

IST 707 – Applied Machine Learning

**Building Machine Learning Models Predicting Water Quality, Cape Coral, FL**

Nikhil Sriram Budamaguntala​

*Abstract*

Water quality has faced a significant decline over the recent years because of growing urbanization, high rates of industrialisation, and improper garbage disposal. Predicting water quality is now an important issue in preserving the aquatic environment because it is necessary for a healthy civilization. In this project we will leverage with the support of machine learning, the possibility of predicting water quality in areas where the water is used for many purposes such as drinking, irrigation, and for a variety of animal species. Chemical compounds that are measured and that will constitute our dataset, pose high difficulty of analysis due to imbalance and missing values. On the other side, for machine learning models to be efficient, balance and full representation of data are equally important as the predicting model selection. No exception was the dataset of Cape Coral, city in southwest Florida, known for its many canals, where the issues of imbalance and missing values were prominent. Through this report we will explain why a combination of Denoising Autoencoder, undersampling and ANN for imputation, balancing and predicting respectively, are a good fit for the selected dataset on building a final model for water quality prediction. Existing methods use algorithms that decrease the accuracy of unbalanced datasets and increase computational complexity, with a primary focus on either water quality. This study explores existing methods and new ones for data balance with a higher accuracy. The techniques used for are SMOTE and undersampling where the one with better result is undersampling. After having a better-balanced dataset, Bagging Classifier + SVM, Bagging Classifier, KNN, Random Forests, Naive Bayes, Decision Tree models and artificial neural network (ANN) are explored for forecasting water quality. The results indicate that ANN model performs better with an accuracy of 95%.

## Introduction

Since we have high levels of pollution in nearly half of US (United States) rivers and more than one-third of lakes are polluted and unfit for swimming, fishing, and drinking. Water pollution is damaging to living species, and it causes a variety of illnesses. It has caused damage to the aquatic ecosystem, necessitating the search for potential solutions. Several global laws and guidelines have been established to protect water quality. This would also serve as a supporting tool for government and agencies on decision-making for water resources.  In Florida, “CERP dictates a complex funding arrangement whereby the federal and state governments shared construction costs equally, subjecting the plan to the unpredictable nature of the legislative budgeting process at both the federal and state levels”[[1]](#footnote-1)

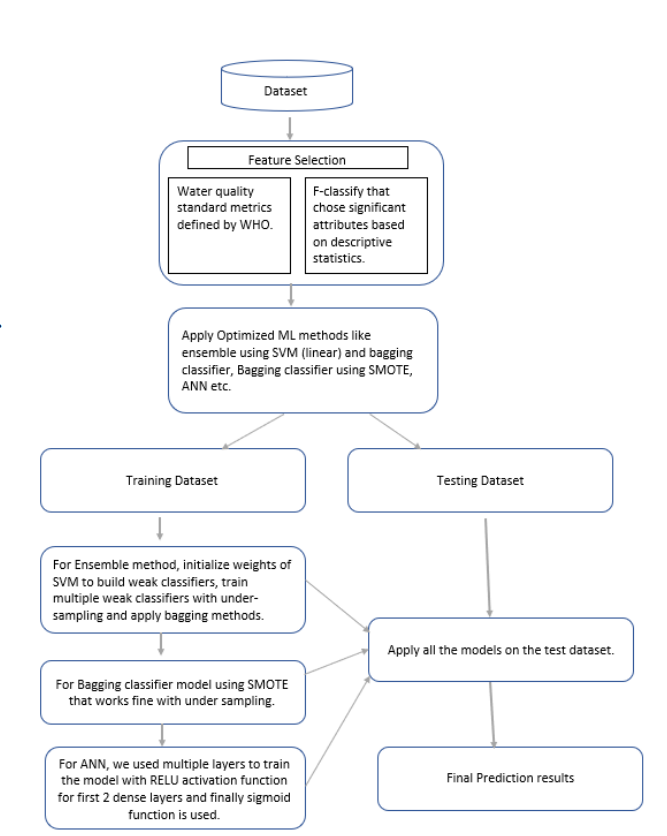
The goal for our project is to build a model for water quality prediction using a dataset of Cape Coral, Florida and give better insights on what range of chemical factors contribute on deciding the water quality levels and how much contaminated they are.

The selected dataset contains 18148 observations(records/entities) with 48 variables (attributes). Considering the more attributes, a dataset has lessened the chances that the predictive model would be stable. The initial filtering of attributes was performed by considering the standards defined by the World Health Organization (WHO). The following table shows the list of attributes selected which have a higher degree of importance in measuring water quality, and their range of values indicating a good quality of water.

In the following section are describe can be found an analysis of data to understand better their characteristics. Then we try to understand if there is a specific pattern for the missing values, which will then be helpful to define the model that should be used for the data imputation. After imputing the missing data, the next phase concerns balancing. Since one of the reasons for bias in machine learning models is unbalanced data, we try resolve this issue by trying several algorithms. The final step is building the machine learning models that predict whether the quality of water is “poor” or “good”.

**Project Workflow**

To implement this project, we followed specific steps. Initially we had to define the dataset to be used. The second step consisted in working with the dataset which consist of: selecting the important features, impute missing data and balance the water quality categories. For each of the steps of working with the dataset, several models were tested and the best performing one is chosen. Next, several training models are tested, such as naïve bayes, knn, random forest and finally artificial neural network. The project workflow is shown in figure 1.



*Figure 1: Project workflow*

## Dataset Description

We have explored three datasets for testing water quality, they are:

* <https://hub.arcgis.com/datasets/b0579ba7aa1145e090c3a74e295564df/explore?location=26.646897%2C-81.929073%2C11.71&showTable=true>
* <https://data.doi.gov/dataset/water-quality-data>
* <https://redivis.com/datasets/g8je-49bxaag3y>

Among these websites we have selected data provided by the city of Cape Coral because it has more significant features and helps to predict our model more accurately. The data set contains 18148 observations with 48 variables, it also has missing values 229 in water temperature, 385 in specific conductance, 299 in dissolved oxygen, 567 in pH, 293 in sampling depth, 303 in salinity, 18148 in dissolved oxygen saturation, 6956 in total dissolved solids, 12073 in Fecal coliforms, 18148 in organic nitrogen,15858 in volatile suspended solids, 15893 in volatile dissolved solids. Among all the features we need to select more significant and eliminate remaining fields. The data set contains 18148 observations with 48 variables.

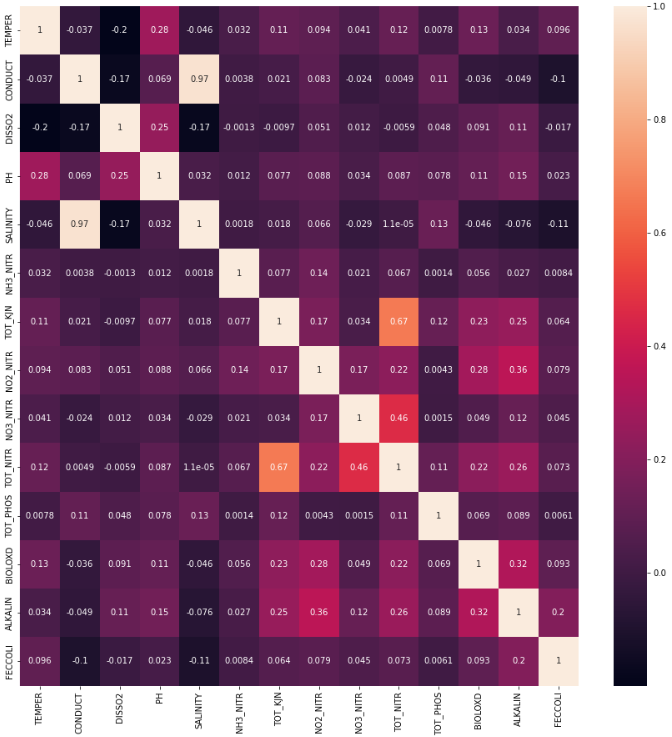
# **Data Pre-processing**

Before delving into defining and building the models for predicting the water quality, it is important to initially understand the data and perform necessary data clean up and preparation.

The selected dataset contains 18148 observations(records/entities) with 48 variables (attributes). Considering the more attributes, a dataset has lessened the chances that the predictive model would be stable. The initial filtering of attributes was performed by considering the standards defined by the World Health Organization (WHO). The following table shows the list of attributes selected which have a higher degree of importance in measuring water quality, and their range of values indicating a good quality of water.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attributes** | | **Description** | **Threshold Values** | | |
| Water Temperature | | Temp of water collected at certain point of time | | | 30-35 Celsius | |
| Specific Conductance | Collective concentration of ions in water | | | <=50 | | |
| Dissolve Oxygen | Amount of oxygen present in water | | | <6 | | |
| Ph | Measure of acididty/basicity in water | | | 6.5-8.0 | | |
| NH3 | Amount of combined nitrogen and hydrogen present | | | <0.2 | | |
| NO2 | Amount of Nitrites present in water | | | <1 | | |
| NO3 | Amount of Nitrates present in water | | | <10 | | |
| Salinity | Salt content dissolved in water | | | <7 | | |
| Total Kjeldahl Nitrogen | Total sum of ammonia nitrogen and organic nitrogenous compounds | | | <1 | | |
| Total Nitrogen | Amount of Nitrites and Nitrates contaminant present in water | | | <10 | | |
| Total Phosphates | Total amount of phosphorus present in water | | | <10 | | |
| Bioloxd | Dissolved gas molecules held in water | | | <5 | | |
| Alkalinity | Amount of neutralizing power of acids and bases | | | <200 | | |
| Fecal Coliforms | Amount of harmful pathogens present in water | | | <200 | | |

Selecting Features on basis of domain knowledge: There are different chemical components present in the water with multiple combinations that affect the quality of water. So, by intruding further with the domain knowledge specific to chemicals we have taken 12 variables refined out of 48 as those features are considered important as per the bio chemist’s knowledge and reference articles [1] related to water quality and its contamination. Also, the lab recorded water samples by force fitting the chemicals are eliminated for model training as it might create a bias while training the models.



*Figure 2: Attribute correlation heatmap*

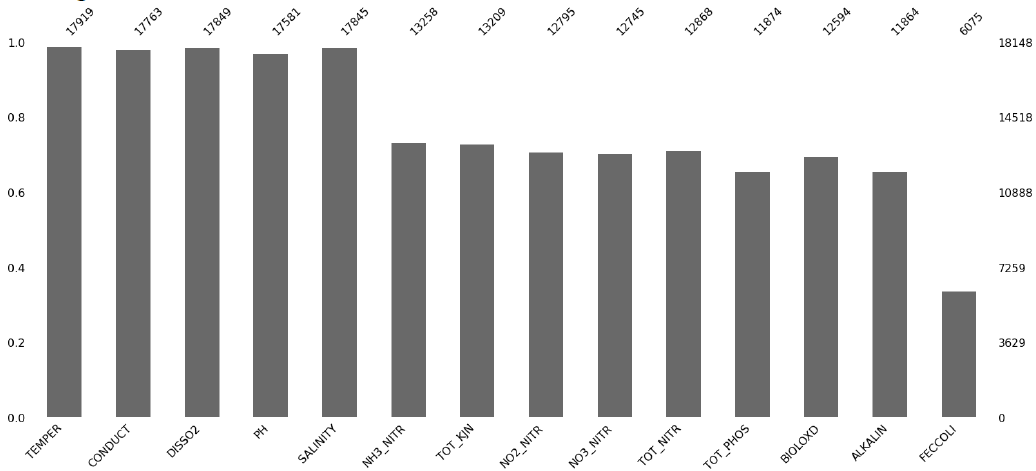
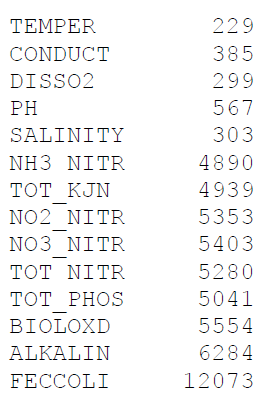
### Defining a Target Variable

For defining the target variable, we used the water quality standards defined by the WHO and created a new variable named water quality which has two attributes “good” and “bad”. The new variable was computed using:

* Conduct < 50.0
* Dissolved Oxygen < 6
* 6.5 <= PH <= 8.0
* Salinity < 7
* NH3 < 0.2
* Total Kjeldahl Nitrogen < 1.0
* NO2 < 1
* NO3 < 10
* Total Nitrogen < 10
* Total Phosphate <10
* Bioloxd < 5
* Alakinity < 200
* Fecal coliform < 200

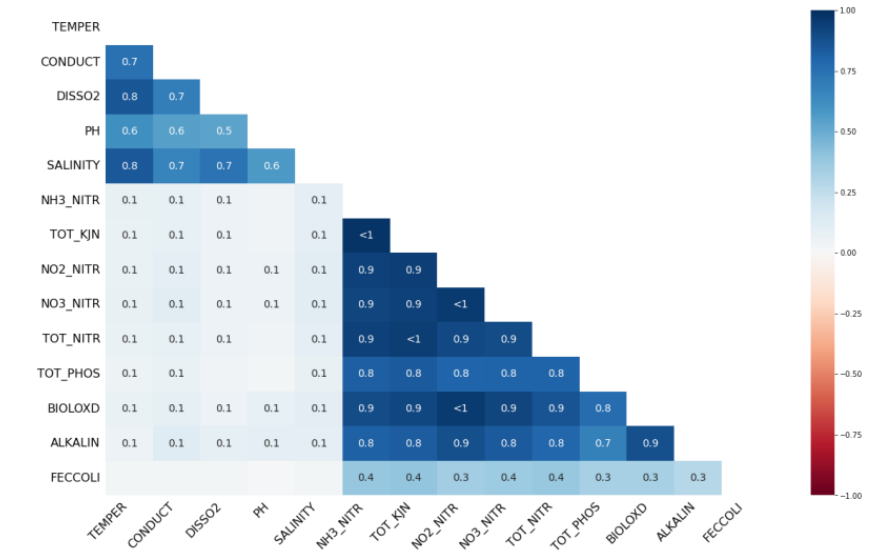
## Handling Missing Values

As presented earlier in this report, in the selected dataset are detected missing variables in some of the attributes. Following, an information about the number of missing values for each attribute, taking in consideration that the overall dataset consists of 18, 148 observations.



*Figure 3: Missing data for each attribute*

The following bar chart shows the completeness of each attribute. Here it can be identified that the attribute with the most missing data is FECCOLI, where more than 50% of values are missing.

Next, we try to identify any correlation of missing values between variables. The following heat map serves this purpose. As it is shown, there are several pairs of attributes that seem to be correlated regarding the missing values, such as TOT\_KJN and NH3\_NITR, or NO2\_NITR and NH3\_NITR. We can say that the missing values are not completely at random, however we cannot identify a specific pattern. Therefore, we are assuming that we are dealing with MAR data.

*Figure 4: Missing data heatmap*

### Denoising Autoencoder

The Denoising Autoencoder consists of multiple layers starting from encoder given with input dimensions and activation layer – ‘rectified linear unit’ along with input data. Decoder with the same number of input dimensions and activation layer as linear along with encoded input. The dataset is spitted in 80% training and 20% testing. The resulting loss is: 6.7029e-07 and the dataset has no missing data. These resulted in the 17164 observations that recorded poor water and 984 as good water.

### KNN Impute

Next, we have implemented KNN Impute by defining the k neighbours as 4 and train the model. The K-Nearest Neighbours algorithm is used by the KNN Impute method to fill in missing numbers in a dataset. KNN Impute assumes that the missing values are missing at random (MAR), and the estimated values assume that the missing values are like the observed values. This rule set reported in 16708 observations as poor water and 1440 as good.

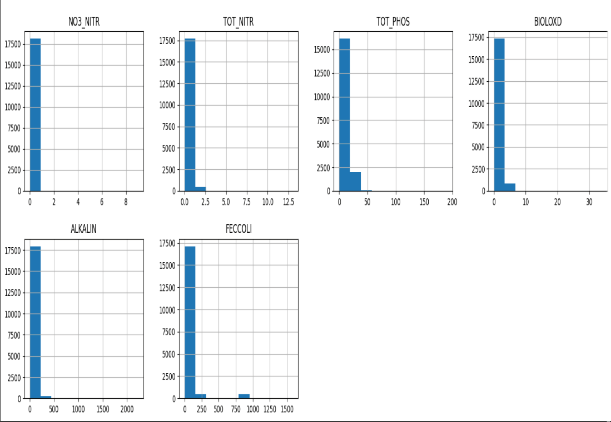
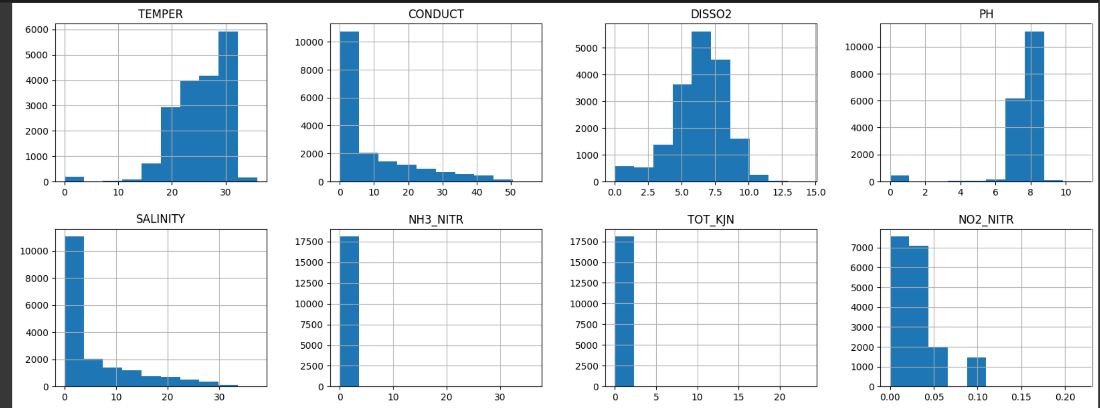
### Soft Impute

Soft Impute is a method for filling in missing data in matrices. Soft Impute can deal with missing data in many ways, such as data that is missing totally at random (MCAR), missing at random (MAR), or missing not at random (MNAR). Soft Impute can also handle datasets with a lot of missing data and can be used to fill in missing values for both continuous and categorical data. The execution of this method gave a distribution of 15618 observations as poor water and 2350 as good water.

For our case, **Denoising Auto Encoder** worked better as it gave a RMSE (Root Mean Squared Error) value better than other imputation methods.

|  |  |
| --- | --- |
| **Method** | **RMSE** |
| Denoising Auto Encoder | 2.2 |
| Soft Impute | 0.0 |
| KNN Impute | 0.0 |

In the following image is shown the distribution of values for each attribute selected, after the imputation.



*Figure 5: Value distribution for each attribute*

## Scaling Features

Minmax Scalar () function: Defining Train and Test Set: The dataset is split into training and testing dataset with 80% of train data and 20% of test data.

Handling Imbalanced Data

SMOTE

It stands for Synthetic Minority Over-sampling Technique. SMOTE works by creating synthetic examples of the minority class by interpolating between existing minority class samples.

Undersampling

In undersampling, the majority class is reduced by randomly removing some of its instances from the training dataset, so that the number of instances in the majority class is similar to the number of instances in the minority class. We investigated different techniques to deal with imbalanced data, and in our situation, we found that SMOTE was causing overfitting, while **undersampling** produced better results.

**Predicting Models**

We tested seven predictive models and found the artificial neural network to have acceptable parameters. Details and results for each model are provided below.

### Bagging Classifier

This approach used in machine learning to address imbalanced datasets where one class has considerably fewer samples than the other. It is a data augmentation technique that generates synthetic examples of the minority class by producing new samples that fall between existing minority class samples. This technique aims to create a more balanced class distribution by providing additional representative samples for the minority class. Ultimately, using SMOTE can enhance machine learning models' performance when working with imbalanced datasets.

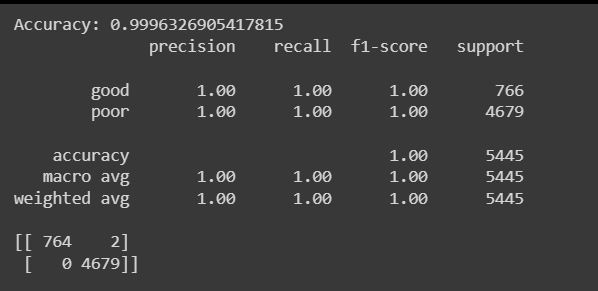
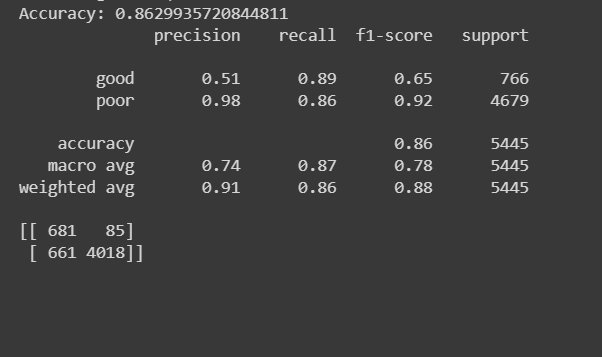


Figure 6: Bagging Classifier Model Output

### Ensemble Method (Bagging Classifier + SVM)

We utilized the f\_classif score to choose the most important features and assigned weights to two classes. Then, we developed an SVM model to tackle overfitting and enhance the model's stability. Additionally, we created a bagging classifier model and conducted under sampling to balance the class distribution. Lastly, we trained the model.

The Accuracy of Ensemble turned out to be 86.2% which seems to be decent prediction as compared to the above SMOTE method and could be improved by tuning of hyperparameters with different range of values.



*Figure 7: Ensemble model output*

However, we will be working on this data with certain models that gives best results for the imbalanced dataset.

## Naïve Bayes

Naive Bayes is a probabilistic algorithm used in machine learning for classification. It is based on Bayes' theorem and the idea that traits don't depend on each other. It's called "naive" because it assumes that each trait is separate from the others. The idea that features are independent can sometimes be taken too far, which makes the program less accurate than others.

## KNN

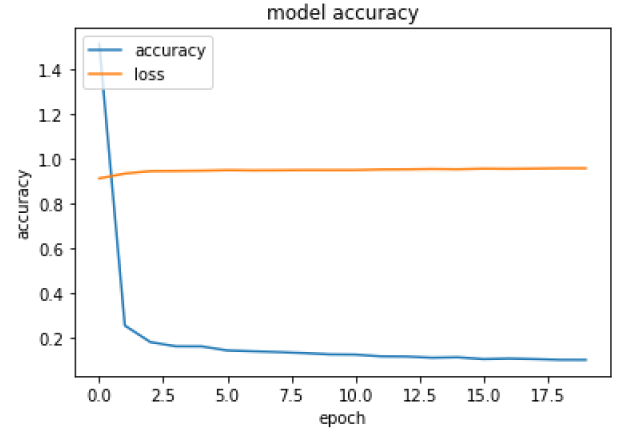
KNN, which stands for "K-Nearest Neighbours," is a method for machine learning that can be used for both classification and regression. It is a non-parametric method, which means that it doesn't make any assumptions about how the data are spread out.

## Random Forest

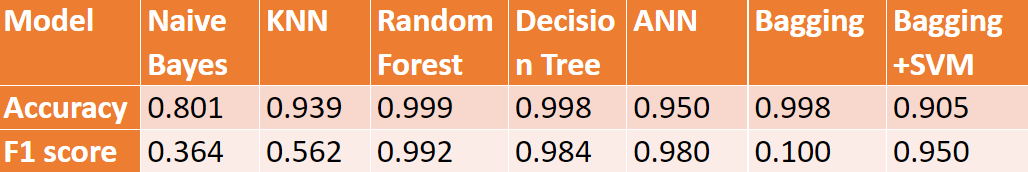
Random forest is a method for machine learning that is used for both classification and regression tasks when it is supervised. It is a type of learning that uses various decision trees to improve the model's ability to predict. The algorithm works by making a forest of decision trees. We have got an accuracy of 99.9% which is an overfit model in predicting the water quality.

## Artificial Neural Network

## The ANN model architecture, suggested by Rustam et al. (2022), uses an initial activation layer followed by a dropout layer and another activation layer with a dropout. The model employs a ReLU activation layer after the dense layer of 256 neurons and a 0.5 dropout rate to randomly eliminate 50% of neurons. The second dense layer has 256 neurons, followed by a ReLU activation layer and a 0.5 dropout layer. The model predicts water quality using a dense two-neuron layer, Adam optimizer, and binary crossentropy loss function, achieving 95% accuracy after being trained for 20 epochs.

.

**Results**



**Conclusion**

There are several approaches to studying water quality. There are studies which analyze pollution levels and focus only on a specific purpose of using water (for example only drinkable water or agriculture purpose). Other studies analyze water parameters and give recommendations on what purpose the water is suitable for. In this project proposal we have come up with a generalized idea of the topic and we gained domain knowledge that we believe will help us in further narrowing down our project area. Also, we can perceive through ANN model that achieved a 95% accuracy which accurately predicted the false negatives (i.e) correctly identifying the poor water as poor which is very important in our case study for chemical based sensitive data in predicting water quality.

# **References**

[1]. Ubah, J.I., Orakwe, L.C., Ogbu, K.N. *et al.* Forecasting water quality parameters using artificial neural network for irrigation purposes. *Sci Rep* **11**, 24438 (2021). <https://doi.org/10.1038/s41598-021-04062-5>

[2]. Amir Hamzeh Haghiabi; Ali Heidar Nasrolahi; Abbas Parsaie. Water quality prediction using machine learning methods. Water Quality Research Journal (2018) 53 (1): 3–13. <https://doi.org/10.2166/wqrj.2018.025>

[3]. Mengyuan Zhu, Jiawei Wang, Xiao Yang, Yu Zhang, Linyu Zhang, Hongqiang Ren, Bing Wu, Lin Ye,A review of the application of machine learning in water quality evaluation, Eco-Environment & Health,Volume 1, Issue 2,2022, Pages 107-116, ISSN 2772-9850, <https://doi.org/10.1016/j.eehl.2022.06.001>.

[4]. Joslyn, Kathleen, "Water Quality Factor Prediction Using Supervised Machine Learning" (2018). REU Final Reports. 6, <https://pdxscholar.library.pdx.edu/reu_reports/6>

1. https:/www.evergladesfoundation.org/post/8-things-to-know-about-south-florida-s-water?gclid=CjwKCAjw5pShBhB\_EiwAvmnNV4fIMq2-NIVeuiAIB0WYEqKdLNKotoF5ePLr4kYsFfVNRe3-dje0dBoCSpAQAvD\_BwE [↑](#footnote-ref-1)