

CURSOS ANALYTICS

Machine Learning Advanced

- Aprendizaje Supervisado -

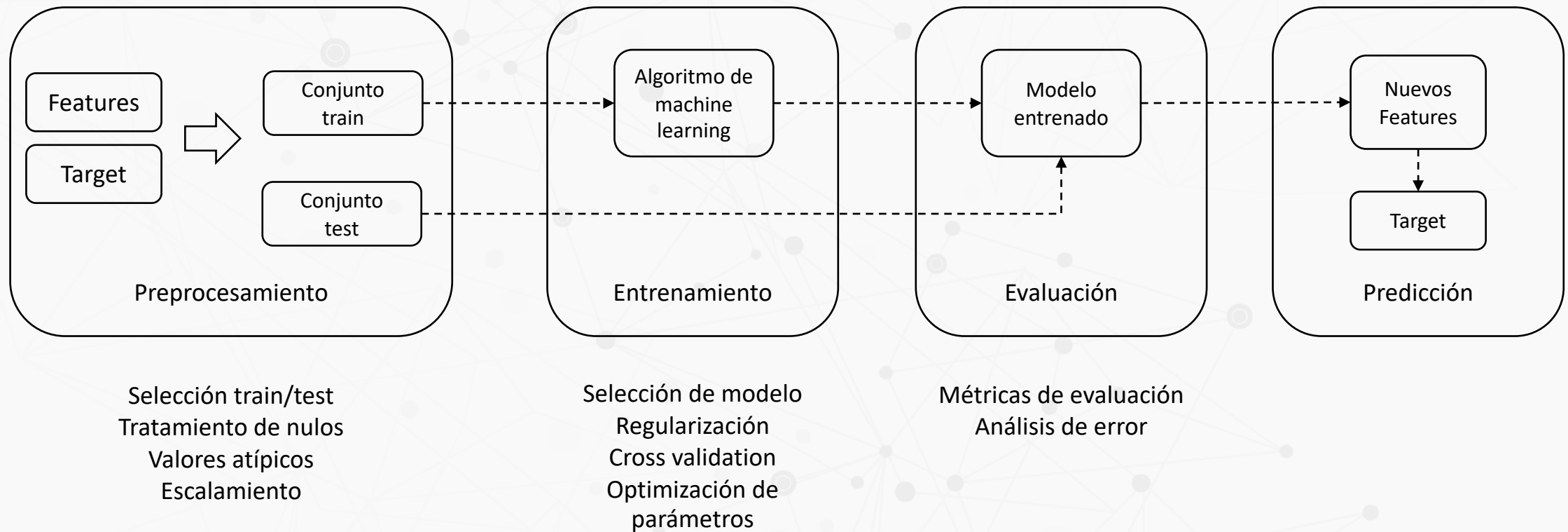
Docente: Manuel Montoya



Preprocesamiento

Cómo construir buenos conjuntos de entrenamiento

Aprendizaje supervisado: Proceso de modelamiento



División en entrenamiento y prueba

- ¿Cómo podemos saber si nuestro modelo va a funcionar bien fuera de la data de entrenamiento?
- ¿Cómo sabemos si no estamos sobreajustando a la data de entrenamiento?
- ¿Cómo podemos evitar los efectos de la estacionalidad?

Entrenamiento

Data de todo el 2020

Test

Primeros meses 2021

- Podemos utilizar la data de un año completo para entrenar nuestro modelo y como prueba “fuera de tiempo” utilizar data de los siguientes tres o seis meses.

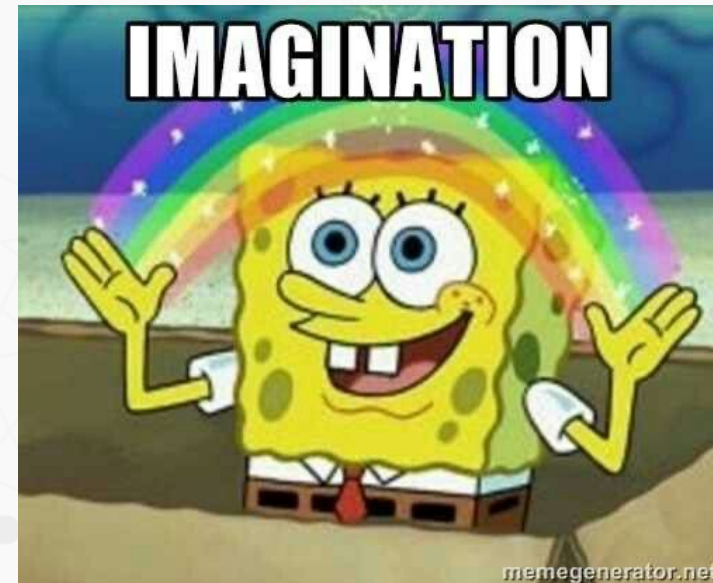
Generación de nuevas features

- Usualmente las variables que tenemos disponibles en nuestra base de datos no son suficientes para el desarrollo de nuestro modelo predictivo.
- ¿Podemos crear nuevas variables a partir de las variable iniciales?

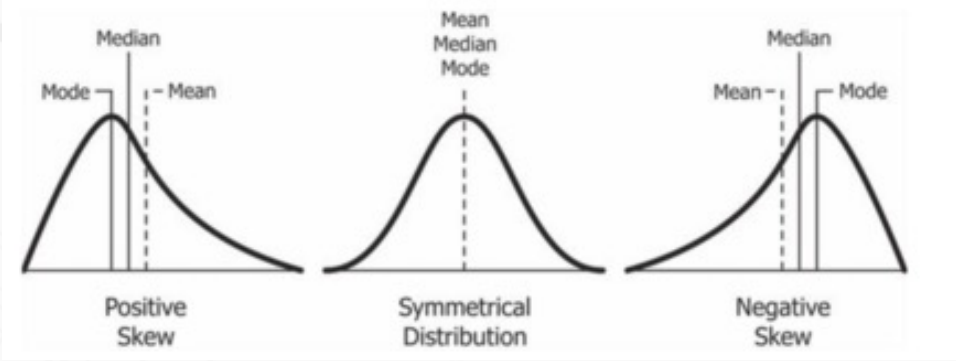
Generación de nuevas features

- Usualmente las variables que tenemos disponibles en nuestra base de datos no son suficientes para el desarrollo de nuestro modelo predictivo.
- ¿Podemos crear nuevas variables a partir de las variable iniciales?
- Promedio de saldo por mes en los últimos x meses
- ¿Nuestro cliente recibe su sueldo en soles o dólares?
- ¿Nuestro cliente suele realizar compras con su tarjeta de crédito en soles o dólares?
- ¿Paga mensualmente algún crédito en dólares?
- ...

Transformaciones: Logaritmos, exponenciación, radicación, normalización, estandarización



Análisis univariado

- De todas las variables que he generado, ¿cuáles contienen información suficiente para mi población?
 - Debemos tener cuidado con aquellas variables que tengan un porcentaje muy alto de valores nulos o ceros
 - Debemos tener cuidado con aquellas variables que no presenten mayor varianza entre los distintos registros
- 
- The diagram shows three bell curves illustrating different types of data distributions:
- Positive Skew:** The curve is skewed to the right. The Mode is at the peak, followed by the Median, and then the Mean is furthest to the right.
 - Symmetrical Distribution:** The curve is a standard bell shape. The Mean, Median, and Mode are all at the same central point.
 - Negative Skew:** The curve is skewed to the left. The Mean is furthest to the left, followed by the Median, and then the Mode is at the peak.
- Porcentajes de nulos, porcentajes de ceros
 - **Medidas de centralidad:** media, mediana
 - **Medidas de dispersión:** desviación estándar, coeficiente de variación
 - **Distribución:** percentiles, rango intercuartil

Imputación de datos faltantes

- El método de imputación depende de la **naturaleza** de la variable
- Si tengo valores nulos en la edad de clientes, ¿con qué valor debería imputar?
- Si tengo valores nulos en la cantidad de transacciones de un cliente, ¿con qué valor debería imputar?

	col1	col2	col3	col4	col5		col1	col2	col3	col4	col5	
0	2	5.0	3.0	6	NaN	mean()	0	2.0	5.0	3.0	6.0	7.0
1	9	NaN	9.0	0	7.0		1	9.0	11.0	9.0	0.0	7.0
2	19	17.0	NaN	9	NaN		2	19.0	17.0	6.0	9.0	7.0

Imputación de datos faltantes

Variables numéricas

Media o mediana (cuidado con la media y los valores extremos)

Variables categóricas

Moda

Otras técnicas

- Utilizar valores promedios por grupos
- Imputación por vecinos más cercanos

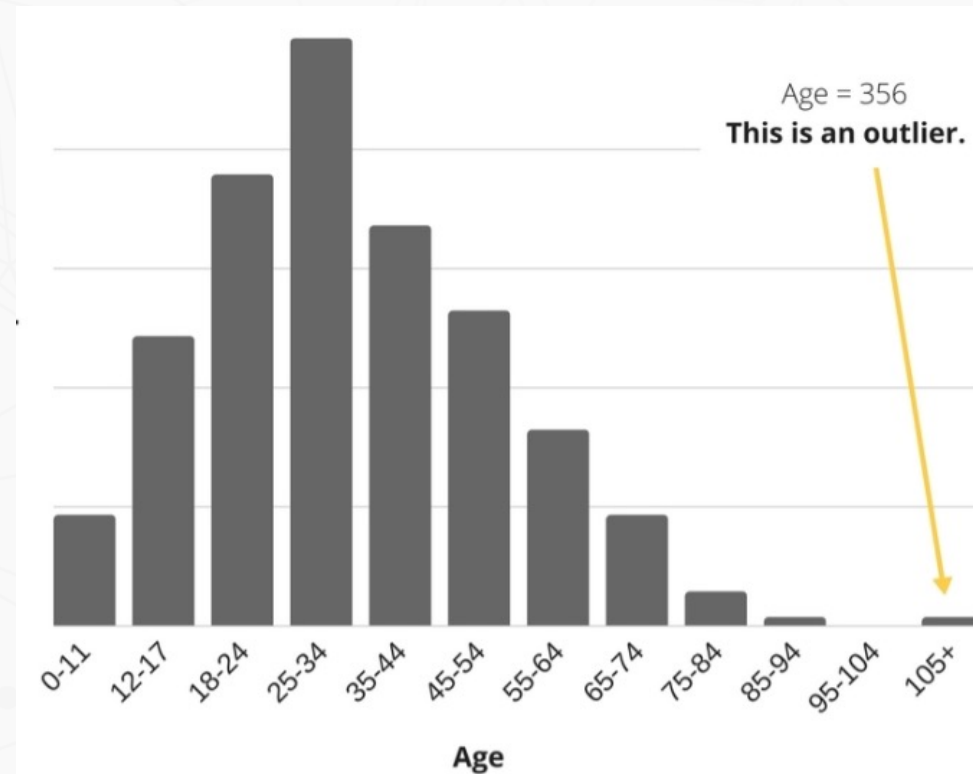
Games	Name	Sex	Height	Team	Sport
1984 Summer	Harri Hilmer Huhtala	M	187.0	Finland	Athletics
1912 Summer	Ladislav emla	M	NaN	Bohemia	Tennis
1972 Summer	Edvard Mikayelyan	M	165.0	Soviet Union	Gymnastics
1980 Winter	Friedrich "Fritz" Fischer	M	181.0	West Germany	Biathlon
1992 Winter	Rachel Lynn Mayer (-Godino)	F	NaN	United States-2	Figure Skating
2008 Summer	Ioanna Samara	F	164.0	Greece	Rhythmic Gymnastics
1972 Summer	Paul Thijs	M	174.0	Belgium	Athletics
1988 Summer	Mria uriinov	F	169.0	Czechoslovakia	Handball
2004 Summer	Corinne Raux	F	164.0	France	Athletics
1972 Summer	Mara Teresa Ramrez Gmez	F	171.0	Mexico	Swimming

Tratamiento de valores atípicos

Los valores atípicos pueden tener efectos negativos en el entrenamiento de modelos, principalmente aquellos que utilizan el cálculo de distancias para el aprendizaje

¿Cómo lidiar con valores extremos?

- **Eliminarlos:** puedo perder información
- **Marcarlos:** opción más segura
- **Reescalar:** reducir el efecto de los valores extremos



Cómo identificar valores atípicos

Validación de valores correctos

- Validación de distribuciones con equipos de negocio
- Validación de que las variable tenga los rangos adecuados

¿De forma automática?

- Observar los percentiles extremos:
Mínimos: 0, 2.5, 5, 10
Máximos: 90, 95, 97.5, 99, 100

	LoanAmount	LoanAmountTerm
count	592.000000	600.00000
mean	146.412162	342.00000
std	85.587325	65.12041
min	9.000000	12.00000
5%	56.000000	180.00000
10%	71.000000	294.00000
15%	84.650000	360.00000
20%	95.000000	360.00000
25%	100.000000	360.00000

Cómo identificar valores atípicos

Validación de valores correctos

- Validación de distribuciones con equipos de negocio
- Validación de que la variable tenga los rangos adecuados

¿De forma automática?

- Observar los percentiles extremos:
Mínimos: 0, 2.5, 5, 10
Máximos: 90, 95, 97.5, 99, 100

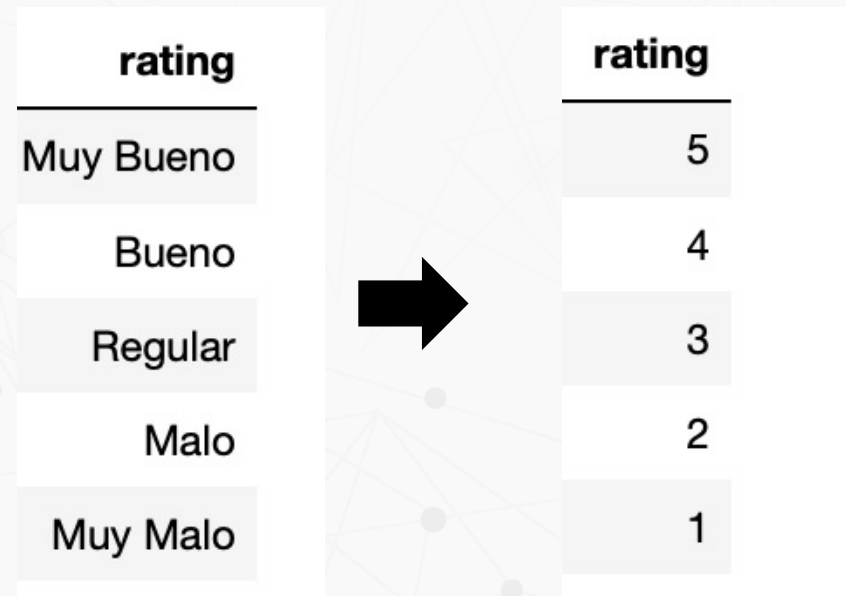
	ApplicantIncome	CoapplicantIncome
91%	9940.900000	3984.020000
92%	10397.240000	4230.560000
93%	11155.360000	4416.000000
94%	12110.000000	4596.100000
95%	14583.000000	4997.400000
96%	15364.320000	5397.040000
97%	16682.250000	5682.670000
98%	19666.040000	7198.560000
99%	32540.410000	8895.890000
max	81000.000000	41667.000000

Labeling

Para convertir una variable categórica **ordinal** a numérica se utiliza el *labeling*

Este proceso se utiliza cuando existe una noción de distancia entre las categorías


Se debe tener cuidado en que el orden de los números utilizados sea el mismo orden que representan las variables categóricas originales



rating	rating
Muy Bueno	5
Bueno	4
Regular	3
Malo	2
Muy Malo	1

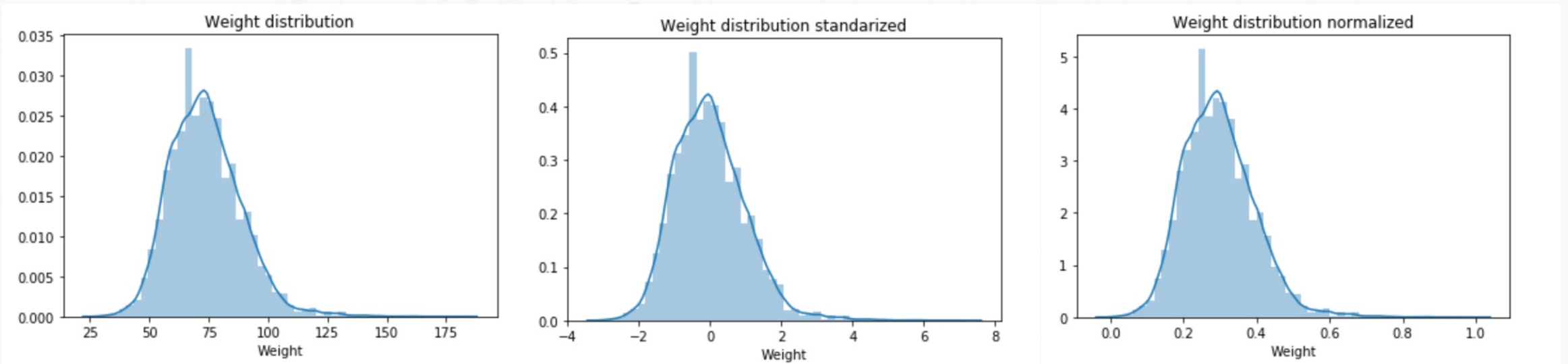
One Hot Encoding

Cuando la variable categórica es **nominal**, se recomienda utilizar one-hot-encoding

department			department_Analytics	department_Operations	department_Sales & Marketing	department_Technology
0	Sales & Marketing		0	0	1	0
1	Operations		0	1	0	0
2	Sales & Marketing		0	0	1	0
3	Sales & Marketing		0	0	1	0
4	Technology		0	0	0	1
5	Analytics		1	0	0	0
6	Operations		0	1	0	0
7	Operations		0	1	0	0
8	Analytics		1	0	0	0
9	Sales & Marketing	0	0	1	0	

Escalamiento

Nos permite que las variables que ingresan al modelo se encuentren en la misma escala.
Se mantiene la distribución pero con los valores transformados a una escala determinada



Variable original x

$$x_z = \frac{x - \mu}{\sigma}$$

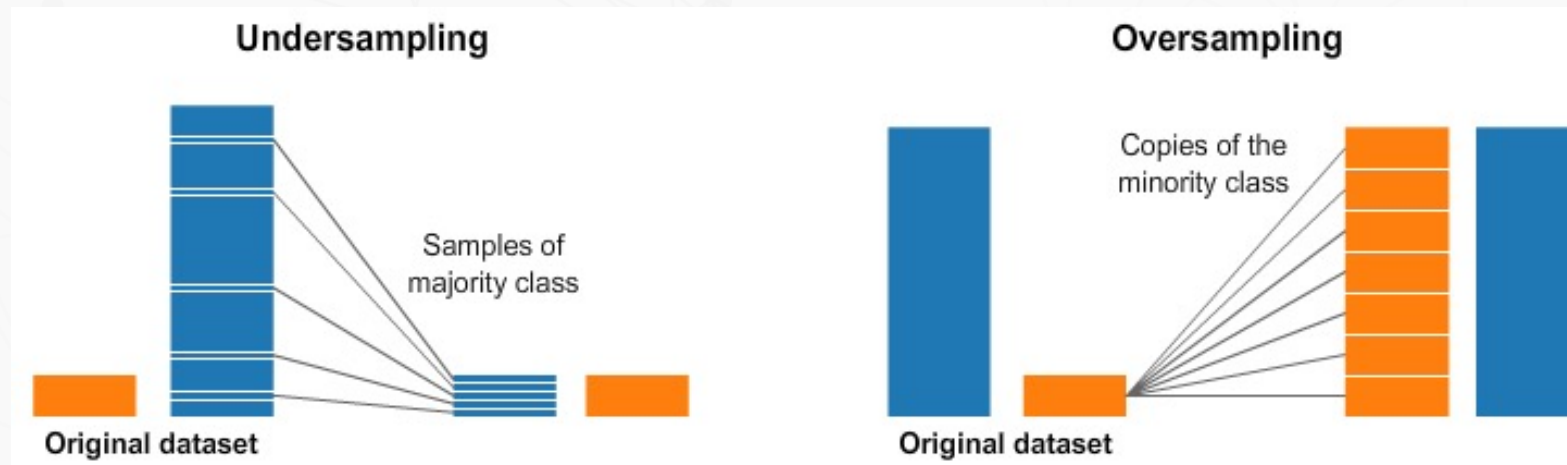
Estandarización normal

$$x_n = \frac{x - \min}{\max - \min}$$

Normalización Min - Max

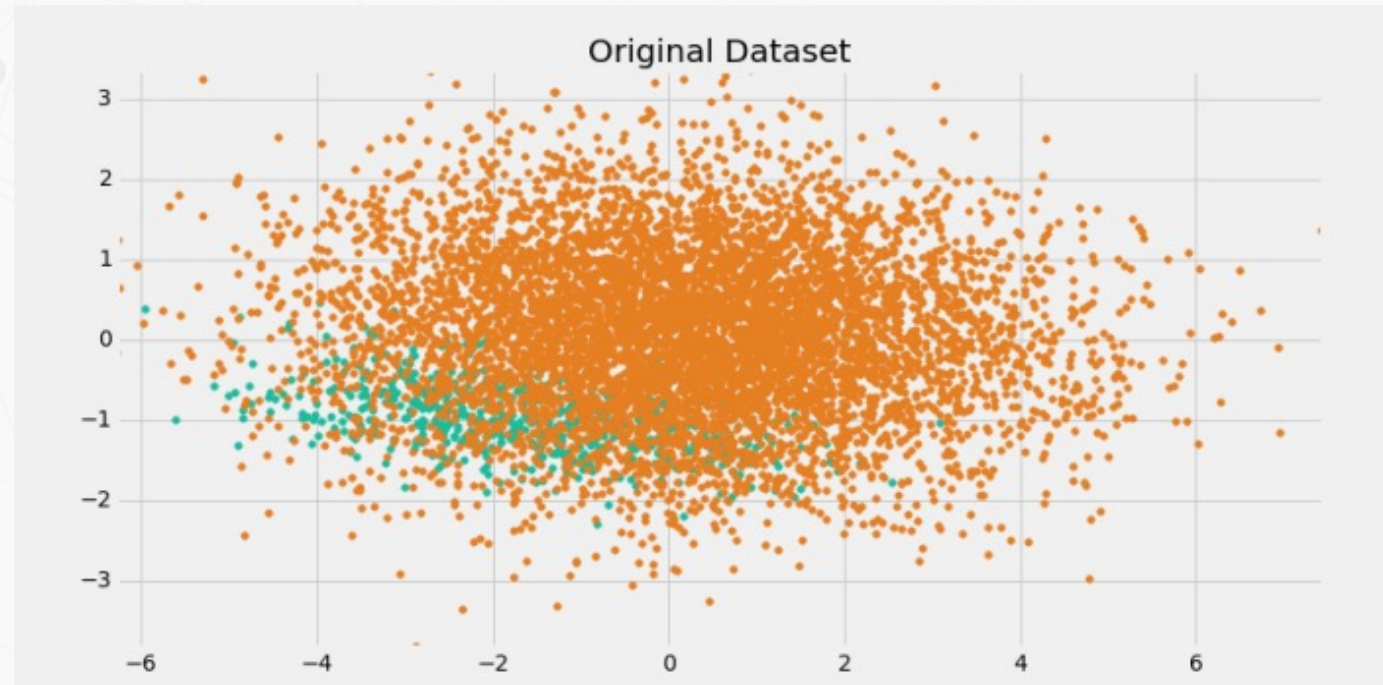
Balanceo de clases

- Si solo el 1% de clientes de nuestra base de datos realiza transacciones de tipo de cambio, ¿podemos entrenar un modelo predictivo?
- Es necesario balancear los datos para que los modelos entrenados puedan detectar los eventos que queremos predecir



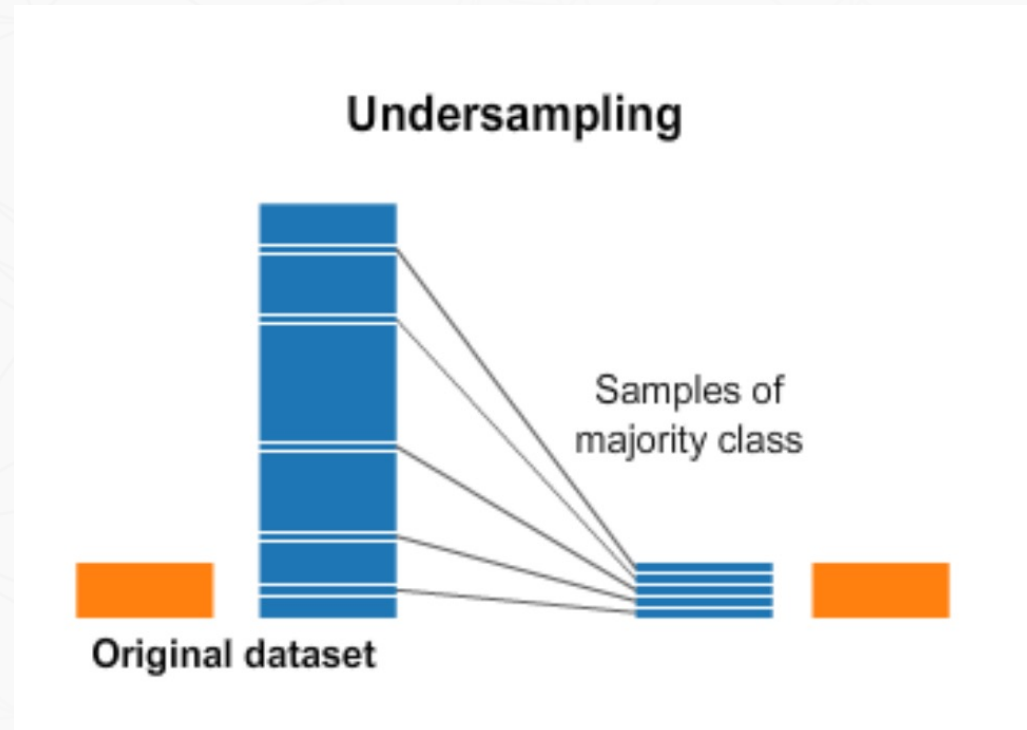
Clases desbalanceadas

- Un dataset está desbalanceado cuando una de las clases está subrepresentada respecto a las otras.
- **Problema:** La mayoría de técnicas de machine learning van a ignorar y tener mala performance en la clase minoritaria cuando comúnmente esta es la clase de interés para el desarrollo del modelo.



Estrategias de balanceo: Undersampling

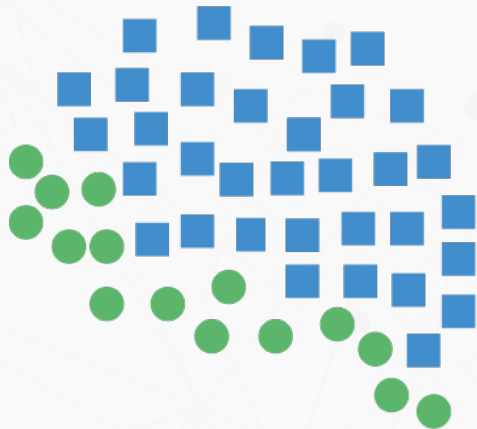
- Se eliminan registros de la clase mayoritaria de forma aleatoria hasta que se obtenga la proporción deseada
- Se podrían perder registros importantes para el desarrollo del modelo



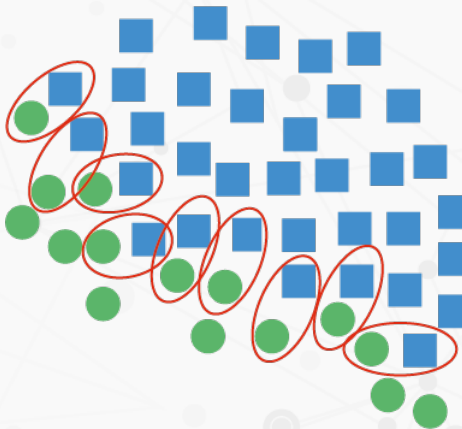
Estrategias de balanceo: Tomek Link

Elimina registros de la clase mayoritaria que se encuentren cerca a la frontera de decisión de las clases

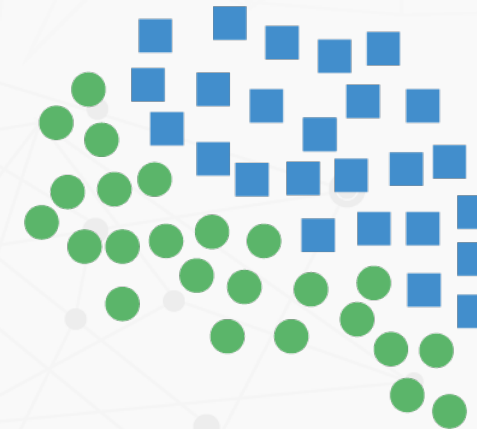
TomekLinks



Original Dataset



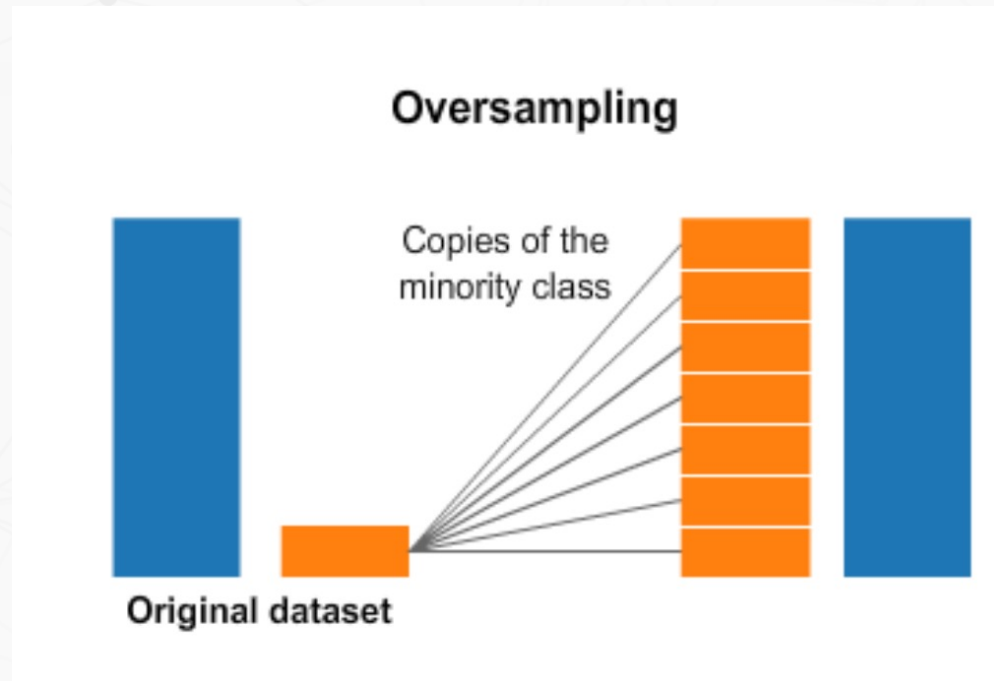
Finding TomekLinks



Resampled Dataset

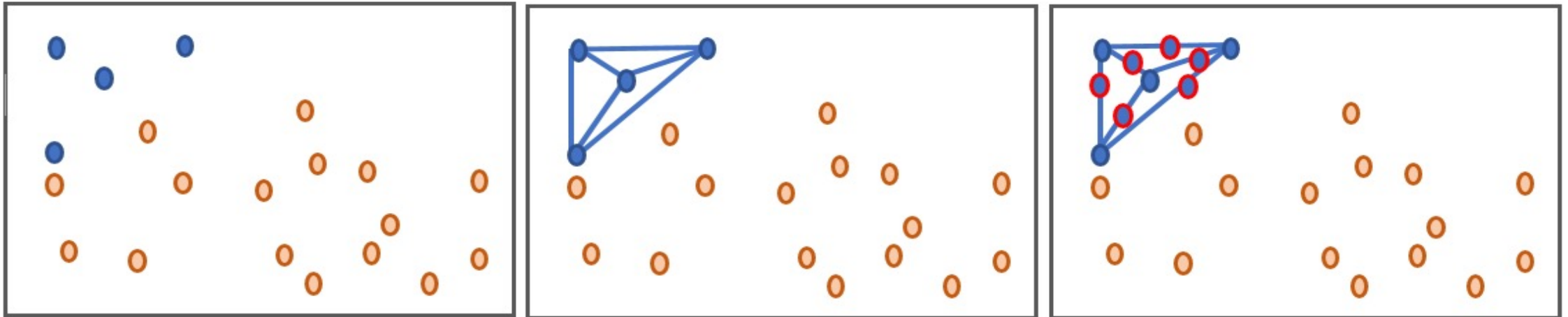
Estrategias de balanceo: Oversampling

- Se replican registros de la clase minoritaria de forma aleatoria hasta que se obtenga la proporción deseada
- Se podrían generar sobreestimación o subestimación dependiendo de la aleatoriedad



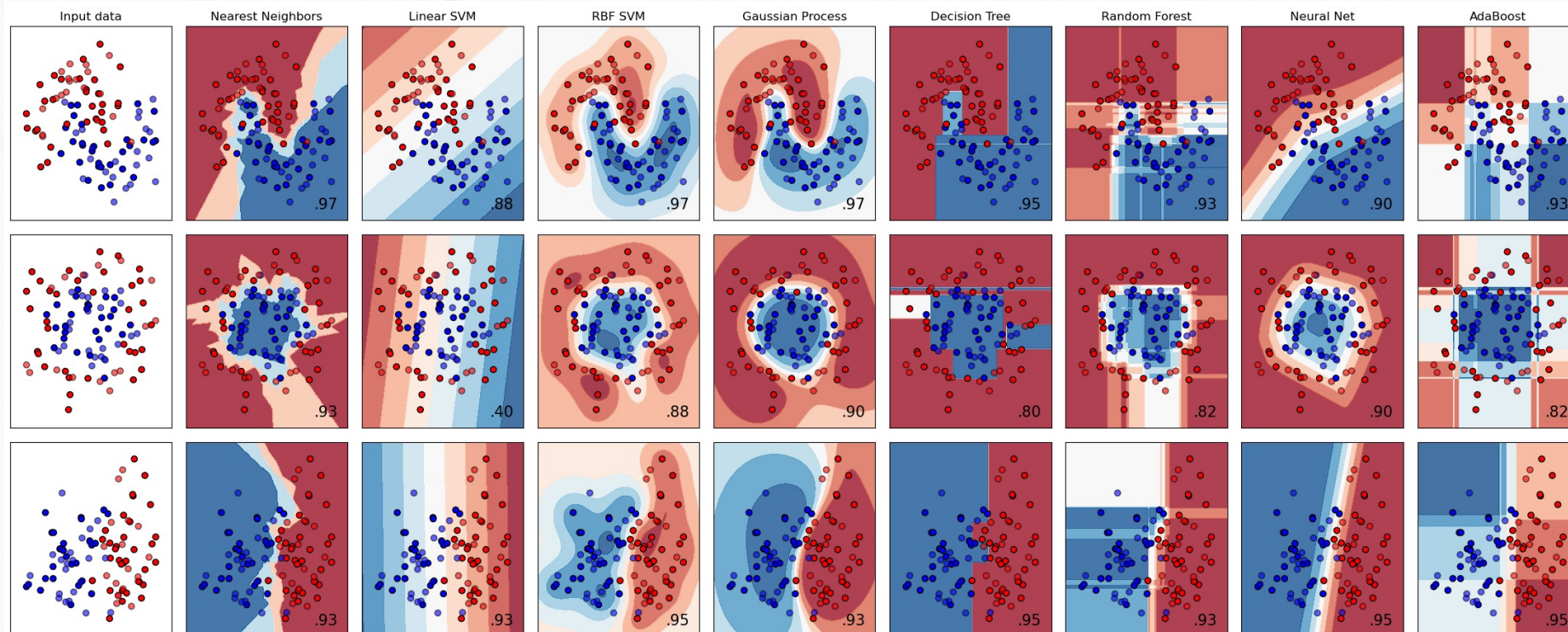
Estrategias de balanceo: Smote

Selecciona una instancia de la clase minoritaria aleatoriamente y encuentra k vecinos cercanos de la misma clase. Los registros nuevos son creados eligiendo uno de los k vecinos y creando una combinación convexa entre ambos.



Entrenamiento

- Con el conjunto de datos preprocesado, las features generadas y seleccionadas, y el target balanceado ya estamos listos para entrenar nuestro modelo predictivo



Referencias

- Python Machine Learning – Sebastian Raschka (Tercera edición):
<https://github.com/rasbt/python-machine-learning-book-3rd-edition>
- Documentación de scikit-learn: <https://scikit-learn.org/stable/>
- Machine learning flashcards – Chris Albon: <https://machinelearningflashcards.com>
- Interpretable Machine Learning – Cristoph Molnar: Interpretable Machine Learning:
<https://christophm.github.io/interpretable-ml-book/>

Lecturas recomendadas

- Python Machine Learning – Capítulo 1: Giving computers the ability to learn from data
- Data Science for Business – Capítulo 1: Introduction: Data-Analytic Thinking
- Data Science for Business – Capítulo 2: Business Problems and Data Science Solutions



¡Gracias!