Journal of Science Research and Reviews
PRINT ISSN: 1595-9074
E-ISSN: 1595-8329
DOI: https://doi.org/10.70882/josrar.2025.v2i1.35
Homepage: https://josrar.esrgngr.org
Original Research Article

Linear Regression Approach to Solving Multicollinearity and Overfitting in Predictive Analysis

***[1]Otse, E. J., [1]Obunadike, G. N. and [2]Abubakar, A.**

[1]Department of Computer Science, Federal University Dutsinma, Katsina State, Nigeria.
[2]Department of Software Engineering, Faculty of Computing FUDMA, Katsina State Nigeria.
*Corresponding Author's email: igkabo@fudutsinma.edu.ng

KEYWORDS
Linear Regression,
Multicollinearity,
Overfitting,
Predictive Analysis,
Exploratory Data Analysis.

ABSTRACT
Multicollinearity and overfitting are ubiquitous problems in predictive analysis, especially in linear regression models, which significantly hinder the precision and interpretability of predicted results providing critical insights for data-driven decision-making in diverse industries. This research examines a linear regression approach to address the dual challenges of multicollinearity and overfitting in predictive analysis. The dataset, sourced from the National Center for Disease Control (NCDC), was analyzed using multiple regression techniques, including Linear Regression, Ridge Regression, LASSO Regression, and Elastic Net Regression. The study aimed to assess and compare the efficacy of these methods in mitigating multicollinearity (measured by Variance Inflation Factor) and reducing overfitting through Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) metrics. Data was analyzed both with all features and after applying feature selection. Results demonstrated that while all models effectively addressed multicollinearity and overfitting, Elastic Net Regression exhibited superior performance, offering the best generalization capabilities with minimal MSE and RMSE discrepancies between internal and external data. These findings highlight the potential of advanced regularization techniques in improving predictive accuracy and interpretability, particularly in high-dimensional data contexts such as those involving COVID-19 outcomes. The study underscores the importance of further research into enhanced machine learning techniques and the inclusion of broader datasets to refine predictive models for practical decision-making across sectors.

CITATION
Otse, E. J., Obunadike, G. N., & Abubakar, A.. (2025). Linear Regression Approach to Solving Multicollinearity and Overfitting in Predictive Analysis. *Journal of Science Research and Reviews*, *2*(1), 108-117. https://doi.org/10.70882/josrar.2025.v2i1.35

INTRODUCTION

Multicollinearity and overfitting are two critical challenges encountered in linear regression analysis. Linear regression is widely used in various fields, including economics, social sciences, and machine learning, to model the relationship between a dependent variable and a set of independent variables. However, the presence of multicollinearity and overfitting can undermine the accuracy and interpretability of regression models (Herawati *et al.*, 2018). Multicollinearity refers to the high correlation among independent variables in a regression model. When multicollinearity is present, it becomes challenging to estimate the individual effects of each predictor accurately. This issue leads to unstable and unreliable coefficient estimates, making it difficult to interpret the impact of specific variables on the dependent variable (Fox, 2015). Multicollinearity can also lead to inflated standard errors and decreased statistical

significance, affecting the reliability of hypothesis tests (Herawati *et al*., 2018). Overfitting, on the other hand, occurs when a regression model becomes excessively complex and captures noise or random fluctuations in the data. An overfit model performs well on the training data but fails to generalize to new, unseen data. Overfitting can result in misleadingly high predictive accuracy during model evaluation, but its performance degrades when applied to real-world scenarios. Overfit models are excessively sensitive to the idiosyncrasies of the training data, leading to poor generalization and unreliable predictions. To address these issues, researchers and practitioners have developed various techniques, including Ridge regression and LASSO regression. These regularization techniques introduce a penalty term to the regression equation, which helps control the coefficients and mitigate the impact of multicollinearity and overfitting. Ridge regression adds a penalty term proportional to the square of the magnitudes of the coefficients, thereby shrinking them towards zero. This regularization constraint reduces the impact of correlated predictors and improves the stability of the coefficient estimates. LASSO regression, on the other hand, employs a penalty term proportional to the absolute values of the coefficient estimates. This penalty has the additional property of performing variable selection by forcing some coefficients to exactly zero, effectively identifying the most relevant predictors. Understanding the effectiveness of Ridge, LASSO and Elastic Net regression in addressing multicollinearity and overfitting is crucial for researchers and practitioners working with regression models. This study aims to analyze Linear Regression Approach to solving the problem of multicollinearity and overfitting in Predictive Analysis. The implications of this study extend to various industries, including finance, healthcare, and marketing, where data-driven decision-making is paramount. By providing insights into how these advanced regression techniques can enhance model performance, the research can inform practitioners on selecting the appropriate methods for predictive analytics. This is particularly relevant in environments with complex datasets, enabling organizations to improve accuracy in forecasts, optimize resource allocation, and ultimately drive better business outcomes.

Related Works

Shuaibu et al. (2024) explores machine learning techniques for predicting agricultural yields in Nigeria, finding that the Decision Tree Regressor achieves a 72% accuracy. It emphasizes feature selection's role in enhancing model performance, highlighting machine learning's potential for improving food security in similar agroecological contexts.

Olatunde et al. (2024) explores the use of machine learning to analyze dietary patterns and their links to health outcomes, aiming to provide personalized dietary recommendations. By utilizing datasets from Kaggle and NHANES, the research emphasizes tailored dietary advice over generic guidelines. The findings advocate for personalized dietary management to mitigate chronic disease risks, showcasing machine learning's potential in improving nutritional science and public health.

Iliyasu et al. (2023) compares Multiple Linear Regression and Artificial Neural Networks for rainfall prediction in Katsina State, Nigeria. The ANN model outperforms MLR in accuracy, precision, and recall, underscoring the significance of accurate rainfall predictions for agricultural planning.

Chakraborty et al. (2023) using data from the U.S., this study applies ridge, LASSO, and elastic net modeling techniques to analyze human mobility factors. Ridge regression shows the best performance with the lowest RMSE, demonstrating its robustness against overfitting.

Chan et al. (2022) explores methods for mitigating multicollinearity in data analysis, emphasizing variable selection and modified estimators, including ridge and Lasso regression. It highlights the advantages of machine learning approaches, which often outperform traditional methods in handling multicollinearity. The authors suggest that combining variable selection with modified estimators can enhance model performance and interpretability.

Abdulmumini et al. (2022) presents a predictive model for child delivery modes using Random Forest, Neural Network, and Naïve Bayes. The Random Forest algorithm shows the best performance, indicating the potential of AI in improving maternal and child healthcare outcomes in Nigeria.

Kumar (2022) employs machine learning to predict COVID-19 trajectories in India, Brazil, Bangladesh, and Italy using LASSO regression. The model forecasts new deaths with an accuracy of 81.2% using Lasso and Ridge Regression, and 90% with linear regression. It suggests integrating support vector machines and ARIMA for improved outcomes.

Noora (2020) investigates multicollinearity in regression analysis, highlighting its impact on statistical significance among predictor variables. It presents three primary detection methods: correlation coefficients, variance inflation factor, and eigenvalue analysis. The study concludes that product attractiveness significantly influences customer satisfaction, with no evidence of multicollinearity among the variables.

Ogundokun et al. (2020) utilizing linear regression, analyzes the influence of travel history and contacts on COVID-19 cases in Nigeria. Findings indicate that travel history and contacts significantly increase infection rates,

reinforcing the importance of monitoring these factors in pandemic management.

Khan et al. (2022) evaluates various regression models for COVID-19 flare-ups, including linear and polynomial regressions. It assesses model effectiveness through metrics like MSE and RMSE, establishing a foundation for future research in machine learning applications.

The above reviewed literatures lack comprehensive exploration of how linear regression, a widely employed predictive modelling technique, can be strategically leveraged to effectively tackle both multicollinearity and overfitting. Existing researches often focus on singular aspects of these problems, leading to fragmented solutions that may not holistically address the intricate relationships between variables and the risk of overfitting. This research "Linear Regression Approach to Solving the Problem of Multicollinearity and Overfitting in Predictive Analysis" aims to bridge this gap by developing a novel linear regression approach that systematically addresses multicollinearity and overfitting concurrently.

MATERIALS AND METHODS

Data collection is the foundational step where relevant information is gathered from various sources to inform analysis. This is followed by Exploratory Data Analysis (EDA) to uncover patterns and insights, detecting multicollinearity to ensure the reliability of predictors, mitigating overfitting to enhance model generalization, and ultimately comparing different models to identify the most effective solution as maybe seen in Figure 1.

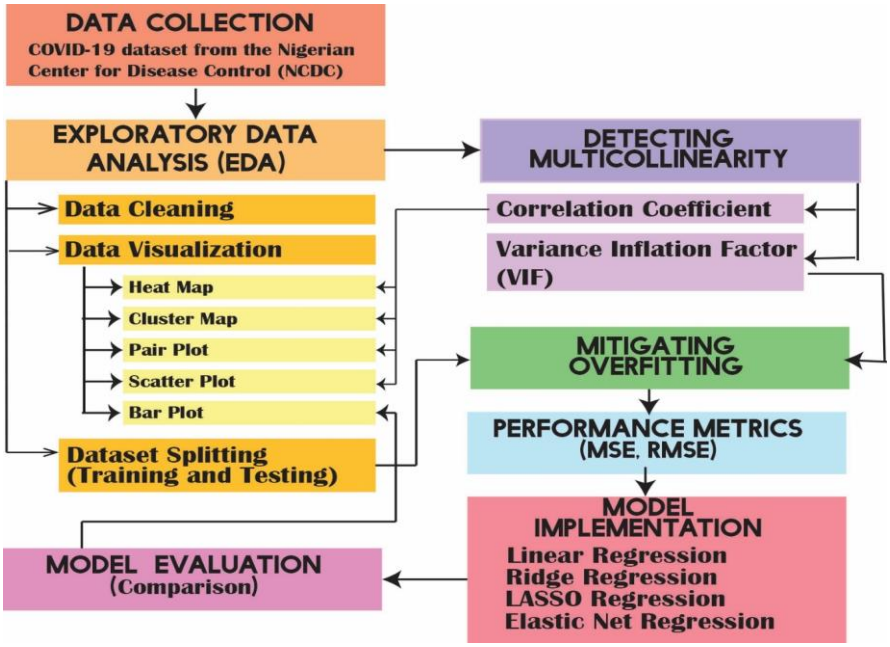

Figure 1: Workflow data frame

Materials

Data was collected from the NCDC repository (https://covid19.ncdc.gov.ng/), covering COVID-19 cases from January 2021 to February 2023. The dataset consists of states, date reported, total confirmed cases, last week confirmed cases, total recovery cases, Last week recovery cases, total death, last week death, active cases, total testing and last week testing as maybe found in Table 1. The dataset has 2960 rows /instances and 11 columns/features. Python and Jupyter Notebooks were employed as the software environment to conduct the analysis efficiently.

Table 1: Features in the Dataset

Feature Number	Feature Name	Data Type
1	Date Reported	datetime64
2	State	Object
3	Total Confirmed Cases	int64
4	Last Week Confirmed Cases	float64
5	Total Recovered Cases	float64
6	Last Week Recovered Cases	float64

7	Total Death Cases	int64
8	Last Week Death Cases	float64
9	Active Cases	float64
10	Total Tested Population	int64
11	Last Week Tested Population	float64

Methods

Detect Multicollinearity

Pairwise Scatterplot and Correlation Coefficients

A scatterplot visually represents the linear relationship between pairs of independent variables, highlighting potential correlations, the high correlation coefficient (close to 0.8 or 0.9) suggests collinearity among the variables (Mason & Perreault 1991 in Chan et al 2022), graphical method that signifies the linear relationship between pairs of independent variables. The correlation coefficient is calculated using the formula in equation (1)

$$r = \frac{n(\sum XY) - (\sum X)(\sum Y)}{\sqrt{[n\sum X^2 - (\sum X)^2][n\sum Y^2 - (\sum Y)^2]}} \quad (1)$$

Where, r = correlation coefficient, n = number of observations, X = first variable in the context, and Y = second variable in the context.

Variance Inflation Factor (VIF)

Variance inflation factor measures the impact of correlated independent variables on the variance of regression coefficient. VIF is calculated as shown in equation (2)

$$VIF = \frac{1}{1-R^2} = \frac{1}{tolerance} \quad (2)$$

Where R^2 is the coefficient of determination, the tolerance is simply the inverse of the VIF, the lower the tolerance, the more likely is the multicollinearity among the variables. VIF values indicate correlation levels: VIF < 5 suggests moderate correlation, while VIF ≥10 signals significant of multicollinearity issues (Belsley, 1991 in Noora, 2020).

Detect Overfitting

Dataset was split into training and testing sets, scores of both the training sets (internal data) and testing sets (External data) was obtained and compare, the performance of the models using Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) to evaluate the risk of overfitting across four regression models: linear regression, ridge regression, lasso regression, and elastic net regression.

Once fine-tuned, the model can predict unseen data using various regression techniques, represented by the formula for Linear Regression:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon \quad (3)$$

where:

y: The dependent variable (what you're trying to predict).

β_0: The y-intercept (the predicted value of y when all x values are zero).

$\beta_1 + \cdots + \beta_p$: Coefficients for each independent variable ($x_1 + \cdots + x_p$), representing the change in y for a one-unit change in each x.

ε: The error term (the difference between the actual and predicted values).

Ridge regression modifies Linear Regression by adding a penalty term, shown as in equation (4):

$$Loss = \sum(y_i - \hat{y}_i)^2 + \lambda \sum \beta_j^2 \quad (4)$$

Where:

The first term $\sum(y_i - \hat{y}_i)^2$: Represents the residual sum of squares (RSS), measuring the fit of the model.

The second term $\lambda \sum \beta_j^2$: Adds a penalty for large coefficients, where λ is a regularization parameter. This helps reduce overfitting by shrinking coefficients toward zero.

LASSO regression incorporates an absolute penalty, represented as in equation (5):

$$Loss = \sum(y_i - \hat{y}_i)^2 + \lambda \sum |\beta_j| \quad (5)$$

Where

The first term is the same as in Ridge Regression (RSS).

The second term $\lambda \sum |\beta_j|$ Adds a penalty based on the absolute values of the coefficients, promoting sparsity (driving some coefficients to zero). This aids in automatic feature selection.'

Elastic Net combines both penalties of Ridge and LASSO in equation (6):

$$Loss = \sum(y_i - \hat{y}_i)^2 + \lambda_1 \sum |\beta_j| + \lambda_2 \sum \beta_j^2 \quad (6)$$

Where:

Combines the penalties from both Ridge and LASSO.

The two regularization parameters $\lambda_1$and λ_2 control the strength of each penalty. This provides flexibility to handle multicollinearity and overfitting effectively.

Performance Metrics

To evaluate models, use Mean Squared Error (MSE):

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y})^2 \quad (7)$$

where y_i is the observed value of the ith attribute, $\hat{y}$ is the predicted value of the ith observation, and N denotes the number of samples in the dataset.

and Root Mean Squared Error (RMSE):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y})^2} \quad (8)$$

where y_i is the actual observation of the ith attribute, $\hat{y}$ is the estimated value of the ith observation, and n denotes the number of samples in the dataset.

RESULTS AND DISCUSSIONS

Exploratory Data Analysis

Key trends in total confirmed cases, recoveries, and deaths were visualized. High correlations were observed among independent variables, indicating multicollinearity in Figures 2 & 3.

Multicollinearity Analysis

A correlation matrix was generated to identify relationships between independent variables. Correlation coefficients revealed moderate to high correlations among several predictors. The Variance Inflation Factor (VIF) was calculated to quantify the severity of multicollinearity, with values exceeding 5 indicating potential issues.

Table 2: Correlation coefficients analysis among independent variables

	Comfirmed LastWeek	Recoveries Total	Recoveries LastWeek	Death Total	Death LastWeek	Active Cases	Testing Total	Testing LastWeek
ComfirmedLastWeek	1.000000	0.378341	0.332257	0.371092	0.342553	0.243770	0.310684	0.565855
RecoveriesTotal	0.378341	1.000000	0.264534	0.934213	0.192236	0.098724	0.941981	0.668652
RecoveriesLastWeek	0.332257	0.264534	1.000000	0.242995	0.318526	0.182327	0.202022	0.335444
DeathTotal	0.371092	0.934213	0.242995	1.000000	0.239091	0.092115	0.892433	0.662918
DeathLastWeek	0.342553	0.192236	0.318526	0.239091	1.000000	0.218572	0.120349	0.302107
ActiveCases	0.243770	0.098724	0.182327	0.092115	0.218572	1.000000	0.025108	0.094200
TestingTotal	0.310684	0.941981	0.202022	0.892433	0.120349	0.025108	1.000000	0.638131
TestingLastWeek	0.565855	0.668652	0.335444	0.662918	0.302107	0.094200	0.638131	1.000000

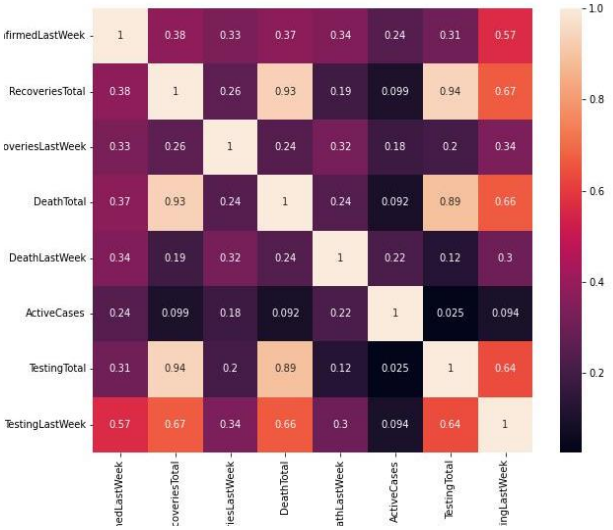

Figure 2: Heat map of independent variables

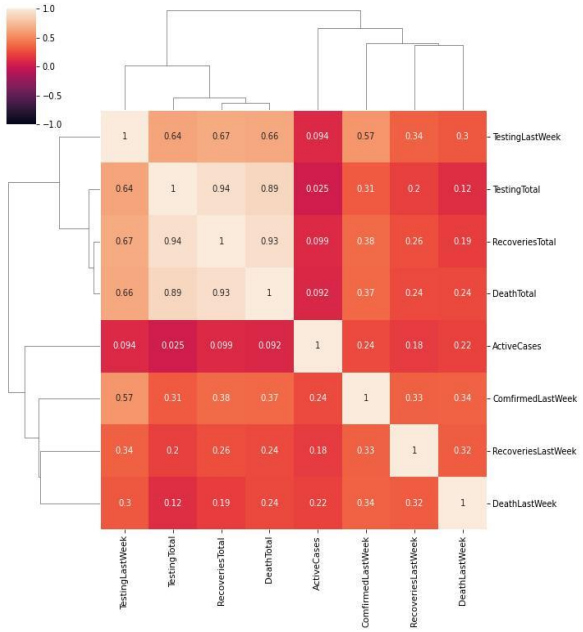

Figure 3: Cluster map of independent variables

Table 3: Variance Inflation Factor showing multicollinearity

	Feature	VIF
0	ComfirmedLastWeek	1.711579
1	RecoveriesTotal	13.998706
2	RecoveriesLastWeek	1.270166
3	DeathTotal	9.447172
4	DeathLastWeek	1.344083
5	ActiveCases	1.128090
6	TestingTotal	11.755597
7	TestingLastWeek	2.960876

Table 4: Variance Inflation Factor showing multicollinearity after feature reduction

	Feature	VIF
0	ComfirmedLastWeek	1.689024
1	RecoveriesLastWeek	1.250680
2	DeathTotal	7.475411
3	DeathLastWeek	1.342467
4	ActiveCases	1.118190
5	TestingTotal	6.954188
6	TestingLastWeek	2.960116

Overfitting Analysis

Four regression models-Linear Regression, Ridge Regression, Lasso Regression, and Elastic Net Regression were used. Each model was analyzed using external data and internal data of MSE and RMSE in two phases; first on dataset with all features and features reduction. The comparative analysis of the four regression models was also carried out.

Table 5: Mean Squared Error for models with all features

Models (MSE)	Internal data	External data	Difference
Linear Regression	583590.3620	939947.2396	356356.8777
Ridge Regression	583590.3620	939945.6569	356355.2950
Lasso Regression	583590.5715	939800.4388	356209.8673
Elastic Net Regression	583613.6920	938426.3681	354812.6761

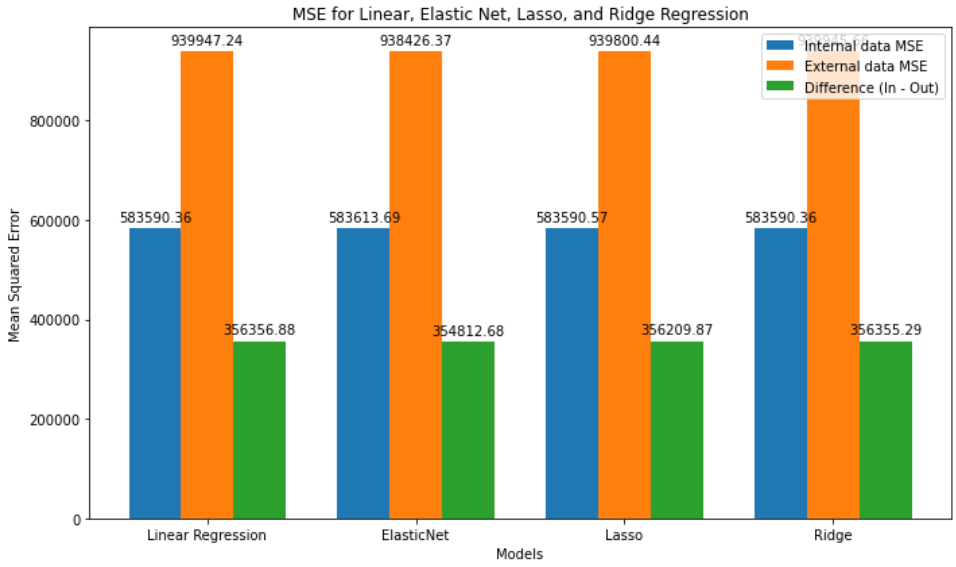

Figure 4: MSE for Linear, Elastic Net, Lasso and Ridge Regression with all features

Table 6: Root Mean Squared Error of models with all features

Models RMSE	Internal data	External data	Difference
Linear Regression	763.9309	969.5088	205.5779
Ridge Regression	763.9309	969.5079	205.5771
Lasso Regression	763.9310	969.4331	205.5021
Elastic Net Regression	763.9461	968.7241	204.7780

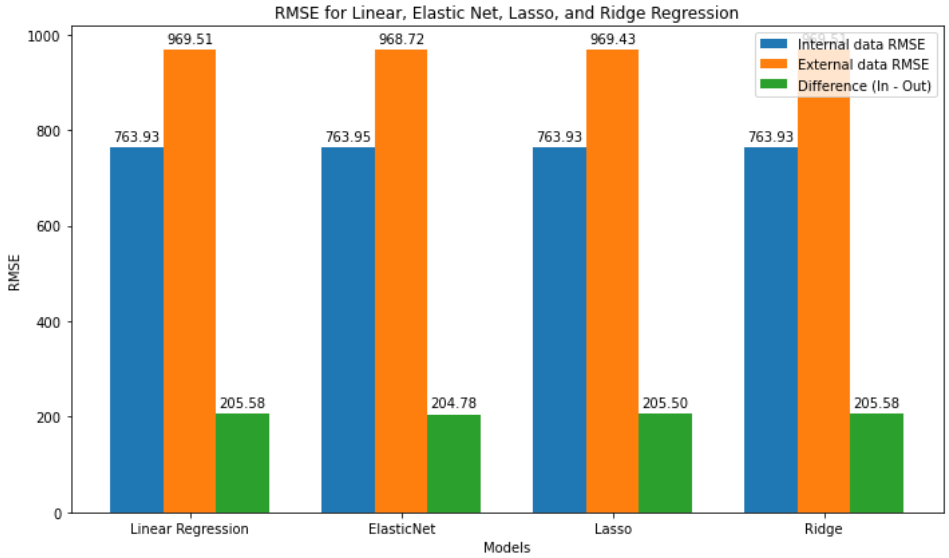

Figure 5: RMSE for Linear, Elastic Net, Lasso and Ridge Regression with all features

Table 7: Mean Squared Error of models After Feature reduction based on VIF

Models MSE	Internal data	External data	Difference
Linear Regression	11860038.9601	12631695.0299	771656.0697
Ridge Regression	11860038.9602	12631691.0249	771652.0646
Lasso Regression	11860039.1383	12631508.8831	771469.7448
Elastic Net Regression	11860136.7399	12627301.5711	767164.8312

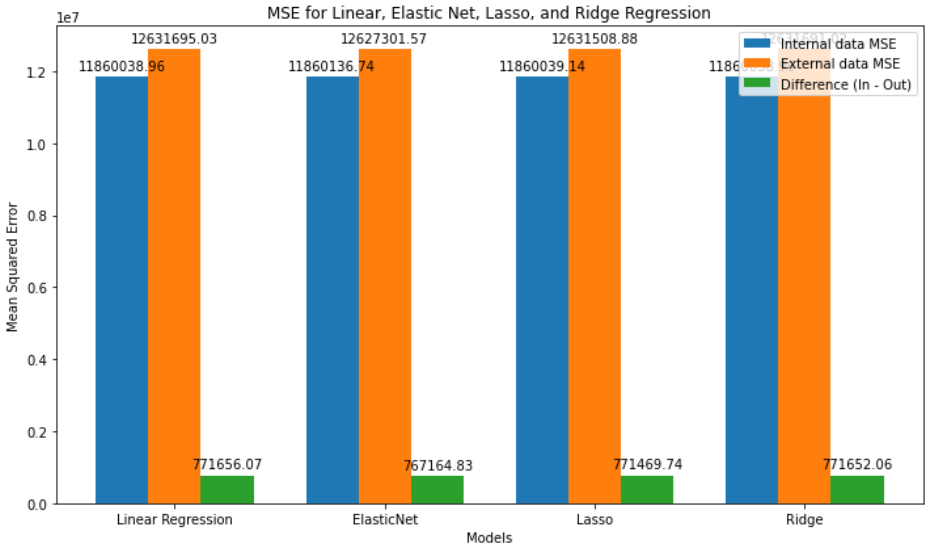

Figure 6: MSE for Linear, Elastic Net, Lasso and Ridge Regression after VIF

Table 8: Root Mean Squared Error of models after Feature reduction based on VIF

Models RMSE	Internal data	External data	Difference
Linear Regression	3443.8407	3554.1096	110.2689
Ridge Regression	3443.8407	3554.1090	110.2683
Lasso Regression	3443.8408	3554.0834	110.2427
Elastic Net Regression	3443.8549	3553.4915	109.6365

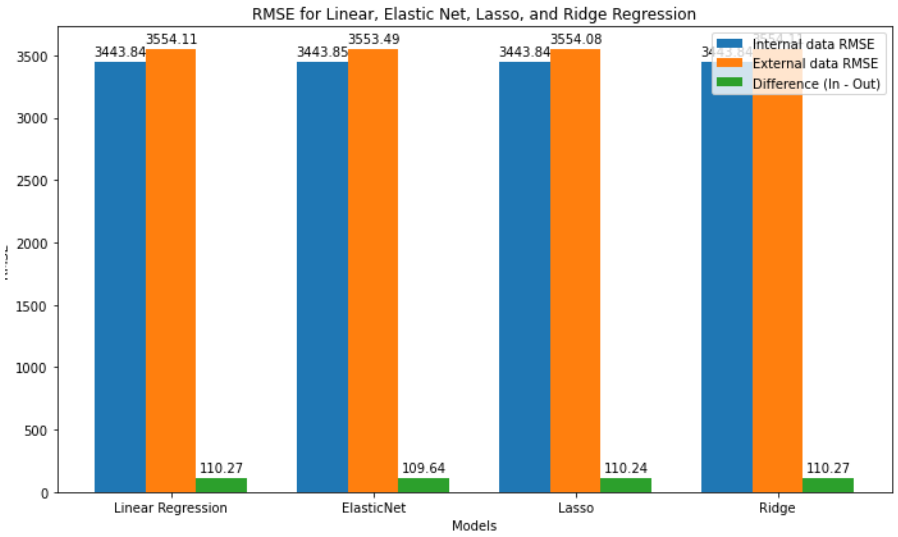

Figure 7: RMSE for Linear, Elastic Net, Lasso and Ridge Regression after VIF

Discussion

The findings of this study underscore the effectiveness of regression techniques in addressing multicollinearity and overfitting. Elastic Net Regression emerged as the most robust method, suggesting its utility in predictive modeling, particularly in public health contexts like COVID-19.

In Table 2, the correlation coefficient of overall total recovered cases with total tested cases has high correlation (0.941981), followed by Total death cases with total recoveries cases (0.934213). this is in agreement with A heatmap and cluster map were also plotted for visual representation of these relationships (see figures 2 & 3). According to Mason & Perreault 1991 in Chan et al 2022, a correlation coefficient of 0.8 suggests a strong positive relationship, indicating that changes in one variable are closely associated with changes in the other. They further note that a coefficient 0.9 implies a even more robust connection, underscoring the reliability of the relationship in predictive modeling.

A correlation matrix and VIF analysis revealed significant multicollinearity, particularly with total recovery and testing cases. Features with high VIF values were removed to improve model robustness.

Total Recovery Cases 13.998706 and Total Testing Cases' 11.755597 have high values of VIF, indicating that these two variables are highly correlated in Table 3. This is expected as the Total Recovery Cases does influence Total Testing Cases. Belsley (1991) in Noora (2020) explains that a variance inflation factor (VIF) exceeding 10 indicates significant multicollinearity among predictors, which distort regression estimates. Hence, considering these two features together still leads to a model with high multicollinearity. Since both correlation matrix and VIF indicate the presence of multicollinearity, we dropped the feature with the highest VIF values.

Furthermore, the comparative analysis of the four regression models—Linear Regression, Ridge Regression, Lasso Regression, and Elastic Net Regression—reveals significant insights regarding their performance under varying conditions. As noted by Khan et al. (2022), various regression models for COVID-19 flare-ups have highlighted the need to evaluate model effectiveness through metrics like MSE and RMSE. This study corroborates those findings, showing that while all models were capable of addressing multicollinearity and overfitting, Elastic Net consistently outperformed the others.

Performance with All Features

Initially, when evaluating the models with all features, the analysis shows that while the internal data Mean Square Error (MSE) values are closely aligned, indicating that all models fit the training data adequately, the external data MSE values demonstrate a notable distinction (Table 5). The Elastic Net Regression outperforms the others with the lowest external data MSE (938426.3681), suggesting it generalizes better to unseen data (Table 5). This performance is further corroborated by the Root Mean Square Error (RMSE) metrics in Table 6, where Elastic Net also exhibits the lowest external data values (968.7241).

The analysis highlights the presence of overfitting, as evidenced by the differences between internal data and external data MSE and RMSE values. However, the Elastic Net model shows the smallest discrepancies, indicating a better balance between fitting the training data and generalizing to new data.

In line with the literature, existing studies have noted the significance of feature selection in enhancing model

performance. For instance, Shuaibu et al. (2024) emphasize how appropriate feature selection can improve prediction accuracy in agricultural yield assessments. Similarly, Iliyasu et al. (2023) highlight the importance of accurate rainfall predictions for effective agricultural planning, underscoring the necessity of robust modeling techniques.

Impact of Feature Reduction

Upon applying feature reduction based on Variance Inflation Factor (VIF) to eliminate multicollinearity, the models still exhibit similar internal data MSE values, underscoring their capability to capture the relevant patterns in the training dataset (Table 7). Nevertheless, the external data MSE values increase significantly, particularly for Linear and Ridge Regression, while Elastic Net maintains its advantage with the lowest external data MSE (12627301.5711). This trend indicates that the feature reduction process may have compromised the generalization capabilities of some models, although Elastic Net continues to demonstrate resilience.

From Tables 7 and 8, the differences in MSE and RMSE values post-feature reduction reveal an increase in overfitting across all models, yet Elastic Net continues to show the smallest gap, reaffirming its robustness in handling new data.

Kumar (2022) and Abdulmumini et al. (2022), have shown that feature selection can lead to improved model robustness. In our analysis, while models with all features demonstrated superior performance, the application of feature reduction led to increased external data MSE values, particularly for Linear and Ridge Regression. This aligns with the observations made by Chakraborty et al. (2023), who noted that model simplicity could sometimes compromise predictive capabilities.

Comparative Insights

When comparing models with all features to those after feature reduction, it becomes evident that models retaining all features exhibit superior performance in both internal data and external data evaluations. The lower MSE and RMSE values for these models suggest that more features contribute beneficially to the model's ability to learn from the training data and generalize effectively.

The findings emphasize the Elastic Net Regression model as the most effective among the tested models, particularly in its ability to generalize to unseen data after both feature inclusion and reduction. These insights are crucial for practitioners in selecting the appropriate modeling approach, especially in contexts where model generalization is critical for predictive accuracy. Future work may focus on exploring other techniques for feature selection and model tuning to further enhance performance.

CONCLUSION

This study underscores the effectiveness of regression techniques in addressing multicollinearity and overfitting. Elastic Net Regression emerged as the most robust method, suggesting its utility in predictive modeling, particularly in public health contexts like COVID-19. Future research should explore broader datasets and advanced machine learning techniques to further enhance predictive accuracy. Data Quality: Improve the accuracy and consistency of COVID-19 data reporting. Further Research: Investigate factors affecting regional variations in COVID-19 outcomes and the impact of public health interventions.

REFERENCES

Abdulmumini A. K., Obunadike G.N., & Jiya E. A. (2022). Predictive Model For Child Delivery. Fudma Journal of Sciences, 6(1), 141 - 145. https://doi.org/10.33003/fjs-2022-0601-885

Belsley, D.A., (1991) Conditioning diagnostics: Collinearity and weak data in regression, John Wiley & Sons, Inc., New York.

Chakraborty, M., Shakir Mahmud, M., Gates, T. J., & Sinha, S. (2023). Analysis and prediction of human mobility in the United States during the early stages of the COVID-19 pandemic using regularized linear models. *Transportation research record*, *2677*(4), 380-395.

Chan, J.Y.-L.; Leow, S.M.H.; Bea, K.T.; Cheng,W.K.; Phoong, S.W.; Hong, Z.-W.; Chen, Y.-L. (2022) Mitigating the Multicollinearity Problem and Its Machine Learning Approach: A Review. Mathematics, 10, 1283. https://doi.org/10.3390/math10081283

Fox, J. (2015). *Applied regression analysis and generalized linear models*. Sage publications.

Herawati, N., Nisa, K., Setiawan, E., & Nusyirwan, T. (2018). Regularized Multiple Regression Methods to Deal with Severe Multicollinearity. *International Journal of Statistics and Applications,* *8*(4), 167-172. https://doi.org/10.5923/j.statistics.20180804.02

Iliyasu U., Obunadike G.N., & Jiya E. A. (2023) Rainfall Prediction Models for Katsina State, Nigeria: Machine Learning Approach. International Journal of Science for Global Sustainability, Vol. 9 No2, pp 151 – 157. DOI: https://doi.org/10.57233/ijsgs.v9i2.473

Khan, M.A., Raza, A., Awais, M., & Iqbal, M. (2022) A survey of machine learning-based methods for covid-19medical image analysis. Medical &Biological Engineering & Computing, 60(1), 1-21. https://doi.org/10.1007/s11517-021-02525-2

Kumar, A., Jain, M., Gupta, A., Chaudhary,P., & Gupta R. (2022) development of machine learning model to Predict COVID-19 mortality : Application of Ensemble Model and Regarding Feature Impacts, PMC. https://doi.org/10.1007/s11517-022-02387-6

Mason, C.H.; Perreault, W.D., Jr. (1991) Collinearity, power, and interpretation of multiple regression analysis. J. Mark. Res. 28, 268–280.

Noora, S. (2020). Detecting Multicollinearity in Regression Analysis. *American Journal of Applied Mathematics and Statistics*, *8*(2), 39-42. https://doi.org/10.12691/ajams-8-2-1

Ogundokun, R. O., Lukman, A. F., Kibria, G. B., Awotunde, J. B., & Aladeitan, B. B. (2020). Predictive modelling of COVID-19 confirmed cases in Nigeria. *Infectious Disease Modelling*, *5*, 543-548

Olutunde, T., Ani, C. L., & Adesue, G. A. (2024). Leveraging Machine Learning for Personalized Dietary Recommendations, Nutritional Patterns, and Health Outcome Predictions. Journal of Science Research and Reviews, 1(2), 43-56. https://doi.org/10.70882/josrar.2024.v1i2.40

Shuaibu N., Obunadike G. N., & Jamilu B. A. (2024). Crop Yield Prediction Using Selected Machine Learning Algorithms. FUDMA Journal of Sciences, 8(1), 61 - 68. https://doi.org/10.33003/fjs-2024-0801-2220