Feature Selection

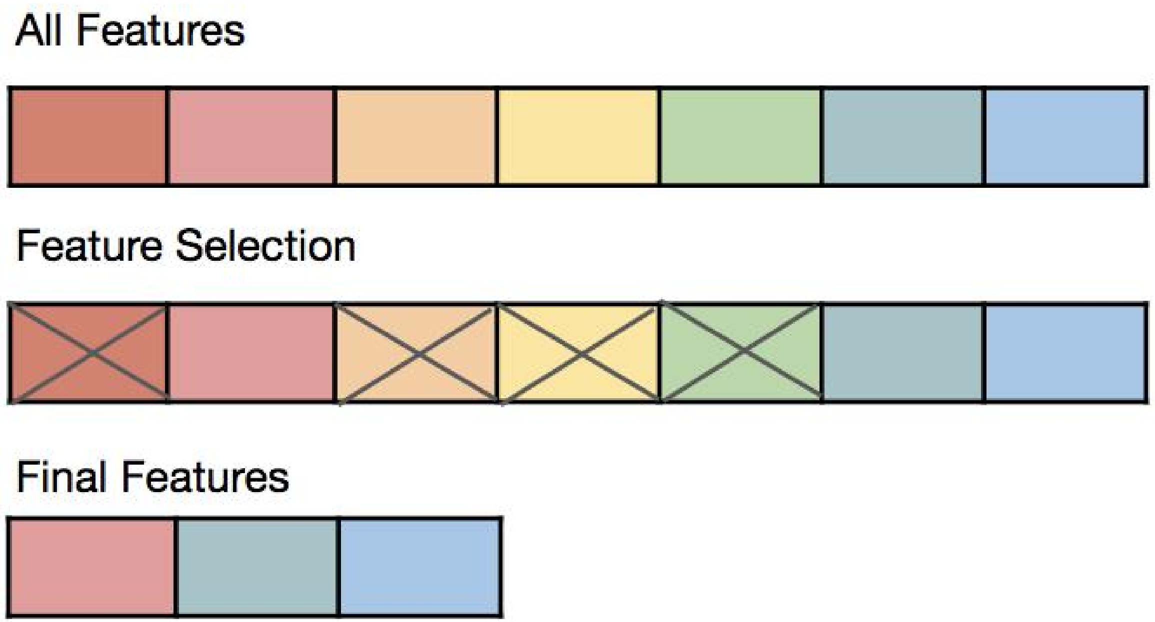
Choosing the Right Variables for Your Analysis

Prepared By: Ashraf Abdulkhaliq

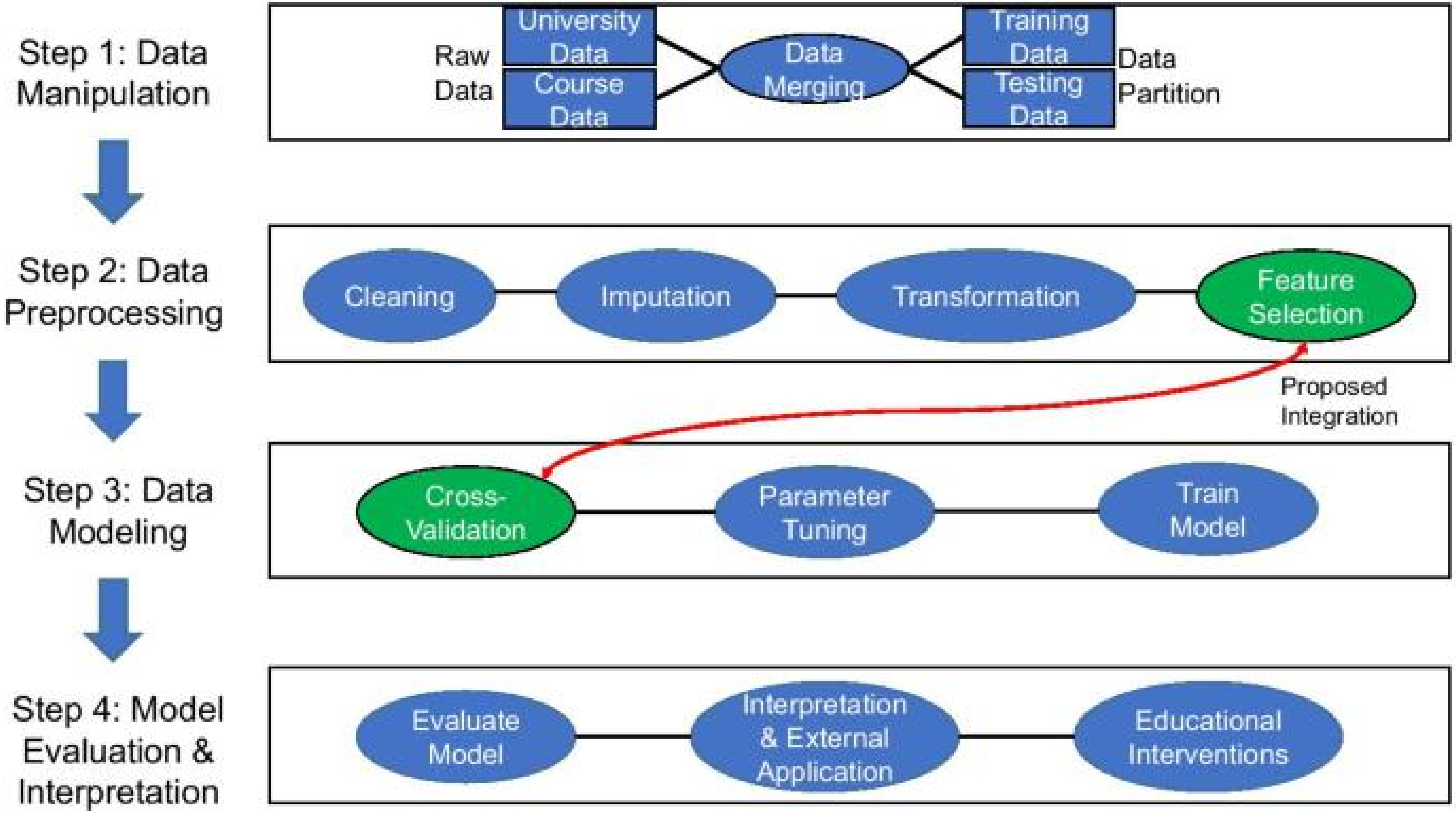
Introduction

- **Feature selection** process is one of the main components of a **feature engineering** process.

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



Introduction



Agenda

- Feature Selection Need
- Types of Features
- Feature Selection Approaches
- Filter Methods
- Wrapper Methods
- Embedded Methods
- Feature Selection Appropriate Techniques
- References

Agenda

- Feature Selection Need
- Types of Features
- Feature Selection Approaches
- Filter Methods
- Wrapper Methods
- Embedded Methods
- Feature Selection Appropriate Techniques
- References

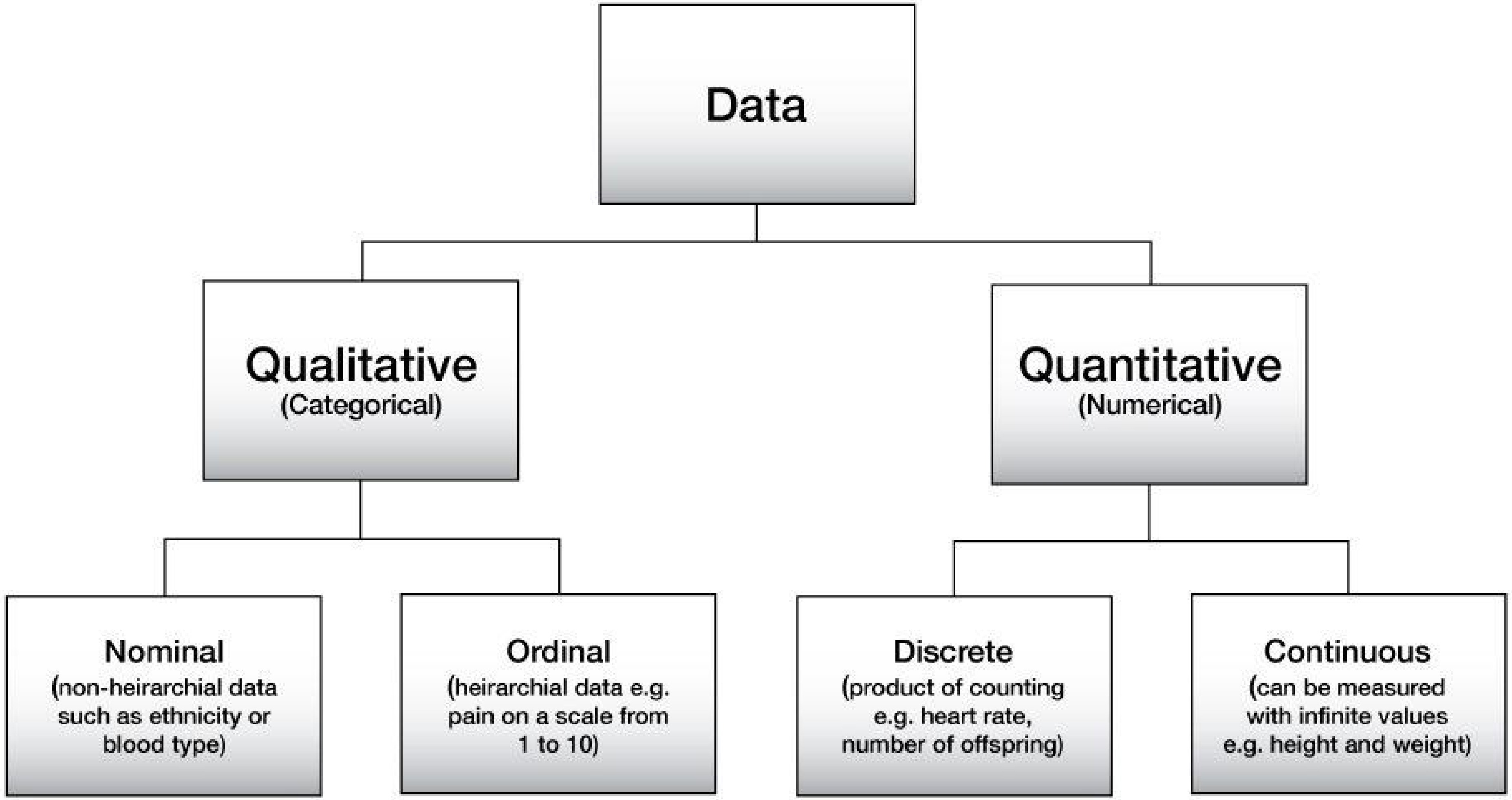
Feature Selection Need

1. It helps in avoiding the curse of **dimensionality**.
2. It helps in the **simplification** of the model so that it can be easily interpreted.
3. It reduces the **training time**.
4. It reduces **overfitting** hence enhance the **generalization**.

Agenda

- Feature Selection Need
- Types of Features
- Feature Selection Approaches
- Filter Methods
- Wrapper Methods
- Embedded Methods
- Feature Selection Appropriate Techniques
- References

Types of Features (DATA)

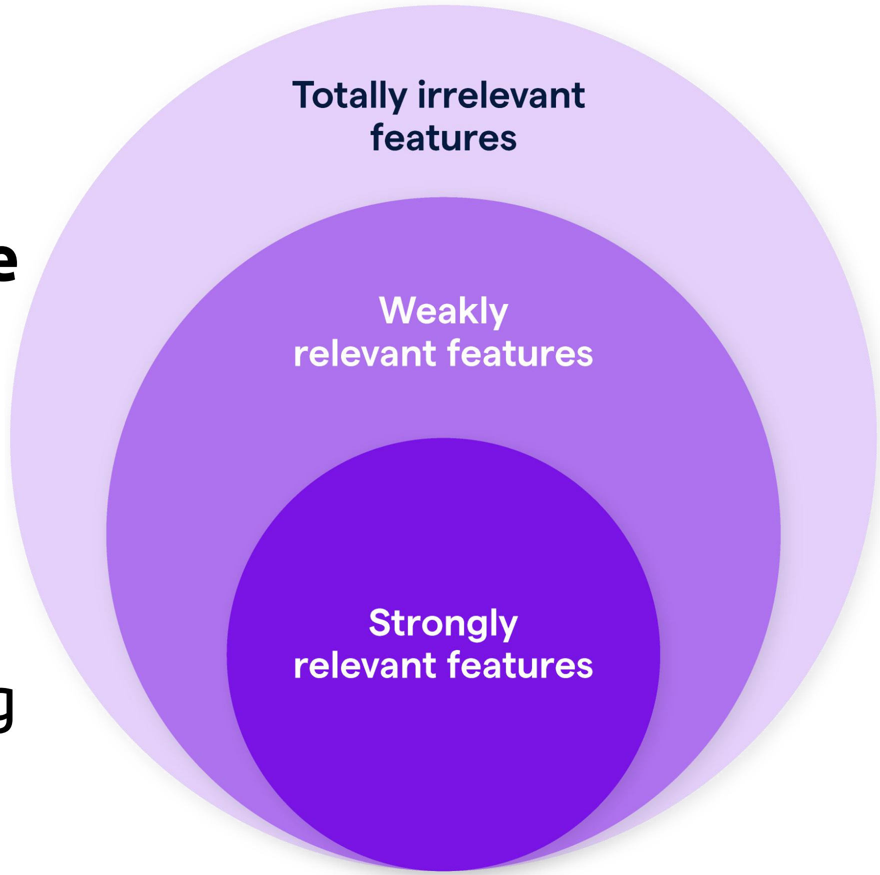


Types of Features (DATA)

Totally Irrelevant Features: These features **lack** any connection or **impact** on the target variable, offering **no valuable information** to the model.

Weakly Relevant Features: While they have **some association** with the target variable, it's provide only minor contributions to the model's performance and contain **limited information**.

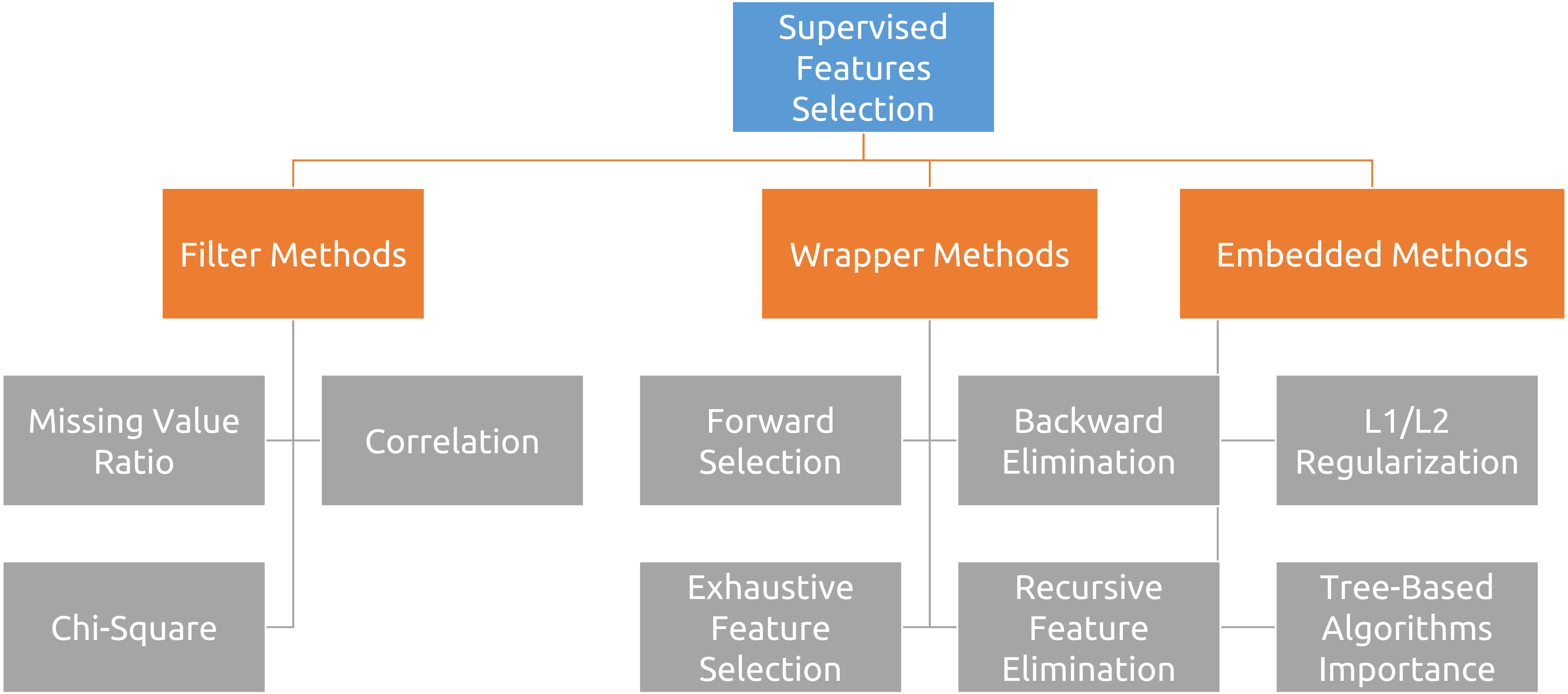
Strongly Relevant Features: These features are **closely linked** to the target variable, significantly **enhancing** the model's performance by containing crucial information necessary for accurate predictions.



Agenda

- Feature Selection Need
- Types of Features
- Feature Selection Approaches
- Filter Methods
- Wrapper Methods
- Embedded Methods
- Feature Selection Appropriate Techniques
- References

Feature Selection Approaches



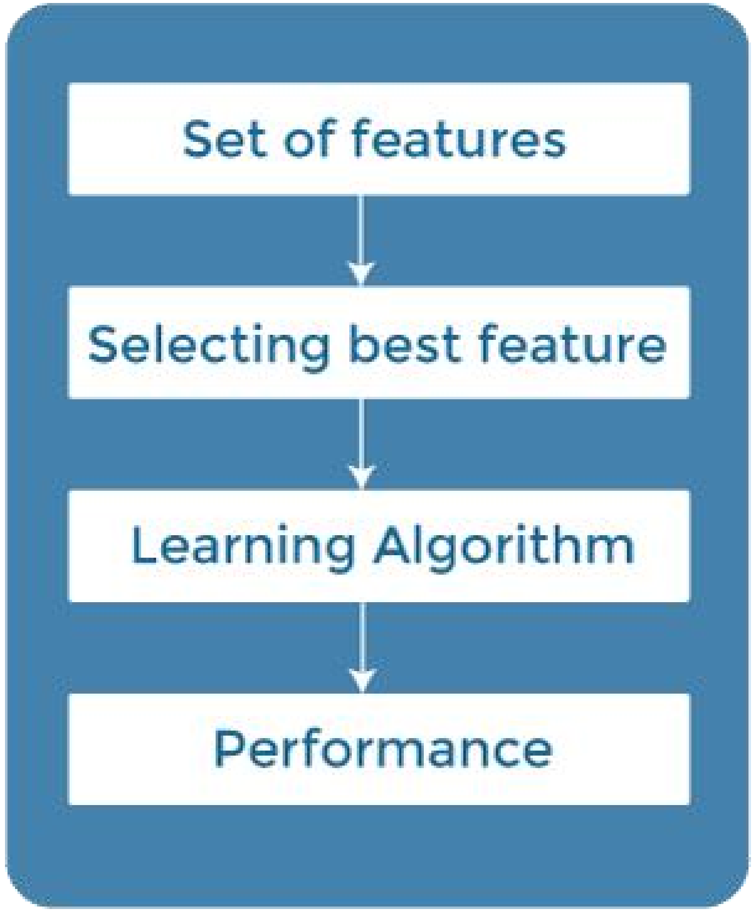
Agenda

- Feature Selection Need
- Types of Features
- Feature Selection Approaches
- Filter Methods
- Wrapper Methods
- Embedded Methods
- Feature Selection Appropriate Techniques
- References

Filter Methods

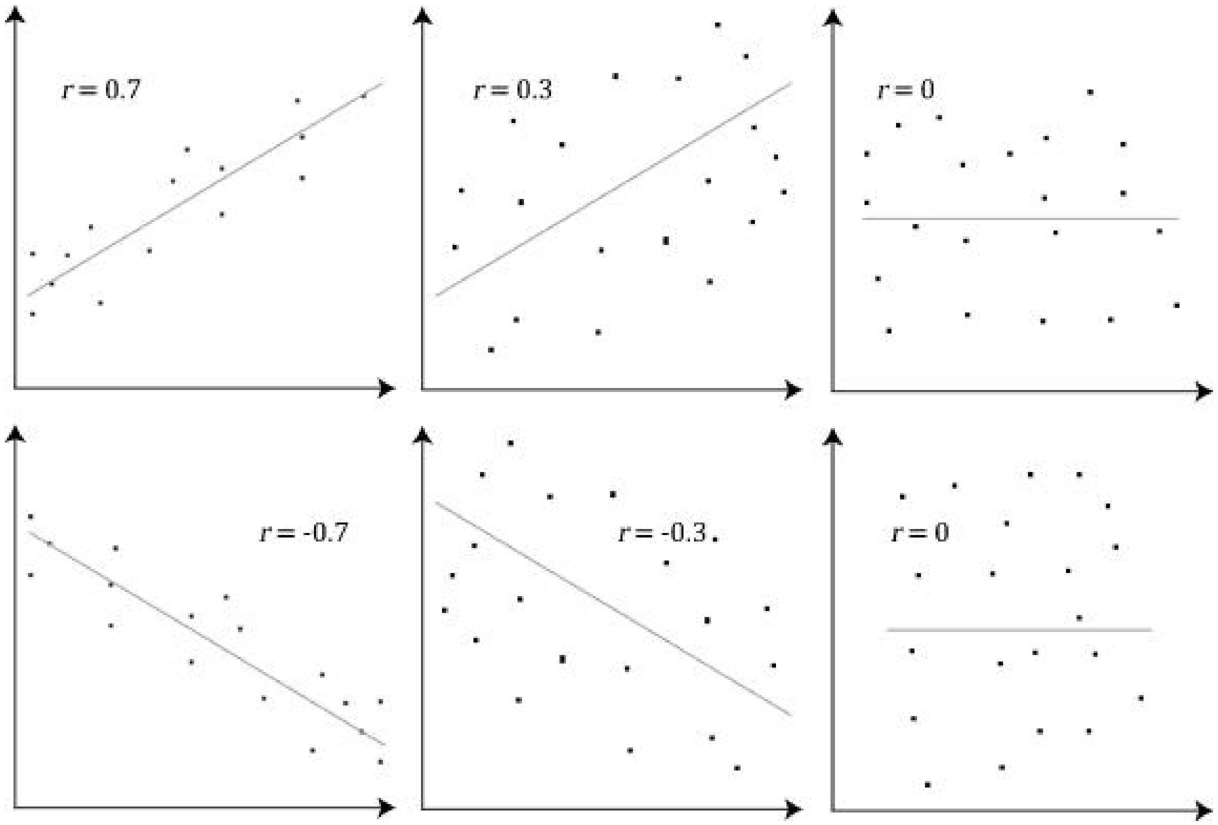
In this method, features are dropped based on their **relation** to the target feature, or how they **correlating** to the target feature.

Advantage of using filter methods is that it needs **low computational time** and **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**.



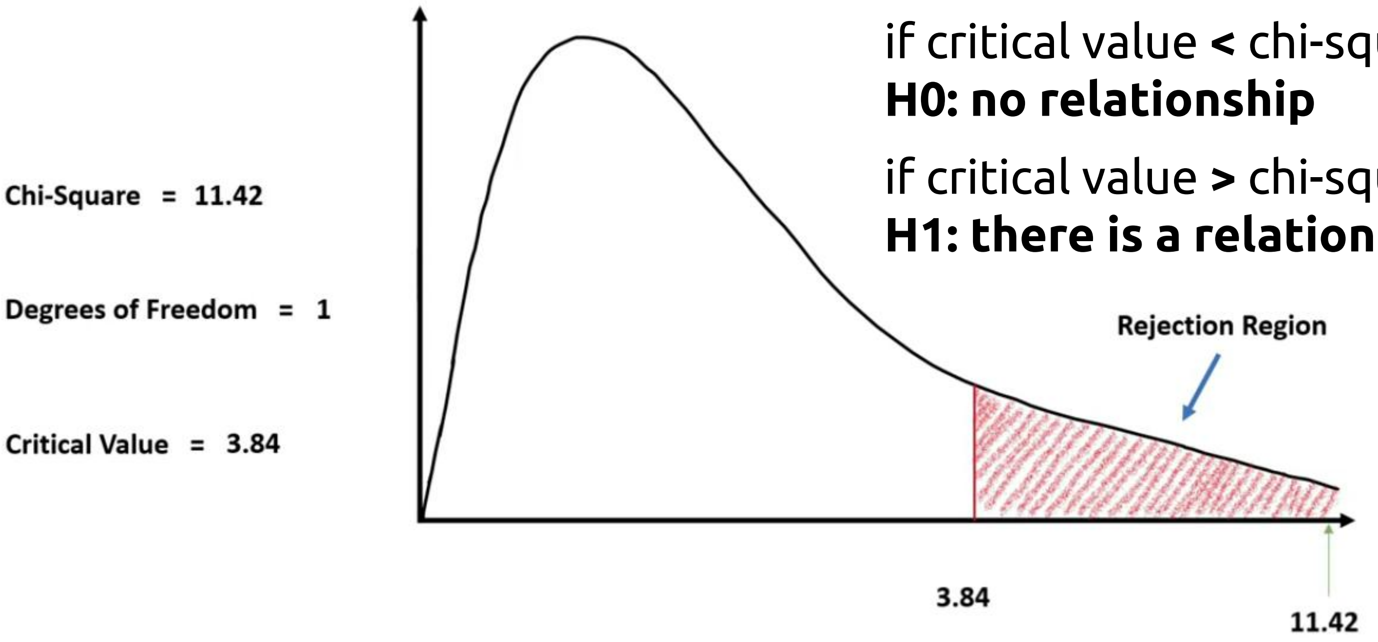
Usage?

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.



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)

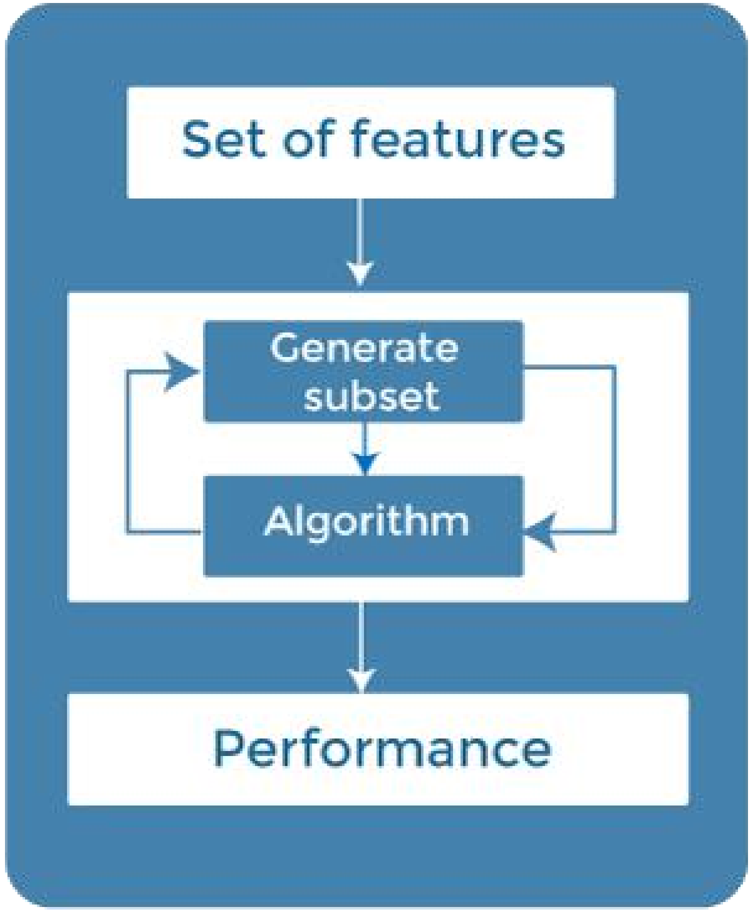
output variable is any (numerical/categorical)

Agenda

- Feature Selection Need
- Types of Features
- Feature Selection Approaches
- Filter Methods
- Wrapper Methods
- Embedded Methods
- Feature Selection Appropriate Techniques
- References

Wrapper Methods

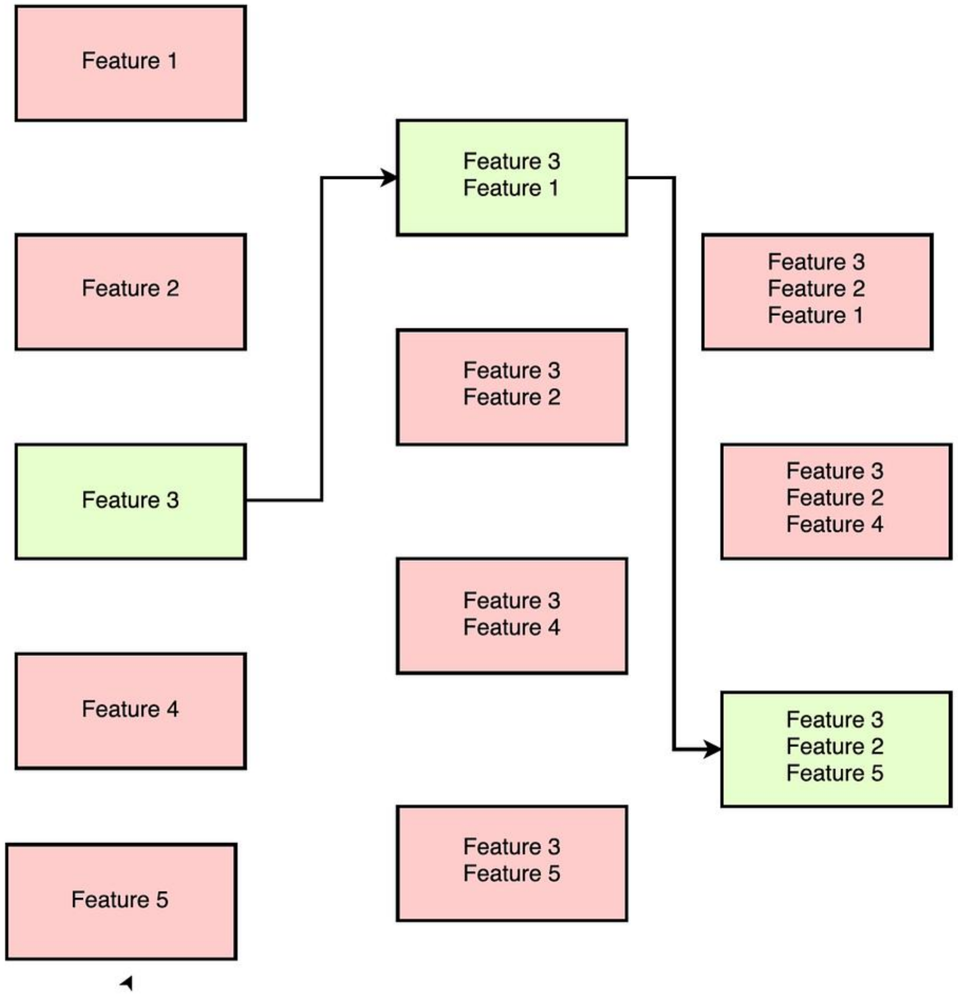
We split our data into **subsets** and train a model using this. Based on the output of the model, then we **add** and **substract** features and train the model again



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.



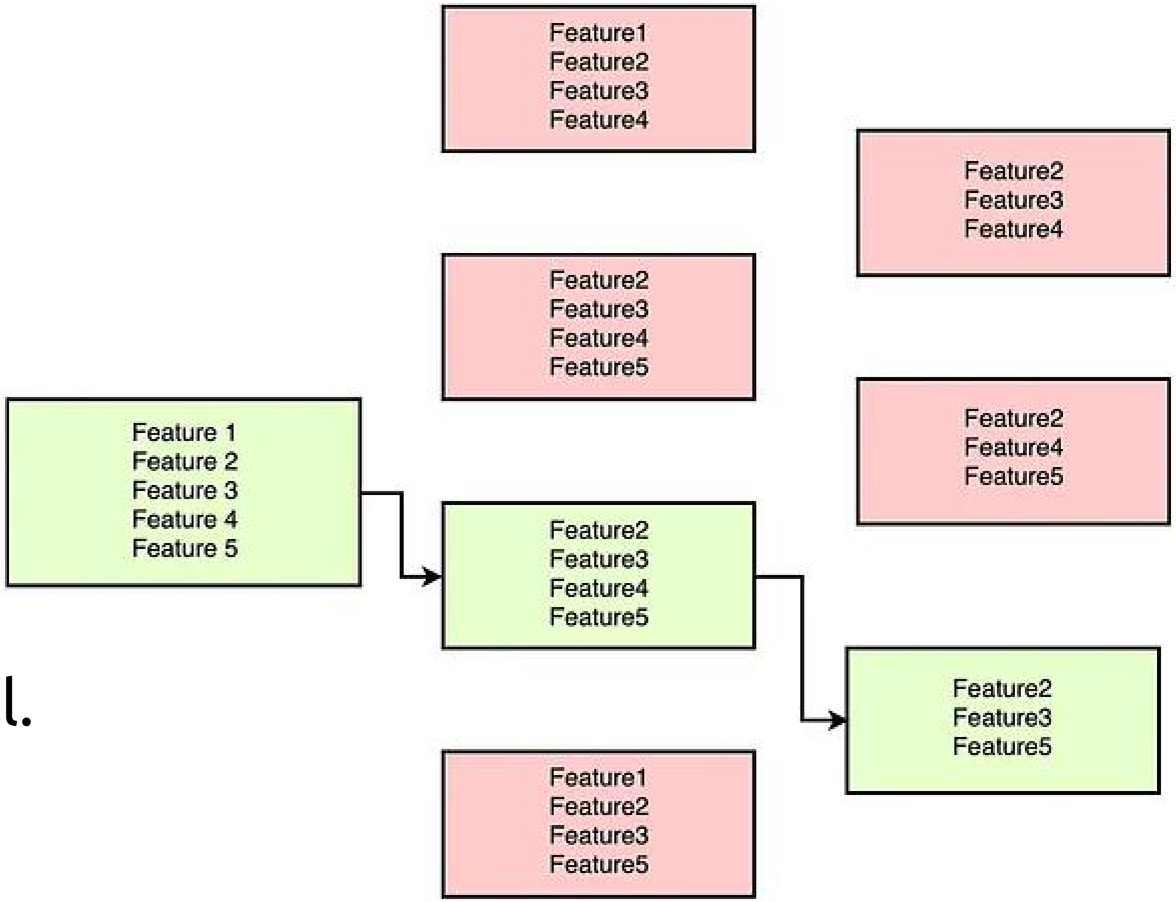
Wrapper Methods (Examples)

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 elimination process continues until removing the **features does not improve the performance** of the model.



Wrapper Methods (Examples)

Exhaustive Feature Selection: One of the best feature selection methods, which evaluates each **feature set as brute-force**. It means this method tries & make each **possible combination** of features and return the best performing feature set.

Recursive Feature Elimination: A recursive greedy optimization approach, where features are selected by **recursively taking a smaller and smaller subset of features**.

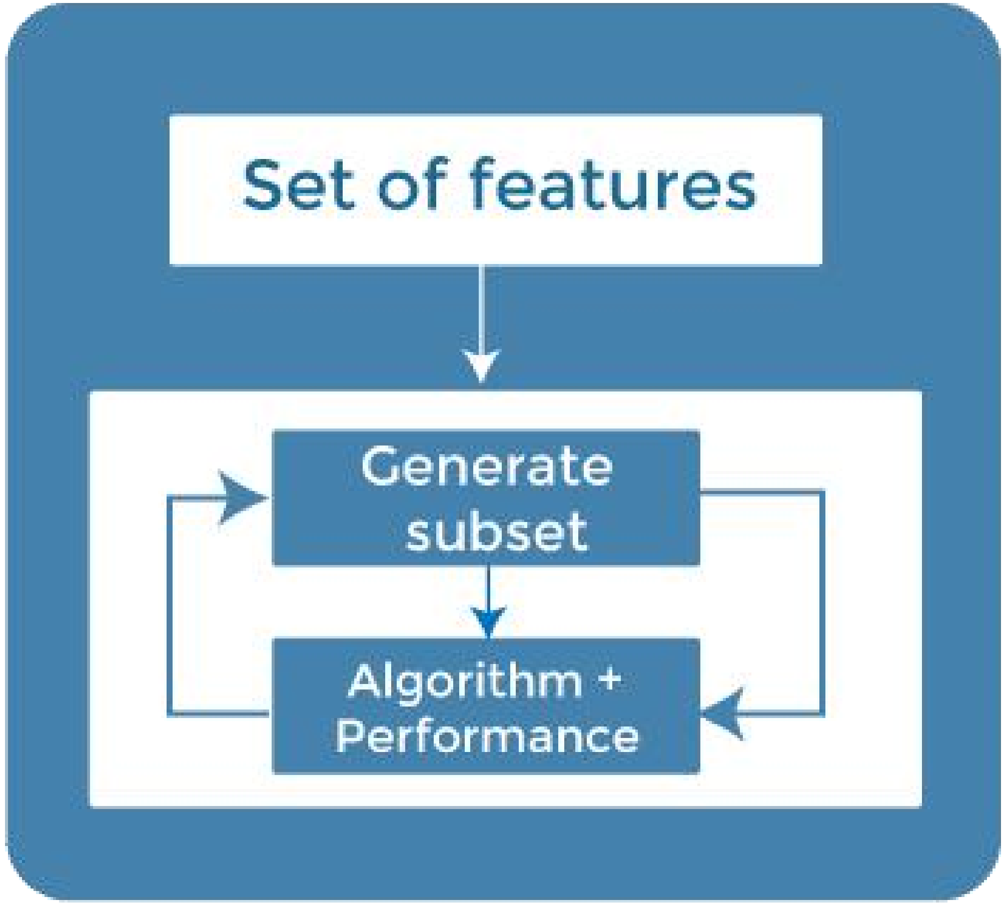
Agenda

- Feature Selection Need
- Types of Features
- Feature Selection Approaches
- Filter Methods
- Wrapper Methods
- Embedded Methods
- Feature Selection Appropriate Techniques
- References

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.

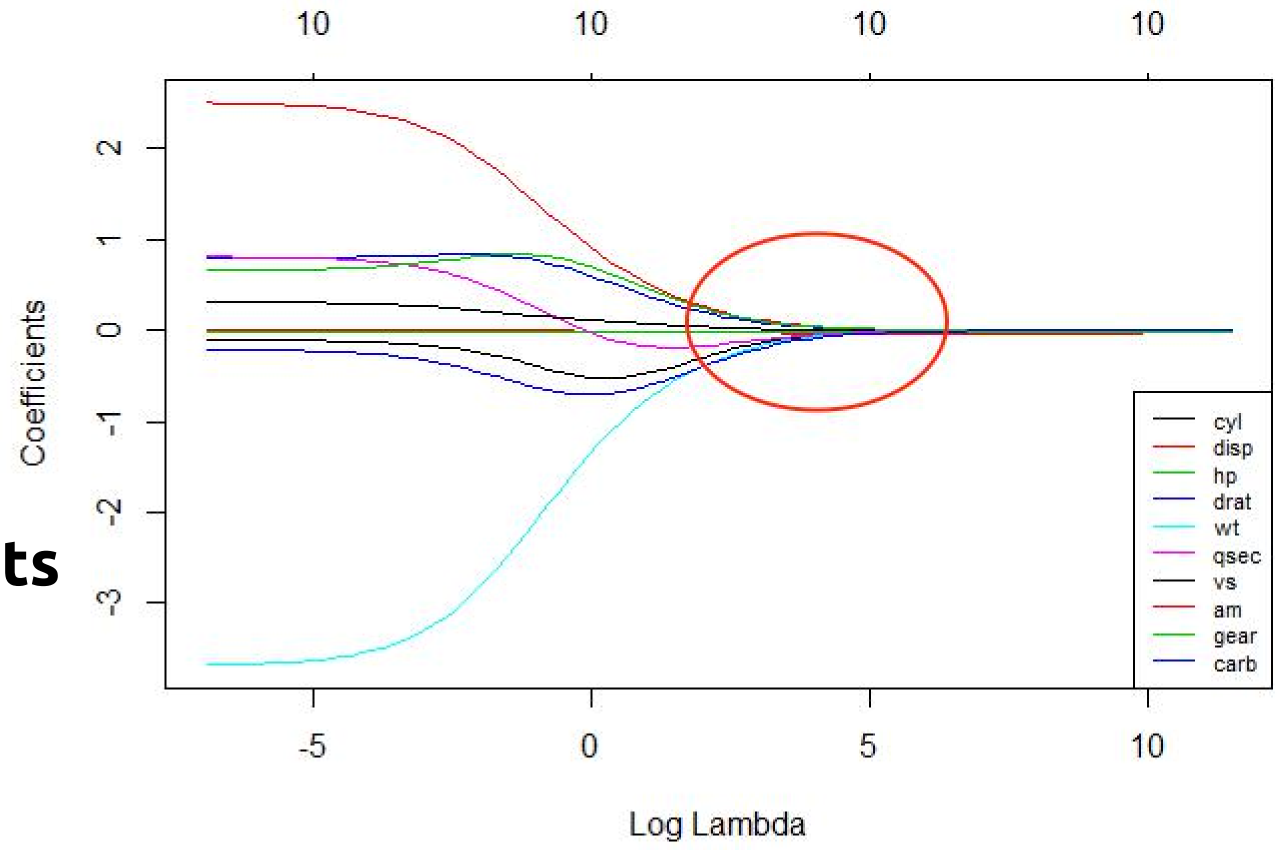


Embedded Methods (Examples)

Usage?
Input variable is numerical
output variable is numerical

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.



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.
- Feature importance specifies which feature has more **importance** in model building or has a **great impact** on the target variable.
- **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**.

Agenda

- Feature Selection Need
- Types of Features
- Feature Selection Approaches
- Filter Methods
- Wrapper Methods
- Embedded Methods
- Feature Selection Appropriate Techniques
- References

Feature Selection Appropriate Techniques

Input Data Type	Output Data Type	Feature Selection Technique
Numerical	Numerical	Correlation
Numerical	Categorical	Chi-Square Test
Numerical	Any	Missing Value Ratio
Categorical	Numerical	Forward Selection
Categorical	Numerical	Backward Elimination
Categorical	Numerical	Exhaustive Feature Selection
Any	Any	Recursive Feature Elimination
Numerical	Any	L1/L2 Regularization (Lasso/Ridge)
Any	Any	Tree-Based Algorithms Importances

Agenda

- Feature Selection Need
- Types of Features
- Feature Selection Approaches
- Filter Methods
- Wrapper Methods
- Embedded Methods
- Feature Selection Appropriate Techniques
- References

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