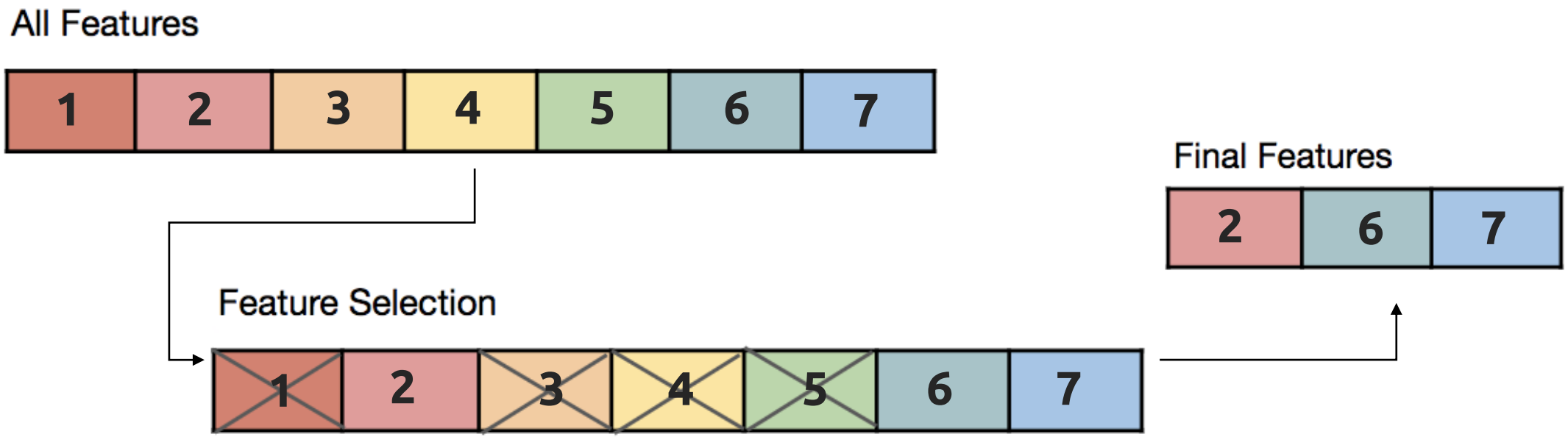
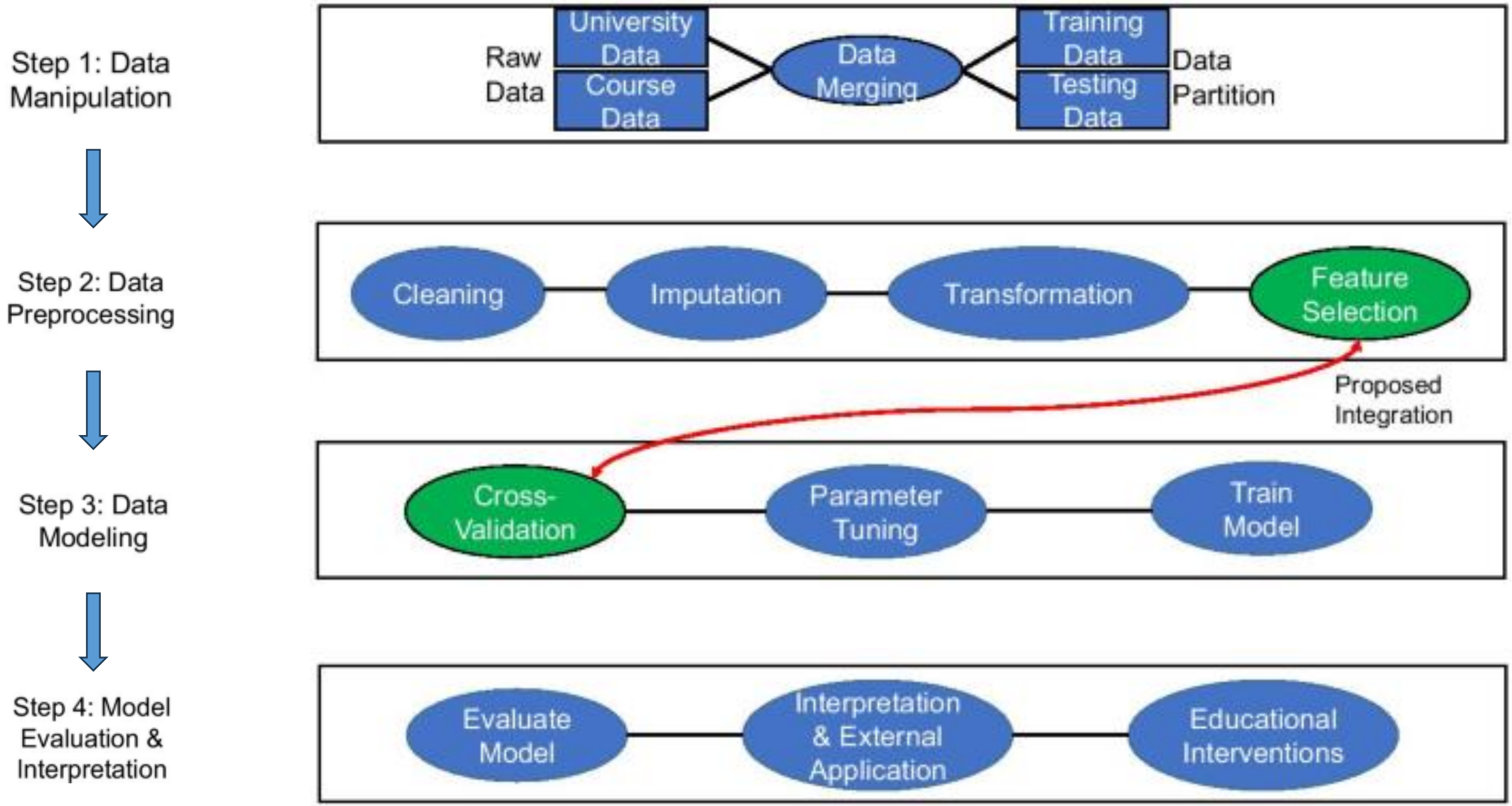
Feature Selection

Choosing the Right Variables for Your Analysis

Prepared By: Ashraf Abdulkhaliq



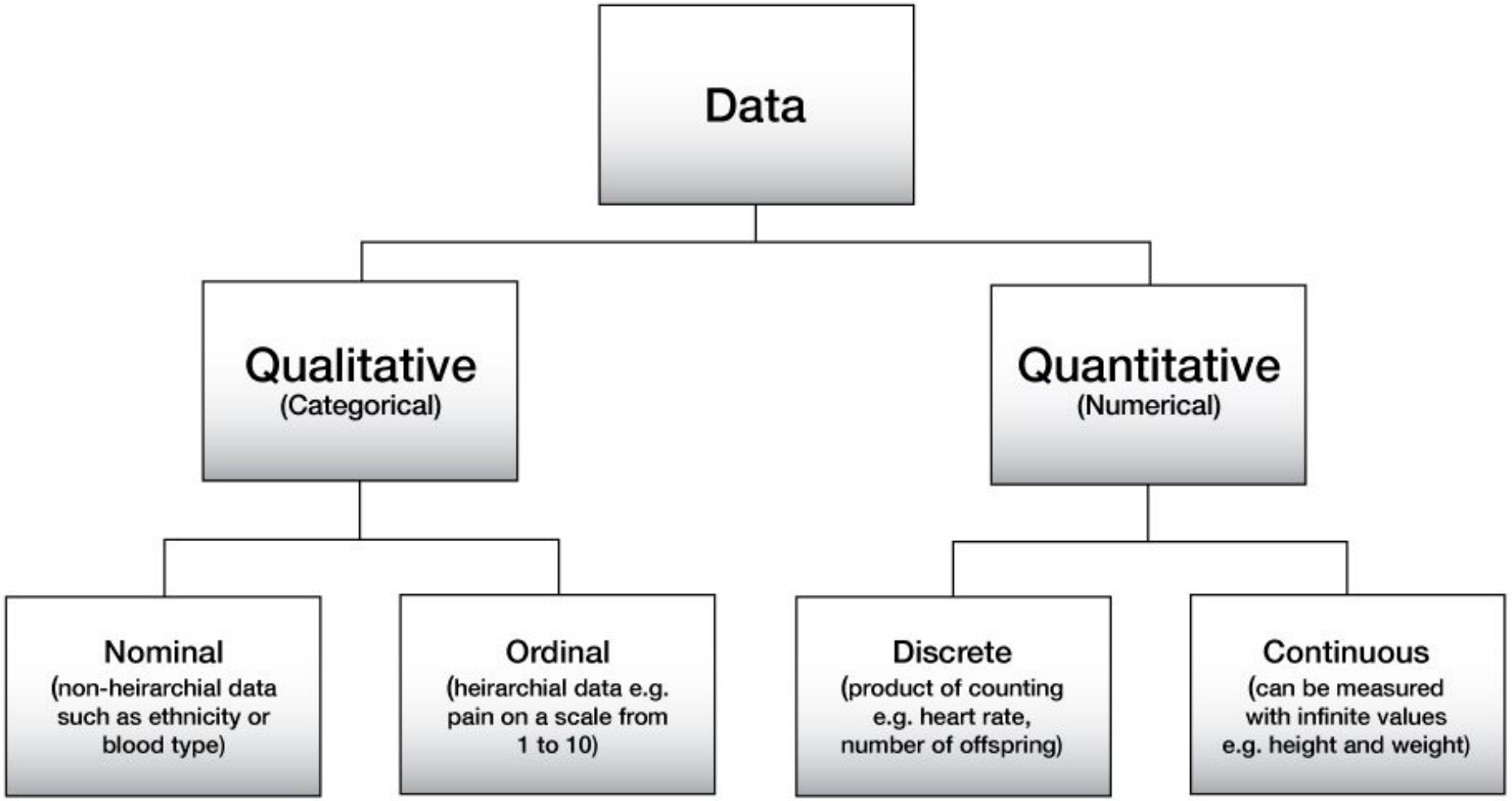
Feature selection techniques are employed to **reduce** the number of input variables/features by eliminating **redundant** or **irrelevant features**.


Step 1: Data Manipulation
Raw Data
University Data
Course Data
Data Merging
Training Data
Testing Data
Data Partition
Step 2: Data Preprocessing
Cleaning
Imputation
Transformation
Feature Selection
Proposed Integration
Step 3: Data Modeling
Cross-Validation
Parameter Tuning
Train Model
Step 4: Model Evaluation & Interpretation
Evaluate Model
Interpretation & External Application
Educational Interventions

Why Feature Selection

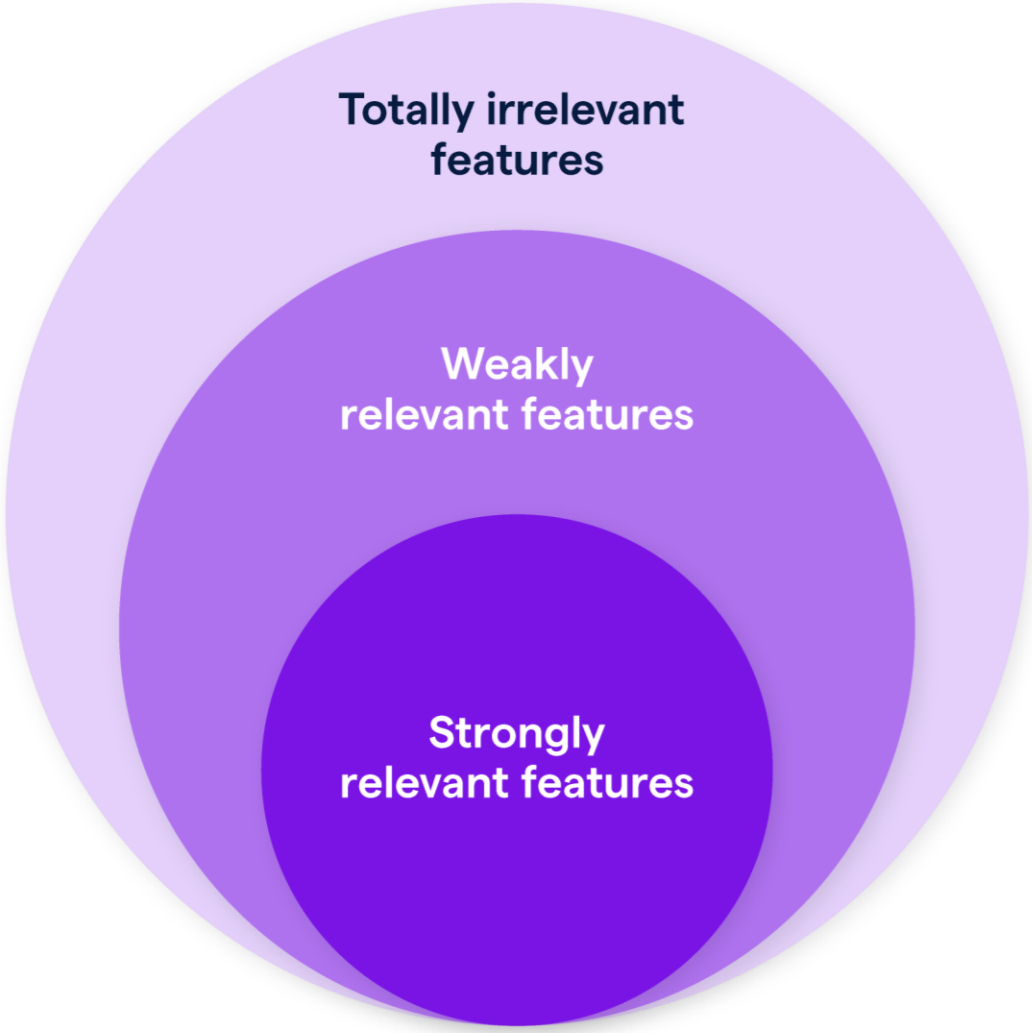
1. Helps reduce **dimensions**.
2. Reduces the **training time**.
3. Helps to make **simple** model that can be easily interpreted.
4. Reduces **overfitting** hence enhance the **generalization**.

Types of Features/DATA

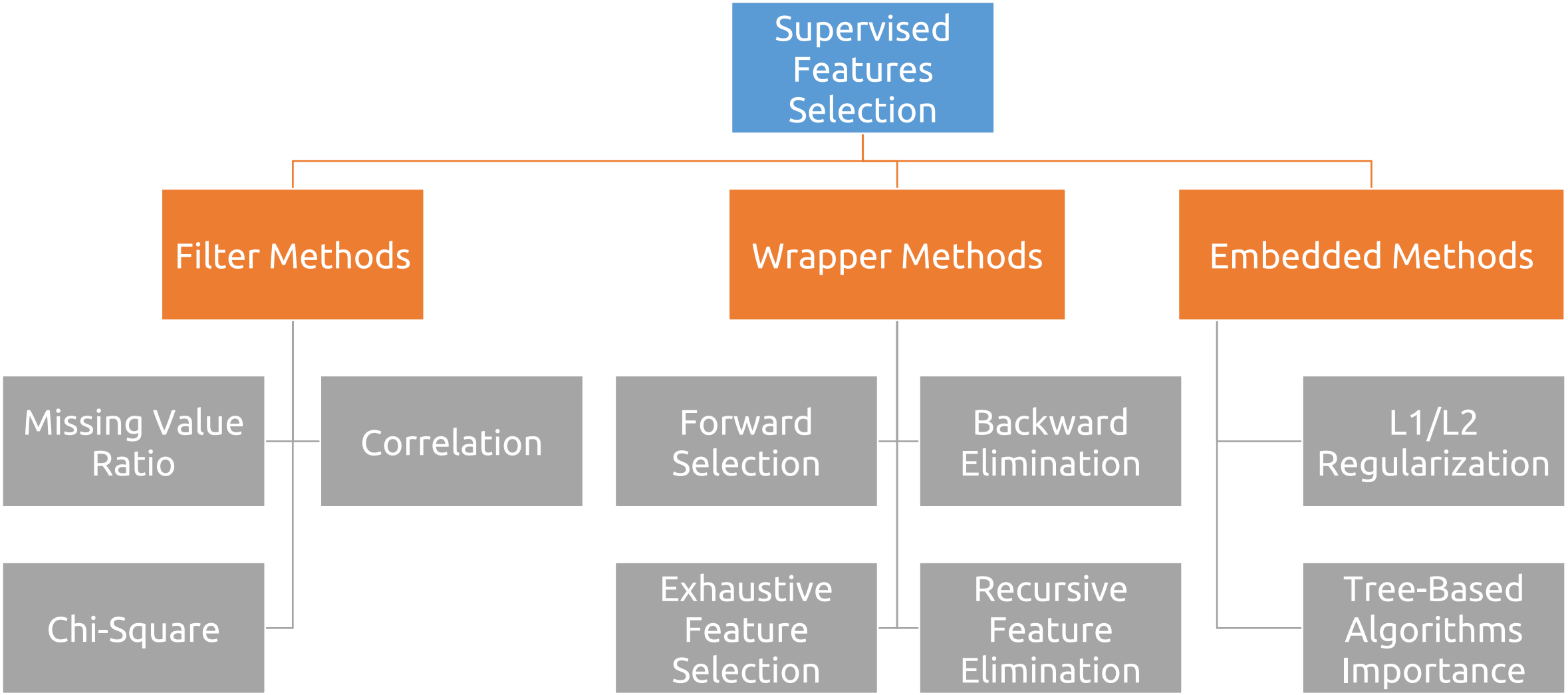


Features Relevance

The relevance of a feature depends on how much it affects the target variable and how much information it provides to the model



Approaches


Supervised Features Selection
Filter Methods
Wrapper Methods
Embedded Methods
Missing Value Ratio
Correlation
Chi-Square
Forward Selection
Backward Elimination
Exhaustive Feature Selection
Recursive Feature Elimination
L1/L2 Regularization
Tree-Based Algorithms Importance

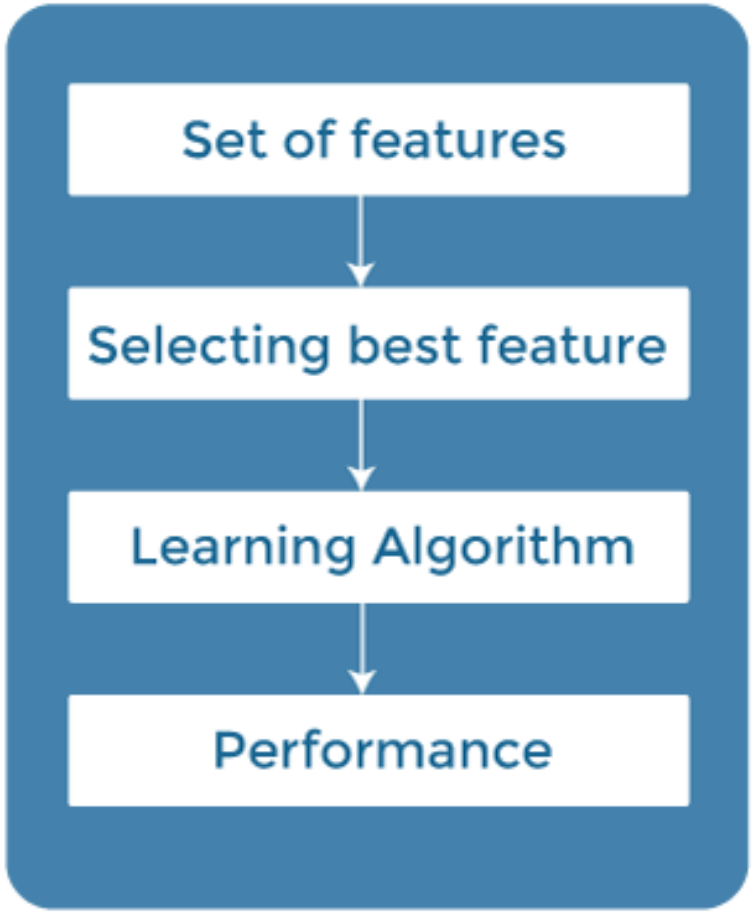
Filter Methods

Filter Methods

In this method, features are dropped based on their **relation** to the target feature.

Advantages:

- low computational time
- does not overfit the data

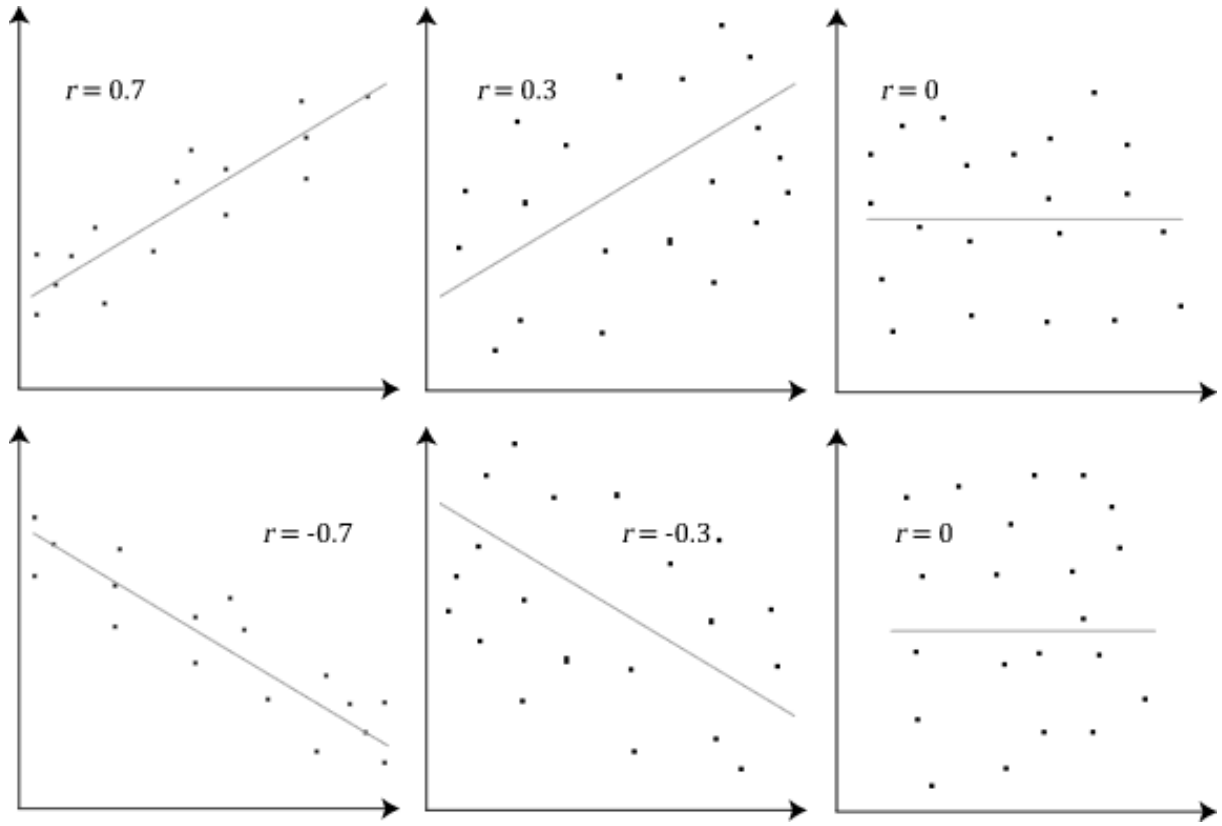


Filter Methods (Examples)

Correlation

It is used to quantify **linear dependence** between two continuous variables, X and Y. Its value ranges from -1 to 1.

Usage?
Input variable is numerical
output variable is numerical



Filter Methods (Examples)

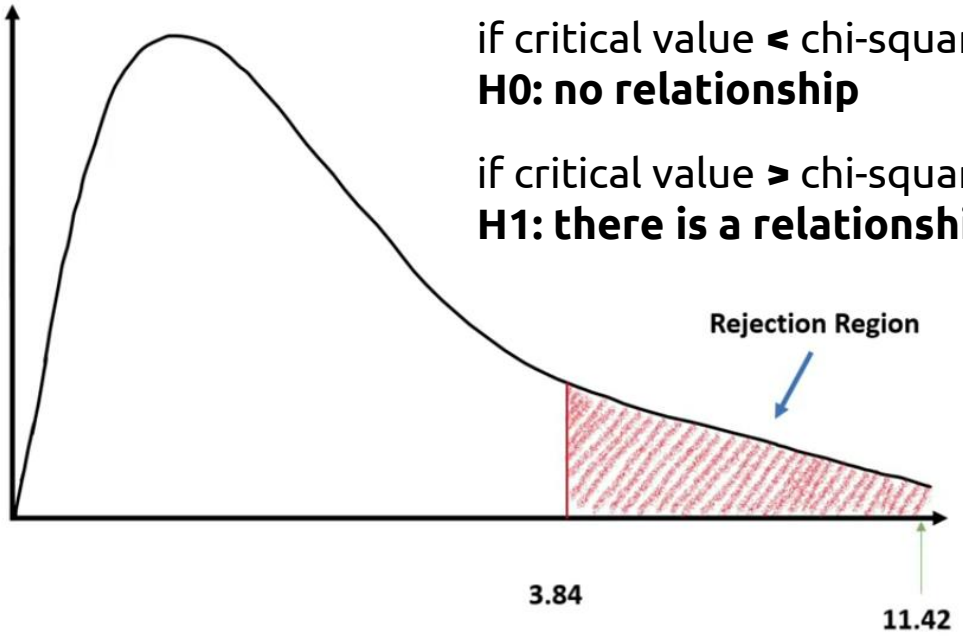
Chi-Square

Chi-square test is a technique to determine the relationship between the **categorical variables**. Chi-Square value is calculated between each feature and the target variable.

Chi-Square = 11.42

Degrees of Freedom = 1

Critical Value = 3.84



if critical value **<** chi-square, then
H0: no relationship

if critical value **>** chi-square, then
H1: there is a relationship

Usage?
Input variable is categorical
output variable is categorical

Tutorial: https://www.youtube.com/watch?v=L6zWgsilOAs

Filter Methods (Examples)

Missing Value Ratio

The value of the missing value ratio can be used for evaluating the feature set against the **threshold** value.

$$Missing\ Value\ Ratio\ = \frac{Number\ of\ Missing\ Values}{Total\ Number\ of\ Obdervations} \times 100$$

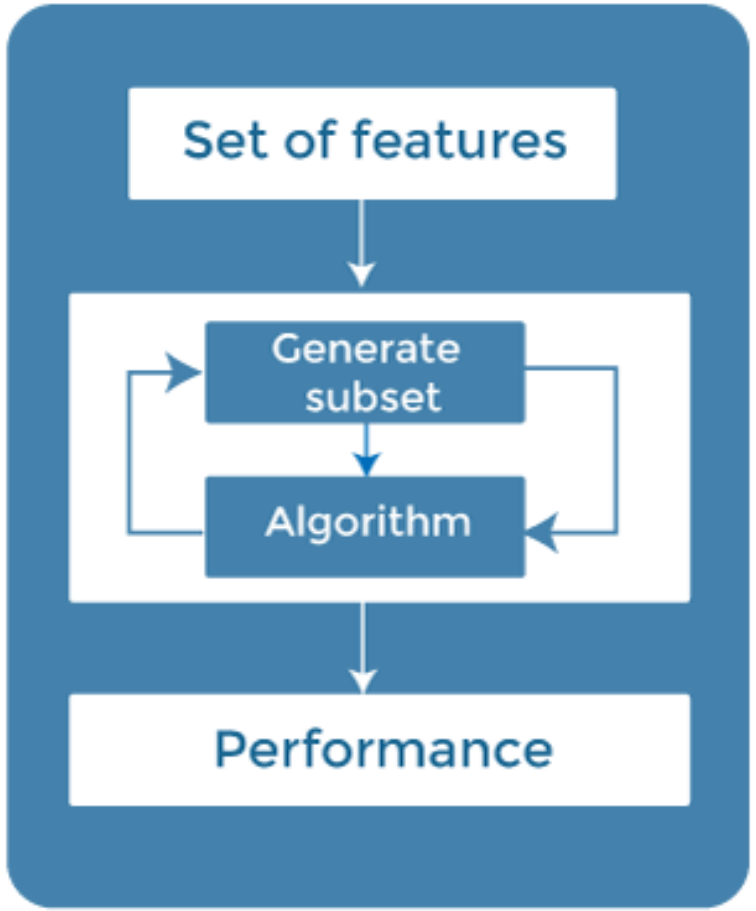
Usage?
Input variable is any (numerical/categorical)
output variable is any (numerical/categorical)

Wrapper Methods

Wrapper Methods

We split our data into **subsets** and train a model using this.

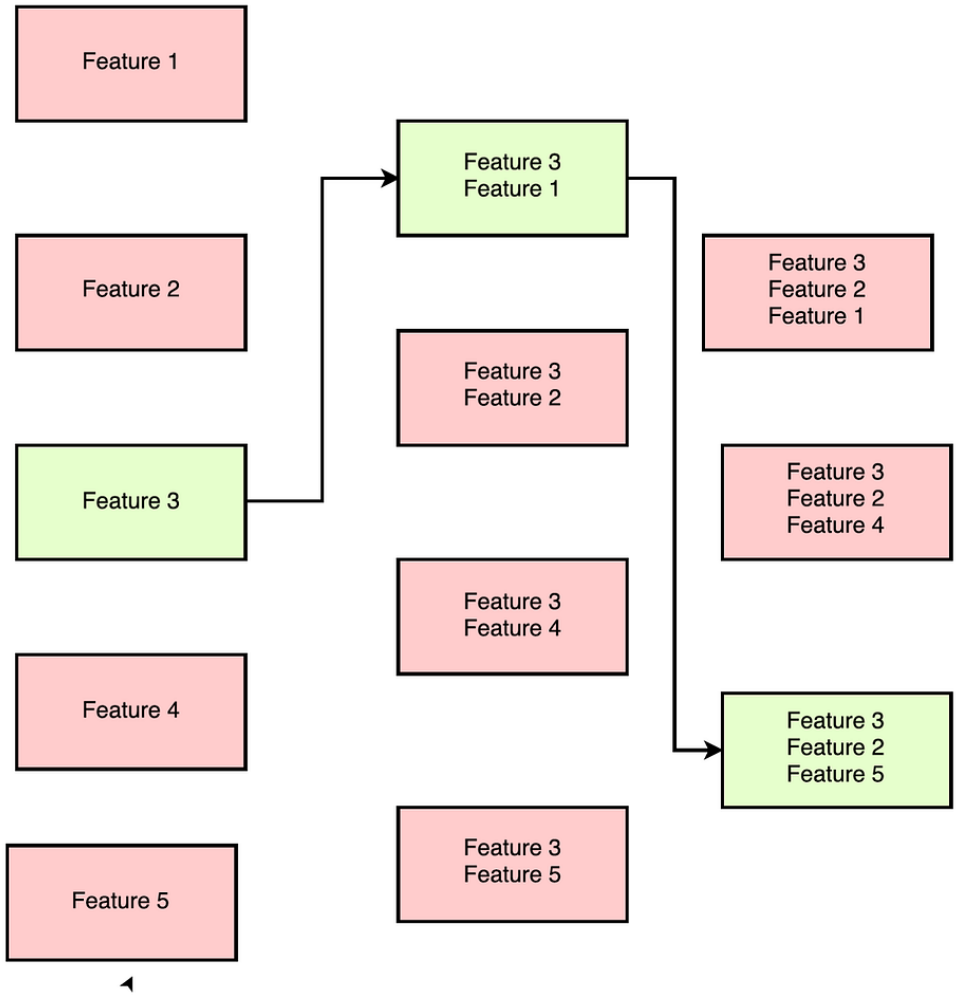
Based on the output of the model, then we **add** and **substract** features and train the model again.



Wrapper Methods (Examples)

Forward selection

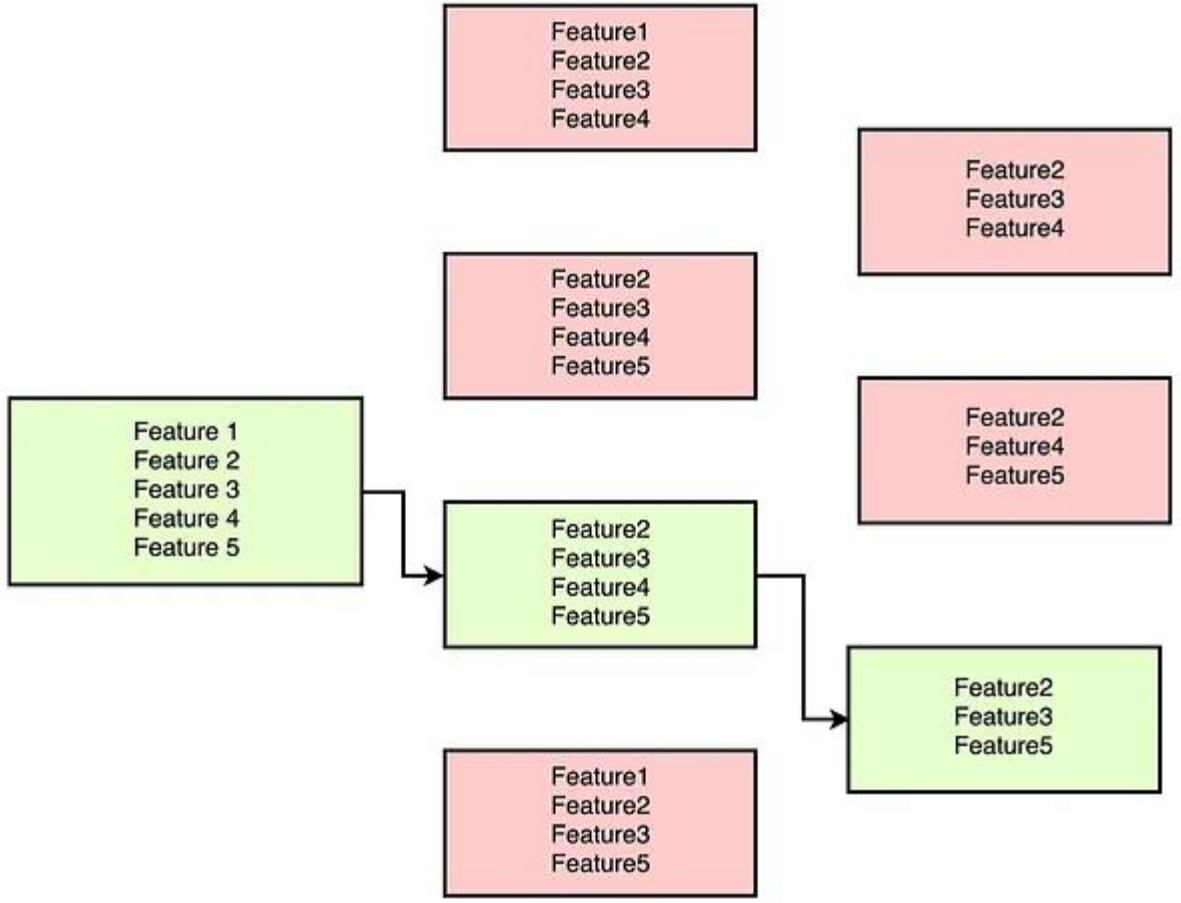
- An iterative process, which begins with an **empty** set of features.
- After each iteration, it keeps adding on a **feature** and **evaluates the performance** to check whether it is improving the performance or not.
- The process continues until the addition of a **new variable/feature does not improve** the performance of the model.



Wrapper Methods (Examples)

Backward elimination

- Also an iterative approach, but it is the opposite of forward selection.
- This technique begins the process by considering all the features and **removes** the least significant feature.
- This process continues until removing the features does not improve the performance of the model.

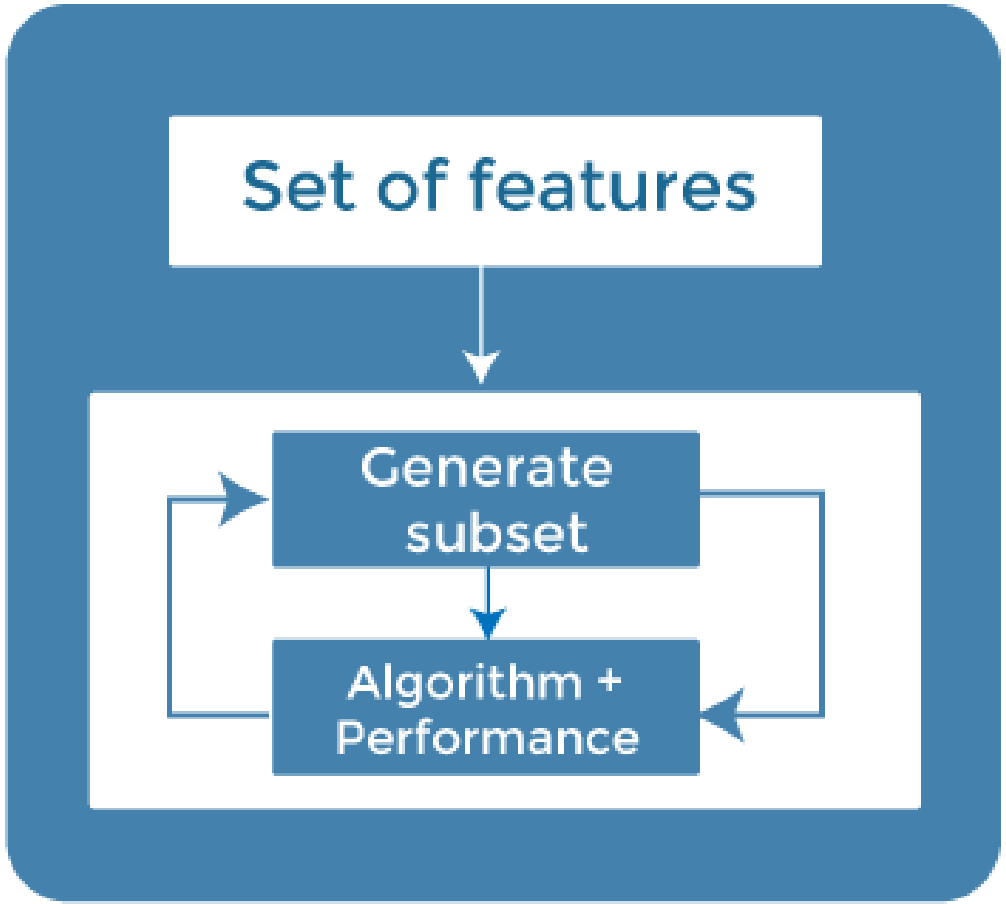


Embedded Methods

Embedded Methods

This method combines the qualities of both **filter** and **wrapper methods** to create the best subset.

Model will train and check the accuracy of **different subsets** and select the best among them.

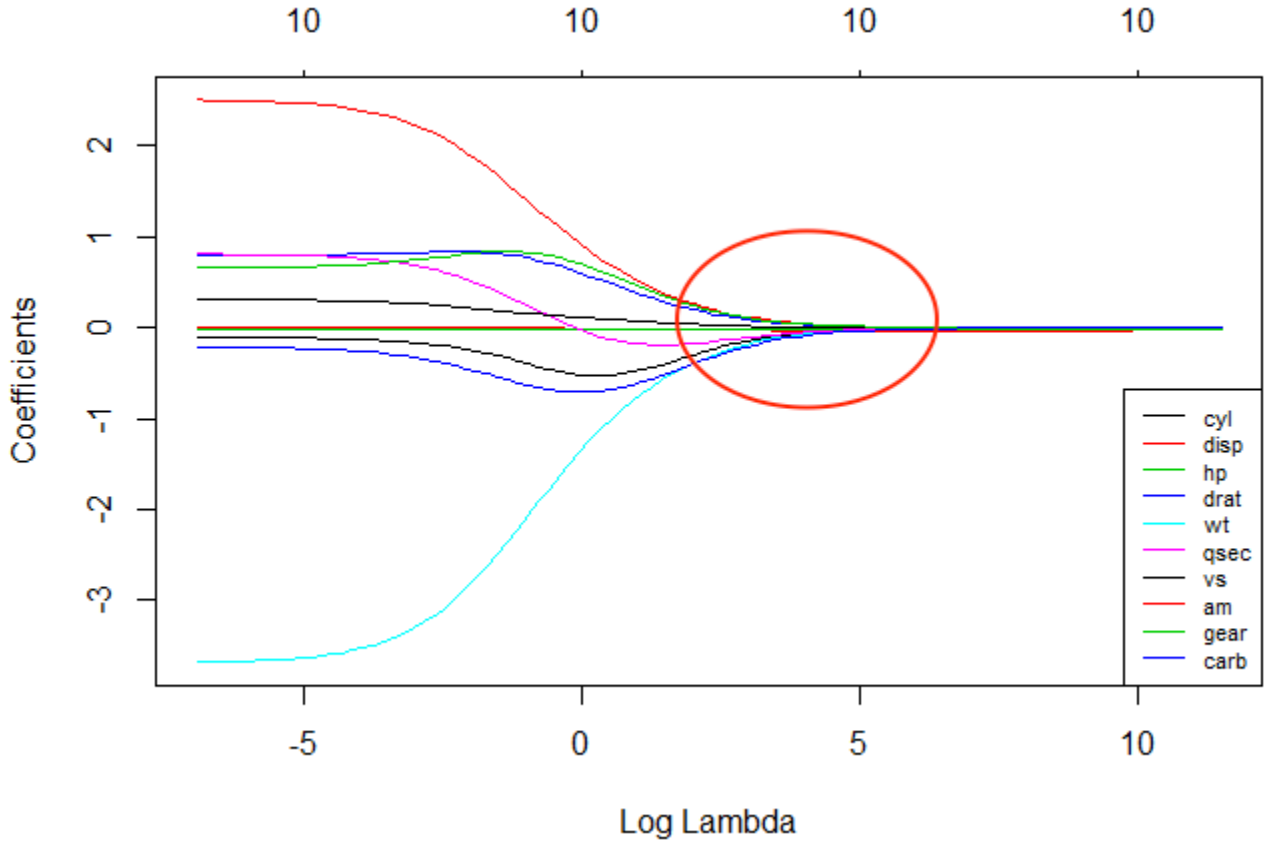


Embedded Methods (Examples)

L1/L2 Regularization

- adds a **penalty term** to different parameters of the machine learning model for **avoiding overfitting**.
- This penalty term is added to the **coefficients**; hence it shrinks some coefficients to zero.
- Those features with **zero coefficients** can be removed from the dataset.

Usage?
Input variable is numerical
output variable is numerical



Embedded Methods (Examples)

Tree-Based Algorithms Importance

- Different tree-based methods of feature selection help us with **feature importance** to provide a way of selecting features.
- **Random Forest** is such a tree-based method, which is a type of **bagging algorithm** that aggregates a different number of **decision trees**.
- It automatically ranks the nodes by **their performance** or **decrease** in the impurity **(Gini impurity)** over all the trees.
- Nodes are arranged as per the **impurity values**, and thus it allows to **pruning** of trees below a specific node. The remaining nodes create a subset of the most **important features**.

References

- https://oercollective.caul.edu.au/foundations-of-biomedical-science/chapter/9-1-types-of-data/
- https://www.featureform.com/post/feature-engineering-guide
- https://h2o.ai/wiki/feature-selection/
- https://www.javatpoint.com/feature-selection-techniques-in-machine-learning
- https://medium.com/@sagar.rawale3/feature-selection-methods-in-machine-learning-eaeef12019cc