Notas sobre Análisis de Datos

Cristo Daniel Alvarado

30 de enero de 2025

Cristo Daniel Alvarado ESFM

Índice general

Capítulo 1

Conceptos Fundamnetales e Introuctorios

El objetivo de este primer capítulo será el de dar una versión descrpitiva del análisis de datos, su importancia y relevancia así como sus posibles aplicaciones.

§1.1 ¿Para qué sirve?

Básicamente el análisis de datos se basa en filtrar, ordenar y presentar información (datos) obtenida de un negocio / empresa / lugar de trabajo con el fin de obtener una predicción (como lo hacen las matemáticas) de lo que nos puede interesar en el futuro.

Nos puede servir para lo siguiente:

- Seguimiento de inventario.
- Identificación de hábitos de compra.
- Detección de tendencias y patrones de usuarios.
- Recomendaciones de compras.
- Definición de optimizaciones de precios.
- Identificación y detección de fraude.

El seguimiento de patrones puede hacerse diariamente, semanalmente, mensualmente o anualmente (dependiendo de lo que se desee analizar).

Quien sepa analizar los datos conocerá el futuro

El reto subyacente al que se enfrentan las empresas actuales es comprender y usar sus datos de forma que afecten a su negocio y, en última instancia, a sus beneficios. Debe ser capaz de examinar los datos y facilitar decisiones empresariales de confianza. Después, necesitará la capacidad de examinar las métricas y comprender claramente su significado.

§1.2 Análisis de Datos

Definición 1.2.1
El **análisis de datos** es el proceso de identificar, limpiar, transformar y modelar los datos para detectar información significativa y útil.

Los datos luego se convierten en una historia a través de **informes** para el análisis con el fin de admitir el proceso crítico de toma de desiciones.

Aunque el proceso de análisis de datos se centra en las tareas de limpieza, modelado y visualización de datos, el concepto de análisis de datos y su importancia para las empresas no se debe subestimar. Para analizar los datos, los componentes principales del análisis se dividen en las siguientes categorías:

- *Descriptivo*: Ayuda a responder preguntas sobre lo que ha sucedido en función de datos históricos.

  Mediante el desarrollo de indicadores clave de rendimiento (KPI), estas estrategias pueden facilitar el seguimiento del éxito o el fracaso de los objetivos clave. En muchos sectores se usan métricas como la rentabilidad de la inversión (ROI), y las métricas especializadas se desarrollan para realizar un seguimiento del rendimiento en sectores específicos.

  Un ejemplo es la generación de informes para proporcionar una visión de los datos financieros y ventas de una organización.

- *Diagnóstico*: Ayuda a responder preguntas sobre por qué se ha producido un evento. Complementa al análisis descriptivo, usan sus resultados para dar una causa del evento.

  Continúa con la investigación de los indicadores de rendimiento. El proceso puede realizarse en tres pasos:

  1. Identificación de las anomalías en los datos (cambios inesperados).
  2. Recopilación de datos relacionados con las anomalías.
  3. Uso de técnicas estadísticas para detectar relaciones y tendencias para explicar anomalías.

- *Predictivo*: Ayuda a respoder las preguntas de lo que ocurrirá en el futuro. Usan datos históricos para identificar tendencias y determinar probabilidad de repetición. Es la herramienta más valiosa.

  Usan técnicas de estadística y aprendizaje automático, como redes neuronales, árboles de desición y regresión.

- *Prescriptivo*: Ayuda a responder las preguntas sobre qué acciones deben llevarse a cabo para lograr un objetivo. Permite tomar desiciones basadas en datos.

  En caso de incertidumbre, permite que una empresa tome desiciones con fundamento. Depende del uso del aprendizaje automático como estrategia para buscar patrones en modelos semánticos.

- *Cognitivo*: Intenta obtener inferencias a partir de datos y patrones existentes, derivar conclusiones en función de bases de conocimiento ya existentes y, después, devolver estos resultados a la base de conocimiento para futuras inferencias, un bucle de comentarios de autoaprendizaje.

  El análisis cognitivo ayuda a saber lo que podría ocurrir si cambiaran las circunstancias y a determinar cómo se podrían controlar estas situaciones.

  Las inferencias no son consultas estructuradas basadas en una base de datos de reglas, sino supuestos no estructurados que se recopilan de varios orígenes y se expresan con distintos grados de confianza. El análisis cognitivo eficaz depende de algoritmos de aprendizaje automático y

usa varios conceptos del procesamiento de lenguaje natural para entender orígenes de datos desaprovechados anteriormente, como los registros de conversaciones de centros de llamadas y revisiones de productos.

Una faceta subyacente del análisis de datos es que las empresas necesitan ser capaces de confiar en sus datos. Como práctica, el proceso de análisis de datos toma datos de fuentes de confianza y los convierte en algo que es consumible, significativo y fácil de comprender para ayudar con el proceso de toma de decisiones. El análisis de datos permite a las empresas comprender sus datos de manera integral a través de procesos y decisiones controladas por datos, para de ese modo confiar en sus decisiones.

A medida que la cantidad de datos crece, también lo hace la necesidad de analistas de datos. Un analista de datos sabe cómo organizar la información y sintetizarla en algo relevante y comprensible. También sabe cómo recopilar los datos correctos y qué hacer con ellos, es decir, dar sentido a los datos a pesar de la sobrecarga de datos.

§1.3 Roles en los Datos

La narración de historias mediante datos es un recorrido que normalmente no comienza con usted. Los datos tienen que provenir de alguna parte. Llevar esos datos a un lugar que pueda usar requiere esfuerzo, y es probable que se escape a su ámbito, sobre todo cuando se trata de la empresa.

Hay un espectro en la detección y comprensión de los datos:

- *Analista de negocios*: Está más cerca de la empresa y análiza los datos que provienen de la visualización. Puede que sea similar al análista de datos.
- *Analista de datos*: Permite a las empresas maximizar el valor de sus recursos de datos a través de herramientas de visualización y creación de informes como Microsoft Power BI. El analista de datos es responsable de la generación de perfiles, la limpieza y la transformación de los datos.

  Trabaja con las partes interesadas para identificar requisitos necesarios y creación de informes necesarios para después dar conclusiones sobre estos datos.

  Se le encomienda la implementación y configuración de los procedimientos de seguridad adecuados, junto con los requisitos de las partes interesadas.

  Generalmente trabaja con ingenieros de datos para determinar y localizar los orígenes de datos adecuados que satisfagan los requisitos de las partes interesadas.

  También con los administradores de bases de datos para tener acceso adecuado a los orígenes de datos necesarios. También se enfoca en mejorar los procesos ya existentes.
- *Ingeniero de datos*: aprovisionan y configuran las tecnologías de plataforma de datos locales y en la nube. Administran y protegen el flujo de datos estructurados y no estructurados procedentes de múltiples orígenes. Se aseguran de que los servicios de datos se integren de forma segura y sin problemas en las plataformas de datos.

  responsabilidades principales se incluye el uso de servicios datos locales y en la nube, y herramientas para la ingesta, la salida y la transformación de datos procedentes de múltiples orígenes.
- *Científico de datos*: Realizan un análisis avanzado para extraer valor de los datos. Su trabajo puede variar del análisis descriptivo al análisis predictivo.

  Un científico de datos examina los datos para determinar las preguntas que necesitan respuestas y, a menudo, diseñará una hipótesis o un experimento, y luego recurrirá al analista de datos para que le ayude con la visualización de datos y la creación de informes.

- *Administrador de base de datos*: Un administrador de bases de datos implementa y administra los aspectos operativos de las soluciones de plataforma de datos híbridas y nativas de la nube que se basan en servicios de datos de Microsoft Azure y Microsoft SQL Server.

Definición 1.3.1
El **análisis descriptivo** evalúa los datos a través de un proceso conocido como análisis de datos exploratorio (EDA). El **análisis predictivo** se usa en el aprendizaje automático para aplicar técnicas de modelado que pueden detectar anomalías o patrones. Estos análisis son una parte importante de los modelos de previsión.

Hay cinco áreas clave en las que participa un analista de datos:

§1.3.1 Preparación

Antes de que se pueda crear un informe, es necesario preparar los datos. La preparación de datos es el proceso de generación de perfiles y de limpieza y transformación de los datos para prepararlos para el modelado y la visualización.

Consiste en tomar datos sin procesar y convertirlos en información de confianza y comprensible. Implica, entre otras cosas, garantizar la integridad de los datos, corregir datos incorrectos o inexactos, identificar los datos que falten, convertir datos de una estructura a otra o de un tipo a otro, o incluso una tarea tan sencilla como hacer que los datos sean más legibles.

También implica comprender cómo va a obtener los datos y a conectarse a ellos, y conocer las implicaciones de rendimiento de las decisiones. Al conectarse a los datos, necesita tomar decisiones para asegurarse de que los modelos y los informes cumplen y llevan a cabo las expectativas y los requisitos confirmados.

Las garantías de privacidad y seguridad también son importantes. Estas pueden incluir la anonimización de los datos para evitar que se compartan en exceso o impedir que los usuarios vean información de identificación personal cuando no es necesario. Ayudar a garantizar que la privacidad y la seguridad también puede implicar la eliminación completa de los datos si no se ajustan a la historia que está intentando narrar.

§1.3.2 Modelado

Cuando los datos están en un estado correcto, están listos para modelarse. El modelado de datos es el proceso de determinar cómo se relacionan las tablas entre sí. Este proceso se realiza mediante la definición y creación de relaciones entre las tablas. A partir de ahí, puede mejorar el modelo si define métricas y agrega cálculos personalizados para enriquecer los datos.

Un modelo semántico eficaz hace que los informes sean más precisos, permite que los datos se exploren de manera más rápida y eficaz, reduce la duración del proceso de creación de informes y simplifica el mantenimiento futuro del informe.

El proceso de preparación y modelado de datos es iterativo. La preparación de datos es la primera tarea en el análisis de datos. Comprender y preparar los datos antes de modelarlos hará que el paso de modelado sea mucho más fácil.

§1.3.3 Visualización

En la tarea de visualización es donde se hace que los datos cobren vida. El objetivo final de la tarea de visualización es solucionar los problemas de la empresa. Un informe bien diseñado debe contar una historia atractiva sobre estos datos. Puede proporcionar un informe eficiente que guíe al lector a través del contenido de forma rápida y eficaz, lo que le permitirá seguir una narrativa en los datos.

Como analista de datos, debe dedicar tiempo a comprender por completo el problema que la empresa intenta resolver.

Determine si todos los puntos de datos son necesarios, ya que un exceso de datos puede dificultar la detección de los puntos clave. Una historia de datos pequeña y concisa puede ayudar a encontrar la información rápidamente.

Los informes se deben diseñar pensando en la accesibilidad desde el principio, de modo que no se necesite ninguna modificación especial en el futuro.

§1.3.4 Análisis

El objetivo es entender e interpretar la información que se muestra en el informe. En su rol como analista de datos, debe comprender las funciones analíticas de Power BI y usarlas para buscar conclusiones, identificar patrones y tendencias, predecir resultados y, después, comunicar esas conclusiones de una forma comprensible para todos.

Con el análisis avanzado, las organizaciones pueden profundizar en los datos para predecir patrones y tendencias futuros, identificar actividades y comportamientos, y permitir a las empresas formular las preguntas adecuadas sobre sus datos.

§1.3.5 Administración

Power BI consta de muchos componentes, como informes, paneles, áreas de trabajo, modelos semánticos y mucho más. Como analista de datos, es responsable de administrar estos recursos de Power BI, de supervisar el uso compartido y la distribución de elementos como informes y paneles, y de garantizar la seguridad de los recursos de Power BI.

La administración del contenido ayuda a fomentar la colaboración entre equipos y usuarios. El uso compartido y la detección de contenido es importante para que las personas adecuadas obtengan las respuestas que necesitan. También es importante asegurarse de que los elementos sean seguros. Querrá asegurarse de que las personas adecuadas tienen acceso y de que los datos no se pierden más allá de las partes interesadas correctas.

§1.4 Tools to analyze data

There are a variety of tools you can use to analyze data, depending on the type of analysis you're performing, the complexity of your data, and your familiarity with different software. Here are some popular tools across different categories:

§1.4.1 Data Cleaning and Manipulation

- **Excel**: A basic but powerful tool for data manipulation, analysis, and visualization.
- **Google Sheets**: Similar to Excel, but cloud-based and great for collaboration.
- **OpenRefine**: An open-source tool for cleaning messy data, especially for larger datasets.
- **Pandas (Python library)**: A powerful library for data manipulation, especially for tabular data.

§1.4.2 Statistical Analysis

- **R**: An open-source programming language and environment for statistical computing and graphics. It's widely used in academia and data science.
- **SPSS**: A statistical software package that's often used in social sciences for data analysis.
- **SAS**: A software suite used for advanced analytics, business intelligence, and data management.
- **Stata**: A software package used for statistics and data analysis.

§1.4.3 Data Visualization

- **Tableau**: A powerful tool for creating interactive data visualizations. It's user-friendly and widely used in business intelligence.
- **Power BI**: A Microsoft tool for data visualization and business analytics, great for connecting to multiple data sources.
- **Matplotlib/Seaborn (Python libraries)**: For creating static, animated, and interactive visualizations in Python.
- **ggplot2 (R package)**: A widely used library in R for creating complex visualizations.

§1.4.4 Machine Learning and Predictive Analytics

- **Scikit-learn (Python library)**: A machine learning library for Python that offers simple and efficient tools for data mining and data analysis.
- **TensorFlow/PyTorch**: Frameworks for machine learning and deep learning, often used for more complex analyses.
- **H2O.ai**: An open-source platform for machine learning and AI.
- **RapidMiner**: A visual data science workflow tool for machine learning, predictive analytics, and data mining.

§1.4.5 Big Data and Distributed Computing

- **Apache Hadoop**: A framework for processing large datasets in a distributed computing environment.
- **Apache Spark**: A unified analytics engine for big data processing, with built-in modules for streaming, machine learning, and graph processing.
- **Google BigQuery**: A fully-managed data warehouse for large-scale data analysis.

- **Amazon Redshift**: A data warehouse solution that allows for fast querying and analysis of large datasets.

§1.4.6 Database Management and Analysis

- **SQL (Structured Query Language)**: A standard language for querying and managing databases.
- **MySQL/PostgreSQL**: Open-source relational database management systems.
- **MongoDB**: A NoSQL database used for large volumes of unstructured data.
- **SQLite**: A lightweight relational database management system.

§1.4.7 Collaboration and Reporting

- **Jupyter Notebooks**: A popular tool for data analysis and sharing code and visualizations interactively in a notebook format, using Python.
- **Google Colab**: A cloud-based Jupyter notebook environment that allows you to run Python code with free access to GPUs.
- **Looker**: A data exploration tool for business intelligence, which helps to visualize and analyze data.

modelos semánticos

aprenizaje automático

estadística

probabilidad

indicadores clave de rendimiento