PUCE Virtual
EDUCACIÓN FLEXIBLE

Adquisición, gestión y gobernanza de datos

Metodologías del Proceso de Descubrimiento del Conocimiento

Clase 7

MAESTRÍA EN
SISTEMAS DE INFORMACIÓN
Mención Data Science

La excelencia no se improvisa

INTRODUCCIÓN

La exploración del proceso de descubrimiento del conocimiento ha adquirido una relevancia fundamental en el contexto académico y profesional actual. Las metodologías que sustentan este proceso no solo facilitan la generación de nuevas ideas, sino que también permiten estructurar el conocimiento de manera que sea accesible y aplicable en diversas disciplinas. Desde enfoques cualitativos hasta técnicas cuantitativas, cada metodología ofrece herramientas específicas que ayudan a los investigadores a formular hipótesis, recolectar datos y analizar resultados, promoviendo un ciclo continuo de aprendizaje y mejora.

A medida que nos adentramos en este tema, es crucial considerar cómo estas metodologías se adaptan a distintos contextos y problemáticas. La interdisciplinariedad y la colaboración se han vuelto esenciales, permitiendo que diversas perspectivas se integren en el proceso de descubrimiento. Este análisis no solo ampliará nuestra comprensión sobre cómo se genera y valida el conocimiento, sino que también nos ofrecerá perspectivas sobre las mejores prácticas y las tendencias emergentes en la investigación. Así, nos proponemos explorar las metodologías más efectivas, sus aplicaciones y su impacto en el avance del conocimiento en la era contemporánea.

RDA2

Esquematizar metodologías apropiadas para la obtención de modelos inductivos (predictivos) y deductivos (descriptivos) que faciliten los procesos de toma de decisiones basadas en datos en las instituciones.

Clase 7: Metodologías del Proceso de Descubrimiento del Conocimiento

7.1 CRISP – DM

CRISP-DM (Cross-Industry Standard Process for Data Mining) es una metodología ampliamente adoptada para la minería de datos que proporciona un marco estructurado para llevar a cabo proyectos de análisis de datos. Desarrollada en la década de 1990, esta metodología se caracteriza por su enfoque iterativo y su adaptabilidad a diferentes industrias y contextos (Shearer, 2000). CRISP-DM consta de seis fases: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación y despliegue. Cada una de estas etapas es crucial para garantizar que el proceso de minería de datos no solo sea técnico, sino que también esté alineado con los objetivos estratégicos de la organización (Chapman et al., 2000).

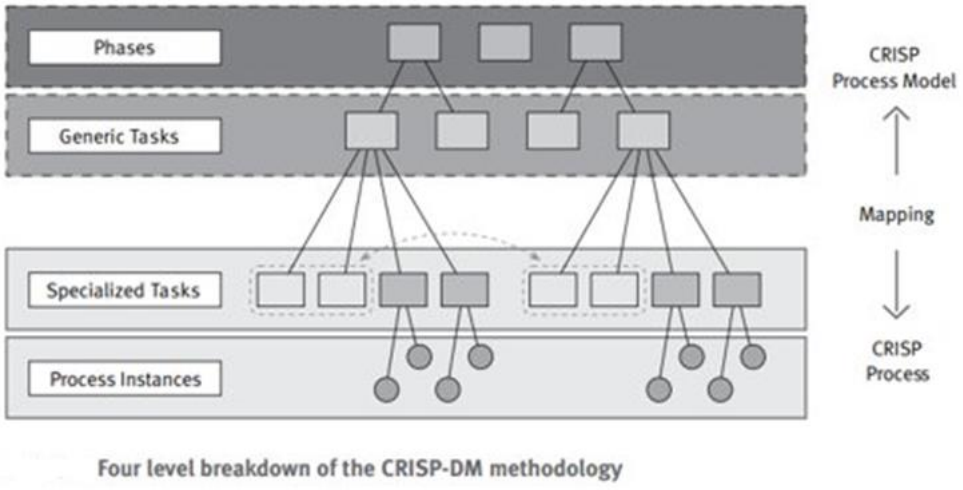
La implementación de CRISP-DM permite a las organizaciones abordar proyectos de minería de datos de manera sistemática, reduciendo el riesgo de fracaso y mejorando la calidad de los resultados. Su enfoque en la colaboración multidisciplinaria y la iteración continua fomenta una cultura de aprendizaje dentro de los equipos, lo que, a su vez, mejora la capacidad de adaptación a cambios y la posibilidad de aprovechar nuevas oportunidades (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). A través de esta introducción a CRISP-DM, se busca proporcionar una comprensión profunda de cómo esta metodología puede ser aplicada para transformar datos en conocimiento útil y significativo en una variedad de contextos.

7.1.1 CRISP como metodología

CRISP, como metodología, se entiende como un modelo jerárquico de procesos. Consiste en cuatro niveles de abstracción: fases, tareas genéricas, tareas especializadas e instancias de proceso (Montero, 2024).

Figura 1

CRISP como metodología



Fuente: https://cic-smartbase.blogspot.com/2016/05/metodologia-crisp-dm-parte-i.html

Las cuatro fases avanzan desde actividades más generales hasta tareas cada vez más específicas. Finalmente, la cuarta fase se refiere a los procesos de instancia, que consisten en registrar las acciones, decisiones y resultados de un proyecto de minería de datos. A continuación, se presenta un ejemplo:

- **Fases:**
  Preprocesamiento de datos
- **Tareas Genéricas:**

  Limpieza de datos
- **Tareas Específicas:**

  Manejo de valores faltantes (missing values)
- **Procesos de Instancia:**

  Calcular la media para las variables numéricas y el valor más frecuente para las variables categóricas.

Capacidad de Adaptación de CRISP-DM

CRISP-DM (Cross-Industry Standard Process for Data Mining) es ampliamente reconocido por su notable capacidad de adaptación, lo que lo convierte en una metodología versátil aplicable a diversas industrias y contextos de investigación. Su estructura flexible

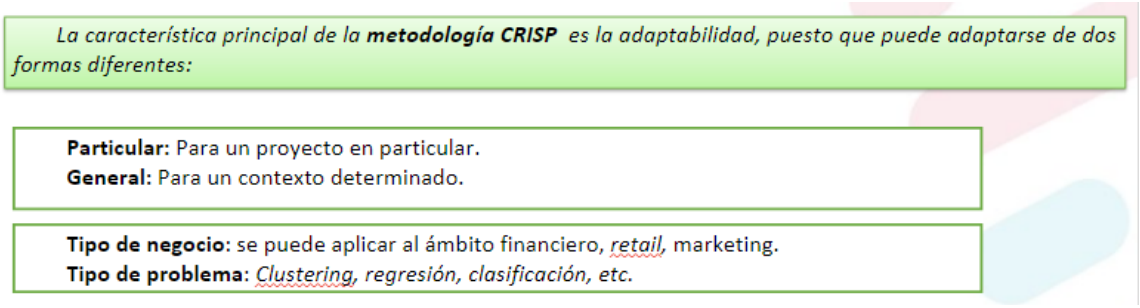
permite a los equipos de trabajo ajustar cada fase del proceso según las necesidades específicas del proyecto, facilitando la incorporación de nuevas técnicas y herramientas conforme surgen (Shearer, 2000). Esta adaptabilidad no solo optimiza el proceso de minería de datos, sino que también facilita una integración más eficaz de diferentes disciplinas, promoviendo la colaboración multidisciplinaria.

Además, la capacidad de adaptación de CRISP-DM se extiende a su enfoque iterativo, que permite a los equipos regresar y realizar ajustes en fases anteriores basándose en nuevos hallazgos o cambios en el contexto del negocio. Esta característica es esencial en un entorno donde los datos y las necesidades del mercado pueden evolucionar rápidamente (Chapman et al., 2000). De este modo, CRISP-DM no solo proporciona un marco estructurado para la minería de datos, sino que también fomenta una cultura de aprendizaje continuo y mejora, lo que resulta crucial para afrontar los desafíos de la toma de decisiones basada en datos en la actualidad.

A continuación, se presenta un ejemplo de la capacidad de adaptación de CRISP-DM:

Figura 1

Adaptación CRISP DM



Fuente: Elaboración Propia

7.1.2 CRISP como modelo de proceso

Este modelo de proceso ofrece una visión general del ciclo de vida de un proyecto de minería de datos, describiendo las fases y las tareas correspondientes, así como las relaciones entre ellas (Chapman, 2000).

CRISP-DM (Cross-Industry Standard Process for Data Mining) es un modelo integral que guía a los profesionales a lo largo del proceso de minería de datos. Compuesto por seis etapas interrelacionadas, este modelo proporciona un enfoque sistemático para convertir datos en conocimiento útil. Las etapas son: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación y despliegue (Shearer, 2000).

Comprensión del negocio: En esta etapa, se definen los objetivos del proyecto y se identifican las necesidades del negocio. Se busca comprender el contexto en el que se realizará el análisis. Por ejemplo, una empresa de telecomunicaciones puede querer reducir la tasa de cancelación de clientes. En este punto, se establecerían metas claras y medibles para guiar el análisis (Chapman et al., 2000).

Comprensión de los datos: Después de definir los objetivos, se realiza una exploración inicial de los datos disponibles. Esto incluye la recolección de datos relevantes y su evaluación para identificar patrones, anomalías o características importantes. Por ejemplo, se pueden revisar datos sobre el historial de clientes, patrones de uso y datos demográficos para entender mejor el comportamiento de los clientes.

Preparación de los datos: Esta fase abarca la limpieza y transformación de los datos recolectados. Se deben manejar valores faltantes, eliminar duplicados y crear nuevas variables que faciliten el análisis. Por ejemplo, en el caso de la empresa de telecomunicaciones, se podrían generar variables que representen el tiempo promedio de llamadas o el número de quejas recibidas por cliente.

Modelado: En esta etapa, se aplican técnicas de modelado para crear algoritmos que identifiquen patrones en los datos. Se seleccionan los métodos de minería de datos adecuados, como árboles de decisión, regresión o redes neuronales, y se ajustan los parámetros necesarios. Por ejemplo, se podría usar un modelo de clasificación para predecir qué clientes tienen más probabilidades de cancelar su servicio.

Evaluación: Una vez que se ha creado el modelo, es crucial evaluar su rendimiento. Se comparan los resultados obtenidos con los objetivos del negocio establecidos en la primera etapa. En nuestro ejemplo, se mediría la precisión del modelo para predecir cancelaciones y se determinaría si se cumplen los objetivos iniciales.

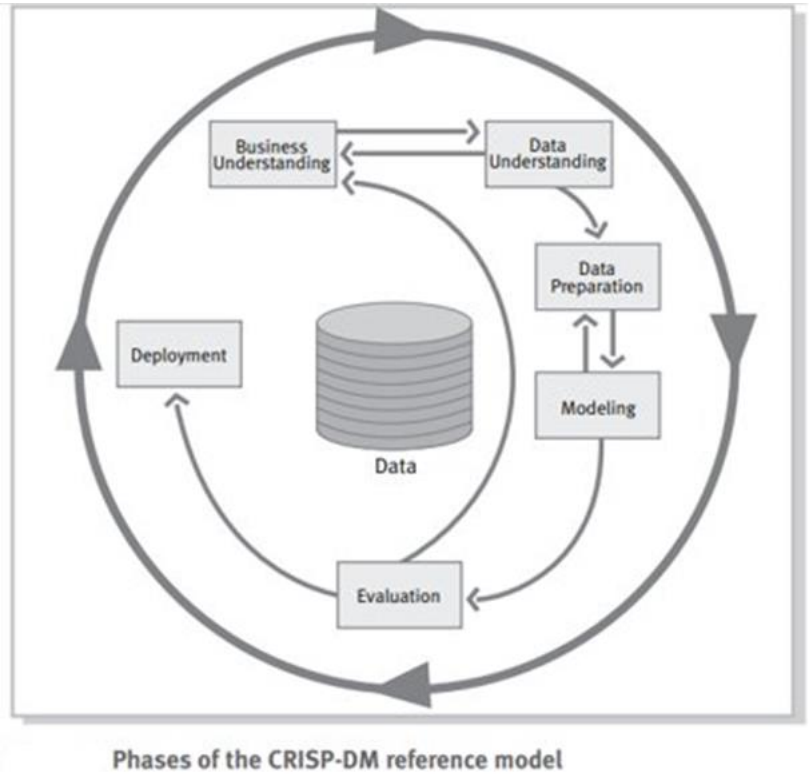
Despliegue: La etapa final implica implementar los resultados del análisis en la práctica. Esto puede incluir la creación de informes, la automatización de procesos o la

implementación de estrategias basadas en los hallazgos del modelo. Por ejemplo, si el modelo identifica a los clientes en riesgo de cancelación, la empresa podría lanzar campañas específicas para retener a esos clientes.

En conjunto, estas etapas ofrecen un enfoque holístico para la minería de datos, permitiendo a las organizaciones convertir datos en decisiones estratégicas efectivas.

Figura 2

Fases del modelo CRISP-DM



Fuente: Chapman (2000)

El siguiente video explica detalladamente en qué consiste la metodología CRISP, desde el minuto 7 hasta el minuto 28: Explicación Metodología CRISP.

7.2 SEMMA

SEMMA (Sample, Explore, Modify, Model, Assess) es una metodología desarrollada por SAS para el proceso de minería de datos. Esta metodología se enfoca en las prácticas de análisis y permite a los analistas de datos crear modelos predictivos de manera eficiente. SEMMA está compuesta por cinco etapas que, aunque se presentan de manera secuencial, a menudo requieren iteraciones y revisiones para obtener los mejores resultados (SAS Institute, 2021).

Sample (Muestreo): La primera etapa consiste en seleccionar una muestra representativa de los datos disponibles. Este paso es crucial porque trabajar con grandes volúmenes de datos puede ser ineficiente y costoso. Al elegir un subconjunto adecuado, los analistas pueden realizar exploraciones iniciales sin comprometer la integridad del análisis posterior. Por ejemplo, en un estudio de comportamiento del cliente, se podría seleccionar un 10% de los datos de transacciones de un año para realizar un análisis preliminar (Delen & Rosati, 2018).

Explore (Exploración): En esta etapa, se examinan los datos seleccionados para identificar patrones, tendencias y relaciones significativas. Utilizando técnicas de visualización y estadísticas descriptivas, los analistas buscan entender la estructura de los datos y detectar cualquier anomalía. Por ejemplo, se pueden crear gráficos de dispersión para visualizar la relación entre el gasto del cliente y la frecuencia de compra, lo que podría revelar insights valiosos sobre el comportamiento de compra (SAS Institute, 2021).

Modify (Modificación): Esta fase implica la limpieza y transformación de los datos para mejorar su calidad y relevancia para el análisis. Esto puede incluir la creación de nuevas variables, la normalización de datos o el manejo de valores faltantes. Por ejemplo, si se detectan errores en los datos de edad de los clientes, se corregirían o imputarían los valores faltantes para asegurar que el análisis sea preciso (Delen & Rosati, 2018).

Model (Modelado): En esta etapa, se aplican técnicas de modelado estadístico y de minería de datos para crear modelos predictivos. Dependiendo del objetivo del análisis, se pueden utilizar métodos como regresiones, árboles de decisión o redes neuronales. Por ejemplo, en un proyecto de segmentación de clientes, se podría aplicar un modelo de clustering para identificar grupos de clientes con características similares, lo que permitiría personalizar estrategias de marketing (SAS Institute, 2021).

Assess (Evaluación): La etapa final implica la evaluación del modelo desarrollado para asegurar que cumpla con los objetivos planteados. Esto incluye medir su precisión, capacidad de generalización y su rendimiento en comparación con otros modelos. Por ejemplo, se podrían utilizar métricas como la precisión, el recall y el área bajo la curva ROC para evaluar un modelo de clasificación. Si el modelo no cumple con las expectativas, se puede regresar a etapas anteriores para realizar ajustes (Delen & Rosati, 2018).

El siguiente video explica detalladamente en qué consiste la metodología SEMMA. Desde el minuto 28 hasta el minuto 38: Explicación Metodología SEMMA

7.3 Elementos de la Ciencia de Datos

La ciencia de datos es un campo multidisciplinario que combina diversas disciplinas y técnicas para extraer conocimientos significativos a partir de grandes volúmenes de datos. Los elementos fundamentales de la ciencia de datos incluyen la recolección de datos, el

almacenamiento y la gestión de datos, el análisis de datos, la visualización y la comunicación de resultados. Cada uno de estos componentes juega un papel crucial en el proceso de transformación de datos en información útil para la toma de decisiones (Provost & Fawcett, 2013).

Recolección de datos: Este es el primer paso en la ciencia de datos, donde se obtienen datos de diversas fuentes, como bases de datos, sensores, APIs y archivos. La calidad y relevancia de los datos recolectados son esenciales, ya que afectan directamente los resultados del análisis. Por ejemplo, en un estudio de mercado, se pueden recolectar datos de encuestas, redes sociales y ventas anteriores (Marr, 2016).

Almacenamiento y gestión de datos: Una vez recolectados, los datos deben ser almacenados de manera eficiente y segura. Esto implica el uso de sistemas de gestión de bases de datos, almacenamiento en la nube y soluciones de big data, como Hadoop o Spark. La organización y estructuración de los datos son cruciales para facilitar su acceso y análisis posterior (Kelleher & Tierney, 2018).

Análisis de datos: En esta etapa, se aplican diversas técnicas estadísticas y algoritmos de aprendizaje automático para identificar patrones y tendencias en los datos. El análisis puede ser descriptivo, predictivo o prescriptivo, dependiendo de los objetivos del proyecto. Por ejemplo, un análisis predictivo podría utilizar datos históricos para prever futuras tendencias de ventas (Hastie, Tibshirani, & Friedman, 2009).

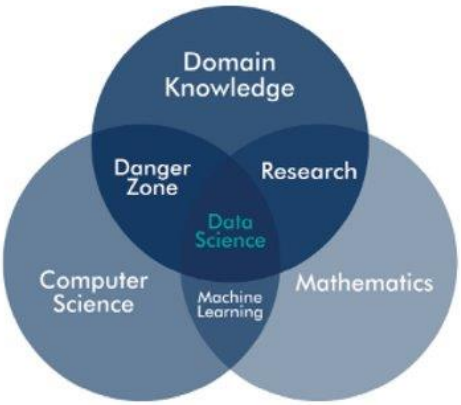
Visualización: La visualización de datos es una parte fundamental de la ciencia de datos, ya que ayuda a comunicar los hallazgos de manera clara y efectiva. Utilizando herramientas como Tableau, Power BI o matplotlib en Python, los científicos de datos pueden crear gráficos e infografías que faciliten la comprensión de los resultados (Few, 2012).

Comunicación de resultados: Finalmente, es vital presentar los resultados de manera comprensible para las partes interesadas. Esto implica la creación de informes y presentaciones que no solo muestren los hallazgos, sino que también expliquen su relevancia y cómo pueden influir en la toma de decisiones estratégicas (Nussbaumer Knaflic, 2015).

La siguiente figura muestra la relación de conocimientos y disciplinas que debe existir para poder desarrollar proyectos de ciencia de datos eficientes.

Figura 3

Relación de conocimientos y disciplinas para desarrollar proyectos de ciencia de datos eficientes



Fuente: Prevos (2019)

La *Danger Zone* es la combinación entre un buen conocimiento del negocio y habilidades computacionales sin un adecuado trasfondo matemático. Esto puede generar una "caja negra" que ofrece una falsa sensación de precisión. Por ejemplo, en el caso de Business Intelligence (What if).

Research se refiere a la combinación entre el conocimiento del negocio y las habilidades matemáticas, pero sin incorporar las habilidades computacionales. Al carecer de la parte computacional, no se podría compartir el código ni garantizar que sea fácilmente reproducible y automatizable.

En el caso de *Machine Learning*, se carece del conocimiento del negocio necesario para dar un uso adecuado al modelo generado. Es fundamental entender la problemática a resolver; de lo contrario, no se podrá implementar el modelo de manera satisfactoria.

La ciencia de datos se trata de personas y problemas del mundo real, no solo de números y algoritmos. Solo un conocimiento adecuado de computación (programación, automatización), habilidades matemáticas (algoritmos predictivos/descriptivos, medidas de rendimiento, etc.) y un profundo entendimiento del negocio (claridad en las variables y problemas del negocio) garantizará el éxito de los proyectos de ciencia de datos.

7.3.1 Análisis de Correlaciones

El análisis de correlaciones es una técnica estadística fundamental que permite examinar la relación entre dos o más variables. Esta metodología se utiliza ampliamente en diversas disciplinas, como la psicología, la economía y la biología, para identificar patrones y

relaciones que pueden influir en la interpretación de los datos. El coeficiente de correlación de Pearson es uno de los métodos más comunes para cuantificar la relación lineal entre dos variables. Su valor varía entre -1 y 1, donde 1 indica una correlación positiva perfecta, -1 una correlación negativa perfecta, y 0 indica ausencia de correlación (Field, 2013).

Un ejemplo práctico del análisis de correlaciones se puede encontrar en el estudio de la relación entre el tiempo de estudio y el rendimiento académico de los estudiantes. Supongamos que se recopilan datos sobre el número de horas que un grupo de estudiantes dedica al estudio y sus calificaciones finales en un examen. Al calcular el coeficiente de correlación de Pearson, se puede determinar si existe una relación significativa entre estas dos variables. Si se obtiene un coeficiente de 0.85, esto indicaría una fuerte correlación positiva, sugiriendo que, a medida que aumenta el tiempo de estudio, también lo hace el rendimiento académico (Cohen et al., 2013).

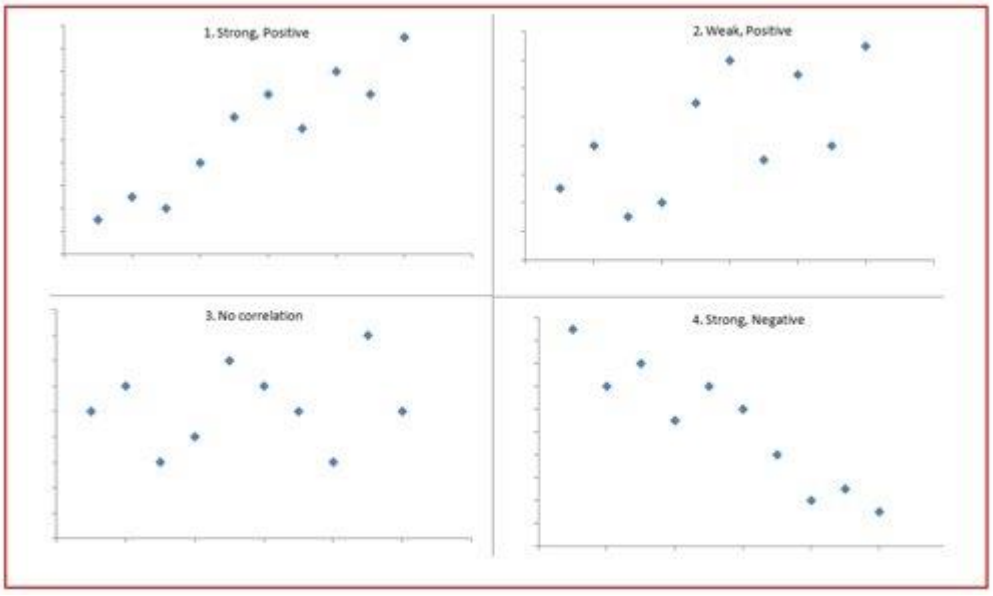
Otro ejemplo relevante se da en el ámbito de la salud, donde se puede investigar la relación entre el consumo de frutas y verduras y el índice de masa corporal (IMC) de los individuos. Si se obtiene un coeficiente de correlación de -0.65, esto sugeriría una correlación negativa moderada; es decir, a medida que aumenta el consumo de frutas y verduras, el IMC tiende a disminuir. Este tipo de análisis podría tener implicaciones significativas para las campañas de salud pública centradas en la nutrición (Hinkle et al., 2003).

Es crucial recordar que la correlación no implica causalidad. Aunque dos variables pueden estar correlacionadas, esto no significa que una cause la otra. Por ejemplo, en el análisis de la relación entre el consumo de helado y el aumento de las temperaturas, se podría encontrar una correlación positiva, pero esto no significa que el consumo de helado cause el aumento de temperatura; en cambio, ambas variables pueden estar influenciadas por una tercera variable, como el clima cálido del verano (Field, 2013).

El siguiente gráfico ilustra de manera visual los tipos y la intensidad de las correlaciones que pueden existir entre las variables.

Figura 4

Tipos e intensidad de correlaciones entre variables

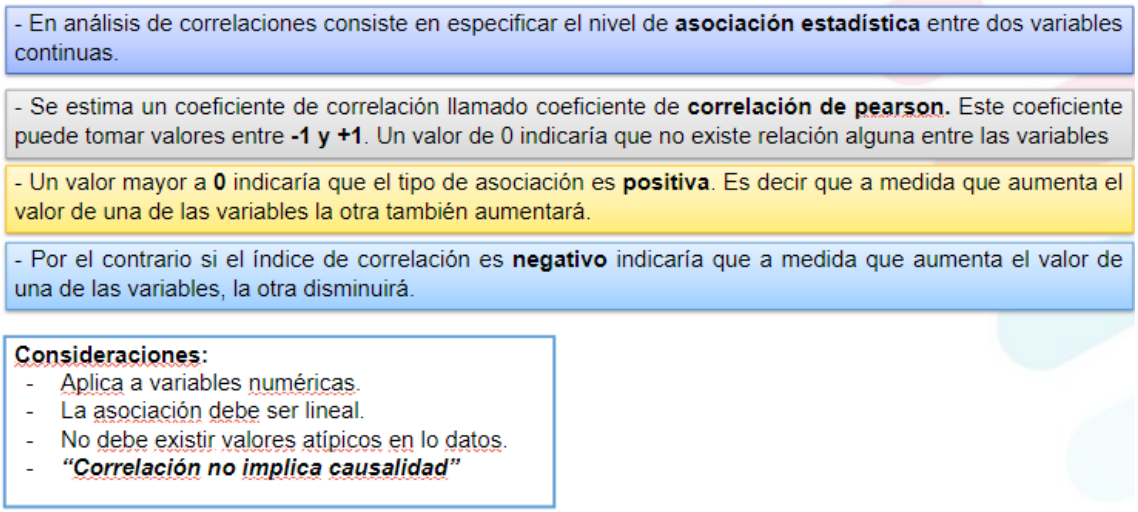


Fuente: Boston University School of Public Health (2013)

La siguiente figura resume algunos aspectos clave del análisis de correlaciones:

Figura 5

Infografía sobre análisis de correlaciones.



Fuente: Elaboración Propia

7.3.2 Principales Métricas de Performance de los Modelos Binarios

La evaluación del rendimiento de los modelos binarios es esencial para garantizar que estos modelos sean efectivos en la predicción de resultados. En el contexto de la clasificación binaria, donde las observaciones se dividen en dos clases (por ejemplo, positivo/negativo, éxito/fracaso), existen varias métricas clave utilizadas para medir la eficacia de estos modelos. Entre las métricas más comunes se encuentran la precisión, el recall (o sensibilidad), la especificidad, el valor predictivo positivo, el valor predictivo negativo y la curva ROC (Receiver Operating Characteristic) con su área bajo la curva (AUC) (Saito & Rehmsmeier, 2015).

Precisión: La precisión se refiere a la proporción de predicciones correctas entre todas las predicciones realizadas. Se calcula como:

$$\text{Precisión} = \frac{\text{Verdaderos Positivos (VP)}}{\text{Verdaderos Positivos (VP)} + \text{Falsos Positivos (FP)}}$$

Por ejemplo, si un modelo clasifica correctamente 70 de 100 casos positivos y 30 de 50 casos negativos, su precisión sería:

$$\text{Precisión} = \frac{70}{70 + 30} = 0.70 \text{ o } 70\%$$

Recall (Sensibilidad): El recall mide la capacidad del modelo para identificar correctamente los casos positivos. Se calcula como:

$$\text{Recall} = \frac{\text{Verdaderos Positivos (VP)}}{\text{Verdaderos Positivos (VP)} + \text{Falsos Negativos (FN)}}$$

Usando el mismo ejemplo, si el modelo tiene 70 verdaderos positivos y 30 falsos negativos, el recall sería:

$$\text{Recall} = \frac{70}{70 + 30} = 0.70 \text{ o } 70\%$$

Especificidad: Esta métrica mide la proporción de verdaderos negativos que se identifican correctamente. Se calcula como:

$$\text{Especificidad} = \frac{\text{Verdaderos Negativos (VN)}}{\text{Verdaderos Negativos (VN)} + \text{Falsos Positivos (FP)}}$$

Siguiendo con el ejemplo anterior, si el modelo clasifica correctamente 30 de 50 negativos, la especificidad sería:

$$\text{Especificidad} = \frac{30}{30 + 30} = 0.50 \text{ o } 50\%$$

Valor Predictivo Positivo (VPP) y Valor Predictivo Negativo (VPN): Estas métricas evalúan la precisión de las predicciones positivas y negativas, respectivamente. Se calculan de la siguiente manera:

$$\text{VPP} = \frac{VP}{VP + FP}, \quad \text{VPN} = \frac{VN}{VN + FN}$$

Por ejemplo, con 70 verdaderos positivos y 30 falsos positivos, el VPP sería:

$$\text{VPP} = \frac{70}{70 + 30} = 0.70 \text{ o } 70\%$$

7.3.2.1 Curva ROC y AUC

La curva ROC traza la tasa de verdaderos positivos frente a la tasa de falsos positivos a diferentes umbrales de clasificación. El área bajo la curva (AUC) mide la capacidad del modelo para distinguir entre las clases. Un AUC de 0.5 indica un modelo sin poder predictivo, mientras que un AUC de 1.0 indica un modelo perfecto (Fawcett, 2006).

7.3.3 Algoritmos de Aprendizaje Supervisado

Los algoritmos de aprendizaje supervisado son métodos de **machine learning** en los que el modelo se entrena utilizando un conjunto de datos etiquetados, es decir, ejemplos que contienen tanto entradas como salidas conocidas. Durante el entrenamiento, el modelo aprende la relación entre las variables de entrada y las etiquetas de salida, lo que le permite hacer **predicciones sobre datos nuevos** (Goodfellow et al., 2016). Este tipo de aprendizaje se utiliza en dos tipos principales de problemas: **clasificación** (donde el modelo asigna datos a categorías específicas) y **regresión** (donde el modelo predice valores continuos) (Murphy, 2012).

Algunos algoritmos representativos de aprendizaje supervisado incluyen el árbol de decisión, la **regresión logística** y las **máquinas de soporte vectorial** (SVM), los cuales

han demostrado ser efectivos en aplicaciones como el reconocimiento de imágenes, procesamiento de lenguaje natural y detección de fraude (Hastie, Tibshirani, & Friedman, 2009).

7.3.3.1 Árbol de Decisión

Un árbol de decisión es una herramienta de modelado predictivo utilizada tanto en clasificación como en regresión. Esta técnica visualiza decisiones y sus posibles consecuencias en forma de un árbol, donde cada nodo representa una pregunta sobre una variable, y cada rama representa el resultado de esa pregunta. El objetivo es dividir los datos en grupos homogéneos, facilitando la toma de decisiones basada en datos (Breiman et al., 1986).

Estructura del árbol

La estructura de un árbol de decisión consta de tres componentes principales:

- **Nodos de decisión**: Representan preguntas o condiciones basadas en las características de los datos. Por ejemplo, en un modelo que predice si un cliente comprará un producto, un nodo de decisión podría ser "¿El cliente tiene un ingreso superior a $50,000?".
- **Ramas**: Indican el resultado de las decisiones tomadas en los nodos. Por ejemplo, si la respuesta a la pregunta sobre el ingreso es "Sí", se sigue por una rama; si es "No", se sigue por otra.
- **Nodos terminales**: Son los resultados finales del árbol, que indican la clase o el valor predicho. En el ejemplo anterior, un nodo terminal podría indicar "Compra" o "No compra".

Ejemplo Práctico

Imaginemos que queremos predecir si un cliente comprará un coche basado en dos características: su ingreso y su edad. Un posible árbol de decisión podría verse así:

Nodo 1: ¿Ingreso > $50,000?

Sí:

Nodo 2: ¿Edad > 30 años?

Sí: Nodo terminal "Compra"

No: Nodo terminal "No compra"

No:

Nodo terminal "No compra"

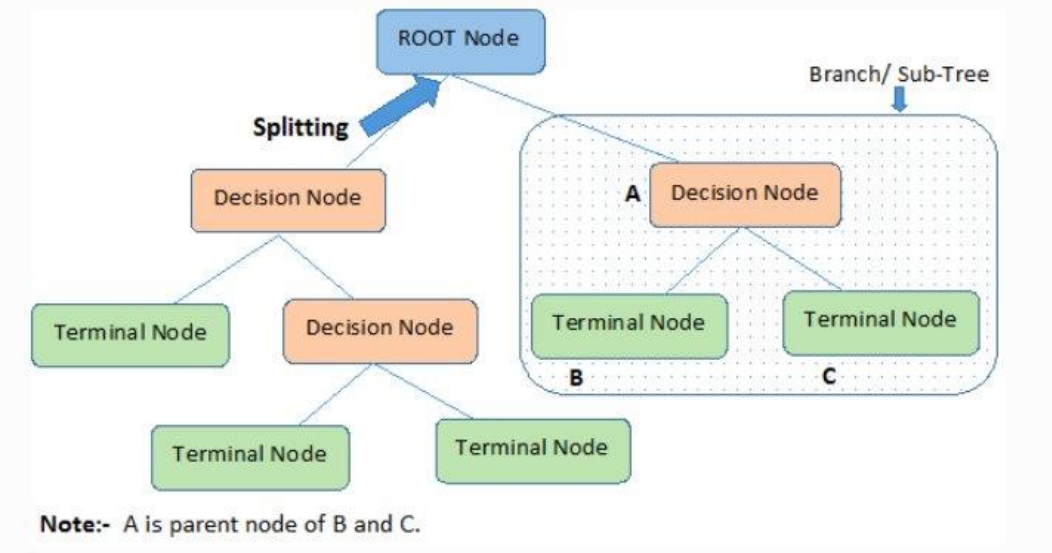
En este árbol, comenzamos con la pregunta sobre el ingreso. Si el ingreso es mayor a $50,000, pasamos a la siguiente pregunta sobre la edad. Este tipo de análisis ayuda a segmentar a los clientes y predecir comportamientos de manera clara y comprensible (Quinlan, 1986).

Ventajas y desventajas

Los árboles de decisión son fáciles de interpretar y pueden manejar tanto variables categóricas como numéricas. Sin embargo, también presentan desventajas, como la tendencia a sobreajustar los datos y ser sensibles a pequeñas variaciones en los datos de entrenamiento (Loh, 2011). Por esta razón, a menudo se combinan con técnicas como la *poda* (pruning) para mejorar su capacidad de generalización.

Figura 6

Árbol de Decisión



Fuente: Elaboración Propia

El siguiente código muestra cómo se puede implementar un árbol de decisión en Python utilizando un conjunto de datos de ejemplo de la librería scikit-learn.

Figura 7

Código para implementar un árbol de decisión

```python
import numpy as np
import pandas as pd
from sklearn.datasets import load_iris
from sklearn.model_selection import train_test_split
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import classification_report, confusion_matrix, accuracy_score

# Cargar el dataset de Iris
iris = load_iris()
X = iris.data
y = iris.target

# Dividir el dataset en conjunto de entrenamiento y prueba
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

```python
# Crear el modelo de árbol de decisión
clf = DecisionTreeClassifier(random_state=42)
clf.fit(X_train, y_train)

# Hacer predicciones en el conjunto de prueba
y_pred = clf.predict(X_test)

# Evaluar el rendimiento del modelo
print("Matriz de confusión:")
print(confusion_matrix(y_test, y_pred))

print("\nReporte de clasificación:")
print(classification_report(y_test, y_pred))

accuracy = accuracy_score(y_test, y_pred)
print(f"\nPrecisión del modelo: {accuracy:.2f}")
```

Fuente: Elaboración Propia

Referencias

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1986). *Classification and regression trees*. Wadsworth.

Boston University School of Public Health. (2013). *Multivariate Analysis*. https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_multivariable/bs704_multivariable5.html

Chapman, P. (2000). *Step-by-step data mining guide, 10*. https://www.the-modeling-agency.com/crisp-dm.pdf

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., & Shearer, C. (2000). *CRISP-DM 1.0: Step-by-step data mining guide*. https://www.crisp-dm.org

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*(8), 861-874. https://doi.org/10.1016/j.patrec.2005.10.010

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine, 17*(3), 37-54.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.

Loh, W. Y. (2011). Fifty years of classification and regression trees. *International Statistical Review, 79*(3), 329-348. https://doi.org/10.1111/j.1751-5823.2011.00105.x

Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. MIT Press.

Prevos, Peter. (2019). *Principles of Strategic Data Science*. https://subscription.packtpub.com/book/data/9781838985295/1/ch01lvl1sec04/the-elements-of-data-science

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning, 1*(1), 81-106. https://doi.org/10.1007/BF00116251

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot. *Nature Methods, 12*(3), 209-210. https://doi.org/10.1038/nmeth.4487

Shearer, C. (2000). The CRISP-DM model: A comprehensive process model for data mining. *Journal of Data Warehousing, 5*(4), 13-22.

PUCE Virtual
EDUCACIÓN FLEXIBLE

La excelencia no se improvisa

síguenos

MDASC - 03 - 251