**Caracterización de las variables**

**Tipos de variables**:

Numéricas

○ Discretas

○ Continuas

● Categóricas

○ Nominales

○ Ordinales

● Fecha/hora

● Compuestas

**Aspectos de variables**:

**1.Datos faltantes.**

Clasificación de Rubin de valores faltantes (1976):

○**Missing Completely at Random (MCAR)**: significa que la razón por la cuál el

dato no está es completamente aleatoria, y que probablemente no podamos

predecir el valor a partir de otro valor en los datos. Omitir estas observaciones no implicaría un sesgo en las inferencias.

○ **Missing at Random (MAR):** los datos faltantes pueden ser explicados por

valores en las otras columnas, pero no por valores de esa columna. Por

ejemplo, si cada columna representa una elección excluyente. Es decir, el azar no es la única causa por la que faltan esos datos.

○ **Missing not at Random (MNAR):** significa que es probable que la falta de ese

dato no sea al azar. En este caso tenemos que investigar la causa por la que

falta ese valor. →Ejemplos: En un estudio sobre depresión, los pacientes con

depresión son menos propensos a responder algunas preguntas.

**2. Variables categóricas.** Cardinalidad y etiquetas raras.

Los valores que puede tomar una variable categórica forman un grupo

de categorías (también llamadas etiquetas).

● Se denomina cardinalidad al número de categorías existentes.

Cardinalidad. Problemas:

● La mayoría de los modelos de aprendizaje automático no aceptan cadenas

de texto como entradas, por lo tanto, las categorías deben ser codificadas

numéricamente.

● Las técnicas de codificación pueden tener un efecto secundario

indeseado, como aumentar la dimensión del espacio de las variables de

entrada.

● Pueden ocurrir errores en la partición de entrenamiento y validación:

○ Valores que sólo estén disponibles en train set → overfitting.

○ Valores que sólo estén disponibles en test set → el modelo no podrá interpretarlos.

Overfitting.

● Las variables con muchas etiquetas dominarán sobre las que tengan

menos etiquetas, especialmente en los algoritmos basados en árboles de

decisión.

● Un número grande de etiquetas empeora la relación señal/ruido.

● Reducir la cardinalidad puede contribuir a mejorar el desempeño de

algunos modelos.

**3. Cumplimiento de los supuestos de linealidad.**

Ver correlación y distribuición normal de datos.

**4. Distribución de los valores de las independientes.**

Se estudiarán dos grupos de métodos para llevar una distribución a una

forma normal:

○ Transformaciones

■ ln(x)

■ exp(x)

■ 1/x

■ Box-Cox, Yeo Johnson

○ Discretización

■ Bins de frecuencia fija (suelen mejorar la distribución)

■ Bins de ancho fijo (por lo general no mejoran la distribución)

**5. Valores extremos (outliers).**

**6. Magnitud/escala.**

Normalización

**Imputación de datos faltantes:**

Pueden agruparse las técnicas de imputación en dos categorías:

○ Univariada: cuando la imputación se realiza de manera independiente para cada variable de entrada, sin considerar las otras variables.

○ Multivariada: cuando el valor imputado para cada variable es una función de dos o más variables.

**Metodos:**

**Variables numéricas:**

■ Imputación por promedio/mediana.

■ Imputación por valor arbitrario.

■ Imputación de “fin de cola” (end of tail).

**Variables categóricas:**

■ Imputación por categoría frecuente.

■ Agregar categoría “FALTANTE”.

**Ambas:**

■ Análisis de caso completo.

■ Agregar indicador “FALTANTE”.

■ Imputación por muestreo aleatorio.

**Codificación de variables categóricas:**

**Alternativas**

Binary encoding

Feature hashing

Otros

**Tradicionales**

One Hot Encoding

Count/frequency encoding

Ordinal/label encoding

**Relación monotónica**

Label encoding ordenado.

Encoding por promedio

Peso de la evidencia (WoE)

| Metodo | Ventajas | Limitaciones |
| --- | --- | --- |
| **One Hot Encoding.** | ●No realiza ningún supuesto sobre la distribución de las categorías de la variable.  ●Mantiene toda la información de la variable categórica.  ●Es apta para modelos lineales. | ● Aumenta la dimensión del espacio de variables de entrada.  ● No agrega información.  ● Introduce muchas variables con información redundante. |
| **Label/integer encoding** | ●Fácil de implementar,  ●No expande la dimensión del espacio de variables  de entrada.  ●Puede funcionar bien con algoritmos basados en árboles. | ●La codificación no aporta información.  ● No es apropiado para modelos lineales.  ● En producción, no maneja automáticamente nuevas categorías. |
| **Count/Frequency encoding**  Cada categoría se reemplaza por el valor o porcentaje de observaciones  en que aparece en el dataset.  Se hace un supuesto importante: la cantidad de observaciones para cada  categoría está relacionadacon la variable a predecir. | ● Fácil de implementar.  ● No expande las dimensiones del espacio de variables de entrada.  ● Puede tener un desempeño aceptable con modelos basados en árboles.  ● Se captura la representación de cada categoría. | ● No apropiado para modelos lineales.  ● No maneja automáticamente nuevas categorías en test set/producción.  ● Si aparecen dos categorías la misma cantidad de veces, pueden ser  reemplazadas por el mismo número, con la consecuente pérdida de  información. |
| **Binary encoding / feature hashing.** | ● Eficiencia en la representación | ●Pérdida de interpretabilidad. |