PREDICTION OF WATER QUALITY USING MACHINE LEARNING ALGORITHMS AND ROUGH SET THEORY

D. Venkata Vara Prasad [1],Suresh Jagannathan [2] Santosh Sivan[3]
[1,2,3]Department of Computer Science and Engineering, Sri Sivasubramaniya Nadar College of Engineering, Kalavakkm-603110, Chennai, India
Email ID: dvvprasad@ssn.edu.in[1], sureshj@ssn.edu.in[2],santhosh2120029@ssn.edu.in[3]

ABSTRACT

Water, an essential resource, has got stern importance in checking its quality due to the influence of various external factors like industrial emissions, acid rain, dumping of degradable and non-degradable wastes. Drinking impure water has a direct impact on public health and life. Hence, it is essential to ensure the quality of drinking water before consumption. This project work focuses on predicting the water quality of well water in Chengalpattu district. The work in this project is two-fold, i) classify the water quality using machine learning models such as Support Vector Machine (SVM), Decision Tree (DT) and Random Forest (RF) and ii) predict the quality using rough set theory (RST). Rough set theory is used to identify dependencies within the data and to handle uncertainty and incomplete information in water quality datasets. Rough set theory facilitates the identification of indiscernibility relations and boundary regions within the data, enabling the classification of water samples that are on the edge of different quality categories. The rough set principles are used to handle uncertainty and ambiguity in water quality classification. The water quality can be classified into four classes excellent, good, poor, and very poor. The results of this study demonstrate the efficiency of machine learning algorithms and rough set theory in water quality analysis.The models are then tested and evaluated to find the best suitable model by analysing the accuracy of prediction, the precision and recall of all the models.

1.1. INTRODUCTION

Bad water has an impact on every living thing that consumes it. A polluted water source can lead to a number of imbalances, that in many cases, can be lethal. The majority of ambient water bodies, such rivers, lakes, and streams, have certain quality requirements that indicate how well they are made. Moreover, water specifications for different uses have their own criteria. For instance, irrigation water shouldn't be overly salty or contaminated substances that can harm the soil or plants, ruining ecosystems. Depend- ing on the particular industrial processes, water quality for industrial usage also requires different properties. Natural water resources, such ground and surface water, are some of the less expensive sources of fresh water. But still, human/industrial activity, as well as other natural processes, can occupy such resources. To determine the purity of the water, the quality can be evaluated. Because traditional methods require a lot of labour and time, an automated model must be created. Various machine learning and deep learn- ing algorithms has been built to predict the quality of water to date but they weren't accurate enough. The parameters considered were also not sufficient. They were not able to handle the multidimensional and imbalanced datasets. Hence, the work involves machine learning algorithms to meet the requirements that the previously used models failed to achieve. Due to its nonlinear nature, the prediction of water quality becomes a difficult task. But the application of various machine learning techniques has been becoming a powerful source for prediction. Quantities like pH, Alkalinity, Total Hardness, Calcium, Magnesium, Chloride, Total Dissolved Solids, Potassium, Sodium, Nitrate, Sulphate, Sodium Adsorption Ratio and Fluoride were extracted from a well in Chengalpattu Town are used in this research project. The goal of this work is to predict the water quality using Machine Learning Model and Rough Set Theory. Rough Sets Theory (RST) can be particularly useful for dealing with vague, imprecise, inconsistent and uncertain knowledge involving data to enable the classification and prediction. The water quality can be classified into four classes such as Excellent water, Good water, Poor water and Very poor water based on water quality index. However, the contribution of the study can be summarised as follows. Applying machine learning algorithms, namely Support Vector Machine, Decision tree and Random Forest for the prediction of Water Quality Classification (WQC). This research aims to explore the application of machine learning algorithms and rough set theory in water quality analysis. The study leverages machine learning algorithms such as decision trees, random forests, support vector machines to classify and predict water quality parameters based on available data. These algorithms analyze multiple water quality attributes and their interrelationships to identify patterns that correlate with specific water quality conditions. Furthermore, rough set theory is employed to handle uncertainty and incomplete information in water quality datasets. Rough set theory facilitates the identification of indiscernibility relations and boundary regions within the data, enabling the classification of water samples that are on the edge of different quality categories. The research methodology involves collecting water quality data from various sources and preprocessing the data to remove noise, handle missing values, and normalize variables. Machine learning algorithms are trained and validated to ensure robust and accurate models. The rough set-based classifier is developed using the principles of lower and upper approximations to handle uncertainty and ambiguity in water quality classification.

The results of this study demonstrate the efficiency of machine learning algorithms and rough set theory in water quality analysis. The developed models provide accurate predictions and classifications, helping to identify potential water quality issues and support decision-making for water resource management and conservation. The research contributes to advancing the field of water quality analysis by incorporating intelligent techniques that can streamline the process and improve the overall assessment accuracy. Based on the predicted values from these models, the accuracy of prediction, the precision rate and recall is analyzed to find the most suitable model.

1.2. MOTIVATION

The goal of this activity is to improve everyone's quality of life while reducing any negative effects on the economy or the environment and ensuring the safety and health of both individuals and communities. In order to understand the significance of water quality analysis, consider the following:

Human health: Drinking water is essential for human survival, and the quality of the water consume can significantly impact our health. Waterborne diseases like cholera, typhoid, and dysentery are caused by consuming contaminated water. Regular water quality analysis can help identify potential threats to human health and prevent the spread of waterborne diseases.

Economic impact: Poor water quality can have a significant economic impact on communities. Contaminated water can lead to illnesses and dis- eases, causing lost productivity, increased healthcare costs, and reduced economic activity.

Environmental impact: The quality of drinking water is also closely tied to the health of the environment. Water quality analysis can help identify potential sources of contamination and prevent further damage to the environment.

1.3. LITERATURE SURVEY

This study's primary goal is to demonstrate how to forecast water quality in order to ensure that it is safe to consume. Traditional techniques, which are inaccurate and time-consuming, are used in laboratories to analyse the quality of the water. This work carries out ML techniques for removing this problem. In the work, Chengalpattu Well Water, Water samples were collected in Tamil Nadu, and they were examined using various ML like SVM, DT as well as RF. DT was found to be best suitable algorithm among the three ML algorithms having the highest precision of 96%.

Advanced artificial intelligence (AI) algorithms are created in Aldhyani, Al-Yaari, Alkahtani, and Maashi (2020) to predict water quality index (WQI) and water quality classification (WQC). Artificial neural network models, specifically the long short-term memory (LSTM) deep learning algorithm and nonlinear autoregressive neural network (NARNET), have been created for the WQI prediction. Also, for the WQC forecasting, three machine learning algorithms—namely, support vector machine (SVM), K- closest neighbour (KNN), and Naive Bayes—have been applied. The created models were assessed based on several statistical criteria, and the employed dataset has 7 significant parameters. The findings showed that the suggested models can forecast WQI properly and categorise the water quality according to superior robustness. The WQI values were predicted by the NARNET model, which performed marginally better than the LSTM, and the WQC values were predicted by the SVM method, which had the best accuracy (97.01 percent). The accuracy for the testing phase was also similar for the NARNET and LSTM models, with only a small difference in the regression coefficient (RNARNET = 96:17 percent and RLSTM = 94:21 percent). The management of water resources can benefit greatly from this kind of promising research. The author of Zavareh and Maggioni (2018) discusses the applications of the rough set. A novel approach that addresses the ambiguity and uncertainty that are highlighted in decision-making is known as rough set theory. Data analysis, the finding of new, useful information, and autonomous decision-making are all significantly impacted by the discipline of data mining. The rough set theory provides a workable method for extracting decision rules from data. A discussion of data representation using rough set theory, including pairs of attribute-value blocks, information table reductions, indiscernibility relations, and decision tables, is presented in this paper. It also introduces the basic concepts of rough set theory and other aspects of data mining. Moreover, notions for several potential rule sets are discussed, as well as the rough set approach to lower and higher approximations. Last but not least, a brief explanation of the data mining system's applications is provided.

A recently created time-series data-based water quality prediction system was discussed by Muharemi, Leon (2019). The information was gathered from a German public water company. SVMs, linear discriminant analysis (LDA),

logistic regression, artificial neural network (ANN), long short-term memory (LSTM), deep neural network (DNN), and recurrent neural network (RNN) were some of the deep learning and machine learning models that were implemented. Time, turbidity, pH, electrical conductivity, water temperature, chloride (Cl), redox chlorine dioxide, and flow rate are the variables they selected. The necessity of water quality analysis utilising various methodologies is discussed in Ritabrata Roy (2019). To determine if a water supply is suitable for the intended application, an assessment of the water quality is required. To establish if the water is suitable for usage, a number of water quality characteristics are evaluated and compared to their standard values. The processes for assessing the water have been standardised as a result of extensive research. These recommendations are succinctly covered in one place in this article for the researchers' and analysts' convenience. Thus, getting a general idea of the standards and practises for water quality assessment may be useful for them.

Ahmed, Mumtaz, Anwar and Irfan (2019) introduced a new methodology that takes into account the variables pH, temperature, total dissolved solids (TDS), and turbidity. To predict the water quality of Rawal Water Lake, researchers used 15 supervised machine learning algorithms, including random forest, multiple linear regression, polynomial regression, gradient boosting algorithm, SVMs, ridge regression, lasso regression, elastic net regression, neural net/ multi-layer perceptrons (MLP), logistic regression, stochastic gradient descent, K nearest neighbour, decision tree, and bagging classifier. The major goal of this research is to use a machine learning algorithm to forecast the hydro-chemical parameters that will determine the water quality in a particular area of Chennai. For the machine learning procedure, SVM, decision trees, random forests, logistic regression, and naive Bayesian were selected. The programme was developed to include the water quality index (WQI) as one of the primary limits while evaluating the quality of the water for each algorithm. 10 years of data were gathered to evaluate the hydro-chemical makeup of lake water. On a time scale of 10 years, the lake's water is sampled every month. Based on Bureau of Indian Standards/American Public Health Association (BIS/APHA) standards, the chosen parameters were investigated and estimated. The WQI used as a basic limit in coding to anticipate the current level of water quality was evaluated using these standards. The research area is the Korattur Lake, which is situated north of the railway line connecting Chennai and Arakkonam. In the western portion of the city, it is one of the biggest lakes. Ambattur Lake, Madhavaram Lake, and Korattur Lake are part of a chain of three lakes. The accuracy of the water quality, precision, and execution time were predicted using machine learning models.

In Zavareh and Maggioni (2018) the author explains about the applications of Rough set. Rough set theory is a new method that deals with vagueness and uncertainty emphasized in decision making. Data mining is a discipline that has an important contribution to data analysis, discovery of new meaningful knowledge, and autonomous decision making. The rough set theory offers a viable approach for decision rule extraction from data. This paper, introduces the fundamental concepts of rough set theory and other aspects of data mining, a discussion of data representation with rough set theory including pairs of attribute-value blocks, information tables reducts, indiscernibility relation and decision tables. Additionally, the rough set approach to lower and upper approximations and certain possible rule sets concepts are introduced. Finally, some description about applications of the data mining system with rough set theory is included.

In Madeira, Tasci and Celebi (2021) the proposed work is on Student Performance Prediction Using Rough Set Theory And Back propagation Neural Networks. This work which focuses to evaluate the usefulness of a model using Rough Set Theory (RST) and Back propagation Neural Network (BPNN) in effectively predicting the students' overall performance. The dataset used consists of 10 different attributes and one decision factor belonging to 53 students collected from a language course which administers in-person education with the aid of an online platform for assignments. RST was implemented in order to reduce the number of attributes used as input in the neural network and the BPNN made an accurate prediction using only 5 of the initial attributes. Thus outperforming a model based solely on BPNN used on the original dataset and reducing computational costs.

In Ritabrata Roy (2019) gives the importance of Water Quality Analysis using different techniques. Assessment of water quality is essential to check the suitability of a water source for the designated use. Several water quality parameters are assessed and compared with their standard values to determine the acceptability of the water to be used. After prolonged research, the procedures for the assessment of the water have been standardized. This article such guidelines are discussed concisely in one place for the convenience of the researchers and analysts. Thus, it may be helpful for them to get an overview of the water quality assessment standards and procedures.

In Chelly Dagdia, Zarges, Beck and Lebbah gives the in- sights about data preprocessing, namely feature selection, on a vast amount of data and high dimensional attribute set, is a significant problem in the knowledge discovery process. Several approaches have been put forth in the literature to address this problem, with varying degrees of success, as the majority of these approaches require additional information about the in- put data being used for thresholding, the specification of noise levels, or the use of feature ranking techniques. Rough set theory (RST) can be used to identify dependencies within data and decrease the amount of attributes contained in an input data set while using just the data

itself and requiring no additional information, so overcoming these restrictions. RST is extremely computationally demanding, hence it is limited when dealing with large data sets. This research focuses to provide an efficient and scalable rough set theory-based method for large scale data preprocessing, particularly for feature selection, within the Spark framework. Our thorough tests, which took into account data sets with up to 10,000 characteristics, showed that our suggested method achieves a good speedup and successfully completes its feature selection work without sacrificing performance. Thus, it is relevant to big data.

In Akther and Tharani gives about the impact of ground water due to the anthropogenic activities such as rapid urbanization, industrialization, pollution, heavy agricultural activities the quality of ground water is diminishing. Therefore, assessment of quality of the ground water resource and the related hydro chemical study is inevitable to undertake suitable management strategies to ensure the water resource is fit for human needs. Geographic Information System (GIS) an effective tool for storing, managing, and displaying spatial data which encountered in water resources management. Spatial interpolation is tool in GIS used to discover the values of unknown points. It is a technique of estimating the values of properties at unsampled locations based on the set of observed values at known locations. Understanding the groundwater quality is important, as it is the main factor determining its suitability for drinking, agricultural and industrial purposes. WQI is defined as a technique of rating that provides the composite influence of individual water quality parameter on the overall quality of water. Water quality and its suitability for drinking purpose can be examined by determining its quality index for human consumption. It is an arithmetical tool used to convert large number of water quality parameters into a single cumulatively derived number. Water quality indices are such approaches which minimize the data volume to a great extent and simplify the expression of water quality status. Objectives of this study are to analyze and assess the distribution pattern of each water quality parameters and to explore the water quality index of Vengalacheddikulam DS (Divisional Secretariat) division.

2. METHODOLOGY

The methodology used in this research work is shown in the following Fig 1. Chengalpattu well data about water was gathered and preprocessed using SMOTE to address the class imbalance in the initial dataset. The proposed work involved machine learning methods such as DT, SVM & RF and Rough Set Theory to predict the water quality. Metrics including accuracy, precision, and recall were used to assess and analyse the built-in models.

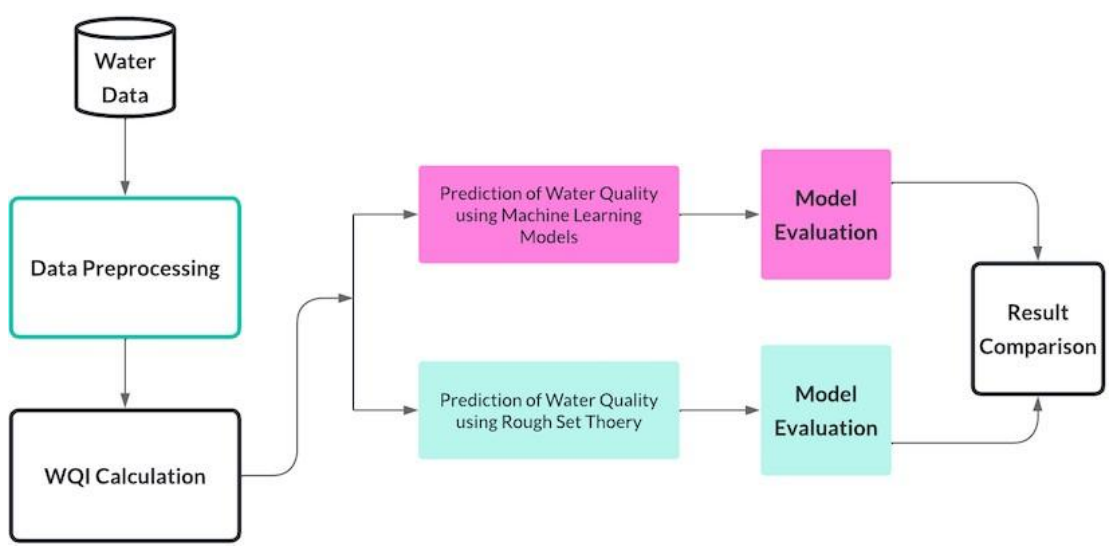

Fig 1. Architecture Diagram

2.1. Dataset Collection

The Chengalpattu Town, Tamil Nadu, India, well was used to obtain the dataset for this study. It has 13 parameters and 10248 samples. pH, Alkalinity, Total Hardness, Calcium, Magnesium, Chloride, Sulfate, Total Dissolved Solids, Potassium, Sodium, Nitrate, Sodium Adsorption Ratio, and Fluoride are the parameters for the dataset. The dataset is unlabeled. Based on the Water Quality Index, these data are divided into several categories, including Excellent water, Good water, Poor water, and Extremely Poor water.

2.2. Dataset Description

The distribution of both classes with more than two can be seen in Fig 2. and Table 1 shows the Chengalpattu well used to describe water data for multiple class classification.

Table 1 - Description of the dataset

Description of the dataset				
Name of the source	**Records quantity**	**Total parameters**	**Classes no**	**Class distribution**
Chengalpattu Well Water	10248	13	4	Excellent - 1420 Good - 2201 Poor - 4299 Very Poor - 2328

2.3. Data Splitting

It is required to divide the data into training and testing sets prior to training the machine learning model. After dividing the data, the model is trained and tested using a specific subset of the data to determine how accurate it performs. During training and testing, the data were divided in a 4:1 ratio. A total of 10,248 samples were used, of which 2562 were used for testing and 7686 for training.

2.4. Water Quality Index (WQI)

The Water Quality Index, or WQI, provides a numerical representation of the overall quality of water for any intended purpose. It is characterised as a score that reflects the combined impact of many water parameters that were taken into account while calculating the Water Quality Index (WQI). The indices are among the best tools for informing the public, policymakers, and those in charge of managing water quality about trends in water quality. The intended use of the water determines the relative importance of several parameters in the construction of the water quality index. The main consideration is if it is fit for human consumption. WQI was determined using the weighed arithmetic index approach.

Then, the WQI is found by:

$$WQI = \sum Q_i * W_i \quad (3.1)$$

where Q_i is the quality rating scale for each parameter i, calculated by equation (3.2) S_i is the recommended standard value of parameter i C is the parameter i.

$$Q_i = C/S_i \quad (3.2)$$

W_i is the relative weight of each parameter and it is given by weight of the each parameter to the total weight.

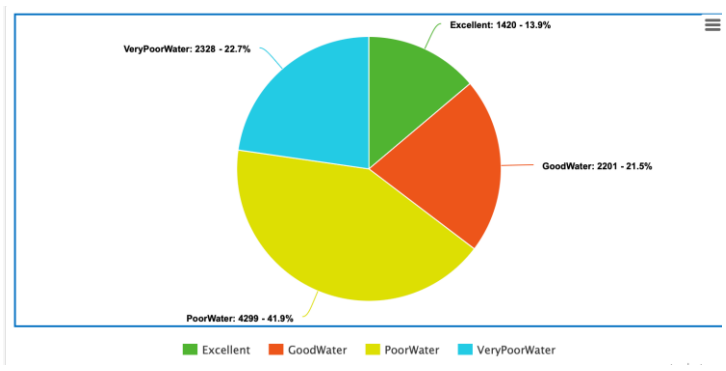

Fig 2. Distribution of multi class dataset

2.5. Data Preprocessing:

2.5.1. Data Preparation

Data preparation is the process of transforming raw data into a very specific format that is required for machine learning algorithms so that they produce useful insights. Good data preparation produces clean and well- curated data

which leads to more practical, accurate model outcomes.

2.5.2. Handling missing values

There are several missing values in the dataset. Missing values are distributed in various columns. There are 2 approaches taken, if all attributes have 0 or NaN value then every value in the data row will be deleted. If the missing value in a row is only in one or two columns, data imputation is carried out by looking for the average value before and after the missing values.

2.5.3. Balancing Classes

One of the most popular oversampling techniques to address the class imbalance issue is SMOTE (synthetic minority oversampling technique). By randomly boosting the minority class by replicating them, it seeks to balance the distribution of classes. SMOTE creates new minority instances by combining minority instances that already exist. For the minority class, it creates virtual training records using linear interpolation. For each example in the minority class, one or more of the k-nearest neighbours are randomly chosen to serve as these synthetic training records. Following the oversampling procedure, the data is rebuilt and can be subjected to several categorization models.

3. METHODS

The well water sample data used to categorise the models is computed using the machine learning techniques listed below. The study of water quality prediction utilised the DT, SVM, RF and RSC models.

3.1. Support Vector Machine

A well-liked machine learning approach called Support Vector Machine (SVM) is utilised for both classification and regression applications. It operates by locating the ideal hyperplane in a high dimensional space that divides the data into various classes. SVM searches for the hyper-plane that optimises the margin between the several classes of data while classifying data. The margin is the separation between the nearest data points from each class and the hyperplane. The SVM can generalise to new data more effectively the wider the margin is. A method known as the kernel trick is used by SVM to translate the data into a higher-dimensional space where it is more likely to be separable because not all data can be properly separated by a hyperplane. There are various kernel function types, including the linear, polynomial, and radial bas function (RBF). SVM can be used for regression tasks in addition to classification jobs by locating a hyperplane that roughly approximates the association between the input variables and the output variable. Over other machine learning algorithms, SVM has a number of advantages, including the capacity to handle high-dimensional data, robustness to outliers, and the ability to operate with both linear and non-linear data. SVM can overfit when the data is noisy or the sample size is short, and it can also be sensitive to the parameters and kernel function selection.

3.2. Decision Tree

A common machine learning approach for solving classification and regression issues is the decision tree. It is a kind of supervised learning technique in which the algorithm creates a tree-like model of choices and potential outcomes. Recursively segmenting the dataset into smaller subsets based on the values of one or more input features, the decision tree works until a stopping requirement is satisfied. The method makes a judgement at each internal node of the tree depending on the value of a specific feature, then proceeds on to the following node until it reaches a leaf node that reflects the model's output. Decision trees are appealing because they are simple to understand and depict. They are relatively quick to learn and can handle both categorical and numerical data. They may not work well on datasets with a lot of features or if the link between features and output is complex, and they can be prone to overfitting, especially if the tree is deep or complex.

3.3. Random Forest

Machine learning uses the well-known ensemble learning algorithm random forest for both classification and regression problems. A large number of decision trees are built during the training phase of this supervised learning method, which then outputs the class or mean prediction of each tree. A huge number of distinct decision trees work together as an ensemble in a random forest. The findings are aggregated after the full dataset is sampled into smaller subsets, each of which is then trained using a decision tree. The class with the highest votes becomes the prediction made

by our model. The random forest's individual trees each spit forth a class prediction. The model will be more accurate if there are more trees in it. By integrating the output of various decision trees on various samples of the data set, random forest is used to increase accuracy while reducing variation in the forecasts.

3.4. Rough Set Theory

Rough Set Theory (RST) is a mathematical framework for dealing with uncertainty and vagueness in data. It was introduced by Polish computer scientist Zdzislaw Pawlak in the early 1980s. The basic idea behind RST is to represent the knowledge about a set of objects or attributes using a collection of decision rules. These rules are defined in terms of lower and upper approximations of the set, which are subsets of the set that provide lower and upper bounds on the objects or attributes that satisfy the rule. RST can be used for feature selection, attribute reduction, and data analysis in various domains, including machine learning, data mining, and artificial intelligence. It has been applied in many real-world applications, such as image process- ing, bioinformatics and business intelligence. One of the advantages of RST is its ability to handle incomplete and uncertain data, which is common in many real-world applications. It also provides a natural way to deal with imprecision and vagueness in data, which is not possible in traditional set theory. However, RST has some limitations, including its sensitivity to noise and the need for domain knowledge to define the decision rules. It also has limited scalability when dealing with large datasets, which can be a challenge in some applications.

3.5. Concepts of Rough Set Theory

Information System: An information system consists of a universe of objects and a set of attributes that describe these objects. It is represented as a table or matrix where rows represent objects and columns represent attributes. In rough set theory, the lower approximation of a set A is the set of all objects that definitely belong to A, while the upper approximation is the set of all objects that possibly belong to A. The boundary region of a set A is the set of objects that are neither definitely in A nor definitely outside A. The indiscernibility relation is a fundamental concept in rough set theory. It defines the similarity or equivalence relation between objects based on the values of their attributes. Objects that cannot be distinguished based on their attribute values are considered indiscernible. A reduct is a minimal subset of attributes that preserves the discernibility information of the original set of attributes. It means that removing any attribute from the reduct would lead to loss of information. Reducts are used to simplify the decision rules generated from rough sets. Decision rules in rough set theory are if-then statements that describe relationships between the attribute values of objects and their corresponding class labels. These rules are derived from the lower and upper approximations of the decision attribute.

4. RESULTS

As observed in Table 3, the models perform well in multi-class classification with RSC having highest accuracy of 100%.

Table 3 - Results of multi-class classification

Algorithms	Precision	Recall	Accuracy
SVM	0.26	0.25	0.26
DT	0.96	0.95	0.96
RF	0.92	0.91	0.92
RSC	1.00	1.00	1.00

5. DISCUSSION

Chengalpattu well water data from Tamil Nadu, India was used to execute the DT, SVM, RF and RSC models for water quality prediction. On the basis of measures like accuracy, precision, and recall, the models were assessed. For multi class classification of Chengalpattu well water data, machine learning methods like Decision Tree (DT), Support Vector Machine (SVM), Random Forest (RF) and Rough Set Classifier (RSC) were investigated. DT produced 96%, SVM produced 26%, RF produced 92% and RSC produced 100%. The results were displayed in the Fig 3.

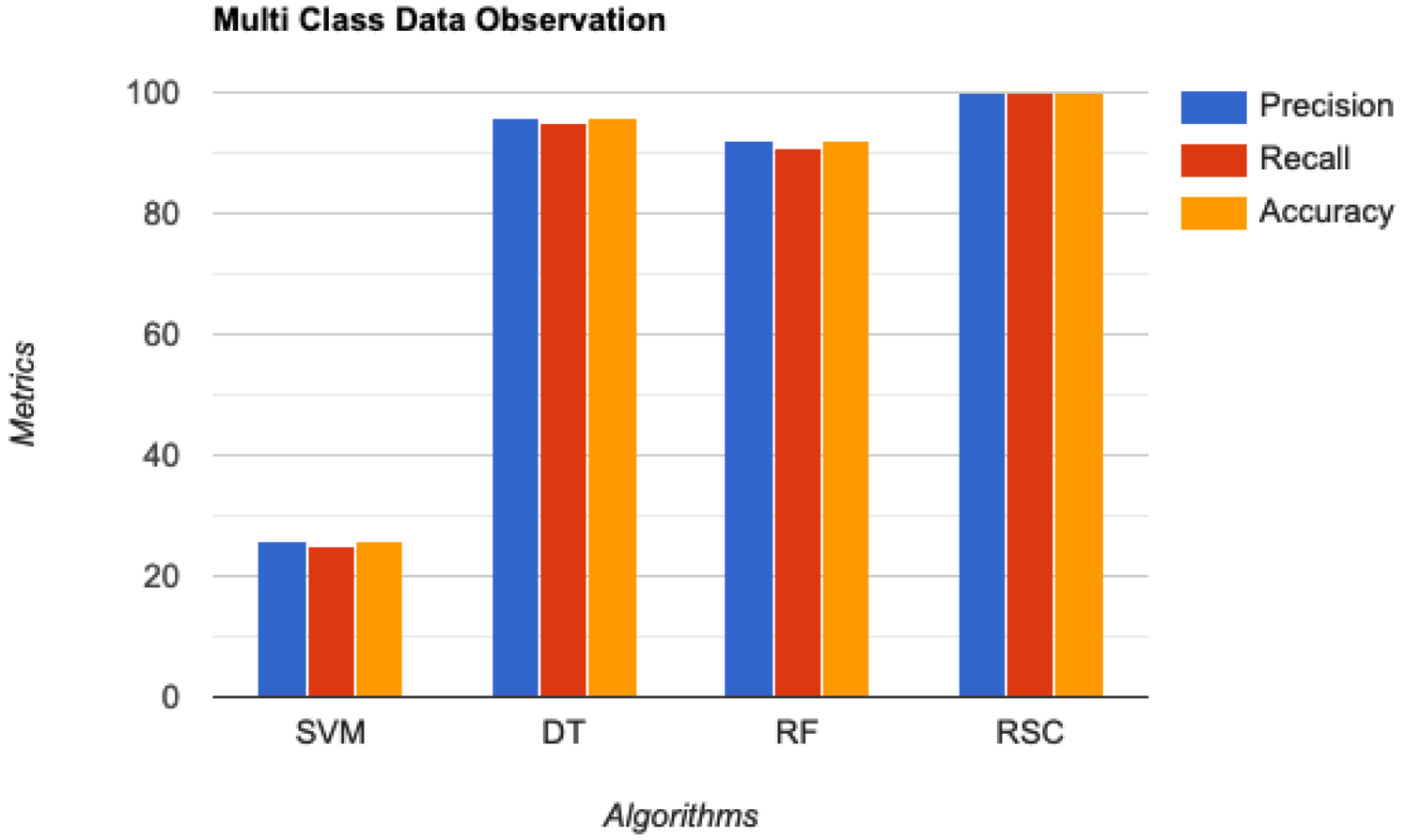

Fig 3. Bar chart comparison of ML Algorithms

6. CONCLUSION

This project aimed to explore the capabilities of machine learning algorithms and rough set theory in effectively predicting and classifying the quality of well water in the Chengalpattu district. In the initial phase, three popular machine learning models, namely Decision Tree, Random Forest, and Support Vector Machine were employed to perform classification tasks on the data collected. In the second phase of the project, a basic set theory-based classifier specifically the Rough Set Classifier is utilized to predict and classify well water quality. Set theory provides a mathematical foundation for understanding and analyzing uncertainty in data, making it a suitable choice for this classification task. The results indicated that the Rough Set Classifier and Decision Tree Model yielded more precise outcomes than the Random Forest and Support Vector Machine models. Although the rough set classifier and decision tree model demonstrated similar levels of accuracy, there was a notable difference in their execution times. The decision tree model takes only 0.05 seconds for execution where the rough set classifier takes 1484 seconds for execution. The decision tree model exhibited significantly lesser execution time compared to the rough set classifier. This means that the decision tree model required fewer computational resources and offered faster predictions and classifications.

7. REFERENCES

1. Zavareh, M., Maggioni, V. (2018). Application of rough set theory to water quality analysis: A case study. Data, Vol.3, No.4, pp.103-111.

2. Dorugade, S. P., Sawant, R. S., Godghate, A. G. (2017). Analysis of Water Quality of Dug Wells from Ajara Town, Western Maharashtra,
India. Imperial Journal of Interdisciplinary Research, Vol. 3, No. 3, pp. 780-784.

3. Alaluddin Khan, Ghufran Ahmad Khan, Mohammad Shahid, Jian Ping Li, Asad Malik, Shadma Parveen (2019). Application of Rough Set The-
ory in Data Mining.Conference paper in Journal of Convergence Information Technology, Vol. 5, pp. 22-36.

4. Inghua Zhang, Qin Xie, Guoyin Wang (2016). A survey on rough set theory and its applications. CAAI Transactions on Intelligence Technology,
Vol. 1, pp. 323-333.

5. Pawlak, Z. (1982). Rough sets. International journal of computer information sciences, Vol. 11, pp. 341-356.

6. Prasad, D., Vara, V., Venkataramana, L. Y., Kumar, P. S., Prasannamedha, G., Soumya, K., Poornema, A. J. (2021). Prediction on water quality of a lake in Chennai, India using machine learning algorithms.Desalination and Water Treatment, Vol. 218, pp. 44-51.

7. Madeira, B., Tasci, T., Celebi, N. (2021). Prediction of student performance using rough set theory and backpropagation neural networks. Eu-
ropean Scientific Journal, Vol. 17, No. 7.

8. Prasad, D. V. V., Kumar, P. S., Venkataramana, L. Y., Prasannamedha, G., Harshana, S., Srividya, S. J., Indraganti, S. (2021). Automating wa-
ter quality analysis using ML and auto ML techniques.Environmental Research 202, https://doi.org/10.1016/ j.envres.2021.111720.

9. Aldhyani, T. H., Al-Yaari, M., Alkahtani, H., Maashi, M. (2020). Water quality prediction using artificial intelligence algorithms.Applied Bionics
and Biomechanics, Vol. 2020, pp. 12-14.

10. Ritabrata Roy, R. (2019). An introduction to water quality analysis.ESSENCE Int. J. Env. Rehab. Conserv, Vol. 9, No. 1, pp. 94-100.

11. Akther, M. S. R., Tharani, G. (2017). Assessment of water quality parameters and determination of water quality index of tube well water in
Vengalacheddikulam DS division, Vavuniya District, Sri Lanka. International Journal of Sciences: Basic and Applied Research (IJSBAR), Vol.
32, No. 3, pp. 317-328.

12. Chelly Dagdia, Z., Zarges, C., Beck, G., Lebbah, M. (2020). A scalable and effective rough set theory-based approach for big data pre- process-
ing.Knowledge and Information Systems, Vol. 62, pp. 3321- 3386.

13. Bpredki, B., Slowinski, R., Stefanowski, J., Susmaga, R., Wilk, S.(1998). ROSE-software implementation of the rough set theory.In Rough Sets and Current Trends in Computing: First International Conference, RSCTC, pp. 605-608.

14. Sikder, I. U. (2010). A granular computing approach to decision analysis using rough set theory. In Proceedings of the 2010 International
Conference on Industrial Energy and Operations Management, Vol. 08, pp. 9-10.

15. Pawlak, Z. (2004). Some issues on rough sets.In Transactions on Rough Sets I, pp. 1-58.

16. Thangavel, K., Pethalakshmi, A. (2009). Dimensionality reduction based on rough set theory: A review. Applied soft computing, Vol. 9, No. 1, pp.
1-12.
17. Albuquerque,L.G., DeOliveiraRoque,F., Valente-Neto,F., Koroiva,R., Buss, D. F., Baptista, D. F., Pinto, J. O.(2021). Large-scale prediction
of tropical stream water quality using Rough Sets Theory.Ecological Informatics, Vol. 61, pp. 101226-101232.

18. Pai, P. F., Lee, F. C. (2010). A rough set based model in water quality analysis.Water resources management, Vol. 24, pp. 2405-2418.

19. Ren, F., Zhou, X., Zhang, C. (2010). Application of rough set svm model in the evaluation on water quality. International Conference on Commu-
nications and Intelligence Information Security, pp. 156-159.

20. Huang, H., Liang, X., Xiao, C., Wang, Z. (2015). Analysis and assessment of confined and phreatic water quality using a rough set theory
method in Jilin City, China. Water Science and Technology: Water Supply, Vol. 15, No. 4, pp. 773-783.

21. Roy, R. and Majumder, M. (2017). Comparison of surface water quality to land use: a case study from Tripura, India. Desalination and Water
Treatment, Vol. 85, pp. 147-153.

22. Bello, R., Falcon, R., 2017. In: Wang, G., Skowron, A., Yao, Y., S lezak, D., Polkowski, L. (Eds.), Rough sets in machine learning: a review.
Springer, Thriving Rough Sets. Studies in Computational Intelligence, pp. 87–118. https://doi.org/ 10.1007/978-3-319-54966-8 5

23. Mac Parthalain, N. and Jensen, R. (2013). Unsupervised rough set-based dimensionality reduction. Information Sciences, Vol. 229, pp. 106–121.

24. Velayutham, C. and Thangavel, K. (2011). Unsupervised quick reduct algorithm using rough set theory. Journal Electronics Science Technology,
Vol. 9, pp. 193–201.

25. Azar, A.T., Anter, A.M. and Fouad, K.M. (2020). Intelligent system for feature selection based on rough set and chaotic binary grey wolf op-
timization. International Journal Computer Appl. Technology, Vol. 63, pp. 4–24.

26. Bagyamathi, M. and Inbarani, H.H. (2017). Prediction of Protein Structural Classes using Rough Set based Feature Selection and Classification
Framework. Journal Recent Res. Engineering Technology, Vol. 4, pp. 1–9.

27. Beniwal, S. and Arora, J. (2019). Classification and feature selection techniques in data Mining. Int. Jorunal Eng. Research Technology, Vol.1, pp. 2278–2284.

28. Inghua Zhang, Qin Xie, Guoyin Wang (2016). A survey on rough set theory and its applications. CAAI Transactions on Intelligence Technology,
Vol. 1, pp. 323-333.

29. Boundary Region Figure: https://www.geeksforgeeks.org/rough-set-theory-an-introduction/

30. Upper and Lower Approximation Equations: https://www.geeksforgeeks.org/rough-set-theory-an-introduction/