Proceso de Minería de datos

Wilmer Gonzalez

6213 - Minería de datos
Facultad de ciencias
Universidad Central de Venezuela

25 de abril de 2024

Contenido

"Notoriamente no hay clasificación del universo que no sea arbitraria y conjetural."[a]

[a] El idioma analítico de John Wilkins - Jorge Luis Borges

KDD

Knowledge Discovery from Data

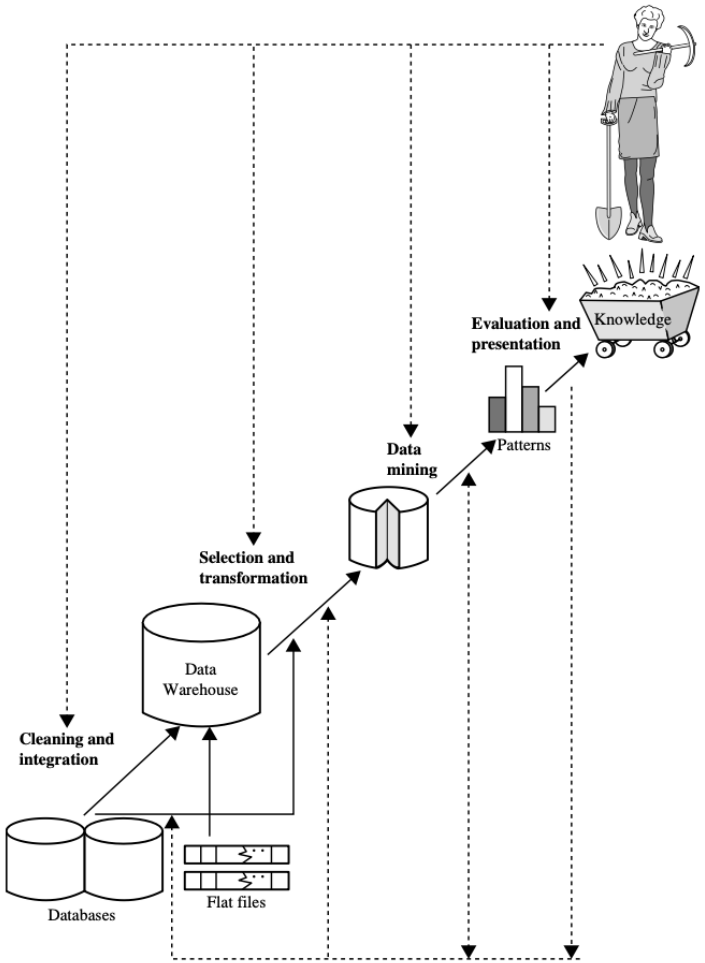

Figura: Modelo de proceso en KDD[1]

SEMMA

Sample, Explore, Modify, Model, and Assess

Muestreo Selección de datos, lo suficientemente grande para contener conocimiento, pero lo suficientemente pequeños para poder ser usados eficientemente.

Exploración Entendimiento de los datos para detectar relaciones entre observaciones y variables, así como valores anómalos, típicamente con la ayuda de visualización de datos.

Modificación Selección, creación y transformación de atributos como preparación al modelado.

Modelado Se aplican distintas técnicas de modelos en los atributos seleccionados con el fin de capturar el comportamiento deseado.

Evaluación Se evalúan modelos con respecto a consistencia y utilidad de acuerdo al problema.

Sample, Explore, Modify, Model, and Assess

Muestreo Selección de datos, lo suficientemente grande para contener conocimiento, pero lo suficientemente pequeños para poder ser usados eficientemente.

Exploración Entendimiento de los datos para detectar relaciones entre observaciones y variables, así como valores anómalos, típicamente con la ayuda de visualización de datos.

Modificación Selección, creación y transformación de atributos como preparación al modelado.

Modelado Se aplican distintas técnicas de modelos en los atributos seleccionados con el fin de capturar el comportamiento deseado.

Evaluación Se evalúan modelos con respecto a consistencia y utilidad de acuerdo al problema.

Sample, Explore, Modify, Model, and Assess

Muestreo Selección de datos, lo suficientemente grande para contener conocimiento, pero lo suficientemente pequeños para poder ser usados eficientemente.

Exploración Entendimiento de los datos para detectar relaciones entre observaciones y variables, así como valores anómalos, típicamente con la ayuda de visualización de datos.

Modificación Selección, creación y transformación de atributos como preparación al modelado.

Modelado Se aplican distintas técnicas de modelos en los atributos seleccionados con el fin de capturar el comportamiento deseado.

Evaluación Se evalúan modelos con respecto a consistencia y utilidad de acuerdo al problema.

Sample, Explore, Modify, Model, and Assess

Muestreo Selección de datos, lo suficientemente grande para contener conocimiento, pero lo suficientemente pequeños para poder ser usados eficientemente.

Exploración Entendimiento de los datos para detectar relaciones entre observaciones y variables, así como valores anómalos, típicamente con la ayuda de visualización de datos.

Modificación Selección, creación y transformación de atributos como preparación al modelado.

Modelado Se aplican distintas técnicas de modelos en los atributos seleccionados con el fin de capturar el comportamiento deseado.

Evaluación Se evalúan modelos con respecto a consistencia y utilidad de acuerdo al problema.

Sample, Explore, Modify, Model, and Assess

Muestreo Selección de datos, lo suficientemente grande para contener conocimiento, pero lo suficientemente pequeños para poder ser usados eficientemente.

Exploración Entendimiento de los datos para detectar relaciones entre observaciones y variables, así como valores anómalos, típicamente con la ayuda de visualización de datos.

Modificación Selección, creación y transformación de atributos como preparación al modelado.

Modelado Se aplican distintas técnicas de modelos en los atributos seleccionados con el fin de capturar el comportamiento deseado.

Evaluación Se evalúan modelos con respecto a consistencia y utilidad de acuerdo al problema.

CRISP-DM

CRoss Industry Standard Process for Data Mining

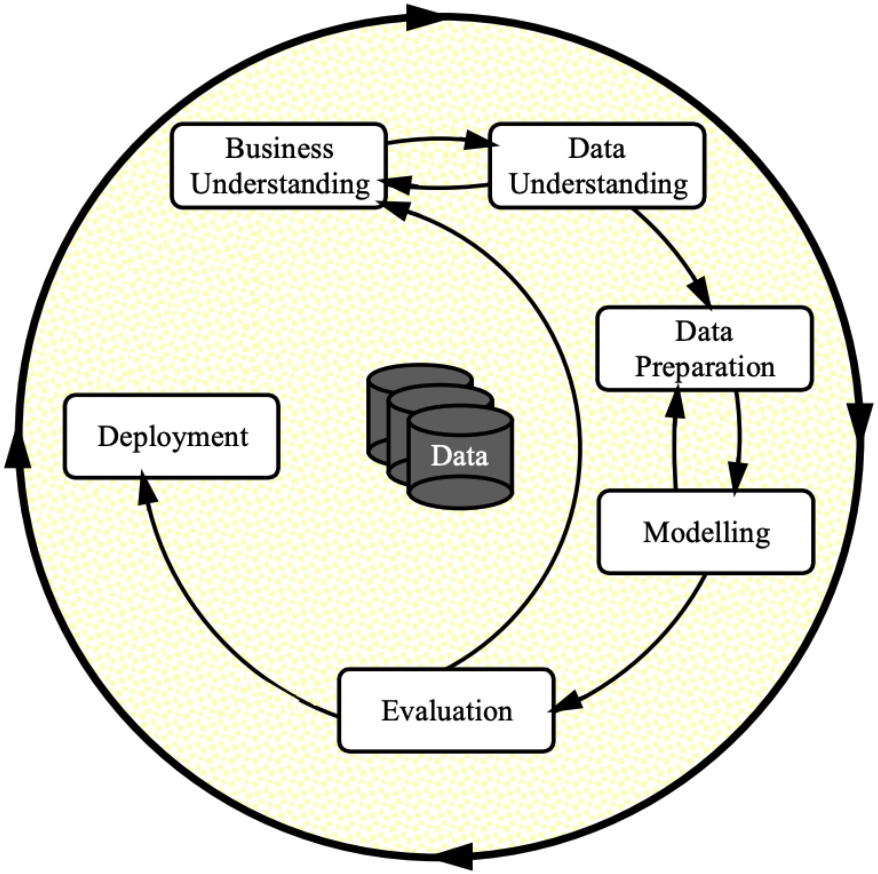

Figura: Modelo de proceso en CRISP-DM[5]

Etapas del proceso CRISP-DM

- *Entendimiento del negocio*
- *Entendimiento de los datos*
- *Preparación de datos*
- *Modelado*
- *Evaluación del modelo*
- *Despliegue*

Objetivos

- Establecimiento de objetivos del negocio
- Establecimiento de criterios de éxito
- Definir:
  - Requerimientos
  - Asunciones
  - Restricciones
  - Riesgos
  - Costo/Beneficio

Etapas del proceso CRISP-DM

- *Entendimiento del negocio*
- *Entendimiento de los datos*
- *Preparación de datos*
- *Modelado*
- *Evaluación del modelo*
- *Despliegue*

Objetivos

- Reporte de colección de datos
- Reporte de descripción de datos
- Reporte de calidad de los datos

Etapas del proceso CRISP-DM

- *Entendimiento del negocio*
- *Entendimiento de los datos*
- *Preparación de datos*
- *Modelado*
- *Evaluación del modelo*
- *Despliegue*

Objetivos

- Selección de datos
- Formato de los datos
- Limpieza de datos:
  - Valores ausentes
  - Valores erróneos
  - Valores anómalos
  - Valores duplicados

Etapas del proceso CRISP-DM

- *Entendimiento del negocio*
- *Entendimiento de los datos*
- *Preparación de datos*
- *Modelado*
- *Evaluación del modelo*
- *Despliegue*

Objetivos

- Selección de técnicas de modelado
- Especificación de asunciones
- Diseño de pruebas
- Configuración de *hiperparámetros*
- Evaluación del modelo
- Optimización de *hiperparámetros*

Etapas del proceso CRISP-DM

- *Entendimiento del negocio*
- *Entendimiento de los datos*
- *Preparación de datos*
- *Modelado*
- *Evaluación del modelo*
- *Despliegue*

Objetivos

- Evaluación de resultados con respecto a los objetivos del negocio
- Evaluación de posibles decisiones

Etapas del proceso CRISP-DM

- *Entendimiento del negocio*
- *Entendimiento de los datos*
- *Preparación de datos*
- *Modelado*
- *Evaluación del modelo*
- *Despliegue*

Objetivos

- Plan de despliegue
- Plan de monitoreo y mantenimiento
- Reporte final
- Documentación

Análisis comparativo

Análisis comparativo

KDD	SEMMA	CRISP-DM
Pre-KDD	—	Entendimiento del negocio
Selección	Muestreo	Entendimiento de datos
Pre-procesamiento	Exploración	Entendimiento de datos
Transformación	Modificación	Preparación de datos
Minería de datoa	Modelado	Modelado
Interpretación/Evaluación	Evaluación	Evaluación
Post-KDD	—	Despliegue

Estado del arte

Estado del arte

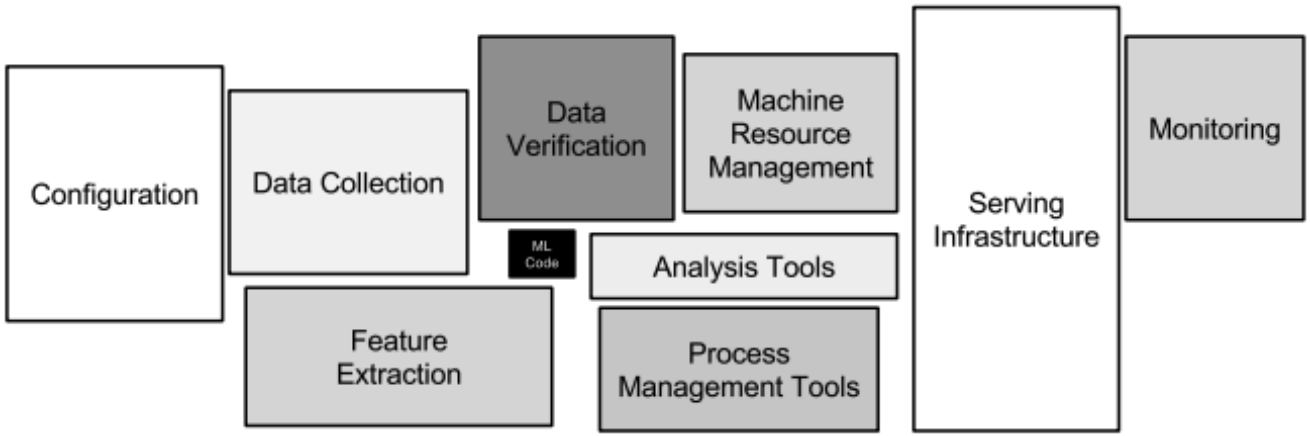

Figura: ML code in ML systems[4][2]

Estado del arte

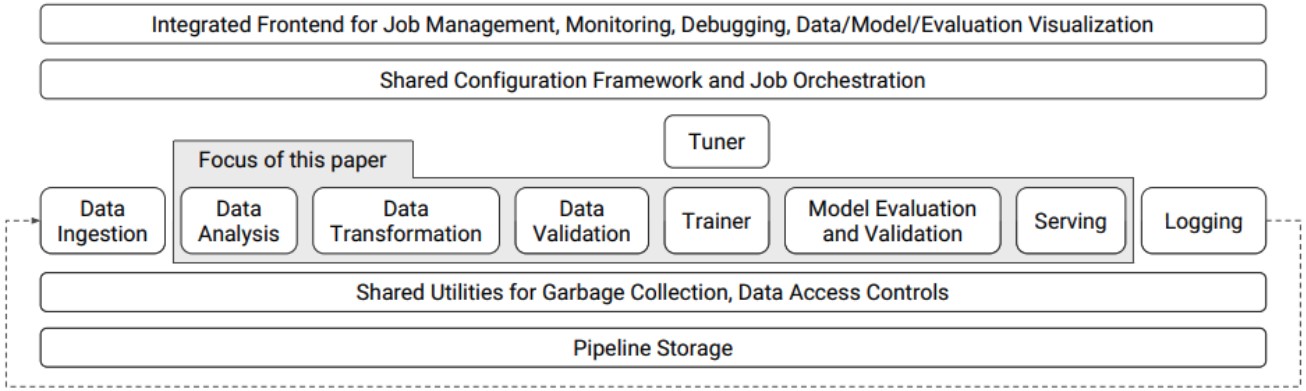

Figura: TFX[3]

Algunas herramientas

- **Fuentes de Datos:** *Pentaho Data Integration*, *Segment*, *AWS Athena*, *GCP BigQuery*
- **Exploración de datos:** *Apache Zeppelin*, *AWS Sagemaker*, *Google Colab Deepnote*, *SQL*, *HEX*
- **Preprocesamiento de datos:** *scikit-learn*, *dataprep*, *pandas*, *PySpark*, *Scala*, *dvc.org*
- **Modelado:** *XGBoost*, *Keras*, *Tensorflow*, *Prophet*, *Orange*, *mlxtend*
- **Evaluación:** *evidentlyai*, *MLFlow*, *neptune.ai*
- **Despliegue:** *(HF) Inference endpoints*, *Modal*, *Groq*, *AWS Lambda FastAPI*...

Referencias

[1] Jiawei Han, Micheline Kamber y Jian Pei. *Data mining concepts and techniques, third edition*. 2012.

[2] Jimmy Lin y Dmitriy Ryaboy. "Scaling big data mining infrastructure: the twitter experience". En: (2013).

[3] Akshay Naresh Modi et al. "TFX: A TensorFlow-Based Production-Scale Machine Learning Platform". En: 2017.

[4] D. Sculley et al. "Hidden Technical Debt in Machine Learning Systems". En: 2015.

[5] Rüdiger Wirth y Jochen Hipp. "CRISP-DM: Towards a Standard Process Model for Data Mining". En: 2000.

¡Gracias!

github.com/ucvia/dm