Syracuse University
School of Information Studies

Predicting Water Quality using Machine Learning

Instructor: Dr.Kelvin K King

Maruthamuthu Kanagarathinam Sowmeya, Shriya Rajendra Gawade, Nikhil Sriram Budamaguntala, Besjana Muraku

OBJECTIVE

Water, as one of the most essential resource on Earth for the survival of existing organisms, is currently one of the most contaminated elements.

Having a strong impact for life and ecosystem, it's quality presents a crucial role for humans.

Water quality directly impacts the health of human beings as it can be used for cooking, drinking, agriculture, etc.

Finding a machine learning model that will help on predicting the water quality.

To validate the existing and provide new models on water quality prediction using Cape Coral, FL reported data.

LITERATURE REVIEW

The term "water quality" refers to the state or condition of water, which takes into account the physical, chemical, and biological properties of the water.(Ahmed, et al. 2019).

From results of the review, it can be concluded that the ANN models are capable of dealing with different modeling problems in rivers, lakes, reservoirs, wastewater treatment plants (WWTPs), groundwater, ponds, and streams.(Chen, Y, et al, 2020).

An increasing number of studies on water quality have turned into data-driven analysis, it has become urgent to solve the missing data problem in this domain.

DATA DESCRIPTION

Attributes	Description	Threshold Values
Water Temperature	Temp of water collected at certain point of time	30-35 deg celcius
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

MODELS

Imputation: denoising autoencoder, soft impute, KNN impute, MICE

Balancing: SMOTE (Synthetic Minority Oversampling Technique), Undersampling

Prediction: SVM with Bagging, SMOTE-Bagging, Artificial Neural Network

KEY FINDINGS

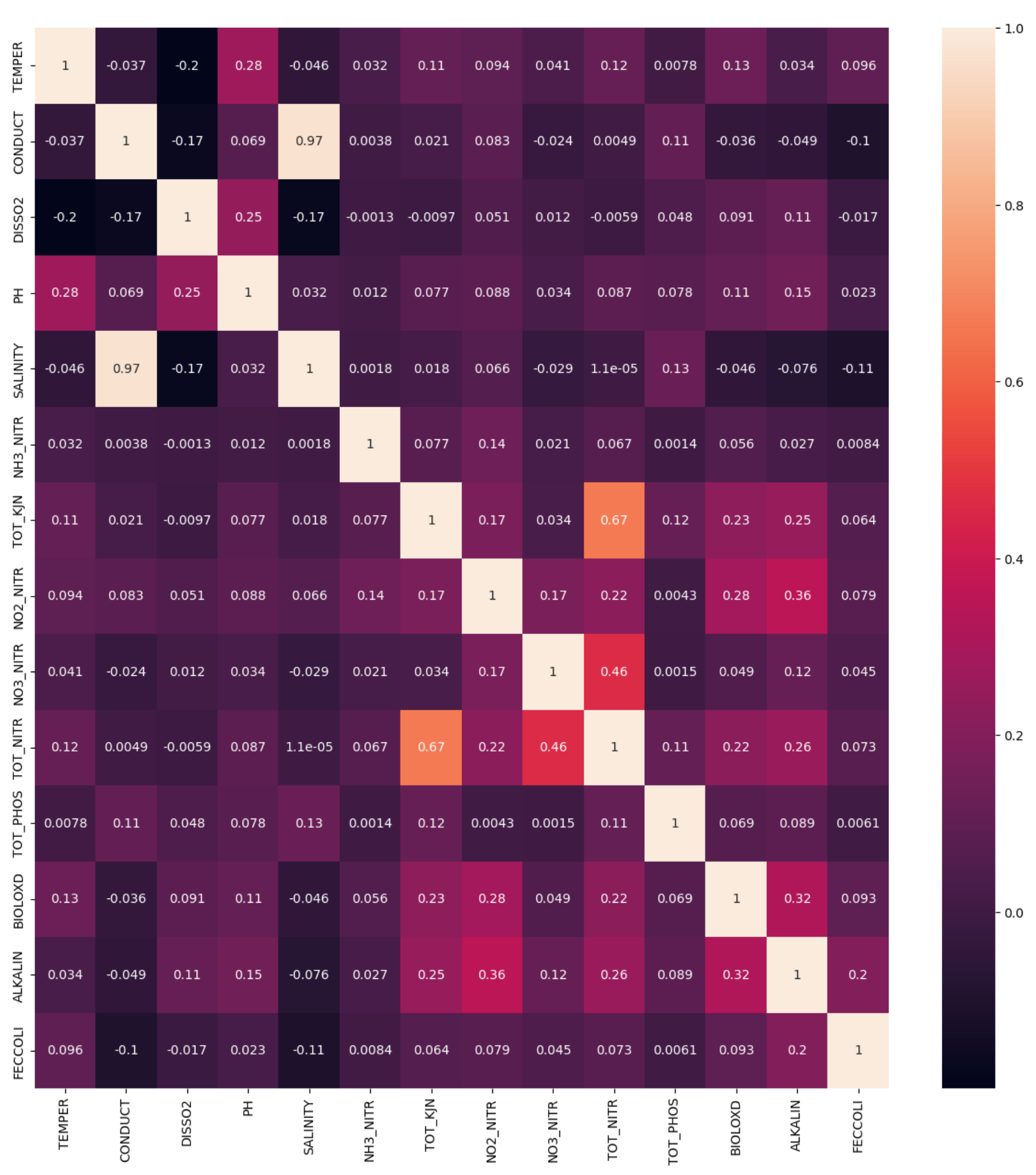
The artificial neural network (ANN) model outperformed other models, achieving a 96% accuracy for instances classified as class "poor" and the precision for predicting "good" quality is 86%.

The missing values can be predicted using 'Denoising Autoencoder' with a Root Mean Squared Error lower than other model.

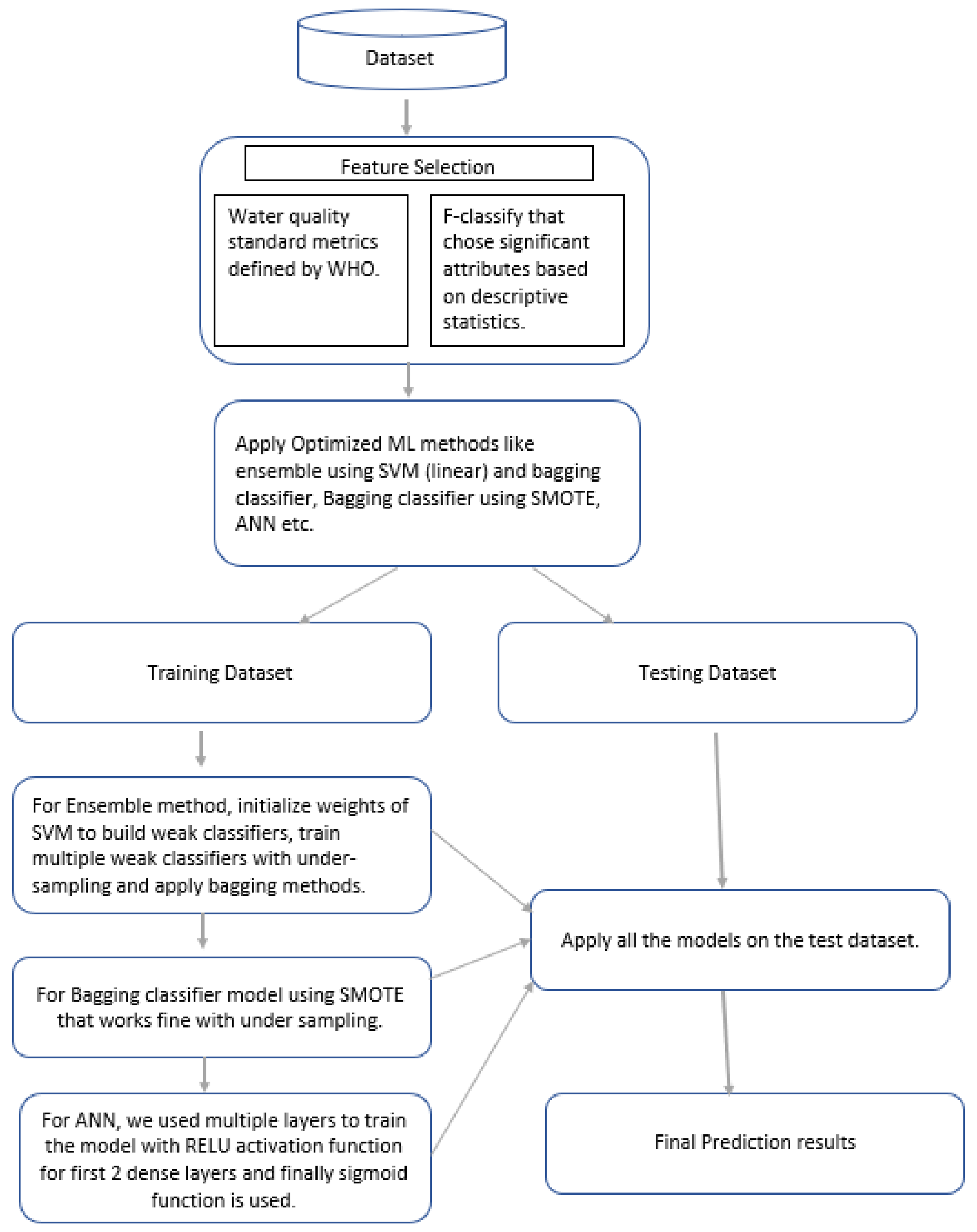
Undersampling model proved to perform better when dealing with imbalanced data.

RESULTS

Attributes correlation HeatMap



Project workflow



Dataset

Feature Selection

Water quality standard metrics defined by WHO.

F-classify that chose significant attributes based on descriptive statistics.

Apply Optimized ML methods like ensemble using SVM (linear) and bagging classifier, Bagging classifier using SMOTE, ANN etc.

Training Dataset

Testing Dataset

For Ensemble method, initialize weights of SVM to build weak classifiers, train multiple weak classifiers with under-sampling and apply bagging methods.

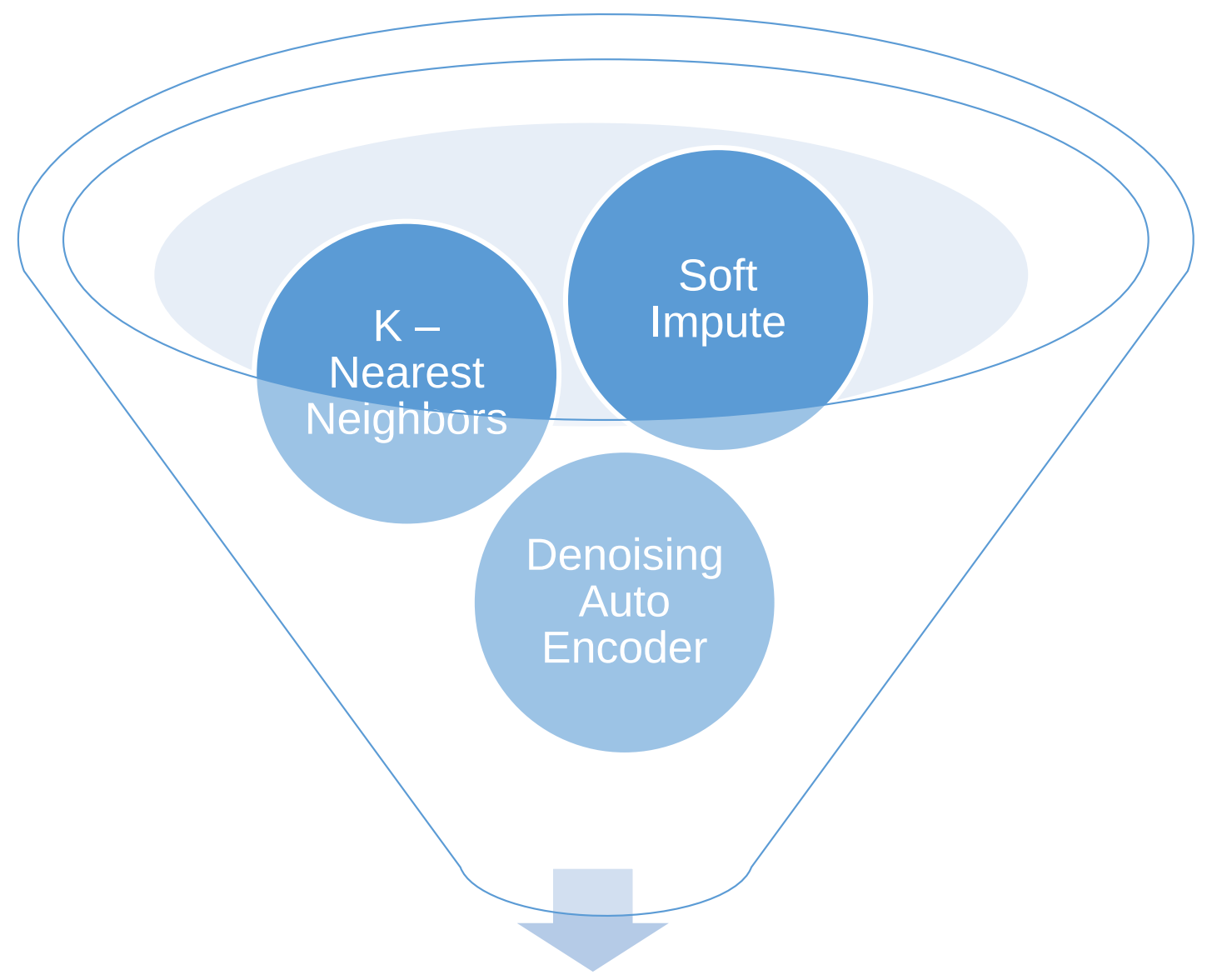
For Bagging classifier model using SMOTE that works fine with under sampling.

For ANN, we used multiple layers to train the model with RELU activation function for first 2 dense layers and finally sigmoid function is used.

Apply all the models on the test dataset.

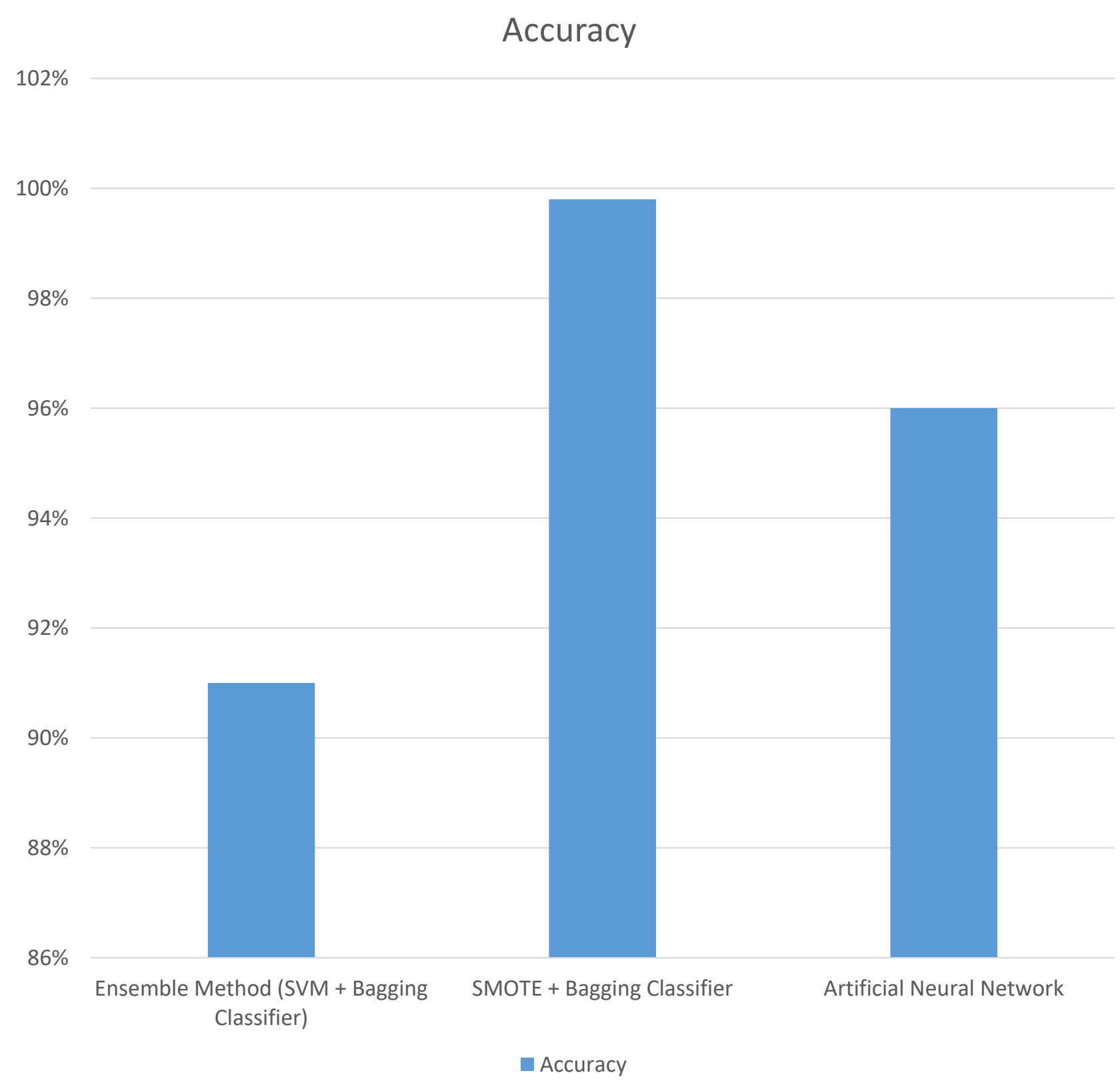
Final Prediction results

Treating Missing Values



Denoising Auto Encoder

Overall Model Performance



Dataset

April 17, 2023

Last Update

June 5,2019

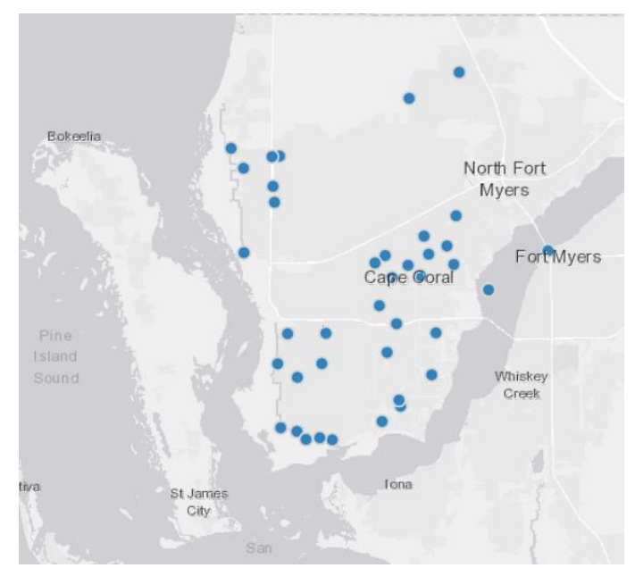
Published Date

36,464 Records

44 Attributes

Public

Anyone can see this content



CONCLUSION

Our study on water quality assessment faced challenges with missing values, which we addressed using denoising autoencoder with the lowest RMSE compared to other methods.

Additionally, our dataset had imbalanced class distribution, for which undersampling was more effective than SMOTE.

Among the three models we tested, ANN demonstrated better performance with higher F1 score and accuracy.

Our findings highlights

- the importance of addressing missing values and
- imbalanced class distribution in water quality assessment studies, and

Denoising autoencoder and undersampling techniques, along with ANN as a predictive model, could be effective approaches for improving water quality prediction models in future research.

LIMITATIONS

This study was conducted in a limited timeframe.

Current sample size can not be a true representation of the whole population.

The current study has been conducted in self-rated manner.

The framework for this study is explored only in the context of Cape Coral, Florida area.

REFERENCES

Ali Najah Ahmed, Faridah Binti Othman, Haitham Abdulmohsin Afan, Rusul Khaleel Ibrahim, Chow Ming Fai, Md Shabbir Hossain, Mohammad Ehteram, Ahmed Elshafie (2019) "Machine learning methods for better water quality prediction", *Journal of Hydrology*, Volume 578, 0022-1694

Chen, Y., Song, L., Liu, Y., Yang, L., & Li, D. (2020). A Review of the Artificial Neural Network Models for Water Quality Prediction. Applied Sciences, 10(17), 5776. MDPI AG. Retrieved from http://dx.doi.org/10.3390/app10175776

Zeng Chen, Huan Xu, Peng Jiang, Shanen Yu, Guang Lin, Igor Bychkov, Alexey Hmelnov, Gennady Ruzhnikov, Ning Zhu, Zhen Liu, (2021), "A transfer Learning-Based LSTM strategy for imputing Large-Scale consecutive missing data and its application in a water quality prediction system", *Journal of Hydrology,* 0022-1694,

sgawade@syr.edu, nbudamag@syr.edu, mksowmey@syr.edu, bmuraku@syr.edu