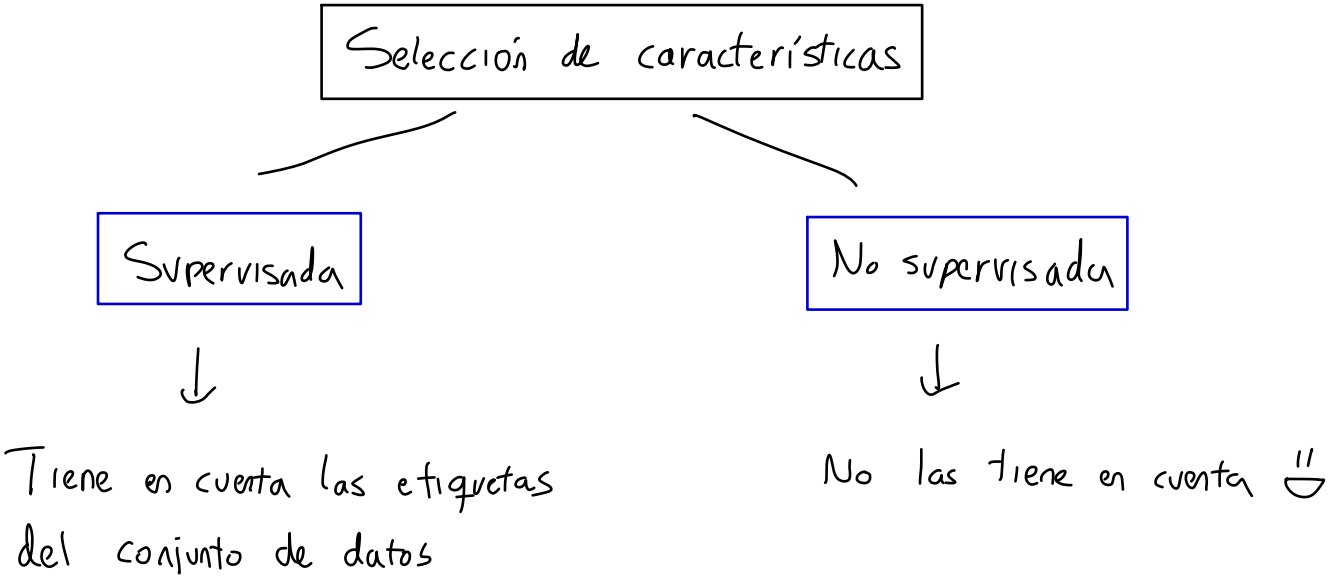
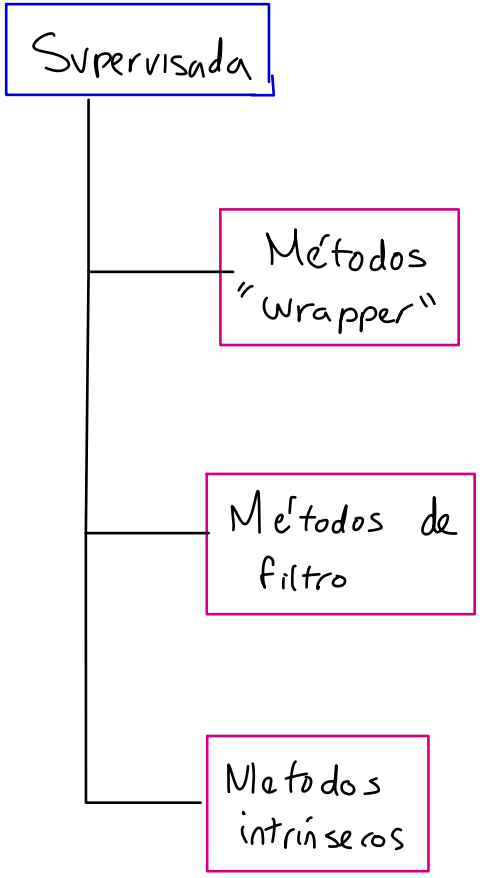
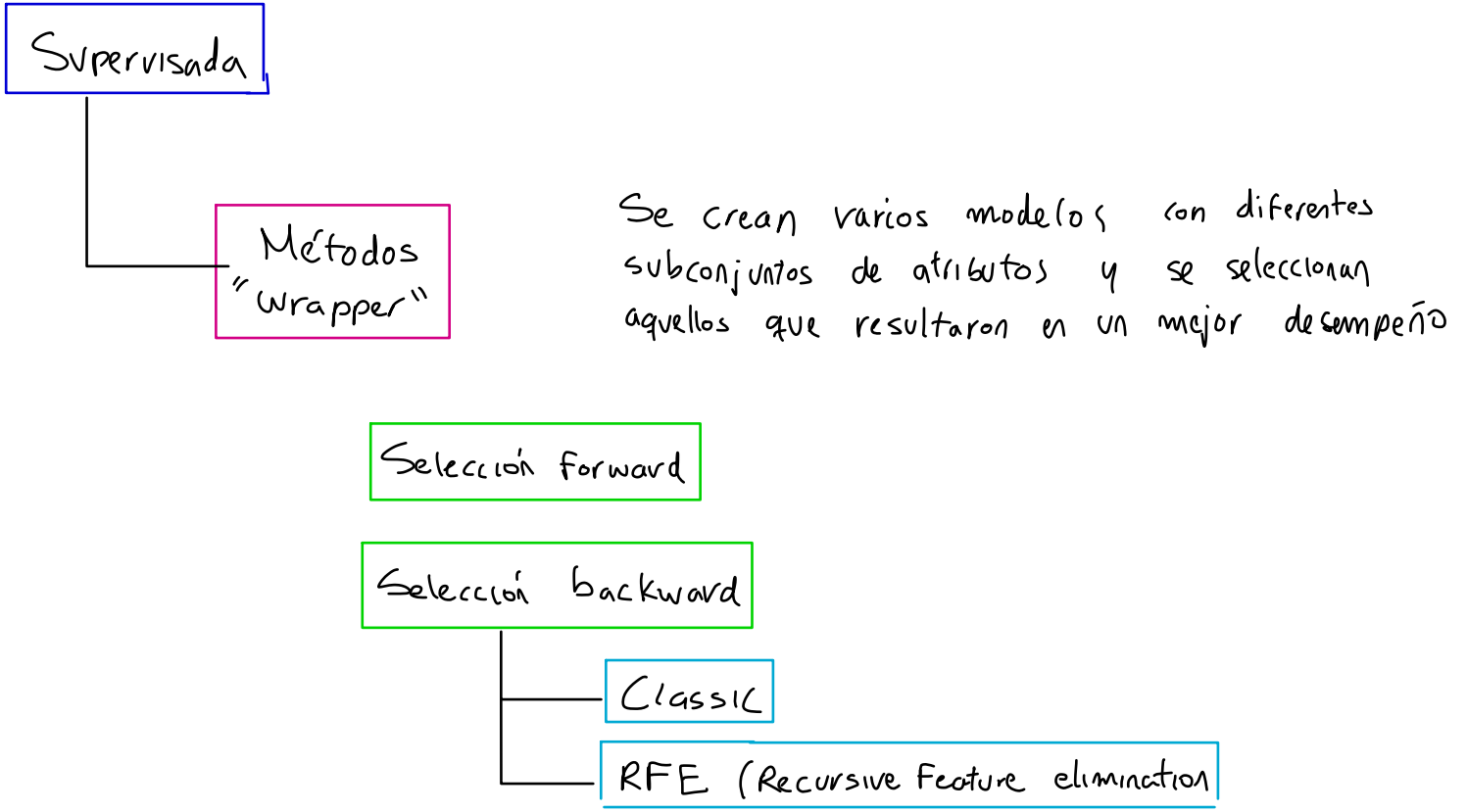
Selección de características

Metodos para restringir el conjunto de atributos con el cuál se entrenará un modelo de Machine Learning con diferentes fines como

- ☐ Reducir dimensionalidad
- ☐ Evitar información redundante, irrelevante o que "engañe" al modelo
- ☐ Disminuir el tiempo de entrenamiento


Selección de características
Supervisada
No supervisada
Tiene en cuenta las etiquetas del conjunto de datos
No las tiene en cuenta


Supervisada
Métodos "wrapper"
Métodos de filtro
Métodos intrínsecos


Supervisada
Métodos "wrapper"
Se crean varios modelos con diferentes subconjuntos de atributos y se seleccionan aquellos que resultaron en un mejor desempeño
Selección forward
Selección backward
Classic
RFE (Recursive Feature elimination

Selección forward

Inicia con un conjunto vacío de atributos y va añadiendo atributos nuevos de forma iterativa

En cada iteración prueba entrenar el modelo con cada uno de los atributos faltantes y se queda con el que se desempeñe mejor

esto ocurre hasta que se cumpla un criterio de mejoramiento en el desempeño del modelo

Sklearn: SequentialFeatureSelector

Selección forward

Sklearn

SequentialFeatureSelector

```
SequentialFeatureSelector ( model,
                            k_features = 5,
                            tol = incremento mínimo del score para continuar
                            direction = forward
                            scoring = 'accuracy',
                            cv = 5 )
```

Selección backward

- Classic
- RFE (Recursive Feature elimination)

Classic Igual que el forward pero inicia con todos los atributos y va quitando uno a uno

RFE (Recursive Feature elimination ⟶

RFE (Recursive Feature elimination)

Evita entrenar muchos modelos en cada paso de la búsqueda.

Se le da una medida de importancia a las variables. Esto las organiza en un ranking

Top 1	X_{i1}
2	X_{i2}
⋮	⋮
k	X_{ik}

En cada paso de la iteración, los atributos menos importantes se eliminan antes de reconstruir el modelo

RFE (Recursive Feature elimination)

1 Tune/train the model on the training set using all P predictors
2 Calculate model performance
3 Calculate variable importance or rankings
4 **for** *each subset size* S_i, $i = 1 \ldots S$ **do**
5 | Keep the S_i most important variables
6 | [Optional] Pre-process the data
7 | Tune/train the model on the training set using S_i predictors
8 | Calculate model performance
9 | [Optional] Recalculate the rankings for each predictor
10 **end**
11 Calculate the performance profile over the S_i
12 Determine the appropriate number of predictors (i.e. the S_i associated with the best performance)
13 Fit the final model based on the optimal S_i

Supervisada

- Métodos de filtro

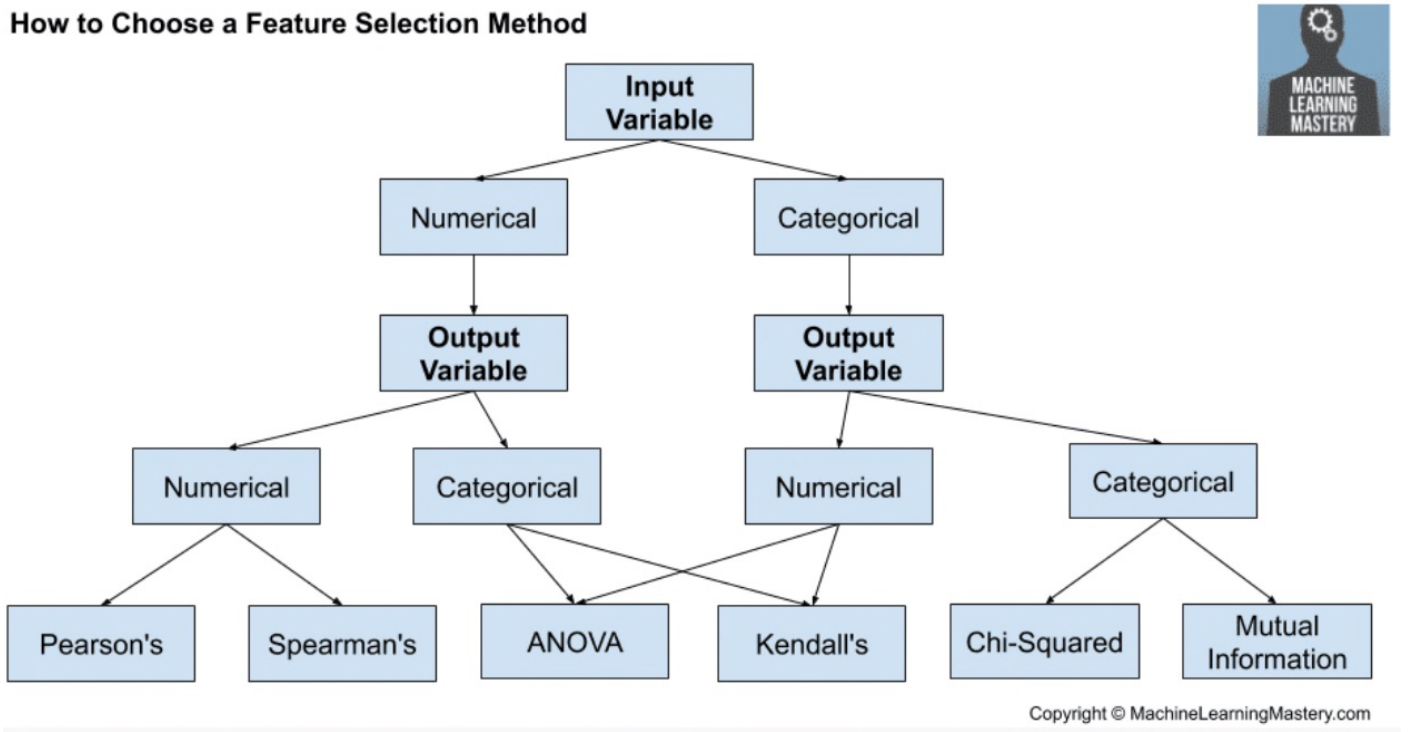
como la correlación

Se utilizan medidas estadística para evaluar la relación entre cada atributo y la variable de salida.

Se mide con un atributo a la vez

Sólo los mejores atributos de acuerdo a un criterio determinado se utilizarán para entrenar el modelo

Con cuál medida estadística medir la relacion de las variables con la variable objetivo?



Pearson

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{Cov(X,Y)}{\sqrt{Var(X)\,Var(Y)}}$$

$$Cov(X,Y) = \frac{1}{n} \sum_{i=1}^{n} (x_i - E[X])(y_i - E(y))$$

Spearman

$$\rho_{R(X),R(Y)} = \frac{Cov(R(X), R(Y))}{\sigma_{R(X)}\,\sigma_{R(Y)}}$$

R es un ranking entre las variables

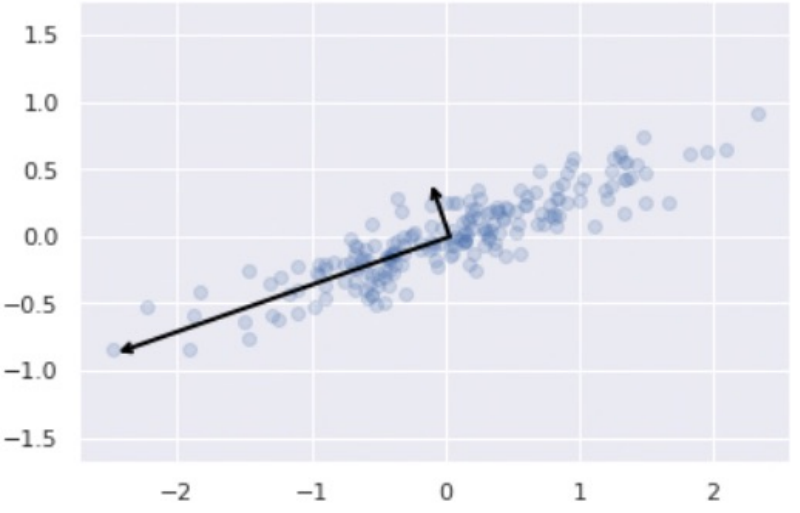
Supervisada

- Métodos intrínsecos

Son aquellos modelos de Machine Learning que durante el entrenamiento seleccionan atributos de manera "natural".

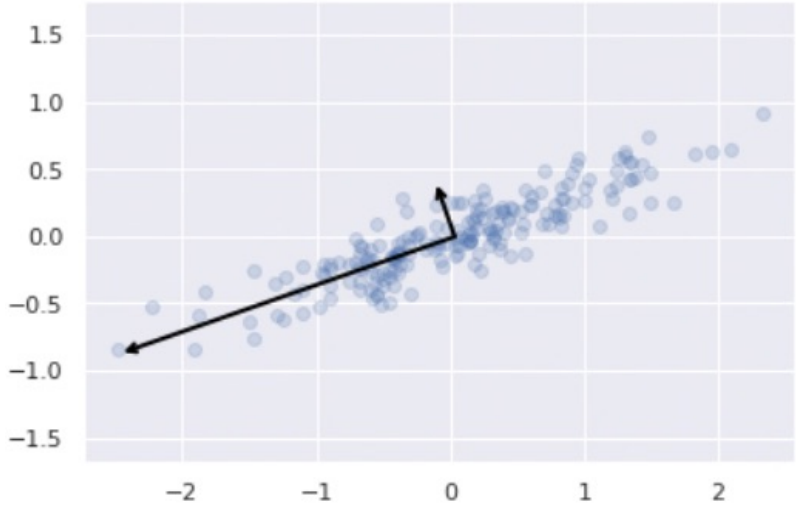
Ej: - Árboles de decisión
- LASSO

Principal Component Analysis



No supervisada

Principal Component Analysis



Dirección de máxima varianza

Nuevas coordenadas

PCA permite:

Reducir la dimensionalidad de los datos

Cuando hay muchas variables correlacionadas, se reduce a un numero bajo de variables independientes

¿Cómo se calculan las componentes?

- Se calcula el promedio de las filas

$$\bar{X} = \frac{1}{n} \sum_{j=1}^{n} X_j$$

X matriz de características

$$X = \begin{bmatrix} — & x_1 & — \\ & \vdots & \\ — & x_m & — \end{bmatrix}$$

Cómo se calculan las componentes?

X matriz de características

- Se calcula el promedio de las filas

$$\bar{X} = \frac{1}{n} \sum_{j=1}^{n} X_j$$

$$X = \begin{bmatrix} — & X_1 & — \\ & \vdots & \\ — & X_m & — \end{bmatrix}$$

- Se resta la media $\bar{X}$ a cada fila

$$B = X - \bar{X}$$ [pequeño abuso de notación]

¿Cómo se calculan las componentes?

X matriz de características

- Se calcula el promedio de las filas

$$\bar{X} = \frac{1}{n} \sum_{j=1}^{n} X_j$$

$$X = \begin{bmatrix} — & x_1 & — \\ & \vdots & \\ — & x_m & — \end{bmatrix}$$

- Se resta la media $\bar{X}$ a cada fila

$$B = X - \bar{X} \qquad [\text{pequeño abuso de notación}]$$

- Consideremos la matriz de covarianza de las filas de B

$$C = B^T B$$

¿Cómo se calculan las componentes?

X matriz de características

$$X = \begin{bmatrix} — & x_1 & — \\ & \vdots & \\ — & x_m & — \end{bmatrix}$$

- Se calcula el promedio de las filas

$$\bar{x} = \frac{1}{n} \sum_{j=1}^{n} x_j$$

- Se resta la media $\bar{x}$ a cada fila

$$B = X - \bar{x} \quad \text{[pequeño abuso de notación]}$$

- Consideremos la matriz de covarianza de las filas de B

$$C = B^T B$$

- Se calculan los vectores / valores propios.

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$$

$$C = V D V^{-1}$$

$$\begin{bmatrix} | & & | \\ e_1 & \cdots & e_n \\ | & & | \end{bmatrix} \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix} \begin{bmatrix} & V^{-1} & \end{bmatrix}$$

V

El vector propio e_1 asociado al mayor valor propio indica la dirección de la primera componente, de la misma forma los demás.