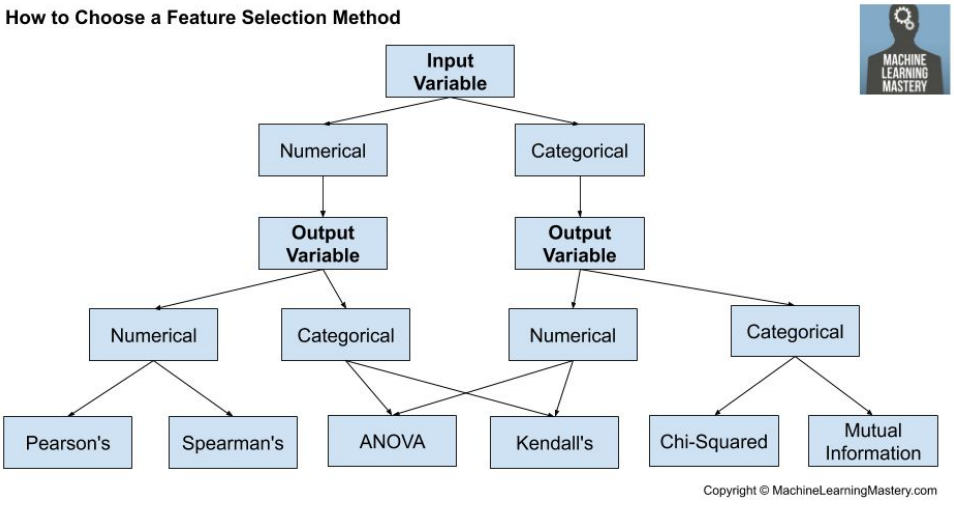
FEATURES SELECTION

Motivación:

1. Reducir coste computacional
2. Reducir overfitting

Tipos:

1. **Supervisado**: Buscamos las variables que tengan una mayor relación con la variable objetivo.
    1. **Wrapper**
        1. Creamos diferentes subconjuntos de variables, evaluamos el comportamiento de cada uno de los subconjuntos y nos quedamos con el subconjunto de variables que mejor resultado haya dado. El problema de este tipo de técnicas es que suelen ser computacionalmente costosas.
    2. **Filter**
        1. Este tipo de técnicas se basan en evaluar la relación de cada variable con la target y quedarnos únicamente con las que superen el umbral que hayamos decidido (Por ejemplo, las variables que tengan más de un 0.8 de correlación con la target). Dependiendo del tipo de variable debemos utilizar distintos tipos de estadísticos.



    3. **Intrinsic**
        1. Hay algoritmos que utilizan únicamente las variables que mejoran la precisión de sus predicciones por lo que la selección de variables la realizan de manera interna. Ejemplos de esto podrían ser Lasso o los arboles de decisión.
            1. Sklearn nos permite hacer uso de esa característica para filtrar columnas atraves de **SelectFromModel**
2. **No supervisado**: No utilizan (o no tienen acceso) a la variable objetivo. Suelen tener el objetivo de eliminar variables redundantes mediante correlación u otro tipo de técnica. Si dos o más variables tienen una correlación muy alta la información que

aportan es redundante y queremos evitar eso por lo que normalmente nos quedaremos con una única variable de esas.

    1. Filtro de varianza: Si nuestras columnas tienen una varianza baja significa que los valores son "parecidos" por lo que seguramente nos aportará poca información. Y podremos filtrar las columnas que menos varianza tengan (**from sklearn.feature_selection import VarianceThreshold**)

3. **Reducción de dimensionalidad**: A diferencia de la selección de variables la reducción de la dimensionalidad no se queda con los datos originales sino con una proyección, el algoritmo más habitual es el PCA