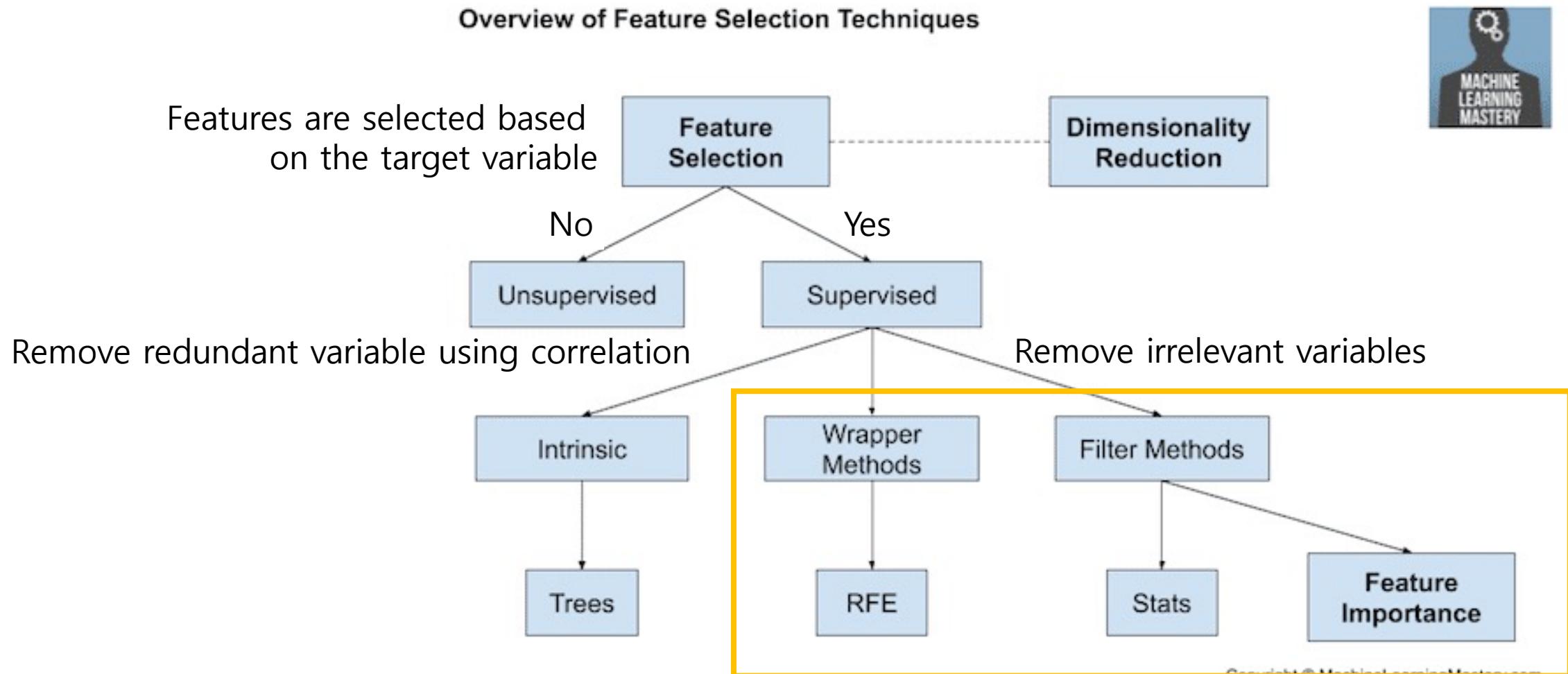


1. Feature Selection

특성추출 (Feature selection)

: intend to reduce the number of input variables to those that are believed to be most useful to a model in order to predict the target value



2. Supervised Feature Selection

Labeled Data	
Training	Test
80 %	20 %

Wrapper and Filter methods

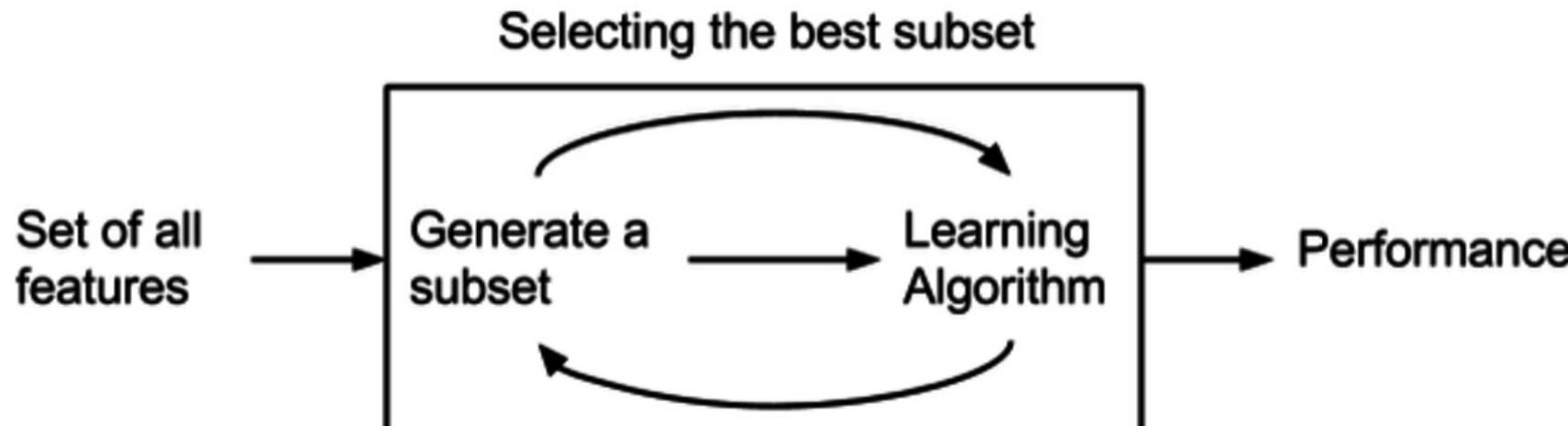
: always supervised and are evaluated based on the performance of a resulting model on a hold out dataset

01. Wrapper

: create many models with different subsets of input features and select those features that result in the best performing model according to a performance metric.

- unconcerned with the variable types, although they can be computationally expensive(need many times).

Ex) [RFE](#), Forward Greedy, Backward Greedy, Genetic Search, Local Search



2. Supervised Feature Selection

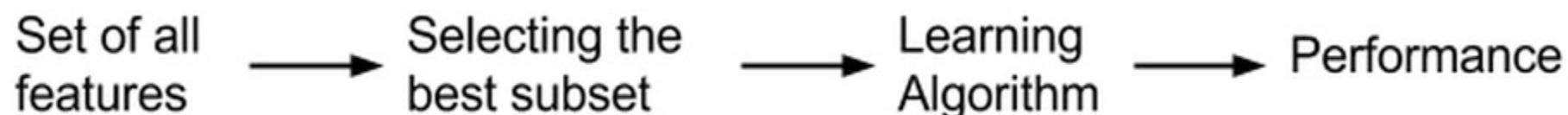
Labeled Data	
Training	Test
80 %	20 %

Wrapper and Filter methods

: always supervised and are evaluated based on the performance of a resulting model on a hold out dataset

02. Filter methods

: use statistical techniques to evaluate the relationship between each input variable and the target variable, and these scores are used as the basis to choose (filter) those input variables that will be used in the model.
Ex) t-test, chi-square test, Information Gain



1. Filter는 종속변수와의 상관 관계에 의해 피처의 관련성을 측정하지만 Wrapper는 실제 모델을 학습하여 피처의 부분집합의 유용성을 측정합니다.
2. Filter는 모델을 학습하지 않기 때문에 Wrapper보다 속도가 빠르며 Wrapper는 계산적이기에 비용이 많이 들고 속도가 느립니다.
3. Filter는 통계 방법을 사용하여 피처의 부부 집합을 만들지만 Wrapper는 교차 검증을 활용하여 피처의 부분 집합을 만듭니다.

2. Supervised Feature Selection

Wrapper and Filter methods

: always supervised and are evaluated based on the performance of a resulting model on a hold out dataset

(+) Wrapper and Filter methods

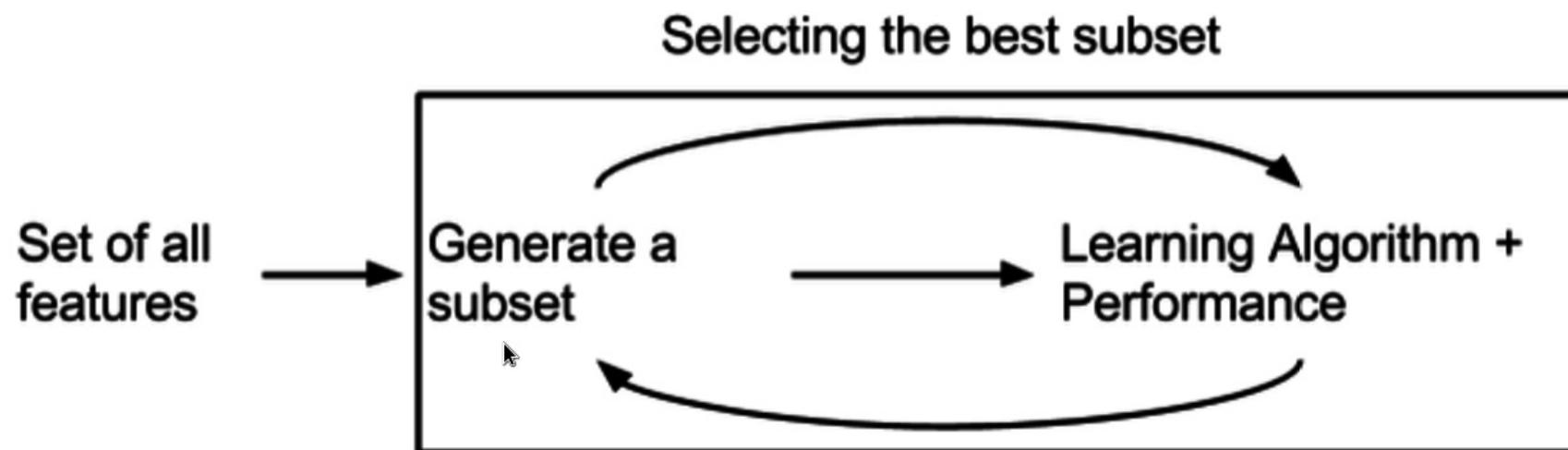
1. Filter는 종속변수와의 상관 관계에 의해 피처의 관련성을 측정하지만 Wrapper는 실제 모델을 학습하여 피처의 부분집합의 유용성을 측정합니다.
2. Filter는 모델을 학습하지 않기 때문에 Wrapper보다 속도가 빠르며 Wrapper는 계산적이기에 비용이 많이 들고 속도가 느립니다.
3. Filter는 통계 방법을 사용하여 피처의 부부 집합을 만들지만 Wrapper는 교차 검증을 활용하여 피처의 부분 집합을 만듭니다.
4. Filter는 항상 최적의 피처 부분 집합을 선택할 수는 없지만 Wrapper는 항상 최적의 피처 부분 집합을 선택할 수 있다.
5. Wrapper 방법은 Filter 방법 보다 overfitting 되기 쉽다.

Labeled Data	
Training	Test
80 %	20 %

2. Supervised Feature Selection

03. Intrinsic method(Embedded Method)

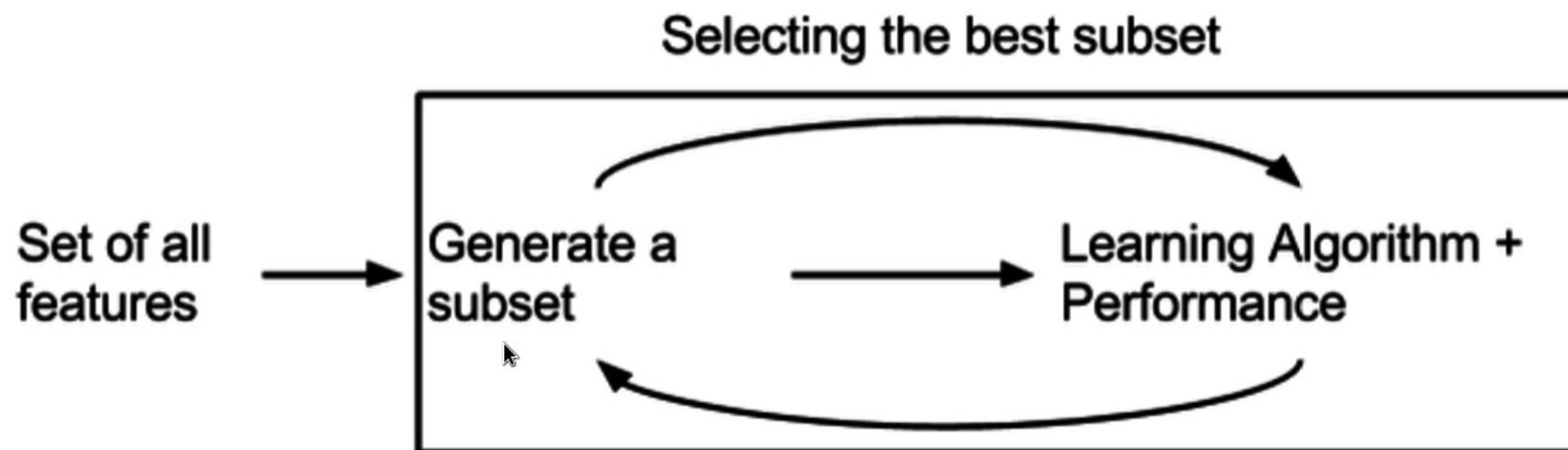
: machine learning algorithms that perform feature selection automatically as part of learning the model
Ex) LASSO = L1 regularization, RIDGE = L2 regularization, Tree



2. Supervised Feature Selection

03. Intrinsic method(Embedded Method)

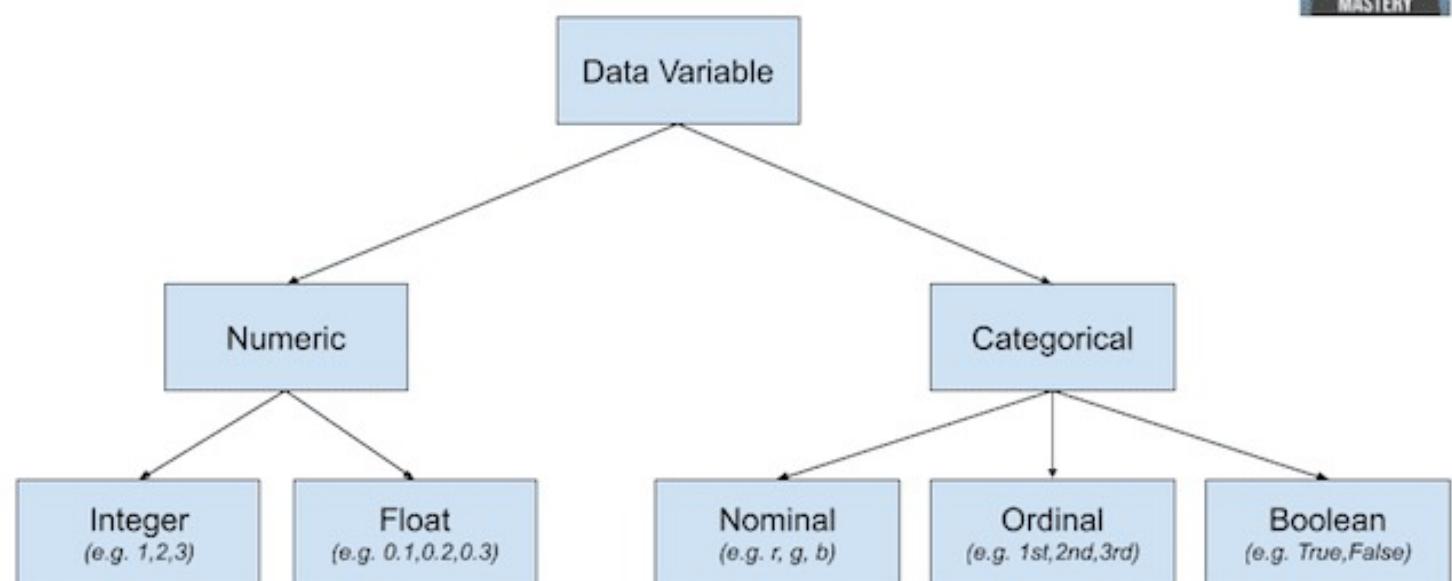
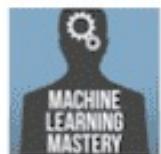
: machine learning algorithms that perform feature selection automatically as part of learning the model
Ex) LASSO = L1 regularization, RIDGE = L2 regularization, Tree



2. Statistics for Filter-Based Feature Selection Methods

- Input variables : provided as input to a model → to reduce in size
- Output variables(response variable) : a model is intended to predict
 - Numerical output : Regression predictive modeling problem
 - Categorical output : Classification predictive modeling problem

Overview of Data Variable Types



Copyright © MachineLearningMastery.com

2. Statistics for Filter-Based Feature Selection Methods

(1) Numerical Input, Numerical Output

: regression predictive modeling problem with numerical input variables.

- Pearson's correlation coefficient (linear)
- Spearman's rank coefficient (nonlinear)

(2) Numerical Input, Categorical Output

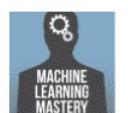
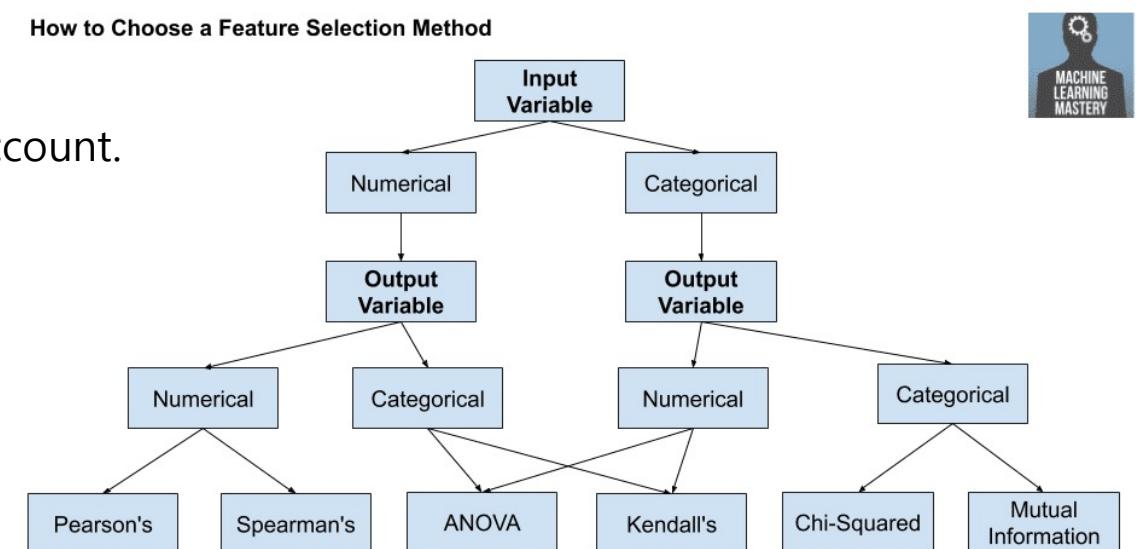
: classification predictive modeling problem with numerical input variables.

(most common example of a classification problem)

Again, the most common techniques are correlation based,
although in this case, they must take the categorical target into account.

- ANOVA correlation coefficient (linear)
- Kendall's rank coefficient (nonlinear)

* Kendall does assume that the categorical variable is ordinal



2. Statistics for Filter-Based Feature Selection Methods

(3) Categorical Input, Numerical Output

: regression predictive modeling problem with categorical input variables.

This is a strange example of a regression problem (e.g. you would not encounter it often).

use the same “Numerical Input, Categorical Output” methods (described above), but in reverse.

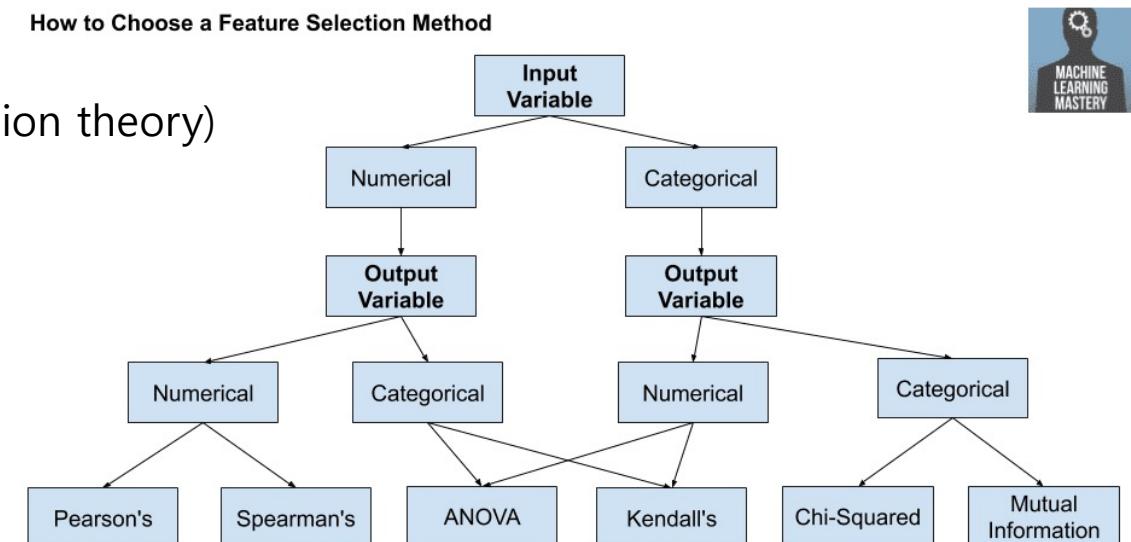
- ANOVA correlation coefficient (linear)
- Kendall’s rank coefficient (nonlinear)

(4) Categorical Input, Categorical Output

: classification predictive modeling problem with categorical input variables.

- Chi-Squared test (contingency tables)
- Mutual Information(information gain, from the field of information theory)

In fact, mutual information is a powerful method
that may prove useful for both categorical and numerical data,
e.g. it is agnostic to the data types.



3. Tips and Tricks for Feature Selection

Correlation Statistics

: The scikit-learn library provides an implementation of most of the useful statistical measures.

- Pearson's Correlation Coefficient: `f_regression()`
- ANOVA: `f_classif()`
- Chi-Squared: `chi2()`
- Mutual Information: `mutual_info_classif()` and `mutual_info_regression()`

Also, the SciPy library provides an implementation of many more statistics, such as Kendall's tau (`kendalltau`) and Spearman's rank correlation (`spearmanr`).

3. Tips and Tricks for Feature Selection

Selection Method

: The scikit-learn library also provides many different filtering methods once statistics have been calculated for each input variable with the target.

Two of the more popular methods include:

- Select the top k variables: [SelectKBest](#)
- Select the top percentile variables: [SelectPercentile](#)

Transform Variables

Consider transforming the variables in order to access different statistical methods.

For example, you can transform a categorical variable to ordinal, even if it is not.

Some statistical measures assume properties of the variables,

such as Pearson's that assumes a Gaussian probability distribution to the observations and a linear relationship.

You can transform the data to meet the expectations of the test and try the test regardless of the expectations and compare results.