INTELIGENCIA DE NEGOCIOS

Tema de la sesión: CRISP DM - Fase 4 Modelado - I

Carrera: Diseño y Desarrollo de Software

Capacidades Terminales

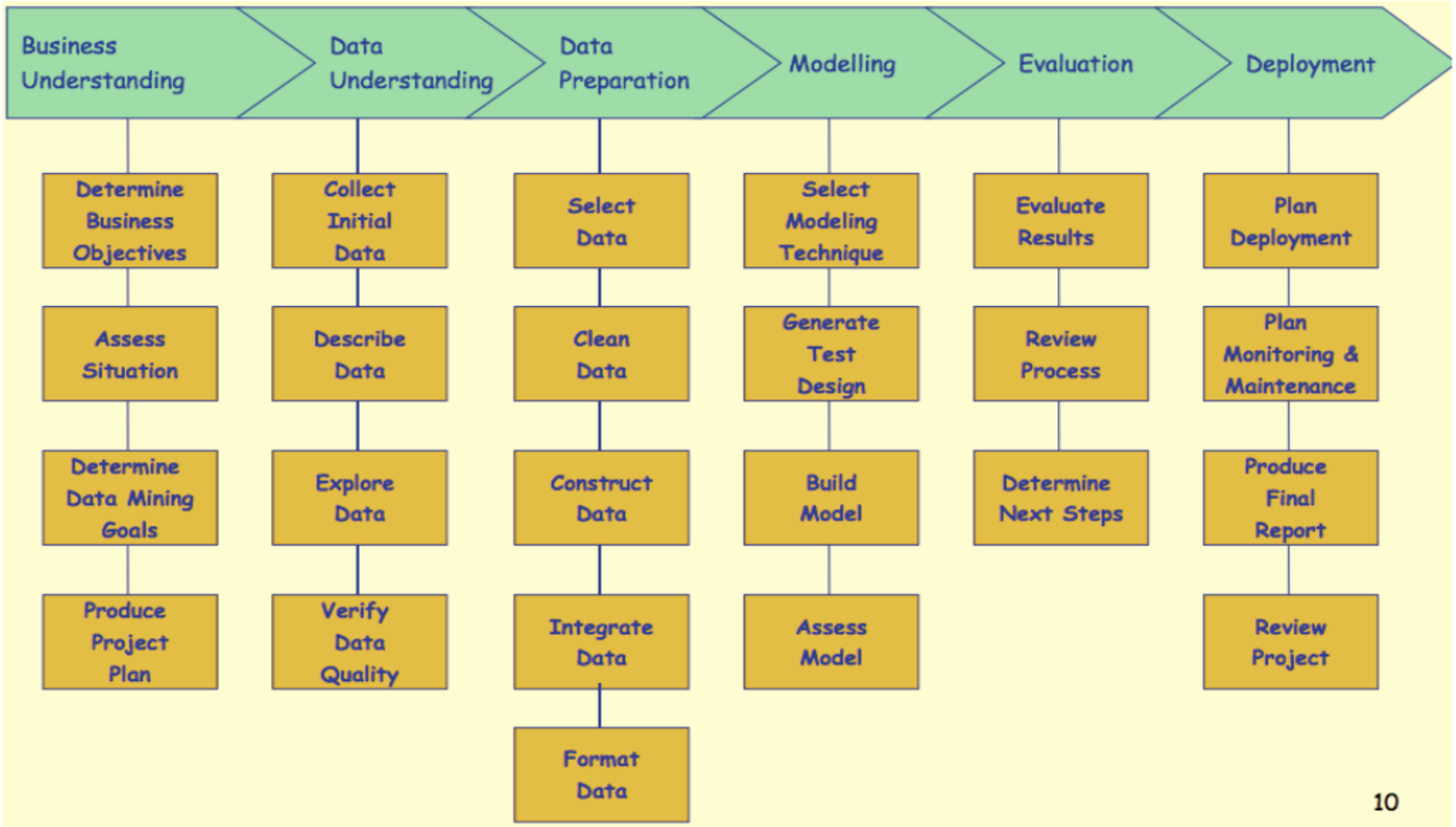
- Conocer la Metodología Crisp DM Fase 4 y su importancia
- Diferencias técnicas de modelado como regresión lineal, árboles de decision y redes neuronales
- Indentificar las tareas de un plan de pruebas

Contenido

- Tareas y fases del modelo CRISP-DM
- Fase 4: Modelado
- Seleccionar técnica de modelado
- Generar plan de pruebas
- Conclusiones
- Referencias Bibliográficas

Motivación

Tareas y fases del modelo CRISP-DM

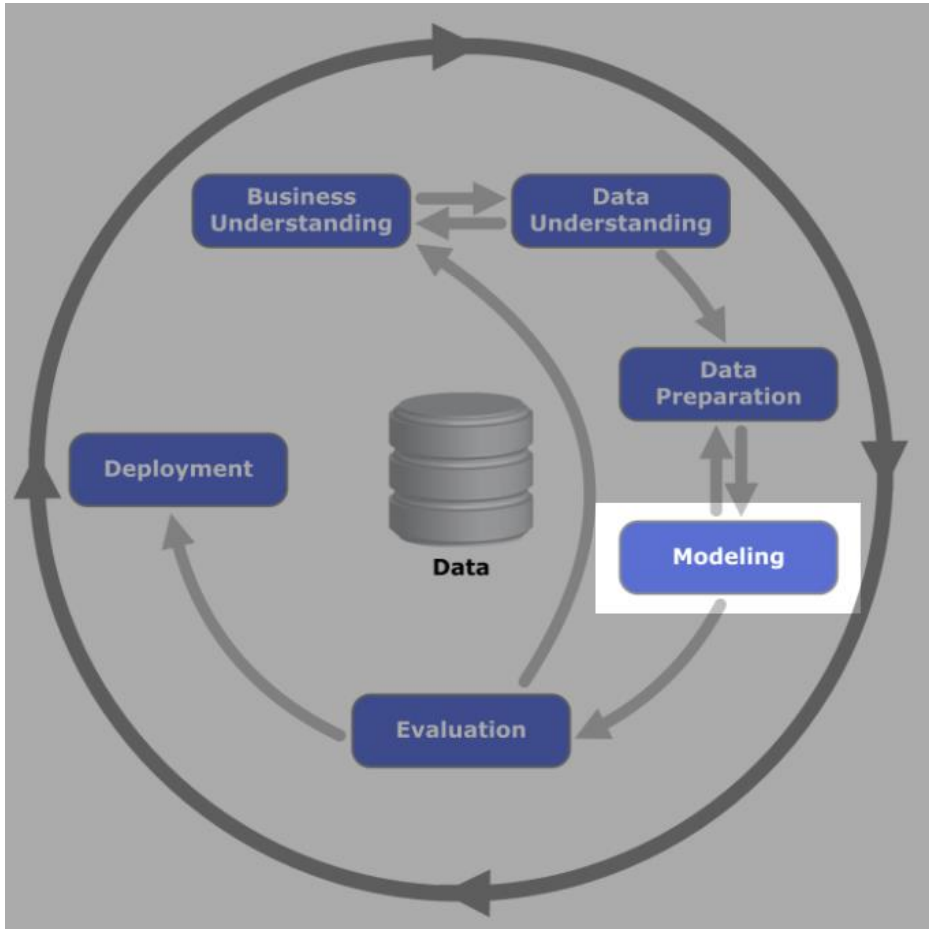


Fase 4: Modelado

El modelado a menudo se considera el trabajo más interesante de la ciencia de datos. En esta fase, el equipo construye y evalúa varios modelos, a menudo utilizando varias técnicas de modelado diferentes.

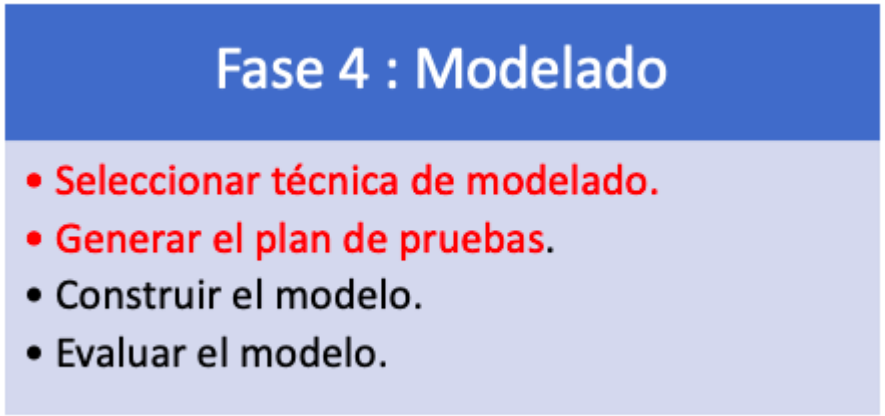
Aunque la guía CRISP-DM sugiere "iterar la construcción y evaluación de modelos hasta que crea firmemente que ha encontrado los mejores modelos", en la práctica, los equipos pueden iterar hasta que tengan un modelo "suficientemente bueno".

Esta fase tiene cuatro tareas (en esta sesión revisaremos las dos primeras).

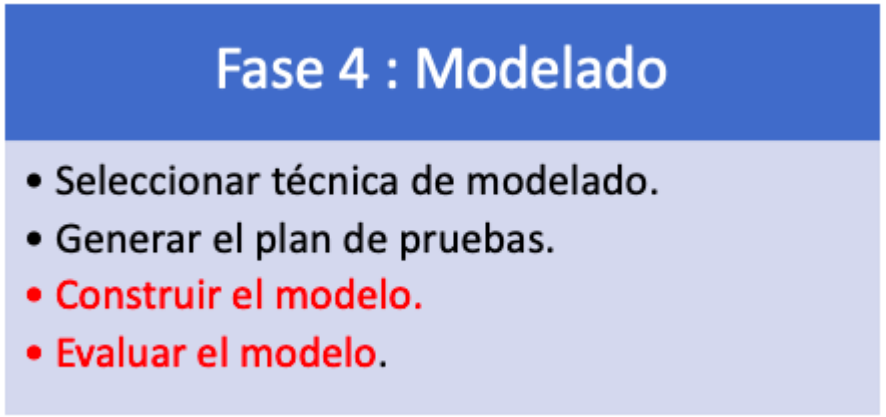


Fase 4: Modelado

1. **Seleccionar técnica de modelado:** Se determina que algoritmo usar, por ejemplo regresión lineal o redes neuronales.
2. **Generar el plan de pruebas:** Dependiendo del enfoque del modelado, es posible que se deba dividir los datos en conjuntos de entrenamiento, prueba y validación.



3. **Construir el modelo**: Es la construcción del modelo usando alguna herramienta o lenguaje de programación con librerías especializadas.
4. **Revisar el modelo:** Por lo general, varios modelos compiten entre sí y el científico de datos debe interpretar los resultados del modelo en función del conocimiento del dominio, los criterios de éxito predefinidos y el diseño de la prueba.



Selección de técnicas de modelado

- Se determina que técnica de modelado o algoritmo se utilizará, por ejemplo, regresión lineal, árboles de decisión, redes neuronales etc. Esto se realiza en función al objetivo del análisis y del tipo de datos disponibles.
- También se prueban varios algoritmos para comparar su rendimiento.

Regresión lineal simple

- La regresión lineal es una técnica estadística que se utiliza para modelar la relación entre una o más variables independientes (predictoras) y una variable dependiente (respuesta), asumiendo que esta relación es lineal.

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Regresión Lineal Múltiple

La regresión lineal múltiple es un método estadístico utilizado para predecir el valor de una variable dependiente en función de dos o más variables independientes. Es una extensión de la regresión lineal simple, donde se tiene una variable independiente.

En la regresión lineal múltiple, el objetivo es encontrar la mejor combinación lineal de las variables independientes para predecir la variable dependiente. Esto se logra ajustando los coeficientes de la ecuación de regresión, que es de la forma:

$$Y=\beta 0+\beta 1X1+\beta 2X2+\ldots+\beta nXn+\varepsilon$$

Donde:

- Y es la variable dependiente que se está prediciendo.
- $X1, X2, \ldots, Xn$ son las variables independientes.
- $\beta 0, \beta 1, \beta 2, \ldots, \beta n$ son los coeficientes que representan el efecto de cada variable independiente en la variable dependiente.
- ε es el término de error, que representa la variabilidad no explicada por el modelo.

Los coeficientes de la regresión lineal múltiple se pueden estimar utilizando diversos métodos, como el método de mínimos cuadrados ordinarios (MCO), que busca minimizar la suma de los cuadrados de las diferencias entre los valores observados y los valores predichos.

Supuestos

Linealidad: La relación entre las variables independientes y la variable dependiente debe ser aproximadamente lineal. Esto significa que los cambios en las variables independientes están asociados con cambios proporcionales en la variable dependiente.

Normalidad: Los errores de predicción deben seguir una distribución normal. Idealmente, los residuos (diferencias entre los valores observados y los valores predichos) deberían tener una distribución normal alrededor de cero.

Homocedasticidad: La varianza de los errores debe ser constante en todos los niveles de las variables independientes. En otras palabras, la dispersión de los errores no debe cambiar a medida que cambian los valores de las variables independientes.

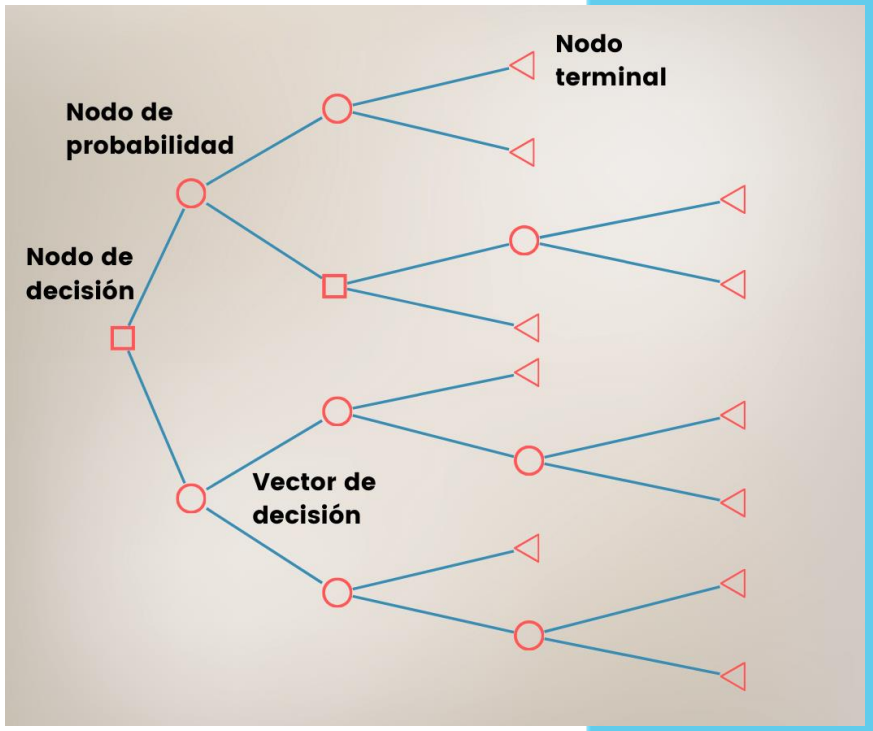
Supuestos

Independencia de los errores: Los errores de predicción deben ser independientes entre sí. Esto significa que el error asociado con una observación no debe estar correlacionado con el error asociado con otra observación.

Ausencia de multicolinealidad: Las variables independientes no deben estar altamente correlacionadas entre sí. La multicolinealidad puede dificultar la interpretación de los coeficientes de regresión y puede afectar la precisión de las predicciones.

Árboles de Decisión

- Un árbol de decisión es un modelo de tipo jerárquico que divide un conjunto de datos en subconjuntos más pequeños a partir de preguntas lógicas sobre los atributos de entrada. Su estructura es similar a un árbol:
  - Raíz (root): es la primera decisión.
  - Nodos internos: representan condiciones (por ejemplo, "¿edad > 30?")
  - Hojas: contienen las predicciones finales (clases o valores).



Random Forest

- Random Forest (Bosque Aleatorio) es un algoritmo de aprendizaje supervisado que combina muchos árboles de decisión individuales para obtener un modelo más robusto, preciso y generalizable.
- Se puede usar para:
  - **Clasificación** (predecir categorías)
  - **Regresión** (predecir valores continuos)

Random Forest

Funcionamiento:

- Bootstrapping: Se crean múltiples subconjuntos aleatorios del conjunto de datos original (con reemplazo).
- Entrenamiento: Para cada subconjunto se entrena un árbol de decisión diferente.
- Selección aleatoria de variables: En cada división del árbol, se selecciona un subconjunto aleatorio de variables para evaluar la mejor separación (esto introduce diversidad entre árboles).
- Agregación del resultado:
  - En clasificación, se toma el voto mayoritario entre todos los árboles.
  - En regresión, se calcula el promedio de las predicciones de todos los árboles.

Random Forest

- Reduce el sobreajuste promediando múltiples árboles independientes.
- Mejora la precisión y estabilidad del modelo.

Ventajas:

- Alta precisión y robustez
- Reduce el sobreajuste
- Funciona con datos numéricos y categóricos
- Maneja bien valores faltantes o ruido
- Permite medir la importancia de las variables

Desventajas:

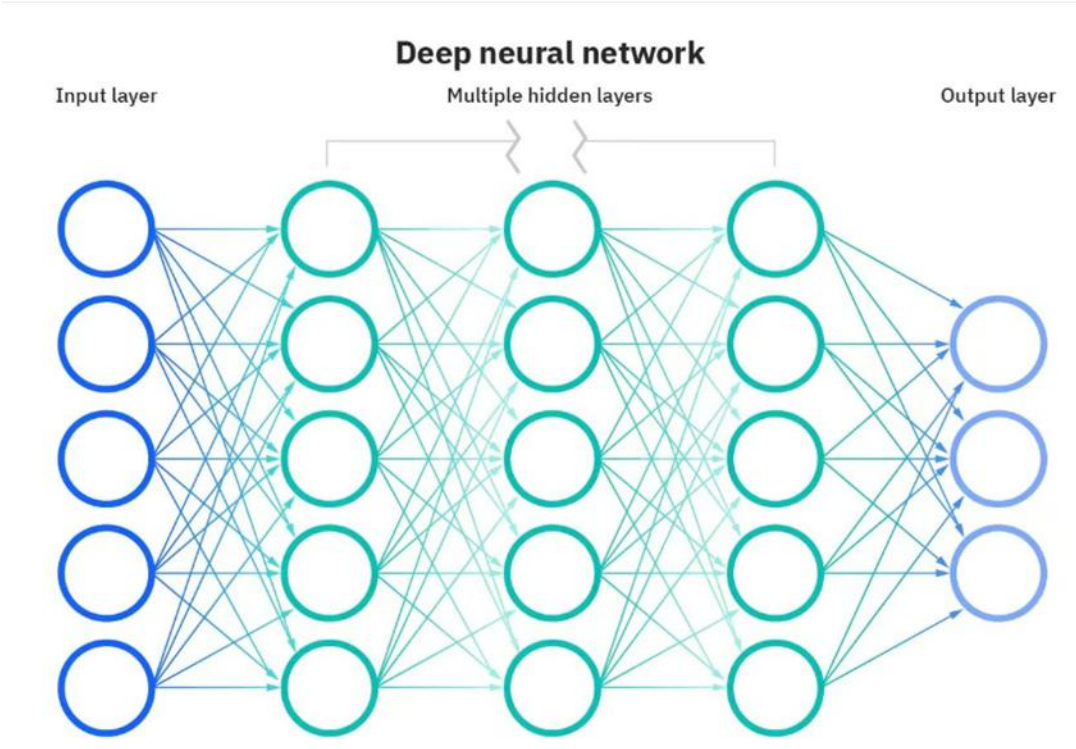
- Difícil de interpretar (no tan intuitivo como un solo árbol)
- Más costoso computacionalmente
- Puede volverse lento con muchos árboles o datos muy grandes

Redes Neuronales

- Una red neuronal artificial (RNA) es un modelo de machine learning inspirado en el funcionamiento del cerebro humano. Se compone de unidades básicas llamadas neuronas artificiales, organizadas en capas que procesan datos y aprenden patrones complejos a través de un proceso iterativo.
- Se basan en datos de entrenamiento con los que aprenden y mejoran su precisión.

Redes Neuronales

- Una red neuronal típica tiene tres tipos de capas:
  - Capa de entrada: recibe los datos de entrada.
  - Capas ocultas (1 o más): procesan la información aprendiendo representaciones abstractas.
  - Capa de salida: entrega el resultado (una clase o un valor continuo).



Redes Neuronales

- Cada neurona realiza una operación matemática:

$$z = w_1 x_1 + w_2 x_2 + \ldots + w_n x_n + b$$

- Luego se aplica una **función de activación** (como ReLU o sigmoide) para introducir no linealidades:

$$a = f(z)$$

- Entrenamiento con retropropagación
- La red ajusta los pesos w y sesgos b minimizando un error (por ejemplo, con función de pérdida MSE o cross-entropy) usando un algoritmo como gradiente descendente.

Redes Neuronales

Recomendadas cuando:

- Cuando los datos tienen patrones no lineales complejos.
- Cuando hay gran cantidad de datos
- Tareas como:
  - Clasificación de imágenes.
  - Reconocimiento de voz o texto.
  - Predicción de series temporales.
  - Procesamiento de lenguaje natural.

Redes Neuronales

Ventajas

- Modelan relaciones no lineales complejas
- Funciona bien con grandes volúmenes de datos
- Se adapta a muchos tipos de datos (imagen, texto, etc.)
- Se puede mejorar con arquitecturas avanzadas

- Desventajas
- Necesitan más datos que modelos simples
- Mayor costo computacional (CPU/GPU)
- Difíciles de interpretar ("caja negra")
- Mayor riesgo de sobreajuste si no se regula bien

Generar el plan de pruebas

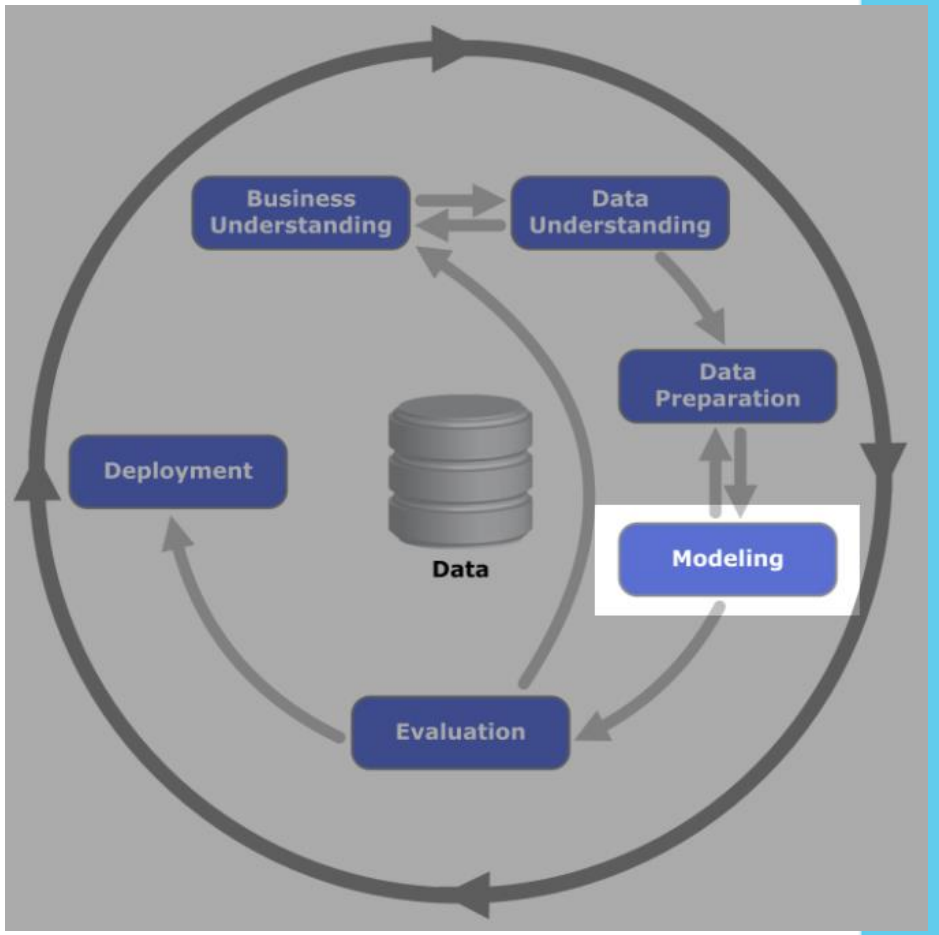
- Es el proceso mediante el cual se define cómo se va a evaluar el rendimiento y la validez de los modelos desarrollados. No se refiere solo a probar el código, sino a probar la calidad de los modelos predictivos y asegurarse de que sus resultados sean confiables antes de ponerlos en producción.
- Se realiza principalmente antes y durante la Fase 4 (Modelado) y se complementa en la Fase 5 (Evaluación) de CRISP-DM.

Generar el plan de pruebas

- Definición del objetivo del modelo
  - ¿Qué se busca hacer, clasificar, predecir, segmentar?
  - ¿Qué métricas utilizar?
- Estrategia de validación
  - División del dataset (entrenamiento / prueba / validación)
  - Validación cruzada (k-fold)
- Métricas de evaluación
  - Clasificación: precisión, recall, F1-score, matriz de confusión, ROC-AUC.
  - Regresión: RMSE, MAE, R^2.
- Herramientas y procedimientos (Librerias como sklearn, keras, etc)

Conclusiones

- Es la cuarta fase de CRISP-DM: Modelado.
- En esta primera parte de la fase del modelado se han vista las actividades de la elección del algoritmo y la organización de los datos para el plan de pruebas.



Referencias Bibliográficas

- Espinosa-Zúñiga, J. J. (2020). Aplicación de metodología CRISP-DM para segmentación geográfica de una base de datos pública. Ingeniería, investigación y tecnología, 21(1).

- Fernández, D. B., & Mora, S. L. (2016). Uso de la metodología CRISP-DM para guiar el proceso de minería de datos en LMS. In Tecnologia, innovación e investigación en los procesos de enseñanza-aprendizaje (pp. 2385-2393). Octaedro.

- Sifuentes, M. S. G. C., Pérez, L. G. V., Cantabrana, M. G. N., Acosta, I. F. O., Santana, F. A. Á., & Fierro, M. D. L. Á. S. (2023). Modelo Predictivo de la Deserción Escolar en Educación Superior: una Aproximación desde la Minería de Datos Utilizando la Metodología CRISP-DM. Ciencia Latina Revista Científica Multidisciplinar, 7(5), 7797-7812.

Tecsup®

TECNOLOGÍA
CON SENTIDO