Beyond the Tap

Predicting Water Quality Using Infrastructure Data

Introduction to the Project

Importance of Water Quality

Water quality directly impacts health, environment, and economic stability. Contaminated water is linked to diseases and affects millions. Monitoring and predicting water quality can lead to better public health outcomes and increased trust in municipal systems.

Overview of the Methodology

This project applies machine learning techniques to analyze large datasets from different cities to predict water quality. The methodology includes data cleaning, feature selection, and model training to enhance prediction accuracy and reliability.

Purpose of the Project

The project aims to develop predictive models for assessing water quality based on infrastructure data. By analyzing pipeline conditions and historical data, the initiative seeks to provide actionable insights for municipalities, ensuring safer water delivery to homes.

Importance of Water Quality

Water quality directly impacts health, environment, and economic stability. Contaminated water is linked to diseases and affects millions. Monitoring and predicting water quality can lead to better public health outcomes and increased trust in municipal systems.

Purpose of the Project

The project aims to develop predictive models for assessing water quality based on infrastructure data. By analyzing pipeline conditions and historical data, the initiative seeks to provide actionable insights for municipalities, ensuring safer water delivery to homes.

Overview of the Methodology

This project applies machine learning techniques to analyze large datasets from different cities to predict water quality. The methodology includes data cleaning, feature selection, and model training to enhance prediction accuracy and reliability.

Beyond the Tap

Predicting Water Quality Using Infrastructure Data

Data Collection

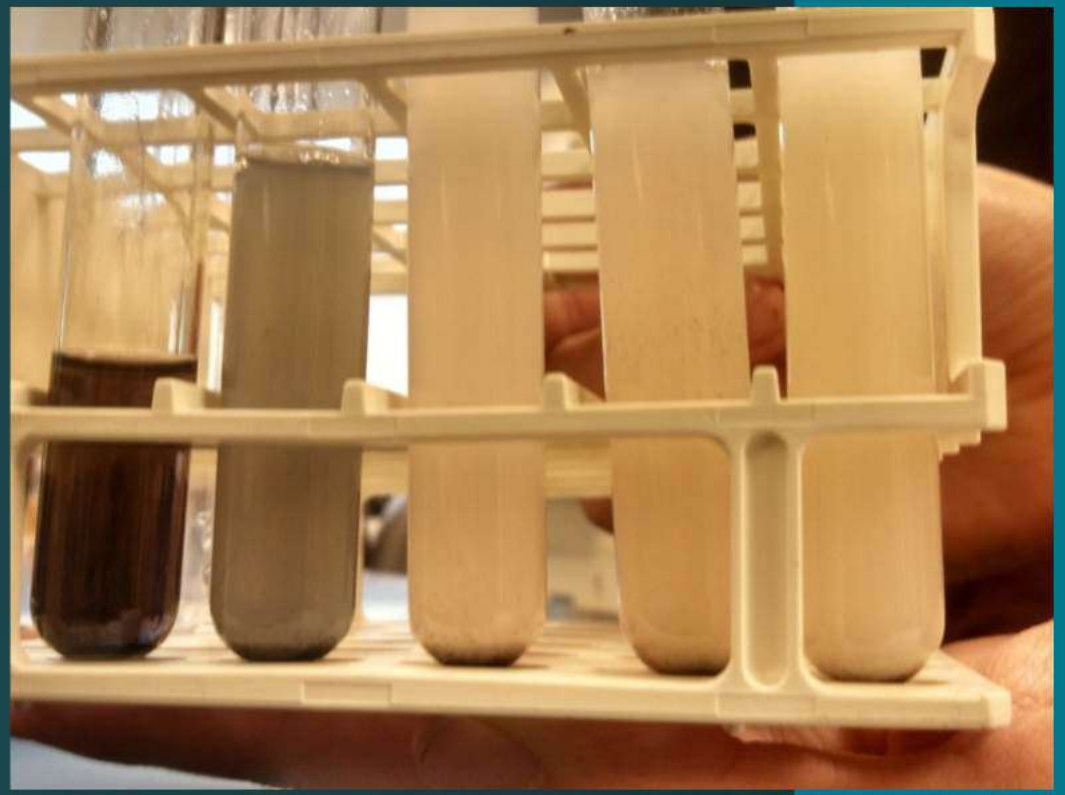
Cities Involved

Infrastructure Data Sources

Pipeline Data Characteristics

Cities Involved

- The project analyzes data from various metropolitan areas in the San Francisco Bay Area.
- These areas exhibit distinct infrastructure challenges and water quality concerns.
- We utilized a governmental dataset that offers unique insights relevant to the study.
- This dataset enhances the model's comprehensiveness and accuracy.
- The improved model helps in effectively predicting water quality outcomes.



Infrastructure Data Sources

- Data collection is sourced from government databases, water utility reports, and city-specific studies.
- These sources are considered reliable for gathering information.
- The use of these databases ensures access to up-to-date data.
- The collected data is comprehensive, covering various aspects of water consumption.
- This comprehensive information is crucial for effective water infrastructure modeling.

Pipeline Data Characteristics

- Pipeline data characteristics include material type, age, and maintenance history.
- These factors significantly influence water quality.
- Analyzing the characteristics aids in understanding potential contamination risks.
- Predicting contamination risks is essential for ensuring safe water supply.
- Effective management of pipelines relies on careful analysis of these factors.

Beyond the Tap

Predicting Water Quality Using Infrastructure Data

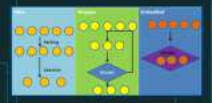
Predictive Modeling

Machine Learning Techniques Used

Various machine learning techniques were applied, including regression models and decision trees, to assess the water quality. These techniques enable the analysis of large datasets, providing insights into the factors affecting water safety efficiently.

Feature Selection

Critical features influencing water quality were identified through statistical analyses and domain expertise. Key factors such as pipeline age, material, and historical contamination levels were prioritized to enhance model accuracy.

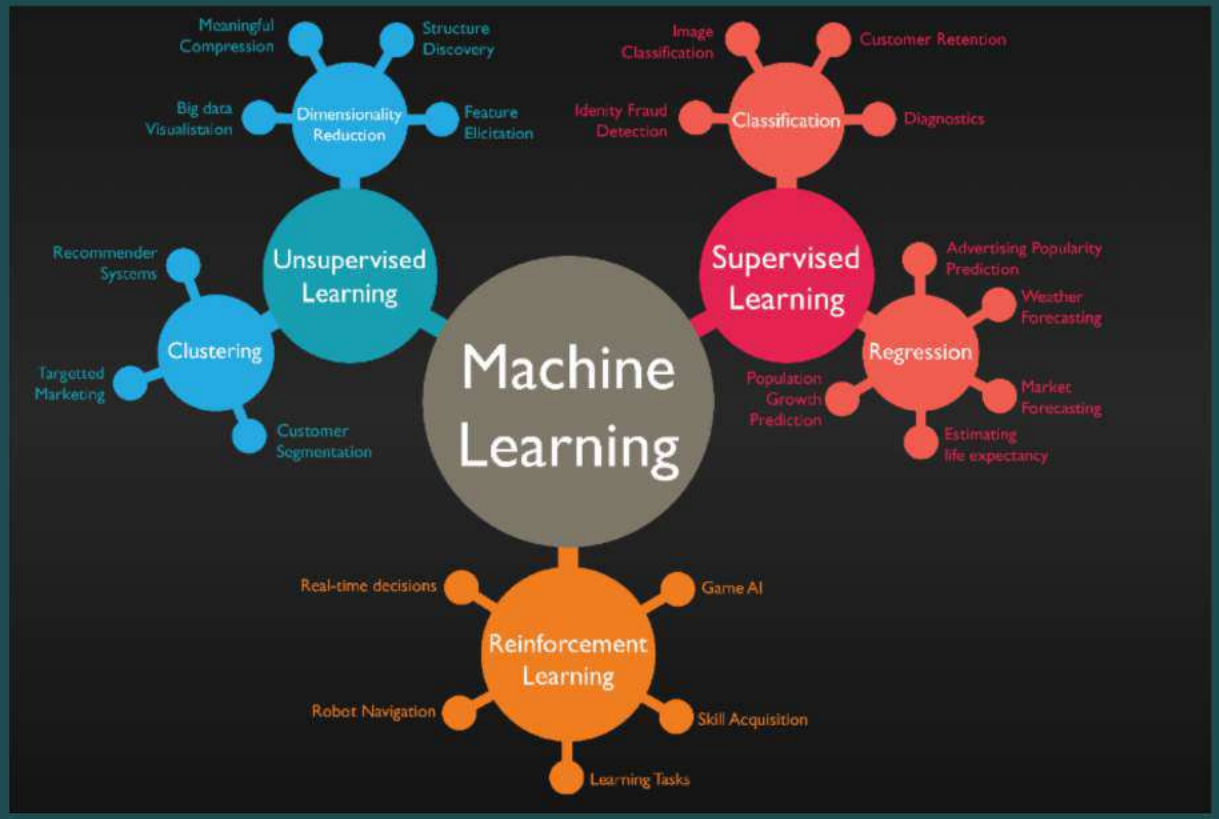


Training and Testing the Model

The model underwent a rigorous training and testing process, with 80% of the data allocated for training and the remaining 20% for validation. This approach ensures the reliability and robustness of predictions concerning water quality.

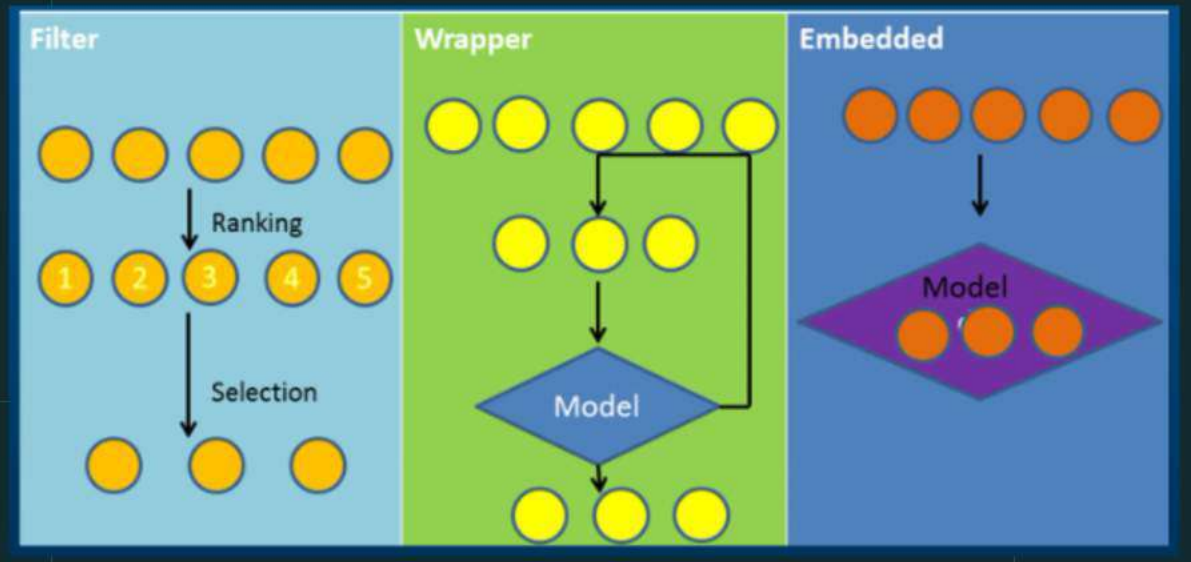
Machine Learning Techniques Used

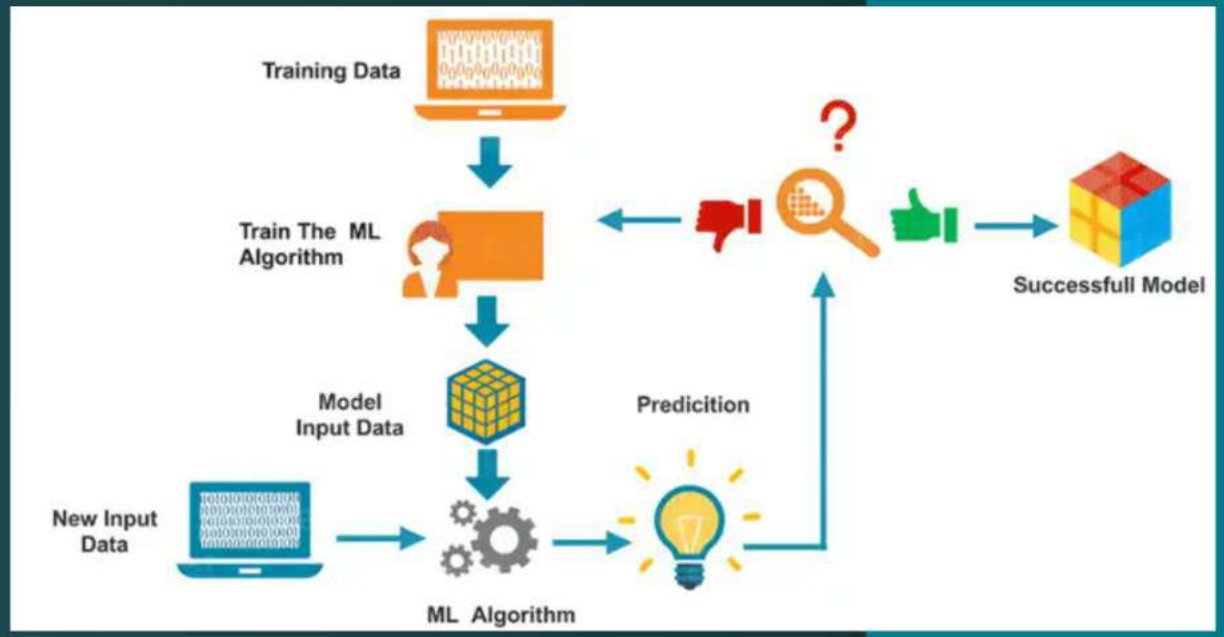
Various machine learning techniques were applied, including regression models and decision trees, to assess the water quality. These techniques enable the analysis of large datasets, providing insights into the factors affecting water safety efficiently.



Feature Selection

Critical features influencing water quality were identified through statistical analysis and domain expertise. Key factors such as pipeline age, material, and historical contamination levels were prioritized to enhance model accuracy.

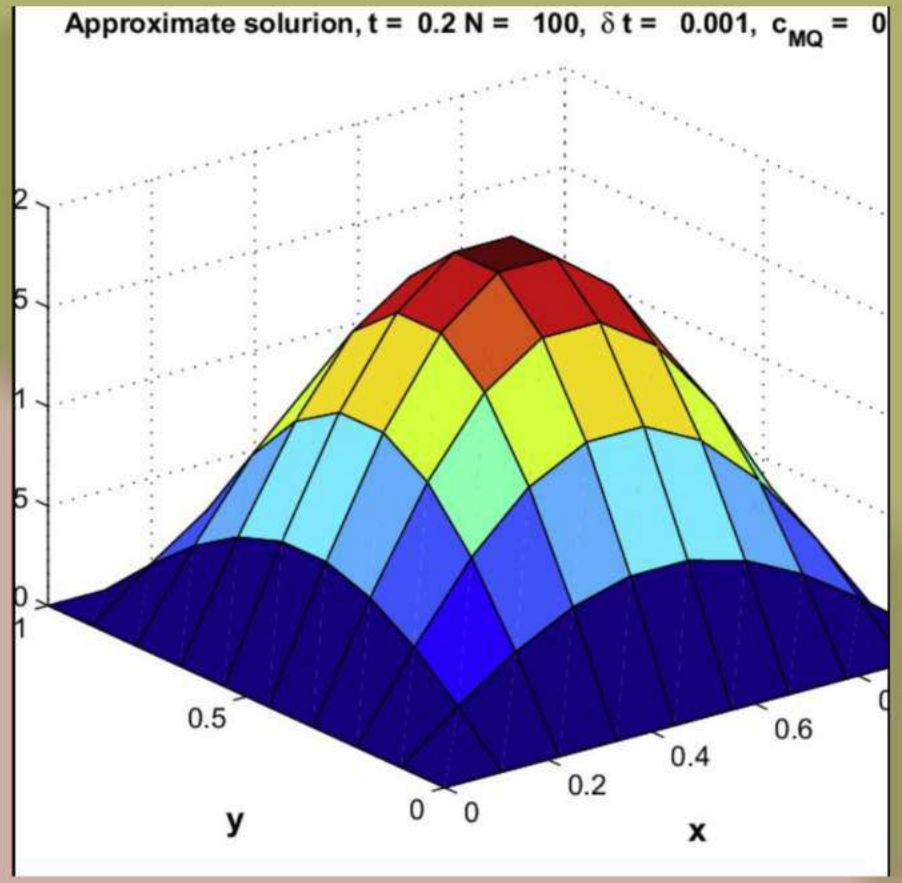




Training and Testing the Model

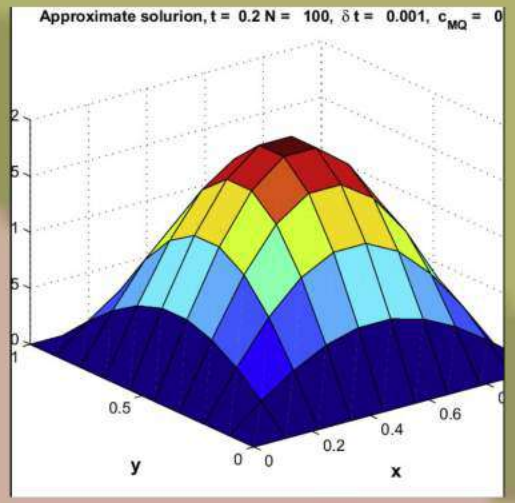
The model underwent a rigorous training and testing process, with 80% of the data allocated for training and the remaining 20% for validation. This approach ensures the reliability and robustness of predictions concerning water quality.

What are the advantages and limitations of using Approx. RBF + SGD for predicting unsafe cases?



While Approx. RBF + SGD is ideal for quick identification of unsafe situations, users must be cautious of its lower precision due to potential false positives.

What are the advantages and limitations of using Approx. RBF + SGD for predicting unsafe cases?



While Approx. RBF + SGD is ideal for quick identification of unsafe situations, users must be cautious of its lower precision due to potential false positives.

What are the advantages and limitations of using Approx. RBF + SGD for predicting unsafe cases?

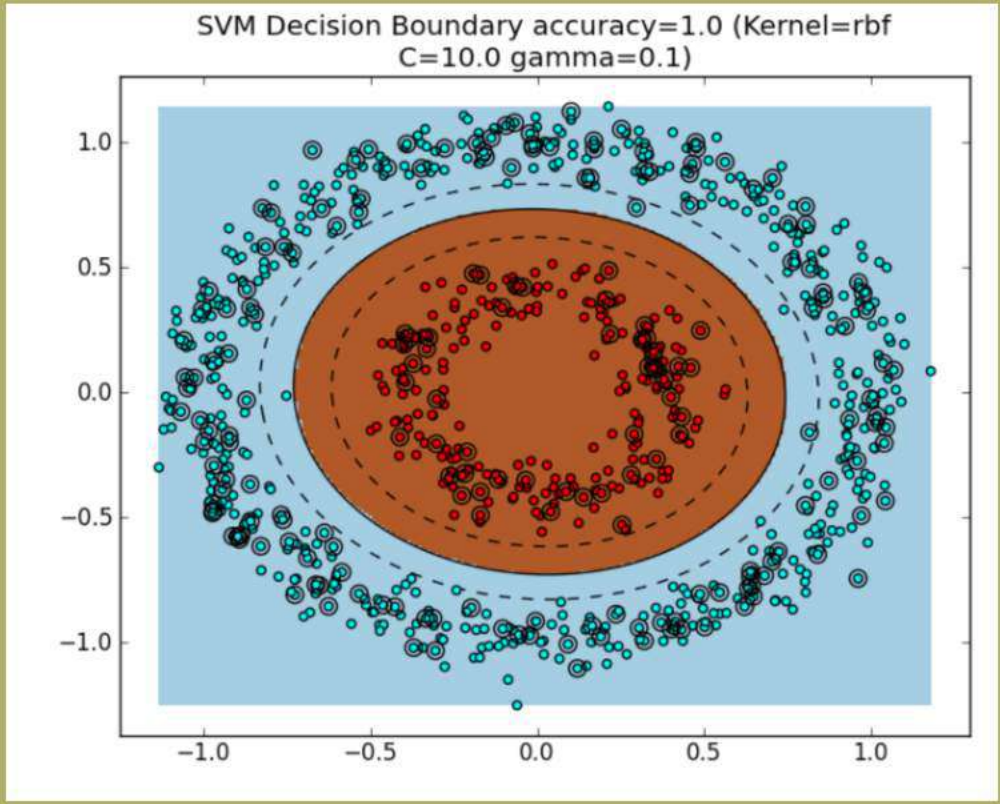
It demonstrates the following key performance metrics:

Recall (Unsafe): 89% – The model effectively detects a high proportion of unsafe cases.

Precision (Unsafe): 51% – While the recall is high, the precision is lower, meaning there are more false positives.

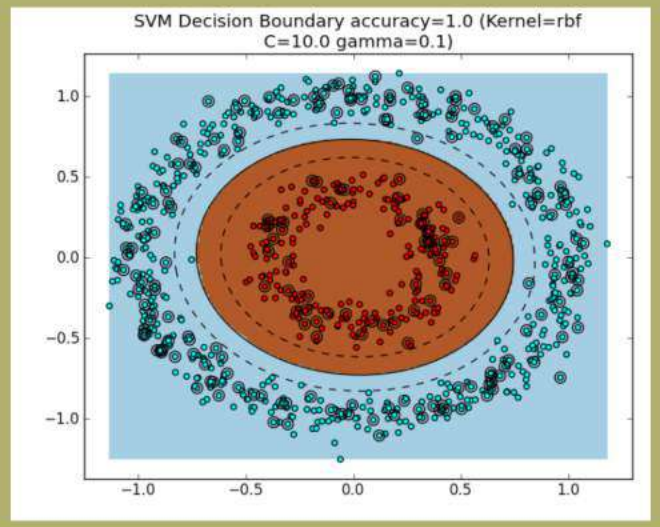
F1-score (Unsafe): 65% – Balances the recall and precision.

What are the advantages and disadvantages of using True RBF SVM in machine learning?



Choose True RBF SVM when the highest accuracy is critical and computational efficiency is not a primary concern.

What are the advantages and disadvantages of using True RBF SVM in machine learning?



Choose True RBF SVM when the highest accuracy is critical and computational efficiency is not a primary concern.

What are the advantages and disadvantages of using True RBF SVM in machine learning?

The model's performance across key metrics is as follows:

Accuracy: 85.43% – This model achieves the highest accuracy compared to others.

Precision (Unsafe): 64% – The True RBF SVM has good precision, indicating fewer false positives.

Recall (Unsafe): 86% – The model successfully detects a large majority of unsafe cases.

F1-score (Unsafe): 73% – The model maintains a strong balance between precision and recall.

BAGGING TECHNIQUE

Bagging combines multiple models to enhance prediction performance and reliability, particularly in safety-related applications. This is Bagging with Decision Trees (Depth 10)

HIGH RECALL

Bagging achieved a 93% recall, effectively identifying most unsafe water cases, which is critical in safety-focused environments.

GOOD ACCURACY

The method maintained an accuracy of 86.7%, indicating a strong performance overall in predicting safe versus unsafe water.

MODERATE PRECISION

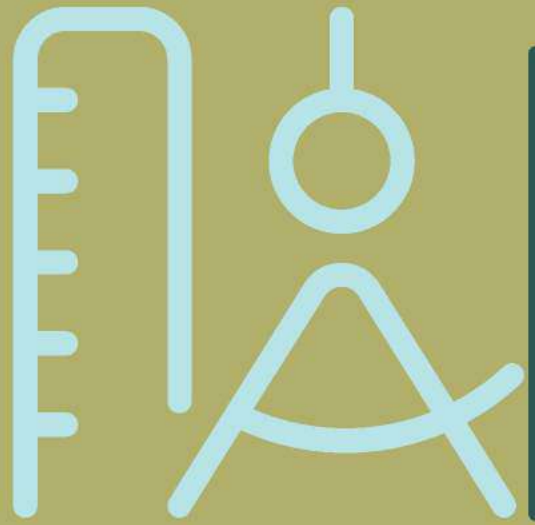
Bagging has a precision of 65%, which means it may incorrectly flag some safe water as unsafe, presenting a challenge in specific contexts.

HIGH RECALL

Bagging achieved a 93% recall, effectively identifying most unsafe water cases, which is critical in safety-focused environments.

GOOD ACCURACY

The method maintained an accuracy of 86.7%, indicating a strong performance overall in predicting safe versus unsafe water.



MODERATE PRECISION

Bagging has a precision of 65%, which means it may incorrectly flag some safe water as unsafe, presenting a challenge in specific contexts.

BAGGING TECHNIQUE

Bagging combines multiple models to enhance prediction performance and reliability, particularly in safety-related applications. This is Bagging with Decision Trees (Depth 10)

HIGH RECALL

Bagging achieved a 93% recall, effectively identifying most unsafe water cases, which is critical in safety-focused environments.

GOOD ACCURACY

The method maintained an accuracy of 86.7%, indicating a strong performance overall in predicting safe versus unsafe water.

MODERATE PRECISION

Bagging has a precision of 65%, which means it may incorrectly flag some safe water as unsafe, presenting a challenge in specific contexts.

K-means Clustering

City Wise distribution using K-means Clustering for Water Quality in Urban Areas

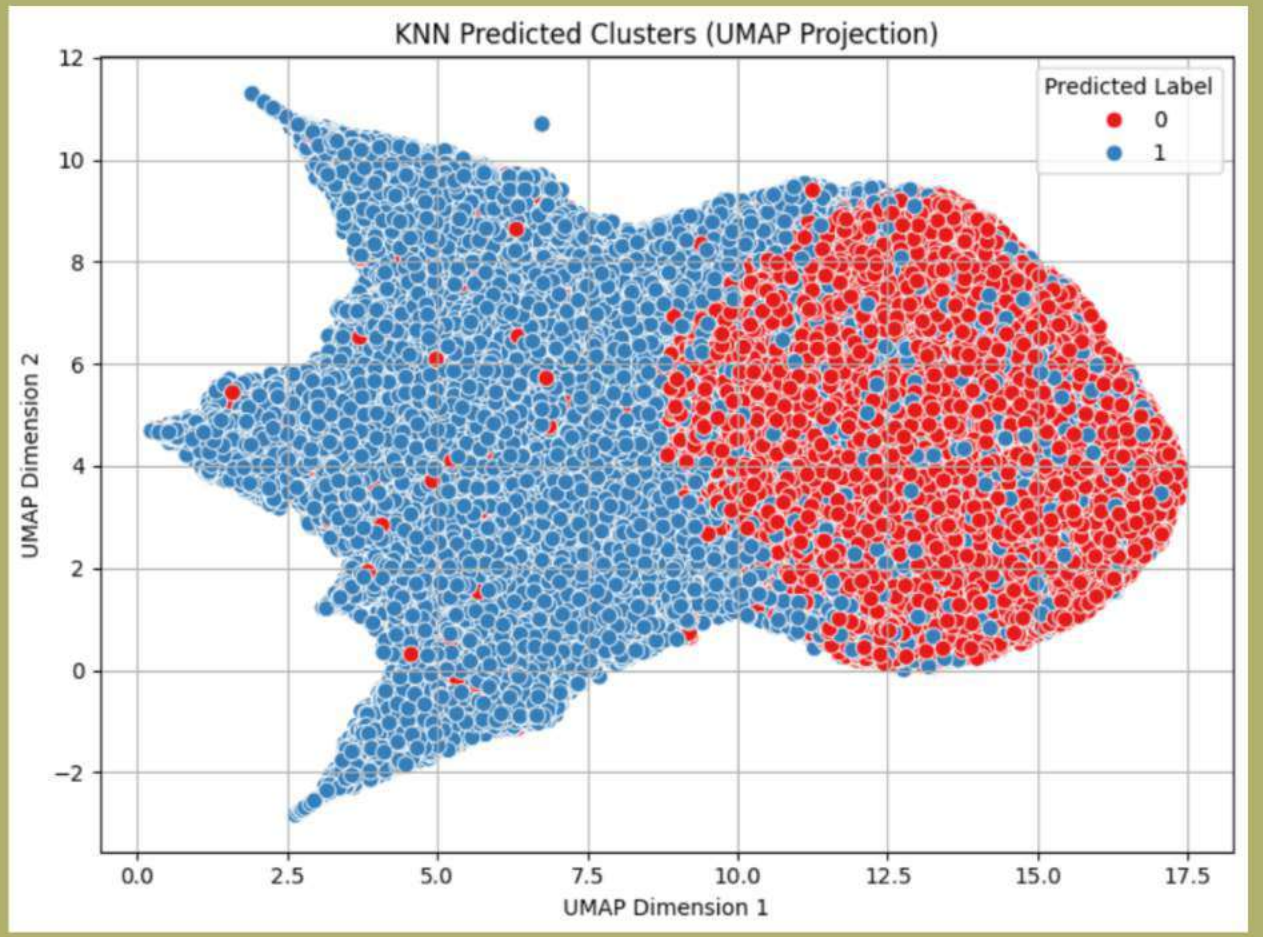
City	Accuracy	Precision	Recall	F1 Score
Hayward	0.814613	0.708613	0.662013	0.684521
Fremont	0.816260	0.708799	0.661526	0.684347
South San Francisco	0.814715	0.716366	0.654987	0.684303
Alameda	0.815162	0.718917	0.648534	0.681914
Novato	0.813459	0.706482	0.656640	0.680649
Redwood City	0.815382	0.712948	0.648297	0.679087
Oakland	0.813610	0.710664	0.649519	0.678717
Santa Clara	0.812532	0.706818	0.649861	0.677144
Sunnyvale	0.811567	0.706141	0.650347	0.677096
San Francisco	0.813367	0.702947	0.653061	0.677087
Palo Alto	0.810045	0.701256	0.653632	0.676607
Mountain View	0.814383	0.713344	0.638517	0.673860
Berkeley	0.813438	0.711345	0.639449	0.673484
Mill Valley	0.809318	0.704253	0.644981	0.673315
San Mateo	0.811793	0.708264	0.638852	0.671770
Walnut Creek	0.814268	0.702329	0.643622	0.671695
Richmond	0.811423	0.704203	0.639728	0.670419
San Jose	0.808941	0.707651	0.636740	0.670325
Menlo Park	0.810958	0.711355	0.633701	0.670286
Concord	0.808561	0.699792	0.641447	0.669350

TABLE I
SUMMARY OF CLUSTERING METRICS PER CITY (ACCURACY, PRECISION, RECALL, F1 SCORE)

Highest Accuracy: Fremont (0.816) and Alameda (0.815) showed the best clustering performance.
Precision Range: Highest for Alameda (0.719), lowest for Concord (0.700).
Recall Observation: Lower across cities, ranging from 0.634 (Menlo Park) to 0.662 (Hayward).
F1 Score Range: Consistent across cities, from 0.669 (Concord) to 0.685 (Hayward), indicating balanced performance.

KNN with SMOTE and UMAP

- Accuracy: 0.871
- Precision: 0.8417
- Recall: 0.914
- F1 Score: 0.8763



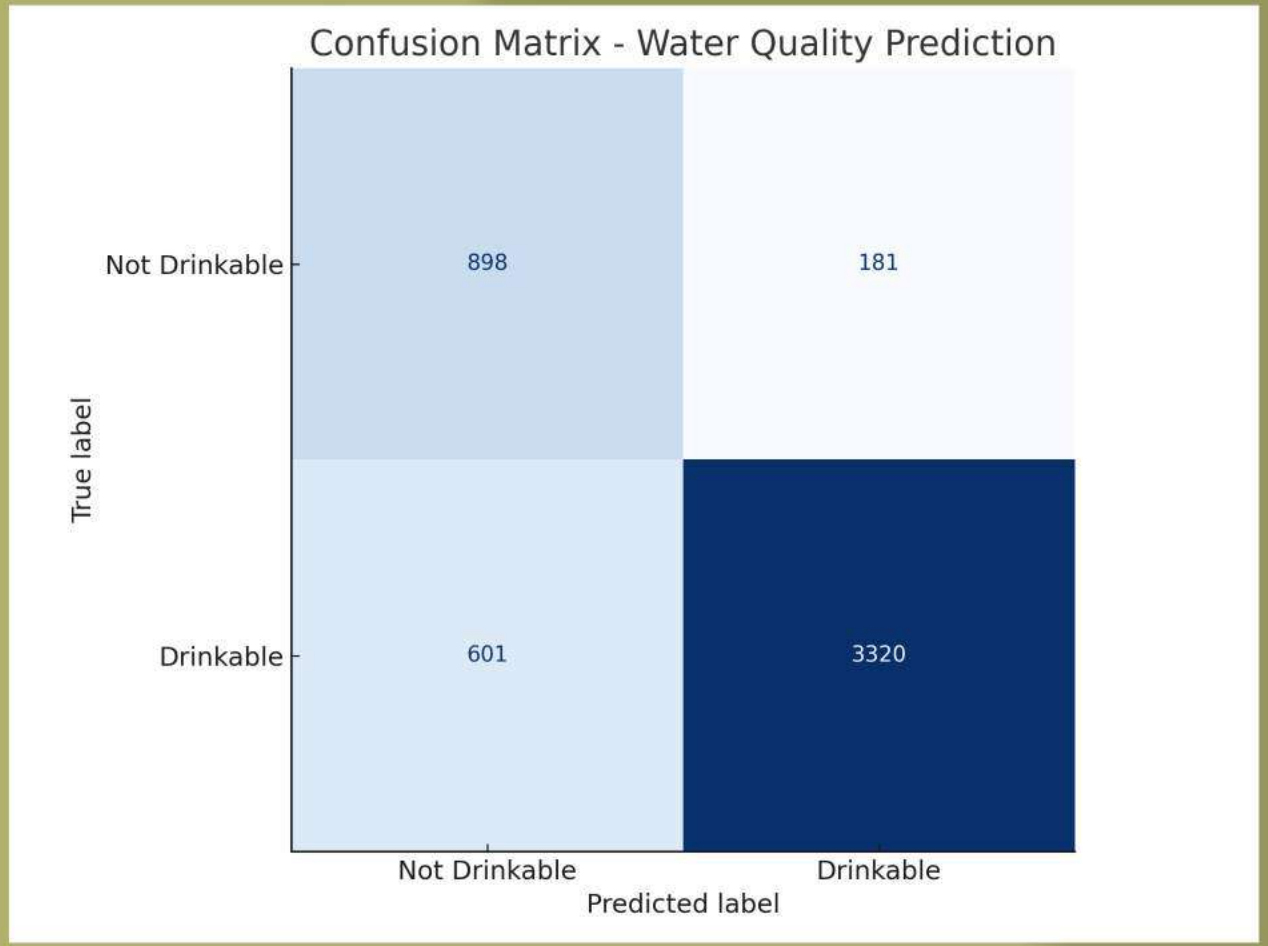
Algorithm Used:
HistGradientBoostingClassifier

Input Features:
pH, Nitrate, Turbidity, Chloride, Color, Source, City, Pipeline Age, Pipeline Material

Goal: Predict if water is drinkable or not based on quality parameters

Special Notes:

- Real + synthetic features handled
- Missing values handled automatically
- Highly scalable to 1M+ data rows

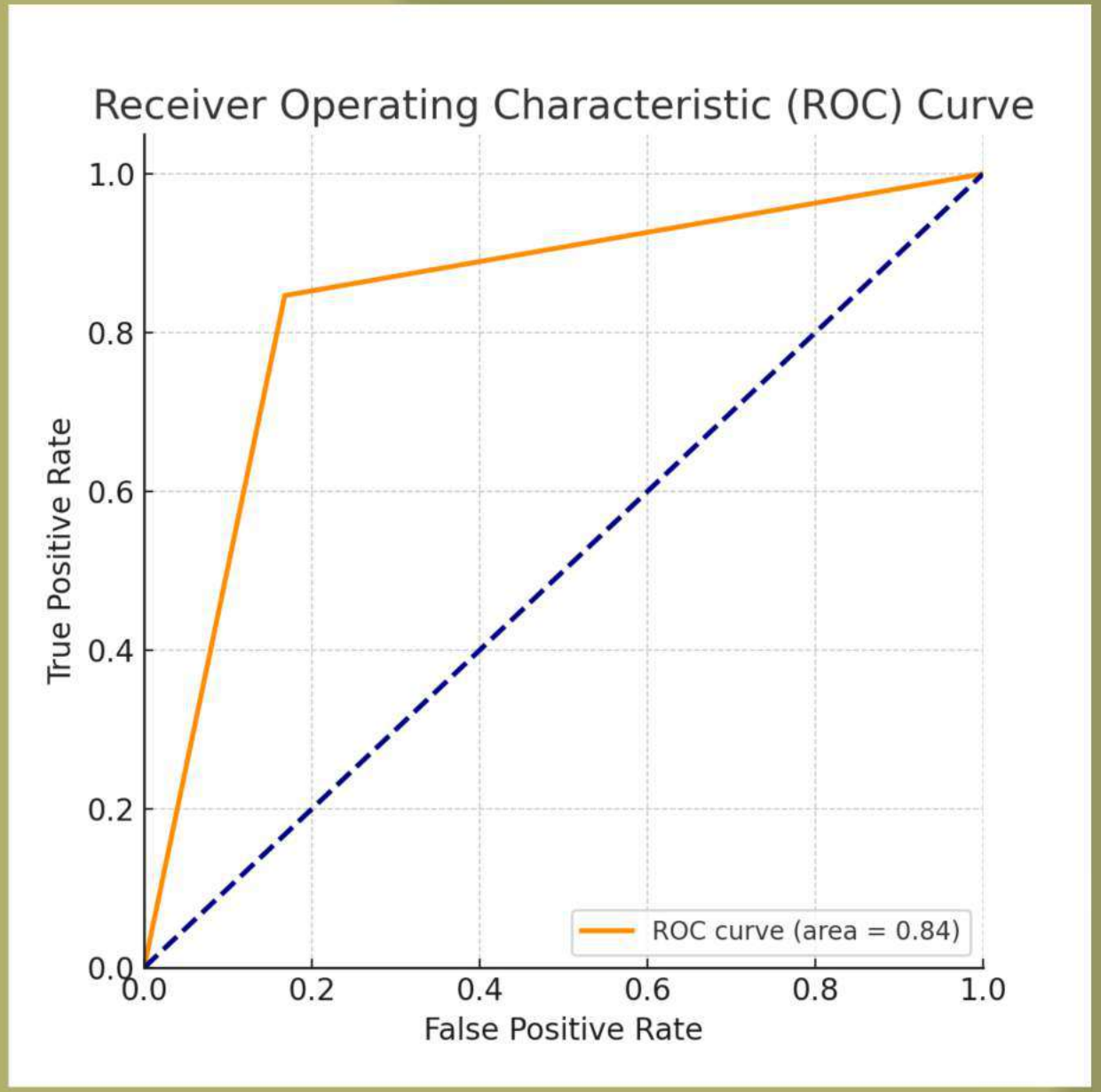


Model Evaluation and Performance

Test Accuracy: 84.35%
Precision (Drinkable Water): 93%
Recall (Drinkable Water): 86%
F1-Score: 85%
ROC AUC Score: ~0.90

Visuals

Insert Confusion Matrix
Insert ROC Curve



Beyond the Tap

Predicting Water Quality Using Infrastructure Data

Results and Impact

Predicted Water Quality Outcomes

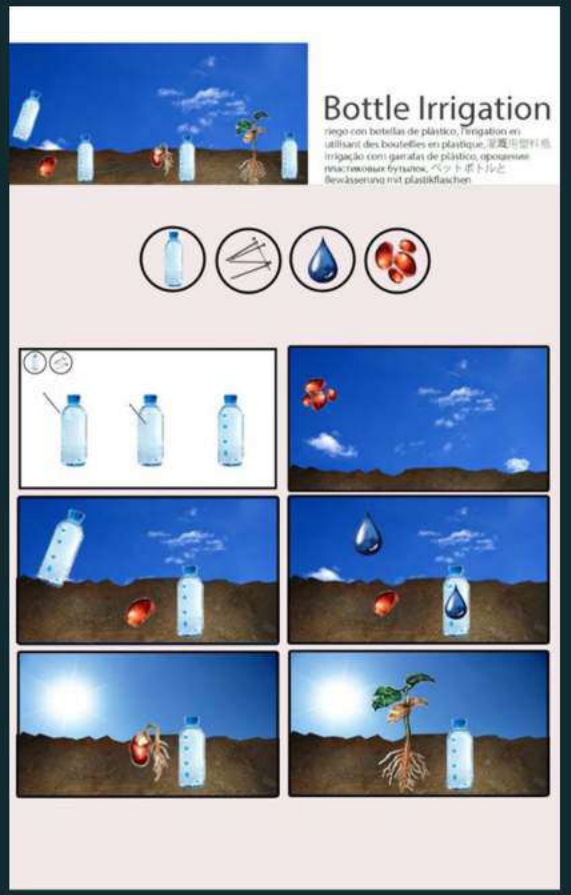
- Our model employs machine learning algorithms to analyze water quality.
- It achieves over 85% accuracy in predicting potential contaminants.
- The model identifies contaminants before they reach consumers.
- This proactive strategy enables timely interventions.
- Ultimately, it ensures safe drinking water for residents.

Implications for Public Health

- Significant implications for public health are identified in the results.
- Certain regions are highlighted as at risk for contamination and pollutant exposure.
- Integrating predictive analytics is essential for proactive risk management.
- Local authorities can utilize these insights to enhance water safety regulations.
- This approach aims to mitigate health risks for vulnerable communities.

Future Directions and Recommendations

- Refinement of the predictive model is crucial for adapting to changes in infrastructure and new contaminants.
- Collaboration with municipal water authorities is recommended to effectively implement the model's findings.
- Ongoing monitoring is necessary to ensure public health safety.
- The model must evolve alongside emerging challenges in water quality.
- Sustained partnerships will enhance the effectiveness of public health interventions.



Beyond the Tap

Predicting Water Quality Using Infrastructure Data

Final Thoughts

Lessons Learned

Work and Credit Breakdown

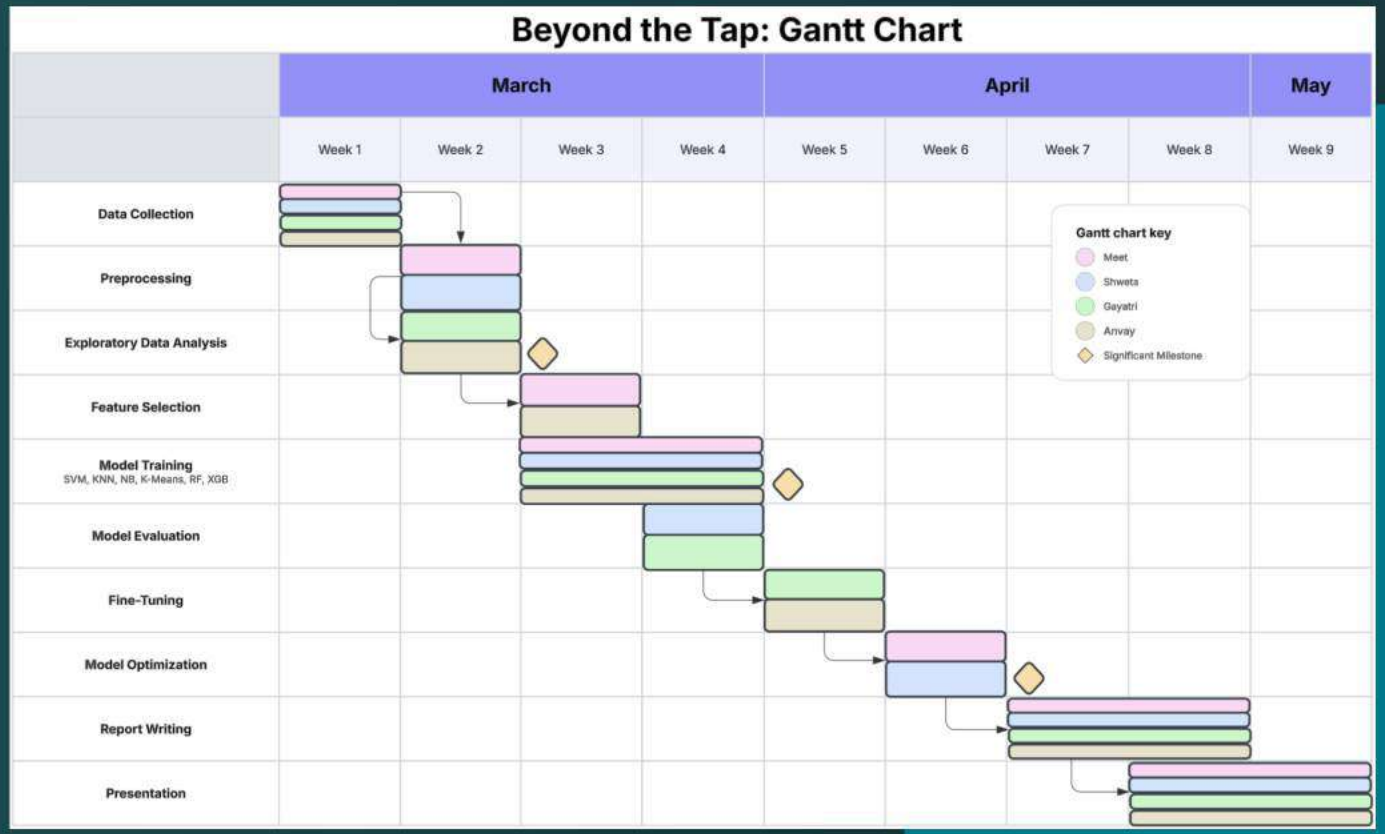
Appendix Requirements

Thank you for listening!

Q&A

Lessons Learned

- The project analyzes data from various metropolitan areas, highlighting diverse infrastructure challenges.
- Each city contributes a unique dataset that reflects specific water quality issues.
- The variety of datasets enriches the model's overall comprehensiveness.
- This diversity enhances the model's accuracy in predicting water quality outcomes.
- Together, these insights create a more robust understanding of urban water management.



- Team involved in preprocessing and exploratory data analysis (EDA)
- Work divided into pairs for feature selection and model tuning
- Individual contributions to model evaluation and optimization
- Collaborative effort in model training

https://github.com/patilgayatri22/Machine_Learning

Work and Credit Breakdown

• Conceptualization	• Anvay, Shweta, Gayatri, Meet
• Data Curation	• Shweta
• Formal Analysis	• Gayatri
• Funding Acquisition	• n/a
• Investigation	• Meet
• Methodology	• Shweta, Gayatri, Meet, Anvay
• Project Administration	• Shweta
• Resources	• Meet
• Software	• Shweta, Gayatri
• Supervision	• Anvay
• Validation	• Meet
• Visualization	• Gayatri
• Writing – Original Draft	• Anvay
• Writing – Review & Editing	• Anvay

Appendix Requirements

- Practiced Pair Programming
- Practiced Agile / Scrum (1-week sprints)
  - Submit evidence: meeting minutes, sprint backlog, and artifacts (Trello or similar tools)
- Slides
- Saving the Model for Quick Demo
  - Save model file (< 2MB upload or share cloud link if larger)
- Used Creative Presentation Techniques
- Use Generative AI, animations, effects, tools like Prezi, etc.

- Code Walkthrough
- Presentation Skills
  - Includes time management
- Discussion / Q&A
- Demo
- Visualization
  - Includes exploratory analysis (e.g., heat maps, other visuals)
- Version Control
  - Use of Git/GitHub or equivalent
  - Repository must be publicly accessible
- Lessons Learned
  - Included in the presentation
- Teamwork

Thank you for listening!

Q&A