



Curso especialización en inteligencia artificial y big data

Tratamiento de características

Profesor: Sebastián Rubio Valero

Noviembre 2025



En muchas actividades prácticas de Ciencia de Datos, el conjunto de datos contendrá variables categóricas. Estas variables se suelen almacenar como valores de texto que representan varios rasgos. Algunos ejemplos incluyen el color ("Rojo", "Amarillo", "Azul"), el tamaño ("Pequeño", "Mediano", "Grande") o designaciones geográficas (Estado o País). Independientemente de para qué se utilice el valor, el desafío es determinar cómo usar estos datos en el análisis

Muchos algoritmos de aprendizaje automático (machine learning) pueden admitir valores categóricos sin más manipulación, pero hay muchos más algoritmos que no lo hacen. Por lo tanto, el analista se enfrenta al desafío de descubrir cómo convertir estos atributos de texto en valores numéricos para su posterior procesamiento.

Afortunadamente, las herramientas de Python como pandas y scikit-learn proporcionan varios enfoques que pueden aplicarse para transformar los datos categóricos en valores numéricos adecuados. Este artículo será un estudio de algunas de las técnicas más comunes (y algunas más complejas), con la esperanza de que ayude a otros a aplicar estas técnicas a sus problemas del mundo real.

Las características que nos vamos a encontrar en nuestros conjuntos de datos los podemos clasificar en:

- Características categóricas: Pueden ser nominales (sin orden inherentes) y ordinales (tienen un orden por ejemplo las tallas)
- Características numéricas: Pueden ser discretas o continuas
- Temporales
- De texto
- Geográficas

No todas las características añaden información, para detectar estas características hay que hacer un primer análisis de datos y dependiendo del tipo genera unos problemas.

El proceso de selección de características racionaliza un modelo identificando las más importantes, impactantes y no redundantes del conjunto de datos. Al reducir el número de características, se mejora la eficacia del modelo y su rendimiento.

Los beneficios de la selección de características incluyen:

- La precisión del modelo
- La interpretabilidad

- El rendimiento computacional
- Menores costes
- Reducir el overfitting
- Implementación más fluida
- Reducción de la dimensionalidad

1 Métodos de selección en modelos supervisados

La selección de características del aprendizaje supervisado utiliza la variable objetivo para determinar cuáles son las más importantes. Dado que las características de los datos ya están identificadas, la tarea consiste en identificar qué variables de entrada tienen un impacto más directo en la variable objetivo. La correlación es el criterio principal para evaluar las características más importantes.

Entre los métodos supervisados de selección de características se incluyen:

- Métodos de filtrado
- Métodos de envoltura
- Métodos incrustados

También es posible utilizar métodos híbridos que combinen dos o más métodos supervisados de selección de características.

1.1 Métodos de filtrado

Los métodos de filtrado son un grupo de técnicas de selección de características que se ocupan únicamente de los datos y no tienen en cuenta la optimización del rendimiento del modelo. Las variables de entrada se evalúan de forma independiente con respecto a la variable objetivo para determinar cuál tiene la mayor correlación. Los métodos que evalúan una característica cada vez se conocen como métodos de selección de características univariantes.

Los métodos de filtrado más comunes son:

- **Ganancia de información:** mide lo importante que es la presencia o ausencia de una característica para determinar la variable objetivo mediante la reducción de entropía.

- **Información mutua:** evalúa la dependencia entre variables al medir la información obtenida sobre una variable a partir de la otra.
- **Prueba de chi-cuadrado:** evalúa la relación entre dos variables categóricas comparando los valores observados con los esperados.
- **Puntuación de Fisher:** utiliza derivadas para calcular la importancia relativa de cada característica a la hora de clasificar los datos. Cuanto mayor sea la puntuación, mayor será su influencia.
- **Coefficiente de correlación de Pearson:** cuantifica la relación entre dos variables continuas con una puntuación que oscila entre -1 y 1.
- **Umbral de varianza:** elimina todas las características que están por debajo de un umbral mínimo de varianza, ya que es probable que las características con mayor varianza contengan más información útil. Un método relacionado es la diferencia media absoluta (DMA).
- **Proporción de valores perdidos:** calcula los porcentajes de casos de un conjunto de datos en los que falta una característica o tiene un valor nulo. Si falta una instancia de una característica, es probable que no sea útil.
- **Coefficiente de dispersión:** relación entre la varianza y el valor medio de una característica. Cuanto mayor sea la dispersión, más información habrá.
- **ANOVA (análisis de varianza):** determina si los distintos valores de las características afectan al valor de la variable objetivo.

<https://www.ibm.com/es-es/think/topics/feature-selection>

1.2 Métodos de envoltura (wrappers)

Los métodos envoltentes entrenan el algoritmo de machine learning con varios subconjuntos de características, y añaden o eliminan características y comprueban los resultados en cada iteración. El objetivo de todos ellos es encontrar el conjunto de características que permita alcanzar el rendimiento óptimo del modelo.

Los métodos envoltentes que prueban todas las combinaciones posibles de características se conocen como algoritmos codiciosos. Dado que su búsqueda del mejor conjunto de características requiere muchos recursos informáticos y tiempo, son más adecuados para conjuntos de datos con pocos espacios de características.

Los científicos de datos pueden configurar el algoritmo para que se detenga cuando disminuya el rendimiento del modelo o cuando se alcance un número determinado de características.

Los métodos de envoltura incluyen:

- **Selección hacia adelante:** se comienza con un conjunto de características vacío y se van añadiendo nuevas características gradualmente hasta encontrar el conjunto óptimo. La selección del modelo se lleva a cabo cuando el rendimiento del algoritmo no mejora tras una iteración específica.
- **Selección regresiva:** entrena un modelo con todas las características originales y elimina iterativamente la característica menos importante del conjunto.
- **Selección exhaustiva de características:** se prueban todas las combinaciones posibles de características para encontrar la mejor en general y optimizar una métrica de rendimiento especificada. Un modelo de regresión logística que utiliza la selección exhaustiva de características prueba todas las combinaciones posibles de todas las características.
- **Eliminación recursiva de características (RFE):** un tipo de selección regresiva que comienza con un espacio de características inicial y elimina o añade características después de cada iteración en función de su importancia relativa.
- **Eliminación recursiva de características con validación cruzada:** una variante de la eliminación recursiva que utiliza la validación cruzada para probar un modelo con datos no vistos y seleccionar el conjunto de características de mejor rendimiento. La validación cruzada es una técnica habitual para evaluar modelos de lenguaje de gran tamaño (LLM).

1.3 Métodos embebidos

Los métodos incrustados incluyen la selección de características en el proceso de entrenamiento del modelo. A medida que el modelo se entrena, utiliza diversos mecanismos para detectar las características de bajo rendimiento y descartarlas en futuras iteraciones.

Muchos métodos incrustados se centran en la regularización, que penaliza las características en función de un umbral de coeficiente preestablecido. Los modelos cambian cierto grado de exactitud por una mayor precisión. El resultado es que los modelos rinden un poco menos durante el entrenamiento, pero se vuelven más generalizables

al reducir el sobreajuste.

Entre los métodos incrustados se incluyen:

- **Regresión LASSO (regresión L1):** añade una penalización a la función de pérdida para coeficientes correlacionados de alto valor, moviéndolos hacia un valor de 0, que se traduce en su eliminación. Cuanto mayor es la penalización, más características se eliminan del espacio de características. El uso eficaz del LASSO consiste en equilibrar la penalización para eliminar suficientes características irrelevantes sin dejar de lado las importantes.
- **Importancia del bosque aleatorio:** genera cientos de árboles de decisión, cada uno con una selección aleatoria de puntos de datos y características. Se evalúa cada árbol en función de la capacidad de división de los puntos de datos que muestra. Cuanto mejores sean los resultados, más importante se considerará la característica o características de ese árbol. Los clasificadores miden la "impureza" de las agrupaciones mediante la impureza de Gini o la ganancia de información, mientras que los modelos de regresión utilizan la varianza.
- **Aumento de gradiente:** añade predictores en secuencia a un conjunto y cada iteración corrige los errores de la anterior. De este modo, puede identificar qué características conducen a resultados óptimos.

2 Métodos de selección en modelos no supervisados

En el aprendizaje no supervisado, los modelos descubren por sí solos las características, los patrones y las relaciones de los datos. No es posible adaptar las variables de entrada a una variable objetivo conocida. Los métodos de selección de características sin supervisión utilizan otras técnicas para simplificar y racionalizar el espacio de características.

Un método de selección de características sin supervisión es el análisis de componentes principales (PCA, por sus siglas en inglés). El PCA reduce la dimensionalidad de los grandes conjuntos de datos transformando las variables potencialmente correlacionadas en un conjunto más pequeño de variables. Estos componentes principales conservan la mayor parte de la información del conjunto de datos original. El PCA contrarresta la "maldición de la dimensionalidad" y también reduce el sobreajuste.

Otros incluyen el análisis de componentes independientes (ICA), que separa los datos

multivariantes en componentes individuales que son estadísticamente independientes, y autocodificadores.

Los autocodificadores se utilizan mucho en arquitecturas de transformadores y son un tipo de red neuronal que aprende a comprimir y luego reconstruir datos. Al hacerlo, los autocodificadores descubren variables latentes, es decir, aquellas que no son directamente observables, pero que afectan en gran medida a la distribución de los datos.

3 Eliminación de características por filtrado

Los métodos de filtro se utilizan generalmente como un paso de preprocesamiento de datos, la selección de características es independiente de cualquier algoritmo de Machine Learning.

Las características se clasifican según ciertas métricas estadísticas, que son independientes del modelo de machine learning que se vaya a utilizar posteriormente

Estos métodos evalúan la relevancia de cada característica (o atributo) basándose en su relación con la variable objetivo (la que se quiere predecir) o su relación con otras características, usando una medida estadística.

- Cálculo de una puntuación: Se elige una medida estadística (como correlación, ganancia de información, o la prueba χ^2) y se calcula una puntuación para cada característica.
- Clasificación y Selección: Las características se clasifican según estas puntuaciones. Solo aquellas que superan un cierto umbral o el grupo con las mejores puntuaciones se seleccionan para el entrenamiento del modelo.
- Filtrado de Redundancia: El objetivo es eliminar las características irrelevantes o redundantes (que proporcionan poca o ninguna información nueva).

Las métricas que se suelen usar son:

- Coeficiente de correlación de Pearson: Mide la relación entre una característica numérica y la variable objetivo numérica
- χ^2 (Chi-cuadrado): Se usa para evaluar la dependencia entre dos variables categóricas (como una característica categórica y la variable objetivo).

- Ganancia de Información/Entropía. Mide la reducción de la incertidumbre (entropía) sobre la variable objetivo que se obtiene al conocer el valor de una característica

Estos métodos se suelen utilizar en los siguientes casos:

- Cuando tenemos un número muy grande de características ≥ 50
- Como un primer paso para aplicar otros métodos de selección de características

A continuación se estudian algunos de los métodos más usados

3.1 Coeficiente de Correlación de Pearson (r)

El Coeficiente de Correlación de Pearson es una medida estadística que cuantifica la fuerza y la dirección de la relación lineal entre dos variables cuantitativas (numéricas).

En el contexto de la selección de características por filtrado, se calcula la correlación entre cada característica numérica y la variable objetivo numérica.

Determinar qué características afectan a la selección de una clase. En estos casos vamos a usar el coeficiente de correlación de Pearson, que mide el grado de relación entre dos variables. Este coeficiente varía entre $[-1, 1]$:

- si $r = 1$ la correlación positiva perfecta La característica es altamente relevante. Un aumento en ella se asocia con un aumento proporcional en el objetivo
- si $0 < r < 1$ la correlación positiva. Relevancia alta
- si $r = 0$ no existe relación lineal
- si $-1 < r < 0$ la correlación es negativa Relevancia alta
- si $r = -1$ la correlación negativa perfecta

Las características optimas son aquellas cuya correlación son las más altas, cuanto más cerca de uno mejor

La única característica que nos va predecir mejor la clase es la que tiene la mayor correlación. Para la eliminación de características que pueden desvirtuar la selección de la clase, serán aquellas que están altamente correlacionadas. Una vez que hemos seleccionado las características que están altamente correlacionadas, ¿cual de las dos eliminamos?, eliminamos la que menos correlación tenga con las demás. ¿Qué otras dos características puedo eliminar según mi diagrama de correlación?, aquellas tengan la menor correlación con el atributo clase.

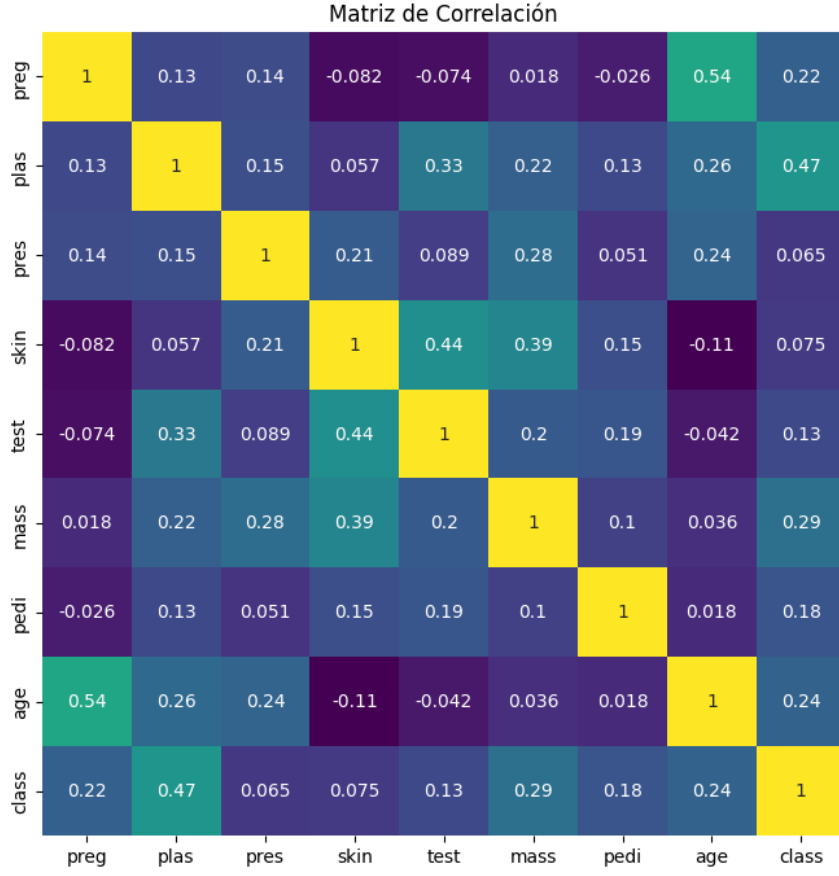


Figure 1: Matriz de Correlación

La principal desventaja es que estos métodos evalúan las características de forma individual o por pares y no consideran las interacciones complejas o sinérgicas entre las diferentes características. Esto puede llevar a la omisión de un subconjunto de características que, aunque individualmente sean débiles, juntas podrían ser muy predictivas.

La fórmula para el cálculo se basa en la Covarianza de las dos variables, normalizado por el producto de sus desviaciones estándar. Esto asegura que el valor de r sea independiente de las unidades de medida.

Para una muestra de N pares de datos (x_i, y_i) , la fórmula es

$$r = \frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^N (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^N (y_i - \bar{y})^2}} \quad (1)$$

La aplicación en la selección de características se utiliza de dos formas:

- La correlación entre una característica y la variable objetivo. Se calcula r entre cada características y la variable objetivo. Las características con valor $|r|$ más cercano a uno se consideran las más relevantes
- La correlación entre características
 - Se calcula r entre todos los pares de características
 - Si dos características están altamente correlacionadas significa que son redundantes
 - En este caso, se suele eliminar una de ellas

El filtrado por Varianza, se basa en la suposición: si una característica no cambia mucho de un ejemplo a otro(es decir tiene una varianza baja), es poco probable que sea útil para distinguir entre las diferentes clases o para modelar la variable objetivo.

- Calcula varianza de cada característica en el total del conjunto de datos
- Se estable un umbral mínimo
- Se eliminan todas las características cuya varianza caiga por debajo de ese umbral

Si la varianza es cero o cerca de cero entonces esa característica se trata de una constante

4 Métodos wrappers

Debido al alto costo computacional, en Big Data con un número muy grande de características, los métodos wrapper a menudo se aplican después de una etapa inicial de filtrado (métodos filter) que reduce el conjunto de datos a un número más manejable de variables.

Estos método necesitan de un algoritmo de Machine Learning y utiliza su rendimiento como criterio de evaluación. Estos métodos busca características que sea más adecuada para el algoritmo y tiene como objetivo mejorar el rendimiento.

Usaremos esto métodos cuando

- Cuando tratamos con modelos específicos de optimización
- Cuando el conjunto de datos es pequeño o medio

4.1 Eliminación Forward

Es una técnica secuencial y greedy (ambicioso) de selección de características. Comienza sin ninguna característica y va añadiendo una a una las que más mejoren el modelo, hasta que agregar nuevas variables ya no produzcan una mejora significativa

- Partimos de un conjunto de características $F = \{\}$ vacío
- Entrenamos nuestro modelo con todas las características de forma individual f_1, f_2, \dots, f_n , y nos quedamos con aquella que produce el mejor rendimiento
- Añadimos esta variable al modelo
- Ahora con las variables restantes, evaluamos nuestro modelo tomando la seleccionada anterior más una nueva del resto, y aquella que dé mejor rendimiento con la seleccionada anterior forma el nuevo conjunto
- El proceso continúa hasta que ninguna variable mejore al conjunto anterior o bien si se alcanza un número máximo de características

Ejemplo

Sean las características $F = \{x_1, x_2, x_3, x_4\}$

1. Inicio: Modelo vacío

2. Paso 1: Pruebas con cada variable sola, usamos la métrica R^2 :

- $x_1 = 0.60$
- $x_2 = 0.45$
- $x_3 = 0.30$
- $x_4 = 0.50$

Elegimos x_1 (el mejor resultado)

3. Paso 2: Probamos añadir otra característica a x_1

- $x_1 + x_2 = 0.68$
- $x_1 + x_3 = 0.63$
- $x_1 + x_4 = 0.70$

Elegimos x_4 (mayor mejora)

4. Paso 3: Intenta añadir una tercera

- $x_1 + x_4 + x_2 = 0.71$ (poca mejora)
- $x_1 + x_4 + x_3 = 0.705$ (menor mejora)

La mejora es mínima, así que se detiene el proceso

4.2 Eliminación Backward

En este método inicialmente partimos con todas las características y evaluamos el modelo, a continuación se quita una característica y se vuelve a evaluar, si el resultado de evaluarlo es mayor entonces eliminamos la característica. Se realiza este procedimiento de forma iterativa, es decir vamos eliminando características, de tal forma que al evaluar el modelo me dé un valor mayor de rendimiento, es decir si al quitar una característica obtengo mejor rendimiento, pues entonces la quito, hasta que me quede con k características, es el valor que selecciona el usuario.

La métrica de rendimiento utilizada es el p -value. Consiste en eliminar aquellas características cuyo valor p está por encima de 0.05, de lo contrario los conservamos.

- Se entrena un modelo con todas las características
- Se analiza el impacto o relevancia de cada variable (p.e, su significancia estadística, su contribución R^2 o su efecto sobre el error)
- Se elimina la variable menos relevante, aquella que reduce menos la calidad del modelo
- Se vuelve a ajustar el modelo con las variables restantes
- El proceso se repite - en cada paso se elimina una variable- hasta que se alcanza un número de características o al quitar más variables se degrada el rendimiento

Ejemplo

Sean las características $F = \{x_1, x_2, x_3, x_4\}$

1. Inicio: Modelo con todas las características, supón que $R^2 = 0.72$
2. Paso 1: Analizamos qué variable tiene menor impacto o mayor p -valor.

Descubres que x_3 no contribuye mucho. La eliminas y el nuevo modelo tiene $R^2 = 0.71$ (pequeña pérdida, aceptable)

3. Paso 2: Revisamos el conjunto restante $\{x_1, x_2, x_4\}$, ahora nos fijamos que x_2 aporta muy poco y su eliminación aumenta $R^3 = 0.73$ (mejor rendimiento). Eliminamos x_2
4. Paso 3: Quedan x_1 y x_4 , si eliminamos alguna de ellas, el rendimiento baja. Por lo tanto se detiene el proceso

Los dos modelos anteriores se suelen usar cuando el número de características es menor de 50

4.3 Eliminación recursiva de características

Funciona eliminando recursivamente los atributos y construyendo un modelo sobre los atributos que quedan, es decir realiza combinaciones entre los atributos y se queda con aquella que mejor resultados se obtiene. Empieza con todas las características, y en cada iteración ajusta un modelo (p.e regresión lineal, SVM, árbol, etc), mide la importancia de cada variable según el modelo y quita la(s) menos importantes.

1. Elegir el modelo base (estimador): p.e regresión lineal, random forest, SVM
2. Decidir criterio de importancia según el modelo:
 - Regresión lineal: magnitud absoluta del coeficiente
 - Árboles/RandomFores: Importancia basada en reducción de impurity o ganancia
 - SVM: magnitud del peso
3. Decidir el número a eliminar por iteración
4. Decidir el criterio de parada:
 - Quedar con k variables predefinidas
 - Parar cuando la métrica de validación empeora
 - Parar según el umbral de importancia
5. Iterar
 - Ajustar el modelo con las características actuales.

- Calcular importancias
 - Eliminar la(s) variable(s) menos importantes
 - (Opcional) evaluar el rendimiento por cross-validation en cada paso y anotar la métrica
6. Seleccionar el conjunto final. Normalmente el número de características que dio el mejor rendimiento en CV, o el conjunto más simple dentro de una tolerancia

Este método se puede usar con diferentes modelos y tiene en cuenta la interacción entre características. Por el contrario, es costoso computacionalmente (hay que ajustar el modelo muchas veces), el resultado depende el estimador base

Para usarlo en clase se debe proceder de la siguiente forma

- Usar la validación cruzada en cada paso al evaluar rendimiento para evitar overfitting durante la selección
- Estandarizar variables si el estimador usa magnitudes
- Registrar la métrica por número de variables y mostrar la curva
- Hacer un análisis de estabilidad. Repetir RFE en diferentes folds o submuestras y ver qué variable se repiten
- Si hay muchísimas variables, considerar una preselección

Ejemplo

Sean las características $F = \{x_1, x_2, x_3, x_4, x_5, x_6\}$ y queremos predecir las ventas. La decisión inicial: Usamos regresión lineal como estimador y eliminamos 1 variable por iteración. Mediremos R^2 en la validación cruzada (CV) tras cada eliminación.

1. Paso 0

- Modelo ajustado con $x_1 \dots x_6$ $\rightarrow R^2(cv) = 0.68$
- Coeficiente (valor absoluto, ordenados): $|x_1| = 0.45, |x_4| = 0.40, |x_5| = 0.10, |x_6| = 0.08, |x_2| = 0.06, |x_3| = 0.03$
- Eliminamos x_3

2. Paso 1 (queda x_1, x_2, x_4, x_5, x_6)

- Ajustar de nuevo $R^2 = 0.685$ (ligera mejoría)

- Importancia $|x_1| = 0.46, |x_4| = 0.41, |x_5| = 0.11, |x_6| = 0.09, |x_2| = 0.05$
 - Eliminar x_2
3. Paso 2 (quedan x_1, x_4, x_5, x_6)
- Ajusta $R^2(CV) = 0.69$ (mejora)
 - Importancia $|x_1| = 0.48, |x_4| = 0.43, |x_5| = 0.12, |x_6| = 0.04$
 - Eliminar x_6
4. Paso 3 (quedan x_1, x_4, x_5)
- Ajusta $R^2(CV) = 0.695$ (mejora pequeña)
 - Importancia $|x_1| = 0.49, |x_4| = 0.44, |x_5| = 0.13$
 - Si tu criterio es "mejor R^2 ", puedes seguir o detenerte. Con tres variables es casi máximo
5. Paso 4 (Si eliminamos x_5)
- Ajustar con x_1, x_4 $R^2(CV) = 0.693$ (ligera caída)
 - Como el R^2 empeora al quitar x_5 , el conjunto óptimo es el anterior

Es un método en el que prueba todas las combinaciones posibles de todas las características tomada en k elementos $\binom{n}{k} = C(n, k) = \frac{n!}{k!(n-k)!}$. Si tenemos 8 características y queremos quedarnos que cuatro realizará 70 combinaciones y en cada una obtendrá un valor de rendimiento, por supuesto nos quedaremos con el mayor rendimiento

5 Métodos embebidos

La selección de las características se realiza durante el proceso de entrenamiento del modelo de Machine Learning. No son ni puramente métodos de filtrado (pre-entrenamiento) ni puramente métodos de envoltura (wrapper, que iteran sobre sub-conjuntos), sino que combinan la eficiencia de los primeros con la precisión de los segundos.

La selección de características se integra en la función de coste (o pérdida) del modelo. A medida que el modelo aprende a hacer predicciones, simultáneamente aprende a identificar y penalizar las características irrelevantes.

El mecanismo más común y más poderoso es la Regularización, que añade un término de penalización a la función de pérdida del modelo. Esta penalización se aplica a los coeficientes (pesos) de las características. Entonces el objetivo es forzar que los coeficientes de las características irrelevantes se acerquen a cero o sean cero, estas serán las ignoradas.

Los métodos embebidos más comunes son:

- Regresión LASSO: Utiliza la norma L_1 como penalización
- Árboles (de Decisión, Random Forest, GBDT, XGBoost)
- Cuando tratamos con modelos lineales
- Con un conjunto de datos muy grande