Ciencia y analítica de datos

DRA. GRETTEL BARCELÓ ALONSO

1

MÓDULO 3

ANÁLISIS, VISUALIZACIÓN Y TRANSFORMACIÓN DE DATOS

2

Problemas detectados con EDA

EXPLORATORY DATA ANALYSIS

Preprocesamiento

- Alto porcentaje de valores faltantes
  - `isna()`
- Valores atípicos
  - `boxplot, método IQR`
- Alta cardinalidad de atributos categóricos
  - `nunique()`

Ingeniería y selección de características

- Distribución sesgada de atributos numéricos
  - `skew(), histogramas`
- Alta correlación entre características (redundancia)
  - `corr(), diagramas de dispersión, mapas de calor`

3

Manejo de valores faltantes

NAN

Estrategia	Definición	Pros	Contras	Implementación con pandas
Preservar	Mantener los valores faltantes.	El conjunto se conserva su estado original.	Sólo algunas herramientas de análisis de datos lo permiten	
Eliminación por lista	Excluir todos los casos (en lista) que tienen valores faltantes. Pueden ser filas o columnas*	Preserva la distribución si MCAR.	1. Puede descartar demasiados datos y dañar el modelo. 2. Puede generar estimaciones sesgadas si no es MCAR (ya que mantenemos una submuestra especial de la población)	`dropna()`
Imputación media/mediana/moda	Reemplazar el NaN por la media/mediana/moda (para características categóricas) de esa variable**	Buena práctica si MCAR.	1. Puede distorsionar la distribución. 2. Puede distorsionar la relación con otras variables.	`fillna()`

* Cuando la cantidad de valores faltantes en una variable es lo suficientemente grande (aproximadamente más del 25 %), eliminar el atributo es mejor que estimar los valores faltantes.

** Cuando la variable tiene una distribución normal, usar la media. Si está sesgada, usar la mediana.

4

Valores faltantes

CON SKLEARN

- Es una biblioteca para aprendizaje automático
- Incluye varios algoritmos de clasificación, regresión y análisis de grupos.
- Está diseñada para interoperar con las bibliotecas numéricas y científicas *NumPy* y *SciPy*.

```
from sklearn.impute import SimpleImputer
imputer = SimpleImputer(strategy='median')
no_missing_data = imputer.fit_transform(data)
```

`strategy` = {mean, median, most_frequent, constant}

Otros métodos

`IterativeImputer()` - Imputador multivariado que estima cada característica a partir de todas las demás.

`KNNImputer()` - Imputación para completar valores faltantes utilizando k-vecinos más cercanos.

5

Canalizaciones

PIPELINES

Una **canalización** o *pipeline* consiste en dividir una tarea de aprendizaje automático completa en un flujo de trabajo de varios pasos.


```
model =
make_pipeline(SimpleImputer(),StandarScaler(),PCA(),LinearRegression())
```

6

Valores atípicos

OUTLIERS

- Un valor atípico es un punto de datos que es significativamente diferente de los datos restantes.
- Pueden afectar el rendimiento de algunos modelos de aprendizaje automático.

Métodos para la detección de atípicos

Límites	Identificar valores atípicos basados en límites con conocimiento de dominio
Media y desviación estándar	`lower_limit = df[variable].mean() - 3 * df[variable].std()` `upper_limit = df[variable].mean() + 3 * df[variable].std()`

Método IQR	`IQR = df[variable].quantile(0.75) - df[variable].quantile(0.25)`
	`lower_limit = df[variable].quantile(0.25) - (IQR * 1.5)`
	`upper_limit = df[variable].quantile(0.75) + (IQR * 1.5)`

7

Manejo de valores atípicos

OUTLIERS

Estrategia	Definición	Pros	Contras	Implementación con pandas
Imputación media/mediana /moda	Reemplazar el valor atípico por la media/mediana/moda de esa variable.	Preservar la distribución.	Se pierde información de valores atípicos si hay uno.	`df.loc[outliers.index, 'var_with_outliers'] = df['var_with_outliers'].median()`
Límites (Winsorización)	Limitar el máximo y mínimo de una distribución en un valor establecido.	Evita el sobreajuste del modelo.	Distorsiona la distribución.	`df['var_with_outliers'] = df['var_with_outliers'].clip( lower=lower_limit, upper=upper_limit)`
Descarte	Eliminar todas las observaciones que son valores atípicos.		Se pierde información de valores atípicos si hay uno.	`df.drop(outliers.index)`
Transformación	Aplicar una función matemática para reducir el efecto de los valores atípicos.	Mejora la normalidad de los datos.	Puede complicar la interpretación	Ejemplo con logaritmo `df['var_transformed'] = np.log1p(df['var_with_outliers'])`

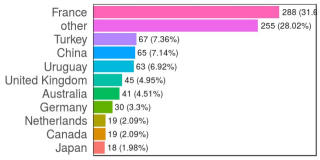
* Cuando la cantidad de outliers es relativamente grande (aunque deberían estar alrededor del 5%), se debe investigar el origen para tomar mejores decisiones.

** Se recomienda hacer varios modelos y comparar resultados.

8

Alta cardinalidad

CATEGÓRICAS



- El número de etiquetas dentro de una variable categórica se conoce como cardinalidad.
- Un alto número de etiquetas dentro de una variable (cientos de valores únicos) se conoce como alta cardinalidad.

Problemas

- Las variables con demasiadas etiquetas tienden a dominar sobre aquellas con solo unas pocas etiquetas.
- Una gran cantidad de etiquetas dentro de una variable puede introducir ruido con poca o ninguna información, lo que hace que los modelos de aprendizaje automático sean propensos a sobreajustarse.

Estrategias

- Agrupación de categorías con conocimiento empresarial.
- Agrupación de categorías con poca ocurrencia en una categoría única.

¿Con qué método de *pandas*?

`groupby()`

9

Características

Y OBJETIVO

Una característica es un atributo de datos que es significativo para el proceso de aprendizaje automático. También conocida como:

- variable independiente
- predictor
- variable de entrada

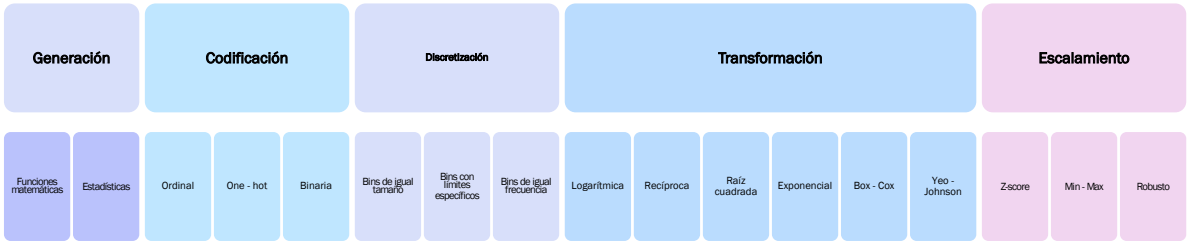
El objetivo, será la variable que se predice en el aprendizaje supervisado. También conocido como:

- variable dependiente
- variable de respuesta
- variable de salida

10

Ingeniería de características

OPERACIONES



11

Codificación

CATEGÓRICAS

fuel
gas
diesel
gas
gas

→

gas	diesel
1	0
0	1
1	0
1	0

Se transforman las variables categóricas a números para que puedan ser procesadas por algoritmos de aprendizaje automático y otras técnicas estadísticas.

Aplicar a: Variables categóricas

¿Por qué es necesario?

- Para que los algoritmos puedan manejar esos valores.
- Incluso si ve que un algoritmo puede tomar entradas categóricas, lo más probable es que el algoritmo incorpore el proceso de codificación en su interior.

12

Codificación
CATEGÓRICAS

Estrategia	Definición	Pros	Contras	Implementación con pandas
One - hot	Crear nuevas variables binarias para indicar si cierta etiqueta es verdadera o no para esa observación. A estas variables también se les conocen como variables *dummies*.	Mantiene toda la información de esa variable.	Expande el espacio de características dramáticamente si hay demasiadas etiquetas en esa variable.	`get_dummies()` `drop_first=True`
Ordinal	Reemplazar las etiquetas por algún número ordinal si el orden es significativo.	Es una transformación muy sencilla.	Puede ingresar sesgo al modelo.	`map()`
Binaria	Reemplazar las etiquetas por un código binario que representa las diferentes categorías de la variable.	Codifica los datos en menos dimensiones que one-hot (log2 # de categorías)	Puede introducir sesgo, especialmente en modelos que son sensibles a las relaciones numéricas entre las variables.	

13

Codificación
CATEGÓRICAS

Naturaleza nominal – Codificación OHE

Categoría del producto – Alimentos, Electrónica, Juguetes, Ropa

Nombre del producto	Categoría	Categoría_Alimentos	Categoría_Electrónica	Categoría_Juguetes	Categoría_Ropa
Audífonos Bluetooth	Electrónica	0	1	0	0
Camiseta estampada	Ropa	0	0	0	1
Caja de cereal integral	Alimentos	1	0	0	0
Muñeca articulada	Juguetes	0	0	1	0
Cargador USB-C	Electrónica	0	1	0	0
Pantalón de mezclilla	Ropa	0	0	0	1
Galletas surtidas	Alimentos	1	0	0	0
Rompecabezas 500 piezas	Juguetes	0	0	1	0

`drop_first`

Naturaleza ordinal – Codificación ordinal

Talla – chica, mediana, grande

chica ⟶ 0
mediana ⟶ 1
grande ⟶ 2

Naturaleza nominal – Codificación binaria

Alimentos ⟶ [0 0]
Electrónica ⟶ [0 1]
Jueguetes ⟶ [1 0]
Ropa ⟶ [1 1]

14

Codificación

CON SKLEARN

One - hot

```
from sklearn.preprocessing import OneHotEncoder
encoder = OneHotEncoder(drop='first')
encoded_data = encoder.fit_transform(data)
```

Ordinal

```
from sklearn.preprocessing import OrdinalEncoder
encoder = OrdinalEncoder(categories=[['bajo','medio','alto']])
encoded_data = encoder.fit_transform(data)
```

Binaria

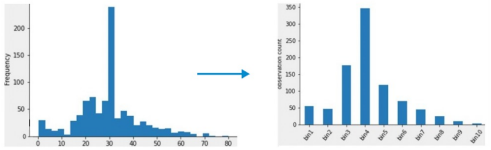
```
!pip install category_encoders
from category_encoders.binary import BinaryEncoder
encoder = BinaryEncoder()
encoded_data = encoder.fit_transform(data)
```

`* category_encoders` Un conjunto de transformadores de estilo *scikit-learn* para codificar variables categóricas en numéricas con diferentes técnicas.

15

Discretización

BINNING



Se transforman las variables continuas en discretas mediante la creación de un conjunto de intervalos contiguos que abarca el rango de valores de la variable.

Aplicar a: Variables numéricas continuas

¿Por qué es necesario?

- Ayuda a mejorar el rendimiento del modelo mediante la agrupación de atributos similares.
- Mejora la interpretabilidad con valores agrupados.
- Minimizar el impacto de los valores extremos.
- Evitar el sobreajuste posible con variables numéricas.

16

Discretización

BINNING

Estrategia	Definición	Pros	Contras	Implementación con pandas
Bins de igual tamaño	Divide los valores de la variable en N contenedores del mismo ancho o con límites establecidos		Sensible a la distribución sesgada.	`cut()`
Bins de igual frecuencia	Divide los valores de la variable en N contenedores, donde cada contenedor contiene la misma cantidad de observaciones.	Puede ayudar a mejorar el rendimiento del algoritmo.	Este agrupamiento arbitrario puede interrumpir la relación con el objetivo.	`qcut()` `bins=N`

data = [2, 4, 7, 9, 10, 11]

Bins de igual tamaño
- cut(data, bins = 3)
  ancho de cada bin = (max – min) / bins
  ancho de cada bin = (11 – 2) / 3 = 3

Se crean 3 intervalos de ancho 3
- [2, 5] - 2, 4
- (5, 8] - 7
- (8, 11] - 9, 10, 11

Bins de igual frecuencia
- qcut(data, bins = 3)
  datos en cada bin = len / bins
  datos en cada bin = 6 / 3 = 2

Se crean 3 intervalos asegurando 2 datos en cada uno
- [2, 4] - 2, 4
- (4, 9] - 7, 9
- (9, 11] - 10, 11

17

Discretización

CON SKLEARN


```
from sklearn.preprocessing import KBinsDiscretizer
grouper = KBinsDiscretizer(n_bins=3, encode='ordinal',
                           strategy='uniform')
grouped_data = grouper.fit_transform(data)
```

`encode`= {onehot, onehot-dense, ordinal}
`strategy` = {uniform, quantile}

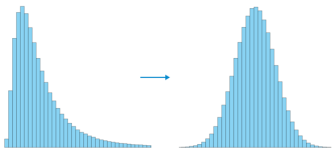
Bins de igual tamaño

Bins de igual frecuencia

18

Transformación

VARIABLES SESGADAS



Se reemplazan los valores originales de las variables con una función matemática de esa variable. Las transformaciones intentan llevar la distribución de la variable a una forma más simétrica, es decir, normal o gaussiana.

Aplicar a: Variables numéricas continuas. Aunque es usual aplicar también estas transformaciones a variables numéricas discretas.

¿Por qué es necesario?

- En la regresión lineal y logística se asume normalidad, que significa que cada variable X debe seguir una distribución normal.
- Los modelos restantes, incluidas las redes neuronales, SVM, los métodos basados en árboles y PCA, no hacen ninguna suposición sobre la distribución de las variables. Sin embargo, en muchas ocasiones el rendimiento del modelo puede beneficiarse de una distribución normal.
- Como se modifica la distribución de los datos de entrada, ayuda a manejar mejor los datos de la cola, los cuales se estarían comportando como valores extremos (outliers)

19

Transformación

VARIABLES SESGADAS

Estrategia	Definición	Implementación con numpy
Logarítmica	$X_{transf} = \log(X)$	`transformed_data = np.log(data)`
Recíproca	$X_{transf} = \frac{1}{X}$	`transformed_data = np.reciprocal(data)`
Raíz cuadrada	$X_{transf} = \sqrt{X}$	`transformed_data = np.sqrt(data)`
Exponencial	$X_{transf} = X^m$	`transformed_data = np.power(data, exponent)`
Box - cox	$X_{transf} = \begin{cases} \frac{X^\lambda - 1}{\lambda} & si\ \lambda > 0\ y\ X > 0 \\ log(X) & si\ \lambda = 0\ y\ X > 0 \end{cases}$	
Yeo - Johnson	$X_{transf} = \begin{cases} \frac{(X+1)^\lambda - 1}{\lambda} & si\ \lambda \neq 0\ y\ X \geq 0 \\ \log(X+1) & si\ \lambda = 0\ y\ X \geq 0 \\ -\frac{(-X+1)^{2-\lambda} - 1}{2-\lambda} & si\ \lambda \neq 2\ y\ X < 0 \\ -\log(-X+1) & si\ \lambda = 2\ y\ X < 0 \end{cases}$	

20

Transformación

CON SKLEARN

```
from sklearn.preprocessing import FunctionTransformer
```

Logarítmica
```
transformer = FunctionTransformer(func=np.log)
```
Recíproca
```
transformer = FunctionTransformer(func=np.reciprocal)
```
Raíz cuadrada
```
transformer = FunctionTransformer(func=np.sqrt)
```
Exponencial
```
transformer = FunctionTransformer(lambda x:
                                  np.power(x, 0.3))
```

```
from sklearn.preprocessing import PowerTransformer
```

Box - cox
```
transformer = PowerTransformer(method='box-cox',
                               standardize=False)
```
Yeo - Johnson
```
transformer = PowerTransformer(method='yeo-johnson',
                               standardize=False)
```

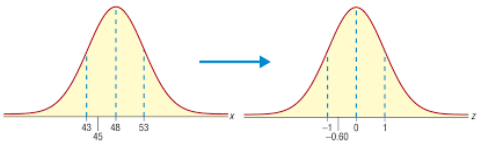
```
transformed_data = transformer.fit_transform(data)
```

Nota. `standarize = true` por defecto. Así, después de la transformación, escalaría con *z-score*

21

Escalamiento

VARIABLES NUMÉRICAS



Se estandariza el rango de las variables independientes o características para tener escalas similares. También se conoce como normalización.

Aplicar a: Variables numéricas continuas. Aunque es usual aplicar también estas transformaciones a variables numéricas discretas.

¿Por qué es necesario?

- Para que todas las variables sean, en principio, igualmente competitivas en cuanto a su relevancia en la construcción del modelo.
- Ayuda a que los métodos de minimización del error como el gradiente descendente, no oscilen demasiado alrededor del valor mínimo buscado.
- Los algoritmos que implican el cálculo de distancias también se ven afectados por la magnitud de la característica.

22

Escalamiento

VARIABLES NUMÉRICAS

Estrategia	Definición	Pros	Contras
Z-score / estandarización / transformación gaussiana	Resta la media y escala los datos a la varianza unitaria. $X_{scaled} = \frac{X - X.mean}{X.std}$	La característica se reescala para tener media 0 y desviación estándar 1.	El uso de la media no permite aminorar el efecto negativo de los outliers.
Min-Max	Transforma características escalando cada característica a un rango dado. Predeterminado a [0,1]. $X_{scaled} = \frac{X - X.min}{X.max - X.min}$	Es la que menos distorsiona los datos originales y hace lo mínimo para que sean competitvas las variables entre sí.	Comprime las observaciones en el rango estrecho si la variable está sesgada o tiene valores atípicos, lo que perjudica el poder predictivo.
Robusto	Elimina la mediana y escala los datos de acuerdo con el rango de cuantiles (el valor predeterminado es IQR) $X_{scaled} = \frac{X - X.median}{IQR}$ $IQR = Q3 - Q1$	El uso de la mediana y el rango intercuartil ayuda a reducir el efecto de outliers.	Robusto

Muchos outliers → **RobustScaler**
Distribución normal y pocos outliers → **StandardScaler**
Valores positivos y rango acotado → **MinMaxScaler**

Algunas veces conviene transformar y entrenar modelos con cada scaler y comparar métricas

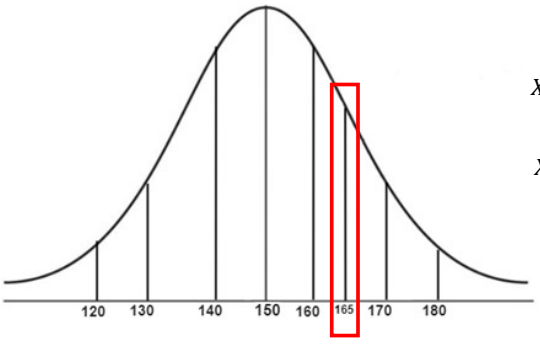
23

Z - score

ESTANDARIZACIÓN

Alturas de los estudiantes de una clase:

- media 150 cm
- desviación estándar 10 cm



$$X_{scaled} = \frac{X - X.mean}{X.std}$$

$$X_{scaled} = \frac{165-150}{10} = 1.5$$

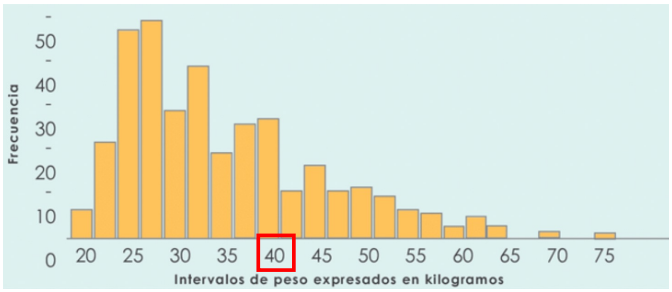
24

Min – Max

ESCALAMIENTO

Pesos de niñas entre 5 y 11 años:

- mínimo 17.5 Kg
- máximo 74.4 Kg



G. Avilés, G. Chávez, y M. Almendarez Hernández, *Análisis de antropometría y de factores determinantes de la prevalencia de obesidad y sobrepeso infantil*, pp. 80-195, Nov. 2017. ISBN 978-607-7777-83-0.

$$X_{scaled} = \frac{X - X.min}{X.\max - X.min}$$

$$X_{scaled} = \frac{40 - 17.5}{74.4 - 17.5} = 0.395$$

25

Escalamiento

CON SKLEARN

Estandarización

```
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
scaled_data = scaler.fit_transform(data)
```

MinMax

```
from sklearn.preprocessing import MinMaxScaler
scaler = MinMaxScaler()
scaled_data = scaler.fit_transform(data)
```

Robusto

```
from sklearn.preprocessing import RobustScaler
scaler = RobustScaler()
scaled_data = scaler.fit_transform(data)
```

Nota. Como la salida la devuelve en formato array, para convertirla a dataframe:

```
scaled_df = pd.DataFrame(
                scaled_data,
                columns=scaler.get_feature_names_out())
```

26

27