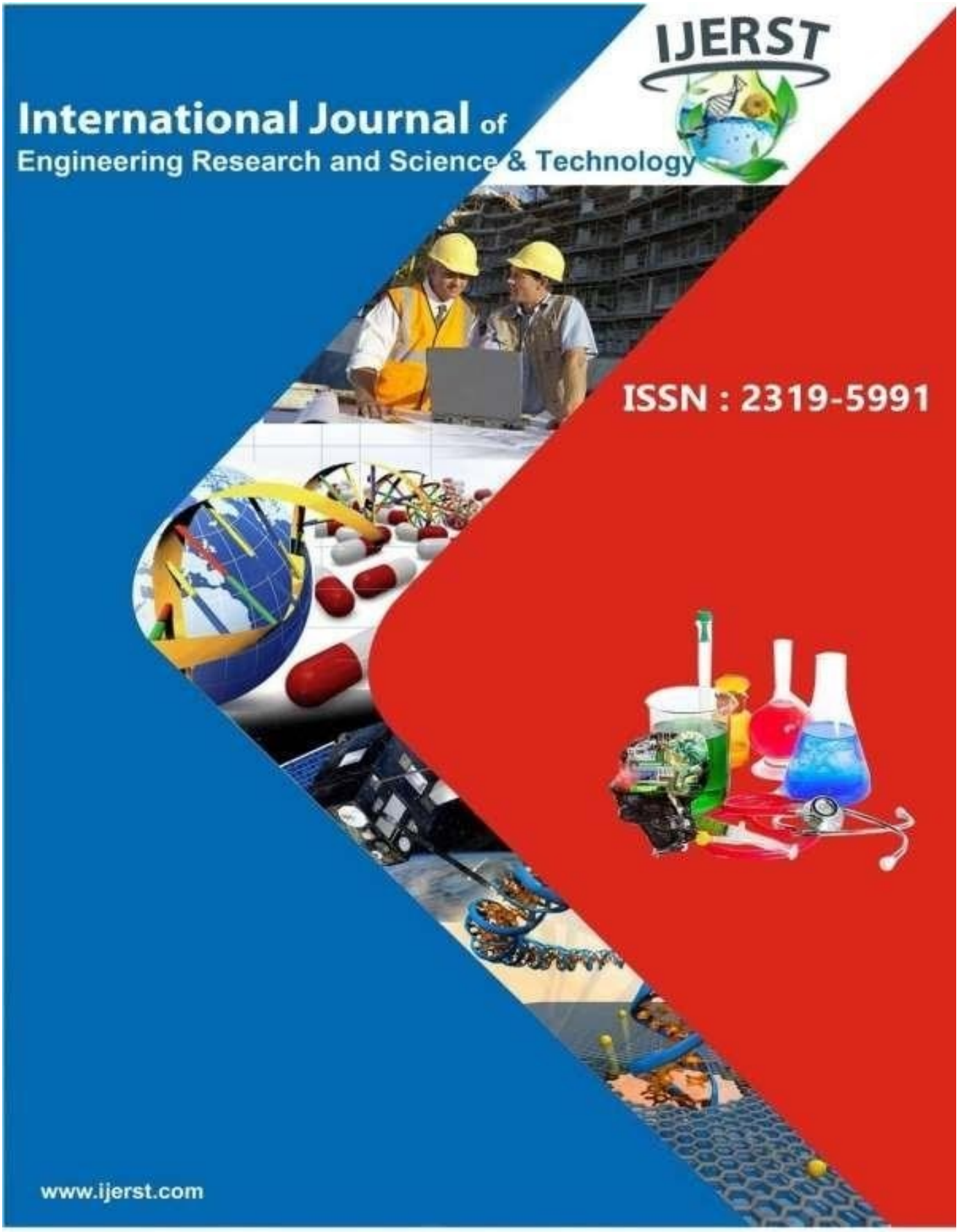

Email: editor@ijerst.com or editor.ijerst@gmail.com

International Journal of Engineering Research and Science & Technology

https://doi.org/10.62643/ijitce.2024.v20.i3.pp159-164

ISSN 2319-5991 www.ijerst.com

Vol. 20, Issue 3, 2024

Water Quality And Potability Using Machine Learning

[1] Mrs. D. Naga Malika, [2] D. Ambika, [3] K. Manga Tayaru, [4] G. Swathi Naga Pujitha

[1] Assistant Professor, Dept. of CSE, Ramachandra College of Engineering (A) ,Affiliated to JNTUK Kakinada , Eluru Andhra Pradesh, India.

[2,3,4,5] UG Students, , Dept. of CSE, Ramachandra College of Engineering (A) ,Affiliated to JNTUK Kakinada , Eluru Andhra Pradesh, India.

To Cite this Article

Y Nagendra Kumar et. al., Unlocking Potential : The Imperative for Research Facilities in Higher Education Institutions, International Journal for Modern Trends in Science and Technology, 2024, 10(04), pages. 12-20.https://doi.org/10.46501/IJMTST1004002

Article Info

Received: 16 March 2024; Accepted: 02 April 2024; Published: 03 April 2024.

ABSTRACT

Access to clean water to drink is crucial for good health, a fundamental human right, and an essential aspect of any health protection strategy. On a national, regional, and local level, this is significant as a development and healthcare issue. Access to safe drinking water is a global challenge. Traditional methods for water quality assessment can be time-consuming or require specialized equipment. This study explores the application of machine learning (ML) for portable water quality prediction. As a result, monitoring water quality now heavily relies on modelling and forecasting water quality. In this work, Machine learning algorithms are developed namely Decision tree, Random Forest, and Naïve Bayes algorithms for the dataset to predict the safety for human consumption. The study begins by compiling a comprehensive dataset comprising diverse water quality indicators such as pH levels, dissolved oxygen, turbidity, heavy metal concentrations, and microbial contaminants. Historical data from different sources, including public databases and local monitoring stations, are collected and preprocessed to ensure consistency and reliability. The used dataset has significant parameters and the developed model will be evaluated based on some statistical parameters. The goal of this algorithm is to create a model that predicts the value of a target variable
(Decision tree, Random Forest, and Naïve Bayes score) and whether the water is potable or not

KEYWORDS- research, imperative, unlocking, potential, higher, education, institutions

1. INTRODUCTION

All life on earth relies on water, making it one of our fundamental necessities. With a surface area of roughly 71% of the entire planet, it controls the majority of the available space. Water that is continuously extracted from water. Additionally, industries contribute to this the surface or the ground and used to such an extent to where it can no longer be used is sometimes referred to as water consumption and use. Water that has been contaminated by anthropogenic contaminants and is

International Journal of Engineering Research and Science & Technology

https://doi.org/10.62643/ijitce.2024.v20.i3.pp159-164

ISSN 2319-5991 www.ijerst.com

Vol. 20, Issue 3, 2024

unfit for human consumption is referred to as contaminated pollution. Waterborne diseases in aquatic organisms can be brought on by pathogens in this polluted water. Several main causes of water contamination are population growth and advancements in technology. If the present state of affairs persists, life on Earth will be untenable since there will be a tremendous demand for water and a potential shortage. As a result, maintaining water potability now heavily relies on the prediction and modeling of water quality. In this study, decision trees, random forests, and simple Bayes algorithms are built for the dataset to predict food safety for human consumption. The created model will be assessed using some statistical variables, and the employed dataset has 10 significant parameters. With the help of decision trees, a random forest, and a naive Bayes score, this methodology tries to produce a model that can predict the value of the desired variable and whether or not the water is potable. The primary focus given here is on predicting the potability of drinking water based on its physicochemical characteristics. The research seeks to develop a model that can provide accurate and timely data on the quality of drinking water, enabling policymakers and water resource managers to implement preventive measures and ensure the availability of safe drinking water for the general public. Also, the research aims to compare the performance and accuracy of various algorithms used for predicting water quality. Water quality and potability are critical issues that impact public health, agriculture, industry, and the environment. Ensuring access to safe drinking water is a fundamental necessity, and monitoring water quality is essential to prevent waterborne diseases and ecological damage. Traditional methods for assessing water quality involve physical, chemical, and biological testing, which can be timeconsuming, costly, and require significant human effort. Machine learning (ML) offers a transformative approach to enhance the efficiency, accuracy, and predictive capabilities of water quality assessment. The Drinkable water quality prediction is essential to ensure safe public health. It is a very much serious issue for a person to survive healthy life. Polluted drinking water can cause various kinds of health diseases. According to the survey, almost 3,575,000 people are died every year due to waterrelated diseases Predicting drinkable water is difficult for those countries that have limited drinkable water sources. In the industrial revolution, chemical dust causes the most water pollution. There is various kind of predicting methods to predict the drinkable water. Among those, neural network gray theory, statistical analysis, and chaos theory are the most useable techniques. For ideal model designing, statistical analysis is very much superior. For better prediction and research, a neural network delivers better performance. Drinking water quality mainly depends on essential measures, such as pH, hardness, sulfate, organic carbon, turbidity, and a few more . Machine learning techniques show significant prediction results in water quality prediction.

2. DISCUSSION

This paper also provides a comprehensive overview of machine learning applications in predicting specific water quality parameters such as chlorophyll-a, salinity, and dissolved oxygen. We provide a thorough examination of various machine learning algorithms' performance in predicting these parameters. Noteworthy is the recognition that the choice of the appropriate algorithm depends on data characteristics and local conditions. The inclusion of studies comparing different algorithms for specific parameters adds valuable insights for researchers seeking optimal prediction models.

Different machine learning algorithms can be selected based on the simulated water quality parameters. The Classification and Regression Tree method can accurately identify macroscopic bloom locations [71]. Decision Forest, Decision Jungle, and Boosted Decision Tree can be used to predict Escherichia Coli and enterococci levels [47]. Extreme Gradient Boosting, combined with the single model Support Vector Regression and Long ShortTerm Memory Long Short-Term Memory is better when estimating Chl-a concentrations. Artificial Neural Network, Gaussian Process, and Support Vector Regression can be used in predicting salinity concentrations. Random Forest performing slightly better than Support Vector Machine in the prediction of dissolved oxygen. Support Vector Regression, Extreme Gradient Boost, and Random Forest can be use in predicting multiple water quality parameters, and

International Journal of Engineering Research and Science & Technology

https://doi.org/10.62643/ijitce.2024.v20.i3.pp159-164

ISSN 2319-5991 www.ijerst.com

Vol. 20, Issue 3, 2024

ensemble machine learning model is also a good choice. Random Forests, Extreme Gradient Boosting, Multi-layer Perceptron, Convolutional Neural Network, and Shortterm Memory are all models which have shown good performance in predicting WQI, while the Random Forest algorithm can effectively select key water quality indicators. The Deep Neural Network algorithm predicts a subset more accurately.

The water quality index (WQI) is a crucial tool for assessing overall water quality, and this paper addresses the uncertainties associated with traditional WQI models. Machine learning algorithms, including Random Forest, Support Vector Machine, and Decision Tree, have been successfully applied to enhance WQI prediction. However, we point out that while these methods achieve effective results, they fall short of fundamentally improving WQI. Our discussion on ongoing efforts to reduce model uncertainty and improve architecture adds depth to the exploration of WQI prediction. The presented research underscores the transformative impact of machine learning on water quality prediction in coastal areas. Our review on the limitations of current models, the need for diverse datasets, and the consideration of evolving environmental conditions points to avenues for future research.

Fig: 1 Input Values For Water Quality

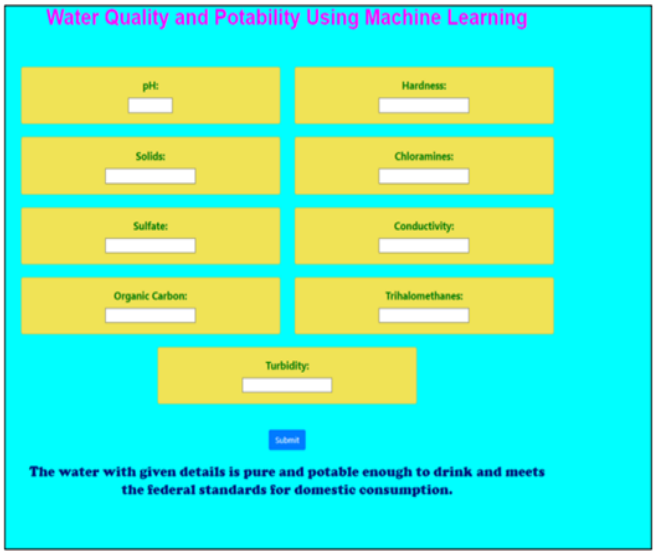

Fig: 2 Pure & Suitable For Drinking Water

3.RESULTS

Machine learning offers a promising approach to water quality and potability assessment. A system can be built to analyze various water quality parameters like pH, presence of bacteria, and minerals. By training the system on a large dataset of water samples with known potability, the model can learn to identify patterns that differentiate safe and unsafe water. This allows for quick and potentially cost effective prediction of potability compared to traditional lab tests. The system could be implemented as a user-friendly app or integrated with real-time sensors for continuous monitoring. However, it's important to remember that the model's accuracy depends on the quality of the training data, and traditional testing remains crucial for regulatory purposes.

1. Correlation Heatmap:

A correlation heatmap, displaying the relationships between different features within a dataset. It provides insights into the strength and direction of the relationship between variables. The correlation coefficient, ranging from -1 to 1, measures the extent of the linear relationship between two variables. A value between 0 and 1 suggests a positive correlation, indicating that the variables move in the same direction. On the contrary, a value between 1

International Journal of Engineering Research and Science & Technology

https://doi.org/10.62643/ijitce.2024.v20.i3.pp159-164

ISSN 2319-5991 www.ijerst.com

Vol. 20, Issue 3, 2024

and 0 indicates a negative correlation, signifying that the variables move in opposite directions. A value of 0 indicates no correlation between the variables. In this study, a correlation heatmap was employed to analyze the relationship among ten water quality parameters. The heatmap reveals a positive correlation of 0.082 between pH and Hardness, as well as a negative correlation of 0.17 between Sulphate and Solids. Detailed analysis can be observed.

Fig: 3 Input Values For Water Quality

Fig: 4 Impure & May Not Suitable For Drinking Water

2. Potability feature distribution:

A significant proportion of the collected samples is shown in Figure 4, approximately 60%, fall under the category of not potable, implying that they are unsuitable for human consumption. The causes of poor water quality can be attributed to various factors, including natural phenomena like erosion, and climate change, and anthropogenic sources like industrial effluents, agricultural drainage, and sewage disposal.

4. CONCLUSION

In conclusion, the development of a water quality assessment and potability determination system is a critical endeavor aimed at safeguarding public health, ensuring environmental sustainability, and facilitating informed decision-making in water resource management. Throughout the project, various aspects were addressed, including data collection, preprocessing, model development, testing, and evaluation. In this paper, the performance of Machine learning including Decision trees, Random Forest, and Naïve Bayes was evaluated to predict the water quality of drinking water. To this end, most dataset related well known components, such as pH, SO4, Na, Ca, Cl, Mg, HCO3, etc., were collected. Results indicated that the applied models have suitable performance for predicting water quality components, however, the best performance was related to the Random Forest. Results of Naïve Bayes indicated that its accuracy is acceptable for practical purposes. The lowest accuracy of models was related to Decision trees. The index of results of applied models shows that all three models slightly overestimate. Although the accuracy of model Naïve Bayes is less than that of model Random Forest, their DDR indices are close together. Furthermore, a comparison of the performance of applied models indicated that the outcomes of Random Forest and Naïve Bayes models were more reliable in comparison with the Decision tree. Predicting drinkable water is essential for environmental preservation and pollution prevention. It is necessary to provide clean drinking water in order to maintain excellent public health. Drinking water from safe sources can ensure the potability of the water. It becomes difficult to predict drinkable water accurately. The ideal learning algorithm is needed to prevent prediction errors. An intelligent model based on five different machine learning algorithms may be used to predict the potability of drinking water based on 10 standard parameters such as pH, hardness, organic carbon, and other factors. In this current work, artificial neural network achieved 98.12 percent accuracy with 0.75 percent training error. In future, the proposed model will be implemented to predict and analysis of different region drinking water.

https://doi.org/10.62643/ijitce.2024.v20.i3.pp159-164

ISSN 2319-5991 www.ijerst.com

Vol. 20, Issue 3, 2024

The high accuracy rates of these models underscore their potential for precise water quality assessment. Feature importance analysis highlights the critical role of specific variables, emphasizing the need for targeted monitoring and management. This study's findings contribute to the broader discourse on machine learning applications in environmental science. The identified variables and models can serve as valuable tools for water resource management, aiding in informed decision-making. Despite the promising results, it is crucial to acknowledge the study's limitations, including dataset size and variable scope. Future research should explore advanced strategies and incorporate additional parameters for a more comprehensive understanding of water quality dynamics in the region. Overall, this study not only showcases the capabilities of machine learning in water quality prediction but also underscores the importance of considering uncertainties for robust environmental assessments.

Conflict of interest statement:

Authors declare that they do not have any conflict of interest.

REFERENCES

[1] ^WATER QUALITY ASSESSMENT USING MACHINE LEARNING TECHNIQUES: A REVIEW" BY F. BEGUM ET AL.IN WATER SCIENCE AND TECHNOLOGY: WATER SUPPLY (2019). THIS PAPER PROVIDES AN OVERVIEW OF VARIOUS MACHINE LEARNING TECHNIQUES USED FOR WATER QUALITY ASSESSMENT AND THEIR APPLICATIONS.

[2] ^PREDICTION OF DRINKING WATER QUALITY PARAMETERS USING MACHINE LEARNING ALGORITHMS" BY M. ARSLAN ET AL. IN ENVIRONMENTAL MONITORING AND ASSESSMENT (2018). THE STUDY EXPLORES THE USE OF MACHINE LEARNING ALGORITHMS SUCH AS ARTIFICIAL NEURAL NETWORKS (ANN) AND SUPPORT VECTOR MACHINES (SVM) TO PREDICT THE QUALITY OF DRINKING WATER.

[3] ^MACHINE LEARNING APPROACH FOR WATER QUALITY PREDICTION IN DRINKING WATER TREATMENT PLANTS" BY M. KAYA ET AL. IN WATER SCIENCE AND TECHNOLOGY: WATER SUPPLY (2020). THE STUDY USES MACHINE LEARNING TECHNIQUES TO PREDICT THE QUALITY OF WATER IN A DRINKING WATER TREATMENT PLANT.

[4] ^A COMPARATIVE STUDY OF MACHINE LEARNING ALGORITHMS FOR PREDICTING WATER QUALITY PARAMETERS" BY S. SINGH ET AL. IN ENVIRONMENTAL MONITORING AND ASSESSMENT (2019). THE STUDY COMPARES THE PERFORMANCE OF VARIOUS MACHINE LEARNING ALGORITHMS IN PREDICTING WATER QUALITY PARAMETERS.

[5] ^MACHINE LEARNING APPROACH FOR PREDICTION OF WATER QUALITY INDEX IN A RIVER BASIN" BY S. PAL ET AL. IN WATER SCIENCE AND TECHNOLOGY: WATER SUPPLY (2020). THE STUDY USES MACHINE LEARNING ALGORITHMS TO PREDICT THE WATER QUALITY INDEX (WQI) OF A RIVER BASIN.

[6] ^LU, H., & MA, X. (2020). HYBRID DECISION TREE-BASED MACHINE LEARNING MODELS FOR SHORT-TERM WATER QUALITYPREDICTION. CHEMOSPHERE, 249, 126169. HTTPS://DOI.ORG/10.1016/J.CHEMOSPHERE.2020.126169

[7] ^ABBA, S. I., & PHAM, Q. B. (2020). IMPLEMENTATION OF DATA INTELLIGENCE MODELS COUPLED WITH ENSEMBLE MACHINE LEARNING FOR PREDICTION OF WATER QUALITY INDEX. ENVIRONMENTAL SCIENCE AND POLLUTION RESEARCH,27(33),41524-41539.HTTPS://DOI.ORG/10.1007/S11356-020-09689-X

[8] ^BUI, D. T., KHOSRAVI, K., & TIEFENBACHER, J. (2020). IMPROVING PREDICTION OF WATER QUALITY INDICES USING NOVEL HYBRID MACHINE-LEARNING ALGORITHMS. SCIENCE OF THE TOTAL ENVIRONMENT, 721,137612.HTTPS://DOI.ORG/10.1016/J.SCITOTENV.2020.137612

International Journal of Engineering Research and Science & Technology

https://doi.org/10.62643/ijitce.2024.v20.i3.pp159-164

ISSN 2319-5991 www.ijerst.com

Vol. 20, Issue 3, 2024

[9] ^DPHE.RAJSHAHI.GOV.BD. 2022. DEPARTMENT OF PUBLIC HEALTH ENGINEERING, RAJSHAHI.. [ONLINE] AVAILABLE AT: <HTTP://DPHE.RAJSHAHI.GOV.BD/> [ACCESSED 3 JUNE 2022].

[10] ^AKTER, T., JHOHURA, F. T., & AKTER, F. (2016). WATER QUALITY INDEX FOR MEASURING DRINKING WATER QUALITY IN RURAL BANGLADESH: A CROSS-SECTIONAL STUDY. JOURNAL OF HEALTH, POPULATION AND NUTRITION, 35(1). HTTPS://DOI.ORG/10.1186/S41043-016-0041-5

[11] ^GUIDELINES FOR DRINKING-WATER QUALITY, FOURTH EDITION. RETRIEVED APRIL 3, 2022, FROM HTTPS://APPS.WHO.INT/IRIS/BITSTREAM/HANDLE/10665/44584/9789241548151_ENG.PDF

[12] ^SARKER, I. H. (2021). MACHINE LEARNING: ALGORITHMS, REAL-WORLD APPLICATIONS AND RESEARCH DIRECTIONS. SN COMPUTER SCIENCE, 2(3). HTTPS://DOI.ORG/10.1007/S42979-021-00592-X 2

[13] ^EXPLAINING FEATURE IMPORTANCE BY EXAMPLE OF A RANDOM FOREST | BY . RETRIEVED APRIL 3, 2022, FROM HTTPS://TOWARDSDATASCIENCE.COM/EXPLAININGFEATURE-IMPORTANCE-BY-EXAMPLE-OF-A-RANDOMFOREST-D9166011959E

[14] ^DARMAWAN, M. F., ZAINAL ABIDIN, A. F., & KASIM,S. (2020). RANDOM FOREST AGE ESTIMATION MODEL BASED ON LENGTH OF LEFT HAND BONE FOR ASIAN POPULATION. INTERNATIONAL JOURNAL OF ELECTRICAL AND COMPUTER ENGINEERING (IJECE), 10(1), 549. HTTPS://DOI.ORG/10.11591/IJECE.V10I1.PP549-558

[15] ^DEEP NEURAL NETWORKS - KDNUGGETS. RETRIEVED APRIL3,2022,FROMHTTPS://WWW.KDNUGGETS.COM/2020/02/DEEP-NEURAL-NETWORKS.HTML

[16] ^UNDERSTANDING SUPPORT VECTOR MACHINE (SVM) ALGORITHM FROM. RETRIEVED APRIL 5, 2022, FROM HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2017/09/UNDERSTAING-SUPPORT-VECTOR-MACHINE-EXAMPLECODE/

[17] ^MATHEMATICS BEHIND SVM | MATH BEHIND SUPPORT VECTOR MACHINE. RETRIEVED APRIL 6, 2022,FROMHTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2020/10/THE-MATHEMATICS-BEHIND-SVM/

[18] ^ASRIADIE, M. S., MUBAROK, M. S., ADIWIJAYA,.(2018).CLASSIFYING EMOTION IN TWITTER USING BAYESIAN NETWORK. JOURNAL OF PHYSICS: CONFERENCE SERIES, 971, 012041. HTTPS://DOI.ORG/10.1088/1742-6596/971/1/012041