Minería de Datos

Introducción

MSC. RENZO CLAURE ARACENA

1

Evolución en el consumo de datos

- Década de 1940: La creación de las primeras computadoras electrónicas marca el inicio de la era digital y la generación de datos en formato electrónico.
- Década de 1970: Nacen las bases de datos relacionales, lo que facilita la organización y el acceso a grandes volúmenes de datos de manera estructurada.
- Década de 1990: La popularización de Internet y la World Wide Web lleva a un explosivo crecimiento en la generación y el intercambio de datos en línea.
- Década de 2000: El auge de las redes sociales, los dispositivos móviles y el comercio electrónico contribuye a un aumento exponencial en la cantidad de datos generados por usuarios y dispositivos conectados.
- Hoy en día: Se estima que se consumen varios exabytes (1 exabyte = 1.000 petabytes) de datos por día, provenientes de una variedad de fuentes como redes sociales, transacciones comerciales, dispositivos conectados, sensores y más.
- Este crecimiento exponencial del consumo de datos ha llevado a la necesidad de desarrollar técnicas avanzadas de minería de datos para extraer información valiosa y conocimiento útil a partir de esta vasta cantidad de información disponible.

MSC. RENZO CLAURE ARACENA

2

Tipos de datos

- Datos Demográficos
- Datos Transaccionales
- Datos contextuales

3

Definición

- La **Minería de Datos** (en inglés, *Data Mining*) es un proceso interdisciplinario que utiliza técnicas matemáticas, estadísticas y de aprendizaje automático para **extraer conocimiento útil de grandes conjuntos de datos**. Este conocimiento puede ser utilizado para tomar mejores decisiones, comprender mejor los fenómenos del mundo real y descubrir nuevas oportunidades.

4

¿Qué tipo de conocimiento?

No Evidente
Novedoso

Preciso
Válido

Comprensible
Reproducible

Útil

MSC. RENZO CLAURE ARACENA

5

¿Qué NO es minería de datos?

- Hacer querys con SQL
- Usar tablas dinámicas y/o filtros en Excel
- Presentar gráficos
- Obtener promedios
- Preguntar a ChatGPT

MSC. RENZO CLAURE ARACENA

6

Objetivos

- Los principales objetivos de la minería de datos son:
  - **Descubrir patrones y tendencias** en los datos que no son evidentes a simple vista.
  - **Predecir** el comportamiento futuro de los datos.
  - **Clasificar** los datos en diferentes categorías.
  - **Agrupar** los datos en grupos similares.
  - **Encontrar relaciones** entre diferentes variables.

7

Aplicaciones

- La minería de datos se aplica en una amplia variedad de áreas, incluyendo:
  - **Negocios:** marketing, finanzas, gestión de riesgos, análisis de clientes.
  - **Ciencia:** medicina, biología, química, física.
  - **Ingeniería:** telecomunicaciones, energía, manufactura.
  - **Gobierno:** seguridad pública, salud pública, economía.

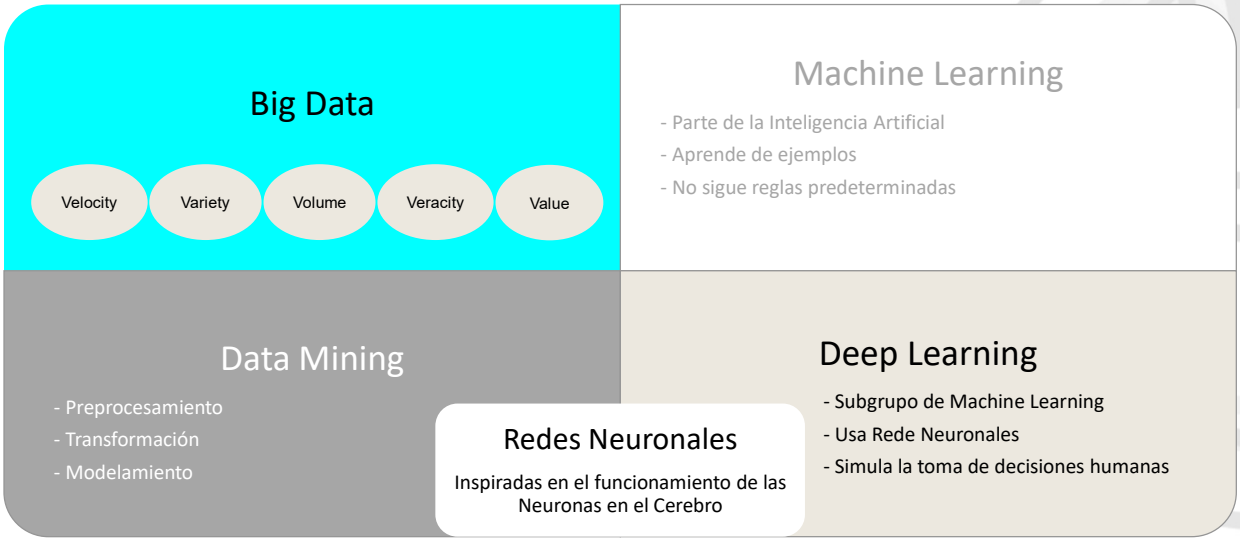
8

Beneficios

- La minería de datos puede proporcionar una serie de beneficios, incluyendo:
  - Mejora en la toma de decisiones: al proporcionar información y conocimiento que no se puede obtener de otras maneras.
  - Reducción de costos: al identificar ineficiencias y oportunidades de ahorro.
  - Aumento de los ingresos: al identificar nuevas oportunidades de negocio y mejorar la satisfacción del cliente.
  - Mejora en la comprensión del cliente: al identificar las necesidades y preferencias de los clientes.
  - Descubrimiento de nuevos conocimientos: al identificar patrones y tendencias que no se conocían previamente.

9

Las diferentes disciplinas



10

Caso de estudio

Una simple aplicación

11

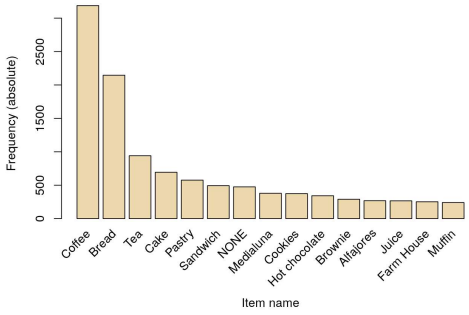
Caso de estudio

- Análisis de Market Basket:
  - El resultado de este tipo de técnica es, en términos simples, un conjunto de "reglas" que pueden entenderse como "si esto, entonces aquello".
  - Es una técnica utilizada por grandes minoristas para descubrir asociaciones entre productos. Funciona buscando combinaciones de artículos que aparecen juntos frecuentemente en las transacciones.
- Datos:
  - Una tienda de retail esta interesada en determinar los patrones de compra màs comunes
  - Tamaño del dataset: 15K observaciones
  - **Fecha:** Variable categórica que indica la fecha de las transacciones (formato YYYY-MM-DD). La columna incluye fechas desde el 30/10/2016 hasta el 09/04/2017.
  - **Hora:** Variable categórica que indica la hora de las transacciones (formato HH:MM:SS).
  - **Transacción:** Variable cuantitativa que permite diferenciar las transacciones. Las filas que comparten el mismo valor en este campo pertenecen a la misma transacción, por lo que el conjunto de datos tiene menos transacciones que observaciones.
  - **Producto:** Variable categórica que contiene los productos comprados.

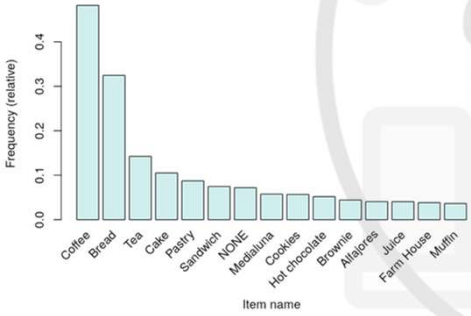
12

Análisis de los datos

- Frecuencia absoluta



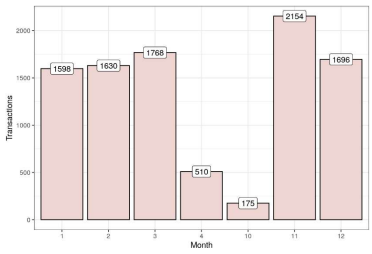
- Frecuencia relativa



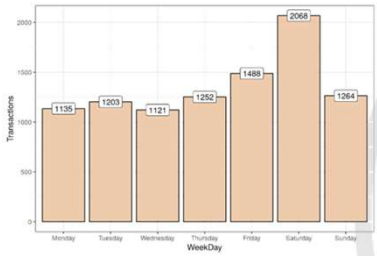
13

Transacciones

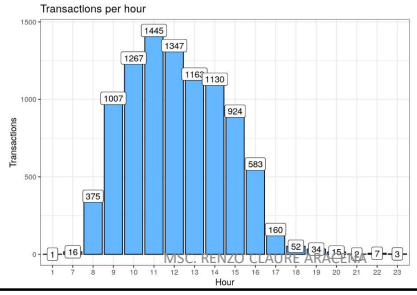
- Mensuales



- Diarias



- Por hora



14

Métricas utilizadas

- Soporte: indica que tan frecuente está un par de ítems (X y Y) en un set de datos

$$supp(X \Rightarrow Y) = \frac{|X \cup Y|}{n}$$

- Confianza: Es el porcentaje en el que Y es comprado con X, es un indicador de cuan a menudo la regla o par es encontrado:

$$conf(X \Rightarrow Y) = \frac{supp(X \cup Y)}{supp(X)}$$

- Lift: Es el la relación entre el soporte observado y el esperado (supuesto de que X y Y son independientes). Elevados valores indican fuerte asociación.

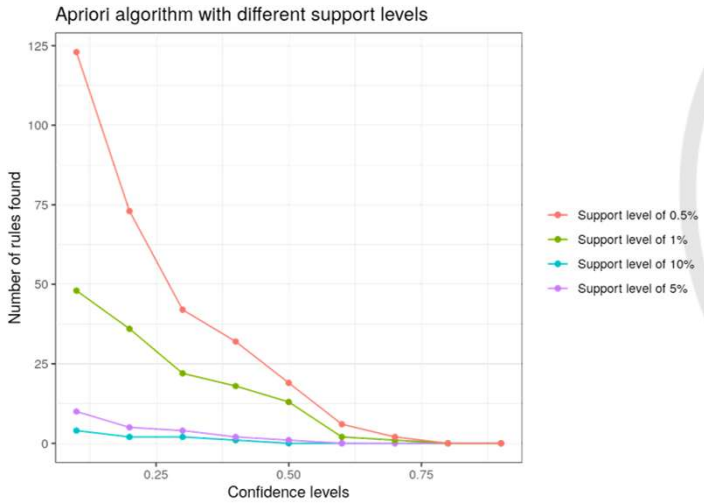
$$lift(X \Rightarrow Y) = \frac{supp(X \cup Y)}{supp(X)supp(Y)}$$

- Conviction: Un elevado valor indica que el consecuente ocurre fuertemente después del precedente.

$$conv(X \Rightarrow Y) = \frac{1 - supp(Y)}{1 - conf(X \Rightarrow Y)}$$

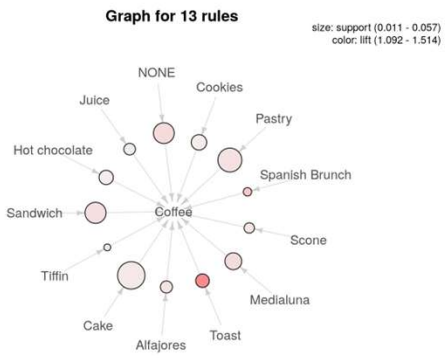
15

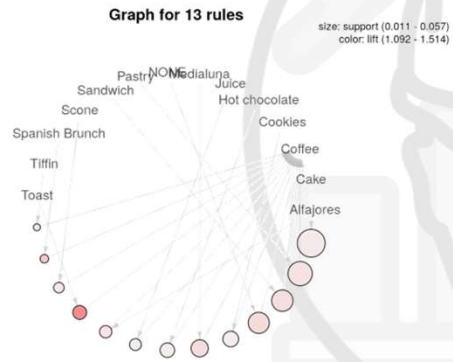
Reglas y soporte



16

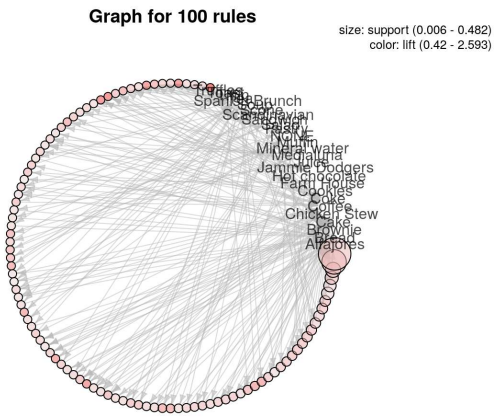
Representación de las reglas





17

Demasiadas reglas



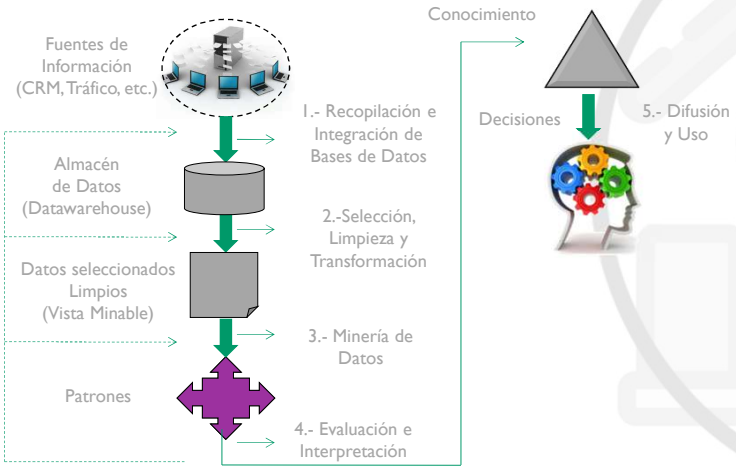
18

El proceso de la minería de datos

Como aplicarla

19

Proceso de extracción de conocimiento



20

1. Definición del Problema

- El primer paso es definir claramente el problema que se quiere resolver con la minería de datos. Esto implica identificar los objetivos del proyecto, las preguntas que se quieren responder y las variables que se van a analizar.

21

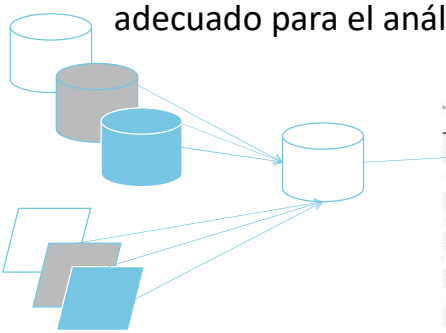
2. Recolección de Datos

- El siguiente paso es recolectar los datos que se van a utilizar en el proyecto. Estos datos pueden provenir de diferentes fuentes, como bases de datos, archivos CSV, APIs, etc. Es importante que los datos sean de alta calidad y relevantes para el problema que se quiere resolver.

22

3. Preprocesamiento de Datos

- Los datos recolectados generalmente no están listos para ser analizados. Es necesario preprocesarlos para eliminar errores, corregir inconsistencias, normalizar variables y transformarlas en un formato adecuado para el análisis.



	symboling	normalized-losses	make	fuel-type	aspiration	num-of-doors	body-style	drive-wheels	engine-location	wheel-base	...	engine-size	fuel-system	bore	stroke	compression-ratio	horsepower
0	3	NaN	alfa-romero	gas	std	two	convertible	rwd	front	88.6	...	130.0	mpfi	3.47	2.68	9.0	111.0
1	1	NaN	alfa-romero	gas	std	two	hatchback	rwd	front	94.5	...	152.0	mpfi	2.68	3.47	9.0	154.0
2	2	164	audi	gas	std	four	sedan	fwd	front	99.8	...	109.0	mpfi	3.19	3.40	10.0	102.0
3	2	164	audi	gas	std	four	sedan	4wd	front	99.4	...	136.0	mpfi	3.19	3.40	8.0	115.0
4	2	NaN	audi	gas	std	two	sedan	fwd	front	99.8	...	136.0	mpfi	3.19	3.40	8.5	110.0
5	1	158	audi	gas	std	four	sedan	fwd	front	105.8	...	136.0	mpfi	3.19	3.40	8.5	110.0
6	1	NaN	audi	gas	std	four	wagon	fwd	front	105.8	...	136.0	mpfi	3.19	3.40	8.5	110.0
7	1	158	audi	gas	turbo	four	sedan	fwd	front	105.8	...	131.0	mpfi	3.13	3.40	8.3	140.0
8	0	NaN	audi	gas	turbo	two	hatchback	4wd	front	99.5	...	131.0	mpfi	3.13	3.40	7.0	160.0
9	2	192	bmw	gas	std	two	sedan	rwd	front	101.2	...	108.0	mpfi	3.50	2.80	8.8	101.0

23

4. Análisis de Datos

- En esta etapa, se aplican técnicas de minería de datos para extraer conocimiento de los datos. Algunas de las técnicas más comunes son:
  - Regresión: para predecir una variable a partir de otras variables.
  - Clasificación: para clasificar los datos en diferentes categorías.
  - Agrupamiento: para agrupar los datos en grupos similares.
  - Asociación: para encontrar relaciones entre diferentes variables.

24

5. Evaluación de Modelos

- Es importante evaluar la precisión y confiabilidad de los modelos que se han desarrollado. Esto se puede hacer mediante diferentes técnicas, como la validación cruzada y el test de holdout.

25

6. Implementación de Resultados

- Finalmente, los resultados del proyecto de minería de datos deben ser implementados en la práctica. Esto puede implicar desarrollar nuevos productos o servicios, mejorar procesos existentes o tomar decisiones más informadas.

26

Metodologías para la minería de datos

- **CRISP-DM:** Es una metodología ampliamente utilizada que define seis fases:
  - Definición del problema: identificar los objetivos del proyecto, las preguntas que se quieren responder y las variables que se van a analizar.
  - Comprensión del problema: recopilar información sobre el problema y los datos disponibles.
  - Preparación de datos: limpiar, transformar y normalizar los datos.
  - Modelado: seleccionar y aplicar técnicas de minería de datos.
  - Evaluación: evaluar la precisión y confiabilidad de los modelos.
  - Implementación: implementar los resultados del proyecto en la práctica.

27

Metodologías para la minería de datos

- **SEMMA:** Es una metodología similar a CRISP-DM, pero que se centra en cinco fases:
  - Muestreo: seleccionar una muestra representativa de los datos.
  - Exploración: analizar los datos para identificar patrones y tendencias.
  - Modificación: preparar los datos para el análisis.
  - Modelado: seleccionar y aplicar técnicas de minería de datos.
  - Evaluación: evaluar la precisión y confiabilidad de los modelos.

28

Preprocesamiento de los datos

Los primeros pasos

MSC. RENZO CLAURE ARACENA

29

Feature engineering

- El preprocesamiento de datos es una etapa crucial en cualquier proyecto de minería de datos. Su objetivo es preparar los datos para el análisis, de modo que sean limpios, consistentes y estén en un formato adecuado para las técnicas de minería de datos que se van a utilizar.
- Las principales técnicas de preprocesamiento de datos son:
  1. Limpieza de datos.
  2. Transformación de datos.
  3. Creación de datos.
  4. Reducción de dimensionalidad.
  5. Imputación de valores faltantes.

MSC. RENZO CLAURE ARACENA

30

1. Limpieza de datos

- Manejo de valores perdidos.
- Detección y corrección de errores: eliminar valores inconsistentes o valores atípicos.

31

1. Limpieza de datos

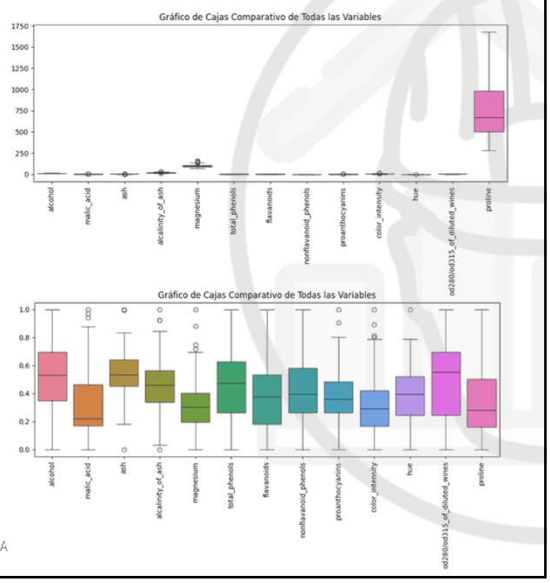
Tratamiento de datos incompletos

- Los datos no siempre están disponibles
- La pérdida de datos puede deberse a:
  - Mal funcionamiento de los equipos de recolección
  - Inconsistencia con otros datos seleccionados y por lo tanto eliminada
  - Data mal comprendida o no completada
  - Ciertos datos pueden no ser considerados importantes al momento de la recolección
  - No existen cambios históricos en los datos
- Los datos perdidos pueden ser inferidos
- Ignorar la variable, ignorar el caso
- Llenar de forma manual los datos faltantes (tedioso, poco confiable)
- Llenar automáticamente con:
  - Un valor constante (cero, desconocido, etc.)
  - Media, moda de la población
  - Media o moda de los miembros de la misma clase
  - Valor más probable: basado en la inferencia Bayesiana.

32

2. Transformación de datos

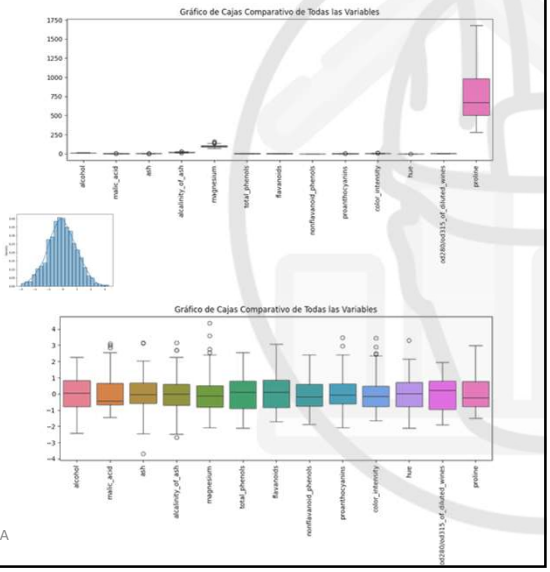
- **Normalización:**
  - Es el re-escalado de los datos a un rango común, por ejemplo: 0 a 1 o -1 a 1. Es utilizado principalmente cuando se tienen muchas unidades y escalas diferentes.
  - Cuando aplicarla:
    - Cuando las distribuciones de los datos son desconocidas o se sospecha que no son Normales.
    - Es adecuada cuando la distribución de los datos no es gaussiana o cuando los algoritmos se benefician de datos escalados en un rango definido. Por ejemplo, es comúnmente utilizada en el procesamiento de imágenes donde los valores de los píxeles deben estar en un rango específico.
    - Es sensible a los valores atípicos (outliers). Si hay valores extremos, pueden comprimir el resto de los datos en un rango muy estrecho.



33

2. Transformación de datos

- **Estandarización:**
  - Es un caso especial de normalización, donde se escalan los datos con una media de 0 y una desv. est. de 1.
  - Cuando aplicarla:
    - Variables de comportamiento natural.
    - Algoritmos de base de gradiente, optimización .
    - Algoritmos que involucran distribuciones, como Regresión, la regresión logística y el análisis de componentes principales (PCA).



34

2. Transformación de datos

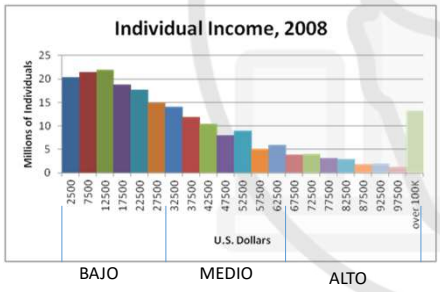
En el caso de datos contínuos:

- **Transformación logarítmica: El logaritmo es excelente para reducir el sesgo positivo en datos con distribuciones sesgadas a la derecha. Es útil cuando los datos tienen un amplio rango de valores y una cola larga. Es importante tener en cuenta que esta transformación solo se puede aplicar a valores positivos.**
- **Transformación de raíz cuadrada: Similar al logaritmo, la raíz cuadrada también reduce el sesgo positivo, pero de manera menos drástica.Es útil para datos de conteo o cuando se desea una transformación más suave.**

35

2. Transformación de datos

- **Discretización (Bining):**
  - Dividir variables continuas en intervalos discretos (bins).
  - Útil para simplificar modelos y manejar no linealidades.
  - Binning de igual ancho, o de igual frecuencia.



36

3. Codificación de datos

- **Categorical encoding:**
  - Label Encoding:
    - Asignar un número entero único a cada categoría.
    - Adecuado para variables ordinales (con un orden inherente).
  - One-Hot Encoding:
    - Crear una columna binaria para cada categoría.
    - Adecuado para variables nominales (sin un orden inherente).
  - Ordinal Encoding:
    - Similar a Label Encoding, pero se realiza de manera explicita el orden de las categorias.
  - Target Encoding:
    - Reemplazar cada categoría con la media de la variable objetivo para esa categoría.
    - Puede ser muy efectivo, pero propenso al sobreajuste (overfitting).
  - Frequency Encoding:
    - Reemplazar cada categoría con su frecuencia en el conjunto de datos.

Variable Original
perro
gato
caballo
perro

Variable Transformada
1
2
3
1

Variable Original
Medicina 1
Medicina 2
Medicina 1
Medicina 3

Variable Transformada		
Medicina 1	Medicina 2	Medicina 3
1	0	0
0	1	0
1	0	0
0	0	1

Variable Original
2. Licenciado
1. Escuela Basica
3. Masterado
4. Doctorado

Variable Transformada
2
1
3
4

Variable Original	V. O.
azul	20
azul	30
amarillo	10
verde	50
verde	60

Variable Transformada
25
25
10
55
55

37

4. Creación de Nuevas Características

- Feature Creation/Generation
  - Extracción de Características de Fecha y Hora:
    - Extraer componentes como día de la semana, mes, año, hora, etc.
  - Creación de Características de Interacción:
    - Combinar dos o más características existentes (e.g., multiplicación, división).
  - Extracción de Características de Texto:
    - TF-IDF (Term Frequency-Inverse Document Frequency): Ponderar la importancia de las palabras en un documento.
    - Word Embeddings (e.g., Word2Vec, GloVe): Representar palabras como vectores numéricos.

38

5. Selección de Características

- **Feature Selection**
  - Métodos de Filtro (Filter Methods): Seleccionar características basadas en su correlación con la variable objetivo. Ejemplos: Correlación de Pearson, Chi-cuadrado.
  - Métodos de Envoltura (Wrapper Methods): Evaluar subconjuntos de características utilizando un modelo predictivo. Ejemplos: Selección hacia adelante, selección hacia atrás.
  - Métodos Integrados (Embedded Methods):Seleccionar características como parte del proceso de entrenamiento del modelo. Ejemplos: Regularización L1 (Lasso), importancia de características de árboles de decisión.
  - Reducción de Dimensionalidad:
    - PCA (Principal Component Analysis): Transformar los datos a un espacio de menor dimensión.
    - t-SNE (t-distributed Stochastic Neighbor Embedding): Reducir la dimensionalidad para visualización.

39

Laboratorio Python/Limpieza

40

Laboratorio
Python/Transformación

- Min_NB_2_DATA_TRANSFORMATION

41

Spark / PysPark

42

Apache Spark: Descripción y Características

- Apache Spark es un framework de procesamiento de datos distribuido diseñado para ser rápido y fácil de usar. Es especialmente útil para manejar grandes volúmenes de datos y realizar tareas de procesamiento en paralelo. Spark es compatible con varios lenguajes de programación, incluyendo Python (a través de PySpark), Scala, Java y R.

43

Características principales de Spark

- **Velocidad**
- **Facilidad de uso**
- **Unificación**
- **Escalabilidad**
- **Tolerancia a fallos**

44

Fundamentos

- Lazy Evaluation (Evaluación perezosa):
  - Spark utiliza un enfoque de "**evaluación perezosa**" para optimizar el procesamiento de datos. Esto significa que las transformaciones (como map, filter, join, etc.) no se ejecutan inmediatamente cuando se llaman. En su lugar, Spark construye un DAG (Directed Acyclic Graph) que representa las operaciones que se deben realizar. La ejecución real solo ocurre cuando se llama a una acción (como collect, count, save, etc.).
  - Esto permite a Spark **optimizar** el plan de ejecución y minimizar el número de pasos necesarios para completar una tarea.
- Transformaciones vs. Acciones:
  - **Transformaciones**: Son operaciones que crean un nuevo **RDD (Resilient Distributed Dataset)** a partir de uno existente. Ejemplos comunes incluyen **map, filter, flatMap, join, groupBy**, etc. Las transformaciones son lazy, lo que significa que no se ejecutan hasta que se llama a una acción.
  - **Acciones**: Son operaciones que **devuelven un valor** al programa principal o escriben datos en un sistema de almacenamiento externo. Ejemplos incluyen **collect, count, take, saveAsTextFile**, etc. Las acciones disparan la ejecución de las transformaciones acumuladas hasta ese punto.

45

Laboratorio Python

- Min_NB_1_FEATURE_ENG2_pyspark

46

Análisis descriptivo

47

Análisis descriptivo de Variables

- Motivación
  - Para mejor comprensión de los datos, tendencia central, variación y dispersión
- Características de la dispersión de los datos
  - max, min, rango, cuantiles, outliers, varianza, etc
- Dimensiones numéricas
  - Dispersión: medido con distintos niveles de granularidad
  - Boxplot, muy útil para ver la distribución de los datos

48

Análisis descriptivo de variables

Medidas de tendencia central

- Media
  - Muestral y Poblacional
  - Ponderada
  - Recortada
- Mediana

$$mediana = L_1 + (\frac{n/2-(\sum freq)l}{freq_{mediana}})\text{ancho}$$

- Moda

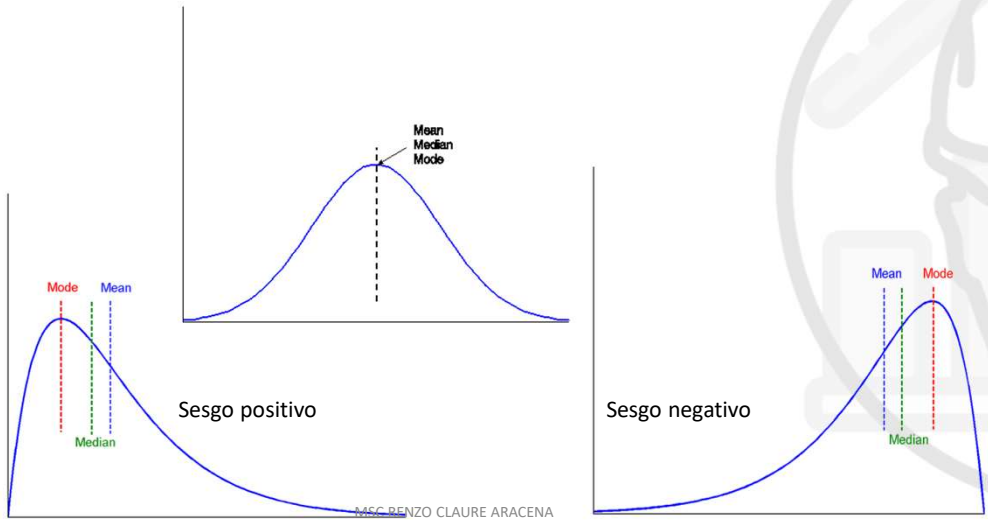
$$mean - mode = 3 \times (mean - median)$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

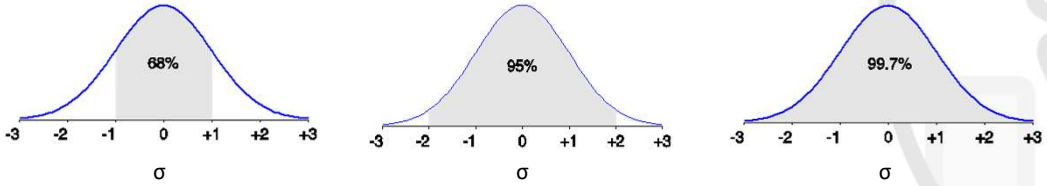
Análisis descriptivo de Variables

Sesgo de los datos



Análisis descriptivo de Variables

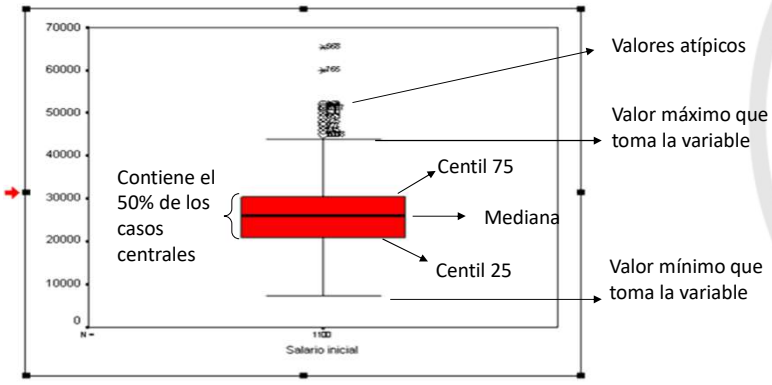
Propiedades de la distribución normal



51

Análisis descriptivo de Variables

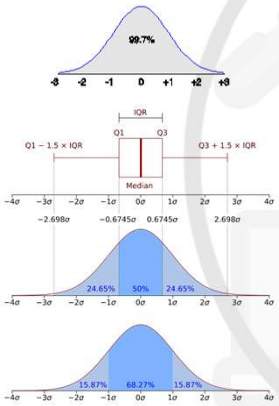
Box Plot



52

Detección de anomalías

- Z-score, para distribuciones normales.
- Intervalo Intercuartílico.
- Local Outlier Factor (LOF).



Análisis de correlación

El coeficiente de correlación lineal de Pearson

- ¿Qué es correlación?
- ¿Qué diferencia hay entre correlación y causalidad?
- ¿Cómo se calcula e interpreta el coeficiente de correlación de Pearson?

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} \qquad \rho_{X,Y} = \frac{\mathbb{E}[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

Análisis de correlación

La significancia estadística

- ¿Qué tan probable es que el coeficiente de correlación sea por azar?
- La Hipótesis nula indica que mientras más grande sea esta probabilidad, es más probable (valga la redundancia), que los resultados se hayan dado meramente por el azar
- Rangos comunes para P value
  - <0,001 hay una alta significancia estadística
  - <0,05 la significancia estadística es moderada
  - <0,1 la significancia es débil
  - >0,1 se descarta la evidencia de que la correlación sea significativa

$$t = \frac{r * \sqrt{n-2}}{\sqrt{1-r^2}}$$

MSC RENZO CLAURE ARACENA

55

Transformaciones de variables

- Transformación logarítmica: El logaritmo es excelente para reducir el sesgo positivo en datos con distribuciones sesgadas a la derecha. Es útil cuando los datos tienen un amplio rango de valores y una cola larga. Es importante tener en cuenta que esta transformación solo se puede aplicar a valores positivos.
- Transformación de raíz cuadrada: Similar al logaritmo, la raíz cuadrada también reduce el sesgo positivo, pero de manera menos drástica.Es útil para datos de conteo o cuando se desea una transformación más suave.

MSC. RENZO CLAURE ARACENA

56

Comparación de Medias

- Comparación de una muestra
  - Sirve para comparar la media de una muestra (aleatoria de una población), contra un valor simple que representa la media hipotética de la población. Por ejemplo, el diámetro de un balón es de 50cm, se desea saber si una muestra de la producción de balones difiere de este diámetro.
  - Hipótesis Nula, la media muestral es igual a la media hipotética o conocida de la población
- Comparación de medias de grupos independientes
  - Se usa para comparar muestras de grupos independientes, por ejemplo comprara si las dimensiones de dos flores de diferente especie son significativamente distintas
  - Hipótesis Nula: Ambos grupos tienen medias iguales
- Comparación de muestras dependientes
  - Se usa para comparar muestras de un mismo grupo antes y después de haber aplicado un tratamiento. Por ejemplo, deseamos saber si la media de peso de un grupo de sujetos a los que se les aplico un tratamiento para adelgazar, es istinta, antes y después del tratamiento.
  - Hipótesis Nula: Las medias muestrales de ambos grupos son iguales

57

Análisis Anova

- Es una técnica estadística que nos ayuda a identificar si existen diferencias estadísticamente significativas en las medias entre grupos
- La hipótesis nula es que las medias son iguales
- Utiliza el estadístico F (Fisher), que mide la diferencia entre las medias, mientras más grande es, mayor es la diferencia
- P Value, mide la significancia estadística de la diferencia de las medias
- El análisis ANOVA de dos entradas servirá para evaluar la interacción de dos factores sobre una variable objetivo
- NB_4

58

Análisis Visual de los Datos

Análisis visual de los datos

MSC. RENZO CLAURE ARACENA

59

Introducción

- **Definición y Alcance**
  - Qué es Visual Data Mining y Visual Analytics?
  - Diferencias y similitudes entre visualización de datos, análisis visual y minería de datos.
- **Importancia en el Proceso de Data Mining**
  - Cómo la visualización facilita la exploración y comprensión de grandes volúmenes de datos.
  - Ejemplos de aplicaciones en la industria (finanzas, salud, marketing, etc.).
- **Objetivos de la Sesión**
  - Comprender los fundamentos teóricos.
  - Conocer las principales herramientas y técnicas.
  - Aplicar conceptos mediante ejemplos prácticos en Python.

MSC. RENZO CLAURE ARACENA

60

Tipos de Gráficos

- **Visualización de Datos:**
  - **Definición:** Es el proceso de representar datos gráficamente para comunicar información de manera clara y efectiva.
  - **Enfoque:** Se centra en la estética, el diseño y la claridad del mensaje.
  - **Ejemplo:** Gráficos de barras, líneas, histogramas, mapas de calor, etc.
- **Análisis Visual:**
  - **Definición:** Es la exploración interactiva de datos a través de visualizaciones, donde el usuario aplica su capacidad interpretativa para identificar patrones, tendencias y anomalías.
  - **Enfoque:** La interacción y el descubrimiento de insights a partir de la exploración visual.
  - **Ejemplo:** Uso de dashboards interactivos que permiten filtrar y profundizar en la información en tiempo real.
- **Minería de Datos:**
  - **Definición:** Es el proceso de aplicar algoritmos y técnicas estadísticas para extraer patrones o conocimiento de grandes conjuntos de datos.
  - **Enfoque:** Automatización y modelado de datos para la detección de relaciones complejas.
  - **Ejemplo:** Algoritmos de clustering, clasificación, regresión y detección de outliers.

61

Principios de Visualización de Datos

- **Claridad:**
  - El gráfico debe ser fácil de entender y no debe tener elementos que distraigan al lector.
- **Precisión:**
  - Los datos deben ser representados de forma precisa y no debe haber distorsiones.
- **Concisión:**
  - El gráfico debe mostrar solo la información relevante y no debe ser demasiado complejo.
- **Eficacia:**
  - El gráfico debe comunicar el mensaje de forma efectiva y no debe ser confuso.

62

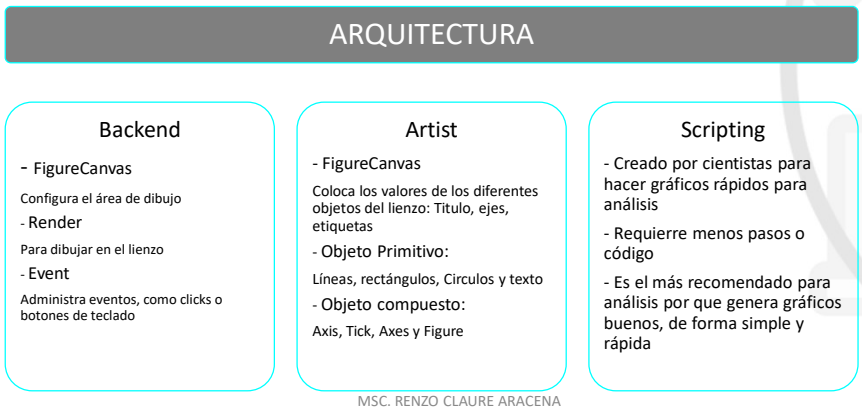
Herramientas y Frameworks

- **Bibliotecas en Python**
  - Matplotlib y Seaborn: Gráficos estáticos y exploratorios.
  - Plotly y Bokeh: Creación de visualizaciones interactivas.
  - Altair: Enfoque declarativo para gráficos.
- **Plataformas Complementarias**
  - Uso de Jupyter Notebook / Lab para integración interactiva.
  - Mención breve de herramientas BI (Tableau, PowerBI) y su integración con Python.

63

Matplot

- Módulo popular para gráficos basados en Matlab.
- Jhon Hunter .
- Arquitectura:



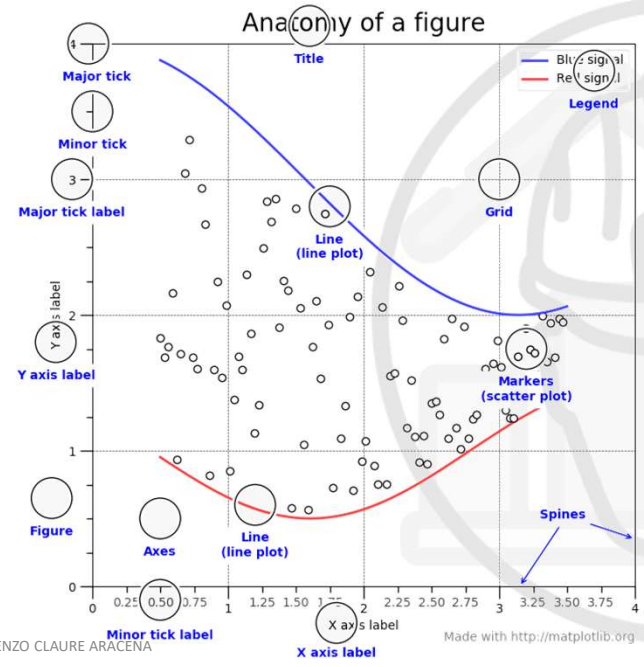
64

Arquitectura

- **Scripting (Pyplot):**
  - **Ventajas:** Rapidez y facilidad de uso, ideal para prototipos y análisis exploratorios.
  - **Limitación:** Menor control sobre los detalles de los objetos gráficos, ya que se maneja de forma implícita.
- **API Orientada a Objetos (Artist):**
  - **Ventajas:** Control total sobre cada elemento de la visualización. Es ideal para aplicaciones complejas o cuando se requiere personalización avanzada.
  - **Desventaja:** Requiere más código y una comprensión más profunda de la estructura interna de Matplotlib.

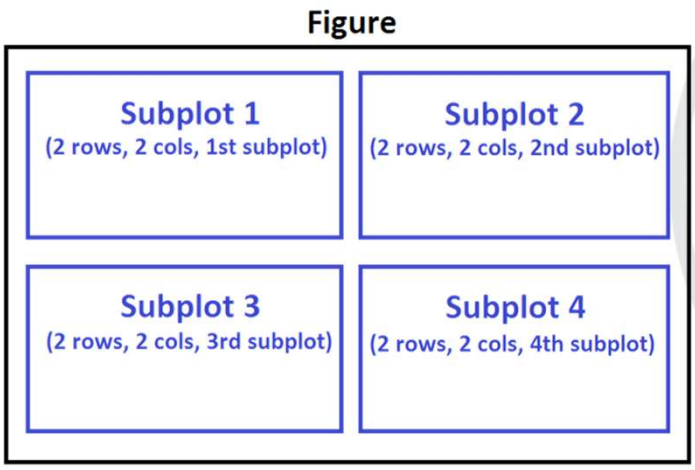
65

Terminología de un gráfico



66

Subplots



Plotly

- Plotly es una biblioteca de visualización de datos interactiva para Python, R y JavaScript. Se destaca por generar gráficos dinámicos que permiten zoom, desplazamiento y visualización mejorada de datos.

Características principales:

- **Interactividad:** Soporta zoom, hover y selección de datos.
- **Facilidad de uso:** Permite crear gráficos con pocas líneas de código.
- **Compatibilidad con frameworks:** Se integra con Dash, Jupyter Notebook y Streamlit.
- **Variedad de gráficos:** Soporta gráficos de barras, líneas, dispersión, mapas geoespaciales, gráficos 3D y más.
- **Personalización avanzada:** Permite modificar colores, fuentes, etiquetas, estilos de líneas y más.

Ventajas de Plotly vs Matplotlib

Característica	Plotly	Matplotlib
Interactividad	Sí, permite zoom y selección	No, estático por defecto
Gráficos 3D	Soportados con facilidad	Limitados y menos eficientes
Personalización	Intuitiva y flexible	Requiere más código
Integración Web	Compatible con Dash y Streamlit	No nativamente

69

Laboratorio Python/Visualizaciòn

70

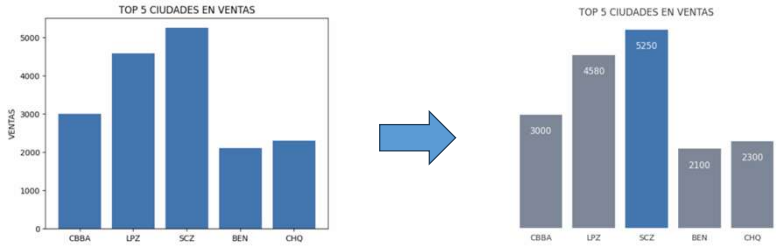
Consejos finales

- Utiliza un título claro y conciso que describa el contenido del gráfico.
- Etiqueta los ejes del gráfico con las variables que se representan.
- Utiliza una escala adecuada para los ejes.
- Utiliza colores y leyendas de forma efectiva.
- No sobrecargues el gráfico con demasiada información.
- Prueba diferentes tipos de gráficos para encontrar el que mejor se adapte a tus necesidades.
- Lectura complementaria:
  - Ten Simple Rules for Better Figures

Trabajo para entregar/Python

- Cree un grafico de barras de las ventas por ciudad de una empresa:

```
ciudades =['CBBA', 'LPZ', 'SCZ', 'BEN', 'CHQ']
pos = np.arange(len(ciudades))
ventas = [3000, 4580, 5250, 2100, 2300]
```



Prueba en grupo de Visualizacion

- Descarga los datos:
  - Usa el dataset "penguins" de la librería seaborn.
  - Limpieza básica: Realiza un tratamiento de valores nulos si es necesario.
  - Crea dos gráficos usando Matplotlib:
    - Gráfico de dispersión (scatter plot): Relación entre el tamaño del pico y la masa corporal de los pingüinos. Distinga los colores de las burbujas por especie.
    - Histograma: Distribución de la masa corporal de los pingüinos. Cree dos gráficos en el mismo lienzo, uno para cada sexo, con colores distintos por especie.
  - Personaliza los gráficos:
    - Agrega títulos, etiquetas en los ejes y leyendas.
    - Usa colores diferentes para cada especie.
    - Expliquen brevemente qué observan en los gráficos.
- El dataset "penguins" contiene las siguientes columnas relevantes:
  - species: Especie del pingüino.
  - bill_length_mm: Longitud del pico (en mm).
  - bill_depth_mm: Profundidad del pico (en mm).
  - body_mass_g: Masa corporal (en gramos).