Northeastern University
College *of* Engineering

IE 7280: Statistical Methods in Engineering
Fall 2024

Topic 1: Crop Recommendation System ANOVA
Topic 2: Water Quality Prediction Using Bayesian Ridge Regression

Group Members:

Akshara Reddy Patlanagari
Reddy Rukmini Reddy

Crop Recommendation System ANOVA

Introduction

The "Crop Recommendation Dataset" aims to provide agricultural insights by suggesting the most suitable crop types based on various soil and environmental parameters. The dataset helps in understanding how factors like nitrogen (N), phosphorus (P), potassium (K), temperature, humidity, pH, and rainfall influence the choice of crops for optimal growth.

Objective

The primary objective of using this dataset is to predict and recommend the best crop based on the provided soil and environmental conditions. The analysis focuses on performing statistical tests such as ANOVA to assess the relationship between soil pH levels and different crop types.

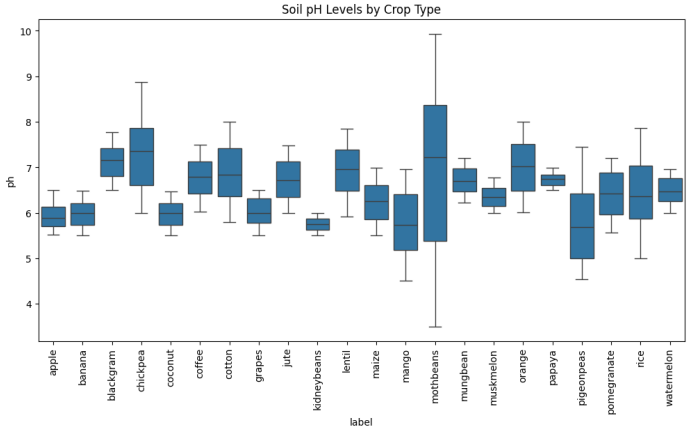
Dataset Information

The dataset consists of 2,200 entries with the following features:

- **N (Nitrogen content in the soil)**: The amount of nitrogen in the soil, which affects plant growth.
- **P (Phosphorus content in the soil)**: The level of phosphorus, essential for energy transfer in plants.
- **K (Potassium content in the soil)**: Potassium is vital for plant metabolism and stress resistance.
- **Temperature**: The temperature in degrees Celsius, which directly impacts crop development.
- **Humidity**: The relative humidity in the environment, influencing plant transpiration.
- **pH**: The soil pH value, a crucial factor affecting nutrient availability.
- **Rainfall**: The amount of rainfall in millimeters, influencing irrigation requirements.
- **Label (Crop recommended)**: The crop type recommended based on the conditions.

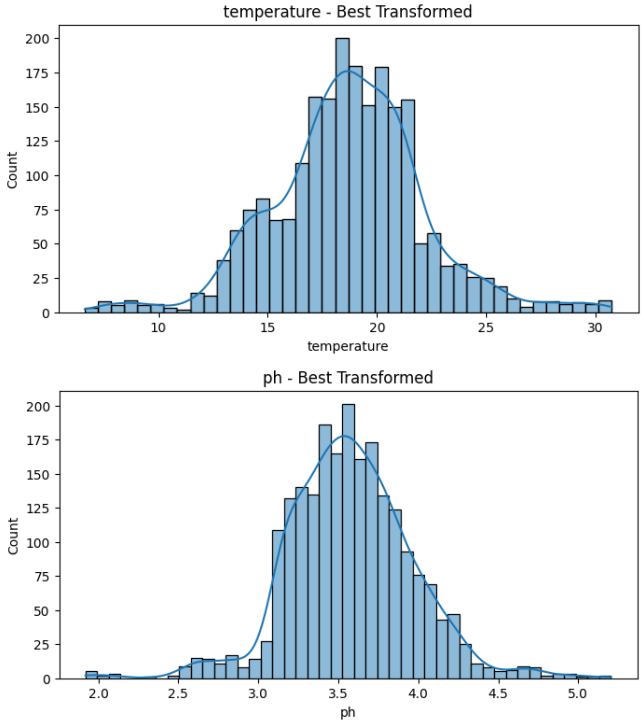
ANOVA Test Steps:

1. **Data Preparation:**
   - The dataset did not contain any missing values.
   - **label** (crop type), is treated as a categorical variable.



2. Assumption Checks:

- **Normality:** Performed Shapiro-Wilk test and if normality is violated, a transformation (e.g., log or square root) was applied to the pH data to achieve normality.



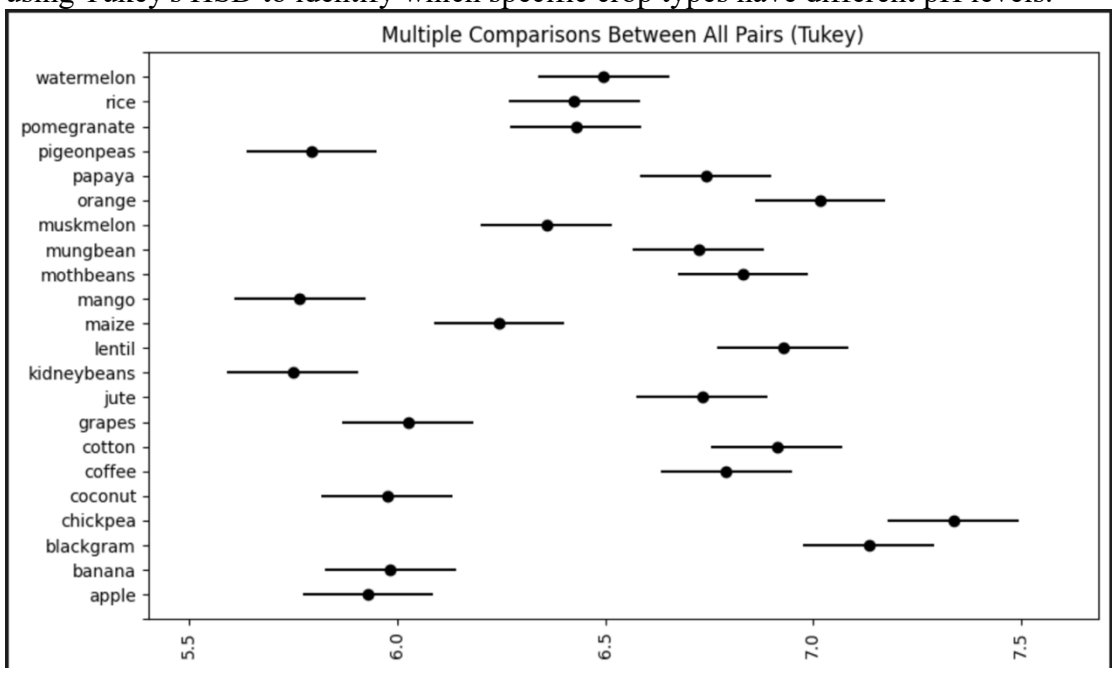
- **Homogeneity of Variances:** Using Levene's test we confirmed that homogeneity of variances assumption is met, since p-value was greater than 0.05.

3. **Performed ANOVA:**

	sum_sq	df	F	PR(>F)
C(label)	123.659317	21.0	59.471899	1.974110e-196
Residual	215.652063	2178.0	NaN	NaN

The p-value is significantly lower than 0.05, so we reject the null hypothesis, which means there is a statistically significant difference in the mean soil pH levels among the different crop types. This suggests that crop type has a significant effect on soil pH.

4. **Post-hoc Analysis:**
   Since the ANOVA test indicated significant differences, post-hoc tests were performed using Tukey's HSD to identify which specific crop types have different pH levels.



Conclusion:

In conclusion, the results of the one-way ANOVA test reveal a statistically significant difference in the mean soil pH levels among the different crop types, as indicated by the p-value being substantially lower than the 0.05 threshold. Consequently, we reject the null hypothesis, which posited that there were no differences in soil pH levels across crop types. This finding suggests that the crop type significantly influences the soil pH, highlighting the importance of considering crop-specific requirements when managing soil conditions for optimal agricultural outcomes.

Water Quality Prediction Using Bayesian Ridge Regression

Introduction

In this project, we aim to forecast the spatio-temporal water quality by predicting the "power of hydrogen (pH)" value . This water quality prediction project utilizes Bayesian ridge regression to model and predict pH levels based on historical water measurement indices. The project focuses on monitoring pH, a crucial metric for assessing water quality, which plays a vital role in ensuring environmental compliance and the health of aquatic ecosystems.

Objective

To develop a regression model that accurately predicts the next day's median pH value at 36 monitoring sites in Georgia, USA, using 11 water quality indices, while accounting for spatio-temporal dependencies to enhance water quality monitoring and support environmental management efforts.

Dataset Information

The input dataset comprises daily samples from 36 monitoring sites in Georgia, USA, providing comprehensive measurements related to pH levels. It contains 11 key indices, including:

- Dissolved Oxygen (DO): Critical for aquatic life.
- Temperature: Influences solubility and reaction rates.
- Specific Conductance: Indicates the water's ability to conduct electricity.

The target variable is the median pH value, labeled as "pH, water, unfiltered, field, standard units (Median)."

The dataset also exhibits spatial dependency with two distinct water systems:

1. One centered around the city of Atlanta.
2. The other located along the eastern coast of Georgia.

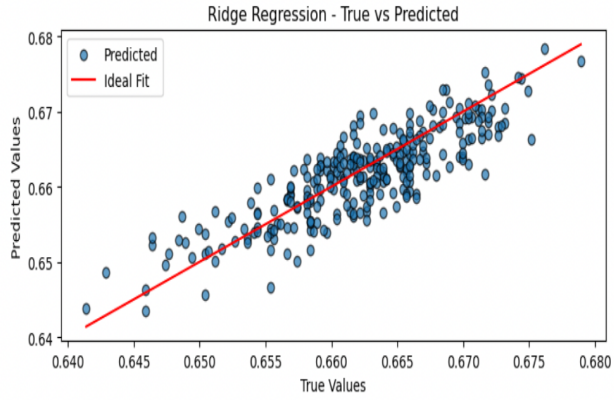
Additional detailed measurements for each feature, such as maximum, minimum, and mean values, are also available, including:

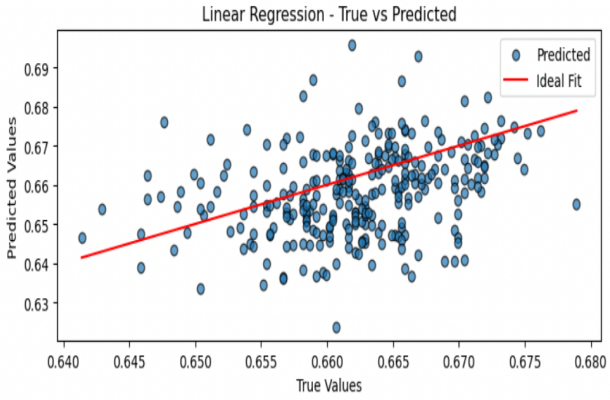
- Specific conductance (microsiemens per centimeter at 25°C).
- Dissolved oxygen (milligrams per liter).
- Temperature (degrees Celsius).

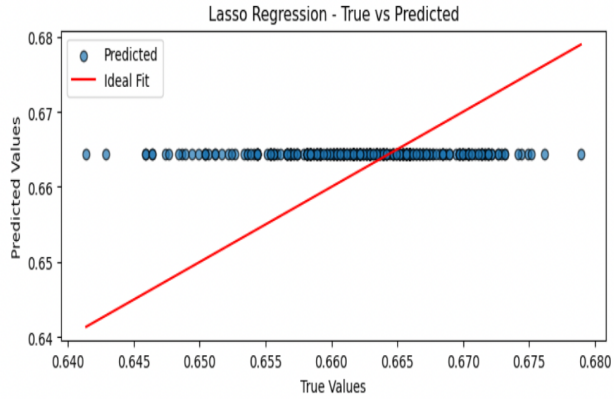
Regression models

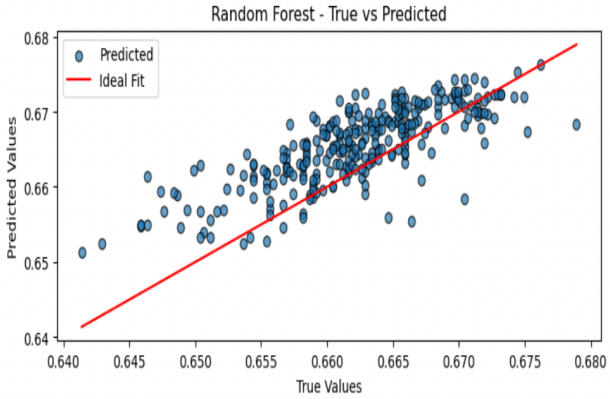
- **Ridge Regression** achieved an MSE of 1.231e-05 and an R^2 score of 0.701. While effective, it was outperformed by Bayesian Ridge in both metrics.

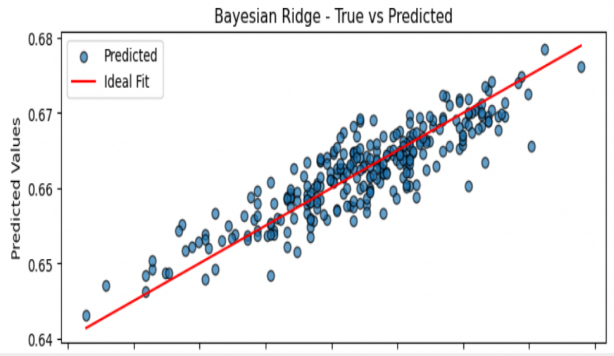
- **Linear Regression** and **Lasso Regression** performed poorly, with high MSE values and negative R^2 scores, indicating poor fit and prediction accuracy.
- **Random Forest Regression** performed moderately well, with an MSE of 2.676e-05 and an R^2 score of 0.351. While it captured some variance, its performance lagged the Ridge-based models.
- **Bayesian Ridge Regression** demonstrated the best performance, achieving the lowest MSE (9.266e-06) and the highest R^2 score (0.775). These results indicate that it made the most accurate predictions and explained the most variance in the target variable.











Through a systematic comparison of performance metrics, Bayesian Ridge Regression emerged as the most accurate and robust model for predicting pH values. Its superior MSE and R^2 score indicate that it outperforms other models in both prediction accuracy and explanatory power, making it the best choice for this analysis. We will now deep dive into Bayesian Ridge Regression.

Bayesian Ridge Regression

Here we are using Bayesian Ridge Regression to model the relationship between input features and the target variable, pH. Bayesian Ridge Regression adopts a probabilistic approach, estimating the posterior distribution of the model's parameters. It incorporates priors for the regression coefficients and noise precision, enabling the model to account for uncertainty in parameter estimates.

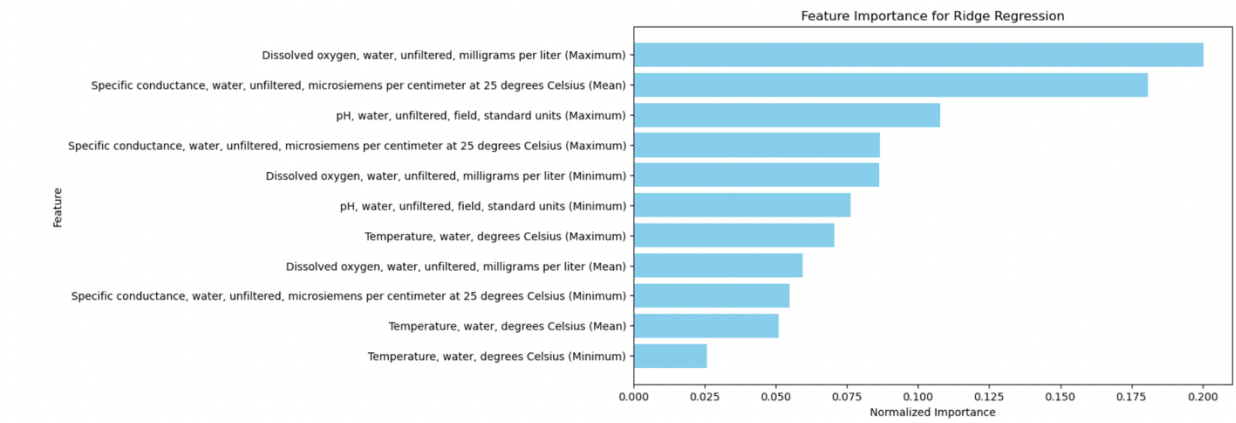
This regression technique balances the trade-off between bias and variance by automatically adjusting regularization through its hyperparameters (alpha_1, alpha_2, lambda_1, and lambda_2). The model is trained on the scaled training dataset to capture the relationship between features and target values.

Bayesian Ridge Regression is particularly effective for datasets with multicollinearity or noise, offering robust and interpretable predictions. It is ideal for analyzing water quality indices, such as dissolved oxygen and temperature, by incorporating the inherent uncertainty in the data while providing confidence intervals for predictions. This makes it a strong choice for ensuring robustness and interpretability in water quality analysis, even in challenging, noisy environments.

```
: from sklearn.linear_model import BayesianRidge
  from sklearn.metrics import mean_squared_error, r2_score

  # Bayesian Ridge Regression
  bayesian_ridge_model = BayesianRidge()
  bayesian_ridge_model.fit(X_train_scaled, y_train)
```

This visualization shows the relative importance of features in predicting pH. Dissolved Oxygen (Maximum) is the most significant predictor, followed by Specific Conductance (Mean)and pH (Maximum). Features related to maximum, minimum, and mean values of dissolved oxygen, specific conductance, and temperature are key drivers. This helps identify critical factors for pH prediction and guides feature prioritization for model optimization.



Performance metrics

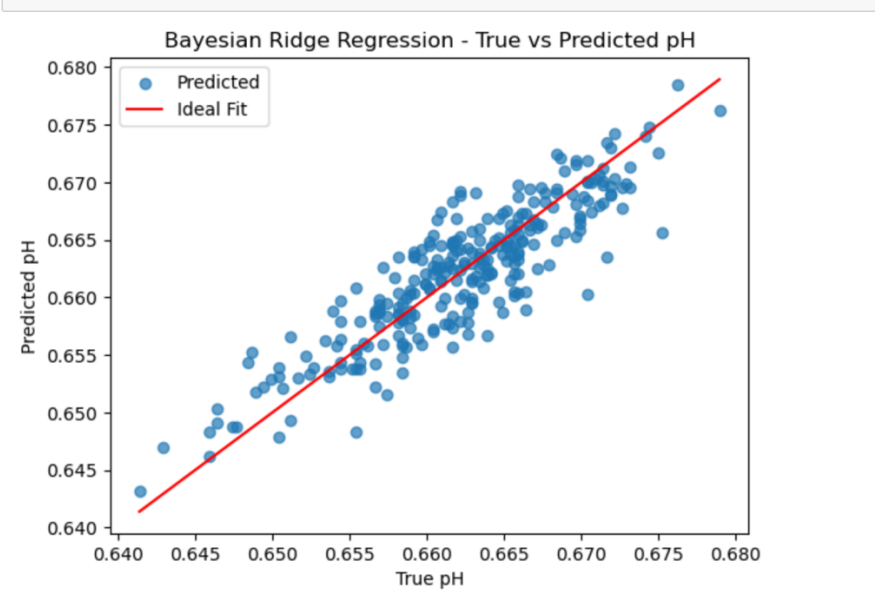
Once the model is trained, it is used to predict pH values for the scaled test dataset to evaluate its performance, we calculate two key metrics: Mean Squared Error (MSE), which quantifies the average squared difference between the actual and predicted values, and \(R^2 \), which measures the proportion of variance in the target variable explained by the model. These metrics provide insights into the model's accuracy and its ability to generalize to unseen data.

```
y_pred_bayesian_ridge = bayesian_ridge_model.predict(X_test_scaled)
mse_bayesian_ridge = mean_squared_error(y_test, y_pred_bayesian_ridge)
r2_bayesian_ridge = r2_score(y_test, y_pred_bayesian_ridge)

print(f"Bayesian Ridge Regression - MSE: {mse_bayesian_ridge}, R^2: {r2_bayesian_ridge}")
```

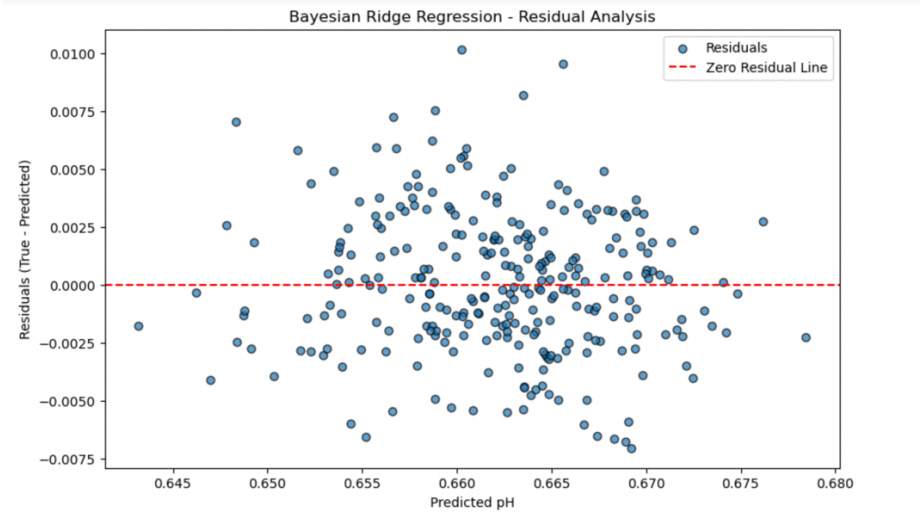
```
Bayesian Ridge Regression - MSE: 9.266098324760234e-06, R^2: 0.775123620808926
```

The Bayesian Ridge Regression model demonstrates excellent performance, achieving a very low Mean Squared Error (MSE) of `9.27e-06`, indicating minimal prediction error and high accuracy in estimating the target variable. Additionally, the model achieved an R^2 score of `0.775`, explaining 77.5% of the variance in the target variable, which reflects a strong fit and reliability. The model's probabilistic framework allows it to effectively balance bias and variance while incorporating uncertainty in parameter estimates. This makes it particularly well-suited for datasets with noise or multicollinearity, ensuring robust and accurate predictions. Overall, Bayesian Ridge Regression outperformed other models tested, making it the most suitable choice for this analysis.



The scatter plot demonstrates the effectiveness of Bayesian Ridge Regression in predicting pH values. The true pH values are plotted on the x-axis, while the predicted values are on the y-axis, with the red line representing the ideal fit where predictions perfectly match the actual values. The blue scatter points are closely aligned with this ideal line, indicating that the model achieves high accuracy in its predictions. The even distribution of points along the line suggests that the model generalizes well across the dataset, with only minor deviations observed for a few predictions.
This alignment supports the model's strong performance metrics, including a very low Mean Squared Error (MSE) and a high R^2 score of 0.775. The minimal scatter around the ideal line confirms that Bayesian Ridge Regression effectively captures the linear relationship between the input features and the target variable. Overall, the plot visually validates the model's reliability and suitability for accurately predicting pH values in this dataset.

Residual Analysis



The residual plot for Bayesian Ridge Regression provides valuable insights into the model's performance. The residuals, which represent the difference between the true and predicted values, are distributed symmetrically around the zero line, as indicated by the red dashed line. This symmetry suggests that the model's predictions are unbiased and free from systematic errors. Additionally, the residuals are randomly scattered, with no discernible patterns, indicating that the model has effectively captured the underlying linear relationships in the data and there are no significant issues such as non-linearity or heteroscedasticity.

Most residuals are close to zero, confirming the model's high prediction accuracy, with only a few larger residuals that appear as outliers. These larger residuals are not excessive and do not indicate significant performance problems. Overall, the residual analysis reinforces that Bayesian Ridge Regression is a robust and reliable choice for this dataset, aligning with its strong statistical metrics, such as the low MSE and high R^2 score.

Conclusion

Regression included:

Model	MSE	R^2 Score
Ridge Regression	1.2314613440510751e-05	0.7011400500424273
Linear Regression	0.00013497720426379774	-2.2757244639924794
Lasso Regression	4.5854134365762554e-05	-0.11282131332018785
Random Forest	2.6761780331683713e-05	0.35052574980617257
Bayesian Ridge	9.266098324760234e-06	0.7751236208088926

The comparison of true vs. predicted plots for all models highlights the superior performance of Bayesian Ridge Regression. While Ridge Regression and Random Forest also demonstrated reasonable accuracy, Bayesian Ridge consistently produced predictions that aligned most closely with the ideal fit line, indicating higher precision. On the other hand, Linear Regression and Lasso Regression showed significant deviations from the ideal fit, suggesting poor model fit and unreliable predictions.

Bayesian Ridge Regression's probabilistic framework allows it to effectively handle noise and multicollinearity in the dataset, leading to more accurate and robust predictions. This is further validated by its lowest Mean Squared Error (MSE) and highest R^2 score among all models tested. The random scatter around the ideal fit line in the residual plots further confirms the absence of systematic bias in its predictions.

Overall, Bayesian Ridge Regression stands out as the most reliable and accurate model for this dataset, offering superior generalization and predictive performance. It is the best choice for modeling the pH values, ensuring robust and interpretable results even in the presence of noise or correlated features.