Kasu Journal of Computer Science Vol.2 No.2 [June, 2025], pp. 133-143
https://doi.org//10.47514/kjcs/2025.2.2.001

Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

Effects of Combination of Novel Feature Engineering Technique with Random Forest Feature Selection and Variance Inflation Factor Reduction on Prediction Accuracy

Johnson Oluwaseyi Fashanu[1], Usman Ismail Abdulmalik[2] and Mustapha Kassim[3]
[1]Department of Computer Science, Federal Polytechnic, Kaura Namoda, Nigeria
[2]Department of Computer Science, Federal Polytechnic, Kaura Namoda, Nigeria
[3]Department of Computer Science, Federal Polytechnic, Mubi, Nigeria
Johnson.cs@fedponam.edu.ng[1], usmaniamailbox@yahoo.com[2], mustaphakassim@rocketmail.com[3]

Abstract

Feature engineering plays a crucial role in improving machine learning model performance. This study presents a novel feature engineering technique using Fibonacci linear and polynomial equations of degree 7 to enhance predictive accuracy, the linear equations are employed to retain the structure of the original glass dataset which determines which of the series of polynomial equations will be used to transform the dataset. The aim is to compare the results of this technique with further combining it with statistical method. The engineered features were combined with the original features, random forest feature selection was used to select features from the combined feature and was evaluated, this achieved 81% accuracy. The polynomial equation used are variant polynomial regression which are susceptible to multicollinearity, the random forest selected features was further subjected to reduction using Variance Inflation Factor (VIF) to eliminate multicollinearity while maintaining interpretability, this achieved an accuracy of 86%. The final model achieved significant accuracy improvements while ensuring reduced redundancy among features. The findings highlight the comparative effectiveness of combining Fibonacci-derived polynomial feature expansion, Random Forest selection; and Fibonacci-derived polynomial feature expansion, Random Forest selection with VIF reduction in predictive modelling.

Keywords: Glass classification, Random Forest, Fibonacci sequence, Feature engineering, Variance inflation factor

1. Introduction

Machine learning models rely heavily on effective feature engineering to optimize performance. Traditional feature transformation techniques often introduce multicollinearity, leading to model instability. To address this, we propose a novel feature engineering approach that leverages Fibonacci polynomial equations to generate transformed features. However, increasing the number of features can introduce redundancy and correlation issues, necessitating feature selection techniques such as Random Forest importance ranking, and VIF reduction. This study investigates the impact of combining these techniques on prediction accuracy. The objective is to assess how Fibonacci-derived polynomial features (degree 7) influence machine learning models and how Random Forest selections as well as VIF help in refining the feature set for optimal performance. A new LASSO based feature selection technique which selects important features picked out significant factors was used in a study, this method (M-estimate-based LASSO) works well than conventional LASSO. (Muthukrishnan, & James. 2024). A study argues that classical VIF based on ordinary least squares are sensitive to outliers and leverage points, therefore an enhanced traditional VIF was introduced using a robust estimator that mitigates the influence of good leverage points. (Ekiz, 2023).

2. Related Works

2.1 Feature Engineering in Machine Learning

Feature engineering is the process of transforming raw data into meaningful representations that enhance predictive performance. Polynomial features have been widely used for capturing non-linear relationships in data (Fashanu, Abdulmalik & Kassim 2025; Ameen & Fashanu, 2025). However, higher-degree polynomials often lead to multicollinearity, necessitating feature selection methods to improve model generalization (Nelson, 2020). Feature engineering is pivotal in enhancing machine learning model performance by transforming raw data into representations that capture complex, non-linear relationships (Ameen & Fashanu, 2025; Fashanu, Abdulmalik & Kassim 2025; James, Witten, Hastie, & Tibshirani, 2013; Hastie, Tibshirani, & Friedman, 2009). One common approach is to generate polynomial features; however, doing so often introduces multicollinearity. High multicollinearity, quantified using the Variance Inflation Factor (VIF), can inflate the variance of coefficient estimates and lead to unstable models (O'Brien, 2007). Glass is a transparent, non-crystalline, amorphous solid that are produced in various varieties which can be classified using various machine learning algorithms. Examples of the algorithms used includes Support Vector Machine (SVM) and Random Forest among others, in a study where Principal Component Analysis (PCA) was used for dimensionality reduction for glass classification, the Random Forest algorithm fared best with 79.62% accuracy (Nujumudeen, Faizal & G, Adarsh 2022). The public and companies in this era of industrial revolution are familiarizing themselves with data science, a key part of data science is data mining, it involves collecting information from large dataset to retrieve patterns. A study using glass dataset from UCI machine learning repository for K-Nearest Neighbour (KNN) algorithm achieved an accuracy of 72.90%. (Suppa, 2023).

2.2 Fibonacci-Based Feature Engineering

The integration of Fibonacci-derived polynomial features adds a novel twist to traditional feature engineering. The Fibonacci sequence has long intrigued researchers due to its appearance in various natural and scientific phenomena (Ameen & Fashanu, 2025; Livio, 2003). Recent work has begun to explore its potential in generating meaningful transformations that capture hidden patterns in data (Fashanu, Abdulmalik, & Kassim, 2025; Cinar & Koklu, 2019). By applying Fibonacci-based transformations to generate higher-degree polynomial features (up to degree 7 in our study), it is possible to enhance the feature space in a structured manner. This enriched feature set, when subsequently refined by Random Forest selection and VIF, leads to a more parsimonious and interpretable model. This study extends the Fibonacci approach to polynomial feature expansion, testing its effects on model accuracy.

2.3 Variance Inflation Factor (VIF) for Feature Selection

VIF is a well-known technique for detecting and removing highly correlated features to prevent multicollinearity. A VIF score above 10 indicates a strong correlation, which can lead to unstable regression models. Removing such features improves the interpretability and robustness of the model. (Cheng et al., 2022)

2.4 Random Forest Feature Selection

Random Forest is widely used for ranking feature importance in classification and regression tasks. It provides an interpretable method for selecting the most influential features while maintaining predictive performance (Fashanu, Abdulmalik, & Kassim, 2025; Nujumudeen, Faizal & G, Adarsh 2022). Combining VIF and Random Forest selection ensures that only relevant, uncorrelated features are retained. Random Forest models have emerged as a robust and interpretable solution for both classification and feature selection tasks (Breiman, 2001). By ranking features based on their contribution to prediction accuracy, Random Forests help identify the most influential variables even in complex, high-dimensional spaces (Biau, 2012). When combined with VIF-based reduction, these methods can effectively prune redundant features while preserving or even improving overall model performance. Visualization tools such as correlation matrices are essential for diagnosing relationships among features. Well-constructed correlation matrices help identify clusters of highly correlated variables, guiding the application of VIF reduction (Hair, Black, Babin, & Anderson, 2010). Furthermore, confusion matrices provide a detailed breakdown of classifier performance by comparing predicted versus actual labels, thus offering insights beyond overall accuracy (Sokolova& Lapalme, 2009). Metrics derived from confusion matrices precision, recall, and F1-score are crucial for a nuanced evaluation of model effectiveness. Together, these studies underscore the importance of a combined approach leveraging advanced feature engineering, statistical diagnostics, and robust classification methods to achieve superior predictive performance while maintaining model interpretability. The present study builds on this foundation by examining the effect of combining Fibonacci-derived feature expansion with Random Forest feature selection and VIF reduction and on prediction accuracy. Therefore, the study aims to fill the gap in feature engineering by introducing a novel approach using Fibonacci polynomial equations of degree 7 to enhance predictive accuracy while addressing multicollinearity issues. Traditional polynomial feature engineering methods often introduce excessive multicollinearity, leading to model instability and reduced interpretability. While feature selection techniques like Random Forest and Variance Inflation Factor (VIF) have been used separately, their combined impact on optimizing feature space remains underexplored. This study bridges this gap by integrating Fibonacci-based polynomial feature expansion with Random Forest selection and VIF reduction, ensuring improved model accuracy, reduced redundancy, and enhanced interpretability in predictive modelling.

3. Methodology

3.1 Dataset Description

The study utilizes glass classification donated by USA Forensic Science Service, it was donated on 8/31/1987, 6 types of glass based on their oxide composition were identified in the dataset hence multivariate which contains 214instances with 9 original features. Samples of the dataset are shown in Figure 11. The dataset has no missing values. The link to the dataset on UC Irvine Machine Learning Repository is found below:
https://archive.ics.uci.edu/dataset/42/glass+identification

3.2 Fibonacci Polynomial Feature Engineering

A total of 9 new features were generated using Fibonacci-derived polynomial equations of degree 7. These features are combined with the original features which increased the feature space to 18. The transformation follows the equation:

$aK + b$ (for linear equations)

$((aK + b)K^{j}) - ((cK + d)K^{j-1}) - (6(aK + e)K^{j-2}) + (5(cK + d)K^{j-3}) + (10(aK + e)K^{j-4}) - (6(cK + d)K^{j-5}) - (4(aK + e)K) + (cK + d)$ (for polynomial equations)

Kasu Journal of Computer Science Vol.2 No.2 [June, 2025], pp. 133-143
https://doi.org//10.47514/kjcs/2025.2.2.001
Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

Where ‘a’, ‘b’, ‘c’, ‘d’, and ‘e’ are Fibonacci sequence-derived constants, j = 7, and K represents the feature value.

3.3 Feature Selection Using VIF and Random Forest

Random Forest Selection: The 18 features were ranked based on importance, with17 top features selected and used in the model.

VIF Reduction: Features with VIF > 10 were removed to reduce multicollinearity. The features were reduced to 14 features. The features importance figures before VIF and after VIF are shown in Figure 1 and Figure 2 respectively.

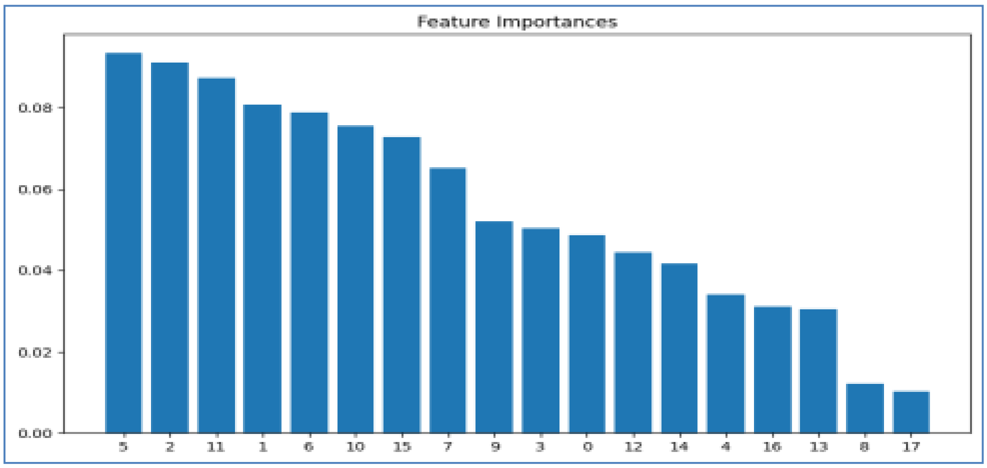


Figure 1: Feature Importance showing 18 selected features before VIF

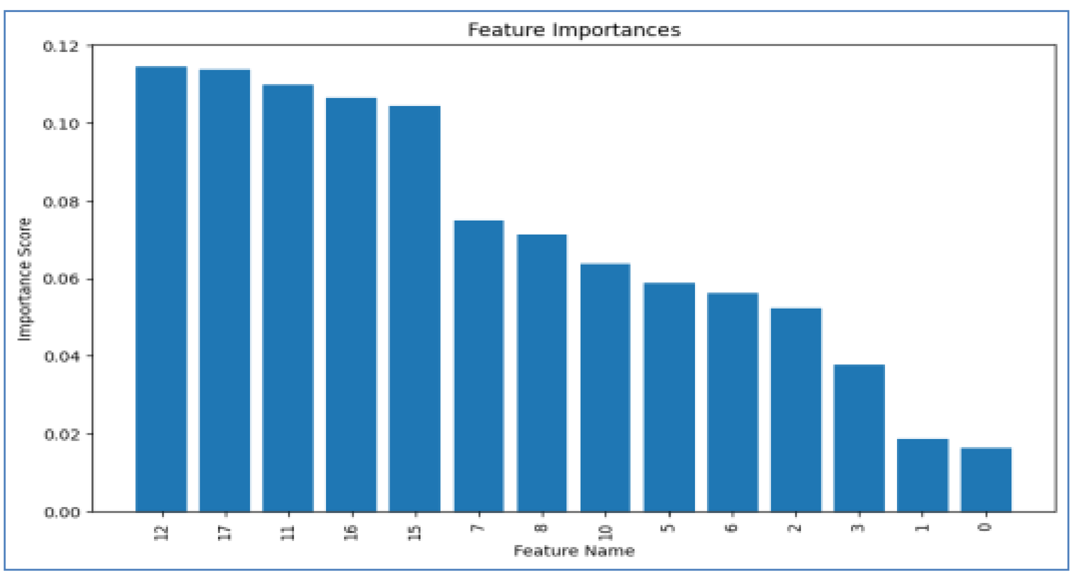


Figure 2: Feature Importance showing 14 selected features after VIF

3.4 Model Training and Evaluation

The model for experimental research with Random Forest classifier feature selection and VIF reduction is shown in Figure 3. The target variable which is multivariate was converted to numeric values, the instances of the dataset were transformed using linear and polynomial equations, combined with the original dataset before applying Random Forest feature selection, the selected features were ran on the model, thereafter, VIF was applied and the new selected dataset were ran. The results were compared.

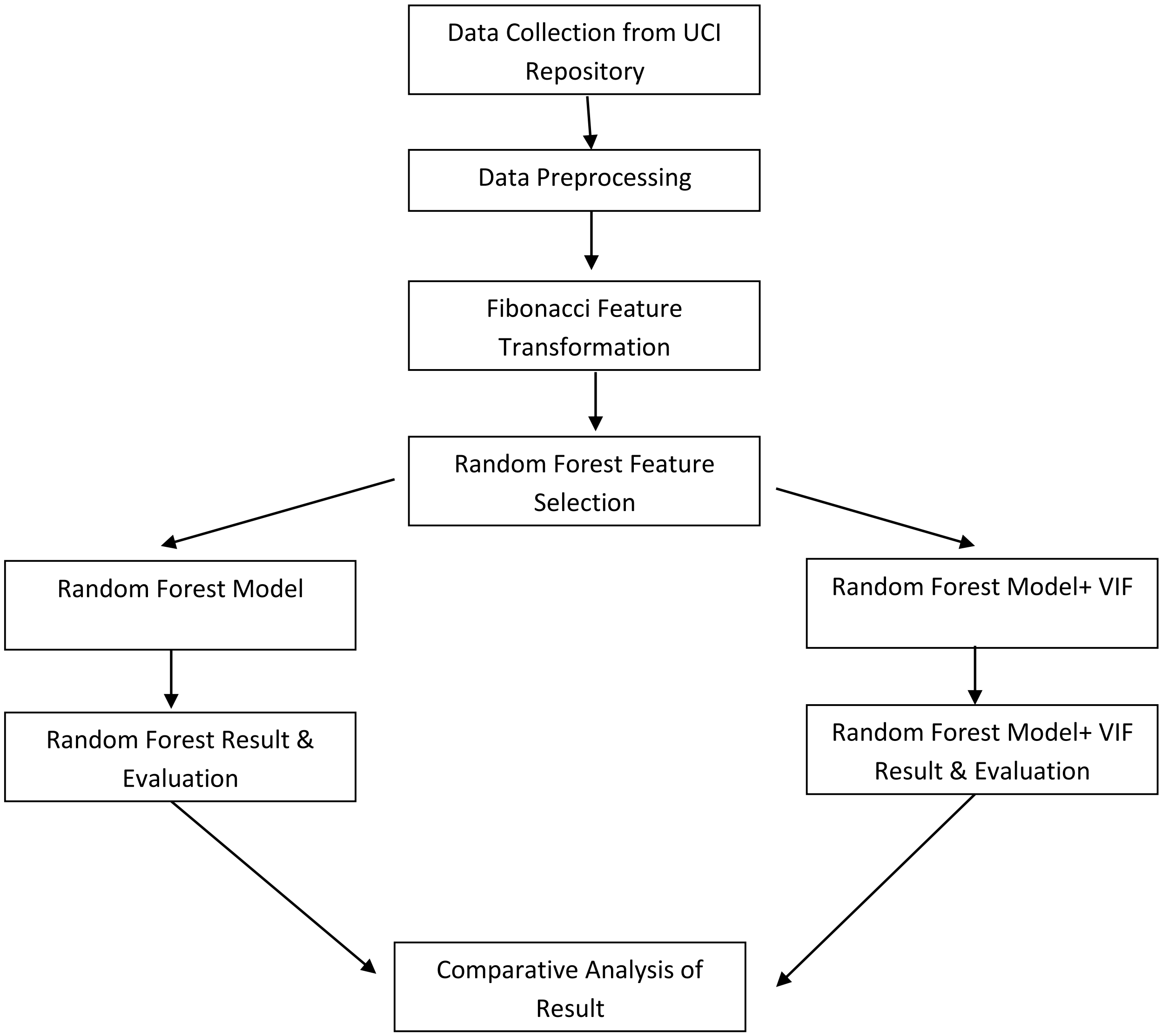


Figure 3: Framework with Random Forest feature selection and VIF

A Random Forest classifier was trained and performance was evaluated for features selected before and after VIF. The model was assessed based on accuracy, Precision, Recall, macro average, weighted average, and F1-Score. The Figure 4 and Figure 5 displace the result respectively.

	precision	recall	f1-score	support
0	0.73	0.73	0.73	11
1	0.91	0.71	0.80	14
2	1.00	0.75	0.86	4
3	0.89	1.00	0.94	8
4	1.00	1.00	1.00	3
5	0.50	1.00	0.67	3
accuracy			0.81	43

Figure 4: Performance Metrics after Random Forest Feature Selection

	precision	recall	f1-score	support
0	0.75	0.82	0.78	11
1	1.00	0.71	0.83	14
2	1.00	1.00	1.00	4
3	0.89	1.00	0.94	8
4	1.00	1.00	1.00	3
5	0.60	1.00	0.75	3
accuracy			0.86	43

Figure 5: Performance Metrics after VIF

To assess the model's performance as shown in Figure 4 and Figure 5, the following evaluation metrics were used:

i. **Accuracy:** Measures the proportion of correctly classified glass. Accuracy = (TP+TN) ÷ (FP+FN+TP+TN), where: TP (True Positives) = correctly classified glass of a specific variety. TN (True Negatives) = correctly classified glass of other varieties. FP (False Positives) = incorrectly classified glass. FN (False Negatives) = glass misclassified as another variety.

ii. **Precision and Recall:** Precision measures the proportion of correctly classified glass among predicted positive instances. Recall measures the proportion of actual positive glass that was correctly classified.

iii. **F1-Score:** Balances Precision and Recall, calculated as: F1= 2× ((Precision x Recall) ÷ (Precision + Recall))

The confusion matrix for the glass classification which displays the prediction of each glass type before and after VIF reduction is shown in Figure 6 and Figure 7 respectively. Furthermore, samples of dataset after feature engineering and VIF reduction is shown in Figure 8.

Kasu Journal of Computer Science Vol.2 No.2 [June, 2025], pp. 133-143
https://doi.org//10.47514/kjcs/2025.2.2.001

Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

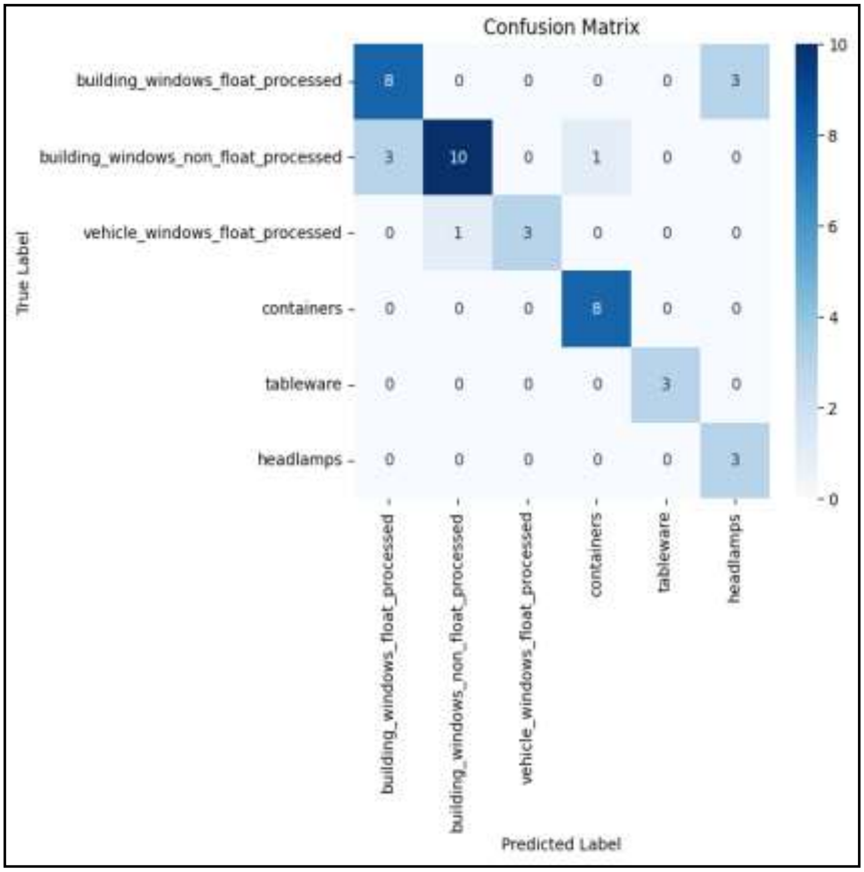


Figure 6: Confusion Matrix of Glass types after Random Forest Feature Selection.

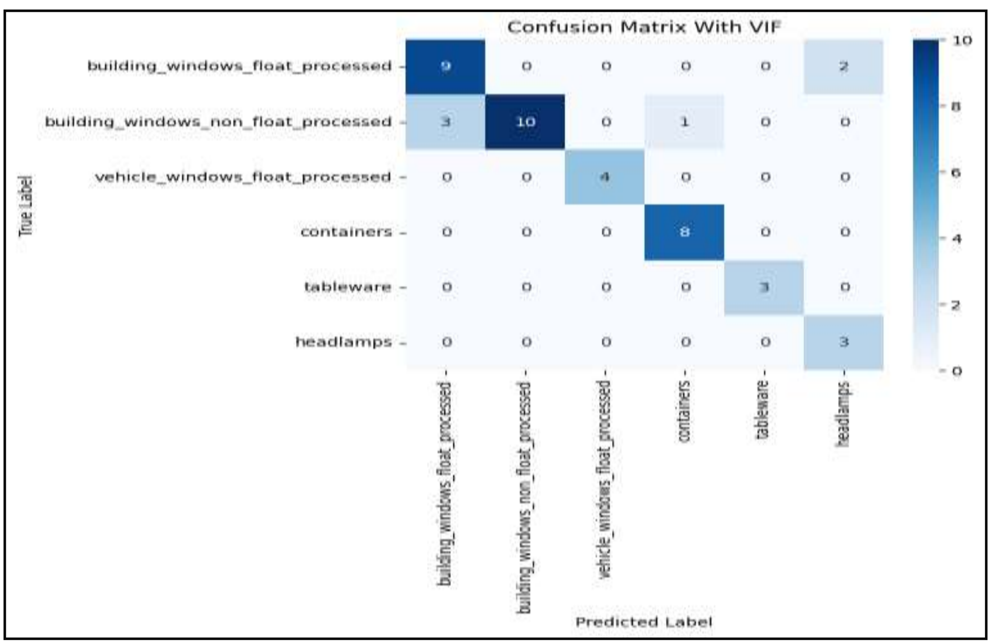


Figure 7: Confusion Matrix of Glass types after VIF Selection.

Kasu Journal of Computer Science Vol.2 No.2 [June, 2025], pp. 133-143
https://doi.org//10.47514/kjcs/2025.2.2.001
Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

3	-0.806923	1.926046	-2.030349	2.021196	0.894259	-0.895953	-0.128932
4	-0.459646	0.439388	0.642576	0.823810	-0.698640	0.249795	-0.751777

	7	8	9	10	11	12	13
0	-0.337611	0.492399	0.305377	-0.577487	0.054464	-0.134998	0.502166
1	0.947042	-0.601110	0.397362	1.953753	-0.413057	0.204641	0.911371
2	-0.337611	1.088858	-0.997165	-0.530089	0.114670	-0.135858	-0.530747
3	0.908113	0.293579	0.808176	2.294690	-0.413057	0.282434	0.911371
4	-0.337611	-0.601110	0.459789	0.240523	0.153868	-0.140356	-0.735929

	14	15	16	17
0	-0.076708	-0.216123	-0.101212	0.028153
1	-0.076701	-0.177371	-0.104381	0.364426
2	-0.076706	-0.065612	-0.101212	-0.147470
3	-0.076701	-0.192719	-0.104846	-2.675399
4	-0.076704	-0.239652	-0.101212	0.364426

Figure 8: Samples of derived and original Experimental Dataset.

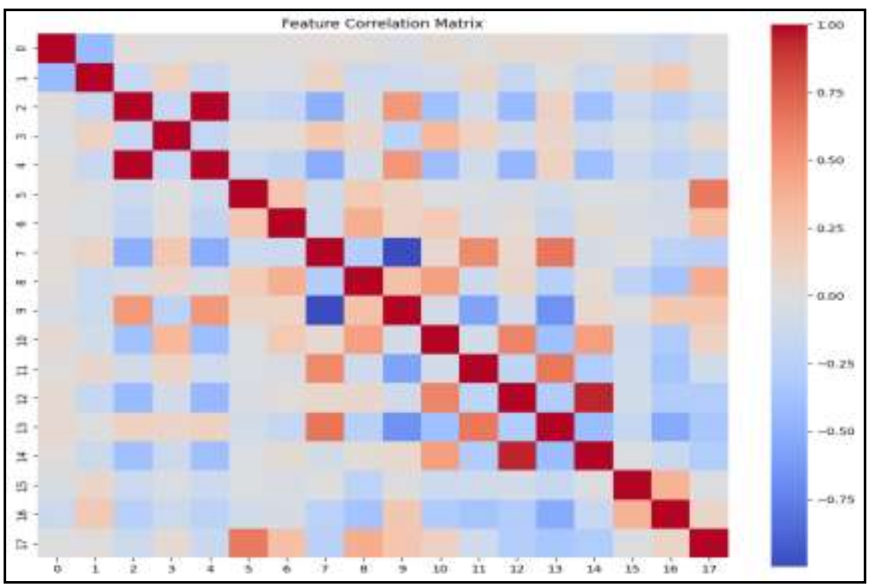
4. Results and Discussion

4.1 Correlation Matrix Analysis

The initial correlation matrix showed higher multicollinearity among transformed features when Random Forest Feature Selection was applied. This claim is supported by the dominance of dark colours in Figure 9 compared to Figure 10 after VIF-based reduction, where the correlation matrix displayed lighter colours, indicating reduced feature redundancy.

4.2 Model's Experimentation Set Up

The dataset has 9 features as shown in Figure 11; these were transformed and combined with the original dataset which expanded the feature spaces to 18, Random Forest feature selection was setup to features importance ranking of the combined 18 features as shown in Figure 1, thereafter VIF reduced the features to 14 as shown in Figure 2.



Kasu Journal of Computer Science Vol.2 No.2 [June, 2025], pp. 133-143
https://doi.org//10.47514/kjcs/2025.2.2.001
Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

Figure 9: Correlation Matrixes after Random Forest Feature Selection

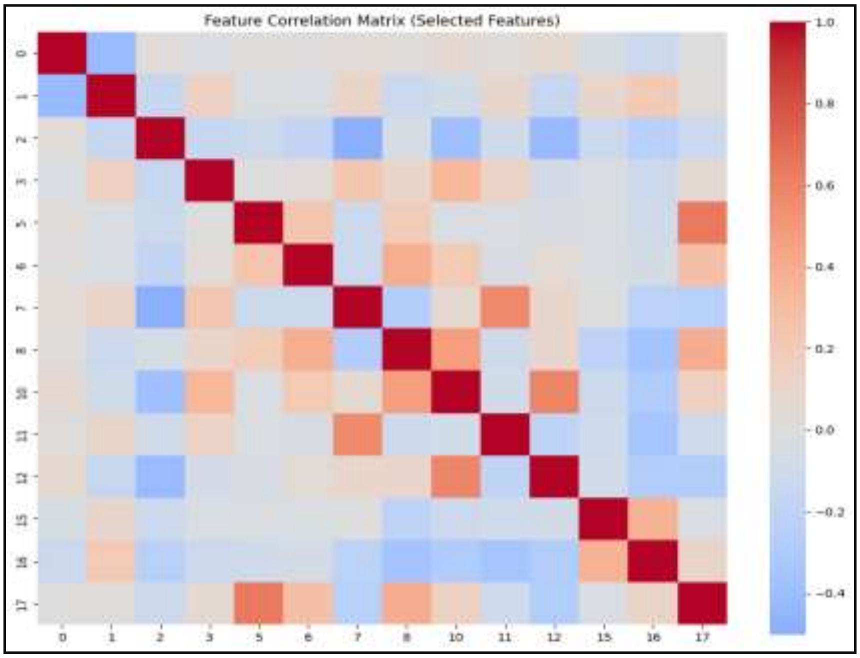
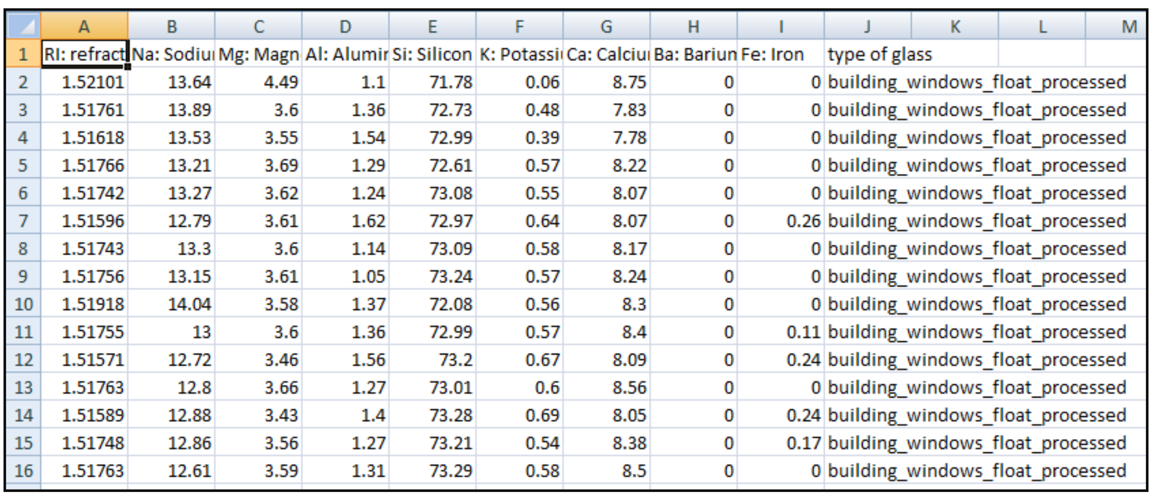


Figure 10: Correlation Matrix after VIF



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	RI: refract	Na: Sodiu	Mg: Magn	Al: Alumir	Si: Silicon	K: Potassi	Ca: Calciu	Ba: Bariun	Fe: Iron	type of glass			
2	1.52101	13.64	4.49	1.1	71.78	0.06	8.75	0	0	building_windows_float_processed			
3	1.51761	13.89	3.6	1.36	72.73	0.48	7.83	0	0	building_windows_float_processed			
4	1.51618	13.53	3.55	1.54	72.99	0.39	7.78	0	0	building_windows_float_processed			
5	1.51766	13.21	3.69	1.29	72.61	0.57	8.22	0	0	building_windows_float_processed			
6	1.51742	13.27	3.62	1.24	73.08	0.55	8.07	0	0	building_windows_float_processed			
7	1.51596	12.79	3.61	1.62	72.97	0.64	8.07	0	0.26	building_windows_float_processed			
8	1.51743	13.3	3.6	1.14	73.09	0.58	8.17	0	0	building_windows_float_processed			
9	1.51756	13.15	3.61	1.05	73.24	0.57	8.24	0	0	building_windows_float_processed			
10	1.51918	14.04	3.58	1.37	72.08	0.56	8.3	0	0	building_windows_float_processed			
11	1.51755	13	3.6	1.36	72.99	0.57	8.4	0	0.11	building_windows_float_processed			
12	1.51571	12.72	3.46	1.56	73.2	0.67	8.09	0	0.24	building_windows_float_processed			
13	1.51763	12.8	3.66	1.27	73.01	0.6	8.56	0	0	building_windows_float_processed			
14	1.51589	12.88	3.43	1.4	73.28	0.69	8.05	0	0.24	building_windows_float_processed			
15	1.51748	12.86	3.56	1.27	73.21	0.54	8.38	0	0.17	building_windows_float_processed			
16	1.51763	12.61	3.59	1.31	73.29	0.58	8.5	0	0	building_windows_float_processed			

Figure 11: Original Glass Dataset from UCI Machine Learning Repository

Kasu Journal of Computer Science Vol.2 No.2 [June, 2025], pp. 133-143
https://doi.org//10.47514/kjcs/2025.2.2.001

Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

Table 1: Performance Metric with Random Forest Feature Selection

Class	Precision	Recall	f1-score	Support
0	0.73	0.73	0.73	11
1	0.91	0.71	0.80	14
2	1.00	0.75	0.86	4
3	0.89	1.00	0.94	8
4	1.00	1.00	1.00	3
5	0.50	1.00	0.67	3
Accuracy			0.81	43

Table 2: Performance Metric with VIF Reduced Features

Class	Precision	Recall	f1-score	Support
0	0.75	0.82	0.78	11
1	1.00	0.71	0.83	14
2	1.00	1.00	1.00	4
3	0.89	1.00	0.94	8
4	1.00	1.00	1.00	3
5	0.60	1.00	0.75	3
Accuracy			0.86	43

The combination of Fibonacci-derived polynomial features, Random Forest selection and VIF reduction resulted in significant accuracy improvements. By removing redundant and highly correlated features, the model achieved better interpretability. While Random Forest Feature Selection achieved accuracy of 81%, VIF further increased the accuracy to 86%, improving performance by 5%. The correlation matrix analysis confirmed the effectiveness of VIF reduction, ensuring a stable and efficient feature set. Key observations of the model are stated below:

- Fibonacci polynomial features enhanced non-linear relationships.
- Random Forest selection retained the most important predictive features.
- VIF removal reduced redundancy and improved model stability.

5. Conclusion

This study demonstrated that Fibonacci-derived polynomial feature engineering, combined with Random Forest selection and VIF reduction, significantly improves machine learning model accuracy while maintaining interpretability. Random Forest feature selection achieved accuracy of 81%, VIF reduction increased the accuracy by 5%, indicating improvement in performance after dropping of redundant features. The combination of derived features and the original feature increased the feature space, the polynomial ensured capturing nonlinear relationship amongst features. Future work will explore extending this method to other datasets and classification problems.

Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

References

Ameen, A.O., Fashanu, J.O. (2025). Linear Regression with Fibonacci-derived Polynomials for Temperature Prediction Model. Dicov Data 3, 7 (2025). https://doi.org/10.1007/s44248-025-00025-7

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://link.springer.com/article/10.1023/A:1010933404324

Biau, G. (2012). Analysis of a random forests model. Journal of Machine Learning Research, 13, 1063–1095.

Cheng J, Sun J, Yao K, Xu M, Cao Y. (2022). A variable selection method based on mutual information and variance inflation factor. *Spectrochim Acta A Mol Biomol Spectrosc*. 10.1016/j.saa.2021.120652. Epub 2021 Nov 20. PMID: 34896682.

Cinar, I., &Koklu, M. (2019). Classification of rice varieties using artificial intelligence methods. International *Journal of Intelligent Systems and Applications in Engineering*, 7(3), 188–194. https://doi.org/10.18201/ijisae.2019355381

Ekiz, O. U. (2023). An improved robust variance inflation factor: Reducing the negative effects of good leverage points. Kuwait Journal of Science, 50(2A). https://doi.org/10.48129/kjs.15533

Fashanu, J. O., Abdulmalik, U. I., & Kassim, M. (2025). An enhanced rice grain classification using a novel feature engineering and random forest modelling. KASU Journal of Computer Science, 2(1), 33–40. https://www.kjcs.com.ng/view_paper/92

Hair, J., Anderson, R., Black, B., & Babin, B. (2016). Multivariate data analysis. Pearson Higher Ed.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition. Springer Science & Business Media.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning: With applications in R. Springer Science & Business Media.

Livio, M. (2003). The Golden Ratio: The Story of Phi, the extra ordinary [i.e. extraordinary] number of nature, art and beauty. Headline Review.

Nelson, D. (2020). What is ensemble learning? Retrieved from https://www.unite.ai/what-is-ensemble-learning/

Nujumudeen, F., & G, Adarsh. (2022). Comparative study of Glass Type Classification using Machine Learning. Ymer. 21. 111. 10.37896/YMER21.03/13.

O'Brien, R. M. (2007). A caution regarding rules of thumb for variance inflation factors. Quality & amp; Quantity, 41(5), 673–690. https://doi.org/10.1007/s11135-006-9018-6

Muthukrishnan, R., & James, C. K. (2024). The effect of multicollinearity on feature selection. Indian Journal Of Science and Technology, 17(35), 3664–3668. https://doi.org/10.17485/ijst/v17i35.1876

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information Processing & Management, 45(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

Suppa, R. (2023). Comparative performance evaluation results of classification algorithm in data mining to identify types of glass based on refractive index and its elements. PENA TEKNIK: Jurnal Ilmiah Ilmu-Ilmu Teknik, 8(1), 20. https://doi.org/10.51557/pt_jiit.v8i1.1705