KKU SCIENCE JOURNAL
Journal Home Page : https://ph01.tci-thaijo.org/index.php/KKUSciJ
Published by the Faculty of Science, Khon Kaen University, Thailand

การติดตามและทำนายการทดสอบทางเคมีของประเภทน้ำ: ระบบ LIMS

Monitoring and Predicting Chemical Testing of Water Types: LIMS System

สมศรี จารุผดุง[1*] วรารัตน์ จักรหวัด[1*] โสรยา สิงสาโร[2] และ ผุสดี มุหะหมัด[2]

Somsri Jarupadung[1*], Wararat Jakawat[1*], Soraya Singsaro[2] and Phusadee Muhamud[2]

[1]สาขาวิทยาศาสตร์การคำนวณ คณะวิทยาศาสตร์ มหาวิทยาลัยสงขลานครินทร์ จังหวัดสงขลา 90110

[2]ศูนย์บริการตรวจสอบและรับรองมาตรฐาน คณะวิทยาศาสตร์ มหาวิทยาลัยสงขลานครินทร์ จังหวัดสงขลา 90110

[1]Division of Computational Science, Faculty of Science, Prince of Songkla University, Songkhla, 90110, Thailand

[2]Standard Inspection and Certification Service Center, Faculty of Science, Prince of Songkla University, Songkhla, 90110, Thailand

บทคัดย่อ

การจัดการข้อมูลมีบทบาทสำคัญในระบบจัดการข้อมูลห้องปฏิบัติการ (Laboratory Information Management Systems: LIMS) ซึ่งมีบทบาทในการจัดเก็บข้อมูลลูกค้าและสนับสนุนกระบวนการวิเคราะห์ทางเคมีของตัวอย่างน้ำ อย่างไรก็ตาม การบริหารจัดการกระบวนการทดสอบสารเคมีอย่างมีประสิทธิภาพนั้นยังคงเผชิญกับความท้าทายหลายประการ โดยเฉพาะอย่างยิ่งด้านค่าใช้จ่าย การจัดสรรทรัพยากร และการจัดการของเสีย แดชบอร์ดที่ออกแบบอย่างเหมาะสมสามารถช่วยเพิ่มประสิทธิภาพของกระบวนการใช้สารเคมีและการวางแผนการใช้ทรัพยากรอย่างเหมาะสม งานวิจัยนี้นำเสนอข้อมูลในรูปแบบของแดชบอร์ดที่ประกอบด้วยการวิเคราะห์ข้อมูลเชิงประวัติศาสตร์ควบคู่กับการทำนายข้อมูล โดยมีวัตถุประสงค์หลักเพื่อวิเคราะห์และแสดงข้อมูลประวัติการทดสอบตัวอย่างน้ำ อันเป็นพื้นฐานสำหรับการพัฒนาแบบจำลองเชิงพยากรณ์ด้วยเทคนิค Gradient Boosted Regression กระบวนการวิจัยเริ่มจากการรวบรวมข้อมูลจากระบบ LIMS ของศูนย์เครื่องมือวิทยาศาสตร์และรับรองมาตรฐาน คณะวิทยาศาสตร์ มหาวิทยาลัยสงขลานครินทร์ จากนั้นจึงดำเนินการเตรียมข้อมูล ซึ่งครอบคลุมถึงการทำความสะอาดและจัดโครงสร้างข้อมูลให้เหมาะสม ก่อนนำไปใช้ในการสร้างแบบจำลองและนำเสนอผลลัพธ์ในรูปแบบภาพข้อมูลเชิงวิเคราะห์ (data visualization) ที่สะท้อนทั้งข้อมูลจริงและข้อมูลที่ได้จากการพยากรณ์ โดยโมเดลดังกล่าวมีค่าความแม่นยำที่วัดจาก R-squared เท่ากับ 0.82 การศึกษานี้มีส่วนช่วยในการพัฒนาความแม่นยำของการคาดการณ์ปริมาณตัวอย่างน้ำที่ส่งเข้าทดสอบ ซึ่งนำไปสู่การจัดสรรทรัพยากรอย่างมีประสิทธิภาพ และสามารถลดการสูญเสียจากการสั่งซื้อสารเคมีที่มากเกินความจำเป็น หรือน้อยจนไม่เพียงพอกับความต้องการ ทั้งนี้แนวทางดังกล่าวยังส่งผลเชิงบวกต่อการลดต้นทุนและผลกระทบต่อสิ่งแวดล้อม อันเป็นการยกระดับคุณภาพของการให้บริการด้านการทดสอบคุณภาพน้ำในภาพรวม

*Corresponding Author, E-mail: somsri.ja@psu.ac.th, wararat.j@psu.ac.th

Received date: 2 April 2025 | Revised date: 4 June 2025 | Accepted date: 20 June 2025

doi: 10.14456/kkuscij.2025.22

ABSTRACT

Data management plays a crucial role in Laboratory Information Management Systems, which are used to collect customer information and conduct chemical tests for water samples. However, successfully managing chemical testing processes presents obstacles, particularly in terms of financial expenditures, resource allocation, and waste management. A well-designed dashboard can help optimize chemical testing and improve resource planning. This study presents data in the form of dashboards that feature both historical and predictive maintenance observations. The primary goal is to evaluate and show historical water sample data, which will serve as the foundation for developing a prediction model using Gradient Boosted Regression. The study process consists of collecting data from the water sample testing service from Laboratory Information Management System (LIMS), Center of Measurement and Standard Accreditation, Faculty of Science, Prince of Songkla University, followed by data processing, which involves data cleaning. This cleaned data is then used to create a model and present the visualization from both the predictions and the actual data. The R-squared score from this model is 0.82. By analyzing and forecasting the volume of water samples, the study helps improve the accuracy of resource allocation, allowing for more efficient planning and management of chemical testing operations. Ultimately, this approach contributes to reducing waste and minimizing unnecessary expenses, optimizing both financial and environmental outcomes in water quality testing.

คำสำคัญ: การแสดงข้อมูลด้วยภาพ แดชบอร์ด การจัดการข้อมูล
Keywords: Data Visualization, Dashboard, Data Management

INTRODUCTION

Data plays an important role in business operations by enabling organizations to extract insights from historical records and align them with strategic objectives. The water quality analysis, understanding customer behavior and sample characteristics is essential for planning chemical use efficiently. Since the types of chemicals vary by water sample, improper stock management can hinder operations. Overstocking can lead to chemical degradation over time, while understocking risks insufficient supply for operational needs. Predictive modeling offers a solution by forecasting sample types and volumes based on historical trends. This helps laboratories prepare the right amount of chemicals while also identifying customer usage patterns for targeted marketing and retention efforts.

This research utilizes data from Center of Measurement and Standard Accreditation, Faculty of Science, Prince of Songkla University, which provides analysis and testing services to industrial sectors, external government agencies, independent local and provincial organizations, and other public entities. The center operates under a Laboratory Information Management System (LIMS) which has been developed by them. LIMS supports the management of data in scientific laboratories. At the center, it is used for the water quality testing system for customer-submitted samples. All relevant data is recorded and stored in the system. The data is used to develop a dashboard for analysis and visualization, leveraging water sample data categorized by time and customer. The study employs the Gradient Boosting Regression technique to

construct a predictive model for water quality testing. Predictive modeling in the water industry has increasingly relied on data-driven methodologies to support proactive decision-making, particularly in water quality monitoring and forecasting. The integration of machine learning (ML) and big data analytics has demonstrated considerable potential in enhancing prediction accuracy, reducing operational costs, and addressing the limitations of traditional statistical or mechanistic models.

Several studies have leveraged diverse datasets for modeling water quality parameters, including sensor-based real-time data, such as pH, dissolved oxygen, turbidity, and chlorine levels (Wei *et al*., 2024), meteorological and hydrological information such as rainfall, temperature and river discharge (Yang, 2023; Wang *et al*., 2016), geospatial data on land use and industrial activity (Nair and Vijaya, 2021), and historical laboratory records similar to our LIMS-based dataset (Khan and Chai, 2016). These datasets are typically characterized by missing values, seasonal fluctuations, and temporal inconsistencies, all of which require extensive preprocessing and transformation. Addressing these challenges through comprehensive preprocessing and transformation represents a significant focus of our current study.

Various machine learning models have been applied in this domain. Tree-based models like Random Forest and Gradient Boosting Regression have demonstrated robustness to missing values and the ability to capture nonlinear relationships (Li *et al*., 2023; Wei *et al*., 2024). Neural networks, particularly Artificial Neural Networks (ANN) and Long Short-Term Memory (LSTM), are frequently employed for modeling time-dependent variables like dissolved oxygen or chlorine levels (Yang, 2023) while Adaptive Neuro-Fuzzy Inference Systems (ANFIS) combine neural learning with fuzzy logic to handle uncertainty in environmental data (Yang, 2023). Aligned with these trends, this study employs Gradient Boosting Regression, which effectively handles mixed-type tabular data and captures complex patterns in water quality variables.

The literature also highlights the importance of structured modeling workflows, especially data preprocessing and feature engineering, such as temporal decomposition, normalization, and regional categorization (Khan and Chai, 2016; Nair and Vijaya, 2021), along with model validation methods like k-fold cross-validation to ensure generalizability and reduce overfitting (Li *et al*., 2023; Wei *et al*., 2024). For example, Wei *et al*. (2024) applied Random Forest and LSTM models to predict chlorine levels in drinking water and validated the models using real-time monitoring feedback. Their integration of model outputs into dashboards for decision support aligns with our use of Looker Studio for data visualization.

In terms of performance, machine learning models have achieved high predictive accuracy in water quality applications. Common metrics include Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2). Previous work by Khan and Chai (2016) reported over a 30% reduction in RMSE compared to traditional regression models, while Li *et al*. (2023) achieved R^2 values exceeding 0.90 in predicting parameters such as total suspended solids using Gradient Boosting Regression. These comparative benchmarks provide a reference point for assessing the effectiveness of our model, which is evaluated using similar metrics and tested on a geographically imbalanced dataset primarily composed of samples from the southern provinces of Thailand.

Although the aforementioned studies have explored various machine learning techniques in water quality contexts, few have explicitly addressed the challenges of laboratory-sourced, historical datasets or examined the implications of regional sampling imbalances. Our study contributes to the field by utilization of structured LIMS-derived historical data in place of real-time sensor data making the approach applicable in resource-limited settings. Geographic attributes are incorporated to reflect regional water quality variation, supporting more localized predictions. The model is designed using Gradient Boosting Regression and is further enhanced through interpretability tools such as feature importance analysis and deployment via interactive dashboards. These elements collectively support informed decision-making in laboratory operations and water quality management.

Collectively, these contributions position the present work as both complementary to and distinct from existing studies, by extending the applicability of machine learning techniques using historical laboratory data as the data source for water quality management. This research adopts the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology to guide the design and implementation of the predictive analytics and visualization system.

MATERIALS AND METHODS

The system development process is structured around a six-stage framework, depicted in Figure 1, and is elaborated in detail in the subsequent section.

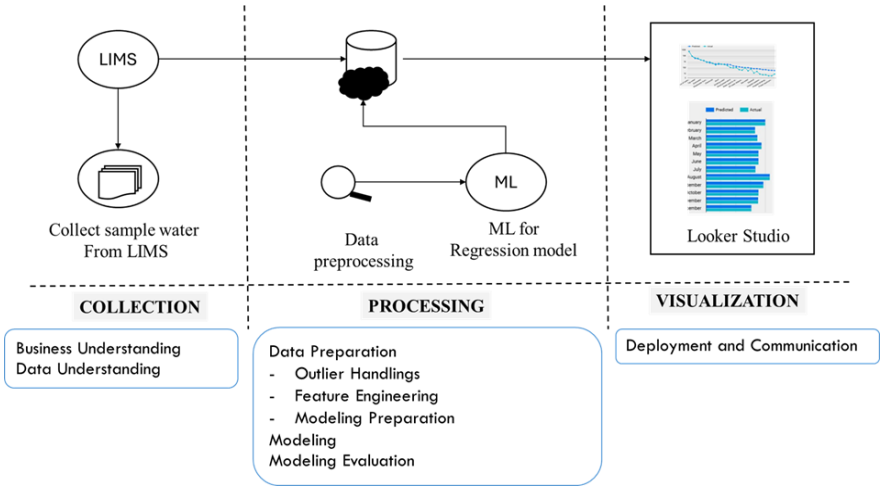

Figure 1 The proposed framework

1. Data Collection (Business Understanding and Data Understanding)

This study aims to enhance water quality monitoring and optimize chemical testing through a predictive system that identifies trends, detects potential issues at an early stage, and supports data-driven decision-making. It also offers insights into customer engagement with the service. Water sample data from LIMS includes information on water samples, consumers, locations, testing dates, and chemicals. Initial exploration was conducted to assess its structure, quality, and distribution. The dataset employed in this study was obtained from LIMS, which is used to monitor and manage water samples submitted for chemical

analysis. Prior to the implementation of LIMS, data were manually recorded using Microsoft Excel spreadsheets, resulting in inconsistencies due to human error and the absence of standardized formatting protocols. Based on these understanding, two key points can be summarized as data characteristics and data limitations.

The exported dataset includes approximately 10,000 records with five key features such as sample collection date, province, and sample type, organized in relational tables spanning 2013 to 2023, primarily covering southern provinces. During data exploration, issues such as inconsistent naming (e.g., “Songkla” vs. “Songkhla”), missing or incomplete data entries, and underrepresentation of certain provinces, leading to geographic imbalance. This stage is critical for understanding the business context and assessing data quality. It establishes the foundation for effective preprocessing, model development, and accurate interpretation of analytical results in the subsequent phases of the CRISP-DM framework.

2. Data Preparation

This phase involves data cleaning, transformation, and integration to ensure quality and consistency. Missing values are handled, formats standardized, and the structured dataset is prepared for machine learning input.

Data cleaning

Data cleaning is a critical step to ensure the accuracy, consistency, and reliability of the dataset. In this study, missing values in numerical fields were imputed using either the mean or median, depending on the distribution skewness, while categorical fields such as province were manually completed using domain knowledge. Inconsistent spellings were standardized through predefined mapping dictionaries (e.g., “Songkla” corrected to “Songkhla”) to maintain uniformity. Irrelevant attributes, including sample IDs and technician comments, were excluded to reduce noise and improve model efficiency. Temporal filtering was also applied by removing records outside the 2013 – 2023 period to ensure consistency. These preprocessing steps collectively enhanced data integrity and provided a solid foundation for effective feature engineering and model development.

Data Transformation (Outlier Handlings and Feature Engineering)

To facilitate analysis and modeling, the data was transformed into a suitable format through various techniques aimed at enhancing quality and extracting meaningful features. Regional data was derived from provincial information to support models that analyze sample submissions by geographic zone. Outlier detection, a critical step in ensuring the reliability of chemical testing, was conducted using both visual tools such as box plots and statistical methods like the 3-sigma rule. We establish a robust approach to identifying these unusual data points. A comparison between the box plots and 3-sigma methods revealed overlapping as well as distinct sets of outliers. While some observations were identified by both methods, others were uniquely flagged by the 3-sigma rule, demonstrating its sensitivity in detecting extreme values that might not fall outside using box plots. Continuous monitoring through LIMS helped maintain data integrity throughout the process. Feature engineering was also applied to enrich the dataset. Temporal features were created by decomposing date fields into components such as month and year, allowing for

seasonal pattern detection and trend analysis. Additionally, provinces were grouped into five regions to add spatial context. These transformations improved the dataset's analytical depth and supported more accurate exploratory and predictive modeling.

Data Splitting (Modeling Preparation)

To assess the performance of the model, the dataset is partitioned into distinct subsets in order to ensure a fair and unbiased evaluation of the predictive models. The dataset was partitioned into training and testing subsets. The training set, comprising 80% of the data, is utilized to train the machine learning regression model. The testing set, consisting of the remaining 20% of the dataset, is reserved for evaluating the model's accuracy. This division ensures a reliable assessment of the model's performance. This splitting approach ensures that the model's performance metrics reflect its generalizability to unseen data.

3. Model Creation (Modeling)

A regression model is employed due to its effectiveness in capturing numerical patterns and forecasting trends in water quality testing. The model is trained on historical data to predict future values, allowing stakeholders to anticipate potential risks or anomalies. In this study, we employed Gradient Boosting Regression, a technique that constructs an ensemble of decision trees in a sequential manner. Each tree is designed to correct the errors of its predecessor by minimizing a predefined loss function. This approach enables the model to improve predictive accuracy iteratively and effectively capture complex nonlinear relationships within the data. The system was implemented using Python 3 within the PyCharm development environment on the Windows operating system. To enhance the robustness and generalizability of the model, 4-fold cross-validation was employed. This procedure divides the dataset into four equal subsets. During each iteration, one subset was used for testing while the remaining three were used for training. This process was repeated four times, with each subset serving as the test set while the others are used for training. The results were averaged across all folds to provide a comprehensive and unbiased estimate of the model's predictive capability.

4. Model Evaluation (Evaluation)

The model's performance is evaluated using standard regression metrics, including Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), correlation coefficient (R), and the coefficient of determination (R^2). This evaluation ensures that the model aligns with the project objectives and delivers actionable insights. The model performance was assessed using several key metrics. The R^2 score indicates the proportion of variance in the dependent variable explained by the independent variables, with higher values signifying better model performance. The R score measures the strength and direction of the linear relationship between predicted and actual values, where values close to +1 indicate a strong positive correlation. Mean Absolute Error (MAE) represents the average absolute difference between predicted and actual values, offering a clear measure of overall prediction accuracy. Root Mean Squared Error (RMSE), which calculates the square root of the average squared differences between predicted and actual values, is more sensitive to large errors and highlights variance in predictive accuracy. These metrics collectively provide a robust basis for evaluating the accuracy, consistency, and generalizability of the predictive model.

5. Data Visualization (Deployment and Communication)

The final model is deployed via Looker Studio, Google's cloud-based visualization platform, enabling real-time dashboards and integration with multiple data sources. These visualizations support continuous monitoring, strategic planning, and resource allocation in water quality management. By following the CRISP-DM methodology, the study ensures a structured and practical approach to predictive analytics, enhancing both interpretability and the resulting system in real-world water quality management contexts. Data visualization plays a crucial role in identifying patterns, trends, and insights within datasets. Various visualization techniques are utilized to effectively represent both the data and the model's performance. In this study, Looker Studio was used to create interactive visualizations that enable dynamic exploration of the results. Time series plots illustrate temporal variations in water sample quantities, allowing for the identification of seasonal patterns and long-term trends. Scatter plots show the relationship between actual and predicted values, providing insight into the model's accuracy and performance. Feature importance plots highlight the variables that significantly influence water sample quantities, shedding light on the key drivers in the dataset.

Although the implementation process involves multiple complex steps, it presents a valuable challenge by utilizing real operational data to make predictions based on desired attributes. For example, preparing appropriate quantities of related chemicals based on the demand for each month or quarter of the year, in order to avoid overstocking that could lead to waste or understocking that could result in insufficient supply.

RESULTS AND DISCUSSION

The Gradient Boosting Regression model had a test set R^2 score of 0.82, suggesting strong predictive performance. The feature importance plot demonstrated that the most important parameters influencing water sample volumes were location and season (month and year). The model accurately predicted historical water sample volumes and recognized trends over time. In this study, we found that predicted values were higher than actual observations (Over-predictions) and predicted values were lower than actual observations (Under-predictions). Over-predictions were observed in water types with complex contamination patterns, such as frozen squid products and hospital drinking water. Under-predictions, in contrast, were more frequently associated with water types like tap water and drinking water in sealed containers.

We applied our dataset to four models: Linear Regression, Ridge Regression, Random Forest Regression, and Gradient Boosting Regression. R^2 are computed to assess model performance, as shown in Table 1. These regression techniques were chosen to assess the performance of various approaches, ranging from simple linear models to more complex ensemble methods, on this prediction task. Both Linear and Ridge Regression had low R^2 values of 0.338, meaning they explained only about 34% of the variance in the data. In contrast, Random Forest and Gradient Boosting achieved much higher R^2 values of 0.815 and 0.819, respectively, indicating that they explained over 81% of the variance.

The correlation coefficient for the ensemble models also indicates a strong positive relationship between the predicted and actual values (0.916 for Random Forest and 0.915 for Gradient Boosting), in contrast to the weaker correlations observed for Linear and Ridge Regression (both approximately 0.58).

Gradient Boosting Regression has the lowest RMSE (107.294), with Random Forest slightly behind at 108.467. This indicates that their predictions are much closer to the actual values compared to the linear models, which had RMSE values exceeding 205. In terms of MAE, Random Forest achieved the lowest MAE (30.139), followed by Gradient Boosting (31.745), whereas the linear models reported much higher MAE values (approximately 89).

This suggests that Random Forest and Gradient Boosting are far better at capturing complex patterns in the data. Their ability to model complex, nonlinear relationships makes them more suitable for datasets with intricate patterns. These results highlight the advantage of ensemble tree-based methods in regression tasks that go beyond the capacity of simple linear approaches.

This predictive feature can be used to optimize chemical testing resources and improve overall management planning in the Laboratory Information Management System. The following visualizations are provided to demonstrate these findings.

Table 1 Comparison for different models

Regression model	R^2	R	RMSE	MAE
Linear Regression	0.338	0.582	205.587	89.462
Ridge Regression	0.338	0.583	205.484	89.035
Random Forest Regression	0.815	0.916	108.467	30.139
Gradient Boosting Regression	0.819	0.915	107.294	31.745

The dashboard is shown in two pages, prediction and historical view. The prediction view (Figure 2a), the data can be filtered by year, region, province and water sample. The dashboard in historical view is demonstrated in Figure 2b. The data on both pages can be filtered by clicking directly on the graph, as shown in Figure 2b, which only shows groundwater sample data. Each graph shows the number of water sample data recorded from 2012 - 2023. The privileged people in the center can access the pages by login to their account.

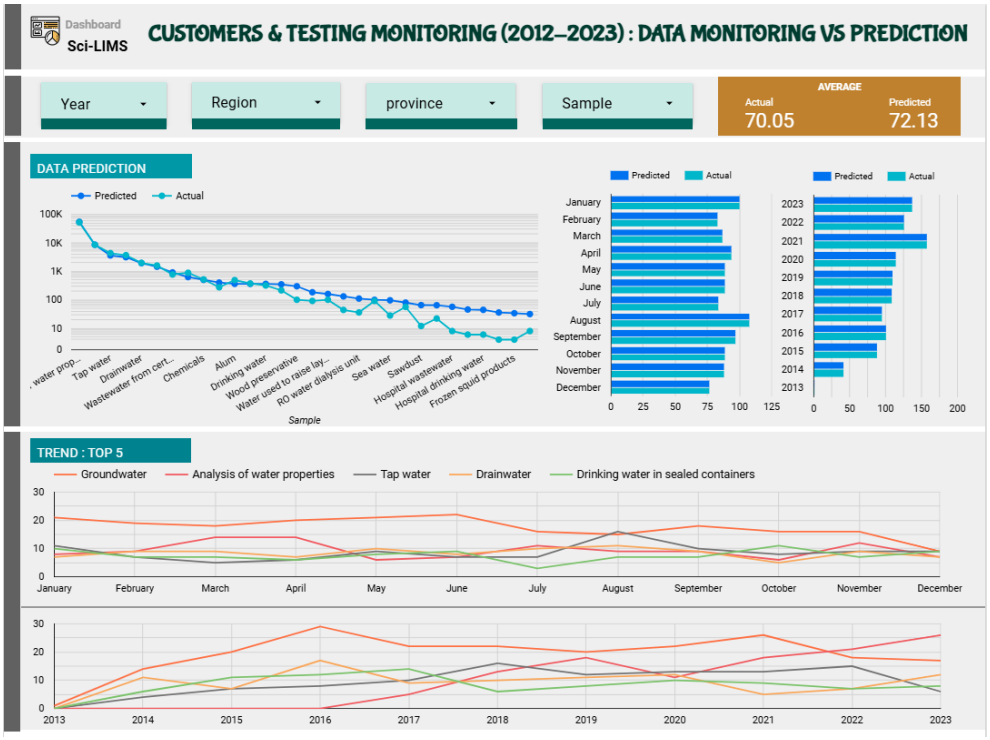

Figure 2a LIMS Dashboard Page 1 (Prediction View)

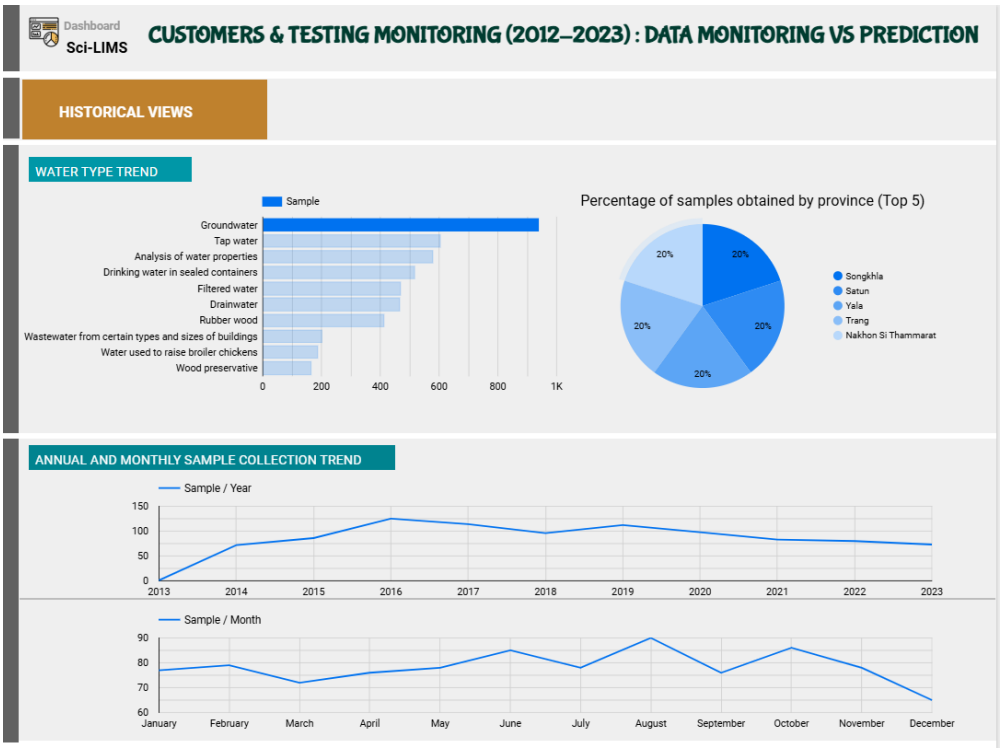

Figure 2b LIMS Dashboard Page 2 (Historical View)

Prediction View

The following chart provides a visual comparison between the predicted and actual values across different types of water samples, showcasing the performance and accuracy of the predictive model.

The figure compares projected and actual values for several water samples, showing key patterns and inconsistencies (Figure 3). Predicted values are generally lower than actual values, especially in high-point samples like tap water and drinking water in sealed containers, implying that the model may be underestimating. In mid-range samples, such as chemicals, projected values are more closely aligned with actual readings, indicating greater accuracy. In contrast, in low-point samples, such as hospital drinking water and frozen squid items, real levels are sometimes lower than expected, indicating that the model have overestimated.

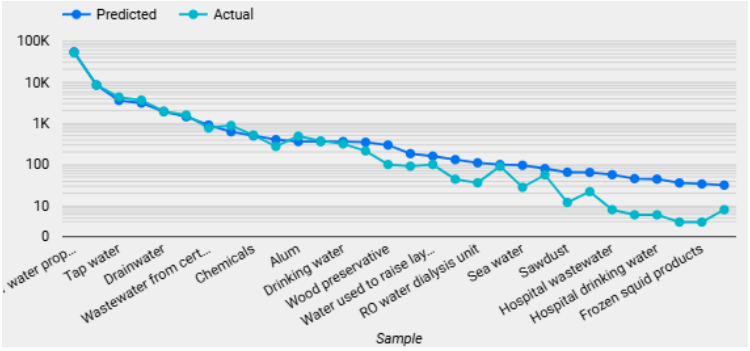

Figure 3 Analysis of Predicted vs Actual Data for Water Samples

Figure 4 displays two comparative bar charts that indicate anticipated actual values across two time periods: months (January to December) on the left and years (2013 to 2023) on the right. Dark bars represent predicted values, whereas bright bars reveal actual values. From an analytical standpoint, the charts clearly show disparities between predicted and actual outcomes. The monthly figure shows probable patterns of overestimation or underestimation over specific time periods, which may reflect seasonal impacts or limitations in the model's sensitivity to monthly fluctuations. Similarly, the annual chart displays patterns over time, indicating potential improvements or consistent biases in the model's performance.

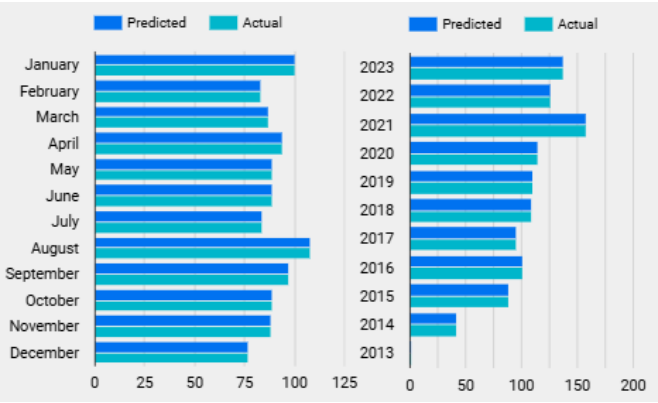

Figure 4 Comparison of Predicted vs Actual Values Across Time (Monthly and Yearly)

Figure 5 depicts a comparative study of data from the top five water types: groundwater, water property analysis, tap water, drainwater, and drinking water in sealed containers, as measured monthly from January to December. The given values most likely represent the volumes of each water type that the organization received from all clients each month. The figure shows that each water type has distinct patterns over time. For example, groundwater readings are rather steady over the first half of the year, with a considerable spike around midyear. Tap water and drinking water in sealed containers, on the other hand, tend to follow a more regular pattern over time. These variations may reflect seasonal causes or changes in water management systems.

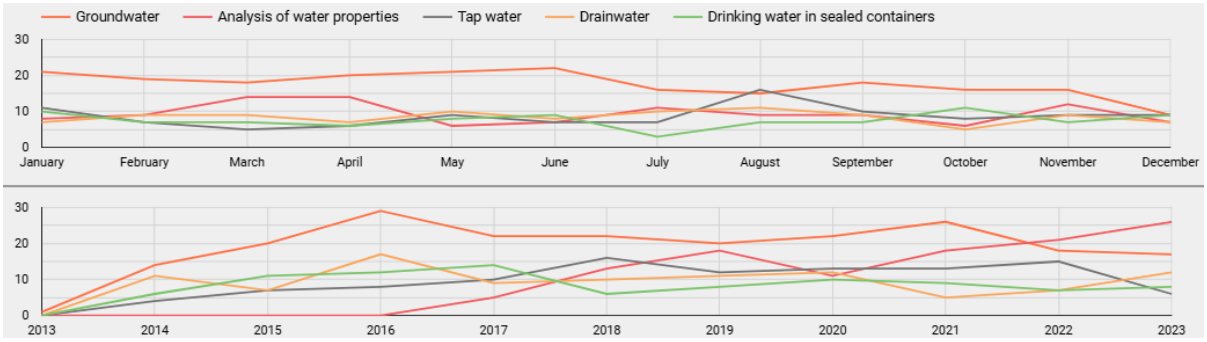

Figure 5 Comparative Analysis of Water Types Over a Year (Top 5)

However, the charts in Figures 4 and 5 show specific months of the year when a considerable volume of sample water is expected. This information allows staff to plan of time, ensuring that adequate testing materials and resources are available. These insights are critical for improving operational readiness and ensuring smooth testing operations during busy months. Furthermore, these insights provide useful suggestions for enhancing water resource management and improving future planning methods.

Historical View

The charts' historical view allows us to track changes and patterns over time in the number of water kinds. Figure 6 shows a bar chart of water types based on the number of samples taken between 2012 and 2023, with the top categories highlighted. Groundwater has the most samples (more than 1,000), followed by tap water, water property analysis, and drinking water in sealed containers, all of which have comparable sample numbers. Other categories, such as filtered water and drainwater, follow closely behind, while less common varieties, such as rubber wood water, wastewater, and wood preservatives, have much less samples. This graph depicts the distribution of sample efforts across various water types, emphasizing the dominance of specific categories and offering information on the focus areas for water study.

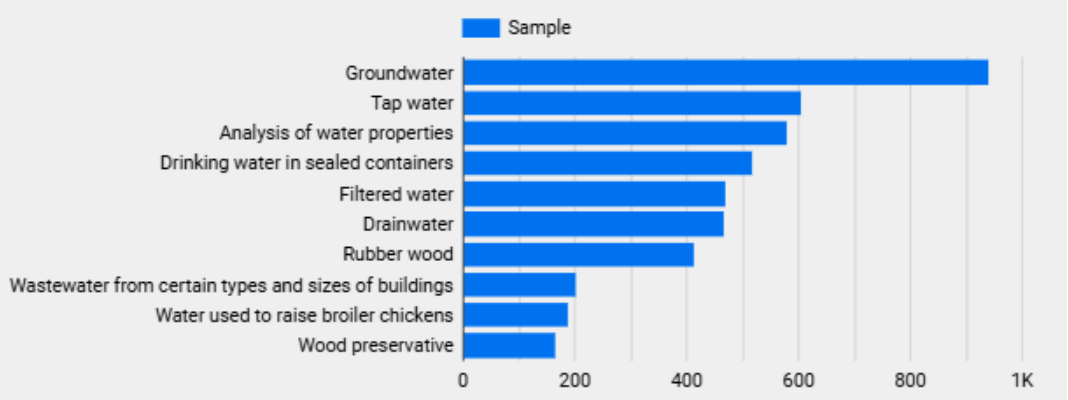

Figure 6 Sample different water types

The pie chart in Figure 7 depicts the percentage distribution of samples obtained from the top five provinces. Songkhla leads with 25.8%, followed by Nakhon Si Thammarat (20.4%), Bangkok (19.4%), Phatthalung (18.3%), and Trang (16.1%). Songkhla and Nakhon Si Thammarat account for 46.2% of the total samples, demonstrating the considerable contribution of the southern provinces.

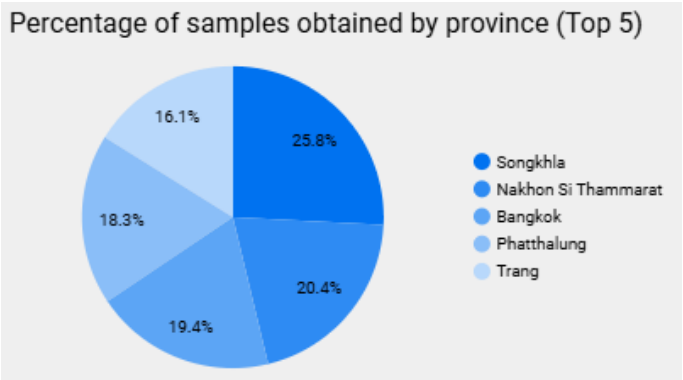

Figure 7 Percentage of Samples Obtained by Province (Top 5)

Figure 8 depicts annual (2012 - 2023) and monthly sample collection trends. The annual data show a continuous increase in sample collecting beginning in 2012, peaking in 2021, then slightly declining in subsequent years. The monthly statistics show moderate seasonal fluctuations, with more samples obtained between January and March and a noticeable reduction near the end of the year, particularly in December. These tendencies could be attributed to operational limitations, environmental conditions, or seasonal variations in sampling attempts.

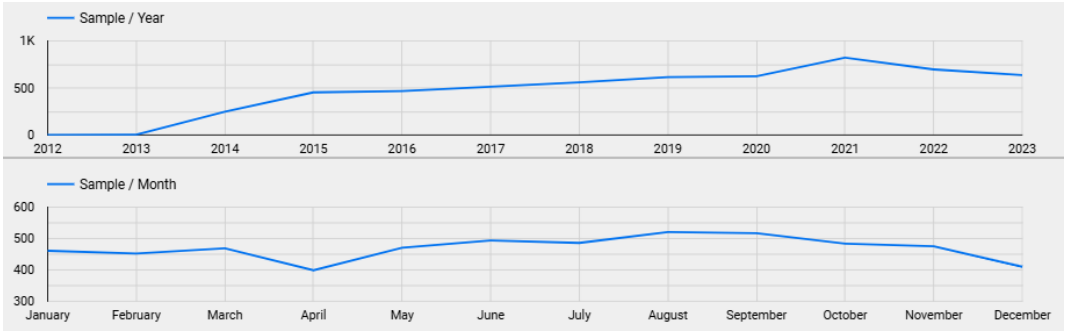

Figure 8 Sample Collection Trends by Year and Month (2012 - 2023)

Based on the insights presented above, it can be observed that the chemicals used for water quality analysis are mainly applied to the testing of groundwater and tap water. Seasonal trends also influence when water samples are submitted, with fluctuations varying by month and year. For instance, the number of groundwater samples remains relatively stable throughout the year, with noticeable peaks occurring around the middle of the year. In contrast, tap water and drinking water samples stored in sealed containers exhibit a consistent annual pattern.

These findings can assist laboratories in forecasting resource requirements and optimizing operational planning ahead of peak testing periods, thereby improving efficiency and ensuring timely processing of water samples. For example, the laboratory staff can prepare sufficient chemicals for groundwater quality analysis throughout the year by planning to increase chemical orders before the mid-year peak period. On the other hand, during the third quarter of the year, overstocking these chemicals should be avoided, as prolonged storage can lead to deterioration in quality, resulting in the need to dispose of the expired chemicals.

CONCLUSIONS

This study proposed a strategy for employing Gradient Boosted Regression to optimize chemical testing for all sorts of water and resource management within a Laboratory Information Management System. The goal is to produce good prediction performance using the Gradient Boosted Regression model, which has a test set R-squared score of 0.82 Location and season were shown to be the most important factors influencing water sample quantity. The prediction algorithm can properly forecast past water sample volumes and identify long-term trends. This data-driven strategy can help firms make better decisions about chemical testing resources and management planning within their LIMS systems.

Future work can focus on expanding the scope of prediction to other water quality parameters, larger datasets or integrating additional data sources such as weather forecasts to improve prediction accuracy that may better capture the complexities of water testing operations. In addition to cross-validation, the external validation is an essential step in assessing the model's generalizability.

REFERENCES

Wang, D., Singhasemanon, N. and Goh, K. (2016). A statistical assessment of pesticide pollution in surface waters using environmental monitoring data: Chlorpyrifos in Central Valley. Science of The Total Environment 571: 332 - 341. doi: 10.1016/j.scitotenv.2016.07.159.

Yang, J. (2023). Predicting water quality through daily concentration of dissolved oxygen using improved artificial intelligence. In Scientific Reports, 13. doi: 10.1038/s41598-023-47060-5.

Nair, J.P. and Vijaya, P.M.S. (2021). Predictive models for river water quality using machine learning and big data techniques - A survey. In: Proceedings of the 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS) Coimbatore, India: IEEE. 1747 - 1753. doi: 10.1109/ICAIS50930.2021.9395832.

Khan, Y. and Chai, S.S. (2016) . Predicting and analyzing water quality using machine learning: A comprehensive model. In: Proceedings of the IEEE Long Island Systems, Applications and Technology Conference (LISAT). Farmingdale, New York: IEEE. 1 - 6. doi: 10.1109/LISAT.2016.7494106.

Li, Y., Mao, S., Yuan, Y., Wang, Z., Kang, Y. and Yao, Y. (2023). Beyond tides and time: Machine learning's triumph in water quality forecasting. American Journal of Applied Mathematics and Statistics 11(3): 89 - 97. doi: 10.12691/ajams-11-3-2.

Wei, S., Richard, R., Hogue, D., Mondal, I., Xu, T., Boyer, T. and Hamilton, K. (2024). High resolution data visualization and machine learning prediction of free chlorine residual in a green building water system. Water Research X 24(11): 100244. doi: 10.1016/j.wroa.2024.100244.

❑❑❑❑❑