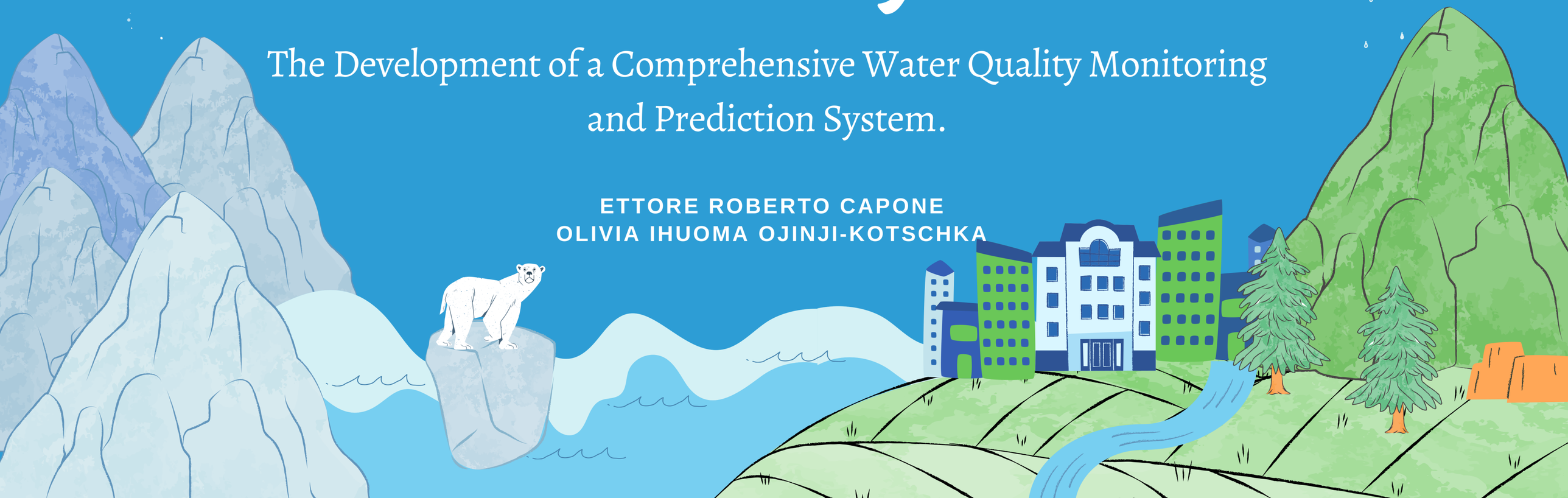
Ironhack Final Project

Water Everywhere

The Development of a Comprehensive Water Quality Monitoring and Prediction System.

ETTORE ROBERTO CAPONE
OLIVIA IHUOMA OJINJI-KOTSCHKA



Problem & Mission Statement

In the water services sector, ensuring access to high-quality drinking water is crucial to public health, environmental protection and community well-being. However, the challenges of monitoring and managing water resources can be complex and costly.

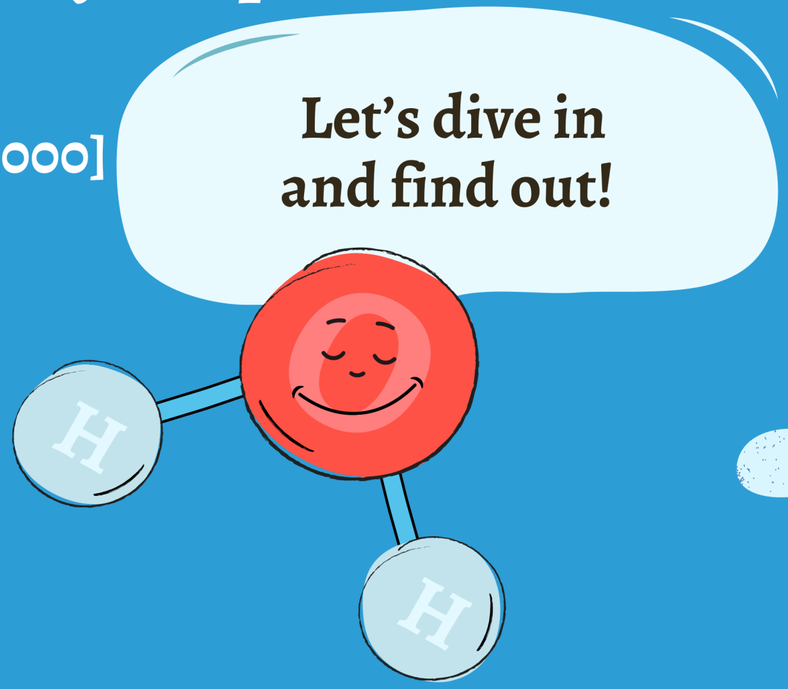
In this business case, we will explore how the application of machine learning models can revolutionize the approach to water potability, offering not only public health benefits but also opportunities for operational improvement and financial savings.

Project Research & Development

- Decision making
- Domain knowledge
- WHO treshold features establishment
- Data research and collection
- Data preprocessing
- Machine learning models application
- Cross validation
- Visualization
- Conclusion
- SWOT analysis
- References
- Models simulation

Our planet is about 71% covered in water. Are they all potable? WHO knows!

Conductivity: [0,400]
Total Dissolved Solids - [0,1000]
Sulfate: [0,1000]
Iron: [0,0.3]
Copper: [0,2]
Sulfate: [0,250]
Chloride: [0,250]
pH: [6.5,8.5]
Nitrate: [0,50]
Lead: [0,0.1]
Fluoride: [0,1.5]
Turbidity: [0,1000]
Chlorine: [0.02,5]
Manganese: [0,0.08]
Zinc: [0,3]



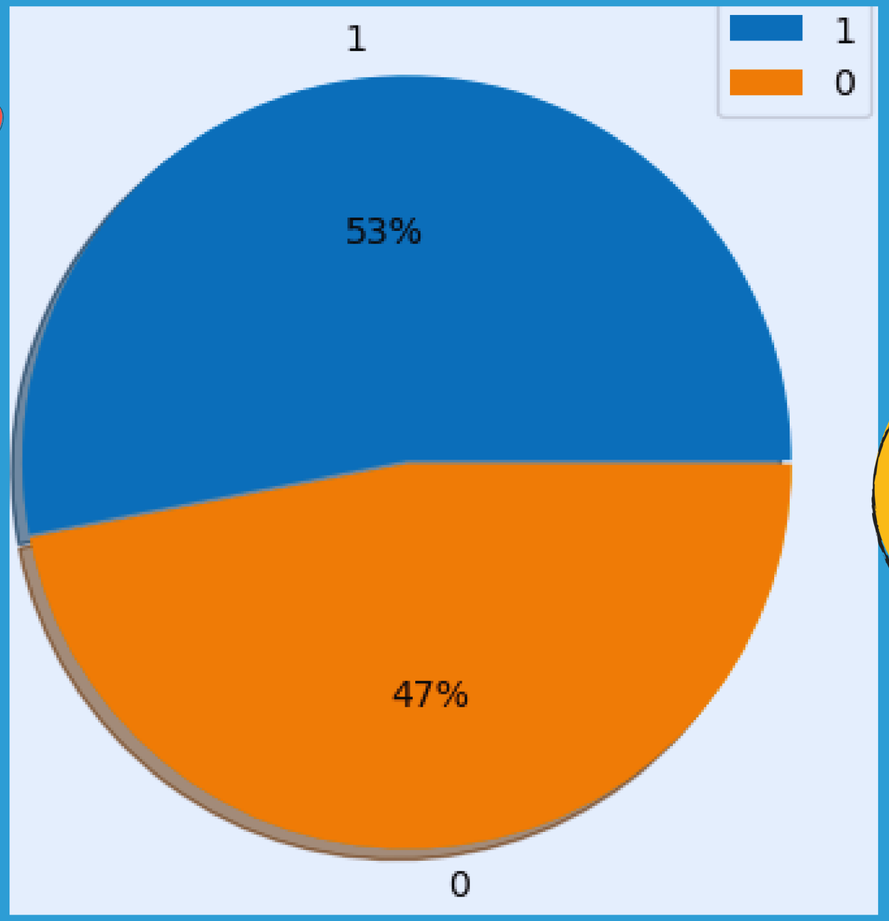
Data Collection & Preprocessing

Collection: Kaggle
WHO (2022 & 2024)

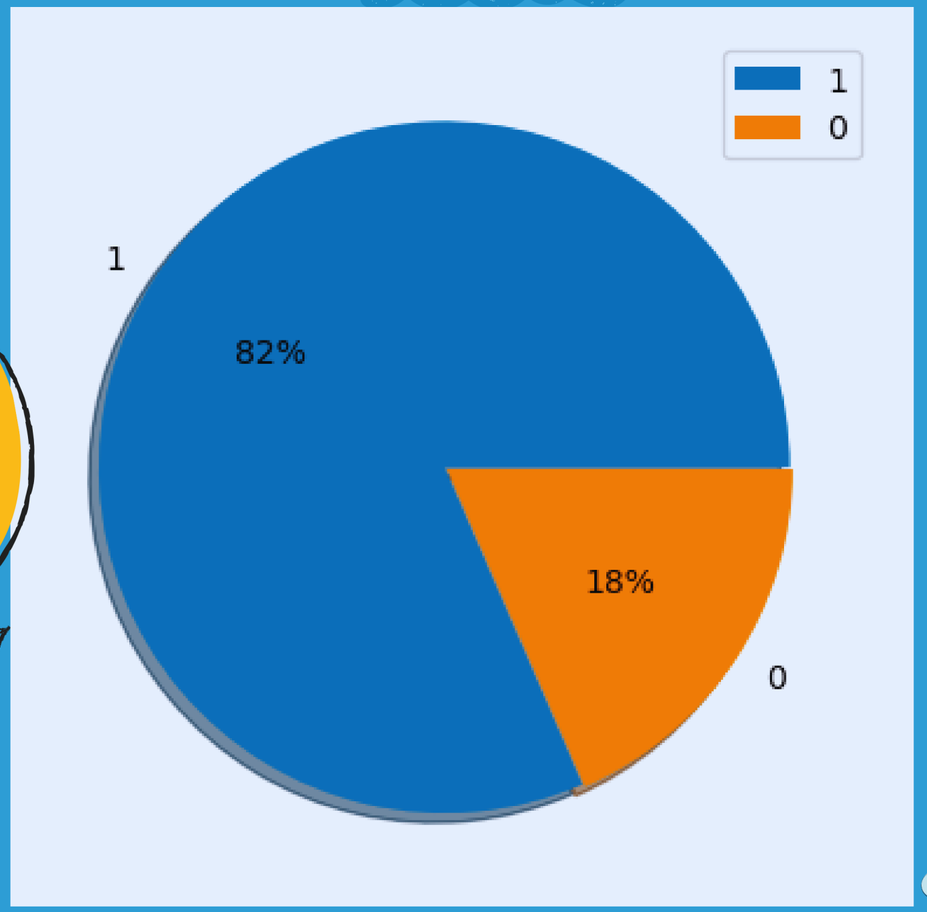
Preprocessing:

- data cleaning:
  - 349,440 rows × 16 columns
  - 258,395 rows × 15 columns
- WHO treshold establishment
- data integration
- data transformation:
  - SMOTE
- data exploration

Comparing the Potability Pre & Post WHO Guidelines



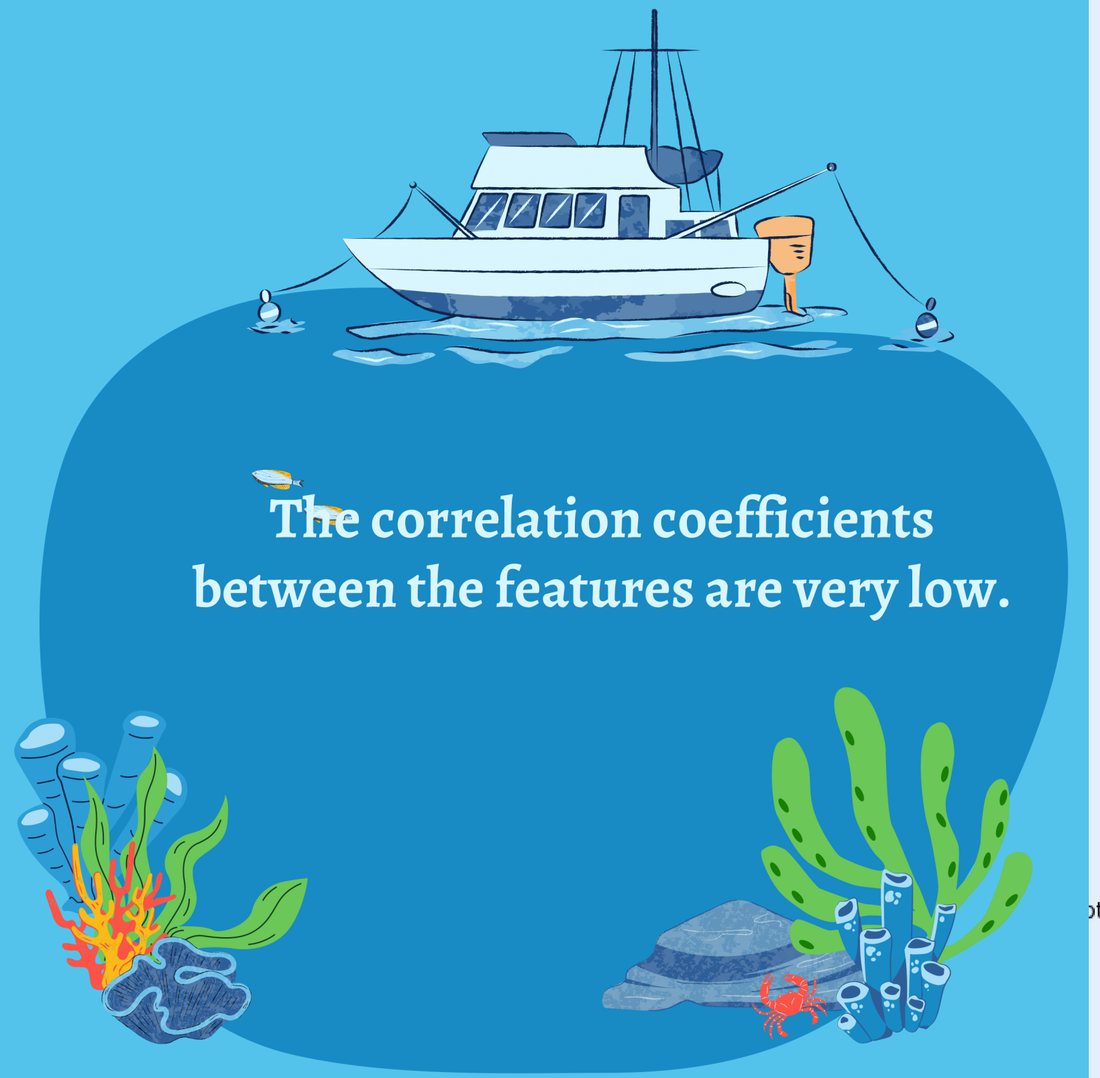
post mapping with WHO Potablility guidelines



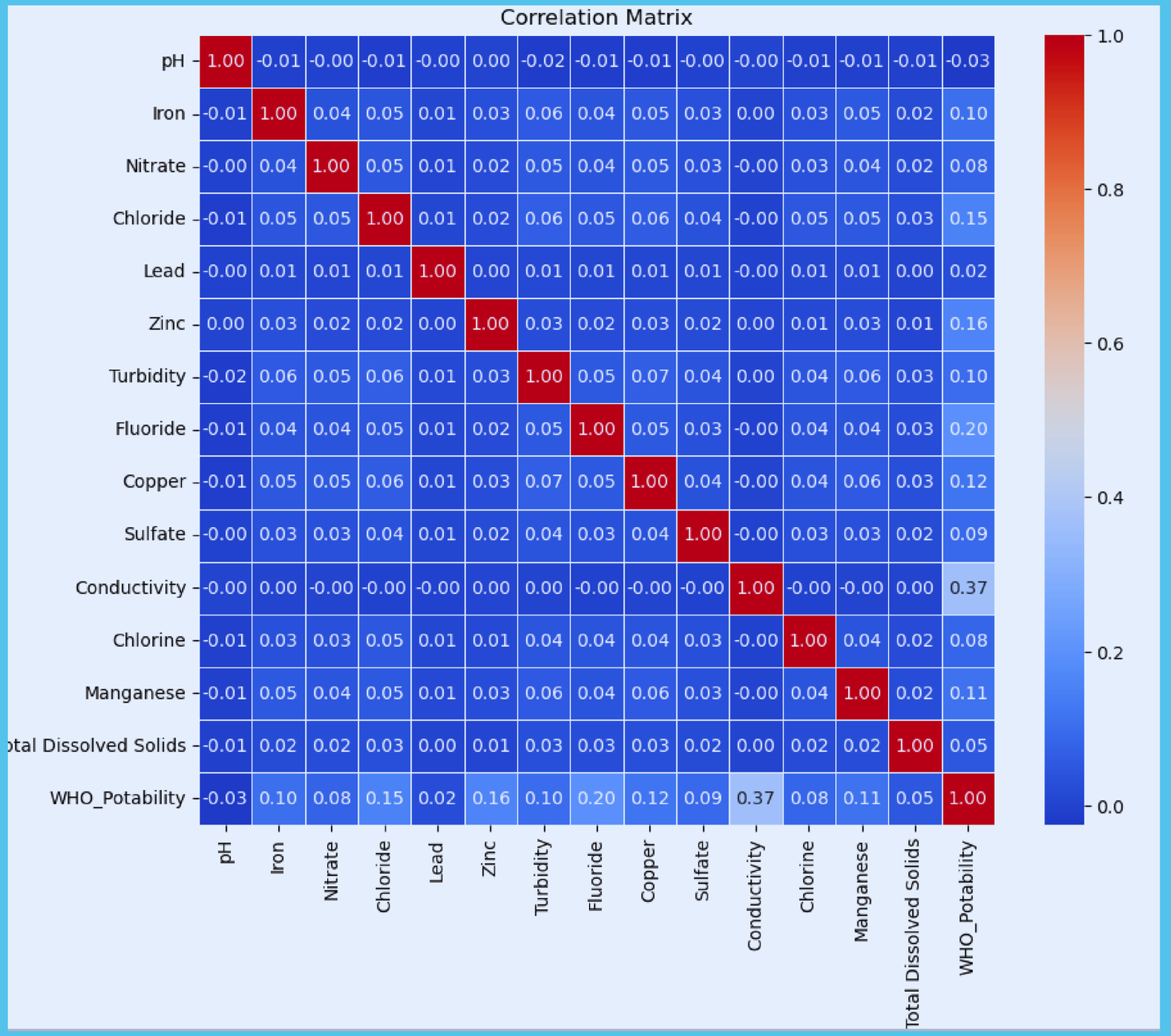
0 = Undefined
1 = Undefined

0 = Potable
1 = Non-Potable

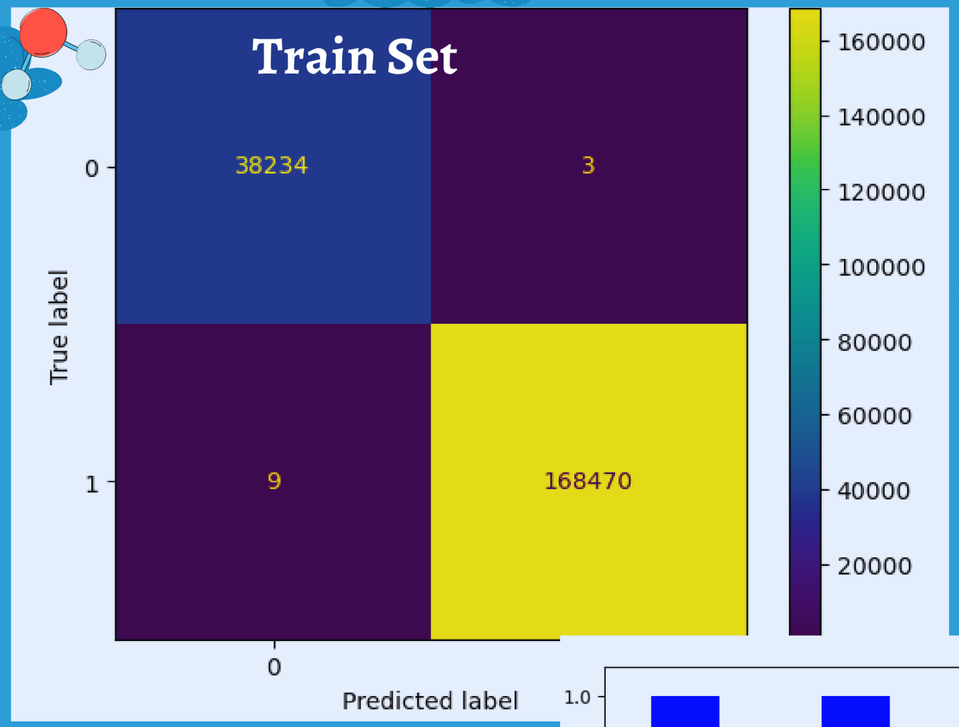
Correlation Matrix

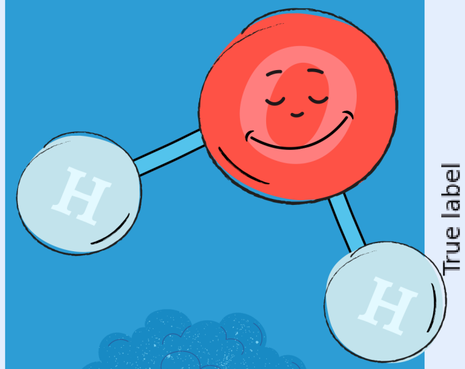


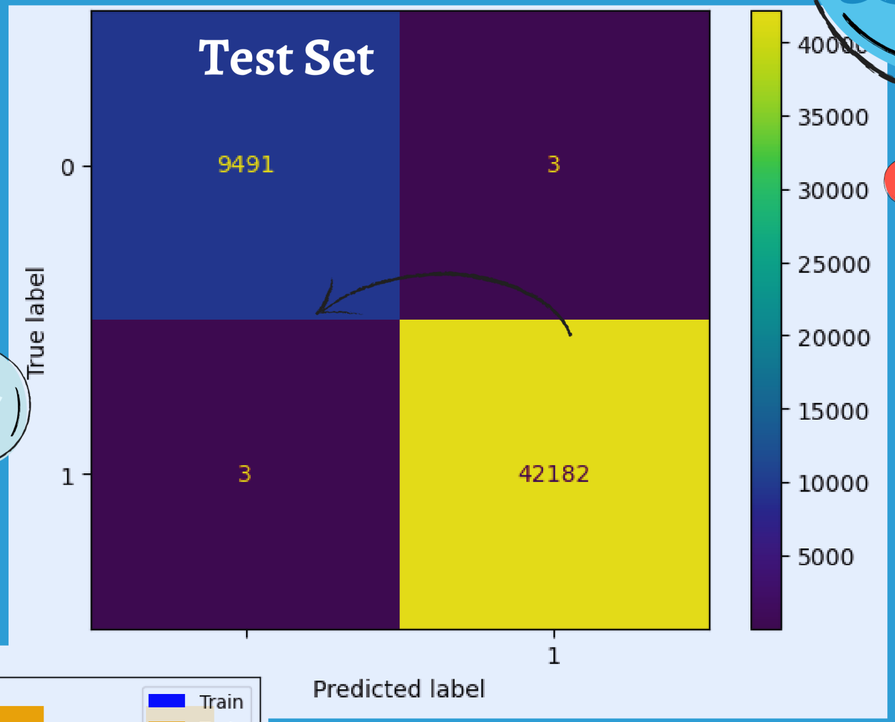
The correlation coefficients between the features are very low.

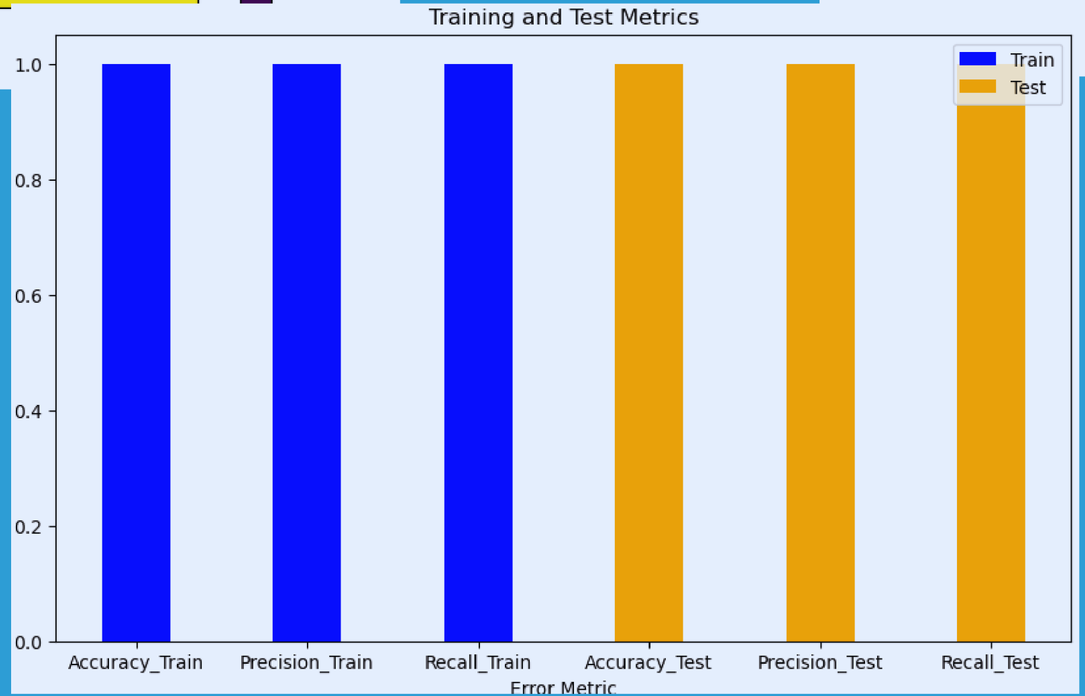


Confusion Matrix Randon Forest Classifier

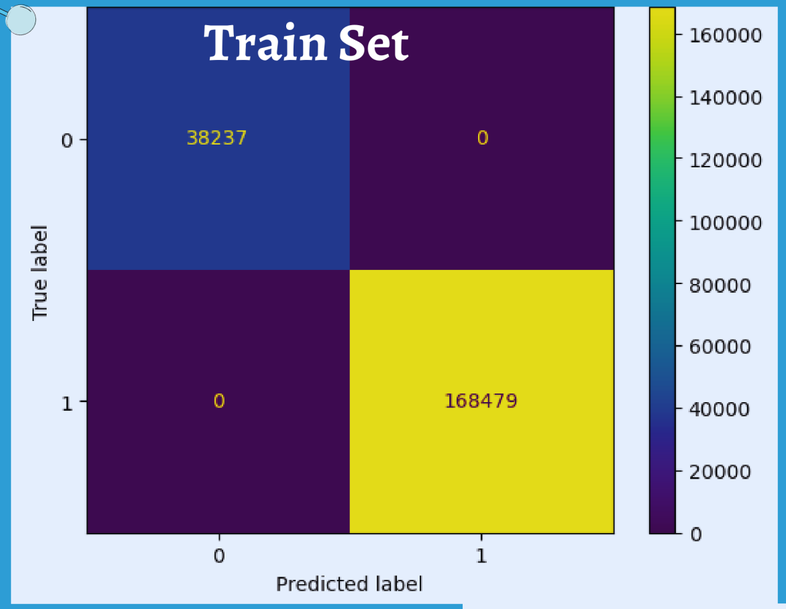


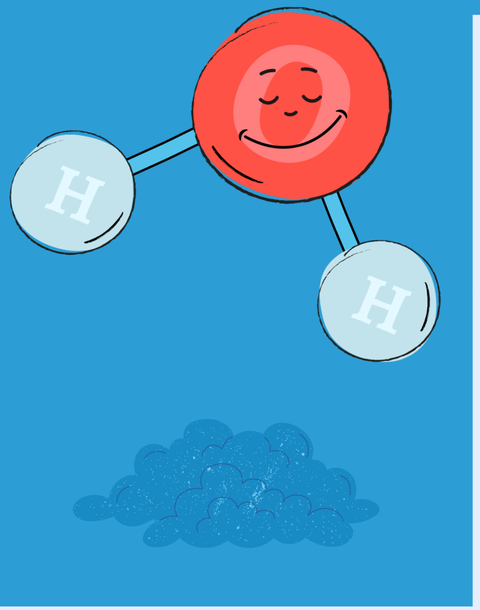


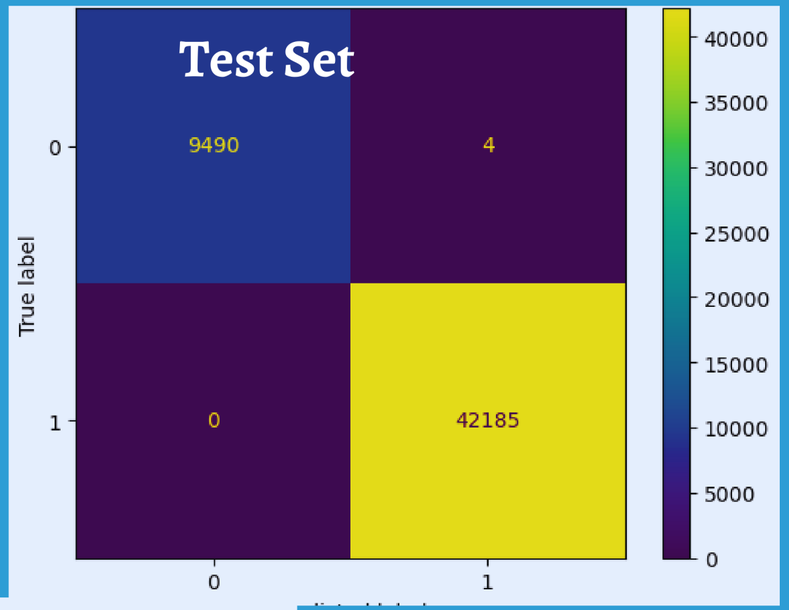


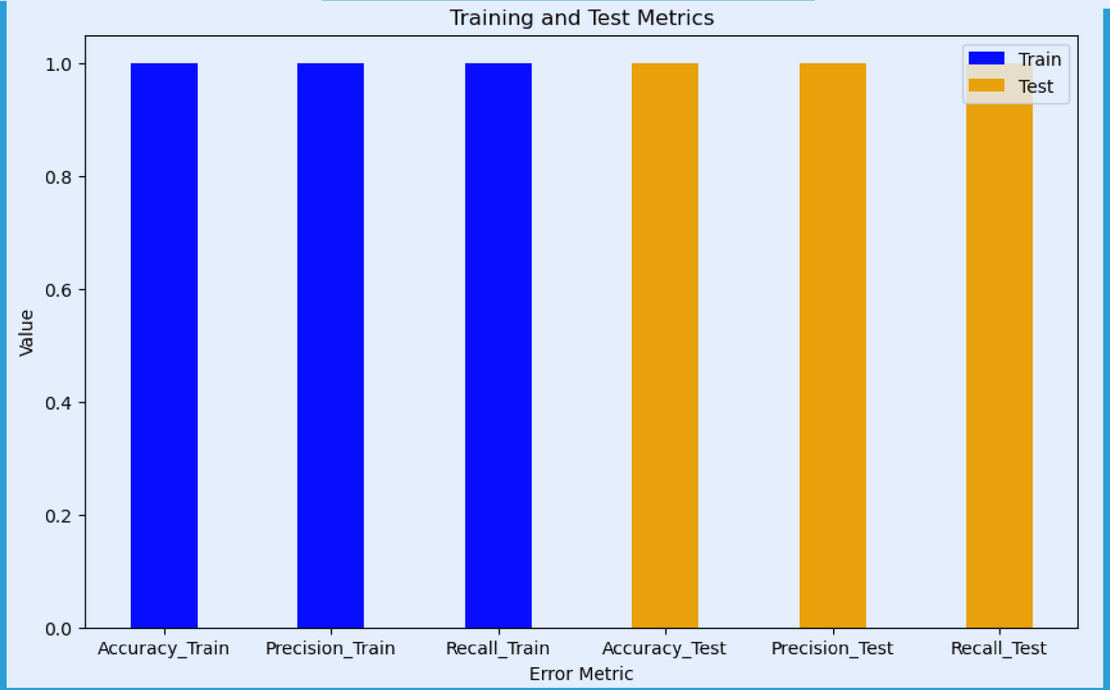


Confusion Matrix Decision Tree Classifier

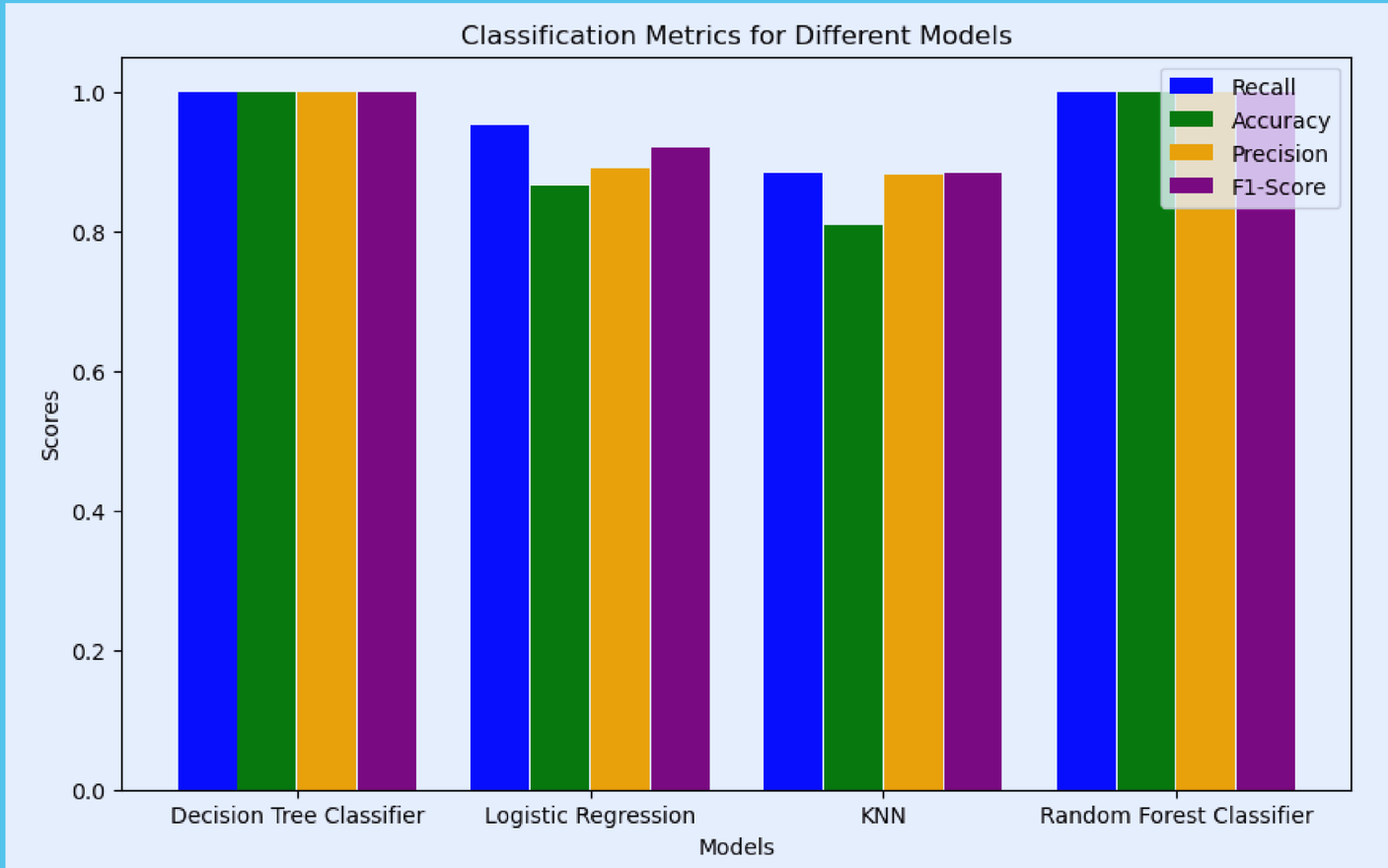


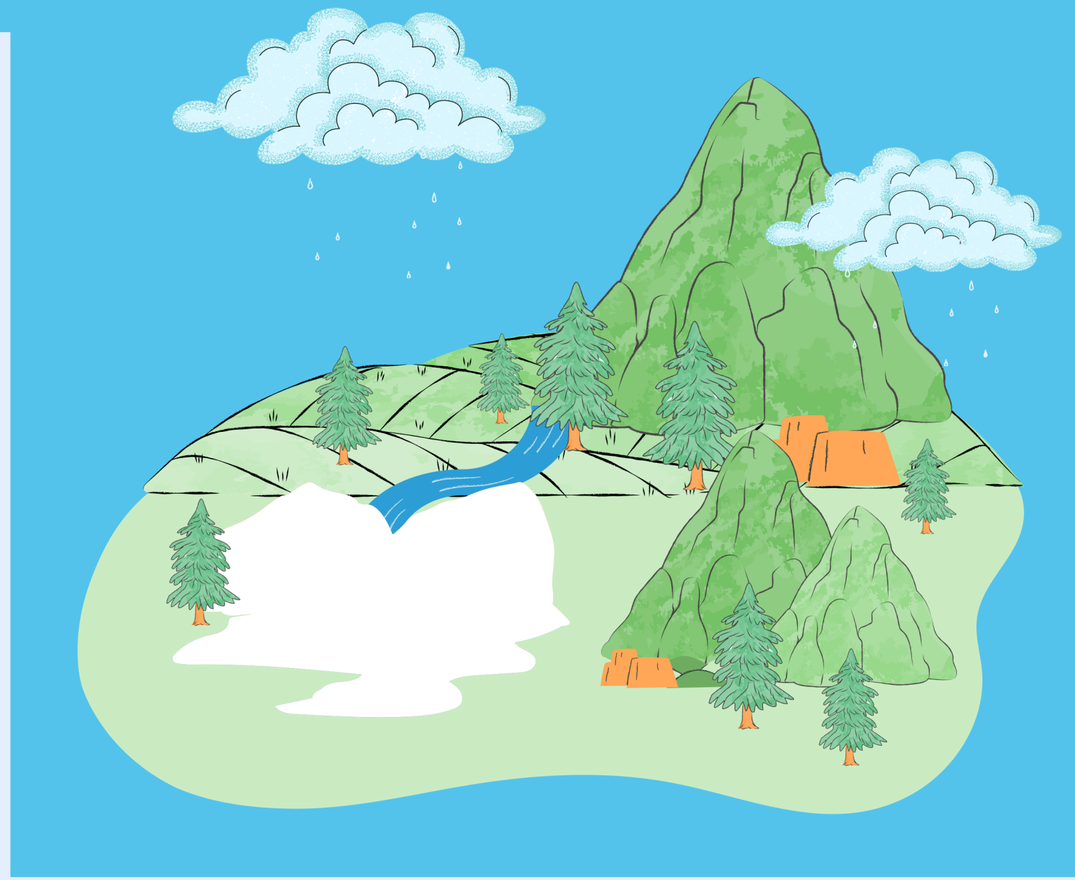






Performance Metrics Comparison on Models





	Recall	Accuracy	Precision	F1-Score
Classification Tree	0.999964	0.999923	0.999941	0.999953
LogisticRegression	0.952386	0.865497	0.890341	0.920319
KNN	0.884164	0.809033	0.881842	0.883001
RandomForestClassifier	0.999976	0.999918	0.999899	0.999938

Quick recap on how it works!

Smart Sensors:

- Small devices are placed in water sources or specific points like taps.
- These devices can measure different things in the water, like how acidic it is, the amount of minerals and other factors that affect its safety.

Data Collection and Transmission:

- The information from these sensors is sent to a central computer system.
- This happens in real-time, collecting updated data.

Machine Learning Models:

- The computer system uses advanced algorithms (machine learning) to analyze the data.
- These algorithms learn and get better over time, making the system smarter as it processes more information.

Result Interpretation:

- The recall value is critical in the model as it is important to avoid a false negative result (Type II error).
- False Negatives (fn): Instances where the model incorrectly predicts non-potable water as potable (actual class is 1, but predicted as 0).

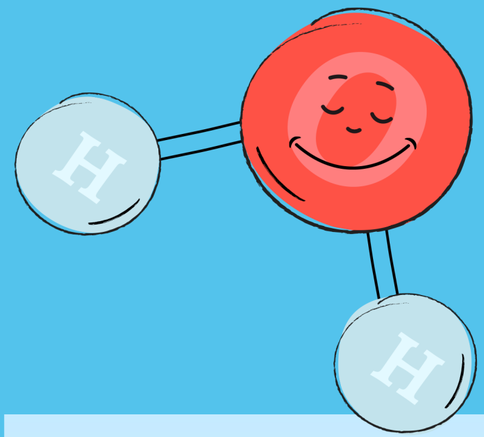
Conclusion

The implementation of machine learning-based solutions for monitoring water potability represents a significant opportunity for water service agencies to improve the quality of services provided, reduce operational costs and protect public health.

Through effective data collection and analysis, these solutions enable agencies to make more informed and timely decisions, helping to ensure a safe and sustainable water supply for communities around the world.

The decision tree and random forest classifiers are the best models for this project as they both give high performaces in recall value of 99%.

SWOT Analysis



Strength	Weakness	Opportunities	Threats
Team work Available data Large data set Time management Model accuracy	Takes time to run a model Data imbalance	Adaptable in various sectors: • Public health • Environmental • Education • Early detection of water contamination	Different locations have defined thresholds

References

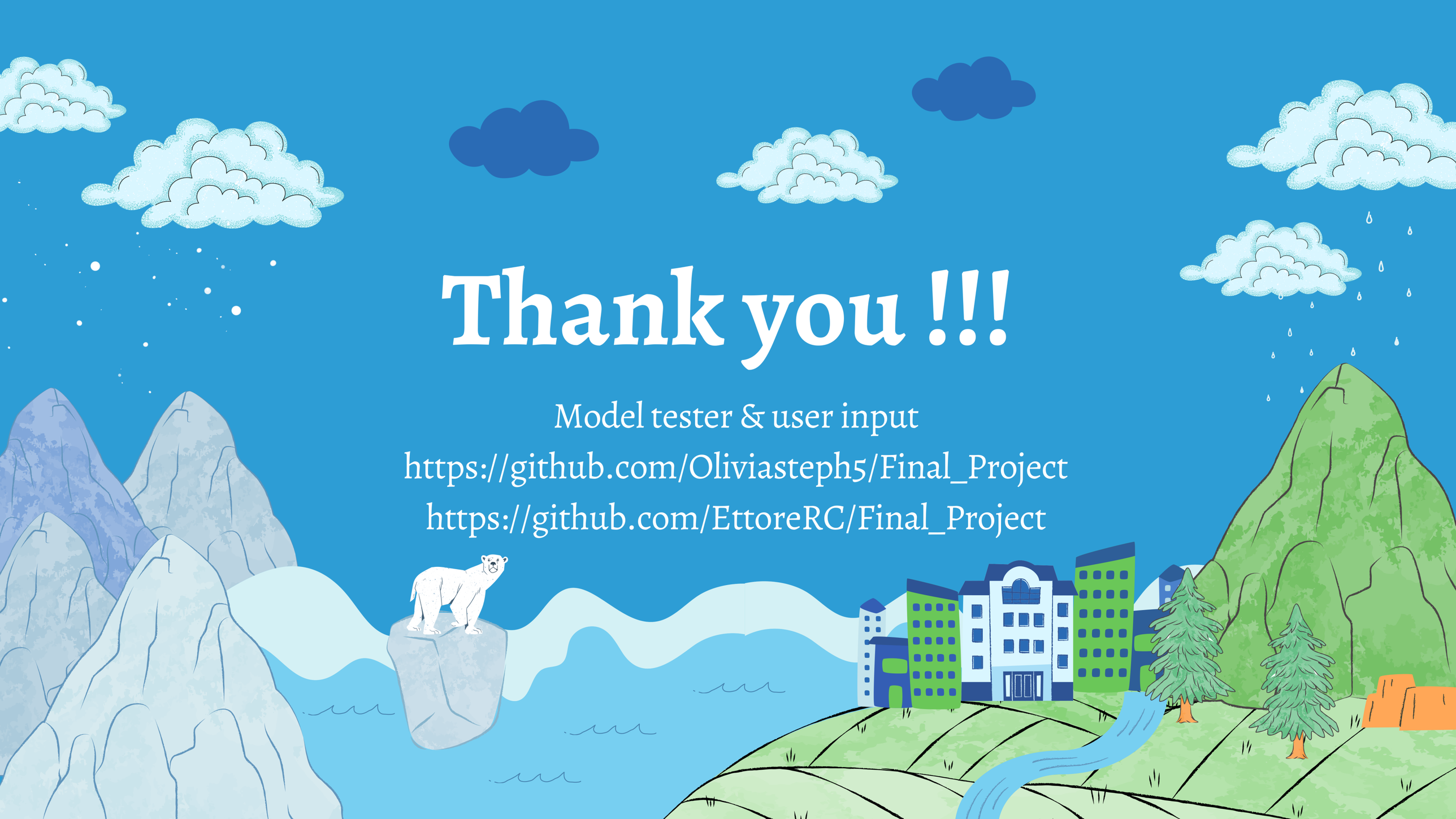
HTTPS://WWW.KAGGLE.COM/DATASETS/XIAOXIAOLIANGZI/WATER-POTABILITY-PREDICTIONHTTPS://WWW.WHO.INT/PUBLICATIONS/I/ITEM/9789240045064
HTTPS://WWW.WHO.INT/PUBLICATIONS/I/ITEM/9789240088740

Thank you !!!

Model tester & user input

https://github.com/Oliviasteph5/Final_Project

https://github.com/EttoreRC/Final_Project



Press these keys while on Present mode!

for blur

for a drumroll

for bubbles

for unveil

for confetti

for mic drop

for quiet

any number from 0-9 for a timer