diplomatura universitaria en
inteligencia artificial

FACULTAD DE CIENCIAS EXACTAS
UNIVERSIDAD NACIONAL DEL CENTRO DE LA PROVINCIA DE BUENOS AIRES

Procesamiento de Lenguaje Natural

Reducción de Dimensionalidad

Reducción de Dimensionalidad

Definición

Las colecciones de texto tienen un gran número de características:

- Entre cientos y millones de palabras únicas

Muchos términos pueden ser irrelevantes o parcialmente relevantes para discriminar entre clases o contenidos

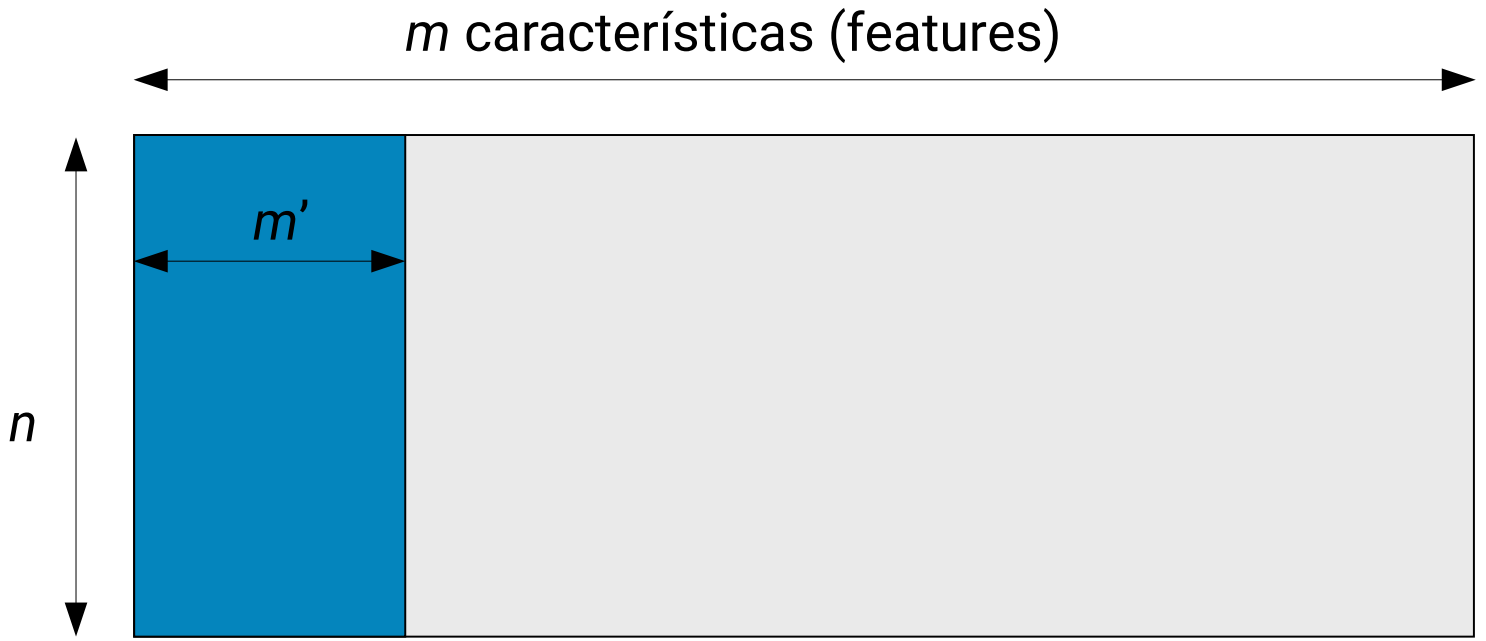
No todos los algoritmos soportan un alto número de características

- Se degrada de la efectividad de los métodos de clasificación
- Los tiempos de aprendizaje se incrementan exponencialmente

Reducción de Dimensionalidad

Definición

Curse of Dimensionality: muchas variables independientes en relación al número de eventos observados



Reducción de Dimensionalidad

Tipos de Reducción

La reducción de dimensionalidad es el proceso a través del cual el número de dimensiones del espacio de vectores se reduce de T a T’, donde T’<<T

Los esquemas de reducción pueden clasificarse en dos categorías:

- Reducción por **selección** de características
- Reducción por **extracción** de características

Reducción de Dimensionalidad

Tipos de Reducción

La reducción por **selección** de características busca reducir los términos de T a T1 de manera que:

- T1 es un subconjunto de términos de T
- T1 << T
- Usando T1 en lugar de T se alcanzan resultados más efectivos en comparación con cualquier otro subconjunto T2

Reducción de Dimensionalidad

Tipos de Reducción

La reducción por **extracción** de características se refiere al mapeo de los datos altamente dimensionales a un espacio de menos dimensiones

- Los métodos de extracción intentan generar a partir de T otro conjunto T1 donde los términos en T1 no necesariamente existen en T
- Un término en T1 puede ser una combinación de términos de T o una transformación de algún término (o grupo de términos) en T

Los criterios para la reducción pueden variar:

- En un contexto no supervisado: minimizar la pérdida de información
- En un contexto supervisado: maximizar la discriminación entre clases

Reducción de Dimensionalidad

Tipos de Reducción

Reducción por selección vs. reducción por extracción

- Extracción de características
  - ➔ Todas las características son utilizadas
  - ➔ Las características resultantes son transformaciones de las originales
- Selección de características:
  - ➔ Solo queda un subconjunto de las características originales

Selección de Características

Definición

Proceso por el cual se elige un subconjunto óptimo de características de acuerdo a una función objetivo

- Características **relevantes**: aquellas que se necesitan para obtener un buen modelo
- Características **irrelevantes**: aquellas que son simplemente innecesarias
- Características **redundantes**: aquellas que se vuelven irrelevantes en la presencia de otras características

Selección de Características

Definición

Objetivos de la selección:

- Reducir la dimensionalidad y remover ruido
- Evitar overfitting y alcanzar una mejor generalización
- Reducir requerimientos de almacenamiento y tiempo de aprendizaje
- Mayor simplicidad y comprensibilidad de los resultados

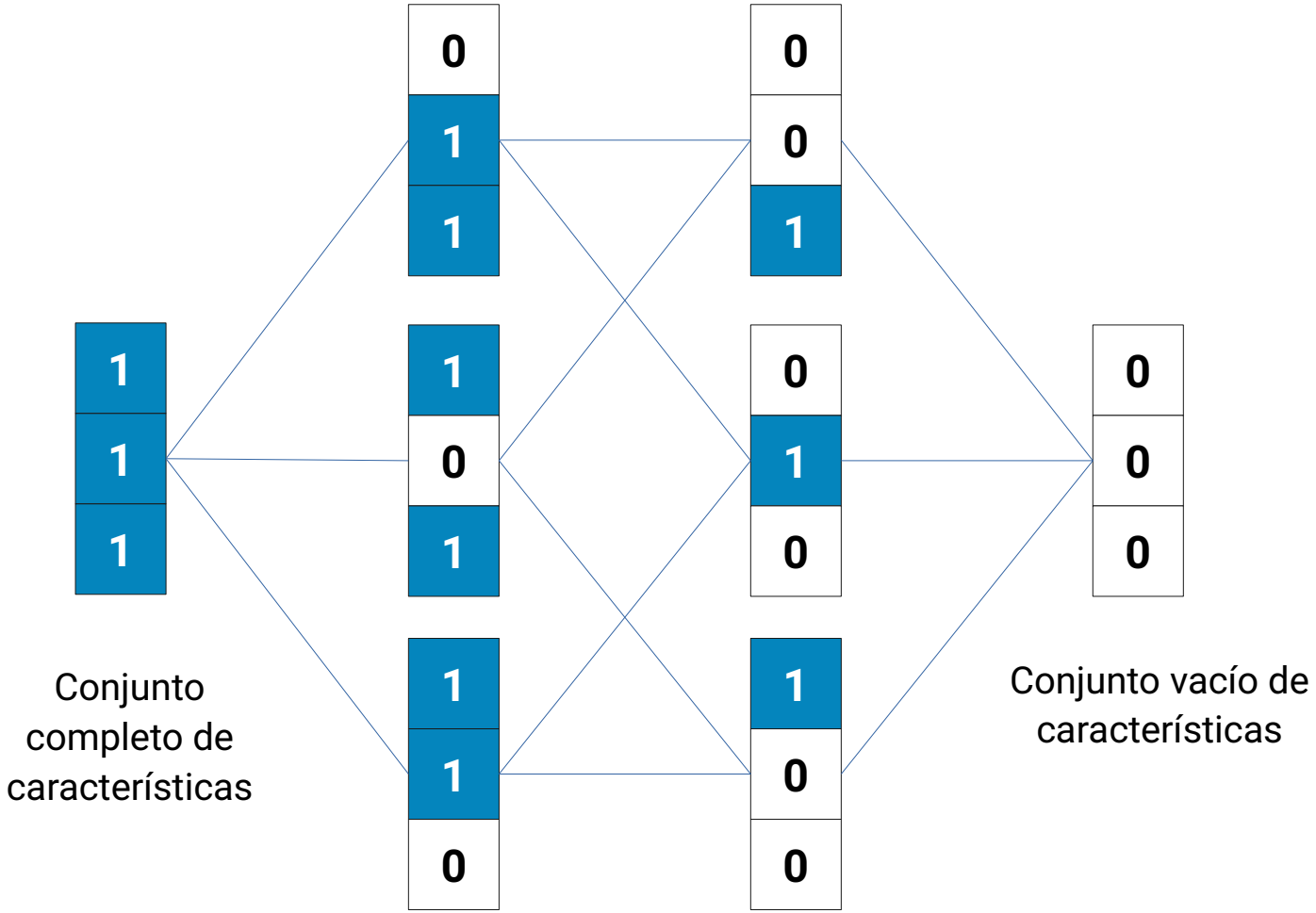
Selección de Características

Definición

La selección de características puede verse como un problema de búsqueda, donde cada estado del espacio de búsqueda corresponde a un subconjunto de características

Selección de Características

Definición

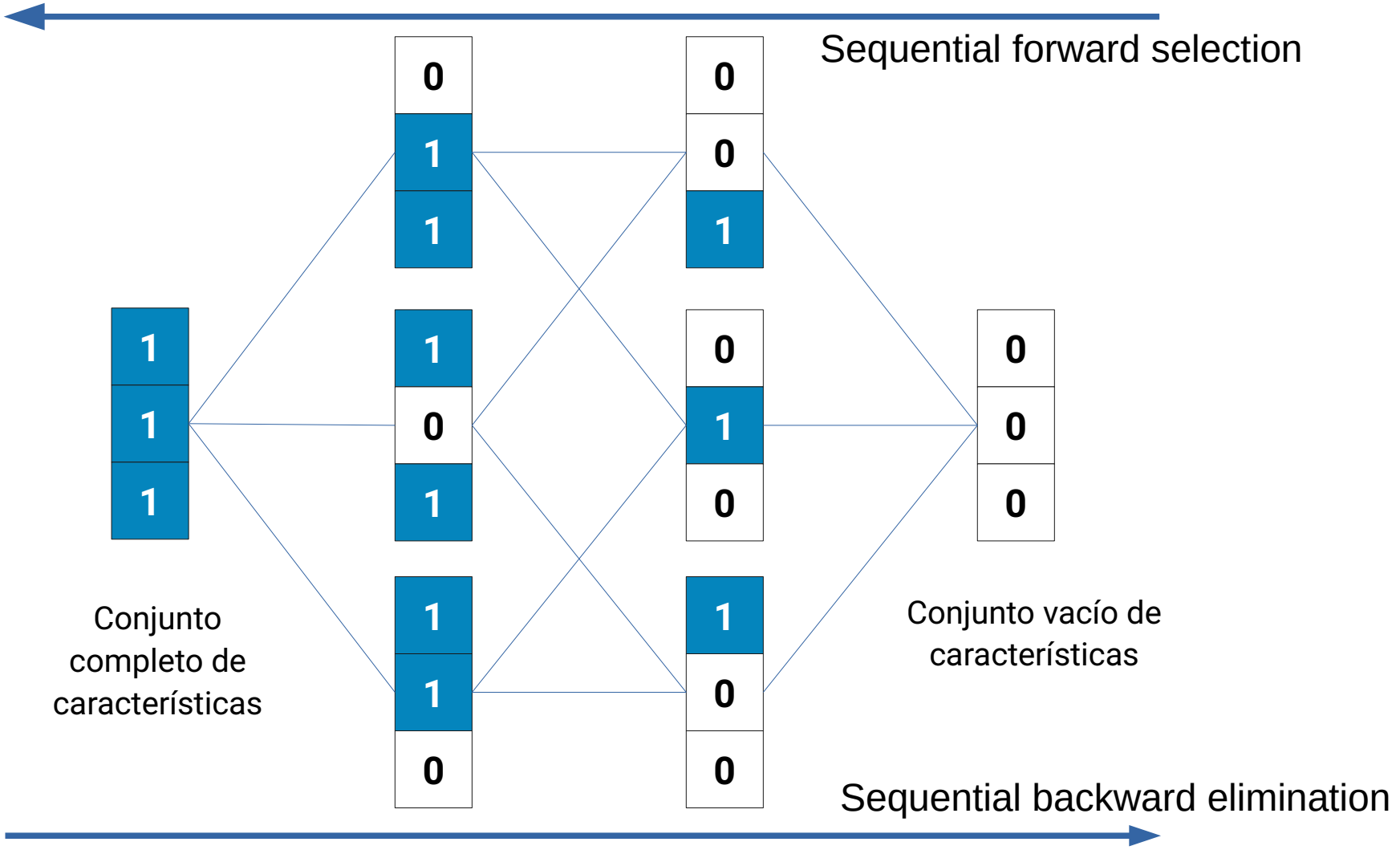


2^M subconjuntos! donde M es el número de características únicas

→ Estrategia de búsqueda

Selección de Características

Definición



Selección de Características

Enfoques

Enfoque Multivariado: selecciona **subconjuntos** de características que combinadas tienen un buen poder predictivo

- Partiendo de un subconjunto inicial se agregan o quitan características, siguiendo un criterio de evaluación:

  → Sequential forward selection

  → Recursive backward elimination

  → Algoritmos genéticos

  → Métodos heurísticos

Selección de Características

Enfoques

Enfoque Univariado: considera las características de a una, en forma independiente de las otras

- Métricas de evaluación de características individuales:
  - → Frecuencia
  - → Teoría de la información
  - → Dependencia
  - → Consistencia
  - → Exactitud

Selección de Características

Enfoques

Métricas de evaluación de características individuales:

- Medidas **supervisadas**:

  → Information gain, cross entropy, mutual information

- Medidas **supervisadas para clases binarias**:

  → Odds ratio (target class vs. the rest), bi-normal separation

- Medidas **no supervisadas**:

  → Term frequency, document frequency

Selección de Características

Enfoques

Ranking de características:

- Se evalúan y rankean las características individuales
- Se seleccionan las top-N del ranking
- Requiere un umbral de selección
- Método simple y que funciona bien en la práctica

Selección de Características

Enfoques

Enfoque Univariado:

- Ignora la relación entre características, se consideran independientes
- Las características pueden ser redundantes
- Una característica que es irrelevante por sí misma, puede ofrecer una mejora significativa en combinación con otra

Enfoque Multivariado:

- Computacionalmente costoso

Selección de Características

Enfoques

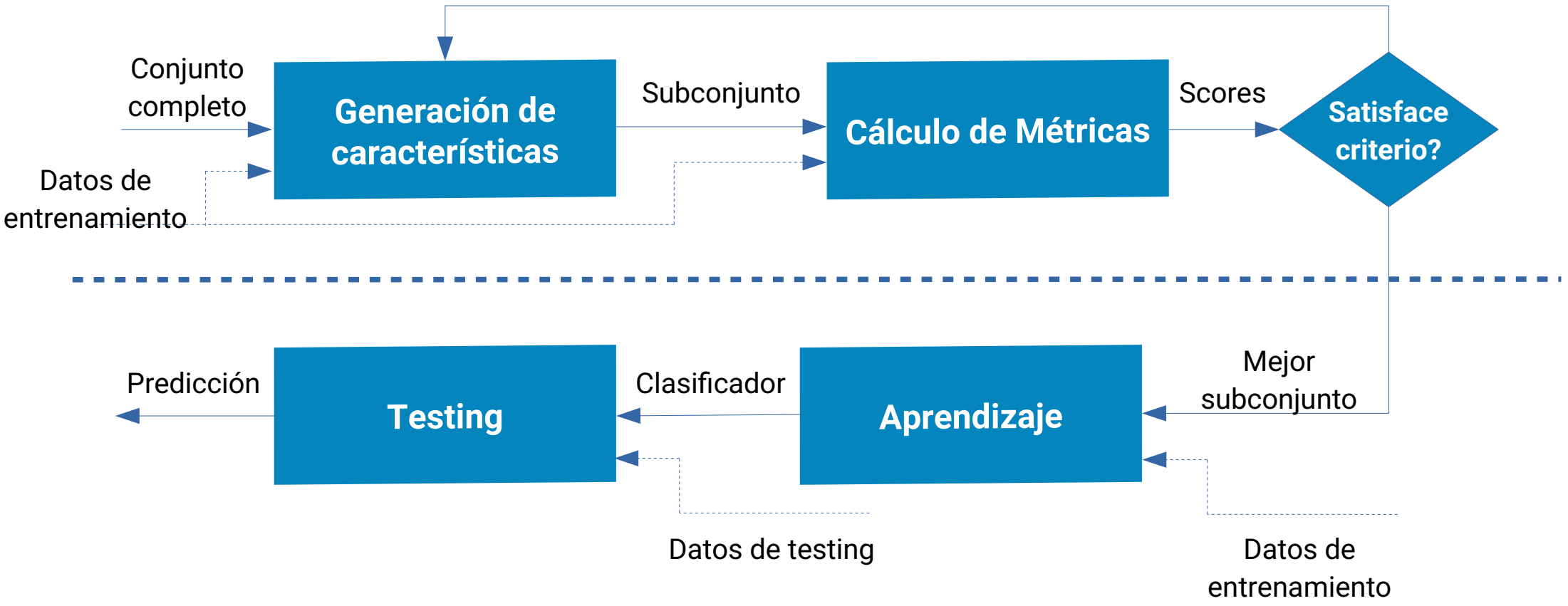
Enfoque de Filtro: rankea características individuales o subconjuntos independientemente del predictor (clasificador)

Enfoque de Wrapper: usa al clasificador para evaluar las características individuales o subconjuntos

Enfoque Embebido: la selección se realiza como parte del entrenamiento del clasificador

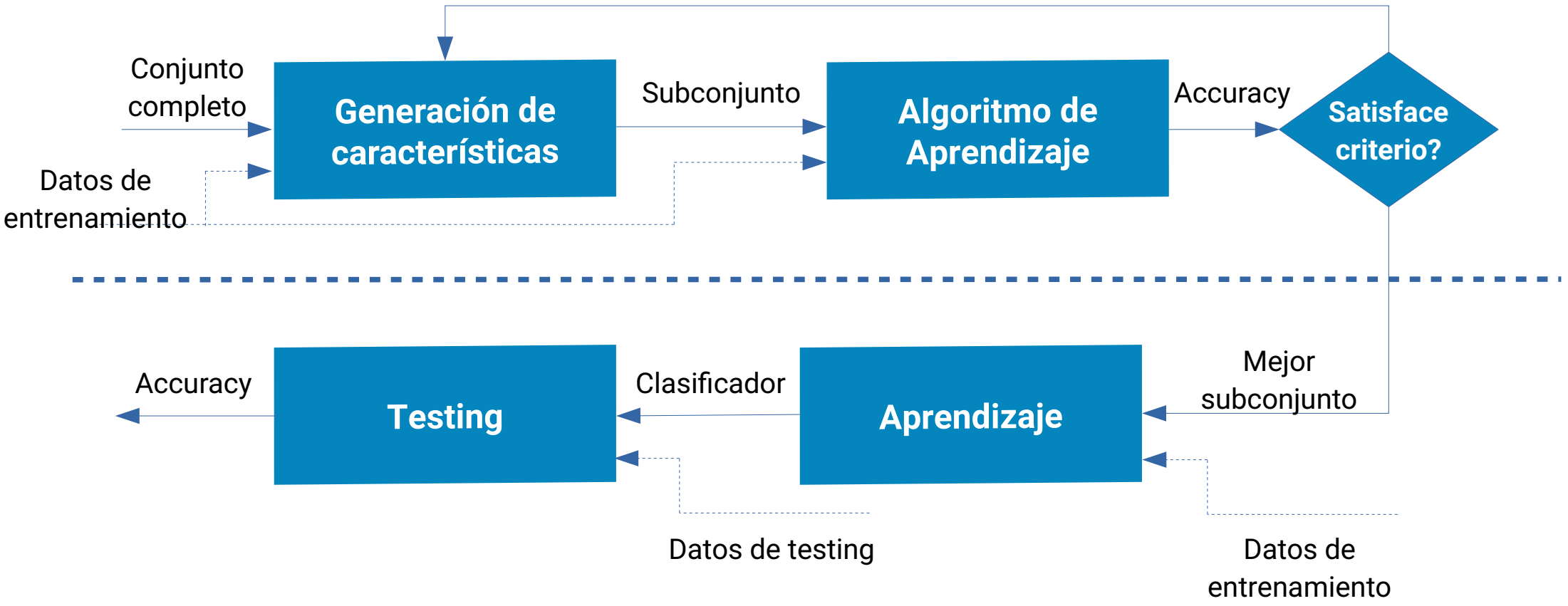
Selección de Características

Enfoque de Filtro



Selección de Características

Enfoque de Wrapper



Selección de Características

Enfoques

Enfoque de Filtro:

- Separa la selección de características del aprendizaje
- Se basa en propiedades generales de los datos (información, distancia, dependencia)
- No está sesgado hacia un algoritmo en particular, las características pueden usarse para aprender distintos modelos
- Eficiente, puede manejar mayor número de características

Enfoque de Wrapper:

- Involucra un algoritmo de aprendizaje determinado
- Usa la exactitud en la predicción como medida de bondad
- Alta exactitud, caro computacionalmente

Selección de Características

Enfoques

	Univariado	Multivariado
Filtro	MI, IG, OR, Freq,...	Category distance
Wrapper	Ranking accuracy	SFS, BFS, RFE,...