Feature Selection with Filter Methods

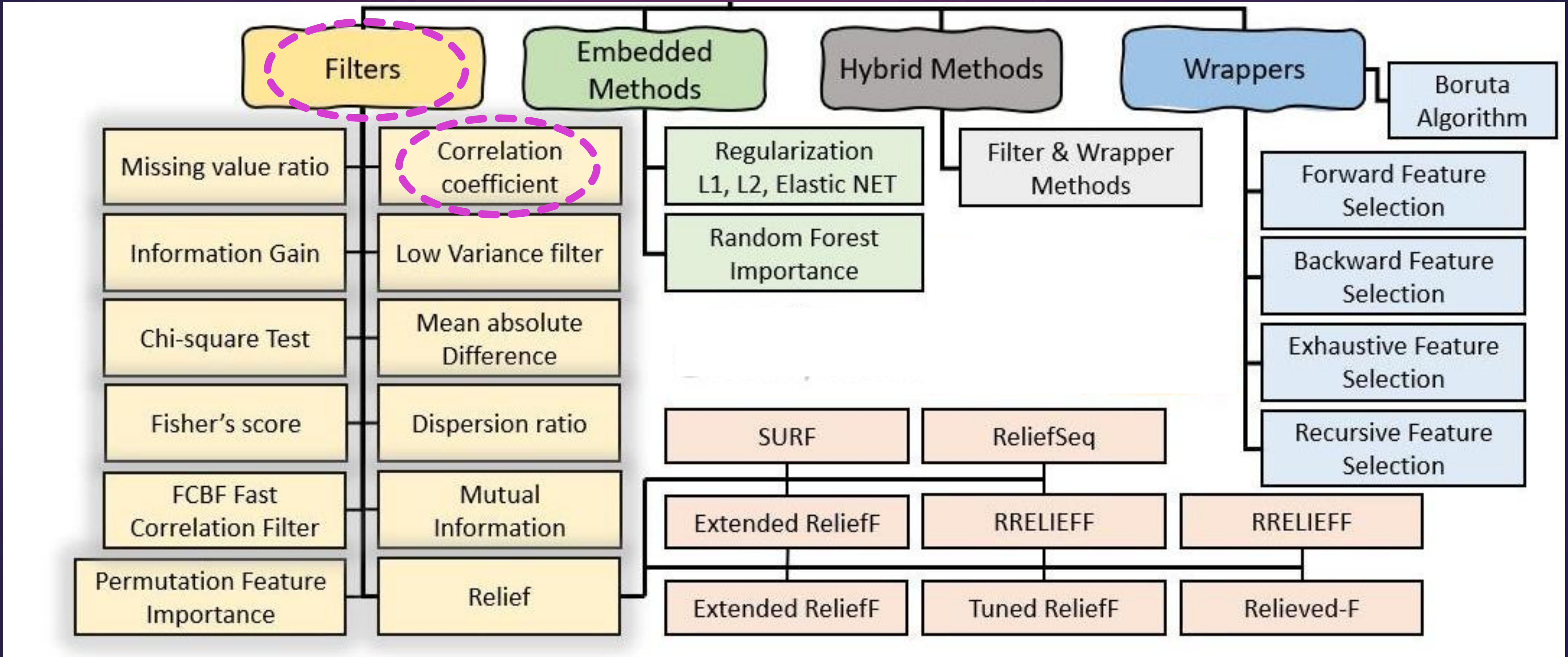
Nov 08 2025

Giảng viên: TS. Lưu Phúc Lợi
Email: Luu.p.loi@googlemail.com
Zalo: 0901802182

Content

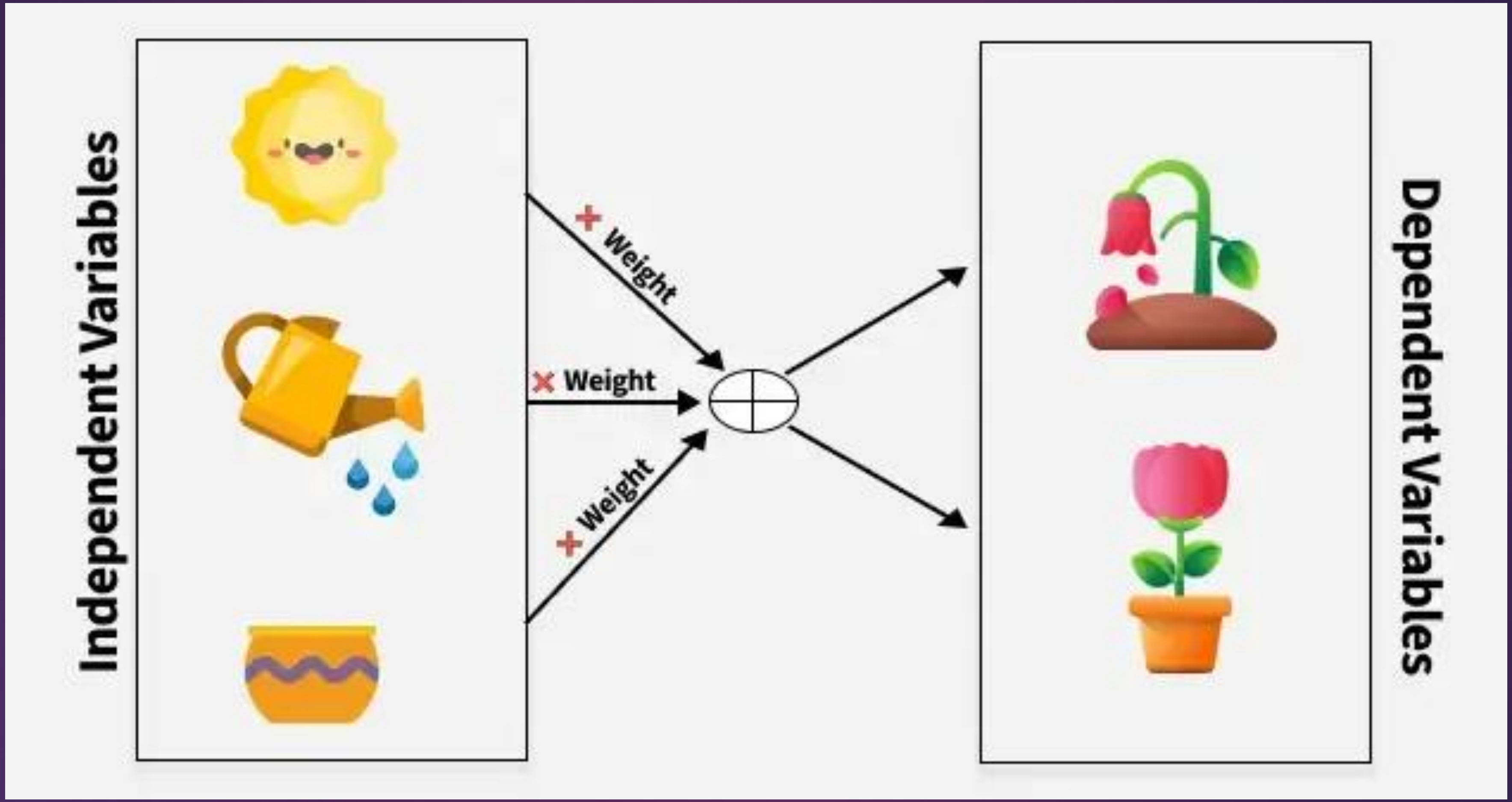
1. Recap of Covariance and Correlation
2. Coefficient correlation filter methods: Principle and Process
3. Advantage and limitation
4. Practice

Supervised Feature Selection: Filter Methods



https://medium.com/analytics-vidhya/feature-selection-extended-overview-b58f1d524c1c

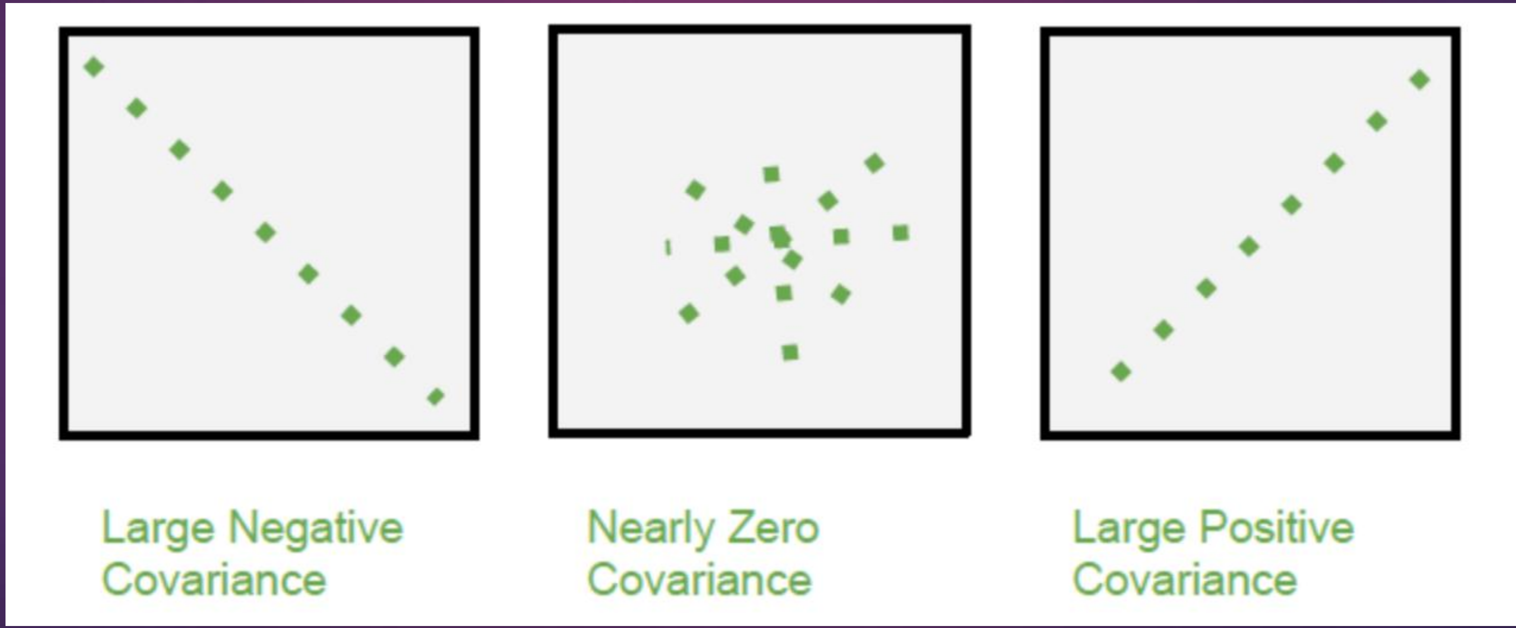
Relationship between Independent and dependent variables



https://www.geeksforgeeks.org/engineering-mathematics/mathematics-covariance-and-correlation/

Covariance

- It can take any value between - infinity to +infinity, where the negative value represents the negative relationship whereas a positive value represents the positive relationship.
- It is used for the linear relationship between variables.
- It gives the direction of relationship between variables.



Sample Covariance

$$\text{Cov}_S(X, Y) = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})$$

Where:

- X_i: The i^{th} value of the variable X in the sample.
- Y_i: The i^{th} value of the variable Y in the sample.
- $\overline{X}$: The sample mean of variable X (i.e., the average of all X_i values in the sample).
- $\overline{Y}$: The sample mean of variable Y (i.e., the average of all Y_i values in the sample).
- n: The number of data points in the sample.
- $\sum$: The summation symbol means we sum the products of the deviations for all the data points.
- $n - 1$: This is the degrees of freedom. When working with a sample, we divide by $n - 1$ to correct for the bias introduced by estimating the population covariance based on the sample data. This is known as Bessel's correction.

Population Covariance

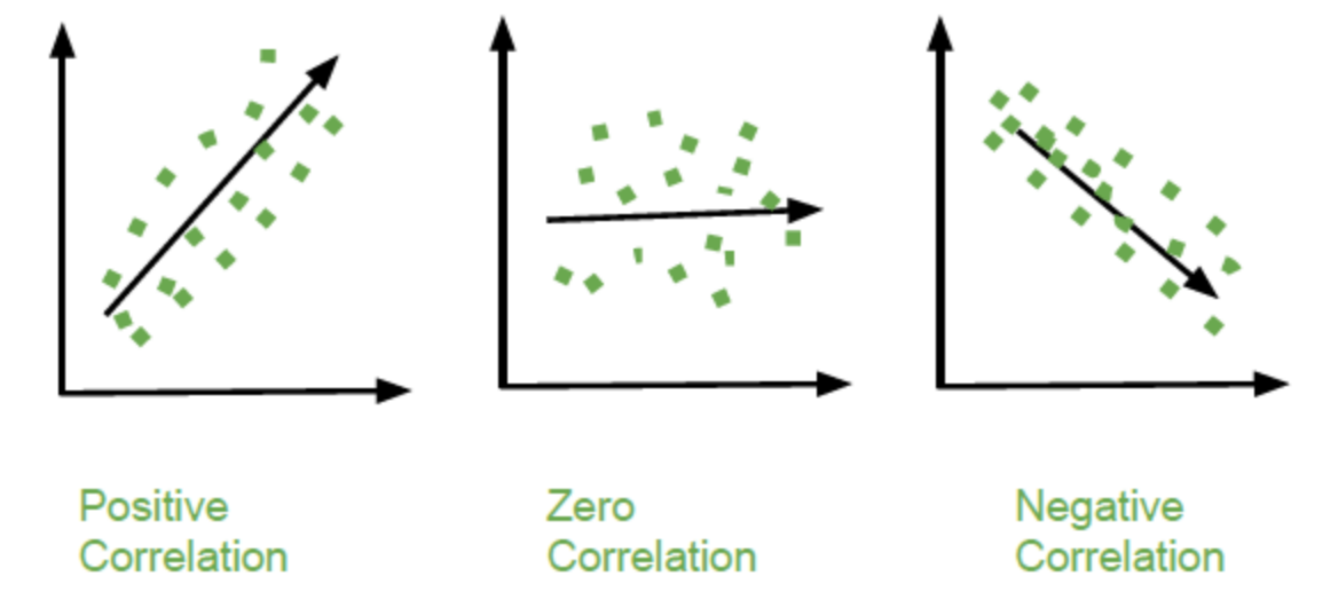
$$\text{Cov}_P(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_X)(Y_i - \mu_Y)$$

Where:

- X_i: The i^{th} value of the variable X in the population.
- Y_i: The i^{th} value of the variable Y in the population.
- μ_X: The population mean of variable X (i.e., the average of all X_i values in the population).
- μ_Y: The population mean of variable Y (i.e., the average of all Y_i values in the population).
- n: The total number of data points in the population.
- $\sum$: The summation symbol means we sum the products of the deviations for all the data points.
- n: In the case of population covariance, we divide by n because we are using the entire population data. There's no need for Bessel's correction since we're not estimating anything.

Correlation

- Correlation is a standardized measure of the strength and direction of the linear relationship between two variables. It is derived from covariance and ranges between -1 and 1. Unlike covariance, which only indicates the direction of the relationship, correlation provides a standardized measure.
- Positive Correlation (close to +1): As one variable increases, the other variable also tends to increase.
- Negative Correlation (close to -1): As one variable increases, the other variable tends to decrease.
- Zero Correlation: There is no linear relationship between the variables.



Correlation Coefficient

The correlation coefficient ρ (rho) for variables X and Y is defined as:

1. Correlation takes values between -1 to +1, wherein values close to +1 represents strong positive correlation and values close to -1 represents strong negative correlation.
2. In this variable are indirectly related to each other.
3. It gives the direction and strength of relationship between variables.

Correlation Formula

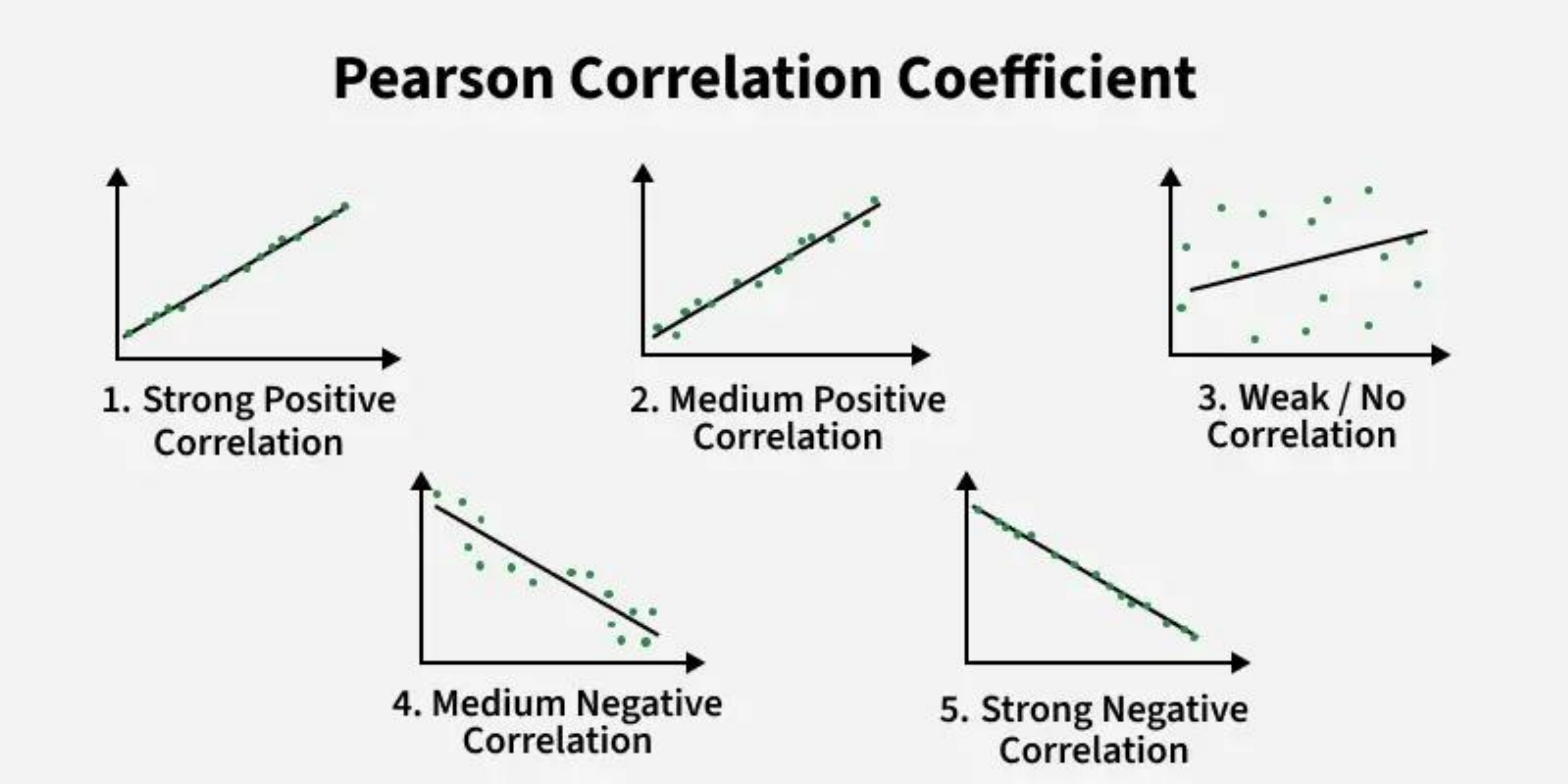
$$Corr(x,y) = \frac{\sum_{i=1}^{n}(x_i - x')\,(y_i - y')}{\sqrt{\sum_{i=1}^{n}(x_i - x')^2 \sum_{i=1}^{n}(y_i - y')^2}}$$

Here,

- x' and y' = mean of given sample set
- n = total no of sample
- x_i and y_i = individual sample of set

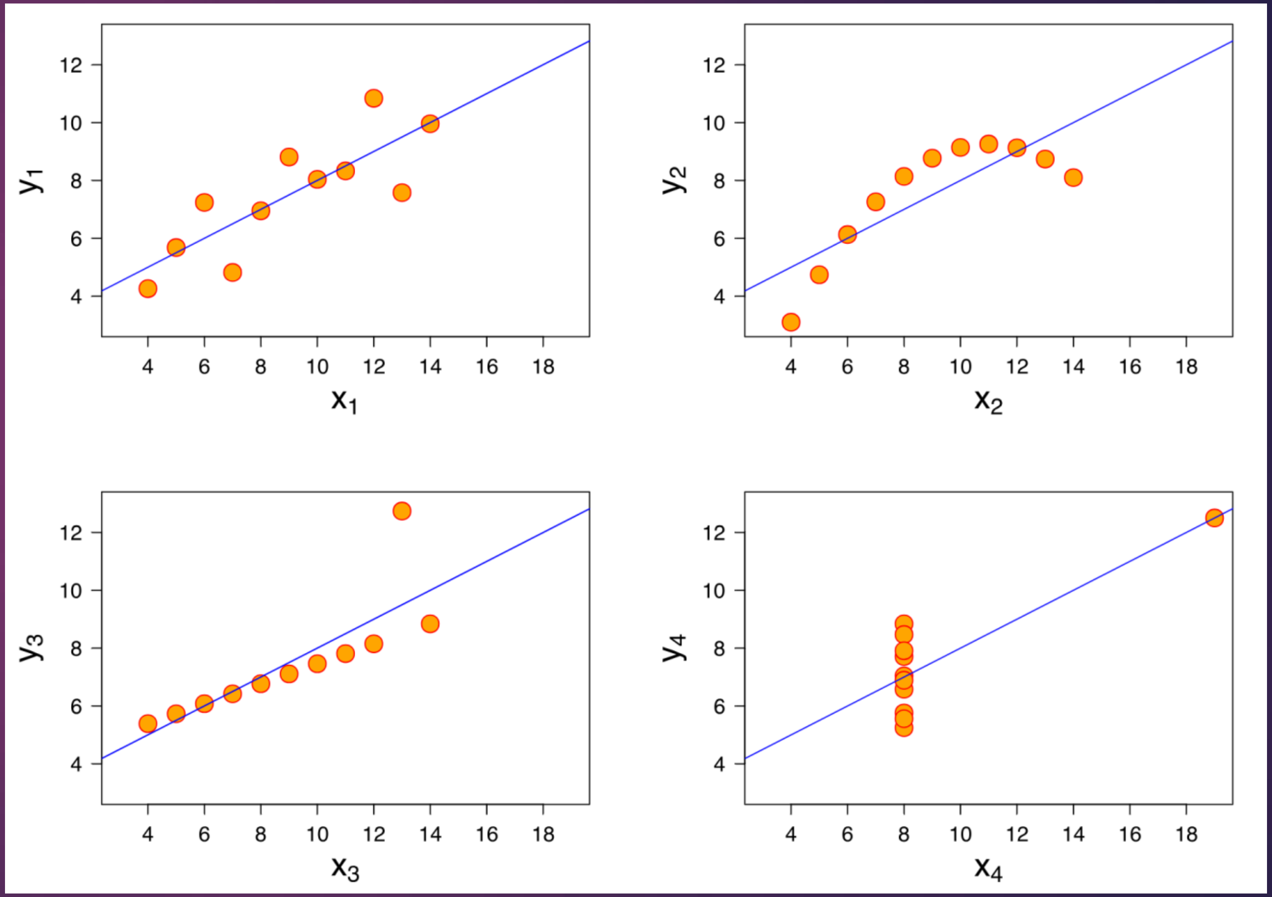
$$\rho_{X_j,y} = \frac{\mathrm{Cov}(X_j,y)}{\sigma_{X_j}\sigma_y}$$

Pearson Correlation Coefficient



Anscombe's quartet

Dataset I		Dataset II		Dataset III		Dataset IV	
x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89



https://en.wikipedia.org/wiki/Anscombe%27s_quartet

Anscombe's quartet

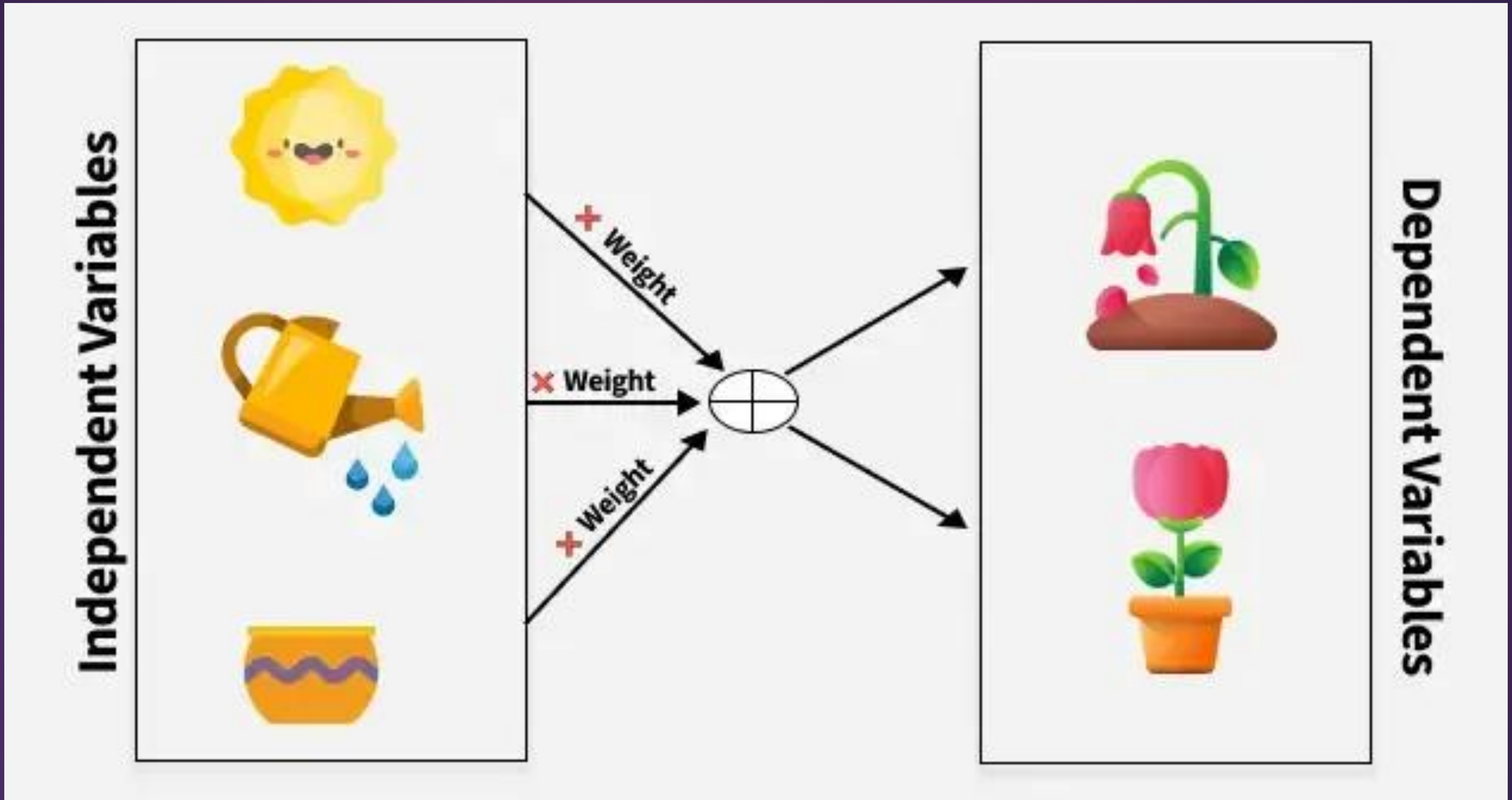
For all four datasets:

Property	Value	Accuracy
Mean of x	9	exact
Sample variance of x: s_x^2	11	exact
Mean of y	7.50	to 2 decimal places
Sample variance of y: s_y^2	4.125	±0.003
Correlation between x and y	0.816	to 3 decimal places
Linear regression line	$y = 3.00 + 0.500x$	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression: R^2	0.67	to 2 decimal places

https://en.wikipedia.org/wiki/Anscombe%27s_quartet

Relationship between Independent and dependent variables

Features



Targets
Or
Response variables

https://www.geeksforgeeks.org/engineering-mathematics/mathematics-covariance-and-correlation/

Coefficient correlation filter methods

Core Principles

1. Relevance to **Target** Variable: **Features** with a strong correlation (positive or negative) to the **target** variable are considered more relevant and informative for predicting the outcome.

2. Minimizing Redundancy (Multicollinearity): Highly correlated **features** among themselves (multicollinearity) can introduce redundancy and potentially hinder model performance. One of the highly correlated features should be removed to reduce this redundancy.

Coefficient correlation filter methods

Calculate Correlation Coefficients

1. **Feature-to-Target** Correlation: Calculate the correlation coefficient between each individual feature and the target variable. **Pearson correlation** is commonly used for continuous variables, while other measures like **Spearman correlation** or **Chi-squared** can be used for other data types.

2. **Feature-to-Feature** Correlation: Calculate the correlation coefficient between all pairs of features to identify **multicollinearity**.

Coefficient correlation filter methods

Rank and Select Features (Relevance)

1. Features are ranked based on the **absolute** value of their correlation with the **target** variable.

2. A threshold can be set, and **features** exceeding this threshold are selected.

Coefficient correlation filter methods

Address Multicollinearity (Redundancy)

1. If two or more **features** are highly correlated with each other, and also with the **target** variable, a decision is made to remove one of them.

2. A common strategy is to keep the **feature** with the higher correlation to the **target** variable and remove the others

Coefficient correlation filter methods

```
import pandas as pd
from sklearn.datasets
import make_regression
# Create a synthetic dataset
X, y = make_regression(n_samples=100, n_features=10, random_state=42)
df = pd.DataFrame(X, columns=[f'feature_{i}' for i in range(10)])
df['target'] = y
# Calculate Pearson correlation between features and target
correlations = df.corr()['target'].abs().sort_values(ascending=False)
print("Feature-to-Target Correlations:\n", correlations)
# Select features with a correlation above a threshold (e.g., 0.5)
selected_features_target = correlations[correlations > 0.5].index.tolist()
print("\nSelected features based on target correlation:", selected_features_target)
# Calculate feature-to-feature correlations to address multicollinearity
feature_corr_matrix = df[selected_features_target].corr().abs()
print("\nFeature-to-Feature Correlation Matrix:\n", feature_corr_matrix)
# Example of removing redundant features (manual for demonstration)
# If feature_1 and feature_2 are highly correlated, and feature_1 has higher target correlation, keep feature_1.
# This step often involves a more systematic approach in practice.
```

Coefficient correlation filter methods

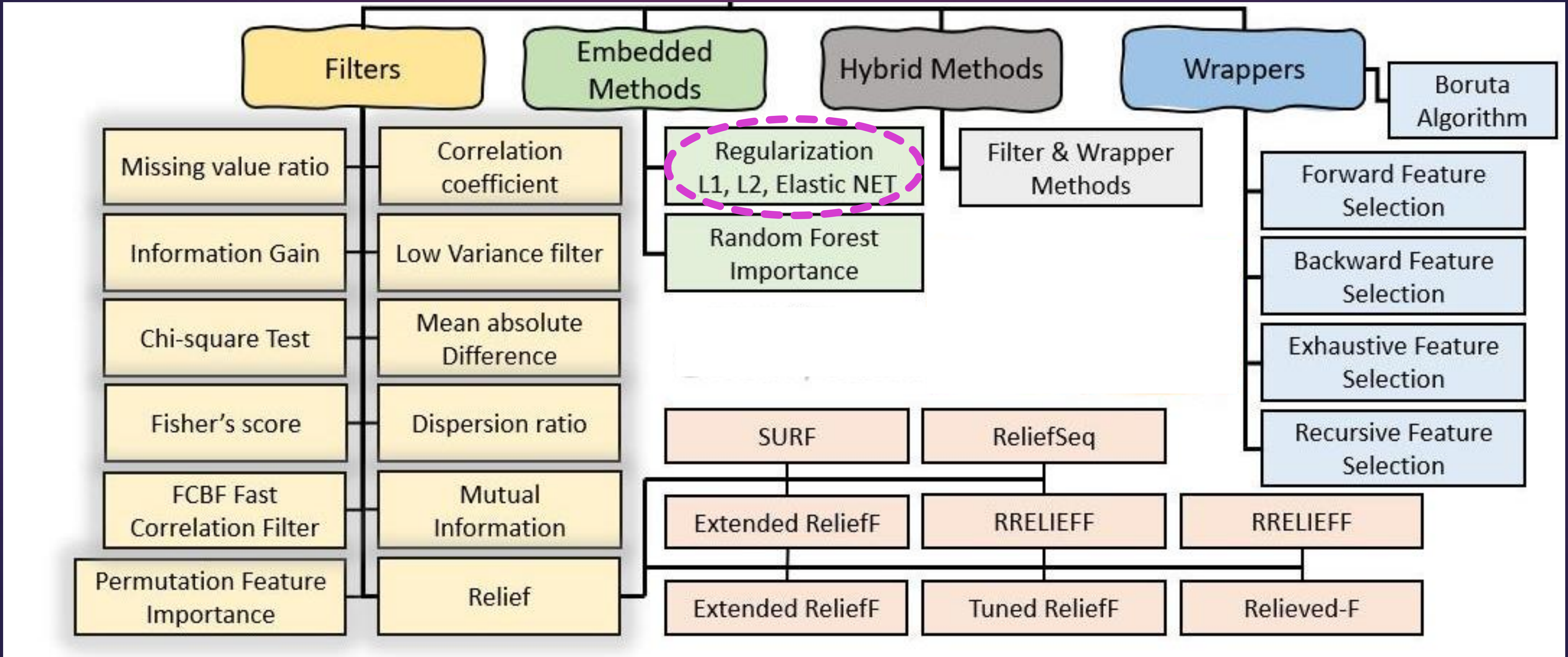
- **Advantages**
  1. **Computational Efficiency**: Filter methods are generally faster than wrapper or embedded methods as they don't involve training and evaluating a model repeatedly.
  2. **Generality**: They can be applied independently of the chosen machine learning algorithm.
  3. **Interpretability**: Correlation coefficients provide a clear understanding of the relationships between variables.

- **Limitations**
  1. **Ignores Feature Interactions**: Filter methods do not inherently consider interactions between features, which might be crucial for model performance.
  2. **Suboptimal Feature Subsets**: The selected feature subset might not be optimal for a specific learning algorithm, as the selection is independent of the model.

Practice

- https://www.datacamp.com/tutorial/tutorial-lasso-ridge-regression
- https://medium.com/@agrawalsam1997/feature-selection-using-lasso-regression-10f49c973f08
- https://www.blog.trainindata.com/lasso-feature-selection-with-python/

What’s next?



https://medium.com/analytics-vidhya/feature-selection-extended-overview-b58f1d524c1c

Xin chân thành cảm ơn!

LUU PHUC LOI, PHD

ZALO: 0901802182

LUU.P.LOI@GOOGLEMAIL.COM