**Water potability prediction**



A Project Report in partial fulfillment of the degree

# Bachelor of Technology

in

# Computer Science & Engineering

## By

**2103A51217** LADE KAVERI

**2103A51161** MALYALA SPANDANA

**Under the Guidance of**

# D. Ramesh

**Submitted to**



# DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

## S R UNIVERSITY, ANANTHASAGAR, WARANGAL



**DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING**

**CERTIFICATE**

This is to certify that the Project Report entitled “MACHINE FAILURE PREDICTION” is a record of Bonafide work carried out by **L.Kaveri and M.Spandana**bearing Roll No(s) **2103A51217,2103A51161**during the academic year 2022-2023 in partial fulfillment of the award of the degree of ***Bachelor of Technology*** in **Computer Science Engineering** by the SR UNIVERSITY, WARANGAL.

# Supervisor Head of the Department

Mr. D. Ramesh Dr. M. Sheshikala

SR University SR University

**External Examiner**

**ACKNOWLEDGEMENT**

We express our thanks to course coordinator Mr. D. Ramesh, Asst. prof. for guiding us from the beginning through the end of the course project. We express our gratitude to head of the department CS&AI, Dr. M. Sheshikala, Associate Professor for encouragement, support and insightful suggestions. We truly value their consistent feedback on our progress, which was always constructive and encouraging and ultimately drove us to the right direction.

We wish to take this opportunity to express our sincere gratitude and deep sense of respect to our beloved Dean, School of Computer Science and Artificial Intelligence, Dr C. V. Guru Rao, for his continuous support and guidance to complete this project in the institute.

Finally, we express our thank to all teaching and non-teaching staff of the department for their suggestions and timely support.

# ABSTRACT

Water potability prediction is the process of determining whether or not water is safe for human consumption. It is a crucial aspect of public health and safety, as contaminated drinking water can lead to various waterborne diseases. Predicting water potability typically involves analyzing water samples for various chemical and microbiological parameters, such as pH, total dissolved solids, turbidity, and bacterial counts. Water potability prediction models can be used in a variety of applications, such as monitoring water quality in drinking water treatment plants, assessing the safety of water sources in rural areas, and providing early warning systems for water contamination events. By enabling early detection of water quality issues, these models can help prevent the spread of waterborne diseases and protect public health.

# Table of Contents

|  |  |  |
| --- | --- | --- |
| **S.NO** | **Content** | **Page No** |
| 1 | Introduction | 1 |
| 2 | Literature Review | 2 |
| 3 | Dataset description | 6 |
| 4 | Data Pre-processing | 7 |
| 5 | Data visualization | 8 |
| 6 | Methodology | 11 |
| 7 | Results | 16 |
| 8 | Conclusion | 17 |
| 9 | References | 18 |

**1.INTRODUCTION:**

Access to safe drinking-water is essential to health, a basic human right and a component of effective policy for health protection. This is important as a health and development issue at a national, regional and local level. In some regions, it has been shown that investments in water supply and sanitation can yield a net economic benefit, since the reductions in adverse health effects and health care costs outweigh the costs of undertaking the interventions.In recent years, machine learning and artificial intelligence have been used to predict water potability. Potable water is water that is safe for drinking and is free from harmful contaminants.

Water potability prediction is crucial in ensuring access to safe drinking water. This approach has the potential to significantly reduce the time and cost associated with traditional water quality testing methods, making it a pro

mising area of research for public health and environmental sustainability.Having an insight into all these algorithms, we have observed that the algorithms work different by generating the pattern among the available dataset and proceeding with prediction

# 2.LITERATURE REVIEW

In the various research papers we have referred that different Machine learning Algorithms have been used. The area of Artificial intelligence has been the suitable criteria to carry out predictions on the datasets by feature extraction and data pre-processing. The various machine learning algorithms that have been used are : Logistic regression, Support Vector Machine, Linear Regression, Discriminant analysis stocastic gradient descent and ridge regression, Naive Bayes classification.

Having an insight into all these algorithms, we have observed that the algorithms work different by generating the pattern among the available dataset and proceeding with prediction. The concepts of logistic regression, ridge regression, stocastic gradient, SVM are used with algorithms especially which follows a close correlation among the variables taken into consideration whereas decision tree, random forest, adaboost works with regenerating similarities by nodes, Deep learning is something that works by generating biases and weights in the layers, rule based takes the bulk values and signifies a rule in it, LSTM takes into account the input, output and forget gates considering the memories of occurrences of the information.

when sentences and language is a mattern of concern, NLP - Natural LanguageProcessing is the best methodology bcause they have those kind of algorithms that infer knowledge and interpretations of the txt and voice data just as we humans do. The packages of NTP are best suitable to proceed with our project and the dataset. In NLP the fed information is first converted into vector form and those numerical figures manipulate thorough the cycles and facilitates interpretations and clssifications. The recommonded algoritm to do the word embedding for data preprocessing is word2vec.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| SI NO DATE OF AUTHORS NAME METHADOLOGY DATASET ACCURACY  PUBLICATION | | | | | | | | | |
| 1 | | 2019/9/1 | | AK Kadam, VM Wagh, AA Muley, BN Umrikar, RN Sankhua | groundwater quality for drinking | ANN | | Water quality prediction | 25.75 to 129.07 |
| 2 | | 2018/9 | | asant Madhav Wagh, Dipak Baburao Panaskar | Health risk assessment of heavy metal contamination in ground water | KNN | | Metal contamination in ground water | 35.67 to 130.42 |
| 3 | | 2018/9/1 | | Vasant Wagh, Dipak Panaskar, Aniket Muley, Shrikant Mukate, Satyajit Gaikwad | Neural network modelling for nitrate concentration in ground water | ANN | | Concentration in ground water | 67.50% to 75% |
| 4 | | 2019/7/2 | | Vasant Madhav Wagh, Dipak Baburao Panaskar, Aniket Avinash Muley | Prediction of ground water suitability using artificial neural network modelling | ANN | | Irrigation using artificial neural network model | 85.2 |
| 5 | | 2018/10/1 | | Shrikant Mukate, Dipak Panaskar, Vasanth | Impact of anthropogenic inputs | ANN | | Water quality in chincholi | 87 |
| 6 | | 2019/1 | | Vasant Madhav Wagh, Dipak Baburao Panaskar, James A Jacobs, Shrikant Vitthal Mukate, Aniket Avinash Muley, Ajaykumar Krushna Kadam | Influence of hydrochemicak process on ground water quality through geostatical process | ANN | | Ground water quality | 93.2 |
| 7 | | 2016/11 | | Vasant Madhav Wagh, Dipak Baburao Panaskar, Abhay Mukund Varade, Shrikant Vitthal Mukate, Satyajit Kundlik Gaikwad, Ranjitsinh | Major ion chemistry and ground water quality | ANN | | Ground water quality | 95.8 |
| 8 | | 2002/2/1 | | Hesham R Lotfy, IG Rashed | Waste water treatment | svm | | Water potability prediction | 31.5 to 125 |
|  | | | | | | | | | |
| 9. | | 2017/5/15 | | Alberto Campisano, David Butler, Sarah Ward | Urban rain water harvesting | Deep learning | Rain water harvesting system | | 115-195 |
| 10 | 2021/1/1 | Saima Sadiq, Arif Mehmood, Saleem Ullah, Maqsood Ahmad | | rain water | CNN | water potable prediction | | 123-189 | |

**3. DATASET DESCRIPTION**

* It contains 2,463 rows 9 features and 1 target values extracted from different water bodies.
* The data was collected from the ministry of water resources,using the parameters:
* Ph
* Hardness
* Solids
* Chloramines
* Sulphate
* Conductivity
* Organic carbon
* Trihalomethanes
* Turbidity

Graphical user interface, application, table, Excel

Description automatically generated

* The water\_potability.csv file contains water quality metrics for 2,463 different water bodies.

**4.DATA PRE-PROCESSING**

The goal of data preprocessing is to ensure that the data is accurate, complete, and consistent before any analysis is performed.

In our dataset we removed a null values and we added zeroes in to it.This is the pre processing we used in our data set.

# 5.DATA VISULAZATION:

# The following are plotting of each feature against the target.

# Ph Hardness

# A picture containing chart Description automatically generated

# Solids Chloramines

A picture containing chart

Description automatically generatedA picture containing diagram

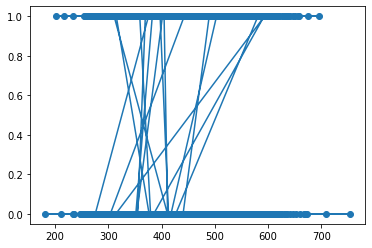
Description automatically generated

**Diagram

Description automatically generatedSalute Conductivity**

A picture containing diagram

Description automatically generated**Organic\_carbon Tryhalomethanes**



A picture containing diagram

Description automatically generated **Turbidity**

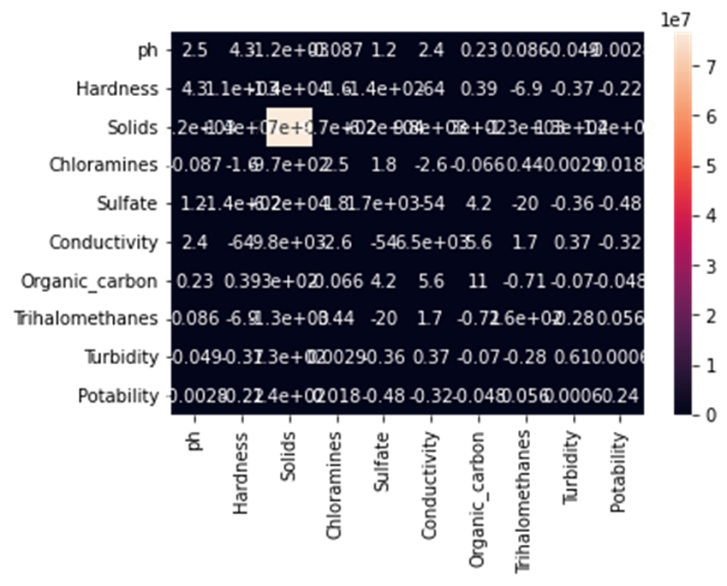
A picture containing diagram

Description automatically generated

**CORRELATION MATRIX:**



**COVARIANCE MATRIX:**

**6.METHADOLOGY:**

**1.Logistic Regression:**

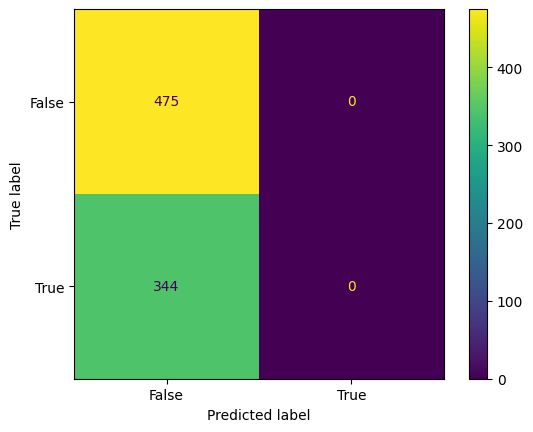
Logistic regression is a statistical method used for binary classificaton problems it works by fitting logistic functions to the input variables,which transfers the input variables into a range between 0 and 1.The output of a logistic regression is binary value.

**Accuracy value for logistic regression is:**

print(accuracy\_score(yp,y\_test))

0.57997557997558

**Graphical confussion matrix display:**



**2.Decision tree:**

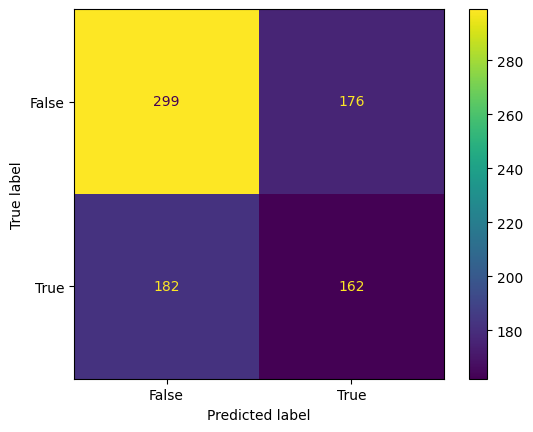
A decision tree algorithm that is used for both classification and regression problem it works by recursively partitioning the data into subsets based on the values of input features the algorithm starts by selecting the best feature to split the data into groups once the tree is bulit new instances can be classified by traversing the tree from the root node to leaf node based on the values of input.

**Accuracy for decision tree:**

print(accuracy\_score(yp,y\_test))

0.5628815628815629

**Graphical confussion matrix display:**



**3.KNN:**

The k nearest neighbours(K-NN) algorithm is a type of supervised machine learning it is used for both classification and regression.

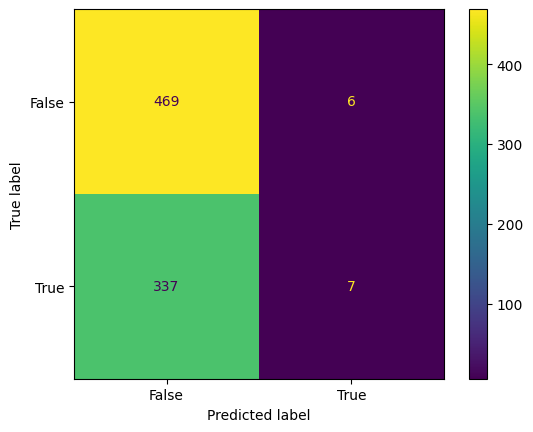
The KNN algorithm works by finding the k nearest data points to a given test point uses the class label to predict the class of the test point it is used to measure similarity between data points by euclidean distance

**Accuracy for knn:**

print(accuracy\_score(yp,y\_test))

0.5518925518925519

**Graphical confussion matrix display:**



**4.Navies-Bayes Theorem:**

Bayes Theorem is a Bayes' Theorem is a fundamental concept in probability theory, named after the Reverend Thomas Bayes. It provides a way to update our beliefs about the likelihood of an event occurring, based on new evidence or information that we receive.Bayes' Theorem is a fundamental concept in probability theory, named after the Reverend Thomas Bayes. It provides a way to update our beliefs about the likelihood of an event occurring, based on new evidence or information that we receive.Bayes' Theorem states that the probability of a hypothesis H, given some observed evidence E, is proportional to the prior p Bayes' Theorem is a fundamental concept in probability theory, named after the Reverend Thomas Bayes. It provides a way to update our beliefs about the likelihood of an event occurring, based on new evidence or information.

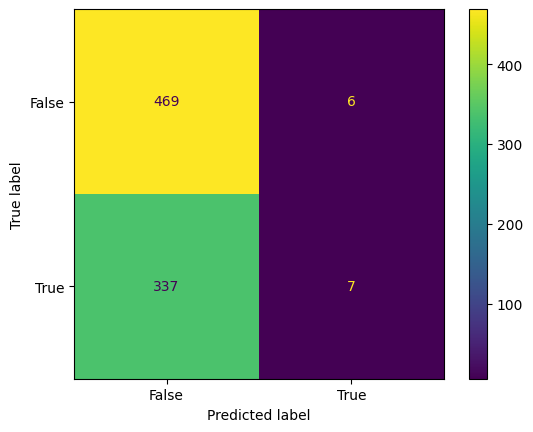
**Accuracy for navies bayes theorem:**

accuracy=accuracy\_score(yp,y\_test)

print(accuracy)

0.5885225885225885

**Graphical confussion matrix display:**



**5.Support vector machine:**

Support Vector Machine (SVM) is a powerful algorithm used in machine learning for classification and regression tasks. The goal of SVM is to find the best possible boundary that separates the data into different classes. This boundary is called a hyperplane, which can be linear or non-linear, depending on the type of SVM used. The SVM algorithm works by mapping the data points into a high-dimensional feature space and then finding the hyperplane that maximizes the distance between the closest points from each class. These closest points are called support vectors, and the distance between them is called the margin. The margin represents the degree of separation between the classes, and the larger it is, the better the classification performance of the SVM.

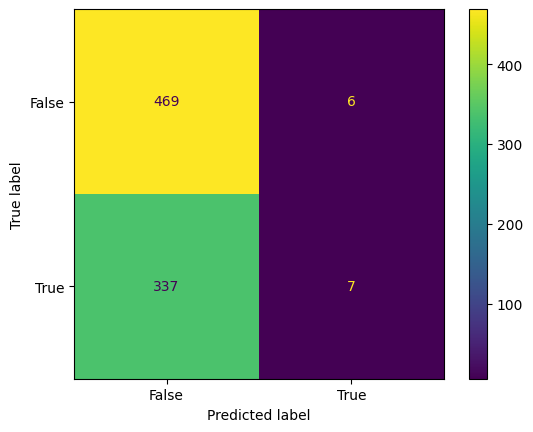
**Accuracy for svm:**

accuracy=accuracy\_score(yp,y\_test)

print(accuracy)

0.5811965811965812

**Graphical confussion matrix display:**



**7.RESULTS:**

**Accuracy through Logistic Regression:**

|  |  |
| --- | --- |
| Ml model | Accuracy |
| Logistic Regression | 0.57997557997558 |

**Accuracy through Decision Tree:**

|  |  |
| --- | --- |
| Ml model | Accuracy |
| Decision tree | 0.5628815628815629 |

**Accuracy through KNN:**

|  |  |
| --- | --- |
| Ml model | Accuracy |
| KNN | 0.5518925518925519 |

**Accuracy through Navies Bayes Theorem:**

|  |  |
| --- | --- |
| Ml model | Accuracy |
| Navies Bayes theorem | 0.5885225885225885 |

**Accuracy through SVM:**

|  |  |
| --- | --- |
| Ml model | Accuracy |
| Support vector machine | 0.5811965811965812 |

**8. CONCLUSION AND FUTURE SCOPE**

**CONCLUSION**

In conclusion, the data provides us the quality of water(potability) whether it is pure enough for the most critical of its indented uses, usually for human consumption.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 SVM.

**FUTURE SCOPE**

Water potability prediction is an important task that can have a significant impact on public health. Future research can focus on collecting more comprehensive and diverse data on water quality parameters from various sources, such as sensors, satellites, and citizen science projects. Future research can focus on identifying new features or combinations of features that can improve the accuracy of water potability prediction models.

**9.REFERENCES:**

[1] https://www.kaggle.com/datasets

[2] https://www.researchgate.net/publication/356666279

[3] https://www.karger.com/Article/Fulltext/505021