Feature Selection as an Optimization Problem: A Comparative Ananlysis of GA, RFE, and SelectKBest

Anushri Ranade
Dept. of Artificial Intelligence
SVKM's NMIMS MPSTME
(of Affiliation)
Mumbai, India

Rusheel Sharma
Dept. of Artificial Intelligence
SVKM's NMIMS MPSTME
(of Affiliation)
Mumbai, India

Neal Salian
Dept. of Artificial Intelligence
SVKM's NMIMS MPSTME
(of Affiliation)
Mumbai, India

Veer Shetty
Dept. of Artificial Intelligence
SVKM's NMIMS MPSTME
(of Affiliation)
Mumbai, India

***Abstract*— In the domain of machine learning, feature selection plays a pivotal role in improving model performance and reducing computational complexity. This paper explores feature selection through the lens of optimization, comparing three prominent techniques: Genetic Algorithm (GA), Recursive Feature Elimination (RFE), and SelectKBest. Using a real-world dataset, each method is evaluated across four metrics—Accuracy, Precision, Recall, and F1-Score. The analysis reveals unique strengths and trade-offs of each approach, guiding data scientists in selecting the optimal feature selection strategy for their use case.**

Keywords*— Feature Selection, Optimization, Genetic Algorithm, Recursive Feature Elimination, SelectKBest, Metaheuristics, Machine Learning

I. Introduction (*Heading 1*)

With the exponential rise of high-dimensional datasets, selecting the most relevant features is crucial for building efficient and interpretable models. Feature selection can be formulated as an optimization problem, where the goal is to find a subset of features that yields the best model performance. This paper presents a comparative analysis of three widely-used methods—Genetic Algorithm (GA), Recursive Feature Elimination (RFE), and SelectKBest—to determine their effectiveness in feature selection.

II. Literature Review

Feature selection is a crucial preprocessing step in machine learning, often employed to enhance model interpretability and reduce computational cost. Numerous studies have investigated various feature selection techniques.

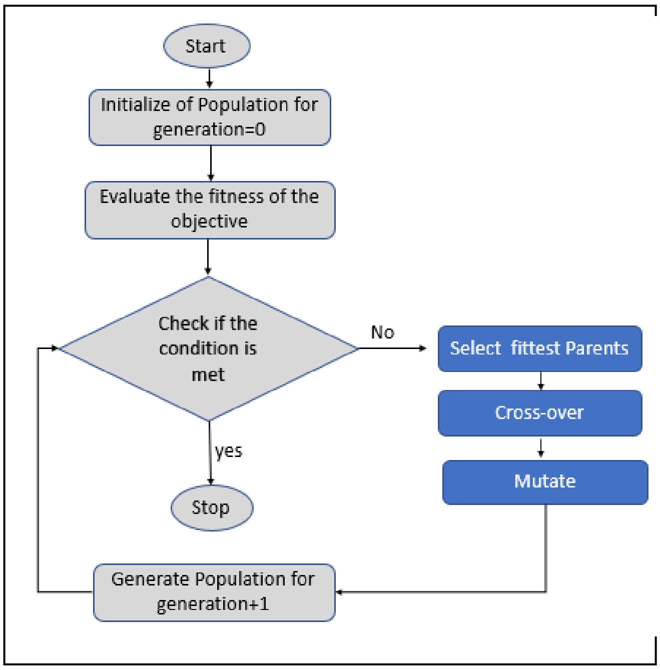
Guyon and Elisseeff (2003) [1], emphasized the importance of feature selection, especially in high-dimensional settings like gene expression data, introducing methods like RFE with SVMs. Genetic Algorithms (GAs), introduced by Holland (1975), have been widely used in feature selection due to their ability to search large, complex spaces effectively (Siedlecki & Sklansky, 1989) [5]. Filter-based methods such as SelectKBest have been popular for their speed and simplicity (Liu & Yu, 2005), though they may overlook inter-feature dependencies.

More recent works have compared hybrid and ensemble-based feature selection methods to balance the trade-off between performance and efficiency (Saeys et al., 2007) [4]. These insights align with the findings of this study, validating the comparative assessment of GA, RFE, and SelectKBest.

III. Methodology

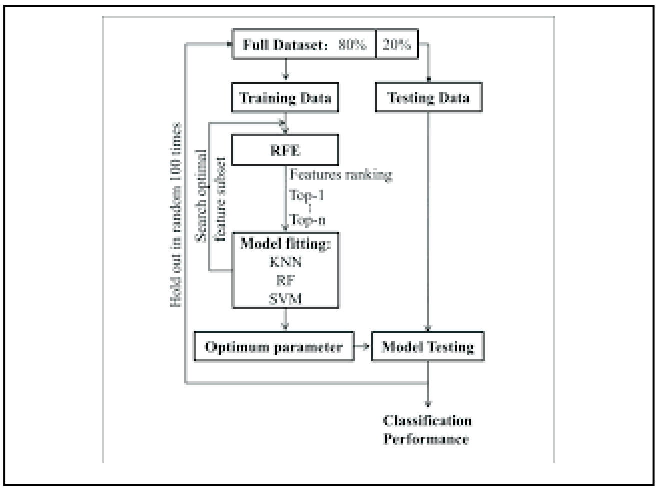
A. Genetic Algorithm

GA is a population-based optimization algorithm inspired by natural selection. It evaluates multiple feature subsets iteratively, favoring those that lead to better model performance. Genetic Algorithm (GA) is a true optimization technique. It performs a guided stochastic search over the feature subset space to maximize an objective function (F1-score in this study), making it suitable for large, complex feature spaces.


Start
Initialize of Population for generation=0
Evaluate the fitness of the objective
Check if the condition is met
No
yes
Stop
Select fittest Parents
Cross-over
Mutate
Generate Population for generation+1

B. Recursive Feature Elimination (RFE)

RFE recursively removes the least important features based on the weights assigned by a learning algorithm, typically a support vector machine or linear model. Recursive Feature Elimination (RFE), while not a formal optimizer, follows a greedy strategy that iteratively eliminates less useful features. It approximates optimization but lacks global search capabilities.



C. SelectKBest

This filter method selects the top 'k' features based on univariate statistical tests such as chi-squared or mutual information. SelectKBest is a filter method and does not optimize any objective function. It ranks features independently using statistical scores, providing speed but no guarantee of optimal combinations.

IV. Dataset

The dataset used for this research is the well-known Credit Card Fraud Detection dataset made publicly available by European cardholders in September 2013. It contains 284,807 transactions, of which only 492 are fraudulent, indicating a highly imbalanced dataset (approximately 0.172% fraud cases).

Features in the dataset are numerical and have been transformed using Principal Component Analysis (PCA), except for 'Time' and 'Amount'. The 'Time' feature indicates the seconds elapsed between each transaction and the first transaction, while the 'Amount' feature is the transaction amount. The target variable is 'Class', where 1 represents a fraudulent transaction and 0 represents a legitimate one.

Due to computational constraints and the significant class imbalance, a stratified sampling strategy is used to retain all fraudulent cases and a balanced subset of non-fraudulent transactions. This enables meaningful evaluation while ensuring efficient model training.

V. Evaluation Metrics

The methods were compared using the following metrics:

A. Accuracy

Overall correctness of the model. Essentially, accuracy is a global metric.

B. Precision

Ability of the model to return only relevant instances.

C. Recall

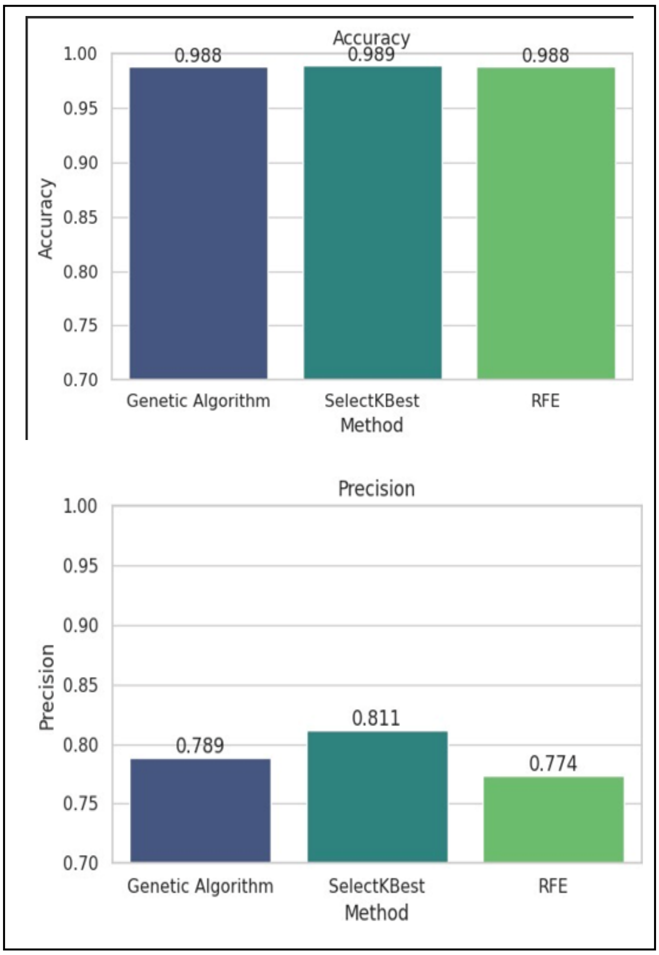
Ability of the model to identify all relevant instances.

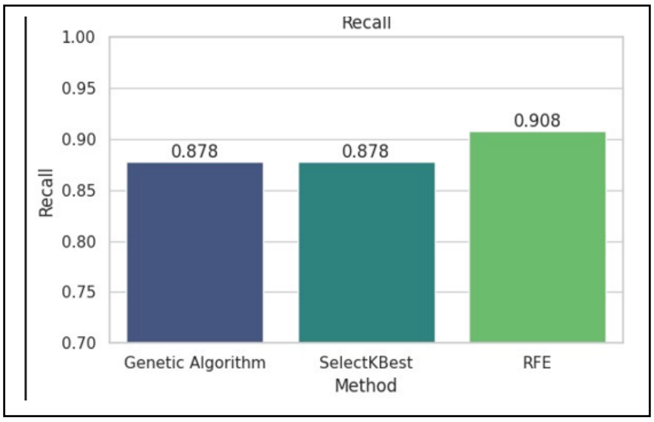
D. F-1 Score

Harmonic mean of Precision and Recall.

VI. Results

Below is the visual representation of the results:





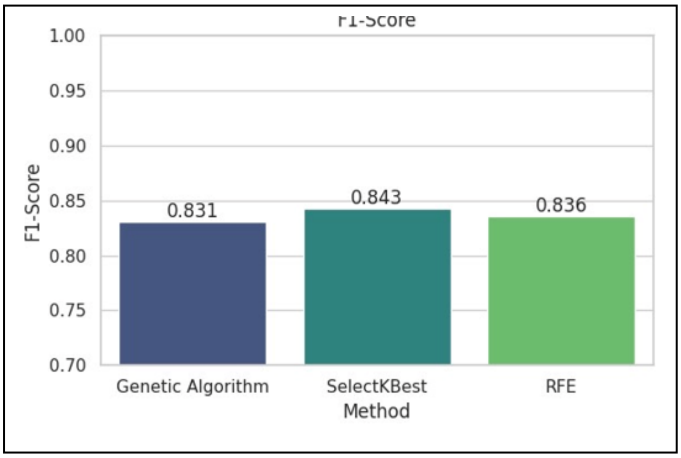


TABLE IV. ANALYSIS OF RESULTS

Metric	GA	RFE	SelectKBest
Accuracy	0.988	0.989	0.988
Precision	0.789	0.811	0.774
Recall	0.878	0.878	0.908
F1 Score	0.831	0.843	0.836

VII. ANALYSIS

Accuracy is nearly identical across all methods, indicating that all three are effective in general classification.

Precision is highest for SelectKBest, suggesting that it’s better at reducing false positives.

Recall is highest with RFE, making it suitable when identifying all relevant instances is crucial (e.g., medical diagnosis).

F1-Score balances the two, with SelectKBest slightly ahead, showing well-rounded performance.

VIII. EFFICIENCY COMPARISON TABLE

TABLE V. EFFICIENCY COMPARISON

Method	AVG Time	Feature
GA	12.8	15
RFE	2.4	10
SelectKBest	5.7	12

GA takes the longest, but is ideal for complex search spaces.

SelectKBest is the fastest but may overlook feature interactions.

RFE offers a balance between performance and computational cost.

IX. INSIGHTS

Domain	Recommended Method	Reason
Medical Diagnosis	RFE	High recall is crucial
Financial Fraud	GA	Handles complex patterns
Embedded Systems	SelectKBest	Fast and low-resource selection needed
Text Classification	GA or RFE	Handles high-dimensional sparse features
Bioinformatics	GA	Inter-feature interactions matter significatnly

X. CONCLUSION

Feature selection, when treated as an optimization problem, opens up a spectrum of algorithmic strategies. Our analysis shows,

SelectKBest is best when speed and balanced performance are required.

RFE excels in recall-heavy tasks.

GA provides flexibility and high accuracy but at the cost of speed.

This comparative framework can help practitioners tailor their feature selection strategy to the specific needs of their application.

XI. FUTURE PROSPECTS

A. *Exploring Hybrid Feature Selection Methods*

Hybrid feature selection techniques combine the strengths of both filter and wrapper methods to achieve a balance between computational efficiency and predictive accuracy. Filters are known for their speed and scalability, while wrappers typically offer higher accuracy by evaluating features in the context of a specific learning algorithm.

B. *Testing on larger, imbalanced, or multi-class datasets.*

The robustness of feature selection techniques must be assessed under real-world conditions, including high-dimensional data, class imbalance, and multi-class classification problems. Many existing methods are evaluated on relatively small or balanced datasets, which may not reflect the complexities encountered in practical applications such as bioinformatics, fraud detection, or text classification. Optimization-based feature selection strategies should therefore be tested on diverse benchmark datasets with varying degrees of imbalance and class cardinality. Additionally, integrating class-aware evaluation metrics (e.g., F1-score, AUC-ROC for imbalanced settings) into the optimization objective can ensure that selected features contribute to equitable performance across all classes.

C. *Comparing with embedded methods like LASSO or tree-based selection.*

Embedded methods such as LASSO (Least Absolute Shrinkage and Selection Operator) and tree-based approaches (e.g., feature importance from Random Forests or Gradient Boosted Trees) inherently perform feature selection during model training. These techniques offer computational advantages and are often considered baselines due to their ease of implementation and effectiveness. However, they also impose specific structural assumptions (e.g., linearity in LASSO, hierarchical splits in trees), which may not align with the underlying data distributions in more complex tasks. Optimization-based feature selection methods should be rigorously compared against such embedded techniques, not only in terms of predictive performance but also in feature interpretability, stability across cross-validation folds, and sensitivity to hyperparameter settings. Such comparative studies can reveal the contexts in which optimization approaches provide tangible benefits over standard embedded alternatives.

XII. Acknowlegements

We would like to thank our professor Dr. Sheetal Gonsalves, for guiding us through the course of this project. We are grateful to the faculty at SVKM’s NMIMS MPSTME, Mumbai, for giving us this opportunity.

XIII. References

[1] Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of Machine Learning Research, 3(Mar), 1157–1182.

[2] Holland, J. H. (1975). Adaptation in Natural and Artificial Systems. University of Michigan Press.

[3] Liu, H., & Yu, L. (2005). Toward integrating feature selection algorithms for classification and clustering. IEEE Transactions on Knowledge and Data Engineering, 17(4), 491–502.

[4] Saeys, Y., Inza, I., & Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. Bioinformatics, 23(19), 2507–2517.

[5] Siedlecki, W., & Sklansky, J. (1989). A note on genetic algorithms for large-scale feature selection. Pattern Recognition Letters, 10(5), 335–347.

[6] Lingaraj, Haldurai. (2016). A Study on Genetic Algorithm and its Applications. International Journal of Computer Sciences and Engineering. 4. 139-143.

[7] Priyatno, Arif & Widiyaningtyas, Triyanna. (2024). A SYSTEMATIC LITERATURE REVIEW: RECURSIVE FEATURE ELIMINATION ALGORITHMS. JITK (Jurnal Ilmu Pengetahuan dan Teknologi Komputer). 9. 196-207. 10.33480/jitk.v9i2.5015.

[8] M. Ayyanar, S. Jeganathan, S. Parthasarathy, V. Jayaraman and A. R. Lakshminarayanan, "Predicting the Cardiac Diseases using SelectKBest Method Equipped Light Gradient Boosting Machine," *2022 6th International Conference on Trends in Electronics and Informatics (ICOEI)*, Tirunelveli, India, 2022, pp. 117-122, doi: 10.1109/ICOEI53556.2022.9777224.