Lista de herramientas y conceptos

Análisis de Datos

Nicolás García Peñaloza
nicolasgp0109@gmail.com

Resumen

Acá encontraran un resumen a grandes rasgos de una pequña parte de todo el mundo en el análisis de datos, desde el componente estadístico, las áreas en las que se dividen y las herramientas.

1. Conceptos estadíticos

Estadítica: Los métodos estadísticos se utilizan para resolver problemas reales en diversas áreas del conocimiento, como economía, medicina, ingeniería, educación, psicología, biología, sociología, negocios, entre otras. Esta va desde recolectar, organizar, analizar e interpretar datos con el objetivo de tomar decisiones informadas y respaldadas por evidencia. esto incluye: i) Diseño de experimentos y encuestas, ii) Análisis exploratorio de datos, iii) Inferencia estadística (estimaciones, pruebas de hipótesis, intervalos de confianza), iv) Modelado estadístico (regresiones, series de tiempo, modelos multivariantes) y v)Visualización de datos.
Enfoque Bayesiano: Muy usada en inteligencia artificial, medicina personalizada, toma de decisiones con incertidumbre. Empiezas con una creencia previa (prior) sobre un parámetro. Actualizas esa creencia con los datos observados, usando el Teorema de Bayes.
Enfoque Frecuentista: Muy extendida en ciencias naturales, ingeniería, economía clásica. Trabajas con los datos observados sin considerar creencias previas. Las inferencias vienen de repetir el experimento muchas veces hipotéticamente.
Regresión: La regresión es una técnica estadística que se utiliza para analizar la relación entre una variable dependiente (respuesta) y una o más variables independientes (predictoras), con el objetivo de entender, predecir o explicar el comportamiento de la variable dependiente.

Parametro (β μ σ^2): Un parámetro es un valor descriptivo y fijo que resume alguna característica de una población completa, como la media, la varianza o una proporción.

Estimación ($\widehat{\beta}$): Una estimación es un valor calculado a partir de una muestra que sirve como aproximación de un parámetro poblacional.

Coeficiente: En un modelo estadístico (como una regresión), un coeficiente es el **número** que mide la relación entre una variable independiente y la variable dependiente.

tipos de regresión:

- Univariante: Sólo una variable de respuesta cuantitativa.
- Multivariante: Dos o más variables cuantitativas de respuesta.
- Simple: Una sola variable predictiva
- Multiple: Dos o más variables predictoras
- Lineal: Todos los parámetros entran linealmente en la ecuación, posiblemente tras la transformación de los datos.
- No-lineal: La relación entre la respuesta y algunos de los predictores es no lineal o algunos de los parámetros aparecen de forma no lineal, pero no es posible realizar ninguna transformación para que los parámetros aparezcan de forma lineal.
- Análisis de varianza: Todos los predictores son variables cualitativas.
- Análisis de covarianza: Algunos predictores son variables cuantitativas y otros son variables cualitativas.
- Logística: La variable de respuesta es cualitativa.
- Regresión penalizada (ridge, lasso): Es una variante de la regresión lineal en la que, además de minimizar el error entre los valores observados y predichos, también se penaliza el tamaño de los coeficientes. Esto ayuda a: Evitar el sobreajuste (overfitting), Seleccionar variables relevantes, Mejorar la estabilidad del modelo. Regresión Ridge: cuando tienes muchas variables y colinealidad, pero no te interesa eliminar ninguna. Regresión Lasso: cuando quieres identificar las variables más importantes (elimina las irrelevantes). Elastic Net: cuando hay colinealidad y sospechas que solo algunas variables son útiles.

Colinealidad (o multicolinealidad): Es cuando dos o más variables independientes están altamente correlacionadas entre sí. Esto causa que el modelo tenga dificultad para estimar de forma precisa los coeficientes de esas variables.

Heterocedasticidad: Ocurre cuando la varianza de los errores del modelo no es constante a lo largo de los valores de la variable independiente.

Autocorrelación: Es cuando los errores del modelo están correlacionados entre sí, es decir, no son independientes. Es común en datos de series temporales.

Estandarizar: significa transformar una variable para que tenga media 0 y desviación estándar 1. Sirve para comparar variables que tienen unidades o escalas diferentes (por ejemplo: ingresos en pesos y edad en años). Muy útil en análisis multivariado.

$Z = \frac{x-\mu}{\sigma}$

X: valor original

μ: media de la variable

σ: desviación estándar de la variable
Normalizar: consiste en transformar una variable para que sus valores estén entre 0 y 1 (o entre -1 y 1, según el método). Sirve para escalar datos sin cambiar la forma de la distribución. Ideal cuando se requiere que todos los valores estén en un mismo rango (por ejemplo, en redes neuronales o algoritmos de distancia como k-NN).

$X = \frac{x-x_{min}}{x_{max}-x_{min}}$

Datos de sección cruzada (Cross-Sectional Data): Son datos recolectados en un solo punto en el tiempo (Pensemos en una encuesta donde i es la observación) para varias unidades (personas, hogares, empresas, regiones, etc.).

Datos de series de tiempo (Time Series Data): Son observaciones de una sola unidad (país, empresa, variable económica) a lo largo del tiempo.

Datos Panel (Panel Data o Longitudinal Data): Son una combinación de sección cruzada y serie temporal: observas las mismas unidades (personas, empresas, países) en varios momentos del tiempo.

2. Campos de acción en el análisis de datos

Data Analyst (Analista de Datos): Limpia, transforma y analiza datos para responder preguntas de negocio. El Data Analyst extrae, limpia, analiza e interpreta datos para descubrir patrones, tendencias y relaciones relevantes. Su labor ayuda a responder preguntas específicas del negocio y comunicar hallazgos de forma clara mediante visualizaciones. Usa herramientas como Excel, SQL, Python o R, y colabora estrechamente con áreas funcionales. Este perfil actúa como el puente entre los datos y las áreas funcionales: plantea hipótesis, realiza análisis comparativos o segmentaciones, y comunica hallazgos con claridad, muchas veces usando gráficos o storytelling con datos.

BI Analyst (Analista de Inteligencia de Negocios): Crea dashboards e informes para apoyar decisiones estratégicas. El Analista BI transforma datos brutos en información valiosa para la toma de decisiones empresariales. Su enfoque está en el presente y pasado del negocio: diseña dashboards, reportes e indicadores clave de rendimiento (KPIs), y utiliza herramientas como Power BI, Tableau o Qlik, junto con SQL, para monitorear y optimizar procesos. Su enfoque está centrado en la visualización y el monitoreo del desempeño organizacional, integrando datos de diferentes fuentes y automatizando informes periódicos. A menudo trabaja con herramientas de visual analytics y colabora estrechamente con áreas como gerencia, ventas, operaciones y finanzas.

Data Architect: Diseña la arquitectura de almacenamiento y flujo de datos a nivel organizacional. El Data Architect diseña la arquitectura global del ecosistema de datos de una organización. Define cómo se recolectan, almacenan, integran, acceden y protegen los datos a lo largo de su ciclo de vida. Este rol tiene una perspectiva estratégica y tecnológica, asegurando que la infraestructura de datos sea escalable, segura, interoperable y alineada con las metas del negocio. Actúa como una figura de alto nivel que traduce los requerimientos organizacionales y regulatorios en decisiones técnicas, colaborando con data engineers, DevOps, analistas y especialistas en gobernanza de datos.

Data Engineer: Construye pipelines para la recolección, limpieza y almacenamiento de grandes volúmenes de datos.El Data Engineer es el arquitecto de los cimientos sobre los que se construyen los análisis de datos. Su trabajo consiste en diseñar, desarrollar y mantener infraestructuras robustas que permitan recolectar, transformar, almacenar y mover grandes volúmenes de datos desde múltiples fuentes hacia entornos analíticos o de modelado. Se encarga de construir pipelines de datos (ETL/ELT), integrando datos estructurados y no estructurados desde APIs, bases de datos transaccionales, sensores, archivos planos o fuentes en la nube. Su objetivo es garantizar que los datos estén disponibles, limpios, organizados y accesibles para los analistas, científicos de datos y aplicaciones empresariales.

Data Scientist: Desarrolla modelos predictivos y de clasificación, experimenta con machine learning. El Científico de Datos va más allá del análisis descriptivo: construye modelos predictivos y algoritmos de aprendizaje automático para responder preguntas complejas del negocio. Trabaja con grandes volúmenes de datos (big data), domina lenguajes como Python o R, y tiene una sólida base en estadística, programación y pensamiento crítico.

3. Conceptos en el análisis de datos

IDE: Es un Entorno de Desarrollo Integrado que ayuda a programar más fácilmente, ya que combina varias herramientas en un solo lugar: editor de código, consola, depurador, gestor de paquetes, y más. Todo estp es en aras de una mejor experiencia.
Terminal: Es una interfaz de texto donde puedes interactuar con tu sistema operativo escribiendo comandos.
Consola: Es el espacio donde el programa ejecuta directamente los comandos.
API (Application Programming Interface): Es un puente que permite que dos aplicaciones, sistemas o servicios se comuniquen entre sí, compartiendo datos o funcionalidades de manera estructurada y segura. Imagina una API como el menú de un restaurante: tú ves qué platos puedes pedir (los .endpoints") sin saber cómo los cocinan. Solo necesitas saber qué pedir, cómo pedirlo y qué esperar como respuesta. Así funcionan las APIs entre programas.

Storytelling: Es la técnica para transmitir mensajes con datos.

ETL (Extract – Transform – Load): Es el proceso fundamental en la ingeniería y análisis de datos que permite mover datos desde diferentes fuentes (PostgreSQL, MySQL , Excel) hacia un sistema centralizado (como un data warehouse).

- Transform (Transformación): Es la etapa más compleja. i) Se limpian (eliminan duplicados, corrigen errores). ii) Se normalizan (fechas, formatos, unidades).iii) Se combinan (joins, uniones).iv) Se agregan o calculan (totales, promedios, KPIs). v) E incluso se enriquecen (por ejemplo, cruzando con datos geográficos o demográficos).
- Los datos transformados se cargan en el sistema de destino: Data warehouse i) (ej. Google BigQuery, Amazon Redshift). ii) Base de datos analítica. iii) Dashboard o herramienta BI.

Data Lake: es un gran repositorio donde puedes almacenar todo tipo de datos (estructurados, semi-estructurados y no estructurados) en su forma original, es decir, sin procesar o sin necesidad de transformarlos previamente.

Data Warehouse: es un sistema diseñado para almacenar datos ya limpios, organizados y estructurados, optimizados para consulta y análisis empresarial.

4. Herramientas en el análisis

Git: es un sistema de control de versiones distribuido. Sirve para guardar, rastrear y administrar los cambios que haces en tu código (o cualquier archivo de texto) a lo largo del tiempo.

GitHub: es una plataforma en la nube que usa Git por debajo. Es como una red social para desarrolladores que permite: i) Guardar tus repositorios (proyectos) en línea, ii)Colaborar con otras personas en tiempo real, iii)Revisar, comentar y fusionar cambios, iv)Hacer proyectos públicos o privados.

Docker: Docker es una plataforma que te permite empaquetar, distribuir y ejecutar aplicaciones en contenedores. Un contenedor Docker es como una "caja" que guarda todo lo que tu proyecto necesita para funcionar: código, librerías, dependencias, sistema operativo, etc.

Apache Airflow: Es una plataforma de código abierto desarrollada por Airbnb para crear, programar y monitorear flujos de trabajo complejos de datos. Es uno de los orquestadores de tareas más populares en ingeniería de datos.

Kedro: es un framework de desarrollo estructurado para proyectos de ciencia de datos y machine learning, creado por QuantumBlack (subsidiaria de McKinsey). Su objetivo es ayudarte a organizar tu código, tus datos y tus pipelines de forma limpia, reproducible y modular.

Apache Sparkes un motor de procesamiento de datos distribuido y de código abierto diseñado para trabajar con grandes volúmenes de datos, de forma rápida, escalable y en memoria (in-memory). Fue desarrollado originalmente por la Universidad de California en Berkeley, y ahora es parte de la Apache Software Foundation.

5. Herramientas empresariales para la gestión y coordinación

Blog para ciencia de datos:

- towards datas cience
- Medium
- Walter Sosa

Slack: es una plataforma de mensajería y colaboración en línea que permite a los equipos trabajar de forma más organizada y unificada. Reúne conversaciones, aplicaciones y clientes en un solo lugar.
ClickUp: es una plataforma de gestión de proyectos "todo en uno"que permite a los equipos organizar, colaborar y gestionar tareas, flujos de trabajo y proyectos de forma eficiente, ofreciendo una amplia gama de funcionalidades y personalización.