TALLER DE MODELADO DE DATOS

Sara Eugenia Rodríguez Reyes

SELECCIÓN DE VARIABLES

Razones Fundamentales

Dimensionalidad y Complejidad

Reducción de Ruido y Mejora de Precisión

Evitar el overfitting (sobre-ajuste)

Interpretabilidad del modelo

Eficiencia Computacional

Selección de variables ≠ Reducción Dimensionalidad

Selección de variables

Se queda con un subconjunto de las variables originales.

Reducción Dimensionalidad

- Crea una proyección de datos que da como resultado características completamente nuevas.
- Transforma todas las variables en un nuevo espacio de menor dimensión.

Cuándo uso cada uno?

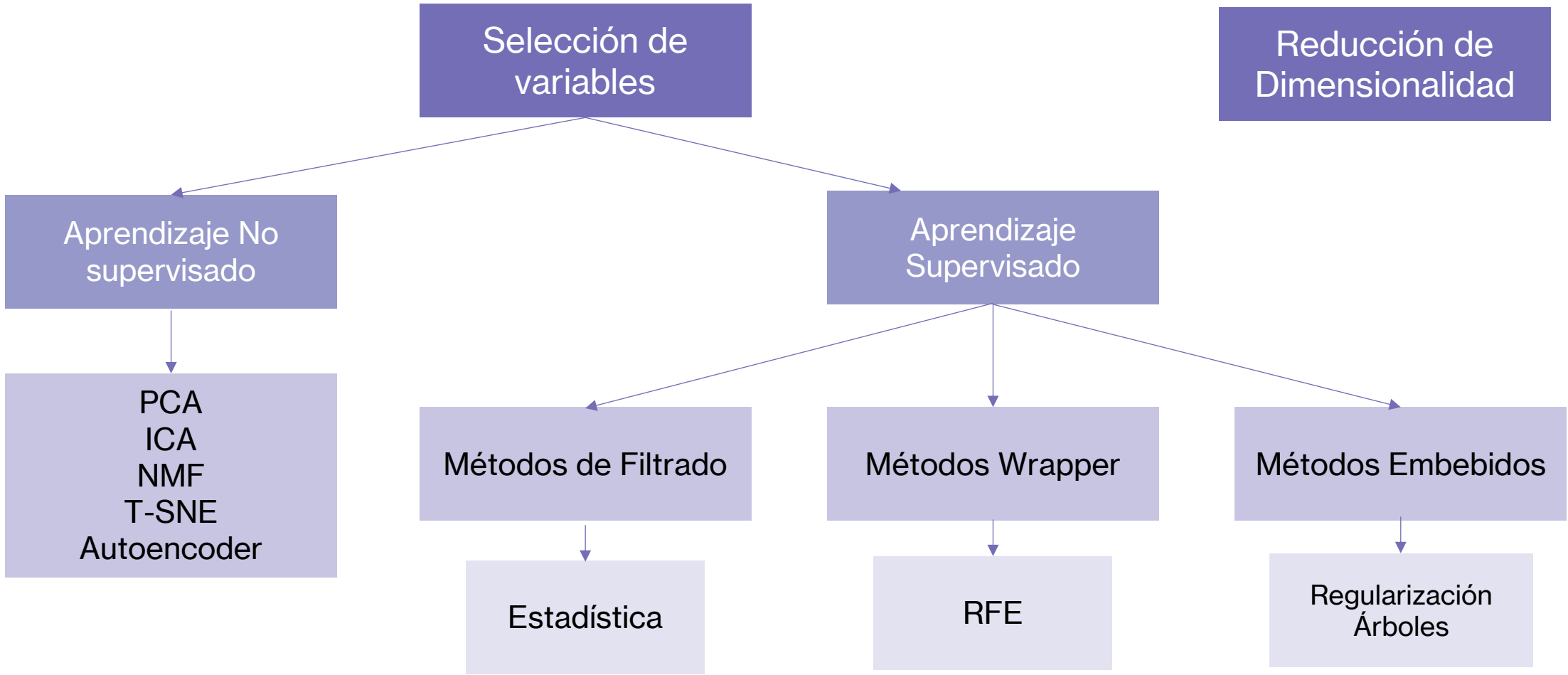
Selección de variables

- Cuando necesitas **interpretabilidad**.
- Cuando sospechas que hay muchas **variables ruido** (que no aportan).
- Cuando las variables ya tienen **sentido propio** (edad, ingresos, etc.) y es valioso entender cuáles importan.

Reducción Dimensionalidad

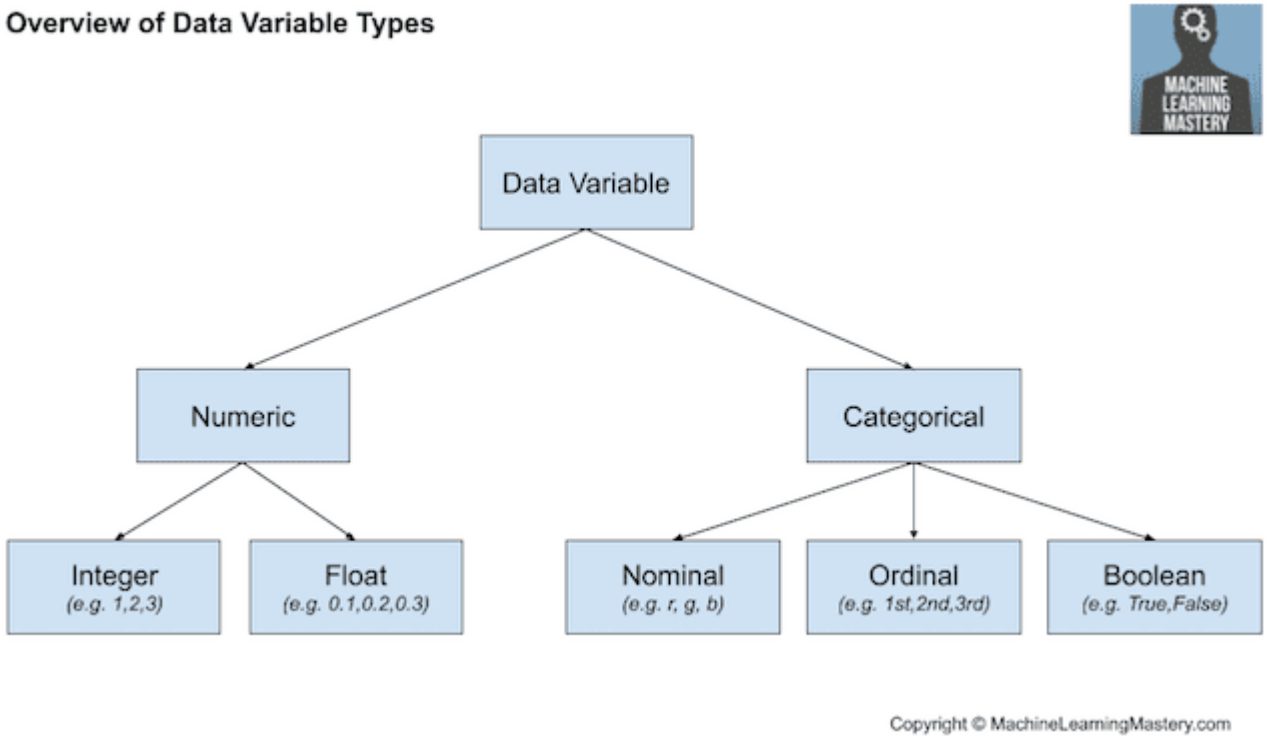
- Cuando las variables originales son **muchísimas** y altamente correlacionadas.
- Cuando la **interpretabilidad no es crítica** (ej. reconocimiento de imágenes, texto, señales biomédicas)

Técnicas de Selección de Variables

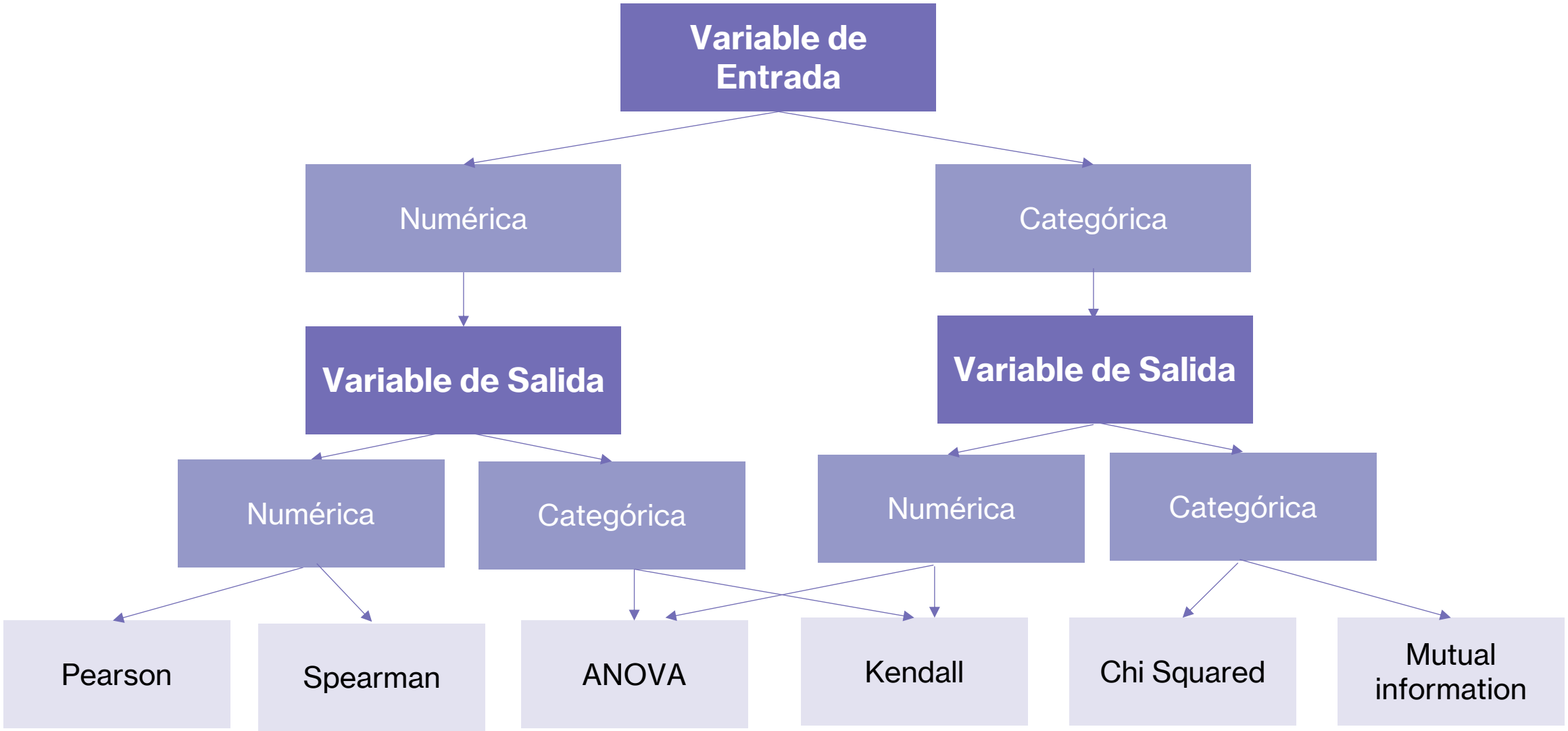


1. Métodos de Filtrado

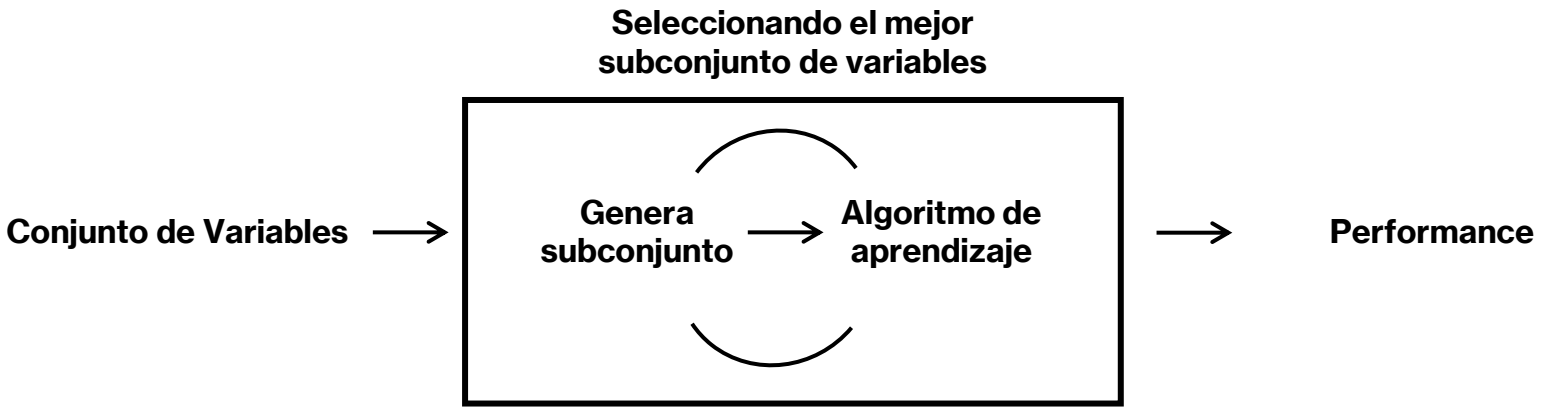
- Es muy común utilizar medidas estadísticas de tipo “correlación” entre variables de entrada y salida como base para la selección de variables.
- **La elección de medidas estadísticas depende en gran medida de los tipos de datos**



Cómo seleccionar el método



2. Métodos Wrapper



- Tratan a la selección de variables como un problema de búsqueda.
- Evalúan diferentes subconjuntos de variables y miden su impacto en el performance del modelo de machine learning específico.

Estrategias comunes:

- **Forward Selection**: Se empieza con un conjunto de características vacío y se le van añadiendo características iterativamente.
- **Backward Selection:** Se empieza con todo el conjunto de características y se van eliminando gradualmente.
- **Recursive Feature Elimination**: Se empieza con todo el conjunto de características y se eliminan características repetidamente según su relevancia, según lo determine el algoritmo de aprendizaje.

3. Métodos Embebidos

- Combinan las ventajas de los métodos de filtrado y los métodos de envoltura
- Método de selección de variables que ocurre **dentro del entrenamiento del modelo.**
- Son **métodos específicos de cada modelo.**
  - Ejemplos:
  1. Lasso
  2. Métodos basados en árboles (Random forest, Xgboost, Decision Tree)

LASSO

- Es un método de reducción que realiza tanto la selección de variables como la regularización al mismo tiempo.

- La regularización es un proceso que reduce los coeficientes (pesos) hacia cero. Significa que estás penalizando modelos más complejos para evitar el sobreajuste.

- LASSO permite que los coeficientes se establezcan en 0. Si un coeficiente es cero, entonces la característica no se toma en cuenta y, por lo tanto, en cierto modo se descarta.

$$L_{lasso}(\hat{\beta}) = \sum_{i=1}^{n}(y_i - x_i'\hat{\beta})^2 + \lambda \sum_{j=1}^{m} |\hat{\beta}_j|.$$

Feature Importance

- Los métodos más conocidos son los basados en árboles.
- La mayoría utiliza el Gini index (reducción promedio de la impureza) para estimar la importancia de cada variable, mientras más pequeño el valor, más importante es la variable.

$$Gini = 1 - \sum_{i=1}^{n} (P_i)^2$$

¿Cuál es el mejor método?

- No hay.... Pero hay algunos tips que pueden ayudar

Método	Enfoque	Utiliza un modelo?	Pro	Cons
Filtrado	Cuando estás en el pre-procesamiento o tienes gran cantidad de variables	No	Rápido, no es específico a un modelo	Ignora interacción entre variables
Wrapper	Cuando quieres mayor precision y tienes tiempo de entrenar	Si	Considera interacción de variables	Computacionalmente caro
Forward		Si	Simple, para datasets chicos	Puede ignorar combos interesantes
Backward		Si	Considera todas las variables (para modelos lineales)	Lento, sensible a multicolinealidad
RFE		Si	Funciona con cualquier modelo	Debe re-entrenar cada modelo
Embebido	Durante el entrenamiento de un modelo específico	Si	Eficiente, integrado	Específico a cada modelo

Ejercicio en python