Analyzing and Predicting Water Quality in Israel's Streams Using Machine Learning

Faculty of Engineering, Ariel University

Department of Industrial Engineering

Written by: Shiraz Hemo and Daniel Yesharim

Lecturer: Mr. Chen Hajaj

Date: 13/01/2025

https://github.com/shiraz3389/water-quality-analysis_project

Introduction:

Water quality is essential for public health, the environment, and the economy, especially in a country like Israel, where freshwater resources are limited. Streams play a key role in replenishing groundwater and supporting ecosystems, but they are under constant threat from pollution caused by human activity and climate change. Contaminated water can harm people, damage ecosystems, and disrupt industries like agriculture. Monitoring water quality and understanding its patterns over time are critical for maintaining ecological balance and ensuring safe water for consumption. This project focuses on analyzing water samples, with an emphasis on chemical parameters such as chloride, nitrate, and sulfate, to assess compliance with environmental standards and identify key factors influencing water quality.

Project Objectives

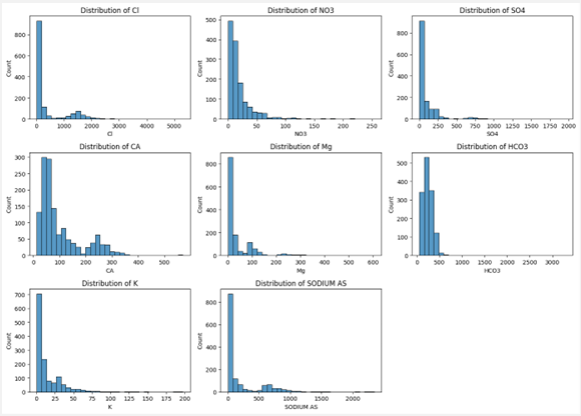
This project aims to analyze water quality in streams across Israel using a dataset containing chemical parameters, sampling station details, and compliance information. The objectives of this study are as follows:

1. Evaluate Water Quality Trends: Analyze the chemical composition of water samples over time to identify patterns, seasonal variations, and long-term trends.
2. Assess Compliance: Determine station compliance with predefined water quality standards to pinpoint stations with potential issues.
3. Predict Station Compliance: Build predictive models to classify stations as compliant or non-compliant based on chemical parameters. This will aid in proactively identifying stations that may require intervention.
4. Cluster Analysis: Group sampling stations based on water quality characteristics to uncover hidden patterns and categorize stations with similar profiles.
5. Provide Actionable Insights: Identify specific parameters contributing to non-compliance and recommend areas for further investigation or remediation.

By combining exploratory data analysis, machine learning, and clustering techniques, this project aims to provide a comprehensive understanding of water quality across Israel's streams.

Dataset and Features :

The dataset used in this project, titled "Water Quality Sampling Stations," contains detailed data collected from streams across Israel. It consists of 48,798 rows and 11 columns, offering a comprehensive view of water quality parameters, station information, and sampling metadata. The dataset includes critical chemical parameters such as chloride (Cl), nitrate (NO3), sulfate (SO4), calcium (CA), bicarbonate (HCO3), magnesium (Mg), potassium (K), and sodium (SODIUM AS). Additionally, it provides sampling-related metadata, including station names and sampling dates, making it a robust foundation for analyzing water quality.

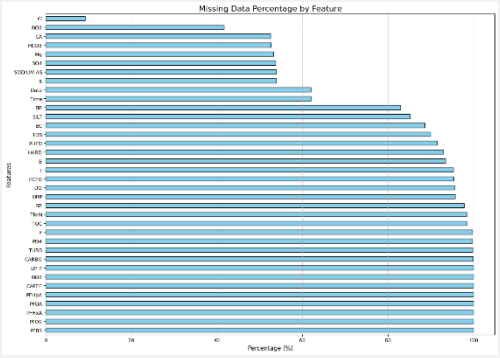


The distributions of the key parameters (e.g., Cl, NO3, SO4, etc.) show that most values are concentrated in a smaller range, with a few outliers in higher ranges. This suggests that while the majority of stations meet standard levels, some stations may have extreme values, possibly due to being non-compliant with water quality .standards

Data Preprocessing- Handling Missing Data

The dataset initially contained several missing values, particularly in parameters like PFHpA, PFBS, TURB, and others. Missing values were handled as follows:

- Columns with over 80% missing values were dropped to retain the integrity of the analysis.



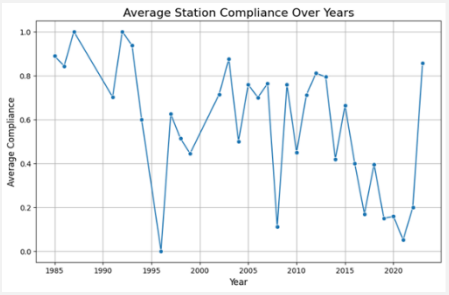
This bar chart visualizes the percentage of missing data for each feature in the dataset. It highlights that some features, such as PFBS, PFOS, and others, have nearly 100% missing values, while key features like Cl and NO3 have significantly lower percentages, indicating their better usability for analysis.

- Missing values in key columns, such as סמל פרמטר, were filled using corresponding values from the "תאור מקוצר" column when appropriate.
- Rows with any remaining missing values after these steps were removed to ensure clean and reliable data for modeling.

Data Preprocessing: Data Transformation-

To structure the data more effectively, a pivot table was created. This transformation organized the data by station ID, station name, sampling date, and parameter values, allowing for a more intuitive analysis. The "תאריך דגימה" (sampling date) column was also converted into a datetime format. Separate columns for the date and time were generated, enabling temporal analysis and providing insights into seasonal and annual trends in water quality.

The graph illustrates the average station compliance over the years



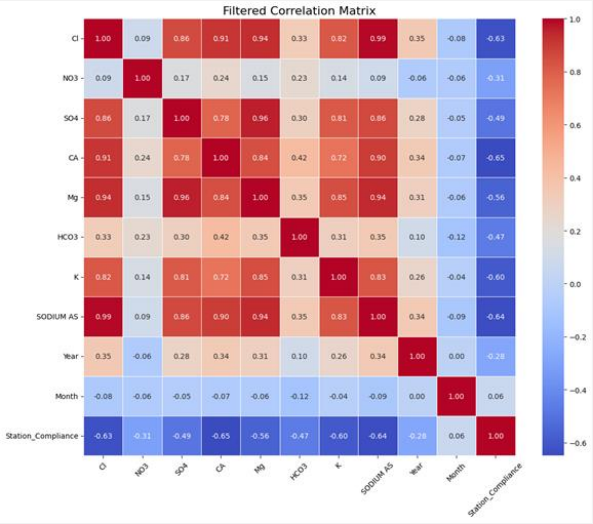
Adding a New Column: Station Compliance

We faced a challenge in the dataset due to the lack of a clear column for water quality compliance. To resolve this, we researched acceptable ranges for key chemical parameters from credible sources and created a new column, "Station Compliance," to indicate whether each station meets the predefined standards. These standards, based on research such as the "Analysis of Ground Water Quality Parameters," included thresholds like 0-250 mg/L for chloride and up to 50 mg/L for nitrate. This addition transformed the dataset into a supervised learning problem, enabling us to predict compliance and gain insights into the factors affecting water quality, all while adhering to scientific standards for better water resource management.

Feature Selection and Rationale- Selected Features-

- **Cl (Chloride)**: A key indicator of salinity and potential contamination.
- **NO3 (Nitrate)**: Crucial for assessing nutrient pollution.
- **SO4 (Sulfate)**: Indicates industrial and agricultural pollution.
- **CA (Calcium)** and **Mg (Magnesium)**: Important for determining water hardness.
- **HCO3 (Bicarbonate)**: Reflects alkalinity and buffering capacity.
- **K (Potassium)**: Linked to agricultural runoff and natural processes.

The correlation matrix revealed that parameters like Calcium (CA) and Potassium (K) have a moderate negative correlation with Station_Compliance, highlighting their impact on water quality. Features such as Chloride (Cl) and Sodium (SODIUM AS) were highly correlated and redundant, so we removed them to avoid multicollinearity and improve model performance. Additionally, the low correlations of Month and Year with Station_Compliance (-0.28 and -0.06) indicated that these temporal features don't significantly affect compliance, so they were excluded from the dataset to reduce noise and enhance model accuracy.



Rationale for Choices

1. **Parameter Selection**: Parameters were chosen based on their significance in determining water quality and compliance with environmental standards.
2. **Handling Missing Values**: Dropping highly sparse columns and rows with missing values ensured the reliability of downstream analyses.
3. **Feature Engineering**: Adding the Station Compliance target variable allowed for supervised learning and a clear evaluation of water quality compliance.

Methodology:

To effectively analyze and predict water quality compliance in Israel's streams, we adopted a comprehensive methodology combining supervised and unsupervised learning techniques alongside statistical analysis. This approach was guided by the characteristics of the dataset and the project objectives, ensuring both robust insights and practical applications.

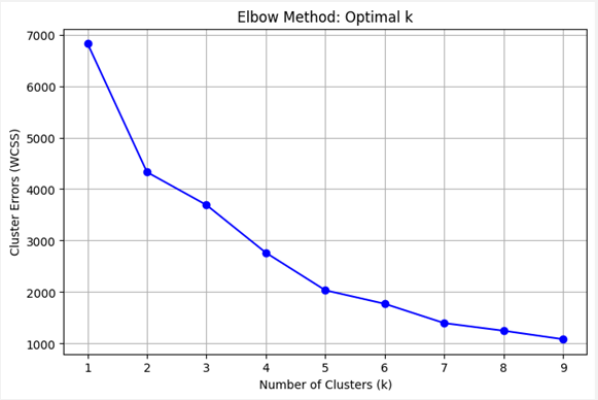
Supervised Learning-

We applied several supervised learning algorithms to predict water quality compliance, including Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, and Support Vector Machine (SVM). Logistic Regression was used as a benchmark model due to its simplicity and interpretability. Random Forest performed excellently with an accuracy of 97.44%, and Decision Tree also showed strong results with 96.34% accuracy. Gradient Boosting provided good performance, achieving 97.07% accuracy. SVM, however, had lower performance with an accuracy of 62.64%, indicating it wasn't well-suited for this dataset. Each model was evaluated using Accuracy, Precision, Recall, and

F1-Score, with Random Forest achieving the highest scores across all metrics, demonstrating its robustness for this problem.

Clustering Analysis-

Unsupervised learning techniques, specifically K-Means and Hierarchical Clustering, were used to identify patterns and group stations based on water quality characteristics. K-Means was chosen for its simplicity and ability to form well-defined clusters, while Hierarchical Clustering provided insights into the data's structure at different levels of granularity. Using the Elbow Method, we determined the optimal number of clusters to be three, balancing clarity and simplicity.



The graph shows how WCSS decreases as the number of clusters (k) increases, meaning the clusters better fit the data. The "elbow" point, around k=3 or k=4, marks where the rate of improvement slows significantly. This suggests that 3 or 4 clusters is the optimal choice, balancing accuracy and simplicity. Using this k will help us create meaningful clusters without overfitting the data.

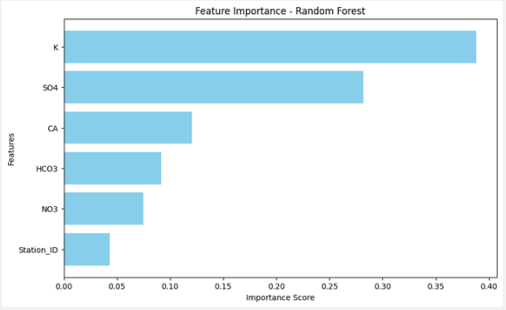
Feature Scaling and Dimensionality Reduction-

To ensure all features contributed equally to the analysis, we normalized the data using StandardScaler. This was particularly critical for clustering and distance-based models, such as SVM and K-Means. Principal Component Analysis (PCA) was employed to reduce dimensionality and visualize the clustering results effectively, helping to interpret the underlying patterns and compare clustering outcomes with true compliance labels.

Feature Importance and Compliance Evaluation-

Through models like Random Forest, we identified the most significant parameters impacting water quality compliance. Parameters such as Calcium (CA) and Potassium (K) emerged as highly influential, highlighting key areas for monitoring and intervention. Additionally, the creation of the Station_Compliance variable enabled a clear evaluation of station performance against predefined environmental standards, transforming the problem into a supervised learning task.

The feature importance plot shows that CA (Calcium) and K (Potassium) are the most significant features contributing to the model's predictions. Other features, such as NO3, SO4, and HCO3, have much lower importance, indicating they play a smaller role in determining station compliance.

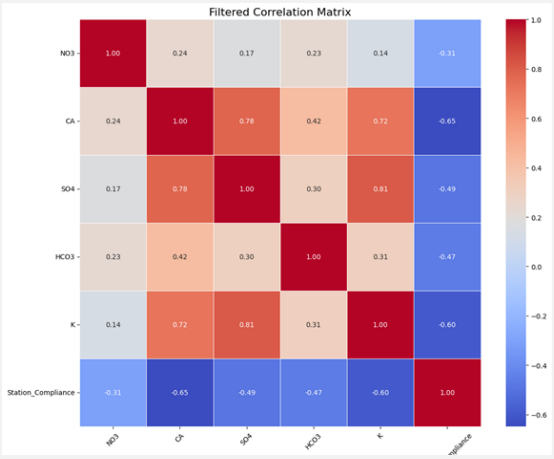


Key Methodological Choices-

Our methodology was shaped by the need for both robust predictive models and actionable insights. Supervised models allowed us to evaluate compliance directly, while clustering revealed hidden patterns in the dataset. Cross-validation ensured consistency and avoided overfitting, while feature engineering and scaling optimized the models' performance and interpretability.

Experiments and Results:

In this experiment, we evaluated the dataset using several supervised learning models. We chose parameters like chloride (Cl), nitrate (NO3), sulfate (SO4), calcium (CA), bicarbonate (HCO3), and potassium (K) for their importance in assessing water quality. All features were scaled using StandardScaler to avoid one feature dominating the results, especially for clustering and distance-based models.



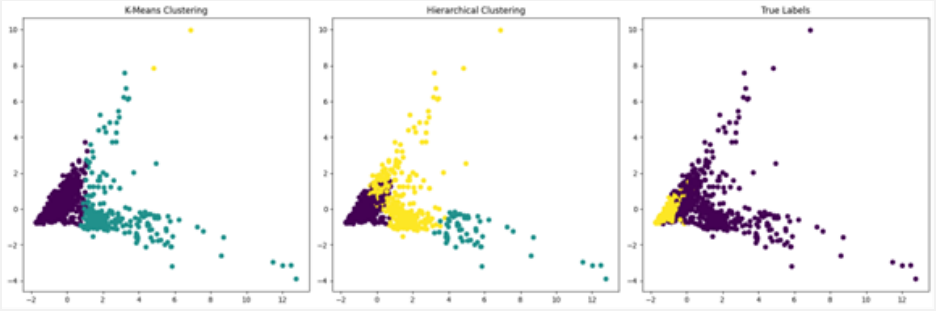
This heatmap will help illustrate the relationships between the features and their influence on Station_Compliance, highlighting key factors like calcium and potassium that affect compliance.

We used Accuracy, Precision, Recall, and F1-Score to evaluate the models. Accuracy gave an overall performance measure, while Precision and Recall helped us understand how well the models identified compliant or non-compliant stations. F1-Score balanced Precision and Recall, accounting for any data imbalances.

Findings and Analysis: The models showed varying results. Logistic Regression was used as a baseline with moderate accuracy. Random Forest (97.44%) and Decision Tree (96.34%) were the best performers, with Random Forest excelling due to its ability to handle non-linear relationships. Gradient Boosting also performed well (97.07%). SVM, however, had low accuracy (62.64%), likely due to its sensitivity to tuning and the dataset's characteristics.

	Model	Accuracy	Precision	Recall	F1-Score
0	Logistic Regression	0.926740	0.928230	0.926740	0.926176
1	Random Forest	0.974359	0.974417	0.974359	0.974375
2	Decision Tree	0.963370	0.963370	0.963370	0.963370
3	Gradient Boosting	0.970696	0.970727	0.970696	0.970658
4	SVM	0.626374	0.772389	0.626374	0.520513

Clustering Analysis using K-Means and Hierarchical Clustering revealed distinct patterns in water quality, identifying groups of stations with similar characteristics. Using the Elbow Method, we found that three clusters were optimal, offering a balance between simplicity and meaningful grouping. K-Means formed more compact clusters, while Hierarchical Clustering offered deeper insights into the data's hierarchical structure. These clusters could help identify regions with common water quality issues, offering a starting point for targeted monitoring and intervention.



The plots show the results of clustering the data using K-Means (left), Hierarchical Clustering (middle), and the true labels (right). While K-Means and Hierarchical Clustering both identified distinct groups, there are differences in the clustering results, with K-Means showing more compact clusters compared to the hierarchical method, highlighting the varying ways these algorithms group the data.

Experimental Successes and Limitations: The experiment successfully identified features influencing water quality compliance, such as calcium (CA) and potassium (K). The creation of the Station_Compliance variable helped make the dataset suitable for supervised learning. However, SVM's poor performance suggested the need for better tuning. While clustering revealed distinct groups, they didn't perfectly align with compliance labels, showing room for improvement.

Algorithm Performance: Random Forest and Gradient Boosting performed best due to their handling of non-linear relationships, while Decision Tree had some risk of overfitting. Logistic Regression worked well for basic analysis but couldn't capture complex patterns. SVM struggled with this dataset due to the high-dimensional data.

In conclusion, the experiments provided insights into water quality compliance, showing the strengths of ensemble models and clustering for pattern discovery. Future work could focus on refining feature engineering and exploring advanced techniques to address limitations.

Conclusion and Discussion:

This project aimed to analyze water quality across Israel's streams using machine learning to evaluate compliance with environmental standards. Despite challenges, particularly with the lack of a clear compliance column, we decided to work with the dataset and used both supervised and unsupervised learning to uncover key patterns and identify non-compliant stations. We emphasized integrating domain knowledge, like predefined parameter ranges, to drive meaningful results.

We tackled issues such as missing data and created the "Station_Compliance" variable to transform the dataset into a supervised learning problem. Algorithms like Random Forest and Gradient Boosting showed the power of ensemble methods in achieving high accuracy, while clustering and PCA visualization helped identify compliance patterns and regional disparities. More detailed location data, including accurate geographic coordinates for each station, would have added another layer to our analysis and helped better understand regional factors. However, we focused mainly on chemical parameters.

Contributions of Team Members:

- **Shiraz Hemo:** Managed dataset preprocessing, feature engineering, and clustering analysis, ensuring clean and structured data. Shiraz also created visualizations and interpreted clustering results.
- **Daniel Yesharim:** Implemented and evaluated supervised models, including Logistic Regression, Random Forest, Gradient Boosting, and SVM. Daniel also analyzed performance and prepared evaluation metrics.

Future Directions:

Future work could explore enhanced feature engineering by adding external data sources like rainfall or land-use patterns. Additional clustering methods, such as DBSCAN or Gaussian Mixture Models, could reveal more complex patterns. Improving model interpretability with SHAP or LIME would clarify predictions for non-compliant stations. Deploying the models in a real-time monitoring system would also offer dynamic insights for timely interventions. Furthermore, incorporating precise geographic location data for each station would help refine the analysis, offering insights into spatial patterns of water quality compliance.

Although we initially hesitated to use this dataset due to the lack of a clear compliance column, we adjusted and made it suitable for analysis. This project provided valuable insights into working with complex datasets and applying machine learning to real-world problems, laying a solid foundation for future efforts in water quality monitoring and environmental protection.