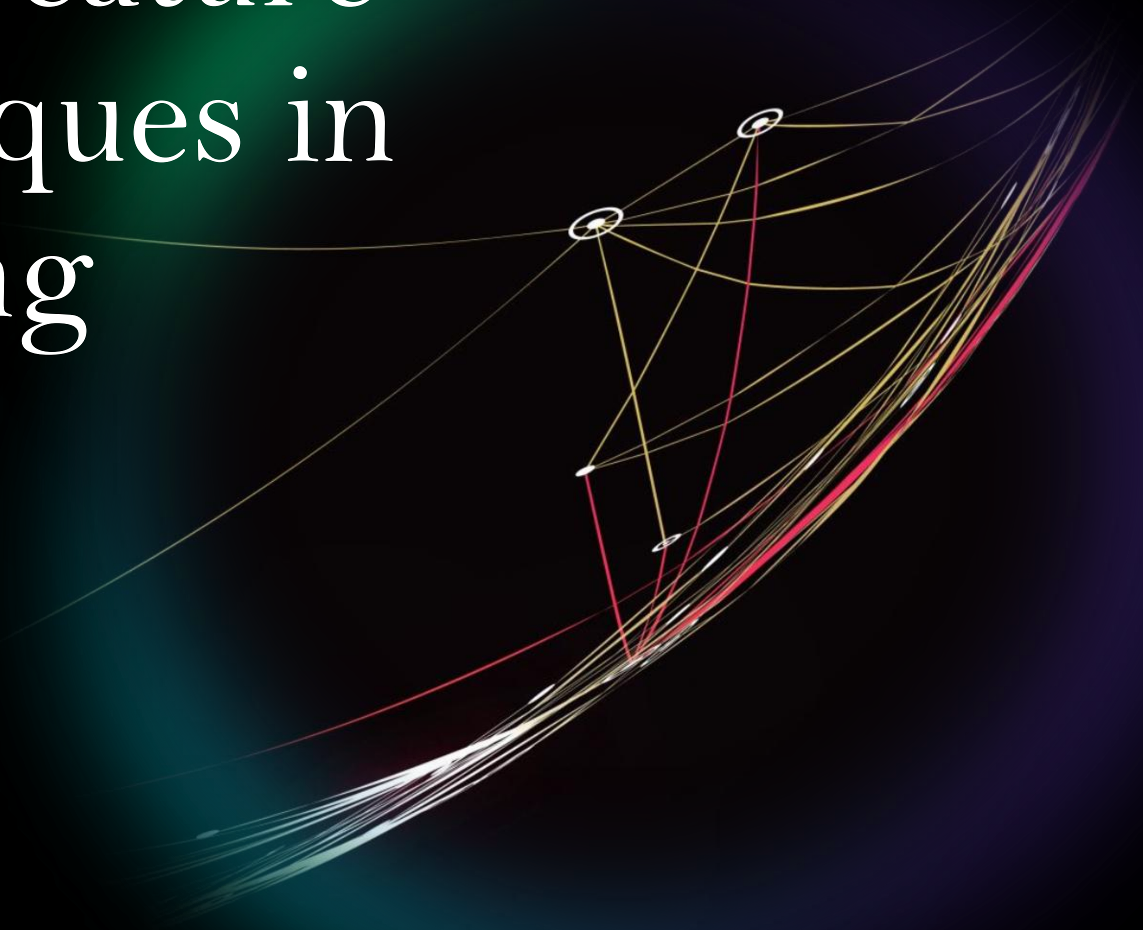
Understanding Feature Selection Techniques in Machine Learning



Anthony Onoja, PhD

School of Health Sciences

What does it mean?

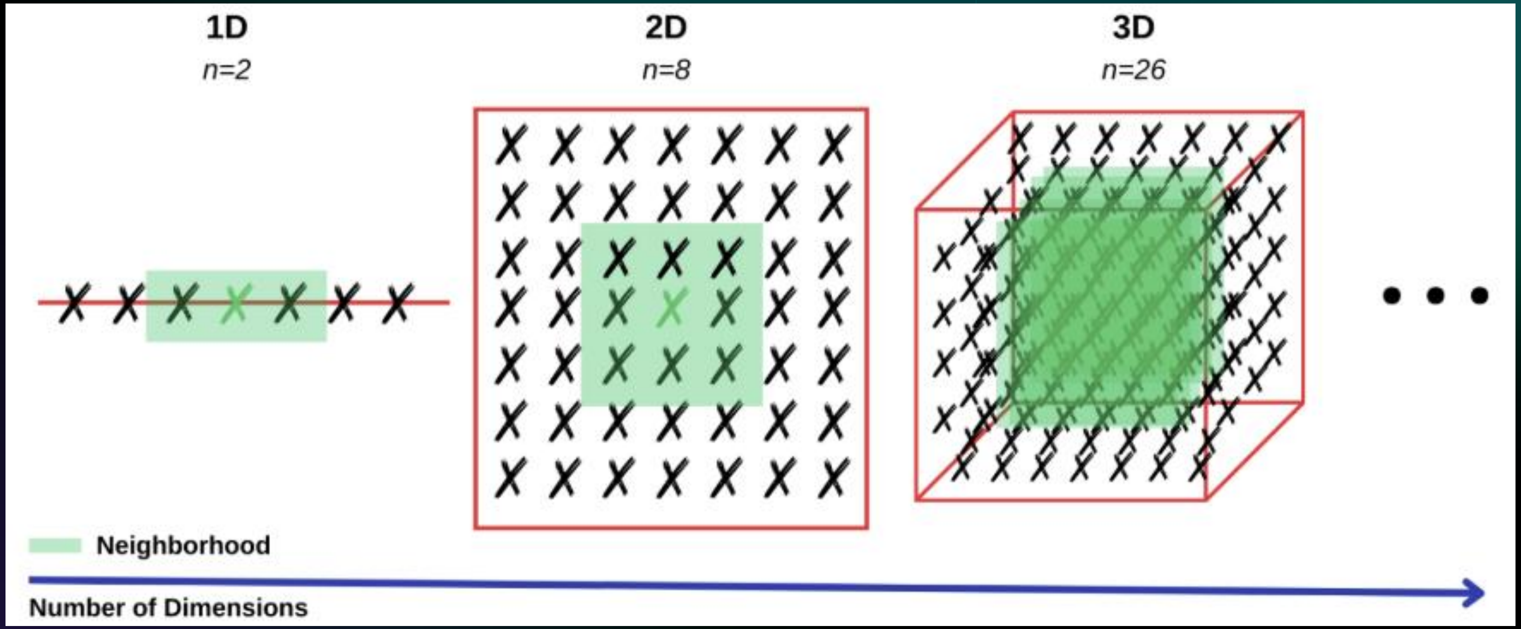
In a typical data science project, datasets often have numerous features, but not all are crucial for prediction. Feature engineering streamlines models by selecting key features.

- **Definition:** The process of selecting the most relevant features for a machine learning model.
- **Importance:** Improves model performance, reduces overfitting, and enhances interpretability.
- **Examples:** Predicting customer churn, medical diagnosis, and financial forecasting.

It's a critical step in the machine learning pipeline especially when dealing with high-dimensional data.

Why is Feature Selection Important?

Reduces Dimensionality, Improving Computational Efficiency



High-dimensional datasets ("the curse of dimensionality") slow down training and increase memory usage. As the number of dimensions increases, the harder it gets to generate meaningful models from the exponential growth in configurations.

By trimming redundant features, models train faster, require less storage, and reduce hardware demands. For example, cutting 100 features to 20 speeds up algorithms like SVM or neural networks exponentially.

Why is Feature Selection Important?

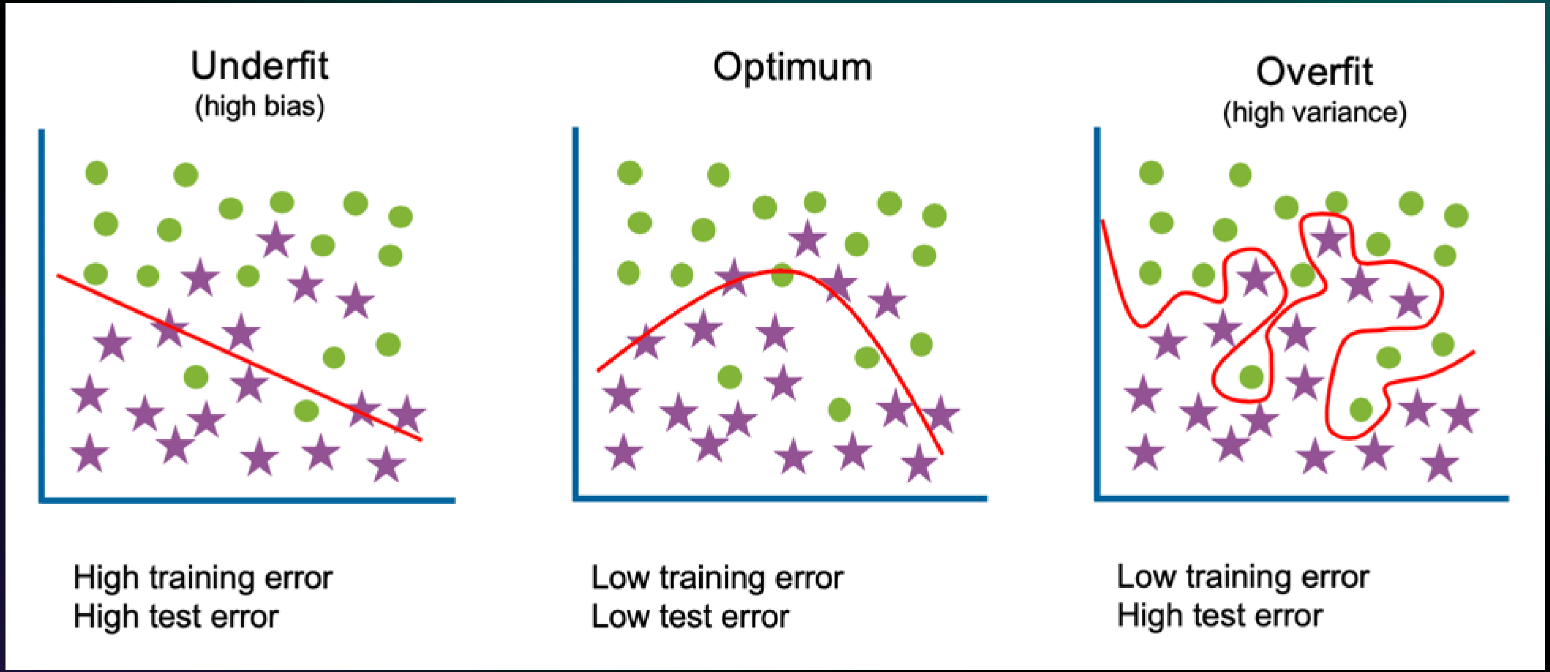
Eliminates Irrelevant/Noisy Features, Enhancing Model Accuracy

Non-predictive features (e.g., random IDs in customer data) act as distractions, while noisy features (e.g., sensor errors) inject false patterns.

Removing irrelevant features sharpens the model's focus on meaningful signals. For instance, dropping "timestamp" from a disease prediction model improves AUC-ROC by 15% if timestamps aren't causal.

Why is Feature Selection Important?

Helps Avoid Overfitting, Improving Generalisation



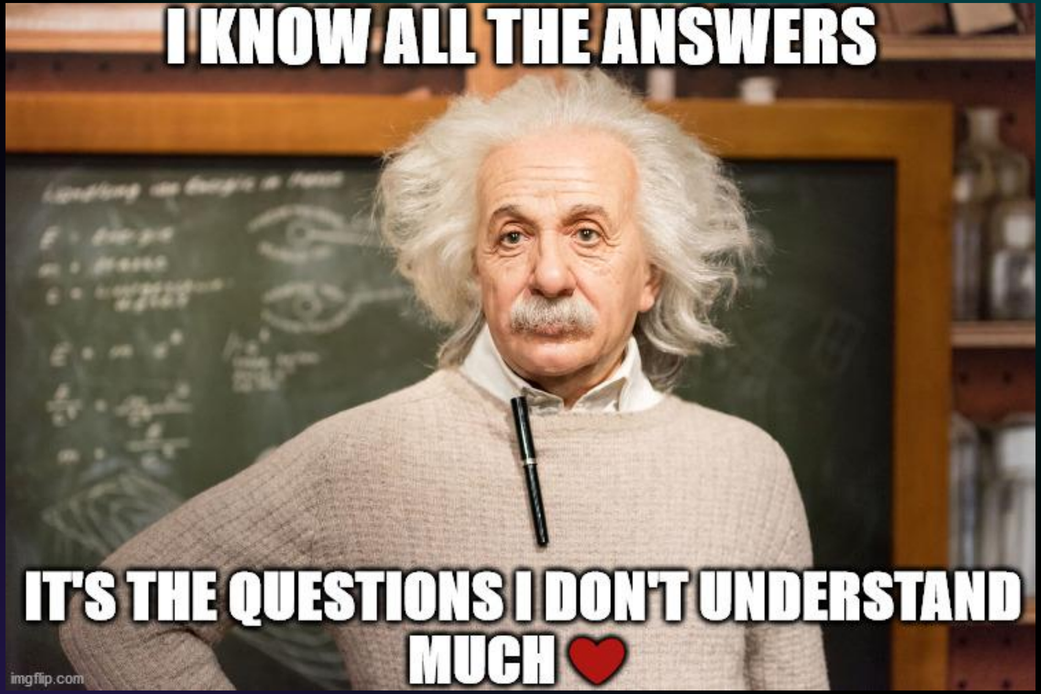
In machine learning, "overfitting" is a common problem that occurs when a model learns the training data too well, including its noise and random fluctuations.

Models with too many features memorise noise instead of learning patterns.

Feature selection simplifies models, reducing their capacity to chase outliers. A random forest with 50 features might achieve 85% test accuracy versus 70% with 200 features due to overfitting on irrelevant correlations.

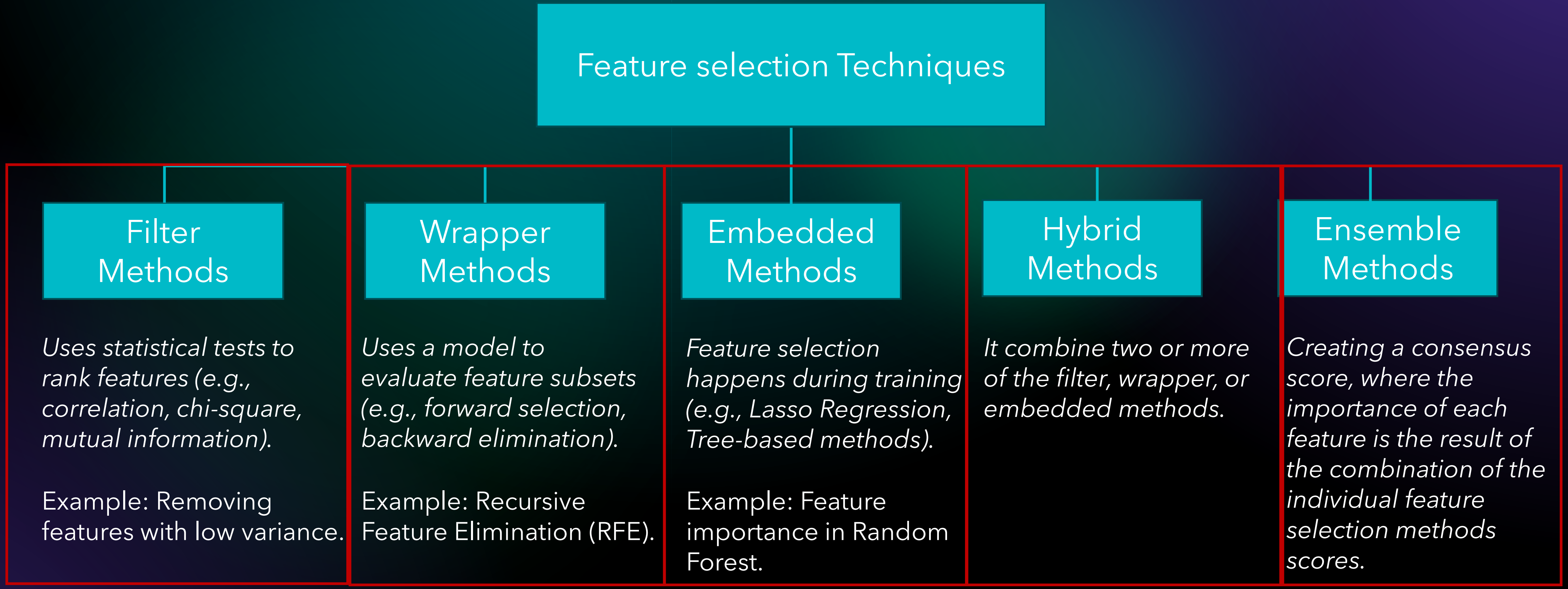
Makes Models Easier to Interpret and Explain

Fewer features = clearer decision logic.



For instance, a loan approval model using only "credit score" and "income" is transparent, while including 50 vague metrics like "social media activity" raises red flags for auditors. Stakeholders trust models when they can trace outcomes to key drivers.

Types of Feature Selection Methods



Feature selection Techniques

Filter Methods

Uses statistical tests to rank features (e.g., correlation, chi-square, mutual information).

Example: Removing features with low variance.

Wrapper Methods

Uses a model to evaluate feature subsets (e.g., forward selection, backward elimination).

Example: Recursive Feature Elimination (RFE).

Embedded Methods

Feature selection happens during training (e.g., Lasso Regression, Tree-based methods).

Example: Feature importance in Random Forest.

Hybrid Methods

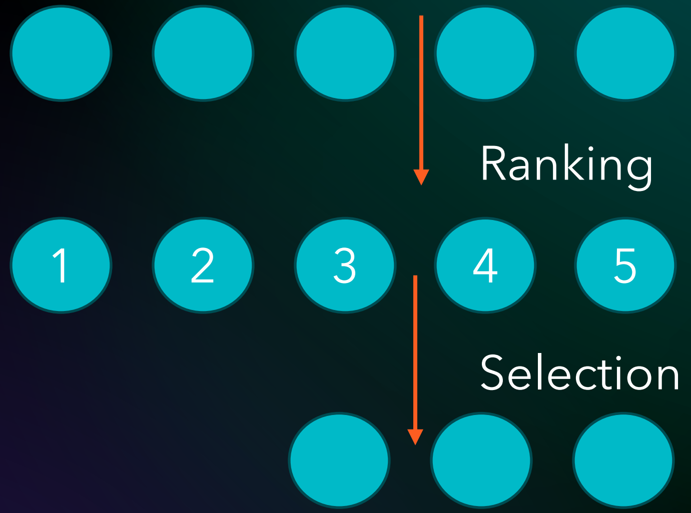
It combine two or more of the filter, wrapper, or embedded methods.

Ensemble Methods

Creating a consensus score, where the importance of each feature is the result of the combination of the individual feature selection methods scores.

Filter Methods (Statistical-based)

Filter methods rank and select features based on their statistical relationship with the target variable, helping remove irrelevant or redundant data before model training.



Example: Removing features with low variance.

Correlation, Chi-square, Mutual information.

They are fast, model-independent, and improve computational efficiency but may overlook feature interactions.

Filter Methods

Advantages:

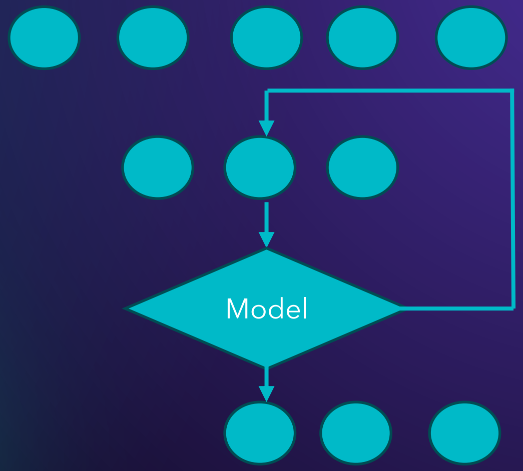
- Fast and inexpensive: Evaluates features without training a model.
- Helps remove redundant or highly correlated features.

Limitations: Does not consider interactions between features, potentially missing important combinations.

Common Techniques

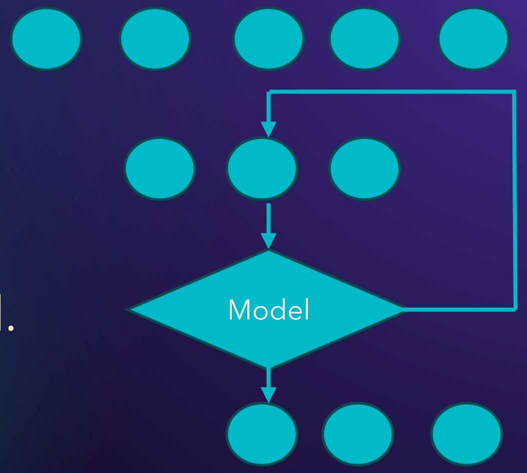
- **Information Gain:** Measures reduction in entropy to assess feature importance.
- **Chi-Square Test:** Evaluates relationships between categorical variables.
- **Fisher's Score:** Ranks features based on their discriminative power.
- **Correlation Coefficient:** Measures linear relationships between continuous variables.
- **Variance Threshold:** Removes low-variance features under a set threshold.
- **Mean Absolute Difference (MAD):** Similar to variance threshold but without squaring.
- **Dispersion Ratio:** Ratio of arithmetic to geometric mean, identifying more relevant features.

Wrapper Methods (Model-based)



- Iteratively selects feature subsets by training and evaluating a model.
- Uses techniques like Forward Selection, Backward Elimination, and Recursive Feature Elimination (RFE) to find the best combination.
- Stops when performance no longer improves, or a predefined number of features is reached.
- **Pros:** Finds the optimal feature subset for a given model.
- **Cons:** Computationally expensive, especially for large datasets.

Wrapper Methods



Advantages

- Improves model performance by selecting features in the context of the model.
- Captures feature dependencies and interactions.

Limitations

Computationally expensive, especially for large datasets.

Techniques

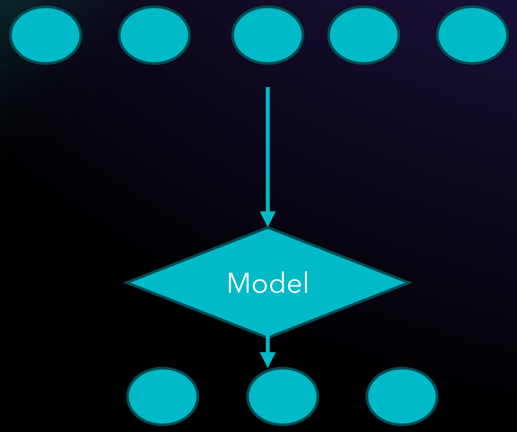
- **Forward Selection:** Starts with no features, adds one at a time until no improvement.
- **Backward Elimination:** Starts with all features, removes the least significant iteratively.
- **Recursive Feature Elimination (RFE):** Eliminates least important features step by step until the optimal subset remains.

Embedded Methods (Integrated into model)

- Performs feature selection during model training, combining filter and wrapper method benefits.
- More efficient than wrapper methods and often scales better to large datasets.
- Works well with specific learning algorithms but may not generalise across all models.

Techniques:

- **Lasso Regression (L1 Regularisation):** Shrinks coefficients, removing less important features.
- **Decision Trees/Random Forest:** Selects features based on node splits (e.g., Gini impurity, information gain).
- **Gradient Boosting:** Prioritises features that reduce model error the most.



How to select the Right Feature Selection Method

- **Dataset Size:** Filter methods work best for large datasets due to efficiency.
- **Feature Interactions:** Wrapper and embedded methods are better for capturing complex relationships.
- **Model Type:** Some techniques are specific to certain models (e.g., Lasso for linear models, Decision Trees for tree-based models).

Example:

- *Use Filter Methods (e.g., correlation, variance threshold) to quickly remove irrelevant features in high-dimensional datasets.*
- *If maximizing model performance is the priority, explore Wrapper (e.g., RFE) or Embedded Methods (e.g., Lasso).*

Key Takeaway: Selecting the right method improves accuracy, reduces overfitting, and makes models more interpretable.

Best Practices

- Avoid removing too many features–retain enough information.
- Check feature correlation to prevent redundancy.
- Use domain knowledge for better feature selection.
- Validate performance using cross-validation.
- Always experiment and validate feature selection choices.

Questions and Answers