

MACHINE LEARNING INMERSION

ANDRÉ OMAR CHÁVEZ PANDURO

« La **virtud de cada ser humano** , es saber mantener el **equilibrio ante sus victorias** y **no caerse** ante sus derrotas»



EXPOSITOR

André Omar Chávez Panduro
UNMSM

MSc in Data Science Candidate
Promotion “Erwin Kraenau Espinal”
Universidad Ricard Palma



**Senior Data
Scientist**



**Customer
Intelligence Analyst**



Data Analyst



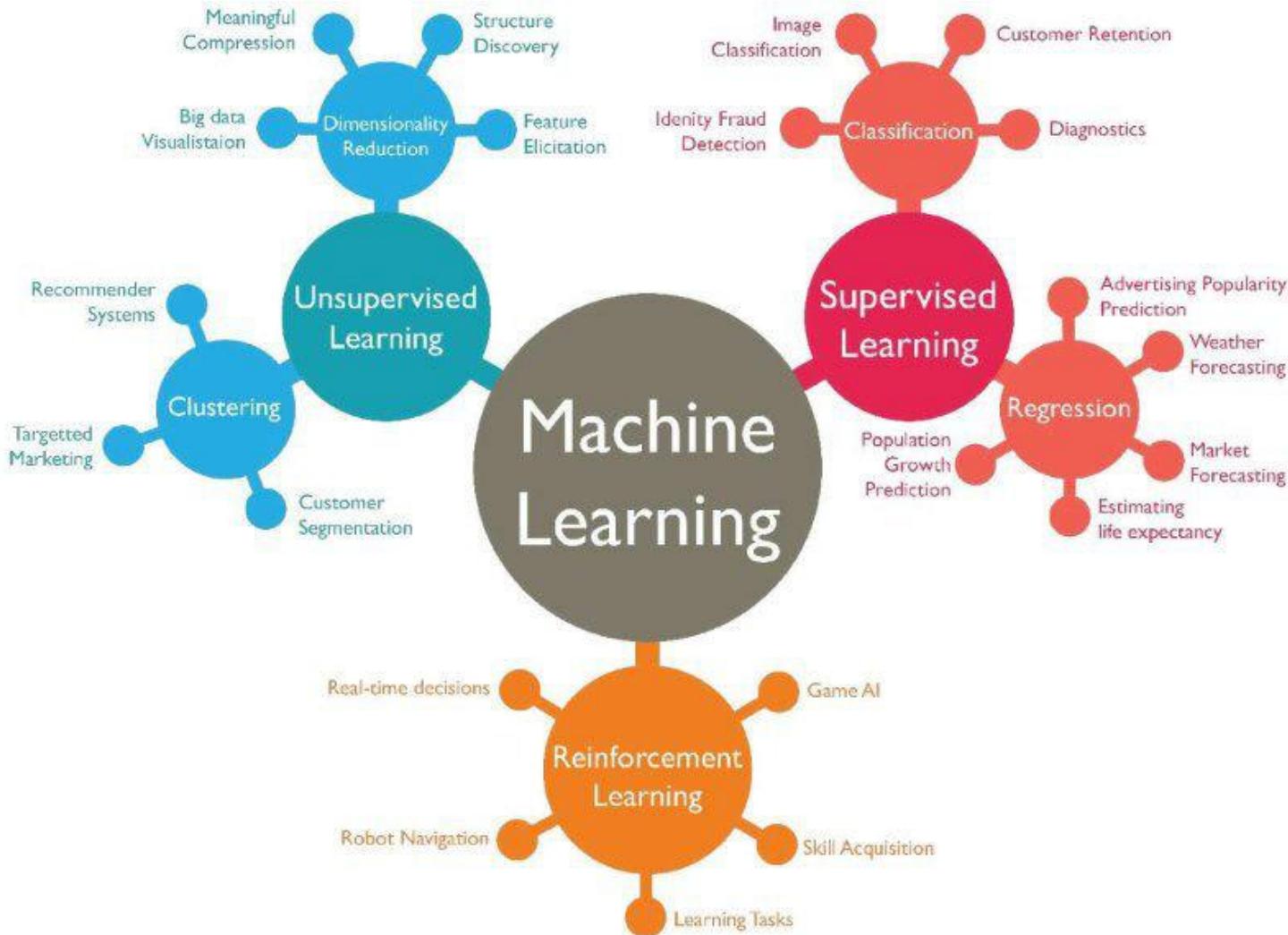
AGENDA



- Entendimiento del Problema.
- Definición de la Variable Target.
- Definición y Creación de Drivers.
- Elección y Definición del Horizonte temporal
(Ventana de Análisis).
- Criterios de Inclusión y Exclusión de la Información
- Mitos y Errores



INTRODUCCIÓN



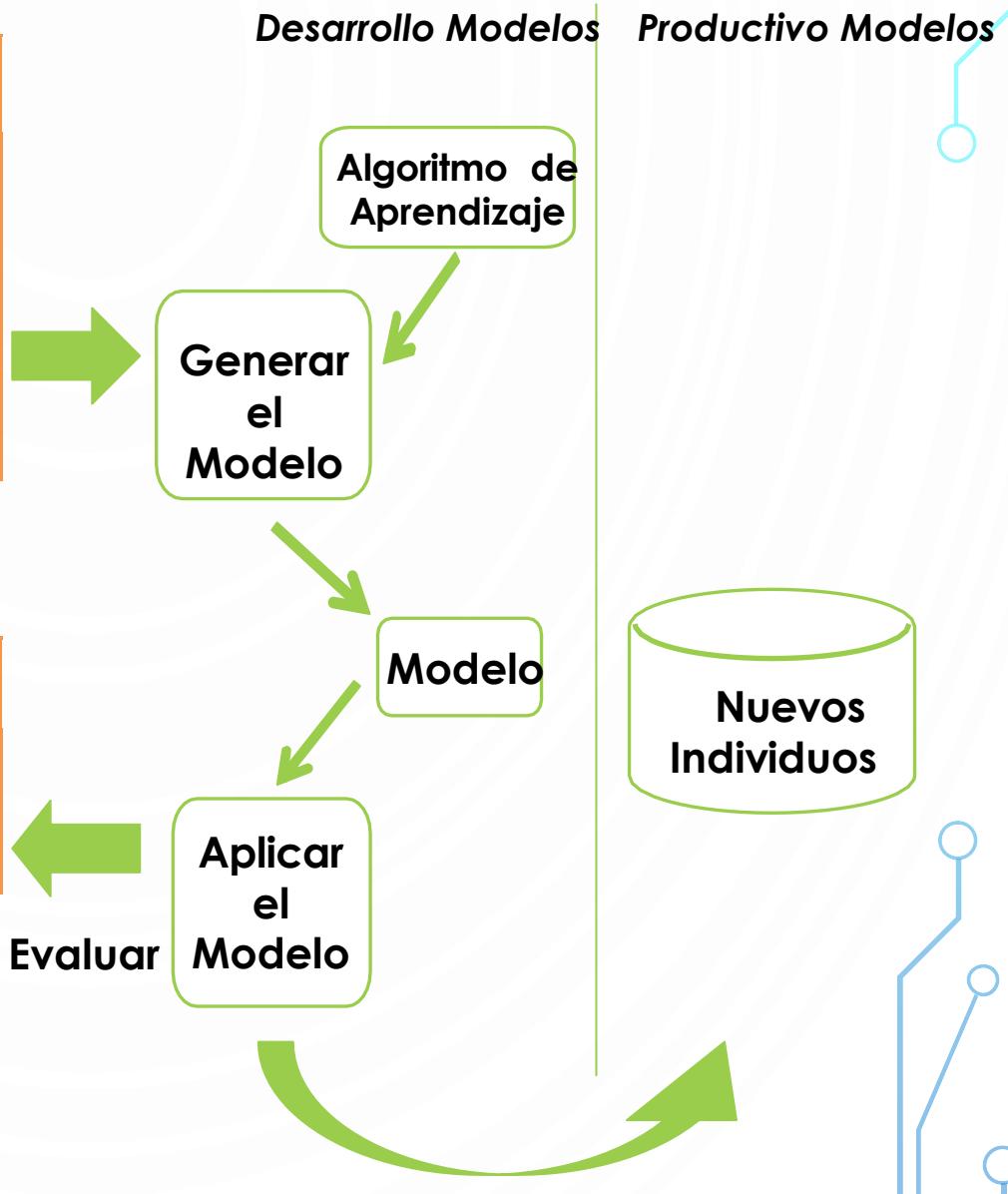
MODELO GENERAL DE LOS MÉTODOS DE CLASIFICACIÓN

ID	REEMBOLSO	ESTADO CIVIL	INGRESOS ANUALES	FRAUDE
1	SI	SOLTERO	S/ 1,000	NO
2	SI	CASADO	S/ 5,000	NO
3	NO	CASADO	S/ 3,500	SI
4	SI	VIUDO	S/ 4,500	NO
5	NO	SOLTERO	S/ 2,000	NO
6	NO	SOLTERO	S/ 1,500	SI

Tabla de Aprendizaje

ID	REEMBOLSO	ESTADO CIVIL	INGRESOS ANUALES	FRAUDE
7	SI	SOLTERO	S/ 4,000	NO
8	SI	CASADO	S/ 5,500	NO
9	NO	CASADO	S/ 6,500	SI

Tabla de Testing



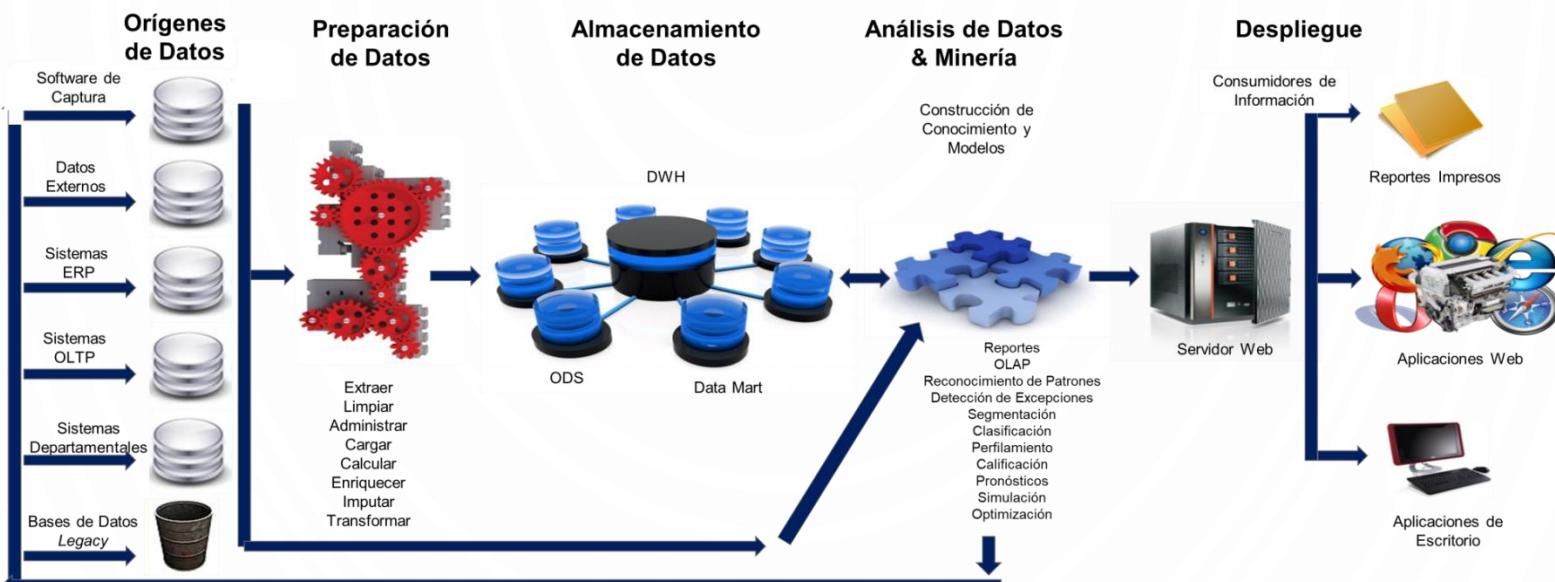
DEFINICIONES BÁSICAS

- **Conjunto de Datos (Data Set):** El total del conjunto de datos sobre los que queremos desarrollar un algoritmo de Machine Learning con el fin de obtener un modelo que lo represente lo mejor posible. Contendrá variables independientes y dependientes.
- **Variables Independientes (Features), (VI):** Aquellas columnas del Data Set que serán usadas por el algoritmo para generar un modelo que prediga lo mejor posible las variables dependientes.
- **Variables dependientes (Labels,Target), (VD):** Columna del data set que responde a una correlación de VI y que debe ser predicha por el futuro modelo
- **Conjunto de Datos de Entrenamiento (Training Set):** Subconjunto del Data Set que será utilizado para entrenar el modelo que se pretende generar.
- **Conjunto de Datos de Test (Test Set):** Subconjunto del data set que se le pasará al modelo una vez haya sido entrenado para comprobar, mediante el uso de diferentes métricas, sus indicadores más importantes de calidad.

ENTENDIMIENTO DEL NEGOCIO

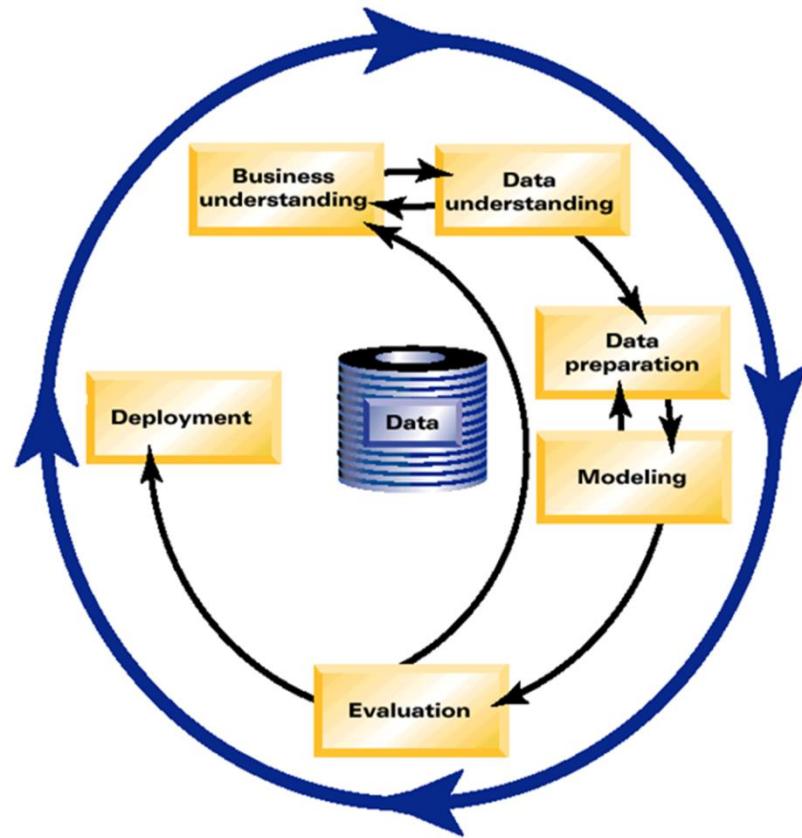
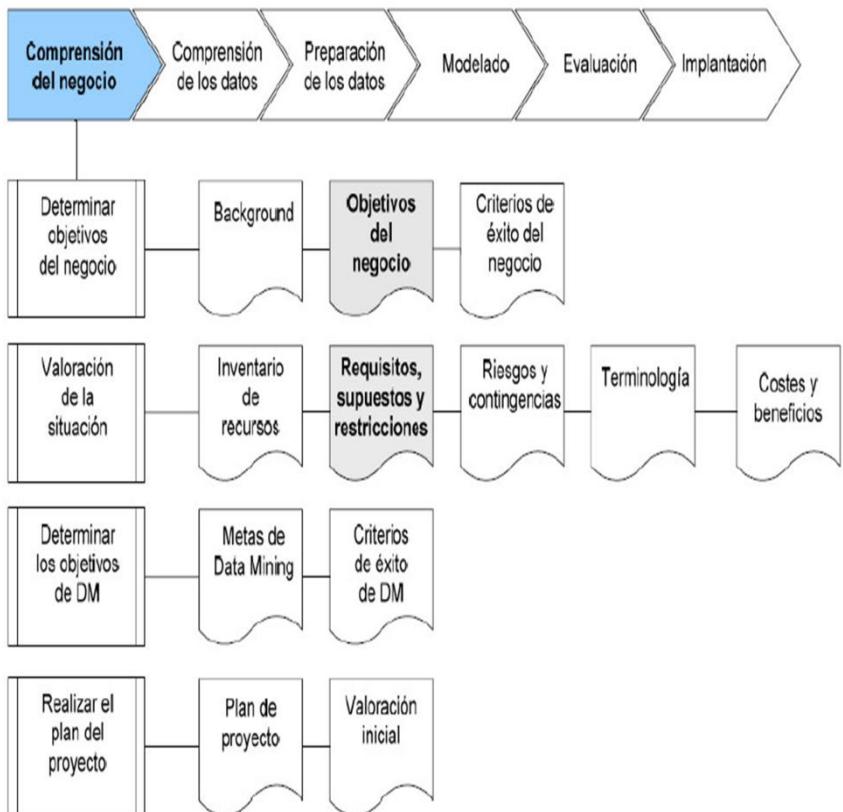


LA ANALÍTICA EN LOS NEGOCIOS



Solución Analítica

METODOLOGÍA CRISP - DM



ETAPA 1: ENTENDIMIENTO CONTEXTUAL DEL PROBLEMA A DESARROLLAR

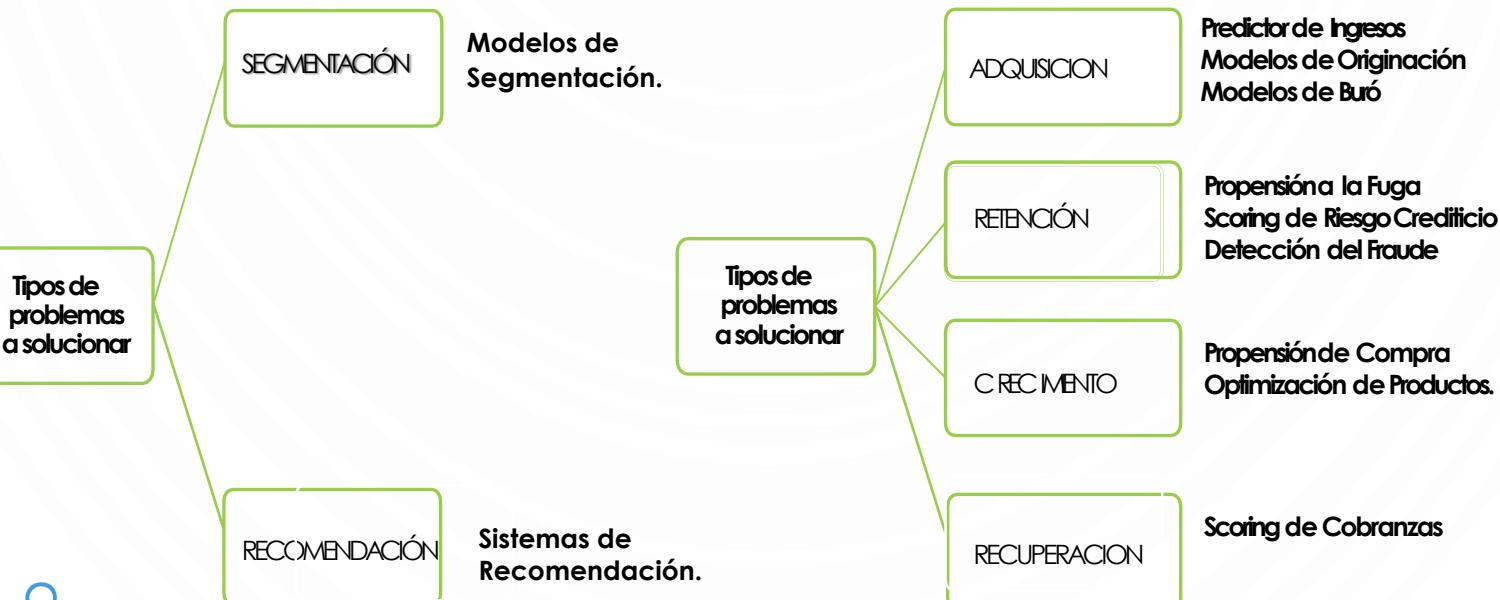


ENTENDIMIENTO DEL NEGOCIO : ENTENDIMIENTO DEL PROBLEMA

PROPÓSITO DEL ANÁLISIS

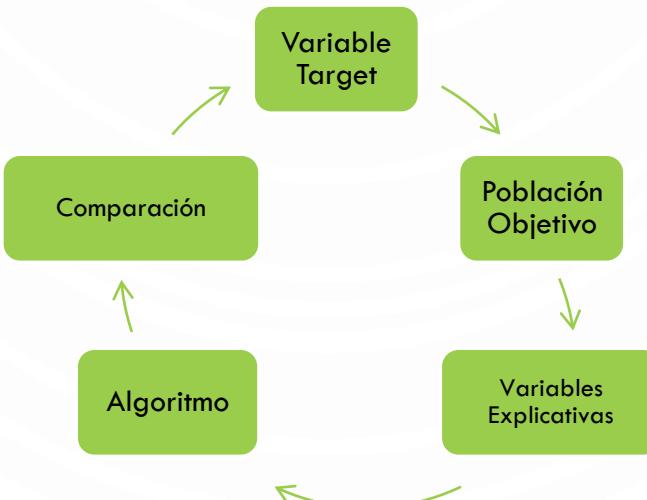
Descubrir eventos o resultados futuros en base al conocimiento previo de los datos, utilizando para ello métodos estadísticos, matemáticos, computacionales y de base de datos, así como de la aplicación de los algoritmos de machine learning. En cualquier negocio el éxito depende de:

- ✓ Capacidad de recopilar y limpiar la información para el análisis.
- ✓ Capacidad de Identificar los patrones y tendencias de los datos en relación a lo que se desea solucionar.
- ✓ Capacidad de crear el modelo que le de valor al negocio.



ENTENDIMIENTO DEL NEGOCIO :ENTENDIMIENTO DEL PROBLEMA

- 1.¿Que problema quiero solucionar? → Variable objetivo o de respuesta (Y) → Ejem: Estimar ingresos de personas no bancarizados
- 2.¿Con qué población analizo el problema? → Población objetivo → Ejem: Dependientes e Independientes
- 3.¿Qué indicadores pueden explicar el problema? → Covariables (Xs) → Ejem: NSE (Reniec), Tipo de automovil (Sunarp)
- 4.¿Qué técnica estadística o informática se ajusta al análisis? → Métrica o algoritmo → Ejem: Arboles de decisión



ENTENDIMIENTO DEL NEGOCIO :ENTENDIMIENTO DEL PROBLEMA

- Es el indicador que se desea predecir y cuyo valor calculado de forma anticipada permitirá optimizar las estrategias del negocio. También se conoce como **Variable de respuesta**.
- Se calcula dentro del periodo de performance y puede ser **real** o **ficticia (latente)**. Si es real queremos decir que existe en la base de datos. Si originalmente no está en nuestra base de datos entonces la creamos (por ello se dice que es ficticia).

Tipos de Target

Un modelo puede ser de clasificación o de regresión, según el tipo de variable target.

- **Target de Clasificación:** Es una categoría o clase.

Ejemplos: Clientes Fuga/No Fuga, Compran/No Compran, Recuperables / No Recuperables, En Mora/Al día, Fraude / No Fraude, etc.

- **Target de Regresión:** Es un valor puntual.

Ejemplos: Pagos, Ingresos, Ventas, etc.

ENTENDIMIENTO DEL NEGOCIO : ENTENDIMIENTO DEL PROBLEMA

Enfoque temporal:

Línea de tiempo

Enfoque matricial:

Información histórica almacenada



- "Futuro": Período de Predicción o Performance
Donde se define a la variable de respuesta
- "Pasado": Período de observación

ID	Segment_Target	Var_Target	Var_X1	Var_X2	Var_X3	Var_X4	Var_X5	Var_X6
1	Segment 1	1	-0.243257655	216	952.4800	1	4	3
2	Segment 2	1	1.696358794	191	633.4949	0	7	2
3	Segment 3	1	0.561226988	192	637.5107	0	6	3
4	Segment 1	1	-1.673888687	205	927.2513	0	8	3
5	Segment 2	0	-0.315746538	200	988.0877	0	2	3
6	Segment 3	0	0.402197729	201	927.5218	1	6	2
7	Segment 1	1	0.668736379	202	582.0028	0	6	2
8	Segment 2	1	1.489475004	197	701.1748	0	6	2
9	Segment 3	0	0.308647509	201	526.3747	0	8	4
10	Segment 1	1	0.090616380	189	989.2571	0	7	4
11	Segment 2	1	0.081223506	200	789.0298	0	8	2
12	Segment 3	1	-0.443663814	207	937.3809	0	2	2
13	Segment 1	1	-1.416088194	220	819.6118	0	9	1
14	Segment 2	1	-0.316298576	187	995.7736	1	2	5

Población objetivo

Variable de respuesta (Y)

Covariables X_i

Métrica
Var_Target =
 $f(Var_X1, Var_X2, Var_X3, Var_X4, Var_X5, Var_X6)$

ENTENDIMIENTO DEL NEGOCIO : DEFINICIÓN DE LA VARIABLE TARGET

Los pasos para crear una variable target de clasificación son:

- Primero: Elegir la variable(es) de interés para crear el target.
- Segundo: Definir el horizonte temporal del periodo de performance o predicción.
- Tercero: Determinar las clases del indicador según los cortes de la variable(es) de interés.

Por ejemplo para un modelo de buró:



➤ Se determinan las clases del Target



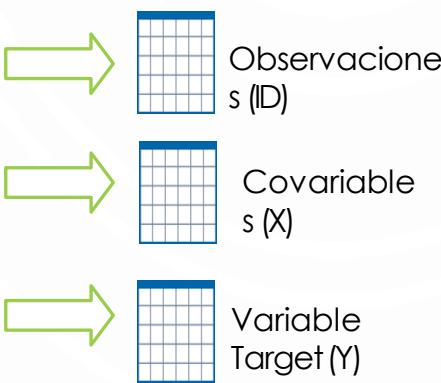
ENTENDIMIENTO DEL NEGOCIO : ELECCIÓN Y DEFINICIÓN DEL HORIZONTE TEMPORAL

- Identificación de las características que debe tener la data para el proceso de modelización, de acuerdo a lo que se **necesita solucionar en el negocio**.
- Se debe identificar las fuentes de información disponible con las que contamos y contaremos en el futuro para que la obtención de los datos sea siempre factible.

Fuentes de Información del Negocio



Extracción de la información



Data consolidada

Recopilación y cálculo de Variables X
Periodo de Observación
Ejemplo: De $t - 12$ a $t - 1$

100100011101000000101000110111010110
100100111101100000011111000110100100
10000110101101111101011100001101001001
1111110100011011100101011100001011
110011111101111111010000110110110
010000110101101100001100000100010000
010101110911001111011011001010010111
001000010110010100000010000010011110
0111010001111100101101010101011110
100010000101100001010101011000101
01001000010101011101110000101000000
01011000010111010101011101001
0110111101011111001010001010000000
011010011010110100010010111001101
0001010000011001100011001000010010110
1001010101000111001010101010111101

Definición de la variable de Respuesta Y
Periodo de Predicción
Ejemplo: De $t+1$ a $t+6$

Data en el corte de tiempo: t

ENTENDIMIENTO DEL NEGOCIO : CRITERIOS DE INCLUSIÓN Y EXCLUSIÓN DE LA INFORMACIÓN

1. Data Original

```
100100011101000000101000110111010110  
10010011110111000000111100110100100  
100001101101111101010011100001101001  
111111010000110111001010111100001011  
11001111110111111100100001110110110  
010000110100110110000110000100010000  
01010111001100111011001110100010111  
001000010101100101000001000010011110  
0111010011111100101110101010111100  
100010000101100010101101010111000101  
0100100001001010111001100001010000  
0101100000100111010100101110110001  
01101111101011100010100010100010000  
011010011011010001000101111001101  
000101000001100110001100100010010110  
10010101010001001110010101010111101
```

Criterios de exclusión



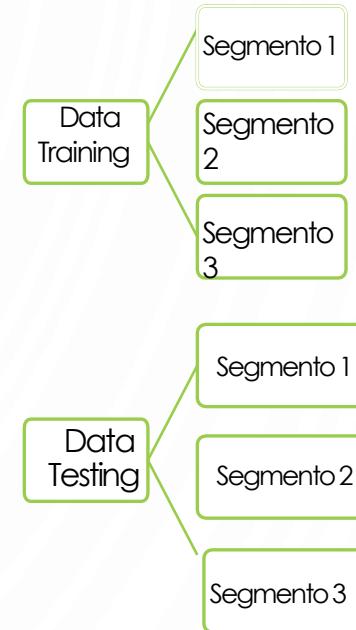
2. Data de Estudio

```
100100011101000000101000110111010110  
10010011110111000000111100110100100  
100001101101111101010011100001101001  
111111010000110111001010111100001011  
11001111110111111100100001110110110  
010000110100110110000110000100010000  
01010111001100111011001110100010111  
001000010101100101000001000010011110  
0111010011111100101110101010111100  
100010000101100010101101010111000101  
0100100001001010111001100001010000  
0101100000100111010100101110110001  
01101111101011100010100010100010000  
011010011011010001000101111001101  
000101000001100110001100100010010110  
10010101010001001110010101010111101
```

Información pulida y lista para el proceso de modelación

Training
Testing

3. Segmentación

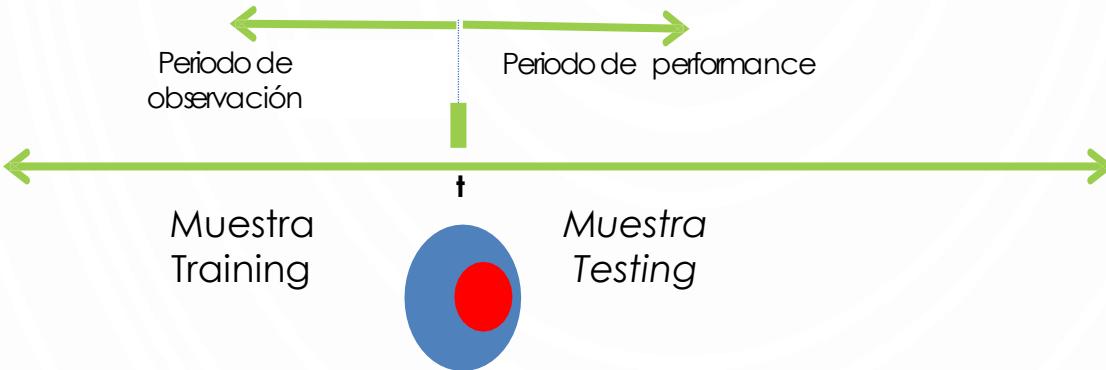


- ✓ Tienen poca información histórica
- ✓ Son riesgosos para el negocio
- ✓ Están fuera de las políticas o de las estrategias del negocio
- ✓ Son datos erróneos
- ✓ Son casos especiales que no se volverán a recopilar
- ✓ Ciclos económicos
- ✓ Temas regulatorios / legales

ENTENDIMIENTO DEL NEGOCIO : VALIDACIÓN DE RESULTADOS

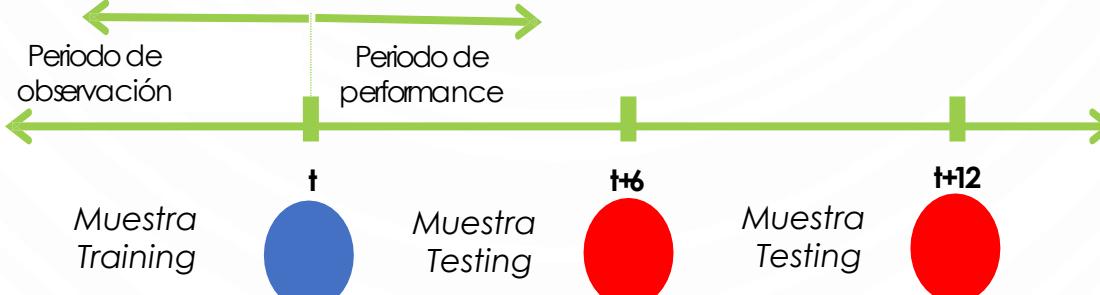
Validación Out of Sample (OOS)

Una vez seleccionada la muestra para la elaboración del modelo, se divide aleatoriamente ésta en 2 sub muestras, una para el entrenamiento del modelo y la otra para su validación. Generalmente la muestra de validación está entre el 10% y 25% del total.

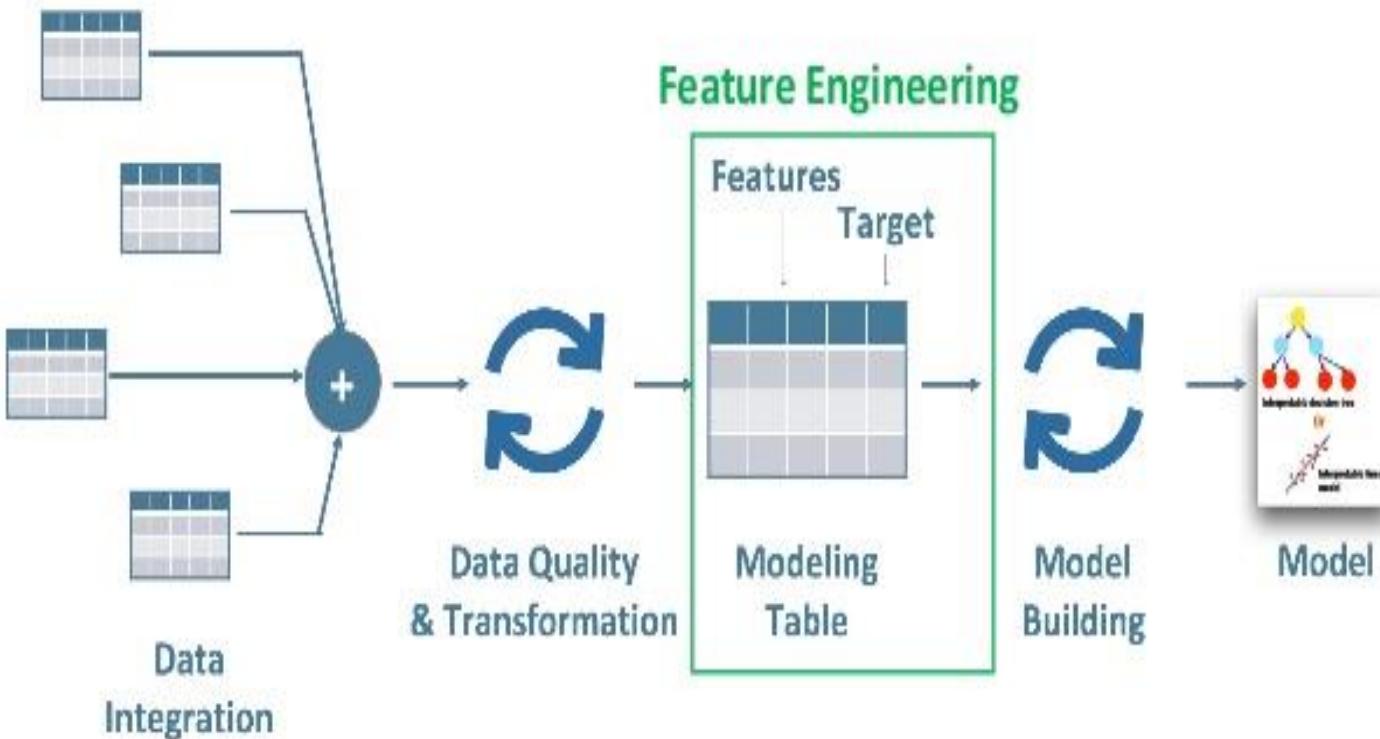


Validación Out of time (OOT)

Se elige una muestra o más muestras fuera del horizonte temporal que se usó en el entrenamiento del modelo.



ETAPA 2: PREPARACIÓN Y VALOR AGREGADO DE LOS DATOS



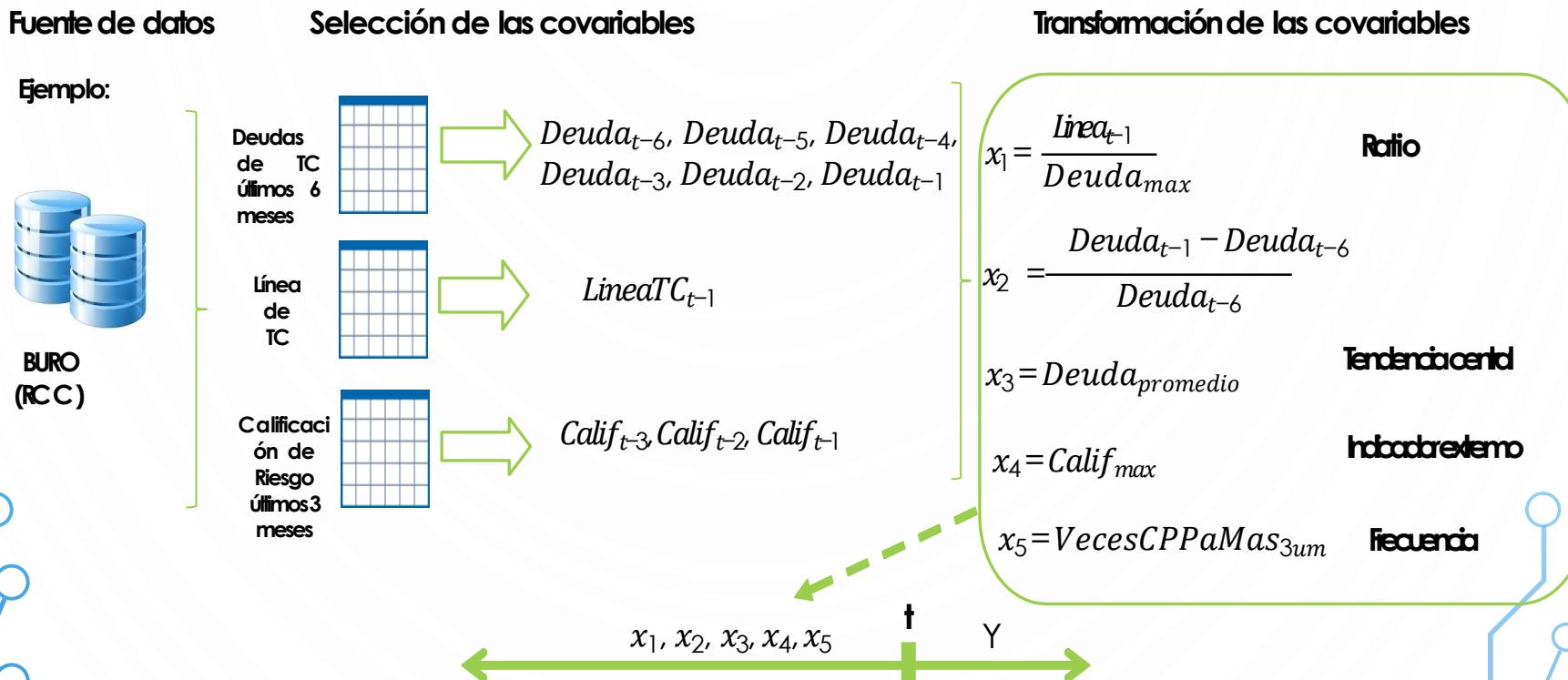
ETAPA 2: PREPARACIÓN Y VALOR AGREGADO DE LOS DATOS

- ✓ Combinar conjuntos de datos de dos archivos distintos
- ✓ Seleccionar subconjuntos de los datos
- ✓ Dividir el archivo de los datos en varias partes
- ✓ Transformar variables
- ✓ Ordenar casos
- ✓ Agregar nuevos datos y/o variable
- ✓ Eliminar datos y/o variables
- ✓ Guardar datos y/o resultados



ENTENDIMIENTO DEL NEGOCIO : DEFINICIÓN Y CREACIÓN DE DRIVERS

Las variables a seleccionar para la solución del problema propuesto deben tener **sentido para el negocio**. En otras palabras al seleccionarlas se espera que estén correlacionadas con la variable de respuesta del modelo. La transformación tiene como propósito optimizar el aporte de las X_i en el modelo.



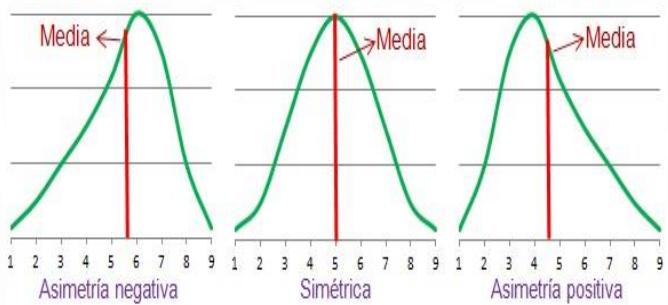
ETAPA 3: ENTENDIMIENTO DE LOS DATOS Y ANÁLISIS EXPLORATORIO DE DATOS



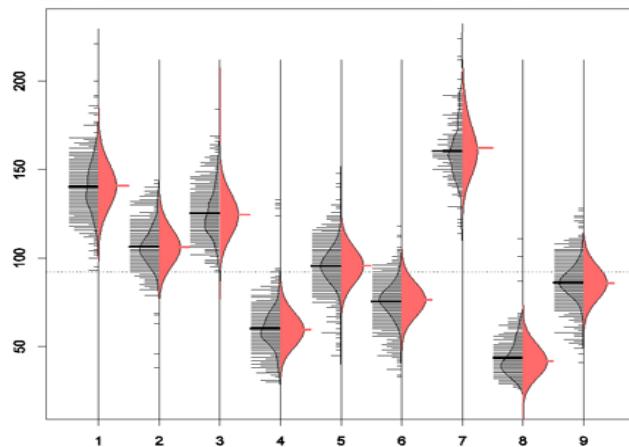
ETAPA 3: EXAMEN GRÁFICO Y DESCRIPTIVO NUMÉRICO

ESCALA DE MEDIDA	REPRESENTACIÓN GRÁFICA	MEDIDA DE TENDENCIA CENTRAL	MEDIDA DE DISPERSIÓN
NOMINAL	Diagrama de barras, líneas , pictograma y sectores	Moda	
ORDINAL	Gráficos de cajas,violín	Moda, Mediana, Media truncada	Rango intercuartílico,
			CVQ
INTERVALO	Histograma, polígonos de frecuencias, Gráficos de cajas,violín	Moda, Media, Mediana	Desviación estándar, Rango intercuartílico, CVQ
RAZÓN	Histograma, polígonos de frecuencias, Gráficos de cajas,violín	Moda, Mediana, Media, Media geométrica	Desviación estándar,
			Coeficiente de variación, Rango intercuartílico,
			CVQ

ETAPA 3: EXAMEN GRÁFICO Y DESCRIPTIVO NUMÉRICO

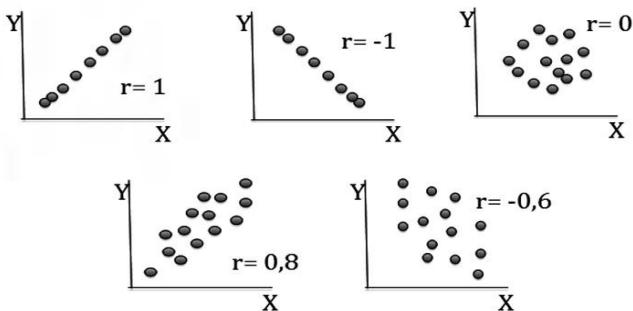


Simetría	Relación
Simétrica o insegada	Moda = Mediana = Media
Asimétrica o con sesgo positivo a la derecha	Moda > Mediana > Media
Asimétrica o con sesgo negativo a la izquierda	Moda < Mediana < Media



ETAPA 3: ASOCIACIONES CUALITATIVAS Y CUANTITATIVAS

Variable X	Variable Y	Coeficiente de correlación
Cualitativa	Cualitativa	Chi-cuadrado Contingencia Phi
Ordinal	Ordinal	Spearman
Cualitativa	Cuantitativa	Biserial-puntual
Cuantitativa	Cuantitativa	Pearson



		EDAD				Total
		MENOS DE 30 AÑOS	ENTRE 30 Y 45	ENTRE 45 Y 60	MÁS DE 60 AÑOS	
IMPRESIÓN	MUY BUENA	40,4%	43,2%	43,2%	46,9%	42,1%
	BUENA	45,6%	43,3%	44,1%	38,9%	44,3%
Total		100,0%	100,0%	100,0%	100,0%	100,0%
IMPRESIÓN		MUY BUENA	40,4%	43,2%	43,2%	46,9%
IMPRESIÓN		BUENA	45,6%	43,3%	44,1%	38,9%
IMPRESIÓN		NORMAL	12,5%	12,3%	11,5%	12,8%
IMPRESIÓN		MALA	1,5%	1,3%	1,2%	1,4%

- ETAPA 3: CASOS ATÍPICOS UNIVARIADOS Y MULTIVARIADOS



ETAPA 4: CASOS ATÍPICOS UNIVARIADOS Y MULTIVARIADOS

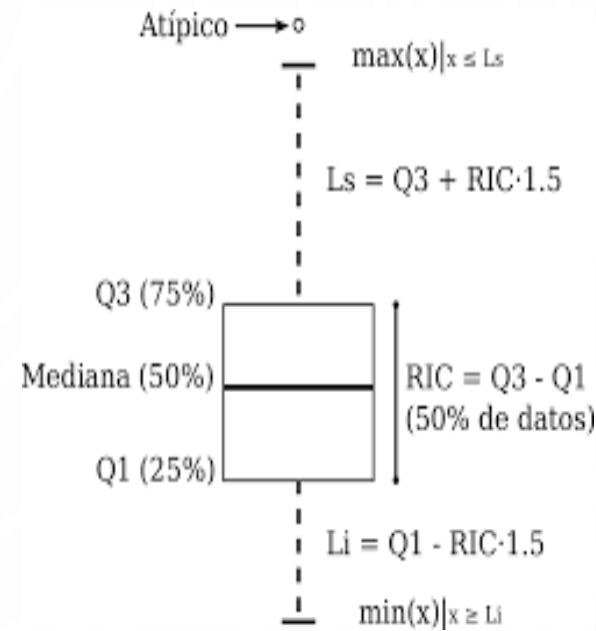
- Los casos atípicos son observaciones con características diferentes de las demás.
- Este tipo de casos no pueden ser caracterizados categóricamente como benéficos o problemáticos sino que deben ser contemplados en el contexto del análisis y debe evaluarse el tipo de información que pueden proporcionar.
- Por otra parte, aunque diferentes a la mayor parte de la muestra, pueden ser indicativos de las características de un segmento válido de la población.

TIPOS DE ATÍPICOS

Los casos atípicos pueden clasificarse en 4 categorías:

➤ **La primera categoría** contiene aquellos casos atípicos que surgen de un error de procedimiento, tales como la entrada de datos o un error de codificación. Estos casos atípicos deberían subsanarse en el filtrado de los datos, y si no se puede, deberían eliminarse del análisis o recodificarse como datos ausentes.

➤ **La segunda clase** es la observación que ocurre como consecuencia de un acontecimiento extraordinario. En este caso, el outlier no representa ningún segmento válido de la población y puede ser eliminado del análisis.



TIPOS DE ATÍPICOS



- **La tercera clase** contiene las observaciones cuyos valores caen dentro del rango de las variables observadas, pero que son únicas en la combinación de los valores de dichas variables. Estas observaciones deberían ser retenidas en el análisis pero estudiando qué influencia ejercen en los procesos de estimación de los modelos considerados.
- **La cuarta y última clase** comprende las observaciones extraordinarias para las que el investigador no tiene explicación. En estos casos lo mejor que se puede hacer es replicar el análisis con y sin dichas observaciones con el fin de analizar su influencia sobre los resultados. Si dichas observaciones son influyentes el analista debería reportarlo en sus conclusiones y debería averiguar el por qué de dichas observaciones.

ETAPA 5: MISSING Y REPRESENTATIVIDAD DE VARIABLES

- Lo primero es determinar las razones que subyacen en el dato ausente buscando entender el proceso principal de esta ausencia para seleccionar la acción más apropiado.
- Se debe determinar cuál es el proceso de datos ausentes, es decir un error de proceso una acción como rehusar al contestar. que da lugar a la ausencia de datos.

ETAPA 5: MISSING Y REPRESENTATIVIDAD DE VARIABLES

- Si se encuentran datos ausentes pueden utilizarse las siguientes aproximaciones:
 - a) Utilizar sólo los casos completos: conveniente si el tamaño de la data no se reduce demasiado.
 - b) Supresión de casos y/o variables con una alta proporción de datos ausentes. Esta supresión deberá basarse en consideraciones teóricas y empíricas. En particular, si algún caso tiene un dato ausente en una variable dependiente, habitualmente excluirlo puesto que cualquier proceso de imputación puede distorsionar los modelos estimados.

- Así mismo una variable independiente con muchos datos ausentes podrá eliminarse si existen otras variables muy similares con datos observados.
- c) Imputar valores a los datos ausentes utilizando valores válidos de otras variables y/o casos de la muestra.



IMPUTACIÓN DE DATOS

- Técnicas de imputación En esta investigación se describen de forma sucinta algunas de las técnicas de imputación existentes que luego serán utilizadas para comparar los resultados por medio de simulaciones.



IMPUTACIÓN DE DATOS MEDIANTE TÉCNICAS PARAMÉTRICAS O UNIVARIADAS

Missing values

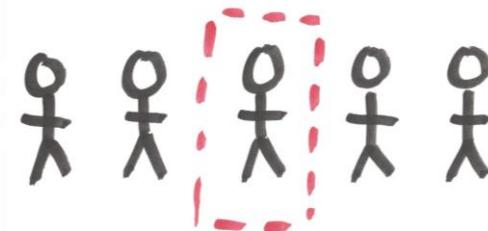
PassengerId	Survived	Pclass	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	male	22	1	0	A/5 21171	7.25		S
2	1	1	female	38	1	0	PC 17599	71.2033	C85	C
3	1	3	female	26	0	0	STON/O2. 3101282	7.925		S
4	1	1	female	35	1	0	113803	53.1	C123	S
5	0	3	male	35	0	0	373450	8.05		S
6	0	3	male		0	0	330877	8.4583		Q

Variables Cuantitativas

* Media o mediana.

Variables Cualitativas

* Moda



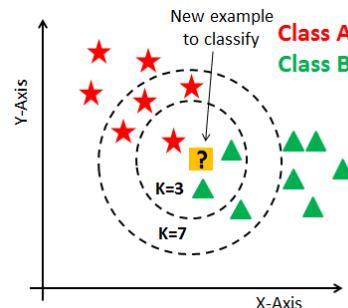
IMPUTACIÓN DE DATOS MEDIANTE TÉCNICAS MULTIVARIABLES O MACHINE LEARNING

Missing values

PassengerId	Survived	Pclass	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	male	22	1	0	A/5 21171	7.25		S
2	1	1	female	38	1	0	PC 17599	71.2033	C85	C
3	1	3	female	26	0	0	STON/O2. 3101282	7.925		S
4	1	1	female	35	1	0	113803	53.1	C123	S
5	0	3	male	35	0	0	373450	8.05		S
6	0	3	male		0	0	330877	8.4583		Q

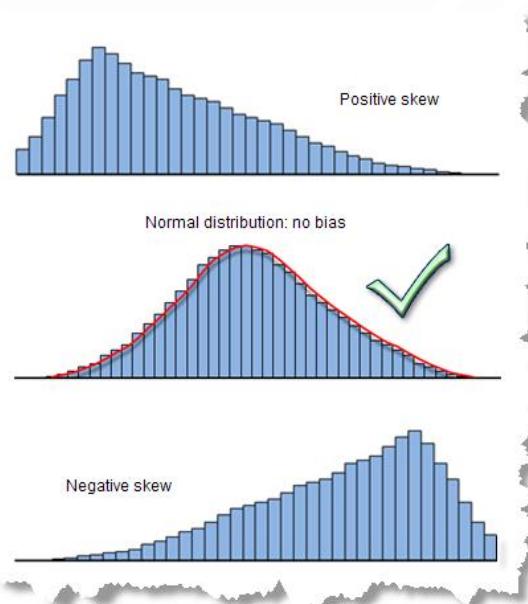
Las técnicas más utilizadas para la Imputación de datos son:

- KNN
- RandomForest
- Técnicas de clustering.



TRANSFORMACIONES DE VARIABLES

- En el modelado de datos, la transformación se refiere al reemplazo de una variable por una función. Por ejemplo, reemplazar una variable x por la raíz cuadrada / cubo o logaritmo x es una transformación.
- En otras palabras, la transformación es un proceso que cambia la distribución o relación de una variable con otras.



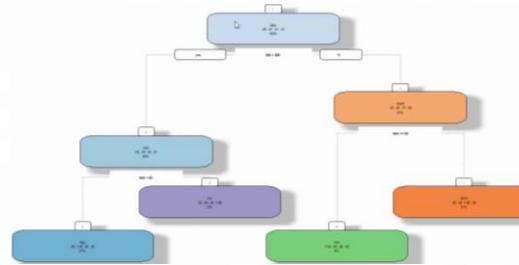
CREACIÓN DE VARIABLES

➤ Creación de Variables mediante criterio de negocio

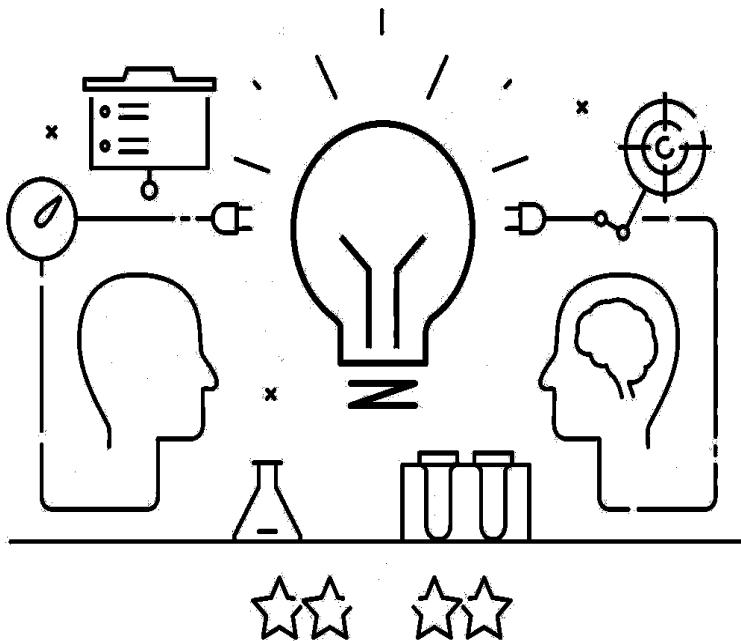
- ✓ Mínimos , máximos.
- ✓ Promedios.
- ✓ Ratios.
- ✓ Variaciones.
- ✓ Desviaciones.

➤ Creación de Variables mediante técnicas avanzadas

- ✓ Arboles de decisión.



ETAPA 4: MODELAMIENTO DE DATOS EN MACHINE LEARNING



SUPERVISED LEARNING (MODELOS SUPERVISADOS)

- Se tiene una **variable objetivo** (Variable de Salida).
- Variables que ayudan a predecir a la variable de salida (Variables de entrada).
- Existe una dependencia de las variables de entrada con las variables de salida.



SUPERVISED LEARNING

- Género.



- Rangos de Edad.



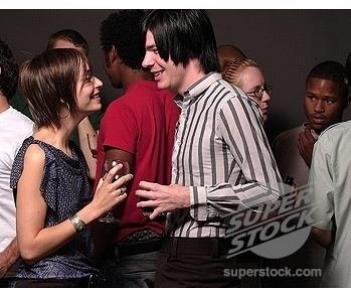
➤ **Si Compra**

- Ingresos.



➤ **No Compra**

- Estado Civil.



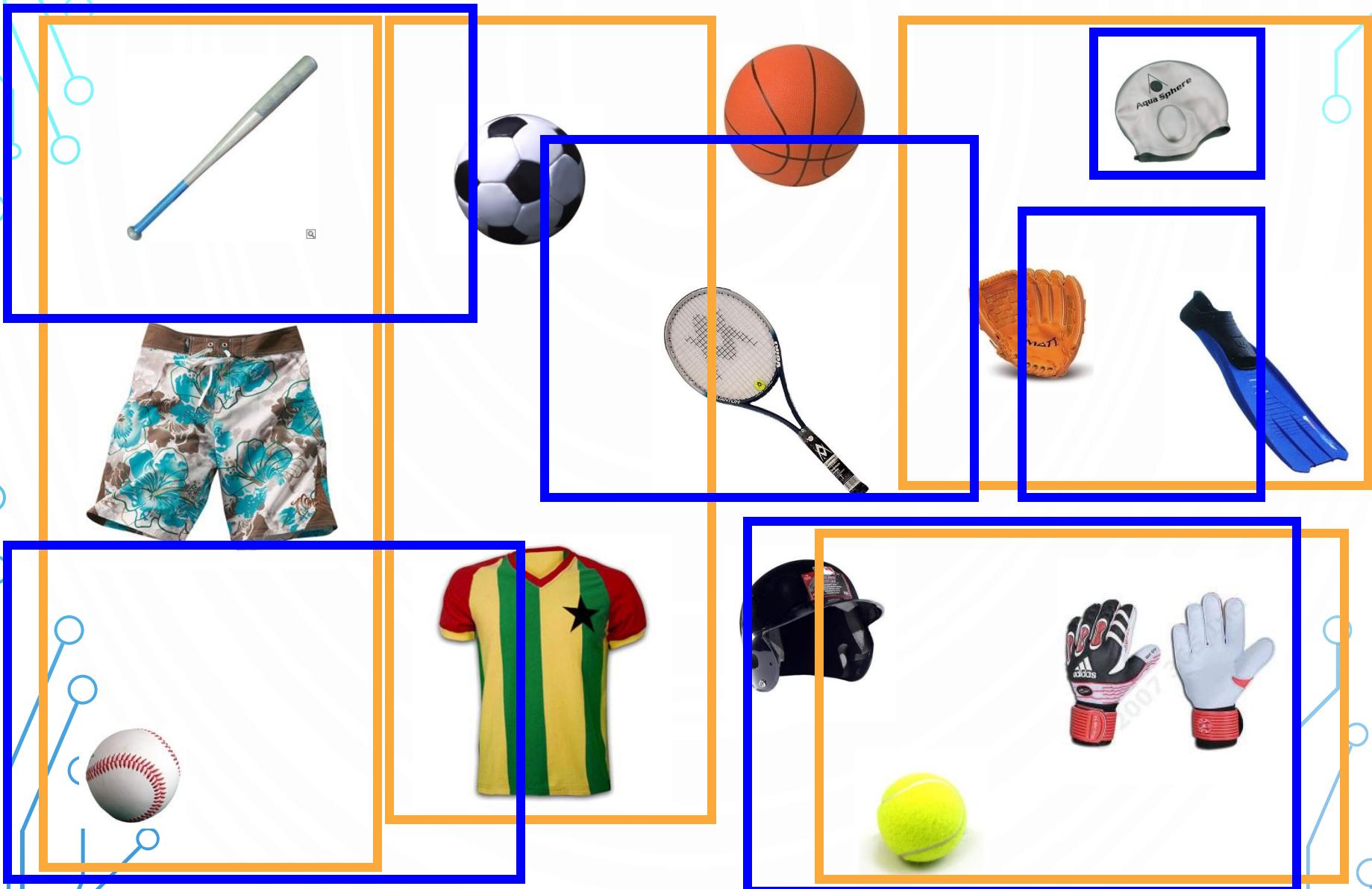
UNSUPERVISED LEARNING (MODELOS NO SUPERVISADOS)

- No hay una variable objetivo (Variable de Salida).
- No hay variables que ayudan a predecir a la variable de salida.



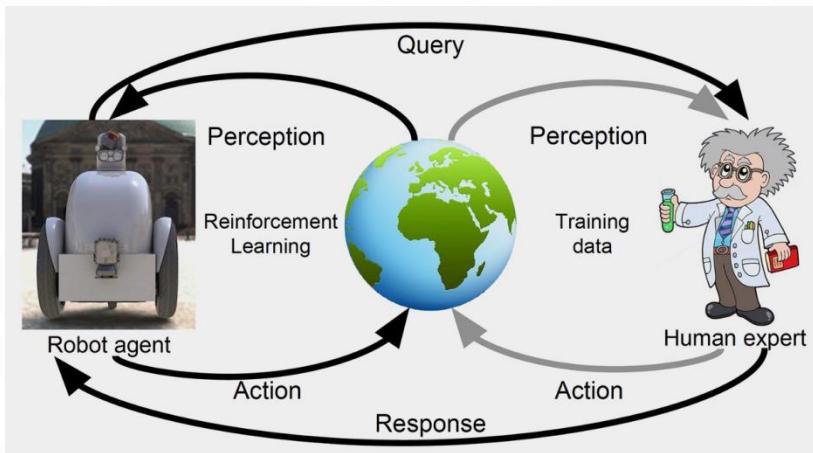
- Todas las variables tienen la misma importancia.
- Se busca la interdependencia de las variables.

MODELOS NO SUPERVISADOS



REINFORCEMENT LEARNING (APRENDIZAJE POR REFUERZO)

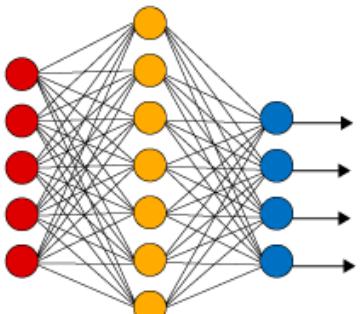
- El algoritmo de aprendizaje recibe un tipo de valoración acerca de la idoneidad de la respuesta dada.
- Cuando la decisión es correcta es muy parecido al aprendizaje supervisado, sin embargo difiere mucho cuando la decisión es incorrecta.



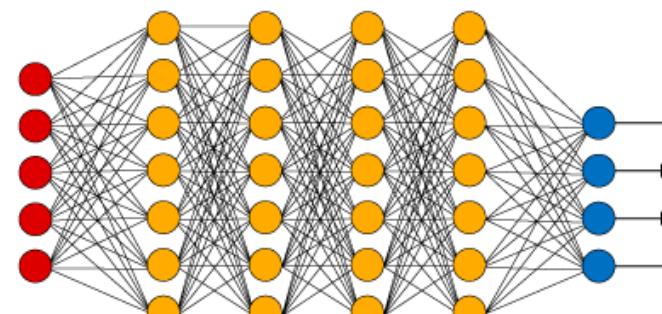
DEEP LEARNING (APRENDIZAJE PROFUNDO)

- Es un conjunto de algoritmos de Machine Learning que intenta modelar abstracciones de alto nivel usando arquitecturas compuestas como redes neuronales profundas, redes neuronales convolucionales y redes de creencia profunda para resolver problemas como visión del computador, reconocimiento automático del habla, reconocimiento del lenguaje escrito.

Simple Neural Network



Deep Learning Neural Network



● Input Layer

● Hidden Layer

● Output Layer

DEEP LEARNING (APRENDIZAJE PROFUNDO)

Neural Networks

©2016 Fjodor van Veen - asimovinstitute.org



APLICACIONES DE MACHINE LEARNING

Deserción Académica



Fuga de Clientes



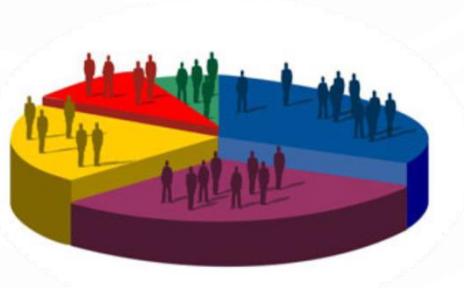
Lavado de Activos



Detección de Fraudes



Generar Perfiles o Grupos



Venta Cruzada

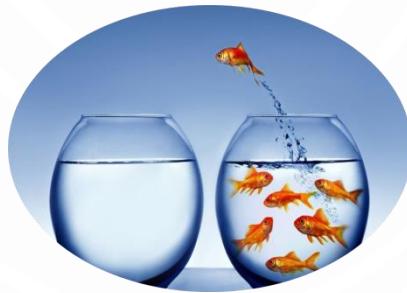


APLICACIÓN DEL MACHINE LEARNING

Deserción Académica



Fuga de Clientes



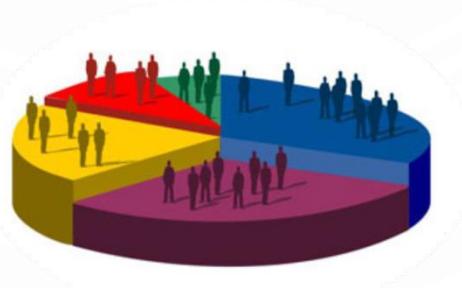
Lavado de Activos



Detección de Fraudes



Generar Perfiles o Grupos



Venta Cruzada



MITOS

Mito:

“Relato o noticia que desfigura lo que realmente es una cosa y le da apariencia de ser más valiosa o más atractiva.”

Diccionario de la Lengua Española.



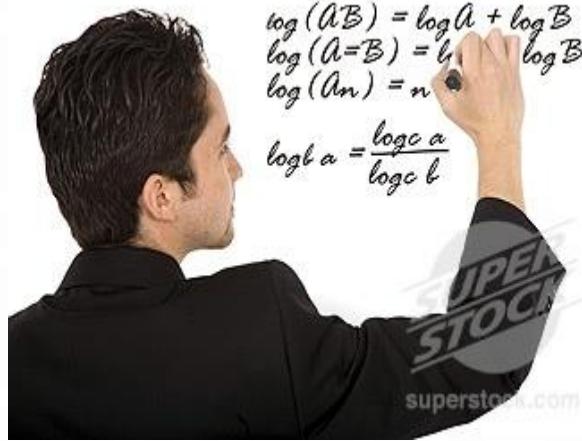
MITO # 1: OLAP/REPORTES ES MACHINE LEARNING

- OLAP/Reportes : Busca responder la pregunta sobre lo que pasó y quedó registrado en la base de datos
- ML:
 - **POR QUÉ** pasaron las cosas
 - **PREDECIR** lo que se espera que pase.
 - **PRESCRIPTIVO**

		Actual		Budget	
		Sales	Margin	Sales	Margin
TV	East				
	West				
VCR	South				
	Total				
TV	East				
	West				
VCR	South				
	Total				

MITO # 2: ESTADÍSTICA ES MACHINE LEARNING

- **Estadística:** Es en esencia el proceso de validar o rechazar hipótesis.
- **ML:**
 - Proceso de GENERAR hipótesis a partir del comportamiento de los datos.
 - Aprender patrones complejos de millones de datos.



MACHINE LEARNING VS. ANÁLISIS ESTADÍSTICO

La diferencia se encuentra en la dirección de la búsqueda

Machine Learning

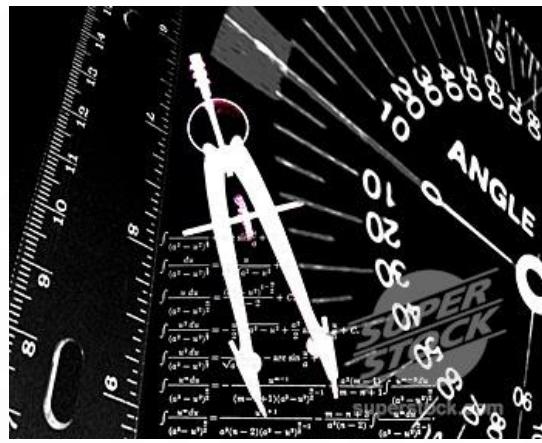
- Originalmente fueron desarrollados para actuar como sistemas expertos que ayudan a resolver problemas.
- Poco interés en la formulación de los algoritmos. Requiere entendimiento de los datos y el problema de negocio.
- Encuentra patrones en grandes cantidades de datos.
- No requiere supuestos sobre los datos. Si tiene sentido se utiliza.

Análisis Estadístico

- Prueba de Hipótesis.
 - Es la relación significativa?
- Requiere de fuertes habilidades estadísticas.
- Trabaja bajo muestras; las técnicas no se encuentran optimizadas para trabajar con grandes cantidades de datos.
- Pruebas estadísticas para la bondad de ajuste.
 - Los supuestos del modelo se cumplen?

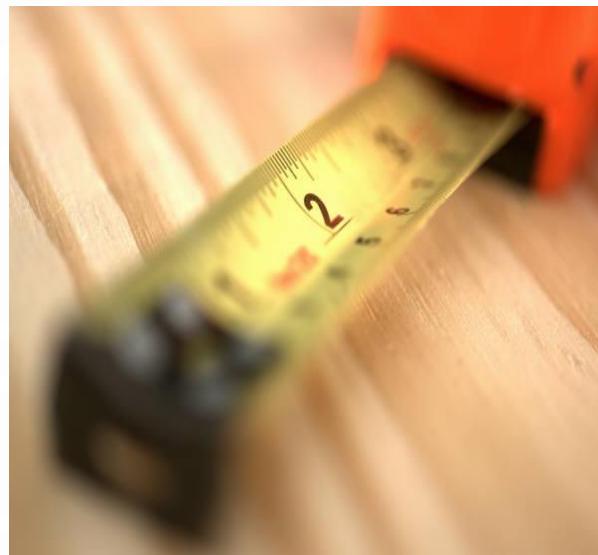
MITO # 3: MACHINE LEARNING SON SÓLO ALGORITMOS

- Todo lo que se requiere para ML son algoritmos sofisticados
- **ML:**
 - Realmente es un proceso complejo, que involucra diversas funciones de negocio,
 - Sólo un 5-10% del tiempo de un proyecto de modelos de ML tiene que ver con los algoritmos.



MITO # 4: MACHINE LEARNING ES EXACTITUD PREDICTIVA

- La calidad de ML se mide exclusivamente con la exactitud predictiva.
- **ML:**
 - Requiere de la exactitud predictiva
 - También el descubrimiento de nuevos patrones o asociaciones le dan validez al análisis



MITO # 5: MACHINE LEARNING REQUIERE DATA WAREHOUSE



- Hasta que no tenga el DW construido no puedo realizar ML.
- **ML:**
 - Estructuras claras, fácil acceso y limpias de datos facilitan análisis.
 - “Esterilización” de datos en los procesos de ETL a veces puede ser contraproducente.

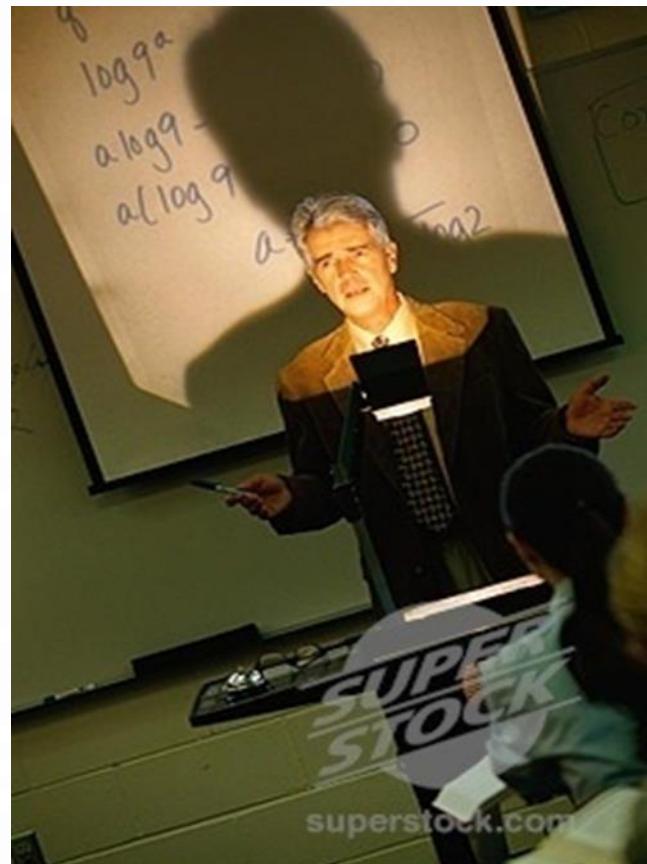
MITO # 6: MACHINE LEARNING SÓLO ES SOBRE GRANDES CANTIDADES DE DATOS

- Si no posee gran cantidad de datos no puede realizar ML.
- **ML:**
 - Grandes cantidades de datos.
 - Los problemas de negocios **realmente no generan tales dimensiones.**
 - Lo que es relevante es optimizar el tiempo del analista, no exclusivamente el de la máquina.



MITO # 8: MACHINE LEARNING REQUIERE EXPERTOS ESTADÍSTICOS Ó SISTEMAS

- Si no tengo un personal experto en los modelos no puedo realizar ML.
- **ML:**
 - Si el proceso o la herramienta de ML no permite que el “doliente” del negocio, generalmente con un conocimiento general de los modelos se favorezca de los mismos, nos estamos enfascando en resultados matemáticos y no de negocio.



MITO # 9: LAS REDES NEURONALES SON OSCURAS, Y POR ENDE INÚTILES

- **RN**, una de las técnicas de ML, no es comprensible ni útil.
- **ML**:
 - No se obtiene una ecuación o reglas.
 - Pueden ser usadas en conjunto con otras técnicas para explicar su principio de funcionamiento.
 - Son potentes para el descubrimiento de patrones o capacidad predictiva.



ERRORES



ERROR #1: ENTERRADO EN LOS DATOS

- Tengo que utilizar TODOS los datos disponibles
- R:
 - Usar sólo los datos relevantes para el problema de negocio.
 - Relevancia tanto en el tiempo como en la pertinencia de los datos.



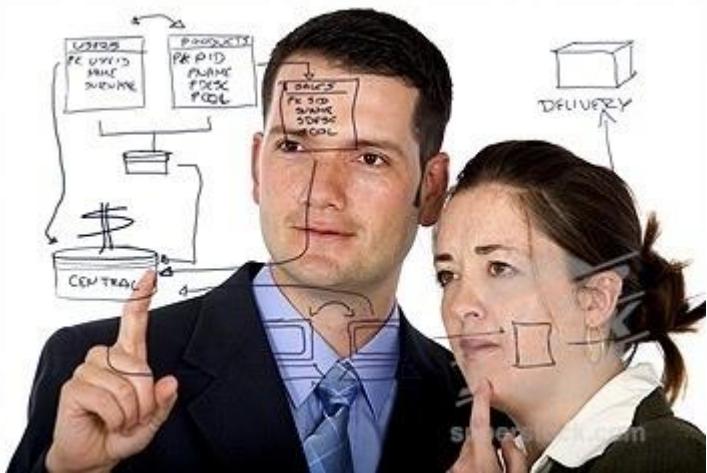
ERROR #2: FALTA DE CONOCIMIENTO DE NEGOCIO

- Tengo unos “datitos”... ¿No será que se los puedo llevar, aplicarle ML, y traerme sus resultados?
- R:
 - El conocimiento general de la industria, o de los modelos, no es suficiente.
 - El proceso de ML exige el conocimiento detallado del problema de negocio para que los resultados sean útiles



ERROR #3: FALTA DE CONOCIMIENTO DE DATOS

- Fechas, campos con información diferente, falta de documentación o desconocimiento del proceso de almacenamiento
- “El 30% de nuestros clientes tienen más de 100 años”
- **R: Se requiere la participación de un experto en las fuentes de datos, claridad en la documentación y cuidado cuando se realicen los supuestos.**



ERROR #4: POR CORTESÍA DE LOS EXPERTOS: SUPUESTOS ERRÓNEOS



- Una persona no puede tener dos cuentas del mismo tipo.
- Una persona aparece únicamente una sola vez en el listado de clientes
- Este es un tipo de teléfono que sólo se activa en el plan A
- No debe haber hombres embarazados
- R:
 - Verificar dichos supuestos
 - Explorar los datos
 - Tomar medidas correctivas

ERROR #5: INCOMPATIBILIDAD DE HERRAMIENTAS

- “Este es el algoritmo perfecto, pero ingresarle los datos tomará 8 días”
- R:
 - Herramientas abiertas
 - Integrables
 - Modelos amplios para múltiples propósitos



ERROR #6: ENCERRADO EN LA PRISIÓN DE DATOS

- “Todas las herramientas se integran perfectamente, sólo se requiere la base de datos AA”
- R:
 - Los datos no pueden estar encarcelados.
 - Busque la flexibilidad
 - Esquemas abiertos



ERROR #7: DESORGANIZACIÓN EN EL PROCESO



- “Primero apliquemos los algoritmos luego veremos qué hacemos con los resultados”.
- R:
 - El seguimiento de un procedimiento, una adecuada metodología le ahorra muchos dolores de cabeza

ERROR #8: MODELAJE SOBRE DATOS INEXISTENTES (EN EL FUTURO)

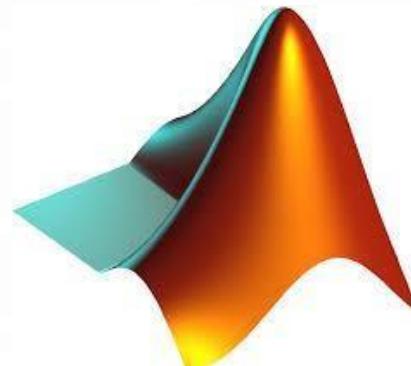
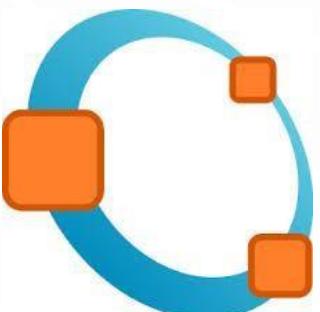
- ¿Qué datos estarán disponibles en el momento de usar los modelos?
- R:
 - No basta con pensar con qué datos se cuenta
 - Se requiere pensar con qué datos se contará.



PANORAMA TECNOLÓGICO : SOFTWARES MACHINE LEARNING



SPSS Modeler



DÓNDE APRENDER MÁS

- ✓ Aprendizaje Automático – Andrew Ng
- ✓ Deep Learning – Andrew Ng
- ✓ Machine Learning for Undergraduates – Nando da Freitas
- ✓ Machine Learning – Tom Mitchell, CMU
- ✓ Learning from Data – Yaser-Abu Mostafa
- ✓ Machine Learning A-Z - Kirill Eremenko y Hadelin de Ponteves



GRACIAS POR SU ATENCIÓN

