Assignment 2: Feature Selection Methods Comparison

1 Introduction

1.1 Background

Feature selection is a crucial process in machine learning that involves identifying and selecting the most relevant features from a dataset. It;s primary goal is to improve model generalization, robustness and interpretability. By identifying and retaining only the most informative features, feature selection helps reduce noise, simplify model structure, and lower computational costs, especially in high-dimensional datasets. For example, in a medical dataset like the Breast Cancer dataset, focusing on key predictive features—such as tumor texture or radius—enhances the model's ability to generalize by preventing overfitting on irrelevant details. Importantly, feature selection does not guarantee an accuracy increase; in fact, accuracy might remain stable or even decrease slightly. However, this trade-off often leads to a more robust model that performs better on unseen data, as it captures the true signal rather than noise.

Various methods exist for feature selection, including filter methods that assess feature relevance based on statistical measures, wrapper methods that evaluate feature subsets using model performance, and embedded methods that perform feature selection during the model training process. Each approach offers distinct advantages and challenges, making the choice of method dependent on the specific characteristics of the dataset and the objectives of the analysis.

1.2 Dataset Overview

The Breast Cancer dataset, commonly used for binary classification tasks, consists of 30 continuous features aimed at distinguishing between malignant and benign tumors. Each feature is derived from digital images of fine-needle aspirates of breast masses, capturing various physical characteristics of cell nuclei. The target variable is binary, indicating the diagnosis: 0 for malignant tumors and 1 for benign ones.

In examining the relationships between features and the target variable, some features exhibit a strong correlation with target, as shown in the Figure. 1 Notably, features like "worst concave points," "worst perimeter," "mean concave points," and "worst radius" display high absolute correlation values with the target, making them potentially valuable predictors. However, many features are also highly correlated with each other, leading to redundancy within the dataset as shown in Figure 2.

This redundancy poses challenges for machine learning models, as correlated features can introduce multicollinearity, potentially destabilizing certain algorithms and obscuring feature importance. Reducing such redundancy through feature selection methods is essential for building a more efficient and interpretable model.

1.3 Objective And Methodology

The objective is to experiment with three feature selection methods on the Breast Cancer dataset: two established techniques from Scikit-learn and a custom Genetic Algorithm. This report will analyze each method's approach to feature correlation, consistency in feature selection, and stability of selected subsets.

The dataset is split into training and testing sets to ensure consistent evaluation. Logistic Regression is used as the classifier for model performance assessment and in feature selection methods that rely on model-exposed coefficients. This classifier is simple and it's linear nature allows for straightforward interpretation of feature coefficients, providing clear insights into the relationship between each feature and the target variable. This interpretability is crucial when assessing the importance of features selected by methods. Since Logistic Regression is also sensitive to scaling, I applied Min-Max scaling to normalize the features. For feature selection, I chose Recursive Feature Elimination (RFE) and SelectFromModel from Scikit-learn, along with a custom Genetic Algorithm (GA). Each method will be discussed in detail in the following sections.

2 Explanation of Feature Selection Methods

2.1 Scikit-Learn Feature Selection Techniques

I chose Scikit-learn feature selection methods that depend on model-driven evaluations: Recursive Feature Elimination (RFE) and SelectFromModel. Unlike filter methods, these techniques do not assess features independently; instead, they evaluate their impact based on model performance. For the Breast Cancer dataset, these methods help ensure that diagnostically relevant features are prioritized, enhancing classification performance and interpretability by retaining features most meaningful for distinguishing malignant from benign tumors.

2.1.1 Recursive Feature Elimination (RFE)

RFE is a wrapper feature selection method, meaning it evaluates feature subsets by assessing their specific contribution to the model's predictive performance. It works by iteratively training a model and ranking features based on model-derived importance scores, such as coefficients in linear models. In each iteration, the least important feature is removed, and the model is retrained on the updated subset. This iterative process continues until the desired number of features remains, allowing RFE to capture combinations of features that may be weak individually but significant together. RFE can be more sensitive to feature interactions because it iteratively refines the feature set, recalculating the importance of remaining features after each elimination. This recalibration helps it capture interactions that may only become apparent after certain features are removed. This iterative refinement process makes RFE potentially more accurate in identifying combinations of features that jointly contribute to model performance, as it continuously adapts to the changing feature set.

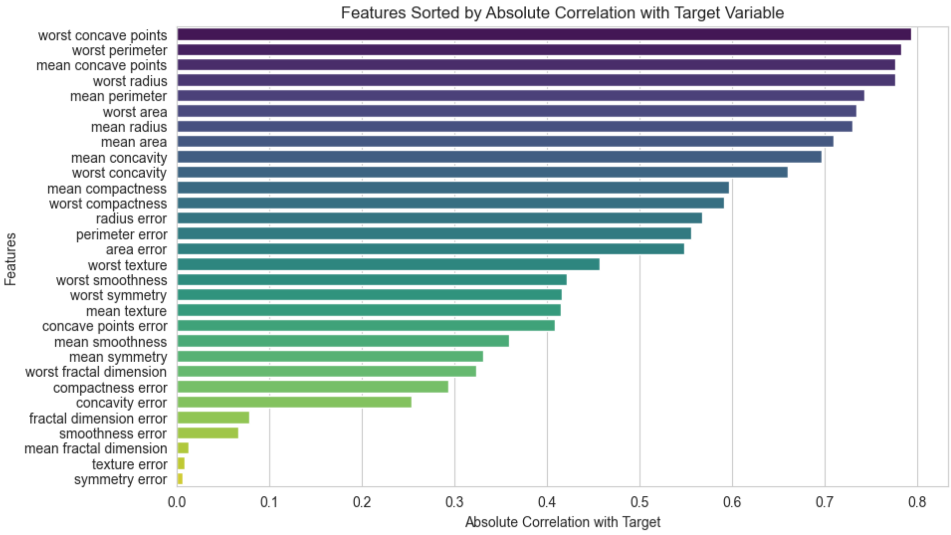

Figure 1: Features sorted by absolute correlation with the target variable.

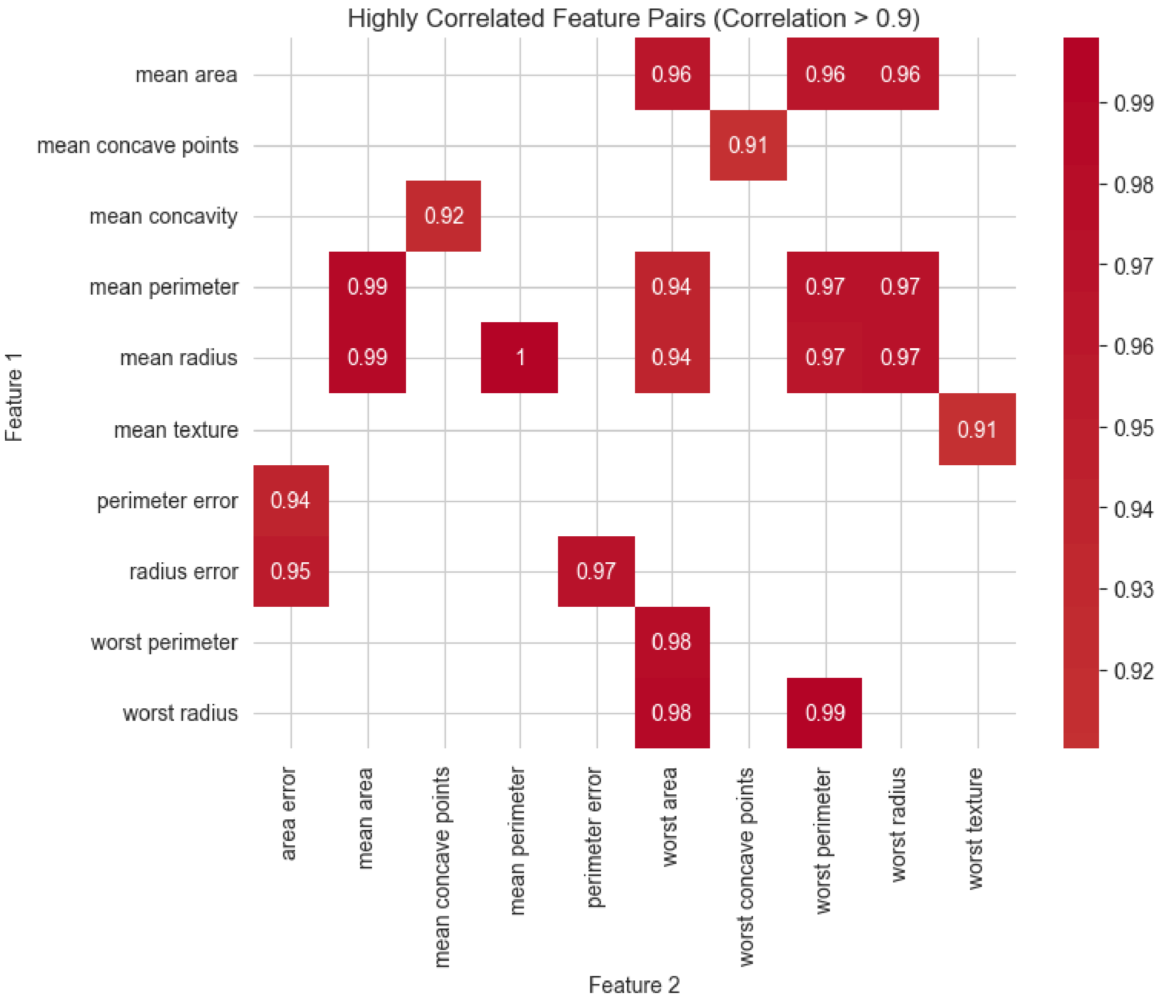

Figure 2: Highly correlated feature pairs (correlation > 0.9).

2.1.2 SelectFromModel

SelectFromModel is an embedded feature selection method, as the name suggests, integrating feature selection directly within the model training process. It selects features by leveraging model-derived importance scores, such as coefficients in linear models or feature importances in tree models, to retain only the features most critical for prediction. Unlike RFE's iterative approach, SelectFromModel performs selection in a single step, applying a threshold to importance scores to identify significant features. This threshold can be user-defined or set based on statistical measures, such as the mean importance score. SelectFromModel is more efficient since it only requires a single training run and operates on the initial set of feature importance scores.

2.2 Genetic Algorithm

I implemented a genetic algorithm that strikes a balance between exploration and exploitation through a fitness function that maximizes recall for the malignant class, as accurately identifying malignant cases is critical in a diagnostic context to minimize false negatives. The fitness function also encourages smaller, diverse feature sets, promoting efficient feature selection. Additionally, I incorporated tournament selection and elitism to further refine this balance. Together, these elements ensure that the genetic algorithm explores the feature space widely without losing focus on high-recall, efficient feature subsets, making it well-suited to the Breast Cancer dataset's diagnostic requirements.

2.2.1 Fitness Function

The "calc_fitness" function calculates a fitness score for each individual in the population by combining performance, sparsity, and correlation metrics. For each individual, features selected by that individual are extracted, and a model is trained using these selected features. The fitness function then calculates the recall for malignant cases (class 0) on a validation set to emphasize minimizing false negatives, which is critical in medical contexts. A feature sparsity penalty is applied by calculating the ratio of selected features to the total features, promoting a compact subset. Additionally, a correlation penalty is computed by examining the average correlation among the selected features, with a higher penalty assigned if the correlation exceeds a set threshold (correlation_threshold=0.9). The final fitness score is calculated as a weighted sum of recall minus the penalties for feature count and correlation, balancing predictive performance with feature efficiency. The formula is structured as follows:

$$\text{Fitness} = \delta \times \text{Recall}_{\text{class } 0} - \gamma \times \text{Feature Count Penalty} - \gamma \times \text{Correlation Penalty} \tag{1}$$

where:

- δ and γ are weighting factors that control the importance of recall relative to the feature count and correlation penalties. In this case $\delta = 1.0$ and $\gamma = 0.01$. So a very small penalty is applied on feature count and correlation, giving more importance to the recall.

The individual components are calculated as follows:

- **Recall for Class 0 ($\text{Recall}_{\text{class } 0}$)**: Measures the model's recall specifically for the malignant class (class 0).
- **Feature Count Penalty**:

$$\text{Feature Count Penalty} = \frac{\text{Number of Selected Features}}{\text{Total Number of Features}} \tag{2}$$

- **Correlation Penalty**:

$$\text{Correlation Penalty} = \begin{cases} \text{Average of Correlations Exceeding Threshold} & \text{if such correlations} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

2.2.2 Tournament Selection

Tournament selection was implemented to select parents for reproduction. In this method, a small subset of individuals is randomly chosen from the population (tournament_size), and the one with the highest fitness score is selected as a parent. This approach provides a balance between exploration and exploitation by allowing higher-fitness individuals a greater chance to reproduce while still giving lower-fitness individuals an occasional opportunity. This diversity in selection helps the algorithm explore a wider range of feature subsets, which is especially beneficial for high-dimensional datasets like the Breast Cancer dataset, where unique feature combinations might be missed with more deterministic selection methods.

2.2.3 Elitism

I implemented Elitism to preserve the best-performing individuals from each generation. By retaining a portion of the top individuals without subjecting them to genetic operations, elitism ensures that high-quality solutions are not lost over generations. This approach accelerates convergence toward optimal solutions and maintains continuity in performance improvement. For the Breast Cancer dataset, where maintaining a high recall and minimal feature redundancy is critical, elitism ensures that high-recall, sparse solutions with low correlation remain in the population, helping the algorithm build on proven feature subsets.

3 Comparing Methods

Table 1: Comparison of Feature Selection Methods

Method	Number of Features Selected	Selected Features	Recall Class 0
SelectFromModel	4	mean concave points, worst radius, worst texture, worst concave points	0.88
RFE	12	mean radius, mean perimeter, mean area, mean concave points, worst radius, worst texture, worst perimeter, worst area, worst smoothness, worst concavity, worst concave points, worst symmetry	0.89
GA	7	mean concavity, mean symmetry, mean fractal dimension, smoothness error, worst perimeter, worst smoothness, worst concave points	0.91

3.1 RFE Vs SelectFromModel

The results for RFE and SelectFromModel reveal distinct differences due to their selection mechanisms. RFE retained 12 features, including some correlated ones like ‘mean radius’, ‘mean perimeter’, and ‘mean area’, because its iterative approach evaluates feature subsets rather than individual importance scores in isolation. This allows RFE to keep a broader set of features that may collectively contribute to model performance, resulting in a slight improvement in recall, especially for the malignant class. In contrast, SelectFromModel, with a mean threshold, retained only 4 features—'mean concave points', 'worst radius', 'worst texture', and 'worst concave points'—by filtering out features with importance scores below the mean in a single pass. This aggressive filtering captured only the top predictors, naturally reducing redundancy, as shown in the Figure 1 where the selected features have strong correlations with the target. SelectFromModel demonstrated greater stability across runs because its threshold approach is less sensitive to minor changes, while RFE's iterative subset evaluation may introduce slight variability. Additionally, RFE does not explicitly address redundancy, so it may retain correlated features that collectively enhance predictive power, whereas SelectFromModel's thresholding avoids redundancy by focusing on the highest individual importance scores. Both methods achieved significant reductions from the original 30 features, with SelectFromModel providing a more compact subset.

3.2 Genetic Algorithm Vs SelectFromModel

The Genetic Algorithm (GA) retained 7 features, selected from the fittest individual in the final run, achieving a significant reduction from the original 30 features with a focus on maximizing recall for the malignant class. Due to its non-deterministic nature, GA was run multiple times, ultimately selecting a subset optimized for recall. Notably, 'worst concave points' was selected in every run, matching with its high relevance with target and consistent impact on recall. GA's exploration-exploitation balance allows it to test diverse feature combinations (exploration) while consistently prioritizing features with high recall impact (exploitation). This flexibility helps GA prioritize essential features without rigidly restricting combinations.

Unlike SelectFromModel (SFM), which strictly removes redundancy through a threshold-based approach, GA applies only a slight penalty to correlated features. This penalty discourages, but does not strictly eliminate, redundancy, resulting in a final subset that includes some moderately correlated features, like 'worst smoothness' and 'mean fractal dimension', when they collectively contribute to recall. This design allows GA to retain features that might be correlated yet provide incremental value in identifying malignant cases. Consequently, GA's subset includes both features with strong correlations to the target (e.g., 'worst concave points') and others with lower correlations, reflecting GA's exploratory nature, which captures feature interactions that collectively enhance recall, even if individual correlations vary.

In contrast, SFM's strict threshold approach selects only the top 4 features, inherently minimizing redundancy by filtering out correlated features with slightly lower importance scores. This leads to a highly interpretable, compact, and consistent subset across runs, focused on individual predictor strength. While GA's subset slightly outperformed SFM in malignant class recall (91% vs. 88%), it includes some redundancy to prioritize recall, while SFM's stable, minimal subset maximizes interpretability and non-redundancy.

4 Conclusion

The Genetic Algorithm (GA) demonstrated clear advantages for feature selection in this dataset, excelling in both model performance and feature reduction, achieving a recall of 91% for the malignant class with a set of 7 features. By optimizing its fitness function specifically for recall, GA selected features that collectively enhance sensitivity to malignant cases, capturing nuanced patterns and interactions. Unlike RFE and SelectFromModel, which tend to focus on top individual predictors or incremental removal, GA's mechanism tests diverse feature combinations, retaining those that support high recall collectively. The selected features highlight GA's exploratory nature, as it uncovered combinations that traditional methods often overlook, ensuring that even moderately important features are retained if they contribute meaningfully to recall when used together.

In terms of feature reduction, GA achieved a balanced subset of 7 features from the original

30, finding an optimal trade-off between dimensionality reduction and retaining essential feature interactions. This adaptability is further supported by GA's flexible fitness function, which can be tailored to the domain's specific needs—here, focusing on high recall in a diagnostic setting. This flexibility allows GA to align with application-specific objectives, adjusting feature selection criteria to support sensitivity in high-stakes contexts. In summary, GA's capacity to maximize recall, capture important feature interactions, and adapt to application demands through customizable fitness functions make it an ideal method for feature selection in this dataset, supporting the sensitivity and interpretability required in medical diagnostics.