

02 – Entendimiento de los Datos



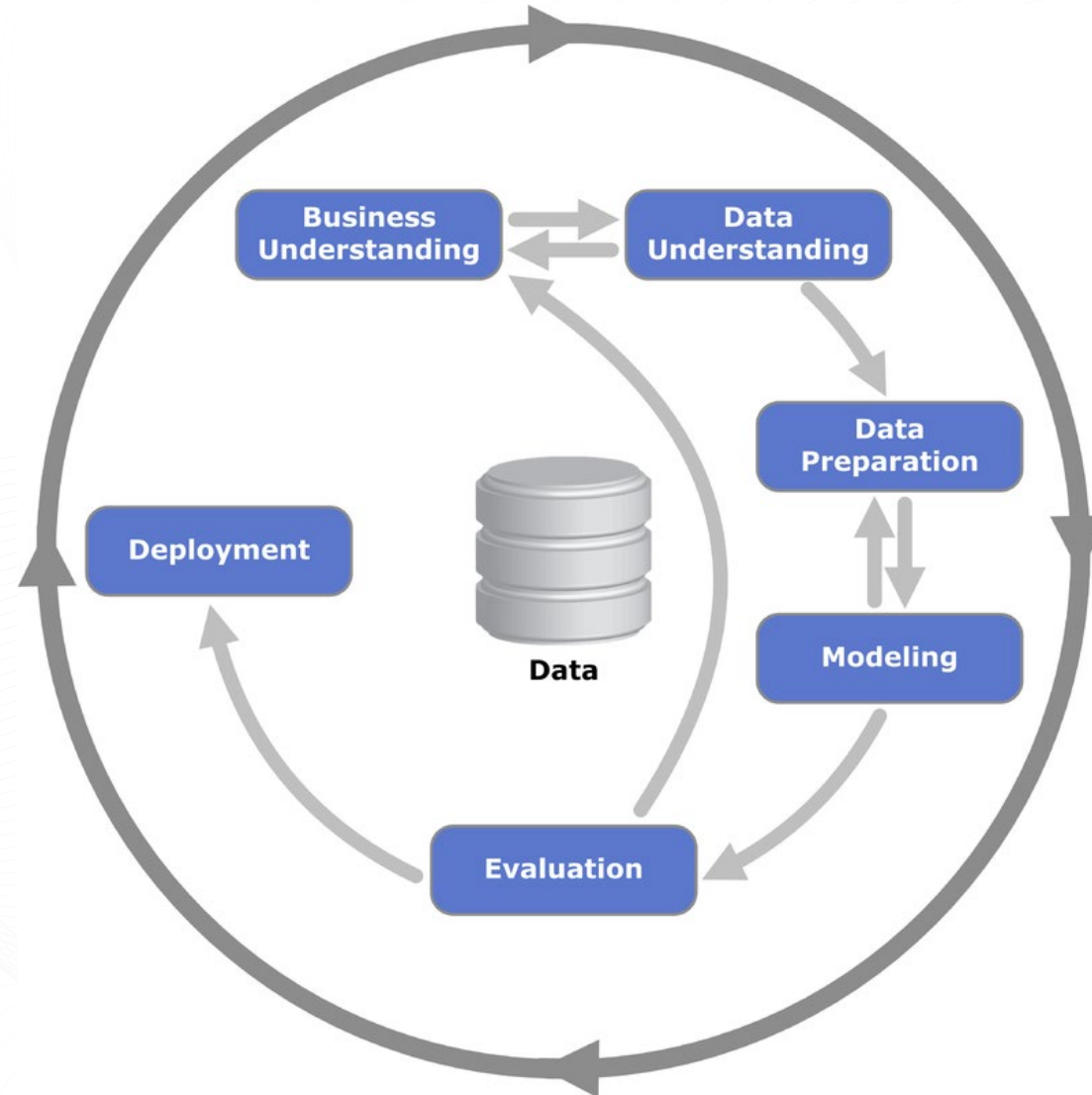
Centro de Servicios y Gestión Empresarial
SENA Regional Antioquia



www.sena.edu.co

Entendimiento de los Datos

Entendimiento de los Datos



Entendimiento de los Datos

La fase de Entendimiento de los Datos es la segunda etapa de la metodología. Esta fase se centra en la **recopilación, exploración y comprensión de los datos disponibles** para garantizar que sean adecuados para cumplir con los objetivos definidos en la fase anterior (Entendimiento del Negocio).

Entendimiento de los Datos



Entendimiento de los Datos

1. Recopilación de Datos Inicial

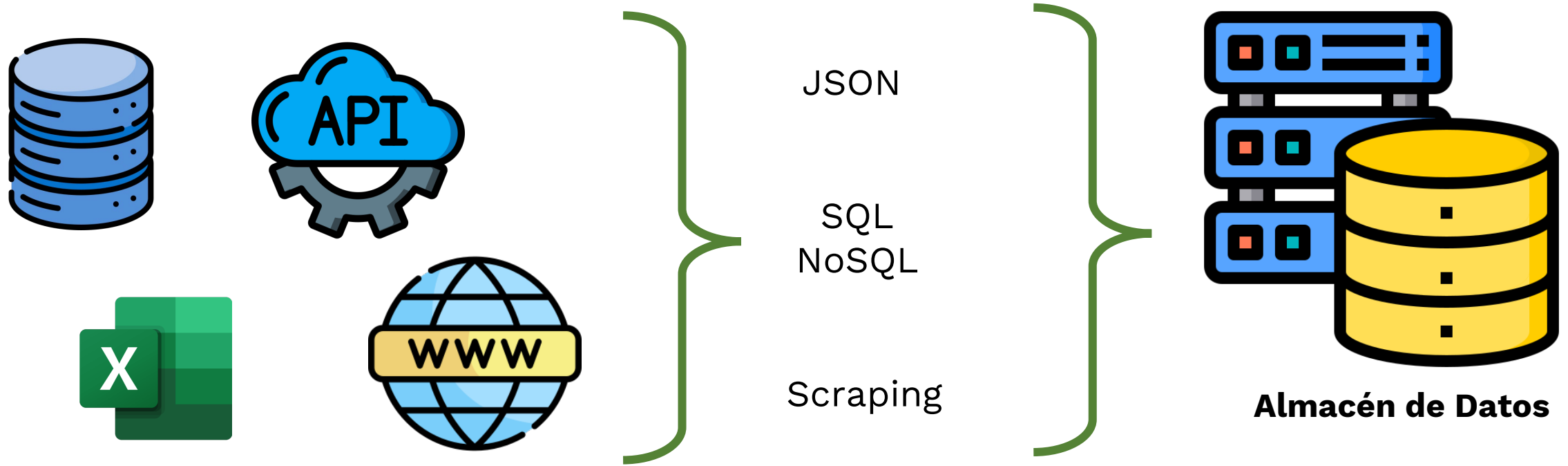
Objetivo: Obtener y organizar los datos que se utilizarán en el proyecto de minería de datos.

Actividades:

- Recolectar los datos de diferentes fuentes (bases de datos, archivos CSV, APIs, etc.).
- Documentar el origen de los datos, su formato y la cantidad de datos disponibles.
- Evaluar la calidad de los datos y determinar si son suficientes y relevantes para los objetivos del negocio.

Entendimiento de los Datos

1. Recopilación de Datos Inicial



Almacén de Datos



Tipo	Estructura de datos	Uso principal	Propósito	Ejemplos
OLTP (Bases de datos operacionales)	Estructurada	Operaciones transaccionales en tiempo real	Garantizar la integridad y consistencia de las transacciones (compras, registros, actualizaciones)	MySQL, PostgreSQL, SQL Server
Data Warehouse (DW)	Estructurada	Análisis histórico y generación de reportes	Integrar y centralizar datos de múltiples fuentes para apoyar decisiones estratégicas	Google BigQuery, Amazon Redshift, Snowflake
Data Lake	Estructurada, semiestructurada y no estructurada	Almacenamiento masivo de datos en bruto	Conservar grandes volúmenes de datos sin transformar para análisis futuros o modelos de IA	Amazon S3, Azure Data Lake, Hadoop HDFS
Data Lakehouse	Híbrida	Analítica y aprendizaje automático	Unificar el almacenamiento transaccional (Data Warehouse) y el analítico (Data Lake) en una sola arquitectura	Databricks Delta Lake, Apache Iceberg, Snowflake

Almacén de Datos



Tipo	Estructura de datos	Uso principal	Propósito	Ejemplos
Data Mart	Estructurada	Análisis departamental o temático	Proporcionar acceso rápido a subconjuntos de datos relevantes para un área específica (marketing, ventas, finanzas)	Snowflake Data Mart, Oracle Data Mart
Base de Datos en Memoria	Estructurada	Procesamiento de datos en tiempo real	Reducir la latencia en consultas y análisis mediante almacenamiento directo en RAM	Redis, SAP HANA, MemSQL
Repositorio de Ciencia de Datos	Variable (estructurada o no)	Gestión de datasets, experimentos y modelos ML	Facilitar la colaboración, trazabilidad y reutilización de datos y modelos en proyectos de ciencia de datos	Kaggle, MLflow, DVC, Vertex AI



Explicación Python

Entendimiento de los Datos

2. Descripción de los Datos

Objetivo: Comprender las propiedades generales de los datos, como la distribución de valores, las medidas estadísticas básicas y las relaciones entre variables.

Actividades:

- Calcular estadísticas descriptivas (medias, medianas, desviaciones estándar, etc.).
- Explorar la distribución de los datos (histogramas, gráficos de barras).
- Identificar el rango de valores y detectar valores extremos o atípicos.



Etiqueta (Variable)	Descripción	Tipo de Dato / Escala	Posibles valores o rango
Steroid	Indica si el paciente recibió tratamiento con esteroides.	Categórica (binaria)	1 = Sí, 2 = No
Antivirals	Indica si el paciente recibió tratamiento con antivirales.	Categórica (binaria)	1 = Sí, 2 = No
Fatigue	Presencia o ausencia de fatiga en el paciente.	Categórica (binaria)	1 = Sí, 2 = No
Malaise	Sensación general de malestar físico o debilidad.	Categórica (binaria)	1 = Sí, 2 = No
Anorexia	Pérdida de apetito.	Categórica (binaria)	1 = Sí, 2 = No
Liver Big	Indica si el hígado está agrandado (hepatomegalia).	Categórica (binaria)	1 = Sí, 2 = No
Liver Firm	Indica si el hígado presenta consistencia firme (signo de daño crónico).	Categórica (binaria)	1 = Sí, 2 = No
Spleen Palpable	Indica si el bazo es palpable (esplenomegalia).	Categórica (binaria)	1 = Sí, 2 = No
Spiders	Presencia de arañas vasculares en la piel (signo de enfermedad hepática).	Categórica (binaria)	1 = Sí, 2 = No



Etiqueta (Variable)	Descripción	Tipo de Dato / Escala	Posibles valores o rango
Ascites	Acumulación de líquido en la cavidad abdominal (signo de cirrosis avanzada).	Categórica (binaria)	1 = Sí, 2 = No
Varices	Presencia de várices esofágicas (dilatación de venas por hipertensión portal).	Categórica (binaria)	1 = Sí, 2 = No
Bilirubin	Nivel de bilirrubina en sangre (mide daño hepático).	Numérica (continua)	Valor ≥ 0.0 (mg/dL)
Alk Phosphate	Nivel de fosfatasa alcalina en sangre (indica obstrucción o lesión hepática).	Numérica (continua)	Valor ≥ 0.0
Sgot	Enzima SGOT o AST (aspartato aminotransferasa); elevada en daño hepático.	Numérica (continua)	Valor ≥ 0.0
Albumin	Nivel de albúmina (proteína sintetizada por el hígado).	Numérica (continua)	0.0 – 6.0 g/dL
Prottime	Tiempo de protrombina (tiempo de coagulación; aumenta con daño hepático).	Numérica (continua)	0 – 100 (segundos o porcentaje normalizado)
Histology	Resultado de la biopsia del hígado (si se confirma daño histológico).	Categórica (binaria)	1 = Sí, 2 = No
Class	Variable objetivo: estado final del paciente.	Categórica (binaria)	1 = Muere, 2 = Vive

Entendimiento de los Datos

2. Descripción de los Datos

- Las variables Fatigue, Malaise, Anorexia, Liver Big, Ascites, etc., son síntomas o signos clínicos observados por el médico.
- Las variables Bilirubin, Albumin, Protine, Sgot, Alk Phosphate son biomarcadores de laboratorio.
- La variable Histology proviene del examen microscópico del tejido hepático.
- La variable Class representa el resultado clínico final y es la variable dependiente (target) en los modelos de predicción.



Explicación Python

Entendimiento de los Datos

3. Exploración de los Datos

Objetivo: Profundizar en los datos para descubrir patrones, relaciones y posibles problemas que puedan afectar el análisis posterior.

Actividades:

- Realizar análisis exploratorio de datos (EDA) utilizando visualizaciones como diagramas de dispersión, mapas de calor, y gráficos de correlación.
- Detectar y comprender relaciones entre variables (por ejemplo, cómo el precio varía con la ubicación o las características de la propiedad).
- Identificar posibles problemas como datos faltantes, valores duplicados, y la presencia de outliers (valores atípicos).



Explicación Python

Entendimiento de los Datos

4. Verificación de la Calidad de los Datos

Objetivo: Asegurar que los datos sean precisos, completos y adecuados para su uso en el análisis.

Actividades:

- Identificar y manejar datos faltantes (imputación de valores, eliminación de registros).
- Detectar y corregir inconsistencias en los datos (por ejemplo, unidades de medida incorrectas o datos mal etiquetados).
- Verificar la integridad de los datos (consistencia entre diferentes fuentes de datos, ausencia de duplicados).



Practicas de Python

4. Párrafo de análisis final



Origen y alcance

- Fuente del conjunto de datos (p. ej. "CSV exportado del CRM", "API interna", "encuesta X").
- Periodo temporal y tamaño del dataset (n registros, m columnas).

Objetivo / variable objetivo

- Qué se intenta predecir o entender (si aplica).

Estructura y tipos de variables

- Variables numéricas / categóricas / fechas / texto; columnas clave.

Balance y representatividad

- Balance de clases en la variable objetivo (si aplica).
- Sesgos o muestreo no aleatorio detectado.

Estadísticas resumen

- Medias, medianas, desviaciones; percentiles relevantes; rangos.
- Distribuciones (asimetría, curtosis) y transformaciones necesarias.

Calidad de los datos

- Conteo de valores faltantes y su distribución por columna.
- Duplicados (si los hay) y si se eliminaron o no.

Outliers y valores atípicos

- Presencia, magnitud y si se excluyen/transforman.

Correlaciones y relaciones clave

- Correlaciones notables entre variables predictoras y con la variable objetivo.

Conclusión / siguientes pasos

- Breve resumen de acciones inmediatas (limpieza, imputación, feature engineering, muestreo).



GRACIAS

Presentó: Alvaro Pérez Niño
Instructor Técnico

Correo: aperezn@misena.edu.co

<http://centrodeserviciosygestionempresarial.blogspot.com/>

Línea de atención al ciudadano: 01 8000 910270

Línea de atención al empresario: 01 8000 910682



@SENAComunica

www.sena.edu.co