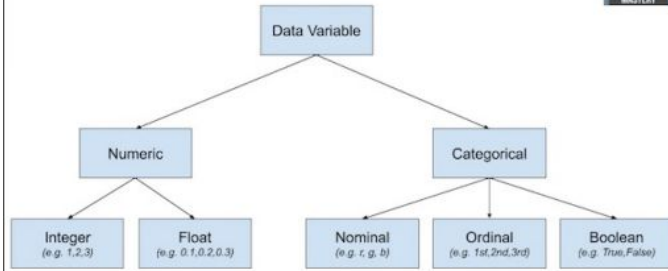


# Feature selection

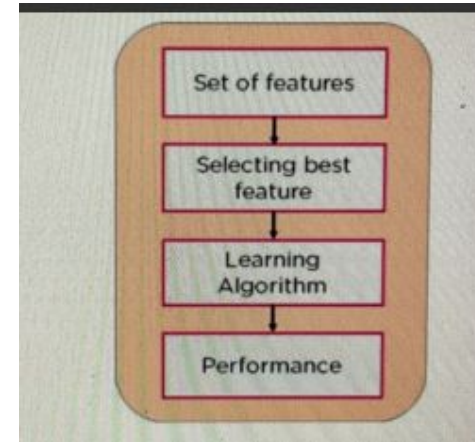
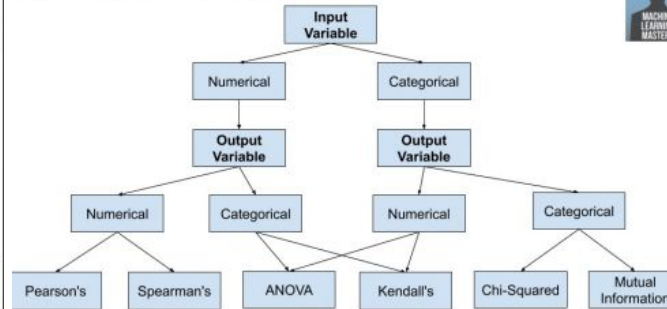
**Feature selection:** is a process of selection a subset of relevant features for use in model construction.

- **Techniques:**
  1. **Filter methods:** features are dropped based on their relation to the output.

Overview of Data Variable Types



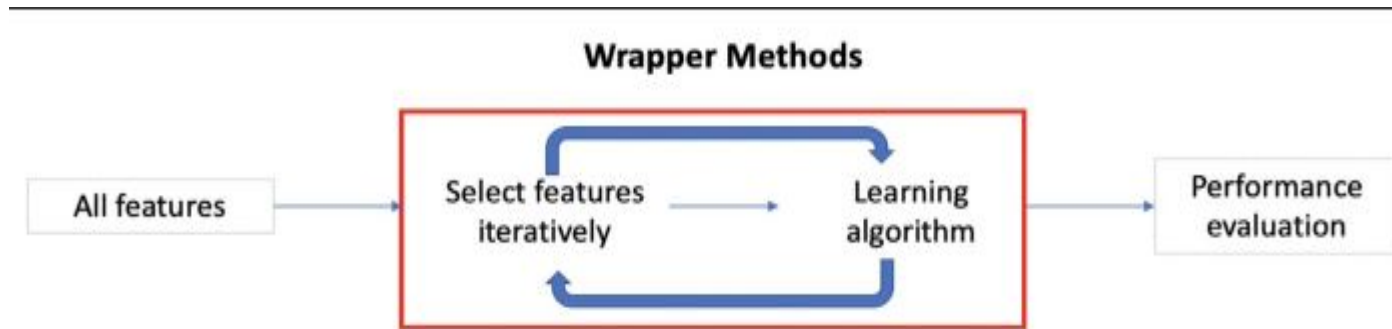
How to Choose a Feature Selection Method



- **Numerical Input, Numerical output:** (Regression):
  - linear: Person's correlation coefficient.
  - non-linear: Spearman's rank coefficient.
- **Numerical Input, Categorical output:** (Classification):
  - linear: ANOVA correlation coefficient.
  - non-linear: Kendall's rank coefficient.
- **Categorical Input, Numerical output:**
  - I can use the same "Num. Input, cat output" methods, but in reserve.
- **Categorical Input, Categorical output:** (Classification):
  - Chi-squared test.
  - Mutual Information.

For example, consider a dataset containing two variables:  $X$  and  $Y$ . If the values of  $X$  are completely independent of the values of  $Y$ , then the mutual information between  $X$  and  $Y$  will be zero, because knowing the value of  $X$  does not provide any additional information about the value of  $Y$ . On the other hand, if the values of  $X$  and  $Y$  are perfectly correlated, then the mutual information between  $X$  and  $Y$  will be maximal, because knowing the value of  $X$  allows you to perfectly predict the value of  $Y$ .

2. **Wrapper Methods:** use a predictive model to evaluate the utility of each feature, and select the features that improve the performance of the model.



- **Person's correlation coefficient:**

- has value [-1,1]
- 0: no correlation.
- A value closer to 1 stronger positive correlation.
- A value closer to -1 stronger negative correlation.

- **spearman's rank:**

statistical measure that quantifies the strength and direction of the relationship between two variables. Range [-1,1]

Where:

- D: is the difference in ranks between two variable.  
 $d_i = R(x_i) - R(y_i)$ .
- n: number of samples.
- 6: constant.

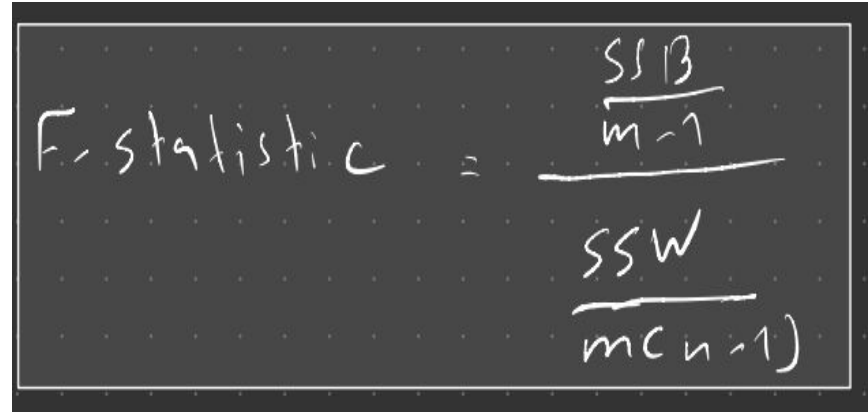
$$r_s = 1 - \frac{6 \sum D^2}{n(n^2 - 1)}$$

- **ANOVA: Analysis of Variance:**

is to calculate the SST, SSW, SSB and corresponding degrees of freedom.

1. SSA: is the sum of the square distance between each data point ( $n$ ) and the mean of the dataset.  
The degrees of freedom is the number of groups ( $m$ ) .
  - degree of freedom:  $m * (n-1)$ .
2. SSW: the sum of the squared distance between each data point and respective group mean.
  - degree of freedom:  $m * (n-1)$ .
3. SSB: the sum of the squared distance between each group mean and the mean of the dataset for each data point.
  - degree of freedom:  $(m - 1)$ .

- P-Value:- is the probability of obtaining an F as larger or larger than the observed.
- if  $P > 0.05$ : we accept the probability.



A chalkboard with a dark background and a light grid. The text 'F-statistic' is written in white chalk on the left. To its right is an equals sign followed by a fraction. The numerator of the fraction is  $\frac{SSB}{m-1}$ , with 'SSB' above a horizontal line and 'm-1' below it. The denominator of the fraction is  $\frac{SSW}{m(n-1)}$ , with 'SSW' above a horizontal line and 'm(n-1)' below it.

$$F\text{-statistic} = \frac{\frac{SSB}{m-1}}{\frac{SSW}{m(n-1)}}$$

- **Kendall Rank:**

$$T = \frac{\text{Number of concordant Pairs} - \text{Number of discordant Pairs}}{n(n-1)/2}$$

- **Concordant pair:**  
( $X_1 > X_2$  &  $Y_1 > Y_2$ ). or ( $X_1 < X_2$  &  $Y_1 < Y_2$ ).
- **Discordant pair:**  
( $X_1 > X_2$  &  $Y_1 < Y_2$ ). or ( $X_1 < X_2$  &  $Y_1 > Y_2$ ).

- **Chi- Squared test:**

Where:

c: degrees of freedom.

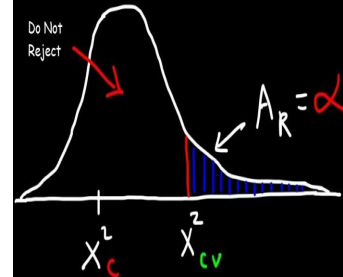
o: observed values.

e: expected values.

## The Chi-Square Test

	Monday	Tuesday	Wednesday	Thursday	Friday
Observed Absences	23	16	14	19	28
Expected Absences	20	20	20	20	20

$$\chi^2_c = \frac{\sum (o - e)^2}{E}$$



$$df = n - 1$$

- **Mutual information:** is non-negative value, it measures the amount of information we can know from one variable by observing the values of the second variable.
- **Entropy:** is the measure of the information.

$$H(x) = - \sum p(x) \log(p(x))$$

$P(x)$ : Probability of the values of  $x$ .

- **Relative Entropy:** measures the distance between two distributions.

$$D(p|q) = \sum p(x) \frac{p(x)}{q(x)}$$

$p(x)$ ,  $q(x)$  are two probability distribution.

- **MI:** is the relative entropy between the joint distribution of the two variables and the product of their marginal distribution.

$$I(x, y) = \sum_x \sum_y p(x, y) \log \left( \frac{p(x, y)}{p(x) \cdot p(y)} \right)$$

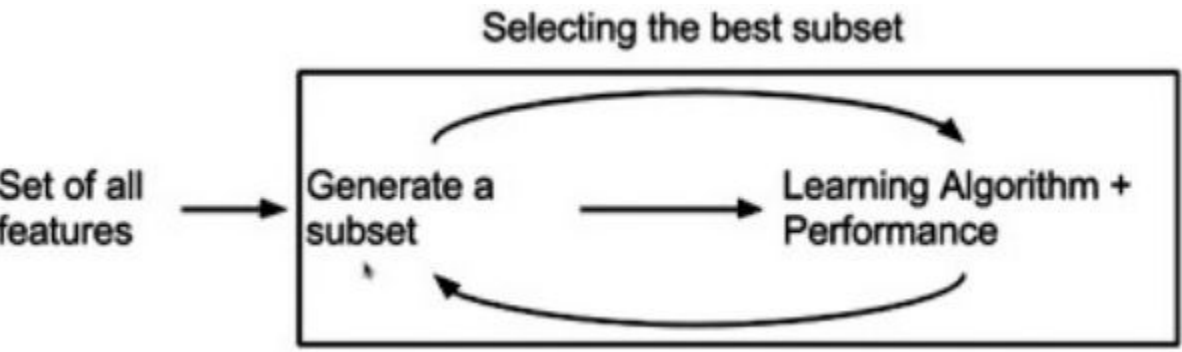
$I(x, y)$ :- is MI between variables  $x$  and  $y$ .

$P(x, y)$ :- is joint probability of the two variables.

- **Forward selection:** is iterative process, which begins with an empty set of features. After each iteration, it keeps adding on a feature and evaluates the performance to check whether it is improving the performance or not. The process continues until the addition of variable does not improve the performance of model.
- **Backward elimination:** begins by considering all the features and removes the worst feature. This continues until removing the features does not improve the performance of model.
- **Exhaustive feature selection:** considers all possible combinations of features and selects the combination that gives the best model performance.  
It is usually used for small datasets with small number of features.
- **Recursive feature Elimination (RFE):** that recursively removes features from the model, based on the improvement in the model performance. It starts with all features and removes the least important features one by one, until the desired number of features is reached.



3. **Embedded Methods:** combined the advantages of both filler and wrapper by considering the interaction of features with low computational.



- Regularization: adds a penalty to different parameters of models for avoiding overfitting. This penalty term is added to the coefficients. Hence it shrinks some coefficients to zeros. Those features with zero coefficients can be removed from the dataset.

Types:

- L1-regularization (lasso Regularization).
- Elastic-Net, (L1 and L2 Regularization).

**L1**

$$Loss = Error(y, \hat{y}) + \lambda \sum_{i=1}^N |w_i|$$

- **Random Forest Importance:**

Different tree- based methods of feature selection help us with feature importance to provide a way of selecting. feature importance specifies with feature has a great impact on the large variable. RF ranks automatically the nodes by their performance. Nodes are arranged as per the impurity values, and thus it allows to pruning of trees below a specific mode. The remaining nodes create a subset of the most important features.

- For classification: the measure of impurity either the Gini or information gain/entropy.
- For regression: the measure of impurity is variance.

	Filter method	Wrapper method	Embedded method
What is it?	Uses proxy measure	Uses predictive model	Feature selection is embedded in the model building phase
Speed	Computationally faster	Slower	Medium
Overfitting	Avoids overfitting	Prone to overfitting	Less prone to overfitting
Performance	Sometimes may fail to select best features	Better performance	Good performance

- **4 best way of feature selection:**
  1. Select K-Best.
  2. Recursive Feature Elimination.
  3. Correlation-matrix with Heatmap.
  4. Random-forest Importance.

4. **Hybrid method:** combine multiple feature selection methods in order to improve accuracy.

- Divided into two main categories:
  1. **Fillers-wrapper:** These method, combine filter method, which uses a statistical measure to evaluate importance of each filter, with a wrapper method wish uses model to evaluate the performance of features subset. Example: Recursive feature (Elimination with cross-validation (RFCCV), which uses (RFE) to select subset of feature, and cross-validation to evaluate the performance of selected features.
  2. **Embedded-wrapper:** uses the model itself to select features with wrapper method. Example: select from model with cross validation. (SSCV).