National College of Ireland

Enhancing Water Quality Prediction Through Explainable Machine Learning

Configuration Manual

Data Analytics

Pramod Akuthota

Student ID: 23408979

School of Computing

National College of Ireland

Supervisor: Arghir Nicolae Moldovan

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Pramod Akuthota
Student ID:	23408979
Programme:	MSc in Data Analytics **Year:** 2025-2026
Module:	Research Practicum Part - 2
Lecturer:	Arghir Nicolae Moldovan
Submission Due Date:	11-12-2025
Project Title:	Enhancing Water Quality Prediction Through Explainable Machine Learning
Word Count:	663 **Page Count:** 5

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Pramod Akuthota

Date: 11-12-2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Enhancing Water Quality Prediction Through Explainable Machine Learning

Pramod Akuthota
23408979

1. Introduction

The objective of this research project is to utilize physicochemical characteristics to forecast the Water Quality of total CCME Water Quality Index predictions (range from 0 to 100) and generate CCME_WQI classification categories (from Excellent to Poor). Due to the inherent challenges associated with water quality data collection, model accuracy and interpretability can be adversely affected by the high level of noise from the data as well as the nonlinear relationships among water quality parameters meaning that water quality data may display different levels of variability, correlation, and plausibility across different geographic regions during different periods of time. The overall process of the project includes dataset preparation through cleaning and formatting, training various classical machine-learning, deep-learning models, evaluating and comparing model performance based on established model evaluation metrics, and producing interpretable predictions.

2. System Specifications

- **Hardware requirements**
  - **CPU:** Minimum 4 cores; recommended 8 cores for faster preprocessing and model training.
  - **GPU:** Optional; NVIDIA GPU with ≥8 GB VRAM recommended for deep-learning experiments.
  - **RAM:** Minimum 8 GB; recommended 16 GB for smooth handling of large datasets.
  - **Storage:** Minimum 5 GB free disk space; recommended 10 GB+ SSD for datasets, checkpoints, and experiment artifacts.
- **Software requirements**
  - **Operating system:** Windows 10/11 or Ubuntu 20.04/22.04
  - **Python:** Python 3.9–3.11 (recommended 3.10).
  - **Environment:** Conda/Miniconda or venv; Jupyter Notebook optional for experimentation.
  - **CUDA/cuDNN (if GPU is used):** CUDA 11.x and cuDNN compatible with the installed TensorFlow version.

3. Required Libraries

Utilizing a typical scientific Python stack, reproducibility and compatibility are ensured in this project through the use of NumPy and Pandas for numerical computations, table format processing, and dataset checks. When needed, statistical-related functions are included through the Scipy library. Visual representation of comparative performance figures are produced with Matplotlib. Preprocessing, train-test splits, feature selection, evaluation metrics, and the foundations of traditional machine learning models are included in this project by Scikit-learn, such as Decision Tree and Random Forest. SMOTE-based oversampling, used to mitigate class imbalance. Deep learning algorithms, namely ANN and LSTM, were implemented using Tensorflow/Keras. LIME was applied to produce as much explanation as possible as to why the model made specific predictions, thus producing more interpretable results as well as enabling an analysis that is transparent.

The necessary libraries can be installed using the following commands:

a) **pip install numpy==1.26.4**
b) **pip install pandas==2.2.2**
c) **pip install matplotlib==3.8.4**
d) **pip install scikit-learn==1.4.2**
e) **pip install imbalanced-learn==0.12.3**
f) **pip install tensorflow==2.15.1**
g) **pip install lime==0.2.0.1**
h) **pip install imbalanced-learn xgboost tensorflow**

4. Dataset description

A total of 2,827,977 records of water quality were examined in five countries including Canada, China, England, Ireland, and the United States of America (USA) and contain information spanning various water body types (e.g., rivers, lakes, canals, coastal/estuarine) covering the period of 1940-01-02 through 2023-12-10. The measured parameters of Ammonia, BOD, Dissolved Oxygen, Orthophosphate, pH, Temperature, Nitrogen, and Nitrate comprise an index on the chemical and physical properties of water. The dataset also provides a CCME Water Quality Index, a numeric value (CCME_Values) rating between 0 and 100, as well as a numerical value designation for overall quality (CCME_WQI) which identifies five water quality categories: excellent, good, fair, marginal, and poor.

5. Models Implemented

The principal and central predictive model used in this study is Random Forest due to its high predictive ability and robustness of generalizing from structured water-quality data. Decision Trees were used as simple baselines to compare with and evaluate Random Forest performance. The primary objective of utilizing deep-learning models including ANN and LSTM, was to assess their usefulness for benchmarking, compared with the tree-based Random Forest. To improve interpretability of the Random Forest models, LIME will be applied, as a means of providing an explanation for each individual prediction by

highlighting which of the various physicochemical attributes were most important for the generation of that specific prediction.

References:

1. https://scikit-learn.org/stable/
2. https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.SMOTE.html
3. https://lime-ml.readthedocs.io/en/latest/
4. https://numpy.org/doc/
5. https://pandas.pydata.org/docs/