Dimensionality Reduction
Part 2: Feature Selection

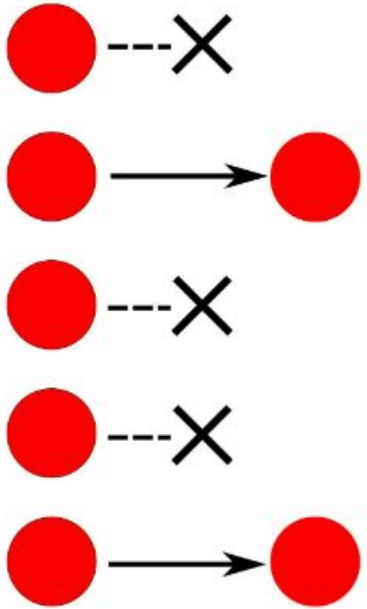
Mohammed Brahimi & Sami Belkacem

Outline

- What is Feature Subset Selection (FSS)?
- Why doing Feature Subset Selection?
- Taxonomy of Feature Selection methods
  - **Unsupervised** Feature Selection
  - **Supervised** Feature Selection
    - Filter methods for FSS
    - Wrapper methods for FSS
- Comparison of Feature Selection methods

What is Feature Subset Selection (FSS)?

*Feature Selection (FSS) is the process of **selecting a subset of features** from a given feature set.*
*The goal is to **optimize an objective function** for the chosen features.*



- FSS can be seen as a form of feature extraction.
- However, the methods and goals differ significantly.

- **Feature extraction:**
  - Transforming the existing features into a lower dimensional space.
- **Feature selection:**
  - Selecting a subset of the existing features without a transformation.

What is Feature Subset Selection (FSS)?

Feature Set:

$$X = \{x_i \mid i = 1, \ldots, n\}$$

Objective:

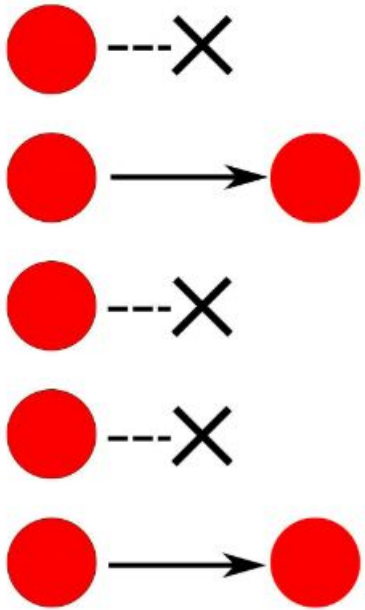
Find a subset:

$$Y_m = \{x_{i_1}, x_{i_2}, \ldots, x_{i_m}\} : m < n$$

that optimizes the objective function $J(Y_m)$

Optimization:

$$Y = \underset{m, i_m}{\arg\max}\, J(Y_m)$$



Why doing Feature Subset Selection?

Why not stick to feature extraction, as it essentially accomplishes the same task of dimensionality reduction?

Why doing Feature Subset Selection?

- **Features may be expensive to obtain**
  - Evaluate a large number of features (sensors) in the test bed and select a few for the final implementation
- **You may want to extract meaningful rules from your classifier**
  - When you transform or project, the measurement units (length, weight, etc.) of your features are lost.
- **Features may not be numeric**
  - A typical situation in the data mining.
- **Fewer features means fewer parameters for pattern recognition**
  - Improved the generalization capabilities and avoid overfitting.
  - Reduced complexity and run-time.

Why Feature selection is challenging?

- Number of features = 3
  - {A1, A2, A3}

- How many features subsets?
  - {A1}, {A2}, {A3}
  - {A1, A2}, {A1, A3}, {A2, A3}
  - {A1, A2, A3}

$$7 = 2^3 - 1$$

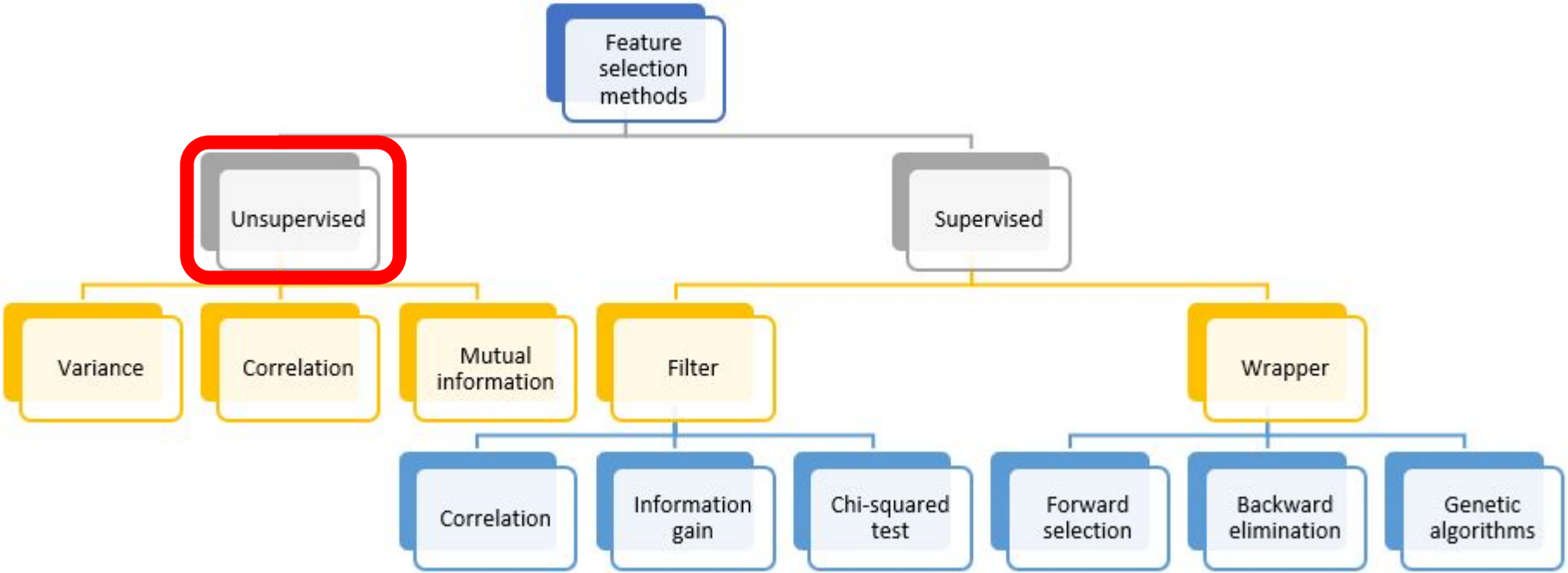
Why Feature selection is challenging?

- Number of features = ***n***
  - {A1, A2, A3, …, An}
- How many features subsets?
  - {A1}, {A2}, {A3} …..
  - {A1, A2}, {A1, A3}, {A2, A3}....
  - {A1, A2, A3} ….
  - ……
- For ***n***=100

1 267 650 600 228 229 401 496 703 205 376

$$2^n - 1$$

Taxonomy of Feature Selection methods



Unsupervised Feature Selection

*Unsupervised feature selection typically uses **statistical measures** to evaluate features. The most common statistical measures include:*

- **Variance:** Features with high variance are more likely to be informative and useful for prediction.
- **Correlation:** Features that are highly correlated with each other are likely to contain redundant information.
- **Mutual information:** Measures the amount of information that two features share. Features with high mutual information are more likely to be predictive of each other.

Variance Filter

- Remove descriptors with a high percentage of identical values for all objects.

	Feature 1	Feature 2	Feature 3	Feature 4
Object 1	4	8	0	16
Object 2	17	34	0	7
Object 3	4	8	0	3
Object 4	15	30	0	6
Object 5	8	16	0	4
Object 6	15	30	0	7

- Why ?

Absence of information in the constant descriptors

Correlation Filter

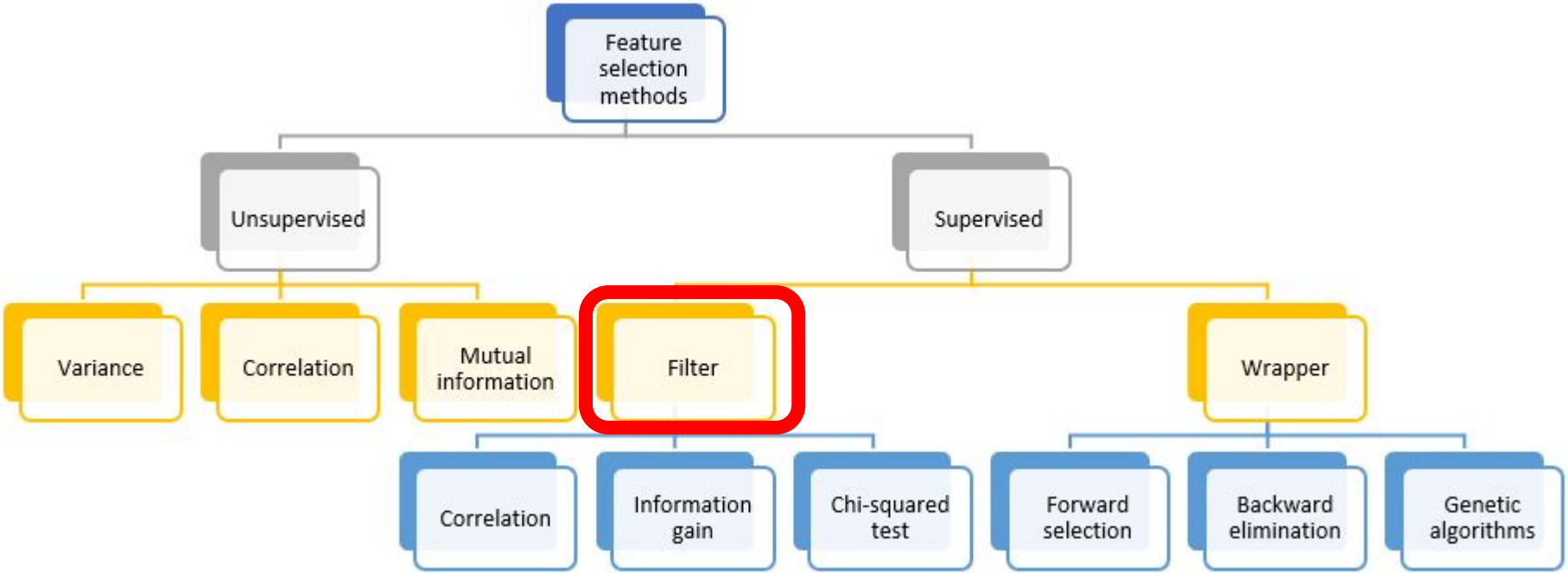
- Remove correlated filters

	Feature 1	Feature 2	Feature 3	Feature 4
Object 1	4	8	0	16
Object 2	17	34	0	7
Object 3	4	8	0	3
Object 4	15	30	0	6
Object 5	8	16	0	4
Object 6	15	30	0	7

- Why ?

Redundant information in descriptors (Attribute 2 = 2 * Attribute 1)

Taxonomy of Feature Selection methods



Filter methods for FSS

*Filter methods evaluate **features independently** and select the **most important features** based on **statistical measures**, such as:*
***correlation** with the target variable, **information gain**, or **chi-squared test**.*

How to use filter methods for FSS

1. Choose a filter metric (correlation, information gain, or chi-squared test).
2. Evaluate each feature using the chosen filter metric and rank them accordingly.
3. Select the features with the highest filter metric scores.

Filter methods: Correlation

- Correlation is a widely used filter method in feature selection.
- It assesses the strength and direction of a relationship between two variables.
- It measures how well a feature relates to the target variable.

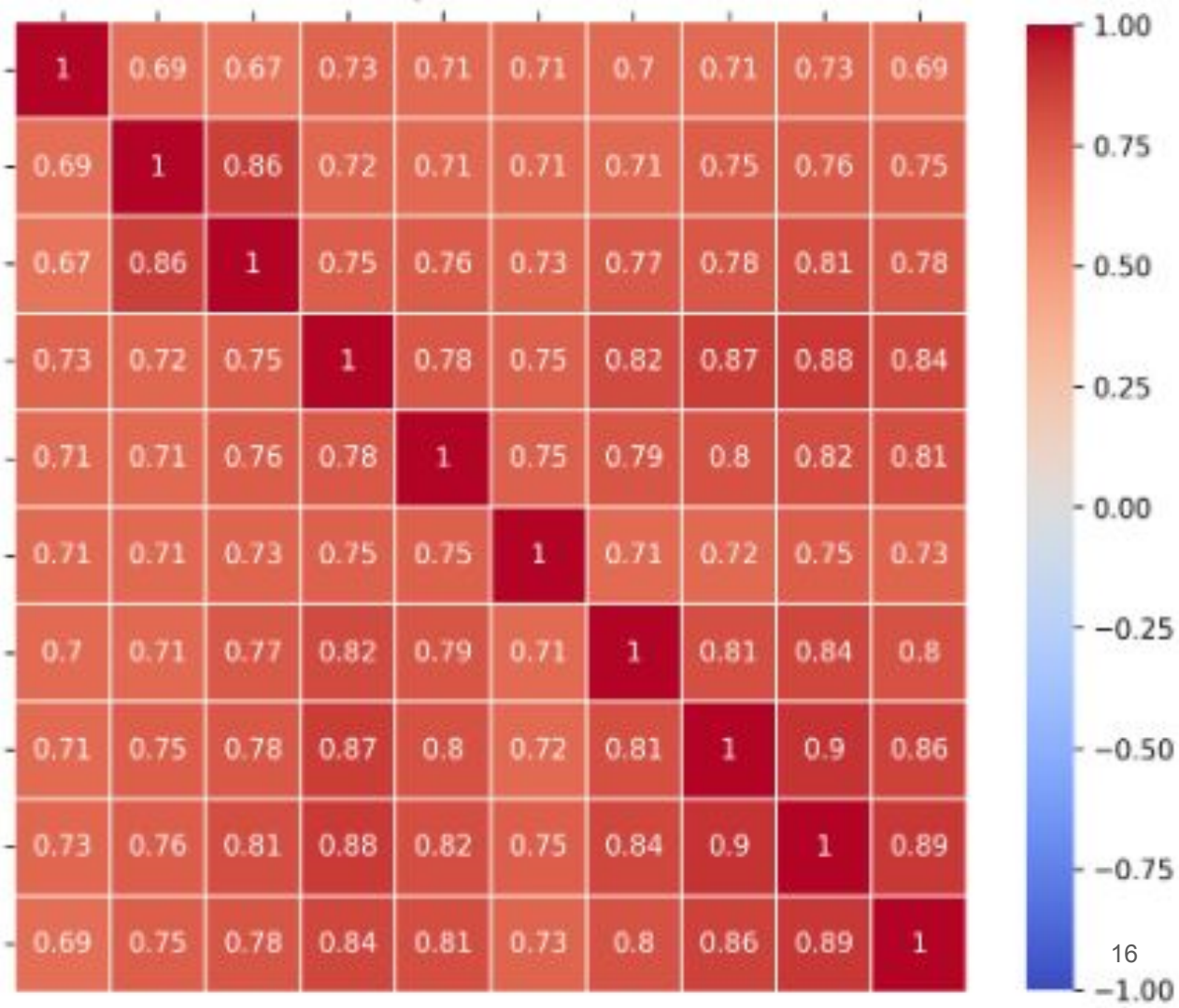
- **High correlation with the target:**
  - Features strongly correlated with the target variable are often good candidates for predictive models
- **Low inter-feature correlation:**
  - It's also essential to consider inter-feature correlations to avoid redundancy in the feature set

The correlation does not consider the non-linear correlation with the target and the interaction between features

Correlation Heatmap

- Warm colors for positive.
- Cool colors. for negative.
- Neutral for no correlation.

Helps in visualizing the correlation between attributes



Filter methods: Information Gain

Information gain is a measure of ***how much information is gained about the target categorical variable when a dataset is split on a given categorical feature****.*

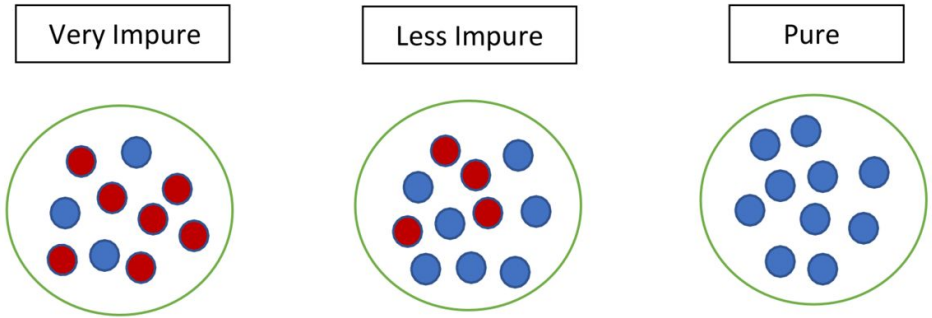
It is calculated as the difference between the ***entropy*** *of the* ***target variable*** *before and after the split.*

What is entropy?

What is entropy ?

- It measures the level of **uncertainty**, **randomness**, or **disorder** in a system or dataset.

- In the context of data, entropy quantifies the **impurity** of a set.



What is entropy ?

- It measures the level of **uncertainty**, **randomness**, or **disorder** in a system or dataset.

- In the context of data, entropy quantifies the **impurity** of a set.

$$H(S) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i)$$

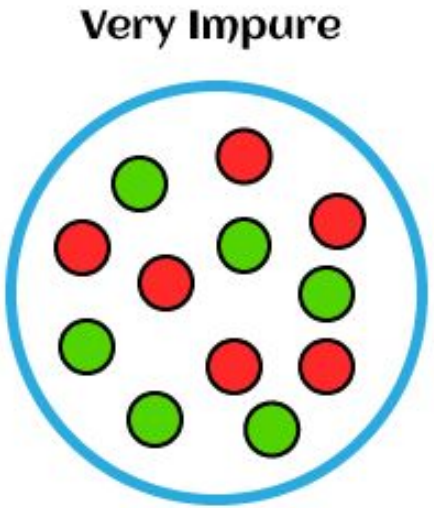
- ***S=*** *{x1,x2 ….xn},* which is the set of elements.

What is entropy ?

$$H(S) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i)$$

- p(red) = 6/12=0.5
- p(green) = 6/12=0.5
- H(S) = -(0.5*log(0.5)+0.5*log(0.5)) = -(-0.5-0.5)

H(S) = 1

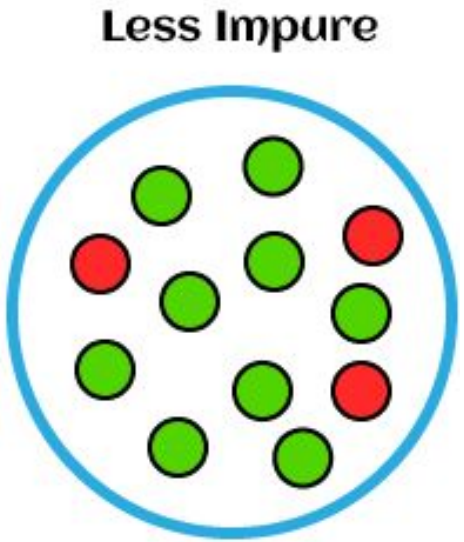


What is entropy ?

$$H(S) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i)$$

- p(red) = 3/12
- p(green) = 9/12
- H(S) = -((3/12)*log(3/12)+9/12*log(9/12))

H(S) = 0.811278

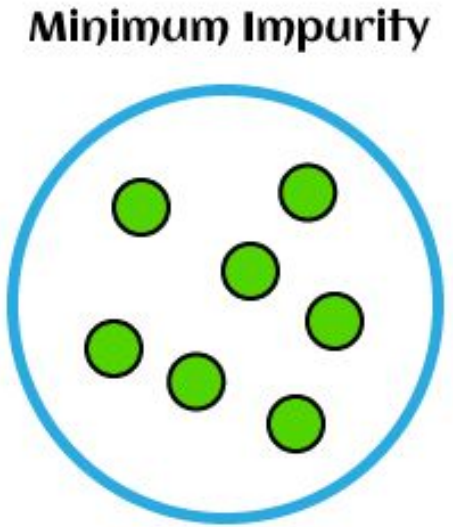


What is entropy ?

$$H(S) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i)$$

- p(red) = 0
- p(green) = 1
- H(S) = -(0+1*log(1))

H(S) = 0



Filter methods: Information Gain

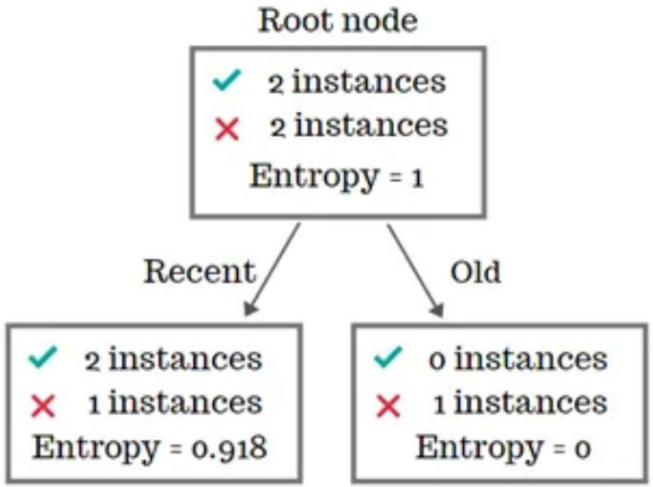
Information gain is a measure of ***how much information is gained about the target <u>categorical</u> variable when a dataset is split on a given <u>categorical</u> feature****.*

It is calculated as the difference between the ***entropy*** *of the* ***target variable*** *before and after the split.*

$$IG(X) = H(S) - \sum_{i=1}^{n} p(x_i) H(S|x = x_i)$$

Example: Information gain for Age

Age	Mileage	Road Tested	Buy
Recent	Low	Yes	Buy
Recent	High	Yes	Buy
Old	Low	No	Don't buy
Recent	High	No	Don't buy



- IG(Age) = H(Root node) - p(Age = Recent) *H(Recent) - p(Age = Old) *H(Old)

  = 1 - (¾)*0.918 – (¼)*0

  IG(Age) = 0.3115

Example: Information gain for Road Tested

Age	Mileage	Road Tested	Buy
Recent	Low	Yes	Buy
Recent	High	Yes	Buy
Old	Low	No	Don't buy
Recent	High	No	Don't buy

Pure set

Pure set

- IG(Road Tested) = 1 - 0.5 *0 - 0.5 *0

  = 1 - 0

IG(Age) = 1 (Perfect predictor)

Filter methods for FSS

Advantages

- **Computational Efficiency**

Filter methods are computationally efficient and scalable to large datasets and high dimensionality as no machine learning model is trained.

- **Algorithm Independence**

Independent of any machine learning algorithm, making them unbiased toward any particular algorithm.

Disadvantages

- **Suboptimal Selection**

May not always identify the best feature subset for a specific machine learning algorithm because it does not consider the model performance.

- **Complex Relationships**

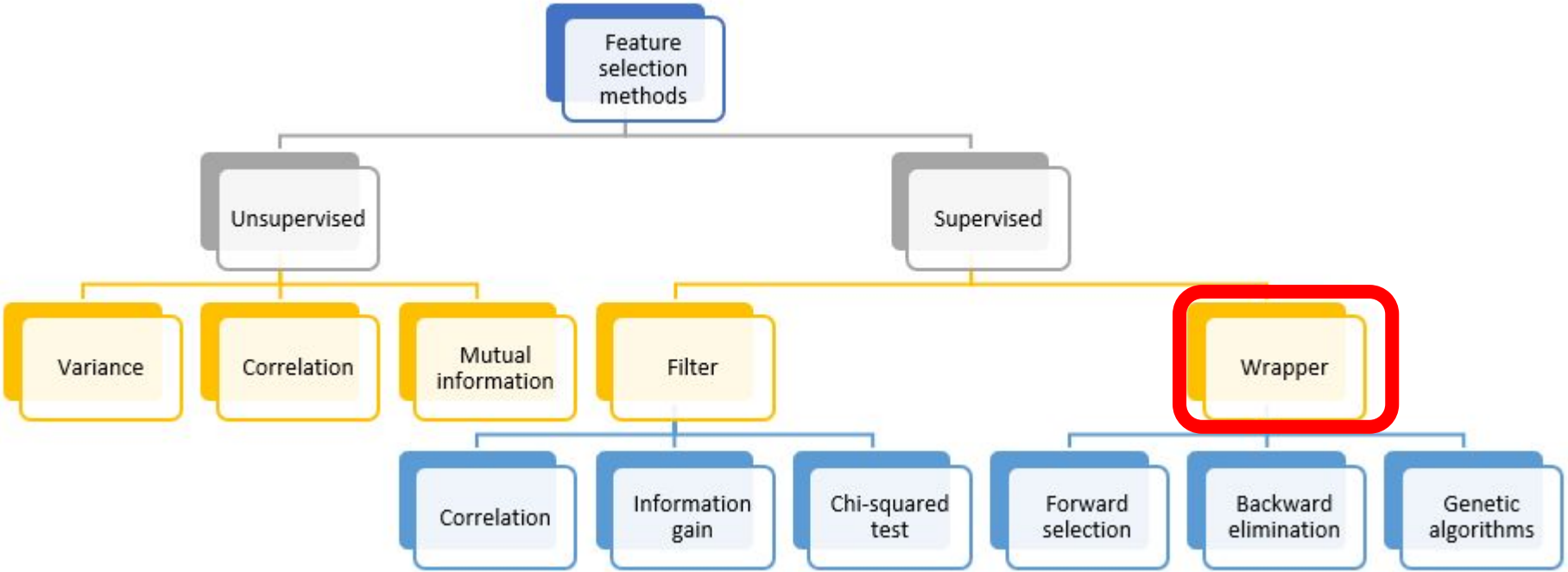
May not be able to handle complex relationships and interactions between features.

How to use filter methods for FSS

- **Filter Methods as an initial step**
  - Filter methods serve as an excellent initial step in the feature selection process.
  - Particularly valuable when dealing with large datasets.

- **Hybridization is key**
  - It's advisable to combine filter methods with other feature selection techniques.
  - Wrapper methods and embedded methods complement filter methods effectively.

Taxonomy of Feature Selection methods



Wrapper methods for FSS

*Wrapper methods are feature selection methods that **evaluate features by using a machine learning model as a black box**.*

*The model is trained on different subsets of features, and **the subset that produces the best performance is selected.***

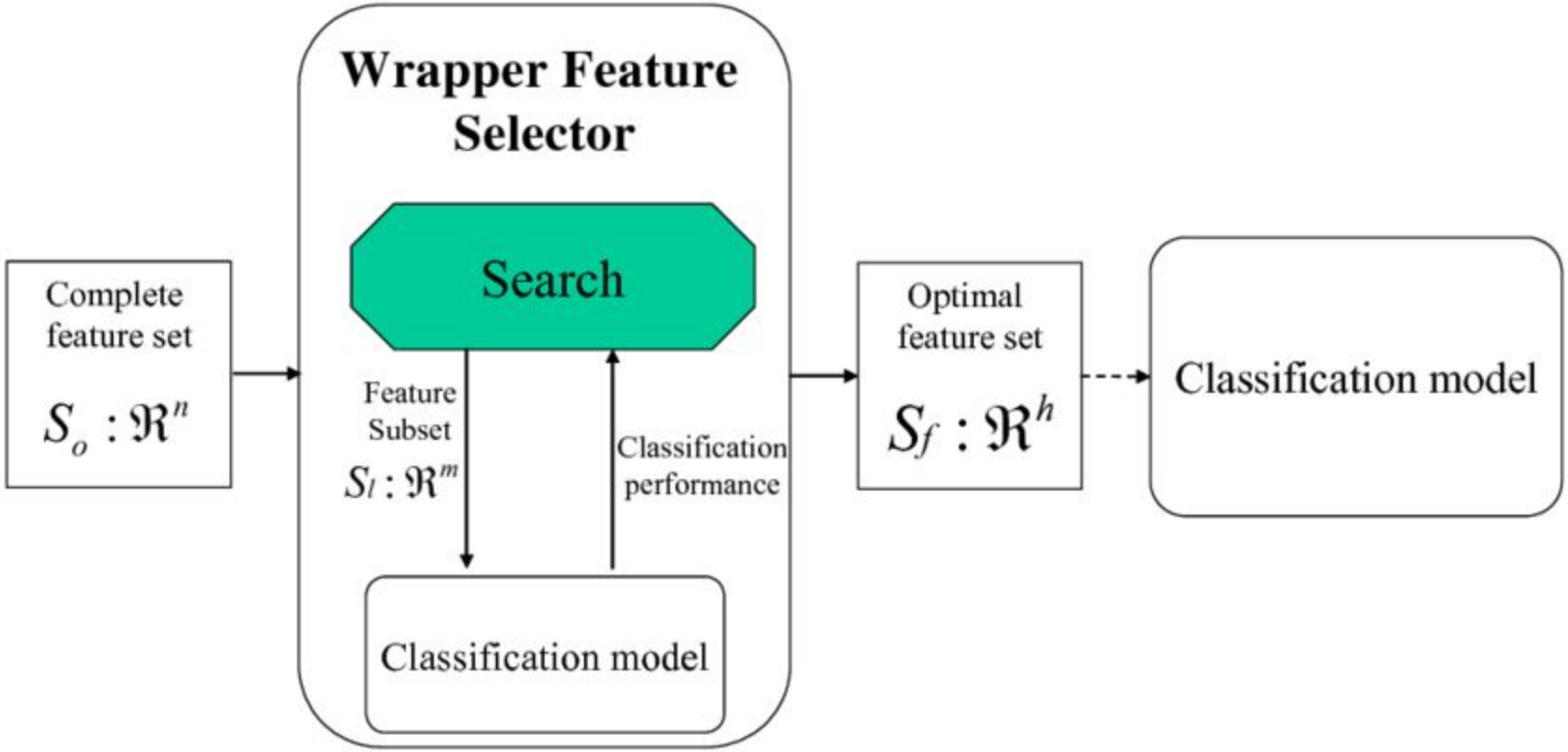
How to use filter methods for FSS

Feature Subset Generation: The search algorithm generates various feature subsets for evaluation

Model Evaluation: Each subset is input to the model and its performance is evaluated via model performance

Selection Criterion: A selection criterion (e.g. accuracy) guides the choice of the best-performing feature subset

Iteration: The process is repeated with different subsets until an optimal set is found

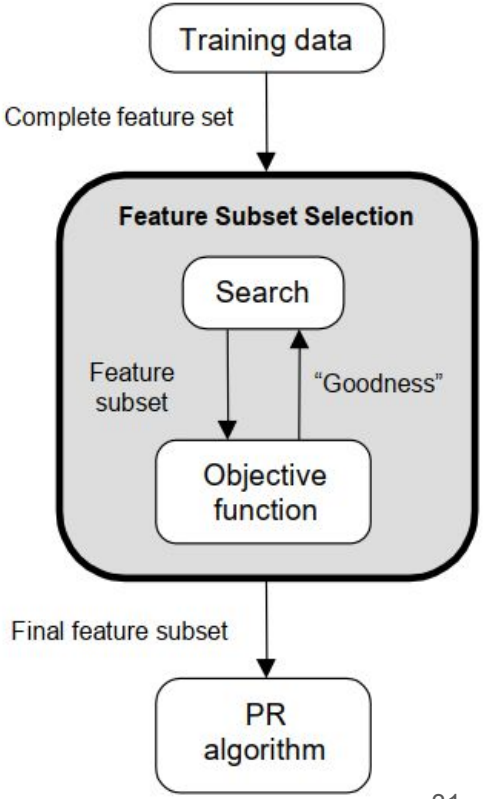

Wrapper Feature Selector
Search
Complete feature set
$S_o : \Re^n$
Feature Subset
$S_l : \Re^m$
Classification performance
Classification model
Optimal feature set
$S_f : \Re^h$
Classification model

Search strategy and objective function

Feature Subset Selection :

- **Search Strategy:** A strategy to select candidate subsets
  - Exhaustive evaluation of all subsets is impossible.
  - A search strategy navigates the vast feature subset space efficiently

- **Objective Function:** Function to evaluate these candidates
  - Evaluates candidate feature subsets
  - Quantitative measure of the "goodness" or quality of a subset
  - Feedback from the objective function to guide the search strategy



Search strategy and objective functions

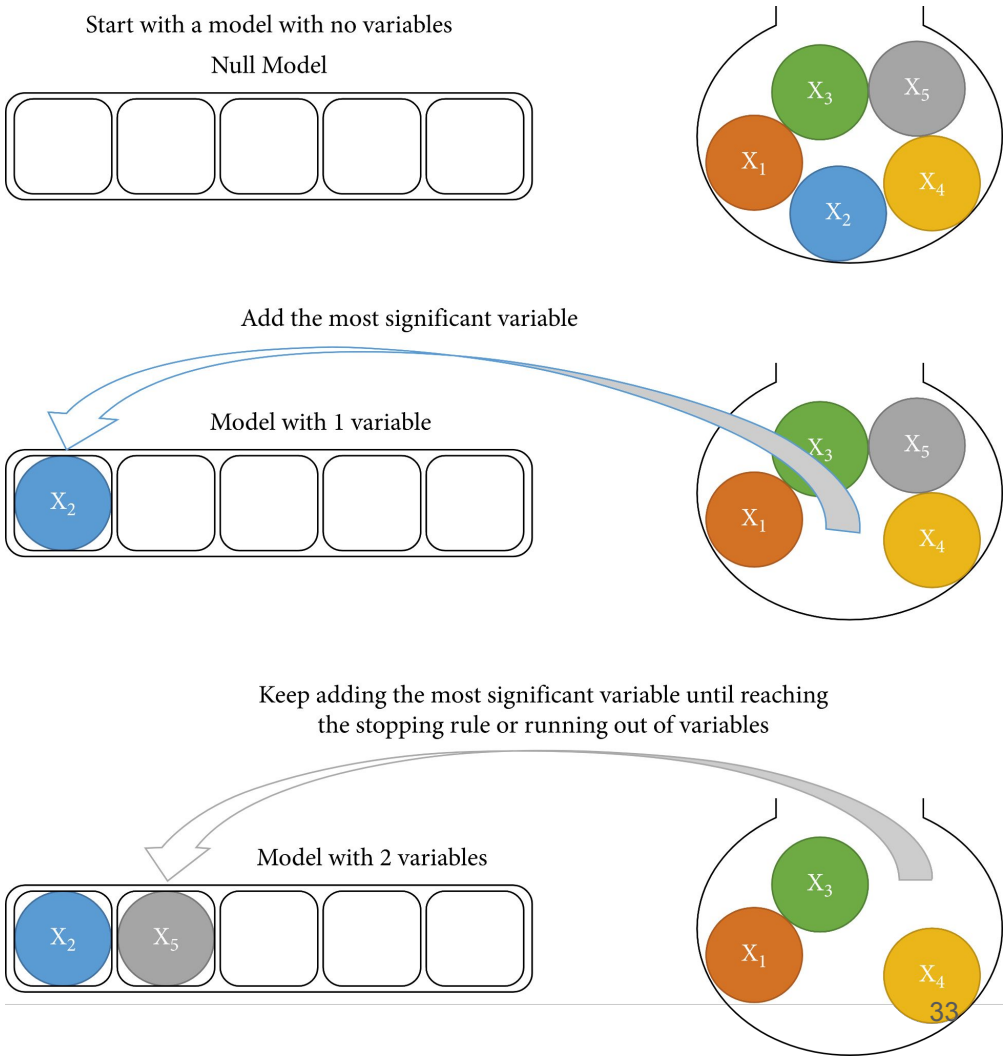
- **Forward selection**
  - Add features step by step, choosing the one that improve the model the most at each iteration.
- **Backward elimination**
  - Remove features step by step, choosing the one that does not improve the model the at each iteration.
- **Genetic Algorithm (GA)**
  - An optimization technique, inspired by natural selection, used to evolve feature subsets for improved model performance.

Forward Selection

1. Start with 0 features

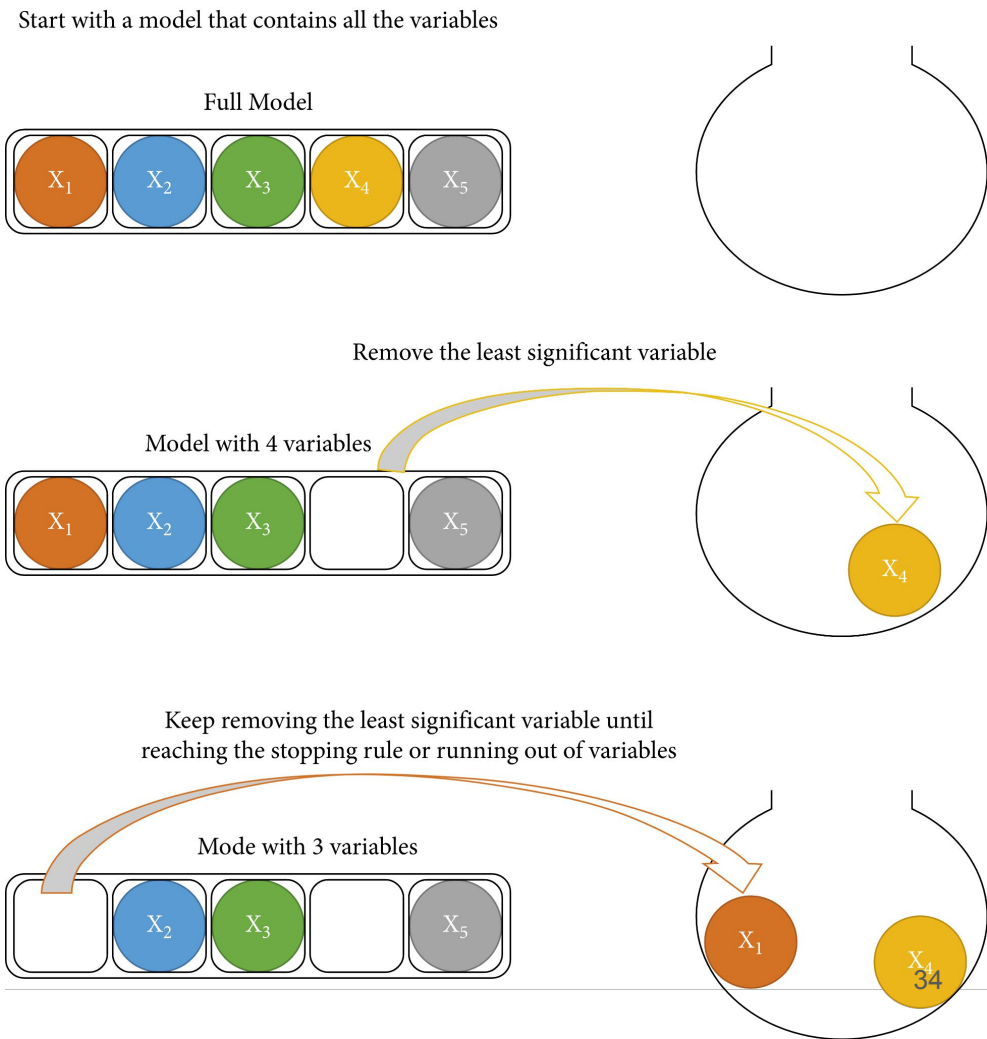
2. Add the best feature

3. Stop adding features
   a. Reach the number of features.
   b. Performance threshold.
   c. Elbow methods.



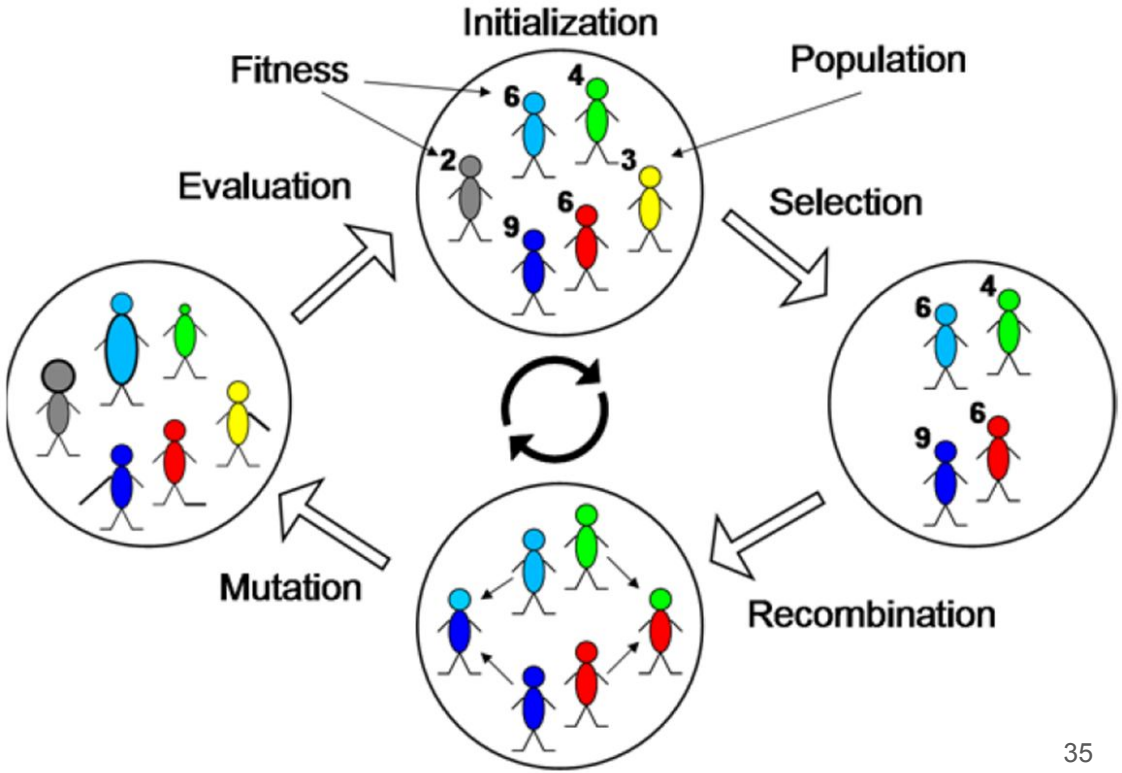
Backward elimination

1. Start with all feature
2. Remove the worst feature
3. Stop removing features
   a. Reach the number of features.
   b. Performance threshold.
   c. Elbow methods.



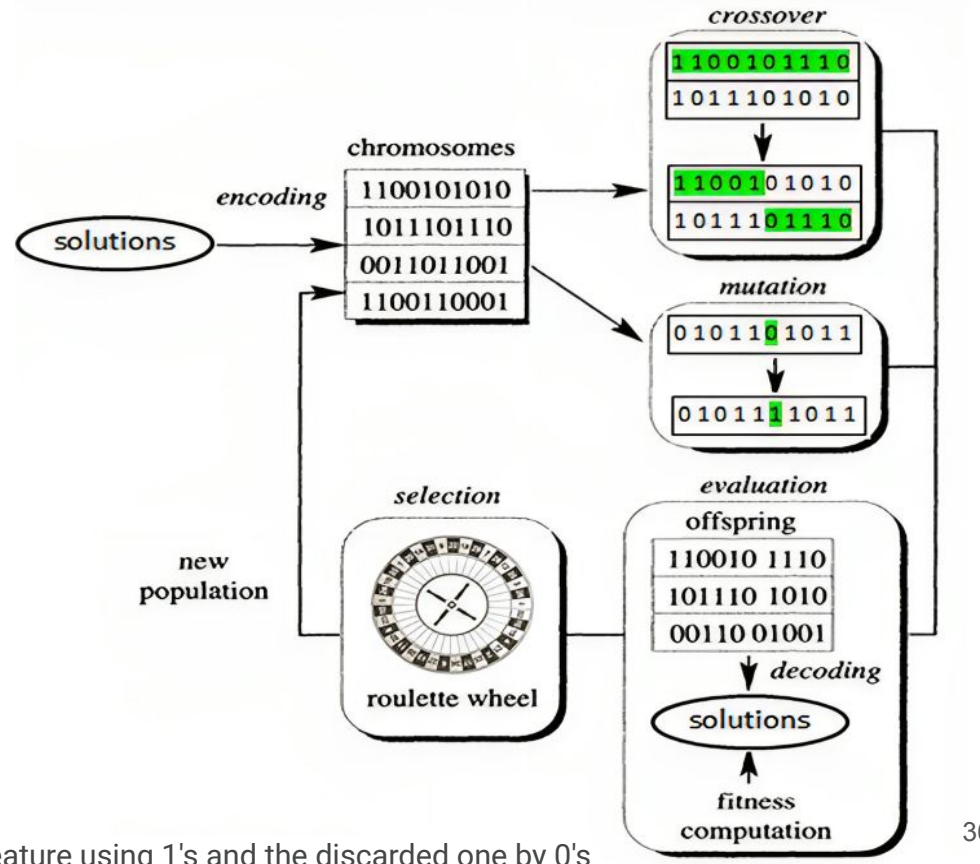
Genetic Algorithm (GA) for Feature Selection

What is a genetic algorithm?



Genetic Algorithm (GA) for Feature Selection

- **Crossover**
  - Combine two solutions
  - Produce good children
- **Mutation**
  - Random change
  - Exploration
- **Selection**
  - Select the best
  - Survival of fittest



N.B. In feature selection and GA, we can encode the selected feature using 1's and the discarded one by 0's

Wrapper methods for FSS

Advantages

- **Optimal Subset Selection**

Can detect the ideal feature subset for a given machine learning algorithm.

- **Managing Complex Relationships**

They are effective in handling relationships between features by evaluating a subset, not a feature.

Disadvantages

- **Computational Intensity**

Tend to be computationally demanding, particularly with large datasets.

- **Algorithm Bias**

They may exhibit bias towards the machine learning algorithm used for feature evaluation.

Comparison of Feature Selection methods

- **Unsupervised methods**
  - **Advantages:** Simple, fast, no need for labels
  - **Disadvantages:** May miss complex relationships, doesn't optimize for prediction
- **Filter methods**
  - **Advantages:** Fast, scalable, algorithm independent
  - **Disadvantages:** May miss complex relationships, suboptimal for specific algorithms
- **Wrapper methods**
  - **Advantages:** Finds optimal subset for algorithm, handles complex relationships
  - **Disadvantages:** Slow, algorithm bias
- **Forward selection**
  - **Advantages:** Starts with no features, adds features incrementally
  - **Disadvantages:** suboptimal and does not consider feature interaction.
- **Backward elimination**
  - **Advantages:** Starts with all features, removes features incrementally
  - **Disadvantages:** Can be slow for high dimensional data
- **Genetic algorithms**
  - **Advantages:** Searches and optimizes feature sets through evolution
  - **Disadvantages:** Computationally intensive but thorough search

Comparison of Feature Selection methods

Method	Advantages	Disadvantages
Unsupervised methods	• Simple & Fast • No need for labels	• May miss complex relationships. • Doesn't optimize for prediction
Filter methods	• Fast • Scalable • algorithm independent	• May miss complex relationships • Suboptimal for specific algorithms
Wrapper methods	• Finds optimal subset for algorithm • Handles complex relationships	• Slow • algorithm bias
Forward selection	• Fast and simple	• Can be suboptimal for high-dimensional data
Backward elimination	• Can be optimal for high dimensional data	• Can be slow
Genetic algorithms	• Thorough search	• Computationally intensive

How to choose a Feature Selection method?

- **Size of dataset**
  - Filter methods work well for large datasets
  - Wrapper methods work well for small datasets
- **Computational budget**
  - Filter methods are fast
  - Wrapper methods are slow
- **Need for interpretability**
  - Filter methods allow inspection of feature importance
  - Wrapper methods act as black box
- **Domain knowledge**
  - Unsupervised methods do not use labels
  - Supervised methods use domain labels
- **Algorithm fit**
  - Wrapper methods tailor to specific ML algorithm
  - Filter methods are independent of any algorithm
- **Feature interactions**
  - Wrapper methods handle feature interactions
  - Filter methods assess features independently

Summary

The goal is to select an optimal subset of features

- **Main approaches:**
  - Unsupervised**:** Evaluate features independently using statistical measures
  - Supervised Filter**:** Fast evaluation of features using metrics like correlation
  - Supervised Wrapper**:** Use model performance to search feature subsets, slow but thorough
- **Considerations:**
  - Dataset size and dimensionality
  - Computational budget and time constraints
  - Need for interpretability vs performance
  - Domain knowledge of features
  - Relationships between features
- No one-size-fits-all method - choose based on goals, data, and constraints
- Hybrid approaches combine strengths of different techniques
- Feature selection critical step for ML pipelines, balances model performance and efficiency