# Features Selection

20CP401T

Himanshu K. Gajera
Department of Computer Science & Engineering
Pandit Deendayal Energy University, Gandhinagar

# Feature Selection

- Developing the machine learning model, only a few variables in the dataset are useful for building the model, and the rest features are either redundant or irrelevant.

- If we input the dataset with all these redundant and irrelevant features, it may negatively impact and reduce the overall performance and accuracy of the model.

- It is very important to identify and select the most appropriate features from the data and remove the irrelevant or less important features, which is done with the help of feature selection in machine learning.

# What is Feature Selection?

- A feature is an attribute that has an impact on a problem or is useful for the problem, and choosing the important features for the model is known as feature selection.

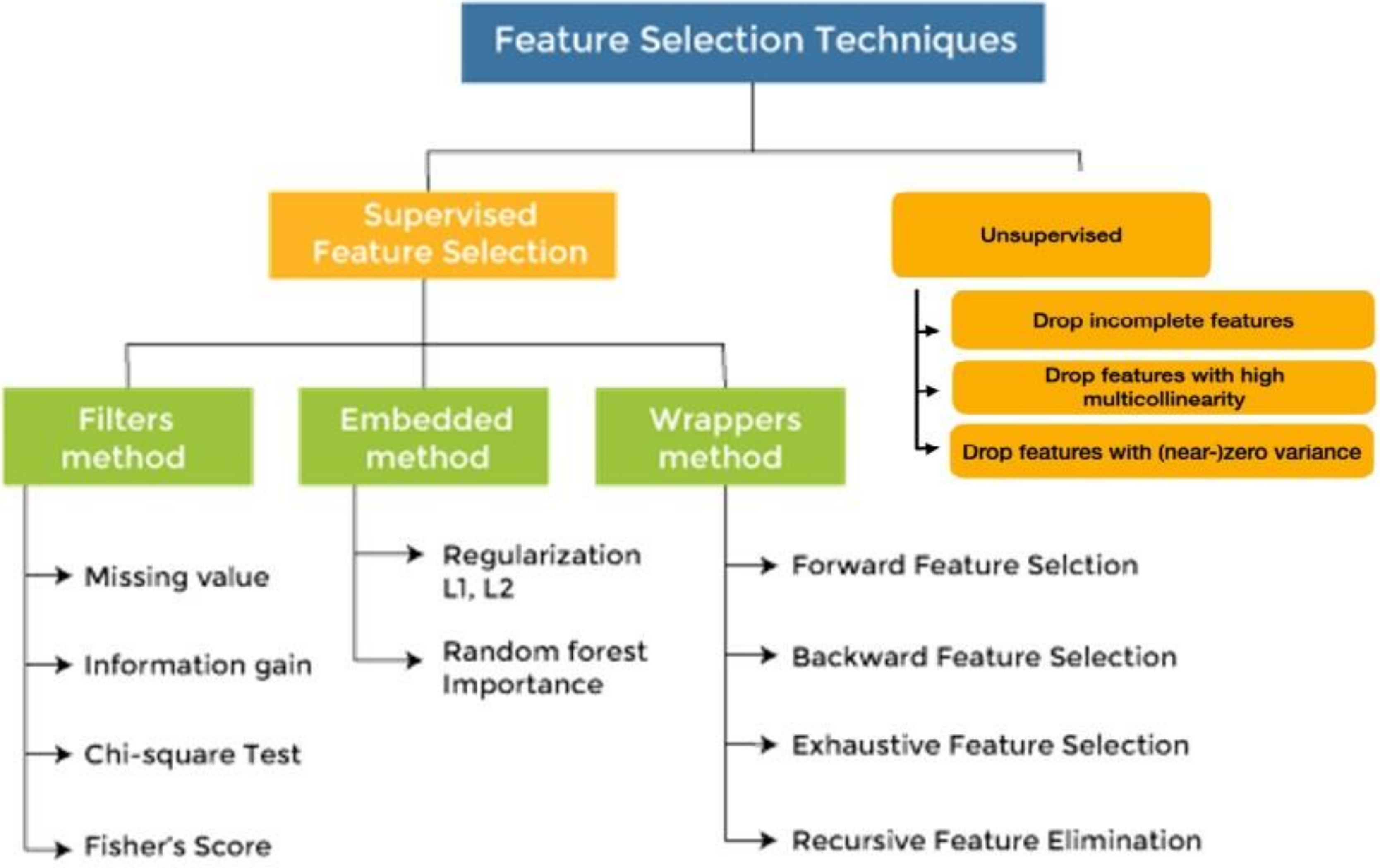
- Feature Selection and Feature Extraction. Although feature selection and extraction processes may have the same objective, both are completely different from each other.

- we can define feature Selection as, "***It is a process of automatically or manually selecting the subset of most appropriate and relevant features to be used in model building***."

# Need for Feature Selection

- It helps in avoiding the curse of dimensionality.
- It helps in the simplification of the model so that it can be easily interpreted by the researchers.
- It reduces the training time.
- It reduces overfitting hence enhance the generalization.

# Feature Selection Techniques



Feature Selection Techniques

Supervised Feature Selection

Unsupervised

- Drop incomplete features
- Drop features with high multicollinearity
- Drop features with (near-)zero variance

Filters method

- Missing value
- Information gain
- Chi-square Test
- Fisher's Score

Embedded method

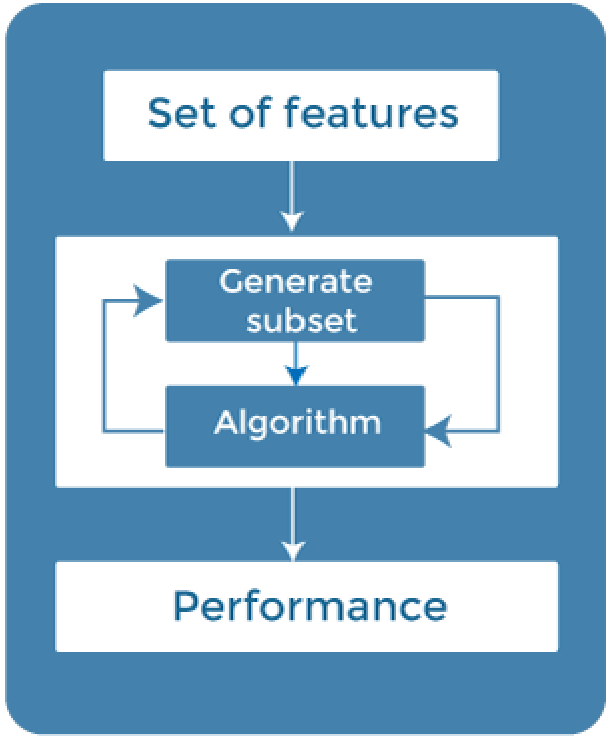
- Regularization L1, L2
- Random forest Importance

Wrappers method

- Forward Feature Selction
- Backward Feature Selection
- Exhaustive Feature Selection
- Recursive Feature Elimination

# Wrapper Methods

- In wrapper methodology, selection of features is done by considering it as a search problem, in which different combinations are made, evaluated, and compared with other combinations.



# Wrapper Methods

**Forward selection** – It is an iterative process. After each iteration, it keeps adding on a feature and evaluates the performance to check whether it is improving the performance or not.

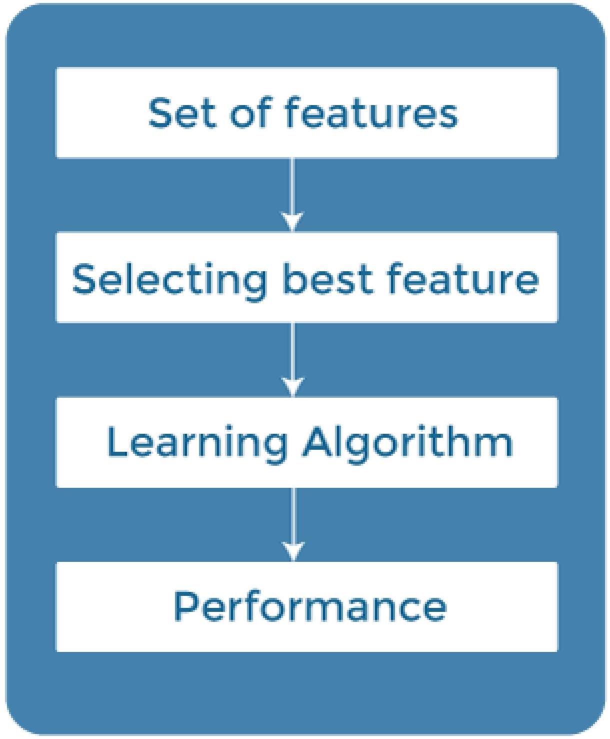
**Backward elimination** – It is also an iterative approach, but it is the opposite of forward selection. This technique begins the process by considering all the features and removes the least significant feature.

**Exhaustive Feature Selection-** It is one of the best feature selection methods, which evaluates each feature set as brute-force.

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

# Filter Methods

- In Filter Method, features are selected on the basis of statistics measures. This method does not depend on the learning algorithm and chooses the features as a pre-processing step.
- The filter method filters out the irrelevant feature and redundant columns from the model by using different metrics through ranking.



Set of features

Selecting best feature

Learning Algorithm

Performance

# Filter Methods

**Information Gain:** Information gain determines the reduction in entropy while transforming the dataset.

**Chi-square Test:** Chi-square test is a technique to determine the relationship between the categorical variables.

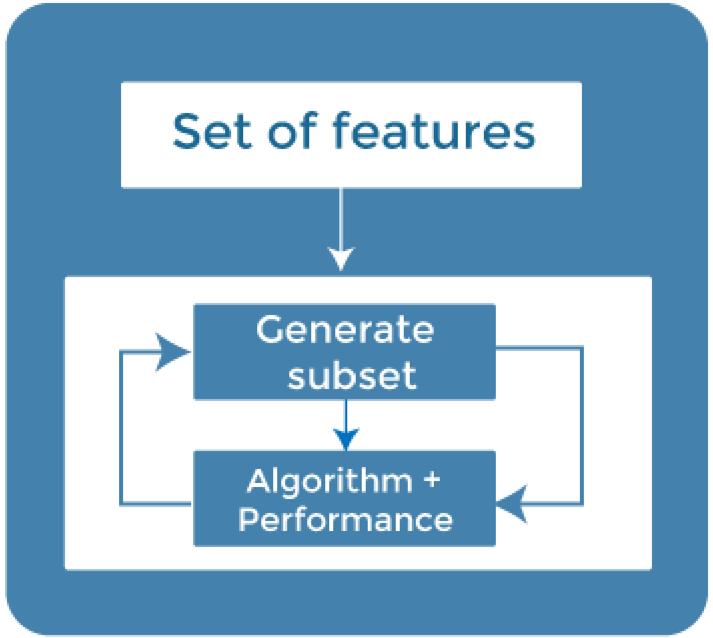
**Fisher's Score:** Fisher's score is one of the popular supervised technique of features selection. It returns the rank of the variable on the fisher's criteria in descending order.

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

$$\text{Missing Value Ratio}=\frac{Number\ of\ Missing\ values * 100}{Total\ number\ of\ observations}$$

# Embedded Methods

- Embedded methods combined the advantages of both filter and wrapper methods by considering the interaction of features along with low computational cost.

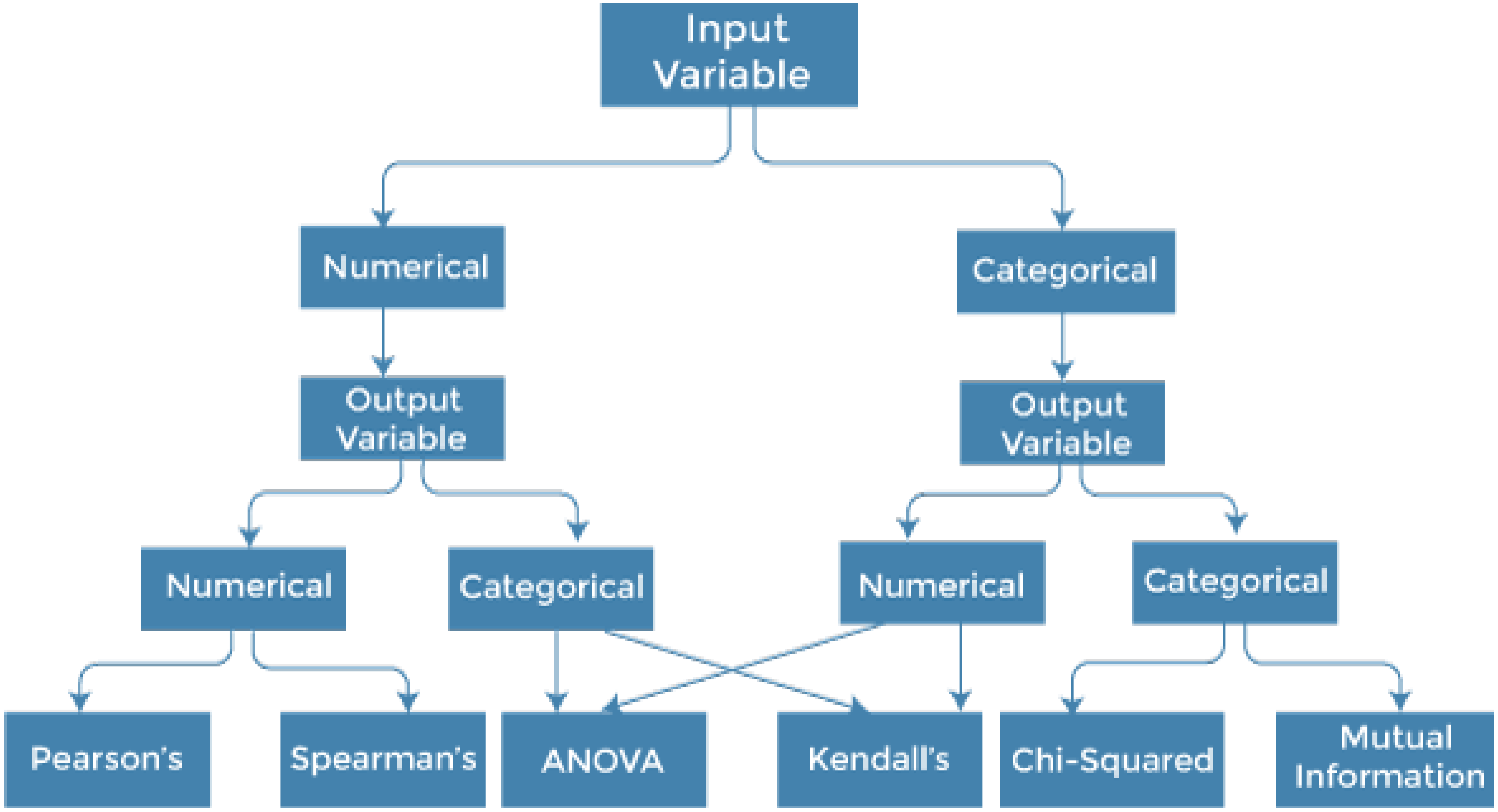


# Embedded Methods

- **Regularization**:
  - This method adds a penalty to different parameters of the machine learning model to avoid over-fitting of the model.
  - Feature selection uses Lasso (L1 regularization) and Elastic nets (L1 and L2 regularization).
  - The penalty is applied over the coefficients, thus bringing down some coefficients to zero. The features having zero coefficient can be removed from the dataset.

- **Tree-based methods** – These methods such as Random Forest, Gradient Boosting provides us feature importance as a way to select features as well.

# How to choose a Feature Selection Method?



# Conclusion

- Feature selection is a very complicated and vast field of machine learning.
- Lots of studies are already made to discover the best methods.
- There is no fixed rule of the best feature selection method.
- Choosing the method depend on a machine learning engineer who can combine and innovate approaches to find the best method for a specific problem.
- Try a variety of model fits on different subsets of features selected through different statistical Measures.
- Dimensionality reduction techniques such as Principal Component Analysis (PCA), Heuristic Search Algorithms, etc. don’t work in the way as to feature selection techniques but can help us to reduce the number of features.