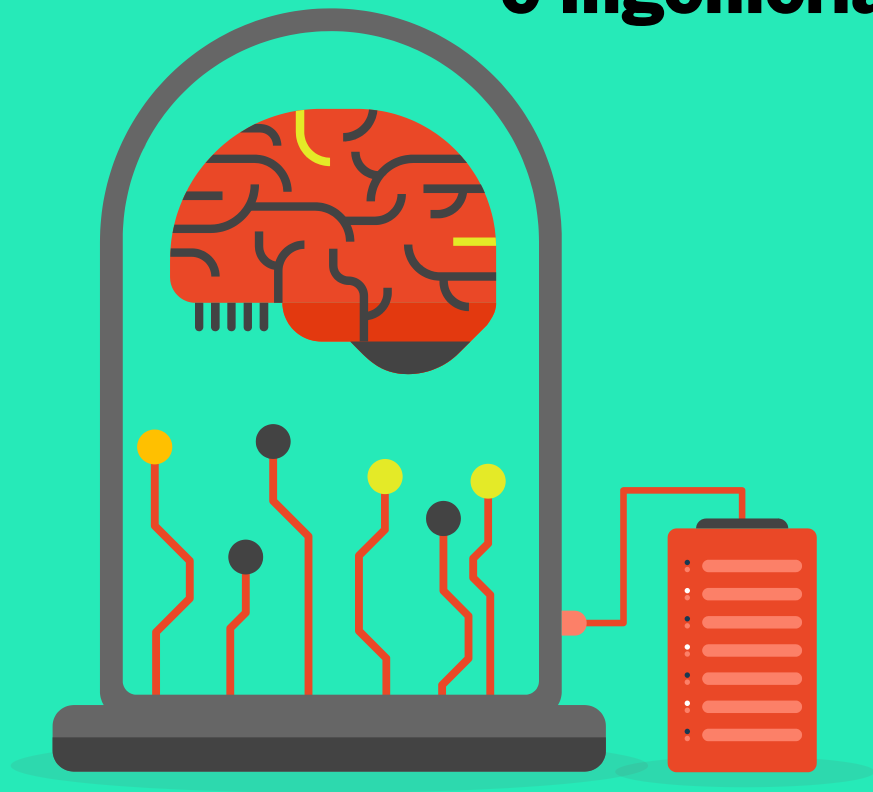


5. Imputación de Variables, Datos Categóricos e Ingeniería de Variables



Imputación de Variables

Cuando se trabaja con datos (sobre todo, con datos reales obtenidos de fenómenos reales), resulta frecuente encontrarse con fuentes de información, tablas o bases de datos que contienen información incompleta, ya sea por problemas durante la recolección de esos datos, o bien por la naturaleza propia de los mismos. Sin embargo, a fin de poder aplicar análisis y algoritmos de Machine Learning, es conveniente realizar la limpieza de estos datos, y eliminar de alguna manera estos registros faltantes.

A este procedimiento se le llama **Imputación de Variables**, es decir, cualquier método que nos permita rellenar, sustituir o incluso eliminar variables incompletas o faltantes en nuestro conjunto de datos.

Imputación de Variables

Un dato faltante es cualquier registro que aparezca como vacío, *NULL* o *NaN* en un conjunto de datos:

	Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
2	LP001002	Male	No	0	Graduate	No	5849	0		360	1	Urban	Y
3	LP001003	Male	Yes	1	Graduate	No	4583	1508	128	360	1	Rural	N
4	LP001005	Male	Yes	0	Graduate	Yes	3000	0	66	360	1	Urban	Y
5	LP001006	Male	Yes	0	Not Graduate	No	2583	2358	120	360	1	Urban	Y
6	LP001008	Male	No	0	Graduate	No	6000	0	141	360	1	Urban	Y
7	LP001011	Male	Yes	2	Graduate	Yes	5417	4196	267	360	1	Urban	Y
8	LP001013	Male	Yes	0	Not Graduate	No	2333	1516	95	360	1	Urban	Y
9	LP001014	Male	Yes	3+	Graduate	No	3036	2504	158	360	0	Semiurban	N
10	LP001018	Male	Yes	2	Graduate	No	4006	1526	168	360	1	Urban	Y
11	LP001020	Male	Yes	1	Graduate	No	12841	10968	349	360	1	Semiurban	N
12	LP001024	Male	Yes	2	Graduate	No	3200	700	70	360	1	Urban	Y
13	LP001027	Male	Yes	2	Graduate		2500	1840	109	360	1	Urban	Y
14	LP001028	Male	Yes	2	Graduate	No	3073	8106	200	360	1	Urban	Y
15	LP001029	Male	No	0	Graduate	No	1853	2840	114	360	1	Rural	N
16	LP001030	Male	Yes	2	Graduate	No	1299	1086	17	120	1	Urban	Y
17	LP001032	Male	No	0	Graduate	No	4950	0	125	360	1	Urban	Y
18	LP001034	Male	No	1	Not Graduate	No	3596	0	100	240		Urban	Y
19	LP001036	Female	No	0	Graduate	No	3510	0	76	360	0	Urban	N
20	LP001038	Male	Yes	0	Not Graduate	No	4887	0	133	360	1	Rural	N
21	LP001041	Male	Yes	0	Graduate		2600	3500	115		1	Urban	Y
22	LP001043	Male	Yes	0	Not Graduate	No	7660	0	104	360	0	Urban	N

Imputación de Variables

Dichos datos pueden aparecer en columnas con cualquier tipo de datos: numéricos enteros, decimales, *categoricos*, fechas, códigos, etc. Y aunque podría considerarse válido el eliminar cualquier registro que presente valores faltantes, dependiendo de la cantidad y sus características, esto podría ser contraproducente para nuestro modelo de Machine Learning.

De manera que es importante conocer y tener en cuenta algunas de las técnicas más comunes de Imputación de Variables. Entre estas, tenemos: eliminar las filas completas, eliminar la variable completa, reemplazar por valor arbitrario, reemplazar por el promedio, la moda o la mediana, reemplazar con el valor anterior o con el valor siguiente, interpolar, reemplazar por el valor más frecuente, o reemplazar con nueva categoría.

Veamos cada una de ellas.

Eliminación de Filas

Si la cantidad de registros (filas) con datos faltantes es considerablemente inferior al total de filas (por ejemplo, 50 filas sin registros, en un total de 50.000 datos) una opción válida podría ser eliminar las filas por completo:

1	Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
2	LP001002	Male	No	0	Graduate	No	5049	0	360	1	1	Urban	Y
3	LP001003	Male	Yes	1	Graduate	No	4583	1508	128	360	1	Rural	N
4	LP001005	Male	Yes	0	Graduate	Yes	3000	0	66	360	1	Urban	Y
5	LP001006	Male	Yes	0	Not Graduate	No	2583	2358	120	360	1	Urban	Y
6	LP001008	Male	No	0	Graduate	No	6000	0	141	360	1	Urban	Y
7	LP001011	Male	Yes	2	Graduate	Yes	5417	4196	267	360	1	Urban	Y
8	LP001013	Male	Yes	0	Not Graduate	No	2333	1516	95	360	1	Urban	Y
9	LP001014	Male	Yes	3+	Graduate	No	3036	2504	158	360	0	Semiurban	N
10	LP001018	Male	Yes	2	Graduate	No	4006	1526	168	360	1	Urban	Y
11	LP001020	Male	Yes	1	Graduate	No	12841	10968	349	360	1	Semiurban	N
12	LP001024	Male	Yes	2	Graduate	No	3200	700	70	360	1	Urban	Y
13	LP001037	Male	Yes	2	Graduate	No	3500	1810	100	360	1	Urban	Y
14	LP001028	Male	Yes	2	Graduate	No	3073	8106	200	360	1	Urban	Y
15	LP001029	Male	No	0	Graduate	No	1853	2840	114	360	1	Rural	N
16	LP001030	Male	Yes	2	Graduate	No	1299	1086	17	120	1	Urban	Y
17	LP001032	Male	No	0	Graduate	No	4950	0	125	360	1	Urban	Y
18	LP001034	Male	No	1	Not Graduate	No	3596	0	100	240		Urban	Y
19	LP001036	Female	No	0	Graduate	No	3510	0	76	360	0	Urban	N
20	LP001038	Male	Yes	0	Not Graduate	No	4887	0	133	360	1	Rural	N
21	LP001041	Male	Yes	0	Graduate	No	2600	3500	113	360	1	Urban	Y
22	LP001043	Male	Yes	0	Not Graduate	No	7660	0	104	360	0	Urban	N

Eliminación de Columnas (variable)

Si existe la sospecha de que la variable faltante no representa un buen predictor, es posible considerar el eliminarla por completo:

	Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
2	LP001002	Male	No	0	Graduate	No	5849	0		360	1	Urban	Y
3	LP001003	Male	Yes	1	Graduate	No	4583	1508	128	360	1	Rural	N
4	LP001005	Male	Yes	0	Graduate	Yes	3000	0	66	360	1	Urban	Y
5	LP001006	Male	Yes	0	Not Graduate	No	2583	2358	120	360	1	Urban	Y
6	LP001008	Male	No	0	Graduate	No	6000	0	141	360	1	Urban	Y
7	LP001011	Male	Yes	2	Graduate	Yes	5417	4196	267	360	1	Urban	Y
8	LP001013	Male	Yes	0	Not Graduate	No	2333	1516	95	360	1	Urban	Y
9	LP001014	Male	Yes	3+	Graduate	No	3036	2504	158	360	0	Semiurban	N
10	LP001018	Male	Yes	2	Graduate	No	4006	1526	168	360	1	Urban	Y
11	LP001020	Male	Yes	1	Graduate	No	12841	10968	349	360	1	Semiurban	N
12	LP001024	Male	Yes	2	Graduate	No	3200	700	70	360	1	Urban	Y
13	LP001027	Male	Yes	2	Graduate		2500	1840	109	360	1	Urban	Y
14	LP001028	Male	Yes	2	Graduate	No	3073	8106	200	360	1	Urban	Y
15	LP001029	Male	No	0	Graduate	No	1853	2840	114	360	1	Rural	N
16	LP001030	Male	Yes	2	Graduate	No	1299	1086	17	120	1	Urban	Y
17	LP001032	Male	No	0	Graduate	No	4950	0	125	360	1	Urban	Y
18	LP001034	Male	No	1	Not Graduate	No	3596	0	100	240		Urban	Y
19	LP001036	Female	No	0	Graduate	No	3510	0	76	360	0	Urban	N
20	LP001038	Male	Yes	0	Not Graduate	No	4887	0	133	360	1	Rural	N
21	LP001041	Male	Yes	0	Graduate		2600	3500	115		1	Urban	Y
22	LP001043	Male	Yes	0	Not Graduate	No	7660	0	104	360	0	Urban	N

Reemplazar por Valor Arbitrario

Dependiendo de la naturaleza de la variable, podría reemplazarse el dato faltante por un valor arbitrario acorde con su rango, escala, clase, etc:

1	Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
2	LP001002	Male	No	0	Graduate	No	5849	0		360	1	Urban	Y
3	LP001003	Male	Yes	1	Graduate	No	4583	1508	128	360	1	Rural	N
4	LP001005	Male	Yes	0	Graduate	Yes	3000	0	66	360	1	Urban	Y
5	LP001006	Male	Yes	0	Not Graduate	No	2583	2358	120	360	1	Urban	Y
6	LP001008	Male	No	0	Graduate	No	6000	0	141	360	1	Urban	Y
7	LP001011	Male	Yes	2	Graduate	Yes	5417	4196	267	360	1	Urban	Y
8	LP001013	Male	Yes	0	Not Graduate	No	2333	1516	95	360	1	Urban	Y
9	LP001014	Male	Yes	3+	Graduate	No	3036	2504	158	360	0	Semiurban	N
10	LP001018	Male	Yes	2	Graduate	No	4006	1526	168	360	1	Urban	Y
11	LP001020	Male	Yes	1	Graduate	No	12841	10968	349	360	1	Semiurban	N
12	LP001024	Male	Yes	2	Graduate	No	3200	700	70	360	1	Urban	Y
13	LP001027	Male	Yes	2	Graduate		2500	1840	109	360	1	Urban	Y
14	LP001028	Male	Yes	2	Graduate	No	3073	8106	200	360	1	Urban	Y
15	LP001029	Male	No	0	Graduate	No	1853	2840	114	360	1	Rural	N
16	LP001030	Male	Yes	2	Graduate	No	1299	1086	17	120	1	Urban	Y
17	LP001032	Male	No	0	Graduate	No	4950	0	125	360	1	Urban	Y
18	LP001034	Male	No	1	Not Graduate	No	3596	0	100	240	0	Urban	Y
19	LP001036	Female	No	0	Graduate	No	3510	0	76	360	0	Urban	N
20	LP001038	Male	Yes	0	Not Graduate	No	4887	0	133	360	1	Rural	N
21	LP001041	Male	Yes	0	Graduate		2600	3500	115		1	Urban	Y
22	LP001043	Male	Yes	0	Not Graduate	No	7660	0	104	360	0	Urban	N

Reemplazar por Media, Moda o Mediana

Del mismo modo, para variables numéricas pueden considerarse la media, moda o mediana como valores a sustituir en los datos faltantes:

1	Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
2	LP001002	Male	No	0	Graduate	No	5849	0	Media	360	1	Urban	Y
3	LP001003	Male	Yes	1	Graduate	No	4583	1508	128	360	1	Rural	N
4	LP001005	Male	Yes	0	Graduate	Yes	3000	0	66	360	1	Urban	Y
5	LP001006	Male	Yes	0	Not Graduate	No	2583	2358	120	360	1	Urban	Y
6	LP001008	Male	No	0	Graduate	No	6000	0	141	360	1	Urban	Y
7	LP001011	Male	Yes	2	Graduate	Yes	5417	4196	267	360	1	Urban	Y
8	LP001013	Male	Yes	0	Not Graduate	No	2333	1516	95	360	1	Urban	Y
9	LP001014	Male	Yes	3+	Graduate	No	3036	2504	158	360	0	Semiurban	N
10	LP001018	Male	Yes	2	Graduate	No	4006	1526	168	360	1	Urban	Y
11	LP001020	Male	Yes	1	Graduate	No	12841	10968	349	360	1	Semiurban	N
12	LP001024	Male	Yes	2	Graduate	No	3200	700	70	360	1	Urban	Y
13	LP001027	Male	Yes	2	Graduate		2500	1840	109	360	1	Urban	Y
14	LP001028	Male	Yes	2	Graduate	No	3073	8106	200	360	1	Urban	Y
15	LP001029	Male	No	0	Graduate	No	1853	2840	114	360	1	Rural	N
16	LP001030	Male	Yes	2	Graduate	No	1299	1086	17	120	1	Urban	Y
17	LP001032	Male	No	0	Graduate	No	4950	0	125	360	1	Urban	Y
18	LP001034	Male	No	1	Not Graduate	No	3596	0	100	240		Urban	Y
19	LP001036	Female	No	0	Graduate	No	3510	0	76	360	0	Urban	N
20	LP001038	Male	Yes	0	Not Graduate	No	4887	0	133	360	1	Rural	N
21	LP001041	Male	Yes	0	Graduate		2600	3500	115	Moda	1	Urban	Y
22	LP001043	Male	Yes	0	Not Graduate	No	7660	0	104	360	0	Urban	N

Reemplazar por Valor Anterior o Siguiente

Se puede considerar reemplazar un dato en función de su valor anterior o siguiente en el conjunto:

	Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
2	LP001002	Male	No	0	Graduate	No	5849	0	128	360	1	Urban	Y
3	LP001003	Male	Yes	1	Graduate	No	4583	1508	128	360	1	Rural	N
4	LP001005	Male	Yes	0	Graduate	Yes	3000	0	66	360	1	Urban	Y
5	LP001006	Male	Yes	0	Not Graduate	No	2583	2358	120	360	1	Urban	Y
6	LP001008	Male	No	0	Graduate	No	6000	0	141	360	1	Urban	Y
7	LP001011	Male	Yes	2	Graduate	Yes	5417	4196	267	360	1	Urban	Y
8	LP001013	Male	Yes	0	Not Graduate	No	2333	1516	95	360	1	Urban	Y
9	LP001014	Male	Yes	3+	Graduate	No	3036	2504	158	360	0	Semiurban	N
10	LP001018	Male	Yes	2	Graduate	No	4006	1526	168	360	1	Urban	Y
11	LP001020	Male	Yes	1	Graduate	No	12841	10968	349	360	1	Semiurban	N
12	LP001024	Male	Yes	2	Graduate	No	3200	700	70	360	1	Urban	Y
13	LP001027	Male	Yes	2	Graduate		2500	1840	109	360	1	Urban	Y
14	LP001028	Male	Yes	2	Graduate	No	3073	8106	200	360	1	Urban	Y
15	LP001029	Male	No	0	Graduate	No	1853	2840	114	360	1	Rural	N
16	LP001030	Male	Yes	2	Graduate	No	1299	1086	17	120	1	Urban	Y
17	LP001032	Male	No	0	Graduate	No	4950	0	125	360	1	Urban	Y
18	LP001034	Male	No	1	Not Graduate	No	3596	0	100	240	1	Urban	Y
19	LP001036	Female	No	0	Graduate	No	3510	0	76	360	0	Urban	N
20	LP001038	Male	Yes	0	Not Graduate	No	4887	0	133	360	1	Rural	N
21	LP001041	Male	Yes	0	Graduate		2600	3500	115		1	Urban	Y
22	LP001043	Male	Yes	0	Not Graduate	No	7660	0	104	360	0	Urban	N

Reemplazar con Interpolación

Para variables numéricas, podría interpolarse un valor en función de su entorno:

	Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
1													
2	LP001002	Male	No	0	Graduate	No	5849	0	102.3	360	1	Urban	Y
3	LP001003	Male	Yes	1	Graduate	No	4583	1508	128	360	1	Rural	N
4	LP001005	Male	Yes	0	Graduate	Yes	3000	0	66	360	1	Urban	Y
5	LP001006	Male	Yes	0	Not Graduate	No	2583	2358	120	360	1	Urban	Y
6	LP001008	Male	No	0	Graduate	No	6000	0	141	360	1	Urban	Y
7	LP001011	Male	Yes	2	Graduate	Yes	5417	4196	267	360	1	Urban	Y
8	LP001013	Male	Yes	0	Not Graduate	No	2333	1516	95	360	1	Urban	Y
9	LP001014	Male	Yes	3+	Graduate	No	3036	2504	158	360	0	Semiurban	N
10	LP001018	Male	Yes	2	Graduate	No	4006	1526	168	360	1	Urban	Y
11	LP001020	Male	Yes	1	Graduate	No	12841	10968	349	360	1	Semiurban	N
12	LP001024	Male	Yes	2	Graduate	No	3200	700	70	360	1	Urban	Y
13	LP001027	Male	Yes	2	Graduate		2500	1840	109	360	1	Urban	Y
14	LP001028	Male	Yes	2	Graduate	No	3073	8106	200	360	1	Urban	Y
15	LP001029	Male	No	0	Graduate	No	1853	2840	114	360	1	Rural	N
16	LP001030	Male	Yes	2	Graduate	No	1299	1086	17	120	1	Urban	Y
17	LP001032	Male	No	0	Graduate	No	4950	0	125	360	1	Urban	Y
18	LP001034	Male	No	1	Not Graduate	No	3596	0	100	240		Urban	Y
19	LP001036	Female	No	0	Graduate	No	3510	0	76	360	0	Urban	N
20	LP001038	Male	Yes	0	Not Graduate	No	4887	0	133	360	1	Rural	N
21	LP001041	Male	Yes	0	Graduate		2600	3500	115	360	1	Urban	Y
22	LP001043	Male	Yes	0	Not Graduate	No	7660	0	104	360	0	Urban	N

Reemplazar con Valor Frecuente

Para variables categóricas, se puede considerar el reemplazar siempre la clase más frecuente:

	Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
1													
2	LP001002	Male	No	0	Graduate	No	5849	0		360	1	Urban	Y
3	LP001003	Male	Yes	1	Graduate	No	4583	1508	128	360	1	Rural	N
4	LP001005	Male	Yes	0	Graduate	Yes	3000	0	66	360	1	Urban	Y
5	LP001006	Male	Yes	0	Not Graduate	No	2583	2358	120	360	1	Urban	Y
6	LP001008	Male	No	0	Graduate	No	6000	0	141	360	1	Urban	Y
7	LP001011	Male	Yes	2	Graduate	Yes	5417	4196	267	360	1	Urban	Y
8	LP001013	Male	Yes	0	Not Graduate	No	2333	1516	95	360	1	Urban	Y
9	LP001014	Male	Yes	3+	Graduate	No	3036	2504	158	360	0	Semiurban	N
10	LP001018	Male	Yes	2	Graduate	No	4006	1526	168	360	1	Urban	Y
11	LP001020	Male	Yes	1	Graduate	No	12841	10968	349	360	1	Semiurban	N
12	LP001024	Male	Yes	2	Graduate	No	3200	700	70	360	1	Urban	Y
13	LP001027	Male	Yes	2	Graduate	No	2500	1840	109	360	1	Urban	Y
14	LP001028	Male	Yes	2	Graduate	No	3073	8106	200	360	1	Urban	Y
15	LP001029	Male	No	0	Graduate	No	1853	2840	114	360	1	Rural	N
16	LP001030	Male	Yes	2	Graduate	No	1299	1086	17	120	1	Urban	Y
17	LP001032	Male	No	0	Graduate	No	4950	0	125	360	1	Urban	Y
18	LP001034	Male	No	1	Not Graduate	No	3596	0	100	240		Urban	Y
19	LP001036	Female	No	0	Graduate	No	3510	0	76	360	0	Urban	N
20	LP001038	Male	Yes	0	Not Graduate	No	4887	0	133	360	1	Rural	N
21	LP001041	Male	Yes	0	Graduate	No	2600	3500	115		1	Urban	Y
22	LP001043	Male	Yes	0	Not Graduate	No	7660	0	104	360	0	Urban	N

Reemplazar con Nueva Categoría

O también, considerar los valores faltantes como una nueva categoría separada y etiquetarla de esa manera:

	Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
1	LP001002	Male	No	0	Graduate	No	5849	0		360	1	Urban	Y
2	LP001003	Male	Yes	1	Graduate	No	4583	1508	128	360	1	Rural	N
3	LP001005	Male	Yes	0	Graduate	Yes	3000	0	66	360	1	Urban	Y
4	LP001006	Male	Yes	0	Not Graduate	No	2583	2358	120	360	1	Urban	Y
5	LP001008	Male	No	0	Graduate	No	6000	0	141	360	1	Urban	Y
6	LP001011	Male	Yes	2	Graduate	Yes	5417	4196	267	360	1	Urban	Y
7	LP001013	Male	Yes	0	Not Graduate	No	2333	1516	95	360	1	Urban	Y
8	LP001014	Male	Yes	3+	Graduate	No	3036	2504	158	360	0	Semiurban	N
9	LP001018	Male	Yes	2	Graduate	No	4006	1526	168	360	1	Urban	Y
10	LP001020	Male	Yes	1	Graduate	No	12841	10968	349	360	1	Semiurban	N
11	LP001024	Male	Yes	2	Graduate	No	3200	700	70	360	1	Urban	Y
12	LP001027	Male	Yes	2	Graduate	Faltante	2500	1840	109	360	1	Urban	Y
13	LP001028	Male	Yes	2	Graduate	No	3073	8106	200	360	1	Urban	Y
14	LP001029	Male	No	0	Graduate	No	1853	2840	114	360	1	Rural	N
15	LP001030	Male	Yes	2	Graduate	No	1299	1086	17	120	1	Urban	Y
16	LP001032	Male	No	0	Graduate	No	4950	0	125	360	1	Urban	Y
17	LP001034	Male	No	1	Not Graduate	No	3596	0	100	240		Urban	Y
18	LP001036	Female	No	0	Graduate	No	3510	0	76	360	0	Urban	N
19	LP001038	Male	Yes	0	Not Graduate	No	4887	0	133	360	1	Rural	N
20	LP001041	Male	Yes	0	Graduate	Faltante	2600	3500	115		1	Urban	Y
21	LP001043	Male	Yes	0	Not Graduate	No	7660	0	104	360	0	Urban	N

Sexto Notebook Práctico

Veamos algunos ejemplos de cómo se aplican estas técnicas en Python.

Sexto Notebook Práctico

Datos Categóricos

En los ejemplos realizados sobre Regresión Lineal, trabajamos con datos cuyas variables eran numéricas (números enteros o decimales). Sin embargo, resulta común encontrar conjuntos de datos en donde no necesariamente todos los tipos de datos presentes serán numéricos. Por ejemplo, en la tabla mostrada en este tema, tenemos variables como *Gender*, *Married* o *Education*, en donde los valores no representan número, sino mas bien clases o *Categorías* finitas.

	Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
1	LP001002	Male	No	0	Graduate	No	5849	0		360	1	Urban	Y
2	LP001003	Male	Yes	1	Graduate	No	4583	1508	128	360	1	Rural	N
3	LP001005	Male	Yes	0	Graduate	Yes	3000	0	66	360	1	Urban	Y
4	LP001006	Male	Yes	0	Not Graduate	No	2583	2358	120	360	1	Urban	Y
5	LP001008	Male	No	0	Graduate	No	6000	0	141	360	1	Urban	Y
6	LP001011	Male	Yes	2	Graduate	Yes	5417	4196	267	360	1	Urban	Y
7	LP001013	Male	Yes	0	Not Graduate	No	2333	1516	95	360	1	Urban	Y
8	LP001014	Male	Yes	3+	Graduate	No	3036	2504	158	360	0	Semiurban	N

Datos Categóricos

Pero ya que los algoritmos de Machine Learning se basan en la aplicación de algoritmos y fórmulas matemáticas sobre las variables de entrada, no es posible entonces emplear los datos categóricos tal cual se presentan, sino que los mismos deben ser codificados de alguna manera en valores numéricos (aunque existen ciertas librerías y algoritmos que se encargan de manera intrínseca de esta codificación).

Ahora bien, los datos Categóricos pueden ser divididos en dos clases principales:

Datos Nominales y Datos Ordinales

Datos Categóricos

Los **Datos Categóricos Nominales** son aquellos que presentan clases que no tienen ningún orden en particular. Por ejemplo, valores como *Masculino* y *Femenino*, *Empleado* y *No Empleado*, *Si* y *No*, o incluso clases como nombres de lugares, países, marcas, etc. En este tipo de datos, la precedencia es irrelevante.

Los **Datos Categóricos Ordinales** son aquellos en donde la precedencia u orden sí es relevante. Por ejemplo, si hablamos de tallas de ropa, el orden de las clases es importante: *XS*, *S*, *M*, *L*, *XL*. O si tenemos una descripción en clases de un rango de temperaturas: *Muy Frío*, *Frío*, *Tibio*, *Caliente*, *Muy Caliente*.

Dependiendo del tipo de dato, es necesario codificarlos de formas diferentes. En este curso veremos las dos formas más comunes de realizarlo: cuando se trata de datos Nominales, usaremos lo que se conoce como ***One-Hot Encoding***, mientras que para datos Ordinales, usaremos el ***Label Encoding***.


Label Encoding

El caso más sencillo de codificación de variables categóricas es el ***Label Encoding***. Como se mencionó, este se emplea cuando se trabaja con datos Ordinales, y consiste en sustituir la variable categórica por una secuencia numérica, respetando el orden o precedencia de los datos originales. En este caso, es necesario conocer la cantidad de clases (cardinalidad) de la variable, y asignar la numeración de manera adecuada:

Variable Original		Variable Codificada
Temperatura		Temp Codificada
Muy Frío		1
Frío	→	2
Tibio		3
Caliente		4
Muy Caliente		5

One-Hot Encoding

El **One-Hot Encoding**, empleado cuando los datos son Nominales, consiste en mapear los datos originales a un vector que contiene 1 y 0 denotando la ausencia o presencia de la variable:

Variable Original		Variable Codificada	
Sexo		Sexo_0	Sexo_1
Femenino		1	0
Masculino		0	1
Masculino		0	1
Femenino		1	0
Masculino		0	1

One-Hot Encoding

Dependiendo de la cantidad de categorías por variable, será el tamaño de dicho vector de unos y ceros:

Variable Original		Variable Codificada		
País		Pais_0	Pais_1	Pais_2
Francia		1	0	0
China		0	1	0
Argentina		0	0	1
Argentina		0	0	1
China		0	1	0

Continuación del Notebook Práctico

Veamos algunos ejemplos de cómo se aplican las técnicas de codificación de datos Cateóricos en Python.

Sexto Notebook Práctico

Ingeniería de Variables

En términos generales, todas las técnicas de imputación y codificación de variables presentadas anteriormente pueden considerarse como métodos de **Ingeniería de Variables (Feature Engineering)**, es decir, maneras de transformar los datos de entrada para obtener variables más adecuadas para los modelos de Machine Learning. Sin embargo, una última técnica fundamental consiste en la creación de variables nuevas a partir de las ya existentes, que puedan enriquecer el conjunto de datos original e, incluso, que puedan servir como mejores predictores que las variables originales.

Es importante mencionar que las variables creadas deben tener significados relevantes dentro del propio conjunto de datos, y el conocimiento del tema o datos en cuestión es esencial para la correcta creación de dichas variables.

Ingeniería de Variables

Veamos un ejemplo. Podríamos partir del siguiente conjunto de datos:

País	Área km ²	Población	PIB MM USD
Venezuela	916.445	28.866.000	142.416
Francia	675.417	67.407.241	3.547.962
China	9.596.960	1.403.500.365	29.375.296
Argentina	2.780.400	47.327.407	1.104.860

Y de manera natural podríamos crear variables que aporten mayor información sobre las variables, como la Densidad de Población ($\text{Población}/\text{Área}$) o el PIB per cápita ($\text{PIB}/\text{Población}$).

Ingeniería de Variables

De modo que convertimos el conjunto de datos al siguiente:

País	Área km ²	Población	Densidad (hab/km ²)	PIB MM USD	PIB per cápita USD
Venezuela	916.445	28.866.000	31,49	142.416	4.933,69
Francia	675.417	67.407.241	99,80	3.547.962	52.634,73
China	9.596.960	1.403.500.365	146,24	29.375.296	20.930,02
Argentina	2.780.400	47.327.407	17,02	1.104.860	23.345,03

En la mayoría de los casos, los modelos de Machine Learning se beneficiarán de este tipo de transformaciones ya que ganan información sobre las variables intrínsecas del conjunto de datos. En consecuencia, los errores de predicción tienden a disminuir, o bien aumentar la precisión de los modelos.

Continuación del Notebook Práctico

Por último, veamos el cómo se implementan este tipo de transformaciones y creación de variables en Python.

Sexto Notebook Práctico