* Feature Selection *

⊙ **Feature Selection:-**

↳ In datasets, some features are useful & important, others redundant / less important / irrelevant.

↳ These have -ve impact / reduce overall performance & model accuracy.

↳ Remove them using Feature selection.

↳ Automatic / manual way of selecting subset of most relevant features from original features set by removing redundant, irrelevant / noisy features.

⊙ **Need for Feature Selection:-**

① Dimensionality Reduction — High-dim dataset lead to increased computation time & overfitting.
   ↳ FS avoid curse of dimensionality.

② Improved Model Performance — select most relevant features, enhance model accuracy & generalization to training data.
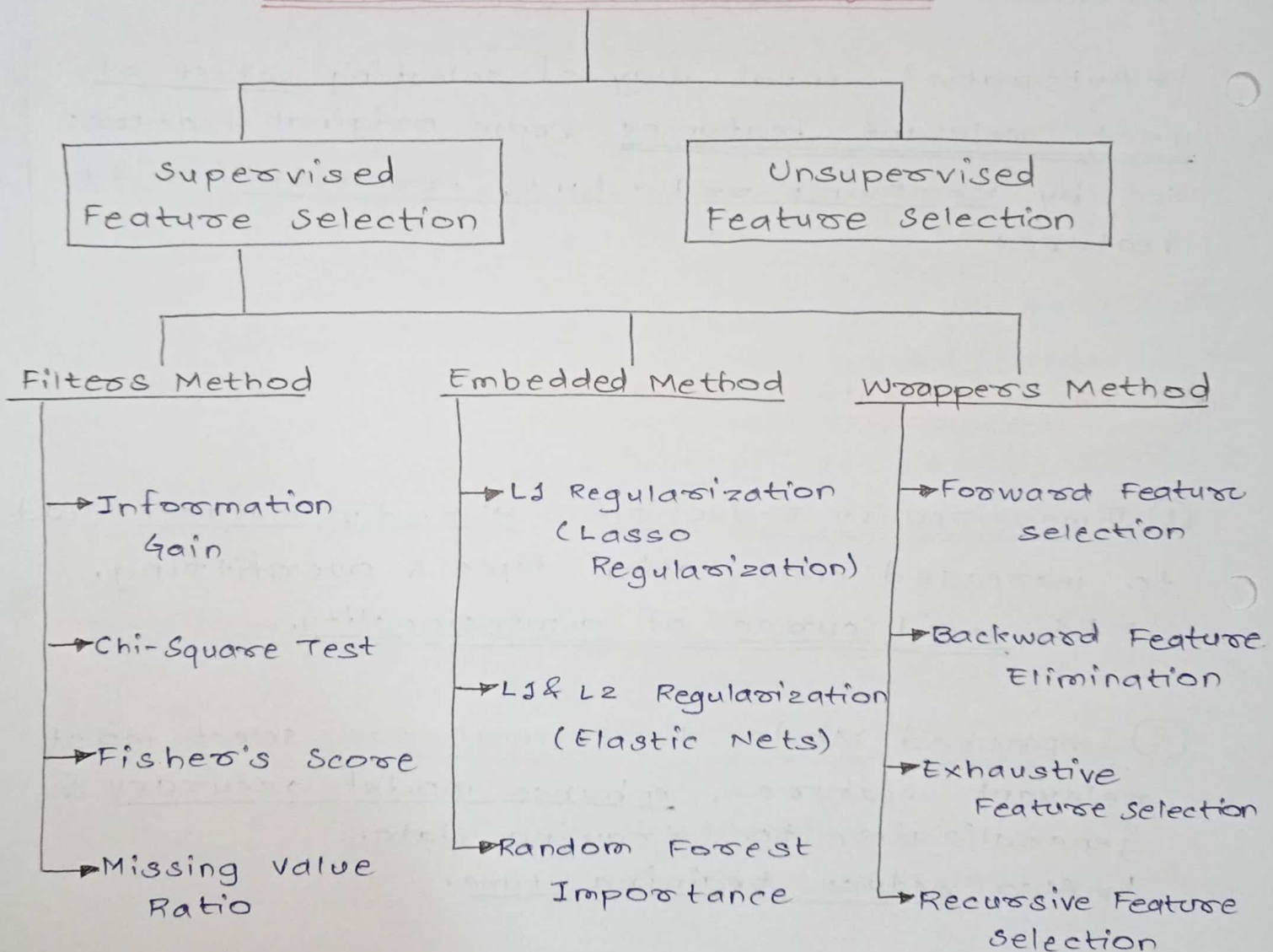   ↳ Also reduce training time.

③ Interpretability — Helps model simplication, thus easy to interpret & understand.

⊙ Feature Selection Techniques:-

① <u>Supervised Feature Selection Techniques:</u> consider target variable & can use for labelled dataset.

② <u>Unsupervised Feature Selection Technique:</u> Ignore target variable & can use for unlabelled dataset.

Feature Selection Techniques

```
                    Feature Selection Techniques
                              │
          ┌───────────────────┴───────────────────┐
    ┌─────┴──────┐                          ┌──────┴──────┐
    │ Supervised │                          │ Unsupervised│
    │ Feature    │                          │ Feature     │
    │ Selection  │                          │ Selection   │
    └─────┬──────┘                          └─────────────┘
          │
   ┌──────┼──────────────────┐
```

Filters Method	Embedded Method	Wrappers Method
→ Information Gain	→ L1 Regularization (Lasso Regularization)	→ Forward Feature selection
→ Chi-Square Test		→ Backward Feature Elimination
→ Fisher's Score	→ L1 & L2 Regularization (Elastic Nets)	→ Exhaustive Feature Selection
→ Missing value Ratio	→ Random Forest Importance	→ Recursive Feature Selection

Steps in Feature Selection :-

① **Data Pre-Processing** — Clean & prepare data.

② **Feature Scoring** — Compute each feature score to reflect its importance to target variable.

③ **Selection** — Select subset of most imp features based on their score & use them for training predictive model.

Advantages of Feature Selection :-

① **Improved model performance** — select relevant features & remove noisy ones, better performance & accuracy. Reduce overfitting.

② **Efficiency** — Reducing features lead to faster model performance, especially complex model / large dataset.

③ **Enhanced Model Interpretability** — Fewer features easy to understand & interpret.

④ **Reduced data storage & Processing costs** — Large datasets costly. Feature selection help reduce cost by eliminate redundent features.

⑤ **Remove curse of Dimensionality** — Feature Selection reduce dimensionality while preserving information.

⑥ **Better Data Visualization** — Fewer features make easy visualize data & Explore relationship betn variables, using EDA.

⊙ Limitations of Feature selection :—

① <u>Information loss</u> — Removing feature leads info loss from dataset.

② <u>Interactions & Non-linearity</u> — Feature selection assume linear relationship bet" feature & target.
↳ complicated to capture non-linear relationship.

③ <u>Increased Model Bias</u> — Feature selection may lead underfitting, reduce performance.

④ <u>Computational Cost</u> — computationally expensive, for large datasets.

⑤ <u>Data Variability</u> — sensitive to data variations, i.e. change in data/ outliers. So poor generalization.

⑥ <u>Domain Knowledge Requirement</u> — Domain knowledge required for feature selection.