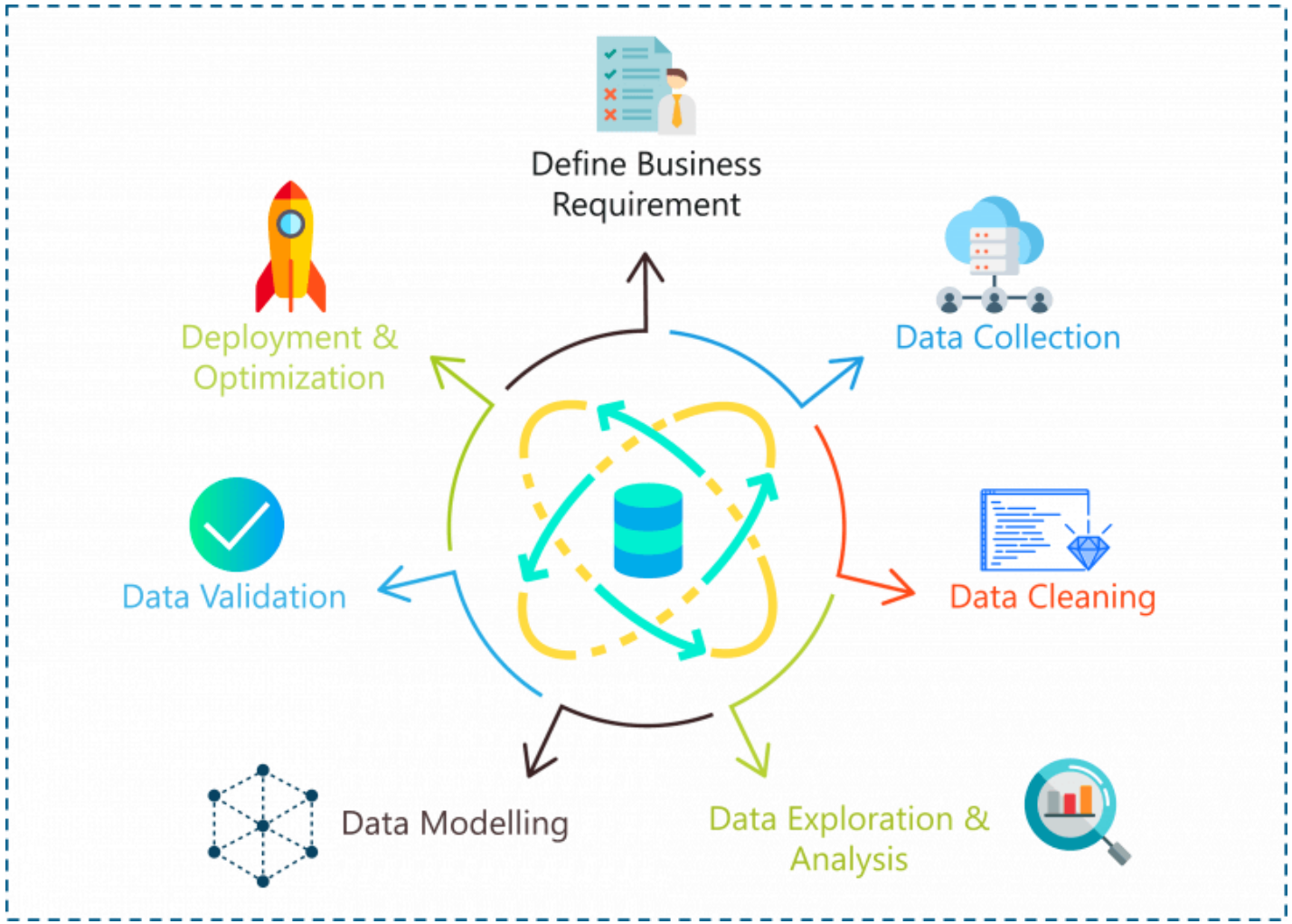
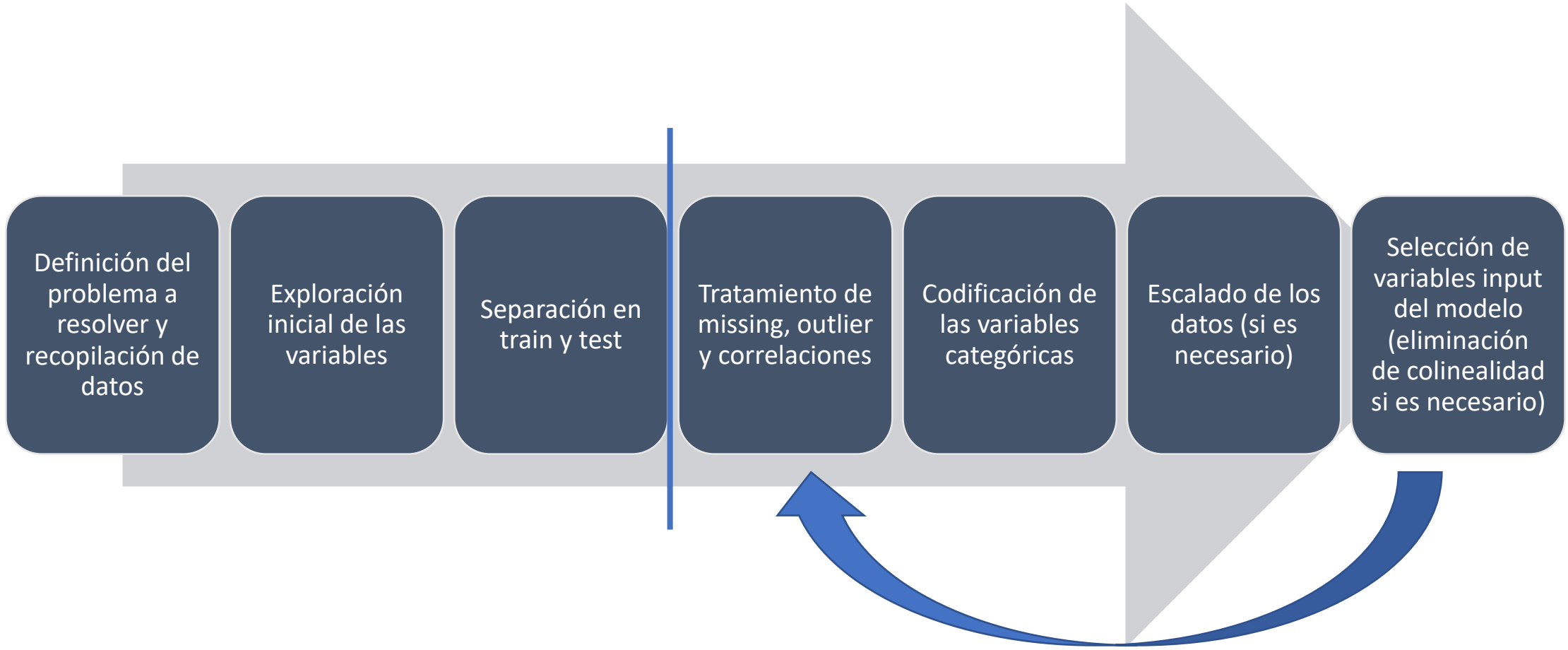
Pasos a seguir hasta la realización de un modelo de ML

Etapas de un proyecto de ciencia de datos



Pasos hasta la realización de un modelo de predicción



Proceso iterativo:
se suele cambiar el preprocesado de datos para comparar diferentes resultados del modelo

1. Definición del problema a resolver

- ¿Cuál es el problema?
- Acción que buscamos hacer para solucionar el problema
- ¿Cuáles son las variables disponibles?
- ¿En qué momento se va a implantar el modelo?¿Qué variables hay disponibles en el momento de llamada al modelo?
- ¿Cómo se va a validar el modelo?

1. Definición del problema a resolver

Ejemplo: Modelo para conceder un préstamo bancario

¿Cuál es el problema a resolver?

¿Qué variables tendré disponibles cuando en producción se ejecute mi modelo?

¿Cómo voy a evaluar el modelo?

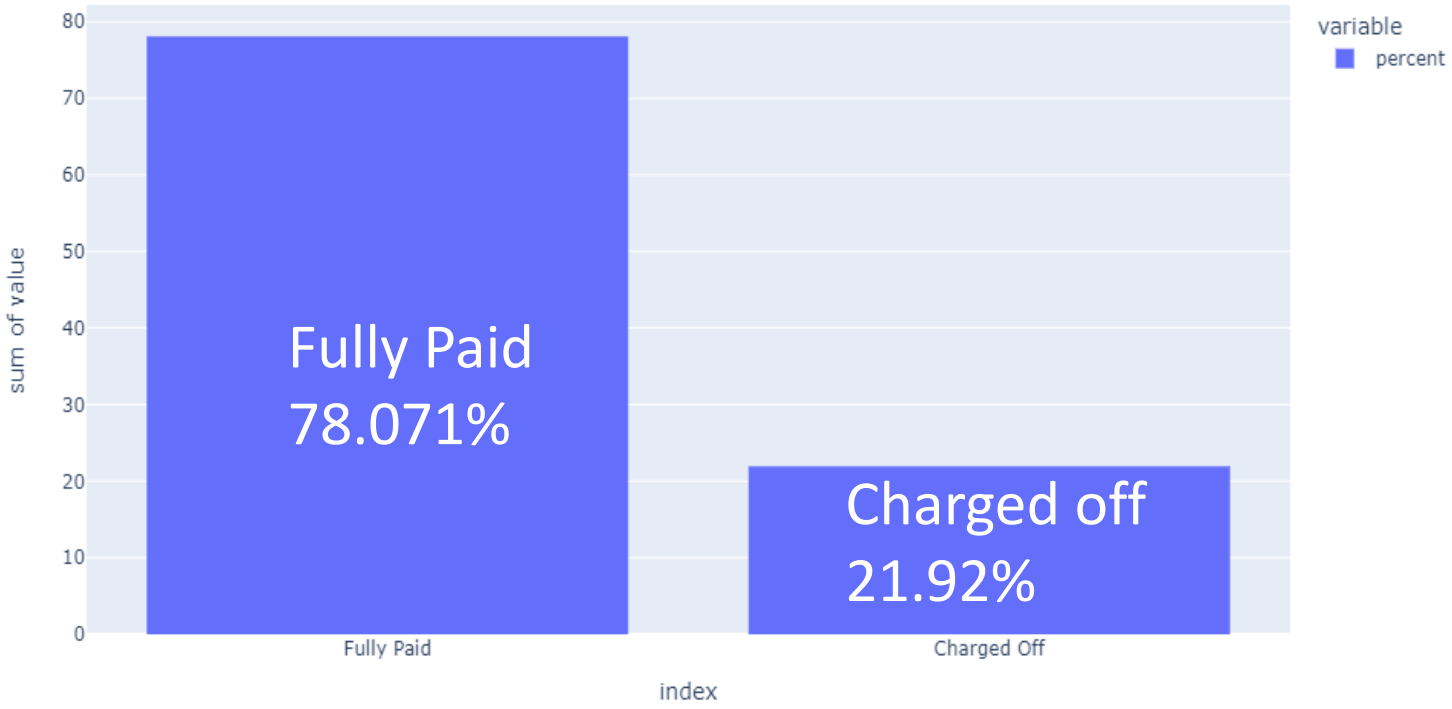
2. Exploración general de la tabla

- Dimensiones de la tabla y variables
- Analizar si las variables estarán disponibles en el momento de la llamada al modelo (sino se estaría introduciendo información a futuro en el modelo)
- Exploración de la variable objetivo
- Rápido análisis de valores nulos
- Se explora el número de variables numéricas y categóricas y se decide qué proceso realizar para tratarlas
- Transformaciones iniciales de algunas variables: formato de fechas, eliminar espacios de una variable string, etc

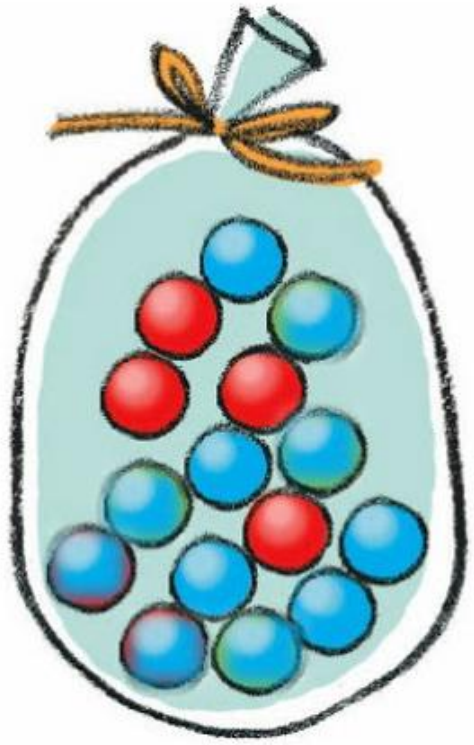
Análisis de la variable objetivo

Ejemplo: Modelo para conceder un préstamo bancario

Variable: loan_status



La probabilidad de sacar una observación aleatoria que tenga la etiqueta Charged off es 21.92%

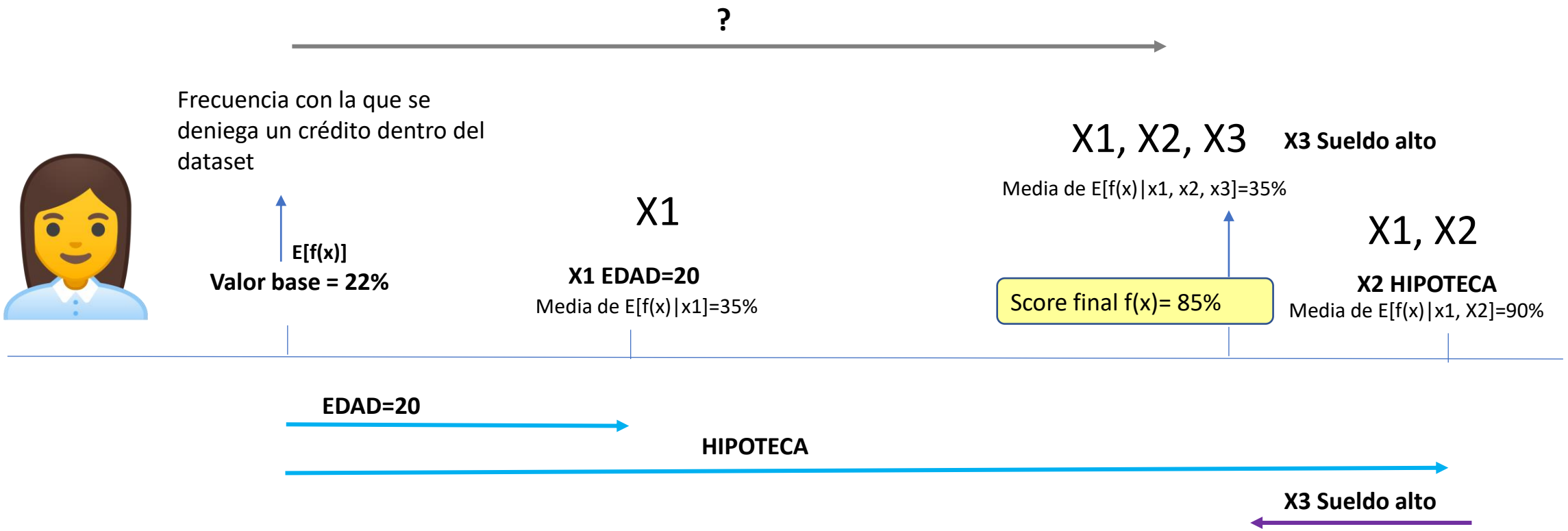


Fully Paid

Charged off

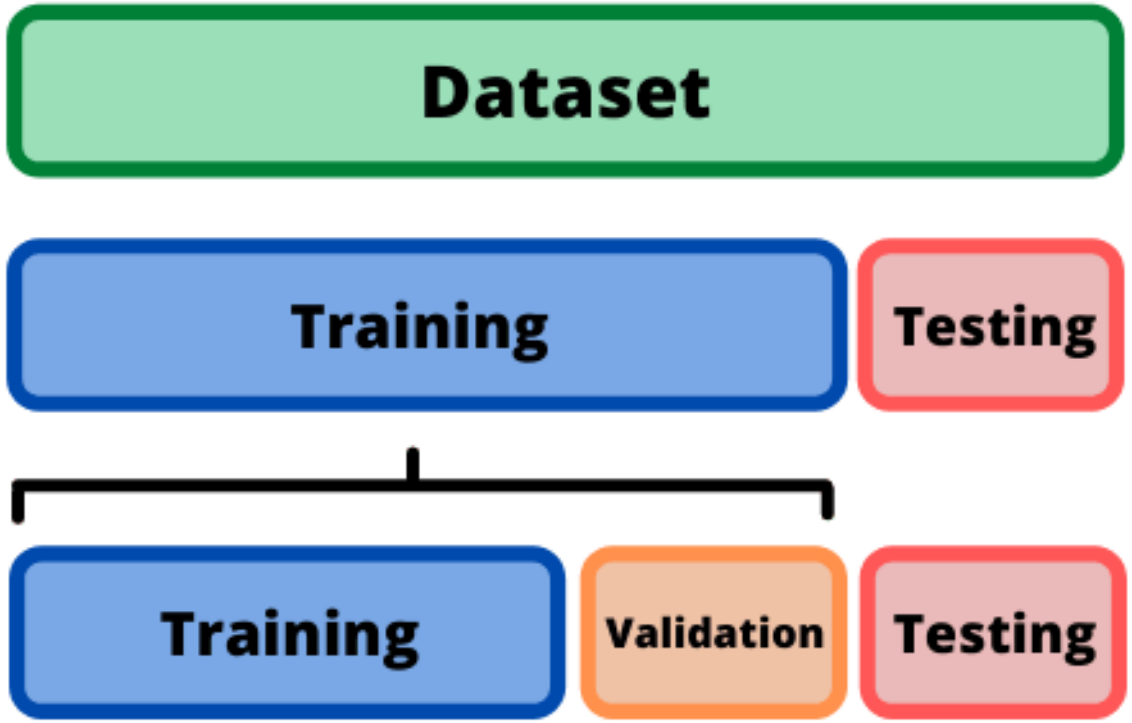
¿Cómo interpretar las probabilidades de un modelo?

- Con las variables explicativas, conseguimos que el porcentaje inicial aumente o disminuya, en función de las características de la persona
- Los modelos buscan la relación de las variables que mejor explican la variable objetivo y devuelven una probabilidad



3. Separación en Train y Test

¡Cuidado! En problemas desbalanceados, la separación de train y test hay que hacerla manteniendo porentajes de la variable objetivo (separación estratificada)



3. Separación en Train y Test

- Es necesario separar el dataset en train y test **antes** de realizar cualquier transformación que requiera de cálculos de la columna.
- Los valores missing y outlier se deben sustituir después de haber dividido en train y test.

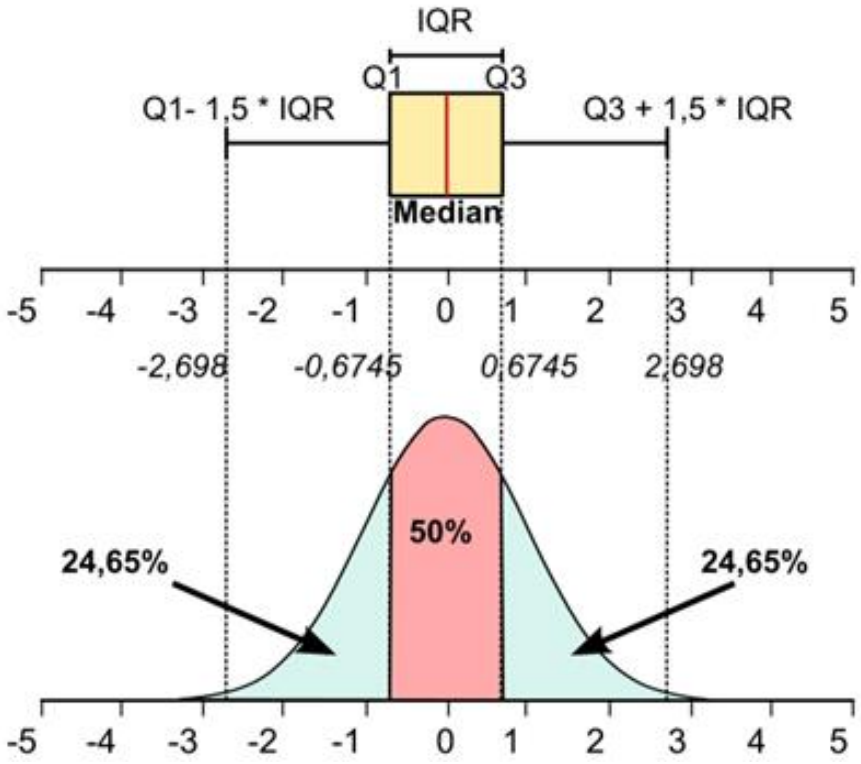
- Por ejemplo, si queremos imputar los valores missing por la media, los pasos a realizar son:

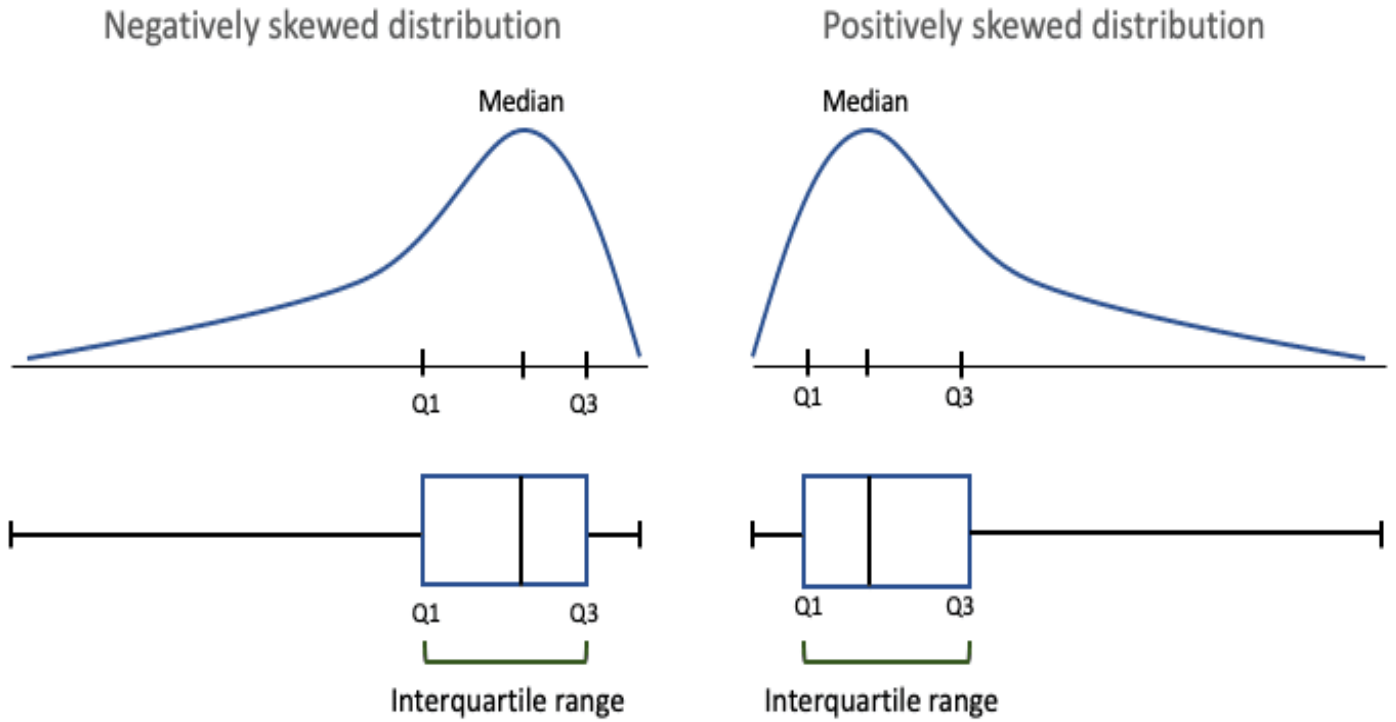
  1. Obtener la media del conjunto de train
  2. Reemplazar los valores missing por la media obtenida en la muestra de train en los conjuntos de train y test

4. Tratamiento de Outlier en datos estáticos

- Antes de tratarlos es necesario analizar la relación que tienen con la variable objetivo y analizar su contexto
- Es posible que algunos puedan ser valores mal introducidos (Por ejemplo: edad 400 años)

Detección de outliers por columnas (variables)

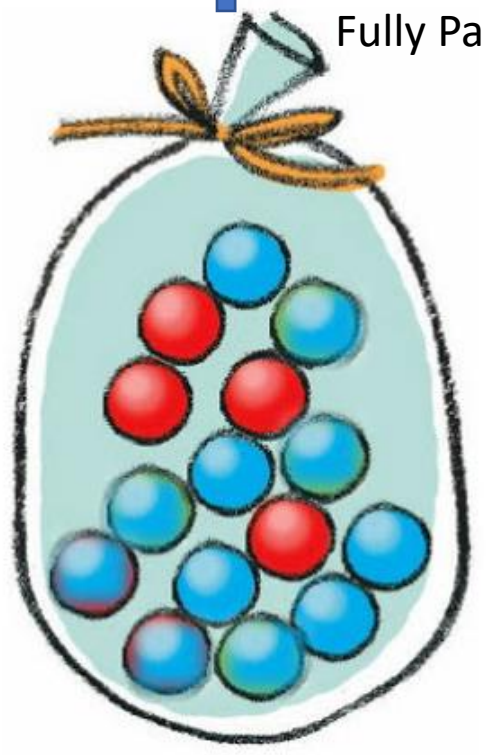




4. Tratamiento de Outlier en datos estáticos

- ¿En qué consiste analizar la relación que tienen los outlier con la variable objetivo?

La probabilidad de sacar una observación aleatoria que tenga la etiqueta Fully Paid es 78%



Fully Paid

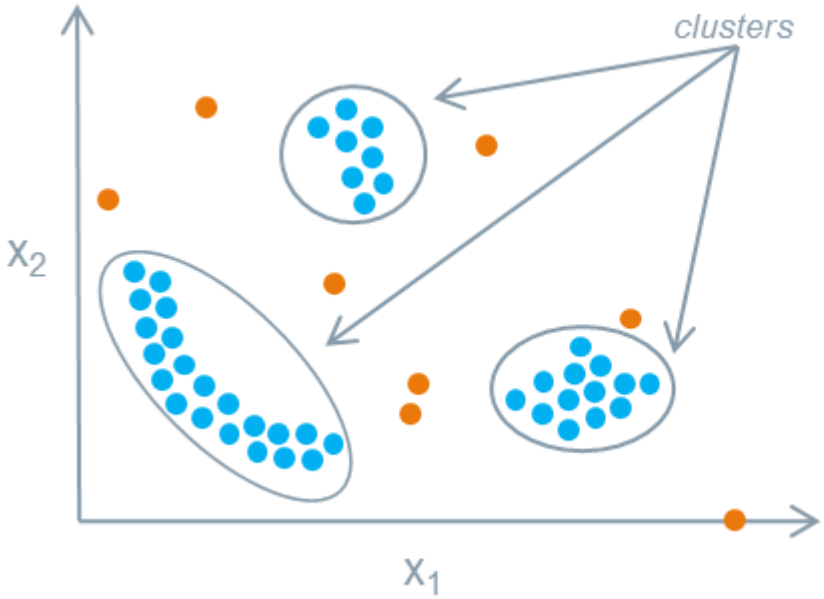
Charged off

Dentro de las filas que tienen outliers para una variable,
¿modifican los porcentajes de la variable objetivo de manera considerable?
En caso de que lo modifiquen, sirven para distinguir la variable objetivo.

4. Tratamiento de Outlier en datos estáticos

- Detectar anomalías por filas (por ejemplo clientes) es un proceso más arduo que por columnas
- Antes de tratarlos es necesario analizar la relación que tienen con la variable objetivo
- Hay muchos proyectos cuyo objetivo es detectar anomalías: anomalías en clientes, facturas, transferencias, movimientos en cuentas

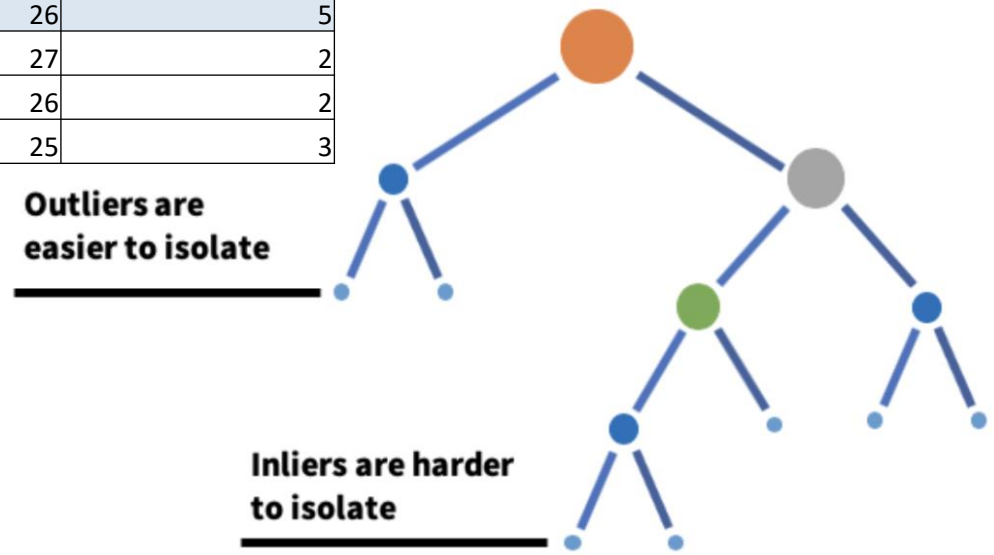
Clustering: DBSCAN



Detección de outliers por filas

Customer	Salary	age	number_accounts
Ana	30K	25	1
Sergio	60K	26	5
Nuria	35K	27	2
David	25K	26	2
Natalia	32K	25	3

Isolation Forest



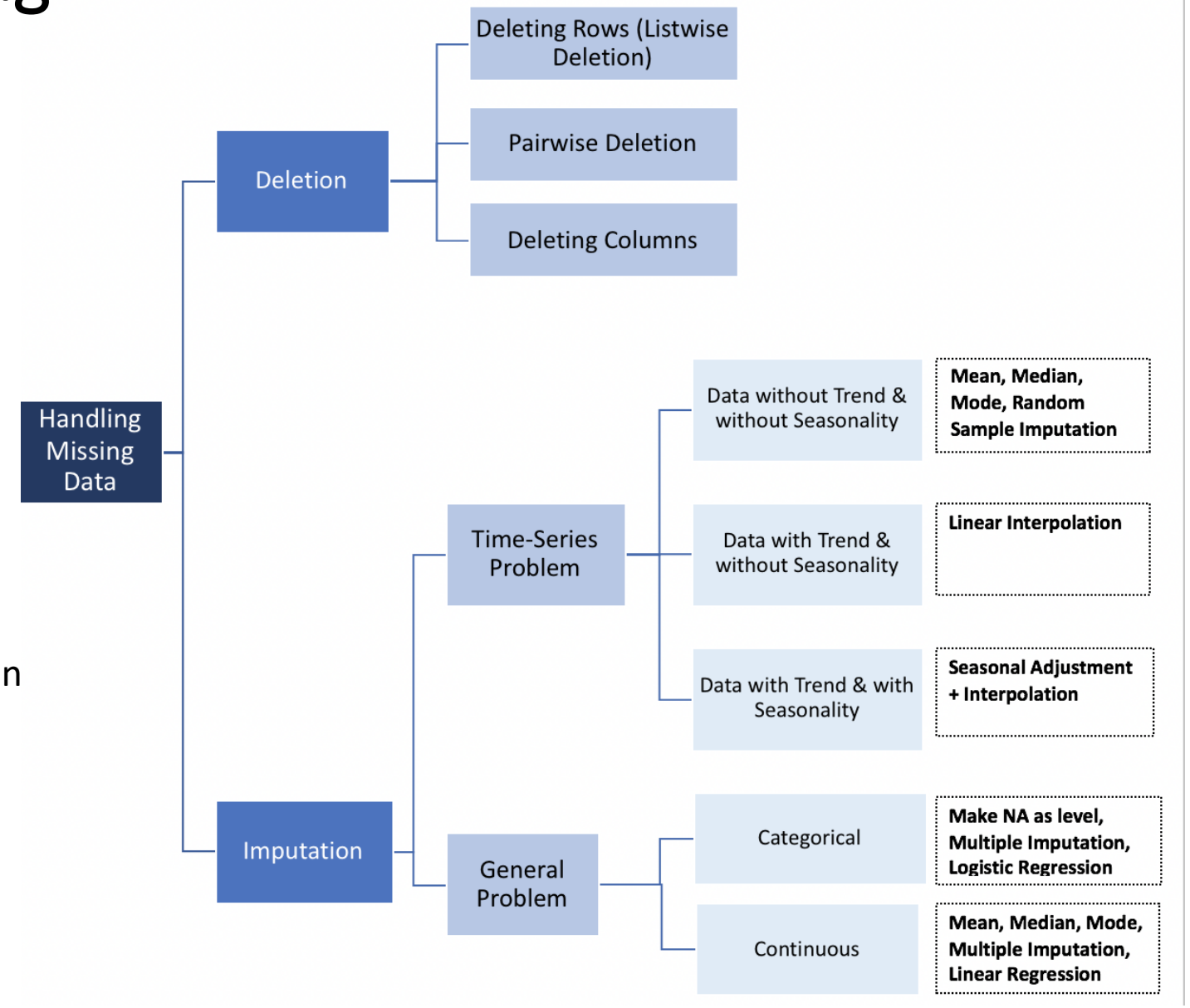
5. Tratamiento de missing

Row no	State	Salary	Yrs of Experience
1	NY	57400	Mid
2	TX		Entry
3	NJ	90000	High
4	VT	36900	Entry
5	TX		Mid
6	CA	76600	High
7	NY	85000	High
8	CA		Entry
9	CT	45000	Entry

Missing values

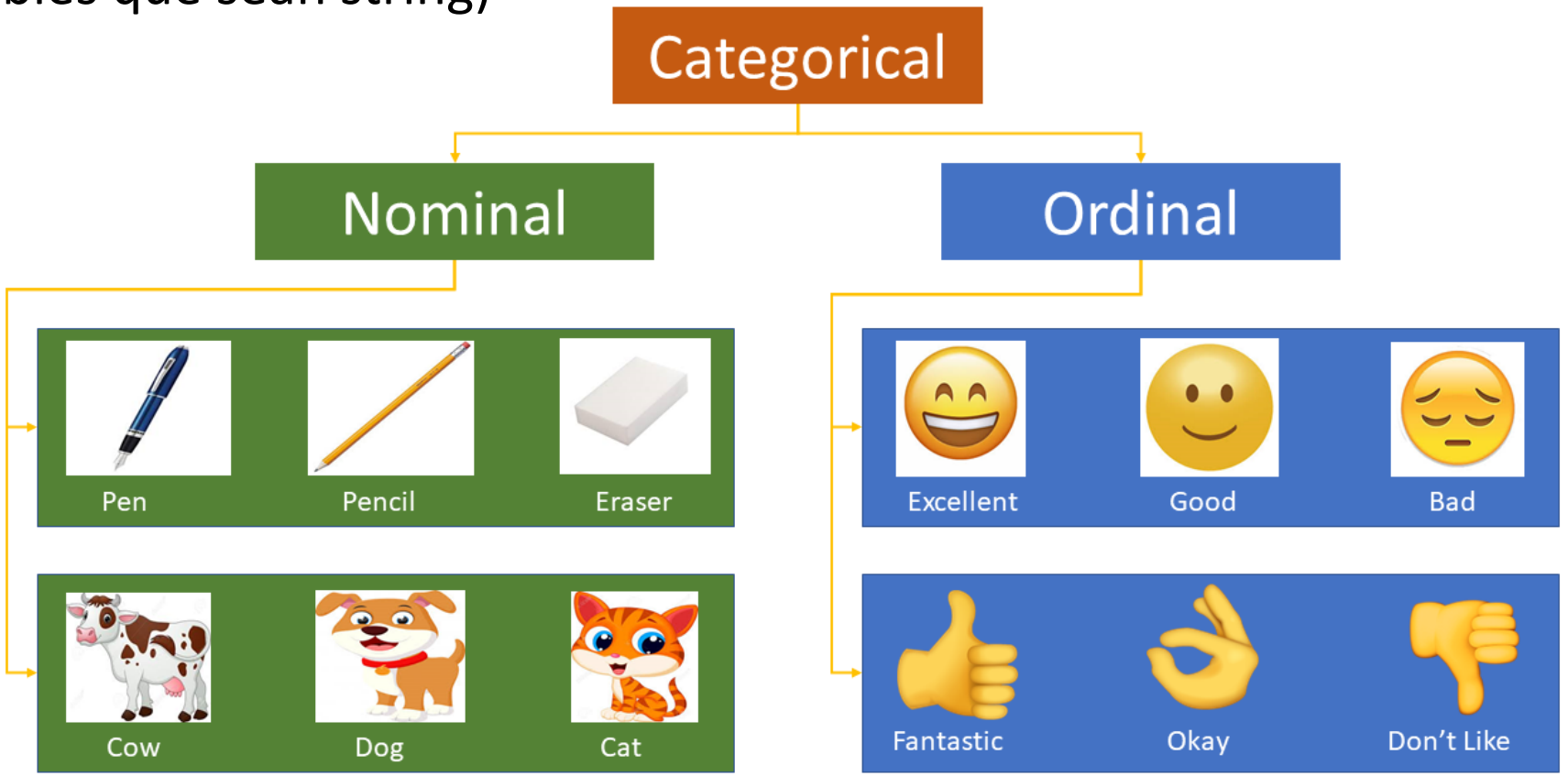
- Antes de imputarlos o eliminarlos, es necesario analizar la relación que tienen con la variable objetivo

- También es interesante analizar los valores missing por filas



6. Codificación de variables categóricas

- Muchos algoritmos requieren que su input sea completamente numérico (sin tener variables que sean string)



*Se deben usar técnicas de codificación para tratarlas

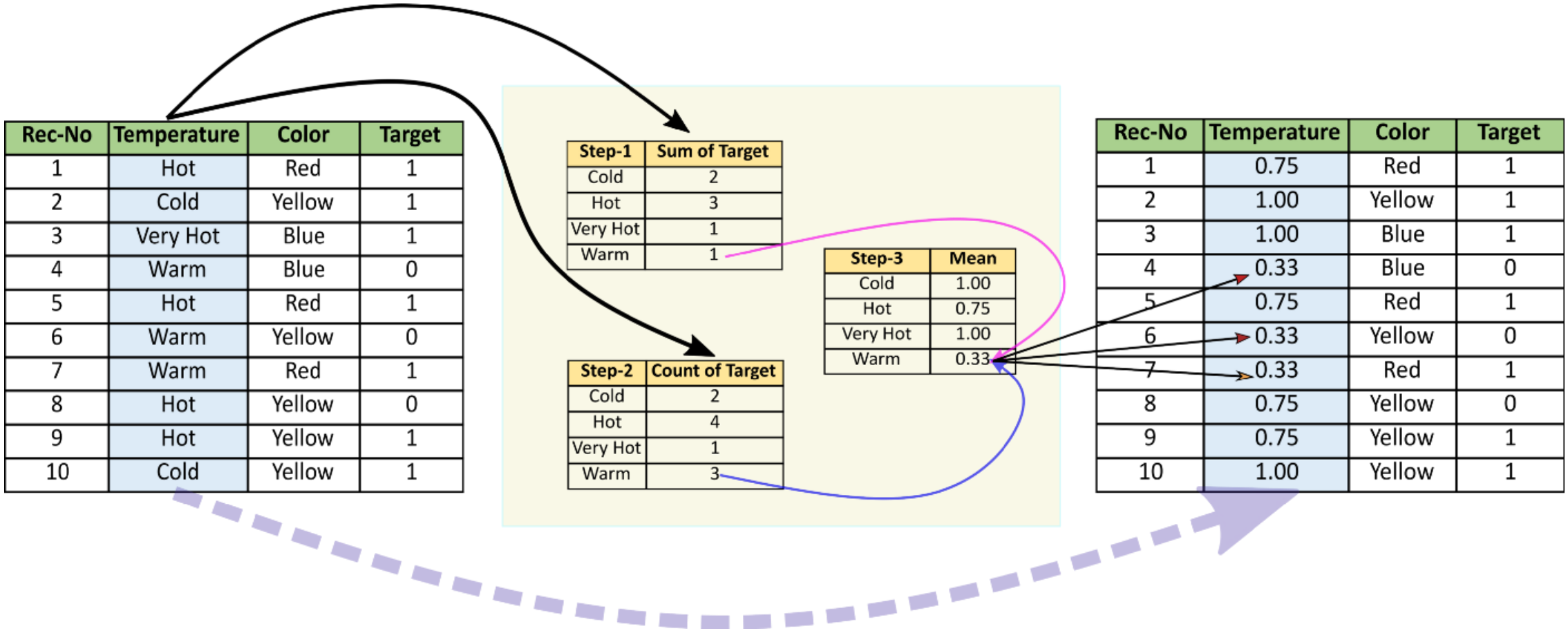
Se pueden convertir a numéricas poniendo un número ordenado. Por ejemplo: Excellent=3, Good=2, bad=1

6. One hot encoding

Temperature	Color	Target
Hot	Red	1
Cold	Yellow	1
Very Hot	Blue	1
Warm	Blue	0
Hot	Red	1
Warm	Yellow	0
Warm	Red	1
Hot	Yellow	0
Hot	Yellow	1
Cold	Yellow	1

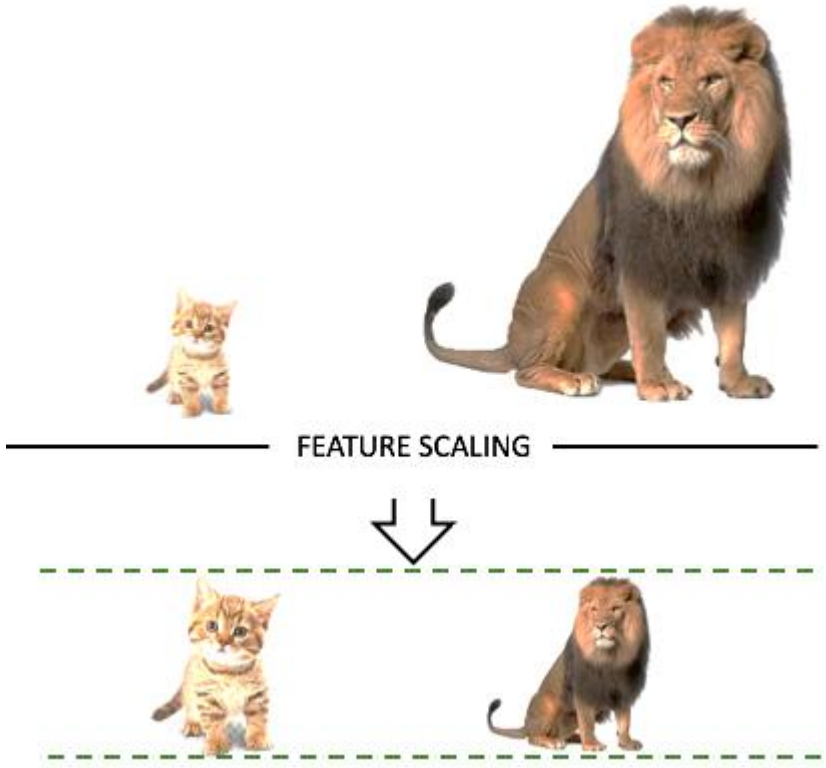
Temperature	Color	Target	Temp_Cold	Temp_Hot	Temp_Very_Hot	Temp_Warm
Hot	Red	1	0.0	1.0	0.0	0.0
Cold	Yellow	1	1.0	0.0	0.0	0.0
Very Hot	Blue	1	0.0	0.0	1.0	0.0
Warm	Blue	0	0.0	0.0	0.0	1.0
Hot	Red	1	0.0	1.0	0.0	0.0
Warm	Yellow	0	0.0	0.0	0.0	1.0
Warm	Red	1	0.0	0.0	0.0	1.0
Hot	Yellow	0	0.0	1.0	0.0	0.0
Hot	Yellow	1	0.0	1.0	0.0	0.0
Cold	Yellow	1	1.0	0.0	0.0	0.0

6. Target Encoding



7. Escalado de las variables

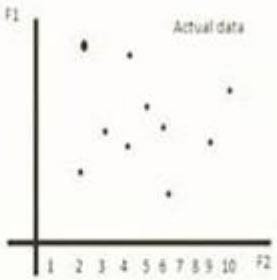
- Una vez que todas las variables son numéricas, algunos algoritmos necesitan que las variables estén escaladas
- Escalar las variables consiste en poner todas en la misma escala numérica. Es decir, por ejemplo, que todos los valores de todas las variables estén entre 0 y 1

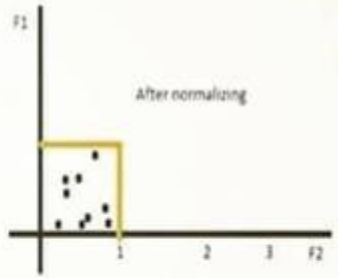


7. Escalado de las variables

❑Normalization

$$X_{changed} = \frac{X - X_{min}}{X_{max} - X_{min}}$$





❑Standardization

$$X_{changed} = \frac{X - \mu}{\sigma}$$

bibliografía

https://www.avenga.com/magazine/anomaly-detection/

https://towardsdatascience.com/dbscan-a-density-based-unsupervised-algorithm-for-fraud-detection-887c0f1016e9

https://cs.nju.edu.cn/zhouzh/zhouzh.files/publication/icdm08b.pdf?q=isolation-forest

https://medium.com/analytics-vidhya/target-encoding-vs-one-hot-encoding-with-simple-examples-276a7e7b3e64

https://scikit-learn.org/stable/

https://pypi.org/project/category-encoders/