FEATURE SELECTION

Feature selection is the process of selecting a subset of relevant features (variables, predictors) for use in model construction.

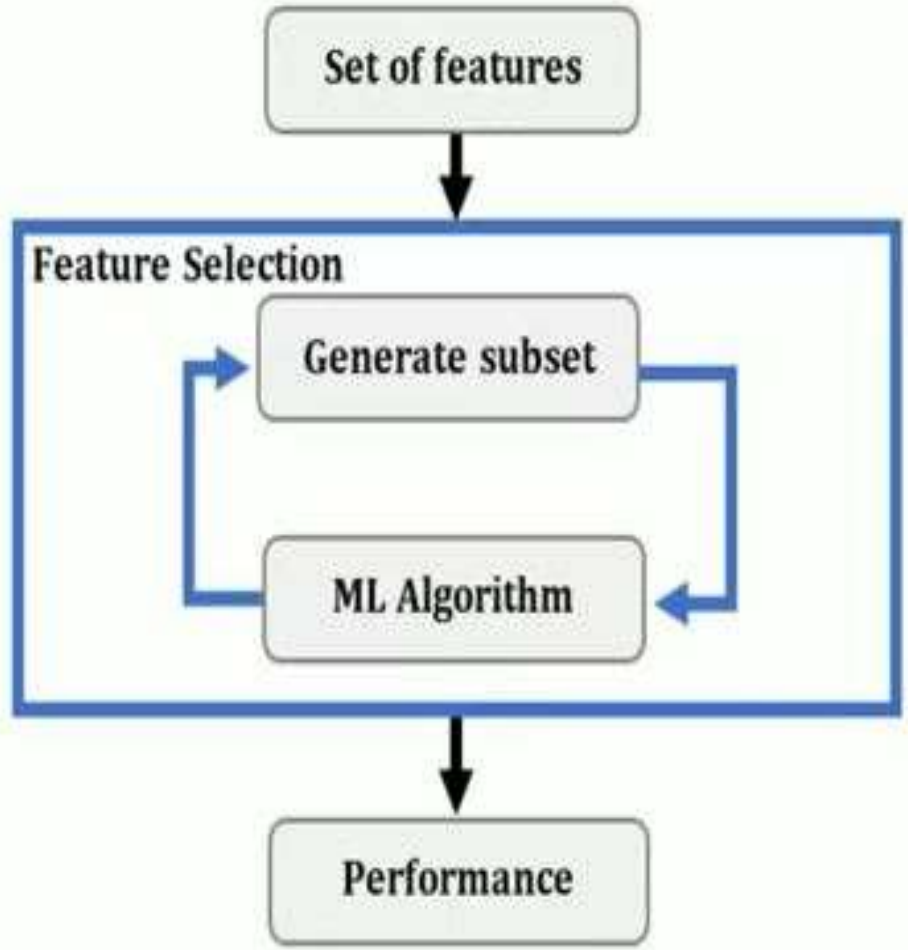
- **Unsupervised Methods**

- Unsupervised feature selection methods are applied to unlabeled data.

- An unsupervised selection method rates each feature dimension according to a number of factors, including entropy, variance, and the capacity to maintain local similarity.

- **Supervised Methods**
- On the other hand, we use supervised feature selection methods on labeled data.
- They determine the features that are expected to maximize the supervised model's performance.
- Supervised feature selection methods can be split into three primary categories based on the feature selection strategy.

- **Wrapper Methods**
- We use a wrapper method after choosing the ML algorithm to use.
- For each feature subset, we estimate the algorithm's performance by training and evaluating it using only the features in a subset.
- Then, we add or remove features based on the estimate.
- This is an iterative process.



We use a greedy strategy to form feature subsets.

- In **forward wrapper methods**, we start from an empty feature set and add the feature maximizing the performance in each step until no substantial improvement is observed.
- So, if there are **n** features, we build **n** ML models in the first iteration.
- Then, we select the feature corresponding to the model with the best performance.
- In the second iteration, we repeat the process with the remaining **n-1** features.

We use a greedy strategy to form feature subsets.

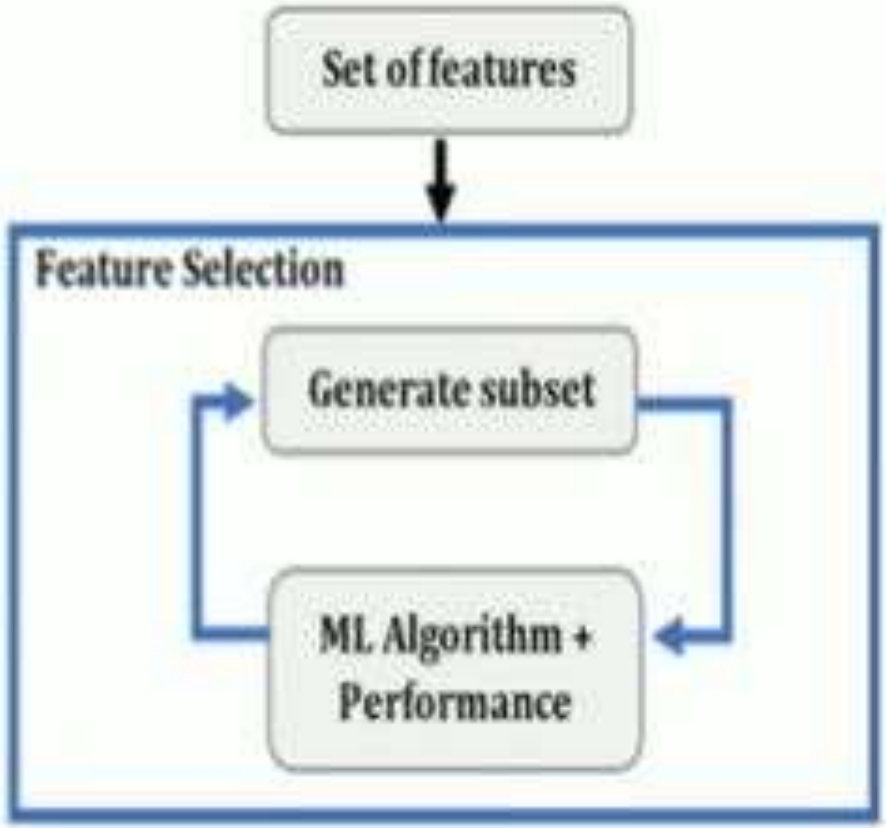
- **Backward methods** work the opposite way.
- They start from the full feature set and remove them one by one.
- Finally, stepwise methods reconsider features.
- So, in each iteration, they can remove a feature previously added as well as add a feature discarded in a previous step.

Filter Methods

- Filter methods use statistical tools to select feature subsets based on their relationship with the target.
- These methods remove features with low correlation with the target variable before training the final ML model.
- In doing so, they compute correlation and estimate the strength of the relationship using the, and other statistical tools.

Chi-Square Test, Information Gain, Fisher's Score, Pearson correlation, ANOVA, variance thresholding

- **Intrinsic (or Embedded) Methods**
- Here selection of feature happens simultaneously with and is performed implicitly by the ML algorithm of our choice.
- During training, some steps of the ML algorithm do feature selection:
- For instance, this is the case with decision trees.
- At each node split, they choose the best feature to split the data by.
- Those choices represent feature selection.



Feature selection methods allow us to:

- Reduce overfitting as less redundant data means less chance to make decisions based on noise;
- Improve accuracy by removing misleading and unimportant data;
- Reduce training time since data with fewer columns mean faster training.
- However, feature selection methods are hard to apply to high-dimensional data.
- The more features we have, the longer it takes for selection to complete.
- Also, there's the risk of overfitting when there aren't enough observations.

- Regularization is a technique used to reduce errors by fitting the function appropriately on the given training set and avoiding overfitting.

The commonly used regularization techniques are :

- Lasso Regularization – L1 Regularization
- Ridge Regularization – L2 Regularization
- Elastic Net Regularization – L1 and L2 Regularization

Lasso Regression

- A regression model which uses the L1 Regularization technique is called LASSO (Least Absolute Shrinkage and Selection Operator) regression.
- Lasso Regression adds the “absolute value of magnitude” of the coefficient as a penalty term to the loss function(L).

Lasso Regression

$$\text{Cost} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{i=1}^{m} |w_i|$$

Ridge Regression

- A regression model that uses the **L2 regularization** technique is called **Ridge regression**.
- **Ridge regression** adds the "*squared magnitude*" of the coefficient as a penalty term to the loss function(L).

$$\text{Cost} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda\sum_{i=1}^{m} w_i^2$$

Elastic Net Regression

- This model is a combination of L1 as well as L2 regularization.
- That implies that we add the absolute norm of the weights as well as the squared measure of the weights.
- With the help of an extra hyperparameter that controls the ratio of the L1 and L2 regularization.

$$\text{Cost} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda\left((1-\alpha)\sum_{i=1}^{m}|w_i| + \alpha\sum_{i=1}^{m}w_i^2\right)$$