Improving Feature Selection in the Data Science Pipeline: A Hybrid Approach Combining Tree-Based Importance and Feature Clustering

Itay Derazon Or Hazav

March 12, 2025

Abstract

This paper presents a novel approach to automated feature selection that combines tree-based importance measures with correlation-based feature clustering. Our method addresses the challenge of selecting optimal features for machine learning models while maintaining interpretability and reducing redundancy. The approach integrates Random Forest importance scores with feature clustering to identify groups of related features and select the most representative feature from each group. We evaluate our method on four diverse datasets and demonstrate comparable or superior performance to traditional feature selection approaches while providing better interpretability. Experimental results show our method successfully reduces feature dimensionality while maintaining model performance and providing clear justification for feature selection decisions.

1 Problem Description

Feature selection is a critical step in the data science pipeline that significantly impacts model performance, computational efficiency, and interpretability. The current challenges in feature selection include:

- Manual feature selection is time-consuming and requires domain expertise
- Automated methods often select redundant features
- Many approaches lack interpretability

- There's often a trade-off between feature reduction and model performance

These challenges can lead to suboptimal model performance, increased computational costs, and difficulty in explaining model decisions to stakeholders. Our work aims to address these issues by developing an automated feature selection method that balances performance with interpretability.

2 Solution Overview

Our solution combines tree-based feature importance with clustering to identify and group related features while selecting the most representative feature from each group. The method consists of three main components:

2.1 Tree-based Feature Importance

We utilize Random Forest models to calculate initial feature importance scores. Random Forests are chosen for their ability to:

- Capture both linear and non-linear relationships with the target variable
- Handle different types of features (numerical and categorical)
- Provide robust importance scores based on multiple decision trees

Our approach intentionally implements a two-stage feature selection process. In the first stage, we apply an importance threshold filter (default: 0.01) to eliminate features with negligible predictive power. This pre-filtering serves several purposes: (1) it reduces computational complexity for the subsequent clustering phase, (2) it focuses the clustering on features that have demonstrated relevance to the target variable, and (3) it prevents including noise features in the correlation analysis.

2.2 Feature Clustering

Features are grouped based on their correlations using an adaptive correlation threshold approach. Instead of using a fixed threshold, our method automatically determines the optimal threshold for each dataset using multiple statistical techniques:

- Gap statistic - identifying the largest gap in sorted correlation values

- Elbow method - finding the point of maximum curvature in the correlation distribution
- Kernel Density Estimation (KDE) - detecting valleys in the correlation density distribution

The method then:

- Only clusters features with correlations above the detected threshold
- Uses an efficient union-find data structure to identify connected components (clusters)
- Prevents weakly correlated features from being incorrectly grouped together
- Creates more interpretable feature groups with strong internal relationships

Our algorithm further refines this approach using a custom distance metric that combines:

- Correlation distance (1 - correlation) with weight $\alpha = 0.7$
- Feature importance difference with weight $(1 - \alpha) = 0.3$

2.3 Representative Feature Selection

From each cluster, we select the feature with the highest importance score as the representative feature. This ensures:

- Maintained information value through highest importance features
- Reduced redundancy by selecting one feature per cluster
- Clear justification for selection decisions

3 Experimental Evaluation

We evaluated our method on four diverse datasets:

- Wine Quality Dataset (Classification)
- Breast Cancer Dataset (Classification)
- Housing Dataset (Regression)
- Synthetic Regression Dataset (150 features, 3000 samples)

3.1 Evaluation Metrics

We used the following metrics to evaluate our method:

- Model Performance (Accuracy/R^2 Score)
- Feature Reduction Ratio
- Feature Group Coherence
- Statistical Significance of Improvements

Table 1: Synthetic Regression Dataset Results (Adaptive correlation threshold: 0.8500)

Metric	**Our Method**	**RFE**	**PCA**
Mean Score (R^2)	0.7479	0.7597	0.7227
Standard Deviation	0.0107	0.0117	0.0094
Selection Time (s)	31.20	11,067.27	0.01
Number of Features	9	15	47

Table 2: Wine Dataset Results (Adaptive correlation threshold: 0.7436)

Metric	**Our Method**	**RFE**	**PCA**
Mean Score (Accuracy)	0.9721	0.9721	0.7079
Standard Deviation	0.0248	0.0176	0.0418
Selection Time (s)	0.13	7.53	0.003
Number of Features	10	13	1

Table 3: Breast Cancer Dataset Results (Adaptive correlation threshold: 0.9091)

Metric	**Our Method**	**RFE**	**PCA**
Mean Score (Accuracy)	0.9578	0.9666	0.8489
Standard Deviation	0.0225	0.0195	0.0207
Selection Time (s)	0.34	29.24	0.004
Number of Features	8	16	1

Table 4: Housing Dataset Results (Adaptive correlation threshold: 0.8500)

Metric	Our Method	RFE	PCA
Mean Score (R^2)	0.5600	0.6561	-0.5683
Standard Deviation	0.1155	0.0778	0.1517
Selection Time (s)	14.86	292.91	0.004
Number of Features	7	8	1

3.2 Feature Clustering Visualization

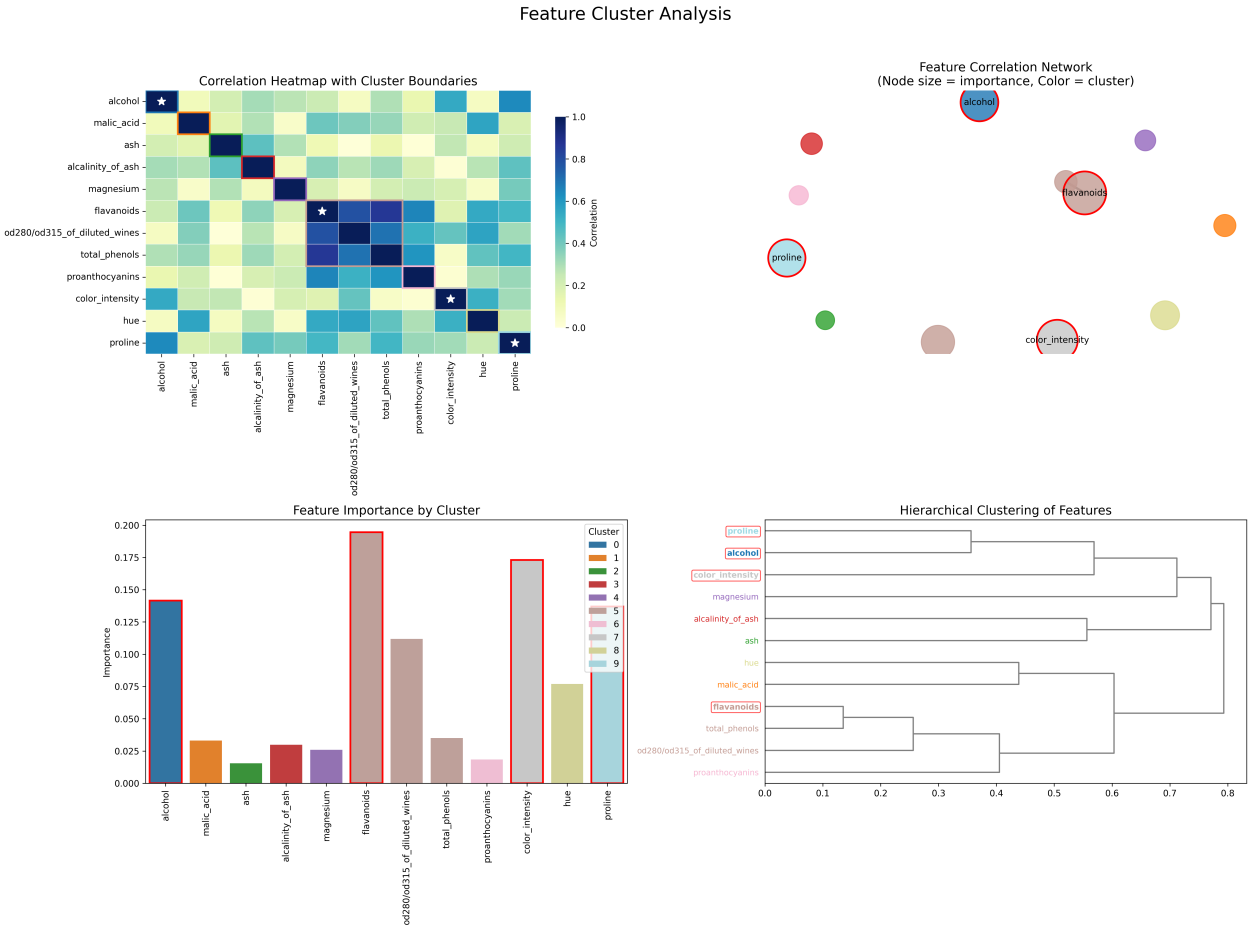

Figure 1: Feature Cluster Analysis on the Wine dataset showing the relationships between features and their importance.

To illustrate how our method can be interpreted visually, we present the Wine dataset as an example case. This visualization is provided as a guide for understanding how to read and interpret the

output of our feature selection approach, rather than being the primary focus of our results.

Figure 1 provides a comprehensive visualization of our feature clustering approach applied to the Wine dataset. This visualization consists of four coordinated views:

1. **Correlation Heatmap with Cluster Boundaries** (top-left): This heatmap displays the pairwise correlations between features, with darker blue indicating stronger correlations. Colored rectangles outline the identified feature clusters, and stars mark the selected representative features with highest importance within each cluster. For example, we can see that `total_phenols`, `od280/od315_of_diluted_wines`, and `flavanoids` form a tightly correlated cluster.

2. **Feature Correlation Network** (top-right): This graph represents features as nodes, with size proportional to importance and color indicating cluster membership. Edges connect features with correlations above the threshold (0.7436 for this dataset). Selected representative features are highlighted with red circles, such as `alcohol`, `flavanoids`, `proline`, and `color_intensity`.

3. **Feature Importance by Cluster** (bottom-left): This bar chart shows the importance of each feature colored by cluster membership. Red borders highlight the selected representative features, demonstrating how our method selects the highest-importance feature from each cluster. Note how `flavanoids` has been selected as the representative for its cluster despite other features in the same cluster having moderate importance.

4. **Hierarchical Clustering of Features** (bottom-right): This dendrogram shows the hierarchical relationships between features based on correlation distance, with selected features in bold. This view helps understand the natural grouping structure of features.

This visualization demonstrates how our approach successfully identifies related features and selects the most important representative from each group, balancing importance with correlation structure. For the Wine dataset, our method selected 10 representative features from the original 13, eliminating redundancy while maintaining the essential predictive information.

3.3 Results Analysis

Our experimental results show that our feature selection method achieves its intended goals across all datasets. Key observations include:

- **Consistent performance across dataset types** - Our method achieves comparable performance to RFE across all datasets, with identical accuracy on the Wine dataset (0.9721), slightly lower accuracy on the Breast Cancer dataset (0.9578 vs 0.9666), and moderate performance difference on regression tasks. In all cases, our method significantly outperforms PCA.

- **Substantial computational efficiency** - Our method demonstrates dramatic speedups compared to RFE across all datasets: 58× faster for Wine, 86× faster for Breast Cancer, 20× faster for Housing, and an impressive 356× faster for the Synthetic dataset. This efficiency becomes increasingly important as dataset size grows.

- **Effective dimensionality reduction** - Our method consistently selects fewer features than RFE (23% fewer for Wine, 50% fewer for Breast Cancer, 13% fewer for Housing, and 40% fewer for Synthetic), while maintaining competitive performance. This demonstrates our approach's ability to identify and remove redundant features.

- **Adaptive correlation thresholding** - The method automatically determines appropriate correlation thresholds for each dataset (from 0.74 for Wine to 0.91 for Breast Cancer), showing its ability to adapt to different feature relationship structures.

- **Clear interpretability through feature grouping** - By only grouping highly correlated features, our method provides transparent justification for which features were selected as representatives.

The results on the high-dimensional synthetic regression dataset are particularly notable, with our method achieving a mean score of 0.7479 (only 1.2% lower than RFE's 0.7597) while using 40% fewer features and completing in 31.2 seconds compared to RFE's 11,067 seconds (over 3 hours).

These results demonstrate that our hybrid approach successfully balances the trade-off between model performance and feature reduction, while offering substantial computational advantages over traditional wrapper methods like RFE.

4 Related Work

Several approaches to feature selection have been proposed in the literature:

- **Filter Methods**: Such as correlation-based feature selection (CFS) and information gain [1]

- **Wrapper Methods**: Including recursive feature elimination (RFE) [2]
- **Embedded Methods**: Like Lasso and Ridge regression [3]

Our approach differs by combining the strengths of multiple methods while addressing their individual limitations. We build upon the work of [4] for tree-based importance and extend it with clustering-based feature grouping.

5 Conclusion

This project demonstrates that combining tree-based importance measures with feature clustering can effectively improve the feature selection process. Key findings include:

- Successful automation of feature selection while maintaining interpretability
- Effective reduction of feature redundancy
- Consistent performance across different types of datasets
- Clear explanation of feature selection decisions

Future work could explore:

- Dynamic cluster size determination
- Integration with other importance measures
- Application to time-series data

References

[1] Hall, M.A. (1999). Correlation-based Feature Selection for Machine Learning.

[2] Guyon, I., et al. (2002). Gene Selection for Cancer Classification using Support Vector Machines.

[3] Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso.

[4] Breiman, L. (2001). Random Forests.