An ML Approach for Assessing Quality Retention and F-value during Sterilization of Canned Peas in Water

Soumadip Das

200107085

Submission Date: April 25, 2024

Final Project submission

Course Name : Applications of Al and ML in chemical engineering

Course Code: CL653

Contents

1 Executive Summary....3
2 Introduction....3
3 Methodology....4
4 Implementation Plan....8
5 Testing and Deployment....8
6 Results and Discussion....9
7 Conclusion and Future Work....15
8 References....15
9 Auxiliaries....16

1 Executive Summary

Thermal sterilization is a crucial method in the food industry for preserving and extending the shelf life of perishable food products by eliminating pathogenic microorganisms or inhibiting enzyme denaturation. Conventionally, this process involves subjecting the food product to higher temperatures for a specified duration, which can potentially lead to nutrient degradation, color alterations, and flavor loss. However, modeling the thermal sterilization process using computational fluid dynamics (CFD) techniques is computationally expensive.

In this study, we propose the integration of machine learning (ML) models as an alternative to the traditional CFD approach. Specifically, linear regression, random forest regression, and neural network regression techniques are employed to predict nutrient retention levels and F-value of predominant microorganisms in peas at the end of sterilization, considering various values of retort temperatures and processing durations. This ML-based approach offers a more efficient means of estimating nutrient retention and F-value, facilitating the development of comprehensive guidelines for maintaining optimal nutrient levels in canned pea products. Our findings revealed that the neural network regressor exhibited the highest R2 scores among the three algorithms. Conversely, linear regression yielded the lowest R2 scores. These results suggest that the neural network regression model outperformed both linear regression and random forest regression in predicting nutrient retention levels and F-value of microorganisms in canned pea products following thermal sterilization.

2 Introduction

2.1 Background

Thermal sterilization stands as one of the most critical methods in the food industry, to preserve and prolong the shelf life of perishable food products by eliminating pathogenic microorganisms or inhibiting enzyme denaturation (Holdsworth & Simpson, 2016). It involves the application of higher temperatures (above 100 °C) to the food product for a specified duration. Thermal sterilization can also potentially lead to nutrient degradation, alterations in color, and flavor loss in the food product Conventionally, the thermal sterilization process is modeled using computational fluid dynamics (CFD) techniques. However, this method is hindered by its significant computational expenses. Hence, instead of the traditional CFD approach, an ML model can be integrated to predict the various parameters post thermal sterilization easily without much computational expense.

2.2 Problem statement

Peas immersed in water is considered for thermal sterilization. The geometry of peas in the can is shown in **Figure 1**. The model depicting the arrangement of peas in the can has been reproduced from Kiziltaş et al. (2010). The diameter and height of the can are 73 mm and 104.5 mm, respectively. A gap of 0.5 mm is assumed between the peas to allow the free movement of water in the interstices. The peas are assumed as spherical balls of diameter 7.5 mm. Thermal sterilization is performed at a particular retort temperature (T_{retort}) at the walls for a specific processing time.

2.3 Governing equations (CFD approach)

The traditional CFD approach uses the following governing equations to predict the temperature and velocity contours in the peas and water. The temperature distribution in solid food (peas) is calculated by solving the energy equation given in Eq. (1).

$$\rho C_p \frac{\partial T}{\partial t} = \nabla \cdot (k \nabla T), \tag{1}$$

Similarly, the energy equation mentioned below is solved for estimating temperature distribution in water,

$$\rho \cdot C_p \left(\frac{\partial T}{\partial t} + \mathbf{U} \cdot \nabla T \right) = \nabla \cdot (k \nabla T) \tag{2}$$

where $\rho, C_p, \mathbf{U}, k, T$ are the density (kg m^{-3}), specific heat capacity (J kg^{-1} K^{-1}), velocity field (m s^{-1}), thermal conductivity (W m^{-1} K^{-1}) and the temperature field (K) of the food products, respectively. The velocity distribution in water is obtained by solving the continuity and momentum equations as,

$$\frac{\partial \rho}{\partial t} + \nabla \cdot \rho \mathbf{U} = 0, \tag{3}$$

$$\rho \left(\frac{\partial \mathbf{U}}{\partial t} + (\mathbf{U} \cdot \nabla)\mathbf{U} \right) = -\nabla p + \mu \left[\nabla \cdot \left(\nabla \mathbf{U} + (\nabla \mathbf{U})^T \right) \right] + \rho g \tag{4}$$

where p, μ, g are the pressure (Pa), dynamic viscosity (Pa s), and acceleration due to gravity (m s^{-2}), respectively.

F-value or the accumulated lethality of the most predominant microorganism in peas, i.e *Clostridium botulinum* at the SHZ can be calculated as,

$$F(t)=\int_0^t 10^{\frac{(T_{shz}-T_{reff})}{z_f}}\,dt \tag{5}$$

where t and T_{shz} represent the total duration of thermal sterilization (min) and the temperature (°C) of the SHZ within the food. The reference temperature and the thermal resistance of the microorganism is denoted by T_{reff} and z_f.

The quality retention of the most heat-susceptible nutrient, i.e, ascorbic acid in peas in the 2-D axis-symmetric plane is defined as,

$$C_{av}(t)=\frac{1}{S_T}\int_0^{S_T}\exp\left[-\frac{\ln 10}{D_{refs}}\int_0^{t_f}10^{\frac{(T-T_{refs})}{z_s}}\,dt\right]dS_T, \tag{6}$$

where S_T is the total surface area of the 2-D axis-symmetric plane, T_{refs} is the reference temperature, D_{refs} is the decimal reduction time (min), and z_s is the thermal resistance for nutrient degradation of the most heat-labile nutrient (°C). The thermophysical properties of peas and water, and the nutrient degradation parameters are adopted from Kiziltaş et. al. (2010).

2.4 Retort temperature profile

Figure 2 shows the retort temperature profile is applied at the walls of the can. The come-up time, heating time, and cooling time is represented using t_{CUT}, t_{h}, and t_{c}, respectively. The heating temperature is shown as T_{wall} °C.

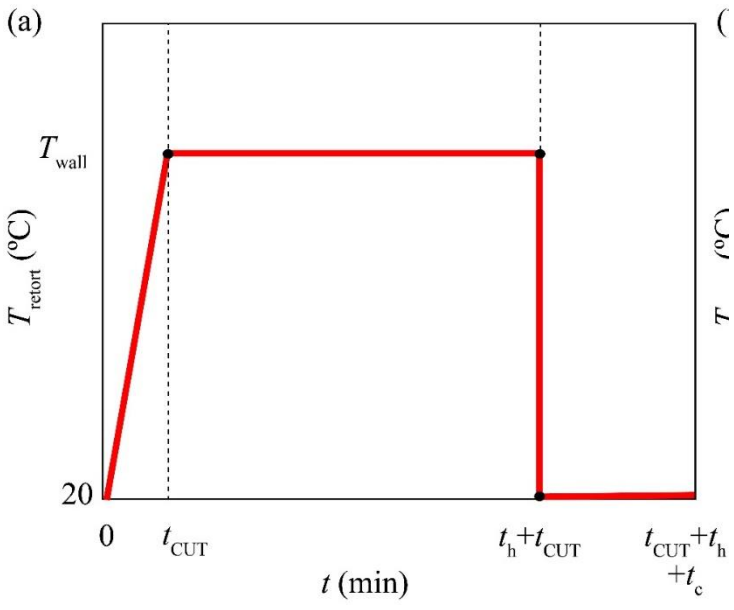

Figure 2. Variation of the retort temperature at the walls with processing time

3 Methodology

3.1 Dataset

The data is generated by solving the equations (1)-(6) for 2195 cases on a powerful computer. The data used for the model is static. The primary input variables in the data are come-up time (t_{CUT}), heating time (t_h), cooling time (t_c) and heating temperature (T_{wall}) for sterilization. The output is the quality retention of ascorbic acid (C_{av}) and F-value (F) of *Clostridium botulinum* in the peas. It is to be noted that the data is devoid of any null values and outliers since it is generated by solving the mathematical equations. The first five elements of the dataset are shown in **Table 1**. The description of dataset is provided in **Table 2**.

Since the data is devoid of any null values and outliers, only scaling of data is performed in the pre-processing stage.

Table 1. First five elements of the dataset

F value (min)	Quality retention (%)	Come up time (min)	Heating time (min)	Cooling time (min)	Retort temperature (degree C)
816.279553	59.876044	5	19	1	139
14.217668	94.183276	4	6	2	131
1113.047961	54.21355	5	20	1	140
6.955381	96.267563	5	12	5	122
1017.976194	56.200988	3	19	4	140

Table 2. Description of the dataset

	F value (min)	Quality retention (%)	Come up time (min)	Heating time (min)	Cooling time (min)	Retort temperature (degree C)
count	2195	2195	2195	2195	2195	2195
mean	130.2487	87.11471	2.971298	12.54442	3.006378	130.0692
std	201.9201	10.03766	1.413439	4.517585	1.412909	6.123519
min	0.326198	54.1846	1	5	1	120
25%	10.77364	82.27947	2	9	2	125
50%	43.49809	90.15044	3	12	3	130
75%	144.645	95.10631	4	16	4	135
max	1113.048	98.82775	5	20	5	140

3.2 Model architecture

The block diagram for the ML model is shown in **Figure 3**. Two ML models are constructed for predicting the F-value of *Clostridium botulinum* and quality retention of ascorbic acid during sterilization of canned peas.

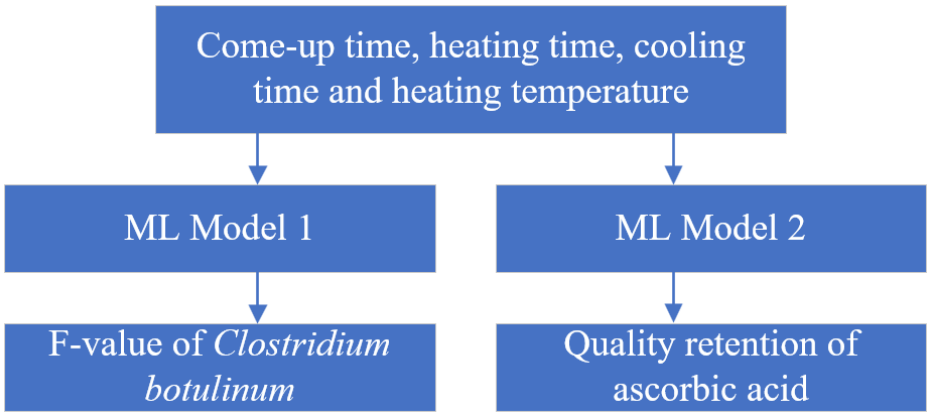

Figure 3. Flowchart depicting the input and output parameters of the model

4 Implementation Plan

4.1 Model selection

For model selection, we employed linear regressor, random forest regressor, and neural networks regressor due to their robustness, flexibility, and effectiveness in handling regression tasks. Linear Regression is a simple yet powerful method that establishes a linear relationship between the independent and dependent variables. The Random Forest Regressor, an ensemble learning method from the scikit-learn package, aggregates predictions from multiple decision trees, reducing overfitting and capturing complex relationships within the data. Moreover, it provides feature importance metrics, facilitating feature selection and interpretation, which is crucial for our predictive modeling task. Neural Networks, on the other hand, are complex models inspired by the human brain's neural structure, capable of capturing intricate patterns and nonlinear relationships in data.

4.2 Training

Following preprocessing, the dataset is split into training and testing sets using sci-kit-learn's train_test_split function. The size of the testing set is specified as 20% of the total dataset and set in a random state to ensure reproducibility. Then the linear regressor from scikit-learn's LinearRegression class, the random forest regressor from scikit-learn's RandomForestRegressor class, and the neural network regressor from the TensorFlow construct are employed to train the model. Following training hyperparameter tuning is also performed with GridSearchCV to find the optimal model settings. This iterative process enables us to fine-tune the model's hyperparameters, enhancing its predictive performance.

4.3 Evaluation Metrics

For evaluation purposes, Mean Squared Error (MSE) and R-squared (R2) scores were selected as the primary evaluation metrics. MSE measures the average squared difference between the actual and predicted values, indicating the model's accuracy. R2 score measures the proportion of the variance in the target variable that is explained by the model, indicating how well the model fits the data.

6.4. Validation Strategy

The validation strategy involves multiple steps to ensure the model's generalizability and robustness. Initially, the dataset is split into training and testing sets using the train-test split method. The model is then trained on the training set and evaluated on the testing set to assess its performance. Additionally, k-fold cross-validation using sci-kit-learn's cross_val_score function is employed to further validate the model. This technique partitions the dataset into k subsets, trains the model on k-1 subsets, and validates it on the remaining subset, repeated k times. After evaluating the model's performance, we further validate it using cross-validation on the entire dataset, providing a more robust estimate of its generalization performance. External validation on unseen datasets, if available, is also conducted to ensure the model's performance on new data. These validation strategies contribute to determining the model's reliability and effectiveness.

5 Testing and Deployment

5.1 Testing strategy

The model's performance metrics in terms of testing are determined using mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), and R-squared. Additionally, hyperparameter tuning techniques like grid search or random search will be applied to optimize the model's parameters, mitigating overfitting and enhancing its robustness.

5.2 Deployment strategy

For the deployment strategy, careful consideration will be given to scalability, performance, and maintenance aspects. The model's scalability will be evaluated to ensure it can handle increasing data volume or user demand efficiently. Continuous monitoring and optimization protocols will be implemented to uphold the model's performance in real-world scenarios. Regular updates and maintenance routines will also be scheduled to address software updates, changing data distributions, or evolving business requirements.

5.3 Ethical considerations

Ethical considerations will play a crucial role in the deployment process. Measures will be taken to identify and mitigate potential biases in the data or model predictions to ensure fairness and prevent discrimination. Transparency will be maintained regarding the model's decision-

making process, explaining how predictions are generated and any underlying assumptions. Clear accountability and responsibility frameworks will be established to address any unintended consequences or errors in model predictions. Moreover, careful consideration will be given to how the model's outputs may be used, ensuring they align with ethical guidelines and societal values.

6 Results and Discussion

7.1. Residuals

The plot of residuals is quantified by homoscedasticity and heteroscedasticity. **Figure 4** and **Figure 5** show the plot of the residuals and the absolute residuals respectively for linear regressor. Similarly, **Figure 6** and **Figure 7** show the plot of the residuals and the absolute residuals respectively for Random Forest Regressor. Lastly, **Figure 8** and **Figure 9** show the plot of the residuals and the absolute residuals respectively. Large values of residuals are reported for linear regressors. On the other hand, the values of residuals are substantially reduced for random forest and neural networks regressor. Moreover, the random forest regressor slightly outperforms the neural networks regressor, thereby reporter slightly lesser values of residuals.

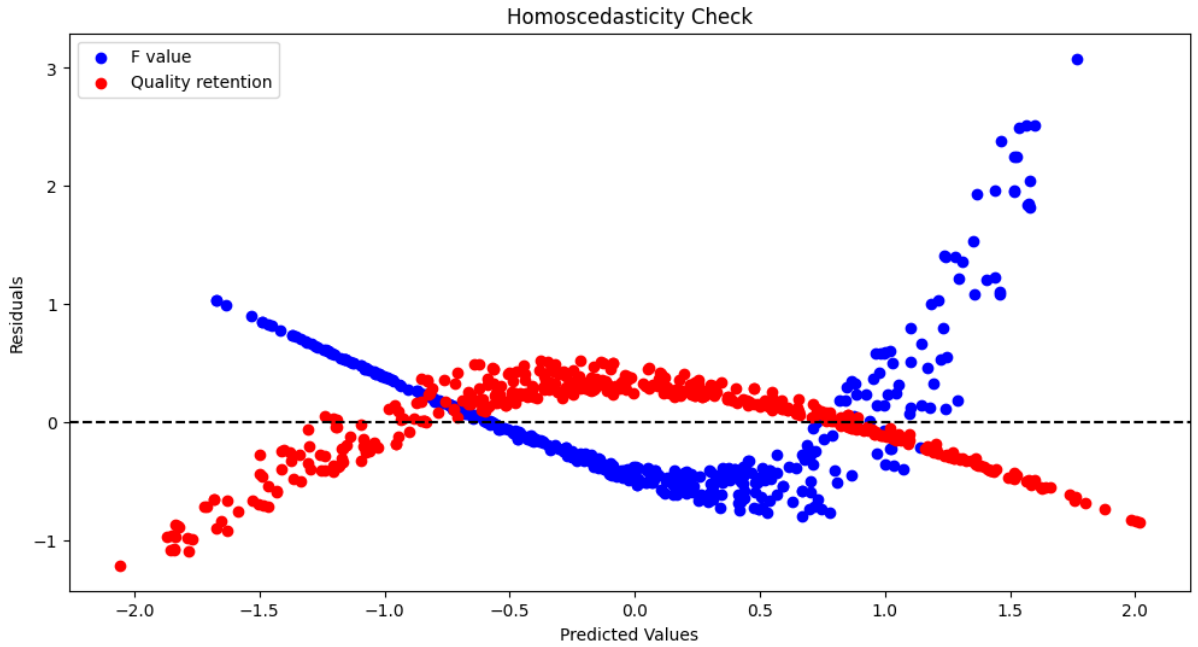

Figure 4. Homoscedasticity plot for linear regressor

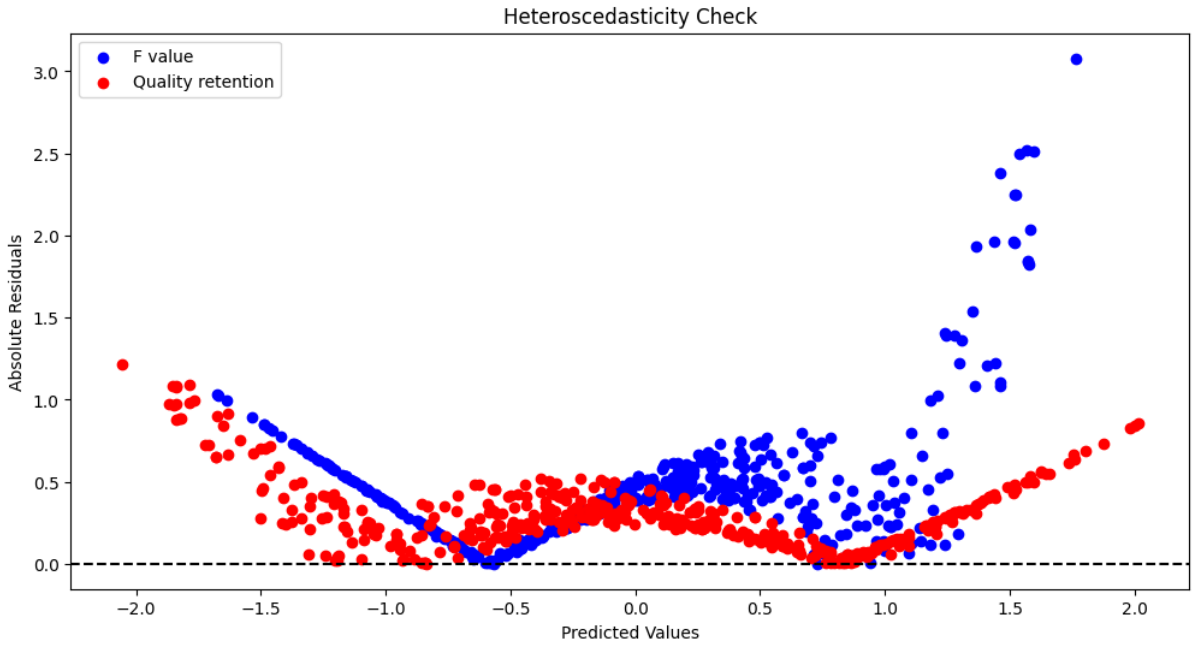

Figure 5. Heteroscedasticity plot for linear regressor

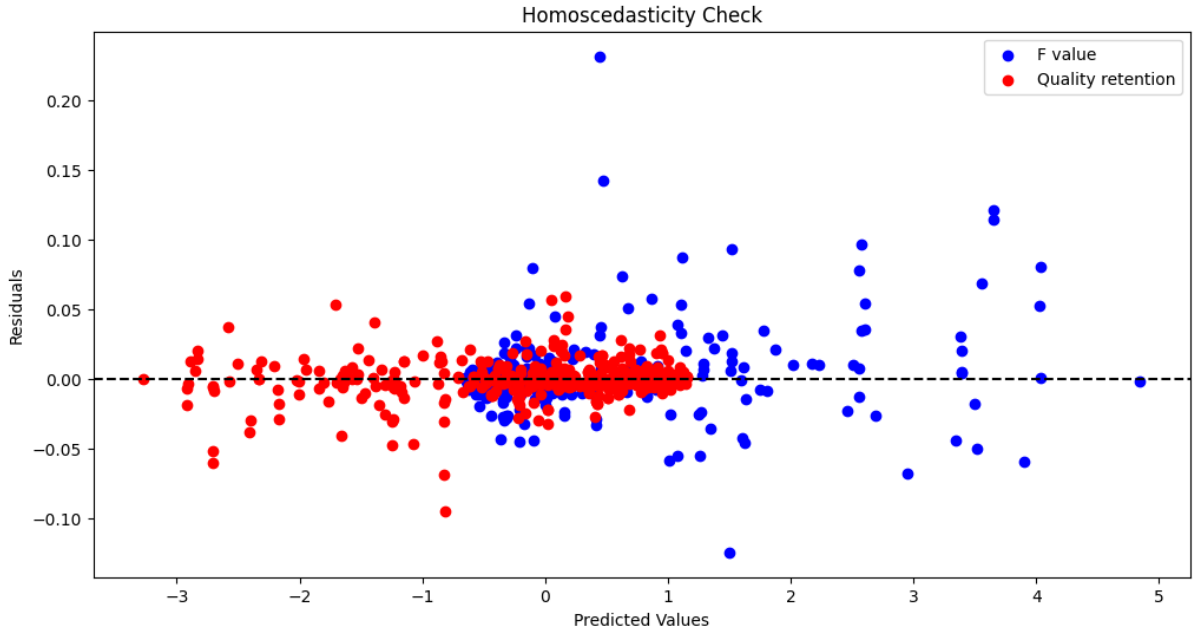

Figure 6. Homoscedasticity plot for random forest regressor

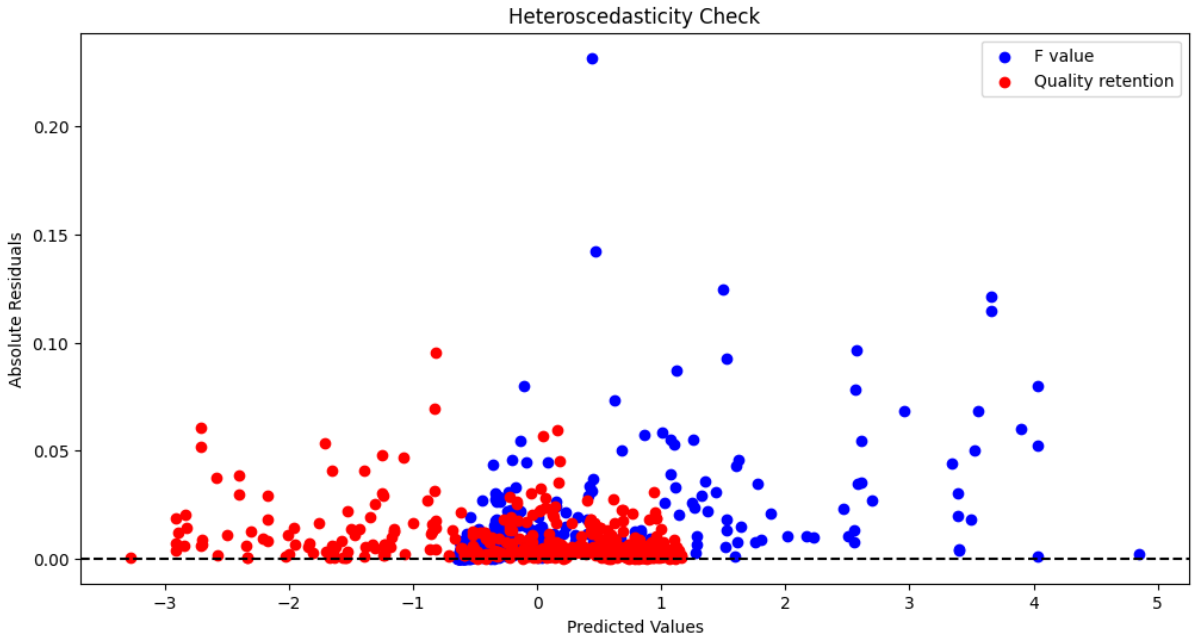

Figure 7. Heteroscedasticity plot for random forest regressor

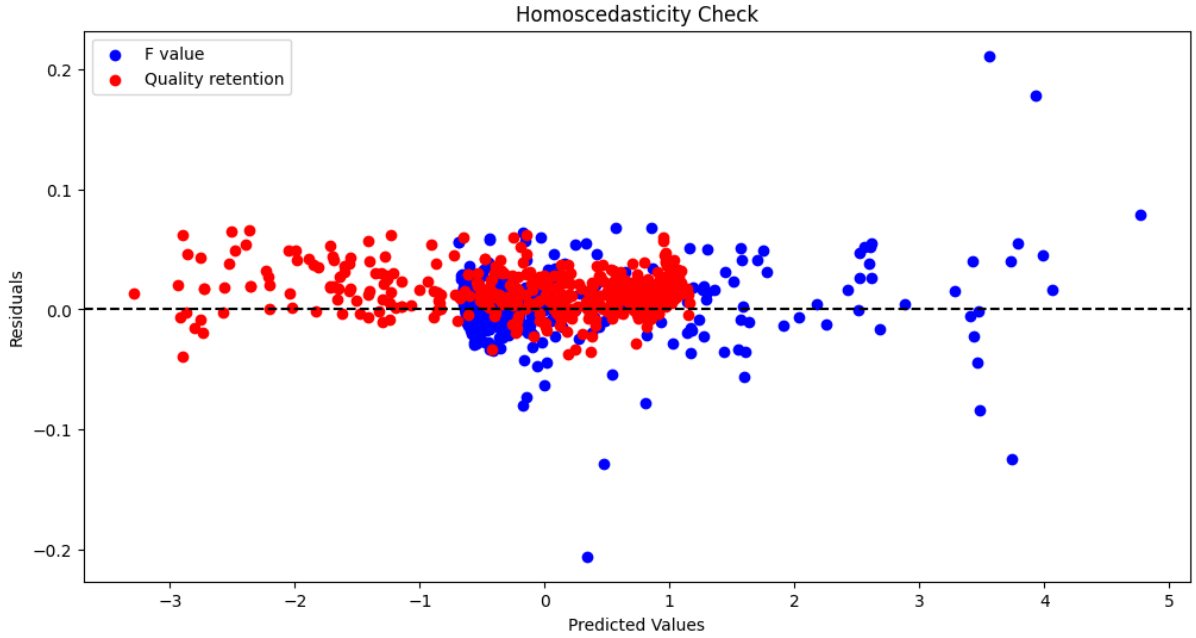

Figure 8. Homoscedasticity plot for neural networks regressor

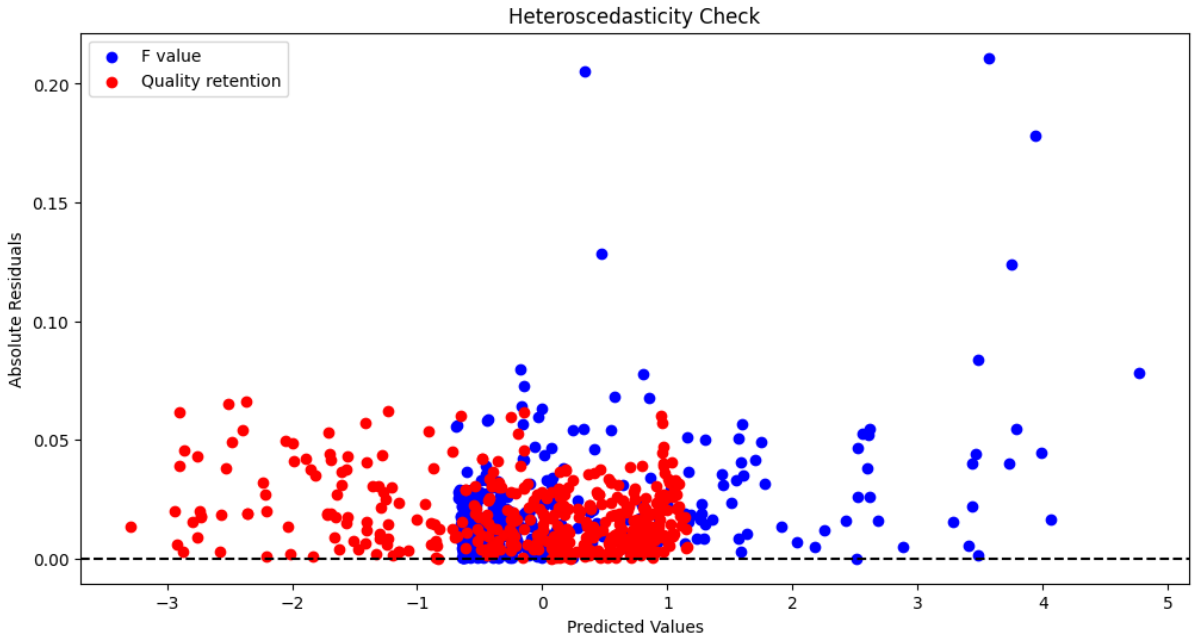

Figure 9. Heteroscedasticity plot for neural networks regressor

7.2. Evaluation and cross-validation

The evaluation and cross-validation metrics and the results are shown in **Table 3**.

Table 3. Evaluation and cross-validation metrics and the corresponding values

Metric	**Linear regressor**	**Random forest regressor**	**Neural networks regressor**
MSE (Model 1)	0.3935	0.0006	0.0010
MSE (Model 2)	0.1314	0.0001	0.0003
R2 Score (Model 1)	0.6226	0.9994	0.9989
R2 Score (Model 2)	0.8707	0.9998	0.9996
Cross-Validation R2 Scores (Model 1)	[0.59773109 0.62640991 0.62323045	[0.99964609, 0.99936063, 0.99877663,	[0.9990, 0.9992, 0.9993, 0.9995, 0.9991]

	0.61903893 0.59990063]	0.99919765, 0.99976843]	
Cross-Validation R2 Scores (Model 2)	[0.87146878 0.87406194 0.86983964 0.87684776 0.86745902]	[0.99985365, 0.9998092, 0.99963089, 0.99969444, 0.99992751]	[0.9997, 0.9997, 0.9996, 0.9998, 0.9998]
Mean Cross-Validation R2 Score (Model 1)	0.6132	0.9993	0.9992
Standard Deviation of Cross-Validation R2 Score (Model 1)	0.012	0.0003	0.0002
Mean Cross-Validation R2 Score (Model 2)	0.8719	0.9997	0.9997
Standard Deviation of Cross-Validation R2 Score (Model 2)	0.0032	0.0001	0

7.3. Hyperparameter tuning

Hyperparameter tuning is not supported for linear regressors or neural network regressors. Hence it is only performed for random forest regressor. The results of hyperparameter tuning are shown in **Table 4**.

Table 4. Results of hyperparameter tuning for random forest regressor

Metric	Value
Best hyperparameters (Model 1)	{'max_depth': 10, 'n_estimators': 100}
Best hyperparameters (Model 2)	{'max_depth': None, 'n_estimators': 150}
R2 Score with Best Hyperparameters (Model 1)	0.9994
R2 Score with Best Hyperparameters (Model 2)	0.9998

7 Conclusion and Future Work

The study aims to predict the values of the F-value of *Clostridium botulinum* and the quality retention of ascorbic acid during thermal sterilization of canned peas. For this purpose, a dataset of 2195 points is generated by solving the governing equations. Employing the dataset, two ML models are trained for predicting the F-value and quality retention. Based on the evaluation of three regression algorithms—Linear Regressor, Random Forest Regressor, and Neural Networks Regressor—several conclusions can be drawn regarding their performance. The Linear regression yielded moderate predictive capability, as evidenced by an R2 score of 0.6226 for Model 1 and 0.8707 for Model 2. However, these scores were comparatively lower than those achieved by the Random Forest Regressor and Neural Networks Regressor. The Random Forest Regressor demonstrated excellent predictive capability, with Model 1 achieving an impressive R2 score of 0.9994 and Model 2 further improving to 0.9998. Moreover, both models exhibited high consistency in performance, as indicated by mean cross-validation R2 scores of 0.9993 and 0.9997, respectively. Additionally, the standard deviations of their cross-validation R2 scores were notably low, highlighting the robustness and stability of the models. Similarly, the Neural Networks Regressor showcased strong predictive capability, albeit slightly lower than the Random Forest Regressor. Model 1 achieved an R2 score of 0.9989, while Model 2 demonstrated improved performance with a score of 0.9996. The mean cross-validation R2 scores for both models were 0.9992, indicating high consistency in performance. Moreover, Model 2 exhibited a very low standard deviation in its cross-validation R2 scores, akin to the Random Forest Regressor, further underscoring its robustness and stability. Furthermore, hyperparameter tuning was exclusively performed for the Random Forest Regresso. The optimized hyperparameters led to enhanced R2 scores of 0.9994 for

Model 1 and 0.9998 for Model 2, affirming the efficacy of hyperparameter tuning in improving predictive performance. Overall, the Random Forest Regressor emerged as the top-performing algorithm, followed closely by the Neural Networks Regressor, while the Linear Regressor exhibited comparatively lower predictive capability.

8 References

Holdsworth, S. D., & Simpson, R. (2016). Thermal Processing of Packaged Foods. Springer International Publishing. https://doi.org/10.1007/978-3-319-24904-9

Das, S., Baro, R. K., Kotecha, P., & Anandalakshmi, R. (2023). Numerical studies on thermal processing of solid, liquid, and solid–liquid food products: A comprehensive analysis. Journal of Food Process Engineering, e14484. https://doi.org/10.1111/jfpe.14484

Kiziltaş, S., Erdoğdu, F., & Palazoğlu, T. K. (2010). Simulation of heat transfer for solid-liquid food mixtures in cans and model validation under pasteurization conditions. Journal of Food Engineering, 97(4), 449–456. https://doi.org/10.1016/j.jfoodeng.2009.10.042

9 Auxiliaries

Please add the below mentioned links.

Data Source:

https://raw.githubusercontent.com/Soumadipdas18/CL653_Project/main/data_ml.csv

Python file:

https://github.com/Soumadipdas18/CL653_Project/blob/main/Soumadip_Das_ML_Project.ipynb