Freshwater Health Assessment Application

Process Documentation

Project source code: https://github.com/Techietash/RRC_DSML

# FRESHWATER HEALTH ASSESSMENT APPLICATION

PROCESS DOCUMENTATION

**Understand the Problem Statement**

- We obtained the dataset from IISD-ELA, which was supplied by the organizer of the Hackathon for challenge 1. Using our interpretation of the problem statement, we crafted a proposal outlining our intended solution.

**Understanding the data set**

- Combine the relevant datasets, such as the chemical datasets.
- Outline and establish comprehensive guidelines for determining the parameter ranges for both healthy drinking water and optimal water conditions for aquatic life.
- Specify and choose the architecture and components for the development of the preferred mobile application.

**Data cleaning and preprocessing**

- Transferred CSV data into an SQL database.
- Employed SQL and Excel to identify and eliminate duplicate records.
- Utilized a pivot on the chemical dataset, organizing characteristics by date for simplified analysis.
- Substituted null values with the median.

**Data exploration analysis**

- Utilized PowerBI to unveil connections within the dataset and generate various graphs, focusing on determining the most informative visuals for the application.
- Using our findings and additional insights, we successfully replicated the graph onto the application dashboard.

**Data transformation/encoding**

- Prepared the data for analysis by transforming variables, normalizing, scaling, and encoding categorical variables.

**Choosing analysis techniques**

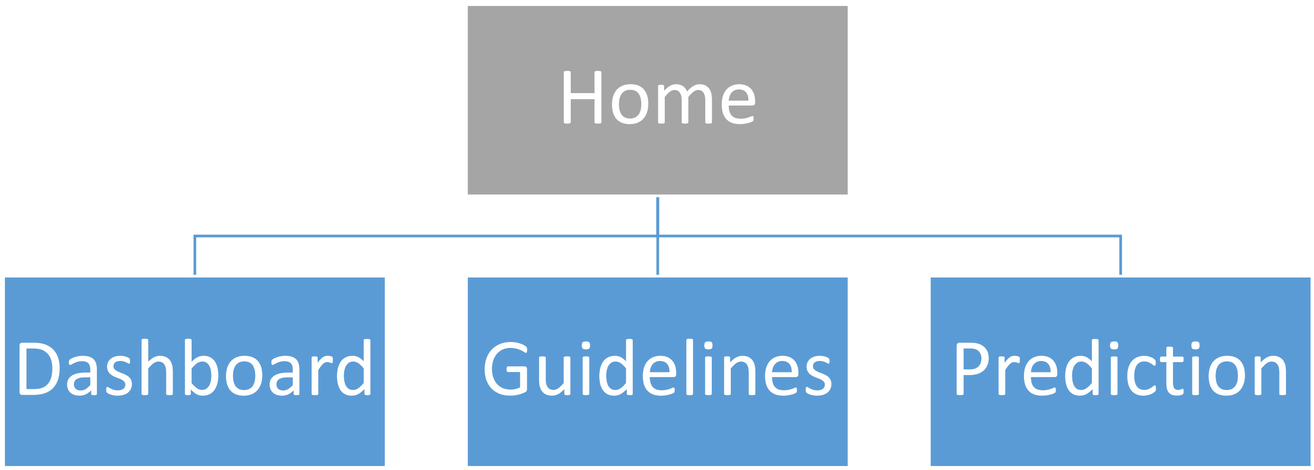
- After assessing the accessible data, we observed that supervised learning techniques might not yield the anticipated results. Consequently, we opted for an unsupervised machine learning model. Subsequently, we experimented with several models like KMeans, Hierarchical Clustering, and DBSCAN. Our analysis led us to favor KMeans as it offered more comprehensive insights into the dataset and delivered more accurate predictions.

**Analysis techniques application**

- We ultimately decided to use K-means for clustering analysis and selected 5 clusters based on an elbow curve analysis. Following the clustering, we developed a supervised learning model to forecast cluster assignments for a specific variable. After testing several classification algorithms and evaluating their performance using different metrics, we concluded that the DecisionTreeClassifier was the most effective choice among the options we explored

**Result interpretation**

- Based on our findings, we interpreted the distinct clusters according to established criteria to assess the water quality as either:
- very healthy
- healthy
- somewhat healthy
- not quite
- something is not right!
- Users are then refered to guideline for a detailed explanation of each of these result.

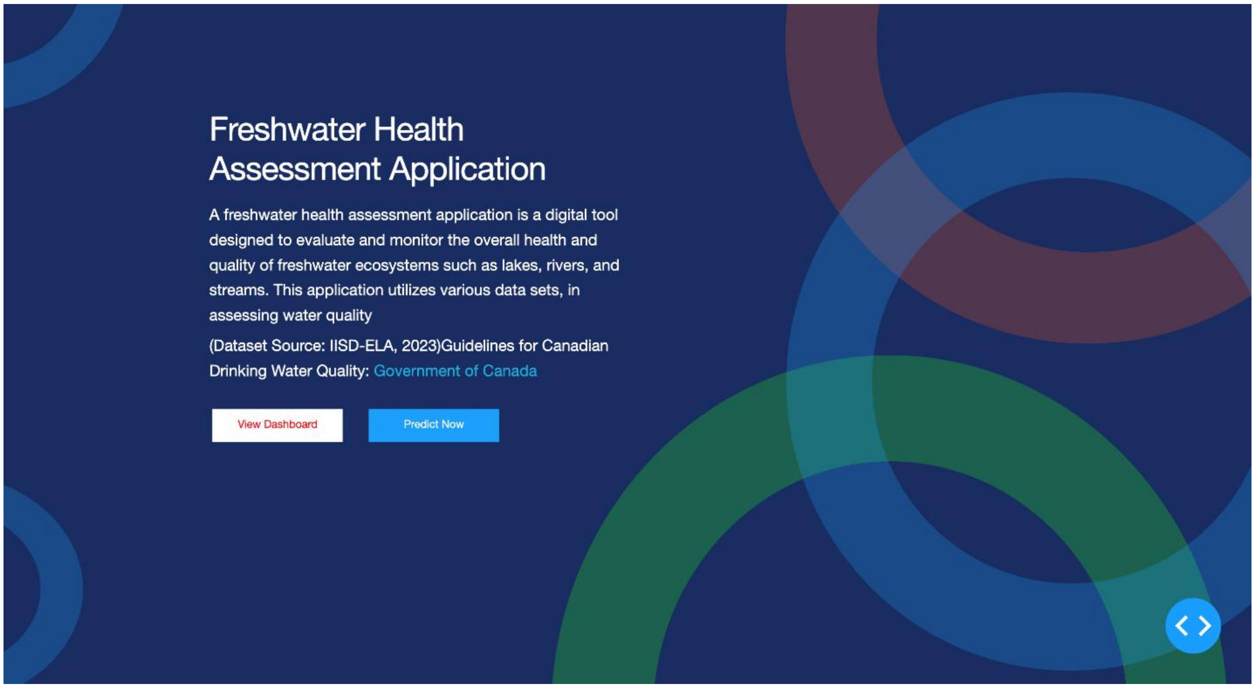



Module Description

Dashboard: Exploratory data analysis, view different chemical parameters results via interactive charts

Guidelines: The description of chemical properties as well as their corresponding guidelines for drinking water and aquatic live

Prediction: predict the state and health of drinking water using unsupervised and supervised machine learning based on the chemical parameter, hydrology, and luminology result values




The initial page acts as an introduction, providing a comprehensive overview of the application's goals and abilities. It gives an explanation that summarizes the fundamental purpose, objectives, features, and the sources of data within the application.


The dashboard serves as the primary interface displaying a variety of charts illustrating the interrelationships among distinct chemical parameters. It offers an overview of seasonal fluctuations and demonstrates how these variations correspond to different water types.


| SN | Water Type | PARAMETERS | Meaning | Definition | Range for drinking | Health Consideration |
|---|---|---|---|---|---|---|
| 1 | Drinking water | ALK | Alkalinity | water body's ability to resist pH changes when acidic substances are introduced. | 20 to 200 | |
| 2 | Drinking water | CL | chloride | Chloride; often expressed as Cl-; is a common anion in water chemistry that comes from dissolved sodium chloride (table salt) and other sources. It's a key parameter measured to assess water quality; especially its salinity; taste; and potential corrosive effects in drinking water and environmental systems. | <= 250 | A guideline value is not necessary as health effects are not of concern at levels found in drinking water. |
| 3 | Drinking water | SO4 | sulphate | In water chemistry; sulfate refers to the sulfate ion (SO4^2-); a chemical species that can be naturally present or introduced through human activities. Monitoring sulfate levels is important to assess water quality and potential sources of contamination; as high concentrations can impact water's taste; odor; and potential for corrosion. | <= 500 | High levels (above 500 mg/L) can cause physiological effects such as diarrhoea or dehydration |
| 4 | Drinking water | CA | Calcium | Calcium (Ca) is measured to assess water hardness and its potential impact on scaling and corrosion in plumbing and industrial processes | None | guideline value not necessary; calcium contributes to hardness. |
| 5 | Drinking water | MG | Magnesium | It plays a vital role in water chemistry and is a major component of water hardness. Magnesium levels in water can impact taste; the formation of scale in pipes; and the suitability of water for various industrial and domestic uses. Monitoring magnesium concentration is important for managing water quality and optimizing water treatment processes | None | No evidence of adverse health effects from magnesium in drinking water; therefore a guideline value is not necessary. |
| 6 | Drinking water | | Sodium | Sodium is a common element found in natural water sources; and its presence can come from various sources; including natural mineral deposits and human activities. Monitoring sodium levels is important; as excessive sodium in drinking water can have health implications; particularly for individuals on low-sodium diets or those with specific health conditions. | None | For persons on strict sodium reduced diets applying to all sources; levels in drinking water should be below 20 mg/L |

The guideline comprehensively comprises six columns explaining specific details:

1. Water Type: This denotes the intended use of the water, distinguishing between its purpose for Drinking or sustaining Aquatic Life.

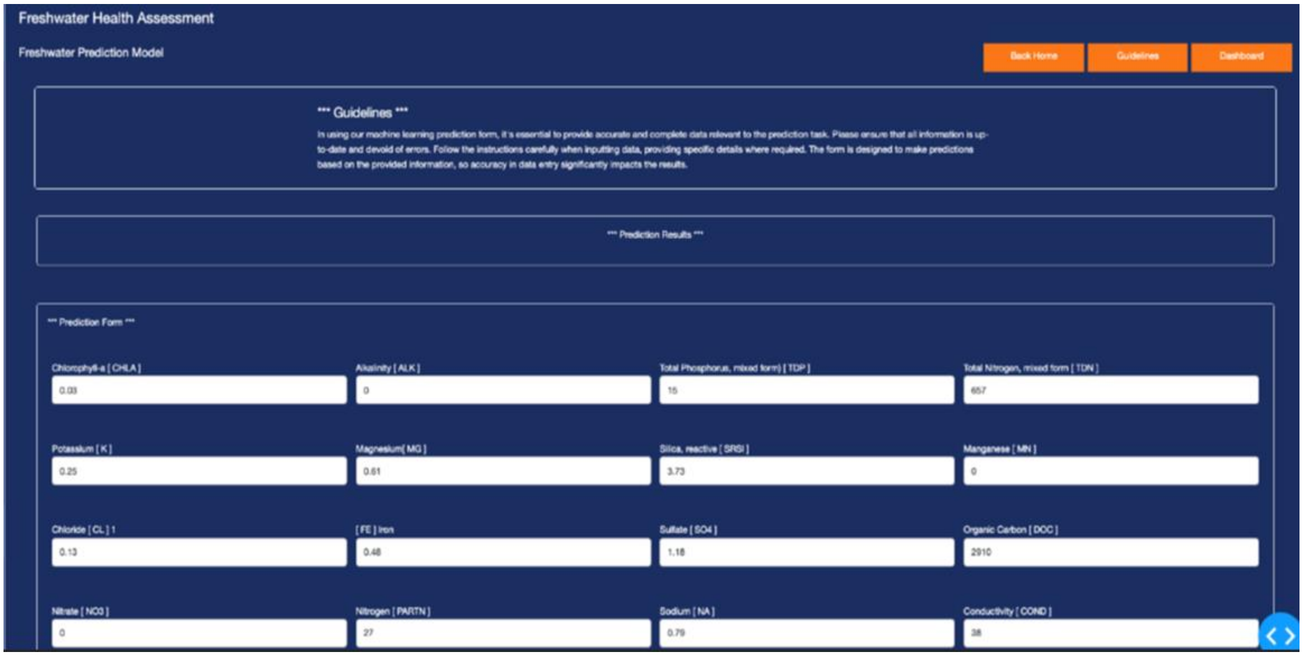
2. Parameters: This column lists the chemical symbols representing various chemical parameters found within the water.

3. Meaning: It explains the significance of each chemical symbol, clarifying the intended chemical parameter.

4. Definition: This section provides detailed explanations and definitions for each parameter, elucidating their relevance to water quality.

5. Range: It outlines the acceptable ranges within which these chemical parameters should ideally exist to maintain healthy water quality.

6. Health Consideration: This portion elaborates on the health implications associated with each chemical parameter if it falls outside the acceptable range, highlighting potential health impacts or concerns.




The Freshwater prediction model is a platform designed for users to input different chemical ranges, predicting the water condition, and providing outcomes categorized as:

- very healthy
- healthy
- somewhat healthy
- not quite
- something is not right!

There's a dedicated guideline page to assist users in comprehending these results.

# Chart Definition and Description

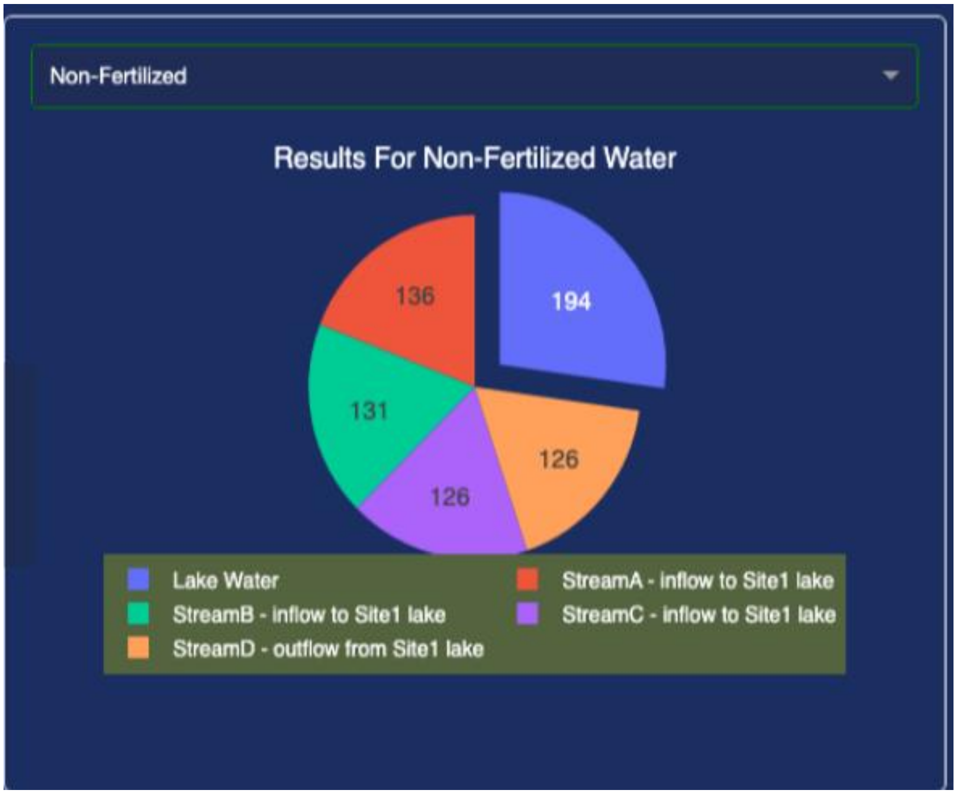
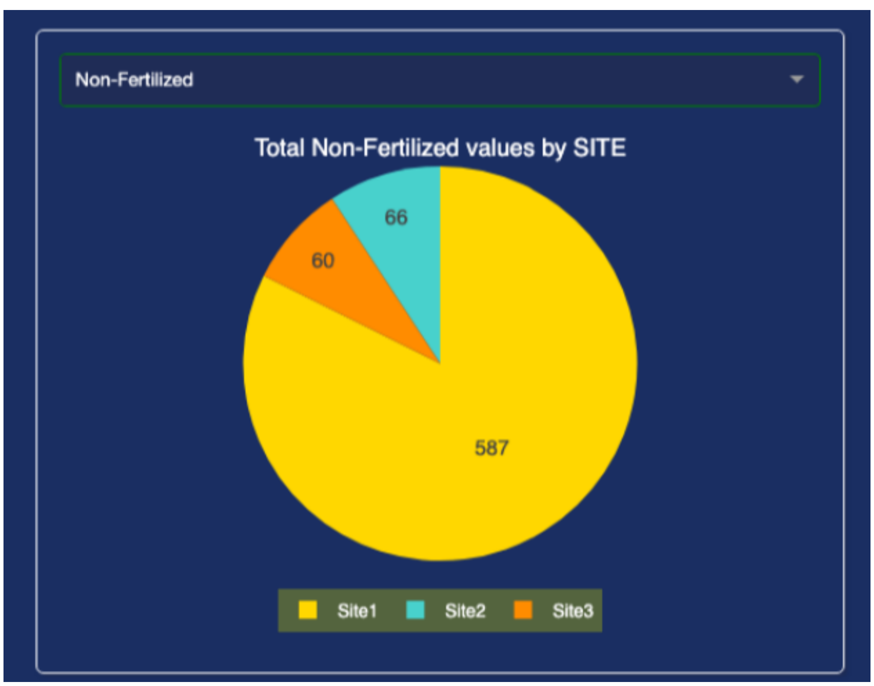



Chart 1: A pie-chart showing the total of each water type based on treatment. It gives a clear visual representation of the proportion and percentage of different water types based on fertilized or non-fertilized.




2. A pie-chart showing the total of each site based on treatment.

It visually represents the proportions of treatments across different sites to easily display the relative sizes of each treatment category in relation to the total, it allows for a quick comparison between different sites and their respective treatment allocations.


Chart 3: A histogram showing year against chemical parameter (Fertilized) to show trend analysis and seasonal patterns.


Chart 4: A histogram showing year against chemical parameter(unfertilized) to give insights to trend analysis and seasonal pattern.


Chart 5: A scatterplot which display the chemical parameters by year and treatment, to give detailed analysis of correlation among three chemical parameters.


Chart 6: A scatterplot showing the relationship between two chemical parameters based on the treatment.


Chart 7: A line chart showing the analysis of chemical parameters over the years with the site, water type and treatment type as available options.

REFERENCES

(*Https://Www.Epa.Gov/Wqc*, n.d.)

(*Https://Www.Canada.ca/En/Health-Canada/Services/Environmental-Workplace-Health/Reports-Publications/Water-Quality/Guidelines-Canadian-Drinking-Water-Quality-Summary-Table.Html*, n.d.)

(*Iisd.Org/Ela*, n.d.)

Appendix: SQL Codes

```
select --abs(id),
dataset_name, site, treatment, activity_media_name,
activity_start_date,
result_sample_fraction,characteristic_name,count(isnull(result_value,0
)) [count], sum(result_value) [sum],
sum(result_value)/count(isnull(result_value,0)) average
into avgAggregate
from ChemicalBackup where result_value is not null
group by
dataset_name, site, treatment, activity_media_name,
activity_start_date,
result_sample_fraction,characteristic_name
having count(isnull(result_value,0)) >1
order by average desc
```

*Code 1.0 discover records that are duplicate using activity start date*

```
update Chemical
set Chemical.result_value = (select average from avgAggregate
where
Chemical.dataset_name = avgAggregate.dataset_name and
Chemical.site = avgAggregate.site and
Chemical.treatment = avgAggregate.treatment and
Chemical.activity_media_name = avgAggregate.activity_media_name and
Chemical.activity_start_date = avgAggregate.activity_start_date and
Chemical.result_sample_fraction = avgAggregate.result_sample_fraction
and
Chemical.characteristic_name = avgAggregate.characteristic_name )
```

*Code 1.1 Update duplicate records with the average result value for each duplicates set*

```
select org.dataset_name, org.site, org.treatment,
org.activity_media_name, org.activity_start_date,
org.result_sample_fraction,org.characteristic_name, org.result_value
```

```
,ou.*
from
Chemical org
inner join
(select dataset_name, site, treatment, activity_media_name,
activity_start_date,
result_sample_fraction,characteristic_name,count(isnull(result_value,0
)) [count], sum(result_value) [sum],
sum(result_value)/count(isnull(result_value,0)) average
from Chemical where result_value is not null
group by
dataset_name, site, treatment, activity_media_name,
activity_start_date,
result_sample_fraction,characteristic_name
having count(isnull(result_value,0)) >1) ou
on org.dataset_name = ou.dataset_name and
org.site = ou.site and
org.treatment = ou.treatment and
org.activity_media_name = ou.activity_media_name and
org.activity_start_date = ou.activity_start_date and
org.result_sample_fraction = ou.result_sample_fraction and
org.characteristic_name = ou.characteristic_name
order by ou.[count] desc
```

*Code 1.2 merging chemical characteristic name result set with secchi (Luminology) and mean daily depth (Hydrology) record*

```
declare @cols as varchar(max)
declare @query as nvarchar(max)

SET @cols = STUFF((SELECT distinct ',' +
QUOTENAME(characteristic_name)
FROM chemical_unique c
FOR XML PATH(''), TYPE
```

```
).value('.', 'NVARCHAR(MAX)')
,1,1,'')
--PRINT @COLS;
set @query = 'SELECT DISTINCT CAST(ACTIVITY_START_DATE AS DATE)
ACTIVITY_START_DATE, DATASET_NAME, SITE, TREATMENT,
ACTIVITY_MEDIA_NAME ,' + @cols + ' from
(
SELECT ACTIVITY_START_DATE, DATASET_NAME, SITE, TREATMENT,
ACTIVITY_MEDIA_NAME , CHARACTERISTIC_NAME, SUM(RESULT_VALUE) AS
RESULT_VALUE
from CHEMICAL_UNIQUE
GROUP BY ACTIVITY_START_DATE, DATASET_NAME, SITE, TREATMENT,
ACTIVITY_MEDIA_NAME , CHARACTERISTIC_NAME
) x
pivot
(
max(RESULT_VALUE)
for CHARACTERISTIC_NAME in (' + @cols + ')
) p'


execute(@query)
```

*Code 1.3 pivoting the unique chemical records ( Pivoted data is in dataset.csv )*

```
SELECT C.*, S.SECCHI_DEPTHS, UN.MEAN_DAILY_DISCHARGE
INTO MASTER_RECORD
FROM CHEMICAL_MERGED C
LEFT OUTER JOIN
(SELECT CAST(ACTIVITY_START_DATE AS DATE) AS ACTIVITY_START_DATE ,
UPPER(DATASET_NAME) DATASET_NAME, SITE, TREATMENT, ACTIVITY_MEDIA_NAME
, RESULT_VALUE AS SECCHI_DEPTHS
FROM Secchi_merged_UNIQUE
) AS S
ON C.ACTIVITY_START_DATE = S.ACTIVITY_START_DATE
```

```sql
AND C.SITE = S.SITE AND
C.TREATMENT= S.TREATMENT
AND C.ACTIVITY_MEDIA_NAME = S.ACTIVITY_MEDIA_NAME
LEFT OUTER JOIN
(SELECT CAST(ACTIVITY_START_DATE AS DATE) AS ACTIVITY_START_DATE,
'MEAN_DAILY_DISCHARGES' AS DATASET_NAME, SITE, '' AS TREATMENT,
STREAM_NAME AS ACTIVITY_MEDIA_NAME, MEAN_DAILY_DISCHARGE
FROM Hydrology --WHERE ACTIVITY_START_DATE = '2015-05-19'
) UN
ON C.ACTIVITY_START_DATE = UN.ACTIVITY_START_DATE
AND C.SITE = UN.SITE
AND C.ACTIVITY_MEDIA_NAME = UN.ACTIVITY_MEDIA_NAME
```

*Code 1.4 join the pivoted chemical data results with secchi (Luminology) and mean daily depth (Hydrology) record [ This produced the final transposed csv that was used for analysis ]*