal, M. (2014). Ground water quality classification by decision tree method in Ardebil region, Iran. Arabian Journal of Geosciences, 7(11), 4767-4777.
18. Sakizadeh, M., &Mirzaei, R. (2016). A comparative study of performance of K-nearest neighbors and support vector machines for classification of groundwater. Journal of Mining and Environment, 7(2), 149-164.
19. Sakizadeh, M. (2015). Assessment the performance of classification methods in water quality studies, A case study in Karaj River. Environmental monitoring and assessment, 187(9), 573.
20. Singh, K. P., Basant, N., & Gupta, S. (2011). Support vector machines in water quality management. Analyticachimicaacta, 703(2), 152-162.

## AUTHORS PROFILE

**Aiswarya Vijayakumar** is a MPhil Scholar at Department of Computer Science and IT, in Amrita School of Arts and Sciences Kochi, India. She completed her BCA from Bharathiar University, Tamil Nadu and Post-graduation (MCA from Amrita Vishwa Vidyapeetham, Tamil Nadu. Her interested areas includes Data Mining, Machine learning, Hadoop and Cloud Computing.

**A S Mahesh** is a Assistant Professor in Department of Computer Science and IT, in Amrita School of Arts and Sciences Kochi, India. His qualifications include M. Sc. (CS), MBA (Systems and Marketing), and M.Phil. (CS). He has more than 19 years of teaching including 6 years in research. His areas of interest includes Cloud Computing, Networking and Programming.