# Water Quality Improvement Project in California

**Final Report**

**Jesus Suarez**

## Problem Statement

Water quality varies statewide in California. To reduce the impact of poor water purity on the state population by 20% in the next two years, it is crucial to understand the chemical and physical properties of water and their correlations.

## Scope of Solution Space

This project aims to:

- Correlate the chemical and physical properties of water (such as Electrical Conductance, Dissolved Oxygen, Temperature, and pH).
- Establish relationships, specifications, and critical properties of water quality.
- Identify key parameters and counties with significant water quality .

## Introduction

Water quality is a fundamental aspect of public health and environmental sustainability, particularly in a diverse state like California, where variations in water purity can have significant implications. Dissolved Oxygen (DO) levels are a key indicator of water quality, influencing both aquatic life and overall ecosystem health. To address the challenge of poor water quality and reduce its impact on the state population by 20% over the next two years, it is essential to understand the chemical and physical properties of water and their correlations.

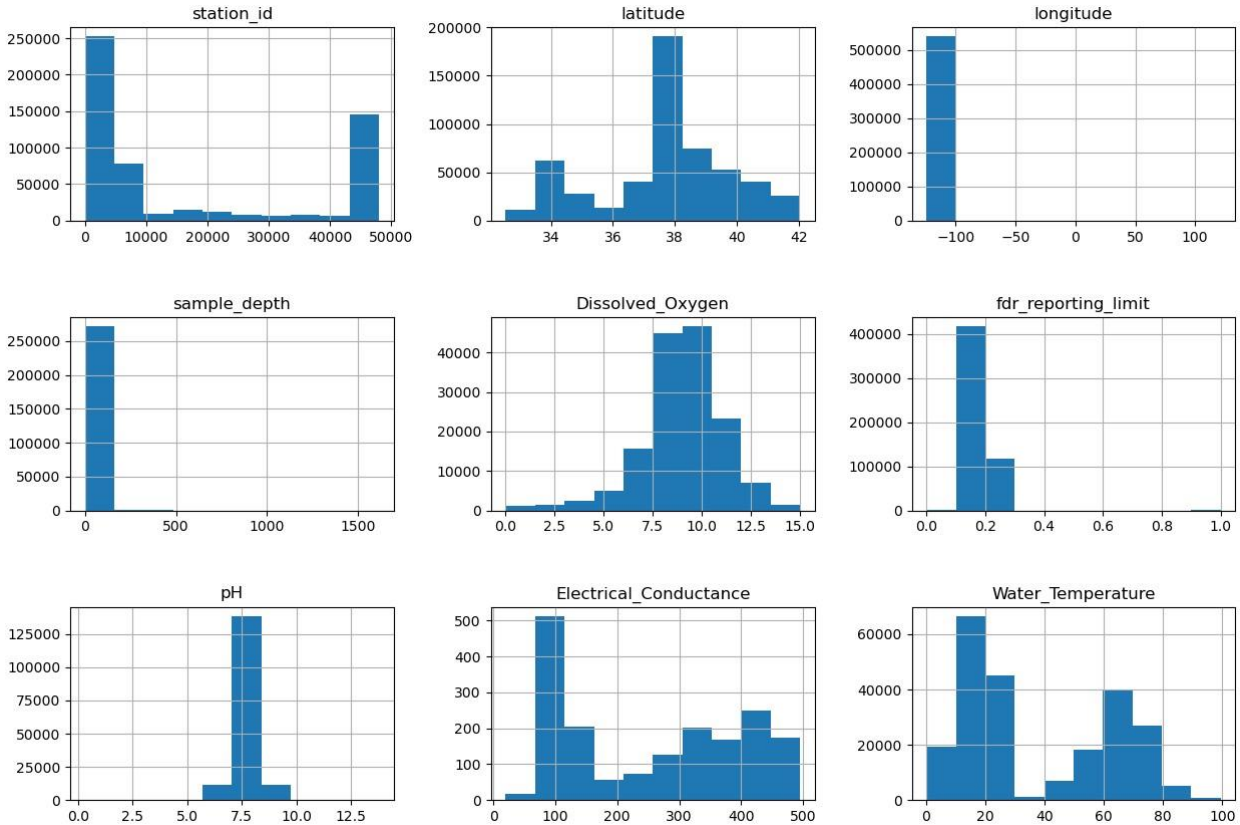
## Data Wrangling

The data wrangling process involved several steps to clean, transform, and prepare the data for analysis. The main dataset used in this project was the field data result of the state of California

Water Quality Analysis, which contained various measurements from different stations. First, an initial exploration of the data set was performed to understand its structure and contents. The dataset contains 1,048,575 entries with 22 columns, including station information, sample details, and field results for various parameters like Dissolved Oxygen, pH, Electrical Conductance, and Water Temperature. Some columns had missing values which required handling during the data cleaning process. To focus on key parameters, the dataset was filtered for Dissolved Oxygen, pH, Electrical Conductance, and Water Temperature. Specific criteria were applied to remove outliers and erroneous data points.

- Dissolved Oxygen: Filtered to include values greater than 0 and less than 15.
- pH: Filtered to include values greater than 0 and less than 14.
- Electrical Conductance: Filtered to include values greater than 0 and less than 500.
- Water Temperature: Filtered to include values greater than 0 and less than 100.

Several visualizations were created to explore and understand the data distributions and relationships.



The cleaned and transformed data was consolidated for further analysis and modeling. This comprehensive data wrangling process ensured that the data was clean, consistent, and ready for subsequent analysis, providing a solid foundation for the project's insights and conclusions.

# Exploratory Data Analysis (EDA)

This EDA provides an initial understanding of the dataset, highlighting important water quality parameters and their relationships.

- The mean dissolved oxygen across counties ranges from approximately 6 to 12 mg/L, and its distribution is approximately normal.
- The mean pH across counties varies from 7 to 9, and its distribution is slightly right skewed.
- The mean electrical conductance across counties ranges widely, and its distribution is right-skewed, indicating some outliers.
- The mean water temperature across counties ranges from approximately 15°C to 25°C, and its distribution is approximately normal.

The correlation analysis from the heatmap chart (Figure I) shows correlations between different water quality parameters. Dissolved Oxygen and pH illustrate a moderate positive correlation. Similarly, pH and Water Temperature show a weak positive correlation. Finally, Electrical Conductance doesn't show a clear correlation with other parameters.

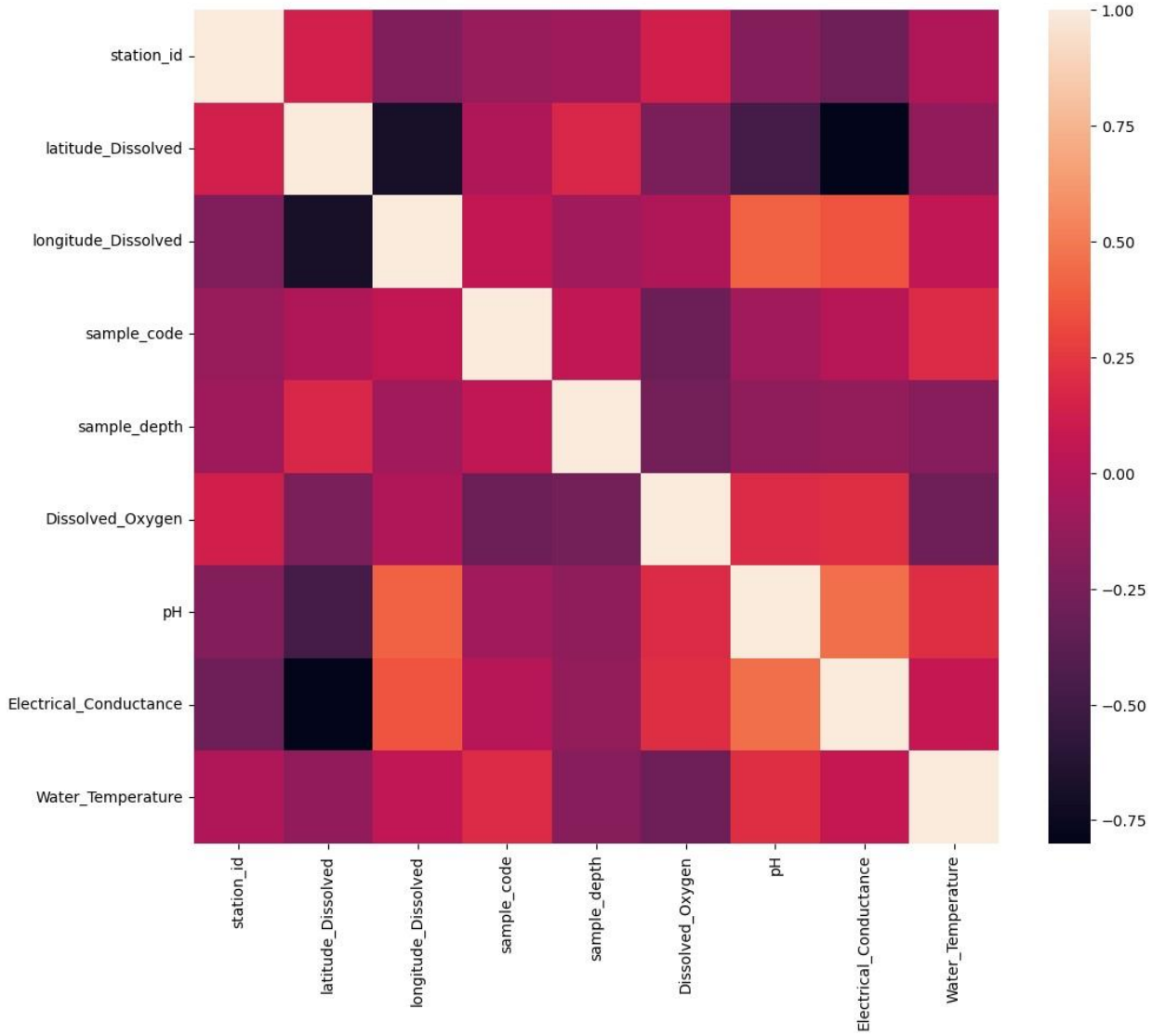



**Figure I:** Heatmap Chart

# Preprocessing and Training Data

Data preprocessing is essential to ensure the dataset is clean, consistent, and suitable for training machine learning models. The data was split into training and testing sets using a 70-30 ratio. The target variable for this analysis was Dissolved Oxygen in water. The Dummy Regressor confirmed that predicting the mean Dissolved Oxygen yields poor results, with $R^2$ close to 0. Both mean and median imputations resulted in a noticeable improvement over the baseline, with $R^2$ scores indicating that the model explains about 34-38% of the variance in the target variable. The importance of features in predicting Dissolved Oxygen can be inferred from the model

coefficients, though this analysis was not explicitly covered in the outputs. The Linear Regression model provided a reasonable starting point for predicting Dissolved Oxygen levels, significantly outperforming the baseline model. Further steps could involve trying more sophisticated models like Random Forest Regressor or Linear Regression, performing hyperparameter tuning, and exploring additional feature engineering techniques.

# Modeling

The modeling phase involved selecting and evaluating different regression models to predict the Dissolved Oxygen level in water samples. The modeling process began with a baseline model and progressed to more complex models. The following models were considered:

1. Dummy Regressor
2. Linear Regression
3. Random Forest Regressor

The Dummy Regressor served as our baseline model, predicting the meaning of the training target values for all instances. A Linear Regression model was employed using a pipeline to handle missing values and feature scaling. In addition, a Random Forest Regressor model was explored, which is a more complex ensemble model. The table below summarizes the performance of the models.

| Model | Training MAE | Test MAE | Training MSE | Test MSE | Training $R^2$ | Test $R^2$ |
|---|---|---|---|---|---|---|
| Dummy Regressor | 1.48 | 1.47 | 4.26 | 4.12 | 0.00 | -0.004 |
| Linear Regression | 1.14 | 1.11 | 2.63 | 2.70 | 0.38 | 0.34 |
| Random Forest | 0.40 | 1.00 | 0.29 | 2.09 | 0.97 | 0.51 |

The Random Forest Regressor outperformed the Dummy Regressor and Linear Regression models, showing the highest $R^2$ and lowest error metrics. This model is recommended for predicting Dissolved Oxygen levels based on the given features.

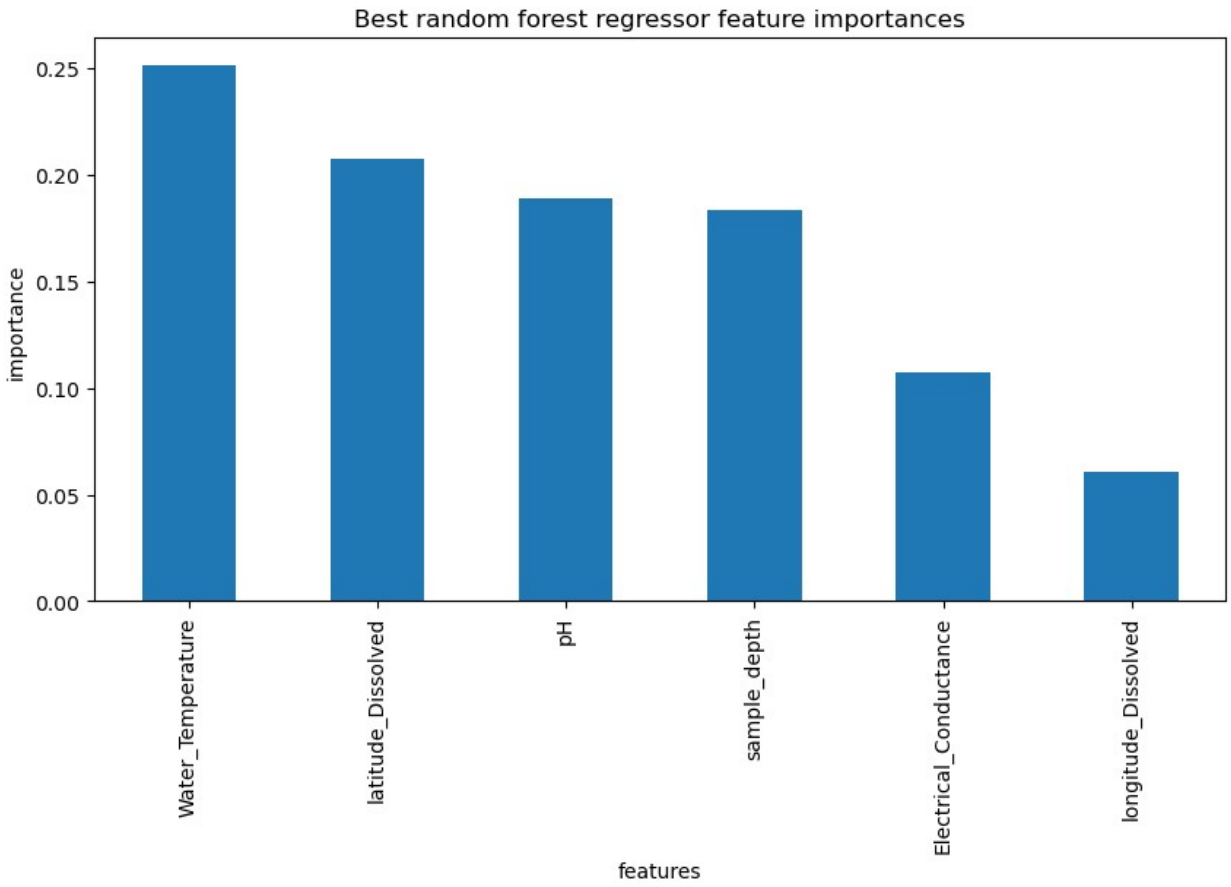



**Figure II:** Importance Chart for Forest Regressor Model

The feature importance chart reveals that Electrical Conductance, Water Temperature, and pH are the top three influential features for predicting Dissolved Oxygen levels. These features are crucial in determining the model's predictions, while geographical factors like latitude and longitude also contribute but to a lesser extent. The sample depth appears to have minimal impact on the model's output. This analysis can help in refining the model by potentially focusing on the most impactful features or further investigating why certain features are more predictive than others.

# Conclusion

This report highlights the critical steps undertaken to understand and improve water quality across California, focusing on predicting Dissolved Oxygen (DO) levels using various chemical

and physical water parameters. Through comprehensive data wrangling, exploratory data analysis, and advanced modeling techniques, we have identified key factors that influence water quality and proposed actionable insights to address the identified issues. The analysis demonstrated that Dissolved Oxygen levels are significantly influenced by Electrical Conductance, Water Temperature, and pH. The Random Forest Regressor model emerged as the most effective in predicting DO levels, outperforming simpler models like the Dummy Regressor and Linear Regression. This model's higher $R^2$ and lower error metrics underscore its reliability and accuracy in forecasting DO levels based on the given features. Future steps should include further refinement of predictive models, exploration of additional features, and targeted interventions in counties with the most severe water quality issues. This approach will contribute significantly to achieving the goal of reducing the impact of poor water purity by 20% within the next two years.