Feature Selection Report

Quan Shi

Section 1: Explanation of Feature Selection Methods

1.1 Recursive Feature Elimination (RFE)

Method Description: Recursive Feature Elimination (RFE) is a wrapper method that performs feature selection by iteratively removing unimportant features. It works by training a model, identifying features with the smallest weights and removing them in each iteration until the number of remaining features reaches a predetermined value. The main goal of RFE is to find the subset of features from the feature space that can best improve model performance.

Reason for selection: RFE selects features based on model feedback (such as accuracy), so it is more efficient in tasks that require evaluation of feature importance. In this experiment, a logistic regression model is selected so that RFE can use model coefficients as a measure of feature importance to help identify features that contribute to prediction accuracy. However, RFE is computationally intensive because it requires repeatedly training the model to optimize the feature combination.

1.2 SelectFromModel

Method Description: SelectFromModel is an embedding method that uses a model-based feature importance metric to filter features. Specifically, it judges feature importance based on the coefficients of the logistic regression model and selects features whose importance exceeds a preset threshold. SelectFromModel only needs one model training to obtain feature importance, which is more efficient on high-dimensional data.

Reason for selection: SelectFromModel strikes a balance between computational efficiency and effective feature screening. Unlike RFE, which requires multiple trainings, SelectFromModel can screen features based on model coefficients after a single training, significantly reducing computational overhead while effectively filtering out features that have little impact on model performance.

1.3 Genetic Algorithm (GA)

Method description: Genetic algorithm (GA) is an optimization algorithm that simulates evolution. In the feature selection task, GA regards it as a binary-coded combinatorial optimization problem, where each feature is represented as selection (1) or exclusion (0). Through genetic operations such as selection, crossover, and mutation, GA gradually improves the feature set to find the optimal subset.

Fitness function: The fitness function is used to evaluate the quality of individuals in GA. In this experiment, the fitness function uses model accuracy as the main indicator, and adds a penalty term for the number of features to encourage the selection of fewer and important features.

Method improvement: To optimize GA, we introduced tournament selection and elite strategy. Tournament selection improves diversity by selecting the best individuals from random samples. The elite strategy ensures that the best individuals in each generation are directly passed to the next generation, avoiding the loss of the optimal solution due to crossover or mutation.

Section 2: Comparing Methods

2.1 RFE vs SelectFromModel

Number of Features
RFE: Retained 29 features, almost all the initial features. This method did not lead to a significant reduction, as only one feature was removed.
SelectFromModel: Retained 9 features, reducing the original set by around 70%. This method effectively filtered out less important features, focusing on the most impactful subset.

Analysis: The primary reason for this difference lies in the selection approach of each method. RFE removes features one-by-one, focusing on the least important feature in each iteration. This iterative approach does not prioritize overall feature reduction, which can lead to a larger feature set that retains multiple correlated features. SelectFromModel, however, uses model coefficients and a threshold to filter features in one pass, allowing it to effectively reduce the feature set by removing less important or redundant features in a single step. This thresholding mechanism makes SelectFromModel more aggressive in reducing feature count, resulting in a concise and interpretable feature subset.

Model Performance
RFE: Accuracy 0.9561; Precision 0.9459; Recall 0.9859; F1 Score 0.9655
SelectFromModel: Accuracy 0.9474; Precision 0.9577; Recall 0.9577; F1 Score 0.9577

Analysis: While RFE achieves slightly better performance metrics, it retains a larger feature set, which provides the model with additional information at the expense of simplicity. SelectFromModel achieves comparable performance but with significantly fewer features, striking a balance between accuracy and model complexity. The slight decrease in accuracy and recall observed in SelectFromModel is due to the more aggressive feature reduction, which may lead to the loss of some features with minor contributions to model performance. This trade-off, however, results in a simpler, more interpretable model with lower computational requirements, which can be advantageous depending on the application.

Feature Importance
RFE: Selected features include a broad set of indicators such as "mean radius," "mean texture," and "mean perimeter," which encompass tumor shape, size, and edge characteristics. This comprehensive feature set covers multiple aspects of the data.
SelectFromModel: Selected fewer features, primarily focusing on size and edge characteristics, including key indicators like "mean radius" and "mean perimeter."

Analysis: The difference in feature selection is due to RFE’s iterative approach versus SelectFromModel’s threshold-based filtering. RFE gradually removes features, aiming to retain those that individually contribute to model accuracy, even if they provide similar information to other features. As a result, RFE may retain multiple correlated features that collectively describe similar characteristics. SelectFromModel, on the other hand, applies a threshold on model coefficients to filter out features in one pass. This approach favors selecting only the most influential features, leading to a more concise feature set that represents essential characteristics like tumor size and edge, while avoiding redundancy.

In contrast, SelectFromModel uses model coefficients in a single pass to filter features based on a predefined importance threshold. This method is more aggressive, as it can eliminate all less important features at once. By setting a threshold, SelectFromModel can focus on the most influential features and discard those that offer marginal value or are highly correlated with the selected features. This results in a more concise and representative feature set that captures essential information about tumor size and edge characteristics while avoiding redundancy.

In summary, while RFE’s iterative approach aims to maximize performance by retaining all potentially useful features, it sacrifices simplicity and conciseness. SelectFromModel, on the other hand, prioritizes simplicity and efficiency, creating a model that is easier to interpret and computationally less expensive, even if it means a slight trade-off in model performance. This makes SelectFromModel particularly suitable for scenarios where interpretability and computational efficiency are key, while RFE may be preferable when model performance is the primary objective.

Stability of Feature Selection

Analysis: The feature sets selected by RFE and SelectFromModel partially overlap. However, the feature set provided by SelectFromModel is more stable, as it consistently selects a concise, representative set of features across different data samples. In contrast, RFE’s results show more variation and retain a larger number of features.

RFE’s iterative selection process involves retraining the model and re-evaluating feature importance in each iteration. This makes it sensitive to small changes in the data, as each pass can shift the importance of remaining features. Consequently, different data samples can lead to different paths of feature elimination, resulting in variability in the selected features. SelectFromModel, by contrast, performs a single-pass selection using model coefficients and a fixed threshold. This single-pass approach is less sensitive to minor data variations, resulting in more stable and consistent feature sets across different samples.

Handling of Feature Correlation

Analysis: SelectFromModel effectively handles feature correlation by setting a threshold to filter out less informative, redundant features. RFE, however, may retain multiple correlated features due to the limitations of its iterative selection mechanism.

SelectFromModel’s threshold-based filtering enables it to remove all features with lower importance in a single step, which helps eliminate correlated features that do not add unique information to the model. This approach creates a more concise feature set by focusing on the features with the highest impact. RFE, on the other hand, assesses features individually at each step and does not explicitly address correlation among features. As a result, RFE may retain several correlated features that provide overlapping information, leading to a larger, potentially redundant feature set.

2.2 Genetic Algorithm (GA) vs SelectFromModel

Number of features
GA: Retained 13 features, significantly reducing the number of features compared to the original set.
SelectFromModel: Retained 9 features, resulting in a smaller, more concise feature set.

Analysis: While both methods effectively reduce the number of features, **SelectFromModel** is more aggressive in feature reduction, yielding a smaller feature subset. This is because SelectFromModel uses a fixed importance threshold based on model coefficients, which allows it to directly filter out less important and redundant features in one pass. GA, on the other hand, optimizes feature selection by balancing model performance and feature count but may retain slightly more features. The GA’s optimization process aims to maximize performance while penalizing larger feature sets, resulting in a balanced reduction that might include additional features to support model generalization. This approach reflects GA’s focus on performance optimization rather than strict feature minimization, making it less aggressive in eliminating features compared to SelectFromModel.

Model performance
GA: Accuracy 0.9546; Precision 0.9577; Recall 0.9577; F1 score 0.9577
SelectFromModel: Accuracy 0.9474; Precision 0.9577; Recall 0.9577; F1 score 0.9577

Analysis: The model performance of GA and SelectFromModel is very similar, with GA achieving slightly higher accuracy. This performance edge may result from GA retaining a few more features, which could provide the model with additional information that aids in more accurate predictions. GA's selection process involves a fitness function that balances accuracy with feature count, allowing it to retain certain features that enhance performance even if they marginally increase the feature set size. SelectFromModel, in contrast, applies a stricter feature selection based solely on importance thresholding, which may exclude some features that could slightly boost performance. This trade-off makes SelectFromModel more efficient but may result in a minor loss in accuracy.

Feature Importance
GA: Selected features include tumor size, margin, and shape information, such as "mean radius," "mean perimeter," and "worst compactness."
SelectFromModel: Primarily focuses on tumor size and margin, selecting the most representative features.

Analysis: GA and SelectFromModel exhibit high consistency in identifying essential features like tumor size and margin characteristics, reflecting both methods’ ability to capture critical predictors. However, GA retains a few additional features, including shape-related characteristics, which can enhance the model's ability to generalize. GA’s inclusion of a more diverse feature set stems from its iterative optimization process, which evaluates different feature combinations for overall performance. This results in the retention of additional features that may support the model in generalizing across different samples, albeit with a larger feature set. SelectFromModel, being more selective, targets only the most representative features with the highest importance scores, resulting in a feature set that is more streamlined and focused on primary predictors.

Stability of Feature Selection
Analysis: GA's introduction of randomness in selection, crossover, and mutation steps can result in variability in the feature set across different runs, affecting the stability and reproducibility of selected features. SelectFromModel, in contrast, uses a deterministic process based on model coefficients, yielding a more stable and reproducible feature set.
GA's process involves stochastic elements, such as random selection and mutation, which means that each run can yield a slightly different feature set. This randomness is intrinsic to GA's design, aiming to explore a wide range of feature combinations to avoid local optima. While this increases the chances of finding a robust feature set, it also reduces stability, as different runs may select slightly different features.

Handling of Feature Correlation
Analysis: SelectFromModel leverages model coefficients to filter out correlated and less informative features more effectively, while GA, despite using penalty terms, may retain some correlated features in its selected set.
GA uses a fitness function that balances feature count and model performance, but it does not explicitly target correlation among features. While penalty terms in GA's fitness function discourage large feature sets, they may not always eliminate correlated features if those features contribute positively to model performance. This limitation can lead to a slightly larger feature set with some redundancy, as GA prioritizes performance over strict redundancy elimination.

Section 3: Conclusion

Best Method Selection:
Considering the number of features and model performance, SelectFromModel is the best feature selection method. It significantly reduces the number of features by selecting 9 features, and the model performance is close to RFE and GA.

Reasons for selection:
1. Control of the number of features: SelectFromModel only needs 9 features to achieve high model performance, which significantly simplifies the feature set and improves the interpretability of the model.
2. Model performance: RFE and GA have slightly higher performance, but more features are retained. The performance difference is small and it is difficult to make up for the lack of feature simplification.
3. Reproducibility: The feature selection process of SelectFromModel is more deterministic, the results are stable and easy to reproduce, and it is suitable for practical applications.
Summary: For this dataset, SelectFromModel achieves the best balance between the number of features and model performance through effective feature screening. Therefore, in practical applications, we recommend SelectFromModel as the preferred choice for this dataset.