Data&Analytics
INNOVACIÓN Y TECNOLOGÍA

Data&Analytics
INNOVACIÓN Y TECNOLOGÍA

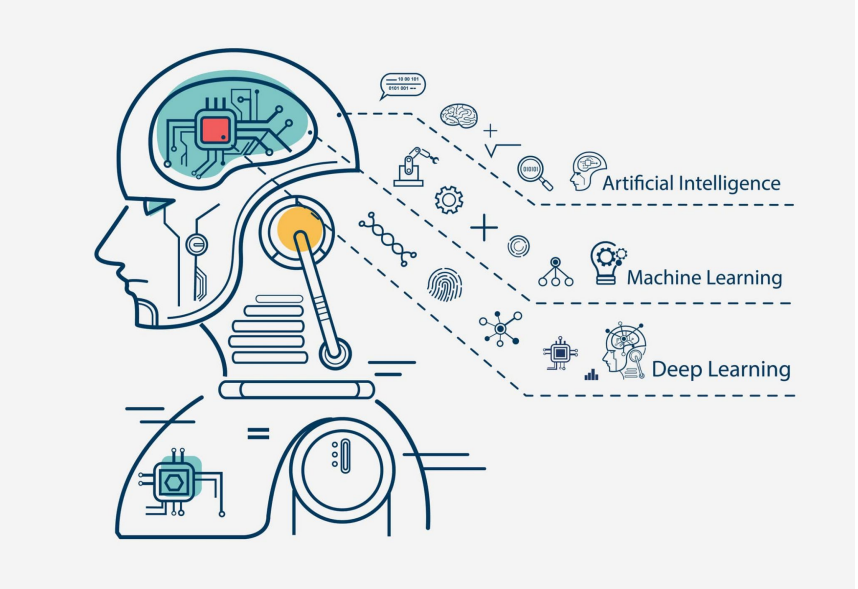
Introducción a machine learning y algunos casos de uso

M.Sc. Angelo Jonathan Diaz Soto

Enero del 2025

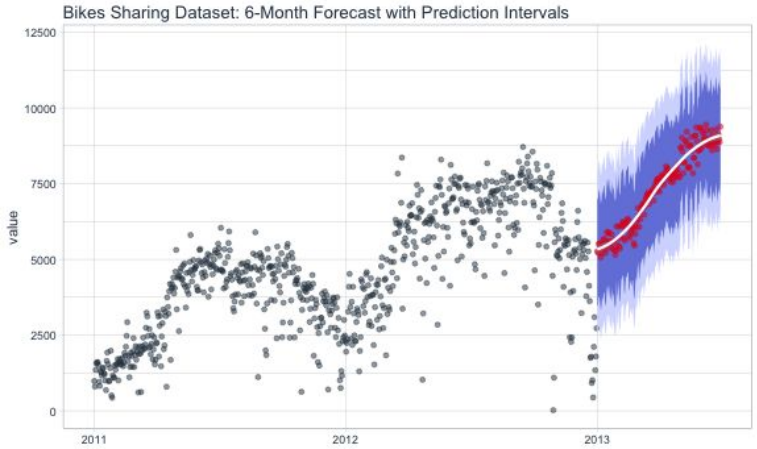
¿Qué es el machine learning?

- ✔ Es una rama de la inteligencia artificial que permite a las computadoras aprender a partir de datos sin ser explícitamente programadas, mejorando su capacidad de realizar tareas y tomar decisiones sin intervención humana directa.



¿Donde se aplica en la industria y los negocios?

- Análisis de datos y pronóstico de ventas: el machine learning puede ser utilizado para analizar grandes cantidades de datos de ventas y predecir las tendencias futuras, lo que puede ayudar a las empresas a tomar decisiones informadas sobre sus estrategias de marketing y ventas.

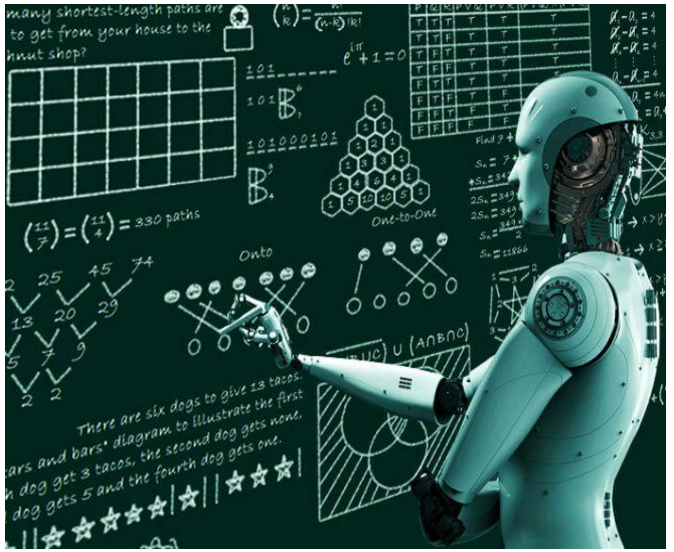


- Personalización de productos y servicios: el machine learning puede ser utilizado para analizar los datos de los clientes, sus preferencias y hábitos de compra, para ofrecer recomendaciones personalizadas de productos y servicios.

¿Donde se aplica en la industria y los negocios?

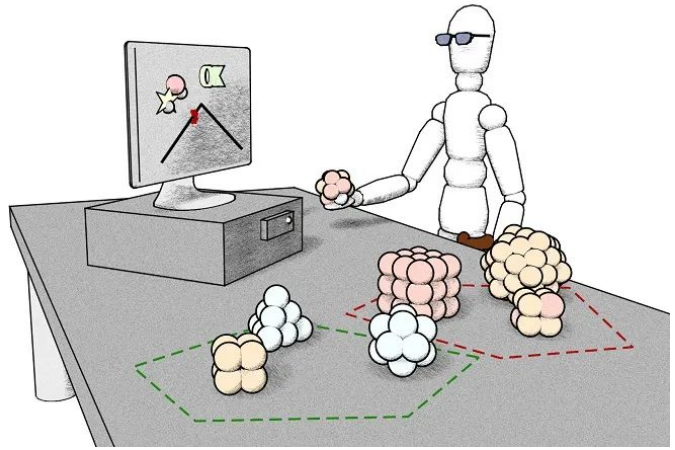
- Detección de fraudes: el machine learning puede ser utilizado para detectar patrones inusuales de transacciones financieras y prevenir fraudes en tiempo real.

- Optimización de precios: el machine learning puede ser utilizado para ajustar los precios en tiempo real en función de la demanda y otros factores relevantes, lo que puede aumentar la eficiencia y la rentabilidad de las empresas.



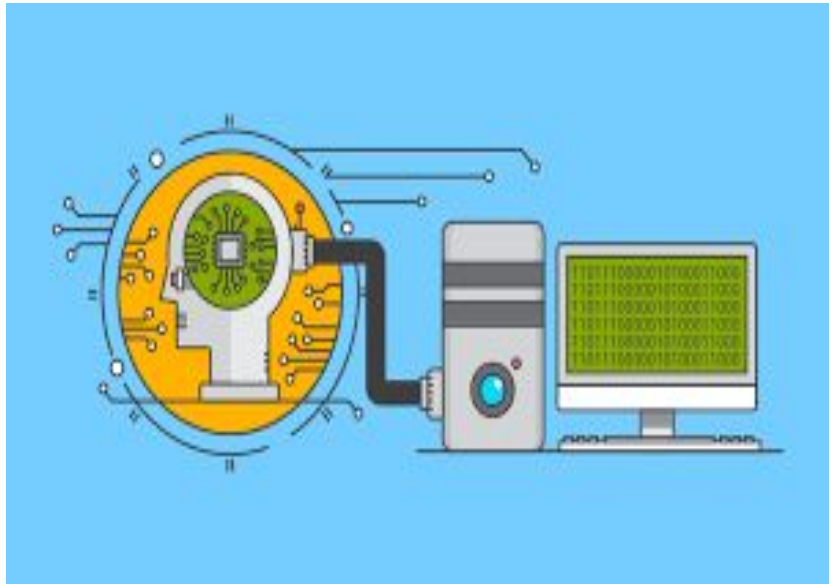
Tipos de análisis

- Análisis supervisado: en el análisis supervisado, el modelo de machine learning aprende a partir de un conjunto de datos etiquetados, es decir, datos que ya han sido clasificados o categorizados previamente. El modelo utiliza esta información para hacer predicciones o clasificaciones sobre nuevos datos. Por ejemplo, el análisis supervisado puede ser utilizado para predecir el precio de una casa a partir de datos históricos de ventas de viviendas.



Tipos de análisis

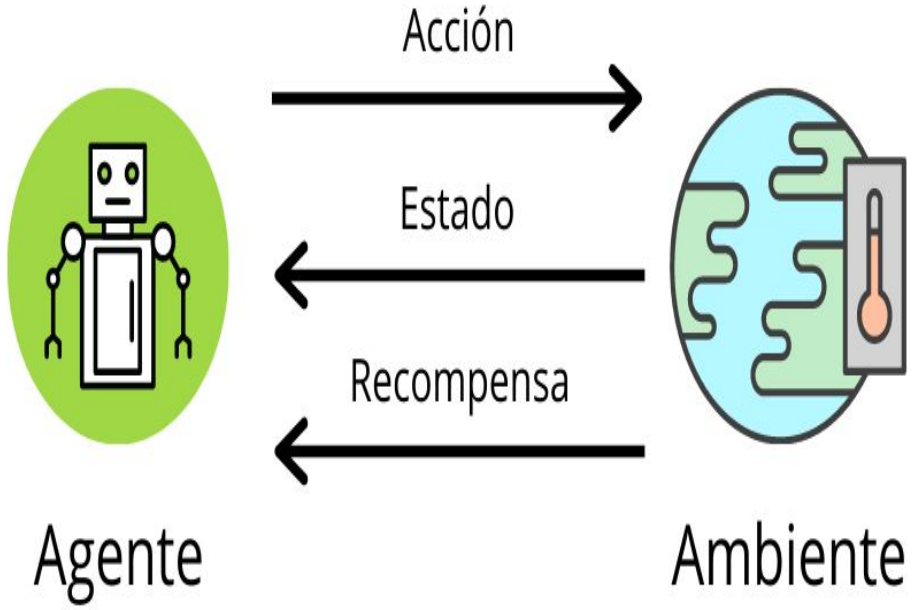
- Análisis no supervisado: en el análisis no supervisado, el modelo de machine learning aprende a partir de un conjunto de datos sin etiquetar, es decir, datos que no tienen una categoría o etiqueta definida. El modelo utiliza técnicas de clustering o agrupamiento para encontrar patrones y estructuras en los datos. Por ejemplo, el análisis no supervisado puede ser utilizado para identificar grupos de clientes con características similares a partir de los datos de sus compras.



Tipos de análisis

Data&Analytics
INNOVACIÓN Y TECNOLOGÍA

Análisis por refuerzo: en el análisis por refuerzo, el modelo de machine learning aprende a partir de la interacción con un ambiente o sistema. El modelo recibe una señal de recompensa o castigo en función de sus acciones y utiliza esta información para mejorar su desempeño. Por ejemplo, el análisis por refuerzo puede ser utilizado para entrenar a un robot para que aprenda a realizar tareas complejas en un entorno físico.



Data&Analytics
INNOVACIÓN Y TECNOLOGÍA

Data&Analytics
INNOVACIÓN Y TECNOLOGÍA

Sesión N° 2:

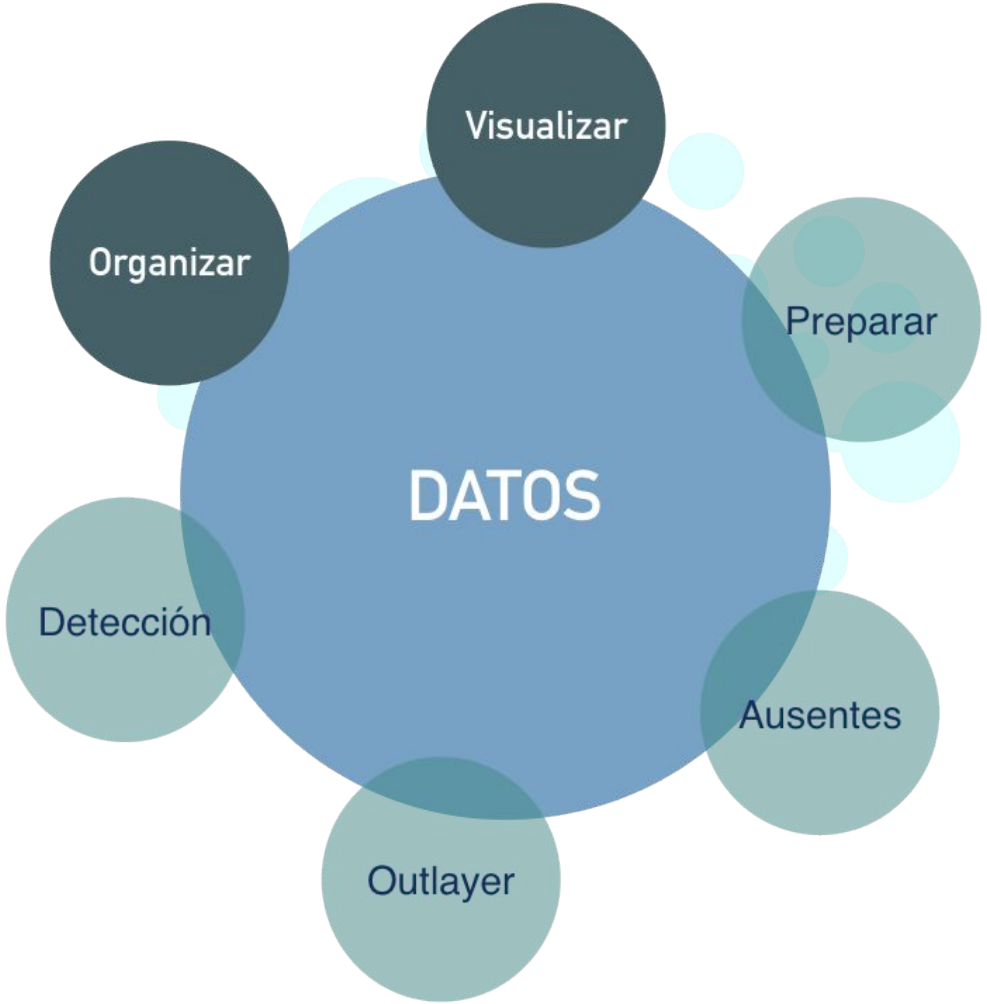
Análisis Exploratorio y Tratamiento de Datos

¿Qué es el EDA?

Data&Analytics
INNOVACIÓN Y TECNOLOGÍA

- ✔ Eda es la sigla en inglés para **Exploratory Data Analysis** y consiste en una de las primeras tareas que tiene que desempeñar el Científico de Datos.
- ✔ Es cuando revisamos por primera vez los datos y debemos intentar comprender "¿de qué se trata?", vislumbrar posibles patrones y reconociendo distribuciones estadísticas que puedan ser útiles en el futuro.
- ✔ **El objetivo principal** es entender los datos y sus variables antes de proceder con algún tipo de análisis más detallado.

¿Qué es el EDA?


Visualizar
Organizar
Preparar
DATOS
Detección
Ausentes
Outlayer

¿Qué sacamos del EDA?

Data&Analytics
INNOVACIÓN Y TECNOLOGÍA

- ✔ El EDA será entonces una primera aproximación a los datos.
- ✔ Nos permitirá concluir que tenemos datos suficientes o son de muy mala calidad.
- ✔ Es todo tan aleatorio que no habrá manera de detectar patrones.
- ✔ **Hay datos suficientes y de buena calidad como para seguir a la próxima etapa.**
- ✔ Una vez que el EDA se haya completado y se hayan extraído insights, sus resultados se pueden utilizar para un análisis o modelado de datos más sofisticado, incluyendo machine learning.

Herramientas para el EDA

Data&Analytics
INNOVACIÓN Y TECNOLOGÍA

- ✔ Técnicas de agrupación en clúster y reducción de dimensiones.
- ✔ Visualización univariante de cada variable en el conjunto de datos sin formato, con estadísticas de resumen.
- ✔ Visualizaciones bivariantes y estadísticas de resumen que le permiten evaluar la relación entre cada variable del conjunto.
- ✔ Visualizaciones multivariantes para correlacionar y comprender interacciones entre diferentes variables.
- ✔ visualizaciones gráficas de datos de alta dimensión que contienen muchas variables.
- ✔ Los modelos predictivos como, por ejemplo, la regresión lineal, utilizan estadísticas y datos para predecir los resultados.

Tratamiento de los datos

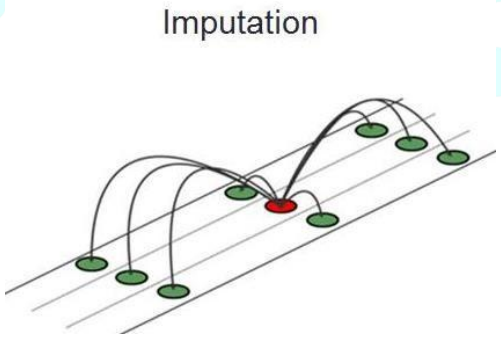
El **tratamiento de datos** engloba a todas aquellas técnicas de análisis de datos que permite mejorar la calidad de un conjunto de datos de modo que las técnicas de extracción de conocimiento puedan obtener mayor y mejor información .

Tratamiento de datos perdidos

Data&Analytics
INNOVACIÓN Y TECNOLOGÍA

Eliminación

Imputación



Valores faltantes (perdidos)

Data&Analytics
INNOVACIÓN Y TECNOLOGÍA

Si has encontrado valores inusuales en tu set de datos y simplemente quieres seguir con el resto de tu análisis, tienes dos opciones.

- ✔ Desecha la fila completa donde están los valores inusuales
- ✔ Reemplazar los valores inusuales con valores faltantes

Nota:

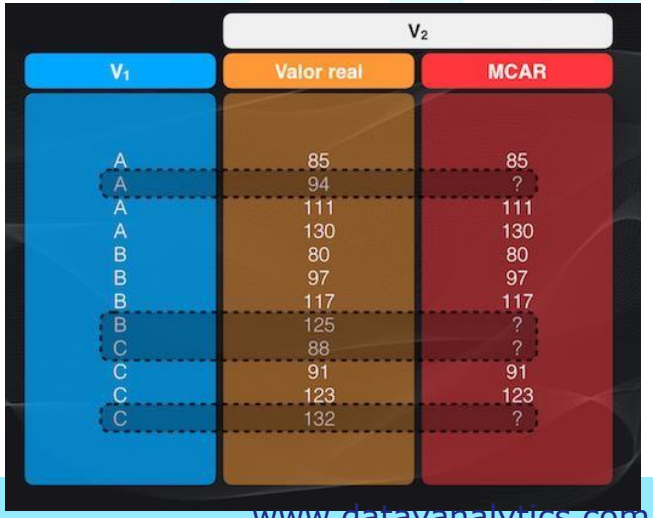
De todos modos sugiero usar estos métodos con pinzas, porque si tenemos demasiados datos faltantes (generalmente más del 10%) estaremos cambiando la distribución de nuestro dataset y esto puede afectar el modelo de Machine Learning que estemos entrenando o el análisis que hagamos posteriormente.

Mecanismos de pérdida de datos

Little y Rubin (2002) propusieron tres tipos de pérdida de datos de acuerdo con el grado de aleatoriedad. Los cuales de detallan a continuación.

- **MCAR (Missing completely at random)**

Una variable es **MCAR** si la probabilidad de pérdida de una observación para todos los individuos es la misma y no depende de las medidas de otras variables. Por ejemplo, un tubo que contiene una muestra de sangre de un individuo es roto por accidente o un cuestionario de individuo se pierde accidentalmente.



	V_2	
V_1	Valor real	MCAR
A	85	85
A	94	?
A	111	111
A	130	130
B	80	80
B	97	97
B	117	117
B	125	?
C	88	?
C	91	91
C	123	123
C	132	?

Mecanismos de pérdida de datos

Data&Analytics
INNOVACIÓN Y TECNOLOGÍA

- **MAR (Missing at random)**

Una variable es **MAR** si la probabilidad de pérdida de la observación de un individuo **depende de la información observada**. Por ejemplo, si se hace un test de aptitud a unos alumnos y a los que superan una nota de corte establecida se les hace otro más difícil mientras que a los demás no, por tanto estos tienen datos perdidos para la segunda variable y se debe a las observaciones de la primera.

V_1	V_2 Valor real	V_2 MAR
A	85	85
A	94	94
A	111	111
A	130	130
B	80	?
B	97	?
B	117	?
B	125	?
C	88	88
C	91	91
C	123	123
C	132	132

Mecanismos de pérdida de datos

- **MNAR (Missing not at random)**

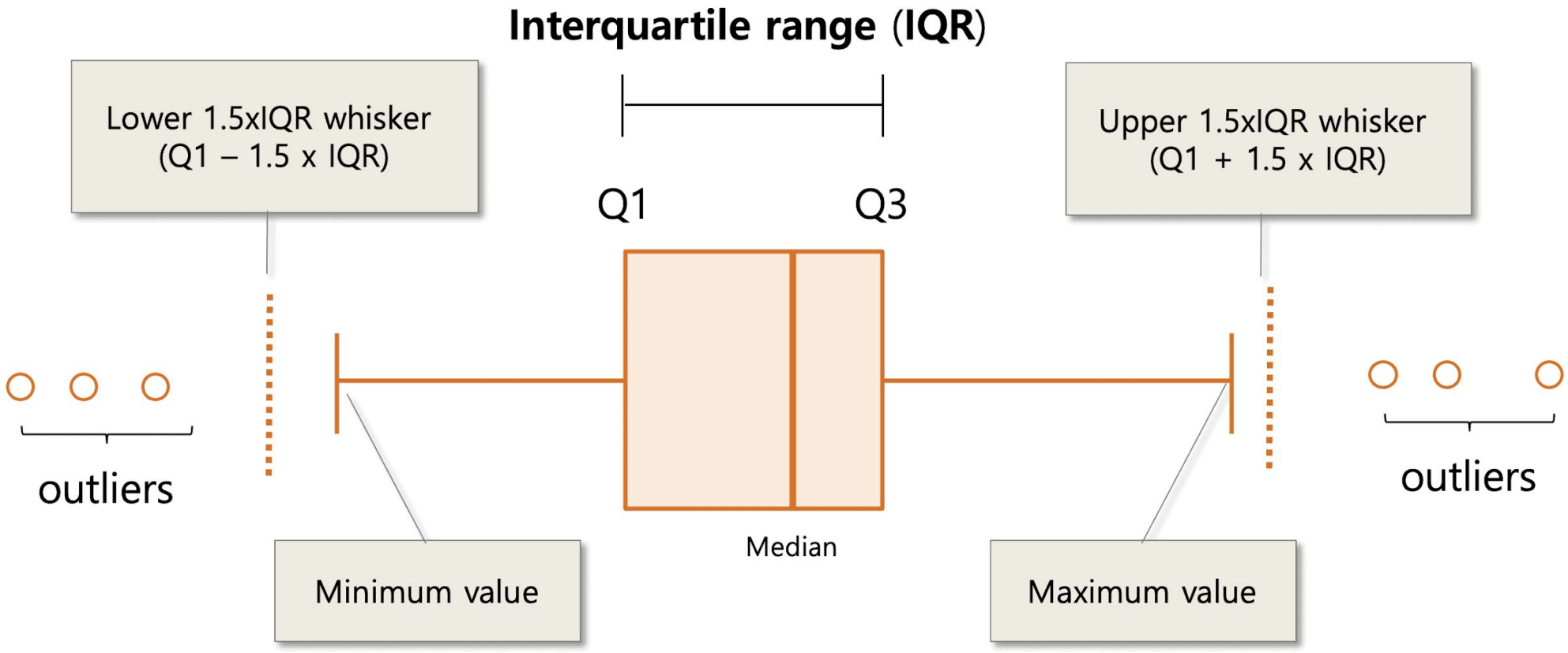
En este mecanismo la probabilidad de pérdida de datos de una variable **X <u>depende de los valores de dicha variable</u>** y también puede depender de los valores observados de las demás variables. Por ejemplo, en un ensayo clínico en el que se pruebe la eficacia de un medicamento contra la hipertensión y se realicen medidas a lo largo del tiempo, si al cabo de cierto tiempo el paciente se encuentra bien, puede decidir no acudir al lugar en el que se le toman las medidas de presión sanguínea, perdiéndose los datos de la parte final del estudio.

Valores atípicos (outliers)

Data&Analytics
INNOVACIÓN Y TECNOLOGÍA

- ✔ Los valores atípicos, conocidos en inglés como **outliers**, son puntos en los datos que parecen no ajustarse al patrón.
- ✔ Algunas veces dichos valores atípicos son errores cometidos durante el recojo de datos.
- ✔ *Es un buen hábito repetir tu análisis con y sin los valores inusuales. Si tienen un efecto mínimo en los resultados y no puedes descubrir por qué están en los datos, es razonable reemplazarlos con valores ausentes y seguir adelante con tu análisis. Sin embargo, si tienen un efecto sustancial en tus resultados, no deberías ignorarlos sin justificación.*

Valores atípicos (outliers)



¿Qué hacemos con los valores atípicos ?

Data&Analytics
INNOVACIÓN Y TECNOLOGÍA

Causa	Acciones posibles
Error de entrada de datos	Corregir el error y volver a analizar los datos.
Problema del proceso	Investigar el proceso para determinar la causa del valor atípico.
Probabilidad aleatoria	Investigar el proceso y el valor atípico para determinar si este se produjo en virtud de las probabilidades; realice el análisis con y sin el valor atípico para ver su impacto en los resultados.

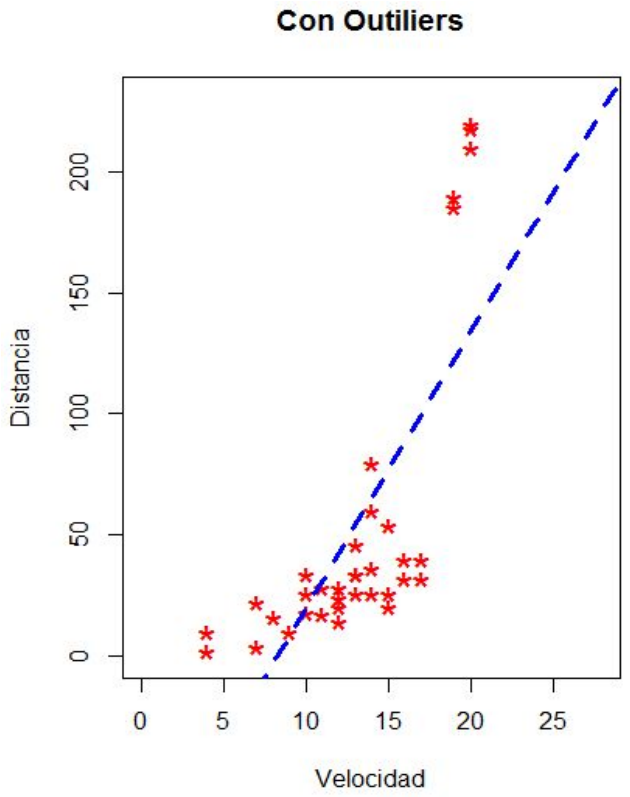
Impacto de los outliers

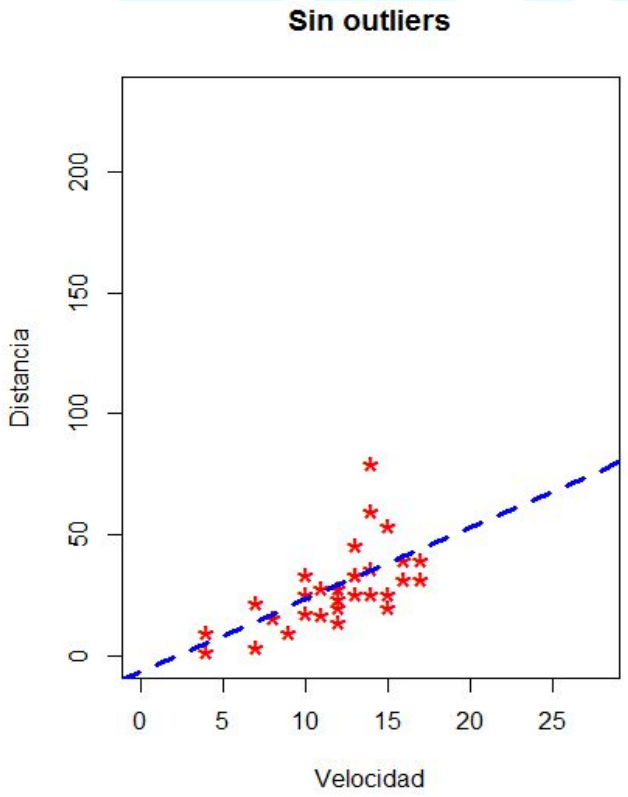
- ✔ Aumenta la varianza del error y reduce la potencia de las pruebas estadísticas.
- ✔ Si los valores atípicos no están distribuidos aleatoriamente, pueden disminuir la normalidad.
- ✔ Pueden sesgar o influir en estimaciones que pueden ser de interés sustantivo.
- ✔ También pueden influir en la hipótesis básica de regresión, ANOVA y otros supuestos del modelo estadístico.

Efecto de los outliers en el modelado

Data&Analytics
INNOVACIÓN Y TECNOLOGÍA

Es importante tener en cuenta que un outlier puede sesgar drásticamente las estimaciones y predicciones de modelo ajustado. Por ejemplo:





Se observa el cambio en la pendiente de la recta de mejor ajuste después de eliminar los valores atípicos. Si hubiéramos usado los valores atípicos para estimar el modelo, nuestras predicciones se sobreestimarían para grandes valores de velocidad, debido a que tiene mayor pendiente.

¿Cómo tratar los outliers?

- ✔ Eliminado observaciones
- ✔ Imputando los valores
- ✔ **Tratándolo como una población independiente (cluster)**

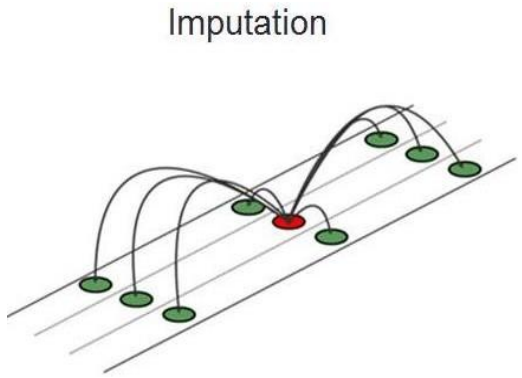
Efecto de los outliers en el modelado

- ✔ Muchas técnicas de Machine Learning (como los árboles de decisión, Random Forest, método de ensamble y las Máquinas de Soporte Vectorial) son robustos a la presencia de outliers.

- ✔ Los modelos de regresión lineal en cambio, sus estimaciones son afectadas por la presencia de outliers.

La imputación

- ✔ Para evitar una pérdida significativa de datos lo mejor es usar la imputación
- ✔ El objetivo de cualquier técnica de imputación es producir un dataset completo que después pueda ser utilizado para el aprendizaje automático.
- ✔ Existen dos técnicas: la imputación simple y la imputación múltiple.



Imputación simple: Media o mediana

Data&Analytics
INNOVACIÓN Y TECNOLOGÍA

- ✔ **En la imputación simple** se usa un algoritmo para hacer una única estimación y el valor obtenido se usa para reemplazar el dato faltante correspondiente. En este caso las tres técnicas más usadas en el Machine Learning y la Ciencia de Datos son: la imputación por la media o la mediana; la imputación por regresión y la imputación hot-deck.

- ✔ **La imputación por la media o la mediana** es la más sencilla de todas: simplemente se toman los valores conocidos en la variable donde están los datos faltantes, se calculan la media o la mediana y se reemplazan los datos faltantes con cualquiera de estos dos valores.

- ✔ El método de imputación por la media o mediana tiene la desventaja de que al reemplazar muchos datos faltantes con un único valor **estaremos cambiando la distribución de los datos**

Imputación simple: Por regresión

Data&Analytics
INNOVACIÓN Y TECNOLOGÍA

- ✔ Una alternativa es hacer la imputación por regresión. En este caso cada dato faltante es reemplazado con el valor predicho por un modelo de regresión.
- ✔ Para esto primero se combina la información de la columna con los datos faltantes con columnas en donde los datos están completos para así ajustar un modelo de regresión. Y luego se usa este modelo para predecir los datos faltantes.
- ✔ **La desventaja** es que para poder realizar cualquier tipo de regresión (regresión lineal, polinómica o regresión logística) debe haber algún tipo de correlación entre las variables que estamos usando para construir este modelo.

Imputación simple: hot-deck

- ✔ En este caso, el dato faltante es reemplazado con valores tomados de datos "cercanos" al dato faltante.
- ✔ Dentro de esta categoría el método más usado es el de k-vecinos más cercanos.
- ✔ Este algoritmo busca los k valores más cercanos (donde k es un número entero, como 2, 3, o 10 por ejemplo) y reemplaza el valor faltante con el promedio de estos vecinos.
- ✔ **La ventaja** de este método es que es mucho más preciso que la media o la mediana, y puede funcionar en lugar de la regresión cuando los datos no están correlacionados.
- ✔ **La desventaja** es que si tenemos muchos datos se requiere bastante tiempo de cómputo, porque para completar cada dato faltante se debe calcular su distancia con respecto a cada uno de los demás datos del set.

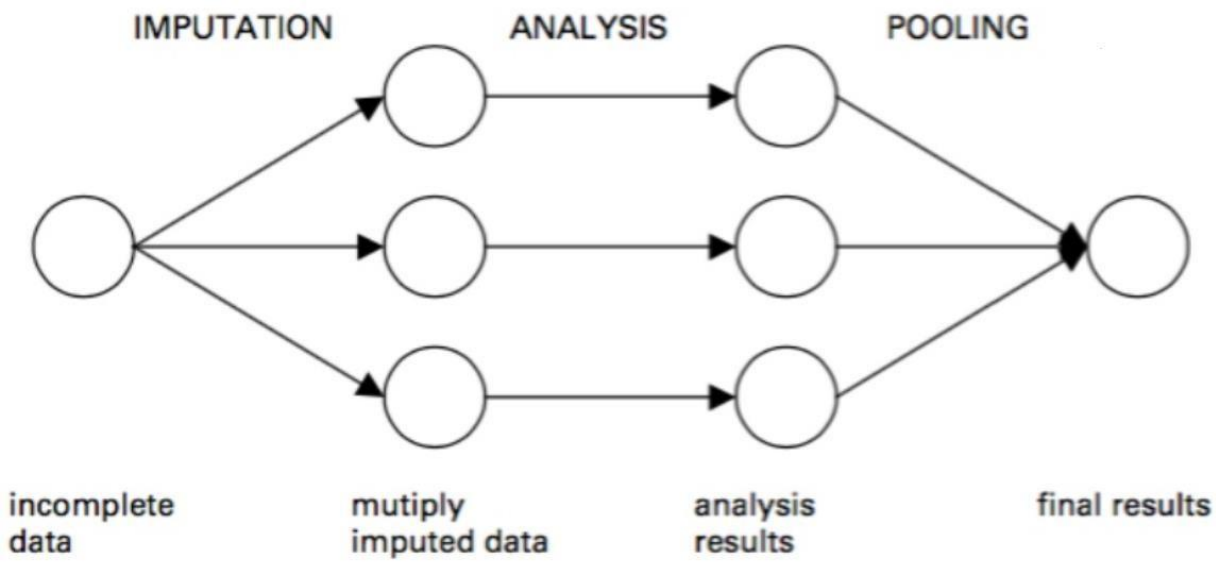
Imputación múltiple: Algoritmo MICE

Data&Analytics
INNOVACIÓN Y TECNOLOGÍA

- ✔ Una alternativa más robusta que todas las técnicas que vimos anteriormente es la imputación múltiple, que de hecho es una de las más usadas en la actualidad.
- ✔ La imputación múltiple hace múltiples estimaciones, que luego se combinan para producir un único valor, que será el usado para reemplazar el dato faltante correspondiente, con lo cual se puede disminuir el sesgo de la estimación.
- ✔ El método de imputación múltiple más usado es el algoritmo de Imputación Múltiple con Ecuaciones Encadenadas (MICE).
- ✔ Este algoritmo MICE es mucho más robusto que cualquiera de los de imputación simple, y en la práctica no cambia la distribución obtenida.

Imputación múltiple: Algoritmo MICE

La estrategia más popular que se sigue es MICE (Multiple Imputation with Chained Equations), que básicamente actualiza una a una las variables con datos faltantes según series completas de distribuciones condicionadas.



Métodos para imputación de datos

Data&Analytics
INNOVACIÓN Y TECNOLOGÍA

- ✔ Regresión lineal
- ✔ Bosque aleatorio
- ✔ k-NN (k Vecino más cercano)
- ✔ Expectativa-Maximización (EM)
- ✔ Algoritmos que funcionan con valores faltantes (XGBoost, k-NN, Árboles, etc.)
- ✔ Etc.

Ejemplo de Imputación múltiple

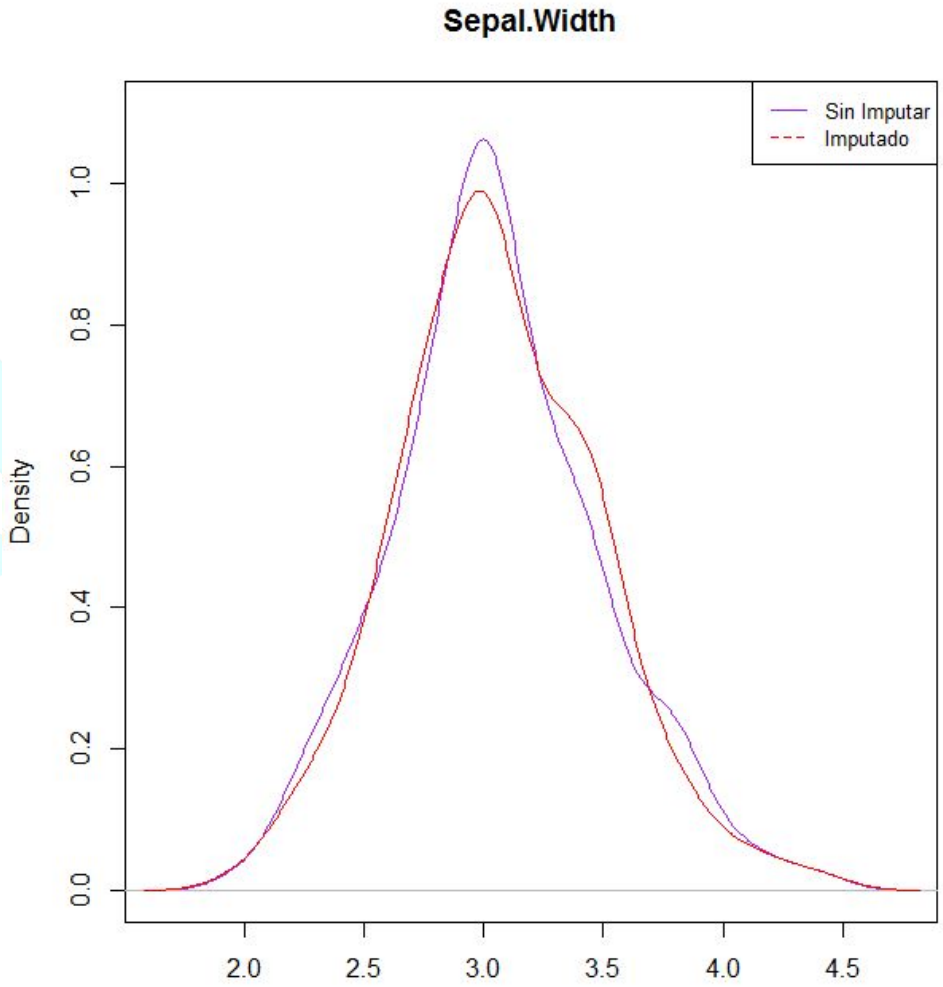
Data Set Original

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
NA	3.5	1.4	NA	setosa
NA	3	NA	NA	setosa
4.7	3.2	1.3	NA	setosa
4.6	3.1	1.5	0.2	setosa
NA	3.6	NA	0.2	setosa
5.4	NA	1.7	0.4	setosa
NA	3.4	1.4	0.3	setosa
4.8	3.4	1.5	0.2	setosa
4.4	2.9	1.4	0.2	setosa

rfImpute(Species ~ ., iris.na)

nuevo Data set Imputado

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
4.3	3.5	1.4	0.1	setosa
4.4	3	1.1	0.3	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
4.5	3.6	1.7	0.2	setosa
5.4	3.9	1.7	0.4	setosa
5.1	3.4	1.4	0.3	setosa
4.8	3.4	1.5	0.2	setosa
4.4	2.9	1.4	0.2	setosa



Transformación de datos

Data&Analytics
INNOVACIÓN Y TECNOLOGÍA

✔ Normalización

- Consiste en reescalar los valores de los datos a un rango pre-especificado.
- Normalizar los datos de entrada ayudará a acelerar la fase de aprendizaje.
- Los atributos con rangos grandes de valores tendrán más peso que los atributos con rangos de valores más pequeños, y entonces dominarán la medida de distancia.

Transformación de datos

Data&Analytics
INNOVACIÓN Y TECNOLOGÍA

✔ Normalización Z-score

$$Z = \frac{X - X^{\omega}}{\sigma} = \frac{X - center(X)}{scale(X)}$$

Este tipo de normalización funciona adecuadamente cuando:

- No se conoce el mínimo ni el máximo de los datos originales.
- Valores outlier pueden afectar el rango de los datos (pero no los elimina).
- Los datos tienen poca variabilidad.

Transformación de datos

Data&Analytics
INNOVACIÓN Y TECNOLOGÍA

✔ Transformación de Box-Cox

Las transformaciones de **Box** y **Cox** son una familia de transformaciones potenciales usadas para corregir sesgos en la distribución de errores, para corregir varianzas desiguales (para diferentes valores de la variable predictora) y principalmente para corregir la no linealidad en la relación (mejorar correlación entre las variables).

$$y_i^{(\lambda)} = \begin{cases} \dfrac{y_i^{\lambda} - 1}{\lambda} & \text{if } \lambda \neq 0, \\ \ln(y_i) & \text{if } \lambda = 0, \end{cases}$$

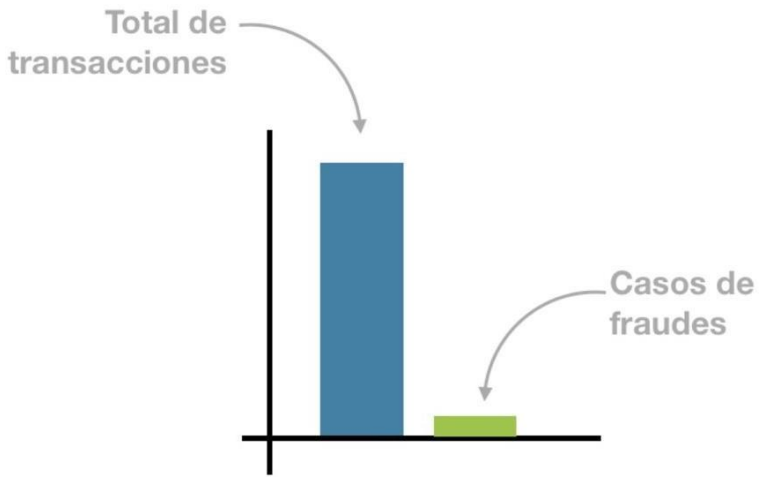
Balanceo de Datos

- ✔ La presencia de **clases desbalanceadas** es el día a día de la mayoría de científicos de datos. Este hecho es algo que ocurre muy a menudo en problemas de clasificación donde hay una diferencia muy grande entre el número de elementos de cada clase.

- ✔ El desbalanceo de clases aparece en entornos variados como pueden ser la detección de fraude, enfermedades o spam.

¿Entonces, qué podemos hacer en estos casos?

- ✔ **No está todo perdido**, existen una serie de procedimientos que podemos usar para predecir correctamente esas clases minoritarias poco representadas en nuestros datos. Alguno de los métodos consisten en cambiar los datos o la métrica a evaluar



Métodos para tratar con clases desbalanceadas

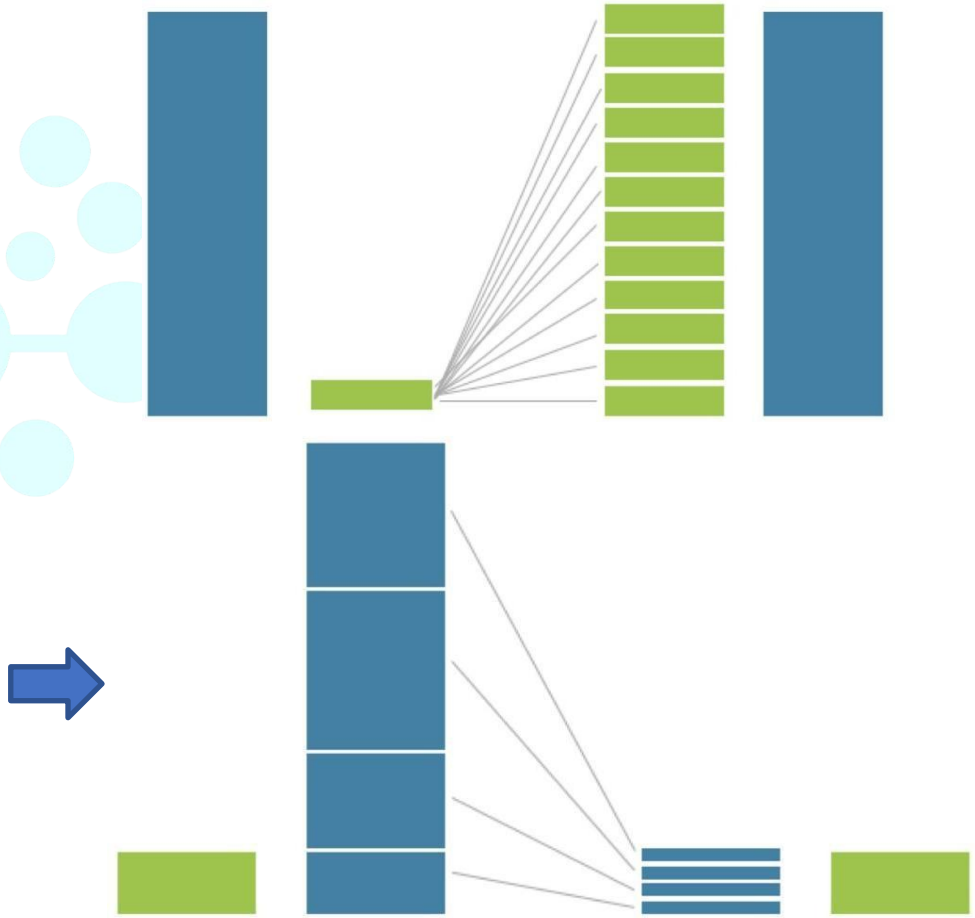
Data&Analytics
INNOVACIÓN Y TECNOLOGÍA

✔ **<u>Cambiar la métrica de evaluación</u>**: podemos usar métricas que tengan más en cuenta los datos de las clases minoritarias como son la f1, la sensitividad o la precisión.

Métodos para tratar con clases desbalanceadas

Data&Analytics
INNOVACIÓN Y TECNOLOGÍA

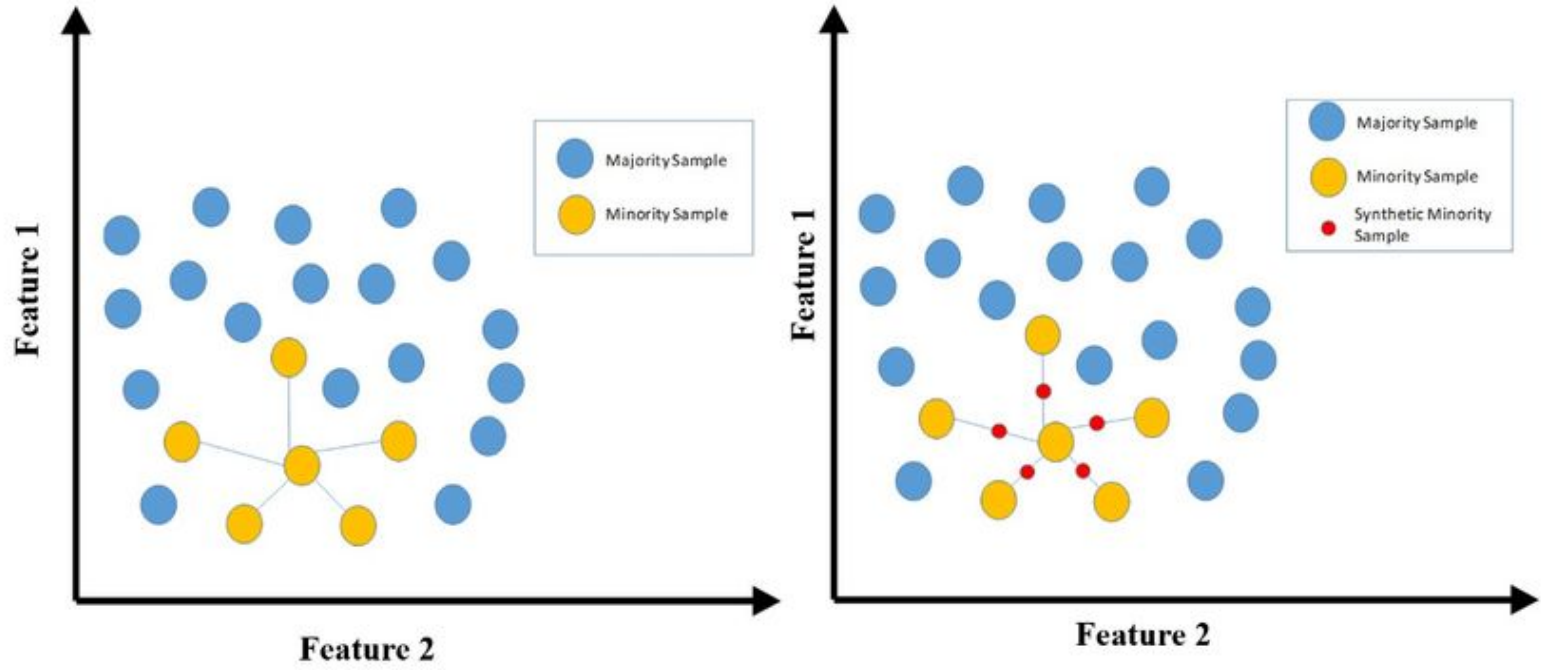
- ✔ **Muestrear:** básicamente hay dos técnicas comúnmente aceptadas.
  - ❖ La primera es el sobremuestreo de la clase minoritaria: consiste en añadir copias de la clase minoritaria para aumentar su peso sobre el total.
  - ❖ La segunda es el submuestreo de la clase mayoritaria: se basa en quitar muestra de la clase mayoritaria para intentar equilibrar el número de muestras en cada clase.



Métodos para tratar con clases desbalanceadas

Data&Analytics
INNOVACIÓN Y TECNOLOGÍA

- ✔ **Generación de muestras sintéticas:** en este caso, utilizando algoritmos como el SMOTE, somos capaces de generar más muestras de la clase minoritaria a partir de las que ya tenemos.



Métodos para tratar con clases desbalanceadas

Data&Analytics
INNOVACIÓN Y TECNOLOGÍA

Por otro lado, **existen otros métodos** que suelo usar bastante aunque no sean tan comunes como los anteriores. Son métodos que dependen del modelo estadístico a utilizar:

- ✔ **Usar algoritmos de Boosting:** estos modelos por definición suelen centrarse en mejorar los errores que cometen. Por ejemplo, en un XGBoost, aumentando el número de árboles, podemos ir corrigiendo los errores de los árboles anteriores.

- ✔ **Darle más peso a las muestras de la clase minoritaria**: algoritmos como la regresión logística permiten ponderar en mayor medida los elementos según la clase que sean. Dándole mayor peso a los elementos de la clase minoritaria se centrará en ajustarse mejor a esa clase y, de este modo, predecir mejor.

- ✔ **Usar algoritmos de Stacking y algoritmos de aprendizaje por refuerzo:** del mismo modo que los Boosting, estos algoritmos permiten ir mejorando los aciertos de la clase minoritaria.

SMOTE

- ✔ **SMOTE** es una técnica estadística de sobremuestreo de minorías sintéticas para aumentar el número de casos de un conjunto de datos de forma equilibrada.
- ✔ El componente funciona cuando genera nuevas instancias a partir de casos minoritarios existentes que se proporcionan como entrada.
- ✔ Esta implementación de SMOTE **no cambia** el número de casos de mayoría.
- ✔ Las nuevas instancias ***no son simples copias*** de los casos minoritarios existentes. En su lugar, el algoritmo toma muestras del espacio de características de cada clase de destino y de sus vecinos más próximos.
- ✔ SMOTE toma todo el conjunto de datos como una entrada, pero **solo aumenta el porcentaje de los casos minoritarios.**

Ventajas de SMOTE

Data&Analytics
INNOVACIÓN Y TECNOLOGÍA

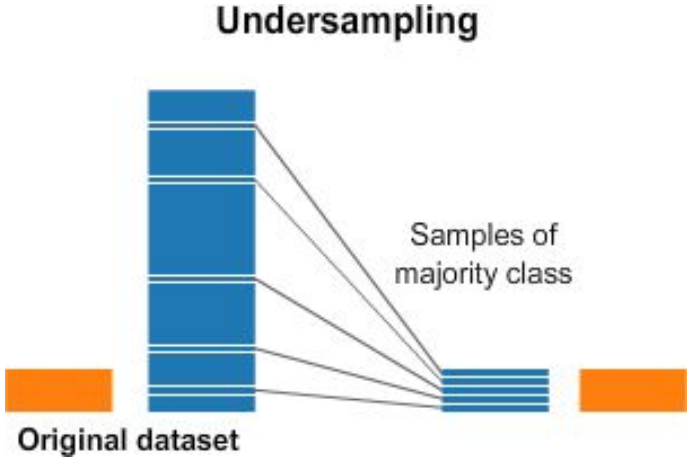
- ✔ La información no se pierde.
- ✔ Esta técnica es sencilla y se puede interpretar e implementar fácilmente en el modelo.
- ✔ Mejora el overfitting como ejemplos sintéticos. Esto ayudará a generar nuevas instancias en lugar de replicarlas.

Técnicas de remuestreo

Data&Analytics
INNOVACIÓN Y TECNOLOGÍA

Submuestreo aleatorio (Random Under-Sampling)

- ✔ El submuestreo se puede definir como eliminar algunas observaciones de la clase mayoritaria . Esto se hace hasta que se equilibre la clase mayoritaria y minoritaria.
- ✔ Significa que la clase minoritaria será la misma cantidad (1 a 1) que la clase mayoritaria, la clase minoritaria copiará sus filas.
- ✔ El submuestreo puede ser una buena opción cuando tiene una tonelada de datos, piense en millones de filas. Pero un inconveniente del submuestreo es que estamos eliminando información que puede ser valiosa.

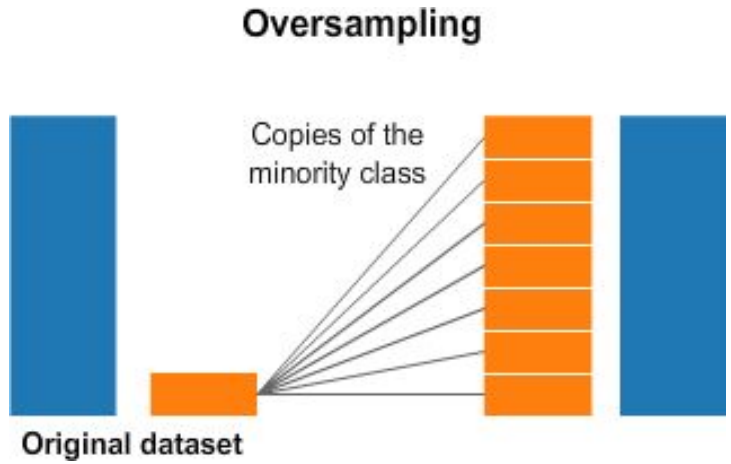


Técnicas de remuestreo

Data&Analytics
INNOVACIÓN Y TECNOLOGÍA

Sobremuestreo aleatorio(Random Over-Sampling)

- ✔ El sobremuestreo se puede definir como agregar más copias a la clase minoritaria. El sobremuestreo puede ser una buena opción cuando no tiene una gran cantidad de datos con los que trabajar.
- ✔ Establecemos la estrategia de muestreo en 1. Significa que la clase mayoritaria será la misma cantidad (1 a 1) que la clase minoritaria, la clase mayoritaria copiará sus filas.
- ✔ Una desventaja a tener en cuenta al realizar un submuestreo es que puede provocar un sobreajuste y una mala generalización en su conjunto de prueba.



Selección de variables

Data&Analytics
INNOVACIÓN Y TECNOLOGÍA

- ✔ Todos hemos visto los conjuntos de datos, en ocasiones pueden ser pequeños mientras que otros son tremendamente grandes en tamaño, en especial cuando cuentan con un gran número de características, ocasionando que sean muy difícil de procesar.

- ✔ Cuando se tiene este de tipo de conjuntos de datos de alta dimensión y se utilizan todas para la creación de modelos de Machine Learning puede ocasionar lo siguiente:

  - ❖ Las características adicionales actúan como un ruido para el cual el modelo de Machine Learning puede tener un rendimiento extremadamente bajo.

  - ❖ El modelo tarda más tiempo en entrenarse.

  - ❖ Asignación de recursos innecesarios para estas características.

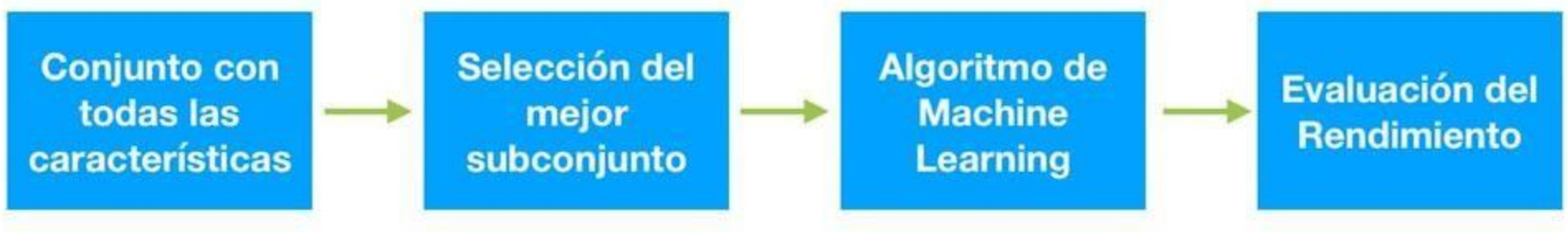
Selección de variables

Data&Analytics
INNOVACIÓN Y TECNOLOGÍA

- ✔ La Selección de Características es el proceso de seleccionar las más importantes y/o relevantes características de un conjunto de datos, con el objetivo de mejorar el rendimiento de predicción de los predictores, proporcionar predictores más rápidos y más rentables y proporcionar una mejor comprensión del proceso subyacente que generó los datos.
- ✔ A continuación, analizaremos varias metodologías y técnicas que puedes utilizar para que tus modelos funcionen mejor y de manera más eficiente.

Selección de variables: Método filtro

La siguiente imagen describe mejor los métodos de selección de características basados en filtros:



- ✔ Los métodos de filtro se utilizan generalmente como un paso de preprocesamiento de datos, la selección de características es independiente de cualquier algoritmo de Machine Learning.

- ✔ Las características se clasifican según los puntajes estadísticos que tienden a determinar la correlación de las características con la variable de resultado, ten en cuenta que la correlación es un término muy contextual y varía de un trabajo a otro.

Selección de variables: Método filtro

En la siguiente tabla puedes utilizarla para definir los coeficientes de correlación para diferentes tipos de datos, en este caso, continuo y categórico.

↓Características/Predicción →	Continuo	Categórico
Continuo	Correlación de Pearson	LDA
Categórico	Anova	Chi-cuadrado

Selección de variables: Métodos de envoltura

- ✔ Como puedes observar, un método de envoltura **necesita un algoritmo de Machine Learning** y utiliza su rendimiento como criterio de evaluación.
- ✔ Este método busca una característica que sea más adecuada para el algoritmo y tiene como objetivo mejorar el rendimiento.
- ✔ Por lo tanto, tratamos de usar un subconjunto de características y entrenamos un modelo usándolos, basándonos en las inferencias que extraemos del modelo anterior, decidimos agregar o eliminar características de su subconjunto.

Selección de variables: Métodos de envoltura

Data&Analytics
INNOVACIÓN Y TECNOLOGÍA

- ✔ Estos métodos suelen ser computacionalmente muy caros.
- ✔ Algunos ejemplos comunes de Métodos de Envoltura son los siguientes:

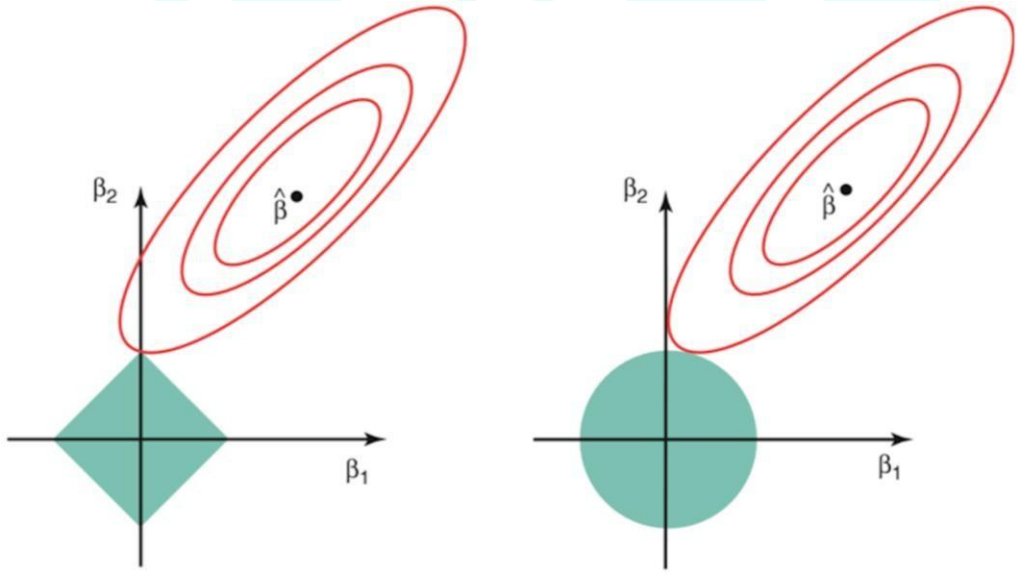
  - ❖ **Selección hacia delante (Forward Selection):** es un método iterativo en el que comenzamos sin tener ninguna característica en el modelo. En cada iteración, seguimos agregando la función que mejor mejora nuestro modelo hasta que la adición de una nueva variable no mejore el rendimiento del modelo.

  - ❖ **Eliminación hacia atrás (Backward Selection):** comenzamos con todas las características y eliminamos la característica menos significativa en cada iteración, lo que mejora el rendimiento del modelo. Repetimos esto hasta que no se observe ninguna mejora en la eliminación de características.

  - ❖ **Eliminación de características recursivas (Recursive Feature Elimination):** es un algoritmo de optimización que busca encontrar el subconjunto de funciones con mejor rendimiento. Crea repetidamente modelos y deja de lado la mejor o la peor característica de rendimiento en cada iteración. Construye el siguiente modelo con las características de la izquierda hasta que se agotan todas las características, luego clasifica las características según el orden de su eliminación.:

Selección de variables: Métodos Integrados

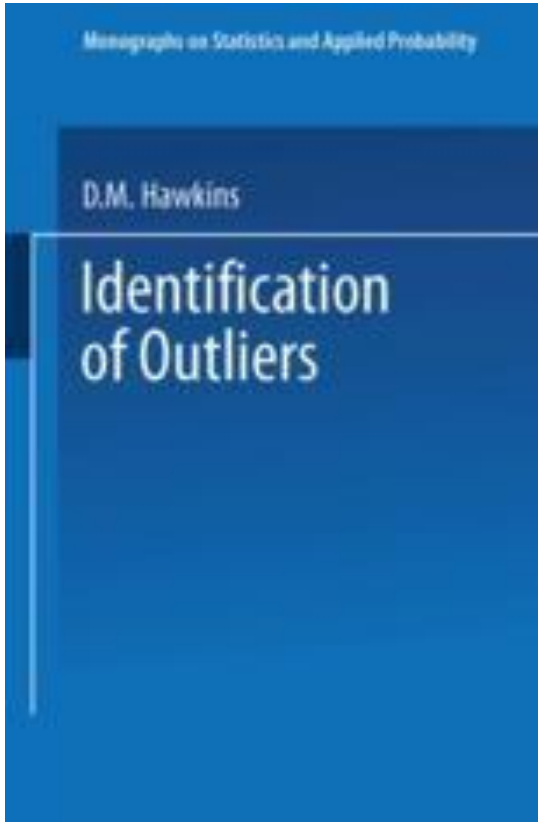
- ✔ Combina las cualidades de los métodos de filtro y envoltura. Se implementa mediante algoritmos que tienes sus propios métodos de selección de características incorporados.
- ✔ Algunos de los ejemplos más populares de estos métodos son la regresión **LASSO** y **RIDGE**, que tienen funciones de penalización incorporadas para reducir el sobreajuste.

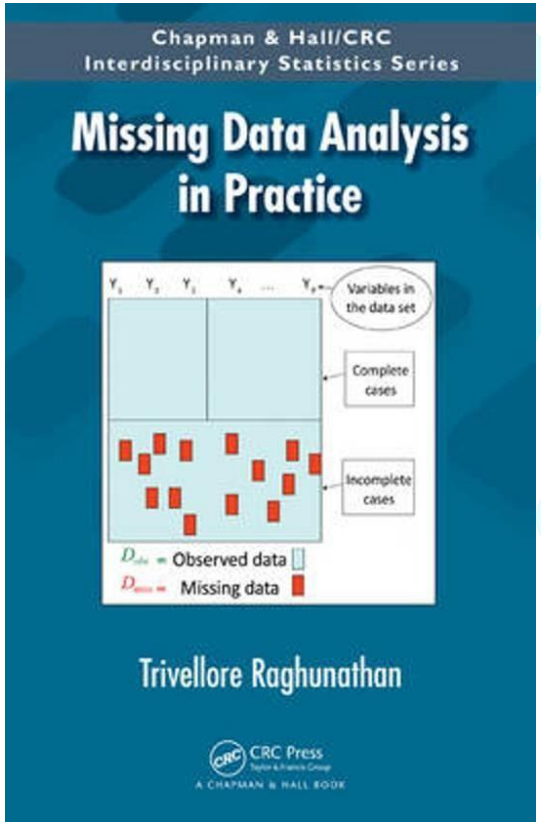


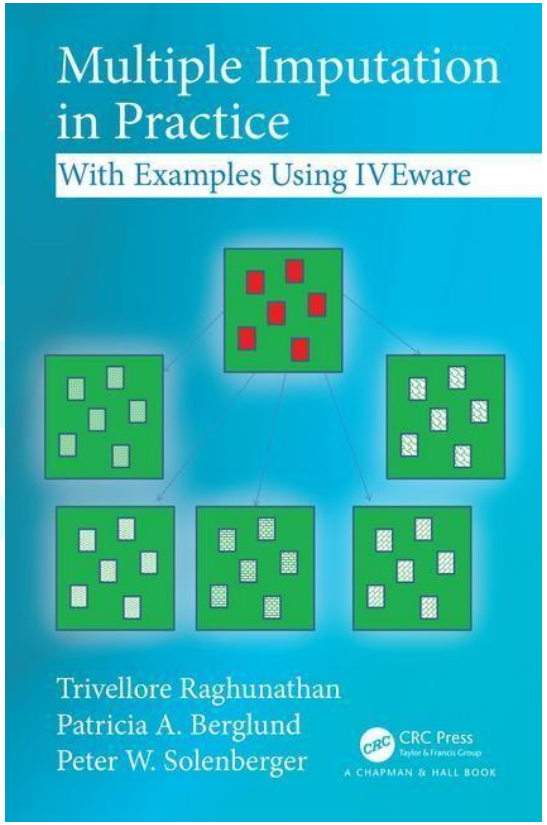
Lasso regression Ridge regression

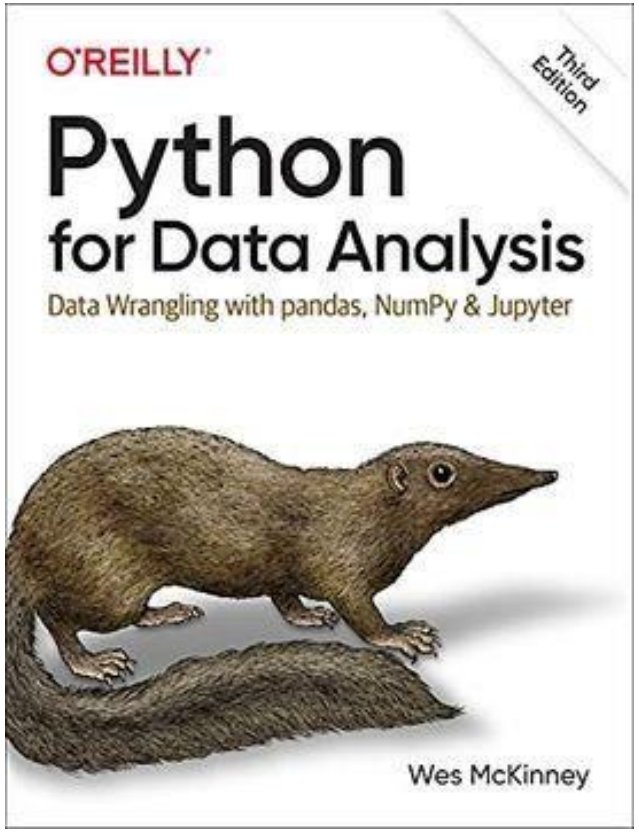
Refereficias bibliográficas

Data&Analytics
INNOVACIÓN Y TECNOLOGÍA









Data&Analytics
INNOVACIÓN Y TECNOLOGÍA

Gracias por su atención…!!!