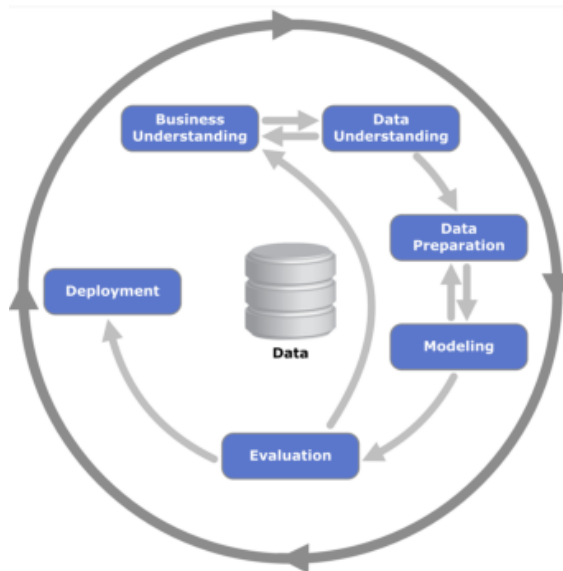


1 Taller02: Modelo CRISP-DM

1.1 Modelo CRISP-DM



El modelo CRISP-DM (*Cross Industry Standard Process for Data Mining*) es un enfoque estándar utilizado en la minería de datos y el análisis de datos. Proporciona una metodología estructurada para guiar el proceso de descubrimiento de conocimiento a partir de datos.

CRISP-DM consta de seis fases principales:

- (1) Comprensión del negocio: Se busca comprender los objetivos y requisitos del negocio, así como determinar qué preguntas deben responderse mediante el análisis de datos.
- (2) Comprensión de los datos: Se recopila y explora el conjunto de datos disponible, evaluando su calidad, familiarizándose con su contenido y buscando patrones o problemas potenciales.
- (3) Preparación de los datos: Se llevan a cabo tareas de limpieza, transformación y selección de los datos para que sean adecuados para el análisis posterior. Esto implica la eliminación de valores atípicos, la normalización de variables y la creación de nuevas características si es necesario.
- (4) Modelado: Se seleccionan y aplican técnicas de modelado apropiadas para descubrir patrones y relaciones en los datos. Esto puede incluir el uso de algoritmos de aprendizaje automático, estadísticas u otras técnicas analíticas.
- (5) Evaluación: Se evalúan los modelos generados en términos de su calidad y relevancia para los objetivos del negocio. Se pueden realizar ajustes o mejoras en los modelos según sea necesario.
- (6) Despliegue: Los resultados del análisis se presentan a las partes interesadas y se implementan en la práctica empresarial. Esto puede implicar la creación de informes, la integración de modelos en sistemas existentes o la toma de decisiones basadas en los resultados obtenidos.

El modelo CRISP-DM proporciona un marco flexible que puede adaptarse a diferentes proyectos y contextos. Permite un enfoque sistemático para abordar problemas complejos de análisis de datos y garantiza un proceso estructurado y repetible para obtener conocimientos valiosos a partir de los datos.

La analítica descriptiva es una rama de la analítica de datos que se centra en la descripción, resumen y visualización de datos para comprender mejor los patrones, tendencias y características presentes en ellos. Su objetivo principal es responder a preguntas como "¿Qué sucedió?" y "¿Qué está ocurriendo actualmente?".

1.2. Analítica Descriptiva

La analítica descriptiva utiliza técnicas y métodos estadísticos para analizar datos históricos y presentarlos de manera comprensible. Estas técnicas pueden incluir medidas de resumen, como promedios, medianas, desviaciones estándar y percentiles, así como gráficos y visualizaciones, como histogramas, gráficos de dispersión y gráficos de líneas.

Al aplicar la analítica descriptiva, se pueden descubrir patrones y tendencias en los datos, identificar valores atípicos, comprender la distribución de los datos y resumir la información de manera concisa. Esto proporciona una base sólida para obtener información valiosa y tomar decisiones fundamentadas.

La analítica descriptiva se utiliza en una amplia gama de industrias y aplicaciones, desde el análisis de datos de ventas y marketing hasta el monitoreo de operaciones y la gestión de recursos humanos. Es una etapa fundamental en el proceso de analítica de datos y sienta las bases para etapas más avanzadas, como la analítica predictiva y la analítica prescriptiva.

Un proyecto de analítica descriptiva típicamente sigue varias fases clave para alcanzar sus objetivos. Estas fases pueden variar según el contexto y la metodología utilizada, pero a continuación se presentan las etapas generales de un proyecto de analítica descriptiva:

- (1) Definición de objetivos: En esta fase inicial, se identifican y establecen los objetivos del proyecto de analítica descriptiva. Esto implica comprender las necesidades y preguntas del negocio que se buscan responder a través del análisis de datos. Es importante definir claramente los objetivos para orientar todo el proceso.
- (2) Recopilación y preparación de datos: En esta fase, se recopilan los datos necesarios para el análisis. Esto puede implicar extraer datos de diversas fuentes, como bases de datos, archivos CSV o sistemas en tiempo real. También se realiza la limpieza y preparación de los datos, incluyendo la eliminación de valores faltantes o erróneos, la estandarización de formatos y la transformación de datos si es necesario.
- (3) Análisis exploratorio de datos: En esta etapa, se lleva a cabo un análisis exploratorio de los datos para comprender su estructura, características y relaciones. Se utilizan técnicas estadísticas y visualizaciones para identificar patrones, tendencias y valores atípicos en los datos. Esto ayuda a obtener una visión inicial de los datos y a generar hipótesis que se abordarán en las siguientes etapas.
- (4) Análisis descriptivo: En esta fase, se aplican técnicas de analítica descriptiva para describir y resumir los datos de manera sistemática. Se utilizan medidas estadísticas, gráficos y visualizaciones para resumir la distribución de los datos, calcular medidas de tendencia central y dispersión, y proporcionar una comprensión más profunda de los datos en relación con los objetivos establecidos.
- (5) Interpretación de resultados: En esta etapa, se interpretan los resultados obtenidos en el análisis descriptivo y se vinculan con los objetivos del proyecto. Se buscan respuestas a las preguntas iniciales y se extraen conclusiones significativas a partir de los hallazgos. Se pueden generar informes y visualizaciones para comunicar los resultados a las partes interesadas.
- (6) Presentación y comunicación: En esta fase final, se presenta y comunica el análisis descriptivo a las partes interesadas. Esto implica la preparación de informes, presentaciones y visualizaciones que resuman los hallazgos clave de manera clara y comprensible. Se facilita la discusión y se brinda oportunidad para aclarar dudas y obtener retroalimentación.

Estas fases no son necesariamente lineales y pueden requerir iteraciones y ajustes a lo largo del proyecto. Además, la analítica descriptiva sienta las bases para etapas más avanzadas, como la analítica predictiva y la analítica prescriptiva, que buscan utilizar los conocimientos obtenidos para predecir y prescribir acciones futuras.

Bibliografía: Shearer, C. (2000). The CRISP-DM model: the new blueprint for data mining. *Journal of data warehousing*, 5(4), 13-22

1.3. Definición de un proyecto de Analítica Descriptiva:

(1) Problema a resolver:

Realizar un análisis del tiempo que tardaron en graduarse los estudiantes politécnicos; primera promoción del SENESCYT en el año 2012 .

(2) Objetivos:

Este proyecto tiene como objetivo medir el tiempo promedio que los estudiantes de la EPN que ingresaron en el semestre 2012-A (pertenecientes a la primera promoción del SENEYC)se tardaron en culminar su carrera y encontrar si existe una relación entre la carrera y el tiempo de graduación. Se debe tener en cuenta que como se trabaja con la lista de planes de tesis, los estudiantes considerados en el análisis son solo los que se gradúan mediante el proceso de titulación por proyecto.

(3) Datos:

a. Fuentes de datos:

Registro de estudiantes que ingresaron en 2012 a la Escuela Politécnica Nacional.
Registro de todos los proyectos integradores SAEW.

b. Tipos de datos:

Registro de proyectos integradores es un archivo .html.
Registro de estudiantes ingresados en 2012 .pdf

c. Procesamiento:

Leer el archivo .pdf y transformar a un Dataset.
Limpiar los datos, eliminar registros sin datos.
Crear un Dataset con los campos relevantes para el análisis.

Leer el archivo .html y transformar a un Dataset.
Crear un Dataset con los campos relevantes para el análisis.
Unir los Dataset y obtener los registros a ser analizados.
Obtener el promedio y definir las carreras con mayor tiempo para graduarse.